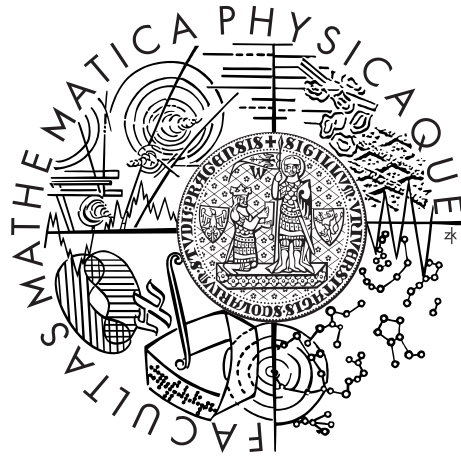


Univerzita Karlova v Praze  
Matematicko-fyzikální fakulta

## DISERTAČNÍ PRÁCE



Vítězslav Línek

## Geometrie lineárního modelu

Katedra didaktiky matematiky

Vedoucí disertační práce: RNDr. Magdalena Hykšová, Ph.D.

Studijní program: matematika

Studijní obor: 4M8 Obecné otázky matematiky  
a informatiky

Praha 2016

Své školitelce, RNDr. Magdaleně Hykšové, Ph. D, jsem nesmírně zavázán za pomoc při vedení dizertační práce, množství cenných připomínek a vstřícnost při konzultacích.

RNDr. Mariánu Rybářovi patří mé díky za kritické přečtení několika kapitol a konzultace ohledně srozumitelnosti práce.

Nakonec bych rád poděkoval Mgr. Zdeňku Halasovi, DiS., Ph.D., za čas strávený diskusemi o tématu mé práce; jeho zájem pro mne mnoho znamenal.

Prohlašuji, že jsem tuto disertační práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V Praze dne 13. června 2016

Vítězslav Línek

Název práce: Geometrie lineárního modelu

Autor: Vítězslav Línek

Katedra: Katedra didaktiky matematiky

Vedoucí disertační práce: RNDr. Magdalena Hykšová, Ph.D., katedra didaktiky matematiky

Abstrakt: Cílem práce je ukázat možnosti využití mnohorozměrné geometrie při výkladu lineárního modelu. Východiskem je tzv. „free-coordinate approach“, tj. pojetí náhodného vektoru jako geometrického objektu, jehož vlastnosti nezávisí na zvolené soustavě souřadnic. S pomocí elementárních geometrických představ a základních statistických pojmů jsou pak odvozeny nejdůležitější vlastnosti lineárního modelu a řada známých aplikací, především statistických testů. Součástí práce je i krátké historické pojednání o počátcích matematické statistiky a rozbor vybraných prací R. A. Fishera, ze kterých je patrné, že geometrický přístup k lineárnímu modelu má i své historické opodstatnění. Text je určen především zájemcům o alternativní vzhled do této problematiky, ale také studentům matematických oborů, kterým matematická statistika působí obtíže; z toho důvodu je doplněn značným množstvím příkladů.

Klíčová slova: lineární model, vícerozměrná geometrie, didaktika statistiky, historie statistiky

Title: Geometry of Linear Model

Author: Vítězslav Línek

Department: Department of Mathematics Education

Supervisor: RNDr. Magdalena Hykšová, Ph.D., Department of Mathematics Education

Abstract: The advantage of the geometric approach to linear model and its applications is known to many authors. In spite of that, it still remains to be rather unpopular in teaching statistics around the world and is almost missing in the Czech Republic. In this work, we use geometry of multidimensional vector spaces to derive some well-known properties of the linear model and to explain some of the most familiar statistical methods to show usefulness of this approach, also known as „free-coordinate“. Besides, historical background including selected results of R. A. Fisher is briefly discussed; it follows that the geometry approach to linear model is justifiable from the historical point of view, too.

Keywords: linear model, multidimensional geometry, statistics education, history of statistics

# Obsah

Předmluva . . . . .	4
Přehled použitého značení . . . . .	6
<b>1 Praktická část</b>	<b>8</b>
1.1 Náhodný vektor . . . . .	8
1.1.1 Střední hodnota, varianční matice . . . . .	8
1.1.2 Rozdělení a hustota náhodného vektoru . . . . .	9
1.1.3 Mnohorozměrné normální rozdělení . . . . .	9
1.1.4 Některá další rozdělení . . . . .	11
1.2 Lineární model . . . . .	13
1.2.1 Výběr z normálního rozdělení . . . . .	13
1.2.2 Jednoduché třídění . . . . .	13
1.2.3 Regresní přímka . . . . .	14
1.3 Odhad střední hodnoty v lineárním modelu . . . . .	14
1.3.1 Výpočet pravoúhlého průmětu . . . . .	15
1.3.2 Proč právě pravoúhlý průmět? . . . . .	16
1.3.3 Metoda nejmenších čtverců . . . . .	16
1.3.4 Výběr z normálního rozdělení (pokračování ze str. 13) . . . . .	17
1.3.5 Jednoduché třídění (pokračování ze str. 13) . . . . .	18
1.3.6 Regresní přímka (pokračování ze str. 14) . . . . .	19
1.4 Rozdělení pravoúhlých průmětů . . . . .	20
1.5 Odhad rozptylu . . . . .	23
1.5.1 Výběr z normálního rozdělení (pokračování ze str. 17) . . . . .	26
1.5.2 Jednoduché třídění (pokračování ze str. 18) . . . . .	27
1.5.3 Regresní přímka (pokračování ze str. 19) . . . . .	27
1.6 Submodel . . . . .	28
1.6.1 Náhodný vektor $\widehat{\mathbf{Y}} - \mathbf{Y}_S$ . . . . .	28
1.6.2 Test submodelu . . . . .	29
1.6.3 Užití Pythagorovy věty . . . . .	30
1.6.4 Tabulka analýzy rozptylu . . . . .	30
1.6.5 Výběr z normálního rozdělení (pokračování ze str. 26) . . . . .	31
1.6.6 Jednoduché třídění (pokračování ze str. 27) . . . . .	32
1.6.7 Jednoduché třídění – obecná formulace . . . . .	33
1.6.8 Regresní přímka (pokračování ze str. 27) . . . . .	35
1.6.9 Mnohonásobná regrese . . . . .	38
1.6.10 Dvě regresní přímky . . . . .	39
1.7 Lineární množina jako submodel . . . . .	40
1.7.1 Pravoúhlý průmět do lineární množiny . . . . .	41
1.7.2 Vlastnosti vektoru $\widehat{\mathbf{Y}} - \mathbf{Y}_{S'}$ . . . . .	42

1.7.3	Užití Pythagorovy věty . . . . .	42
1.7.4	Výběr z normálního rozdělení (pokračování ze str. 31) . . . . .	43
1.7.5	Regresní přímka (pokračování ze str. 35) . . . . .	45
1.7.6	Mnohonásobná regrese (pokračování ze str. 38) . . . . .	50
1.8	Více submodelů . . . . .	53
1.8.1	Posloupnost do sebe vnořených podprostorů . . . . .	53
1.8.2	Systém navzájem kolmých podprostorů . . . . .	55
1.8.3	Dvojné třídění bez interakcí . . . . .	56
1.8.4	Dvojné třídění s interakcemi . . . . .	62
1.8.5	Dvojné třídění – obecná formulace . . . . .	65
1.9	Aplikace rozdělení $t$ . . . . .	69
1.9.1	Výběr z normálního rozdělení (pokračování ze str. 31) . . . . .	70
1.9.2	Dvouvýběrový $t$ -test . . . . .	72
1.9.3	Mnohonásobná regrese (pokračování ze str. 50) . . . . .	76
1.9.4	Regresní přímka (pokračování ze str. 35) . . . . .	80
1.9.5	Jednoduché třídění (pokračování ze str. 33) . . . . .	84
1.10	Korelační koeficienty, koeficienty spolehlivosti . . . . .	85
1.10.1	Regresní přímka (pokračování ze str. 80) . . . . .	85
1.10.2	Mnohonásobná regrese (pokračování ze str. 76) . . . . .	88
1.10.3	Výběrový parciální korelační koeficient . . . . .	89
1.10.4	Koeficient korelace v modelu bez absolutního členu . . . . .	92
1.11	Vazba regresních koeficientů v modelu s neúplnou hodnotí . . . . .	93
1.11.1	Jednoduché třídění (pokračování ze str. 84) . . . . .	93
1.11.2	Dvojné třídění (pokračování ze str. 65) . . . . .	95
1.12	Aplikace Scheffého věty . . . . .	97
1.12.1	Jednoduché třídění (pokračování ze str. 93) . . . . .	99
1.12.2	Pás spolehlivosti pro regresní přímku . . . . .	102
<b>2</b>	<b>Teoretická část</b>	<b>108</b>
2.1	Geometrické důsledky zavedení skalárního součinu . . . . .	108
2.2	Kolmost podprostorů . . . . .	108
2.2.1	Základní definice kolmosti podprostorů . . . . .	110
2.2.2	Zobecnění pojmu kolmosti, knižní kolmost . . . . .	112
2.2.3	Knižní kolmost dvou podprostorů . . . . .	113
2.2.4	Knižní kolmost tří podprostorů . . . . .	114
2.2.5	Knižní kolmost více podprostorů . . . . .	117
2.3	Pravoúhlý průmět . . . . .	118
2.3.1	Existence a jednoznačnost . . . . .	118
2.3.2	Nejbližší prvek . . . . .	119
2.3.3	Metody výpočtu . . . . .	119
2.3.4	Ortogonální projekce a její vlastnosti . . . . .	121
2.4	Tjurův systém . . . . .	124
2.4.1	Rozklad Tjurova systému na kolmé podprostory . . . . .	125
2.4.2	Tjurův systém a třídění . . . . .	126
2.4.3	Trojné třídění . . . . .	126
2.4.4	Trojné třídění s jednou interakcí prvního řádu . . . . .	128
2.5	Náhodný vektor a jeho charakteristiky . . . . .	130

2.5.1	Poznámky k definici náhodného vektoru . . . . .	130
2.5.2	Střední hodnota, varianční operátor . . . . .	131
2.5.3	Obraz variančního operátoru . . . . .	132
2.5.4	Přechod k jinému skalárnímu součinu . . . . .	133
2.5.5	Rozklad samoadjungovaného zobrazení na součet ortogonálních projekcí . . . . .	134
2.6	Geometrické vlastnosti mnohorozměrného normálního rozdělení . . . . .	135
2.6.1	Rotace soustavy souřadnic . . . . .	135
2.6.2	Rozdělení podmíněného pravoúhlého průmětu . . . . .	136
2.6.3	Nezávislost na směru a nezávislost souřadnic . . . . .	137
2.6.4	Rozdělení homomorfismu $H(\mathbf{Y})$ . . . . .	138
2.7	Odvození vybraných rozdělení . . . . .	139
2.7.1	Rozdělení $\chi^2$ . . . . .	139
2.7.2	Rozdělení $t$ . . . . .	140
2.7.3	Rozdělení $F$ . . . . .	143
2.8	Gaussova-Markovova věta . . . . .	145
2.8.1	Funkcionální a vektorová verze Gaussovy-Markovovy věty . . . . .	145
2.8.2	Ekvivalence obou definic . . . . .	146
2.8.3	Důkaz Gaussovy-Markovovy věty . . . . .	147
2.8.4	Důsledky Gaussovy-Markovovy věty a její zobecnění . . . . .	148
2.9	Lineární vazba regresních koeficientů – obecné poznámky . . . . .	148
<b>3</b>	<b>Historická část</b> . . . . .	<b>151</b>
3.1	Počátky moderní matematické statistiky . . . . .	151
3.1.1	William Sealy Gosset (1876 – 1937) . . . . .	151
3.1.2	Ronald Aylmer Fisher (1890 – 1962) . . . . .	151
3.1.3	Gossetův článek <i>The Probable Error of a Mean</i> [28] . . . . .	152
3.1.4	Fisherovo „Studentovo“ rozdělení . . . . .	153
3.2	Geometrie v díle R. A. Fishera . . . . .	154
3.2.1	Sdružené rozdělení průměru a výběrové směrodatné odchylky . . . . .	154
3.2.2	Rozdělení výběrového korelačního koeficientu . . . . .	156
3.2.3	Rozdělení odchylky od průměru . . . . .	158
3.2.4	Sdružené rozdělení dvou odchylek od průměru . . . . .	159
3.2.5	Sdružené rozdělení průměrné odchylky a výběrové směrodatné odchylky . . . . .	162
3.2.6	Test významnosti . . . . .	166
3.3	Další osudy geometrického přístupu . . . . .	167
	Shrnutí . . . . .	170
	Přehled použité literatury . . . . .	170

## Předmluva

Lineární model je jedním ze základních pilířů matematické statistiky. To, že jej lze interpretovat geometricky, není pro statistiky žádným tajemstvím; této možnosti se však při výuce statistiky téměř nevyužívá, a je proto pro nezanedbatelnou část matematické veřejnosti prakticky neznámá, o studentech matematicky zaměřených oborů ani nemluvě. Tento přístup má přitom výhody, které by řada z nich mohla ocenit: názornost, možnost odvození mnoha důležitých výsledků z několika málo jednoduchých geometrických představ, komplexní pohled na celou problematiku a v poslední řadě i jistou estetickou hodnotu.

Cílem této práce je seznámit s geometrickým přístupem ty čtenáře, kteří nejsou spokojeni se svým dosavadním porozuměním statistické teorii a pro které je tradiční maticově-algebraický výklad příliš abstraktní a nepřehledný. U nich předpokládám jednak dobrou geometrickou představivost a znalost teorie vektorových prostorů, jednak alespoň hrubou představu o elementárních statistických pojmech (například rozdělení náhodné veličiny či hladina významnosti).

S ohledem na tento záměr jsem uspořádal text tak, že nejdůležitější myšlenky a výsledky jsou vysvětleny bez zbytečného otálení hned v první praktické části, a to často bez důkazů, aby případný čtenář nebyl odrazen opakováním známé teorie nebo rozebíráním pro něho nepodstatných detailů. Podrobnější rozbor jsem pak spolu s připomenutím některých teoretických základů umístil do druhé, tzv. teoretické části. Považuji však za nutné upozornit, že ačkoli jsem se zde snažil být důkladný, bylo mým cílem spíše jen naznačit možnosti hlubšího uchopení celé problematiky, než budovat rigorózní teorii. Tomu by jednak nevyhovovalo uvedené uspořádání (proto jsem také nepoužil formu „definice – věta – důkaz“), ale především bych v takovém případě nevyhnutelně kopíroval jiné autory. Náročnější čtenář tedy může v mém textu některá témata postrádat, neboť jejich výběr je mimo jiné podřízen také tomu, do jaké míry jsem byl schopen je originálně zpracovat. Místy jsem ovšem považoval za nezbytné použít cizí formulaci či důkaz; v takových případech to je samozřejmě vždy důsledně citováno.

Geometrická interpretace lineárního modelu však není pozoruhodná jen z didaktického hlediska. Existuje totiž řada přesvědčivých dokladů toho, že je to zároveň pojetí historicky původní a že autor lineárního modelu, geniální britský matematik a biolog Ronald Aylmer Fisher, jeho teorii vypracoval právě díky své schopnosti geometrické vizualizace. Těmito souvislostmi se zabývá třetí, historická část práce.

Čtenářům toužícím po důkladnějším studiu mohu vřele doporučit vynikající knihu [31], kde je velice solidně vybudována kompletní teorie lineárního modelu na geometrickém základě. Ti, kteří naopak moji práci shledají příliš obtížnou, ale přesto se i nadále domnívají, že geometrie by pro ně mohla být vhodnou cestou ke statistice, snad naleznou názornější výklad v knize [27]. Co se týče potřebných základů geometrie a algebry, lze je nalézt například v učebnicích [21] a [4], za hlavní zdroj znalostí z matematické statistiky mi sloužily knihy [1], [2], [3], [34] a [35]. Informace o díle a životě R. A. Fishera shrnuje nejúplněji životopis [5].

Závěrem bych rád poznamenal, že si jsem vědom, že vhodnost či nevhodnost určité didaktické metody je záležitost subjektivní a obtížně prokazatelná. Na základě vlastní zkušenosti se však domnívám, že by si geometrický přístup mohl najít své příznivce; a protože je matematická statistika všeobecně považována

za obtížné odvětví matematiky, odvažuji se zároveň doufat, že má snaha o její zpřístupnění širšímu okruhu čtenářů nebude přijata nevlídně těmi, kteří dávají přednost tradičnějším metodám.

## Přehled použitého značení

$\mathbf{Y}, \boldsymbol{\beta}, \dots$	vektor, náhodný vektor (tučným patkovým písmem), resp. jeho souřadnice zapsané do sloupce
$\mathbf{X}, \mathbf{M}, \dots$	matice (tučným bezpatkovým písmem)
$\mathbf{x}^T$	transpozice matice $\mathbf{A}$
$\mathbf{I}_n$	jednotková matice o rozměrech $n \times n$
$\boldsymbol{\mu}$	střední hodnota náhodného vektoru popsaného lineárním modelem
$\Sigma_{\mathbf{Y}}$	varianční operátor náhodného vektoru $\mathbf{Y}$ (str. 131)
$\mathbf{e}$	vektor o souřadnicích $(1, \dots, 1)^T$
$\mathbf{e}_1, \dots, \mathbf{e}_n$	vektory zavedené ortonormální báze
$E$	podprostor generovaný vektorem $(1, \dots, 1)^T$
$M$	podprostor určený modelem
$\bar{\mathbf{Y}}$	pravoúhlý průmět vektoru $\mathbf{Y}$ do podprostoru $E$ , tj. vektor $(\bar{Y}, \dots, \bar{Y})^T$
$\widehat{\mathbf{Y}}$	pravoúhlý průmět náhodného vektoru $\mathbf{Y}$ do podprostoru určeného modelem
$\mathbf{Y}_M$	pravoúhlý průmět náhodného vektoru $\mathbf{Y}$ do podprostoru $M$
$P_M$	ortogonální projekce do podprostoru $M$ (str. 121)
$Q_M$	ortogonální projekce do podprostoru $M^\perp$
$[\mathbf{x}_1, \dots, \mathbf{x}_k]$	lineární obal vektorů $\mathbf{x}_1, \dots, \mathbf{x}_k$
$\chi_n^2(\alpha)$	kritická hodnota rozdělení $\chi_n^2$ na hladině $\alpha$ (str. 11)
$t_n(\alpha)$	kritická hodnota rozdělení $t_n$ na hladině $\alpha$ (str. 11)
$F_{m,n}(\alpha)$	kritická hodnota rozdělení $F_{m,n}$ na hladině $\alpha$ (str. 11)
$S^2$	reziduální rozptyl (str. 24)
$\mathbf{u} + M$	lineární množina určená vektorem $\mathbf{u}$ a podprostorem $M$ (str. 40)
$A + B$	součet podprostorů $A, B$ (str. 109)

$A - B$	ortogonální doplněk podprostoru $B$ v podprostoru $A$ (str. 109)
$M^\perp$	ortogonální doplněk podprostoru $M$ (str. 23)
$A \perp B$	podprostory $A, B$ jsou navzájem kolmé (str. 110)
$A \perp\!\!\!\perp B$	alternativní definice kolmosti podprostorů $A, B$ (str. 112)
$A \sqcup B$	podprostory $A, B$ jsou navzájem knižně kolmé (str. 112)
$\{A_1, \dots, A_k\} \in \mathcal{P}^\perp$	podprostory $A_1, \dots, A_k$ jsou po dvojicích navzájem kolmé
$\{A_1, \dots, A_k\} \in \mathcal{P}^\sqcup$	podprostory $A_1, \dots, A_k$ jsou po dvojicích navzájem knižně kolmé (str. 112)
$A = L_1 \oplus \dots \oplus L_k$	podprostory $L_1, \dots, L_k$ tvoří ortogonální rozklad podprostoru $L$ (str. 110)
$\text{Im } T$	obraz zobrazení $T$
$\text{Ker } T$	jádro zobrazení $T$
$S_n(r)$	objem $n$ -rozměrné sféry o poloměru $r$
$S_n$	objem $n$ -rozměrné sféry o poloměru 1

# 1. Praktická část

## 1.1 Náhodný vektor

Základní charakteristikou geometrického přístupu k lineárnímu modelu je způsob, jakým je zde chápán *náhodný vektor*. Zatímco při standardním výkladu je definován jako uspořádaná  $n$ -tice náhodných veličin  $(Y_1, \dots, Y_n)^T$ , při geometrickém pohledu jej – zjednodušeně řečeno – považujeme za výsledek náhodného pokusu, jehož možnými realizacemi jsou vektory, tj. prvky nějakého reálného vektorového prostoru  $V_n$  konečné dimenze  $n$ , na kterém je definován skalární součin  $\circ$ , a tedy i norma  $\|\cdot\|$  a úhel. Vektory si tedy můžeme představovat ve středoškolském smyslu jako „množiny všech uspořádaných úseček stejné délky a stejného směru“. Pokud jde o vyšší dimenze, není třeba se nijak zvlášť znepokojovat nedostatkem představivosti: pro intuitivní porozumění dokážeme většinu úvah zjednodušit na trojrozměrný případ a u formálních důkazů na ni stejně nemůžeme spoléhat.

Když nyní v tomto vektorovém prostoru  $V_n$  zvolíme nějakou bázi, souřadnice náhodného vektoru vzhledem k této bázi budou zřejmě náhodné veličiny a vytvoří náhodný vektor v obvyklém smyslu. Důležitý důsledek geometrické definice se však ukáže v situaci, kdy se rozhodneme soustavu souřadnic změnit: tehdy totiž budeme moci hovořit o stále stejném náhodném vektoru, ačkoli jeho souřadnice se změní. Tím se tento přístup podstatně liší od od obvyklého pojetí, kde změna souřadnic náhodného vektoru automaticky znamená, že se jedná o jiný vektor.

Tak se zcela přirozeně nabízí možnost studovat vlastnosti náhodného vektoru z pouhého geometrického názoru, bez použití souřadnic, a souřadnicový zápis aplikovat až na dosažené závěry. To je důvod, proč se tento přístup nazývá v angličtině *coordinate-free*.

V důsledném výkladu založeném na geometrické definici by měla být zavedena různá označení pro vektor a pro jeho souřadnicovou reprezentaci. Tím by se však naše zápisy vzdálily od podoby, na kterou je pravděpodobně většina čtenářů zvyklá; zvolíme proto cestu kompromisu. Na vektorovém prostoru  $V_n$  zavedeme nějakou ortonormální bázi, a aniž bychom ztráceli ze zřetele výchozí geometrickou představu, budeme používat stejná označení pro prvky vektorového prostoru  $V_n$  a pro jejich souřadnicové reprezentace vzhledem k této bázi; v druhém případě budeme tyto souřadnice psát do sloupce a v té podobě je příležitostně zapojíme i do maticových formulí. Podobně nebudeme rozlišovat mezi operacemi a relacemi definovanými na  $V_n$  a jim odpovídajícími protějšky na  $\mathbb{R}^n$ . Pokud nebudeme měnit soustavu souřadnic, nebude hrozit nedorozumění; v opačném případě na to vždy výslovně upozorníme.

### 1.1.1 Střední hodnota, varianční matice

*Střední hodnotu a varianční matici* náhodného vektoru  $\mathbf{Y}$  zavedeme obvyklým způsobem, tj.

$$\begin{aligned} \mathbf{E} \mathbf{Y} &\equiv (\mathbf{E}Y_1, \dots, \mathbf{E}Y_n)^T, \\ \mathbf{V}_{\mathbf{Y}} &\equiv (\text{cov}(Y_i, Y_j))_{n \times n}, \end{aligned}$$

kde  $EY_i$  je střední hodnota náhodné veličiny  $Y_i$ ,  $\text{cov}(Y_i, Y_i)$  je její rozptyl (či též variance), značený ovšem obvykle  $\text{var } Y_i$ , a  $\text{cov}(Y_i, Y_j)$  je kovariance náhodných veličin  $Y_i, Y_j$  (podrobněji viz např. [3]). Připomeňme elementární vztahy

$$E(\mathbf{a} + \mathbf{B}\mathbf{Y}) = \mathbf{a} + \mathbf{B} \cdot E\mathbf{Y}, \quad (1.1)$$

$$\mathbf{V}_{\mathbf{a}+\mathbf{B}\mathbf{Y}} = \mathbf{B}\mathbf{V}_Y\mathbf{B}^T, \quad (1.2)$$

platné pro libovolný vektor  $\mathbf{a}$  a matici  $\mathbf{B}$  vhodných rozměrů. Ze vztahu (1.1) plyne, že střední hodnotu lze geometricky interpretovat, protože její souřadnice reprezentují vektor, jehož poloha není na zvolené soustavě souřadnic závislá. Je-li totiž  $\mathbf{B}$  matice přechodu od původní báze prostoru  $V_n$  k nějaké jiné, je střední hodnota nových souřadnic rovna  $E(\mathbf{B}\mathbf{Y})$ , zatímco souřadnice původní střední hodnoty vzhledem k nové bázi jsou  $\mathbf{B} \cdot E\mathbf{Y}$ .

Nebude-li uvedeno jinak, budeme střední hodnotu náhodného vektoru  $\mathbf{Y}$  značit v celé naší práci tradičně zavedeným symbolem  $\boldsymbol{\mu}$ .

### 1.1.2 Rozdělení a hustota náhodného vektoru

Chování náhodného vektoru nejuplněji charakterizuje jeho *rozdělení*. To je funkce, která každé „rozumné“ podmnožině prostoru  $V_n$  přiřazuje pravděpodobnost, že realizace náhodného vektoru bude prvkem této podmnožiny. V našem pojednání se setkáme pouze se spojitým rozdělením, v jehož případě je toto pravidlo obvykle formulováno pomocí tzv. *hustoty rozdělení*, což je funkce  $f : V_n \rightarrow \mathbb{R}$  udávající prostřednictvím vztahu

$$dP = f(\mathbf{y}) dV \quad (1.3)$$

pravděpodobnost  $dP$ , že realizace náhodného vektoru bude ležet v „nekonečně malém“ okolí vektoru  $\mathbf{y} \in V_n$  o  $n$ -rozměrném objemu  $dV$ . V souřadnicích má tento zápis tvar

$$dP = f(y_1, \dots, y_n) \cdot dy_1 \cdots dy_n;$$

$dP$  je tedy pravděpodobnost, že nastane jev

$$\mathbf{Y} \in \langle y_1, y_1 + dy_1 \rangle \times \cdots \times \langle y_n, y_n + dy_n \rangle.$$

Stojí za zmínku, že vztah (1.3) může popisovat rozdělení náhodného vektoru i v takovém případě, kdy je množina jeho možných realizací omezena na nějaký podprostor (obecněji lineární množinu nebo varietu) prostoru  $V_n$ , jehož dimenze je  $k < n$  (například v případě podmíněného rozdělení, kdy jsou souřadnice náhodného vektoru vázány nějakou lineární podmínkou). Pak ovšem  $dV$  představuje  $k$ -rozměrný objem a v obecném případě jej nelze vyjádřit pomocí původních souřadnic.

### 1.1.3 Mnohorozměrné normální rozdělení

Má-li každá lineární funkce náhodného vektoru  $\mathbf{Y}$  normální rozdělení, říkáme, že  $\mathbf{Y}$  má *mnohorozměrné normální rozdělení*; je-li  $\boldsymbol{\mu}$  jeho střední hodnota a  $\mathbf{V}$  varianční matice, vyjadřujeme to zápisem

$$\mathbf{Y} \sim N(\boldsymbol{\mu}, \mathbf{V}).$$

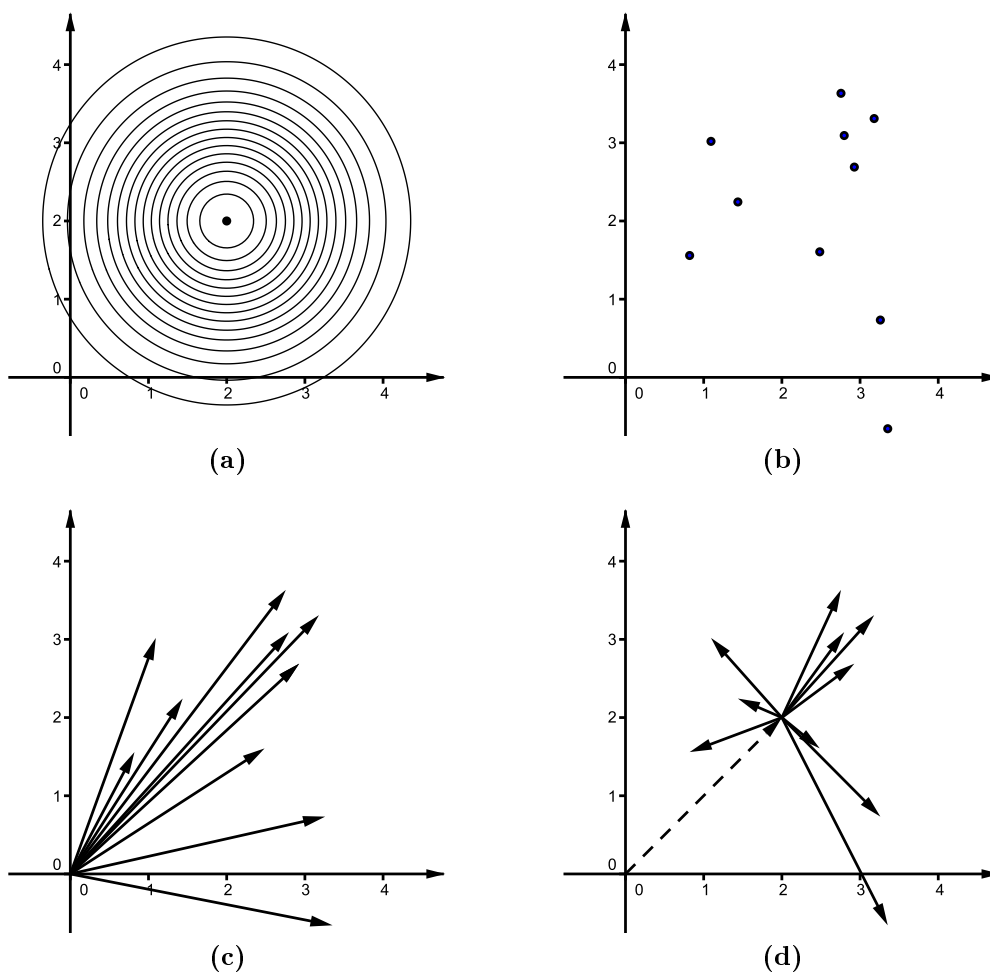
Hustota  $\mathbf{Y}$  je parametry  $\boldsymbol{\mu}$  a  $\mathbf{V}$  jednoznačně určena. Z definice speciálně plyne, že  $\mathbf{Y}$  má mnohorozměrné normální rozdělení právě tehdy, když jeho souřadnice mají normální rozdělení.

V praxi je časté, že varianční matice je  $\sigma^2$ -násobkem jednotkové matice  $\mathbf{I}_n$  typu  $n \times n$ . To znamená, že jednotlivé souřadnice náhodného vektoru mají stejný rozptyl  $\sigma^2$  a kovariance jakýchkoli dvou různých souřadnic je nulová. Lze dokázat (viz [26]), že v případě normálního rozdělení je nulová kovariance ekvivalentní s nezávislostí; jednotlivé souřadnice náhodného vektoru jsou tedy v případě rozdělení  $N(\boldsymbol{\mu}; \sigma^2 \mathbf{I}_n)$  nezávislé.

Hustota je v tomto speciálním případě určena předpisem

$$f(\mathbf{y}) = \left(\sqrt{2\pi} \cdot \sigma\right)^{-n} \exp\left\{-\frac{\|\mathbf{y} - \boldsymbol{\mu}\|^2}{2\sigma^2}\right\}. \quad (1.4)$$

Závisí tedy pouze na vzdálenosti od střední hodnoty, nikoli na směru. Jinými slovy, body se stejnou hustotou tvoří sféry koncentricky uspořádané kolem střední hodnoty. Obrázek 1.1 ilustruje toto uspořádání, chování náhodného vektoru a různé způsoby zobrazování jeho realizací pro případ  $n = 2$ .



**Obrázek 1.1:** (a) Vrstevnice hustoty náhodného vektoru  $\mathbf{Y}$  s rozdělením  $N((2, 2)^T, \mathbf{I}_2)$ , odstupňované po 0,01. Hodnota hustoty v bodě  $(2, 2)^T$  je  $(2\pi)^{-1} \doteq 0,159$ . (b) Znázornění deseti realizací téhož náhodného vektoru. (c) Tytéž realizace znázorněné jako vektory. (d) Tytéž realizace znázorněné jako součet střední hodnoty (čárkovaně) a odchylky od střední hodnoty.

### 1.1.4 Některá další rozdělení

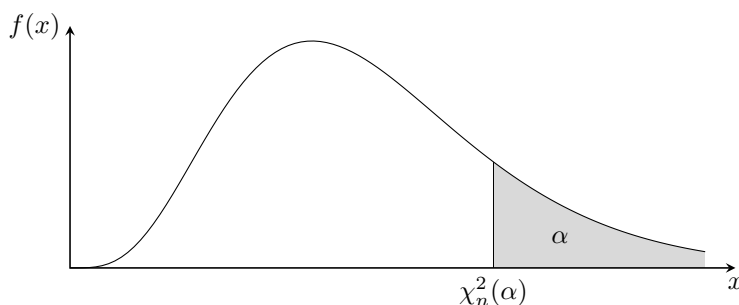
Připomeňme ještě další užitečná rozdělení odvozená od normálního rozdělení. Necht' náhodné veličiny  $Y_1, \dots, Y_n$  a  $Z_1, \dots, Z_m$  mají rozdělení  $N(0, 1)$  a jsou navzájem nezávislé; pak náhodná veličina

$$\sum_{i=1}^n Y_i^2 \sim \chi_n^2 \quad (1.5)$$

má rozdělení zvané  $\chi^2$  o  $n$  stupních volnosti, zkráceně  $\chi_n^2$ . Je-li  $0 < \alpha < 1$ , nazýváme *kritickou hodnotou* tohoto rozdělení na hladině  $\alpha$  takovou hodnotu  $\chi_n^2(\alpha)$ , která splňuje podmínku

$$Y \sim \chi_n^2 \implies P[Y \geq \chi_n^2(\alpha)] = \alpha$$

(viz obr. 1.2). Náhodná veličina



**Obrázek 1.2:** Hustota a kritická hodnota rozdělení  $\chi_n^2$  (zde pro případ  $n = 10$ ).

$$\frac{Z_1}{\sqrt{\sum_{i=1}^n Y_i^2 / n}} \quad (1.6)$$

má rozdělení  $t_n$ , tj. „*Studentovo*“ rozdělení  $t$  o  $n$  stupních volnosti. Jeho kritická hodnota  $t_n(\alpha)$  je definovaná vztahem

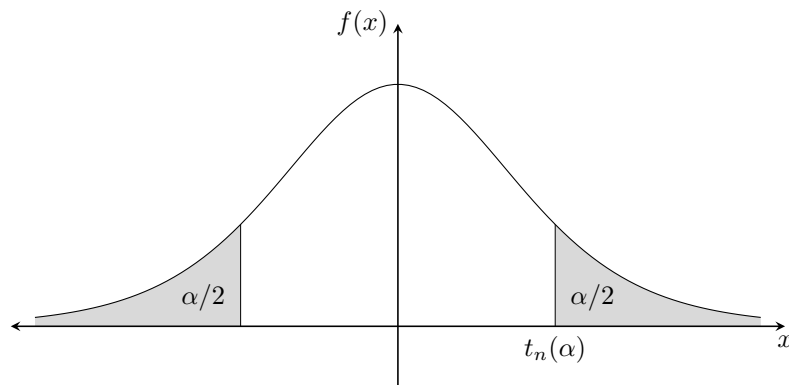
$$X \sim t_n \implies P[X \geq t_n(\alpha)] = \frac{\alpha}{2}, \quad (1.7)$$

resp. ekvivalentní podmínkou

$$X \sim t_n \implies P[|X| \geq t_n(\alpha)] = \alpha$$

(viz obr. 1.3). Konečně náhodná veličina

$$\frac{\sum_{i=1}^m Z_i^2 / m}{\sum_{i=1}^n Y_i^2 / n} \quad (1.8)$$

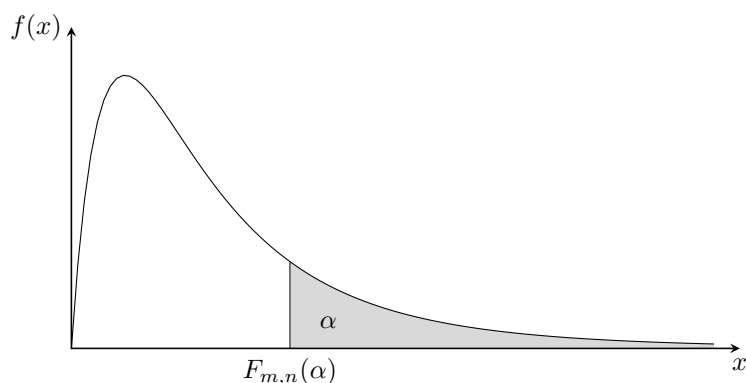


**Obrázek 1.3:** Hustota a kritická hodnota rozdělení  $t_n$  (zde pro  $n = 7$ ).

má tzv. *Fisherovo rozdělení* o  $m$  a  $n$  stupních volnosti, zkráceně  $F_{m,n}$ . Kritická hodnota  $F_{m,n}(\alpha)$  tohoto rozdělení je určena rovností

$$X \sim F_{m,n} \implies \mathbb{P}[X \geq F_{m,n}(\alpha)] = \alpha$$

(viz obr. 1.4).<sup>1</sup>



**Obrázek 1.4:** Hustota a kritická hodnota rozdělení  $F_{m,n}$  (konkrétně pro  $m = 4$ ,  $n = 10$ ).

Povšimněme si ještě celkem prostého vztahu mezi kritickými hodnotami dvou posledně jmenovaných rozdělení. Má-li náhodná veličina  $X$  rozdělení  $t_n$ , má náhodná veličina  $X^2$  zřejmě rozdělení  $F_{1,n-1}$ . Platí tedy

$$\mathbb{P}[|X| \geq t_{n-1}(\alpha)] = \alpha = \mathbb{P}[X^2 \geq F_{1,n-1}(\alpha)],$$

z čehož je patrná rovnost

$$t_{n-1}^2(\alpha) = F_{1,n-1}(\alpha). \tag{1.9}$$

---

<sup>1</sup>Kritické hodnoty výše uvedených rozdělení jsou často uváděny v učebnicích statistiky, také je však lze snadno zjistit pomocí jakéhokoli statistického software nebo aplikace Excel.

## 1.2 Lineární model

O lineárním modelu hovoříme za předpokladu, že pro náhodný vektor  $\mathbf{Y}$  o  $n$  složkách platí

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}, \quad (1.10)$$

kde  $\mathbf{X}$  je známá matice typu  $n \times m$  ( $m < n$ ),  $\boldsymbol{\beta} \equiv (\beta_1, \dots, \beta_m)^T$  je uspořádaná  $m$ -tice neznámých parametrů a  $\mathbf{Z}$  je náhodný vektor o  $n$  složkách, jehož střední hodnota je  $\mathbf{0} \equiv (0, \dots, 0)^T$ . Výraz  $\mathbf{X}\boldsymbol{\beta}$  tudíž představuje střední hodnotu  $\boldsymbol{\mu}$  náhodného vektoru  $\mathbf{Y}$ .

Ačkoli to není všude nezbytné, omezíme se v celém našem pojednání na případ normálního rozdělení. Dále budeme předpokládat – pokud nebude řečeno jinak – že sloupce matice  $\mathbf{X}$  jsou lineárně nezávislé, tj. hodnost této matice je  $m$ , a že varianční matice náhodného vektoru  $\mathbf{Z}$ , resp.  $\mathbf{Y}$ , je  $\sigma^2$ -násobkem jednotkové matice typu  $n \times n$ , kde  $\sigma^2$  je neznámý parametr, tj.

$$\mathbf{Z} \sim \mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n).$$

Obecnější případy budou diskutovány zvlášť.

## Příklady

### 1.2.1 Výběr z normálního rozdělení

Představme si, že  $n$  náhodně vybraným osobám podáme nějaký lék a zjistíme u nich změny v hodnotách nějaké fyziologické veličiny před aplikací tohoto léku a po ní. Předpokládáme-li, že změna této veličiny způsobená podáním léku se řídí normálním rozdělením s neznámou střední hodnotou  $\mu$  a rozptylem  $\sigma^2$ , představuje námi získaná  $n$ -tice hodnot realizací náhodného vektoru  $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ , jehož střední hodnota je  $\boldsymbol{\mu} = (\mu, \dots, \mu)^T$  a jehož chování popisuje model

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \cdot \mu + \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix}, \quad (1.11)$$

kde náhodný vektor  $\mathbf{Z} = (Z_1, \dots, Z_n)^T$  má rozdělení  $\mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ . Matice  $\mathbf{X}$  má v tomto případě pouze jeden sloupec a všechny její prvky jsou rovny jedné.

### 1.2.2 Jednoduché třídění

Podobně jako v předchozím příkladu budeme zaznamenávat reakci pacientů na podání léku, ale tentokrát budeme zkoušet tři různé léky a pro názornost zvolíme konkrétní počet pacientů, řekněme sedm. Prvním dvěma pacientům podáme lék F, druhým dvěma lék G a zbývajícím třem lék H. Za předpokladu, že odezva na všechny použité léky se řídí normálním rozdělením se stejným rozptylem  $\sigma^2$ , ale různými středními hodnotami  $\mu_f, \mu_g, \mu_h$ , získáme měřením realizací náhodného vektoru  $\mathbf{Y} = (Y_1, \dots, Y_7)^T$ , jehož střední hodnota je

$\boldsymbol{\mu} = (\mu_f, \mu_f, \mu_g, \mu_g, \mu_h, \mu_h, \mu_h)^T$  a který lze popsat lineárním modelem

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \mu_f \\ \mu_g \\ \mu_h \end{pmatrix} + \begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \\ Z_5 \\ Z_6 \\ Z_7 \end{pmatrix}, \quad (1.12)$$

kde vektor  $\mathbf{Z} = (Z_1, \dots, Z_7)^T$  má rozdělení  $N(\mathbf{0}, \sigma^2 \mathbf{I}_7)$ .

### 1.2.3 Regresní přímka

Předpokládejme, že hmotnost  $Y$  úrody určité plodiny sklizené z jednoho aru závisí na hmotnosti  $x$  použitého hnojiva (obojí v kilogramech) vztahem

$$Y = \beta_0 + \beta_1 x + Z,$$

kde  $Z \sim N(0, \sigma^2)$ . Koeficienty  $\beta_0, \beta_1$  ani rozptyl  $\sigma^2$  nejsou známé. Máme-li k dispozici údaje ze sedmi různých experimentálních ploch, na kterých byly hmotnosti  $x_i$  použitého hnojiva postupně 3 kg, 2 kg, 4 kg, 3 kg, 5 kg, 4 kg a 7 kg, představují získané hodnoty sklizně realizaci náhodného vektoru  $\mathbf{Y} = (Y_1, \dots, Y_7)^T$ , jehož chování popisuje lineární model

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \end{pmatrix} = \begin{pmatrix} 1 & 3 \\ 1 & 2 \\ 1 & 4 \\ 1 & 3 \\ 1 & 5 \\ 1 & 4 \\ 1 & 7 \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \\ Z_5 \\ Z_6 \\ Z_7 \end{pmatrix}, \quad (1.13)$$

kde vektor  $\mathbf{Z} = (Z_1, \dots, Z_7)^T$  má opět rozdělení  $N(\mathbf{0}, \sigma^2 \mathbf{I}_7)$ .

## 1.3 Odhad střední hodnoty v lineárním modelu

Máme-li zformulovaný lineární model (1.10) a získáme nějakou konkrétní realizaci  $\mathbf{Y}$ , tj. jeden vektor  $\mathbf{y}$  ležící v  $n$ -rozměrném vektorovém prostoru  $V_n$ , je obvykle cílem odhadnout střední hodnotu  $\boldsymbol{\mu}$ , koeficienty  $\beta_i$  a rozptyl  $\sigma^2$ . Dále lze vytvářet a testovat různé hypotézy týkající se těchto parametrů.

Střední hodnota  $\boldsymbol{\mu}$  náhodného vektoru  $\mathbf{Y}$  je rovna výrazu  $\mathbf{X}\boldsymbol{\beta}$ . Použijeme-li označení  $\mathbf{x}_1, \dots, \mathbf{x}_m$  pro sloupce matice  $\mathbf{X}$ , můžeme psát

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} = \beta_1 \mathbf{x}_1 + \dots + \beta_m \mathbf{x}_m.$$

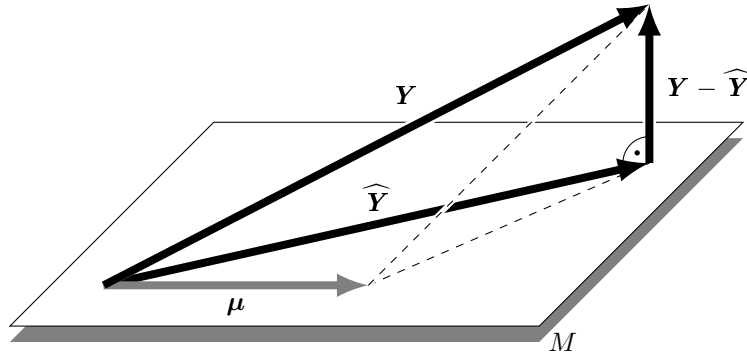
Vektor  $\boldsymbol{\mu}$  je tedy lineární kombinací vektorů  $\mathbf{x}_1, \dots, \mathbf{x}_m$  s neznámými koeficienty  $\beta_1, \dots, \beta_m$ . Musí proto ležet někde ve vektorovém podprostoru generovaném sloupci matice  $\mathbf{X}$ . Tento podprostor budeme značit symbolem  $M$ , tj. platí

$$\boldsymbol{\mu} \in M,$$

kde

$$M \equiv [\mathbf{x}_1, \dots, \mathbf{x}_m].$$

Jelikož předpokládáme, že vektory  $\mathbf{x}_1, \dots, \mathbf{x}_m$  jsou lineárně nezávislé, je dimenze tohoto podprostoru  $m$ . Získáme-li nyní nějakou konkrétní realizaci náhodného vektoru  $\mathbf{Y}$ , pravděpodobně v podprostoru  $M$  ležet nebude. V takové situaci je nejpřirozenějším krokem odhadnout  $\boldsymbol{\mu}$  pomocí pravoúhlého průmětu této realizace do podprostoru  $M$ . Náhodný vektor, jehož realizaci takto získáme, budeme v obecném případě značit symbolem  $\widehat{\mathbf{Y}}$  (viz obr. 1.5).



**Obrázek 1.5:** Náhodný vektor  $\mathbf{Y}$  a jeho pravoúhlý průmět  $\widehat{\mathbf{Y}}$  do podprostoru  $M$ . Náhodný vektor  $\widehat{\mathbf{Y}}$  je odhadem neznámé střední hodnoty  $\boldsymbol{\mu}$ , která leží v  $M$ .

### 1.3.1 Výpočet pravoúhlého průmětu

Z požadavků definujících pravoúhlý průmět (viz kapitola 2.3) lze vektor  $\widehat{\mathbf{Y}}$  snadno určit. Za prvé, aby bylo splněno  $\widehat{\mathbf{Y}} \in M$ ,<sup>2</sup> musí platit

$$\widehat{\mathbf{Y}} = \mathbf{X}\mathbf{b} \quad (1.14)$$

pro nějakou uspořádanou  $m$ -tici náhodných veličin<sup>3</sup>  $\mathbf{b} \equiv (b_1, \dots, b_m)^T$ . Druhým požadavkem je, aby byl vektor  $\mathbf{Y} - \widehat{\mathbf{Y}}$  kolmý na podprostor  $M$ ; musí tedy být kolmý na všechny vektory, které  $M$  generují, tj. na sloupce matice  $\mathbf{X}$ :

$$\mathbf{X}^T \cdot (\mathbf{Y} - \widehat{\mathbf{Y}}) = \mathbf{0}. \quad (1.15)$$

Po dosazení (1.14) do (1.15) dostáváme postupně

$$\begin{aligned} \mathbf{X}^T \cdot (\mathbf{Y} - \mathbf{X}\mathbf{b}) &= \mathbf{0}, \\ \mathbf{X}^T \mathbf{X}\mathbf{b} &= \mathbf{X}^T \mathbf{Y}. \end{aligned} \quad (1.16)$$

Jak je známo z lineární algebry (viz podkapitola 2.3.1), pravoúhlý průmět  $\widehat{\mathbf{Y}}$  vždy existuje a je určen jednoznačně. Soustava (1.16), zvaná *soustava normálních rovnic*, proto musí mít právě jedno řešení  $\mathbf{b}$ . Je totiž důsledkem podmínek

<sup>2</sup>Přesnější by bylo hovořit o „každé možné realizaci náhodného vektoru  $\widehat{\mathbf{Y}}$ “. Pro stručnost nebudeme tento rozdíl v dalším textu zdůrazňovat.

<sup>3</sup>Ve shodě se zvyklostmi budeme nadále hovořit o náhodném vektoru, ačkoli z hlediska geometrické definice se nejedná o stejný případ, jakým jsme zavedli náhodný vektor  $\mathbf{Y}$ .

(1.14) a (1.15), definujících pravoúhlý průmět, takže alespoň jedno řešení musí existovat. Těchto řešení však nemůže být více, neboť to by znamenalo (vzhledem k nezávislosti sloupců matice  $\mathbf{X}$ ) více různých vektorů  $\widehat{\mathbf{Y}} = \mathbf{X}\mathbf{b}$ . Matice  $\mathbf{X}^T\mathbf{X}$  je tudíž regulární, existuje k ní inverzní matice a můžeme dokončit:

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}, \quad (1.17)$$

$$\widehat{\mathbf{Y}} = \mathbf{X} (\mathbf{X}^T\mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}. \quad (1.18)$$

Poznamenejme, že jsou-li sloupce matice  $\mathbf{X}$  lineárně závislé, netvoří bázi podprostoru  $M$ , a koeficienty  $b_i$  proto nejsou jednoznačně určeny. Soustava (1.16) má v takovém případě nekonečně mnoho řešení a matice  $\mathbf{X}^T\mathbf{X}$  není regulární, neexistuje tedy její inverze. K nalezení nějakého řešení je pak třeba použít jinou metodu, např. pseudoinverzní matici; podrobněji viz podkapitola 2.3.3. Náhodný vektor  $\widehat{\mathbf{Y}}$  je však bez ohledu na tyto komplikace určen jednoznačně.

### 1.3.2 Proč právě pravoúhlý průmět?

Správnost naší intuice ohledně odhadu  $\boldsymbol{\mu}$  pomocí pravoúhlého průmětu náhodného vektoru  $\mathbf{Y}$  do podprostoru  $M$  potvrzuje Gaussova-Markovova věta, podle níž je – za předpokladu, že varianční matice je kladným násobkem matice jednotkové – průmět  $\widehat{\mathbf{Y}}$  tzv. *nejlepším nestranným lineárním odhadem* vektoru  $\boldsymbol{\mu}$  (viz kapitola 2.8). Jejím důležitým důsledkem je mj. i to, že v případě lineární nezávislosti sloupců matice  $\mathbf{X}$  jsou náhodné veličiny  $b_i$  stejně vhodnými odhady parametrů  $\beta_i$ .

Názorné ospravedlnění našeho postupu nám poskytuje metoda maximální věrohodnosti. Podle ní je totiž třeba odhadnout střední hodnotu tak, aby hustota pravděpodobnosti v bodě odpovídajícím získané realizaci byla maximální. Protože hustota rozdělení  $\mathbf{N}(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n)$  klesá se vzdáleností od střední hodnoty (viz (1.4)), je jasné, že za její odhad musíme vzít ten vektor ležící v  $M$ , který je k dané realizaci nejblíže – a takovou vlastnost má právě pravoúhlý průmět  $\widehat{\mathbf{Y}}$ .

### 1.3.3 Metoda nejmenších čtverců

Dodejme ještě, že právě zmíněná vlastnost je důvodem, proč se metodě odhadu  $\boldsymbol{\mu}$  pomocí  $\widehat{\mathbf{Y}}$  říká *metoda nejmenších čtverců*. S minimalizací délky  $\|\mathbf{Y} - \widehat{\mathbf{Y}}\|$  totiž minimalizujeme zároveň hodnotu výrazu  $\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2$ , což není nic jiného než zmíněný součet čtverců – naše soustava souřadnic je totiž ortonormální, a pro  $\mathbf{y} \in V$  proto platí

$$\|\mathbf{y}\|^2 = y_1^2 + \dots + y_n^2.$$

Označíme-li tedy  $x_{ij}$  prvky matice  $\mathbf{X}$ , můžeme určit  $\widehat{\mathbf{Y}} = \mathbf{X}\mathbf{b}$  také tak, že hledáme taková  $b_i$ , aby součet

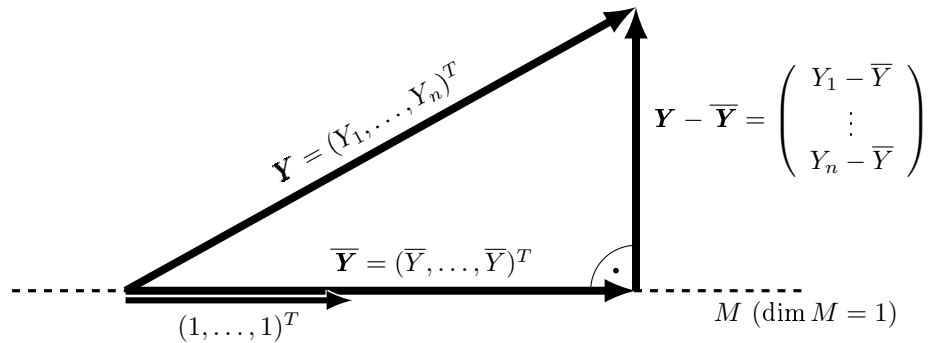
$$\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2 = \sum_{i=1}^n (Y_i - x_{i1}b_1 - \dots - x_{im}b_m)^2$$

byl minimální. Takto formulovanou úlohu je možné řešit užitím diferenciálního počtu; stejně jako podmínky (1.14), (1.15) vede však i tento postup k soustavě rovnic (1.16).

## Příklady

### 1.3.4 Výběr z normálního rozdělení (pokračování ze str. 13)

Podprostor  $M$ , ve kterém se podle modelu (1.11) nachází střední hodnota  $\boldsymbol{\mu}$ , je jednorozměrný a je generován vektorem  $(1, \dots, 1)^T$ . Je to tedy množina všech vektorů o souřadnicích  $(a, \dots, a)^T$ , kde  $a \in \mathbb{R}$ . Snadno se ověří, že pravoúhlým průmětem náhodného vektoru  $\mathbf{Y}$  do podprostoru  $M$  je náhodný vektor  $\widehat{\mathbf{Y}} = (\bar{Y}, \dots, \bar{Y})^T \equiv \bar{\mathbf{Y}}$  (viz obr. 1.6), kde



**Obrázek 1.6:** Pravoúhlým průmětem náhodného vektoru  $(Y_1, \dots, Y_n)^T$  do jednorozměrného podprostoru  $M$  generovaného vektorem  $(1, \dots, 1)^T$  je náhodný vektor  $(\bar{Y}, \dots, \bar{Y})^T$ .

$$\bar{Y} = \frac{Y_1 + \dots + Y_n}{n}.$$

Je totiž zřejmé  $\bar{\mathbf{Y}} \in M$  a zároveň

$$(\mathbf{Y} - \bar{\mathbf{Y}}) \circ (1, \dots, 1) = \sum_{i=1}^n (Y_i - \bar{Y}) = 0,$$

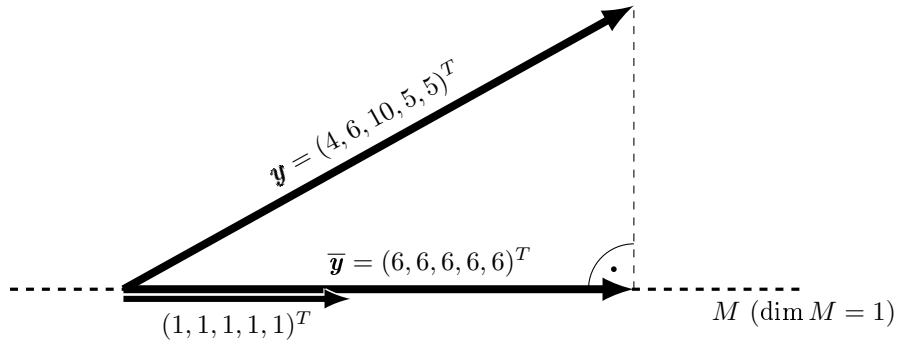
tudíž  $\mathbf{Y} - \bar{\mathbf{Y}} \perp M$ .

Náhodný vektor  $(\bar{Y}, \dots, \bar{Y})^T$  je proto nejlepším nestranným lineárním odhadem vektoru  $(\mu, \dots, \mu)^T$  a náhodná veličina  $\bar{Y}$  je nejlepším nestranným lineárním odhadem hodnoty  $\mu$ .

Představme si například, že u pěti pacientů naměříme hodnoty změny sledované veličiny postupně 4, 6, 10, 5 a 5. Pravoúhlým průmětem vektoru  $\mathbf{y} = (4, 6, 10, 5, 5)^T$  do podprostoru  $M$  je vektor  $\bar{\mathbf{y}} = (6, 6, 6, 6, 6)^T$  (viz obr. 1.7). Pro odhad neznámé střední hodnoty  $\mu$  tedy použijeme hodnotu 6.

Geometrické znázornění všech zmíněných vektorů pěkně ilustruje dva často uváděné vztahy. Skutečnost, že součet odchylek od průměru  $Y_i - \bar{Y}$  je vždy roven nule, odpovídá vlastně tomu, že vektor  $\mathbf{Y} - \bar{\mathbf{Y}}$  je kolmý na vektor  $(1, \dots, 1)^T$ . Podobně názornou interpretaci má i další známá vlastnost průměru: hodnota výrazu

$$\sum_{i=1}^n (Y_i - a)^2$$



**Obrázek 1.7:** Pravoúhlým průmětem náhodného vektoru  $\mathbf{y} = (4, 6, 10, 5, 5)^T$  do jedno-rozměrného podprostoru  $M$  generovaného vektorem  $(1, 1, 1, 1, 1)^T$  je náhodný vektor  $\bar{\mathbf{y}} = (6, 6, 6, 6, 6)^T$ .

nabývá svého minima v případě volby  $a = \bar{Y}$ . Tento výraz totiž představuje čtverec vzdálenosti vektoru  $\mathbf{Y}$  od vektoru  $(a, \dots, a)^T \in M$ ; ze všech takových vektorů má však nejmenší vzdálenost od vektoru  $\mathbf{Y}$  právě jeho pravoúhlý průmět do  $M$ .

### 1.3.5 Jednoduché třídění (pokračování ze str. 13)

V tomto případě je podprostor  $M$  generován vektory

$$\mathbf{a}_f \equiv \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{a}_g \equiv \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{a}_h \equiv \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{pmatrix},$$

a je tedy trojrozměrný. Podobně jako v předchozím příkladu lze snadno ukázat, že pravoúhlým průmětem náhodného vektoru  $\mathbf{Y}$  do  $M$  je náhodný vektor  $\widehat{\mathbf{Y}} = (\bar{Y}_f, \bar{Y}_f, \bar{Y}_g, \bar{Y}_g, \bar{Y}_h, \bar{Y}_h, \bar{Y}_h)^T$ , kde

$$\begin{aligned} \bar{Y}_f &= (Y_1 + Y_2)/2, \\ \bar{Y}_g &= (Y_3 + Y_4)/2, \\ \bar{Y}_h &= (Y_5 + Y_6 + Y_7)/3, \end{aligned}$$

neboť jednak je tento vektor očividně lineární kombinací vektorů  $\mathbf{a}_f, \mathbf{a}_g, \mathbf{a}_h$ , a tudíž leží v podprostoru  $M$ , a jednak platí

$$\begin{aligned} (\mathbf{Y} - \widehat{\mathbf{Y}}) \circ \mathbf{a}_f &= Y_1 + Y_2 - 2 \cdot \bar{Y}_f = 0, \\ (\mathbf{Y} - \widehat{\mathbf{Y}}) \circ \mathbf{a}_g &= Y_3 + Y_4 - 2 \cdot \bar{Y}_g = 0, \\ (\mathbf{Y} - \widehat{\mathbf{Y}}) \circ \mathbf{a}_h &= Y_5 + Y_6 + Y_7 - 3 \cdot \bar{Y}_h = 0, \end{aligned}$$

a tím pádem je splněna i podmínka  $(\mathbf{Y} - \widehat{\mathbf{Y}}) \perp M$ . Nejlepšími nestrannými lineárními odhady neznámých hodnot  $\mu_f, \mu_g, \mu_h$  jsou tedy náhodné veličiny  $\bar{Y}_f, \bar{Y}_g,$

$\bar{Y}_h$ .

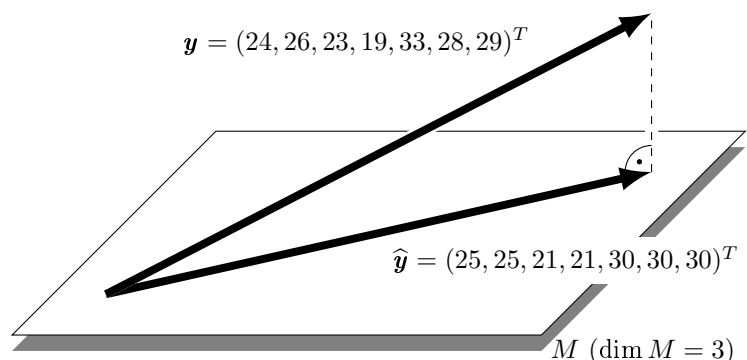
Pokud tedy v konkrétním případě naměříme například vektor hodnot

$$\mathbf{y} = (24, 26, 23, 19, 33, 28, 29)^T,$$

je průmětem této realizace do podprostoru  $M$  vektor

$$\hat{\mathbf{y}} = (25, 25, 21, 21, 30, 30, 30)^T$$

a k odhadu parametrů  $\mu_f, \mu_g, \mu_h$  použijeme hodnoty 25, 21 a 30 (viz obr. 1.8).



**Obrázek 1.8:** Odhad neznámé střední hodnoty  $\mu$  ze získaných dat v případě jednoduchého třídění popsaného modelem 1.12.

### 1.3.6 Regresní přímka (pokračování ze str. 14)

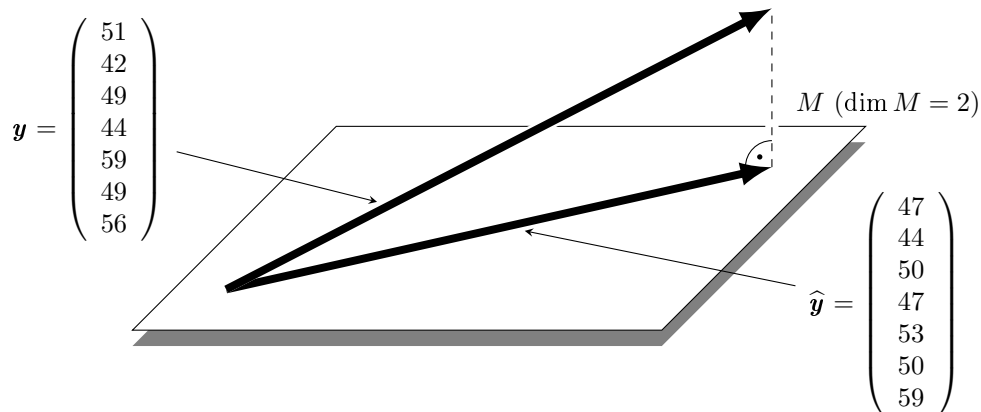
Podprostor  $M$  je v tomto případě dvojrozměrný, je to tedy rovina umístěná v sedmírozměrném prostoru  $V_7$ . Budou-li zjištěné výnosy na našich experimentálních pozemcích (tj. realizované souřadnice vektoru  $\mathbf{Y}$ ) například 51 kg, 42 kg, 49 kg, 44 kg, 59 kg, 49 kg a 56 kg, získáme užitím vzorců (1.17), (1.18) realizace náhodných vektorů  $\mathbf{b} = (b_0, b_1)^T$  a  $\hat{\mathbf{Y}}$  o souřadnicích

$$\mathbf{b} = \begin{pmatrix} 38 \\ 3 \end{pmatrix}, \quad \hat{\mathbf{y}} = \begin{pmatrix} 1 & 3 \\ 1 & 2 \\ 1 & 4 \\ 1 & 3 \\ 1 & 5 \\ 1 & 4 \\ 1 & 7 \end{pmatrix} \cdot \begin{pmatrix} 38 \\ 3 \end{pmatrix} = \begin{pmatrix} 47 \\ 44 \\ 50 \\ 47 \\ 53 \\ 50 \\ 59 \end{pmatrix}$$

(viz obr. 1.9). Složky  $b_0, b_1$  vektoru  $\mathbf{b}$  představují nejlepší nestranné lineární odhady složek  $\beta_0, \beta_1$  neznámého vektoru  $\beta$ . Na základě zjištěných dat tak můžeme odhadnout závislost výnosu  $Y$  na hmotnosti hnojiva  $x$  ve tvaru

$$\hat{Y} = 38 + 3x + Z,$$

kde  $Z \sim N(0, \sigma^2)$ .



**Obrázek 1.9:** Odhad neznámé střední hodnoty  $\mu$  ze získaných dat v případě regresní přímky popsané modelem 1.13.

### Obecná formulace

Uveďme nyní výsledky pro obecný případ. Platí-li pro  $n$  navzájem nezávislých náhodných veličin vztah

$$Y_i \sim \mathbf{N}(\beta_0 + \beta_1 x_i; \sigma^2),$$

kde hodnoty parametrů  $\beta_0, \beta_1$  a  $\sigma^2$  nejsou známé, můžeme náhodný vektor  $\mathbf{Y}$  popsat modelem

$$\mathbf{Y} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} Z_1 \\ \vdots \\ Z_2 \end{pmatrix} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}; \quad (1.19)$$

složky vektoru  $\mathbf{Z}$  se řídí rozdělením  $\mathbf{N}(0; \sigma^2)$  a jsou navzájem nezávislé. Střední hodnota vektoru  $\mathbf{Y}$  tedy leží v rovině  $M$  generované vektory  $\mathbf{e} \equiv (1, \dots, 1)^T$  a  $\mathbf{x} \equiv (x_1, \dots, x_n)^T$ . Řešením soustavy (1.16) získáme známé vzorce pro nejlepší nestranné lineární odhady  $b_0, b_1$  složek vektoru  $\boldsymbol{\beta}$ :

$$b_1 = \frac{\overline{xY} - \bar{x}\bar{Y}}{\overline{x^2} - \bar{x}^2}, \quad b_0 = \bar{Y} - b_1\bar{x}, \quad (1.20)$$

kde

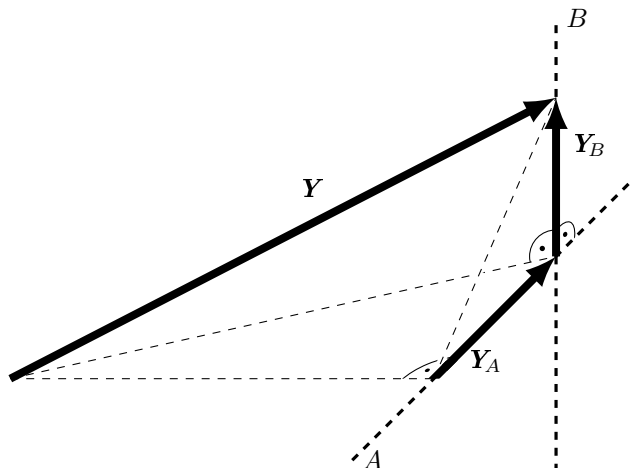
$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}, \quad \bar{Y} = \frac{\sum_{i=1}^n Y_i}{n}, \quad \overline{xY} = \frac{\sum_{i=1}^n x_i Y_i}{n}, \quad \overline{x^2} = \frac{\sum_{i=1}^n x_i^2}{n}.$$

Z nich dále vypočítáme souřadnice  $\hat{Y}_i = b_0 + b_1 x_i$  vektoru  $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b}$ , který je nejlepším nestranným lineárním odhadem střední hodnoty  $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ .

## 1.4 Rozdělení pravoúhlých průmětů

Odbočme nyní na chvíli k úvahám zaměřeným více teoreticky. Nechť  $A, B$  jsou podprostory vektorového prostoru  $V_n$  dimenzí  $a, b$ , které jsou navzájem kolmé,

tj. pro všechna  $\mathbf{a} \in A$ ,  $\mathbf{b} \in B$  platí  $\mathbf{a} \perp \mathbf{b}$ . Odvodíme některé důležité pravděpodobnostní charakteristiky pravoúhlých průmětů  $\mathbf{Y}_A$ ,  $\mathbf{Y}_B$  náhodného vektoru  $\mathbf{Y}$  do podprostorů  $A$  a  $B$  (viz obr. 1.10); nejdříve předpokládejme, že platí  $\mathbf{Y} \sim N(\mathbf{0}, \mathbf{I}_n)$ .



**Obrázek 1.10:** Pravoúhlé průměty náhodného vektoru  $\mathbf{Y}$  do dvou navzájem kolmých podprostorů  $A$  a  $B$ .

### Zavedení alternativní báze

Za tím účelem zavedeme ve  $V_n$  novou ortonormální bázi  $\mathbf{e}_1^*, \dots, \mathbf{e}_n^*$  takovou, že platí

$$\begin{aligned} \mathbf{e}_1^*, \dots, \mathbf{e}_a^* &\in A, \\ \mathbf{e}_{a+1}^*, \dots, \mathbf{e}_{a+b}^* &\in B. \end{aligned}$$

Nechť souřadnice náhodného vektoru  $\mathbf{Y}$  vzhledem k této nové bázi jsou  $(Y_1^*, \dots, Y_n^*)$ . Nové souřadnice průmětů  $\mathbf{Y}_A$  a  $\mathbf{Y}_B$  jsou pak zřejmě

$$\begin{aligned} (Y_1^*, \dots, Y_a^*, 0, \dots, 0, 0, \dots, 0), \\ (0, \dots, 0, Y_{a+1}^*, \dots, Y_{a+b}^*, 0, \dots, 0). \end{aligned}$$

Tyto nové souřadnice musí mít stejné rozdělení jako souřadnice původní. Hustota náhodného vektoru  $\mathbf{Y}$  totiž závisí pouze na vzdálenosti od počátku (viz vzorec (1.4) a obr. 1.1a). Ta se však při použití *jakýchkoli* ortonormálních souřadnic vypočte stejně, tj. ze součtu jejich čtverců, takže vzorec (1.4) se takovou transformací souřadnic nezmění.<sup>4</sup> Nové souřadnice  $Y_i^*$  mají tudíž rozdělení  $N(0, 1)$  a jsou navzájem nezávislé.

Podle definice rozdělení  $\chi^2$  tedy platí

$$\sum_{i=1}^a (Y_i^*)^2 \sim \chi_a^2.$$

<sup>4</sup>Zjednodušeně můžeme říci, že použitím jiné ortonormální báze se soustava souřadnic pouze pootočí, takže hustota  $\mathbf{Y}$  – která je konstantní na sférahách se středem v počátku – bude v nových souřadnicích „vypadat stejně“. Exaktnější zdůvodnění je uvedeno v kapitole 2.6.

Podle definice rozdělení  $F$  dostáváme

$$\frac{\sum_{i=1}^a (Y_i^*)^2 / a}{\sum_{i=a+1}^{a+b} (Y_i^*)^2 / b} \sim F_{a,b},$$

a konečně ve speciálním případě, kdy je podprostor  $A$  jednorozměrný, můžeme aplikovat i definici rozdělení  $t$  a odvodit, že platí

$$\frac{Y_1^*}{\sqrt{\sum_{i=2}^{b+1} (Y_i^*)^2 / b}} \sim t_b. \quad (1.21)$$

### Návrat k původní bázi

Jelikož jsou  $Y_i^*$  souřadnice vzhledem k ortonormální bázi, představují všechny výše uvedené sumy čtverce délek vektorů  $\mathbf{Y}_A$ ,  $\mathbf{Y}_B$ . Proto

$$\|\mathbf{Y}_A\|^2 \sim \chi_a^2, \quad (1.22)$$

$$\frac{\|\mathbf{Y}_A\|^2 / a}{\|\mathbf{Y}_B\|^2 / b} \sim F_{a,b}, \quad (1.23)$$

$$\frac{Y_1^*}{\|\mathbf{Y}_B\| / \sqrt{b}} \sim t_b, \quad (1.24)$$

přičemž při výpočtu délek se můžeme pochopitelně vrátit k původním souřadnicím. Vztah (1.21) je z tohoto hlediska trochu problematičtější, neboť v něm stále figuruje souřadnice  $Y_1^*$ . Prozatím ji ponechme tak, jak je, tj. jako souřadnici průmětu  $\mathbf{Y}$  do jednorozměrného podprostoru  $A$  vzhledem k nějakému jednotkovému vektoru, který  $A$  generuje.

### Zobecnění

Závěrem se podívejme na obecnější situaci, kdy  $\mathbf{Y} \sim N(0, \sigma^2)$ . Tento případ můžeme snadno převést na předchozí, neboť platí  $\mathbf{Y}/\sigma \sim N(0, 1)$ . Ve vztazích (1.22), (1.23) a (1.24) tedy stačí vydělit všechny vektory, resp. jejich souřadnice, hodnotou  $\sigma$ , která se ovšem ve druhém a třetím případě zkrátí.

Zrekapitulujme si dosažené výsledky: došli jsme k tomu, že má-li náhodný vektor  $\mathbf{Y}$  rozdělení  $N(0, \sigma^2)$  a  $A$  je podprostor vektorového prostoru  $V_n$  dimenze  $a$ , platí

$$\frac{\|\mathbf{Y}_A\|^2}{\sigma^2} \sim \chi_a^2. \quad (1.25)$$

Je-li dále  $B$  podprostor vektorového prostoru  $V_n$  dimenze  $b$  kolmý na  $A$ , dostáváme, že

$$\frac{\|\mathbf{Y}_A\|^2 / a}{\|\mathbf{Y}_B\|^2 / b} \sim F_{a,b}. \quad (1.26)$$

Je-li konečně podprostor  $A$  jednorozměrný a  $Y^*$  je souřadnice průmětu  $Y_A$  vzhledem k nějakému jednotkovému vektoru, který  $A$  generuje, platí

$$\frac{Y^*}{\|Y_B\|/\sqrt{b}} \sim t_b. \quad (1.27)$$

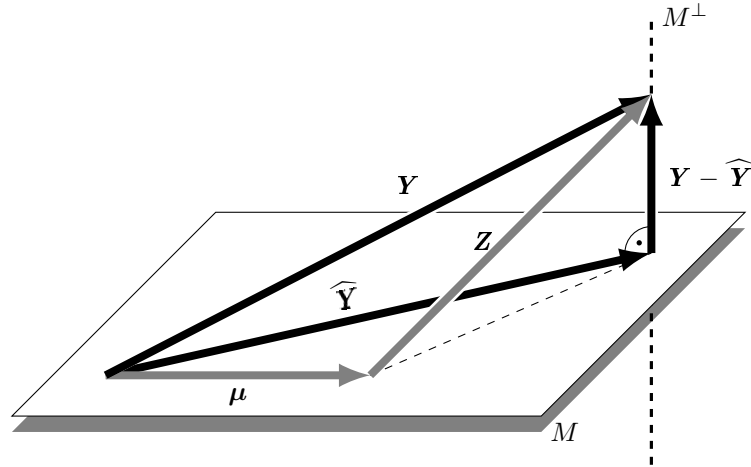
## 1.5 Odhad rozptylu

Je-li  $\widehat{Y}$  pravoúhlým průmětem náhodného vektoru  $Y$  do podprostoru  $M$ , je zároveň  $Y - \widehat{Y}$  pravoúhlým průmětem  $Y$  do *ortogonálního doplňku* podprostoru  $M$ , tj. do podprostoru  $M^\perp \subset V_n$ , definovaného rovností

$$M^\perp \equiv \{x \in V_n : x \perp M\}.$$

Dimenze podprostoru  $M^\perp$  je  $n - m$ .

Pro naše potřeby je však důležitější skutečnost, že řídí-li se náhodný vektor modelem (1.10), je  $Y - \widehat{Y}$  také pravoúhlým průmětem náhodného vektoru  $Z = Y - \mu$  do podprostoru  $M^\perp$  (viz obr. 1.11). Jednak totiž z definice průmětu  $\widehat{Y}$  leží v podprostoru  $M^\perp$ , jednak platí



**Obrázek 1.11:** Náhodný vektor  $Y - \widehat{Y}$  je pravoúhlým průmětem náhodných vektorů  $Y$  a  $Z = Y - \mu$  do ortogonálního doplňku  $M$ , tj. do podprostoru  $M^\perp$ , jehož dimenze je  $n - m$ .

$$\begin{aligned} Z - (Y - \widehat{Y}) &= (Y - \mu) - (Y - \widehat{Y}) = \\ &= \widehat{Y} - \mu. \end{aligned}$$

Jelikož oba vektory  $\widehat{Y}$ ,  $\mu$  leží v podprostoru  $M$ , leží v něm i vektor  $Z - (Y - \widehat{Y})$ , který je proto kolmý na podprostor  $M^\perp$ .

Protože náhodný vektor  $Z$  má rozdělení  $N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ , dostáváme podle (1.25) tvrzení

$$\frac{\|Y - \widehat{Y}\|^2}{\sigma^2} \sim \chi_{n-m}^2. \quad (1.28)$$

Toho můžeme využít v první řadě k bodovému odhadu neznámé hodnoty  $\sigma^2$ . Střední hodnota rozdělení  $\chi_k^2$  je totiž  $k$ , a proto platí

$$\mathbb{E} \left[ \frac{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2}{\sigma^2} \right] = n - m,$$

neboli

$$\mathbb{E} \left[ \frac{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2}{n - m} \right] = \sigma^2.$$

Hodnota

$$S^2 \equiv \frac{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2}{n - m},$$

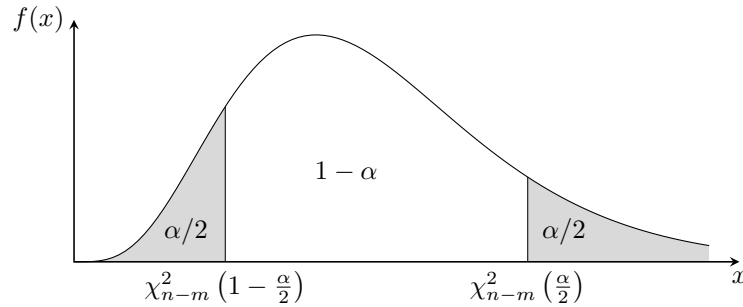
nazývaná *reziduální rozptyl*, je tedy nestranným odhadem rozptylu  $\sigma^2$ . (To mimo-  
chodem platí i bez předpokladu normality – postačuje skutečnost, že varianční  
matice je  $\sigma^2 \mathbf{I}_n$ .)

Poznamenejme ještě, že hodnota  $\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2$  se nazývá *reziduální součet čtverců*  
a k jejímu výpočtu se často užívá Pythagorova věta:

$$\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2 = \|\mathbf{Y}\|^2 - \|\widehat{\mathbf{Y}}\|^2. \quad (1.29)$$

### Intervaly spolehlivosti

Vzhledem k (1.28) lze pro  $\alpha \in (0, 1)$  psát



**Obrázek 1.12:** Hustota  $f$  a kritické hodnoty rozdělení  $\chi_{n-m}^2$ , použité k určení oboustranného  
( $1 - \alpha$ )% intervalu spolehlivosti pro neznámý parametr  $\sigma^2$ .

$$\mathbb{P} \left[ \chi_{n-m}^2 \left( 1 - \frac{\alpha}{2} \right) < \frac{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2}{\sigma^2} < \chi_{n-m}^2 \left( \frac{\alpha}{2} \right) \right] = 1 - \alpha \quad (1.30)$$

(viz obr. 1.12), odkud snadnou úpravou dostaneme, že jev

$$\sigma^2 \in \left( \frac{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2}{\chi_{n-m}^2(\alpha/2)}; \frac{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2}{\chi_{n-m}^2(1 - \alpha/2)} \right),$$

resp.

$$\sigma^2 \in \left( \frac{S^2(n-m)}{\chi_{n-m}^2(\alpha/2)}; \frac{S^2(n-m)}{\chi_{n-m}^2(1-\alpha/2)} \right),$$

nastane s pravděpodobností  $1 - \alpha$ .

Podobným způsobem můžeme z rovností

$$\mathbb{P} \left[ \frac{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2}{\sigma^2} < \chi_{n-m}^2(\alpha) \right] = 1 - \alpha, \quad (1.31)$$

resp.

$$\mathbb{P} \left[ \chi_{n-m}^2(1-\alpha) < \frac{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2}{\sigma^2} \right] = 1 - \alpha, \quad (1.32)$$

odvodit jednostranné intervaly spolehlivosti: pravděpodobnost jevu

$$\sigma^2 \in \left( \frac{S^2(n-m)}{\chi_{n-m}^2(\alpha)}; \infty \right),$$

resp.

$$\sigma^2 \in \left( -\infty; \frac{S^2(n-m)}{\chi_{n-m}^2(1-\alpha)} \right),$$

je  $1 - \alpha$ . Obvyklou praxí je volba  $\alpha = 0,05$ , vedoucí k oboustrannému či jednostrannému 95% intervalu spolehlivosti.

### Test hypotézy $\sigma^2 = \sigma_0^2$

Podle rovností (1.30), resp. (1.31), resp. (1.32), zamítneme nulovou hypotézu  $\sigma^2 = \sigma_0^2$  ve prospěch alternativní hypotézy  $\sigma^2 \neq \sigma_0^2$ , resp.  $\sigma^2 > \sigma_0^2$ , resp.  $\sigma^2 < \sigma_0^2$ , na  $\alpha\%$  hladině významnosti tehdy, když nebude splněna podmínka

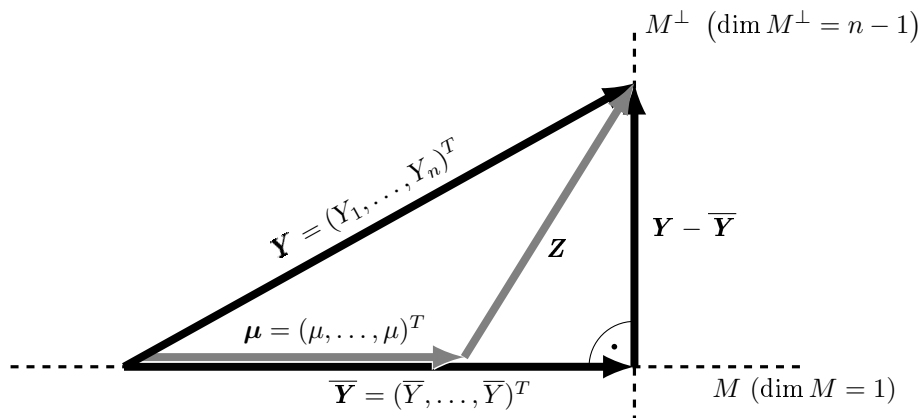
$$\chi_{n-m}^2 \left( 1 - \frac{\alpha}{2} \right) \leq \frac{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2}{\sigma_0^2} \leq \chi_{n-m}^2 \left( \frac{\alpha}{2} \right),$$

resp.

$$\frac{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2}{\sigma_0^2} \leq \chi_{n-m}^2(\alpha),$$

resp.

$$\chi_{n-m}^2(1-\alpha) \leq \frac{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2}{\sigma_0^2}.$$



**Obrázek 1.13:** Náhodný vektor  $\mathbf{Y} - \bar{\mathbf{Y}} = (Y_1 - \bar{Y}, \dots, Y_n - \bar{Y})^T$  je pravoúhlým průmětem náhodných vektorů  $\mathbf{Y}$  a  $\mathbf{Z} = (Y_1 - \mu, \dots, Y_n - \mu)^T$  do podprostoru  $M^\perp$  dimenze  $n - 1$ .

## Příklady

### 1.5.1 Výběr z normálního rozdělení (pokračování ze str. 17)

Náhodný vektor  $\mathbf{Y} - \bar{\mathbf{Y}}$  je pravoúhlým průmětem náhodného vektoru  $\mathbf{Z}$  do podprostoru  $M^\perp$  dimenze  $n - 1$  (viz obr. 1.13).

Rozdělení vektoru  $\mathbf{Z}$  je  $\mathbf{N}(\mathbf{0}, \sigma^2)$ , podle (1.28) proto platí

$$\frac{\|\mathbf{Y} - \bar{\mathbf{Y}}\|^2}{\sigma^2} \sim \chi_{n-1}^2.$$

Náhodná veličina

$$S^2 = \frac{\|\mathbf{Y} - \bar{\mathbf{Y}}\|^2}{n - 1},$$

nazývaná *výběrový rozptyl*, je tedy nestranným odhadem neznámé hodnoty  $\sigma^2$ .

V našem konkrétním případě, kdy platí  $\dim M^\perp = 4$  a získali jsme realizace

$$\begin{aligned} \mathbf{y} &= (4, 6, 10, 5, 5)^T, \\ \bar{\mathbf{y}} &= (6, 6, 6, 6, 6)^T, \end{aligned} \quad (1.33)$$

dostáváme odhad neznámé hodnoty  $\sigma^2$

$$\begin{aligned} s^2 &= \frac{\|(-2, 0, 4, -1, -1)^T\|^2}{4} = \\ &= \frac{22}{4} = 5,5. \end{aligned}$$

V obecném případě se však k výpočtu  $S^2$  zpravidla používá vzorec

$$S^2 = \frac{\sum_{i=1}^n Y_i^2 - n\bar{Y}^2}{n - 1},$$

plynoucí z Pythagorovy věty:

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \|\mathbf{Y} - \bar{\mathbf{Y}}\|^2 = \|\mathbf{Y}\|^2 - \|\bar{\mathbf{Y}}\|^2 = \sum_{i=1}^n Y_i^2 - n\bar{Y}^2.$$

### 1.5.2 Jednoduché třídění (pokračování ze str. 18)

Náhodný vektor  $\mathbf{Y} - \widehat{\mathbf{Y}}$  je pravoúhlým průmětem náhodného vektoru  $\mathbf{Z}$  do podprostoru  $M^\perp$ . Jelikož je  $n = 7$  a  $\dim M = 3$ , je  $\dim M^\perp = 4$ . Platí tedy

$$\frac{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2}{\sigma^2} \sim \chi_4^2$$

a k odhadu  $\sigma^2$  použijeme statistiku

$$S^2 = \frac{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2}{4}.$$

Jelikož jsme získali realizace

$$\begin{aligned} \mathbf{y} &= (24, 26, 23, 19, 33, 28, 29)^T, \\ \widehat{\mathbf{y}} &= (25, 25, 21, 21, 30, 30, 30)^T, \end{aligned}$$

dostáváme hodnotu

$$\begin{aligned} s^2 &= \frac{\|(-1, 1, 2, -2, 3, -2, -1)^T\|^2}{4} = \\ &= \frac{1 + 1 + 4 + 4 + 9 + 4 + 1}{4} = 6. \end{aligned}$$

Také zde jsme k výpočtu čitatele mohli použít Pythagorovu větu

$$\begin{aligned} \|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2 &= \|\mathbf{Y}\|^2 - \|\widehat{\mathbf{Y}}\|^2 = \\ &= \sum_{i=1}^7 Y_i^2 - (2\bar{Y}_f^2 + 2\bar{Y}_g^2 + 3\bar{Y}_h^2). \end{aligned}$$

### 1.5.3 Regresní přímka (pokračování ze str. 19)

Jelikož je  $n = 7$  a  $\dim M = 2$ , je  $\dim M^\perp = 5$ . Platí tedy

$$\frac{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2}{\sigma^2} \sim \chi_5^2$$

a nestranným odhadem  $\sigma^2$  je statistika

$$S^2 = \frac{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2}{5}.$$

V našem konkrétním případě jsme získali realizace

$$\mathbf{y} = \begin{pmatrix} 51 \\ 42 \\ 49 \\ 44 \\ 59 \\ 49 \\ 56 \end{pmatrix}, \quad \widehat{\mathbf{y}} = \begin{pmatrix} 47 \\ 44 \\ 50 \\ 47 \\ 53 \\ 50 \\ 59 \end{pmatrix}, \quad \text{tj. } \mathbf{y} - \widehat{\mathbf{y}} = \begin{pmatrix} 4 \\ -2 \\ -1 \\ -3 \\ 6 \\ -1 \\ -3 \end{pmatrix},$$

a dostáváme tak pro odhad  $\sigma^2$  hodnotu

$$s^2 = \frac{4^2 + 2^2 + 1^2 + 3^2 + 6^2 + 1^2 + 3^2}{5} = 15,2.$$

Podobně jako v předchozích příkladech jsme hodnotu čitatele mohli vypočítat pomocí Pythagorovy věty

$$\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2 = \|\mathbf{Y}\|^2 - \|\widehat{\mathbf{Y}}\|^2 = \sum_{i=1}^7 Y_i^2 - \sum_{i=1}^7 \widehat{Y}_i^2,$$

kde  $\widehat{Y}_i$  jsou souřadnice náhodného vektoru  $\widehat{\mathbf{Y}}$ .

### Obecná formulace

V obecném případě je  $\dim M^\perp = n - 2$ , nestranným odhadem rozptylu je tudíž náhodná veličina

$$S^2 = \frac{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2}{n - 2} = \frac{\sum_{i=1}^n Y_i^2 - \sum_{i=1}^n \widehat{Y}_i^2}{n - 2}.$$

## 1.6 Submodel

Často je třeba zaujmout stanovisko k otázce, zdali nelze model (1.10) nahradit modelem přesnějším, tzv. *submodelem*. Upřesnění spočívá v tzv. *nulové hypotéze*, což je tvrzení, že střední hodnota  $\boldsymbol{\mu}$  náhodného vektoru  $\mathbf{Y}$  neleží v podprostoru  $M$  kdekoli, nýbrž v nějakém jeho podprostoru  $S \subset M$  dimenze  $s < m$ . Necht'  $\mathbf{S}$  je matice typu  $n \times s$ , jejíž lineárně nezávislé sloupce generují  $S$ ; submodel pak můžeme zapsat jako předpoklad, že platí

$$\mathbf{Y} = \mathbf{S}\boldsymbol{\alpha} + \mathbf{Z},$$

kde  $\mathbf{Z} \sim \mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$  a  $\boldsymbol{\alpha}$  je  $s$ -tice neznámých parametrů. Sloupce matice  $\mathbf{S}$  musí být samozřejmě prvky podprostoru  $M$ .

Ze stejných důvodů jako v případě původního modelu je v této situaci nejrozměnějším krokem odhadnout střední hodnotu  $\boldsymbol{\mu}$  pomocí pravoúhlého průmětu  $\mathbf{Y}$  do  $S$ . Označme tento průmět  $\mathbf{Y}_S$  (viz obr. 1.14); určíme jej podobně, jako jsme určili  $\widehat{\mathbf{Y}}$  (viz str. 15).

### 1.6.1 Náhodný vektor $\widehat{\mathbf{Y}} - \mathbf{Y}_S$

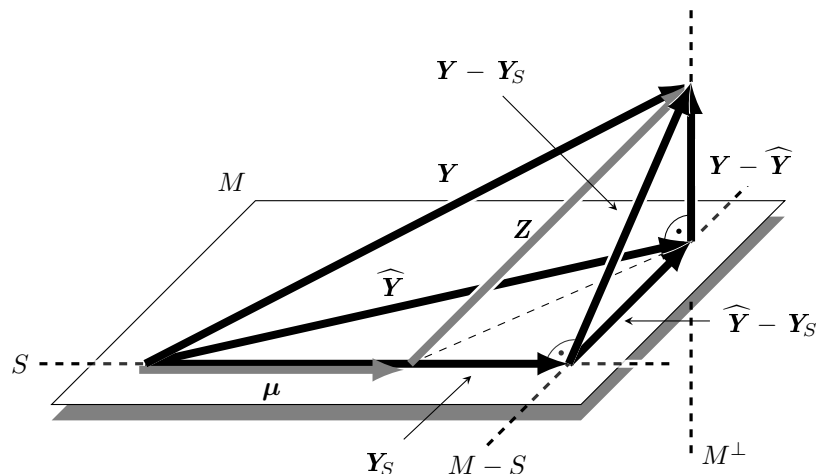
Nyní je důležité si uvědomit, že pokud submodel skutečně platí, je rozdíl  $\widehat{\mathbf{Y}} - \mathbf{Y}_S$  pravoúhlým průmětem náhodného vektoru  $\mathbf{Z}$  do podprostoru  $M - S$ , tj. podprostoru definovaného rovností

$$M - S = \{\mathbf{x} \in M : \mathbf{x} \perp S\}.$$

K důkazu potřebujeme ověřit, že platí

$$\widehat{\mathbf{Y}} - \mathbf{Y}_S \in M - S, \tag{1.34}$$

$$\mathbf{Z} - (\widehat{\mathbf{Y}} - \mathbf{Y}_S) \perp M - S. \tag{1.35}$$



**Obrázek 1.14:** V případě, že platí nulová hypotéza  $\mu \in S$ , je náhodný vektor  $\widehat{\mathbf{Y}} - \mathbf{Y}_S$  pravoúhlým průmětem náhodného vektoru  $\mathbf{Z} = \mathbf{Y} - \mu$  do podprostoru  $M - S$ , tj. do ortogonálního doplňku  $S$  v rámci  $M$ .

Ukažme nejprve, že je splněna první podmínka. Vektor  $\widehat{\mathbf{Y}} - \mathbf{Y}_S$  je roven rozdílu vektorů  $\mathbf{Y} - \mathbf{Y}_S$  a  $\mathbf{Y} - \widehat{\mathbf{Y}}$ . První z nich je z definice kolmý na podprostor  $S$ , druhý na podprostor  $M$ , a tím pádem i na  $S$ , jelikož  $S \subset M$ . Vektor  $\widehat{\mathbf{Y}} - \mathbf{Y}_S$  je tedy také kolmý na podprostor  $S$ . Protože je zároveň zřejmě prvkem podprostoru  $M$ , platí podmínka (1.34).

Dále lze psát

$$\begin{aligned} \mathbf{Z} - (\widehat{\mathbf{Y}} - \mathbf{Y}_S) &= (\mathbf{Y} - \mu) - (\widehat{\mathbf{Y}} - \mathbf{Y}_S) = \\ &= (\mathbf{Y}_S - \mu) + (\mathbf{Y} - \widehat{\mathbf{Y}}); \end{aligned}$$

první ze sčítanců leží v podprostoru  $S$  (zde přichází ke slovu náš předpoklad, že  $\mu \in S$ ), a je tudíž kolmý na podprostor  $M - S$ , druhý z nich je z definice kolmý na  $M$ , a tedy i na  $M - S$ . Je tedy splněna i podmínka (1.35).

## 1.6.2 Test submodelu

Jak toho můžeme využít k ověření nulové hypotézy? Víme, že náhodný vektor  $\mathbf{Z}$  má rozdělení  $\mathbf{N}(\mathbf{0}; \sigma^2 \mathbf{I}_n)$ , jeho pravoúhlým průmětem do podprostoru  $M^\perp$  dimenze  $n - m$  je náhodný vektor  $\mathbf{Y} - \widehat{\mathbf{Y}}$ , jeho pravoúhlým průmětem do podprostoru  $M - S$  dimenze  $m - s$  je (za platnosti nulové hypotézy) náhodný vektor  $\widehat{\mathbf{Y}} - \mathbf{Y}_S$ , a podprostory  $M^\perp$  a  $M - S$  jsou navzájem kolmé; podle (1.26) tedy platí

$$F \equiv \frac{\|\widehat{\mathbf{Y}} - \mathbf{Y}_S\|^2 / (m - s)}{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2 / (n - m)} \sim F_{m-s, n-m}. \quad (1.36)$$

Je zřejmé, že nulovou hypotézu nebudeme považovat za věrohodnou tehdy, když bude odhad  $\mathbf{Y}_S$  od původního odhadu  $\widehat{\mathbf{Y}}$  „příliš daleko“ ve srovnání s hodnotou  $\|\mathbf{Y} - \widehat{\mathbf{Y}}\|$ , tj. když bude podíl na levé straně vztahu (1.36) příliš velký. Nulovou hypotézu tedy zamítneme na  $\alpha\%$  hladině významnosti v tom případě, když

nastane nerovnost

$$\frac{\|\widehat{\mathbf{Y}} - \mathbf{Y}_S\|^2 / (m - s)}{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2 / (n - m)} > F_{m-s, n-m}(\alpha). \quad (1.37)$$

Alternativní hypotézou je v tomto testu zřejmě  $\boldsymbol{\mu} \notin S$ .

Poznamenejme, že jmenovatel zlomku ve vztahu (1.36) představuje veličinu  $S^2$ , kterou jsme zavedli v kapitole 1.5 jako nestranný odhad rozptylu  $\sigma^2$ .

### 1.6.3 Užití Pythagorovy věty

Vytvořením pravoúhlého průmětu  $\mathbf{Y}_S$  se značně rozšířily naše možnosti ohledně využití Pythagorovy věty. V první řadě plyne z definice vektoru  $\mathbf{Y}_S$  vztah  $\mathbf{Y} - \mathbf{Y}_S \perp \mathbf{Y}_S$ ; tyto dva vektory tedy tvoří odvěsny pravoúhlého trojúhelníku, jehož přeponou je jejich součet  $\mathbf{Y}$ . Z toho plyne rovnost

$$\|\mathbf{Y} - \mathbf{Y}_S\|^2 + \|\mathbf{Y}_S\|^2 = \|\mathbf{Y}\|^2. \quad (1.38)$$

Podobně jsou na sebe kolmé vektory  $\widehat{\mathbf{Y}} - \mathbf{Y}_S$  a  $\mathbf{Y} - \widehat{\mathbf{Y}}$ , neboť první z nich leží v podprostoru  $M$  a druhý je na něj kolmý. Přeponou je jejich součet  $\mathbf{Y} - \mathbf{Y}_S$ :

$$\|\mathbf{Y} - \mathbf{Y}_S\|^2 = \|\widehat{\mathbf{Y}} - \mathbf{Y}_S\|^2 + \|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2. \quad (1.39)$$

Ukázali jsme, že vektor  $\widehat{\mathbf{Y}} - \mathbf{Y}_S$  je kolmý na podprostor  $S$ , je tedy kolmý také na vektor  $\mathbf{Y}_S$ . Součtem těchto vektorů je vektor  $\widehat{\mathbf{Y}}$ . Z toho vyplývá další vztah:

$$\|\widehat{\mathbf{Y}} - \mathbf{Y}_S\|^2 + \|\mathbf{Y}_S\|^2 = \|\widehat{\mathbf{Y}}\|^2.$$

Ten jsme ovšem mohli také odvodit z rovností (1.38) a (1.39) spolu s (1.29). Ze vztahů (1.38) a (1.39) plyne ještě vzorec

$$\|\mathbf{Y}\|^2 = \|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2 + \|\widehat{\mathbf{Y}} - \mathbf{Y}_S\|^2 + \|\mathbf{Y}_S\|^2,$$

ilustrující rozklad náhodného vektoru  $\mathbf{Y}$  do tří navzájem kolmých podprostorů  $M^\perp$ ,  $M - S$  a  $S$ .

### 1.6.4 Tabulka analýzy rozptylu

Test založený na statistice  $F$  ze vztahu (1.36) je známý pod názvem *analýza rozptylu* a pod zkratkou ANOVA (z anglického *analysis of variance*). Název je odvozen ze skutečnosti, že – jak je snad zřejmé z kapitol 1.4 a 1.5 – v případě platnosti nulové hypotézy představují čitatel i jmenovatel této statistiky dva nezávislé odhady rozptylu  $\sigma^2$ . Výsledky testu bývají zpravidla uvedeny v tabulce, v níž figurují parametry pravoúhlého trojúhelníku tvořeného odvěsnami  $\widehat{\mathbf{Y}} - \mathbf{Y}_S$ ,  $\mathbf{Y} - \widehat{\mathbf{Y}}$  a přeponou  $\mathbf{Y} - \mathbf{Y}_S$  (viz tab. 1.1).

Zdroj variability	Součet čtverců	Počet stupňů volnosti	Podíl	Testová statistika $F$
Odstraněný vliv	$\ \widehat{\mathbf{Y}} - \mathbf{Y}_S\ ^2$	$m - s$	$\frac{\ \widehat{\mathbf{Y}} - \mathbf{Y}_S\ ^2}{m - s}$	$\frac{\ \widehat{\mathbf{Y}} - \mathbf{Y}_S\ ^2}{m - s}$
Reziduální	$\ \mathbf{Y} - \widehat{\mathbf{Y}}\ ^2$	$n - m$	$\frac{\ \mathbf{Y} - \widehat{\mathbf{Y}}\ ^2}{n - m}$	$\frac{\ \mathbf{Y} - \widehat{\mathbf{Y}}\ ^2}{n - m}$
Celkový	$\ \mathbf{Y} - \mathbf{Y}_S\ ^2$	$n - s$	–	–

**Tabulka 1.1:** Tabulka analýzy rozptylu. Místo názvu „odstraněný vliv“ se zpravidla používá konkrétní popis vlivu, který by byl redukcí modelu na submodel odstraněn. Součet čtverců představuje druhou mocninu délky pravoúhlého průmětu a počet stupňů volnosti je dimenze podprostoru, do kterého se promítá. Hodnoty v posledním řádku jsou součtem hodnot z předchozích řádků – viz Pythagorova věta (1.39). Za platnosti nulové hypotézy představují hodnoty ze sloupce označeného „podíl“ nezávislé odhady rozptylu  $\sigma^2$  a hodnota  $F$  má rozdělení  $F_{m-s, n-m}$ .

## Příklady

### 1.6.5 Výběr z normálního rozdělení (pokračování ze str. 26)

Zkusme navrhnout test hypotézy  $\mu = 0$ . Chceme tedy rozhodnout, zda je přijatelná redukce původního jednorozměrného podprostoru  $M = [e]$  na triviální vektorový podprostor  $S \equiv \{\mathbf{0}\}$  dimenze 0. Pokud nulová hypotéza skutečně platí, řídí se náhodný vektor  $\mathbf{Y}$  rozdělením  $N(\mathbf{0}; \sigma^2 \mathbf{I}_n)$ . Jeho pravoúhlými průměty do navzájem kolmých podprostorů  $M$  a  $M^\perp$  jsou vektory  $\overline{\mathbf{Y}}$  a  $\mathbf{Y} - \overline{\mathbf{Y}}$  (viz obr. 1.6). Dimenze těchto podprostorů jsou 1 a  $n - 1$ . Platí tedy, že náhodná veličina

$$\frac{\|\overline{\mathbf{Y}}\|^2/1}{\|\mathbf{Y} - \overline{\mathbf{Y}}\|^2/(n-1)} = \frac{n\overline{Y}^2}{S^2}$$

má rozdělení  $F_{1, n-1}$ . Pokud tedy pro její konkrétní realizaci nastane nerovnost

$$\frac{n\overline{Y}^2}{S^2} \geq F_{1, n-1}(\alpha),$$

zamítneme nulovou hypotézu na hladině  $\alpha$ .

V našem konkrétním případě, kdy jsme měřili hodnoty u pěti pacientů a získali jsme realizace (1.33), dostáváme

$$\frac{\|\overline{\mathbf{y}}\|^2/1}{\|\mathbf{y} - \overline{\mathbf{y}}\|^2/(n-1)} = \frac{180}{22/4} = 32,7.$$

Protože kritická hodnota je  $F_{1,4}(0,05) = 7,71$ , můžeme nulovou hypotézu  $\mu = 0$  na 5% hladině významnosti zamítnout.

### 1.6.6 Jednoduché třídění (pokračování ze str. 27)

Testujme nulovou hypotézu  $\mu_f = \mu_g = \mu_h = \mu_0$ . Budeme tedy uvažovat o možnosti redukce původního modelu (1.12) na model

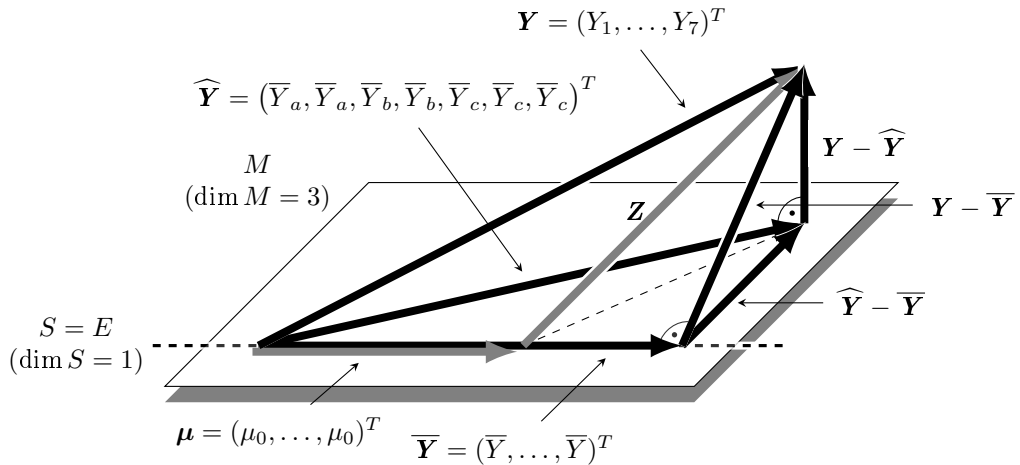
$$\mathbf{Y} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \cdot \mu_0 + \begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \\ Z_5 \\ Z_6 \\ Z_7 \end{pmatrix} \equiv \mathbf{e} \cdot \mu_0 + \mathbf{Z}.$$

To znamená, že původní trojrozměrný podprostor  $M$ , generovaný vektory  $\mathbf{a}_f$ ,  $\mathbf{a}_g$ ,  $\mathbf{a}_h$ , nahradíme podprostorem  $S = E \subset M$ , generovaným vektorem  $\mathbf{e}$ . Tento vektor je lineární kombinací sloupců matice  $\mathbf{X}$  – konkrétně jejich součtem – takže skutečně generuje podprostor podprostoru  $M$ .

Promítneme tedy náhodný vektor  $\mathbf{Y}$  do podprostoru  $E$ , a jak už víme z příkladu 1.3.4 v kapitole 1.3, tímto průmětem je náhodný vektor

$$\bar{\mathbf{Y}} = (\bar{Y}, \dots, \bar{Y})^T$$

(viz obr. 1.15). Vektor  $\widehat{\mathbf{Y}} - \bar{\mathbf{Y}}$  je pak za platnosti nulové hypotézy pravoúhlým



**Obrázek 1.15:** Vektory figurující v případě jednoduchého třídění (1.12), kdy testujeme nulovou hypotézu  $\mu_a = \mu_b = \mu_c \equiv \mu_0$ , tj.  $\boldsymbol{\mu} = (\mu_0, \dots, \mu_0)^T$ , neboli  $\boldsymbol{\mu} \in S$ , kde  $S = E$  je jednorozměrný podprostor generovaný vektorem  $\mathbf{e} = (1, \dots, 1)^T$ .

průmětem vektoru  $\mathbf{Z} = \mathbf{Y} - \boldsymbol{\mu}$  do podprostoru  $M - E$ , jehož dimenze je

$$\dim M - \dim E = 3 - 1 = 2.$$

Vektor  $\mathbf{Y} - \widehat{\mathbf{Y}}$  je naproti tomu průmětem vektoru  $\widehat{\mathbf{Y}} - \bar{\mathbf{Y}}$  do podprostoru  $M^\perp$  dimenze

$$n - \dim M = 7 - 3 = 4.$$

Ve shodě s (1.36) tedy můžeme tvrdit, že za předpokladu platnosti nulové hypotézy platí

$$\frac{\|\widehat{\mathbf{Y}} - \overline{\mathbf{Y}}\|^2 / 2}{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2 / 4} \sim F_{2,4},$$

takže nulovou hypotézu zamítneme na  $\alpha\%$  hladině významnosti tehdy, když nastane nerovnost

$$\frac{\|\widehat{\mathbf{Y}} - \overline{\mathbf{Y}}\|^2 / 2}{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2 / 4} \geq F_{2,4}(\alpha).$$

Ukažme si vše ještě jednou na konkrétních hodnotách. Průměr ze získaných měření je roven 26; realizace vektorů, které potřebujeme k výpočtu statistiky  $F$ , jsou tedy

$$\mathbf{y} = \begin{pmatrix} 24 \\ 26 \\ 23 \\ 19 \\ 33 \\ 28 \\ 29 \end{pmatrix}, \quad \widehat{\mathbf{y}} = \begin{pmatrix} 25 \\ 25 \\ 21 \\ 21 \\ 30 \\ 30 \\ 30 \end{pmatrix}, \quad \overline{\mathbf{y}} = \begin{pmatrix} 26 \\ 26 \\ 26 \\ 26 \\ 26 \\ 26 \\ 26 \end{pmatrix},$$

z čehož vypočteme

$$\mathbf{y} - \widehat{\mathbf{y}} = \begin{pmatrix} -1 \\ 1 \\ 2 \\ -2 \\ 3 \\ -2 \\ -1 \end{pmatrix}, \quad \widehat{\mathbf{y}} - \overline{\mathbf{y}} = \begin{pmatrix} -1 \\ -1 \\ -5 \\ -5 \\ 4 \\ 4 \\ 4 \end{pmatrix},$$

testová statistika je tedy

$$F = \frac{(1 + 1 + 25 + 25 + 16 + 16 + 16)/2}{(1 + 1 + 4 + 4 + 9 + 4 + 1)/4} = 8,33.$$

Jelikož kritická hodnota příslušného rozdělení  $F$  je

$$F_{2,4}(0,05) = 6,94,$$

můžeme nulovou hypotézu na 5% hladině zamítnout.

### 1.6.7 Jednoduché třídění – obecná formulace

Připomeňme nyní značení používané v obecném případě, kdy vyšetřujeme vliv  $I$  různých typů ošetření na  $n$  různých jednotkách, přičemž každou jednotku ošetříme právě jedním typem a  $i$ -tý typ ošetření použijeme celkem u  $n_i$  jednotek

(tj.  $n_1 + \dots + n_I = n$ ). Reakce jednotlivých jednotek, tj. souřadnice náhodného vektoru  $\mathbf{Y}$ , se obvykle značí symboly  $Y_{ij}$ , kde  $i \in \{1, \dots, I\}$ ,  $j \in \{1, \dots, n_i\}$ . Index  $i$  tedy označuje číslo typu ošetření, index  $j$  číslo jednotky ve skupině, u které bylo toto ošetření aplikováno. Předpokládáme, že platí

$$Y_{ij} \sim N(\mu_i, \sigma^2). \quad (1.40)$$

Model charakterizující náhodný vektor  $\mathbf{Y}$  má tedy tvar

$$\mathbf{Y} = \mu_1 \mathbf{a}_1 + \dots + \mu_I \mathbf{a}_I + \mathbf{Z}, \quad (1.41)$$

kde náhodný vektor  $\mathbf{Z}$  má rozdělení  $N(\mathbf{0}; \sigma^2 \mathbf{I}_n)$ . Každý z vektorů  $\mathbf{a}_i$  odpovídá jednomu typu ošetření, přičemž každá z jeho  $n$  souřadnic odpovídá jedné pokusné jednotce a je rovna jedné nebo nule podle toho, zda daný typ byl či nebyl u této jednotky použit. Vektory  $\mathbf{a}_i$  jsou lineárně nezávislé, proto generují podprostor  $M$  dimenze  $I$ . Průmětem vektoru  $\mathbf{Y}$  do podprostoru  $M$  je pak vektor  $\widehat{\mathbf{Y}}$  o souřadnicích  $\widehat{Y}_{ij}$ , kde

$$\widehat{Y}_{ij} = \bar{Y}_i = \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i}.$$

Vektor  $\widehat{\mathbf{Y}}$  tedy vznikne z vektoru  $\mathbf{Y}$  tak, že u každé jednotky nahradíme naměřenou hodnotu průměrem hodnot získaných od všech jednotek, u kterých byl použit stejný typ ošetření.

Vektor  $\mathbf{Y} - \widehat{\mathbf{Y}}$  je pravoúhlým průmětem vektoru  $\mathbf{Y} - \boldsymbol{\mu}$  do podprostoru  $M^\perp$  dimenze  $n - I$ , k odhadu rozptylu  $\sigma^2$  tedy použijeme náhodnou veličinu

$$\begin{aligned} S^2 &= \frac{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2}{n - I} = \\ &= \frac{\sum_{i=1}^I \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2}{n - I}. \end{aligned}$$

K výpočtu čitatele lze využít Pythagorovy věty:

$$\begin{aligned} \|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2 &= \|\mathbf{Y}\|^2 - \|\widehat{\mathbf{Y}}\|^2 = \\ &= \sum_{i=1}^I \sum_{j=1}^{n_i} Y_{ij}^2 - \sum_{i=1}^I n_i \bar{Y}_i^2. \end{aligned}$$

Rozhodneme-li se testovat nulovou hypotézu  $\mu_1 = \dots = \mu_I$ , tj. hypotézu, že  $\boldsymbol{\mu}$  leží v podprostoru  $E$  generovaném vektorem  $\mathbf{e} = (1, \dots, 1)^T$ , promítneme vektor  $\mathbf{Y}$  do podprostoru  $E$  a získáme vektor

$$\bar{\mathbf{Y}} \equiv (\bar{Y}, \dots, \bar{Y})^T,$$

kde

$$\bar{Y} = \frac{\sum_{i=1}^I \sum_{j=1}^{n_i} Y_{ij}}{n}$$

(místo symbolu  $\bar{Y}$  se též někdy používá  $\bar{Y}_.$ ). Protože dimenze podprostoru  $E$  je 1, je za platnosti nulové hypotézy vektor  $\widehat{\mathbf{Y}} - \bar{\mathbf{Y}}$  pravoúhlým průmětem vektoru  $\mathbf{Y} - \boldsymbol{\mu}$  do podprostoru  $M - E$  dimenze  $I - 1$ , a k verifikaci nulové hypotézy tedy použijeme hodnotu náhodné veličiny

$$\begin{aligned} F &= \frac{\|\widehat{\mathbf{Y}} - \bar{\mathbf{Y}}\|^2 / (I - 1)}{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2 / (n - I)} = \\ &= \frac{\sum_{i=1}^I \sum_{j=1}^{n_i} (\bar{Y}_{i.} - \bar{Y}_{i.})^2 / (I - 1)}{\sum_{i=1}^I \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2 / (n - I)}. \end{aligned}$$

K výpočtu čitatele lze opět použít Pythagorovu větu:

$$\begin{aligned} \|\widehat{\mathbf{Y}} - \bar{\mathbf{Y}}\|^2 &= \|\widehat{\mathbf{Y}}\|^2 - \|\bar{\mathbf{Y}}\|^2 = \\ &= \sum_{i=1}^I n_i \bar{Y}_{i.}^2 - n \bar{Y}^2. \end{aligned}$$

Přesáhne-li realizovaná hodnota statistiky  $F$  kritickou hodnotu  $F_{I-1, n-I}$ , můžeme na  $\alpha\%$  hladině nulovou hypotézu zamítnout. Celý postup ještě můžeme zrekapitulovat v tabulce analýzy rozptylu (viz tab. 1.2):

Zdroj variability	Součet čtverců	Počet stupňů volnosti	Podíl	Testová statistika $F$
Vliv ošetření	$\ \widehat{\mathbf{Y}} - \bar{\mathbf{Y}}\ ^2$	$I - 1$	$\frac{\ \widehat{\mathbf{Y}} - \bar{\mathbf{Y}}\ ^2}{I - 1}$	$\frac{\ \widehat{\mathbf{Y}} - \bar{\mathbf{Y}}\ ^2}{I - 1}$
Reziduální	$\ \mathbf{Y} - \widehat{\mathbf{Y}}\ ^2$	$n - I$	$\frac{\ \mathbf{Y} - \widehat{\mathbf{Y}}\ ^2}{n - I}$	$\frac{\ \mathbf{Y} - \widehat{\mathbf{Y}}\ ^2}{n - I}$
Celkový	$\ \mathbf{Y} - \bar{\mathbf{Y}}\ ^2$	$n - 1$	–	–

**Tabulka 1.2:** Tabulka analýzy rozptylu jednoduchého třídění. Za platnosti nulové hypotézy  $\mu_1 = \dots = \mu_I$  má hodnota  $F$  rozdělení  $F_{I-1, n-I}$ .

## 1.6.8 Regresní přímka (pokračování ze str. 27)

Test hypotézy  $\beta_1 = 0$

Testujme nejprve nulovou hypotézu  $\beta_1 = 0$ , tj. hypotézu, že úroda na hmotnosti použitého hnojiva vůbec nezávisí. Platí-li nulová hypotéza, můžeme druhý sloupec matice  $\mathbf{X}$  modelu (1.13) vynechat. To znamená, že uvažujeme o redukci původního

dvojrozměrného modelu na jednorozměrný submodel

$$\mathbf{Y} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \beta_0 + \mathbf{Z} \equiv \mathbf{e}\beta_0 + \mathbf{Z}.$$

Promítneme tedy vektor  $\mathbf{Y}$  do podprostoru  $E$ , generovaného vektorem  $\mathbf{e} = (1, \dots, 1)^T$ . Stejně jako v předchozích příkladech takto získáme vektor  $\widehat{\mathbf{Y}} = (\widehat{Y}, \dots, \widehat{Y})^T$ . Platí-li nulová hypotéza, vektor  $\widehat{\mathbf{Y}} - \overline{\mathbf{Y}}$  je pravoúhlým průmětem vektoru  $\mathbf{Z} = \mathbf{Y} - \boldsymbol{\mu}$  do podprostoru  $M - E$ , jehož dimenze je 1. Již jsme uvedli, že vektor  $\mathbf{Y} - \widehat{\mathbf{Y}}$  je pravoúhlým průmětem vektoru  $\mathbf{Y} - \boldsymbol{\mu}$  do podprostoru  $M^\perp$  dimenze  $7 - 2 = 5$ ; náhodná veličina

$$F = \frac{\|\widehat{\mathbf{Y}} - \overline{\mathbf{Y}}\|^2 / 1}{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2 / 5}$$

má tedy za předpokladu platnosti nulové hypotézy rozdělení  $F_{1,5}$ , takže nulovou hypotézu zamítneme na hladině  $\alpha$  tehdy, když získaná realizace  $F$  přesáhne kritickou hodnotu  $F_{1,5}(\alpha)$ .

Dosaďme nyní získané realizace: kromě vektorů

$$\begin{aligned} \mathbf{y} &= (51, 42, 49, 44, 59, 49, 56)^T, \\ \widehat{\mathbf{y}} &= (47, 44, 50, 47, 53, 50, 59)^T \end{aligned}$$

máme ještě

$$\overline{\mathbf{y}} = (50, 50, 50, 50, 50, 50, 50)^T.$$

Určíme druhé mocniny délek příslušných rozdílů:

$$\begin{aligned} \|\widehat{\mathbf{y}} - \overline{\mathbf{y}}\|^2 &= 144, \\ \|\mathbf{y} - \widehat{\mathbf{y}}\|^2 &= 76, \end{aligned}$$

a vypočteme hodnotu statistiky  $F$ :

$$F = \frac{144/1}{76/5} = 9,47.$$

Protože  $F_{1,5}(0,05) = 6,61$ , můžeme nulovou hypotézu zamítnout.

### Test hypotézy $\beta_0 = 0$

Podobně můžeme testovat nulovou hypotézu  $\beta_0 = 0$ . Protože v takovém případě uvažujeme o redukci dvojrozměrného podprostoru  $M$  na jednorozměrný podprostor  $S$  generovaný vektorem

$$\mathbf{x} = (3, 2, 4, 3, 5, 4, 7)^T,$$

určíme podle vzorce (2.32) pravoúhlý průmět vektoru  $\mathbf{Y}$  do  $S$ :

$$\mathbf{Y}_S = \frac{\mathbf{Y} \circ \mathbf{x}}{\|\mathbf{x}\|^2} \mathbf{x}. \quad (1.42)$$

Pak již můžeme vyjádřit hodnotu statistiky  $F$ :

$$F = \frac{\|\widehat{\mathbf{Y}} - \mathbf{Y}_S\|^2 / 1}{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2 / 5}, \quad (1.43)$$

která se za předpokladu platnosti nulové hypotézy řídí rozdělením  $F_{1,5}$ .

Jelikož v případě našich konkrétních hodnot je

$$\mathbf{y}_S = \frac{\mathbf{x} \circ \mathbf{y}}{\|\mathbf{x}\|^2} \mathbf{x} = 11,3125 \cdot \begin{pmatrix} 3 \\ 2 \\ 4 \\ 3 \\ 5 \\ 4 \\ 7 \end{pmatrix} \doteq \begin{pmatrix} 33,9 \\ 22,6 \\ 45,3 \\ 33,9 \\ 56,6 \\ 45,3 \\ 79,2 \end{pmatrix},$$

dostáváme hodnotu statistiky

$$F = \frac{1263,5/1}{76/5} \doteq 83,1 > F_{1,5}(0,05) \doteq 6,61,$$

takže i hypotézu  $\beta_0 = 0$  můžeme na 5% hladině významnosti zamítnout.

### Obecná formulace

Pokud chceme testovat hypotézu  $\beta_1 = 0$ , uvažujeme o redukci původního podprostoru  $M$  na podprostor  $E \equiv [e]$ . Odhadem střední hodnoty odpovídající příslušnému submodelu  $\mathbf{Y} = \beta_0 \mathbf{e} + \mathbf{Z}$  je pravoúhlý průmět vektoru  $\mathbf{Y}$  do podprostoru  $E$ , což je vektor  $\overline{\mathbf{Y}} \equiv (\overline{Y}, \dots, \overline{Y})^T$ . Protože dimenze tohoto podprostoru je o 1 nižší než dimenze podprostoru  $M$ , testová statistika bude mít tvar

$$F = \frac{\|\widehat{\mathbf{Y}} - \overline{\mathbf{Y}}\|^2 / 1}{\|\widehat{\mathbf{Y}} - \overline{\mathbf{Y}}\|^2 / (n-2)} = \frac{\|\widehat{\mathbf{Y}}\|^2 - \|\overline{\mathbf{Y}}\|^2}{S^2} = \frac{\sum_{i=1}^n \widehat{Y}_i^2 - n\overline{Y}^2}{S^2}, \quad (1.44)$$

a nulovou hypotézu budeme zamítat na  $\alpha\%$  hladině významnosti, bude-li její realizace větší než hodnota  $F_{1,n-2}(\alpha)$ <sup>5</sup>. Test této hypotézy je velmi používaný a tabulka shrnující jeho výsledky bývá často uváděna statistickým softwarem zcela automaticky, proto ji uvedme i zde (viz tabulka (1.3)).

<sup>5</sup>Místo výše popsaného  $F$ -testu se používá obvykle k verifikaci hypotézy  $\beta_1 = 0$  spíše  $t$ -test. Podobně jako v případě testování hypotézy  $\mu = \mu_0$  v modelu (1.11) však lze ukázat, že jeho výsledky jsou v případě oboustranné alternativy ekvivalentní s výsledky zde uvedeného postupu. Srovnání obou testů v případě modelu (1.11) viz kapitola 1.7.4.

Zdroj variability	Součet čtverců	Počet stupňů volnosti	Podíl	Testová statistika $F$
Lineární koeficient	$\ \widehat{\mathbf{Y}} - \overline{\mathbf{Y}}\ ^2$	1	$\frac{\ \widehat{\mathbf{Y}} - \overline{\mathbf{Y}}\ ^2}{1}$	$\frac{\ \widehat{\mathbf{Y}} - \overline{\mathbf{Y}}\ ^2}{\ \mathbf{Y} - \widehat{\mathbf{Y}}\ ^2}$
Reziduální	$\ \mathbf{Y} - \widehat{\mathbf{Y}}\ ^2$	$n - 2$	$\frac{\ \mathbf{Y} - \widehat{\mathbf{Y}}\ ^2}{n - 2}$	$\frac{\ \mathbf{Y} - \widehat{\mathbf{Y}}\ ^2}{n - 2}$
Celkový	$\ \mathbf{Y} - \overline{\mathbf{Y}}\ ^2$	$n - 1$	–	–

**Tabulka 1.3:** Tabulka analýzy rozptylu odpovídající  $F$ -testu hypotézy  $\beta_1 = 0$  v modelu (1.13). Za platnosti nulové hypotézy má hodnota  $F$  rozdělení  $F_{1,n-2}$ .

V případě hypotézy  $\beta_0 = 0$  redukuje podprostor  $M$  na podprostor  $S = [x]$ . Pravoúhlým průmětem vektoru  $\mathbf{Y}$  do tohoto podprostoru je vektor  $\mathbf{Y}_S$  určený vzorcem (1.42); platí

$$\|\mathbf{Y}_S\|^2 = \left\| \frac{\mathbf{Y} \circ \mathbf{x}}{\|\mathbf{x}\|^2} \mathbf{x} \right\|^2 = \frac{(\mathbf{Y} \circ \mathbf{x})^2}{\|\mathbf{x}\|^4} \|\mathbf{x}\|^2 = \frac{\left( \sum_{i=1}^n x_i Y_i \right)^2}{\sum_{i=1}^n x_i^2},$$

a statistiku (1.43) tedy můžeme explicitně vyjádřit ve tvaru

$$\frac{\|\widehat{\mathbf{Y}} - \mathbf{Y}_S\|^2 / 1}{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2 / (n - 2)} = \frac{\|\widehat{\mathbf{Y}}\|^2 - \|\mathbf{Y}_S\|^2}{S^2} = \frac{\sum_{i=1}^n x_i^2 \sum_{i=1}^n \widehat{Y}_i^2 - \left( \sum_{i=1}^n x_i Y_i \right)^2}{S^2 \sum_{i=1}^n x_i^2}.$$

### 1.6.9 Mnohonásobná regrese

Pro úplnost popište ještě obecný případ mnohonásobné regrese. Necht' pro náhodnou veličinu  $Y$  platí

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + Z,$$

kde  $Z \sim N(0, \sigma^2)$ , hodnoty koeficientů  $x_1, \dots, x_k$  jsou známé a hodnoty ostatních parametrů nikoli. Provedeme-li  $n$  ( $n > k + 1$ ) nezávislých měření této veličiny, při nichž jsou hodnoty známých koeficientů v  $i$ -tém měření rovny  $x_{i,1}, \dots, x_{i,k}$ , získáme realizaci náhodného vektoru  $\mathbf{Y} \equiv (Y_1, \dots, Y_n)^T$ , jehož chování popisuje model

$$\mathbf{Y} = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,k} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \mathbf{Z} \equiv \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}, \quad (1.45)$$

kde náhodný vektor  $\mathbf{Z}$  má rozdělení  $\mathbf{N}(\mathbf{0}; \sigma^2 \mathbf{I})$ .

Předpokládáme-li nezávislost sloupců matice  $\mathbf{X}$ , je dimenze podprostoru  $M$  daného tímto modelem  $k + 1$ . Pravoúhlý průmět  $\widehat{\mathbf{Y}}$  náhodného vektoru  $\mathbf{Y}$  do podprostoru  $M$  a odhady  $\widehat{b}_0, \dots, \widehat{b}_k$  koeficientů  $\beta_0, \dots, \beta_k$  získáme řešením soustavy (1.16). Vektor  $\mathbf{Y} - \widehat{\mathbf{Y}}$  je pak pravoúhlým průmětem vektoru  $\mathbf{Y} - \boldsymbol{\mu}$  do podprostoru  $V_n - M$  dimenze  $n - k - 1$ , nestranný odhad rozptylu tedy bude určen vzorcem

$$S^2 = \frac{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2}{n - k - 1}.$$

**Test hypotézy**  $\beta_1 = \dots = \beta_k = 0$

Tato hypotéza představuje redukci podprostoru  $M$  na podprostor generovaný prvním sloupcem matice  $\mathbf{X}$ , tj. podprostor  $E = [e]$ , jehož dimenze je o  $k$  menší než dimenze  $M$ . Protože pravoúhlým průmětem vektoru  $\mathbf{Y}$  do podprostoru  $E$  je vektor  $\overline{\mathbf{Y}} = (\overline{Y}, \dots, \overline{Y})^T$ , má za předpokladu platnosti nulové hypotézy  $\beta_1 = \dots = \beta_k = 0$  náhodná veličina

$$F = \frac{\|\widehat{\mathbf{Y}} - \overline{\mathbf{Y}}\|^2 / k}{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2 / (n - k - 1)} \quad (1.46)$$

rozdělení  $F_{k, n-k-1}$ . Hypotézu zamítneme tedy na hladině významnosti  $\alpha$  tehdy, když nastane nerovnost  $F > F_{k, n-k-1}(\alpha)$ . Tabulka analýzy rozptylu, shrnující výsledky tohoto testu, bývá statistickým softwarem často uváděna automaticky.

**Test hypotézy**  $\beta_i = 0$

V případě, že uvažujeme o možnosti vypuštění některého z parametrů  $\beta_i$  z modelu, chceme omezit podprostor  $M$  na podprostor  $S_i$ , který je generován sloupci matice  $\mathbf{X}$  s výjimkou  $i$ -tého. Označme pravoúhlý průmět náhodného vektoru  $\mathbf{Y}$  do tohoto podprostoru symbolem  $\widehat{\mathbf{Y}}_i$ ; najdeme jej opět řešením soustavy (1.16), kde ovšem v matici  $\mathbf{X}$  vypustíme dotyčný sloupec. Testová statistika má pak tvar

$$F = \frac{\|\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_i\|^2 / 1}{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2 / (n - k - 1)}$$

a nulovou hypotézu zamítneme na hladině  $\alpha$  v případě splnění nerovnosti  $F > F_{1, n-k-1}(\alpha)$ .

### 1.6.10 Dvě regresní přímky

Předpokládejme, že pro náhodné veličiny  $Y_1, \dots, Y_4$  platí vztah

$$Y_i = \beta_0^a + \beta_1^a \cdot x_i + Z_i$$

a pro náhodné veličiny  $Y_5, \dots, Y_9$  platí vztah

$$Y_i = \beta_0^b + \beta_1^b \cdot x_i + Z_i,$$

kde  $\beta_0^a, \beta_1^a, \beta_0^b, \beta_1^b$  jsou neznámé koeficienty,  $x_i$  jsou známé hodnoty a  $Z_i$  jsou navzájem nezávislé náhodné veličiny s rozdělením  $N(0, \sigma^2)$  (rozptyl  $\sigma^2$  rovněž není znám). Náhodný vektor  $\mathbf{Y} \equiv (Y_1, \dots, Y_9)^T$  lze tedy popsat lineárním modelem

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \\ Y_8 \\ Y_9 \end{pmatrix} = \begin{pmatrix} 1 & x_1 & 0 & 0 \\ 1 & x_2 & 0 & 0 \\ 1 & x_3 & 0 & 0 \\ 1 & x_4 & 0 & 0 \\ 0 & 0 & 1 & x_5 \\ 0 & 0 & 1 & x_6 \\ 0 & 0 & 1 & x_7 \\ 0 & 0 & 1 & x_8 \\ 0 & 0 & 1 & x_9 \end{pmatrix} \begin{pmatrix} \beta_0^a \\ \beta_1^a \\ \beta_0^b \\ \beta_1^b \end{pmatrix} + \begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \\ Z_5 \\ Z_6 \\ Z_7 \\ Z_8 \\ Z_9 \end{pmatrix},$$

kde vektor  $\mathbf{Z} \equiv (Z_1, \dots, Z_9)^T$  má rozdělení  $N(\mathbf{0}, \sigma^2 \mathbf{I})$ . To znamená, že střední hodnota náhodného vektoru  $\mathbf{Y}$  leží někde ve čtyřrozměrném podprostoru  $M \subset V_n$ , který je generován sloupci výše uvedené matice. Odhadneme ji tedy pomocí pravoúhlého průmětu  $\widehat{\mathbf{Y}}$  vektoru  $\mathbf{Y}$  do tohoto podprostoru.

Chceme-li testovat nulovou hypotézu  $\beta_1^a = \beta_1^b = \beta_1$ , tj. hypotézu, že regresní přímky v případě obou skupin náhodných veličin  $Y_i$  mají stejný sklon, uvažujeme vlastně o redukci výše uvedeného modelu na submodel

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \\ Y_8 \\ Y_9 \end{pmatrix} = \begin{pmatrix} 1 & 0 & x_1 \\ 1 & 0 & x_2 \\ 1 & 0 & x_3 \\ 1 & 0 & x_4 \\ 0 & 1 & x_5 \\ 0 & 1 & x_6 \\ 0 & 1 & x_7 \\ 0 & 1 & x_8 \\ 0 & 1 & x_9 \end{pmatrix} \begin{pmatrix} \beta_0^a \\ \beta_0^b \\ \beta_1 \end{pmatrix} + \begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \\ Z_5 \\ Z_6 \\ Z_7 \\ Z_8 \\ Z_9 \end{pmatrix}.$$

Navrhujeme tedy původní čtyřrozměrný podprostor  $M$  omezit na trojrozměrný podprostor  $S \subset M$ . Test provedeme tak, že určíme  $\mathbf{Y}_S$ , pravoúhlý průmět  $\mathbf{Y}$  do podprostoru  $S$ , a vypočteme podíl

$$\frac{\|\widehat{\mathbf{Y}} - \mathbf{Y}_S\|^2 / (\dim M - \dim S)}{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2 / (\dim V - \dim M)} = \frac{\|\widehat{\mathbf{Y}} - \mathbf{Y}_S\|^2 / 1}{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2 / 5};$$

platí-li nulová hypotéza, má tento podíl rozdělení  $F_{1,5}$ , takže pokud vypočtená realizace přesáhne hodnotu  $F_{1,5}(\alpha)$ , nulovou hypotézu zamítneme na hladině  $\alpha$ .

## 1.7 Lineární množina jako submodel

Množina možných středních hodnot, na kterou submodel omezuje původní podprostor  $M$ , nemusí nutně tvořit vektorový podprostor. V obecnějším případě se může jednat o *lineární množinu*, tj. množinu  $S' \subset V_n$  definovanou vztahem

$$S' \equiv \mathbf{c} + S \equiv \{\mathbf{c} + \mathbf{x} : \mathbf{x} \in S\}$$



### 1.7.2 Vlastnosti vektoru $\widehat{\mathbf{Y}} - \mathbf{Y}_{S'}$

Dokažme nyní, že v případě platnosti nulové hypotézy  $\boldsymbol{\mu} \in S'$  je vektor  $\widehat{\mathbf{Y}} - \mathbf{Y}_{S'}$  pravoúhlým průmětem vektoru  $\mathbf{Y} - \boldsymbol{\mu}$  do podprostoru  $M - S$ ; potřebujeme tedy ukázat, že jsou splněny podmínky

$$\widehat{\mathbf{Y}} - \mathbf{Y}_{S'} \in M - S, \quad (1.48)$$

$$(\mathbf{Y} - \boldsymbol{\mu}) - (\widehat{\mathbf{Y}} - \mathbf{Y}_{S'}) \perp M - S. \quad (1.49)$$

Abychom dokázali platnost podmínky (1.48), musíme se přesvědčit, že platí

$$\begin{aligned} \widehat{\mathbf{Y}} - \mathbf{Y}_{S'} &\in M, \\ \widehat{\mathbf{Y}} - \mathbf{Y}_{S'} &\perp S. \end{aligned}$$

První z těchto požadavků je zřejmě splněn, neboť je  $\widehat{\mathbf{Y}} \in M$ ,  $\mathbf{Y}_{S'} \in S' \subset M$ . Co se týče druhého, platí

$$\widehat{\mathbf{Y}} - \mathbf{Y}_{S'} = (\mathbf{Y} - \mathbf{Y}_{S'}) - (\mathbf{Y} - \widehat{\mathbf{Y}});$$

vektor v první závorce je z definice kolmý na podprostor  $S$ , vektor v druhé závorce je z definice kolmý na podprostor  $M$ , a tedy i na podprostor  $S \subset M$ . I jejich rozdíl je proto kolmý na tento podprostor. Podmínka (1.48) je tedy splněna.

Abychom dokázali platnost podmínky (1.49), provedeme úpravu

$$(\mathbf{Y} - \boldsymbol{\mu}) - (\widehat{\mathbf{Y}} - \mathbf{Y}_{S'}) = (\mathbf{Y} - \widehat{\mathbf{Y}}) + (\mathbf{Y}_{S'} - \boldsymbol{\mu}).$$

Dostali jsme tak součet dvou vektorů, které jsou oba kolmé na podprostor  $M - S$ : první z nich je z definice kolmý na podprostor  $M$ , a tedy i na podprostor  $M - S \subset M$ . Co se týče druhého vektoru, uvědomme si, že rozdíl dvou vektorů, které jsou prvky množiny  $S'$ , musí být prvkem podprostoru  $S$ . Vektor  $\mathbf{Y}_{S'} - \boldsymbol{\mu}$  tedy za předpokladu platnosti hypotézy  $\boldsymbol{\mu} \in S'$  leží v podprostoru  $S$ , takže je kolmý na podprostor  $M - S$ . Podmínka (1.49) je tudíž rovněž splněna.

Vektory  $\mathbf{Y} - \widehat{\mathbf{Y}}$  a  $\widehat{\mathbf{Y}} - \mathbf{Y}_{S'}$  tedy představují pravoúhlé průměty vektoru  $\mathbf{Y} - \boldsymbol{\mu}$  do dvou navzájem kolmých podprostorů  $V_n - M$  a  $M - S$ , jejichž dimenze jsou  $n - m$  a  $m - s$ . Z toho plyne, že náhodná veličina

$$\frac{\|\widehat{\mathbf{Y}} - \mathbf{Y}_{S'}\|^2 / (m - s)}{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2 / (n - m)} \quad (1.50)$$

má rozdělení  $F_{m-s, n-m}$  a můžeme ji použít k verifikaci nulové hypotézy.

### 1.7.3 Užití Pythagorovy věty

Vektor  $\mathbf{Y}_{S'}$  není prvkem podprostoru  $S$ , nemusí být proto kolmý na vektor  $\mathbf{Y} - \mathbf{Y}_{S'}$ , a tím pádem ani na vektor  $\widehat{\mathbf{Y}} - \mathbf{Y}_{S'}$ . Oproti situaci, kdy submodel určuje vektorový podprostor, zde tedy ke vztahu (1.29) přibývá pouze rovnost

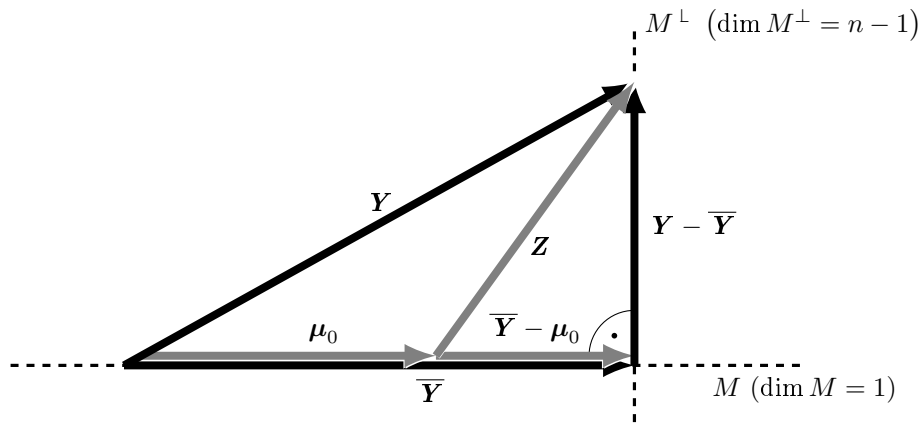
$$\|\mathbf{Y} - \mathbf{Y}_{S'}\|^2 = \|\widehat{\mathbf{Y}} - \mathbf{Y}_{S'}\|^2 + \|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2,$$

plynoucí z toho, že vektor  $\widehat{\mathbf{Y}} - \mathbf{Y}_{S'}$  je prvkem podprostoru  $M$  a vektor  $\mathbf{Y} - \widehat{\mathbf{Y}}$  je na tento podprostor kolmý.

## Příklady

### 1.7.4 Výběr z normálního rozdělení (pokračování ze str. 31)

Vzhledem k jednoduchosti modelu (1.11) přichází v úvahu jedině nulová hypotéza typu  $\mu = \mu_0$ , kde  $\mu_0 \in \mathbb{R}$  je nějaká pevně daná hodnota. To znamená, že uvažujeme o redukcí původního jednorozměrného podprostoru  $M$  na lineární množinu  $S'$  tvořenou jediným vektorem  $\boldsymbol{\mu}_0 \equiv (\mu_0, \dots, \mu_0)^T$ , jejíž dimenze je 0. Aniž bychom se museli odkazovat na obecné závěry odvozené výše, je patrné (viz obr. (1.17)), že platí-li nulová hypotéza, je vektor  $\bar{\mathbf{Y}} - \boldsymbol{\mu}_0$  pravoúhlým průmětem vektoru  $\mathbf{Z}$  do podprostoru  $M$  dimenze 1 a vektor  $\mathbf{Y} - \bar{\mathbf{Y}}$  je pravoúhlým průmětem vektoru  $\mathbf{Z}$  do podprostoru  $M^\perp$  dimenze  $n - 1$ , takže statistika



Obrázek 1.17: Vektory figurující v testu hypotézy  $\mu = \mu_0$  v modelu (1.11).

$$F = \frac{\|\bar{\mathbf{Y}} - \boldsymbol{\mu}_0\|^2 / 1}{\|\mathbf{Y} - \bar{\mathbf{Y}}\|^2 / (n - 1)} = \frac{n (\bar{Y} - \mu_0)^2}{S^2} \quad (1.51)$$

má rozdělení  $F_{1,n-1}$ . Pokud pro konkrétní realizaci nastane nerovnost

$$\frac{n (\bar{Y} - \mu_0)^2}{S^2} \geq F_{1,n-1}(\alpha), \quad (1.52)$$

zamítneme nulovou hypotézu na hladině  $\alpha$ .

Statistiku (1.51) můžeme použít i pro určení intervalu spolehlivosti: je-li skutečná hodnota střední hodnoty  $\mu$ , platí

$$\mathbb{P} \left[ \frac{n (\bar{Y} - \mu)^2}{S^2} < F_{1,n-1}(\alpha) \right] = 1 - \alpha,$$

z čehož vyplývá, že jev

$$\mu \in \left( \bar{Y} - S \sqrt{\frac{F_{1,n-1}(\alpha)}{n}}; \bar{Y} + S \sqrt{\frac{F_{1,n-1}(\alpha)}{n}} \right)$$

nastane s pravděpodobností  $1 - \alpha$ .

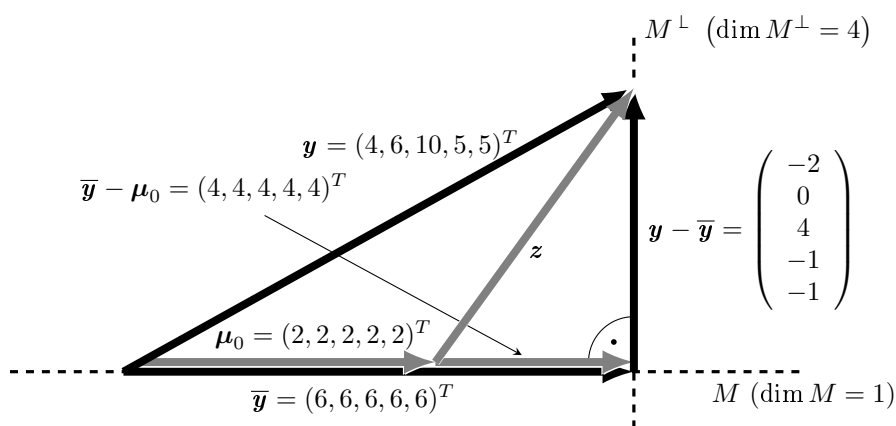
Řekněme, že v našem konkrétním případě, kdy jsme měřili hodnoty u pěti pacientů, chceme testovat hypotézu  $\mu = 2$ , tj.  $\mu_0 = (2, 2, 2, 2, 2)^T$ . Ze získané realizace

$$\mathbf{y} = (4, 6, 10, 5, 5)^T$$

postupně vypočteme

$$\begin{aligned}\bar{\mathbf{y}} &= (6, 6, 6, 6, 6)^T, \\ \mathbf{y} - \bar{\mathbf{y}} &= (-2, 0, 4, -1, -1)^T, \\ \bar{\mathbf{y}} - \mu_0 &= (4, 4, 4, 4, 4)^T\end{aligned}$$

(viz obr. 1.18) a dostaneme hodnotu statistiky



**Obrázek 1.18:** Příklad konkrétních realizací figurujících v testu hypotézy  $\mu = 2$  v modelu (1.11).

$$\begin{aligned}F &= \frac{\|\bar{\mathbf{y}} - \mu_0\|^2 / 1}{\|\mathbf{y} - \bar{\mathbf{y}}\|^2 / 4} = \\ &= \frac{\|(4, 4, 4, 4, 4)^T\|^2 / 1}{\|(-2, 0, 4, -1, -1)^T\|^2 / 4} \doteq \\ &\doteq 14,55.\end{aligned}$$

Protože je  $14,55 \geq F_{1,4}(0,05) \doteq 7,71$ , nulovou hypotézu můžeme zamítnout na hladině významnosti 5%.

### Jednovýběrový $t$ -test versus $F$ -test

Čtenář snad může být výše popsaným testem překvapen, protože k verifikaci uvedené nulové hypotézy se obvykle používá jednovýběrový  $t$ -test. Je ale třeba si uvědomit, že výsledek  $t$ -testu je – alespoň v případě jeho oboustranné varianty – s výsledkem našeho postupu ekvivalentní. Jak je známo (viz např. [3], případně příklad 1.9.1), testová statistika používaná v  $t$ -testu

$$t = \frac{(\bar{Y} - \mu_0)\sqrt{n}}{S}$$

se řídí rozdělením  $t_{n-1}$  a při jejím použití zamítáme nulovou hypotézu  $\mu = \mu_0$  ve prospěch její alternativy  $\mu \neq \mu_0$  na hladině  $\alpha$  právě tehdy, když nastane nerovnost

$$\frac{|\bar{Y} - \mu_0|\sqrt{n}}{S} \geq t_{n-1}(\alpha). \quad (1.53)$$

Ta je však ekvivalentní s nerovností

$$\frac{n(\bar{Y} - \mu_0)^2}{S^2} \geq t_{n-1}^2(\alpha)$$

a vzhledem ke vztahu (1.9) i s nerovností (1.52).

### Jednostranná alternativa $F$ -testu

Výše uvedený  $F$ -test lze ovšem použít i tehdy, když je alternativní hypotéza jednostranná, např.  $\mu > \mu_0$ . V takovém případě zamítneme nulovou hypotézu na hladině  $\alpha$ , pokud bude splněna dvojice nerovností

$$\bar{x} > \mu_0 \quad \wedge \quad \frac{n(\bar{x} - \mu_0)^2}{S^2} \geq F_{1,n-1}(2\alpha).$$

V případě platnosti nulové hypotézy totiž platí

$$\mathbf{P} \left[ \frac{n(\bar{x} - \mu_0)^2}{S^2} \geq F_{1,n-1}(2\alpha) \right] = 2\alpha,$$

tj.

$$\mathbf{P} \left[ \frac{\sqrt{n}|\bar{x} - \mu_0|}{S} \geq \sqrt{F_{1,n-1}(2\alpha)} \right] = 2\alpha,$$

z čehož vzhledem k symetrii rozdělení  $\bar{x}$  kolem  $\mu_0$  a k nezávislosti  $\bar{x}$  na  $S^2$  plyne

$$\mathbf{P} \left[ \bar{x} > \mu_0 \quad \wedge \quad \frac{n(\bar{x} - \mu_0)^2}{S^2} \geq F_{1,n-1}(2\alpha) \right] = \alpha.$$

Tento postup a jeho zdůvodnění není samozřejmě nikterak elegantní; uvádíme jej proto, abychom ukázali univerzálnost  $F$ -testu, který je z geometrického hlediska podstatně názornější než  $t$ -test.

### 1.7.5 Regresní přímka (pokračování ze str. 35)

#### Test hypotézy $\beta_0 = \beta_0^0, \beta_1 = \beta_1^0$

Navrhněme test hypotézy fixující hodnoty obou parametrů vektoru  $\beta$ . Ukážeme si tento postup rovnou na konkrétních hodnotách; dejme tomu, že v modelu (1.13) chceme testovat nulovou hypotézu  $\beta_0 = 30, \beta_1 = 4$ . V takovém případě uvažujeme o redukci původního dvojrozměrného podprostoru na jediný vektor  $\mu_0 \equiv 30\mathbf{e} + 4\mathbf{x}$ , tj. na lineární množinu  $S'$  dimenze 0. K vektoru  $\mathbf{Y}$  je z této množiny nejbližše pochopitelně právě tento jediný vektor, jehož souřadnice jsou

$$\mu_0 = \begin{pmatrix} 1 & 3 \\ 1 & 2 \\ 1 & 4 \\ 1 & 3 \\ 1 & 5 \\ 1 & 4 \\ 1 & 7 \end{pmatrix} \cdot \begin{pmatrix} 30 \\ 4 \end{pmatrix} = \begin{pmatrix} 42 \\ 38 \\ 46 \\ 42 \\ 50 \\ 46 \\ 58 \end{pmatrix}.$$

Pokud nulová hypotéza  $\boldsymbol{\mu} = \boldsymbol{\mu}_0$  skutečně platí, je vektor  $\widehat{\mathbf{Y}} - \boldsymbol{\mu}_0$  pravoúhlým průmětem vektoru  $\mathbf{Y} - \boldsymbol{\mu}_0$  do podprostoru dimenze  $\dim M - \dim S' = 2$ ; vektor  $\mathbf{Y} - \widehat{\mathbf{Y}}$  je zároveň pravoúhlým průmětem vektoru  $\mathbf{Y} - \boldsymbol{\mu}_0$  do podprostoru dimenze  $\dim V_n - \dim M = 5$ , takže náhodná veličina

$$F = \frac{\|\widehat{\mathbf{Y}} - \boldsymbol{\mu}_0\|^2 / 2}{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2 / 5}$$

má rozdělení  $F_{2,5}$ . Hodnota naší realizace je

$$F = \frac{128/2}{76/5} = 4,21;$$

protože kritická hodnota rozdělení  $F_{2,5}(0,05)$  je 5,79, nemůžeme (alespoň ne při požadavku 5% hladiny významnosti) na základě získaných dat zamítnout nulovou hypotézu  $\beta_0 = 30, \beta_1 = 4$ .

### Obecná formulace

Chceme-li v modelu (1.19) testovat nulovou hypotézu  $\beta_0 = \beta_0^0, \beta_1 = \beta_1^0$ , kde  $\beta_0^0, \beta_1^0 \in \mathbb{R}$  jsou nějaké pevně dané hodnoty, uvažujeme o redukci podprostoru daného původním modelem na jediný vektor  $\boldsymbol{\mu}_0 \equiv \beta_0^0 \mathbf{e} + \beta_1^0 \mathbf{x}$ , tj. snižujeme dimenzi původního podprostoru  $M$  o hodnotu 2. Odhadem střední hodnoty odpovídající submodelu je právě tento jediný vektor  $\boldsymbol{\mu}_0$  a testová statistika bude mít tvar

$$F = \frac{\|\widehat{\mathbf{Y}} - \boldsymbol{\mu}_0\|^2 / 2}{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2 / (n-2)} = \frac{\sum_{i=1}^n (\widehat{Y}_i - \beta_0^0 - \beta_1^0 x_i)^2}{2S^2}. \quad (1.54)$$

Nulovou hypotézu budeme zamítat na hladině významnosti  $\alpha$ , bude-li příslušná statistika větší než hodnota  $F_{2,n-2}(\alpha)$ .

### Ortogonalizace generátorů podprostoru $M$

V dalších úvahách již budeme vycházet z obecného případu, tj. z modelu (1.19). Nejdříve uvedeme jeden často používaný postup, který nám značně usnadní výpočty. Spočívá v nahrazení hodnot  $x_i$  hodnotami

$$q_i \equiv x_i - \bar{x}, \quad \text{kde} \quad \bar{x} \equiv \frac{\sum_{i=1}^n x_i}{n}.$$

Z geometrického hlediska se tedy jedná o nahrazení vektoru  $\mathbf{x}$  vektorem

$$\mathbf{q} \equiv \mathbf{x} - \bar{x} \mathbf{e} \equiv \mathbf{x} - \bar{\mathbf{x}}. \quad (1.55)$$

Je zřejmé, že vektory  $\mathbf{e}$  a  $\mathbf{q}$  generují též podprostor  $M$  jako vektory  $\mathbf{e}$  a  $\mathbf{x}$ , model (1.19) tedy můžeme nahradit modelem

$$\mathbf{Y} = \begin{pmatrix} 1 & q_1 \\ \vdots & \vdots \\ 1 & q_n \end{pmatrix} \cdot \begin{pmatrix} \beta_0^* \\ \beta_1^* \end{pmatrix} + \mathbf{Z} \equiv \mathbf{X}^* \boldsymbol{\beta}^* + \mathbf{Z}. \quad (1.56)$$

Souřadnice  $\beta_0^*, \beta_1^*$  střední hodnoty  $\boldsymbol{\mu}$  vzhledem k bázi  $\{\mathbf{e}, \mathbf{q}\}$  podprostoru  $M$  budou sice jiné než původní souřadnice  $\mu, \beta$  vzhledem k bázi  $\{\mathbf{e}, \mathbf{x}\}$ , ze srovnání

$$\begin{aligned}\boldsymbol{\mu} = \beta_0 \mathbf{e} + \beta_1 \mathbf{x} &= \beta_0^* \mathbf{e} + \beta_1^* \mathbf{q} = \\ &= \beta_0^* \mathbf{e} + \beta_1^* (\mathbf{x} - \bar{x} \mathbf{e}) = \\ &= (\beta_0^* - \beta_1^* \bar{x}) \mathbf{e} + \beta_1^* \mathbf{x}\end{aligned}$$

je však patrné, že platí

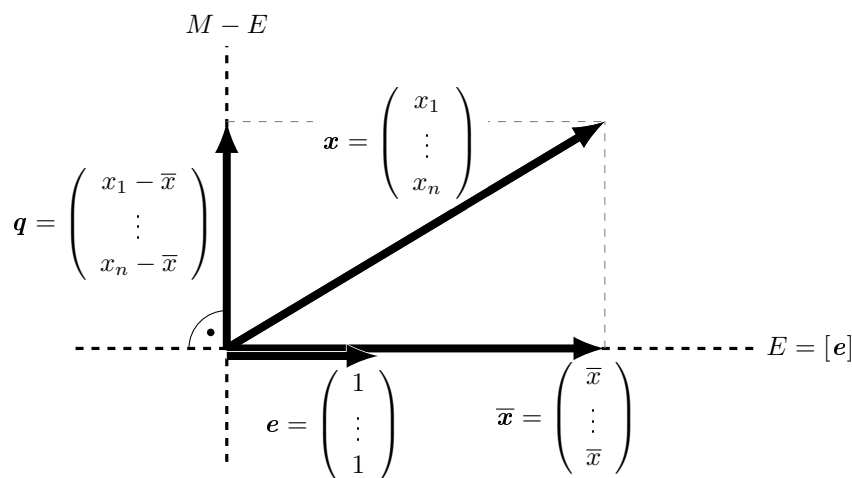
$$\begin{aligned}\beta_0 &= \beta_0^* - \beta_1^* \bar{x}, \\ \beta_1 &= \beta_1^*,\end{aligned}$$

a podobně pro souřadnice  $b_0, b_1$ , resp.  $b_0^*, b_1^*$ , náhodného vektoru  $\widehat{\mathbf{Y}}$  vzhledem k bázi  $\{\mathbf{e}, \mathbf{q}\}$ , resp.  $\{\mathbf{e}, \mathbf{x}\}$ :

$$\begin{aligned}b_0 &= b_0^* - b_1^* \bar{x}, \\ b_1 &= b_1^*\end{aligned}$$

(symboly  $\beta_1^*, b_1^*$  tedy již nebudeme používat).

Poněvadž vektor  $\bar{\mathbf{x}} = (\bar{x}, \dots, \bar{x})^T$  je pravoúhlým průmětem vektoru  $\mathbf{x}$  do podprostoru  $E = [\mathbf{e}]$ , je vektor  $\mathbf{q} = \mathbf{x} - \bar{x} \mathbf{e}$  pravoúhlým průmětem vektoru  $\mathbf{x}$  do podprostoru  $M - E$  (viz obr. 1.19). Vektory  $\mathbf{e}, \mathbf{q}$  tedy generují dvě navzájem



**Obrázek 1.19:** Nahrazením vektoru  $\mathbf{x}$  v modelu (1.19) vektorem  $\mathbf{q} \equiv \mathbf{x} - \bar{x} \mathbf{e}$  získáme ortogonální bázi  $\{\mathbf{e}, \mathbf{q}\}$  roviny  $M$  (ta zde splývá s rovinou nákresny).

kolmé přímky, takže hodnoty  $b_0^*, b_1$  lze určit tak, že promítneme vektor  $\mathbf{Y}$  na každou z těchto přímek zvlášť (viz též vzorec (2.34)):

$$b_0^* = \frac{\mathbf{Y} \circ \mathbf{e}}{\|\mathbf{e}\|^2} = \bar{Y}, \quad (1.57)$$

$$b_1 = \frac{\mathbf{Y} \circ (\mathbf{x} - \bar{x} \mathbf{e})}{\|\mathbf{x} - \bar{x} \mathbf{e}\|^2}. \quad (1.58)$$

Jelikož vektor  $\overline{\mathbf{Y}}$  leží v podprostoru  $E$ , jsou vektory  $\overline{\mathbf{Y}}$  a  $\mathbf{x} - \overline{\mathbf{x}}$  navzájem kolmé, tj. platí  $\overline{\mathbf{Y}} \circ (\mathbf{x} - \overline{\mathbf{x}}) = 0$ . Poslední rovnost lze proto psát také ve tvaru

$$b_1 = \frac{(\mathbf{Y} - \overline{\mathbf{Y}}) \circ (\mathbf{x} - \overline{\mathbf{x}})}{\|\mathbf{x} - \overline{\mathbf{x}}\|^2}, \quad (1.59)$$

který je speciálním případem vzorce (1.63). Podobně jsou kolmé též vektory  $\mathbf{Y} - \overline{\mathbf{Y}}$  a  $\overline{\mathbf{x}}$ , z čehož plyne rovnost  $\mathbf{y} \circ \overline{\mathbf{x}} = \overline{\mathbf{Y}} \circ \overline{\mathbf{x}}$ . Díky tomu lze upravit čítec výrazu (1.58) na tvar

$$\begin{aligned} \mathbf{Y} \circ (\mathbf{x} - \overline{\mathbf{x}}) &= \mathbf{Y} \circ \mathbf{x} - \mathbf{Y} \circ \overline{\mathbf{x}} = \\ &= \mathbf{Y} \circ \mathbf{x} - \overline{\mathbf{Y}} \circ \overline{\mathbf{x}} = \\ &= \sum_{i=1}^n x_i Y_i - n \overline{\mathbf{x}} \overline{\mathbf{Y}}. \end{aligned}$$

Při výpočtu jmenovatele lze pro změnu využít skutečnosti, že vektory  $\mathbf{x} - \overline{\mathbf{x}}$  a  $\overline{\mathbf{x}}$  představují rozklad vektoru  $\mathbf{x} \in M$  do dvou navzájem kolmých podprostorů  $E$  a  $M - E$ , platí tedy

$$\|\mathbf{x} - \overline{\mathbf{x}}\|^2 = \|\mathbf{x}\|^2 - \|\overline{\mathbf{x}}\|^2 = \sum_{i=1}^n x_i^2 - n \overline{\mathbf{x}}^2;$$

po dosazení těchto výrazů do (1.58) a zkrácení zlomku hodnotou  $n$  dojdeme k prvnímu ze vztahů (1.20).

### Test hypotézy $\beta_1 = \beta_1^0$

V případě hypotézy  $\beta_1 = \beta_1^0$ , kde  $\beta_1^0 \in \mathbb{R}$  je nějaká pevně daná hodnota, uvažujeme o redukci roviny  $M$  dané modelem (1.13) na přímku

$$S' \equiv \{\beta_1^0 \mathbf{x} + t \mathbf{e}; t \in \mathbb{R}\} = \beta_1^0 \mathbf{x} + E.$$

Mohli bychom se samozřejmě odkázat na vzorce (1.47), zdá se nám však užitečné podívat se na tuto situaci z jiného úhlu.

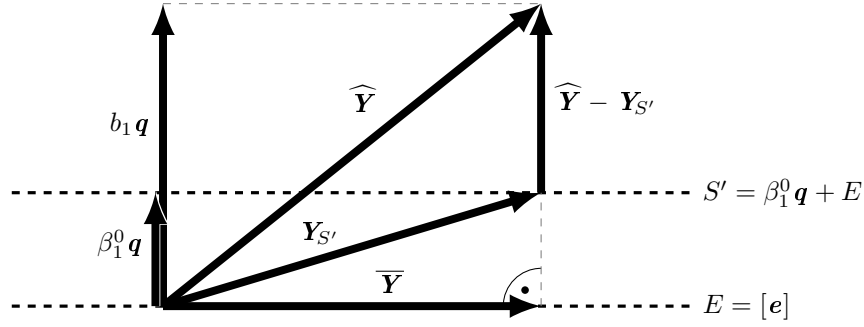
Místo báze  $\{\mathbf{e}, \mathbf{x}\}$  použijeme k orientaci v rovině  $M$  bázi  $\{\mathbf{e}, \mathbf{q}\}$ , kde vektor  $\mathbf{q}$  byl zaveden vztahem (1.55). Pro přímku  $S'$  tak můžeme psát

$$\begin{aligned} S' &= \{\beta_1^0 (\mathbf{q} + \overline{\mathbf{x}} \mathbf{e}) + t \mathbf{e}; t \in \mathbb{R}\} = \\ &= \{\beta_1^0 \mathbf{q} + t \mathbf{e}; t \in \mathbb{R}\} = \\ &= \beta_1^0 \mathbf{q} + E. \end{aligned}$$

Z předchozí podkapitoly známe souřadnice  $b_0^*, b_1$  pravoúhlého průmětu náhodného vektoru  $\mathbf{Y}$  do podprostoru  $M$  vzhledem k bázi  $\{\mathbf{e}, \mathbf{q}\}$ . Tímto průmětem je tedy vektor

$$\widehat{\mathbf{Y}} = b_0^* \mathbf{e} + b_1 \mathbf{q} = \overline{\mathbf{Y}} \mathbf{e} + b_1 \mathbf{q} = \overline{\mathbf{Y}} + b_1 \mathbf{q}.$$

Ukážeme, že pravoúhlým průmětem náhodného vektoru  $\mathbf{Y}$  do lineární množiny  $S'$  je vektor  $\mathbf{Y}_{S'} = \overline{\mathbf{Y}} + \beta_1^0 \mathbf{q}$  (viz obr. 1.20); je totiž zřejmě prvkem množiny  $S'$  a zároveň je vektor  $\mathbf{Y} - \mathbf{Y}_{S'}$  kolmý na podprostor  $E$ , neboť je součtem vektorů



**Obrázek 1.20:** Vektorem, který je ze všech prvků lineární množiny  $S' = \beta_1^0 \mathbf{q} + E$  nejbližší k vektoru  $\mathbf{Y}$ , je vektor  $\mathbf{Y}_{S'} = \bar{\mathbf{Y}} + \beta_1^0 \mathbf{q}$ . Rozdíl vektorů  $\widehat{\mathbf{Y}} - \mathbf{Y}_{S'}$  je pak roven  $(b_1 - \beta_1^0) \mathbf{q}$  a je kolmý na podprostor  $E$  (rovina  $M$  zde splývá s nákresem).

$\mathbf{Y} - \widehat{\mathbf{Y}}$  a  $\widehat{\mathbf{Y}} - \mathbf{Y}_{S'}$ , z nichž první je z definice kolmý na podprostor  $M$ , a tudíž i na podprostor  $E$ , a pro druhý platí

$$\begin{aligned} \widehat{\mathbf{Y}} - \mathbf{Y}_{S'} &= (\bar{\mathbf{Y}} + b_1 \mathbf{q}) - (\bar{\mathbf{Y}} + \beta_1^0 \mathbf{q}) = \\ &= (b_1 - \beta_1^0) \mathbf{q}, \end{aligned}$$

je tedy rovněž kolmý na podprostor  $E$ . Dimenze submodelu je o 1 nižší než dimenze původního modelu, dostáváme tak vzorec pro výpočet statistiky pro test nulové hypotézy:

$$\begin{aligned} F &= \frac{\|\widehat{\mathbf{Y}} - \mathbf{Y}_{S'}\|^2 / 1}{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2 / (n - 2)} = \\ &= \frac{(b_1 - \beta_1^0)^2 \|\mathbf{q}\|^2}{S^2} = \\ &= \frac{(b_1 - \beta_1^0)^2 \left( \sum_{i=1}^n x_i^2 - n\bar{x}^2 \right)}{S^2}. \end{aligned}$$

Nulovou hypotézu zamítneme, nastane-li nerovnost  $F \geq F_{1,n-2}(\alpha)$ . Uvedený výsledek lze použít i k vytvoření intervalu spolehlivosti: je-li skutečná hodnota lineárního koeficientu rovna  $\beta_1$ , platí

$$\mathbb{P} \left[ \frac{(b_1 - \beta_1)^2 \left( \sum_{i=1}^n x_i^2 - n\bar{x}^2 \right)}{S^2} < F_{1,n-2}(\alpha) \right] = 1 - \alpha,$$

což znamená, že jev

$$\beta_1 \in \left( b_1 - S \sqrt{\frac{F_{1,n-2}}{n \sum_{i=1}^n x_i^2 - n\bar{x}}}; b_1 + S \sqrt{\frac{F_{1,n-2}}{n \sum_{i=1}^n x_i^2 - n\bar{x}}} \right)$$

nastane s pravděpodobností  $1 - \alpha$ .

### 1.7.6 Mnohonásobná regrese (pokračování ze str. 38)

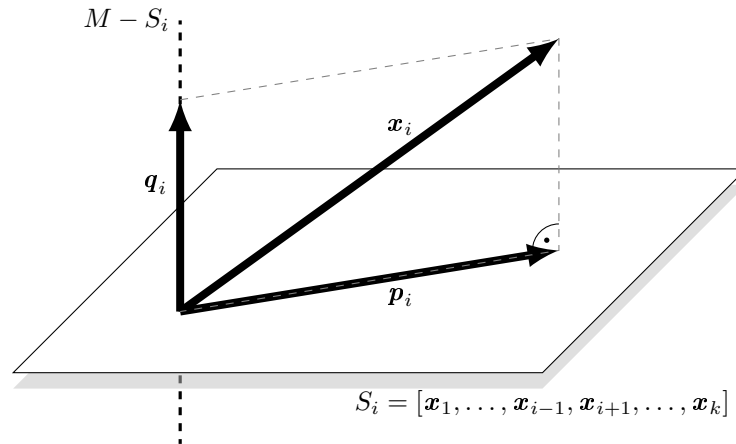
Označme vektory, které reprezentují sloupce matice  $\mathbf{X}$  v modelu (1.45), symboly  $\mathbf{x}_0, \dots, \mathbf{x}_k$  (tj.  $\mathbf{x}_0 \equiv \mathbf{e}$ ). Dále zavedme označení  $S_i$  pro vektorový podprostor generovaný všemi vektory  $\mathbf{x}_0, \dots, \mathbf{x}_k$  s výjimkou  $i$ -tého, tj.

$$S_i \equiv [\mathbf{x}_0, \dots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \dots, \mathbf{x}_k],$$

a označení  $\mathbf{p}_i$  pro pravoúhlý průmět vektoru  $\mathbf{x}_i$  do podprostoru  $S_i$ . Označíme-li ještě symbolem  $\mathbf{q}_i$  vektor  $\mathbf{x}_i - \mathbf{p}_i$ , představuje zápis

$$\mathbf{x}_i = \mathbf{p}_i + \mathbf{q}_i$$

rozklad vektoru  $\mathbf{x}_i$  do dvou navzájem kolmých podprostorů  $S_i$  a  $M - S_i$  (viz obr. 1.21).



**Obrázek 1.21:** Vektory  $\mathbf{q}_i$  a  $\mathbf{p}_i$ , použité při testu hypotézy  $\beta_i = \beta_i^0$  v případě mnohonásobné regrese, představují rozklad vektoru  $\mathbf{x}_i$  do dvou navzájem kolmých podprostorů  $S_i$  a  $M - S_i$ .

#### Test hypotézy $\beta_i = \beta_i^0$

Hypotéza fixující hodnotu parametru  $\beta_i$ , tj. hypotéza  $\beta_i = \beta_i^0$ , kde  $\beta_i^0 \in \mathbb{R}$  je nějaké pevně dané číslo, představuje redukcí podprostoru  $M$  na lineární množinu  $S'_i$  danou předpisem

$$S'_i \equiv \beta_i^0 \mathbf{x}_i + S_i.$$

Protože je  $\mathbf{p}_i \in S_i$ , můžeme tuto lineární množinu  $S'_i$  zapsat ve tvaru

$$S'_i = \beta_i^0 (\mathbf{p}_i + \mathbf{q}_i) + S_i = \beta_i^0 \mathbf{q}_i + S_i$$

a vektor  $\widehat{\mathbf{Y}}$  ve tvaru

$$\begin{aligned} \widehat{\mathbf{Y}} &= b_0 \mathbf{x}_0 + \dots + b_{i-1} \mathbf{x}_{i-1} + b_i \mathbf{x}_i + b_{i+1} \mathbf{x}_{i+1} + \dots + b_k \mathbf{x}_k = \\ &= b_0 \mathbf{x}_0 + \dots + b_{i-1} \mathbf{x}_{i-1} + b_i (\mathbf{p}_i + \mathbf{q}_i) + b_{i+1} \mathbf{x}_{i+1} + \dots + b_k \mathbf{x}_k \equiv \\ &\equiv b_i \mathbf{q}_i + \mathbf{v}, \end{aligned}$$

kde  $\mathbf{v}$  je nějaký vektor ležící ve vektorovém podprostoru  $S_i$ . Nyní lze ukázat, že pravoúhlým průmětem vektoru  $\mathbf{Y}$  do lineární množiny  $S'_i$  je vektor

$$\mathbf{Y}_{S'_i} = \beta_i^0 \mathbf{q}_i + \mathbf{v}.$$

Je totiž zjevně prvkem množiny  $S'_i$  a zároveň platí

$$\begin{aligned} \mathbf{Y} - \mathbf{Y}_{S'_i} &= (\mathbf{Y} - \widehat{\mathbf{Y}}) + (\widehat{\mathbf{Y}} - \mathbf{Y}_{S'_i}) = \\ &= (\mathbf{Y} - \widehat{\mathbf{Y}}) + (b_i - \beta_i^0) \mathbf{q}_i, \end{aligned}$$

přičemž oba sčítance jsou zřejmě kolmé na podprostor  $S_i$ . Dimenze lineární množiny  $S'_i$  je o 1 menší než dimenze podprostoru  $M$ , testová statistika tedy bude mít tvar

$$F = \frac{\|\widehat{\mathbf{Y}} - \mathbf{Y}_{S'_i}\|^2 / 1}{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2 / (n - k - 1)} = \frac{(b_i - \beta_i^0)^2 \|\mathbf{q}_i\|^2}{S^2}. \quad (1.60)$$

Nulovou hypotézu zamítneme v případě, že nastane nerovnost  $F > F_{1,n-k-1}(\alpha)$ .<sup>6</sup>

### Interval spolehlivosti pro $\beta_i$

Výše odvozený výsledek umožňuje též odvození intervalu spolehlivosti: je-li skutečná hodnota testovaného parametru  $\beta_i$ , platí

$$\mathbf{P} \left\{ \frac{(b_i - \beta_i)^2 \|\mathbf{q}_i\|^2}{S^2} > F_{1,n-k-1}(\alpha) \right\} = 1 - \alpha,$$

takže

$$\beta_i \in \left( b_i - \frac{S \sqrt{F_{1,n-k-1}(\alpha)}}{\|\mathbf{q}_i\|}; b_i + \frac{S \sqrt{F_{1,n-k-1}(\alpha)}}{\|\mathbf{q}_i\|} \right)$$

nastane s pravděpodobností  $1 - \alpha$ .

### Srovnání s $t$ -testem

Lze ukázat, že platí

$$\|\mathbf{q}_i\|^2 = 1/c_{ii},$$

kde  $c_{ii}$  je prvek na  $i$ -tém místě diagonály (počítáno od 0) matice  $(\mathbf{X}^T \mathbf{X})^{-1}$ . Vzhledem k již zmíněnému vztahu (1.9) mezi rozdělením  $t$  a  $F$  je tedy výše odvozený test až na znaménko identický se standardně používaným  $t$ -testem, založeným na statistice

$$t = \frac{b_i - \beta_i^0}{S \sqrt{c_{ii}}}.$$

(viz [3], resp. příklad 1.9.3).

<sup>6</sup>Ve speciálním případě  $\beta_i^0 = 0$  dostáváme samozřejmě vzorec pro test popsany v příkladu 1.6.9.

## Vyjádření vektoru $\mathbf{b}$ – alternativní způsob

K výpočtu náhodného vektoru  $\mathbf{b}$  můžeme samozřejmě použít vzorec (1.17), inspirování některými postupy použitými v kapitole 1.7.5 však nabízí i jinou cestu, po níž dojdeme k zajímavému výsledku. Víme, že vektory  $\overline{\mathbf{Y}}$  a  $\widehat{\mathbf{Y}} - \overline{\mathbf{Y}}$  jsou navzájem kolmé, takže zápis

$$\widehat{\mathbf{Y}} = \overline{\mathbf{Y}} + (\widehat{\mathbf{Y}} - \overline{\mathbf{Y}}) \quad (1.61)$$

představuje rozklad vektoru  $\widehat{\mathbf{Y}} \in M$  do dvou navzájem kolmých podprostorů  $E$  a  $M - E$ . Položíme-li dále

$$\bar{x}_j \equiv \sum_{i=1}^n x_{ij} / n,$$

můžeme podobně rozložit vektory  $\mathbf{x}_j$  tvořící sloupce matice  $\mathbf{X}$  na součet  $\mathbf{x}_j = \bar{\mathbf{x}}_j + (\mathbf{x}_j - \bar{\mathbf{x}}_j)$ , kde

$$\bar{\mathbf{x}}_j = (\bar{x}_j, \dots, \bar{x}_j)^T.$$

Vektory  $\bar{\mathbf{x}}_j$  jsou tedy pravoúhlými průměty vektorů  $\mathbf{x}_j$  do podprostoru  $E$  a vektory  $\mathbf{x}_j - \bar{\mathbf{x}}_j$  jsou průměty do podprostoru  $M - E$  (příčemž tvoří zřejmě jeho bázi). Nyní můžeme vektor  $\widehat{\mathbf{Y}}$  zapsat ve tvaru

$$\begin{aligned} \widehat{\mathbf{Y}} &= b_0 \mathbf{e} + \sum_{j=1}^k b_j \mathbf{x}_j = \\ &= b_0 \mathbf{e} + \sum_{j=1}^k b_j \bar{\mathbf{x}}_j + \sum_{j=1}^k b_j (\mathbf{x}_j - \bar{\mathbf{x}}_j), \end{aligned} \quad (1.62)$$

ve kterém jsou zřejmě první dva sčítance prvky podprostoru  $E$  a třetí sčítanec leží v podprostoru  $M - E$ ; ze srovnání tohoto výrazu s rozkladem (1.61) je tak patrné, že platí

$$\widehat{\mathbf{Y}} - \overline{\mathbf{Y}} = \sum_{j=1}^k b_j (\mathbf{x}_j - \bar{\mathbf{x}}_j).$$

Označme symbolem  $\mathbf{W}$  matici, jejíž sloupce jsou tvořeny vektory  $\mathbf{x}_i - \bar{\mathbf{x}}_i$ , a symbolem  $\mathbf{b}^*$  vektor náhodných veličin

$$\mathbf{b}^* \equiv (b_1, \dots, b_k)^T;$$

můžeme tedy psát

$$\widehat{\mathbf{Y}} - \overline{\mathbf{Y}} = \mathbf{W} \mathbf{b}^*.$$

Vektor  $\widehat{\mathbf{Y}} - \overline{\mathbf{Y}}$  je pravoúhlým průmětem vektoru  $\mathbf{Y} - \overline{\mathbf{Y}}$  do podprostoru  $M - E$ , neboť je prvkem tohoto podprostoru a přitom platí

$$(\mathbf{Y} - \overline{\mathbf{Y}}) - (\widehat{\mathbf{Y}} - \overline{\mathbf{Y}}) = \mathbf{Y} - \widehat{\mathbf{Y}} \perp M - E.$$

Vektor  $\mathbf{b}^*$  tedy můžeme podobně jako vektor  $\mathbf{b}$  v kapitole 1.3 určit ze soustavy

$$\mathbf{W}^T [(\mathbf{Y} - \overline{\mathbf{Y}}) - \mathbf{W} \mathbf{b}^*] = \mathbf{0},$$

tj. z podmínky, že vektor

$$(\mathbf{Y} - \bar{\mathbf{Y}}) - (\widehat{\mathbf{Y}} - \bar{\mathbf{Y}}) = (\mathbf{Y} - \bar{\mathbf{Y}}) - \mathbf{W}\mathbf{b}^*$$

je kolmý na všechny sloupce matice  $\mathbf{W}$  (tj. vektory  $\mathbf{x}_i - \bar{\mathbf{x}}_i$ ). Dostáváme

$$\begin{aligned} \mathbf{b}^* &= (\mathbf{W}^T \mathbf{W})^{-1} \mathbf{W}^T (\mathbf{Y} - \bar{\mathbf{Y}}) = \\ &= \left( \frac{\mathbf{W}^T \mathbf{W}}{n-1} \right)^{-1} \cdot \frac{\mathbf{W}^T (\mathbf{Y} - \bar{\mathbf{Y}})}{n-1} = \\ &\equiv \mathbf{C}_{\mathbf{X}, \mathbf{X}}^{-1} \cdot \mathbf{C}_{\mathbf{X}, \mathbf{Y}}. \end{aligned} \tag{1.63}$$

V případě, že vektory

$$\begin{aligned} (x_{11}, \dots, x_{1k}, Y_i), \\ \vdots \\ (x_{n1}, \dots, x_{nk}, Y_i) \end{aligned}$$

představují náhodný výběr z  $(k+1)$ -rozměrného rozdělení náhodného vektoru

$$(X_1, \dots, X_k, Y) \equiv (\mathbf{X}, Y)$$

můžeme matici  $\mathbf{C}_{\mathbf{X}, \mathbf{X}}$ , resp.  $\mathbf{C}_{\mathbf{X}, \mathbf{Y}}$ , interpretovat jako výběrovou kovarianční matici náhodných vektorů  $\mathbf{X}$  a  $\mathbf{X}$ , resp. náhodného vektoru  $\mathbf{X}$  a náhodné veličiny  $Y$ . Tento vztah je obecnějším případem vzorce (1.59). Z rovnosti

$$\bar{\mathbf{Y}} = b_0 \mathbf{e} + \sum_{j=1}^n b_j \bar{\mathbf{x}}_j,$$

která plyne opět ze srovnání vztahů (1.61) a (1.62), dostáváme posléze

$$b_0 = \bar{Y} - \sum_{j=1}^n b_j \bar{x}_j;$$

speciálním případem tohoto vzorce je druhý ze vztahů (1.20).

## 1.8 Více submodelů

### 1.8.1 Posloupnost do sebe vnořených podprostorů

Pokud podprostory určené jednotlivými submodely tvoří posloupnost, ve které je každý podprostor podmnožinou předchozího, je to – přinejmenším z geometrického hlediska – situace poměrně prostá. Pro jednotnost označme podprostor určený výchozím modelem  $M_1$ , podprostory určené následujícími submodely  $M_2, \dots, M_r$  (musí být  $r \leq k+1$ , kde  $k$  je dimenze  $M_1$ , neboť dimenze podprostoru  $M_r$  může být 0),  $d_1, \dots, d_r$  jejich dimenze a  $\mathbf{Y}_1, \dots, \mathbf{Y}_r$  pravoúhlé průměty náhodného vektoru  $\mathbf{Y}$  do těchto podprostorů. Dodefinujme ještě  $\mathbf{Y}_0 \equiv \mathbf{Y}$  a  $M_0 \equiv V_n$ . Platí tedy  $M_0 \supset \dots \supset M_r$  a

$$\begin{aligned} \mathbf{Y}_i &\in M_i, \\ \mathbf{Y} - \mathbf{Y}_i &\perp M_i \end{aligned}$$

pro všechna  $i = 0, \dots, r$ .

Ukažme nejprve, že pro všechna  $i = 1, \dots, r$  platí také

$$\mathbf{Y}_{i-1} - \mathbf{Y}_i \perp M_i. \quad (1.64)$$

Lze totiž psát

$$\mathbf{Y}_{i-1} - \mathbf{Y}_i = (\mathbf{Y} - \mathbf{Y}_i) - (\mathbf{Y} - \mathbf{Y}_{i-1}),$$

kde obě závorky představují vektory kolmé na podprostor  $M_i$  (vektor  $\mathbf{Y} - \mathbf{Y}_{i-1}$  je z definice kolmý na podprostor  $M_{i-1}$ , a tudíž i na podprostor  $M_i$ , neboť je  $M_i \subset M_{i-1}$ ).

Protože každý z vektorů  $\mathbf{Y}_{i-1} - \mathbf{Y}_i$  leží v podprostoru  $M_{i-1}$ , znamená to spolu s tvrzením (1.64), že je prvkem podprostoru  $M_{i-1} - M_i$ . Tyto podprostory jsou navzájem kolmé, takže zápis

$$\mathbf{Y} - \mathbf{Y}_r = (\mathbf{Y} - \mathbf{Y}_1) + (\mathbf{Y}_1 - \mathbf{Y}_2) + \dots + (\mathbf{Y}_{r-1} - \mathbf{Y}_r)$$

představuje rozklad vektoru  $\mathbf{Y} - \mathbf{Y}_r$  do  $r$  navzájem kolmých podprostorů a pro všechna  $0 \leq i < j \leq r$  platí obecná varianta Pythagorovy věty

$$\|\mathbf{Y}_i - \mathbf{Y}_j\|^2 = \|\mathbf{Y}_i - \mathbf{Y}_{i+1}\|^2 + \|\mathbf{Y}_{i+1} - \mathbf{Y}_{i+2}\|^2 + \dots + \|\mathbf{Y}_{j-1} - \mathbf{Y}_j\|^2.$$

Kromě toho samozřejmě platí také pro všechna  $i = 1, \dots, r$

$$\|\mathbf{Y} - \mathbf{Y}_i\|^2 = \|\mathbf{Y}\|^2 - \|\mathbf{Y}_i\|^2; \quad (1.65)$$

z obou výše uvedených rovností lze odvodit nepřeberné množství dalších vztahů<sup>7</sup>.

Leží-li nyní skutečná střední hodnota  $\boldsymbol{\mu}$  v podprostoru  $M_r$ , leží v něm též vektor  $\mathbf{Y}_r - \boldsymbol{\mu}$ . Pro  $i \leq r$  je podprostor  $M_r$  kolmý na všechny podprostory  $M_{i-1} - M_i$ , takže pravá strana výrazu

$$\mathbf{Y} - \boldsymbol{\mu} = (\mathbf{Y}_0 - \mathbf{Y}_1) + (\mathbf{Y}_1 - \mathbf{Y}_2) + \dots + (\mathbf{Y}_{r-1} - \mathbf{Y}_r) + (\mathbf{Y}_r - \boldsymbol{\mu})$$

představuje ortogonální rozklad náhodného vektoru  $\mathbf{Y} - \boldsymbol{\mu}$  do  $r+1$  navzájem kolmých podprostorů. Protože tyto kolmé průměty lze libovolně sdružovat, vyplývá z toho podle (1.26), že za předpokladu platnosti hypotézy  $\boldsymbol{\mu} \in M_r$  má náhodná veličina

$$\frac{\|\mathbf{Y}_k - \mathbf{Y}_l\|^2 / (d_k - d_l)}{\|\mathbf{Y}_i - \mathbf{Y}_j\|^2 / (d_i - d_j)}$$

rozdělení  $F_{d_k-d_l, d_i-d_j}$  pro všechna  $i, j, k, l$  taková, že  $0 \leq i < j \leq k < l \leq r$ .

Netroufneme si zabývat se zde otázkou, který z těchto podílů by se měl používat k testování jaké hypotézy. V publikaci [3], kde je rozebrán případ pro  $r = 3$ , je dokazováno, že podíl

$$\frac{\|\mathbf{Y}_2 - \mathbf{Y}_3\|^2 / (d_2 - d_3)}{\|\mathbf{Y} - \mathbf{Y}_1\|^2 / (n - d_1)} = \frac{\|\mathbf{Y}_2 - \mathbf{Y}_3\|^2 / (d_2 - d_3)}{S^2}$$

<sup>7</sup>Vztah (1.65) ovšem obecně nemusí platit v případě, že  $i$ -tý submodel představuje lineární množinu neobsahující nulový vektor, tj. množinu tvaru  $M'_i \equiv \mathbf{a} + M_i$ , kde  $M_i$  je nějaký vektorový podprostor a  $\mathbf{a}$  je vektor, který v něm neleží. Pak je totiž vektor  $\mathbf{Y} - \mathbf{Y}_i$  kolmý na podprostor  $M_i$ , avšak vektor  $\mathbf{Y}_i$  v tomto podprostoru neleží.

má rozdělení  $F_{d_2-d_3, n-d_1}$ . Není nám však jasné, proč by se k testu nulové hypotézy  $\boldsymbol{\mu} \in M_3$  nemohl zrovna tak používat podíl

$$\frac{\|\mathbf{Y}_2 - \mathbf{Y}_3\|^2 / (d_2 - d_3)}{\|\mathbf{Y} - \mathbf{Y}_2\|^2 / (n - d_2)},$$

mající (za předpokladu platnosti nulové hypotézy) rozdělení  $F_{d_2-d_3, n-d_2}$ , což je postup odpovídající situaci, kdy existenci modelu  $M_1$  vůbec nebereme v potaz. Tento postup by měl být vlastně výhodnější, neboť využívá více informací poskytovaných vektorem  $\mathbf{Y}$  a zde použitá statistika má menší rozptyl, neboť  $n - d_2 > n - d_1$ <sup>8</sup>. Tak či onak, vysoká hodnota použité statistiky svědčí o nevhodnosti modelu  $M_3$ , neboť odpovídá situaci, kdy jsou od sebe pravoúhlé průměty  $\mathbf{Y}_2$  a  $\mathbf{Y}_3$  „příliš daleko“, takže redukce podprostoru  $M_2$  na podprostor  $M_3$  představuje relativně podstatnou změnu v odhadu střední hodnoty.

Pro srovnání ještě uveďme sekvenční postup uvedený v knize [35], která se této problematice věnuje důkladněji. Jednou ze zde zmíněných možností pro vhodný výběr submodelu v případě vícenásobné regrese je tzv. vzestupný výběr, tj. postup, kdy vyjdeme z minimálního modelu (předpokládáme, že je míněn model  $\mathbf{Y} = \beta_0 \cdot \mathbf{e} + \mathbf{Z}$ ; pravoúhlým průmětem vektoru  $\mathbf{Y}$  do podprostoru daného tímto modelem je samozřejmě vektor  $\overline{\mathbf{Y}}$ ) a pak přidáme ten sloupec  $\mathbf{x}_i$ , pro který je statistika  $F$  testující jeho odebrání z takto vzniklého modelu maximální; to znamená, že hledáme ten podprostor  $M_i \equiv [\mathbf{e}, \mathbf{x}_i]$ , který maximalizuje hodnotu výrazu

$$\frac{\|\mathbf{Y}_i - \overline{\mathbf{Y}}\|^2 / 1}{\|\mathbf{Y} - \mathbf{Y}_i\|^2 / (n - 2)}.$$

Vybraný podprostor poté rozšíříme na ten podprostor  $M_{i,j} \equiv [\mathbf{e}, \mathbf{x}_i, \mathbf{x}_j]$ , pro který je maximální hodnota výrazu

$$\frac{\|\mathbf{Y}_{i,j} - \mathbf{Y}_i\|^2 / 1}{\|\mathbf{Y} - \mathbf{Y}_{i,j}\|^2 / (n - 3)}$$

atd<sup>9</sup>. To opakujeme tak dlouho, dokud v nějakém kroku neklesnou všechny vypočtené statistiky  $F$  pod nějakou předem určenou hodnotu. V tomto případě tedy sekvence submodelů není předem dána.

## 1.8.2 Systém navzájem kolmých podprostorů

Neurčí-li submodely posloupnost navzájem vnořených podprostorů, je z hlediska jejich statistického zpracování užitečné, tvoří-li tyto podprostory tzv. *Tjurův*

<sup>8</sup>Rozptyl náhodné veličiny mající rozdělení  $F_{m,n}$  je pro  $n > 4$  určen vzorcem

$$\frac{2n^2}{m(n-2)^2} \cdot \frac{m+n-2}{n-4}$$

(viz [3]), což je funkce klesající vzhledem k proměnné  $n$ , neboť je to součin dvou kladných klesajících funkcí.

<sup>9</sup>V obecnějším případě, kdy jsou součástí regresního modelu i sloupce reprezentující faktory (viz příklad 1.2.2), lze samozřejmě přidat i více sloupců najednou.

*system*;<sup>10</sup> zjednodušeně řečeno to znamená, že jsou v jistém obecnějším smyslu navzájem kolmé. Přesnější formulace a obecnější přístup čtenář nalezne v kapitolách 2.2 až 2.4, popřípadě v publikaci [31]; v tuto chvíli se spokojíme se dvěma příklady zásadního významu, na nichž ilustrujeme podstatné myšlenky a výhodnost tohoto uspořádání.

## Příklady

### 1.8.3 Dvojné třídění bez interakcí

Řekněme, že u dvanácti osob změříme hodnoty nějaké fyziologické veličiny. První polovina z pokusných osob jsou muži, druhá polovina ženy. Kromě toho pochází ze tří různých lokalit: první dva muži jsou z lokality  $L_1$ , další dva z lokality  $L_2$ , další dva z lokality  $L_3$  a stejně je tomu u žen. Naměřené veličiny tedy můžeme označit  $Y_{ijk}$ , kde první index označuje pohlaví (1... muži, 2... ženy), druhý index lokalitu a poslední index je pro rozlišení osob stejného pohlaví, které pocházejí ze stejné lokality. Předpokládáme, že jednotlivá měření jsou nezávislá, mají stejný rozptyl a jejich střední hodnota závisí na pohlaví a lokalitě. To vyjádříme vzorcem

$$EY_{ijk} = \mu + \alpha_i + \beta_j,$$

kde  $\mu$  je jistá bazální úroveň, společná všem osobám,  $\alpha_i$  je vliv pohlaví a  $\beta_j$  vliv lokality. Realizace náhodného vektoru  $\mathbf{Y} \equiv (Y_{111}, \dots, Y_{232})^T$  tedy leží ve vektorovém prostoru  $V_{12}$  a jeho chování lze popsat modelem

$$\mathbf{Y} = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \mathbf{Z} \equiv \mathbf{X} \cdot \boldsymbol{\beta} + \mathbf{Z}, \quad (1.66)$$

kde náhodný vektor  $\mathbf{Z}$  má rozdělení  $N(\mathbf{0}, \sigma^2 \mathbf{I}_{12})$ . Označme vektory, které jsou reprezentovány sloupci matice  $\mathbf{X}$ , postupně  $\mathbf{e}$ ,  $\mathbf{a}_1$ ,  $\mathbf{a}_2$ ,  $\mathbf{b}_1$ ,  $\mathbf{b}_2$ ,  $\mathbf{b}_3$ . Dále zavedeme označení pro některé podprostory, které tyto vektory generují:

$$\begin{aligned} E &\equiv [\mathbf{e}], \\ A &\equiv [\mathbf{a}_1, \mathbf{a}_2], \\ B &\equiv [\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3]. \end{aligned}$$

Jejich dimenze jsou zřejmě postupně 1, 2, 3.

<sup>10</sup>Obecnějšími případy se v této práci nebudeme zabývat.

Vektor  $e$  je evidentně lineární kombinací vektorů  $a_1, a_2$ . Sloupce matice  $\mathbf{X}$  jsou tedy lineárně závislé a složky vektoru  $b$ , pro který platí  $\mathbf{Y}_M = \mathbf{X}b$ , nejsou určeny jednoznačně. To znamená mimo jiné to, že soustava (1.16) má nekonečně mnoho řešení, matice  $\mathbf{X}^T\mathbf{X}$  je singulární a k určení vektoru  $b$  nemůžeme použít vzorce (1.17). Vektor  $\mathbf{Y}_M$ , o který se nám nyní primárně jedná, se tedy pokusíme nalézt jiným způsobem, aniž bychom určovali vektor  $b$ .<sup>11</sup>

### Struktura podprostoru $M$

Vektor  $e$  je prvkem jak podprostoru  $A$ , tak podprostoru  $B$ :

$$a_1 + a_2 = e = b_1 + b_2 + b_3.$$

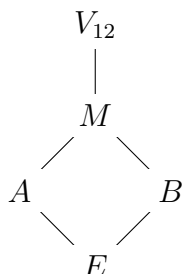
Jiné vektory než jeho násobky však tuto vlastnost mít nemohou – pro dvě různá  $t_1, t_2 \in \mathbb{R}$  totiž nelze vektor tvaru  $t_1 a_1 + t_2 a_2$  sestavit z vektorů  $b_1, b_2, b_3$ . Platí proto

$$E = A \cap B.$$

Protože zřejmě platí také  $M = A + B$ , je dimenze podprostoru  $M$  rovna hodnotě

$$\begin{aligned} \dim M &= \dim(A + B) = \\ &= \dim A + \dim B - \dim(A \cap B) = \\ &= 2 + 3 - 1 = \\ &= 4. \end{aligned}$$

Vztahy mezi zúčastněnými podprostory lze přehledně znázornit pomocí schématu, v němž čára od níže položeného podprostoru  $X$  k výše položenému podprostoru  $Y$  symbolizuje relaci  $X \subset Y$ <sup>12</sup>:



### Rozklad podprostoru $M$

Pokusme se nyní podprostor  $M$  rozložit na součet disjunktních podprostorů  $E$ ,  $A - E$  a  $B - E$ . Nahraďme za tím účelem sloupce  $a_1, a_2, b_1, b_2, b_3$  v matici  $\mathbf{X}$  sloupci

$$\begin{aligned} a^* &\equiv a_1 - a_2, \\ b_1^* &\equiv b_1 - b_2, \\ b_2^* &\equiv b_2 - b_3; \end{aligned}$$

<sup>11</sup>Obvykle se tato překážka řeší přidáním tzv. *reparametrizačních rovnic*, které nějakým způsobem vážou složky vektoru  $\beta$  a tak zajišťují jejich jednoznačnost. Podrobněji se těmto metodám věnujeme v kapitolách 1.11 a 2.9.

<sup>12</sup>Tento způsob znázornění je převzat z publikace [31]; podrobnější výklad o jeho použití lze nalézt v kapitole 2.4.

dostaneme tak matici

$$\begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & 0 & -1 \\ 1 & -1 & 1 & 0 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 0 & -1 \\ 1 & -1 & 0 & -1 \end{pmatrix}. \quad (1.67)$$

Vektor  $\mathbf{a}^*$  je kolmý na podprostor  $E$ , tyto dva vektory jsou tedy lineárně nezávislé a dimenze jejich lineárního obalu je 2. Přitom jsou oba tyto vektory prvkem podprostoru  $A$ , jehož dimenze je rovněž 2, platí tedy

$$[\mathbf{e}, \mathbf{a}^*] = A.$$

Podobně vektory  $\mathbf{b}_1^*, \mathbf{b}_2^*$  jsou navzájem nezávislé a zároveň jsou oba kolmé na vektor  $\mathbf{e}$ , vektory  $\mathbf{e}, \mathbf{b}_1^*, \mathbf{b}_2^*$  jsou tedy lineárně nezávislé a dimenze jejich lineárního obalu je 3; protože jsou všechny tyto vektory prvky podprostoru  $B$ , jehož dimenze je rovněž tři, platí

$$[\mathbf{e}, \mathbf{b}_1^*, \mathbf{b}_2^*] = B.$$

Vzhledem k výše uvedeným kolmostem pak dostáváme

$$\begin{aligned} [\mathbf{a}^*] &= A - E, \\ [\mathbf{b}_1^*, \mathbf{b}_2^*] &= B - E. \end{aligned}$$

Nyní si ale můžeme povšimnout překvapující skutečnosti – vektor  $\mathbf{a}^*$  je kolmý na vektory  $\mathbf{b}_1^*, \mathbf{b}_2^*$ . Rozložili jsme tedy podprostor  $M$  na součet *tří navzájem kolmých podprostorů*:

$$M = E + (A - E) + (B - E).$$

### Pravoúhlý průmět do podprostoru $M$

Značme nadále symbolem  $\mathbf{Y}_X$  pravoúhlý průmět vektoru  $\mathbf{Y}$  do podprostoru  $X$ . Vzhledem k výše uvedenému rozkladu podprostoru  $M$  do tří navzájem kolmých podprostorů můžeme vektor  $\mathbf{Y}_M$  vyjádřit jako

$$\mathbf{Y}_M = \mathbf{Y}_E + \mathbf{Y}_{A-E} + \mathbf{Y}_{B-E};$$

protože je však  $E \subset A$ ,  $E \subset B$ , platí

$$\begin{aligned} \mathbf{Y}_{A-E} &= \mathbf{Y}_A - \mathbf{Y}_E, \\ \mathbf{Y}_{B-E} &= \mathbf{Y}_B - \mathbf{Y}_E \end{aligned}$$

(viz vzorce (2.41) a (2.42)). Průměty  $\mathbf{Y}_A$ ,  $\mathbf{Y}_B$  vypočteme stejně v případě jednoduchého třídění podle pohlaví, resp. lokality (viz příklad 1.3.5); průmětem  $\mathbf{Y}_E$  je samozřejmě vektor  $\bar{\mathbf{Y}}$ :

$$\mathbf{Y}_A = \begin{pmatrix} \bar{Y}_{1..} \\ \bar{Y}_{1..} \\ \bar{Y}_{1..} \\ \bar{Y}_{1..} \\ \bar{Y}_{1..} \\ \bar{Y}_{1..} \\ \bar{Y}_{2..} \\ \bar{Y}_{2..} \\ \bar{Y}_{2..} \\ \bar{Y}_{2..} \\ \bar{Y}_{2..} \\ \bar{Y}_{2..} \end{pmatrix}, \quad \mathbf{Y}_B = \begin{pmatrix} \bar{Y}_{.1} \\ \bar{Y}_{.1} \\ \bar{Y}_{.2} \\ \bar{Y}_{.2} \\ \bar{Y}_{.3} \\ \bar{Y}_{.3} \\ \bar{Y}_{.1} \\ \bar{Y}_{.1} \\ \bar{Y}_{.2} \\ \bar{Y}_{.2} \\ \bar{Y}_{.3} \\ \bar{Y}_{.3} \end{pmatrix}, \quad \mathbf{Y}_E = \bar{\mathbf{Y}} = \begin{pmatrix} \bar{Y}_{...} \\ \bar{Y}_{...} \\ \bar{Y}_{...} \\ \bar{Y}_{...} \\ \bar{Y}_{...} \\ \bar{Y}_{...} \\ \bar{Y}_{...} \\ \bar{Y}_{...} \\ \bar{Y}_{...} \\ \bar{Y}_{...} \\ \bar{Y}_{...} \\ \bar{Y}_{...} \end{pmatrix},$$

kde

$$\bar{Y}_{i..} \equiv \frac{\sum_{j,k} Y_{ijk}}{6}, \quad \bar{Y}_{.j.} \equiv \frac{\sum_{i,k} Y_{ijk}}{4}, \quad \bar{Y}_{...} \equiv \frac{\sum_{i,j,k} Y_{ijk}}{12}.$$

Tj.  $\bar{Y}_{1..}$  je průměr měření získaných od mužů,  $\bar{Y}_{2..}$  průměr měření získaných od žen,  $\bar{Y}_{.1}$  je průměr měření získaných od osob z lokality  $L_1$ ,  $\bar{Y}_{.2}$  průměr měření získaných od osob z lokality  $L_2$  atd. Pravoúhlým průmětem náhodného vektoru  $\mathbf{Y}$  do podprostoru  $M$  je pak vektor

$$\begin{aligned} \mathbf{Y}_M &= (\mathbf{Y}_A - \bar{\mathbf{Y}}) + (\mathbf{Y}_B - \bar{\mathbf{Y}}) + \bar{\mathbf{Y}} = \\ &= \mathbf{Y}_A + \mathbf{Y}_B - \bar{\mathbf{Y}}. \end{aligned} \tag{1.68}$$

Z rovnosti (1.68) plynou mimochodem užitečné vztahy

$$\begin{aligned} \mathbf{Y}_M - \mathbf{Y}_A &= \mathbf{Y}_B - \bar{\mathbf{Y}}, \\ \mathbf{Y}_M - \mathbf{Y}_B &= \mathbf{Y}_A - \bar{\mathbf{Y}}, \end{aligned} \tag{1.69}$$

jejichž geometrický význam lze snadno znázornit (viz obr. 1.22).<sup>13</sup>

### Použití Pythagorovy věty

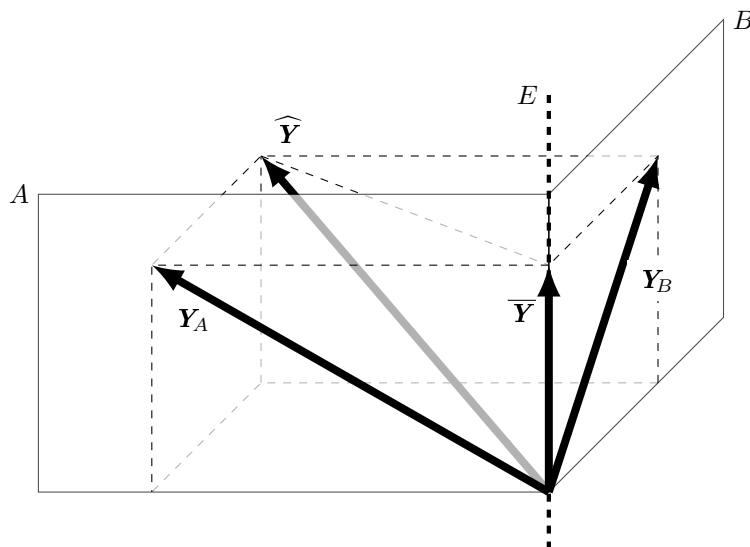
Protože platí

$$\begin{aligned} \mathbf{Y} - \mathbf{Y}_M &\in V_{12} - M, \\ \mathbf{Y}_M - \mathbf{Y}_B &= \mathbf{Y}_A - \bar{\mathbf{Y}} \in A - E, \\ \mathbf{Y}_B - \bar{\mathbf{Y}} &= \mathbf{Y}_M - \mathbf{Y}_A \in B - E, \\ \bar{\mathbf{Y}} &\in E, \end{aligned}$$

<sup>13</sup>Tyto vztahy jsou ekvivalentní se skutečností, že platí

$$\begin{aligned} M - A &= B - E, \\ M - B &= A - E, \end{aligned}$$

což lze ověřit krátkou úvahou nad sloupci matice (1.67).



**Obrázek 1.22:** Znázornění vztahů mezi pravoúhlými průměty do podprostorů  $A$ ,  $B$ ,  $M = A+B$  a  $E = A \cap B$ , kde  $A$ ,  $B$  reprezentují dva různé faktory v případě vyváženého dvojnásobného třídění. Jelikož máme v náčrtu k dispozici pouze tři dimenze, jsou podprostory  $A$ ,  $B$  znázorněny jako roviny, obecně však mohou být i vyšší dimenze. (Vektor  $\mathbf{Y} - \widehat{\mathbf{Y}}$ , který je kolmý na všechny znázorněné podprostory, se už do obrázku „nevešel“.)

a zároveň

$$(\mathbf{Y} - \mathbf{Y}_M) + (\mathbf{Y}_M - \mathbf{Y}_B) + (\mathbf{Y}_B - \overline{\mathbf{Y}}) + \overline{\mathbf{Y}} = \mathbf{Y},$$

tvoří tyto čtyři vektory rozklad vektoru  $\mathbf{Y}$  do čtyř navzájem kolmých podprostorů  $V_{12} - M$ ,  $A - E$ ,  $B - E$  a  $E$ . To umožňuje vytvořit značné množství variant Pythagorovy věty; tak například součtem druhého a čtvrtého vektoru je vektor  $\mathbf{Y}_A$ , který musí být kolmý na třetí vektor  $\mathbf{Y}_M - \mathbf{Y}_A$ ; z toho plyne

$$\begin{aligned} \|\mathbf{Y}_A\|^2 + \|\mathbf{Y}_M - \mathbf{Y}_A\|^2 &= \|\mathbf{Y}_A + (\mathbf{Y}_M - \mathbf{Y}_A)\|^2 = \\ &= \|\mathbf{Y}_M\|^2. \end{aligned}$$

Mnoho takto odvoditelných rovností bylo již ovšem zmíněno dříve. Uveďme tedy bez dalšího odvozování jen některé zajímavější:

$$\begin{aligned} \|\mathbf{Y}_M\|^2 - \|\mathbf{Y}_A\|^2 &= \|\mathbf{Y}_M - \mathbf{Y}_A\|^2 = \|\mathbf{Y}_B - \overline{\mathbf{Y}}\|^2 = \|\mathbf{Y}_B\|^2 - \|\overline{\mathbf{Y}}\|^2, \\ \|\mathbf{Y}_M\|^2 - \|\mathbf{Y}_B\|^2 &= \|\mathbf{Y}_M - \mathbf{Y}_B\|^2 = \|\mathbf{Y}_A - \overline{\mathbf{Y}}\|^2 = \|\mathbf{Y}_A\|^2 - \|\overline{\mathbf{Y}}\|^2, \\ \|\mathbf{Y}_M\|^2 &= \|\mathbf{Y}_A\|^2 + \|\mathbf{Y}_B\|^2 - \|\overline{\mathbf{Y}}\|^2. \end{aligned}$$

### Testy submodelů

Chceme-li posoudit, zdali je vliv pohlaví na měřenou veličinu statisticky významný, zkusíme jej z modelu (1.66) vynechat. Vypuštěním sloupců  $\mathbf{a}_1$ ,  $\mathbf{a}_2$  ze skupiny generátorů redukuje podprostor  $M$  na podprostor  $B$ . Leží-li skutečná střední hodnota v tomto podprostoru (což odpovídá absenci vlivu pohlaví), má statistika

$$\frac{\|\mathbf{Y}_M - \mathbf{Y}_B\|^2 / (\dim M - \dim B)}{\|\mathbf{Y} - \mathbf{Y}_M\|^2 / (\dim V_{12} - \dim M)} = \frac{\|\mathbf{Y}_M - \mathbf{Y}_B\|^2 / 1}{\|\mathbf{Y} - \mathbf{Y}_M\|^2 / 8} \quad (1.70)$$

podle (1.36) rozdělení  $F_{1,8}$ . Vliv pohlaví tedy bude potvrzen na hladině významnosti  $\alpha$ , pokud tato statistika bude vyšší než hodnota  $F_{1,8}(\alpha)$ . Podobně vliv lokality budeme testovat srovnáním původního modelu se submodelem nezahrnujícím vliv lokality, tj. submodelem obsahujícím pouze vliv lokality, který je reprezentován podprostorem  $A$ . Použijeme tedy statistiku

$$\frac{\|\mathbf{Y}_M - \mathbf{Y}_A\|^2 / (\dim M - \dim A)}{\|\mathbf{Y} - \mathbf{Y}_M\|^2 / (\dim V_{12} - \dim M)} = \frac{\|\mathbf{Y}_M - \mathbf{Y}_A\|^2 / 2}{\|\mathbf{Y} - \mathbf{Y}_M\|^2 / 8},$$

kteřá má v případě absence vlivu lokality rozdělení  $F_{2,8}$ . K výpočtu čitatele lze použít vztahy uvedené v předchozím odstavci. Jmenovatel v obou statistikách představuje nestranný odhad parametru  $\sigma^2$ , v této práci označovaný  $S^2$  (viz kapitola 1.5).

Užitečným důsledkem vztahů (1.69) je skutečnost, že čítel obou statistik je stejný, jako kdybychom vliv daného faktoru testovali v modelu, ve kterém druhý faktor vůbec nebereme v úvahu, tj. jako kdybychom pro test vlivu pohlaví (reprezentovaný podprostorem  $A$ ) používali statistiku

$$\frac{\|\mathbf{Y}_A - \bar{\mathbf{Y}}\|^2 / (\dim A - \dim E)}{\|\mathbf{Y} - \mathbf{Y}_M\|^2 / (\dim V_{12} - \dim M)},$$

resp. pro test vlivu lokality (reprezentovaný podprostorem  $B$ ) statistiku

$$\frac{\|\mathbf{Y}_B - \bar{\mathbf{Y}}\|^2 / (\dim B - \dim E)}{\|\mathbf{Y} - \mathbf{Y}_M\|^2 / (\dim V_{12} - \dim M)}. \quad (1.71)$$

To je příznivá situace, neboť v opačném případě by mohla vzniknout pochybnost, která z obou metod je vhodnější. Díky tomuto uspořádání také může být postup testování formulován tak, že nejprve z původního modelu odebereme vliv jednoho faktoru, řekněme pohlaví, čímž redukuje podprostor  $M$  na podprostor  $B$ . Vhodnost této redukce ověříme užitím statistiky (1.70), a poté ze zbylého submodelu odebereme vliv druhého faktoru, tj. lokality, čímž redukuje podprostor  $B$  na podprostor  $E$ ; to otestujeme užitím statistiky (1.71). Takový postup lze považovat za korektní, protože právě díky vztahům (1.69) nezáleží na pořadí, v jakém vliv faktorů odstraňujeme.

Zdůrazněme, že výše uvedené vlastnosti jsou důsledkem rovnosti (1.68), která plyne z toho, že podprostory  $A - E$ ,  $B - E$  jsou navzájem kolmé. Tato kolmost je patrná z jejich generátorů, jak jsou uvedeny v matici (1.67). Podoba této matice je pro tento závěr příznivá díky tomu, že třídění bylo *vyvážené*, tj. že pro každou kombinaci pohlaví a lokality jsme měli k dispozici stejný počet měření. To však není nutná podmínka; podprostory  $A - E$ ,  $B - E$  a  $E$  mohou být navzájem kolmé za obecnějších okolností (podrobněji viz [31]).

### Konkrétní hodnoty

Pro názornost ještě demonstrováme výše odvozené postupy na konkrétních hodnotách. Nechť tedy máme k dispozici realizaci náhodného vektoru

$$\mathbf{y} = (49, 47, 46, 42, 40, 42, 36, 32, 33, 35, 40, 38)^T. \quad (1.72)$$

V následující tabulce je uveden přehled těchto hodnot tak, jak odpovídají úrovním jednotlivých faktorů, včetně průměrů pro tyto úrovně (v pravém dolním rohu

je průměr ze všech měření):

	lokalita 1	lokalita 2	lokalita 3	∅
muži	49; 47	46; 42	40; 42	44 <sup>1</sup> / <sub>3</sub>
ženy	36; 32	33; 35	40; 38	35 <sup>2</sup> / <sub>3</sub>
∅	41	39	40	40

Pravoúhlými průměty získané realizace náhodného vektoru  $\mathbf{Y}$  do podprostorů  $A$ ,  $B$ ,  $E$  a  $M$  jsou tedy vektory

$$\mathbf{y}_A = \begin{pmatrix} 44^{1/3} \\ 44^{1/3} \\ 44^{1/3} \\ 44^{1/3} \\ 44^{1/3} \\ 44^{1/3} \\ 35^{2/3} \\ 35^{2/3} \\ 35^{2/3} \\ 35^{2/3} \\ 35^{2/3} \\ 35^{2/3} \\ 35^{2/3} \end{pmatrix}, \quad \mathbf{y}_B = \begin{pmatrix} 41 \\ 41 \\ 39 \\ 39 \\ 40 \\ 40 \\ 41 \\ 41 \\ 39 \\ 39 \\ 40 \\ 40 \end{pmatrix}, \quad \bar{\mathbf{y}} = \begin{pmatrix} 40 \\ 40 \\ 40 \\ 40 \\ 40 \\ 40 \\ 40 \\ 40 \\ 40 \\ 40 \\ 40 \\ 40 \end{pmatrix}, \quad \mathbf{y}_M = \begin{pmatrix} 45^{1/3} \\ 45^{1/3} \\ 43^{1/3} \\ 43^{1/3} \\ 44^{1/3} \\ 44^{1/3} \\ 36^{2/3} \\ 36^{2/3} \\ 34^{2/3} \\ 34^{2/3} \\ 35^{2/3} \\ 35^{2/3} \end{pmatrix}, \quad (1.73)$$

kde k výpočtu posledního vektoru jsme použili vzorec (1.68). K posouzení vlivu pohlaví vypočteme

$$\frac{\|\mathbf{y}_A - \bar{\mathbf{y}}\|^2/1}{\|\mathbf{y} - \mathbf{y}_M\|^2/8} = \frac{225,33}{98,67/8} = 18,27;$$

jelikož tato statistika má mít v případě absence vlivu pohlaví rozdělení  $F_{1,8}$  a kritická hodnota tohoto rozdělení na hladině 0,05 je 5,32, můžeme považovat vliv pohlaví za statisticky průkazný.

Pro obdobný test vlivu lokality vypočteme hodnotu statistiky

$$\frac{\|\mathbf{y}_B - \bar{\mathbf{y}}\|^2/2}{\|\mathbf{y} - \mathbf{y}_M\|^2/8} = \frac{8/2}{98,67/8} = 0,32,$$

která má mít v případě, že lokality žádný vliv nemají, rozdělení  $F_{2,8}$ . Protože kritická hodnota tohoto rozdělení na hladině 0,05 je 4,46, nelze na základě získaných dat prokázat vliv lokality na hodnotu sledované veličiny.

Dodejme ještě, že hodnota

$$S^2 = \|\mathbf{y} - \mathbf{y}_M\|^2/8 = 12,33$$

je nestranným odhadem rozptylu  $\sigma^2$ .

#### 1.8.4 Dvojné třídění s interakcemi

Zůstaňme u předchozí situace, ale tentokrát zahrňme do modelu i *interakce*, tj. předpoklad, že vzájemné působení vlivů pohlaví a lokality je komplikovanější než jejich pouhý součet. To lze vyjádřit vzorcem

$$EY_{ijk} = \mu + \alpha_i + \beta_j + \lambda_{ij},$$

kde parametr  $\lambda_{ij}$  představuje hodnotu, která je důsledkem interakce pohlaví  $i$  a lokality  $j$ . Za těchto okolností popisuje chování náhodného vektoru  $\mathbf{Y}$  model

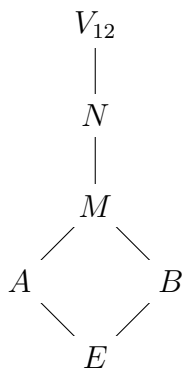
$$\mathbf{Y} = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \lambda_{11} \\ \lambda_{12} \\ \lambda_{13} \\ \lambda_{21} \\ \lambda_{22} \\ \lambda_{23} \end{pmatrix} + \mathbf{Z} \equiv$$

$$\equiv \mathbf{X}^* \cdot \boldsymbol{\beta} + \mathbf{Z}, \quad (1.74)$$

kde náhodný vektor  $\mathbf{Z}$  má opět rozdělení  $\mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{12})$ . Označme posledních šest sloupců matice  $\mathbf{X}^*$  symboly  $\mathbf{c}_{11}, \mathbf{c}_{12}, \mathbf{c}_{13}, \mathbf{c}_{21}, \mathbf{c}_{22}, \mathbf{c}_{23}$ , jejich lineární obal symbolem  $C$  a podprostor určený modelem (1.74) symbolem  $N$ . Ostatní označení vektorů a podprostorů ponechme stejné jako v příkladu 1.8.3.

### Struktura podprostoru $N$

Vektory  $\mathbf{c}_{ij}$  jsou zřejmě lineárně nezávislé a jejich kombinací lze vyjádřit jakýkoli z ostatních vektorů, které představují sloupce matice  $\mathbf{X}^*$ ; je proto  $N = C$  a dimenze podprostoru  $N$  je  $2 \times 3 = 6$ . Podprostor  $M = A + B$  je podmnožinou podprostoru  $N$ . Vztahy mezi zúčastněnými podprostory můžeme znázornit pomocí schématu:



Podprostor  $M$  lze opět rozložit na součet tří navzájem kolmých podprostorů  $A - E$ ,  $B - E$  a  $E$ .

### Pravoúhlý průmět do podprostoru $N$

Jelikož podprostor  $N$  je generován vektory  $\mathbf{c}_{ij}$ , lze jej interpretovat jako podprostor určený modelem jednoduchého třídění, obsahujícího jediný faktor o šesti

úrovních; pravoúhlý průmět náhodného vektoru  $\mathbf{Y}$  do podprostoru  $N$  je tedy určen vztahem

$$\mathbf{Y}_N = (\bar{Y}_{11\cdot}, \bar{Y}_{11\cdot}, \bar{Y}_{12\cdot}, \bar{Y}_{12\cdot}, \bar{Y}_{13\cdot}, \bar{Y}_{13\cdot}, \bar{Y}_{21\cdot}, \bar{Y}_{21\cdot}, \bar{Y}_{22\cdot}, \bar{Y}_{22\cdot}, \bar{Y}_{32\cdot}, \bar{Y}_{32\cdot})^T,$$

kde

$$\bar{Y}_{ij\cdot} \equiv \frac{\sum_k Y_{ijk}}{2},$$

tj.  $\bar{Y}_{ij\cdot}$  je průměr z měření všech osob stejného pohlaví  $i$ , které pochází ze stejné lokality  $L_j$ .

### Testy submodelů

Nejdříve testujeme hypotézu, že interakce jsou nulové, tj. že vzájemný vliv pohlaví a lokality je aditivní; musíme tedy zjistit, zda odstranění interakcí z modelu (1.74) způsobí významnou změnu v pravoúhlém průmětu vektoru  $\mathbf{Y}$  do podprostoru daného modelem. Pokud interakce vynecháme, redukuje podprostor  $N$  na podprostor  $M$ . K testu uvažované hypotézy tedy použijeme statistiku

$$\frac{\|\mathbf{Y}_N - \mathbf{Y}_M\|^2 / (\dim N - \dim M)}{\|\mathbf{Y} - \mathbf{Y}_N\|^2 / (\dim V_{12} - \dim N)} = \frac{\|\mathbf{Y}_N - \mathbf{Y}_M\|^2 / (6 - 4)}{\|\mathbf{Y} - \mathbf{Y}_N\|^2 / (12 - 6)},$$

kteřá má v případě platnosti testované hypotézy rozdělení  $F_{2,6}$ . Hypotézu tedy zamítneme na hladině významnosti  $\alpha$ , pokud získaná realizace této statistiky překročí hodnotu  $F_{2,6}(\alpha)$ .

Jak budeme postupovat v případě testu hypotézy

$$\alpha_1 = \alpha_2 = 0,$$

tj. hypotézy, že vliv pohlaví je nulový? Pokud bychom z matice  $\mathbf{X}^*$  modelu (1.74) vyloučili pouze sloupce  $\mathbf{a}_i$ , nemělo by to žádný podstatný efekt, neboť sloupce  $\mathbf{c}_{ij}$ , které by v matici zbyly, generují tentýž podprostor jako původní sloupce. Na druhou stranu, pokud bychom vynechali jak sloupce  $\mathbf{a}_i$ , tak  $\mathbf{c}_{ij}$ , testovali bychom vlastně hypotézu „vliv pohlaví a interakce jsou nulové“, což není zcela přesně to, co chceme. Proto pro test této hypotézy použijeme statistiku

$$\frac{\|\mathbf{Y}_M - \mathbf{Y}_B\|^2 / (\dim M - \dim B)}{\|\mathbf{Y} - \mathbf{Y}_N\|^2 / (\dim V_{12} - \dim N)} = \frac{\|\mathbf{Y}_M - \mathbf{Y}_B\|^2 / (4 - 3)}{\|\mathbf{Y} - \mathbf{Y}_N\|^2 / (12 - 6)}, \quad (1.75)$$

tj. odstraníme sloupce představující vliv pohlaví ze submodelu zahrnujícího pouze vliv pohlaví a vliv lokality, nikoli interakce (redukuje tedy podprostor  $M$  na podprostor  $B$ ); rozdíl v pravoúhlém průmětu vektoru  $\mathbf{Y}$ , který tato redukce způsobí, porovnáváme s odhadem rozptylu vycházejícím z původního modelu (1.74). Až na jmenovatel je to tedy stejný vzorec jako (1.70). Vliv pohlaví lze považovat za statisticky průkazný na hladině významnosti  $\alpha$ , pokud tato statistika překročí hodnotu  $F_{1,6}(\alpha)$ .

Podobně pro test vlivu lokality použijeme statistiku

$$\frac{\|\mathbf{Y}_M - \mathbf{Y}_A\|^2 / (\dim M - \dim A)}{\|\mathbf{Y} - \mathbf{Y}_N\|^2 / (\dim V_{12} - \dim N)} = \frac{\|\mathbf{Y}_M - \mathbf{Y}_A\|^2 / (4 - 2)}{\|\mathbf{Y} - \mathbf{Y}_N\|^2 / (12 - 6)}$$

a tento vliv budeme považovat za průkazný tehdy, když její získaná realizace překročí hodnotu  $F_{2,6}(\alpha)$ .

Při výpočtu můžeme samozřejmě použít kterýkoli ze vztahů uvedených v příkladu 1.8.3, případně další vztahy plynoucí z Pythagorovy věty:

$$\begin{aligned}\|\mathbf{Y} - \mathbf{Y}_N\|^2 &= \|\mathbf{Y}\|^2 - \|\mathbf{Y}_N\|^2, \\ \|\mathbf{Y}_N - \mathbf{Y}_M\|^2 &= \|\mathbf{Y}_N\|^2 - \|\mathbf{Y}_M\|^2.\end{aligned}$$

### Konkrétní hodnoty

Použijme opět realizaci (1.72). Její pravoúhlé průměty do podprostorů  $M$ ,  $A$ ,  $B$  a  $E$  již známe (viz (1.73)), zbývá určit průmět do podprostoru  $N$ :

$$\mathbf{y}_N = (48, 48, 44, 44, 41, 41, 34, 34, 34, 34, 39, 39)^T.$$

Hodnota statistiky  $F$  pro test hypotézy, že interakce jsou nulové, je tedy

$$\frac{\|\mathbf{y}_N - \mathbf{y}_M\|^2/2}{\|\mathbf{y} - \mathbf{y}_N\|^2/6} = 9,33;$$

protože kritická hodnota rozdělení  $F_{2,6}(0,05)$  je 5,14, můžeme považovat vliv interakcí za prokázaný.

Hodnoty statistik pro test vlivu pohlaví, resp. lokality, jsou pak

$$\frac{\|\mathbf{y}_M - \mathbf{y}_B\|^2/1}{\|\mathbf{y} - \mathbf{y}_N\|^2/6} = 56,33, \quad (1.76)$$

resp.

$$\frac{\|\mathbf{y}_M - \mathbf{y}_A\|^2/2}{\|\mathbf{y} - \mathbf{y}_N\|^2/6} = 1,00; \quad (1.77)$$

vliv pohlaví je tedy statisticky průkazný, zatímco vliv lokality nikoli (příslušné kritické hodnoty pro 5% hladinu významnosti jsou 5,99 a 5,14). Jelikož jsme však již prokázali vliv interakcí, nemá nyní smysl vliv lokality zpochybňovat.<sup>14</sup>

### 1.8.5 Dvojné třídění – obecná formulace

Věnujme se nyní zobecnění výše popsaných metod, a to na případ dvojného vyváženého třídění podle dvou faktorů o libovolných počtech úrovní. Nebudeme již opakovat argumentaci, pomocí které jsme je odvodili v předchozích speciálních případech; je ale důležité si uvědomit, že všechny úvahy, které jsme při té příležitosti provedli, lze aplikovat i v obecné situaci.

Mějme tedy náhodný vektor, jehož realizace leží ve vektorovém prostoru  $V_n$  a jehož souřadnice jsou  $Y_{ijk}$ , kde  $i \in \{1; \dots; I\}$ ,  $j \in \{1; \dots; J\}$  a  $k \in \{1; \dots; K\}$ , tj.  $n = IJK$ . Nechť pro tyto souřadnice v případě modelu bez interakcí platí

$$Y_{ijk} = \mu + \alpha_i + \beta_j + Z_{ijk},$$

<sup>14</sup>Interpretace výsledků statistických testů není předmětem našeho zájmu; poznamenejme však, že pokud by rozsah modelu nebyl předem dán a chtěli bychom se rozhodovat, které podprostory do něj zahrneme, mohli bychom dojít k odlišnému závěru – z hodnot statistik (1.76) a (1.77) bychom usoudili, že vliv pohlaví je třeba do modelu zahrnout, ale vliv lokality nikoli, a uvažovat o interakcích by pak nemělo smysl.

v případě modelu s interakcemi platí

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \lambda_{ij} + Z_{ijk},$$

kde náhodné veličiny  $Z_{ijk}$  jsou nezávislé a řídí se normálním rozdělením s nulovou střední hodnotou a stejným rozptylem  $\sigma^2$ . Koeficienty  $\alpha_i$  představují vliv faktoru  $A$ , který má  $I$  úrovní, koeficienty  $\beta_j$  představují vliv faktoru  $B$  o  $J$  úrovních; pro každou kombinaci úrovní máme k dispozici stejný počet měření  $K$ . Koeficienty  $\lambda_{ij}$  představují vliv interakcí, které můžeme interpretovat jako další faktor, řekněme  $C$ , který má  $IJ$  úrovní. Faktor  $A$  je tedy v matici modelu zastoupen  $I$  sloupci; podprostor, který generují, označme rovněž symbolem  $A$ . Podobně je faktor  $B$  zastoupen  $J$  sloupci generujícími podprostor  $B$ . Koeficient  $\mu$  představuje vliv bazální úrovně, která je společná pro všechny souřadnice vektoru  $\mathbf{Y}$ ; v modelu je zastoupen jediným vektorem  $(1, \dots, 1)^T$ , který generuje podprostor  $E$ . Zahrnujeme-li do modelu i interakce, jsou v něm zastoupeny  $IJ$  sloupci, které generují podprostor  $C$ . Dimenze těchto podprostorů jsou

$$\begin{aligned} \dim E &= 1, \\ \dim A &= I, \\ \dim B &= J, \\ \dim C &= IJ. \end{aligned}$$

Přitom platí, že podprostor  $E$  je průnikem podprostorů  $A$  a  $B$  a podprostory  $E$ ,  $A$  a  $B$  jsou vlastními podmnožinami podprostoru  $C$ .

### Model bez interakcí

V případě, že nepočítáme s vlivem interakcí, je modelem určen podprostor  $M \equiv E + A + B = A + B$ , jehož dimenze je

$$\begin{aligned} \dim M &= \dim A + \dim B - \dim(A \cap B) = \\ &= I + J - 1. \end{aligned}$$

Pravoúhlým průmětem vektoru  $\mathbf{Y}$  do tohoto podprostoru je vektor

$$\mathbf{Y}_M = \mathbf{Y}_A + \mathbf{Y}_B - \bar{\mathbf{Y}}.$$

Souřadnice s indexem  $ijk$  vektoru  $\mathbf{Y}_M$  je tedy rovna hodnotě

$$\bar{Y}_{i..} + \bar{Y}_{.j.} - \bar{Y}_{...},$$

kde

$$\bar{Y}_{i..} \equiv \frac{\sum_{j,k} Y_{ijk}}{JK}, \quad \bar{Y}_{.j.} \equiv \frac{\sum_{i,k} Y_{ijk}}{IK}, \quad \bar{Y}_{...} \equiv \frac{\sum_{i,j,k} Y_{ijk}}{IJK}.$$

Významnost vlivu faktoru  $A$  posoudíme pomocí statistiky

$$\frac{\frac{\|\mathbf{Y}_M - \mathbf{Y}_B\|^2}{\dim A - \dim E}}{\frac{\|\mathbf{Y} - \mathbf{Y}_M\|^2}{\dim V_n - \dim M}} = \frac{\frac{\|\mathbf{Y}_A - \bar{\mathbf{Y}}\|^2}{I - 1}}{\frac{\|\mathbf{Y} - \mathbf{Y}_M\|^2}{n - I - J + 1}}.$$

Zdroj variability	Součet čtverců	Počet stupňů volnosti	Podíl	F
Faktor $A$	$\ \mathbf{Y}_A - \bar{\mathbf{Y}}\ ^2$	$I - 1$	$\frac{\ \mathbf{Y}_A - \bar{\mathbf{Y}}\ ^2}{I - 1}$	$\frac{\frac{\ \mathbf{Y}_A - \bar{\mathbf{Y}}\ ^2}{I - 1}}{\frac{\ \mathbf{Y} - \mathbf{Y}_M\ ^2}{n - I - J + 1}}$
Faktor $B$	$\ \mathbf{Y}_B - \bar{\mathbf{Y}}\ ^2$	$J - 1$	$\frac{\ \mathbf{Y}_B - \bar{\mathbf{Y}}\ ^2}{J - 1}$	$\frac{\frac{\ \mathbf{Y}_B - \bar{\mathbf{Y}}\ ^2}{J - 1}}{\frac{\ \mathbf{Y} - \mathbf{Y}_M\ ^2}{n - I - J + 1}}$
Reziduální	$\ \mathbf{Y} - \mathbf{Y}_M\ ^2$	$n - I - J + 1$	$\frac{\ \mathbf{Y} - \mathbf{Y}_M\ ^2}{n - I - J + 1}$	–

**Tabulka 1.4:** Tabulka analýzy rozptylu pro případ dvojného třídění bez interakcí.

V případě, že vliv faktoru  $A$  je nulový, má (nezávisle na tom, jaký je vliv faktoru  $B$ ) tato statistika rozdělení  $F_{I-1, n-I-J+1}$ ; je-li tedy její realizace větší než příslušná kritická hodnota, můžeme vliv faktoru  $A$  považovat za prokázaný. Podobně vliv faktoru  $B$  budeme testovat pomocí hodnoty

$$\frac{\frac{\|\mathbf{Y}_M - \mathbf{Y}_A\|^2}{\dim B - \dim E}}{\frac{\|\mathbf{Y} - \mathbf{Y}_M\|^2}{\dim V_n - \dim M}} = \frac{\frac{\|\mathbf{Y}_B - \bar{\mathbf{Y}}\|^2}{J - 1}}{\frac{\|\mathbf{Y} - \mathbf{Y}_M\|^2}{n - I - J + 1}},$$

která má v případě absence vlivu faktoru  $B$  rozdělení  $F_{J-1, n-I-J+1}$ . Jmenovatel v obou uvedených výrazech představuje nestranný odhad rozptylu  $\sigma^2$ . Výsledky jsou zpravidla shrnuty ve formě tabulky analýzy rozptylu (viz tab. 1.4).

### Model s interakcemi

Pokud je do počátečního modelu zahrnut i vliv interakcí, lze tyto interakce interpretovat jako třetí faktor, jehož každá hladina odpovídá určité kombinaci hladin faktorů  $A$  a  $B$ . Těchto kombinací je  $IJ$ , takže interakce jsou v matici modelu zastoupeny  $IJ$  sloupci, které generují podprostor  $C$  dimenze  $IJ$ ; protože všechny ostatní sloupce jsou jejich lineárními kombinacemi, platí pro podprostor  $N$  určený modelem

$$N = E + A + B + C = C.$$

Pravoúhlým průmětem náhodného vektoru  $\mathbf{Y}$  do tohoto podprostoru je vektor  $\mathbf{Y}_N$ , jehož souřadnice s indexem  $ijk$  je rovna hodnotě

$$\bar{Y}_{ij\cdot} \equiv \frac{\sum_{k=1}^K Y_{ijk}}{K}.$$

Chceme-li testovat vliv interakcí, z modelu je zkusíme odstranit; tím redukuje podprostor  $N$  dimenze  $IJ$  na podprostor  $M = A + B$  dimenze  $I + J - 1$ . Zjištěný

Zdroj variability	Součet čtverců	Počet stupňů volnosti	Podíl	F
Faktor $A$	$\  \mathbf{Y}_A - \bar{\mathbf{Y}} \ ^2$	$I - 1$	$\frac{\  \mathbf{Y}_A - \bar{\mathbf{Y}} \ ^2}{I - 1}$	$\frac{\frac{\  \mathbf{Y}_A - \bar{\mathbf{Y}} \ ^2}{I - 1}}{\frac{\  \mathbf{Y} - \mathbf{Y}_N \ ^2}{n - IJ}}$
Faktor $B$	$\  \mathbf{Y}_B - \bar{\mathbf{Y}} \ ^2$	$J - 1$	$\frac{\  \mathbf{Y}_B - \bar{\mathbf{Y}} \ ^2}{J - 1}$	$\frac{\frac{\  \mathbf{Y}_B - \bar{\mathbf{Y}} \ ^2}{J - 1}}{\frac{\  \mathbf{Y} - \mathbf{Y}_N \ ^2}{n - IJ}}$
Interakce	$\  \mathbf{Y}_N - \mathbf{Y}_M \ ^2$	$(I - 1)(J - 1)$	$\frac{\  \mathbf{Y}_N - \mathbf{Y}_M \ ^2}{(I - 1)(J - 1)}$	$\frac{\frac{\  \mathbf{Y}_N - \mathbf{Y}_M \ ^2}{(I - 1)(J - 1)}}{\frac{\  \mathbf{Y} - \mathbf{Y}_N \ ^2}{n - IJ}}$
Reziduální	$\  \mathbf{Y} - \mathbf{Y}_N \ ^2$	$n - IJ$	$\frac{\  \mathbf{Y} - \mathbf{Y}_N \ ^2}{n - IJ}$	–

**Tabulka 1.5:** Tabulka analýzy rozptylu pro případ dvojného třídění s interakcemi.

rozdíl v pravoúhlých průmětech vektoru  $\mathbf{Y}$  do těchto podprostorů je pak součástí statistiky

$$\frac{\frac{\| \mathbf{Y}_N - \mathbf{Y}_M \|^2}{\dim N - \dim M}}{\frac{\| \mathbf{Y} - \mathbf{Y}_N \|^2}{\dim V_n - \dim N}} = \frac{\frac{\| \mathbf{Y}_N - \mathbf{Y}_M \|^2}{IJ - (I + J - 1)}}{\frac{\| \mathbf{Y} - \mathbf{Y}_N \|^2}{n - IJ}} = \frac{\frac{\| \mathbf{Y}_N - \mathbf{Y}_M \|^2}{(I - 1)(J - 1)}}{\frac{\| \mathbf{Y} - \mathbf{Y}_N \|^2}{n - IJ}},$$

která má v případě, že je vliv interakcí skutečně nulový, rozdělení  $F_{(I-1)(J-1), n-IJ}$ . Vliv faktoru  $A$ , resp.  $B$ , posoudíme na základě statistik

$$\frac{\frac{\| \mathbf{Y}_A - \bar{\mathbf{Y}} \|^2}{\dim A - \dim E}}{\frac{\| \mathbf{Y} - \mathbf{Y}_N \|^2}{\dim V_n - \dim N}} = \frac{\frac{\| \mathbf{Y}_A - \bar{\mathbf{Y}} \|^2}{I - 1}}{\frac{\| \mathbf{Y} - \mathbf{Y}_N \|^2}{n - IJ}},$$

resp.

$$\frac{\frac{\| \mathbf{Y}_B - \bar{\mathbf{Y}} \|^2}{\dim B - \dim E}}{\frac{\| \mathbf{Y} - \mathbf{Y}_N \|^2}{\dim V_n - \dim N}} = \frac{\frac{\| \mathbf{Y}_B - \bar{\mathbf{Y}} \|^2}{J - 1}}{\frac{\| \mathbf{Y} - \mathbf{Y}_N \|^2}{n - IJ}};$$

pokud je vliv daného faktoru nulový, řídí se tyto statistiky rozdělením  $F_{I-1, n-IJ}$ , resp.  $F_{J-1, n-IJ}$ . Výsledky uvedených testů bývají tradičně shrnuty ve formě tabulky analýzy rozptylu (viz tab. 1.5).

Pro výpočet druhých mocnin délek zúčastněných vektorů můžeme využít vzta-

hy plynoucí z Pythagorovy věty:

$$\begin{aligned}\| \mathbf{Y}_A - \bar{\mathbf{Y}} \|^2 &= JK \sum_i \bar{Y}_{i..}^2 - n\bar{Y}_{...}^2, \\ \| \mathbf{Y}_B - \bar{\mathbf{Y}} \|^2 &= IK \sum_j \bar{Y}_{.j.}^2 - n\bar{Y}_{...}^2, \\ \| \mathbf{Y} - \mathbf{Y}_M \|^2 &= \sum_{i,j,k} Y_{ijk}^2 - JK \sum_i \bar{Y}_{i..}^2 - IK \sum_j \bar{Y}_{.j.}^2 + n\bar{Y}_{...}^2, \\ \| \mathbf{Y}_N - \mathbf{Y}_M \|^2 &= K \sum_{i,j} \bar{Y}_{ij.}^2 - JK \sum_i \bar{Y}_{i..}^2 - IK \sum_j \bar{Y}_{.j.}^2 + n\bar{Y}_{...}^2, \\ \| \mathbf{Y} - \mathbf{Y}_N \|^2 &= \sum_{i,j,k} Y_{ijk}^2 - K \sum_{i,j} \bar{Y}_{ij.}^2.\end{aligned}$$

## 1.9 Aplikace rozdělení $t$

Následující myšlenky jsou převzaty z publikace [31].

Nechť  $\mathbf{z}$  je nějaký pevně daný vektor ležící v podprostoru  $M$  daném modelem (1.10); dimenze tohoto podprostoru nechť je opět  $m$ . Vektor  $\mathbf{z}$  reprezentuje nějakou lineární formu definovanou na podprostoru  $M$ , která každému vektoru  $\boldsymbol{\mu} \in M$  přiřazuje hodnotu  $\mathbf{z} \circ \boldsymbol{\mu} \in \mathbb{R}$ . Protože skutečný vektor  $\boldsymbol{\mu}$  neznáme, neznáme ani hodnotu  $\mathbf{z} \circ \boldsymbol{\mu}$ . Chceme-li tuto hodnotu odhadnout na základě pozorované realizace náhodného vektoru  $\mathbf{Y}$ , zdá se rozumné použít náhodnou veličinu  $\mathbf{z} \circ \mathbf{Y}$ ; ze vzorce (1.1) totiž plyne

$$\mathbb{E}(\mathbf{z} \circ \mathbf{Y}) = \mathbf{z} \circ \mathbb{E} \mathbf{Y} = \mathbf{z} \circ \boldsymbol{\mu};$$

jedná se tedy o nestranný odhad. Navíc ale také platí

$$\mathbf{z} \circ (\mathbf{Y} - \widehat{\mathbf{Y}}) = 0,$$

tj.

$$\mathbf{z} \circ \mathbf{Y} = \mathbf{z} \circ \widehat{\mathbf{Y}}, \quad (1.78)$$

kde  $\widehat{\mathbf{Y}}$  je opět pravoúhlý průmět vektoru  $\mathbf{Y}$  do podprostoru  $M$ . To znamená, že náhodná veličina  $\mathbf{z} \circ \mathbf{Y}$  je lineární funkcí vektoru  $\widehat{\mathbf{Y}}$ , a jako taková je podle Gaussovy-Markovovy věty (viz kapitola 2.8) dokonce *nejlepším nestranným lineárním odhadem* své střední hodnoty  $\mathbf{z} \circ \boldsymbol{\mu}$ .

Nyní můžeme ve vzorci (1.27) položit  $A = [\mathbf{z}]$  a  $B = M^\perp$ ; jelikož ortonormální báze podprostoru  $[\mathbf{z}]$  je tvořena vektorem  $\mathbf{z}/\|\mathbf{z}\|$ , je souřadnice pravoúhlého průmětu vektoru  $\mathbf{Y} - \boldsymbol{\mu}$  do tohoto prostoru vzhledem k této bázi rovna výrazu  $(\mathbf{Y} - \boldsymbol{\mu}) \circ \mathbf{z}/\|\mathbf{z}\|$  (viz vzorec (2.33)), takže dostáváme

$$\frac{(\mathbf{Y} - \boldsymbol{\mu}) \circ \mathbf{z}/\|\mathbf{z}\|}{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|/\sqrt{n-m}} = \frac{(\mathbf{Y} - \boldsymbol{\mu}) \circ \mathbf{z}}{\|\mathbf{z}\|S} \sim t_{n-m}, \quad (1.79)$$

resp. vzhledem ke vztahu (1.78)

$$\frac{(\widehat{\mathbf{Y}} - \boldsymbol{\mu}) \circ \mathbf{z}}{\|\mathbf{z}\|S} \sim t_{n-m}. \quad (1.80)$$

Rozdělení  $t$  tedy použijeme všude tam, kde je třeba odhadnout hodnotu nějaké lineární funkce neznámého vektoru  $\boldsymbol{\mu}$  – typickým příkladem je právě odhad regresních koeficientů  $\beta_i$ . Pozoruhodné je, že není vždy nezbytně nutné mít k dispozici explicitní vyjádření vektoru  $\mathbf{z}$ .

Pro srovnání uveďme ještě tradičnější způsob, kterým lze dojít k výsledku (1.79). Ze vztahů (1.1), (1.2) a z definice mnohorozměrného normálního rozdělení plyne, že náhodná veličina  $\mathbf{Y} \circ \mathbf{z}$  má rozdělení  $N(\boldsymbol{\mu} \circ \mathbf{z}, \|\mathbf{z}\|^2 \sigma^2)$ ; můžeme tedy psát

$$\frac{(\mathbf{Y} - \boldsymbol{\mu}) \circ \mathbf{z}}{\|\mathbf{z}\| \sigma} \sim N(0, 1).$$

Z kapitoly 1.5 dále víme, že platí

$$\frac{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2}{\sigma^2} \sim \chi_{n-m}^2.$$

Nezávislost dvou posledně uvedených veličin se dokazuje obvykle pomocí maticového počtu; šlo by to ale snadno i pomocí změny souřadnic – tak, jak jsme to udělali v kapitole 1.4. Pak už stačí jen použít definici rozdělení  $t$ :

$$\frac{\frac{(\mathbf{Y} - \boldsymbol{\mu}) \circ \mathbf{z}}{\|\mathbf{z}\| \sigma}}{\sqrt{\frac{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2}{\sigma^2} / (n - m)}} = \frac{(\mathbf{Y} - \boldsymbol{\mu}) \circ \mathbf{z}}{\|\mathbf{z}\| S} \sim t_{n-m}.$$

Je patrné, že ve srovnání s rozdělením  $F$  je geometrická interpretace náhodných veličin majících rozdělení  $t$  o něco méně názorná.<sup>15</sup> I tak ji však považujeme za zajímavou alternativu, která nabízí neobvyklý úhel pohledu a odhaluje některé méně známé souvislosti statistiky s geometrií a lineární algebrou.

## Příklady

### 1.9.1 Výběr z normálního rozdělení (pokračování ze str. 31)

**Test hypotézy  $\mu = \mu_0$**

Položme ve vzorci (1.27)  $A = [\mathbf{e}]$ ,  $B = [\mathbf{e}]^\perp$ . Ortonormální bázi podprostoru  $[\mathbf{e}]$  tvoří jediný vektor

$$\frac{\mathbf{e}}{\|\mathbf{e}\|} = \left( \frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}} \right)^T.$$

Pravoúhlým průmětem náhodného vektoru  $\mathbf{Y} - \boldsymbol{\mu}$  do tohoto podprostoru je vektor

$$\overline{\mathbf{Y}} - \boldsymbol{\mu} = (\overline{Y} - \mu, \dots, \overline{Y} - \mu)^T$$

<sup>15</sup>Proto se jím také zabýváme až nyní. V tradičních učebnicích statistiky je naproti tomu vždy uvedeno nejprve rozdělení  $t$  a pak teprve – jako jeho obecnější, tj. obtížnější varianta – rozdělení  $F$ .

(viz obr. 1.13), jehož souřadnice vůči uvedené ortonormální bázi má zřejmě hodnotu

$$Y^* = \sqrt{n}(\bar{Y} - \mu);$$

platí proto

$$\frac{\sqrt{n}(\bar{Y} - \mu)}{\sqrt{\|\mathbf{Y} - \mathbf{Y}_M\|^2 / (n-1)}} = \frac{\sqrt{n}(\bar{Y} - \mu)}{S} \sim t_{n-1}. \quad (1.81)$$

Chceme-li tedy pro nějaké  $\mu_0 \in \mathbb{R}$  testovat nulovou hypotézu  $\mu = \mu_0$  oproti alternativě  $\mu \neq \mu_0$ , určíme hodnoty  $\bar{Y}$  a  $S$  a hypotézu zamítneme na hladině významnosti  $\alpha$  tehdy, když nastane nerovnost

$$\frac{\sqrt{n}|\bar{Y} - \mu_0|}{S} \geq t_{n-1}(\alpha)$$

(viz definice kritických hodnot rozdělení  $t$  na straně 11). V případě jednostranné alternativy  $\mu > \mu_0$ , resp.  $\mu < \mu_0$ , zamítáme nulovou hypotézu při realizaci nerovnosti

$$\frac{\sqrt{n}(\bar{Y} - \mu_0)}{S} > t_{n-1}(2\alpha),$$

resp.

$$\frac{\sqrt{n}(\bar{Y} - \mu_0)}{S} < -t_{n-1}(2\alpha).$$

### Intervaly spolehlivosti pro $\mu$

Rovnost

$$\mathbb{P} \left[ -t_{n-1}(\alpha) < \frac{\sqrt{n}(\bar{Y} - \mu)}{S} < t_{n-1}(\alpha) \right] = 1 - \alpha,$$

která plyne z tvrzení (1.81), je ekvivalentní s rovností

$$\mathbb{P} \left[ \bar{Y} - \frac{t_{n-1}(\alpha)S}{\sqrt{n}} < \mu < \bar{Y} + \frac{t_{n-1}(\alpha)S}{\sqrt{n}} \right] = 1 - \alpha.$$

Dostáváme tak oboustranný interval spolehlivosti

$$\left( \bar{Y} - \frac{t_{n-1}(\alpha)S}{\sqrt{n}}; \bar{Y} + \frac{t_{n-1}(\alpha)S}{\sqrt{n}} \right),$$

který překrývá skutečnou hodnotu parametru  $\mu$  s pravděpodobností  $1 - \alpha$ . Podobně z rovnosti

$$\mathbb{P} \left[ \frac{\sqrt{n}(\bar{Y} - \mu)}{S} < t_{n-1}(2\alpha) \right] = 1 - \alpha,$$

resp.

$$\mathbb{P} \left[ -t_{n-1}(2\alpha) < \frac{\sqrt{n}(\bar{Y} - \mu)}{S} \right] = 1 - \alpha,$$

můžeme odvodit jednostranné intervaly spolehlivosti

$$\left( \bar{Y} - \frac{t_{n-1}(2\alpha)S}{\sqrt{n}}; \infty \right),$$

resp.

$$\left( -\infty; \bar{Y} + \frac{t_{n-1}(2\alpha)S}{\sqrt{n}} \right),$$

v nichž skutečná hodnota  $\mu$  leží rovněž s pravděpodobností  $1 - \alpha$ .

Připomeňme, že vzhledem k rovnosti (1.9) jsou zde uvedené výsledky ekvivalentní s těmi, které byly odvozeny v příkladu 1.6.5 na základě rozdělení F.

### 1.9.2 Dvouvýběrový $t$ -test

Nechť jsou všechny složky náhodného vektoru  $\mathbf{Y}$  opět navzájem nezávislé, řídí se normálním rozdělením se stejným rozptylem a jejich střední hodnota je rovna hodnotě  $\mu_1$  pro prvních  $n_1$  složek a hodnotě  $\mu_2$  pro zbyvajících  $n_2$  složek, kde  $n_1 + n_2 = n$ . Chování náhodného vektoru  $\mathbf{Y}$  tedy popisuje model

$$\mathbf{Y} = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix} \equiv \mathbf{X}\boldsymbol{\mu} + \mathbf{Z},$$

kde matice  $\mathbf{X}$  má v prvních  $n_1$  řádcích v prvním sloupci prvek 1 a v druhém sloupci prvek 0, ve zbylých  $n_2$  řádcích je tomu naopak; náhodný vektor  $\mathbf{Z}$  má rozdělení  $\mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ . Označme sloupce matice  $\mathbf{X}$  symboly  $\mathbf{x}_1$  a  $\mathbf{x}_2$  a položme jako obvykle  $\mathbf{e} = (1, \dots, 1)^T$ .

#### Odhad střední hodnoty $\boldsymbol{\mu}$ a rozptylu $\sigma^2$

Pravoúhlým průmětem vektoru  $\mathbf{Y}$  do podprostoru  $M$  je vektor

$$\widehat{\mathbf{Y}} = (\bar{Y}_1, \dots, \bar{Y}_1, \bar{Y}_2, \dots, \bar{Y}_2)^T,$$

kde výraz

$$\bar{Y}_1 \equiv \frac{\sum_{i=1}^{n_1} Y_i}{n_1}$$

stojí na místě prvních  $n_1$  souřadnic, zatímco zbylých  $n_2$  souřadnic má hodnotu

$$\bar{Y}_2 \equiv \frac{\sum_{i=n_1+1}^n Y_i}{n_2};$$

náhodné veličiny  $\bar{Y}_1, \bar{Y}_2$  jsou nejlepšími nestrannými lineárními odhady parametrů  $\mu_1, \mu_2$ . Vektor  $\mathbf{Y} - \widehat{\mathbf{Y}}$  je pak pravoúhlým průmětem vektoru  $\mathbf{Y} - \boldsymbol{\mu}$  do podprostoru  $V_n - M$  dimenze  $n - 2$ , takže nestranným odhadem rozptylu  $\sigma^2$  je náhodná veličina

$$S^2 = \frac{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2}{n - 2} = \frac{\|\mathbf{Y}\|^2 - \|\widehat{\mathbf{Y}}\|^2}{n - 2} = \frac{\sum_{i=1}^n Y_i^2 - n_1 \bar{Y}_1^2 - n_2 \bar{Y}_2^2}{n - 2}.$$

Někdy se počítá odhad rozptylu pro každou část výběru zvlášť, jako by se jednalo o dva samostatné výběry z normálního rozdělení<sup>16</sup>

$$S_1^2 = \frac{\sum_{i=1}^{n_1} Y_i^2 - n_1 \bar{Y}_1^2}{n_1 - 1}, \quad S_2^2 = \frac{\sum_{i=n_1+1}^n Y_i^2 - n_2 \bar{Y}_2^2}{n_2 - 1}.$$

V tom případě mezi těmito odhady platí vztah

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

### Test hypotézy $\mu_1 = \mu_2$

Chceme-li testovat hypotézu, že střední hodnoty  $\mu_1$  a  $\mu_2$  se od sebe neliší, tj. hypotézu  $\mu_1 - \mu_2 = 0$ , uvažujeme o redukci podprostoru  $M \equiv [\mathbf{x}_1, \mathbf{x}_2]$  dimenze 2 na jednorozměrný podprostor  $E \equiv [\mathbf{e}]$ . Ve vzorci (1.27) tedy můžeme položit  $A = M - E$ ,  $B = M^\perp$ . Ve jmenovateli tak dostaneme odmocninu z výběrového rozptylu  $S^2$ ; zbývá nám určit čitatel.

Jedním z možných generátorů podprostoru  $M - E$  je vektor

$$\mathbf{z} \equiv (n_2, \dots, n_2, -n_1, \dots, -n_1)^T,$$

který má prvních  $n_1$  souřadnic rovných hodnotě  $n_2$  a zbývajících  $n_2$  souřadnic hodnotě  $-n_1$ ; je evidentně lineární kombinací vektorů  $\mathbf{x}_1$  a  $\mathbf{x}_2$  a přitom je kolmý na vektor  $\mathbf{e}$ .<sup>17</sup> Leží-li střední hodnota  $\boldsymbol{\mu}$  skutečně v podprostoru  $E$ , platí  $\boldsymbol{\mu} \circ \mathbf{z} = 0$ , takže souřadnice průmětu vektoru  $\mathbf{Y} - \boldsymbol{\mu}$  do podprostoru  $M - E$  vzhledem k jeho ortonormální bázi  $\{\mathbf{z}/\|\mathbf{z}\|\}$  je

$$\begin{aligned} \frac{(\mathbf{Y} - \boldsymbol{\mu}) \circ \mathbf{z}}{\|\mathbf{z}\|} &= \frac{\mathbf{Y} \circ \mathbf{z}}{\|\mathbf{z}\|} = \frac{n_2 \sum_{i=1}^{n_1} Y_i - n_1 \sum_{i=n_1+1}^n Y_i}{\sqrt{n_1 n_2 (n_1 + n_2)}} = \\ &= \frac{n_1 n_2 \bar{Y}_1 - n_1 n_2 \bar{Y}_2}{\sqrt{n_1 n_2 (n_1 + n_2)}} = \\ &= (\bar{Y}_1 - \bar{Y}_2) \sqrt{\frac{n_1 n_2}{n_1 + n_2}}. \end{aligned}$$

<sup>16</sup>To má samozřejmě opodstatnění především tehdy, když nepředpokládáme shodnost rozptylů u obou částí výběru. Tímto případem se zde však nebudeme zabývat.

<sup>17</sup>Kdyby se nám nepovedlo vektor  $\mathbf{v}$  takto „uhodnout“, mohli bychom jej – nebo nějaký jeho násobek – nalézt tak, že bychom určili pravoúhlý průmět vektoru  $\mathbf{x}_1$  do podprostoru  $E$  a tento průmět, řekněme  $\mathbf{p}_1$ , bychom odečetli od vektoru  $\mathbf{x}_1$ ; výsledný vektor  $\mathbf{x}_1 - \mathbf{p}_1$  by byl pravoúhlým průmětem vektoru  $\mathbf{x}_1$  do jednorozměrného podprostoru  $M - E$ , a tudíž jeho generátorem.

Za předpokladu platnosti hypotézy  $\mu_1 - \mu_2 = 0$  má tedy náhodná veličina

$$\frac{\bar{Y}_1 - \bar{Y}_2}{S} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \quad (1.82)$$

rozdělení  $t_{n-2}$ ; leží-li její získaná realizace mimo interval

$$\left(-t_{n-2}(\alpha); t_{n-2}(\alpha)\right),$$

můžeme testovanou hypotézu zamítnout na hladině významnosti  $\alpha$ .

### Porovnání s $F$ -testem

Výše popsáný test ovšem můžeme snadno provést i užitím analýzy rozptylu, stejně jako v případě jednoduchého třídění (viz příklad 1.6.6). Pravoúhlým průmětem vektoru  $\mathbf{Y}$  do podprostoru  $E$  je vektor  $\bar{\mathbf{Y}} \equiv (\bar{Y}, \dots, \bar{Y})^T$ , kde

$$\bar{Y} = \frac{\sum_{i=1}^n Y_i}{n} = \frac{n_1 \bar{Y}_1 + n_2 \bar{Y}_2}{n}. \quad (1.83)$$

V případě platnosti nulové hypotézy  $\boldsymbol{\mu} \in E$  má náhodná veličina

$$\begin{aligned} \frac{\|\widehat{\mathbf{Y}} - \bar{\mathbf{Y}}\|^2 / 1}{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2 / 6} &= \frac{\|\widehat{\mathbf{Y}}\|^2 - \|\bar{\mathbf{Y}}\|^2}{S^2} = \\ &= \frac{n_1 \bar{Y}_1^2 + n_2 \bar{Y}_2^2 - n \bar{Y}^2}{S^2} = \\ &= \frac{n_1 n_2}{n_1 + n_2} \frac{(\bar{Y}_1 - \bar{Y}_2)^2}{S^2} \end{aligned}$$

rozdělení  $F_{1, n-2}$  (poslední rovnost získáme dosazením pravé strany vztahu (1.83) za  $\bar{Y}$ ). Nulovou hypotézu tedy zamítneme na hladině významnosti  $\alpha$ , bude-li získaná realizace této veličiny větší než hodnota  $F_{1, n-2}(\alpha)$ ; jelikož je tato veličina zřejmě druhou mocninou veličiny (1.82), je vzhledem ke vztahu (1.9) tento test ekvivalentní s výše popsáním  $t$ -testem.

### Odhad parametru $\mu_1 - \mu_2$

Ze vztahu (1.79), resp. (1.80), však můžeme snadno dostat obecnější výsledek: tento vztah totiž platí pro jakýkoli vektor  $\mathbf{z} \in M$  a není nutné, aby skutečná střední hodnota  $\boldsymbol{\mu}$  náhodného vektoru  $\mathbf{Y}$  ležela v podprostoru  $E$ . Nechť je  $\boldsymbol{\mu} = \mu_1 \mathbf{x}_1 + \mu_2 \mathbf{x}_2$ ; pak platí

$$\boldsymbol{\mu} \circ \mathbf{z} = n_1 n_2 (\mu_1 - \mu_2),$$

takže výraz  $\boldsymbol{\mu} \circ \mathbf{z}$  představuje lineární funkci parametru  $\mu_1 - \mu_2$ , který je předmětem našeho zájmu. Po dosazení do vzorce (1.80) a několika estetických úpravách tak dostáváme

$$\frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{S} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \sim t_{n-1}; \quad (1.84)$$

tento vztah můžeme nyní použít k navržení testů nulové hypotézy typu  $\mu_1 - \mu_2 = d_0$  pro nějaké  $d_0 \in \mathbb{R}$ , a to oboustranných i jednostranných. Má-li například alternativní hypotéza tvar  $\mu_1 - \mu_2 > d_0$ , zamítneme nulovou hypotézu ve prospěch hypotézy alternativní na hladině významnosti  $\alpha$ , nastane-li nerovnost

$$\frac{(\bar{Y}_1 - \bar{Y}_2) - d_0}{S} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} > t_{n-1}(2\alpha).$$

Podobně jako v příkladu 1.9.1 můžeme vztah (1.84) použít také k odvození intervalů spolehlivosti pro hodnotu parametru  $\mu_1 - \mu_2$ , a to opět jednostranných i oboustranných.

### Konkrétní hodnoty

Nechť první tři složky náhodného vektoru  $\mathbf{Y}$  představují náhodný výběr z rozdělení  $N(\mu_1, \sigma^2)$ , zbylých pět složek z rozdělení  $N(\mu_2, \sigma^2)$ ; chování náhodného vektoru  $\mathbf{Y}$  tedy popisuje model

$$\mathbf{Y} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \mathbf{Z},$$

kde náhodný vektor  $\mathbf{Z}$  má rozdělení  $N(\mathbf{0}; \sigma^2 \mathbf{I}_8)$ .

Nechť dále získaná realizace náhodného vektoru  $\mathbf{Y}$  je

$$\mathbf{y} = (4, 8, 9, 7, 11, 6, 10, 11)^T.$$

Jelikož průměry z jednotlivých částí výběru jsou  $\bar{y}_1 = 7$  a  $\bar{y}_2 = 9$ , je pravoúhlým průmětem získané realizace do podprostoru  $M$  generovaného sloupci matice  $\mathbf{X}$  vektor

$$\hat{\mathbf{y}} = (7, 7, 7, 9, 9, 9, 9, 9)^T.$$

Vektor  $\mathbf{y} - \hat{\mathbf{y}}$  je pravoúhlým průmětem vektoru  $\mathbf{y} - \boldsymbol{\mu}$  do podprostoru  $V_8 - M$  dimenze 6, takže nestranným odhadem rozptylu  $\sigma^2$  je hodnota

$$\begin{aligned} S^2 &= \frac{\|\mathbf{y} - \hat{\mathbf{y}}\|^2}{6} = \\ &= \frac{\|(-3, 1, 2, -2, 2, -3, 1, 2)^T\|^2}{6} = \\ &= 6. \end{aligned}$$

Generátorem podprostoru  $M - E$  je vektor

$$\mathbf{z} = (5, 5, 5, -3, -3, -3, -3, -3)^T$$

délky  $\sqrt{120}$ , střední hodnotou náhodného vektoru  $\mathbf{Y}$  je vektor

$$\boldsymbol{\mu} = (\mu_1, \mu_1, \mu_1, \mu_2, \mu_2, \mu_2, \mu_2, \mu_2)^T;$$

dosazením do vzorce (1.79) tedy dostáváme

$$\begin{aligned} \frac{(\widehat{\mathbf{y}} - \boldsymbol{\mu}) \circ \mathbf{z}}{\|\mathbf{z}\|S} &= \frac{15(\bar{y}_1 - \bar{y}_2) - 15(\mu_1 - \mu_2)}{\sqrt{120}\sqrt{6}} = \\ &= \frac{-30 - 15(\mu_1 - \mu_2)}{\sqrt{720}}. \end{aligned}$$

Protože tato hodnota je realizací náhodné veličiny mající rozdělení  $t_6$ , leží s pravděpodobností 95 % v intervalu

$$\left(-t_6(0,05); t_6(0,05)\right) \doteq (-2,45; 2,45),$$

z čehož snadno vypočteme, že skutečná hodnota rozdílu  $\mu_1 - \mu_2$  leží v intervalu

$$\left(\frac{-30 - 2,45\sqrt{720}}{15}; \frac{-30 + 2,45\sqrt{720}}{15}\right) \doteq (-6,38; 2,38)$$

s pravděpodobností 95 %.

### 1.9.3 Mnohonásobná regrese (pokračování ze str. 50)

Dimenze podprostoru  $V_n - M$  je v tomto případě  $n - k - 1$  a náhodná veličina  $S$  ve vzorci (1.79) představuje odmocninu z odhadu rozptylu

$$S^2 = \frac{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2}{n - k - 1}.$$

Abychom mohli vzorec použít k odhadu parametru  $\beta_i$ , musíme najít takový vektor  $\mathbf{z}_i \in M$ , pro který je výraz  $\mathbf{z}_i \circ \boldsymbol{\mu}$  lineární funkcí parametru  $\beta_i$ , ale nikoli ostatních.

#### Vyjádření souřadnic pomocí skalárního součinu

Jelikož jsou vektory  $\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_k$  dle našich předpokladů lineárně nezávislé, představují bázi podprostoru  $M$ ; označme tuto bázi symbolem  $\mathcal{B}$ . Každý vektor  $\mathbf{a} \in M$  lze zapsat jako jejich lineární kombinaci:

$$\mathbf{a} = a_0\mathbf{x}_0 + a_1\mathbf{x}_1 + \dots + a_k\mathbf{x}_k.$$

Koeficienty  $a_i$  představují souřadnice vektoru  $\mathbf{a}$  vzhledem k bázi  $\mathcal{B}$  a jsou jednoznačně určeny. Tyto souřadnice jsou lineárními funkcemi vektoru  $\mathbf{a}$ , pro každé  $i = 0, 1, \dots, k$  tedy existuje právě jedna lineární forma  $\psi_i$  definovaná na podprostoru  $M$ , která každému vektoru  $\mathbf{a} \in M$  přiřazuje jeho  $i$ -tou souřadnici  $a_i$ . Jak známo, lze tuto formu vyjádřit pomocí skalárního součinu; existuje tedy právě jeden vektor  $\mathbf{z}_i \in M$  takový, že platí

$$\psi_i(\mathbf{a}) = \mathbf{z}_i \circ \mathbf{a} = a_i$$

pro všechna  $\mathbf{a} \in M$  (podrobnější výklad k tomuto konceptu viz např. publikace [4], [25], nebo [31]). Tento vektor zřejmě splňuje vztahy

$$\mathbf{z}_i \circ \boldsymbol{\mu} = \beta_i, \quad \mathbf{z}_i \circ \widehat{\mathbf{Y}} = b_i.$$

Vzhledem k rovnosti (1.78) navíc platí

$$\mathbf{z}_i \circ \mathbf{Y} = b_i;$$

položíme-li tedy ve vzorci (1.79)  $\mathbf{z} = \mathbf{z}_i$ , dostáváme v čitateli výraz  $b_i - \beta_i$ .

### Výpočet hodnoty $\|\mathbf{z}_i\|$

Každý z vektorů  $\mathbf{z}_i$  leží v podprostoru  $M$ ; označme jeho souřadnice vzhledem k bázi  $\mathcal{B}$  tohoto podprostoru symboly  $c_{ij}$ , tj. pro každé  $i = 0, \dots, k$  platí

$$\mathbf{z}_i = \sum_{j=0}^k c_{ij} \mathbf{x}_j.$$

Víme ovšem, že když nějaký vektor ležící v podprostoru  $M$  vynásobíme vektorem  $\mathbf{z}_j$ , dostaneme jeho souřadnici s indexem  $j$  vzhledem k bázi  $\mathcal{B}$ ; z toho plyne, že pro všechna  $i, j \in \{0, \dots, k\}$  musí být splněna rovnost

$$c_{ij} = \mathbf{z}_i \circ \mathbf{z}_j.$$

Z toho dostáváme

$$\mathbf{z}_i = \sum_{j=0}^k (\mathbf{z}_i \circ \mathbf{z}_j) \mathbf{x}_j. \quad (1.85)$$

Dále, ze stejného důvodu musí pro všechna  $i, l \in \{0, \dots, k\}$  platit

$$\mathbf{z}_i \circ \mathbf{x}_l = \delta_{il}, \quad (1.86)$$

kde  $\delta_{il} = 0$  pro  $i \neq l$ ,  $\delta_{il} = 1$  pro  $i = l$ ; pro každé  $\mathbf{x}_l$  je totiž

$$\mathbf{x}_l = \sum_{i=0}^k \delta_{il} \mathbf{x}_i.$$

Dosazením rovnosti (1.85) do vztahu (1.86) dostáváme

$$\left[ \sum_{j=0}^k (\mathbf{z}_i \circ \mathbf{z}_j) \mathbf{x}_j \right] \circ \mathbf{x}_l = \sum_{j=0}^k (\mathbf{z}_i \circ \mathbf{z}_j) (\mathbf{x}_j \circ \mathbf{x}_l) = \delta_{il}. \quad (1.87)$$

Definujme nyní čtvercové matice  $\mathbf{C}$  a  $\mathbf{A}$  o rozměrech  $(k+1) \times (k+1)$ , jejichž řádky a sloupce jsou indexovány  $0, \dots, k$  a jejichž prvky jsou určeny vztahy

$$c_{ij} = \mathbf{z}_i \circ \mathbf{z}_j, \quad a_{ij} = \mathbf{x}_i \circ \mathbf{x}_j.$$

Prostřední část rovnosti (1.87) tak vlastně představuje maticový součin řádku  $i$  matice  $\mathbf{C}$  se sloupcem  $l$  matice  $\mathbf{A}$ , zatímco výraz  $\delta_{il}$  na pravé straně je prvek

jednotkové matice velikosti  $k+1$  (se stejně indexovanými řádky a sloupci). Protože tato rovnost je splněna pro všechna  $i, l \in \{0, \dots, k\}$ , znamená to, že platí

$$\mathbf{C}\mathbf{A} = \mathbf{I}_{k+1}.$$

Obě matice jsou však symetrické, takže platí též  $\mathbf{A}\mathbf{C} = \mathbf{I}_{k+1}$ ; to znamená, že matice  $\mathbf{A}$  a  $\mathbf{C}$  jsou navzájem inverzní.<sup>18</sup> Matici  $\mathbf{A}$  můžeme snadno určit, protože sloupce matice  $\mathbf{X}$  v modelu (1.45) představují vyjádření vektorů  $\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_k$  vzhledem k ortonormální bázi:

$$\mathbf{A} = \mathbf{X}^T\mathbf{X}.$$

Tím pádem známe i matici  $\mathbf{C}$ :

$$\mathbf{C} = (\mathbf{X}^T\mathbf{X})^{-1},$$

dostáváme tedy, že délka vektoru  $\mathbf{z}_i$  je

$$\|\mathbf{z}_i\| = \sqrt{\mathbf{z}_i \circ \mathbf{z}_i} = \sqrt{c_{ii}},$$

kde  $c_{ii}$  je prvek matice  $(\mathbf{X}^T\mathbf{X})^{-1}$  s indexy  $0 \leq i, j \leq k$ .

### Testy hypotéz a intervalové odhady pro $\beta_i$

Dosažením výsledků z předchozího odstavce do vztahu (1.79) dostáváme, že je-li skutečná hodnota parametru  $\beta_i$  rovna hodnotě  $\beta_i^0$ , platí

$$t \equiv \frac{b_i - \beta_i^0}{S\sqrt{c_{ii}}} \sim t_{n-k-1}. \quad (1.88)$$

Nulovou hypotézu  $\beta_i = \beta_i^0$  tedy zamítneme ve prospěch alternativní hypotézy  $\beta_i \neq \beta_i^0$ , resp.  $\beta_i > \beta_i^0$ , resp.  $\beta_i < \beta_i^0$ , nastane-li nerovnost

$$\begin{aligned} |t| &> t_{n-k-1}(\alpha), \quad \text{resp.} \\ t &> t_{n-k-1}(2\alpha), \quad \text{resp.} \\ t &< -t_{n-k-1}(2\alpha). \end{aligned}$$

Dále můžeme pomocí vztahu (1.88) stanovit intervaly spolehlivosti, které skutečnou hodnotu parametru  $\beta_i$  překrývají s pravděpodobností  $1 - \alpha$ : oboustranný interval je

$$\left( b_i - S\sqrt{c_{ii}}t_{n-k-1}(\alpha); b_i + S\sqrt{c_{ii}}t_{n-k-1}(\alpha) \right),$$

jednostranné intervaly mají tvar

$$\left( b_i - S\sqrt{c_{ii}}t_{n-k-1}(2\alpha); \infty \right),$$

resp.

$$\left( -\infty; b_i + S\sqrt{c_{ii}}t_{n-k-1}(2\alpha) \right).$$

<sup>18</sup>Matice  $\mathbf{A}$ , resp.  $\mathbf{C}$ , se nazývá *Grammovou maticí* vektorů  $\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_k$ , resp.  $\mathbf{z}_0, \mathbf{z}_1, \dots, \mathbf{z}_k$  (viz [4]). Regularitu matice  $\mathbf{A}$  už jsme dokazovali v kapitole 1.3, zde plyne z existence inverzní matice.

## Porovnání s výsledky odvozenými pomocí rozdělení F

V posledním odstavci příkladu 1.7.6 jsme odvodili vzorec (1.60) pro náhodnou veličinu  $F$ , která má za předpokladu platnosti nulové hypotézy  $\beta_i = \beta_i^0$  rozdělení  $F_{1,n-k-1}$ . Uvedli jsme, že výsledky testů a intervalové odhady učiněné na základě této statistiky jsou ekvivalentní s těmi založenými na rozdělení  $t$  (které jsou k tomuto účelu používány tradičně). K důkazu této ekvivalence nám chybí doložit, že pro vektor  $\mathbf{q}_i$ , který jsme v příkladu 1.7.6 definovali jako pravoúhlý průmět vektorů  $\mathbf{x}_i$  do podprostoru

$$M - [\mathbf{x}_0, \dots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \dots, \mathbf{x}_k]$$

(viz obr. 1.21), platí

$$\|\mathbf{q}_i\|^2 = 1/c_{ii},$$

kde hodnota  $c_{ii}$  představuje prvek na  $i$ -tém místě diagonály (počítáno od 0) matice  $(\mathbf{X}^T\mathbf{X})^{-1}$ . Dokončíme nyní tento důkaz.

Z definice vektoru  $\mathbf{q}_i$  plynou vztahy

$$\mathbf{q}_i \circ \mathbf{x}_j = 0$$

pro  $i \neq j$  a

$$\begin{aligned} \mathbf{q}_i \circ \mathbf{x}_i &= \mathbf{q}_i \circ [\mathbf{q}_i + (\mathbf{x}_i - \mathbf{q}_i)] = \\ &= \mathbf{q}_i \circ \mathbf{q}_i + \mathbf{q}_i \circ (\mathbf{x}_i - \mathbf{q}_i) = \\ &= \|\mathbf{q}_i\|^2. \end{aligned}$$

Nechť nyní pro vektor  $\mathbf{a} \in M$  platí

$$\mathbf{a} = \sum_{j=0}^k a_j \mathbf{x}_j;$$

z toho dostáváme

$$\begin{aligned} \mathbf{a} \circ \mathbf{q}_i &= \left( \sum_{j=0}^k a_j \mathbf{x}_j \right) \circ \mathbf{q}_i = \\ &= \sum_{j=0}^k a_j (\mathbf{x}_j \circ \mathbf{q}_i) = \\ &= a_i \mathbf{x}_i \circ \mathbf{q}_i = \\ &= a_i \|\mathbf{q}_i\|^2, \end{aligned}$$

tj.

$$\frac{\mathbf{a} \circ \mathbf{q}_i}{\|\mathbf{q}_i\|^2} = a_i.$$

To ale znamená, že vektor  $\mathbf{q}_i/\|\mathbf{q}_i\|^2$  má právě tu vlastnost, pomocí níž jsme definovali jednoznačně určený vektor  $\mathbf{z}_i$ . Platí tedy

$$\frac{\mathbf{q}_i}{\|\mathbf{q}_i\|^2} = \mathbf{z}_i,$$

a proto je

$$c_{ii} = \|\mathbf{z}_i\|^2 = \left\| \frac{\mathbf{q}_i}{\|\mathbf{q}_i\|^2} \right\|^2 = \frac{\|\mathbf{q}_i\|^2}{\|\mathbf{q}_i\|^4} = \frac{1}{\|\mathbf{q}_i\|^2}.$$

Tím je důkaz hotov.

## Porovnání s maticovým přístupem

Porovnejme ještě na závěr část našich výsledků se standardně používanou metodou využívající maticového počtu. Jsou-li sloupce matice  $\mathbf{X}$  lineárně nezávislé, je náhodný vektor  $\mathbf{b}$  určen vztahem

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

(viz (1.17)). Z toho podle vzorce (1.2) plyne, že jeho varianční matice je rovná matici

$$\begin{aligned} \text{var } \mathbf{b} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \cdot \text{var } \mathbf{Y} \cdot [(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T]^T = \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \cdot \sigma^2 \mathbf{I}_{k+1} \cdot \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} = \\ &= \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} = \\ &= \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} = \\ &= \sigma^2 \mathbf{C} \end{aligned}$$

(při přechodu z prvního řádku na druhý jsme využili skutečnosti, že matice  $\mathbf{X}^T \mathbf{X}$  je symetrická, a proto je symetrická i její inverze). To znamená, že rozptyl náhodné veličiny  $b_i$  je roven hodnotě  $\sigma^2 c_{ii}$ ; a jelikož tato náhodná veličina má normální rozdělení a její střední hodnota je za předpokladu platnosti nulové hypotézy rovna  $\beta_i^0$ , dostáváme

$$\frac{b_i - \beta_i^0}{\sigma \sqrt{c_{ii}}} \sim \text{N}(0; 1).$$

Tak se dostane hodnota  $\sqrt{c_{ii}}$  do jmenovatele vzorce (1.88); další podrobnosti viz [3].

### 1.9.4 Regresní přímka (pokračování ze str. 35)

#### Odhad parametru $\beta_1$

Nechť skutečné hodnoty regresních koeficientů jsou  $\beta_0$  a  $\beta_1$ , tj. střední hodnota náhodného vektoru  $\mathbf{Y}$  je

$$\boldsymbol{\mu} = \beta_0 \mathbf{e} + \beta_1 \mathbf{x}.$$

Hledáme-li vektor  $\mathbf{z}_1 \in Z$  takový, že pro jakékoli hodnoty  $\beta_0, \beta_1 \in \mathbb{R}$  platí

$$\mathbf{z}_1 \circ \boldsymbol{\mu} = \beta_1,$$

je tímto vektorem podle výsledků odvozených v předcházejícím obecnějším příkladu 1.9.3 vektor

$$\mathbf{z}_1 = \frac{\mathbf{q}}{\|\mathbf{q}\|^2},$$

kde vektor  $\mathbf{q}$  je pravoúhlým průmětem vektoru  $\mathbf{x}$  do podprostoru  $M - [\mathbf{e}]$  (viz (1.55) a obrázek 1.19). Nejlepším nestranným lineárním odhadem parametru  $\beta_1$  je pak náhodná veličina

$$\mathbf{z}_1 \circ \mathbf{Y} = \frac{\mathbf{q} \circ \mathbf{Y}}{\|\mathbf{q}\|^2},$$

což ovšem není nic jiného než veličina  $b_1$  (viz vzorec (1.59) na straně 47).

Délka vektoru  $\mathbf{z}_1$  je rovna

$$\|\mathbf{z}_1\| = \frac{\|\mathbf{q}\|}{\|\mathbf{q}\|^2} = \frac{1}{\|\mathbf{q}\|};$$

když tedy položíme ve vztahu (1.79)  $\mathbf{z} = \mathbf{z}_1$ , dostáváme, že náhodná veličina

$$\frac{(\mathbf{Y} - \boldsymbol{\mu}) \circ \mathbf{z}_1}{\|\mathbf{z}_1\|S} = \frac{(b_1 - \beta_1)\|\mathbf{q}\|}{S} \quad (1.89)$$

má rozdělení  $t_{n-2}$ . Ještě zbývá vyjádřit délku vektoru  $\mathbf{q}$ : jelikož vektory  $\bar{x}\mathbf{e}$  a  $\mathbf{q}$  tvoří odvěsny pravoúhlého trojúhelníku, jehož přeponou je vektor  $\mathbf{x}$ , platí

$$\|\mathbf{q}\| = \sqrt{\|\mathbf{x}\|^2 - \|\bar{x}\mathbf{e}\|^2} = \sqrt{\sum_{i=1}^n x_i^2 - n\bar{x}^2},$$

takže dostáváme finální podobu tvrzení

$$\frac{(b_1 - \beta_1) \sqrt{\sum_{i=1}^n x_i^2 - n\bar{x}^2}}{S} \sim t_{n-2}.$$

Na základě toho lze obvyklým způsobem navrhnout testy hypotéz týkajících se parametru  $\beta_1$  a stanovit intervaly spolehlivosti libovolného typu.

### Odhad parametru $\beta_0$

Co se týče druhého parametru, platí

$$\begin{aligned} \beta_0 \mathbf{e} &= \boldsymbol{\mu} - \beta_1 \mathbf{x} = \\ &= \boldsymbol{\mu} - (\mathbf{z}_1 \circ \boldsymbol{\mu}) \mathbf{x}; \end{aligned}$$

skalárním vynásobením vektorem  $\mathbf{e}$  dostaneme

$$\begin{aligned} \beta_0 \|\mathbf{e}\|^2 &= \boldsymbol{\mu} \circ \mathbf{e} - (\mathbf{z}_1 \circ \boldsymbol{\mu}) \mathbf{x} \circ \mathbf{e} = \\ &= \boldsymbol{\mu} \circ [\mathbf{e} - (\mathbf{x} \circ \mathbf{e}) \mathbf{z}_1], \end{aligned}$$

takže hledaný vektor  $\mathbf{z}_0$  je

$$\mathbf{z}_0 = \frac{\mathbf{e} - (\mathbf{x} \circ \mathbf{e})}{\|\mathbf{e}\|^2} = \frac{\mathbf{e} - \left(\sum_{i=1}^n x_i\right) \mathbf{z}_1}{n} = \mathbf{e}/n - \bar{x} \mathbf{z}_1.$$

Nejlepším nestranným lineárním odhadem hodnoty  $\beta_0 = \mathbf{z}_0 \circ \boldsymbol{\mu}$  je tedy náhodná veličina

$$\mathbf{z}_0 \circ \mathbf{Y} = \mathbf{e}/n \circ \mathbf{Y} - \bar{x} \mathbf{z}_1 \circ \mathbf{Y} = \bar{Y} - \bar{x} b_1,$$

což koresponduje s druhým ze vztahů (1.20), neboť se jedná samozřejmě o veličinu  $b_0$ .

Jelikož vektor  $z_1$  je násobkem vektoru  $q$ , jsou vektory  $e$  a  $z_1$  jsou navzájem kolmé; délka vektoru  $z_0$  je proto

$$\|z_0\| = \sqrt{\|e/n\|^2 + \|\bar{x}z_1\|^2} = \sqrt{\frac{\|e\|^2}{n^2} + \frac{\bar{x}^2}{\|q\|^2}} = \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}}.$$

Nyní můžeme ve vztahu (1.79) položit  $z = z_0$  a dostáváme, že náhodná veličina

$$\frac{z_0 \circ (\mathbf{Y} - \boldsymbol{\mu})}{\|z_0\|S} = \frac{b_0 - \beta_0}{S \cdot \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}}} \quad (1.90)$$

má rozdělení  $t_{n-2}$ ; toho lze opět využít při testování hypotéz týkajících se parametru  $\beta_0$  a k odvození intervalů spolehlivosti pro tento parametr.

### Pás spolehlivosti kolem regresní přímky

Nechť  $x \in \mathbb{R}$  je pevně zvolené číslo; chceme odhadnout hodnotu  $\beta_0 + \beta_1 x$ . S využitím výsledků předcházejících odstavců nejprve odvodíme, že je

$$\begin{aligned} \beta_0 + \beta_1 x &= z_0 \circ \boldsymbol{\mu} + (z_1 \circ \boldsymbol{\mu})x = \\ &= (z_0 + xz_1) \circ \boldsymbol{\mu}. \end{aligned}$$

Nejlepším nestranným odhadem této hodnoty je tedy náhodná veličina

$$(z_0 + xz_1) \circ \mathbf{Y} = b_0 + xb_1,$$

což jsme ovšem mohli čekat. Položme

$$\begin{aligned} z &= z_0 + xz_1 = \\ &= e/n - \bar{x}z_1 + xz_1 = \\ &= e/n + (x - \bar{x})z_1; \end{aligned}$$

vzhledem ke kolmosti vektorů  $e$  a  $z_1$  je délka tohoto vektoru rovna

$$\|z\| = \sqrt{\|e/n\|^2 + \|(x - \bar{x})z_1\|^2} = \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}};$$

ze vztahu (1.79) tedy dostáváme, že náhodná veličina

$$\frac{z \circ (\mathbf{Y} - \boldsymbol{\mu})}{\|z\|S} = \frac{(b_0 + b_1 x) - (\beta_0 + \beta_1 x)}{S \cdot \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}}}$$

má rozdělení  $t_{n-2}$ . Z toho lze odvodit, že s pravděpodobností  $1 - \alpha$  leží skutečná hodnota výrazu  $\beta_0 + \beta_1 x$  v intervalu s koncovými body

$$(b_0 + b_1 x) \pm t_{n-2}(\alpha) S \cdot \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}}. \quad (1.91)$$

Vyneseme-li tyto hodnoty pro každé  $x \in \mathbb{R}$  do grafu  $x, y$ , ohraničují oblast zvanou *pás spolehlivosti kolem regresní přímky*.

### Konkrétní hodnoty

Vraťme se ještě jednou ke konkrétním hodnotám, uvedeným v příkladech 1.2.3 a 1.3.6, a ilustrujme na nich výše uvedené postupy. Z vektoru  $\mathbf{x}$  (viz model (1.13)) určíme nejprve vektor  $\mathbf{q}$ :

$$\mathbf{q} = \mathbf{x} - \bar{x}\mathbf{e} = \begin{pmatrix} 3 \\ 2 \\ 4 \\ 3 \\ 5 \\ 4 \\ 7 \end{pmatrix} - \begin{pmatrix} 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \end{pmatrix} = \begin{pmatrix} -1 \\ -2 \\ 0 \\ -1 \\ 1 \\ 0 \\ 3 \end{pmatrix}.$$

Jeho délka je  $\|\mathbf{q}\| = 4$ . Nyní můžeme stanovit bodové odhady hodnot  $\beta_1$  a  $\beta_0$ :

$$b_1 = \frac{\mathbf{q} \circ \mathbf{y}}{\|\mathbf{q}\|^2} = \frac{48}{16} = 3, \quad b_0 = \bar{y} - \bar{x}b_1 = 50 - 4 \cdot 3 = 38;$$

výsledky se samozřejmě shodují s hodnotami vypočtenými v příkladu 1.3.6.

Hodnota veličiny  $S^2 = 15,2$  je vypočtena v příkladu 1.5.3, máme tedy vše, co potřebujeme k dosazení do vztahu (1.89); dostáváme výraz

$$\frac{(b_1 - \beta_1)\|\mathbf{q}\|}{S} = \frac{(3 - \beta_1) \cdot 4}{\sqrt{15,2}}.$$

Podobně po dosazení do vzorce (1.90) dostáváme

$$\frac{38 - \beta_0}{\sqrt{15,2} \cdot \sqrt{\frac{1}{7} + \frac{16}{16}}}.$$

Oba tyto výrazy jsou realizací náhodných veličin majících rozdělení  $t_5$ , takže s pravděpodobností 95 % leží v intervalu

$$\left(-t_5(0,05); t_5(0,05)\right) = (-2,57; 2,57).$$

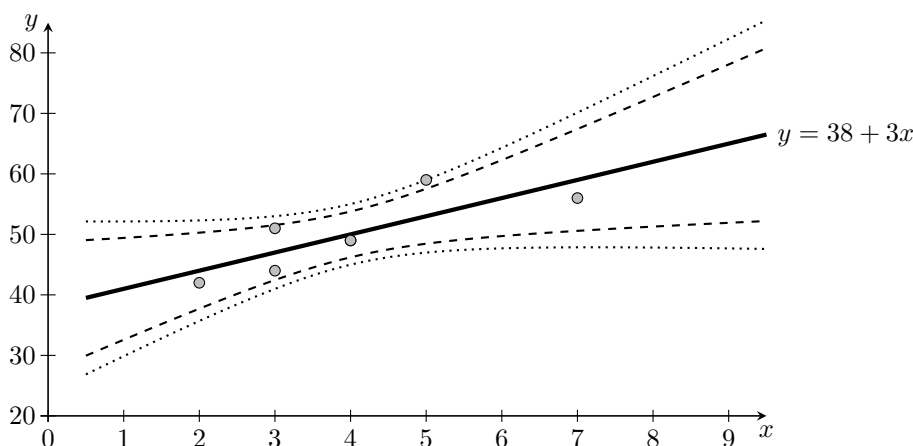
Z toho můžeme usoudit, že platí

$$\begin{aligned} \mathbf{P} \left[ \beta_1 \in (0,49; 5,51) \right] &= 0,95, \\ \mathbf{P} \left[ \beta_0 \in (27,3; 48,7) \right] &= 0,95. \end{aligned}$$

Chceme-li odhadnout např. hodnotu  $y(9) = \beta_0 + 9\beta_1$ , tj. střední hodnotu závislé veličiny  $Y$  v případě, že hodnota nezávislé veličiny  $x$  bude 9, je tímto odhadem hodnota  $b_0 + 9b_1 = 38 + 3 \cdot 9 = 65$  a ze vzorce (1.91) dostáváme, že skutečná hodnota leží s pravděpodobností 95% v intervalu s krajními mezemi

$$65 \pm 2,57 \cdot \sqrt{15,2} \cdot \sqrt{\frac{1}{7} + \frac{(9-4)^2}{16}} = 65 \pm 13,1.$$

Hodnoty  $x_i$  a  $Y_i$ , regresní přímku  $y = b_0 + b_1x$ , pás spolehlivosti kolem regresní přímky, který je vymezen hranicemi (1.91) pro  $x \in \mathbb{R}$  a pás spolehlivosti pro regresní přímku, který je odvozen v příkladu 1.12.2 (viz vzorec (1.117) na str. 105), jsou znázorněny na obrázku 1.23.



**Obrázek 1.23:** Hodnoty  $x_i$  a  $Y_i$ , regresní přímka  $y = b_0 + b_1x$ , pás spolehlivosti kolem regresní přímky (---) a pás spolehlivosti pro regresní přímku (.....) v případě lineárního modelu (1.13) a realizace náhodného vektoru  $\mathbf{Y}$  tak, jak je uvedena v příkladu 1.3.6 (viz strana 19). Bodů  $[x_i, y_i]$  je vidět pouze 6, neboť dva se překrývají.

### 1.9.5 Jednoduché třídění (pokračování ze str. 33)

Jelikož střední hodnota náhodného vektoru  $\mathbf{Y}$  v modelu (1.41) je

$$\boldsymbol{\mu} = \sum_{i=1}^I \mu_i \mathbf{a}_i,$$

platí

$$\begin{aligned} (\mathbf{Y} - \boldsymbol{\mu}) \circ \mathbf{a}_k / n_k &= \left( \mathbf{Y} \circ \mathbf{a}_k - \mathbf{a}_k \circ \sum_{i=1}^I \mu_i \mathbf{a}_i \right) / n_k = \\ &= \left( \sum_{j=1}^{n_k} Y_{kj} - \mu_k \|\mathbf{a}_k\|^2 \right) / n_k = \\ &= \bar{Y}_k - \mu_k. \end{aligned} \tag{1.92}$$

Délka vektoru  $\mathbf{a}_k/n_k$  je

$$\|\mathbf{a}_k/n_k\| = \|\mathbf{a}_k\|/n_k = \sqrt{n_k}/n_k = 1/\sqrt{n_k}.$$

Protože odhad rozptylu  $S^2$  jsme získali z pravoúhlého průmětu vektoru  $\mathbf{Y}$  do podprostoru  $M^\perp$  dimenze  $n - I$ , dostáváme dosazením  $\mathbf{a}_k/n_k$  za vektor  $\mathbf{z}$  do vztahu (1.79) tvrzení

$$\frac{(\mathbf{Y} - \boldsymbol{\mu}) \circ \mathbf{a}_k/n_k}{\|\mathbf{a}_k/n_k\| S} = \frac{(\bar{Y}_{k\cdot} - \mu_k) \sqrt{n_k}}{S} \sim t_{n-I},$$

které lze dle libosti využít k testování hypotéz či určování intervalů spolehlivosti.

## 1.10 Korelační koeficienty, koeficienty spolehlivosti

V případě výběru z vícerozměrného rozdělení představují výběrové korelační koeficienty odhady korelačních koeficientů (někdy pro zdůraznění rozdílu nazývaných populační korelační koeficienty). Ohledně vztahu mezi těmito odhady a skutečnými hodnotami nemá naše metoda co nabídnout. Naproti tomu v případě zformulovaného lineárního modelu umožňuje geometrický přístup poměrně efektivní odvození mnoha užitečných vztahů pro výběrové korelační koeficienty.

### Příklady

#### 1.10.1 Regresní přímka (pokračování ze str. 80)

##### Výběrový korelační koeficient

V modelu (1.19) definujme *výběrový korelační koeficient*  $r_{X,Y}$  vztahem

$$r_{X,Y} \equiv \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 (Y_i - \bar{Y})^2}} = \frac{(\mathbf{x} - \bar{x}) \circ (\mathbf{Y} - \bar{Y})}{\|\mathbf{x} - \bar{x}\| \cdot \|\mathbf{Y} - \bar{Y}\|} \quad (1.93)$$

(v rámci tohoto příkladu značme stručně  $r$ ). Podle této definice je tedy zřejmé

$$r = \cos \beta,$$

kde  $\beta$  je úhel, který svírají vektory  $\mathbf{x} - \bar{x}$  a  $\mathbf{Y} - \bar{Y}$ . Považujeme-li hodnoty  $x_i$  za realizace náhodné veličiny  $X$ , představuje hodnota  $r_{X,Y}$  určité měřítko lineární závislosti veličin  $X$  a  $Y$ . To můžeme velice přibližně vysvětlit tak, že je-li velikost úhlu  $\beta$  blízká hodnotě 0, resp.  $\pi$ , znamená to, že vektory  $\mathbf{x} - \bar{x}$  a  $\mathbf{Y} - \bar{Y}$  mají zhruba stejný, resp. opačný směr. Existuje tedy nějaké  $k \in \mathbb{R}^+$ , resp.  $k \in \mathbb{R}^-$ , pro které platí

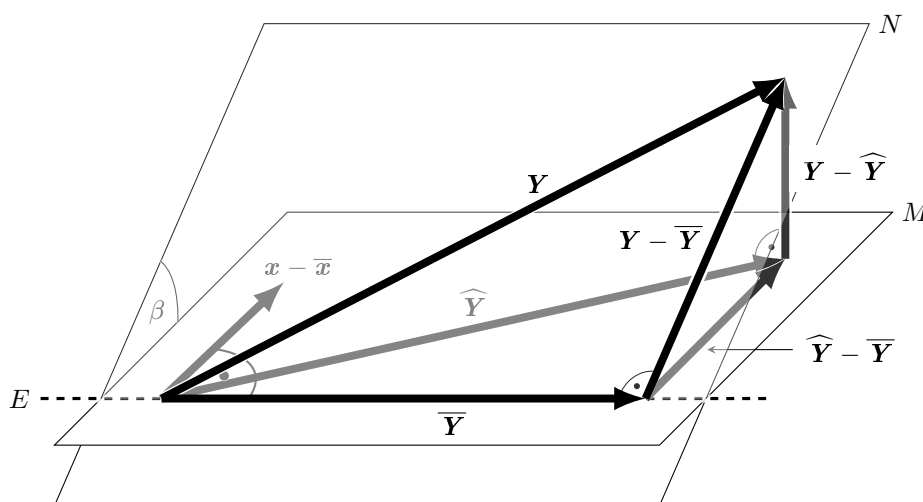
$$Y_i - \bar{Y} \doteq k(x_i - \bar{x}),$$

což je právě to, co si představujeme pod pojmem lineární závislost. O lineární závislosti tedy svědčí hodnoty  $r$  blízké hodnotám  $\pm 1$ .

Jelikož vektory  $\bar{\mathbf{x}}$ , resp.  $\bar{\mathbf{Y}}$ , představují pravoúhlý průmět vektoru  $\mathbf{x}$ , resp.  $\mathbf{Y}$ , do přímky  $E = [\mathbf{e}]$ , jsou vektory  $\mathbf{x} - \bar{\mathbf{x}}$  a  $\mathbf{Y} - \bar{\mathbf{Y}}$  na tuto přímku kolmé. Přímka  $E$  je průsečnicí rovin  $M = [\mathbf{e}, \mathbf{x}]$  a  $N \equiv [\mathbf{e}, \mathbf{Y}]$ , takže ve shodě se středoškolskou definicí odchylky dvou rovin<sup>19</sup> můžeme říci, že  $r$  reprezentuje – až na znaménko – kosinus odchylky rovin  $M, N$ .

### Koeficient determinace

Odchylku rovin  $M, N$  můžeme určit také jiným způsobem – jako úhel, který svírají vektory  $\mathbf{Y} - \bar{\mathbf{Y}}$  a  $\widehat{\mathbf{Y}} - \bar{\mathbf{Y}}$  (viz obr. 1.24). Druhý z těchto vektorů totiž



**Obrázek 1.24:** V případě modelu (1.19) představuje výběrový korelační koeficient  $r$  kosinus úhlu  $\beta$ , který svírají vektory  $\mathbf{x} - \bar{\mathbf{x}}$  a  $\mathbf{Y} - \bar{\mathbf{Y}}$ . To znamená, že až na znaménko je to také kosinus odchylky rovin  $M = [\mathbf{e}, \mathbf{x}]$  a  $N = [\mathbf{e}, \mathbf{Y}]$ , jehož absolutní hodnotu lze zjistit také z rozměrů pravoúhlého trojúhelníku tvořeného odvěsnami  $\mathbf{Y} - \widehat{\mathbf{Y}}$ ,  $\widehat{\mathbf{Y}} - \bar{\mathbf{Y}}$  a přeponou  $\mathbf{Y} - \bar{\mathbf{Y}}$ .

stejně jako vektor  $\mathbf{x} - \bar{\mathbf{x}}$  leží v rovině  $M$  a přitom je kolmý na přímku  $E$ , neboť můžeme psát

$$\widehat{\mathbf{Y}} - \bar{\mathbf{Y}} = (\mathbf{Y} - \bar{\mathbf{Y}}) - (\mathbf{Y} - \widehat{\mathbf{Y}}),$$

kde druhý ze sčítanců je kolmý z definice na přímku  $E$  a první je z definice kolmý na rovinu  $M$ , a tudíž i na přímku  $E \subset M$ . Protože vektory  $\mathbf{Y} - \bar{\mathbf{Y}}$  a  $\widehat{\mathbf{Y}} - \bar{\mathbf{Y}}$  tvoří přeponu a odvěsnu pravoúhlého trojúhelníku, dostáváme

$$r^2 = \cos^2 \beta = \frac{\|\widehat{\mathbf{Y}} - \bar{\mathbf{Y}}\|^2}{\|\mathbf{Y} - \bar{\mathbf{Y}}\|^2} = \frac{\|\widehat{\mathbf{Y}}\|^2 - \|\bar{\mathbf{Y}}\|^2}{\|\mathbf{Y}\|^2 - \|\bar{\mathbf{Y}}\|^2} = \frac{\sum_{i=1}^n \widehat{Y}_i^2 - n\bar{Y}^2}{\sum_{i=1}^n Y_i^2 - n\bar{Y}^2}.$$

<sup>19</sup>Připomeňme: odchylka rovin  $M, N$  je úhel, který svírají přímky  $p \subset M$ ,  $q \subset N$ , které jsou obě kolmé na průsečnici těchto rovin.

Tato hodnota se nazývá *koefficient determinace* a často se označuje též symbolem  $R^2$ . Protože výraz ve jmenovateli představuje celkovou variabilitu vektoru  $\mathbf{Y}$  (ve smyslu součtu čtverců odchylek od průměru), zatímco v čitateli je variabilita vektoru  $\widehat{\mathbf{Y}}$ , která je dána polohou tohoto vektoru v rovině  $M$  určené modelem (1.19), bývá koefficient determinace často interpretován jako podíl variability „vysvětlené modelem“ vůči celkové variabilitě. Opět platí, že čím je tato hodnota bližší jedné, tím průkaznější je lineární závislost veličin  $x$  a  $Y$ . Často bývá koefficient determinace definován vzorcem

$$R^2 = 1 - \frac{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2}{\|\mathbf{Y} - \overline{\mathbf{Y}}\|^2} = 1 - \frac{\sum_{i=1}^n Y_i^2 - \sum_{i=1}^n \widehat{Y}_i^2}{\sum_{i=1}^n Y_i^2 - n\overline{Y}^2};$$

je zřejmé, že tento vztah je variantou rovnosti

$$\cos^2 \beta = 1 - \sin^2 \beta.$$

### Výběrový korelační koeficient a rozdělení $F$

Pomocí veličiny  $r$  lze snadno vyjádřit i statistiku  $F$ , používanou pro test hypotézy  $\beta_1 = 0$  (viz vzorec (1.44)):

$$F = \frac{\frac{\|\widehat{\mathbf{Y}} - \overline{\mathbf{Y}}\|^2}{n-2}}{\frac{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2}{n-2}} = (n-2)\cotg^2 \beta = \frac{(n-2)\cos^2 \beta}{1 - \cos^2 \beta} = \frac{(n-2)r^2}{1 - r^2};$$

víme tedy, že je-li  $\beta_1 = 0$ , platí

$$\frac{(n-2)r^2}{1 - r^2} \sim F_{1, n-2}. \quad (1.94)$$

Toho lze využít k testu uvedené hypotézy: je-li konkrétní realizace této veličiny větší než kritická hodnota  $F_{1, n-2}$ , hypotézu můžeme zamítnout.

### Výběrový korelační koeficient a rozdělení $t$

K testu hypotézy  $\beta_1 = 0$  se častěji používá vztah (1.89); platí-li uvedená hypotéza, má podle tohoto vztahu náhodná veličina

$$\frac{b_1 \|\mathbf{q}\|}{S} = \frac{b_1 \|\mathbf{x} - \overline{\mathbf{x}}\|}{S} \quad (1.95)$$

rozdělení  $t_{n-2}$ . Abychom tuto veličinu vyjádřili pomocí koeficientu  $r$ , upravme nejprve vzorec (1.59) pro výpočet hodnoty  $b_1$ :

$$b_1 = \frac{(\mathbf{Y} - \overline{\mathbf{Y}}) \circ (\mathbf{x} - \overline{\mathbf{x}})}{\|\mathbf{x} - \overline{\mathbf{x}}\|^2} = r \frac{\|\mathbf{Y} - \overline{\mathbf{Y}}\|}{\|\mathbf{x} - \overline{\mathbf{x}}\|}.$$

Po dosazení do (1.95) dostáváme

$$\frac{r \frac{\|\mathbf{Y} - \overline{\mathbf{Y}}\|}{\|\mathbf{x} - \overline{\mathbf{x}}\|} \cdot \|\mathbf{x} - \overline{\mathbf{x}}\|}{\|\mathbf{Y} - \widehat{\mathbf{Y}}\| / \sqrt{n-2}} = \frac{r\sqrt{n-2}}{\sin \beta} = r\sqrt{\frac{n-2}{1-r^2}};$$

můžeme uzavřít, že za předpokladu platnosti hypotézy  $\beta_1 = 0$  platí

$$r\sqrt{\frac{n-2}{1-r^2}} \sim t_{n-2}. \quad (1.96)$$

To je ovšem výsledek, který jsme – vzhledem k již několikrát zmiňovanému vztahu mezi rozdělením  $t$  a  $F$  – mohli tušit již ze vztahu (1.94).

### Výběrový a populační korelační koeficient

Pokud hodnoty  $x_1, \dots, x_n$  v modelu (1.19) představují realizace složek  $X_i$  náhodného výběru

$$\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \dots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$$

z dvojrozměrného normálního rozdělení, lze ukázat (viz [3]), že obě výše odvozená pravidla pro rozdělení funkce koeficientu  $r$  plynou i z hypotézy  $\rho = 0$ , kde

$$\rho \equiv \frac{E(X - EX)(Y - EY)}{\sqrt{\text{var}X \cdot \text{var}Y}}$$

je korelační koeficient náhodných veličin  $X, Y$  (tj. populační korelační koeficient). Proto se v tomto případě mohou vztahy (1.94) a (1.96) použít k testování hypotézy  $\rho = 0$ , a proto jsou také výsledky tohoto testu ekvivalentní s výsledky testu hypotézy  $\beta_1 = 0$ .

### 1.10.2 Mnohonásobná regrese (pokračování ze str. 76)

V případě modelu (1.45) definujme *výběrový koeficient mnohonásobné korelace* vzorcem

$$r_{Y,X} \equiv \frac{(\mathbf{Y} - \bar{\mathbf{Y}}) \circ (\widehat{\mathbf{Y}} - \bar{\mathbf{Y}})}{\|\mathbf{Y} - \bar{\mathbf{Y}}\| \cdot \|\widehat{\mathbf{Y}} - \bar{\mathbf{Y}}\|}.$$

Jedná se tedy o kosinus úhlu  $\beta$ , který svírají vektory  $\mathbf{Y} - \bar{\mathbf{Y}}$  a  $\widehat{\mathbf{Y}} - \bar{\mathbf{Y}}$ ; protože druhý z těchto vektorů je pravoúhlým průmětem prvního do podprostoru  $M$ , lze úhel  $\beta$  interpretovat také jako odchylku vektoru  $\mathbf{Y} - \bar{\mathbf{Y}}$  od podprostoru  $M$ .

Jelikož vektory  $\mathbf{Y} - \bar{\mathbf{Y}}$  a  $\widehat{\mathbf{Y}} - \bar{\mathbf{Y}}$  tvoří přeponu a odvěsnu pravoúhlého trojúhelníku (přílehlou k úhlu  $\beta$ ), je hodnota koeficientu  $r_{Y,X}$  nutně nezáporná a lze ji z tohoto trojúhelníku vyjádřit také ve formě

$$r_{Y,X} = \frac{\|\widehat{\mathbf{Y}} - \bar{\mathbf{Y}}\|}{\|\mathbf{Y} - \bar{\mathbf{Y}}\|}.$$

Hodnota  $r_{Y,X}^2$  se opět nazývá *koeficient determinace*, značí se též  $R^2$  a má podobný význam jako v předchozím příkladu. Někdy bývá její definice uvedena ve formě

$$R^2 = 1 - \sin^2\beta = 1 - \frac{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2}{\|\mathbf{Y} - \bar{\mathbf{Y}}\|^2}.$$

## Výběrový koeficient mnohonásobné korelace a rozdělení F

Výraz (1.46) můžeme upravit na

$$F = \frac{\frac{\|\widehat{\mathbf{Y}} - \overline{\mathbf{Y}}\|^2}{k}}{\frac{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2}{n-k-1}} = \frac{(n-k-1)\cot^2\beta}{k} = \frac{(n-k-1)\cos^2\beta}{k(1-\cos^2\beta)};$$

víme tedy, že za předpokladu platnosti hypotézy  $\beta_1 = \dots = \beta_k = 0$  platí

$$\frac{(n-k-1)r_{\mathbf{Y},\mathbf{X}}^2}{k(1-r_{\mathbf{Y},\mathbf{X}}^2)} \sim F_{k,n-k-1}. \quad (1.97)$$

Opět lze ukázat (viz [1]), že představují-li řádky matice  $\mathbf{X}$  v modelu (1.45) realizace prvních  $k$  složek náhodného výběru

$$\begin{aligned} &(X_{11}, \dots, X_{1k}, Y_1), \\ &(X_{21}, \dots, X_{2k}, Y_2), \\ &\quad \vdots \\ &(X_{n1}, \dots, X_{nk}, Y_n) \end{aligned}$$

z mnohorozměrného normálního rozdělení, lze vztah (1.97) použít i k testování hypotézy  $\rho_{\mathbf{Y},\mathbf{X}} = 0$ , kde  $\rho_{\mathbf{Y},\mathbf{X}}$  je (populační) mnohorozměrný korelační koeficient.

### 1.10.3 Výběrový parciální korelační koeficient

Nechť pro náhodný vektor  $\mathbf{Y}$ , jehož realizace jsou prvky vektorového prostoru  $V_n$ , platí model

$$\mathbf{Y} = \beta_0 \mathbf{e} + \beta_1 \mathbf{x}_1 + \dots + \beta_k \mathbf{x}_k + \omega \mathbf{w} + \mathbf{Z}, \quad (1.98)$$

kde náhodný vektor  $\mathbf{Z}$  se řídí rozdělením  $N(\mathbf{0}; \sigma^2 \mathbf{I}_n)$ ,  $n > k + 2$  a vektory  $\mathbf{e}, \mathbf{x}_1, \dots, \mathbf{x}_k, \mathbf{w}$  jsou lineárně nezávislé. Označme

$$\begin{aligned} L &= [\mathbf{e}, \mathbf{x}_1, \dots, \mathbf{x}_k, \mathbf{w}], \\ M &= [\mathbf{e}, \mathbf{x}_1, \dots, \mathbf{x}_k]. \end{aligned}$$

Nechť  $\widehat{\mathbf{Y}}$  a  $\widehat{\mathbf{w}}$  jsou pravoúhlé průměty vektorů  $\mathbf{Y}$  a  $\mathbf{w}$  do podprostoru  $M$ . Vektory  $\mathbf{Y} - \widehat{\mathbf{Y}}$  a  $\mathbf{w} - \widehat{\mathbf{w}}$  tak představují ty části vektorů  $\mathbf{Y}, \mathbf{w}$ , které se nepovedlo vysvětlit pomocí vektorů  $\mathbf{x}_1, \dots, \mathbf{x}_k$ . Výběrový korelační koeficient vypočtený z těchto reziduí se nazývá *výběrový parciální korelační koeficient*  $r_{\mathbf{Y},\mathbf{w},\mathbf{X}}$ ; je to tedy hodnota, kterou získáme dosazením těchto vektorů do vzorce (1.93):

$$r_{\mathbf{Y},\mathbf{w},\mathbf{X}} = \frac{\left[ (\mathbf{Y} - \widehat{\mathbf{Y}}) - \overline{(\mathbf{Y} - \widehat{\mathbf{Y}})} \right] \circ \left[ (\mathbf{w} - \widehat{\mathbf{w}}) - \overline{(\mathbf{w} - \widehat{\mathbf{w}})} \right]}{\left\| (\mathbf{Y} - \widehat{\mathbf{Y}}) - \overline{(\mathbf{Y} - \widehat{\mathbf{Y}})} \right\| \cdot \left\| (\mathbf{w} - \widehat{\mathbf{w}}) - \overline{(\mathbf{w} - \widehat{\mathbf{w}})} \right\|},$$

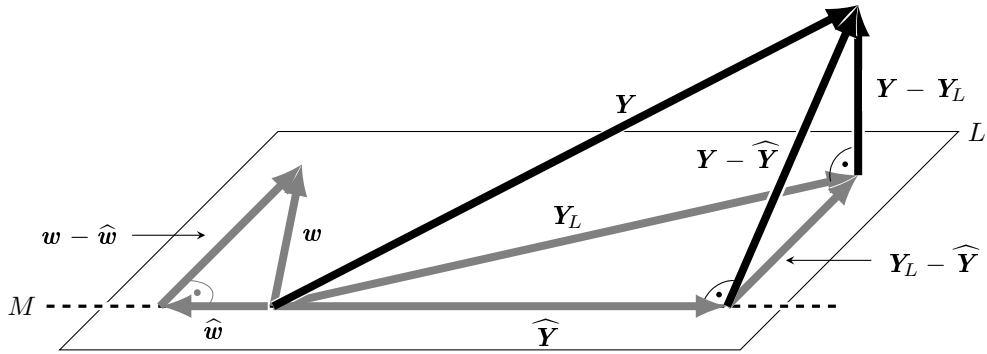
kde všechny souřadnice vektoru  $\overline{(\mathbf{Y} - \widehat{\mathbf{Y}})}$ , resp.  $\overline{(\mathbf{w} - \widehat{\mathbf{w}})}$ , jsou rovny průměrné hodnotě souřadnic vektoru  $\mathbf{Y} - \widehat{\mathbf{Y}}$ , resp. vektoru  $\mathbf{w} - \widehat{\mathbf{w}}$ . Protože však oba

posledně jmenované vektory jsou kolmé na podprostor  $M$ , a tudíž i na vektor  $e$ , je tato průměrná hodnota v obou případech rovna nule; můžeme tedy definici okamžitě zjednodušit na tvar

$$r_{Y,W,X} = \frac{(\mathbf{Y} - \widehat{\mathbf{Y}}) \circ (\mathbf{w} - \widehat{\mathbf{w}})}{\|\mathbf{Y} - \widehat{\mathbf{Y}}\| \cdot \|\mathbf{w} - \widehat{\mathbf{w}}\|}. \quad (1.99)$$

### Výběrový parciální korelační koeficient a rozdělení F

Vztah (1.99) představuje kosinus úhlu  $\beta$ , který svírají vektory  $\mathbf{Y} - \widehat{\mathbf{Y}}$  a  $\mathbf{w} - \widehat{\mathbf{w}}$ . Tento kosinus musí být až na znaménko stejný jako kosinus úhlu sevřeného vektory  $\mathbf{Y} - \widehat{\mathbf{Y}}$  a  $\mathbf{Y}_L - \widehat{\mathbf{Y}}$ , kde  $\mathbf{Y}_L$  je pravoúhlý průmět vektoru  $\mathbf{Y}$  do podprostoru  $L$ ; oba vektory  $\mathbf{w} - \widehat{\mathbf{w}}$  a  $\mathbf{Y}_L - \widehat{\mathbf{Y}}$  jsou totiž prvky podprostoru  $L - M$ , který je jednorozměrný, takže jsou rovnoběžné. Protože vektory  $\mathbf{Y} - \widehat{\mathbf{Y}}$  a  $\mathbf{Y}_L - \widehat{\mathbf{Y}}$  tvoří přeponu a odvěsnu pravoúhlého trojúhelníku (viz obr. 1.25), dostáváme vztahy



**Obrázek 1.25:** Výběrový parciální korelační koeficient  $r_{Y,W,X}$  představuje kosinus úhlu sevřeného vektory  $\mathbf{Y} - \widehat{\mathbf{Y}}$  a  $\mathbf{w} - \widehat{\mathbf{w}}$ , kde vektor  $\widehat{\mathbf{Y}}$ , resp.  $\widehat{\mathbf{w}}$ , představuje pravoúhlý průmět vektoru  $\mathbf{Y}$ , resp.  $\mathbf{w}$  do podprostoru  $M$  generovaného vektory  $e, \mathbf{x}_1, \dots, \mathbf{x}_k$ . Označme symbolem  $\mathbf{Y}_L$  průmět vektoru  $\mathbf{Y}$  do podprostoru  $L \equiv [e, \mathbf{x}_1, \dots, \mathbf{x}_k, \mathbf{w}]$ ; oba vektory  $\mathbf{w} - \widehat{\mathbf{w}}$  a  $\mathbf{Y}_L - \widehat{\mathbf{Y}}$  leží v jednorozměrném podprostoru  $L - M$ , takže jsou rovnoběžné. Hodnota koeficientu je tedy až na znaménko stejná jako kosinus úhlu sevřeného vektory  $\mathbf{Y} - \widehat{\mathbf{Y}}$  a  $\mathbf{Y}_L - \widehat{\mathbf{Y}}$ . Tyto vektory tvoří přeponu a odvěsnu pravoúhlého trojúhelníku. (Pro větší názornost jsou vektory ležící v podprostoru  $L$  vybarveny šedě.)

$$r_{Y,W,X}^2 = \frac{\|\mathbf{Y}_L - \widehat{\mathbf{Y}}\|^2}{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2} = 1 - \frac{\|\mathbf{Y} - \mathbf{Y}_L\|^2}{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2}.$$

Podle kapitoly 1.6 (místo podprostorů  $M$  a  $S$  zde máme podprostory  $L$  a  $M$ ) víme, že leží-li střední hodnota  $\boldsymbol{\mu}$  náhodného vektoru  $\mathbf{Y}$  v podprostoru  $M$  (tj. v modelu (1.98) platí  $\omega = 0$ ), má náhodná veličina

$$\begin{aligned} \frac{\|\mathbf{Y}_L - \widehat{\mathbf{Y}}\|^2 / 1}{\|\mathbf{Y} - \mathbf{Y}_L\|^2 / (n - k - 2)} &= (n - k - 2) \cot^2 \beta = \\ &= (n - k - 2) \frac{\cos^2 \beta}{1 - \cos^2 \beta} = \\ &= (n - k - 2) \frac{r_{Y,W,X}^2}{1 - r_{Y,W,X}^2} \end{aligned} \quad (1.100)$$

rozdělení  $F_{1,n-k-2}$ .

### Výběrový parciální korelační koeficient a rozdělení $t$

Protože dimenze podprostoru  $L$  určeného modelem (1.98) je  $k+2$ , je nestranným odhadem rozptylu náhodná veličina

$$S^2 = \frac{\|\mathbf{Y} - \mathbf{Y}_L\|^2}{n - k - 2}.$$

Jelikož je  $\mathbf{w} - \widehat{\mathbf{w}} \in L$ , vyplývá ze vzorce (1.79), že náhodná veličina

$$\begin{aligned} t &= \frac{(\mathbf{Y} - \boldsymbol{\mu}) \circ (\mathbf{w} - \widehat{\mathbf{w}})}{S \cdot \|\mathbf{w} - \widehat{\mathbf{w}}\|} = \\ &= \sqrt{n - k - 2} \cdot \frac{(\mathbf{Y} - \boldsymbol{\mu}) \circ (\mathbf{w} - \widehat{\mathbf{w}})}{\|\mathbf{Y} - \mathbf{Y}_L\| \cdot \|\mathbf{w} - \widehat{\mathbf{w}}\|} \end{aligned}$$

má rozdělení  $t_{n-k-2}$ . Chceme-li tento výraz vyjádřit pomocí koeficientu  $r_{Y,W,X}$ , povšimněme si nejprve, že v případě platnosti hypotézy  $\boldsymbol{\mu} \in M$  platí

$$(\mathbf{Y} - \boldsymbol{\mu}) \circ (\mathbf{w} - \widehat{\mathbf{w}}) = (\mathbf{Y} - \widehat{\mathbf{Y}}) \circ (\mathbf{w} - \widehat{\mathbf{w}}),$$

neboť tehdy vektor  $\widehat{\mathbf{Y}} - \boldsymbol{\mu}$  leží v podprostoru  $M$ , na který je vektor  $\mathbf{w} - \widehat{\mathbf{w}}$  kolmý, z čehož plyne rovnost

$$\begin{aligned} 0 &= (\widehat{\mathbf{Y}} - \boldsymbol{\mu}) \circ (\mathbf{w} - \widehat{\mathbf{w}}) = \\ &= [(\mathbf{Y} - \boldsymbol{\mu}) - (\mathbf{Y} - \widehat{\mathbf{Y}})] \circ (\mathbf{w} - \widehat{\mathbf{w}}). \end{aligned}$$

Zlomek ve výrazu (1.101) díky tomu můžeme upravit na tvar

$$\begin{aligned} \frac{(\mathbf{Y} - \boldsymbol{\mu}) \circ (\mathbf{w} - \widehat{\mathbf{w}})}{\|\mathbf{Y} - \mathbf{Y}_L\| \cdot \|\mathbf{w} - \widehat{\mathbf{w}}\|} &= \frac{(\mathbf{Y} - \widehat{\mathbf{Y}}) \circ (\mathbf{w} - \widehat{\mathbf{w}})}{\|\mathbf{Y} - \widehat{\mathbf{Y}}\| \cdot \|\mathbf{w} - \widehat{\mathbf{w}}\| \cdot \frac{\|\mathbf{Y} - \mathbf{Y}_L\|}{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|}} = \\ &= \frac{r_{Y,W,X}}{\sin \beta} = \\ &= \frac{r_{Y,W,X}}{\sqrt{1 - r_{Y,W,X}^2}}. \end{aligned}$$

Došli jsme tedy ke zjištění, že je-li v modelu (1.98) parametr  $w$  roven nule, tj. vektor  $\mathbf{Y}$  na hodnotách  $w_i$  nezávisí, má náhodná veličina

$$r_{Y,W,X} \sqrt{\frac{n - k - 2}{1 - r_{Y,W,X}^2}}$$

rozdělení  $t_{n-k-2}$ ; tento vztah se dal ovšem tušit již z rozdělení veličiny (1.100).

Opět platí, že za jistých okolností lze tento výsledek použít k testování hypotézy  $\rho_{Y,W,X} = 0$ , kde  $\rho_{Y,W,X}$  je (populační) parciální korelační koeficient (podrobnosti viz [1]).

### 1.10.4 Koeficient korelace v modelu bez absolutního členu

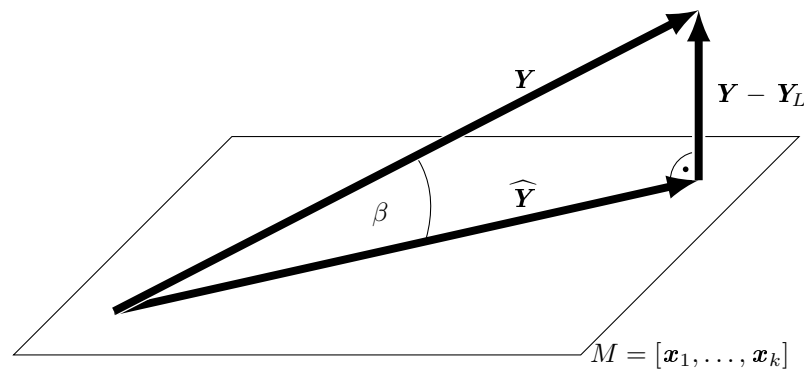
Modelem bez absolutního členu se míní model, ve kterém matice  $\mathbf{X}$  neobsahuje vektor  $\mathbf{e} = (1, \dots, 1)^T$ ; nechtť tedy pro náhodný vektor  $\mathbf{Y}$  platí model

$$\mathbf{Y} = b_1 \mathbf{x}_1 + \dots + b_k \mathbf{x}_k + \mathbf{Z},$$

kde náhodný vektor  $\mathbf{Z}$  má rozdělení  $N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$  a vektory  $\mathbf{x}_1, \dots, \mathbf{x}_k$  jsou lineárně nezávislé a různé od vektoru  $\mathbf{e}$ . Předpokládejme navíc, že vektor  $\mathbf{e}$  není ani lineární kombinací vektorů  $\mathbf{x}_1, \dots, \mathbf{x}_k$ , tj. neleží v podprostoru  $M$ , který tyto vektory generují. Tím pádem neplatí ani  $\overline{\mathbf{Y}} \in M$  a žádná z výše uvedených definic, která s vektorem  $\overline{\mathbf{Y}}$  počítá, nemá k předpokládanému modelu žádný relevantní vztah. Proto se v tomto případě někdy zavádí koeficient determinace vzorcem

$$R^2 = \frac{\|\widehat{\mathbf{Y}}\|^2}{\|\mathbf{Y}\|^2}$$

(viz [35]). Protože vektory  $\widehat{\mathbf{Y}}$  a  $\mathbf{Y}$  tvoří odvěsnu a přeponu pravoúhlého trojúhelníku (viz obr. 1.26), představuje hodnota  $R^2$  čtverec kosinu úhlu  $\beta$  sevřeného



**Obrázek 1.26:** V případě, že vektor  $\mathbf{e}$  neleží v podprostoru  $M$  určeném modelem (tj. model neobsahuje absolutní člen), zavádí se někdy koeficient korelace ve formě  $R^2 = \|\widehat{\mathbf{Y}}\|^2 / \|\mathbf{Y}\|^2$ , což představuje druhou mocninu kosinu úhlu  $\beta$ , který svírají vektory  $\widehat{\mathbf{Y}}$  a  $\mathbf{Y}$ .

těmito vektory a lze ji vyjádřit také ve tvaru

$$R^2 = 1 - \frac{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2}{\|\mathbf{Y}\|^2},$$

který odpovídá vzorci

$$\cos^2 \beta = 1 - \sin^2 \beta.$$

Nejedná se však o druhou mocninu žádného z výše uvedených výběrových korelačních koeficientů. Interpretace této veličiny je podobná jako v předchozích příkladech – představuje poměr variability vysvětlené modelem ku celkové variabilitě. Variabilitou je však míněn součet čtverců odchylek od předpokládané nulové střední hodnoty, nikoli od průměru.

I v tomto případě lze snadno odvodit pravidlo pro rozdělení vhodně zvolené funkce této veličiny: víme, že platí-li nulová hypotéza  $\boldsymbol{\mu} = \mathbf{0}$ , představují vektory

$\widehat{\mathbf{Y}}$  a  $\mathbf{Y} - \widehat{\mathbf{Y}}$  rozklad náhodného vektoru  $\mathbf{Y}$  s rozdělením  $\mathbf{N}(\mathbf{0}; \sigma^2 \mathbf{I})$  do dvou navzájem kolmých podprostorů  $M$  a  $V_n - M$  o dimenzích  $k$  a  $n - k$ , takže náhodná veličina

$$\begin{aligned} \frac{\|\widehat{\mathbf{Y}}\|^2 / k}{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2 / (n - k)} &= \cotg^2 \beta \cdot \frac{n - k}{k} = \\ &= \frac{\cos^2 \beta}{1 - \cos^2 \beta} \cdot \frac{n - k}{k} = \\ &= \frac{(n - k)R^2}{k(1 - R^2)} \end{aligned}$$

má rozdělení  $F_{k, n-k}$ . Možnost použití této veličiny nastává tedy tam, kde přichází v úvahu hypotéza  $\mathbf{E} \mathbf{Y} = \mathbf{0}$ .

## 1.11 Vazba regresních koeficientů v modelu s neúplnou hodnotí

V případě *modelu s neúplnou hodnotí* jsou sloupce matice  $\mathbf{X}$  v modelu (1.10) lineárně závislé, netvoří bázi podprostoru  $M$  a parametry  $\beta_i$  coby souřadnice vektoru  $\boldsymbol{\mu}$  vůči této skupině vektorů nejsou jednoznačně určeny. Totéž platí samozřejmě i pro složky  $b_i$  vektoru  $\mathbf{b}$ , které představují souřadnice vektoru  $\widehat{\mathbf{Y}}$  (který však jednoznačně určen je). Soustava (1.16) má tím pádem nekonečně mnoho řešení  $\mathbf{b}$ ; to nemusí být problém, pokud se zajímáme jen o pravoúhlý průmět vektoru  $\mathbf{Y}$  do podprostoru  $M$ , neboť všechna tato řešení vedou k jedinému vektoru  $\mathbf{X}\mathbf{b} = \widehat{\mathbf{Y}}$ . Pokud jsou ale předmětem našeho zájmu parametry  $\beta_i$ , je třeba tuto nesnáž překonat. Obvyklý postup je doplnit model (1.10) o lineární omezení vazajících koeficienty  $\beta_i$  takovým způsobem, že jsou určeny jednoznačně. Obecnější poznámky k tomuto tématu nalezne čtenář v kapitole 2.9. Na tomto místě rozebereme pouze dva důležité příklady.

### Příklady

#### 1.11.1 Jednoduché třídění (pokračování ze str. 84)

Zapišme vztah (1.40) ve tvaru

$$Y_{ij} \sim \mathbf{N}(\mu + \alpha_i, \sigma^2). \quad (1.101)$$

To znamená, že střední hodnoty jednotlivých souřadnic si představujeme jako součet dvou parametrů – základní úrovně  $\mu$ , společné všem souřadnicím, a hodnoty  $\alpha_i$ , reprezentující vliv  $i$ -té hladiny faktoru (tj.  $i$ -tého léku). Model (1.41) tedy nahrazujeme modelem

$$\mathbf{Y} = \mu \mathbf{e} + \sum_{i=1}^I \alpha_i \mathbf{a}_i + \mathbf{Z}, \quad (1.102)$$

kde vektor  $\mathbf{Z} = (Z_1, \dots, Z_7)^T$  má rozdělení  $\mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{I}_7)$ . Je zřejmé, že z geometrického hlediska se jedná o ekvivalentní modely, neboť vektory  $\mathbf{e}, \mathbf{a}_1, \dots, \mathbf{a}_I$  generují tentýž podprostor  $M$  jako vektory  $\mathbf{a}_1, \dots, \mathbf{a}_I$ . Sloupce matice nového modelu jsou však lineárně závislé, takže hodnoty  $\mu, \alpha_1, \dots, \alpha_I$  nejsou střední hodnotou  $\boldsymbol{\mu}$  jednoznačně určeny.

Přidáme-li ovšem podmínku

$$\sum_{i=1}^I n_i \alpha_i = 0,$$

zjistíme, že platí

$$\begin{aligned} \mathbf{e} \circ \sum_{i=1}^I \alpha_i \mathbf{a}_i &= \sum_{i=1}^I \alpha_i (\mathbf{e} \circ \mathbf{a}_i) = \\ &= \sum_{i=1}^I \alpha_i n_i = \\ &= 0. \end{aligned}$$

To znamená, že výraz

$$\boldsymbol{\mu} = \mu \mathbf{e} + \sum_{i=1}^I \alpha_i \mathbf{a}_i \quad (1.103)$$

představuje rozklad střední hodnoty  $\boldsymbol{\mu}$  do dvou navzájem kolmých podprostorů  $E = [e]$  a  $M - E$ . Protože první z nich je jednorozměrný, tvoří vektor  $\mathbf{e}$  jeho bázi a hodnota  $\mu$  – jakožto souřadnice pravoúhlého průmětu vektoru  $\boldsymbol{\mu}$  do podprostoru  $E$  vzhledem k této bázi – je jednoznačně určena. Ze srovnání (1.40) a (1.101) dále plyne

$$\alpha_i = \mu_i - \mu,$$

kde hodnoty  $\mu_i$  jsou – jakožto souřadnice vektoru  $\boldsymbol{\mu}$  vůči bázi  $\mathbf{a}_1, \dots, \mathbf{a}_I$  podprostoru  $M$  – rovněž jednoznačně určeny; jsou tedy jednoznačně určeny i hodnoty  $\alpha_i$  a je zřejmé, že se jedná o lineární funkce vektoru  $\boldsymbol{\mu}$ .

Nejlepším nestranným lineárním odhadem hodnot  $\alpha_i$  jsou tedy náhodné veličiny

$$a_i \equiv \bar{Y}_{i.} - \bar{Y};$$

odhadem hodnot  $\mu_i$  jsou totiž veličiny  $\bar{Y}_{i.}$  (jakožto souřadnice vektoru  $\widehat{\mathbf{Y}}$  vzhledem k bázi  $\mathbf{a}_1, \dots, \mathbf{a}_I$ ) a odhadem hodnoty  $\mu$  je veličina  $\bar{Y}$  (jakožto souřadnice pravoúhlého průmětu vektoru  $\widehat{\mathbf{Y}}$  do podprostoru  $E$  vzhledem k bázi  $\mathbf{e}$ <sup>20</sup>).

<sup>20</sup>Připomeňme, že platí-li  $\mathbf{e} \in M$ , je vektor  $\bar{\mathbf{Y}} = (\bar{Y}, \dots, \bar{Y})^T$  pravoúhlým průmětem jak vektoru  $\mathbf{Y}$ , tak vektoru  $\widehat{\mathbf{Y}}$  do podprostoru  $E$ . To plyne z toho, že platí

$$\begin{aligned} \mathbf{Y} - \bar{\mathbf{Y}} &\perp E, \\ \mathbf{Y} - \widehat{\mathbf{Y}} &\perp M \supset E, \end{aligned}$$

a tím pádem i

$$(\mathbf{Y} - \bar{\mathbf{Y}}) - (\mathbf{Y} - \widehat{\mathbf{Y}}) = \widehat{\mathbf{Y}} - \bar{\mathbf{Y}} \perp E.$$

## Odhady regresních koeficientů a rozdělení $t$

Abychom mohli použít vztah (1.79), uvědomme si, že platí

$$\begin{aligned} (\mathbf{Y} - \boldsymbol{\mu}) \circ \frac{\mathbf{e}}{n} &= \mathbf{Y} \circ \frac{\mathbf{e}}{n} - \left( \mu \mathbf{e} + \sum_{i=1}^I \alpha_i \mathbf{a}_i \right) \circ \frac{\mathbf{e}}{n} = \\ &= \bar{Y} - \mu \frac{\|\mathbf{e}\|^2}{n} - 0 = \\ &= \bar{Y} - \mu, \end{aligned}$$

a vzhledem k rovnosti (1.92) tedy také

$$\begin{aligned} (\mathbf{Y} - \boldsymbol{\mu}) \circ \left( \frac{\mathbf{a}_k}{n_k} - \frac{\mathbf{e}}{n} \right) &= (\bar{Y}_{k\cdot} - \mu_k) - (\bar{Y} - \mu) = \\ &= (\bar{Y}_{k\cdot} - \bar{Y}) - (\mu_k - \mu) = \\ &= a_k - \alpha_k. \end{aligned}$$

Protože odhad rozptylu  $S^2$  jsme získali z pravoúhlého průmětu vektoru  $\mathbf{Y}$  do podprostoru  $M^\perp$  dimenze  $n - I$ , dostáváme dosazením za vektor  $\mathbf{b}$  do vztahu (1.79) tvrzení

$$\frac{(\mathbf{Y} - \boldsymbol{\mu}) \circ \frac{\mathbf{e}}{n}}{\left\| \frac{\mathbf{e}}{n} \right\| S} = \frac{(\bar{Y} - \mu) \sqrt{n}}{S} \sim t_{n-I},$$

resp.

$$\frac{(\mathbf{Y} - \boldsymbol{\mu}) \circ \left( \frac{\mathbf{a}_k}{n_k} - \frac{\mathbf{e}}{n} \right)}{\left\| \frac{\mathbf{a}_k}{n_k} - \frac{\mathbf{e}}{n} \right\| S} = \frac{a_k - \alpha_k}{S} \sqrt{\frac{nn_k}{n - n_k}} \sim t_{n-I},$$

čehož lze využít k testování hypotéz a určování intervalů spolehlivosti.

### 1.11.2 Dvojné třídění (pokračování ze str. 65)

Sloupce matice modelu popsaného v podkapitole 1.8.5, které generují podprostory  $E$ ,  $A$ ,  $B$  a  $C$ , označme symboly  $\mathbf{e}$ ,  $\mathbf{a}_i$ ,  $\mathbf{b}_j$  a  $\mathbf{c}_{ij}$ . Pro střední hodnotu náhodného vektoru  $\mathbf{Y}$  tedy platí

$$\boldsymbol{\mu} = \mu \mathbf{e} + \sum_{i=1}^I \alpha_i \mathbf{a}_i + \sum_{j=1}^J \beta_j \mathbf{b}_j \quad (1.104)$$

v případě bez interakcí, resp.

$$\boldsymbol{\mu} = \mu \mathbf{e} + \sum_{i=1}^I \alpha_i \mathbf{a}_i + \sum_{j=1}^J \beta_j \mathbf{b}_j + \sum_{i=1}^I \sum_{j=1}^J \lambda_{ij} \mathbf{c}_{ij}, \quad (1.105)$$

jsou-li interakce zahrnuty. Protože se v obou případech jedná o model s neúplnou hodnotí, bývá doplněn soustavou podmínek

$$\sum_{i=1}^I \alpha_i = 0, \quad \sum_{j=1}^J \beta_j = 0,$$

v případě modelu s interakcemi ještě navíc

$$\sum_{i=1}^I \lambda_{ij} = 0 \quad (j = 1, \dots, J), \quad \sum_{j=1}^J \lambda_{ij} = 0 \quad (i = 1, \dots, I).$$

Podobně jako v předchozím příkladu lze snadno ukázat, že v důsledku těchto podmínek jsou vektory

$$\sum_{i=1}^I \alpha_i \mathbf{a}_i, \quad \sum_{j=1}^J \beta_j \mathbf{b}_j \quad \text{a} \quad \sum_{i=1}^I \sum_{j=1}^J \lambda_{ij} \mathbf{c}_{ij}$$

kolmé na podprostor  $E$ , vektory

$$\sum_{i=1}^I \alpha_i \mathbf{a}_i \quad \text{a} \quad \sum_{i=1}^I \sum_{j=1}^J \lambda_{ij} \mathbf{c}_{ij}$$

kolmé na podprostor  $B$  a vektory

$$\sum_{j=1}^J \beta_j \mathbf{b}_j \quad \text{a} \quad \sum_{i=1}^I \sum_{j=1}^J \lambda_{ij} \mathbf{c}_{ij}$$

kolmé na podprostor  $A$ . Sčítance na pravé straně výrazu (1.104), resp. (1.105), tedy představují rozklad vektoru  $\boldsymbol{\mu}$  do navzájem kolmých podprostorů  $E$ ,  $A - E$ ,  $B - E$ , resp.  $E$ ,  $A - E$ ,  $B - E$ ,  $N - (A + B)$ , jejichž součtem je podprostor  $M$ , resp. podprostor  $N$ .

Z toho v první řadě plyne, že vektor  $\mu \mathbf{e}$  představuje pravoúhlý průmět vektoru  $\boldsymbol{\mu}$  do podprostoru  $E$  a hodnota  $\mu$  – jakožto souřadnice tohoto průmětu vzhledem k bázi  $\{\mathbf{e}\}$  tohoto podprostoru – je vektorem  $\boldsymbol{\mu}$  jednoznačně určena. Jejím nejlepším nestranným lineárním odhadem je odpovídající souřadnice pravoúhlého průmětu vektoru  $\widehat{\mathbf{Y}}$  do podprostoru  $E$ . Tímto průmětem je vektor  $\overline{\mathbf{Y}} = (\overline{Y}, \dots, \overline{Y})^T$ ; pro odhad parametru  $\mu$  tedy použijeme náhodnou veličinu  $\overline{Y}$ .

Dále je vektor

$$\mu \mathbf{e} + \sum_{i=1}^I \alpha_i \mathbf{a}_i$$

pravoúhlým průmětem vektoru  $\boldsymbol{\mu}$  do podprostoru  $A$ . Hodnoty  $\mu + \alpha_i$  představují souřadnice tohoto průmětu vzhledem k bázi  $\{\mathbf{a}_1, \dots, \mathbf{a}_I\}$  tohoto podprostoru, jsou tedy jednoznačně určeny. Tím pádem jsou jednoznačně určeny i hodnoty  $\alpha_i$ . Pravoúhlým průmětem vektoru  $\widehat{\mathbf{Y}}$  do podprostoru  $A$  je vektor  $\mathbf{Y}_A$ ; jeho souřadnice  $\overline{Y}_{i..}$  vzhledem k bázi  $\{\mathbf{a}_1, \dots, \mathbf{a}_I\}$  jsou tedy nejlepšími nestrannými lineárními

odhady hodnot  $\mu + \alpha_i$ . Z toho plyne, že nejlepšími nestrannými lineárními odhady parametrů  $\alpha_i$  jsou náhodné veličiny  $\bar{Y}_{i\cdot} - \bar{Y}$ . Po analogické úvaze dojdeme k odhadům  $\bar{Y}_{\cdot j} - \bar{Y}$  pro parametry  $\beta_j$ .

Konečně v případě modelu s interakcemi představují hodnoty  $\mu_{ijk} \equiv \mu + \alpha_i + \beta_j + \lambda_{ij}$  jednoznačně určené souřadnice vektoru  $\boldsymbol{\mu}$  vzhledem k bázi podprostoru  $M$ , tvořené vektory  $\mathbf{c}_{ij}$ , z čehož vzhledem k jednoznačnosti hodnot  $\boldsymbol{\mu}$ ,  $\alpha_i$ ,  $\beta_j$  plyne jednoznačnost hodnot  $\lambda_{ij}$ . Jelikož odhadem hodnot  $\mu_{ijk}$  jsou náhodné veličiny  $\bar{Y}_{ij\cdot}$  (tj. souřadnice vektoru  $\widehat{\mathbf{Y}}$  k téže bázi) a platí

$$\lambda_{ij} = \mu_{ijk} - \mu - \alpha_i - \beta_j,$$

použijeme pro odhad parametrů  $\lambda_{ij}$  náhodné veličiny

$$\bar{Y}_{ij\cdot} - \bar{Y} - (\bar{Y}_{i\cdot} - \bar{Y}) - (\bar{Y}_{\cdot j} - \bar{Y}) = \bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}.$$

Doplňme, že všechny výše uvedené průměry lze snadno vyjádřit pomocí skalárního součinu s vhodně zvoleným vektorem z matice aktuálního modelu:

$$\bar{Y} = \frac{\mathbf{Y} \circ \mathbf{e}}{n}, \quad \bar{Y}_{i\cdot} = \frac{\mathbf{Y} \circ \mathbf{a}_i}{JK}, \quad \bar{Y}_{\cdot j} = \frac{\mathbf{Y} \circ \mathbf{b}_j}{IK}, \quad \bar{Y}_{ij\cdot} = \frac{\mathbf{Y} \circ \mathbf{c}_{ij}}{K},$$

což lze – podobně jako v předcházejícím příkladu – využít k odvození funkcí jednotlivých parametrů, které se řídí rozdělením  $t$ .

## 1.12 Aplikace Scheffého věty

Nechť podprostor  $M$  určený lineárním modelem (1.10) je dimenze  $m$  a  $A \subset M$  je nějaký jeho podprostor dimenze  $a < m$ . Označme symbolem  $\mathbf{Z}_A$  pravouhlý průmět náhodného vektoru  $\mathbf{Z} = \mathbf{Y} - \boldsymbol{\mu}$  do podprostoru  $A$  a symbolem  $\widehat{\mathbf{Y}}$  (jako obvykle) průmět vektoru  $\mathbf{Y}$  do podprostoru  $M$ . Jak víme, vektor  $\mathbf{Y} - \widehat{\mathbf{Y}}$  je v tom případě průmětem vektoru  $\mathbf{Z}$  do podprostoru  $M^\perp$ , jehož dimenze je  $n - m$ .

Podprostory  $A$ ,  $M^\perp$  jsou navzájem kolmé, můžeme tedy uplatnit vzorec (1.26) a usoudit, že platí

$$\frac{\|\mathbf{Z}_A\|^2 / a}{\|\widehat{\mathbf{Y}} - \mathbf{Y}\|^2 / (n - m)} \sim F_{a, n-m}.$$

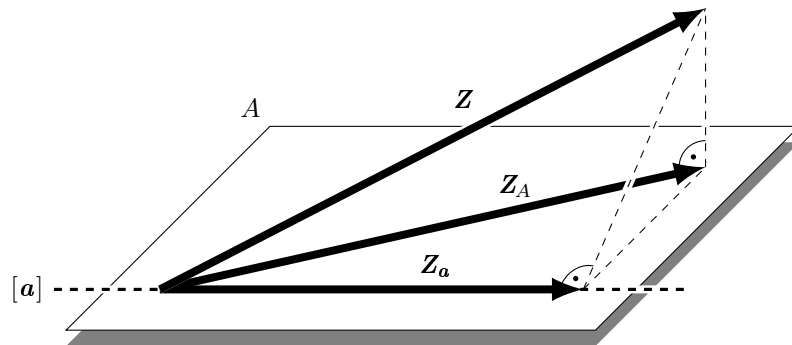
V kapitole 1.5 jsme odvodili, že jmenovatel tohoto výrazu je nestranným odhadem rozptylu  $\sigma^2$  a zavedli jsme pro něj označení  $S^2$ . Platí tedy, že pravděpodobnost splnění nerovnosti

$$\frac{\|\mathbf{Z}_A\|^2 / a}{S^2} \leq F_{a, n-m}(\alpha)$$

ekvivalentní s nerovností

$$\|\mathbf{Z}_A\| \leq S \sqrt{a \cdot F_{a, n-m}(\alpha)} \tag{1.106}$$

je  $1 - \alpha$ .



**Obrázek 1.27:** Je-li vektor  $\mathbf{a}$  prvkem podprostoru  $A$  a  $\mathbf{Z}_A$ , resp.  $\mathbf{Z}_a$ , jsou pravoúhlými průměty vektoru  $\mathbf{Z}$  do podprostoru  $A$ , resp.  $[a]$ , je vektor  $\mathbf{Z}_A - \mathbf{Z}_a$  kolmý na podprostor  $[a]$ ; je totiž rozdílem vektorů  $\mathbf{Z} - \mathbf{Z}_a$  a  $\mathbf{Z} - \mathbf{Z}_A$ , které jsou oba z definice kolmé na podprostor  $[a]$ . To znamená, že vektor  $\mathbf{Z}_a$  je pravoúhlým průmětem vektoru  $\mathbf{Z}_A$  do podprostoru  $[a]$ . Proto musí platit  $\|\mathbf{Z}_a\| \leq \|\mathbf{Z}_A\|$ .

Nechť nyní  $\mathbf{a}$  je libovolný vektor ležící v podprostoru  $A$ . Pravoúhlý průmět vektoru  $\mathbf{Z}$  do jednorozměrného podprostoru  $[a]$  označme  $\mathbf{Z}_a$ . Tento vektor je zároveň pravoúhlým průmětem vektoru  $\mathbf{Z}_A$  do podprostoru  $[a]$  (viz obr. 1.27); to znamená, že musí platit

$$\|\mathbf{Z}_a\| \leq \|\mathbf{Z}_A\|.$$

Zároveň lze však v případě jakékoli realizace náhodného vektoru  $\mathbf{Z}$  najít vektor  $\mathbf{a} \in A$  takový, že  $\|\mathbf{Z}_a\| = \|\mathbf{Z}_A\|$  – stačí jej zvolit tak, aby byl rovnoběžný s danou realizací vektoru  $\mathbf{Z}_A$ . Je tedy

$$\|\mathbf{Z}_A\| = \max \{ \|\mathbf{Z}_a\| : \mathbf{a} \in A \}, \quad (1.107)$$

takže nerovnost (1.106) je ekvivalentní s nerovností

$$\forall \mathbf{a} \in A : \|\mathbf{Z}_a\| \leq S \sqrt{a \cdot F_{a,n-m}(\alpha)}. \quad (1.108)$$

Vektor  $\mathbf{Z}_a$  lze snadno vyjádřit užitím vzorce (2.32):

$$\mathbf{Z}_a = \frac{\mathbf{a} \circ (\mathbf{Y} - \boldsymbol{\mu})}{\|\mathbf{a}\|^2} \mathbf{a},$$

jeho délka je tedy

$$\|\mathbf{Z}_a\| = \frac{|\mathbf{a} \circ (\mathbf{Y} - \boldsymbol{\mu})|}{\|\mathbf{a}\|};$$

po dosazení a jednoduché úpravě tak docházíme k formulaci

$$\mathbb{P} \left[ \forall \mathbf{a} \in A : |\mathbf{a} \circ (\mathbf{Y} - \boldsymbol{\mu})| \leq \|\mathbf{a}\| S \sqrt{a \cdot F_{a,n-m}(\alpha)} \right] = 1 - \alpha,$$

resp. její obměně

$$\mathbb{P} \left[ \exists \mathbf{a} \in A : |\mathbf{a} \circ (\mathbf{Y} - \boldsymbol{\mu})| > \|\mathbf{a}\| S \sqrt{a \cdot F_{a,n-m}(\alpha)} \right] = \alpha. \quad (1.109)$$

Toto tvrzení představuje speciální případ *Scheffého věty*.<sup>21</sup> V následujících příkladech si uvedeme dvě její důležité aplikace.

<sup>21</sup>Obecnější verze pro případ  $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}; \mathbf{V})$ , kde  $\mathbf{V}$  je pozitivně semidefinitní matice, je uvedena v publikaci [3].

## Příklady

### 1.12.1 Jednoduché třídění (pokračování ze str. 93)

Nejprve si připomeneme princip standardního  $F$ -testu hypotézy

$$H_0: \mu_1 = \cdots = \mu_I,$$

který jsme odvodili v příkladu 1.6.6. Označme pro stručnost symbolem  $A$  podprostor  $M - E$ ; jeho dimenze je  $I - 1$ . Platí tedy

$$\frac{\|\mathbf{Z}_A\|^2 / (I - 1)}{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2 / (n - I)} = \frac{\|\mathbf{Z}_A\|^2 / (I - 1)}{S^2} \sim F_{I-1, n-I},$$

tj.

$$\mathbb{P} \left[ \frac{\|\mathbf{Z}_A\|^2}{(I - 1)S^2} \leq F_{I-1, n-I}(\alpha) \right] = 1 - \alpha.$$

Položíme-li  $r \equiv S\sqrt{(I - 1)F_{I-1, n-I}(\alpha)}$ , dostáváme

$$\mathbb{P} \left[ \|\mathbf{Z}_A\| \leq r \right] = 1 - \alpha. \quad (1.110)$$

Hypotézu  $H_0$  lze ekvivalentně vyjádřit ve tvaru

$$H_0: \boldsymbol{\mu} \in E,$$

tj.

$$H_0: \boldsymbol{\mu} \perp A;$$

v případě její platnosti se tedy střední hodnota  $\boldsymbol{\mu}$  při projekci vektoru  $\mathbf{Z} = \mathbf{Y} - \boldsymbol{\mu}$  do podprostoru  $A$  eliminuje a platí  $\mathbf{Z}_A = \mathbf{Y}_A$ . Z tvrzení (1.110) tak dostáváme, že za předpokladu platnosti hypotézy  $H_0$  nastane nerovnost

$$\|\mathbf{Y}_A\| \leq r$$

s pravděpodobností  $1 - \alpha$ ; pokud tedy délka vektoru  $\mathbf{Y}_A$  překročí uvedenou mez  $r$ , hypotézu  $H_0$  zamítneme na hladině  $\alpha$ .

#### Test hypotézy $\mu_i = \mu_j$

Pokud při použití výše popsaného testu zamítneme hypotézu  $H_0$ , vzniká přirozeně otázka, jak rozhodnout, pro které dvojice  $i, j$  ( $1 \leq i \neq j \leq I$ ) není rovnost  $\mu_i = \mu_j$  splněna. Na ni nám však tato metoda nedává žádnou odpověď.

Zkusme tedy zvolit jiný postup a testujme každou z těchto rovností jako samostatnou hypotézu  $H_{ij}$ . Každá z těchto hypotéz představuje redukcí podprostoru  $M$  generovaného vektory  $\mathbf{a}_1, \dots, \mathbf{a}_I$  na podprostor  $S_{ij}$ , který je generován stejnou skupinou vektorů, až na to, že vektory  $\mathbf{a}_i, \mathbf{a}_j$  jsou v ní nahrazeny jediným vektorem  $\mathbf{a}_i + \mathbf{a}_j$ . Dimenze podprostoru  $S_{ij}$  je o jednotku nižší než dimenze

podprostoru  $M$ , podprostor  $M - S_{ij}$  je tedy jednorozměrný. Je generován např. vektorem

$$\mathbf{a}_{ij} \equiv \frac{\mathbf{a}_i}{n_i} - \frac{\mathbf{a}_j}{n_j},$$

neboť tento vektor zřejmě leží v podprostoru  $M$  a přitom – jak lze snadno ověřit – je kolmý na všechny generátory podprostoru  $S_{ij}$ . Náhodná veličina

$$\frac{\|\mathbf{Z}_{\mathbf{a}_{ij}}\|^2/1}{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2/(n-I)} = \frac{\|\mathbf{Z}_{\mathbf{a}_{ij}}\|^2}{S^2}$$

má tím pádem rozdělení  $F_{1,n-I}$ , takže platí

$$P \left[ \|\mathbf{Z}_{\mathbf{a}_{ij}}\| \leq S \sqrt{F_{1,n-I}(\alpha)} \right] = 1 - \alpha. \quad (1.111)$$

Hypotézu  $H_{ij}$ :  $\boldsymbol{\mu} \in S_{ij}$  lze ekvivalentně vyjádřit ve formě

$$H_{ij}: \boldsymbol{\mu} \perp \mathbf{a}_{ij},$$

takže v případě její platnosti se střední hodnota  $\boldsymbol{\mu}$  při projekci vektoru  $\mathbf{Z} = \mathbf{Y} - \boldsymbol{\mu}$  do podprostoru  $[\mathbf{a}_{ij}]$  eliminuje a platí  $\mathbf{Z}_{\mathbf{a}_{ij}} = \mathbf{Y}_{\mathbf{a}_{ij}}$ , kde  $\mathbf{Y}_{\mathbf{a}_{ij}}$  je pravoúhlý průmět vektoru  $\mathbf{Y}$  do podprostoru  $[\mathbf{a}_{ij}]$ . Ze vztahu (1.111) tak dostáváme, že v případě platnosti hypotézy  $H_{ij}$  nastane nerovnost

$$\|\mathbf{Y}_{\mathbf{a}_{ij}}\| \leq S \sqrt{F_{1,n-I}(\alpha)}$$

s pravděpodobností  $1 - \alpha$ ; nebude-li tedy tato nerovnost v případě konkrétní realizace splněna, hypotézu  $H_{ij}$  zamítneme na hladině významnosti  $\alpha$ .

Problém však je v tom, že pokud zamítneme alespoň jednu z hypotéz  $H_{ij}$ , musíme samozřejmě zamítnout i hypotézu  $H_0$ ; a jelikož v případě jednotlivých testů je pravděpodobnost nesprávného zamítnutí platné hypotézy  $\alpha$ , je při provedení více testů a platnosti hypotézy  $H_0$  pravděpodobnost nesprávného zamítnutí alespoň jedné z hypotéz  $H_{ij}$  (a tedy i hypotézy  $H_0$ ) vyšší než  $\alpha$ . To znamená, že co se týče testu hypotézy  $H_0$ , nemáme při použití tohoto postupu pod kontrolou pravděpodobnost chyby prvního druhu. To, co potřebujeme k překonání této obtíže, je otestovat všechny hypotézy  $H_{ij}$  „najednou“.

### Scheffého metoda mnohonásobných porovnávání

Všimněme si tedy dále, že všechny vektory  $\mathbf{a}_{ij}$  jsou kolmé na vektor  $\mathbf{e}$ , leží tudíž v podprostoru  $M - E = A$ . Ze vztahů (1.110) a (1.107) plyne, že pravděpodobnost jevu

$$\forall \mathbf{a} \in A : \|\mathbf{Z}_{\mathbf{a}}\| \leq r, \quad (1.112)$$

je  $1 - \alpha$ . My se však nezajímáme o všechny vektory  $\mathbf{a} \in M - E$ , nýbrž jen o některé z nich – totiž o vektory  $\mathbf{a}_{ij}$ ; jelikož jevu

$$\forall i, j : \|\mathbf{Z}_{\mathbf{a}_{ij}}\| \leq r$$

je nutným důsledkem jevu (1.112), pravděpodobnost jeho uskutečnění je *alespoň*  $1 - \alpha$ . V případě platnosti hypotézy  $H_0$  (a tedy i všech hypotéz  $H_{ij}$ ) platí  $\mathbf{Z}_{\mathbf{a}_{ij}} = \mathbf{Y}_{\mathbf{a}_{ij}}$ , takže nastane-li v případě konkrétní realizace jev

$$\exists i, j : \|\mathbf{Y}_{\mathbf{a}_{ij}}\| > r,$$

opravňuje nás to k zamítnutí hypotézy  $H_0$  na hladině významnosti *nejvýše*  $\alpha$ , a zároveň pro tuto dvojici zamítneme hypotézu  $H_{ij} : \mu_i = \mu_j$ .

Zbývají estetické úpravy: platí

$$\|\mathbf{Y}_{\mathbf{a}_{ij}}\| = \frac{|\mathbf{Y} \circ \mathbf{a}_{ij}|}{\|\mathbf{a}_{ij}\|} = (\bar{Y}_i - \bar{Y}_j) \sqrt{\frac{n_i n_j}{n_i + n_j}},$$

takže  $H_0$  a současně  $H_{ij}$  zamítneme v případě splnění nerovnosti

$$|\bar{Y}_i - \bar{Y}_j| > r \sqrt{\frac{n_i + n_j}{n_i n_j}},$$

tj.

$$|\bar{Y}_i - \bar{Y}_j| > S \sqrt{\frac{(n_i + n_j)(I - 1) F_{I-1, n-m}(\alpha)}{n_i n_j}}.$$

### Porovnání s tradiční analýzou rozptylu

V případě tradičního  $F$ -testu ověřujeme hypotézu  $H_0$  porovnáním hodnoty  $r$  s délkou vektoru

$$\mathbf{Y}_A = \mathbf{Y}_M - \mathbf{Y}_E = \widehat{\mathbf{Y}} - \bar{\mathbf{Y}}.$$

Tento postup si tedy můžeme představit tak, že v  $(I - 1)$ -rozměrném podprostoru  $M - E = A$  vytvoříme kouli  $K$  se středem v počátku soustavy souřadnic a poloměrem  $r$ ; leží-li v případě konkrétní realizace vektor  $\widehat{\mathbf{Y}} - \bar{\mathbf{Y}}$  vně této koule, zamítneme hypotézu  $H_0$  na hladině významnosti  $\alpha$ .

Vzhledem k rovnosti (1.107) bychom stejného výsledku dosáhli, pokud bychom s hodnotou  $r$  poměřovali všechny vektory  $\mathbf{Y}_{\mathbf{a}}$ , kde  $\mathbf{a} \in A$ . V případě Scheffého metody mnohonásobného porovnávání tak však činíme pouze s vektory  $\mathbf{Y}_{\mathbf{a}_{ij}}$ . Tyto vektory můžeme získat jako pravoúhlé průměty vektoru  $\mathbf{Y}_A = \widehat{\mathbf{Y}} - \bar{\mathbf{Y}}$  do podprostorů  $[\mathbf{a}_{ij}]$ ; pokud některý z těchto průmětů zasahuje vně koule  $K$ , zamítneme hypotézu  $H_0$  na hladině významnosti nejvýše  $\alpha$ , a zároveň víme, pro jakou dvojici  $i, j$  bychom měli zamítnout hypotézu  $H_{ij}$ .

K této situaci dojde zřejmě tehdy, když vektor  $\widehat{\mathbf{Y}} - \bar{\mathbf{Y}}$  zasahuje vně  $(I - 1)$ -rozměrného tělesa  $T \subset M - E$ , jehož hranice tvoří části lineárních množin kolmých na přímkách  $[\mathbf{a}_{ij}]$  a ležících ve vzdálenosti  $r$  od počátku. Protože vektorů  $\mathbf{a}_{ij}$  je celkem

$$\binom{I}{2} = \frac{I(I - 1)}{2}$$

a každý určuje dvě stěny tělesa  $T$ , je tímto tělesem  $I(I - 1)$ -stěn opsaný kouli  $K$ .

Tím pádem může nastat taková situace, že platí

$$\widehat{\mathbf{Y}} - \overline{\mathbf{Y}} \notin K, \quad \widehat{\mathbf{Y}} - \overline{\mathbf{Y}} \in T;$$

při použití klasické analýzy rozptylu pak hypotézu  $H_0$  zamítneme, zatímco při použití mnohonásobného porovnávání nikoli (opačný případ ovšem možný není). To je právě projevem toho, že hladina testu prováděného Scheffého metodou je nižší.

Přesná hladina významnosti Scheffého metody je zřejmě pravděpodobnost, že – za předpokladu platnosti hypotézy  $H_0$  – nastane jev

$$\widehat{\mathbf{Y}} - \overline{\mathbf{Y}} \notin T.$$

### Ilustrace pro případ $I = 3$

Tyto úvahy lze snadno znázornit v případě, kdy platí  $I = 3$ ; tehdy si můžeme podprostor  $M$  představit jako trojrozměrný prostor, ve kterém souřadnicové osy představují směry vektorů  $\mathbf{a}_1, \mathbf{a}_2$  a  $\mathbf{a}_3$ . Vektor  $\mathbf{e} = \mathbf{a}_1 + \mathbf{a}_2 + \mathbf{a}_3$  leží v prvním oktantu; ve speciálním případě vyváženého třídění spolu svírají dvojice vektorů  $\mathbf{a}_i$  a  $\mathbf{e}$  úhly stejné velikosti. Podprostor  $M - E$  se redukuje na rovinu kolmou na tento vektor (představujme si ji jako procházející počátkem) a koule  $K$  na kruh ležící v této rovině.

Co se týče vektorů  $\mathbf{a}_{ij}$ , každý z nich leží v rovině  $[\mathbf{a}_i, \mathbf{a}_j]$  a zároveň v rovině  $M - E$ , leží tedy na průsečnicích roviny  $M - E$  se třemi rovinami určenými souřadnicovými osami. Pokud v rovině  $M - E$  sestrojíme přímky, které jsou na tyto průsečnice kolmé a leží ve vzdálenosti  $r$  od počátku souřadnic, ohraničují tyto přímky šestiúhelník  $T$  opsaný kruhu  $K$  (viz obr. 1.28).

Ve speciálním případě vyváženého třídění mají všechny vektory  $\mathbf{a}_{ij}$  stejnou délku a svírají navzájem vždy stejný úhel, tj.  $2\pi/3$ ; šestiúhelník  $T$  je pak pravidelný.

### 1.12.2 Pás spolehlivosti pro regresní přímku

Nechť pro náhodný vektor  $\mathbf{Y}$  platí model (1.19). V příkladu 1.9.4 jsme pro pevně zvolené  $x \in \mathbb{R}$  našli vektor  $\mathbf{z}(x)$  (značme jej takto místo původního symbolu  $\mathbf{z}$ , abychom zdůraznili jeho závislost na hodnotě  $x$ ), pro který platí

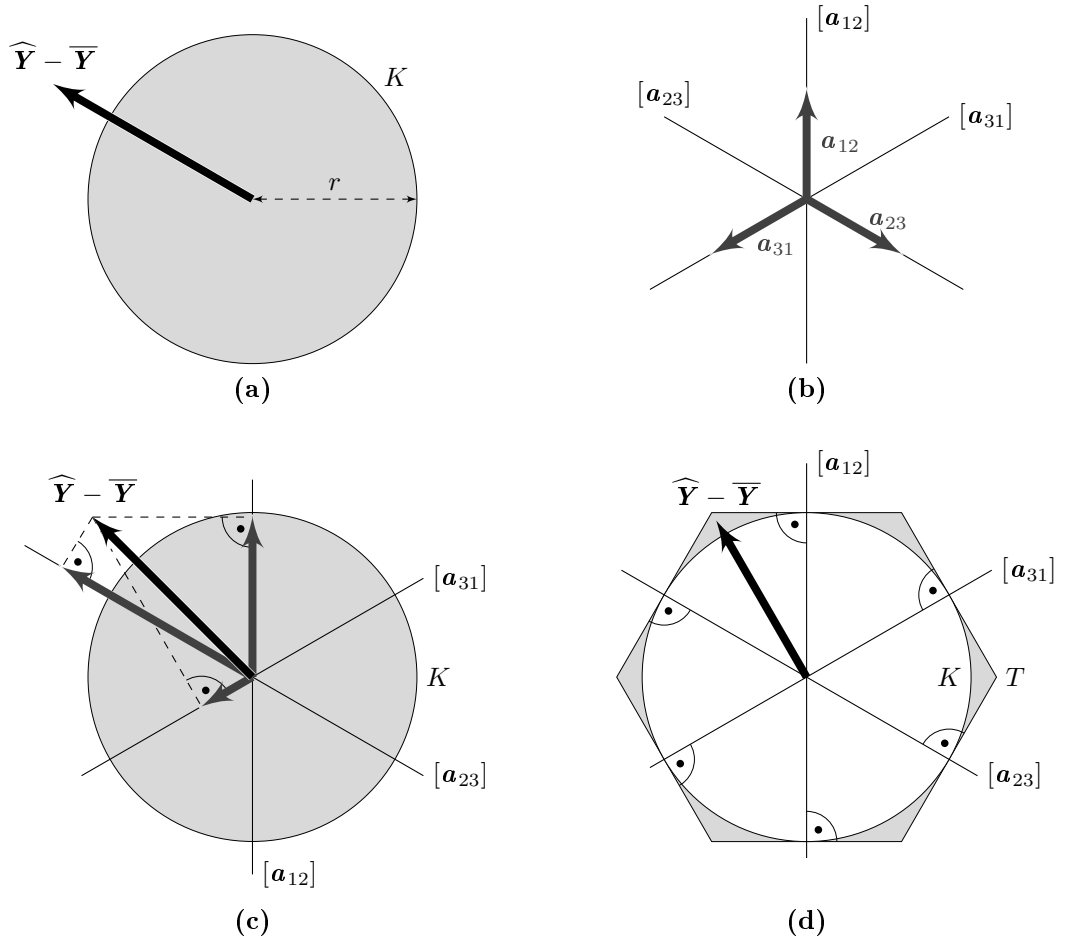
$$\boldsymbol{\mu} \circ \mathbf{z}(x) = \beta_0 + \beta_1 x, \quad \mathbf{Y} \circ \mathbf{z} = b_0 + b_1 x.$$

Tento vektor je určen vzorcem

$$\mathbf{z}(x) = \frac{\mathbf{e}}{n} + (x - \bar{x}) \frac{\mathbf{q}}{\|\mathbf{q}\|^2},$$

kde  $\bar{x}$  je průměrná hodnota souřadnic vektoru  $\mathbf{x}$  a vektor  $\mathbf{q}$  je pravoúhlý průmět vektoru  $\mathbf{x}$  do podprostoru  $M - E$ , je tedy kolmý na vektor  $\mathbf{e}$ . Označme symbolem  $P$  lineární množinu, kterou vytvoří všechny možné vektory  $\mathbf{z}(x)$ , když necháme proměnnou  $x$  probíhat celou množinou  $\mathbb{R}$ . Je zřejmé, že  $P$  je podmnožinou podprostoru  $M$ . Z toho plyne, že pravděpodobnost jevu

$$\forall \mathbf{a} \in P : \|\mathbf{Z}_a\| \leq S \sqrt{2F_{2,n-2}(\alpha)} \quad (1.113)$$



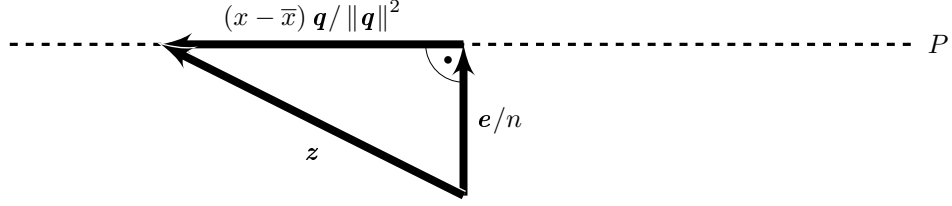
**Obrázek 1.28:** Je-li v případě jednoduchého třídění  $I = 3$ , představuje podprostor  $M - E$  rovinu. (a) Při klasickém  $F$ -testu hypotézy  $H_0: \mu_1 = \mu_2 = \mu_3$  hypotézu zamítneme na hladině  $\alpha$ , pokud vektor  $\widehat{\mathbf{Y}} - \overline{\mathbf{Y}}$  zasahuje vně kruhu  $K$  o poloměru  $r = S\sqrt{2F_{2,n-3}(\alpha)}$ . (b) Při použití Scheffého metody mnohonásobného porovnání promítneme nejdříve vektor  $\widehat{\mathbf{Y}} - \overline{\mathbf{Y}}$  na přímky  $[\mathbf{a}_{ij}]$ , kde  $\mathbf{a}_{ij} = \mathbf{a}_i/n_i - \mathbf{a}_j/n_j \in M - E$ . (c) Hypotézu  $H_0$  pak zamítneme tehdy, když některý z těchto průmětů zasahuje vně kruhu  $K$ ; současně získáváme informaci o tom, pro které dvojice  $i, j$  zřejmě platí  $\mu_i \neq \mu_j$  (podle našeho obrázku bychom tedy došli k závěru, že  $\mu_2 \neq \mu_3$ ). (d) Hladina tohoto testu je tedy rovna pravděpodobnosti, že (za předpokladu platnosti hypotézy  $H_0$ ) bude vektor  $\widehat{\mathbf{Y}} - \overline{\mathbf{Y}}$  ležet vně šestiúhelníku  $T$  opsaného kruhu  $K$ , jehož strany jsou kolmé na přímky  $[\mathbf{a}_{ij}]$ . Leží-li koncový bod vektoru  $\widehat{\mathbf{Y}} - \overline{\mathbf{Y}}$  vně  $K$ , ale uvnitř  $T$  (šedě vybarvená zóna), zamítneme  $H_0$  při použití první metody, avšak nikoli při použití druhé.

je alespoň  $1 - \alpha$ ; tento jev je totiž nutným důsledkem jevu

$$\forall \mathbf{a} \in M : \|\mathbf{Z}_{\mathbf{a}}\| \leq S\sqrt{2F_{2,n-2}(\alpha)}, \quad (1.114)$$

který podle vztahu (1.108) nastane s pravděpodobností právě  $1 - \alpha$ .

Ukážeme však, že zároveň jev (1.113) implikuje jev (1.114). Nejdříve si uvědomme, že pravoúhlá projekce do jednorozměrného podprostoru určeného jediným vektorem závisí pouze na směru tohoto vektoru, nikoli na jeho délce. Prvky množiny  $P$  reprezentují všechny směry podprostoru  $M = [\mathbf{e}, \mathbf{q}]$  s výjimkou směru vektoru  $\mathbf{q}$  (viz obr. 1.29). Platí-li tedy nerovnost



**Obrázek 1.29:** Necháme-li proměnnou  $x$  probíhat množinu všech reálných čísel, tvoří vektory  $\mathbf{z} = \mathbf{e}/n + (x - \bar{x}) \mathbf{q} / \|\mathbf{q}\|^2$  jednorozměrnou lineární množinu  $P$ . Vektory  $\mathbf{z}$  reprezentují všechny směry podprostoru  $M = [\mathbf{e}, \mathbf{q}]$  s výjimkou směru vektoru  $\mathbf{q}$ .

$$\|\mathbf{Z}_a\| \leq S\sqrt{2F_{2,n-2}(\alpha)} \quad (1.115)$$

pro všechny vektory  $\mathbf{a} \in P$ , znamená to, že platí pro všechny vektory  $\mathbf{a} \in M$ , které nejsou rovnoběžné s vektorem  $\mathbf{q}$ . Speciálně se jedná o vektory jednotkové délky, které lze vyjádřit ve tvaru

$$\mathbf{a} = t \frac{\mathbf{e}}{\|\mathbf{e}\|} + s \frac{\mathbf{q}}{\|\mathbf{q}\|}, \quad (1.116)$$

kde  $t, s \in \mathbb{R}$ ,  $t \neq 0$ ,  $t^2 + s^2 = 1$ . Pro tyto vektory můžeme s využitím vzorce (2.33) psát

$$\lim_{t \rightarrow 0} \|\mathbf{Z}_a\| = \lim_{t \rightarrow 0} \left\| \mathbf{Z} \circ \left( t \frac{\mathbf{e}}{\|\mathbf{e}\|} + s \frac{\mathbf{q}}{\|\mathbf{q}\|} \right) \right\| = \left\| \mathbf{Z} \circ \frac{\mathbf{q}}{\|\mathbf{q}\|} \right\| = \|\mathbf{Z}_q\|.$$

To znamená, že platí

$$\|\mathbf{Z}_q\| \leq \sup \left\{ \|\mathbf{Z}_a\| : \mathbf{a} \not\parallel \mathbf{q}, \|\mathbf{a}\| = 1 \right\} = \sup \left\{ \|\mathbf{Z}_a\| : \mathbf{a} \not\parallel \mathbf{q} \right\}.$$

Nerovnost (1.115) tedy platí i pro všechny vektory rovnoběžné s vektorem  $\mathbf{q}$ , takže platí pro všechny vektory  $\mathbf{a} \in M$ .

Jevy (1.113) a (1.114) jsou tím pádem ekvivalentní, a tedy stejně pravděpodobné. Protože prvky množiny  $P$  jsou vektory  $\mathbf{z}(x)$ , znamená to, že platí

$$\mathbb{P} \left[ \forall x \in \mathbb{R}: \|\mathbf{Z}_{\mathbf{z}(x)}\| \leq S\sqrt{2F_{2,n-2}(\alpha)} \right] = 1 - \alpha.$$

Zbývá dosadit

$$\|\mathbf{Z}_{\mathbf{z}(x)}\| = \frac{|\mathbf{z}(x) \circ (\mathbf{Y} - \boldsymbol{\mu})|}{\|\mathbf{z}(x)\|} = \frac{|(b_0 + b_1x) - (\beta_0 + \beta_1x)|}{\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}}};$$

po vynásobení nerovnosti jmenovatelem docházíme k poznatku, že pravděpodobnost jevu

$$\forall x \in R: \left| (b_0 + b_1x) - (\beta_0 + \beta_1x) \right| \leq S \sqrt{2F_{2,n-2}(\alpha) \left( \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \right)}$$

je  $1 - \alpha$ . To ovšem znamená, že s pravděpodobností  $1 - \alpha$  leží přímka  $y = \beta_0 + \beta_1 x$  v části roviny  $x, y$  vymezené hranicemi

$$y = b_0 + b_1 x \pm S \sqrt{2F_{2,n-2}(\alpha) \left( \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \right)}. \quad (1.117)$$

Tato oblast se nazývá *pás spolehlivosti pro regresní přímku* (viz obr. 1.23).

### Alternativní způsob odvození

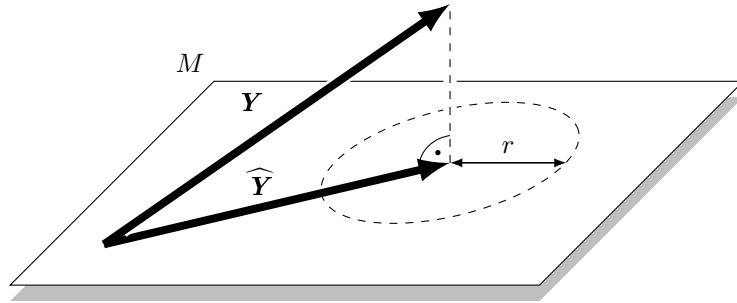
Ukažme si ještě jiný, méně tradiční způsob, jakým lze dojít k hranicím (1.117). Při odvozování testu hypotézy  $\boldsymbol{\mu} = \boldsymbol{\mu}_0$  jsme viděli, že je-li skutečná střední hodnota náhodného vektoru  $\mathbf{Y}$  rovna  $\boldsymbol{\mu}$ , platí

$$\frac{\|\widehat{\mathbf{Y}} - \boldsymbol{\mu}\|^2 / 2}{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2 / (n-2)} = \frac{\|\widehat{\mathbf{Y}} - \boldsymbol{\mu}\|^2}{2S^2} \sim F_{2,n-2}$$

(viz (1.54) na straně 46), tj.

$$\mathbb{P} \left[ \|\widehat{\mathbf{Y}} - \boldsymbol{\mu}\|^2 \leq 2S^2 F_{2,n-2}(\alpha) \right] = 1 - \alpha.$$

To znamená, že s pravděpodobností  $1 - \alpha$  leží střední hodnota  $\boldsymbol{\mu}$  v části roviny  $M$  vymezené kružnicí se středem v koncovém bodě vektoru  $\widehat{\mathbf{Y}}$  a poloměrem  $r = S\sqrt{2F_{2,n-2}(\alpha)}$  (viz obr. 1.30).<sup>22</sup> Označme tuto množinu  $K$ ; každý vektor  $\boldsymbol{\mu}$



**Obrázek 1.30:** Skutečná střední hodnota  $\boldsymbol{\mu}$  leží v případě modelu (1.13), resp. (1.56), s pravděpodobností  $1 - \alpha$  v části roviny  $M$  ohraničené kružnicí se středem v koncovém bodě vektoru  $\widehat{\mathbf{Y}}$  a poloměrem  $r = S\sqrt{2F_{2,n-2}(\alpha)}$ , kde  $S\sqrt{n-2}$  je délka vektoru  $\mathbf{Y} - \widehat{\mathbf{Y}}$ .

této množiny odpovídá právě jedné dvojici hodnot  $\beta_0, \beta_1 \in \mathbb{R}$ , určené vztahem

$$\boldsymbol{\mu} = \beta_0 \mathbf{e} + \beta_1 \mathbf{x},$$

<sup>22</sup>Jiná možná formulace je, že část roviny vymezená touto kružnicí obsahuje právě ty vektory  $\boldsymbol{\mu}_0$ , pro něž bychom nezamítli nulovou hypotézu  $\boldsymbol{\mu} = \boldsymbol{\mu}_0$  na hladině významnosti  $\alpha$ .

a každá taková dvojice odpovídá právě jedné přímce

$$y = \beta_0 + \beta_1 x$$

v rovině  $x, y$ . Všechny tyto přímky pokrývají v rovině  $x, y$  právě pás spolehlivosti pro regresní přímku, tj. množinu, ve které hledaná přímka odpovídající skutečné hodnotě  $\boldsymbol{\mu}$  leží s pravděpodobností  $1 - \alpha$ . Určit hranice tohoto pásu znamená najít pro každé pevně zvolené  $x \in \mathbb{R}$  minimum a maximum výrazu

$$f(\boldsymbol{\mu}) \equiv \beta_0 + \beta_1 x$$

přes všechna  $\boldsymbol{\mu} = \beta_0 \mathbf{e} + \beta_1 \mathbf{x}$  taková, že  $\boldsymbol{\mu} \in K$ .

Funkce  $f$  je lineární, je tedy zřejmé, že hledaného minima a maxima bude nabývat ve dvou protilehlých bodech hraniční kružnice. Hodnotu funkce  $f$  ve středu této kružnice známe:

$$f(\widehat{\mathbf{Y}}) = b_0 + b_1 x,$$

takže stačí určit derivaci ve směru gradientu funkce  $f$  a přičtením či odečtením přírůstku funkční hodnoty, odpovídajícího délce poloměru  $r$ , zjistíme hledané hodnoty. K tomu ovšem potřebujeme vyjádřit  $f$  v souřadnicích vzhledem k nějaké ortonormální bázi roviny  $M$ .

Zavedme tedy bázi

$$\left\{ \frac{\mathbf{e}}{\|\mathbf{e}\|}, \frac{\mathbf{q}}{\|\mathbf{q}\|} \right\},$$

kde  $\mathbf{q}$  je již dříve zavedený vektor definovaný vztahem (1.55). Pokud má pak vektor  $\boldsymbol{\mu} = \beta_0 \mathbf{e} + \beta_1 \mathbf{x} \in K$  vzhledem k této bázi souřadnice  $[t, u]$ , platí

$$\begin{aligned} \boldsymbol{\mu} = t \frac{\mathbf{e}}{\|\mathbf{e}\|} + u \frac{\mathbf{q}}{\|\mathbf{q}\|} &= \beta_0 \mathbf{e} + \beta_1 \mathbf{x} = \\ &= \beta_0 \mathbf{e} + \beta_1 (\bar{x} \mathbf{e} + \mathbf{q}) = \\ &= (\beta_0 + \beta_1 \bar{x}) \mathbf{e} + \beta_1 \mathbf{q}, \end{aligned}$$

tj.

$$\frac{t}{\|\mathbf{e}\|} = \beta_0 + \beta_1 \bar{x}, \quad \frac{u}{\|\mathbf{q}\|} = \beta_1,$$

z čehož dostaneme vyjádření souřadnic  $\beta_0, \beta_1$  pomocí souřadnic  $t, u$ :

$$\beta_0 = \frac{t}{\|\mathbf{e}\|} - \frac{u \bar{x}}{\|\mathbf{q}\|}, \quad \beta_1 = \frac{u}{\|\mathbf{q}\|}.$$

Dosadíme do předpisu funkce  $f$  a dostáváme

$$f(\boldsymbol{\mu}) = \frac{t}{\|\mathbf{e}\|} + \frac{u(x - \bar{x})}{\|\mathbf{q}\|}.$$

Gradient funkce  $f$ , tj. vektor  $\mathbf{g}$ , v jehož směru je růst funkce  $f$  maximální, má tedy vzhledem k zavedené ortonormální bázi souřadnice

$$\left( \frac{\partial f}{\partial t}, \frac{\partial f}{\partial u} \right)^T = \left( \frac{1}{\|\mathbf{e}\|}, \frac{x - \bar{x}}{\|\mathbf{q}\|} \right)^T.$$

Derivace funkce  $f$  ve směru  $\mathbf{s} \in M$  je

$$f'_s = \frac{\mathbf{g} \circ \mathbf{s}}{\|\mathbf{s}\|},$$

takže derivace ve směru gradientu je

$$f'_g = \frac{\mathbf{g} \circ \mathbf{g}}{\|\mathbf{g}\|} = \|\mathbf{g}\| = \sqrt{\frac{1}{\|\mathbf{e}\|^2} + \frac{(x - \bar{x})^2}{\|\mathbf{q}\|^2}} = \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}}.$$

Maximální, resp. minimální hodnota funkce  $f$  je proto

$$f(\widehat{\mathbf{Y}}) \pm f'_g \cdot r = b_0 + b_1 x \pm S \sqrt{2F_{2,n-2}(\alpha) \left( \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \right)};$$

došli jsme ke stejnému výsledku, jako je (1.117).

## 2. Teoretická část

### 2.1 Geometrické důsledky zavedení skalárního součinu

Nechť je na reálném vektorovém prostoru  $V_n$  nad tělesem  $\mathbb{R}$  definován *skalární součin*  $\circ$ .<sup>1</sup> Tím je na něm určena též *norma* (délka) vektoru  $\mathbf{y} \in V_n$ :

$$\|\mathbf{y}\| \equiv \sqrt{\mathbf{y} \circ \mathbf{y}}$$

a *úhel* sevřený dvěma nenulovými vektory  $\mathbf{x}, \mathbf{y} \in V_n$ :

$$\beta \equiv \arccos \frac{\mathbf{x} \circ \mathbf{y}}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|}.$$

Platí-li pro dva vektory  $\mathbf{x}, \mathbf{y} \in V_n$  rovnost  $\mathbf{x} \circ \mathbf{y} = 0$ , říkáme, že jsou *kolmé* a píšeme  $\mathbf{x} \perp \mathbf{y}$ ; to zřejmě nastane právě tehdy, svírají-li úhel  $\pi/2$  nebo je-li aspoň jeden z nich nulový. Z vlastností skalárního součinu plyne, že je-li vektor  $\mathbf{y}$  kolmý na vektory  $\mathbf{a}, \mathbf{b}$ , je kolmý i na jakoukoli jejich lineární kombinaci. Pro skupinu navzájem kolmých vektorů  $\mathbf{x}_1, \dots, \mathbf{x}_k$  platí

$$\begin{aligned} \left\| \sum_{i=1}^k \mathbf{x}_i \right\|^2 &= \sum_{i=1}^k \|\mathbf{x}_i\|^2 + 2 \sum_{i=1}^k \sum_{j=i+1}^k \mathbf{x}_i \circ \mathbf{x}_j = \\ &= \sum_{i=1}^k \|\mathbf{x}_i\|^2, \end{aligned}$$

tj. mnohorozměrná Pythagorova věta.

Nechť  $M$  je podprostor vektorového prostoru  $V_n$  (podobně jako všechny dále uvedené podprostory). Jsou-li všechny vektory nějaké báze tohoto podprostoru navzájem kolmé, nazývá se tato báze *ortogonální*; mají-li navíc také jednotkovou délku, nazývá se *ortonormální*. Je-li dimenze podprostoru  $M$  konečná, lze v něm jakoukoli skupinu navzájem kolmých jednotkových vektorů doplnit na ortonormální bázi tohoto podprostoru. To znamená, že v jakémkoli podprostoru konečné dimenze (včetně prostoru  $V_n$ , je-li  $n < \infty$ ) lze vytvořit ortonormální bázi.

Zdůrazněme, že všechny výše uvedené pojmy jsou závislé na definici skalárního součinu.

### 2.2 Kolmost podprostorů

Nejprve uveďme několik elementárnějších definic a tvrzení.

- Řekneme, že vektor  $\mathbf{y}$  je kolmý na podprostor  $A$  ( $\mathbf{y} \perp A$ ), jestliže platí  $\mathbf{x} \perp \mathbf{y}$  pro všechny vektory  $\mathbf{x} \in A$ . K prokázání kolmosti vektoru  $\mathbf{y}$  na podprostor  $A$  stačí vzhledem k linearitě skalárního součinu doložit jeho kolmost na libovolnou skupinu generátorů  $A$ .

---

<sup>1</sup>Vektorový prostor  $V_n$ , na kterém je definován skalární součin, se nazývá *unitární prostor* (viz [4]).

- Je-li  $A$  libovolný podprostor, platí

$$[\mathbf{x} \perp A \wedge \mathbf{x} \in A] \implies \mathbf{x} = \mathbf{0}. \quad (2.1)$$

Z předpokladu totiž plyne  $\mathbf{x} \perp \mathbf{x}$ , tj.  $\mathbf{x} \circ \mathbf{x} = 0$ , z čehož podle definice skalárního součinu vyplývá  $\mathbf{x} = \mathbf{0}$ .

- *Součtem* vektorových podprostorů  $A, B$  míníme množinu

$$A + B \equiv \{\mathbf{a} + \mathbf{b}; \mathbf{a} \in A, \mathbf{b} \in B\}.$$

- Je-li podprostor  $B$  konečné dimenze  $b$  podmnožinou podprostoru  $A$ , tvoří množina všech vektorů ležících v podprostoru  $A$  a kolmých na podprostor  $B$  vektorový podprostor, který budeme značit  $A - B$  a nazveme jej (*relativním*) *ortogonálním doplňkem* podprostoru  $B$  v podprostoru  $A$ , tj.

$$A - B \equiv \{\mathbf{x} : \mathbf{x} \in A \wedge \mathbf{x} \perp B\}.$$

Platí  $B \cap (A - B) = \{\mathbf{0}\}$ .

- Platí-li navíc  $\dim A = a < \infty$ , můžeme zvolit ortonormální bázi  $\{\mathbf{e}_1, \dots, \mathbf{e}_b\}$  podprostoru  $B$  a doplnit ji na ortonormální bázi  $\{\mathbf{e}_1, \dots, \mathbf{e}_a\}$  podprostoru  $A$ . Pak lze ukázat, že platí

$$A - B = [\mathbf{e}_{b+1}, \dots, \mathbf{e}_a],$$

takže je zřejmé  $\dim(A - B) = a - b$  a lze odvodit vztahy

$$A - (A - B) = B, \quad (2.2)$$

$$B + (A - B) = A. \quad (2.3)$$

- Ne každé tvrzení, které vyhlíží podobně triviálně, je však pravdivé: například rovnost

$$(A + B) - B = A$$

obecně neplatí.

- Jsou-li podprostory  $A, B$  podmnožinami podprostoru  $C$ , platí

$$C - (A + B) = (C - A) \cap (C - B). \quad (2.4)$$

Leží-li totiž vektor  $\mathbf{x}$  v podprostoru  $C - (A + B)$ , znamená to, že je prvkem podprostoru  $C$  a zároveň je kolmý na podprostor  $A + B$ . Je tedy kolmý i na podprostory  $A$  a  $B$ , takže je prvkem jak podprostoru  $C - A$ , tak podprostoru  $C - B$ .

Je-li naopak vektor  $\mathbf{x}$  prvkem podprostoru  $(C - A) \cap (C - B)$ , leží v podprostoru  $C$  a přitom je kolmý jak na podprostor  $A$ , tak na podprostor  $B$ . Je tedy kolmý na všechny vektory  $\mathbf{a} \in A$ ,  $\mathbf{b} \in B$ , takže je kolmý i na jejich lineární kombinace, což ovšem znamená, že platí  $\mathbf{x} \perp (A + B)$ , a tedy  $\mathbf{x} \in C - (A + B)$ .

- Pokud je  $\dim C < \infty$ , můžeme ve vztahu (2.4) díky rovnosti (2.2) zaměnit  $A$  za  $C - A$  a  $B$  za  $C - B$ ; získáme tak tvrzení

$$C - [(C - A) + (C - B)] = A \cap B,$$

které lze opět díky vzorci (2.2) přepsat na tvar

$$C - (A \cap B) = (C - A) + (C - B). \quad (2.5)$$

- Místo  $V_n - A$  píšeme prostě  $A^\perp$ ; z rovností (2.2), (2.3), (2.4) a (2.5) tak dosazením za „menšenec“ dostáváme vztahy

$$(A^\perp)^\perp = A, \quad (2.6)$$

$$A + A^\perp = V_n, \quad (2.7)$$

$$(A + B)^\perp = A^\perp \cap B^\perp, \quad (2.8)$$

$$(A \cap B)^\perp = A^\perp + B^\perp. \quad (2.9)$$

Poznamenejme, že pouze vztah (2.4), resp. (2.8), jsme dokázali odvodit bez předpokladu konečné dimenze prostoru  $C$ , resp.  $V_n$ . Pokud se tedy v dalším textu odvoláváme na některé ze zbývajících vztahů, je tento předpoklad nezbytný. Protože však v našem pojednání hrají roli pouze konečné vektorové prostory, nebudeme tuto skutečnost již připomínat; nadále tedy budeme automaticky předpokládat, že dimenze prostoru  $V_n$  je konečná.

## 2.2.1 Základní definice kolmosti podprostorů

Nejběžnější a nejjednodušší definice kolmosti dvou podprostorů, ze které budeme vycházet, je tato: podprostory  $A, B$  nazýváme *kolmé* a píšeme  $A \perp B$ , jsou-li všechny vektory z jednoho podprostoru kolmé na všechny vektory z druhého podprostoru, tj.

$$A \perp B \iff \forall \mathbf{a} \in A, \mathbf{b} \in B: \mathbf{a} \perp \mathbf{b}. \quad (2.10)$$

Skutečnost, že podprostory  $A_1, \dots, A_k$  jsou navzájem po dvojicích kolmé, budeme značit zápisem

$$\{A_1, \dots, A_k\} \in \mathcal{P}^\perp.$$

Pokud pro navzájem kolmé podprostory  $L_1, \dots, L_k$  platí rovnost  $A = L_1 + \dots + L_k$ , říkáme, že tvoří *ortogonální rozklad* podprostoru  $A$ . Tuto skutečnost budeme vyjadřovat zápisem

$$A = L_1 \oplus \dots \oplus L_k,$$

resp.

$$A = \bigoplus_{i=1}^k L_i.$$

Následuje výběr jednoduchých užitečných tvrzení, které se týkají relace kolmosti:

$$\begin{aligned}
 A \perp B &\implies A \cap B = \{\mathbf{0}\}, \\
 A \perp B &\iff B \subseteq A^\perp. \\
 A \perp B &\iff (A + B) - B = A, \\
 [A \perp B \wedge C \subseteq A] &\implies C \perp B, \\
 [A \perp B \wedge A \perp C] &\iff A \perp (B + C).
 \end{aligned} \tag{2.11}$$

Jsou-li podprostory  $A, B$  kolmé, je rozklad libovolného vektoru  $\mathbf{x} \in A + B$  na část  $\mathbf{a} \in A$  a část  $\mathbf{b} \in B$  jednoznačně určen. Je-li totiž

$$\mathbf{x} = \mathbf{a}_1 + \mathbf{b}_1 = \mathbf{a}_2 + \mathbf{b}_2,$$

musí platit

$$(\mathbf{a}_1 + \mathbf{b}_1) - (\mathbf{a}_2 + \mathbf{b}_2) = (\mathbf{a}_1 - \mathbf{a}_2) + (\mathbf{b}_1 - \mathbf{b}_2) = \mathbf{0}.$$

První sčítanec leží v podprostoru  $A$ , musí v něm tedy ležet i ten druhý. Ten je však zároveň prvkem podprostoru  $B$ . Průnikem podprostorů  $A, B$  je ovšem množina  $\{\mathbf{0}\}$ , takže je

$$\mathbf{a}_1 - \mathbf{a}_2 = \mathbf{b}_1 - \mathbf{b}_2 = \mathbf{0}.$$

Výše uvedený poznatek lze snadno zobecnit na libovolný počet podprostorů, tj.

$$\begin{aligned}
 A &= \bigoplus_{i=1}^k A_i \\
 &\Downarrow \\
 \forall \mathbf{x} \in A &\quad \exists! \mathbf{x}_1 \in A_1, \dots, \mathbf{x}_k \in A_k : \mathbf{x} = \sum_{i=1}^k \mathbf{x}_i.
 \end{aligned}$$

Závěrem dokažme jedno speciální tvrzení, které využijeme později: platí implikace

$$\{A, B, C\} \in \mathcal{P}^\perp \implies (A + B) \cap (A + C) = A. \tag{2.12}$$

Inkluze  $A \subseteq (A + B) \cap (A + C)$  je jistě zřejmá, potřebujeme tedy dokázat, že je-li splněn předpoklad, platí

$$(A + B) \cap (A + C) \subseteq A. \tag{2.13}$$

Nechť tedy vektor  $\mathbf{x}$  leží v podprostoru  $(A + B) \cap (A + C)$ . Pak jej lze psát ve dvou tvarech

$$\begin{aligned}
 \mathbf{x} &= \mathbf{a}_1 + \mathbf{b} = \\
 &= \mathbf{a}_2 + \mathbf{c},
 \end{aligned}$$

kde  $\mathbf{a}_i \in A, \mathbf{b} \in B, \mathbf{c} \in C$ . Z toho plyne

$$0 = (\mathbf{a}_1 + \mathbf{b}) - (\mathbf{a}_2 + \mathbf{c}) = (\mathbf{a}_1 - \mathbf{a}_2) + (\mathbf{b} - \mathbf{c}).$$

První z těchto sčítanců leží v podprostoru  $A$ , musí v něm tedy ležet i ten druhý. Ten je však na tento podprostor zároveň kolmý, takže musí platit  $\mathbf{b} = \mathbf{c}$  a tudíž také  $\mathbf{b} = \mathbf{c} = \mathbf{0}$  (neboť vektor  $\mathbf{0}$  je jediný, který mají podprostory  $B$  a  $C$  společný). Je tedy  $\mathbf{x} = \mathbf{a}_1 = \mathbf{a}_2 \in A$ , takže platí (2.13).

## 2.2.2 Zobecnění pojmu kolmosti, knižní kolmost

Nedostatkem definice (2.10) je, že nezahrnuje kolmost dvou rovin v prostoru  $V_3$  tak, jak je běžně chápána. Z toho důvodu je navržena v publikaci [21] její obecnější varianta:

$$A \perp\!\!\!\perp B \iff [A^\perp \subseteq B \vee B \subseteq A^\perp]. \quad (2.14)$$

Tamtéž je ukázáno, že se jedná o symetrickou definici, ač to není na první pohled patrné. Jiná alternativa je navržena v knize [31]:

$$A \sqcup B \iff A - P \perp B - P, \quad (2.15)$$

kde  $P \equiv A \cap B$  (ponechme toto značení i nadále). Podprostory  $A, B$  se v tomto případě nazývají *knižně kolmé*.<sup>2</sup> Skutečnost, že podprostory  $A_1, \dots, A_k$  jsou navzájem knižně kolmé, budeme vyjadřovat zápisem

$$\{A_1, \dots, A_k\} \in \mathcal{P}^\sqcup.$$

Obě definice (2.14) a (2.15) jsou zřejmě zobecněním definice (2.10), neboť ze vztahu  $A \perp B$  triviálně plynou vztahy  $A \perp\!\!\!\perp B$ , resp.  $A \sqcup B$ . Můžeme však ukázat, že definice (2.15) je obecnější než definice (2.14). Nechť platí  $A \perp\!\!\!\perp B$ ; to znamená, že je buďto  $B \subseteq A^\perp$ , nebo  $A^\perp \subseteq B$ . V prvním případě platí  $A \perp B$ , a tím pádem i  $A \sqcup B$ . V případě druhém je množina  $B - A^\perp$  tvořena všemi prvky podprostoru  $B$ , které jsou kolmé na podprostor  $A^\perp$ , tj. leží v podprostoru  $A$ . Platí tedy

$$B - A^\perp = P,$$

z čehož podle vztahu (2.2) plyne  $B - P = A^\perp$ . To znamená, že podprostory  $B - P$  a  $A$  jsou navzájem kolmé, tím spíše jsou tedy kolmé i podprostory  $B - P$  a  $A - P$ .

Ze vztahu  $A \perp\!\!\!\perp B$  tedy plyne vztah  $A \sqcup B$ . Není tomu však naopak, jak ukazuje následující protipříklad: nechť  $\{e_1, e_2, e_3, e_4\}$  je ortonormální báze prostoru  $V_4$ . Položme

$$\begin{aligned} A &\equiv [e_1, e_2], \\ B &\equiv [e_1, e_3], \end{aligned}$$

takže je

$$\begin{aligned} A^\perp &= [e_3, e_4], \\ P &= [e_1], \\ A - P &= [e_2], \\ B - P &= [e_3]. \end{aligned}$$

Je vidět, že není splněno ani  $A^\perp \subseteq B$ , ani  $B \subseteq A^\perp$ , takže neplatí  $A \perp\!\!\!\perp B$ . Přitom však je  $A - P \perp B - P$ , takže platí  $A \sqcup B$ .

<sup>2</sup>V originále je použit termín „book orthogonal“; lepší překlad nás bohužel nenapadá. Pojem je zde značen symbolem  $\perp_B$ , který však má – kromě toho, že je „nepěkný“ – tu nevýhodu, že není symetrický. Protože se touto relací hodláme zabývat podrobněji, dovolili jsme si pro ni zavést vlastní označení, stejně jako v případě definice (2.14), kde jsme pro změnu museli použít značení odlišné od definice (2.10).

### 2.2.3 Knižní kolmost dvou podprostorů

Uveďme nyní některé užitečné vztahy týkající se relace  $\sqcup$  mezi dvěma podprostory  $A, B$ .

- Platí

$$B \subseteq A \implies A \sqcup B, \quad (2.16)$$

neboť je-li  $B \subseteq A$ , je  $P = B$ , a tedy  $B - P = \{\mathbf{0}\} \perp A - P$ .

- Definici (2.15) lze psát v ekvivalentním tvaru

$$A \sqcup B \iff A - P \perp B. \quad (2.17)$$

Je-li totiž  $A - P \perp B$ , platí tím spíše  $A - P \perp B - P$ , a tedy  $A \sqcup B$ . Co se týče opačné implikace, podle (2.3) platí

$$B = (B - P) + P,$$

takže je-li podprostor  $A - P$  kolmý na podprostor  $B - P$ , a je kolmý i na celý podprostor  $B$  (neboť na podprostor  $P$  je kolmý z definice).

- Platí-li  $A \sqcup B$ , tvoří podprostory  $A - P, B - P$  a  $P$  ortogonální rozklad podprostoru  $A + B$ , tj.

$$A \sqcup B \implies A + B = (A - P) \oplus (B - P) \oplus P. \quad (2.18)$$

Z definice ortogonálního doplňku totiž plyne  $A - P \perp P$  a  $B - P \perp P$ , z předpokladu  $A \sqcup B$  vyplývá  $A - P \perp B - P$  a navíc zřejmě platí

$$\begin{aligned} A + B &= [P + (A - P)] + [P + (B - P)] = \\ &= P + (A - P) + P + (B - P) = \\ &= P + (A - P) + (B - P). \end{aligned}$$

- Platí

$$A \sqcup B \implies A \sqcup B^\perp. \quad (2.19)$$

To lze snadno dokázat pomocí vhodně zvolené ortogonální báze prostoru  $V_n$  takové, že její části generují postupně podprostory  $A - P, B - P, P$  a  $(A + B)^\perp$ .

- Opakovaným použitím předchozího tvrzení a tvrzení (2.6) dostáváme vztahy

$$\begin{aligned} A \sqcup B &\iff A \sqcup B^\perp, \\ A \sqcup B &\iff A^\perp \sqcup B^\perp. \end{aligned}$$

## 2.2.4 Knižní kolmost tří podprostorů

- Dokažme implikaci

$$[(A \sqcup C) \wedge (B \sqcup C)] \implies (A \cap B) \sqcup C. \quad (2.20)$$

Využijeme alternativní definici (2.17): předpoklad implikace je ekvivalentní s tvrzením

$$A \perp C - (A \cap C) \quad \wedge \quad B \perp C - (B \cap C).$$

Z něj vyplývají vztahy

$$\begin{aligned} A \cap B \perp C - (A \cap C), \\ A \cap B \perp C - (B \cap C), \end{aligned}$$

ze kterých plyne, že platí i

$$A \cap B \perp [C - (A \cap C)] + [C - (B \cap C)].$$

Podle vztahu (2.4) je však

$$[C - (A \cap C)] + [C - (B \cap C)] = C - (A \cap B \cap C).$$

Ukázali jsme tedy, že platí

$$A \cap B \perp C - (A \cap B \cap C),$$

což je podle (2.17) ekvivalentní s tvrzením  $(A \cap B) \sqcup C$ .

- Podobně platí implikace

$$(A \sqcup C \wedge B \sqcup C) \implies (A + B) \sqcup C; \quad (2.21)$$

z předpokladu totiž díky vztahu (2.19) vyplývá

$$(A^\perp \sqcup C) \wedge (B^\perp \sqcup C),$$

z čehož dále vzhledem k tvrzení (2.20) plyne

$$(A^\perp \cap B^\perp) \sqcup C.$$

Opět použijeme tvrzení (2.19) a dostaneme vztah

$$(A^\perp \cap B^\perp)^\perp \sqcup C,$$

ze kterého díky rovnostem (2.9) a (2.6) plyne závěr.

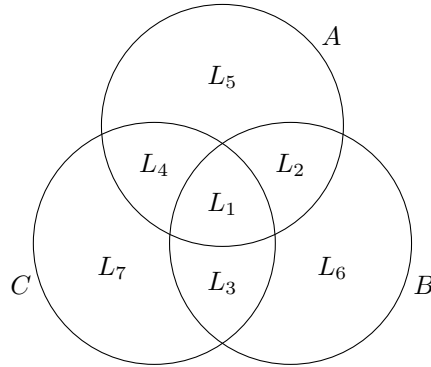
- O něco pracnější bude dokázat následující tvrzení: jsou-li podprostory  $A$ ,  $B$  a  $C$  navzájem knižně kolmé, existují navzájem kolmé podprostory  $L_1, \dots, L_m$  takové, že každý z podprostorů  $A$ ,  $B$  a  $C$  je součtem některých z nich.

Nechť tedy platí  $\{A, B, C\} \in \mathcal{P}^\perp$ . Položíme

$$\begin{aligned}
 L_1 &\equiv A \cap B \cap C, \\
 L_2 &\equiv A \cap B - L_1, \\
 L_3 &\equiv B \cap C - L_1, \\
 L_4 &\equiv A \cap C - L_1, \\
 L_5 &\equiv A - (L_1 + L_2 + L_4), \\
 L_6 &\equiv B - (L_1 + L_2 + L_3), \\
 L_7 &\equiv C - (L_1 + L_3 + L_4),
 \end{aligned} \tag{2.22}$$

takže platí (viz obr. 2.1)

$$\begin{aligned}
 A \cap B &= L_1 + L_2, \\
 B \cap C &= L_1 + L_3, \\
 A \cap C &= L_1 + L_4.
 \end{aligned}$$



**Obrázek 2.1:** Schematické znázornění podprostorů  $L_1, \dots, L_7$  (viz (2.22)).

Podle definice ortogonálního doplňku je zřejmě splněno

$$\begin{aligned}
 L_1 &\perp L_2, L_3, L_4, \\
 L_5 &\perp L_1, L_2, L_4, \\
 L_6 &\perp L_1, L_2, L_3, \\
 L_7 &\perp L_1, L_3, L_4.
 \end{aligned}$$

Z toho je mimo jiné patrné, že podprostor  $L_5 \subseteq A$  je kolmý na podprostor  $A \cap B = L_1 + L_2$ . Je tedy podmnožinou podprostoru  $A - (A \cap B)$ , který je podle definice (2.17) díky předpokladu  $A \perp B$  kolmý na podprostor  $B$ . Platí tedy  $L_5 \perp B$ , a tudíž i

$$\begin{aligned}
 L_5 &\perp L_3, \\
 L_5 &\perp L_6
 \end{aligned}$$

(neboť je  $L_3 \subseteq B$ ,  $L_6 \subseteq B$ ). Analogicky lze dokázat i kolmosti

$$\begin{aligned}
 L_6 &\perp L_4, \\
 L_6 &\perp L_7, \\
 L_7 &\perp L_2, \\
 L_7 &\perp L_5.
 \end{aligned}$$

Dále z předpokladu  $A \sqcup C$ ,  $B \sqcup C$  dostáváme díky tvrzení (2.20) vztah  $(A \cap B) \sqcup C$ ; z tvrzení (2.16) pak plyne  $(A \cap B) \sqcup B$ . Dalším užitím implikace (2.20) z těchto závěrů dostáváme

$$(A \cap B) \sqcup (B \cap C);$$

to ovšem podle definice (2.15) znamená, že platí

$$(A \cap B) - (A \cap B \cap C) \perp (B \cap C) - (A \cap B \cap C),$$

tj.

$$L_2 \perp L_3.$$

Podobně můžeme dokázat vztahy

$$L_3 \perp L_4,$$

$$L_2 \perp L_4.$$

Všechny podprostory  $L_1, \dots, L_7$  jsou tedy navzájem kolmé. Jak z nich lze „složit“ původní podprostory  $A, B, C$ , lze snadno nahlédnout ze způsobu, jakým byly zavedeny; dokázali jsme tedy, že platí

$$\{A, B, C\} \in \mathcal{P}^\sqcup$$



$$\exists \{L_1, \dots, L_7\} \in \mathcal{P}^\perp : \begin{cases} A = L_1 + L_2 + L_4 + L_5, \\ B = L_1 + L_2 + L_3 + L_6, \\ C = L_1 + L_3 + L_4 + L_7. \end{cases}$$

(2.23)

- S pomocí výše uvedeného rozkladu můžeme dokázat další vztah, který nám později přijde vhod: platí

$$\{A, B, C\} \in \mathcal{P}^\sqcup \implies [A - (A \cap C)] \sqcup [B - (B \cap C)]. \quad (2.24)$$

Nejdříve dosadíme:

$$\begin{aligned} A - (A \cap C) &= (L_1 + L_2 + L_4 + L_5) - (L_1 + L_4) = \\ &= L_2 + L_5, \\ B - (B \cap C) &= (L_1 + L_2 + L_3 + L_6) - (L_1 + L_3) = \\ &= L_2 + L_6; \end{aligned}$$

druhé rovnosti plynou z asociativity sčítání podprostorů a ze vztahu (2.11). Podle tvrzení (2.12) dále platí

$$[A - (A \cap C)] \cap [B - (B \cap C)] = L_2.$$

Jelikož podprostory

$$\begin{aligned} [A - (A \cap C)] - L_2 &= L_5, \\ [B - (B \cap C)] - L_2 &= L_6 \end{aligned}$$

jsou navzájem kolmé, je tvrzení (2.24) dokázáno.

## 2.2.5 Knižní kolmost více podprostorů

- Z tvrzení (2.20) a (2.21) lze snadno odvodit obecnější vztahy: jsou-li všechny podprostory  $A_1, \dots, A_k$  navzájem knižně kolmé, jsou knižně kolmé i jakékoli jejich vzájemné průniky či součty, tj.

$$\begin{aligned} & \{A_1, \dots, A_k\} \in \mathcal{P}^\sqcup \\ & \Downarrow \\ \forall I, J \subseteq \{1, \dots, k\} : & \begin{cases} \bigcap_{i \in I} A_i \sqcup \bigcap_{j \in J} A_j, \\ \sum_{i \in I} A_i \sqcup \sum_{j \in J} A_j, \\ \bigcap_{i \in I} A_i \sqcup \sum_{j \in J} A_j. \end{cases} \end{aligned} \quad (2.25)$$

- O něco náročnější je dokázat obecnější verzi tvrzení (2.23): necht' všechny podprostory  $A_1, \dots, A_k$  jsou navzájem knižně kolmé. Pak existují navzájem kolmé podprostory  $L_1, \dots, L_m$  takové, že každý z podprostorů  $A_j$  je součtem některých z nich, tj. formálně

$$\begin{aligned} & \{A_1, \dots, A_k\} \in \mathcal{P}^\sqcup \\ & \Downarrow \\ \exists \{L_1, \dots, L_m\} \in \mathcal{P}^\perp \quad \forall i \in \{1, \dots, k\} \quad \exists M_i \subseteq \{1, \dots, m\} : & \\ & A_i = \sum_{j \in M_i} L_j. \end{aligned} \quad (2.26)$$

Důkaz provedeme matematickou indukcí. Pro  $k = 2$  je tvrzení zřejmé, pro  $k = 3$  jej máme již dokázané. Předpokládejme, že tvrzení platí pro nějaké dané  $k \in \mathbb{N}$  ( $k \geq 3$ ) a ukažme, že platí i pro  $k + 1$ .

Necht' je tedy  $\{A_1, \dots, A_{k+1}\} \in \mathcal{P}^\sqcup$ . Pro  $j = 1, \dots, k$  položme

$$\begin{aligned} P_j & \equiv A_j \cap A_{k+1}, \\ Q_j & \equiv A_j - P_j, \\ D & \equiv A_{k+1} - \sum_{j=1}^k P_j. \end{aligned}$$

Pro všechna  $j \in \{1, \dots, k\}$  zřejmě platí

$$D \perp P_j.$$

Z předpokladu  $A_j \sqcup A_{k+1}$  podle (2.17) dále plyne  $Q_j \perp A_{k+1}$ , což ovšem znamená, že platí také

$$\begin{aligned} Q_j & \perp D, \\ Q_j & \perp P_i \end{aligned}$$

pro všechna  $i, j \in \{1, \dots, k\}$  (neboť  $D, P_i \subseteq A_{k+1}$ ).

Podle (2.25) jsou všechny podprostory  $P_j$  navzájem knižně kolmé a jejich počet je  $k$ , podle indukčního předpokladu je lze tedy rozložit na navzájem kolmé podprostory  $L_1, \dots, L_m$ .

Dále pro všechna  $i, j \in \{1, \dots, k\}$  platí

$$\begin{aligned} Q_i &= A_i - (A_i \cap A_{k+1}), \\ Q_j &= A_j - (A_j \cap A_{k+1}), \end{aligned}$$

z čehož podle (2.24) plyne  $Q_i \perp Q_j$ . Těchto podprostorů je rovněž  $k$ , takže podle indukčního předpokladu i pro ně existují podprostory  $L_{m+1}, \dots, L_{2m}$  patřičných vlastností; kromě toho jsou zřejmě kolmé na všechny podprostory  $L_1, \dots, L_m$ . Definujeme-li konečně  $L_{2m+1} \equiv D$ , je patrné, že podprostory  $L_1, \dots, L_{2m+1}$  jsou všechny navzájem kolmé a přitom platí

$$\begin{aligned} A_i &= P_i + Q_i = \sum_{j \in M_i} L_j + \sum_{j \in N_i} L_j, \\ A_{k+1} &= D + \sum_{j=1}^k P_j = L_{2m+1} + \sum_{j \in M} L_j, \end{aligned}$$

kde  $i \in \{1, \dots, k\}$ ,  $M_i, M \subseteq \{1, \dots, m\}$ ,  $N_i \subseteq \{m+1, \dots, 2m\}$ .

## 2.3 Pravoúhlý průmět

Je-li  $M$   $m$ -rozměrný podprostor prostoru  $V_n$  a  $\mathbf{y} \in V_n$ , nazýváme *pravoúhlým průmětem vektoru  $\mathbf{y}$  do podprostoru  $M$*  takový vektor  $\hat{\mathbf{y}}$ , který splňuje podmínky

$$\hat{\mathbf{y}} \in M, \tag{2.27}$$

$$\mathbf{y} - \hat{\mathbf{y}} \perp M. \tag{2.28}$$

Tento pojem je tedy závislý na definici skalárního součinu.

### 2.3.1 Existence a jednoznačnost

Pravoúhlý průmět vektoru  $\mathbf{y}$  do podprostoru  $M$  vždy existuje a je určen jednoznačně. Existenci můžeme doložit volbou ortonormální báze  $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}$  prostoru  $V_n$  takové, že vektory  $\mathbf{e}_1, \dots, \mathbf{e}_m$  leží v podprostoru  $M$ . Vyjádříme-li pak vektor  $\mathbf{y}$  pomocí této báze:

$$\mathbf{y} = y_1 \mathbf{e}_1 + \dots + y_m \mathbf{e}_m + y_{m+1} \mathbf{e}_{m+1} + \dots + y_n \mathbf{e}_n,$$

je zřejmé, že je

$$\hat{\mathbf{y}} = y_1 \mathbf{e}_1 + \dots + y_m \mathbf{e}_m.$$

Jiným často používaným způsobem, jak dokázat existenci pravoúhlého průmětu, je zvolit ortonormální bázi  $\{\mathbf{e}_1, \dots, \mathbf{e}_m\}$  podprostoru  $M$  a ukázat, že platí

$$\hat{\mathbf{y}} = (\mathbf{y} \circ \mathbf{e}_1) \mathbf{e}_1 + \dots + (\mathbf{y} \circ \mathbf{e}_m) \mathbf{e}_m. \tag{2.29}$$

Podmínku (2.27) totiž tento vektor zřejmě splňuje a dále pro všechna  $i = 1, \dots, m$  platí

$$\begin{aligned}(\mathbf{y} - \widehat{\mathbf{y}}) \circ \mathbf{e}_i &= \mathbf{y} \circ \mathbf{e}_i - (\mathbf{y} \circ \mathbf{e}_1) \mathbf{e}_1 \circ \mathbf{e}_i - \dots - (\mathbf{y} \circ \mathbf{e}_m) \mathbf{e}_m \circ \mathbf{e}_i = \\ &= \mathbf{y} \circ \mathbf{e}_i - \mathbf{y} \circ \mathbf{e}_i = \\ &= 0.\end{aligned}$$

Vektor  $\mathbf{y} - \widehat{\mathbf{y}}$  je tedy kolmý na všechny generátory podprostoru  $M$ , tím pádem je kolmý i na celý podprostor  $M$  a je splněna podmínka (2.28). Výhodou tohoto důkazu je jednak skutečnost, že je v něm zahrnut návod, jak – v případě, že disponujeme ortonormální bází podprostoru  $M$  – vektor  $\widehat{\mathbf{y}}$  nalézt, a jednak to, že nevyžaduje předpoklad konečnosti dimenze prostoru  $V_n$ ; stačí, má-li konečnou dimenzi podprostor  $M$ .

Dokažme dále jednoznačnost průmětu  $\widehat{\mathbf{y}}$ : necht' vedle vektoru  $\widehat{\mathbf{y}}$  existuje ještě jiný vektor  $\widehat{\mathbf{y}}^*$  splňující podmínky (2.27), (2.28). Pro všechny vektory  $\mathbf{x} \in M$  pak platí

$$\begin{aligned}\mathbf{x} \circ (\widehat{\mathbf{y}} - \widehat{\mathbf{y}}^*) &= \mathbf{x} \circ [(\mathbf{y} - \widehat{\mathbf{y}}^*) - (\mathbf{y} - \widehat{\mathbf{y}})] = \\ &= \mathbf{x} \circ (\mathbf{y} - \widehat{\mathbf{y}}^*) - \mathbf{x} \circ (\mathbf{y} - \widehat{\mathbf{y}}) = \\ &= 0.\end{aligned}$$

Vektor  $\widehat{\mathbf{y}} - \widehat{\mathbf{y}}^*$  je tedy kolmý na všechny vektory podprostoru  $M$ , avšak přitom v tomto podprostoru leží, musí proto platit  $\widehat{\mathbf{y}} - \widehat{\mathbf{y}}^* = \mathbf{0}$ .

### 2.3.2 Nejbližší prvek

Důležitou vlastností pravoúhlého průmětu  $\widehat{\mathbf{y}}$  je to, že jako jediný má ze všech prvků podprostoru  $M$  nejmenší vzdálenost<sup>3</sup> od vektoru  $\mathbf{y}$ , tj. platí implikace

$$[\mathbf{x} \in M \wedge \mathbf{x} \neq \widehat{\mathbf{y}}] \implies \|\mathbf{y} - \mathbf{x}\| > \|\mathbf{y} - \widehat{\mathbf{y}}\|.$$

Je tomu tak proto, že pro  $\mathbf{x} \in M$  leží vektor  $\widehat{\mathbf{y}} - \mathbf{x}$  v podprostoru  $M$ , takže vektor  $\mathbf{y} - \widehat{\mathbf{y}}$  je na něj podle podmínky (2.28)) kolmý. Tvoří s ním tedy odvěsny pravoúhlého trojúhelníku, jehož přeponou je vektor  $\mathbf{y} - \mathbf{x}$ ; nerovnost pak plyne z Pythagorovy věty.

### 2.3.3 Metody výpočtu

Předpokládejme, že podprostor  $M$  je určen množinou svých generátorů  $\mathbf{x}_1, \dots, \mathbf{x}_k$ . Jsou nám známy tři způsoby, jak lze určit vektor  $\widehat{\mathbf{y}}$ .

1. Vyjdeme z podmínek (2.27), (2.28). Z první z nich plyne, že vektor  $\widehat{\mathbf{y}}$  musí být lineární kombinací vektorů  $\mathbf{x}_1, \dots, \mathbf{x}_k$ :

$$\widehat{\mathbf{y}} = b_1 \mathbf{x}_1 + \dots + b_k \mathbf{x}_k \quad (b_i \in \mathbb{R}), \quad (2.30)$$

podle druhé podmínky musí být vektor  $\mathbf{y} - \widehat{\mathbf{y}}$  na všechny tyto vektory kolmý. Z toho plynou pro  $1 \leq i \leq k$  takzvané *normální rovnice*

$$(\mathbf{y} - b_1 \mathbf{x}_1 - \dots - b_k \mathbf{x}_k) \circ \mathbf{x}_i = 0. \quad (2.31)$$

Řešením vzniklé soustavy nalezneme koeficienty  $b_i$  a potažmo vektor  $\widehat{\mathbf{y}}$ .

---

<sup>3</sup>Připomeňme, že i vzdálenost je závislá na definici skalárního součinu.

2. Využijeme vlastnosti (2.30): do výrazu

$$\|\mathbf{y} - \hat{\mathbf{y}}\|^2$$

dosadíme za  $\hat{\mathbf{y}}$  lineární kombinaci (2.30) a zjistíme, pro jaké hodnoty  $b_i$  nabývá tato funkce svého minima. Musíme tedy vyřešit soustavu

$$\frac{\partial \|\mathbf{y} - b_1 \mathbf{x}_1 - \dots - b_k \mathbf{x}_k\|^2}{\partial b_i} = 0 \quad (1 \leq i \leq k).$$

Tato soustava je ekvivalentní se soustavou získanou předchozím způsobem.

3. Je-li  $\mathbf{X}$  matice, jejíž sloupce představují souřadnice vektorů  $\mathbf{x}_i$  vzhledem k aktuální ortonormální bázi, platí v maticovém zápisu

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{X}^+ \mathbf{y},$$

kde  $\mathbf{X}^+$  je *Mooreova-Penroseova pseudoinverzní matice* k matici  $\mathbf{X}$  (viz [4]).

### Speciální případy

Je-li podprostor  $M$  generován jediným vektorem  $\mathbf{x}$ , vede rovnice (2.31) k vyjádření

$$\hat{\mathbf{y}} = \frac{\mathbf{y} \circ \mathbf{x}}{\|\mathbf{x}\|^2} \mathbf{x}. \quad (2.32)$$

V případě, že je délka vektoru  $\mathbf{x}$  jednotková, tedy dostáváme vztah

$$\hat{\mathbf{y}} = (\mathbf{y} \circ \mathbf{x}) \mathbf{x}.$$

Souřadnice vektoru  $\hat{\mathbf{y}}$  vzhledem k ortonormální bázi podprostoru  $M = [\mathbf{x}]$  je pak rovna hodnotě

$$y^* = \mathbf{y} \circ \mathbf{x}. \quad (2.33)$$

Pokud vektory  $\mathbf{x}_1, \dots, \mathbf{x}_k$  tvoří ortogonální bázi podprostoru  $M$  (tj.  $k = m$ ), plynou z rovnic (2.31) vztahy

$$b_i = \frac{\mathbf{y} \circ \mathbf{x}_i}{\|\mathbf{x}_i\|^2}. \quad (2.34)$$

To znamená, že hodnoty  $b_i$  získáme promítnutím vektoru  $\mathbf{y}$  do jednotlivých přímk generovaných vektory  $\mathbf{x}_i$ . Ve speciálním případě, kdy je báze  $\{\mathbf{x}_1, \dots, \mathbf{x}_k\}$  dokonce ortonormální, dostáváme rovnost (2.29).

### Řešení singulární soustavy normálních rovnic

V obecném případě se zpravidla soustava rovnic (2.31) zapisuje v maticovém tvaru,<sup>4</sup> který byl uveden již v kapitole 1.3 (viz (1.16); místo  $\mathbf{Y}$  nyní píšeme  $\mathbf{y}$ ). Otázku řešitelnosti a jednoznačnosti řešení této soustavy jsme již při té příležitosti probrali. Shrňme nyní metody, které máme k dispozici v případě, že vektory  $\mathbf{x}_1, \dots, \mathbf{x}_k$  jsou lineárně závislé a tato soustava je neúplná:

<sup>4</sup>Matice její levé strany  $\mathbf{X}^T \mathbf{X}$  má v pozici s indexem  $i, j$  skalární součin  $\mathbf{x}_i \circ \mathbf{x}_j$  a nazývá se *Gramovou maticí* vektorů  $\mathbf{x}_1, \dots, \mathbf{x}_m$ .

1. Nejjednodušší možností je vynechat z původní skupiny generátorů ty vektory  $\mathbf{x}_i$ , které jsou lineární kombinací ostatních; z takového uspořádání již odvodíme soustavu, která je regulární a jednoznačně řešitelná. Tak vypočteme koeficienty  $b_i$ , které odpovídají nevynechaným vektorům, zatímco koeficienty  $b_i$  odpovídající vynechaným vektorům položíme rovny nule.
2. Řešení neúplné soustavy (1.16) lze vyjádřit ve tvaru

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X} \mathbf{y},$$

kde  $\mathbf{A}^{-}$  je matice *pseudoinverzní* k matici  $\mathbf{A}$ , tj. matice splňující rovnost

$$\mathbf{A} \mathbf{A}^{-} \mathbf{A} = \mathbf{A}.$$

Není ovšem určena jednoznačně a různé možnosti její volby vedou k různým hodnotám koeficientů  $b_i$ .

Podrobnější informace o pseudoinverzních maticích lze nalézt v publikacích [3], [4] a [25].

3. Soustavu (1.16) lze doplnit tzv. *reparametrizačními rovnicemi* svazujícími koeficienty  $b_i$  takovým způsobem, aby vzniklá soustava byla úplná. Tato metoda byla ukázána na příkladech v kapitole 1.11, další poznámky k tomuto tématu jsou podány v kapitole 2.9.

### 2.3.4 Ortogonální projekce a její vlastnosti

Zobrazení  $P_M$ , které každému vektoru  $\mathbf{y} \in V_M$  přiřazuje jeho pravoúhlý průmět  $\hat{\mathbf{y}}$  do podprostoru  $M$ , se nazývá *ortogonální projekce do podprostoru  $M$* . Uveďme některé jeho důležité vlastnosti.

- Toto zobrazení je lineární, což je patrné již ze způsobů, jakým lze pravoúhlý průmět vypočítat (řešením soustavy lineárních rovnic), lze to ale snadno ověřit i přímo: je-li  $s, t \in \mathbb{R}$  a  $\mathbf{x}, \mathbf{y} \in V_n$ , je pravoúhlým průmětem vektoru  $s\mathbf{x} + t\mathbf{y}$  do podprostoru  $M$  vektor  $sP_M(\mathbf{x}) + tP_M(\mathbf{y})$ , neboť splňuje obě podmínky (2.27), (2.28). Zřejmě totiž leží v podprostoru  $M$  a dále platí

$$[s\mathbf{x} + t\mathbf{y}] - [sP_M(\mathbf{x}) + tP_M(\mathbf{y})] = s[\mathbf{x} - P_M(\mathbf{x})] + t[\mathbf{y} - P_M(\mathbf{y})],$$

což je lineární kombinace vektorů kolmých na podprostor  $M$ .

- Platí

$$\mathbf{y} \in M \iff P_M(\mathbf{y}) = \mathbf{y}. \quad (2.35)$$

Leží-li totiž vektor  $\mathbf{y}$  v podprostoru  $M$ , splňuje obě podmínky (2.27), (2.28) pravoúhlého průmětu, který je určen jednoznačně; musí tedy být  $\mathbf{y} = P_M(\mathbf{y})$ . Opačná implikace plyne z podmínky (2.27).

- Obrazem zobrazení  $P_M$  je podprostor  $M$ :

$$\text{Im } P_M = M.$$

To snadno plyne z podmínky (2.27) a předchozího tvrzení.

- Pro všechna  $\mathbf{y} \in V_n$  platí

$$\mathbf{y} \perp M \iff P_M(\mathbf{y}) = \mathbf{0}.$$

Je-li  $\mathbf{y} \perp M$ , splňuje jistě vektor  $\mathbf{0}$  obě podmínky (2.27), (2.28) definice pravoúhlého průmětu; je-li naopak  $P_M(\mathbf{y}) = \mathbf{0}$ , vyplývá z podmínky (2.28), že platí  $\mathbf{y} = \mathbf{y} - P_M(\mathbf{y}) \perp M$ .

- Z toho plyne, že jádrem zobrazení  $P_M$  je množina

$$\text{Ker } P_M = M^\perp.$$

- Zobrazení  $P_M$  je dále *idempotentní*, což znamená, že platí

$$P_M P_M = P_M; \tag{2.36}$$

pro všechna  $\mathbf{x} \in V_n$  je totiž  $P_M(\mathbf{x}) \in M$  a pro všechna  $\mathbf{x} \in M$  je  $P_M(\mathbf{x}) = \mathbf{x}$ , z čehož plyne, že pro všechna  $\mathbf{x} \in V_n$  platí  $P_M P_M(\mathbf{x}) = P_M(\mathbf{x})$ .

- Další vlastností ortogonální projekce je to, že je to zobrazení *samoadjungované*, tj. jakékoli dva vektory  $\mathbf{x}, \mathbf{y} \in V_n$  splňují rovnost

$$\mathbf{x} \circ P_M(\mathbf{y}) = P_M(\mathbf{x}) \circ \mathbf{y}.$$

Pro levou stranu totiž platí

$$\begin{aligned} \mathbf{x} \circ P_M(\mathbf{y}) &= [\mathbf{x} - P_M(\mathbf{x}) + P_M(\mathbf{x})] \circ P_M(\mathbf{y}) = \\ &= [\mathbf{x} - P_M(\mathbf{x})] \circ P_M(\mathbf{y}) + P_M(\mathbf{x}) \circ P_M(\mathbf{y}) = \\ &= P_M(\mathbf{x}) \circ P_M(\mathbf{y}), \end{aligned}$$

a na stejný tvar lze analogickým způsobem upravit pravou stranu.

- Zároveň můžeme ukázat, že každé lineární samoadjungované idempotentní zobrazení je ortogonální projekcí. Nechť  $T$  je takové zobrazení a podprostor  $M \equiv \text{Im}(T)$  je jeho obraz. Jakýkoli vektor  $\mathbf{x} \in M$  je obrazem  $T(\mathbf{z})$  nějakého vektoru  $\mathbf{z} \in V_n$ , takže pro jakékoli dva vektory  $\mathbf{y} \in V_n$ ,  $\mathbf{x} \in M$  platí

$$\begin{aligned} [\mathbf{y} - T(\mathbf{y})] \circ \mathbf{x} &= [\mathbf{y} - T(\mathbf{y})] \circ T(\mathbf{z}) = \\ &= T[\mathbf{y} - T(\mathbf{y})] \circ \mathbf{z} = \\ &= [T(\mathbf{y}) - TT(\mathbf{y})] \circ \mathbf{z} = \\ &= [T(\mathbf{y}) - T(\mathbf{y})] \circ \mathbf{z} = \\ &= 0 \end{aligned}$$

(druhá rovnost plyne ze samoadjungovanosti, třetí z linearit a čtvrtá z idempotence). Vektor  $\mathbf{y} - T(\mathbf{y})$  je tedy kolmý na všechny vektory  $\mathbf{x} \in M$ , tím pádem je kolmý na podprostor  $M$ ; to znamená, že vektor  $T(\mathbf{y})$  je pravoúhlým průmětem vektoru  $\mathbf{y}$  do podprostoru  $M$  (volně podle ([31])).

Poznamenejme, že matice samoadjungovaného zobrazení vzhledem k libovolné ortonormální bázi reálného vektorového prostoru je symetrická.

- Nechť  $M$  je podprostor prostoru  $V_n$ ; definujme zobrazení

$$Q_M(\mathbf{y}) \equiv \mathbf{y} - P_M(\mathbf{y}).$$

Je zřejmé, že platí

$$\begin{aligned} Q_M(\mathbf{y}) &\in M^\perp, \\ \mathbf{y} - Q_M(\mathbf{y}) &\perp M^\perp, \end{aligned}$$

z čehož plyne, že  $Q_M$  je ortogonální projekce do podprostoru  $M^\perp$ .

- Pro všechna  $\mathbf{y} \in V_n$  je splněna rovnost

$$\mathbf{y} = P_M(\mathbf{y}) + Q_M(\mathbf{y}),$$

což znamená, že zobrazení  $P_M + Q_M$  je identita.

- Pro všechna  $\mathbf{y} \in V_n$  platí

$$\|P_M(\mathbf{y})\| \leq \|\mathbf{y}\|. \quad (2.37)$$

Vektory  $P_M(\mathbf{y})$  a  $Q_M(\mathbf{y})$  jsou totiž kolmé a podle Pythagorovy věty je

$$\|\mathbf{y}\|^2 = \|P_M(\mathbf{y})\|^2 + \|Q_M(\mathbf{y})\|^2.$$

- Jsou-li podprostory  $A, B$  kolmé, je složením odpovídajících ortogonálních projekcí nulové zobrazení:

$$A \perp B \implies P_A P_B = 0. \quad (2.38)$$

Pro všechna  $\mathbf{y} \in V_n$  je totiž  $P_B(\mathbf{y}) \in B$ , což je podle předpokladu vektor kolmý na podprostor  $A$ , takže je  $P_A P_B(\mathbf{y}) = \mathbf{0}$ .

Lze se snadno přesvědčit, že platí i obrácená implikace, jedná se tedy o ekvivalenci.

- Platí

$$A \perp B \implies P_A + P_B = P_{A+B}. \quad (2.39)$$

Pro všechna  $\mathbf{y} \in V_n$  je totiž  $P_A(\mathbf{y}) \in A$ ,  $P_B(\mathbf{y}) \in B$ , takže vektor  $P_A(\mathbf{y}) + P_B(\mathbf{y})$  je prvkem podprostoru  $A + B$  a splňuje podmínku (2.27) pravoúhlého průmětu. Abychom dokázali, že je splněna i podmínka (2.28), musíme ukázat, že vektor

$$\mathbf{y} - P_A(\mathbf{y}) - P_B(\mathbf{y}) \quad (2.40)$$

je kolmý na podprostor  $A + B$ . Uvědomme si tedy, že oba vektory  $\mathbf{y} - P_A(\mathbf{y})$  a  $P_B(\mathbf{y})$  jsou kolmé na podprostor  $A$  (první z nich z definice, druhý díky předpokladu  $A \perp B$ ). Tím pádem je na podprostor  $A$  kolmý i vektor (2.40). Podobně lze doložit, že je kolmý i na podprostor  $B$ , je tedy kolmý i na podprostor  $A + B$ .

- Poslední tvrzení lze snadno zobecnit:

$$A = \bigoplus_{i=1}^k A_i \implies P_A = \sum_{i=1}^k P_{A_i}. \quad (2.41)$$

I v tomto případě je možné nahradit implikaci ekvivalencí (viz [31]).

- Z implikace (2.39) vyplývá, že platí

$$B \subseteq A \implies P_A = P_{A-B} + P_B. \quad (2.42)$$

Je totiž  $(A - B) \perp B$ , tj.

$$P_{A-B} + P_B = P_{(A-B)+B} = P_A.$$

- Z tvrzení (2.41) a Pythagorovy věty dále plyne

$$A = \bigoplus_{i=1}^k A_i \implies \|P_A(\mathbf{y})\|^2 = \sum_{i=1}^k \|P_{A_i}(\mathbf{y})\|^2. \quad (2.43)$$

- Platí

$$A \sqcup B \implies P_A P_B = P_C,$$

kde  $C \equiv A \cap B$ . Můžeme totiž psát

$$\begin{aligned} P_A P_B &= (P_{A-C} + P_C)(P_{B-C} + P_C) = \\ &= P_{A-C} P_{B-C} + P_{A-C} P_C + P_C P_{B-C} + P_C P_C = \\ &= P_C P_C = \\ &= P_C. \end{aligned}$$

První rovnost plyne z tvrzení (2.42), třetí rovnost z tvrzení (2.38), neboť je  $\{A - C, B - C, C\} \in \mathcal{P}^\perp$ . Poslední rovnost platí díky tomu, že zobrazení  $P_C$  je idempotentní.

- Z předchozího speciálně plyne

$$B \subseteq A \implies P_B P_A = P_B,$$

neboť je-li  $B \subseteq A$ , platí  $A \sqcup B$  a  $A \cap B = B$ . To znamená, že promítneme-li pravoúhlý průmět  $P_A(\mathbf{y})$  do podprostoru  $B \subseteq A$ , dostaneme tentýž výsledek, jako když vektor  $\mathbf{y}$  promítneme přímo do podprostoru  $B$ . Tuto skutečnost jsme využili v mnoha příkladech.

## 2.4 Tjurův systém

Z hlediska lineárního modelu je výhodné takové uspořádání, kdy je podprostor modelu součtem podprostorů tvořících tzv. *Tjurův systém*,<sup>5</sup> tj. konečnou množinu

<sup>5</sup>Tento termín uvádí Wichura v publikaci [31], podle které je tato kapitola volně zpracována. Míněn je dánský matematik Tue Tjur (nar. 1945); Wichura však neuvádí žádný jeho konkrétní počin, který by tento název vysvětloval. Snad by se mohlo jednat o článek [30].

$\mathcal{T}$  podprostorů  $A_i$  splňující podmínky

$$\begin{aligned} \mathcal{T} &\in \mathcal{P}^{\cup}, \\ V_n &\in \mathcal{T}, \\ A_i, A_j \in \mathcal{T} &\implies A_i \cap A_j \in \mathcal{T}. \end{aligned}$$

Relace  $\subseteq$  určuje na množině  $\mathcal{T}$  částečné uspořádání; minimálním prvkem, který je podmnožinou všech ostatních, je zde zřejmě podprostor

$$\bigcap_{A_i \in \mathcal{T}} A_i,$$

maximálním prvkem je prostor  $V_n$ . Vztahy mezi podprostory tvořícími Tjurův systém lze přehledně znázornit pomocí schématu, ve kterém čára od výše ležícího podprostoru  $A_i$  k níže ležícímu podprostoru  $A_j$  znamená, že platí  $A_j \subset A_i$  (viz příklady 1.8.3, 1.8.4 a příklady na konci této kapitoly).

## 2.4.1 Rozklad Tjurova systému na kolmé podprostory

Definujeme-li pro každý podprostor  $A_i \in \mathcal{T}$

$$L_i \equiv A_i - \sum_{A_j \subset A_i} A_j,$$

lze dokázat tzv. *Tjurův teorém* (viz [31]): podprostory  $L_i$  jsou navzájem kolmé a přitom platí

$$A_i = \sum_{A_j \subset A_i} L_j. \quad (2.44)$$

To je vlastně silnější verze tvrzení (2.26), neboť jakoukoli konečnou množinu navzájem knižně kolmých podprostorů lze zřejmě doplnit na Tjurův systém.<sup>6</sup>

Z tvrzení (2.44) plynou díky implikacím (2.41) a (2.43) analogické vztahy pro pravoúhlé průměty a čtverce jejich délek:

$$P_{A_i} = \sum_{A_j \subset A_i} P_{L_j}, \quad (2.45)$$

$$\|P_{A_i}(\mathbf{y})\|^2 = \sum_{A_j \subset A_i} \|P_{L_j}(\mathbf{y})\|^2. \quad (2.46)$$

Podobný vztah platí samozřejmě i pro dimenze:

$$\dim A_i = \sum_{A_j \subset A_i} \dim L_j. \quad (2.47)$$

---

<sup>6</sup>Chceme-li doplnit navzájem knižně kolmé podprostory  $A_1, \dots, A_k$  na Tjurův systém, musíme k nim přidat všechny jejich možné průniky. Výsledný počet podprostorů bude zřejmě nejvýše  $2^k - 1$ ; v tom případě totiž každému průniku odpovídá právě jedna neprázdna podmnožina množiny  $\{A_1, \dots, A_k\}$ . To znamená, že pro  $k$  navzájem knižně kolmých podprostorů existuje nejvýše  $2^k - 1$  podprostorů  $L_i$  vlastností požadovaných v tvrzení (2.26). Položíme-li  $m \equiv 2^k - 1$ , potřebujeme pro  $(k + 1)$  podprostorů  $A_i$  nejvýše

$$2^{k+1} - 1 = 2 \cdot (2^k - 1) + 1 = 2m + 1$$

podprostorů  $L_i$ , což odpovídá počtu použitým v důkazu tvrzení (2.26).

To znamená, že všechny vztahy mezi projekcemi  $P_{A_i}$  a  $P_{L_j}$ , které odvodíme z rovnosti (2.45), platí analogicky i pro dimenze a čtverce délek příslušných pravoúhlých průmětů (samotných podprostorů  $A_i$ ,  $L_j$  se však tato analogie netýká, neboť zde je význam symbolů  $+$ ,  $-$  odlišný a platí pro ně jiná pravidla).

## 2.4.2 Tjurův systém a třídění

Spojovacím článkem mezi pojmem Tjurova systému a lineárním modelem je vícenásobné třídění. Je-li totiž toto třídění vyvážené, tvoří podprostory určené jednotlivými faktory a jejich interakcemi spolu s celým podprostorem modelu a prostorem  $V_n$  Tjurův systém a výše uvedená tvrzení značně usnadňují výpočet všech potřebných pravoúhlých průmětů, resp. jejich délek. Průměty do podprostorů  $A_i$  určených jednotlivými faktory totiž můžeme snadno nalézt (viz příklad 1.3.5), z nich lze dále vypočítat průměty do podprostorů  $L_i$  a z těch je pak možné „poskládat“ průmět do libovolného dalšího relevantního podprostoru. Průměty  $P_{L_i}$  navíc reprezentují příspěvek odpovídajícího faktoru k celkové variabilitě a čtverce jejich délky mohou posloužit k posouzení vlivu příslušného faktoru pomocí  $F$ -testu.

Tento koncept jsme tedy vlastně použili v příkladech 1.8.3, 1.8.4, kde podprostory  $E, A, B, M, V_{12}$ , resp.  $E, A, B, C, M, N, V_{12}$  byly navzájem knižně kolmé. Pro lepší ilustraci výše naznačených metod uvedeme ještě další dva příklady týkající se trojného třídění.

Poznamenejme ještě, že k tomu, aby podprostory určené jednotlivými faktory byly navzájem knižně kolmé, je podmínka vyváženosti postačující, avšak nikoli nutná. Například podprostory

$$A \equiv \left[ \begin{array}{c} \left( \begin{array}{c} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{array} \right) \\ \left( \begin{array}{c} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \end{array} \right) \end{array} \right], \quad B \equiv \left[ \begin{array}{c} \left( \begin{array}{c} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{array} \right) \\ \left( \begin{array}{c} 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 1 \end{array} \right) \end{array} \right]$$

jsou knižně kolmé a přitom je lze interpretovat jako faktory v nevyváženém třídění. Další podrobnosti viz [31].

## Příklady

### 2.4.3 Trojné třídění

Nechť pro navzájem nezávislé složky  $Y_{ijk}$  náhodného vektoru  $\mathbf{Y}$  platí

$$Y_{ijk} \sim N(\mu + \alpha_i + \beta_j + \gamma_k; \sigma^2),$$

kde  $i, j, k \in \{1, 2\}$ . Koeficienty  $\alpha_i, \beta_j, \gamma_k$  tedy představují vliv tří dvouúrovňových faktorů  $A, B, C$ . Náhodný vektor  $\mathbf{Y}$  tedy můžeme popsat modelem

$$\begin{pmatrix} Y_{111} \\ Y_{112} \\ Y_{121} \\ Y_{122} \\ Y_{211} \\ Y_{212} \\ Y_{221} \\ Y_{222} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \underline{\mu} \\ \underline{\alpha_1} \\ \underline{\alpha_2} \\ \underline{\beta_1} \\ \underline{\beta_2} \\ \underline{\gamma_1} \\ \underline{\gamma_2} \end{pmatrix} + \mathbf{Z} \equiv \\ \equiv \mathbf{X}\boldsymbol{\beta} + \mathbf{Z},$$

kde  $\mathbf{Z} \sim \mathbf{N}(\mathbf{0}; \sigma^2 \mathbf{I}_8)$ . Označme sloupce matice  $\mathbf{X}$  postupně

$$\mathbf{e}, \mathbf{a}_1, \mathbf{a}_2, \mathbf{b}_1, \mathbf{b}_2, \mathbf{c}_1, \mathbf{c}_2$$

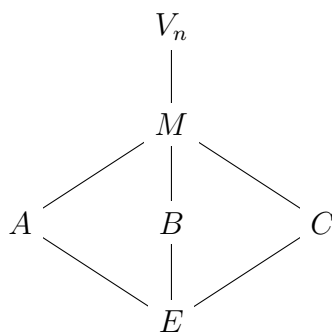
a definujme podprostory

$$\begin{aligned} E &\equiv [\mathbf{e}], \\ A &\equiv [\mathbf{a}_1, \mathbf{a}_2], \\ B &\equiv [\mathbf{b}_1, \mathbf{b}_2], \\ C &\equiv [\mathbf{c}_1, \mathbf{c}_2], \\ M &\equiv [\mathbf{e}, \mathbf{a}_1, \mathbf{a}_2, \mathbf{b}_1, \mathbf{b}_2, \mathbf{c}_1, \mathbf{c}_2] = E + A + B + C = A + B + C; \end{aligned}$$

podprostor  $E$  je jednorozměrný, podprostory  $A, B, C$  jsou dvojrozměrné. Podobně jako v příkladu 1.8.3 lze ukázat, že platí

$$\begin{aligned} A \cap B = B \cap C = A \cap C = A \cap B \cap C &= E, \\ \{A - E, B - E, C - E\} &\in \mathcal{P}^\perp, \end{aligned}$$

takže všechny podprostory  $E, A, B, C, M$  jsou na sebe navzájem knižně kolmé a spolu s prostorem  $V_8$  tvoří Tjurův systém. Vztahy mezi jednotlivými podprostory můžeme znázornit pomocí následujícího schématu:



Nyní položíme

$$\begin{aligned} L_E &\equiv E - \{\} = E, \\ L_A &\equiv A - E, \\ L_B &\equiv B - E, \\ L_C &\equiv C - E, \\ L_M &\equiv M - (A + B + C + E) = \{\mathbf{0}\}; \end{aligned}$$

až na poslední jsou všechny tyto podprostory zřejmě jednorozměrné. Díky tvrzení (2.42) můžeme dále vyjádřit ortogonální projekce do podprostorů  $L_A, L_B, L_C$ :

$$\begin{aligned} P_{L_A} &= P_A - P_E, \\ P_{L_B} &= P_B - P_E, \\ P_{L_C} &= P_C - P_E, \end{aligned}$$

a podle tvrzení (2.45) máme

$$\begin{aligned} P_M &= P_{L_M} + P_{L_A} + P_{L_B} + P_{L_C} + P_{L_E} = \\ &= 0 + (P_A - P_E) + (P_B - P_E) + (P_C - P_E) + P_E = \\ &= P_A + P_B + P_C - 2P_E. \end{aligned}$$

Získáme-li nyní realizaci náhodného vektoru  $\mathbf{Y}$ , snadno určíme (viz příklady 1.3.4, 1.3.5) její pravoúhlé průměty  $\bar{\mathbf{Y}}, \bar{\mathbf{Y}}_A, \bar{\mathbf{Y}}_B, \bar{\mathbf{Y}}_C$  do podprostorů  $E, A, B, C$ ; pravoúhlým průmětem vektoru  $\mathbf{Y}$  do podprostoru  $M$  je pak vektor

$$\mathbf{Y}_M = \bar{\mathbf{Y}}_A + \bar{\mathbf{Y}}_B + \bar{\mathbf{Y}}_C - 2\bar{\mathbf{Y}},$$

pro jehož délku platí analogicky

$$\|\mathbf{Y}_M\|^2 = \|\bar{\mathbf{Y}}_A\|^2 + \|\bar{\mathbf{Y}}_B\|^2 + \|\bar{\mathbf{Y}}_C\|^2 - 2\|\bar{\mathbf{Y}}\|^2$$

(viz (2.46)). Podle vztahu (2.47) dostáváme dále

$$\begin{aligned} \dim M &= \dim L_M + \dim L_A + \dim L_B + \dim L_C + \dim L_E = \\ &= 1 + 1 + 1 + 1 + 0 = \\ &= 4. \end{aligned}$$

Budeme-li chtít posoudit např. významnost vlivu faktoru  $A$ , učiníme tak prostřednictvím statistiky

$$\frac{\|P_{L_A}(\mathbf{Y})\|^2 / \dim L_A}{\|\mathbf{Y} - P_M(\mathbf{Y})\|^2 / (\dim V_8 - \dim M)} = \frac{\|\bar{\mathbf{Y}}_A\|^2 - \|\bar{\mathbf{Y}}\|^2}{[\|\mathbf{Y}\|^2 - \|\bar{\mathbf{Y}}_M\|^2] / 4},$$

která má za předpokladu platnosti hypotézy  $\alpha_1 = \alpha_2$  (tj. absence vlivu faktoru  $A$ ) rozdělení  $F_{1,4}$ .

#### 2.4.4 Trojné třídění s jednou interakcí prvního řádu

Chceme-li zahrnout do předchozího modelu ještě interakce mezi faktory  $A, B$ , musíme pro složky  $Y_{ijk}$  náhodného vektoru  $\mathbf{Y}$  uvažovat vztah

$$Y_{ijk} \sim N(\mu + \alpha_i + \beta_j + \gamma_k + \delta_{ij}; \sigma^2);$$

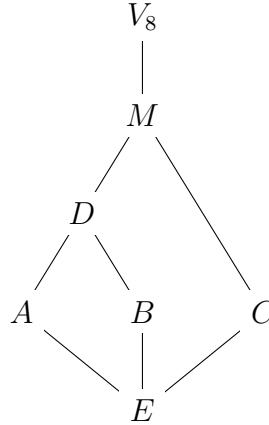
koeficienty  $\delta_{ij}$  zde představují vliv interakcí. Náhodný vektor  $\mathbf{Y}$  tedy můžeme popsat modelem

$$\mathbf{Y} = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} -\frac{\mu}{\alpha_1} \\ -\frac{\alpha_2}{\beta_1} \\ \beta_2 \\ -\frac{\gamma_1}{\gamma_2} \\ \delta_{11} \\ \delta_{12} \\ \delta_{21} \\ \delta_{22} \end{pmatrix} + \mathbf{Z},$$

kde  $\mathbf{Z} \sim \mathbf{N}(\mathbf{0}; \sigma^2 \mathbf{I}_8)$ . Ponechme označení z předchozího příkladu, ale doplňme jej o čtyřrozměrný podprostor  $D$  generovaný posledními čtyřmi sloupci odpovídajícími interakcím a přizpůsobme definici podprostoru  $M$ :

$$M = E + A + B + C + D = C + D$$

(neboť je zřejmé  $A, B, E \subseteq D$ ). Opět lze ukázat, že podprostory  $E, A, B, C, D, M, V_n$  tvoří Tjurův systém, jehož strukturu nám ukazuje následující schéma:



Definice podprostorů  $L_E, L_A, L_B, L_C$  můžeme ponechat beze změny z předchozího příkladu, dále doplňme

$$\begin{aligned} L_D &\equiv D - (A + B + E) = D - (A + B), \\ L_M &\equiv M - (A + B + C + D + E) = \{\mathbf{0}\}. \end{aligned}$$

Aplikací tvrzení (2.45) dostaneme

$$\begin{aligned} P_D &= P_{L_D} + P_{L_A} + P_{L_B} + P_{L_E} = \\ &= P_{L_D} + (P_A - P_E) + (P_B - P_E) + P_E = \\ &= P_{L_D} + P_A + P_B - P_E, \\ P_M &= P_{L_M} + P_{L_D} + P_{L_A} + P_{L_B} + P_{L_C} + P_{L_E} = \\ &= 0 + P_{L_D} + (P_A - P_E) + (P_B - P_E) + (P_C - P_E) + P_E = \\ &= P_{L_D} + P_A + P_B + P_C - 2P_E. \end{aligned}$$

Projekce  $P_A, P_B, P_C, P_E$  a  $P_D$  opět určíme snadno,<sup>7</sup> a jelikož víme, že platí

$$\begin{aligned} P_D &= P_{L_D} + P_{L_A} + P_{L_B} + P_{L_E} = \\ &= P_{L_D} + (P_A - P_E) + (P_B - P_E) + P_E = \\ &= P_{L_D} + P_A + P_B - P_E, \end{aligned}$$

můžeme vyjádřit projekci  $P_{L_D}$ :

$$P_{L_D} = P_D - P_A - P_B + P_E \quad (2.48)$$

a dosadit ji do výrazu pro  $P_M$ :

$$\begin{aligned} P_M &= (P_D - P_A - P_B + P_E) + P_A + P_B + P_C - 2P_E = \\ &= P_D + P_C - P_E. \end{aligned} \quad (2.49)$$

Pravoúhlým průmětem vektoru  $\mathbf{Y}$  do podprostoru  $M$  je tedy vektor

$$\mathbf{Y}_M = \overline{\mathbf{Y}}_D + \overline{\mathbf{Y}}_C - \overline{\mathbf{Y}}.$$

Chceme-li dále například testovat vliv interakcí  $D$ , určíme nejdříve dimenze podprostorů  $L_D$  a  $M$ , a to ze vztahů analogických rovnostem (2.48), (2.49):

$$\begin{aligned} \dim L_D &= \dim D - \dim A - \dim B + \dim E = \\ &= 4 - 2 - 2 + 1 = \\ &= 1, \\ \dim M &= \dim D + \dim C - \dim E = \\ &= 4 + 2 - 1 = \\ &= 5; \end{aligned}$$

pak vypočteme hodnotu statistiky

$$\frac{\|P_{L_D}(\mathbf{Y})\|^2 / \dim L_D}{\|\mathbf{Y} - P_M(\mathbf{Y})\|^2 / (\dim V_8 - \dim M)} = \frac{\|\overline{\mathbf{Y}}_D\|^2 - \|\overline{\mathbf{Y}}_A\|^2 - \|\overline{\mathbf{Y}}_B\|^2 + \|\overline{\mathbf{Y}}\|^2}{[\|\mathbf{Y}\|^2 - \|\overline{\mathbf{Y}}_D\|^2 - \|\overline{\mathbf{Y}}_C\|^2 + \|\overline{\mathbf{Y}}\|^2] / 3},$$

která má v případě absence vlivu interakcí  $D$  rozdělení  $F_{1,3}$ .

## 2.5 Náhodný vektor a jeho charakteristiky

### 2.5.1 Poznámky k definici náhodného vektoru

V kapitole 1.1.4 jsme uvedli geometrickou definici náhodného vektoru způsobem spíše intuitivním. Zmiňme se nyní stručně o dvou způsobech, jak lze k tomuto pojmu přistoupit exaktněji.

První možností je zavést přímo na vektorovém prostoru  $V_n$   $\sigma$ -algebru, na které definujeme pravděpodobnostní míru.

Jiná alternativa je použita v publikaci [31]: nechť trojice  $(\Sigma, \mathcal{A}, \mathbb{P})$  je pravděpodobnostní prostor. Náhodný vektor je pak taková funkce  $\mathbf{Y} : \Sigma \rightarrow V_n$ , že pro všechna  $\mathbf{a} \in V_n$  je výraz  $\mathbf{a} \circ \mathbf{Y}$  náhodná veličina (na prostoru  $V_n$  tedy předpokládáme zavedený skalární součin). To je ekvivalentní požadavku, aby  $\mathbf{Y}$  byla měřitelná funkce vzhledem k nejmenší algebře  $(V_n, \mathcal{B})$ , vzhledem ke které jsou měřitelné všechny lineární formy na prostoru  $V_n$  (tj. funkce  $\mathbf{x} \rightarrow \mathbf{a} \circ \mathbf{x}$ , kde  $\mathbf{a} \in V_n$ ).

<sup>7</sup>Interakce  $D$  vlastně představují samostatný faktor; často se tento faktor označuje symbolem  $A \times B$ .

## 2.5.2 Střední hodnota, varianční operátor

Podobným způsobem jsou v knize [31] zavedeny další elementární pojmy: střední hodnotou  $\mathbf{E} \mathbf{Y}$  náhodného vektoru  $\mathbf{Y}$  je míněn takový vektor  $\boldsymbol{\mu} \in V_n$ , že pro všechna  $\mathbf{a} \in V_n$  platí

$$\mathbf{E}(\mathbf{a} \circ \mathbf{Y}) = \mathbf{a} \circ \boldsymbol{\mu}; \quad (2.50)$$

varianční operátor<sup>8</sup> je zde definován jako takový homomorfismus  $\sum_{\mathbf{Y}} : V_n \rightarrow V_n$ , který splňuje podmínku

$$\text{cov}(\mathbf{a} \circ \mathbf{Y}, \mathbf{b} \circ \mathbf{Y}) = \mathbf{a} \circ \sum_{\mathbf{Y}}(\mathbf{b})$$

pro všechna  $\mathbf{a}, \mathbf{b} \in V_n$ . Zatímco střední hodnota je výše uvedeným způsobem definována jednoznačně, varianční operátor závisí na zavedeném skalárním součinu.

Obě definice považujeme za velmi praktické a jejich aktivní používání se nám osvědčilo. Uveďme jejich použití v důkazech některých elementárních tvrzení; zdrojem všech níže uvedených myšlenek týkajících se tohoto tématu je kniha [31].

- Nechť  $Y_1, \dots, Y_n$  jsou souřadnice náhodného vektoru  $\mathbf{Y}$  vzhledem k ortonormální bázi  $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}$  a  $v_{ij}$  jsou prvky matice homomorfismu  $\sum_{\mathbf{Y}}$  vzhledem k této bázi. Pak platí

$$\begin{aligned} \text{cov}(Y_i, Y_j) &= \text{cov}(\mathbf{e}_i \circ \mathbf{Y}, \mathbf{e}_j \circ \mathbf{Y}_j) = \\ &= \mathbf{e}_i \circ \sum_{\mathbf{Y}}(\mathbf{e}_j) = \\ &= v_{ij}. \end{aligned}$$

Maticí homomorfismu  $\sum_{\mathbf{Y}}$  vzhledem k dané ortonormální bázi je tedy standardní varianční matice, jejíž prvky jsou rozptyly a kovariance souřadnic vektoru  $\mathbf{Y}$  vzhledem k této bázi.

- Zobrazení  $\sum_{\mathbf{Y}}$  je pozitivně semidefinitní a samoadjungované, neboť pro libovolné vektory  $\mathbf{a}, \mathbf{b} \in V_n$  platí

$$\begin{aligned} \mathbf{a} \circ \sum_{\mathbf{Y}}(\mathbf{a}) &= \text{var}(\mathbf{a} \circ \mathbf{Y}) \geq 0, \\ \mathbf{a} \circ \sum_{\mathbf{Y}}(\mathbf{b}) &= \text{cov}(\mathbf{a} \circ \mathbf{Y}, \mathbf{b} \circ \mathbf{Y}) = \\ &= \text{cov}(\mathbf{b} \circ \mathbf{Y}, \mathbf{a} \circ \mathbf{Y}) = \\ &= \mathbf{b} \circ \sum_{\mathbf{Y}}(\mathbf{a}) = \\ &= \sum_{\mathbf{Y}}(\mathbf{a}) \circ \mathbf{b}. \end{aligned}$$

To je jiným vyjádřením skutečnosti, že varianční matice je pozitivně semidefinitní a symetrická (neboť matice samoadjungovaného zobrazení vzhledem k ortonormální bázi je symetrická).

- Je-li  $T : V_n \rightarrow W_k$  libovolný homomorfismus, platí

$$\mathbf{E}T(\mathbf{Y}) = T(\mathbf{E}\mathbf{Y}), \quad (2.51)$$

neboť můžeme psát

$$\mathbf{E}[\mathbf{w} \bullet T(\mathbf{Y})] = \mathbf{E}[T'(\mathbf{w}) \circ \mathbf{Y}] = T'(\mathbf{w}) \circ \mathbf{E}\mathbf{Y} = \mathbf{w} \bullet T(\mathbf{E}\mathbf{Y}),$$

---

<sup>8</sup>V originále „dispersion operator“.

kde  $w$  je libovolný prvek vektorového prostoru  $W_k$ ,  $\bullet$  je skalární součin na tomto prostoru a  $T'$  je homomorfismus adjungovaný k homomorfismu  $T$  vzhledem k použitým skalárním součinům.

V maticovém zápisu má tento vztah podobu (1.1).

- Podobně můžeme odvodit pro libovolné dva vektory  $a, b \in W_k$  rovnosti

$$\begin{aligned} \text{cov}[a \bullet T(\mathbf{Y}), b \bullet T(\mathbf{Y})] &= \text{cov}[T'(a) \circ \mathbf{Y}, T'(b) \circ \mathbf{Y}] = \\ &= T'(a) \circ \sum_{\mathbf{Y}} T'(b) = \\ &= a \bullet T \sum_{\mathbf{Y}} T'(b), \end{aligned}$$

z čehož plyne vztah

$$\sum_{T(\mathbf{Y})} = T \sum_{\mathbf{Y}} T'. \quad (2.52)$$

Maticovou analogií tohoto tvrzení je vzorec (1.2).

### 2.5.3 Obraz variančního operátoru

Nahlížíme-li na varianční operátor  $\sum_{\mathbf{Y}}$  jako na lineární zobrazení, má toto zobrazení zajímavou vlastnost, které bychom si jinak nemuseli povšimnout: jeho obrazem je nejmenší podprostor prostoru  $V_n$ , ve kterém leží náhodný vektor  $\mathbf{Y} - \boldsymbol{\mu}$  s pravděpodobností 1, tj. formálně

$$\mathbb{P}[\mathbf{Y} - \boldsymbol{\mu} \in \text{lm } \sum_{\mathbf{Y}}] = 1, \quad (2.53)$$

$$\mathbb{P}[\mathbf{Y} - \boldsymbol{\mu} \in M] = 1 \implies \text{lm } \sum_{\mathbf{Y}} \subseteq M, \quad (2.54)$$

kde  $M$  je libovolný podprostor prostoru  $V_n$ .

Nejdříve ukažme, že je splněna rovnost (2.53). Pro jakýkoli vektor  $a$  ležící v podprostoru  $(\text{lm } \sum_{\mathbf{Y}})^\perp$  platí

$$\text{var}(a \circ \mathbf{Y}) = a \circ \sum_{\mathbf{Y}}(a) = 0,$$

což znamená, že výraz  $a \circ \mathbf{Y}$  je s pravděpodobností 1 roven nějaké konstantě. Touto konstantou musí být ovšem hodnota  $\mathbb{E}(a \circ \mathbf{Y}) = a \circ \boldsymbol{\mu}$ . Můžeme tedy psát

$$\begin{aligned} 1 &= \mathbb{P}[\forall a \in (\text{lm } \sum_{\mathbf{Y}})^\perp : a \circ \mathbf{Y} = a \circ \boldsymbol{\mu}] = \\ &= \mathbb{P}[\forall a \in (\text{lm } \sum_{\mathbf{Y}})^\perp : a \circ (\mathbf{Y} - \boldsymbol{\mu}) = 0] = \\ &= \mathbb{P}[\mathbf{Y} - \boldsymbol{\mu} \perp (\text{lm } \sum_{\mathbf{Y}})^\perp] = \\ &= \mathbb{P}[\mathbf{Y} - \boldsymbol{\mu} \in \text{lm } \sum_{\mathbf{Y}}]. \end{aligned}$$

Tvrzení (2.54) dokažme sporem. Předpokládejme, že existuje takový podprostor  $M \subset \text{lm } \sum_{\mathbf{Y}}$  (tj.  $M \neq \text{lm } \sum_{\mathbf{Y}}$ ), že pravděpodobnost jevu  $\mathbf{Y} - \boldsymbol{\mu} \in M$  je 1. Podprostor  $\text{lm } \sum_{\mathbf{Y}} - M$  je neprázdný; zvolme tedy nějaký jeho nenulový prvek  $a$ . Tento vektor leží v podprostoru  $\text{lm } \sum_{\mathbf{Y}}$ , takže existuje nějaký vektor  $b \in V_n$ , pro který je  $a = \sum_{\mathbf{Y}}(b)$ . Pro vektory  $a, b$  tedy platí

$$\text{cov}(a \circ \mathbf{Y}, b \circ \mathbf{Y}) = a \circ \sum_{\mathbf{Y}}(b) = \|a\|^2 \neq 0.$$

Vektor  $\mathbf{a}$  je však také kolmý na podprostor  $M$ , takže podle předpokladu platí

$$\mathbb{P} [\mathbf{a} \circ (\mathbf{Y} - \boldsymbol{\mu}) = 0] = 1,$$

z čehož plyne

$$\text{cov}(\mathbf{a} \circ \mathbf{Y}, \mathbf{b} \circ \mathbf{Y}) = \text{cov}[\mathbf{a} \circ (\mathbf{Y} - \boldsymbol{\mu}), \mathbf{b} \circ \mathbf{Y}] = 0,$$

což je spor.

Máme-li tedy varianční operátor vyjádřený pomocí varianční matice, je jakákoli možná realizace náhodného vektoru  $\mathbf{Y}$  lineární kombinací jejích sloupců.

Doplňme, že pokud neplatí rovnost  $\text{lm} \sum_{\mathbf{Y}} = V_n$ , nazývá se rozdělení náhodného vektoru  $\mathbf{Y}$  *singulární*.

#### 2.5.4 Přejchod k jinému skalárnímu součinu

Jak bylo výše uvedeno, závisí varianční operátor definovaný jako lineární zobrazení na aktuálně zavedeném skalárním součinu. Necht  $\bullet$  je jiný skalární součin; pak existuje samoadjungovaný pozitivně definitní homomorfismus  $G$  splňující podmínku

$$\mathbf{a} \bullet \mathbf{b} = \mathbf{a} \circ G(\mathbf{b})$$

pro všechny vektory  $\mathbf{a}, \mathbf{b} \in V_n$  (viz [31]). Platí tedy

$$\begin{aligned} \text{cov}(\mathbf{a} \bullet \mathbf{Y}, \mathbf{b} \bullet \mathbf{Y}) &= \text{cov}[G(\mathbf{a}) \circ \mathbf{Y}, G(\mathbf{b}) \circ \mathbf{Y}] = \\ &= G(\mathbf{a}) \circ \sum_{\mathbf{Y}} G(\mathbf{b}) = \\ &= \mathbf{a} \circ G \sum_{\mathbf{Y}} G(\mathbf{b}) = \\ &= \mathbf{a} \bullet \sum_{\mathbf{Y}} G(\mathbf{b}), \end{aligned}$$

což znamená, že variančním operátorem vzhledem ke skalárnímu součinu  $\bullet$  je zobrazení  $\sum_{\mathbf{Y}} G$ .

Toho lze využít, pokud nám stávající varianční operátor nevyhovuje. Doposud jsme ve všech příkladech předpokládali, že varianční matice náhodného vektoru  $\mathbf{Y}$  je  $\sigma^2$ -násobkem matice jednotkové. Analogií tohoto předpokladu v „coordinate-free“ přístupu je požadavek, aby zobrazení  $\sum_{\mathbf{Y}}$  bylo  $\sigma^2$ -násobkem identity na prostoru  $V_n$ .<sup>9</sup> Je-li tato podmínka splněna, platí pro všechny vektory  $\mathbf{a}, \mathbf{b} \in V_n$

$$\text{var}(\mathbf{a} \circ \mathbf{Y}) = \sigma^2 \|\mathbf{a}\|^2, \quad (2.55)$$

$$\text{cov}(\mathbf{a} \circ \mathbf{Y}, \mathbf{b} \circ \mathbf{Y}) = \sigma^2 \mathbf{a} \circ \mathbf{b}, \quad (2.56)$$

díky čemuž lze odvodit, že pravoúhlý průmět vektoru  $\mathbf{Y}$  do podprostoru daného modelem je nejlepším nestranným lineárním odhadem střední hodnoty (viz kapitolu 2.8) a průměty do navzájem kolmých podprostorů jsou nekorelované; v případě normality jsou pak i nezávislé a jejich příslušné funkce mají rozdělení  $\chi^2$ ,  $F$ , resp.  $t$ .

V praxi se však často vyskytuje obecnější případ  $\sum_{\mathbf{Y}} \equiv \sigma^2 V$ , kde  $V$  je známý pozitivně definitní a samoadjungovaný homomorfismus (v maticové formě mu

<sup>9</sup>V takovém případě se rozdělení vektoru  $\mathbf{Y}$  nazývá *slabě sférické*.

odpovídá symetrická pozitivně definitní matice) a  $\sigma^2$  je neznámý parametr. Tehdy ovšem vztahy (2.55), (2.56) neplatí; jednou z možností, jak toto úskalí obejít, je předefinovat skalární součin. Položíme-li

$$\mathbf{a} \bullet \mathbf{b} \equiv \mathbf{a} \circ V^{-1}(\mathbf{b})$$

(nezbytnou podmínkou je samozřejmě regularita zobrazení  $\sum_{\mathbf{Y}}$ ),<sup>10</sup> je variančním operátorem náhodného vektoru  $\mathbf{Y}$  vzhledem ke skalárnímu součinu  $\bullet$  homomorfismus  $\sigma^2 V V^{-1} = \sigma^2 I$ , tj.  $\sigma^2$ -násobek identity. Tím je problém převeden na předchozí situaci. Nesmíme ovšem zapomenout na to, že se změnou skalárního součinu se mění i pojem kolmosti, a tedy i pravoúhlého průmětu.

V knize [35] je uveden alternativní postup, vycházející ze vztahu (2.52): je-li  $\sum_{\mathbf{Y}} = \sigma^2 V$ , nalezneme zobrazení  $T$ , pro které platí  $T V T' = I$  (toto zobrazení se běžně označuje jako  $V^{-1/2}$ ), a pak pracujeme s transformovaným náhodným vektorem  $T(\mathbf{Y})$ , jehož varianční operátor je  $T \sigma^2 V T' = \sigma^2 I$ . Je ovšem třeba transformovat i podprostor  $M$  daný modelem.<sup>11</sup> Obě metody vedou pochopitelně ke stejným výsledkům.

### 2.5.5 Rozklad samoadjungovaného zobrazení na součet ortogonálních projekcí

V souvislosti s posledně uvedenou metodou uveďme zmiňme jedno užitečné tvrzení: je-li  $V$  samoadjungovaný homomorfismus, lze jej zapsat jako lineární kombinaci navzájem kolmých ortogonálních projekcí, tj. existují navzájem kolmé podprostory  $A_1, \dots, A_k$  ( $k \leq n$ ), které tvoří ortogonální rozklad prostoru  $V_n$  a přitom platí

$$V = \sum_{i=1}^k c_i P_{A_i}, \quad (2.57)$$

kde  $c_i \in \mathbb{R}$  (důkaz viz [31]).

Pro libovolný vektor  $\mathbf{x} \in A_i$  zřejmě platí

$$V(\mathbf{x}) = c_i \mathbf{x},$$

takže koeficienty  $c_i$  jsou vlastní čísla zobrazení  $V$  a vektory ležící v podprostoru  $A_i$  jsou jim odpovídající vlastní vektory. V každém z podprostorů  $A_i$  lze tedy zvolit nějakou ortonormální bázi, jejíž prvky jsou vlastní vektory zobrazení  $V$  příslušné k vlastnímu číslu  $c_i$  (jehož násobnost odpovídá dimenzi příslušného podprostoru). Sjednocení těchto bází tvoří ortonormální bázi  $\mathcal{B}^*$  prostoru  $V_n$ . Je-li nyní  $\mathbf{V}$  matice zobrazení  $V$  vzhledem k nějaké původní ortonormální bázi  $\mathcal{B}$  a  $\mathbf{U}$  matice přechodu od báze  $\mathcal{B}^*$  k bázi  $\mathcal{B}$ <sup>12</sup>, lze rovnost (2.57) vyjádřit ve tvaru

$$\mathbf{V} = \mathbf{U} \mathbf{D} \mathbf{U}^T,$$

<sup>10</sup>V maticovém zápisu se tedy jedná o to, že nahrazujeme původní skalární součin  $\mathbf{a} \circ \mathbf{b} = \mathbf{a}^T \mathbf{b}$  alternativním skalárním součinem  $\mathbf{a} \bullet \mathbf{b} \equiv \mathbf{a}^T \mathbf{V}^{-1} \mathbf{b}$ , kde  $\mathbf{V}$  je varianční matice, resp. její násobek.

<sup>11</sup>Tento postup můžeme interpretovat tak, že zatímco v případě slabě sférického rozdělení tvoří množina vektorů  $\mathbf{x}$ , pro které je hodnota výrazu  $\text{var}(\mathbf{x} \circ \mathbf{Y})$  rovna konstantě, povrch hyperkoule, v obecném případě se jedná o povrch obecného elipsoidu. Abychom tedy získali slabě sférické rozdělení, je třeba vektorový prostor  $V_n$  patřičně „natáhnout“.

<sup>12</sup>Ve smyslu, jak je definována v publikaci [4], tj. matice, jejíž sloupce představují souřadnice vektorů báze  $\mathcal{B}^*$  vzhledem k bázi  $\mathcal{B}$ .

kde  $\mathbf{D}$  je diagonální matice, která má na diagonále vlastní čísla  $c_i$  v počtu a pořadí odpovídajícím pořadí vektorů báze  $\mathcal{B}^*$ . Vskutku, budeme-li násobit výraz na pravé straně zprava sloupcem souřadnic vzhledem k původní bázi  $\mathcal{B}$ , představuje násobení maticí  $\mathbf{U}^T$  transformaci těchto souřadnic na souřadnice vzhledem k bázi  $\mathcal{B}^*$  (obě báze jsou ortonormální, takže platí  $\mathbf{U}^{-1} = \mathbf{U}^T$ ), násobení maticí  $\mathbf{D}$  reprezentuje vynásobení těchto transformovaných souřadnic příslušným koeficientem  $c_i$  a konečně se pomocí matice  $\mathbf{U}$  vrátíme zpět k původní bázi.

Je-li zobrazení  $V$  pozitivně definitní, platí pro libovolný nenulový vektor  $\mathbf{x} \in A_i$

$$0 < \mathbf{x} \circ V(\mathbf{x}) = c_i \|\mathbf{x}\|^2,$$

z čehož plyne, že je  $c_i > 0$  pro všechna  $i = 1, \dots, k$ . Je-li tedy pro regulární (a tím pádem i pozitivně definitní) varianční operátor  $\sigma^2 V$  třeba najít zobrazení  $T$ , pro které platí  $TVT' = I$ , stačí položit

$$T = \sum_{i=1}^k \frac{P_{A_i}}{\sqrt{c_i}}.$$

Požadovaná rovnost pak plyne z toho, že  $T$  je (jakožto součet samoadjungovaných zobrazení) samoadjungované a platí  $P_{A_i} P_{A_j} = 0$  pro  $i \neq j$  a  $P_{A_i} P_{A_i} = P_{A_i}$  (viz tvrzení (2.38) a (2.36)).

V maticovém zápisu to znamená položit

$$\mathbf{T} = \mathbf{U} \mathbf{D}^{-1/2} \mathbf{U}^T,$$

kde  $\mathbf{D}^{-1/2}$  je diagonální matice, jejíž prvky na diagonále jsou rovny hodnotám  $1/\sqrt{c_i}$ , kde  $c_i$  jsou vlastní čísla matice  $\mathbf{V}$  (samozřejmě v pořadí odpovídajícím pořadí normovaných vlastních vektorů reprezentovaných sloupci matice  $\mathbf{U}$ ).

## 2.6 Geometrické vlastnosti mnohorozměrného normálního rozdělení

Definici mnohorozměrného normálního rozdělení jsme uvedli již na straně 9 a nepokládáme za nutné ji doplňovat; upozorníme pouze, že formulací „každá lineární funkce náhodného vektoru  $\mathbf{Y}$ “ míníme jakýkoli výraz tvaru  $\mathbf{a} \circ \mathbf{Y}$ , kde  $\mathbf{a}$  je libovolný vektor prostoru  $V_n$ .

V následujících odstavcích se pro přehlednost omezíme na předpoklad, že náhodný vektor  $\mathbf{Y}$  má rozdělení  $\mathbf{N}(\mathbf{0}; \mathbf{I}_n)$ . Jeho hustota je tedy v prostoru  $V_n$  určena předpisem

$$f(\mathbf{y}) = (2\pi)^{-\frac{n}{2}} \exp\left\{-\frac{r^2}{2}\right\}, \quad (2.58)$$

kde  $r$  je vzdálenost bodu  $\mathbf{y}$  od počátku soustavy souřadnic.

### 2.6.1 Rotace soustavy souřadnic

V kapitole 1.4 jsme zdůvodnili invarianci normálního rozdělení  $\mathbf{N}(\mathbf{0}; \mathbf{I}_n)$  vůči rotaci soustavy souřadnic (tj. vůči přechodu k alternativní ortonormální bázi) víceméně intuitivně odkazem na nezávislost hustoty (2.58) na směru. Považujeme

tento argument za dostačující, pro úplnost ale přesto uvedeme další dvě možnosti, jak lze tuto důležitou skutečnost doložit.

Způsob uvedený např. v [27] nebo [17] využívá vztahů (2.55), (2.56): jelikož nová báze  $\{\mathbf{e}_1^* \dots, \mathbf{e}_n^*\}$  je ortonormální, můžeme získat nové souřadnice  $Y_i^*$  z pravoúhlých průmětů vektoru  $\mathbf{Y}$  do podprostorů  $[\mathbf{e}_i^*]$ :

$$Y_i^* = \mathbf{e}_i^* \circ \mathbf{Y}$$

(viz vztah (2.33), resp. (2.29), kde podprostor  $M$  nahradíme prostorem  $V_n$ ). Pro tyto nové souřadnice tedy platí

$$\begin{aligned} \mathbb{E}Y_i^* &= \mathbf{e}_i^* \circ \mathbf{0} = 0, \\ \text{var } Y_i^* &= \|\mathbf{e}_i^*\|^2 = 1, \\ \text{cov}(Y_i^*, Y_j^*) &= \mathbf{e}_i^* \circ \mathbf{e}_j^* = 0. \end{aligned}$$

Nové souřadnice mají tedy opět rozdělení  $N(\mathbf{0}; \mathbf{I}_n)$ .

Jinou alternativou je využít maticového počtu a vztahů (1.1), (1.2). Je-li  $\mathbf{B}$  matice přechodu od původní ortonormální báze k bázi nové, je tato matice ortonormální, tj. platí  $\mathbf{B}^T \mathbf{B} = \mathbf{I}_n$ . To ovšem znamená, že matice  $\mathbf{B}^T$  je maticí inverzní k matici  $\mathbf{B}$ , a je proto určena jednoznačně. Musí tedy platit také  $\mathbf{B} \mathbf{B}^T = \mathbf{I}_n$ . Pro nové souřadnice  $\mathbf{Y}^* \equiv (Y_1^*, \dots, Y_n^*)$  tak máme

$$\begin{aligned} \mathbb{E} \mathbf{Y}^* &= \mathbf{B} \mathbf{0} = \mathbf{0}, \\ \mathbf{V}_{\mathbf{Y}^*} &= \mathbf{B} \mathbf{I}_n \mathbf{B}^T = \mathbf{I}_n, \end{aligned}$$

takže mají rozdělení  $N(\mathbf{0}; \mathbf{I}_n)$ .

Rozdíl mezi oběma způsoby je ovšem pouze formální.

## 2.6.2 Rozdělení podmíněného pravoúhlého průmětu

Nechť  $M \subset V_n$  je podprostor dimenze  $m$ . Označme symboly  $\mathbf{P}, \mathbf{Q}$  pravoúhlé průměty náhodného vektoru  $\mathbf{Y}$  do podprostorů  $M$  a  $M^\perp$  a určíme hustotu náhodného vektoru  $\mathbf{P}$  za podmínky  $\mathbf{Q} = \mathbf{q}$ , kde  $\mathbf{q} \in M^\perp$ .<sup>13</sup>

Za tím účelem zvolme nejprve ortonormální bázi tak, že vektory  $\mathbf{e}_1, \dots, \mathbf{e}_m$  generují podprostor  $M$  a vektor  $\mathbf{q}$  je  $q$ -násobkem vektoru  $\mathbf{e}_{m+1}$ . Realizace  $\mathbf{y}$  vyhovující podmínce  $\mathbf{Q} = \mathbf{q}$  tvoří lineární množinu, jejíž prvky mají vzhledem k použité bázi souřadnice

$$(y_1, \dots, y_m, q, 0, \dots, 0)^T,$$

příčemž

$$(y_1, \dots, y_m, 0, 0, \dots, 0)^T$$

jsou zřejmě souřadnice odpovídajících realizací  $\mathbf{p}$  vektoru  $\mathbf{P}$ . Pro hustotu na této množině platí

$$f(\mathbf{y}) = (2\pi)^{-\frac{n}{2}} e^{-\frac{y_1^2 + \dots + y_m^2 + q^2}{2}} = (2\pi)^{-\frac{n}{2}} e^{-\frac{\|\mathbf{p}\|^2}{2}} e^{-\frac{\|\mathbf{q}\|^2}{2}},$$

<sup>13</sup>Máme samozřejmě na mysli hustotu ve smyslu podkapitoly 1.1.2, tj. funkci definovanou na podprostoru  $M$ .

takže předpis hledané podmíněné hustoty je

$$g(\mathbf{p}|\mathbf{Q} = \mathbf{q}) = Ce^{-\frac{\|\mathbf{p}\|^2}{2}} e^{-\frac{\|\mathbf{q}\|^2}{2}},$$

kde  $C$  je kladná konstanta. Musí být ovšem splněna podmínka

$$\begin{aligned} 1 &= \int_{\mathbf{p} \in M} g(\mathbf{p}|\mathbf{Q} = \mathbf{q}) dV = \\ &= Ce^{-\frac{\|\mathbf{q}\|^2}{2}} \int_{\mathbf{p} \in M} e^{-\frac{\|\mathbf{p}\|^2}{2}} dV, \end{aligned}$$

kde  $V$  je  $m$ -rozměrný objem definovaný na podprostoru  $M$ . Hodnota posledně uvedeného integrálu závisí pouze na dimenzi podprostoru  $M$ . Z toho plyne, že hodnota výrazu  $Ce^{-\frac{\|\mathbf{q}\|^2}{2}}$ , a tedy ani předpis podmíněné hustoty, nezávisí na vektoru  $\mathbf{q}$ ; ze srovnání se vzorcem (2.58) je zřejmé, že musí vyjít

$$\begin{aligned} g(\mathbf{p}|\mathbf{Q} = \mathbf{q}) &= (2\pi)^{-\frac{m}{2}} \exp\left\{-\frac{\|\mathbf{p}\|^2}{2}\right\} = \\ &= (2\pi)^{-\frac{m}{2}} \exp\left\{-\frac{r^2}{2}\right\}, \end{aligned} \tag{2.59}$$

kde  $r$  je délka vektoru  $\mathbf{p}$ , tj. vzdálenost bodu  $\mathbf{p}$  od počátku soustavy souřadnic. Rozdělení podmíněného pravoúhlého průmětu je tedy – až na konstantu kompenzující vliv dimenze příslušného podprostoru – stejné jako rozdělení původního náhodného vektoru.<sup>14</sup>

Z toho bezprostředně plyne nezávislost jakýchkoli dvou pravoúhlých průmětů do dvou navzájem kolmých podprostorů, speciálně tedy i nezávislost souřadnic vzhledem k libovolné ortogonální bázi. Dalším důsledkem je to, že vzorec (2.59) je zároveň předpisem hustoty *nepodmíněného* pravoúhlého průmětu do podprostoru  $M$ .

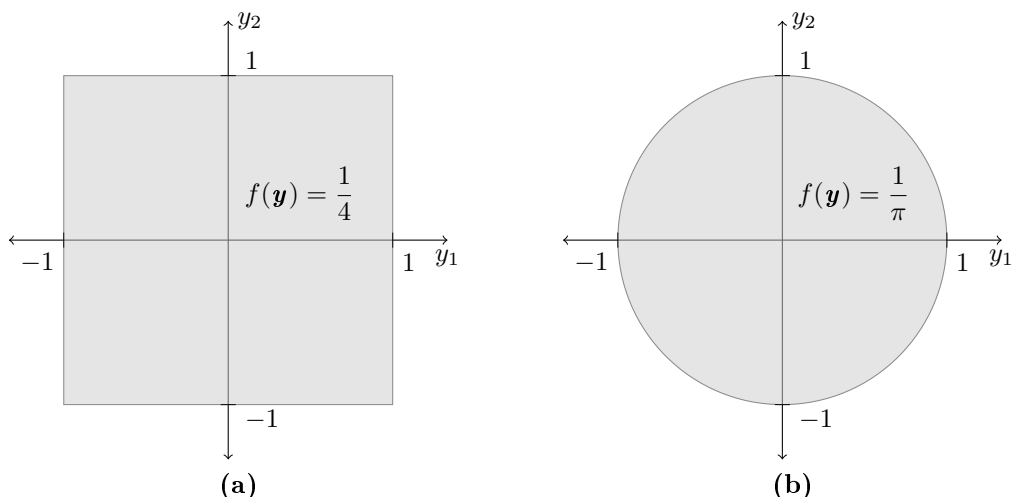
### 2.6.3 Nezávislost na směru a nezávislost souřadnic

Hustota (2.58) mnohorozměrného normálního rozdělení  $\mathbf{N}(\mathbf{0}; \mathbf{I}_n)$  je funkcí délky vektoru  $\mathbf{y}$ , nezávisí tedy na směru. Díky této vlastnosti můžeme nechat libovolně rotovat soustavu souřadnic (viz podkapitolu 2.6.1). Dalším důsledkem je skutečnost, že pravděpodobnost umístění realizace náhodného vektoru v oblasti vymezené (hyper)prostorovým úhlem je rovna relativní velikosti tohoto úhlu. Tuto důležitou vlastnost náležitě využijeme v následující kapitole při odvození hustoty rozdělení  $t$  a  $F$ .

Další charakteristikou vlastností rozdělení  $\mathbf{N}(\mathbf{0}; \mathbf{I}_n)$  je vzájemná nezávislost souřadnic vůči dané ortonormální bázi (viz podkapitolu 2.6.2). Ta má významný vztah k praxi, neboť odráží obvyklé uspořádání náhodných pokusů.

Domníváme se, že mnohorozměrné normální rozdělení je jediné rozdělení, které disponuje oběma těmito vlastnostmi zároveň. Nevíme ovšem o způsobu, jak by bylo možné tuto domněnku – je-li vůbec správná – dokázat; můžeme pouze ilustrovat obtíže, s jakými se lze setkat při pokusu o nalezení jiného rozdělení s těmito vlastnostmi (viz obr. 2.2).

<sup>14</sup>Poněkud hrubě řečeno je to důsledkem té skutečnosti, že rozdělení  $\mathbf{N}(\mathbf{0}; \mathbf{I}_n)$  má na jakémkoli řezu stále stejný „tvar“.



**Obrázek 2.2:** (a) Je-li hustota dvojrozměrného náhodného vektoru  $\mathbf{Y} \equiv (Y_1, Y_2)^T$  rovnoměrně rozdělena na ploše tvaru čtverce  $\langle -1; 1 \rangle \times \langle -1; 1 \rangle$ , jsou souřadnice  $Y_1, Y_2$  zřejmě nezávislé, avšak hustota závisí na směru. (b) Při rovnoměrném rozložení hustoty do oblasti tvaru kruhu se středem v počátku soustavy souřadnic je sice hustota nezávislá na směru, ale souřadnice nezávislé nejsou – fixujeme-li jednu z nich, ovlivníme tím množinu možných realizací druhé.

## 2.6.4 Rozdělení homomorfismu $H(\mathbf{Y})$

Nechť  $V_k$  je vektorový prostor dimenze  $k$ , kde  $k \leq n$ , a  $H$  je homomorfismus dimenze  $k$  zobrazující prostor  $V_n$  na prostor  $V_k$ . Zamysleme se nad hustotou náhodného vektoru  $\mathbf{X} \equiv H(\mathbf{Y})$ .

Pro libovolný vektor  $\mathbf{x} \in V_k$  tvoří množina všech realizací  $\mathbf{y} \in V_n$  náhodného vektoru  $\mathbf{Y}$ , vyhovujících podmínce  $\mathbf{x} = H(\mathbf{y})$ , lineární množinu  $H^{-1}(\mathbf{x}) \equiv \mathbf{y}_0 + J$ , kde  $\mathbf{y}_0$  je libovolný vektor splňující uvedenou podmínku a  $J$  je jádro homomorfismu  $H$ , tj. podprostor dimenze  $n - k$ .

Bez újmy na obecnosti můžeme předpokládat, že vektor  $\mathbf{y}_0$  je kolmý na podprostor  $J$ ; kdyby tomu totiž bylo jinak, stačilo by nahradit v definici množiny  $H^{-1}(\mathbf{x})$  tento vektor jeho pravoúhlým průmětem do podprostoru  $J^\perp$ . Tento vektor je určen jednoznačně.<sup>15</sup> Označme jeho délku  $r$ ; tato hodnota představuje vzdálenost množiny  $H^{-1}(\mathbf{x})$  od počátku soustavy souřadnic.

Nyní zaveďme ortonormální bázi prostoru  $V_n$  takovou, že vektory  $\mathbf{e}_1, \dots, \mathbf{e}_{n-k}$  generují podprostor  $J$  a vektor  $\mathbf{y}_0$  je  $r$ -násobkem vektoru  $\mathbf{e}_{n-k+1}$ . Všechny vektory  $\mathbf{y}$  ležící v množině  $H^{-1}(\mathbf{x})$  mají vzhledem k této bázi souřadnice

$$(\mathbf{y}_1, \dots, \mathbf{y}_{n-k}, r, 0, \dots, 0)^T,$$

<sup>15</sup>Nechť platí  $\mathbf{y}_1, \mathbf{y}_2 \in H^{-1}(\mathbf{x})$ , tj.  $\mathbf{y}_1 = \mathbf{y}_0 + \mathbf{u}_1$ ,  $\mathbf{y}_2 = \mathbf{y}_0 + \mathbf{u}_2$ , kde  $\mathbf{u}_1, \mathbf{u}_2 \in J$ . Označme symbolem  $Q$  ortogonální projekci do podprostoru  $J^\perp$ ; zřejmě platí

$$\begin{aligned} Q(\mathbf{y}_1) &= Q(\mathbf{y}_0) + Q(\mathbf{u}_1) = Q(\mathbf{y}_0), \\ Q(\mathbf{y}_2) &= Q(\mathbf{y}_0) + Q(\mathbf{u}_2) = Q(\mathbf{y}_0), \end{aligned}$$

což znamená, že průmět všech vektorů náležících množině  $H^{-1}(\mathbf{x})$  do podprostoru  $J^\perp$  je tentýž.

takže integrál z hustoty přes celou množinu  $H^{-1}(\mathbf{x})$  je roven hodnotě

$$\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} e^{-\frac{y_1^2 + \dots + y_{n-k}^2 + r^2}{2}} dy_1 \dots dy_{n-k} = e^{-\frac{r^2}{2}} \int_{\mathbf{y} \in J} e^{-\frac{\|\mathbf{y}\|^2}{2}} dV,$$

kde  $V$  je  $(n - k)$ -rozměrný objem na podprostoru  $J$ . Poslední integrál je ovšem konstanta závislá pouze na dimenzi tohoto podprostoru. Hustota náhodného vektoru  $\mathbf{X}$  je tedy dána vzorcem

$$g(\mathbf{x}) = Ce^{-\frac{r^2}{2}},$$

kde  $r$  je vzdálenost množiny  $H^{-1}(\mathbf{x})$  od počátku soustavy souřadnic (je to tedy funkce vektoru  $\mathbf{x}$ ) a  $C$  je konstanta určená podmínkou

$$C \int_{\mathbf{x} \in V_k} g(\mathbf{x}) dV = 1.$$

Tento výsledek využijeme v kapitole 3.2.

## 2.7 Odvození vybraných rozdělení

V této kapitole ukážeme, jak lze s využitím geometrického přístupu odvodit z mnohorozměrného normálního rozdělení hustoty rozdělení  $\chi^2$ ,  $t$  a  $F$ , která jsme v předchozím textu hojně používali. Kromě znalosti derivování a hustoty mnohorozměrného normálního rozdělení budeme potřebovat určit objem  $n$ -rozměrné sféry o poloměru  $r$ ; ten je určen vzorcem<sup>16</sup>

$$S_n(r) = \frac{2\pi^{\frac{n+1}{2}} r^n}{\Gamma\left(\frac{n+1}{2}\right)}.$$

Objem jednotkové  $n$ -rozměrné sféry budeme značit prostě  $S_n$ ; zřejmě platí  $S_n(r) = S_n r^n$ .

Všechny níže uvedené způsoby odvození jsou inspirovány postupy R. A. Fishera, naznačenými např. v publikacích [5], [12], [13] a [14] (podrobnosti viz kapitolu 3.2). Není nám však známo, že by byly v této konkrétní podobě někde explicitně uvedeny.

### 2.7.1 Rozdělení $\chi^2$

Nechť náhodný vektor  $\mathbf{Y} = (Y_1, \dots, Y_n)^T$  má rozdělení  $N(\mathbf{0}, \mathbf{I}_n)$ , jeho hustota v prostoru  $V_n$  je tedy dána vztahem

$$f(\mathbf{y}) = (2\pi)^{-\frac{n}{2}} \exp\left\{-\frac{\|\mathbf{y}\|^2}{2}\right\}. \quad (2.60)$$

<sup>16</sup> $n$ -rozměrnou sférou je míněn povrch  $(n+1)$ -rozměrné koule, tj. množina těch bodů v  $(n+1)$ -rozměrném eukleidovském prostoru, jejichž vzdálenost od daného středu je rovna konstantě. Ve shodě s běžně používanou terminologií bychom tedy veličinu  $S_n$  mohli nazývat povrchem  $(n+1)$ -rozměrné koule.

Podle definice (1.5) má náhodná veličina

$$\chi^2 \equiv \sum_{i=1}^n Y_i^2 = \|\mathbf{Y}\|^2$$

rozdělení  $\chi_n^2$ ; odvodme hustotu  $g(x)$  tohoto rozdělení.

Pro  $\chi^2 = x$ , kde  $x \in \mathbb{R}^+$ , tvoří množina možných realizací náhodného vektoru  $\mathbf{Y}$  povrch  $n$ -rozměrné koule se středem v počátku soustavy souřadnic a poloměrem  $r = \sqrt{x}$ . Jevu  $\chi^2 \in (x; x + dx)$ , kde  $dx \rightarrow \infty$ , tedy odpovídá množina obepínající tuto kouli v podobě „slupky“ o tloušťce

$$dr = \frac{dx}{2\sqrt{x}}.$$

Povrch koule je tvořen  $(n-1)$ -rozměrnou sférou, jejíž  $(n-1)$ -rozměrný objem je

$$S_{n-1}(r) = S_{n-1} r^{n-1} = S_{n-1} x^{\frac{n-1}{2}},$$

$n$ -rozměrný objem „slupky“ je tedy

$$dV = S_{n-1}(r) \cdot dr = \frac{S_{n-1} x^{\frac{n}{2}-1}}{2} dx.$$

Na této množině platí  $\|\mathbf{y}\|^2 = x$ , takže hustota  $f(\mathbf{y})$  je zde konstantní; pravděpodobnost jevu  $\chi^2 \in (x; x + dx)$  je tedy

$$dP = f(\mathbf{y}) \cdot dV = (2\pi)^{-\frac{n}{2}} e^{-\frac{x}{2}} \cdot \frac{S_{n-1} x^{\frac{n}{2}-1}}{2} dx.$$

Část před diferenciálem je hledaná hustota náhodné veličiny  $X$ ; po dosazení za  $S_{n-1}$  a drobných estetických úpravách dostáváme

$$g(x) = \frac{x^{\frac{n}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)}.$$

## 2.7.2 Rozdělení $t$

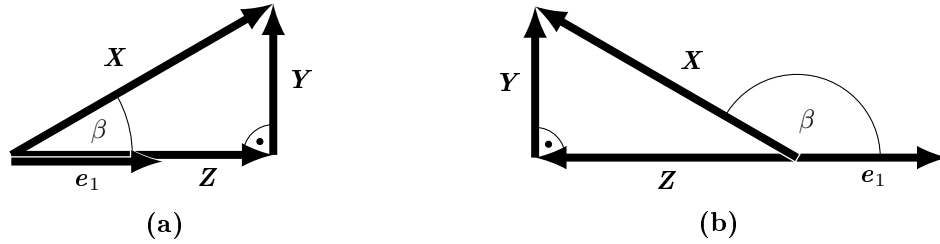
Nechť náhodný vektor  $\mathbf{X} = (Z_1, Y_1, \dots, Y_n)^T$  má  $(n+1)$ -rozměrné normální rozdělení. Položme

$$T \equiv \frac{Z_1}{\sqrt{\frac{\sum_{i=1}^n Y_i^2}{n}}}; \quad (2.61)$$

podle definice (1.6) má tato náhodná veličina rozdělení  $t_n$ . Odvodme její hustotu  $g(t)$ .

Nejprve si povšimněme, že až na případné znaménko představuje výraz

$$\frac{Z_1}{\sqrt{\sum_{i=1}^n Y_i^2}} \quad (2.62)$$



**Obrázek 2.3:** Vektory  $\mathbf{Z} = (Z_1, 0, \dots, 0)^T = Z_1 \mathbf{e}_1$  a  $\mathbf{Y} = (0, Y_1, \dots, Y_n)^T$  tvoří odvěsny pravoúhlého trojúhelníku, jehož přeponou je vektor  $\mathbf{X} = (Z_1, Y_1, \dots, Y_n)^T$ . Podíl  $\|\mathbf{Z}\|/\|\mathbf{Y}\| = |Z_1|/\|\mathbf{Y}\|$  je tedy kotangens úhlu, který svírají vektory  $\mathbf{X}$  a  $\mathbf{Z}$ . To znamená, že ať je  $Z_1 \geq 0$  (a) nebo  $Z_1 \leq 0$  (b), představuje podíl  $Z_1/\|\mathbf{Y}\|$  kotangens úhlu  $\beta$ , který svírá vektor  $\mathbf{X}$  s vektorem  $\mathbf{e}_1 = (1, 0, \dots, 0)^T$ .

podíl délek vektorů

$$\begin{aligned}\mathbf{Z} &\equiv (Z_1, 0, \dots, 0), \\ \mathbf{Y} &\equiv (0, Y_1, \dots, Y_n),\end{aligned}$$

tvořících odvěsny pravoúhlého trojúhelníku s přeponou  $\mathbf{X}$  (viz obr. 2.3). Jeho absolutní hodnotu tedy můžeme interpretovat jako kotangens úhlu, který svírají vektory  $\mathbf{X}$  a  $\mathbf{Z}$ . Vektor  $\mathbf{Z}$  je však rovnoběžný s vektorem  $\mathbf{e}_1 \equiv (1, 0, \dots, 0)^T$ , přičemž tyto dva vektory jsou souhlasně orientované právě tehdy, když  $Z_1 > 0$ . Podíl (2.62) tedy představuje kotangens úhlu, který vektor  $\mathbf{X}$  svírá s vektorem  $(1, 0, \dots, 0)^T$ ; označme jej  $\beta$  a můžeme psát

$$T = \sqrt{n} \cdot \cotg \beta.$$

Pro  $x \in \mathbb{R}$  je tedy množina realizací odpovídajících jevu  $T = x$  tvořena všemi vektory, které svírají s vektorem  $(1, 0, \dots, 0)^T$  úhel

$$\beta = \operatorname{arccotg} \frac{x}{\sqrt{n}}.$$

Přírůstku proměnné  $x$  o hodnotu  $dx$ , kde  $dx \rightarrow 0$ , pak odpovídá přírůstek úhlu  $\beta$  o hodnotu

$$d\beta = -\frac{dx}{\sqrt{n} \left(1 + \frac{x^2}{n}\right)}.$$

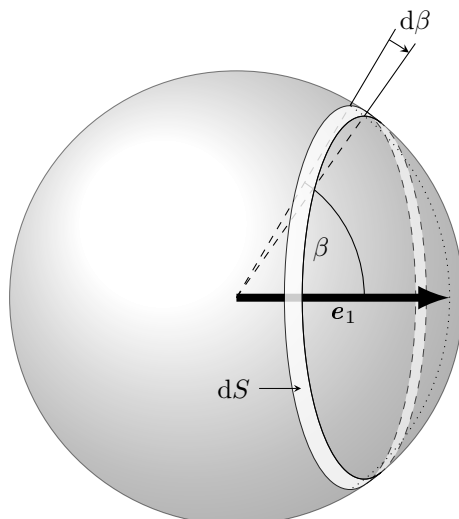
### Trojrozměrný případ

Abychom si trochu usnadnili další výklad, popíšme nejdříve případ pro  $n = 2$ , kdy realizacemi vektoru  $\mathbf{X}$  jsou prvky prostoru  $V_3$ . Uvažujme jednotkovou kouli se středem v počátku soustavy souřadnic. Vektory svírající s vektorem  $(1, 0, 0)^T$  úhel dané velikosti  $\beta$  protínají povrch této koule v kružnici o poloměru  $r = \sin \beta$ , jejíž délka je

$$l = 2\pi \sin \beta = S_1 \sin \beta.$$

Jevu  $T \in (x; x + dx)$  tedy odpovídá množina realizací, které na povrchu jednotkové koule ohraničují oblast o obsahu

$$dS = -l d\beta = -S_1 \sin \beta d\beta$$



**Obrázek 2.4:** Množina všech vektorů, které svírají s vektorem  $e_1$  úhel o velikosti ležící v intervalu  $(\beta; \beta + d\beta)$ , kde  $d\beta \rightarrow 0_-$ , protíná povrch jednotkové koule v pásu o délce  $l = 2\pi \sin \beta$ , šířce  $-d\beta$  a obsahu  $-l d\beta$ .

(viz obr. 2.4). A protože rozdělení náhodného vektoru  $\mathbf{X}$  nezávisí na směru, můžeme pravděpodobnost tohoto jevu vypočítat jako podíl obsahu této množiny vůči povrchu celé koule:

$$dP = \frac{dS}{S_2} = -\frac{S_1 \sin \beta}{S_2} d\beta.$$

Dosazením za  $S_1$ ,  $S_2$ ,  $\beta$  a  $d\beta$  dostáváme po jednoduché úpravě<sup>17</sup>

$$\begin{aligned} dP &= \frac{2\pi \sin\left(\operatorname{arccotg} \frac{x}{\sqrt{2}}\right)}{4\pi} \cdot \frac{dx}{\sqrt{2} \left(1 + \frac{x^2}{2}\right)} = \dots \\ \dots &= 2^{-\frac{3}{2}} \left(1 + \frac{x^2}{2}\right)^{-\frac{3}{2}} dx; \end{aligned}$$

část před diferenciálem je hustota rozdělení  $t_2$ .

### Zobecnění

V obecném případě  $n \in \mathbb{N}$  musíme povrch  $S_2$  jednotkové trojrozměrné koule nahradit objemem jednotkové  $n$ -rozměrné sféry a délku  $S_1 \sin \beta$  kružnice objemem

<sup>17</sup>Připomeňme vztahy

$$\begin{aligned} \sin(\operatorname{arccotg} t) &= \frac{1}{\sqrt{1+t^2}}, \\ \cos(\operatorname{arccotg} t) &= \frac{t}{\sqrt{1+t^2}}. \end{aligned}$$

$(n - 1)$ -rozměrné sféry o poloměru  $r = \sin \beta$ :

$$\begin{aligned} dP &= -\frac{S_{n-1} \sin^{n-1} \beta}{S_n} d\beta = \\ &= \frac{2\pi^{\frac{n}{2}} \sin^{n-1} \left( \operatorname{arccotg} \frac{x}{\sqrt{n}} \right)}{\frac{2\pi^{\frac{n+1}{2}}}{\Gamma\left(\frac{n+1}{2}\right)}} \cdot \frac{dx}{\sqrt{n} \left(1 + \frac{x^2}{n}\right)} = \dots \\ \dots &= \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right) \sqrt{n\pi}} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} dx. \end{aligned}$$

Část před diferenciálem je hledaná hustota rozdělení  $t_n$ .

### 2.7.3 Rozdělení F

Nechť náhodný vektor  $\mathbf{X} = (Z_1, \dots, Z_m, Y_1, \dots, Y_n)^T$  má  $(m + n)$ -rozměrné normální rozdělení. Odvoďme hustotu  $g(x)$  náhodné veličiny

$$F \equiv \frac{\sum_{i=1}^m Z_i^2 / m}{\sum_{i=1}^n Y_i^2 / n},$$

která má podle definice (1.8) rozdělení  $F_{m,n}$ .

Položme

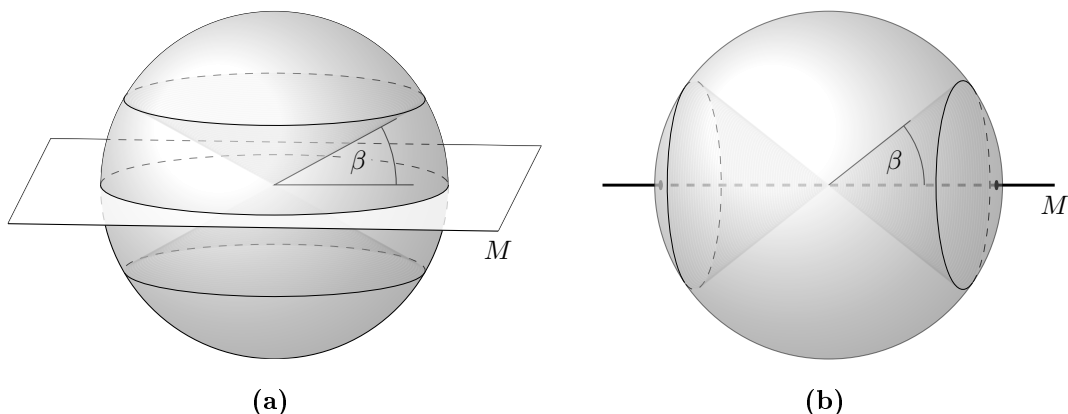
$$\begin{aligned} \mathbf{Z} &\equiv (Z_1, \dots, Z_m, 0, \dots, 0)^T, \\ \mathbf{Y} &\equiv (0, \dots, 0, Y_1, \dots, Y_n)^T. \end{aligned}$$

Tak jako v předchozím příkladu tvoří tyto vektory odvěsny pravoúhlého trojúhelníku, jehož přeponou je vektor  $\mathbf{X}$ , takže výraz

$$\frac{\sum_{i=1}^m Z_i^2}{\sum_{i=1}^n Y_i^2} = \frac{\|\mathbf{Z}\|^2}{\|\mathbf{Y}\|^2}$$

představuje druhou mocninu kotangenty úhlu  $\beta$ , který svírá vektor  $\mathbf{X}$  s vektorem  $\mathbf{Z}$ . Vektor  $\mathbf{Z}$  je však vlastně pravoúhlým průmětem vektoru  $\mathbf{X}$  do  $m$ -rozměrného podprostoru  $M \equiv [\mathbf{e}_1, \dots, \mathbf{e}_m]$ , kde  $\{\mathbf{e}_1, \dots, \mathbf{e}_m, \mathbf{e}_{m+1}, \dots, \mathbf{e}_{m+n}\}$  je aktuální báze prostoru  $V_{m+n}$ . To znamená, že úhel  $\beta$  představuje odchylku vektoru  $\mathbf{X}$  od podprostoru  $M$ . Pro  $x \in \mathbb{R}^+$  je tedy množina realizací odpovídajících jevu  $F = x$  tvořena všemi vektory, které s podprostorem  $A$  svírají úhel  $\beta$  splňující rovnost

$$x = \frac{n}{m} \cotg^2 \beta,$$



**Obrázek 2.5:** V prostoru  $V_3$  protíná množina všech vektorů svírajících s podprostorem  $M$  úhel dané velikosti  $\beta$  povrch jednotkové koule ve dvou kružnicích, jejichž poloměr je buď  $r = \cos \beta$ , je-li  $\dim M = 2$  (a), nebo  $r = \sin \beta$ , je-li  $\dim M = 1$  (b).

tj.

$$\beta = \operatorname{arccotg} \sqrt{\frac{mx}{n}}.$$

Jaká je míra průniku této množiny s jednotkovou  $(m+n-1)$ -rozměrnou sférou se středem v počátku soustavy souřadnic? Prozkoumejme nejprve jediné dva trojrozměrné případy (viz obr. 2.5). Je-li  $m=2$ ,  $n=1$ , jedná se o dvě kružnice (tj. jednorozměrné sféry) o poloměru  $r = \cos \beta$ , ležící v rovinách rovnoběžných s rovinou  $M$ . Jejich celková délka (tj. jednorozměrný objem) je tedy

$$l = 2 \cdot S_1(\cos \beta).$$

V případě  $m=1$ ,  $n=2$  se jedná také o dvě kružnice, ale tentokrát leží v rovinách kolmých na přímku  $M$ , jejich poloměr je  $r = \sin \beta$  a jejich celková délka je

$$l = 2 \cdot S_1(\sin \beta).$$

V obecném případě musíme tyto úvahy zkombinovat: hledaný průnik je vlastně kartézský součin  $(m-1)$ -rozměrné sféry o poloměru  $\cos \beta$ , která leží v podprostoru  $M$ , a  $(n-1)$ -rozměrné sféry o poloměru  $\sin \beta$ , ležící v podprostoru  $N = M^\perp$ . Jeho  $(m+n-2)$ -rozměrný objem je tedy

$$\begin{aligned} l &= S_{m-1}(\cos \beta) \cdot S_{n-1}(\sin \beta) = \\ &= S_{m-1} S_{n-1} \cdot \cos^{m-1} \beta \sin^{n-1} \beta. \end{aligned}$$

Všimněme si, že je  $S_0(\cos \beta) = S_0(\sin \beta) = 2$  (povrch jednorozměrné koule totiž tvoří dva body) – odtud konstanta 2 v obou trojrozměrných případech.

Dále můžeme postupovat podobně jako v předchozím příkladu: jevu  $F \in (x; x+dx)$  odpovídá množina realizací, které vymezují na jednotkové  $(m+n-1)$ -rozměrné sféře oblast o  $(m+n-1)$ -rozměrném objemu

$$\begin{aligned} dS &= -l d\beta = \\ &= -S_{m-1} S_{n-1} \cdot \cos^{m-1} \beta \sin^{n-1} \beta d\beta, \end{aligned}$$

kde

$$d\beta = -\frac{1}{2}\sqrt{\frac{m}{nx}}\left(1 + \frac{mx}{n}\right)^{-1} dx.$$

Vzhledem k nezávislosti rozdělení náhodného vektoru  $\mathbf{Y}$  na směru je pravděpodobnost tohoto jevu rovna poměru tohoto objemu vůči objemu celé sféry. Po dosazení dostáváme

$$\begin{aligned} dP &= \frac{dS}{S_{m+n-1}} = \\ &= \frac{S_{m-1}S_{n-1}\sqrt{\frac{m}{nx}}\left[\cos\left(\operatorname{arccotg}\sqrt{\frac{mx}{n}}\right)\right]^{m-1}\left[\sin\left(\operatorname{arccotg}\sqrt{\frac{mx}{n}}\right)\right]^{n-1}}{2S_{m+n-1}\cdot\left(1 + \frac{mx}{n}\right)} dx = \dots \\ &\dots = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}\left(\frac{m}{n}\right)^{\frac{m}{2}}x^{\frac{m}{2}-1}\left(1 + \frac{mx}{n}\right)^{-\frac{m+n}{2}} dx; \end{aligned}$$

část před diferenciálem je hledaná hustota  $g(x)$ .

Poznamenejme ještě, že k odvození hustoty rozdělení  $t$  a  $F$  jsme vlastně nepotřebovali předpoklad normality; stačila nezávislost rozdělení náhodného vektoru  $\mathbf{Y}$  na směru.

## 2.8 Gaussova-Markovova věta

### 2.8.1 Funkcionální a vektorová verze Gaussovy-Markovovy věty

Ústředním pojmem Gaussovy-Markovovy věty je *nejlepší nestranný lineární odhad*. Ten však v literatuře není definován zcela jednotně. My budeme za výchozí považovat tzv. „funkcionální definici“ tohoto pojmu, převzatou z publikace [31]: platí-li model (1.10) a  $T$  je homomorfismus  $T : V_n \rightarrow V_n$ , nazýváme náhodnou veličinu  $T(\mathbf{Y})$  nejlepším nestranným lineárním odhadem střední hodnoty  $\mathbf{E}\mathbf{Y} = \boldsymbol{\mu}$ , pokud pro všechna  $\mathbf{a} \in V_n$  a všechna  $\boldsymbol{\mu} \in M$  platí jednak

$$\mathbf{E}[\mathbf{a} \circ T(\mathbf{Y})] = \mathbf{a} \circ \boldsymbol{\mu}, \quad (2.63)$$

a dále, je-li pro nějaký vektor  $\mathbf{b} \in V_n$  splněna rovnost  $\mathbf{E}(\mathbf{b} \circ \mathbf{Y}) = \mathbf{a} \circ \boldsymbol{\mu}$ , platí

$$\operatorname{var}(\mathbf{b} \circ \mathbf{Y}) \geq \operatorname{var}[\mathbf{a} \circ T(\mathbf{Y})]. \quad (2.64)$$

To znamená, že pro jakoukoli lineární formu na prostoru  $V_n$  (vyjádřenou skalárním součinem s vektorem  $\mathbf{a}$ ) je náhodná veličina  $\mathbf{a} \circ T(\mathbf{Y})$  nestranným odhadem hodnoty  $\mathbf{a} \circ \boldsymbol{\mu}$  a přitom má ze všech možných nestranných lineárních odhadů této hodnoty nejmenší možný rozptyl.

Tzv. „vektorová definice“, uvedená např. v [35], je mírně odlišná: náhodná veličina  $T(\mathbf{Y})$  se zde nazývá nejlepším nestranným lineárním odhadem střední hodnoty  $\boldsymbol{\mu}$ , pokud platí jednak

$$\mathbf{E}T(\mathbf{Y}) = \boldsymbol{\mu}, \quad (2.65)$$

a dále, splňuje-li nějaký homomorfismus<sup>18</sup>  $B$  podmínku  $EB(\mathbf{Y}) = \boldsymbol{\mu}$ , platí pro jakýkoli vektor  $\mathbf{a} \in V_n$  nerovnost

$$\text{var} [\mathbf{a} \circ B(\mathbf{Y})] \geq \text{var} [\mathbf{a} \circ T(\mathbf{Y})]. \quad (2.66)$$

## 2.8.2 Ekvivalence obou definic

Na první pohled není snadné posoudit, do jaké míry je odlišnost obou definic podstatná. Vyřešme tedy tuto otázku a ukažme, že jsou ve skutečnosti ekvivalentní. Nemusíme se nicméně zabývat podmínkami (2.63) a (2.65), neboť ty jsou ekvivalentní z definice střední hodnoty (2.50); stačí se zaměřit pouze na podmínky (2.64) a (2.66).

Předpokládejme tedy nejprve, že náhodná veličina je  $T(\mathbf{Y})$  je nejlepším nestranným lineárním odhadem střední hodnoty  $\boldsymbol{\mu}$  podle funkcionální definice. Nechť  $B$  je homomorfismus splňující podmínku  $EB(\mathbf{Y}) = \boldsymbol{\mu}$ . Pro libovolný vektor  $\mathbf{a} \in V_n$  pak platí

$$E[B'(\mathbf{a}) \circ \mathbf{Y}] = E[\mathbf{a} \circ B(\mathbf{Y})] = \mathbf{a} \circ \boldsymbol{\mu},$$

kde  $B'$  je homomorfismus adjungovaný k homomorfismu  $B$ . To znamená, že náhodná veličina  $B'(\mathbf{a}) \circ \mathbf{Y}$  je nestranným lineárním odhadem hodnoty  $\mathbf{a} \circ \boldsymbol{\mu}$ ; podle předpokladu (2.64), kde položíme  $\mathbf{b} = B'(\mathbf{a})$ , tedy platí

$$\text{var} [B'(\mathbf{a}) \circ \mathbf{Y}] \geq \text{var} [\mathbf{a} \circ T(\mathbf{Y})],$$

což je ekvivalentní s nerovností (2.66). Náhodná veličina  $T(\mathbf{Y})$  tudíž splňuje požadavky vektorové definice.

Nyní naopak předpokládejme, že  $T(\mathbf{Y})$  je nejlepším nestranným lineárním odhadem střední hodnoty  $\boldsymbol{\mu}$  podle vektorové definice. Nechť pro nějaké dva vektory  $\mathbf{a}, \mathbf{b} \in M$  je náhodná veličina  $\mathbf{b} \circ \mathbf{Y}$  nestranným odhadem hodnoty  $\mathbf{a} \circ \boldsymbol{\mu}$  pro všechna  $\boldsymbol{\mu} \in M$ , tj.

$$E(\mathbf{b} \circ \mathbf{Y}) = \mathbf{a} \circ \boldsymbol{\mu}; \quad (2.67)$$

potřebujeme ukázat, že platí nerovnost (2.64).

Nejprve si povšimněme, že z předpokladu (2.67) plyne  $\mathbf{b} \circ \boldsymbol{\mu} = \mathbf{a} \circ \boldsymbol{\mu}$ , tj.  $(\mathbf{a} - \mathbf{b}) \circ \boldsymbol{\mu}$ ; má-li tato rovnost platit pro všechna  $\boldsymbol{\mu} \in M$ , znamená to, že vektor  $\mathbf{a} - \mathbf{b}$  musí být kolmý na podprostor  $M$ . Z toho dále plyne rovnost  $P_M(\mathbf{a}) = P_M(\mathbf{b})$ , kde  $P_M$  je ortogonální projekce do podprostoru  $M$ .

Zvolme dále libovolný homomorfismus  $A$  tak, aby pro jeho adjungovaný homomorfismus  $A'$  platilo  $A'(\mathbf{a}) = \mathbf{b}$ , a položme

$$B \equiv P_M + AQ_M,$$

kde  $Q_M \equiv I - P_M$ . Protože obě zobrazení  $P_M, Q_M$  jsou ortogonální projekce, jsou samoadjungované, takže homomorfismus adjungovaný k zobrazení  $B$  je

$$B' = P_M + Q_M A';$$

<sup>18</sup>Značení jsme mírně přizpůsobili našim potřebám a matice jsme nahradili homomorfismy. V originále je navíc místo homomorfismu  $B$  uvedeno obecnější lineární zobrazení ve tvaru  $\mathbf{c} + B(\mathbf{Y})$ , kde  $\mathbf{c} \in V_n$ , rychle se však ukáže, že musí platit  $\mathbf{c} = \mathbf{0}$ .

platí totiž vztahy  $(A + B)' = A' + B'$ ,  $(AB)' = B'A'$  (viz [4]). Nyní můžeme nahlédnout, že jednak podle (2.51) platí

$$\mathbb{E}B(\mathbf{Y}) = B(\boldsymbol{\mu}) = P_M(\boldsymbol{\mu}) + AQ_M(\boldsymbol{\mu}) = \boldsymbol{\mu},$$

takže náhodná veličina  $B(\mathbf{Y})$  je nestranným odhadem střední hodnoty  $\boldsymbol{\mu}$  a podle našeho předpokladu (2.66) je splněna nerovnost

$$\text{var} [\mathbf{a} \circ B(\mathbf{Y})] \geq \text{var} [\mathbf{a} \circ T(\mathbf{Y})],$$

a jednak můžeme psát

$$\begin{aligned} \mathbf{a} \circ B(\mathbf{Y}) &= B'(\mathbf{a}) \circ \mathbf{Y} = \\ &= [P_M(\mathbf{a}) + Q_M A'(\mathbf{a})] \circ \mathbf{Y} = \\ &= [P_M(\mathbf{a}) + Q_M(\mathbf{b})] \circ \mathbf{Y} = \\ &= [P_M(\mathbf{b}) + Q_M(\mathbf{b})] \circ \mathbf{Y} = \\ &= \mathbf{b} \circ \mathbf{Y}. \end{aligned}$$

Tím je nerovnost (2.64) dokázána.

Obě definice jsou tedy vskutku ekvivalentní; stojí však za upozornění, že odvodit funkcionální verzi z vektorové je podle všeho podstatně náročnější než naopak. Funkcionální verze se tím pádem může zdát praktičtější. Je ovšem možné, že existuje jednodušší způsob odvození, který nám zůstal utajen.

### 2.8.3 Důkaz Gaussovy-Markovovy věty

Nyní již můžeme přistoupit k důkazu Gaussovy-Markovovy věty: v lineárním modelu (1.10) je pravoúhlý průmět  $P_M(\mathbf{Y})$  náhodného vektoru  $\mathbf{Y}$  do podprostoru  $M$  nejlepším nestranným lineárním odhadem střední hodnoty  $\mathbb{E}\mathbf{Y} = \boldsymbol{\mu}$ . Dokažme tuto skutečnost užitím funkcionální definice. Rovnost (2.63) je ekvivalentní s rovností (2.65), která plyne snadno z tvrzení (2.51), (2.35) a z toho, že  $\boldsymbol{\mu} \in M$ :

$$\mathbb{E}P_M(\mathbf{Y}) = P_M(\boldsymbol{\mu}) = \boldsymbol{\mu};$$

tím je dokázána nestrannost. Co se týče druhé podmínky, nejdříve připomeňme, že z rovnosti  $\mathbb{E}(\mathbf{b} \circ \mathbf{Y}) = \mathbf{a} \circ \boldsymbol{\mu}$  platné pro všechna  $\boldsymbol{\mu} \in M$  plyne, že pravoúhlé průměty  $P_M(\mathbf{a})$  a  $P_M(\mathbf{b})$  jsou totožné. Můžeme tedy psát

$$\begin{aligned} \text{var}(\mathbf{b} \circ \mathbf{Y}) &= \|\mathbf{b}\|^2 \sigma^2 = \\ &\geq \|P_M(\mathbf{b})\|^2 \sigma^2 = \\ &= \|P_M(\mathbf{a})\|^2 \sigma^2 = \\ &= \text{var} [P_M(\mathbf{a}) \circ \mathbf{Y}] = \\ &= \text{var} [\mathbf{a} \circ P_M(\mathbf{Y})]. \end{aligned}$$

První a předposlední rovnost platí díky předpokladu, že varianční matice je  $\sigma^2$ -násobkem matice jednotkové, který ve „free-coordinate“ přístupu odpovídá požadavku, aby varianční operátor  $\sum_{\mathbf{Y}}$  byl  $\sigma^2$ -násobkem identity na prostoru  $V_n$  – viz předcházející kapitolu a vzorec (2.55). Poslední rovnost plyne ze samoadjungovanosti ortogonální projekce, nerovnost vyplývá z tvrzení (2.37).

## 2.8.4 Důsledky Gaussovy-Markovovy věty a její zobecnění

Z Gaussovy-Markovovy věty podle funkcionální definice plyne, že jakoukoli lineární funkci střední hodnoty  $\boldsymbol{\mu}$  je nejlépe odhadnout pomocí odpovídající lineární funkce pravoúhlého průmětu  $P_M(\mathbf{Y})$ . V první řadě se jedná o souřadnice střední hodnoty  $\boldsymbol{\mu}$ , které odhadujeme pomocí souřadnic pravoúhlého průmětu  $P_M(\mathbf{Y})$  – ať již se jedná o souřadnice  $(y_1, \dots, y_n)^T$  vzhledem k původní bázi prostoru  $V_n$ , či o souřadnice  $(\beta_1, \dots, \beta_m)^T$  vzhledem k bázi podprostoru  $M$  představované sloupci matice  $\mathbf{X}$  v modelu (1.10), pokud jsou tyto lineárně nezávislé. V řadě druhé se totéž týká i jakékoli další lineární funkce těchto souřadnic. Poznamenejme, že jakákoli lineární funkce střední hodnoty  $\boldsymbol{\mu}$  se nazývá *odhadnutelný parametr*.

Dodejme ještě, že obvykle je Gaussova-Markovova věta uváděna v obecnější podobě, kdy je varianční operátor libovolný (zpravidla je požadována regularita). Tuto situaci lze předefinováním skalárního součinu převést na předcházející případ (viz kapitola 2.5.4); se změnou skalárního součinu se ovšem změní i pojem ortogonální projekce.

## 2.9 Lineární vazba regresních koeficientů – obecné poznámky

V kapitole 1.11 jsme viděli, že je-li model neúplný, lze nejednoznačnost koeficientů  $\beta_i$ , resp. jejich odhadů  $b_i$ , vyřešit přidáním vhodných lineárních podmínek svazujících tyto koeficienty. Je však zřejmé, že pokud je model úplný a přidáme k němu podmínky tohoto druhu, docílíme tím úplně jiného efektu – omezíme původní podprostor  $M$  daný modelem na nějaký jeho podprostor či lineární podmnožinu, tj. vlastně definujeme submodel (viz kapitoly 1.6 a 1.7). Prozkoumejme podrobněji, co je příčinou rozdílnosti těchto dvou situací.

Uvažujme model (1.10); hodnost matice  $\mathbf{X}$  označíme  $k$  ( $k \leq m$ ); připouštíme tedy i případ neúplného modelu, tj. situaci, kdy jsou sloupce matice  $\mathbf{X}$  lineárně závislé. Za těchto předpokladů platí:

- Dimenze podprostoru možných středních hodnot  $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ , které model připouští, je rovna hodnotě  $k$ .
- Prostor, ze kterého volíme vektor parametrů  $\boldsymbol{\beta}$ , je  $m$ -rozměrný.
- Je-li  $m > k$ , je při známé střední hodnotě  $\boldsymbol{\mu}$  třeba zvolit  $m - k$  parametrů  $\beta_i$ , ostatní jsou touto volbou jednoznačně určeny.

Nyní přidejme k modelu vazební podmínky. Ty mají zpravidla v maticovém zápisu podobu

$$\mathbf{T}\boldsymbol{\beta} = \mathbf{c}, \quad (2.68)$$

kde  $\mathbf{T}$  je matice typu  $t \times m$  a  $\mathbf{c} \in \mathbb{R}^t$  je uspořádaná  $t$ -tice reálných čísel (psána jako sloupec). Předpokládejme, že tato soustava podmínek je splnitelná, tj. existuje  $\boldsymbol{\beta}_0 \in \mathbb{R}^m$  takové, že platí  $\mathbf{T}\boldsymbol{\beta}_0 = \mathbf{c}$ , a dále předpokládejme, že hodnost matice  $\mathbf{T}$

je  $t$ ; to znamená, že je  $t \leq m$  a žádná z rovnic soustavy není nadbytečná. Tato soustava omezuje původní podprostor přijatelných středních hodnot

$$M = \{\mathbf{X}\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^m\}$$

daný modelem (1.10) na nějakou jeho lineární podmnožinu

$$\begin{aligned} S &= \{\mathbf{X}\boldsymbol{\beta} : \mathbf{T}\boldsymbol{\beta} = \mathbf{c}\} = \\ &= \mathbf{X}\boldsymbol{\beta}_0 + \{\mathbf{X}\boldsymbol{\beta} : \mathbf{T}\boldsymbol{\beta} = \mathbf{0}\}. \end{aligned}$$

Je třeba zjistit, jaká je dimenze této množiny, tj. podprostoru  $\{\mathbf{X}\boldsymbol{\beta} : \mathbf{T}\boldsymbol{\beta} = \mathbf{0}\}$ . Definujme tedy na množině určené podmínkou  $\mathbf{T}\boldsymbol{\beta} = \mathbf{0}$  zobrazení

$$X^* : \boldsymbol{\beta} \rightarrow \mathbf{X}\boldsymbol{\beta}.$$

Dimenze definičního oboru tohoto zobrazení je  $m - t$ . Jeho jádrem je množina

$$\begin{aligned} \text{Ker } X^* &= \{\boldsymbol{\beta} \in \mathbb{R}^m : \mathbf{T}\boldsymbol{\beta} = \mathbf{0} \wedge \mathbf{X}\boldsymbol{\beta} = \mathbf{0}\} = \\ &= \left\{ \boldsymbol{\beta} \in \mathbb{R}^m : \begin{pmatrix} \mathbf{X} \\ \mathbf{T} \end{pmatrix} \boldsymbol{\beta} = \mathbf{0} \right\}, \end{aligned}$$

jejíž dimenze je rovna hodnotě  $m - d$ , kde  $d$  je hodnota matice

$$\mathbf{D} \equiv \begin{pmatrix} \mathbf{X} \\ \mathbf{T} \end{pmatrix}.$$

Můžeme tedy určit i dimenzi obrazu homomorfismu  $X^*$ , tj. množiny  $S$ :

$$\begin{aligned} \dim(\text{Im } X^*) &= (m - t) - (m - d) = \\ &= d - t. \end{aligned}$$

Soustava podmínek (2.68) tedy spolu s modelem (1.10) vymezuje pro polohu střední hodnoty  $\boldsymbol{\mu}$  lineární množinu dimenze  $d - t$ . Tato dimenze se liší od dimenze  $k$  původního podprostoru  $M$  v závislosti na hodnotě  $d$ , tj. na tom, do jaké míry jsou řádky matice  $\mathbf{T}$  lineární kombinací řádků matice  $\mathbf{X}$ . Z praktického hlediska jsou typické jsou dva následující případy.

Buďto je každý řádek matice  $\mathbf{T}$  nějakou lineární kombinací matice  $\mathbf{X}$ ; to znamená, že hodnota matice  $\mathbf{D}$  je stejná jako hodnota matice  $\mathbf{X}$ , tj.  $d = k$ . Potom platí:

- Dimenze podprostoru možných středních hodnot  $\boldsymbol{\mu}$ , které model a podmínky (2.68) připouštějí, je rovna hodnotě  $d - t = k - t$ , je tedy o  $t$  menší než dimenze původního podprostoru  $M$ . Podmínky tudíž spolu s původním modelem formulují submodel.
- Lineární množina  $J'$ , ze které volíme vektor parametrů  $\boldsymbol{\beta}$ , je  $(m - t)$ -rozměrná.
- Při známé střední hodnotě  $\boldsymbol{\mu} \in M$  vyhovující podmínkám (2.68) je tedy třeba zvolit  $(m - t) - (k - t) = m - k$  parametrů  $\beta_i$ ; to je stejný počet jako u původního modelu bez podmínek.

Opačnou alternativou je případ, kdy jsou všechny řádky matice  $\mathbf{T}$  nezávislé na řádcích matice  $\mathbf{X}$ ; to znamená, že hodnost matice  $\mathbf{D}$  je rovna součtu hodností matic  $\mathbf{X}$  a  $\mathbf{T}$ , tj.  $d = k + t$ . Potom platí:

- Dimenze podprostoru možných středních hodnot  $\boldsymbol{\mu}$ , které model a podmínky (2.68) připouštějí, je rovna hodnotě  $d - t = k$ , jedná se tedy o tentýž podprostor, jako je podprostor  $M$  daný modelem.
- Lineární množina  $J'$ , ze které volíme vektor parametrů  $\boldsymbol{\beta}$ , je  $(m - t)$ -rozměrná.
- Při známé střední hodnotě  $\boldsymbol{\mu} \in M$  je tedy potřeba zvolit  $m - t - k$  parametrů  $\beta_i$ , ostatní jsou touto volbou jednoznačně určeny. Vhodné je samozřejmě volit  $t = m - k$ .

Zhruba řečeno se jedná o to, že v prvním případě podmínka (2.68) neovlivňuje dimenzi jádra původního homomorfismu  $\boldsymbol{\beta} \rightarrow \mathbf{X}\boldsymbol{\beta}$ , kde  $\boldsymbol{\beta} \in \mathbb{R}^m$ ; snižuje tudíž dimenzi jeho obrazu, a to o hodnotu  $t$ . V druhém případě naopak o tuto hodnotu snižuje dimenzi jádra, takže dimenze obrazu zůstává nezměněna.

Chceme-li tedy pomocí podmínek (2.68) definovat submodel, jehož dimenze je o hodnotu  $t$  menší než dimenze původního modelu, musí být matice  $\mathbf{T}$  tvořena  $t$  řádky závislými na řádcích matice  $\mathbf{X}$ . Pokud je naším cílem dosáhnout jednoznačnosti parametrů  $\beta_i$  v situaci, kdy je hodnost  $k$  matice  $\mathbf{X}$  menší než počet parametrů  $m$ , je třeba volit matici  $\mathbf{T}$  tak, aby měla  $m - k$  řádků lineárně nezávislých na řádcích matice  $\mathbf{X}$ .

## 3. Historická část

### 3.1 Počátky moderní matematické statistiky

Počátek 20. století byl svědkem bouřlivého rozvoje statistiky. V této době byla například odvozena metoda maximální věrohodnosti, analýza rozptylu, byly odvozeny hustoty rozdělení označovaných dnes jako  $t$  a  $F$  a jejich použití ve statistických testech získalo dnešní podobu. Stěžejní postavou tohoto vývoje byl britský matematik a biolog Ronald Aylmer Fisher, který je dnes oprávněně považován za tvůrce moderní statistiky. Podstatným zdrojem inspirace mu ovšem byly práce jiného britského autora, Williama Sealy Gosseta, publikujícího pod pseudonymem „Student“. Ty byly podnětem ke korespondenci, která vyústila v dlouhodobou spolupráci a přátelství obou mužů.

#### 3.1.1 William Sealy Gosset (1876 – 1937)

W. S. Gosset se narodil v roce 1876 jako první z pěti dětí. Vystudoval matematiku a chemii v Oxfordu a v roce 1899 nastoupil na místo sládka v pivovaru Guinness v Dublinu, kde zůstal po zbytek života. Management firmy v té době najal řadu mladých absolventů z Cambridge a Oxfordu s úmyslem zavést do výroby vědecké metody (viz [24]). Zaměstnanci pivovaru obecně neměli dovoleno publikovat své výsledky; Gossetovi však byla povolena výjimka pod podmínkou, že (kvůli utajení před ostatními zaměstnanci) bude publikovat pod pseudonymem (viz [32]). Byla zvolena přezdívka „Student“, pod kterou Gosset publikoval většinu ze svých 21 článků. Jedním z prvních, kterým se ovšem nesmazatelně zapsal do dějin matematiky, bylo pojednání *The Probable Error of a Mean* [28]. Mnoho zdrojů zdůrazňuje Gossetův vynikající charakter, pro který byl respektován svými současníky; o jeho skromné povaze, poctivosti a smyslu pro humor svědčí i dochovaná korespondence (viz [5], [6], [22] a [24]). W. S. Gosset zemřel v roce 1937 ve věku 61 let.

#### 3.1.2 Ronald Aylmer Fisher (1890 – 1962)

R. A. Fisher se narodil v roce 1890 jako nejmladší z osmi dětí. Jeho příchod byl překvapením – matka totiž čtvrt hodiny před ním porodila jeho mrtvého sourozence a další dítě nikdo neočekával. Fisher od dětství projevoval známky matematického nadání: při jedné příležitosti například opravil svou matku, že mu nejsou 3 roky, nýbrž 3 roky, 4 měsíce a 5 dní. Měl vynikající geometrickou představivost, údajně vytříbenou i díky tomu, že byl kvůli svému slabému zraku nucen studovat sférickou trigonometrii bez pomoci papíru a tužky (viz [5]). Vystudoval matematiku v Cambridge, okruh jeho znalostí však byl mnohem širší – mimo statistiku proslul především jako biolog. Jeho bibliografie zahrnuje 4 knihy a téměř 300 článků (matematických i biologických); mnoho z nich lze považovat za zcela zásadní pro rozvoj statistiky. Nejznámějším titulem je pravděpodobně kniha *Statistical Methods for Research Workers* [16]. Jeho dcera jej líčí jako zásadového idealistu, oddaného svým přátelům, ale nesmiřitelného vůči těm, kdo ho (byť jen domněle) zradili (viz [5]). To se zřejmě projevilo v jeho dlouholetém konfliktu

s Karlem Pearsonem. R. A. Fisher zemřel v Adelaide v roce 1962 ve věku 72 let.

### 3.1.3 Gossetův článek *The Probable Error of a Mean* [28]

Gosset při své práci v pivovaru často pracoval s náhodnými výběry malého rozsahu, na jejichž zpracování však tehdejší statistická teorie nenabízela žádný aparát. Dosavadní praxe se zabývala pouze výběry velkého rozsahu, u nichž se výběrová směrodatná odchylka  $s$  dala považovat za dostatečně přesný odhad směrodatné odchylky  $\sigma$ . Bylo tedy legitimní předpokládat, že rozptyl průměru  $\bar{x}$  z výběru z normálního rozdělení  $N(\mu, \sigma^2)$  o rozsahu  $n$  je  $s^2/n$ , a tak testovat hypotézy o střední hodnotě  $\mu$ . Gosset ovšem nechtěl ignorovat, že při malém rozsahu výběru je hodnota  $s$  zatížena chybou, která spolehlivost takových testů podstatně zkresluje. Ve své nejslavnější práci [28] se proto zaměřil na rozdělení náhodné veličiny

$$z = \frac{\bar{x}}{s},$$

kde

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}, \quad s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}}$$

a  $x_1, \dots, x_n$  je výběr z normálního rozdělení  $N(0, \sigma^2)$  ( $z$  tedy představuje vzdálenost výběrového průměru  $\bar{x}$  od populačního průměru měřenou v jednotkách  $s$ ).

V článku je nejprve odvozena hustota rozdělení náhodných veličin  $s^2$  a  $s$ , poté je dokázáno, že  $\bar{x}$  a  $s$  jsou nekorelované, a pak je odvozena hustota rozdělení náhodné veličiny  $z$ . Následuje podrobný rozbor vlastností získaných rozdělení. V další části Gosset porovnává získané výsledky s výsledky experimentu, který podle svých slov provedl ještě dříve, než přistoupil k výpočtům. (Na str. 13 v práci [28] píše: „Before I had succeeded in solving my problem analytically, I had endeavoured to do so empirically.“) Experiment spočíval v tom, že Gosset na 3000 kartiček přepsal délky levého prostředníčku a výšky 3000 trestanců. Kartičky pak promíchal a rozdělil na 750 čtveřic. Tak získal  $2 \times 750$  náhodných výběrů ze dvou rozdělení o známých parametrech (vypočtených z původního souboru 3000 měření). Pro každou čtveřici vypočítal hodnotu  $z$  podle výše uvedeného vztahu (od  $x_i$  ovšem odečetl hodnotu populačního průměru) a získal tak dvě empirické frekvenční křivky, které v článku porovnává s teoretickými hustotami  $z$ ; shodu shledává nanejvýš uspokojivou. Jak uvádí Zabell v pojednání [32], použití takovéto simulace bylo v té době velice neobvyklé. Práci uzavírá tabulka s vybranými hodnotami distribuční funkce  $z$  pro rozsah výběru  $n = 4$  až 10 a čtyři příklady z praxe, ilustrující použití jeho „ $z$ -testu“ při ověření hypotézy o střední hodnotě (tj. ekvivalence dnešního jednovýběrového  $t$ -testu).

Fisher v práci [15] upozorňuje na dva nedostatky textu. Gosset jednak rozdělení  $s^2$  vlastně pouze odhaduje pomocí prvních čtyř centrálních momentů a tzv. Pearsonova systému frekvenčních křivek, jednak místo nezávislosti  $\bar{x}$  a  $s^2$ , která je potřebná k odvození rozdělení  $z$ , dokazuje pouze jejich nekorelovanost. (První chyby si však byl Gosset vědom, sám ji v článku připouští.) Mimoto jsou v jednom z praktických příkladů chybně uvedeny názvy testovaných léků<sup>1</sup> a data použitá

<sup>1</sup>Fisher tento příklad převzal do své knihy [16], aniž si původ dat ověřil. V roce 1934 si

v tomtéž příkladu představují průměry z různých počtů původních měření; nepocházejí tedy ze stejných rozdělení, a tudíž nejsou pro daný test použitelná (viz [32]).

### 3.1.4 Fisherovo „Studentovo“ rozdělení

Gossetův článek zpočátku v akademických kruzích nevyvolal téměř žádnou odezvu (viz [32]); jedinou výjimkou byl R. A. Fisher. Počátečním impulsem ke korespondenci mezi ním a Gossetem byla v roce 1912 podle [5] neshoda ohledně správného jmenovatele ve vzorci pro směrodatnou odchylku.<sup>2</sup> Fishera však zjevně zaujaly Gossetovy výsledky, neboť mu už ve svém třetím dopise poslal důkaz jeho vzorců pro rozdělení  $z$ , uvedených v [28]. Tento dopis se nezachoval, ale z korespondence Gosseta s Karlem Pearsonem je zřejmé, že důkaz byl proveden pomocí geometrické reprezentace náhodného výběru v  $n$ -rozměrném prostoru; tento přístup mu později umožnil dosáhnout mnoha dalších pozoruhodných výsledků, viz [5]. Gosset přeposlal důkaz Karlu Pearsonovi, na kterého však tento neučinil žádný dojem: „I do not follow Mr Fisher’s proof & it is not the kind of proof which appeals to me. (...) I do not see what the writer is doing at all. (...) Of course, if Mr Fisher will write a proof, in which each line flows from the preceding one & define his terms I will gladly consider its publication.“ ([24], str. 47–48). Fisher publikoval geometrický důkaz rozdělení  $z$  až v roce 1920 v článku [13].

Korespondence pokračovala v roce 1915, kdy Fisher publikoval článek [12] o rozdělení korelačního koeficientu (viz [6]). I v tomto případě byl inspirován Gossetovými výsledky, konkrétně článkem [29] z roku 1908, a i zde uplatnil svůj vynikající geometrický vhled. Jedním z dílčích výsledků bylo tvrzení, že je-li populační korelační koeficient  $\rho$  roven nule, má poměr

$$\frac{r}{\sqrt{1-r^2}}$$

stejně rozdělení jako Gossetova veličina  $z$  v případě výběru o rozsahu  $n - 1$ .

V roce 1922 Gosset v dopise Fisherovi nadhodil problém rozdělení regresních koeficientů; Fisher mu obratem zaslal řešení, ve kterém se (k nemalé Gossetově

---

chyby v jeho knize povšiml Dr. Isidor Greenwald a napsal mu dopis, který cituje E. S. Pearson v publikaci [24]. V pozdějších vydáních knihy [16] již jsou léky označeny pouze jako A a B.

<sup>2</sup>Tato záležitost je poněkud nejasná. Gosset ve svém dopise Karlu Pearsonovi z 12. září 1912 citovaném v knize [5] líčí, že mu Fisher poslal důkaz, že správný vzorec pro směrodatnou odchylku je

$$\sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}} \quad \text{a nikoli} \quad \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{(n-1)}}.$$

Fisher ve svém prvním článku [11] skutečně pomocí metody maximální věrohodnosti odvozuje vzorec pro odhad  $\sigma^2$ , kde je ve jmenovateli  $n$ . Jak je však patrné z výše uvedeného textu, Gosset v [28] používá rovněž  $n$ , a není tedy jasné, proč mu Fisher své výsledky posílal. Jedině snad proto, že Gosset mj. uvádí, že střední hodnota výrazu

$$\sum_{i=1}^n (x_i - \bar{x})^2 / n$$

je rovna  $(n-1)\sigma^2/n$  (což je ovšem správně). Dopis navíc pokračuje zmínkou o dalším Fisherově dopise, ve kterém ukázal, že správný jmenovatel je nakonec přece jen  $n - 1$ .

radosti) ukázalo, že i zde je třeba použít jeho  $z$ -rozdělení. V souvislosti s těmito objevy zřejmě vykrystalizoval Fisherův komplexní pohled na celou problematiku, zahrnující kromě předešlého též testy rozdílu dvou průměrů a testy korelačních koeficientů pomocí „Studentova“ rozdělení (viz [10]). Ten byl definitivně shrnut v práci [15], kde však Fisher s ohledem na čtenáře postrádající jeho geometrickou představivost použil k odvozování algebraický přístup (viz [5]). Celá metodika byla zpopularizována také díky Fisherově knize [16], kde již bylo „Studentovo“ rozdělení uvedeno v dnešní podobě, tj. po transformaci

$$t_{n-1} = z_n \sqrt{n-1}.$$

## 3.2 Geometrie v díle R. A. Fishera

V následujícím textu ukážeme na několika příkladech, jakým způsobem R. A. Fisher využíval ve svých statistických pracích svého geometrického vhledu. Je třeba upozornit, že si neděláme nárok na úplnost tohoto přehledu; Fisherovo dílo je značně rozsáhlé a dle našeho názoru i poměrně obtížné. Co se týče samotné geometrie, uvádí své úvahy velice zkratkovitě, často jednou větou, a jejich důkazy se zpravidla vůbec nezabývá. Pro čtenáře, který postrádá jeho sebejistotu v používání  $n$ -rozměrné geometrie a potřebuje si všechny úvahy tohoto druhu formálně ověřit, jsou Fisherovy články nepřiměřeně stručné.

Všechny níže uvedené příklady se týkají náhodného výběru z normálního rozdělení. Po určitém váhání jsme ponechali původní značení, pouze bezvýznamně upravené. Soustředíme se výhradně na geometrické aspekty problematiky.

### 3.2.1 Sdružené rozdělení průměru a výběrové směrodatné odchyly

V článku [14] Fisher odvozuje rozdělení náhodné veličiny

$$z \equiv \frac{\bar{x} - m}{s} \equiv \frac{\bar{x} - m}{\sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}}}$$

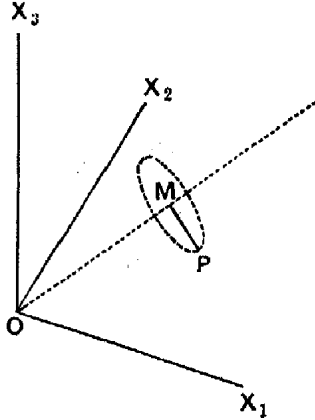
získané z výběru z rozdělení  $N(m; \sigma^2)$  o rozsahu  $n$  (jedná se o reakci na jiný článek z roku 1923, jehož autor uvádí odlišný výsledek). Hned v úvodu upozorňuje, že hledaný vzorec byl publikován Gossetem v práci [28] již v roce 1908. Následuje odvození sdruženého rozdělení průměru  $\bar{x}$  a směrodatné odchyly  $s$  užitím  $n$ -rozměrné geometrie; s největší pravděpodobností je to postup, který použil již v roce 1912 v korespondenci s Gossetem (viz podkapitoly 3.1.3, 3.1.4).<sup>3</sup>

Nejprve Fisher formuluje geometrickou interpretaci hustoty: navrhuje nahlížet na pozorování  $x_1, \dots, x_n$  jako na souřadnice bodu v  $n$ -rozměrném prostoru. Pravděpodobnost, že náhodný vektor bude ležet v okolí tohoto bodu<sup>4</sup> o objemu

$$dx_1 \cdot \dots \cdot dx_n,$$

<sup>3</sup>Tento způsob odvození je ovšem – v poněkud stručnější podobě – shrnut již ve Fisherově článku [12] z roku 1915. Odsud také pochází obrázek 3.1, jeden z mála, kterými Fisher své úvahy ilustroval.

<sup>4</sup>V originále „volume element“.



**Obrázek 3.1:** Fisherova ilustrace z článku [12] ke geometrickému odvození hustoty sdruženého rozdělení výběrového průměru  $\bar{x}$  a směrodatné odchylky  $s$ .

je rovna hodnotě

$$\frac{e^{-\frac{\sum(x_i-m)^2}{2\sigma^2}}}{(\sigma\sqrt{2\pi})^n} \cdot dx_1 \cdot \dots \cdot dx_n.$$

Dále autor upozorňuje na geometrický význam zkoumaných veličin. Při pozorované hodnotě průměru  $\bar{x}$  leží všechny přípustné realizace  $\mathbf{P} = (x_1, \dots, x_n)^T$  v nadrovině (tj. podprostoru dimenze  $n - 1$ ) určené rovnicí

$$x_1 + \dots + x_n = n\bar{x},$$

která je kolmá na vektor  $(1, \dots, 1)^T$  a má od počátku soustavy souřadnic  $\mathbf{O}$  vzdálenost  $\bar{x}\sqrt{n}$ . Hodnota  $\sqrt{n}s$  pak představuje vzdálenost pozorované realizace od bodu  $\mathbf{M} = (\bar{x}, \dots, \bar{x})^T$  (viz obr. 3.1). Při dané hodnotě  $s$  tvoří tedy množina možných pozorování v této nadrovině povrch  $(n - 1)$ -rozměrné koule se středem v bodě  $(\bar{x}, \dots, \bar{x})^T$  a poloměrem úměrným hodnotě  $s$ . Míra této množiny je úměrná hodnotě  $s^{n-2}$ . Přírůstků veličin  $\bar{x}$ ,  $s$  o hodnoty  $d\bar{x}$ ,  $ds$  tedy odpovídá oblast<sup>5</sup> o objemu úměrném výrazu

$$s^{n-2} d\bar{x} ds.$$

Navíc platí

$$e^{-\frac{\sum(x_i-m)^2}{2\sigma^2}} = e^{-\frac{n(\bar{x}-m)^2}{2\sigma^2}} \cdot e^{-\frac{ns^2}{2\sigma^2}},$$

což znamená, že hustota je na výše popsané množině konstantní. Pravděpodobnost, že realizace náhodného výběru bude ležet v této oblasti, je tedy úměrná výrazu

$$e^{-\frac{n(\bar{x}-m)^2}{2\sigma^2}} \cdot e^{-\frac{ns^2}{2\sigma^2}} s^{n-2} ds d\bar{x}, \quad (3.1)$$

<sup>5</sup>Až na konstanty úměrnosti, které Fisher neuvádí, se vlastně jedná o „slupku“ tloušťky  $ds$  obepínající plášť válce výšky  $d\bar{x}$ , jehož podstavou je  $(n - 1)$ -rozměrná koule o poloměru  $s$ .

což je – až na konstantu – hustota hledaného sdruženého rozdělení; konstantu lze určit z podmínky

$$\int_{s \in \mathbb{R}^+, \bar{x} \in \mathbb{R}} dP = 1.$$

Ze vzorce (3.1) je v první řadě patrná nezávislost veličin  $s$  a  $\bar{x}$ ; tímto výsledkem Fisher napravuje jeden z nedostatků Gossetova článku [28] (viz podkapitolu 3.1.3).

V dalším textu autor uvádí marginální hustoty veličin  $\bar{x}$  a  $s$  a pomocí substituce dochází k hledané hustotě náhodné veličiny  $z$ .

### 3.2.2 Rozdělení výběrového korelačního koeficientu

Cílem článku [12] je odvodit hustotu rozdělení výběrového korelačního koeficientu  $r$ , pocházejícího z náhodného výběru z dvojrozměrného normálního rozdělení veličin

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N \left[ \begin{pmatrix} m_1 \\ m_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right]$$

o rozsahu  $n$ . Fisher nejdříve formuluje geometrický význam hustoty zkoumaného  $2n$ -rozměrného náhodného vektoru: výraz

$$\frac{e^{-\frac{1}{1-\rho^2} \sum_{i=1}^n \left( \frac{(x_i-m_1)^2}{2\sigma_1^2} - \frac{2\rho(x_i-m_1)(y_i-m_2)}{2\sigma_1\sigma_2} + \frac{(y_i-m_2)^2}{2\sigma_2^2} \right)}}{(2\pi\sigma_1\sigma_2\sqrt{1-\rho^2})^n} dx_1 dy_1 \dots dx_n dy_n \quad (3.2)$$

představuje pravděpodobnost, že vektor pozorování bude ležet v okolí hodnoty  $(x_1, y_1, \dots, x_n, y_n)^T$  o objemu  $dx_1 dy_1 \dots dx_n dy_n$ . Toho využije k určení hustoty sdruženého rozdělení veličin

$$\begin{aligned} \bar{x} &\equiv \frac{\sum_{i=1}^n x_i}{n}, & \mu_1^2 &\equiv \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}, \\ \bar{y} &\equiv \frac{\sum_{i=1}^n y_i}{n}, & \mu_2^2 &\equiv \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n}, \\ r &\equiv \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n\mu_1\mu_2}, \end{aligned}$$

což je pro něj úkol převážně geometrický:

... it is evident that the only difficulty lies in the expression of an element of volume in  $2n$  dimensional space in terms of these derivatives.

([12], str. 86)

Nejdříve navrhuje promítnout zvlášť  $x$ -ová a zvlášť  $y$ -ová pozorování do téhož  $n$ -rozměrného prostoru. Nyní při pozorovaných hodnotách  $\bar{x}$ ,  $\bar{y}$ ,  $\mu_1$  a  $\mu_2$  tvoří množina možných pozorování povrch dvou  $(n-1)$ -rozměrných koulí ležících v nadrovině kolmé na směr  $(1, \dots, 1)^T$ , jejichž středy jsou v bodech  $(\bar{x}, \dots, \bar{x})^T$  a  $(\bar{y}, \dots, \bar{y})^T$  a jejichž poloměry jsou  $\sqrt{n}\mu_1$  a  $\sqrt{n}\mu_2$ . Výběrový korelační koeficient  $r$  pak představuje kosinus úhlu sevřeného vektory

$$\begin{pmatrix} x_1 - \bar{x}, \dots, x_n - \bar{x} \\ y_1 - \bar{y}, \dots, y_n - \bar{y} \end{pmatrix}^T, \quad (3.3)$$

Dále je Fisher velmi stručný:

Taking one of the projections as fixed at any point on the sphere of radius  $\sqrt{n}\mu_2$ , the region for which  $r$  lies in the range  $dr$ , is a zone, on the other sphere in  $n-1$  dimensions, of radius  $\mu_1\sqrt{n}\sqrt{1-r^2}$ , and of width  $\mu_1\sqrt{n}dr/\sqrt{1-r^2}$ , and therefore having a volume proportional to  $\mu_1^{n-2}(1-r^2)^{\frac{n-4}{2}}dr$ . ([12], str. 87)

Fixuje tedy nejprve hodnoty pozorování  $y_1, \dots, y_n$  (a tedy i  $\bar{Y}$  a  $\mu_2$ ) a hodnoty statistik  $\bar{x}$ ,  $\mu_1$ ; při dané hodnotě koeficientu  $r$  jsou pak přípustná pozorování  $x_1, \dots, x_n$  – bez této podmínky tvořící povrch výše popsané  $(n-1)$ -rozměrné koule určené hodnotami  $\bar{x}$  a  $\mu_1$  – omezena na ty případy, kdy vektory (3.3) spolu svírají úhel  $\alpha = \arccos r$ . Tato pozorování tvoří na povrchu  $(n-1)$ -rozměrné koule povrch  $(n-2)$ -rozměrné koule (představme si jej jako kružnici) o poloměru  $\mu_1\sqrt{n}\sin\alpha = \mu_1\sqrt{n}\sqrt{1-r^2}$ . Přírůstkem koeficientu  $r$  o hodnotu  $dr$  odpovídá změna úhlu

$$d\alpha = -dr/\sqrt{1-r^2},$$

a tedy pás na povrchu  $(n-1)$ -rozměrné koule o „délce“

$$\left[\mu_1\sqrt{n}\sqrt{1-r^2}\right]^{n-3}$$

a šířce

$$\mu_1\sqrt{n}dr/\sqrt{1-r^2}.$$

Míra této množiny je tedy

$$\left[\mu_1\sqrt{n}\sqrt{1-r^2}\right]^{n-3} \cdot \mu_1\sqrt{n}dr/\sqrt{1-r^2} = n^{\frac{n}{2}-1}\mu_1^{n-2}(1-r^2)^{\frac{n-4}{2}}dr,$$

což je – až na konstantu  $n^{\frac{n}{2}-1}$  – Fisherův výsledek.

Přírůstek objemu odpovídající přírůstkům veličin  $\bar{x}$ ,  $\bar{y}$ ,  $\mu_1$ ,  $\mu_2$ ,  $r$  je tedy úměrný výrazu

$$\mu_1^{n-2}\mu_2^{n-2}(1-r^2)^{\frac{n-4}{2}}d\bar{x}d\bar{y}d\mu_1d\mu_2dr$$

(je třeba ještě přidat příslušné diferenciály a činitel  $\mu_2^{n-2}$ , neboť v dosavadních úvahách byly hodnoty  $y_1, \dots, y_n$  fixované). Nyní již můžeme vyjádřit výraz (3.2) pomocí zkoumaných statistik: platí

$$\begin{aligned} \sum_{i=1}^n \left( \frac{(x_i - m_1)^2}{2\sigma_1^2} - \frac{2\rho(x_i - m_1)(y_i - m_2)}{2\sigma_1\sigma_2} + \frac{(y_i - m_2)^2}{2\sigma_2^2} \right) = \\ = n \left[ \frac{(\bar{x} - m_1)^2 + \mu_1^2}{2\sigma_1^2} - \frac{2\rho[r\mu_1\mu_2 + (\bar{x} - m_1)(\bar{y} - m_2)]}{2\sigma_1\sigma_2} + \frac{(\bar{y} - m_2)^2 + \mu_2^2}{2\sigma_2^2} \right], \end{aligned}$$

takže po dosazení za přírůstek objemu dostáváme (až na násobek) výraz

$$e^{-\frac{n}{1-\rho^2} \left[ \frac{(\bar{x}-m_1)^2 + \mu_1^2}{2\sigma_1^2} - \frac{2\rho[r\mu_1\mu_2 + (\bar{x}-m_1)(\bar{y}-m_2)]}{2\sigma_1\sigma_2} + \frac{(\bar{y}-m_2)^2 + \mu_2^2}{2\sigma_2^2} \right]} \cdot \mu_1^{n-2} \mu_2^{n-2} (1-r^2)^{\frac{n-4}{2}} d\bar{x} d\bar{y} d\mu_1 d\mu_2 dr,$$

kde část před diferenciálem je (až na násobek) hledanou hustotou sdruženého rozdělení.

Další text se naštěstí již geometrie netýká. Za upozornění však stojí jeden z výsledků, který je významný z historického hlediska – v případě  $\rho = 0$  má náhodná veličina

$$\frac{r}{\sqrt{1-r^2}}$$

rozdělení identické s rozdělením odvozeným „Studentem“ v článku [28] (viz též podkapitulu 1.10.1); tímto odhalením se Fisherovi potvrdil význam „Studentova“ rozdělení.

### 3.2.3 Rozdělení odchylky od průměru

Fisherův článek [13] je po geometrické stránce mimořádně podnětný. Jedním z prvních dílčích úkolů, kterými se zde autor zabývá, je odvození hustoty rozdělení odchylek  $x_i$  jednotlivých měření od výběrového průměru získaného z náhodného výběru z rozdělení  $N(m; \sigma^2)$  o rozsahu  $n$  (Fisher značí odchylky od průměru symbolem  $x_i$ , stejně jako původní měření v jiných odstavcích téhož článku). Každý bod reprezentovaný těmito hodnotami leží podle něj v podprostoru dimenze  $n-1$  daném rovnicí

$$\sum_{i=1}^n x_i = 0;$$

tento podprostor budeme dále značit symbolem  $M$ . Hustota udávající na tomto podprostoru rozdělení vektoru odchylek je násobkem výrazu

$$e^{-\frac{r^2}{2\sigma^2}}, \quad (3.4)$$

kde  $r^2 = \sum_i x_i^2$ , tj.  $r$  je vzdálenost od počátku soustavy souřadnic.<sup>6</sup> Toho Fisher využije následujícím způsobem:

The region in which any co-ordinate has an assigned value,  $x_1$ , is a plane of  $(n-2)$  dimensions, at a distance

$$x_1 \sqrt{\frac{n}{n-1}} \quad (3.5)$$

from the origin, and the frequency with which  $x_1$  falls into the range  $dx_1$  is therefore proportional to

$$e^{-\frac{nx_1^2}{2(n-1)\sigma^2}}. \quad (3.6)$$

<sup>6</sup>V článku je toto tvrzení uvedeno bez důkazu; my jsme jej dokázali v podkapitole 2.6.2 – vektor odchylek od průměru totiž představuje pravoúhlý průmět původního vektoru do  $(n-1)$ -rozměrného podprostoru kolmého na vektor  $(1, \dots, 1)^T$ .

Thus deviations from the mean of samples of  $n$  of a normal population are themselves normally distributed. ([13], str. 191)

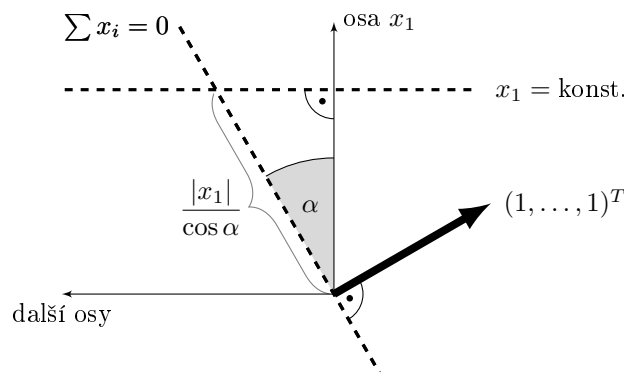
První část této úvahy můžeme snadno osvětlit: normálovým vektorem nadroviny  $\sum_i x_i = 0$  je vektor  $(1, \dots, 1)^T$ , takže pro úhel  $\alpha$ , který svírá tato nadrovina s osou  $x_1$ , platí

$$\sin \alpha = \frac{|(1, 0, \dots, 0)^T \circ (1, \dots, 1)^T|}{\|(1, 0, \dots, 0)^T\| \cdot \|(1, \dots, 1)^T\|} = \frac{1}{\sqrt{n}},$$

z čehož vyplývá

$$\cos \alpha = \sqrt{1 - \sin^2 \alpha} = \sqrt{\frac{n-1}{n}};$$

z toho lze určit vzdálenost (3.5) (viz obr. 3.2).<sup>7</sup> Jak však z této skutečnosti



**Obrázek 3.2:** Vzdálenost lineární množiny určené rovnicemi  $x_1 = \text{konst.}$ ,  $\sum x_i = 0$  od počátku soustavy souřadnic je  $|x_1|/\cos \alpha$ , kde  $\alpha$  je úhel, který osa  $x_1$  svírá s podprostorem určeným rovnicí  $\sum_i x_i = 0$ . To platí díky tomu, že vektor  $(1, \dots, 1)^T$  a osa  $x_1$  jsou na uvedené lineární množinu kolmé, a proto je na ni kolmá i jimi určená rovina, tj. náčrta obrázku.

vyvozuje Fisher tak bezprostředně závěr (3.6), nám není jasné; proto jsme cítili potřebu dokázat správnost tohoto kroku v podkapitole 2.6.4. Nevylučujeme ale možnost existence přímočařejšího náhledu.

### 3.2.4 Sdružené rozdělení dvou odchylek od průměru

Podobným způsobem Fisher dále z výsledku (3.4) odvodí hustoty sdruženého rozdělení dvou odchylek  $x_1$ ,  $x_2$  od výběrového průměru:

<sup>7</sup>Jiným způsobem, jak lze ke vzdálenosti (3.5) dojít, je tato úvaha: ze všech vektorů vyhovujících rovnicím  $\sum_i x_i = 0$ ,  $x_1 = \text{konst.}$ , je zřejmě nejkratší vektor

$$\left( x_1, -\frac{x_1}{n-1}, \dots, -\frac{x_1}{n-1} \right)^T,$$

jehož délka je (3.5). Správně by ovšem mělo být ve výsledném výrazu  $x_1$  v absolutní hodnotě.

Consider the the distribution of pairs of values  $x_1$  and  $x_2$ . The space in which the representative points lie is parallel to the line

$$\begin{aligned} x_1 + x_2 &= 0 \\ x_3 = x_4 = \dots = x_n &= 0, \end{aligned} \quad (3.7)$$

while it makes with the line

$$\begin{aligned} x_1 &= x_2 \\ x_3 = x_4 = \dots = x_n &= 0, \end{aligned} \quad (3.8)$$

an angle the cosine of which is

$$\sqrt{\frac{n-2}{n}}. \quad (3.9)$$

Consequently the frequency in the range  $dx_1 dx_2$  is proportional to

$$e^{-\frac{1}{2\sigma^2} \left\{ \frac{(x_1-x_2)^2}{2} + \frac{n}{n-2} \cdot \frac{(x_1+x_2)^2}{2} \right\}} dx_1 dx_2 = \quad (3.10)$$

$$= e^{-\frac{n-1}{2(n-2)\sigma^2} \left\{ x_1^2 + \frac{2x_1x_2}{n-1} + x_2^2 \right\}} dx_1 dx_2, \quad (3.11)$$

showing a surface of normal correlation<sup>8</sup>, with correlation coefficient,  $-\frac{1}{n-1}$ , between any two deviations. ([13], str. 191)

Fisher tedy v podprostoru  $M$  fixuje hodnoty proměnných  $x_1$  a  $x_2$ , pak odvodí, že vzdálenost takto definované lineární množiny (označme ji  $L'$ ) od počátku soustavy souřadnic je

$$\sqrt{\frac{(x_1 - x_2)^2}{2} + \frac{n}{n-2} \cdot \frac{(x_1 + x_2)^2}{2}} \quad (3.12)$$

(to ovšem není v textu explicitně uvedeno), a to mu umožní – opět pomocí pravidla dokázaného v podkapitole 2.6.4 – formulovat vzorec (3.10). Ten pak upraví na tvar (3.11), ze kterého zřejmě srovnáním s obecným vzorcem dvojrozměrného normálního rozdělení vyčte korelační koeficient mezi odchylkami  $x_1$  a  $x_2$ .

Obtížně pochopitelným krokem se nám v této úvaze zdá způsob, jakým autor dochází k výrazu (3.12). Výsledek je samozřejmě správný – stačí si uvědomit, že ze všech vektorů, které jsou prvkem množiny  $L'$ , je zřejmě nejkratší vektor

$$\left( x_1, x_2, -\frac{x_1 + x_2}{n-2}, \dots, -\frac{x_1 + x_2}{n-2} \right),$$

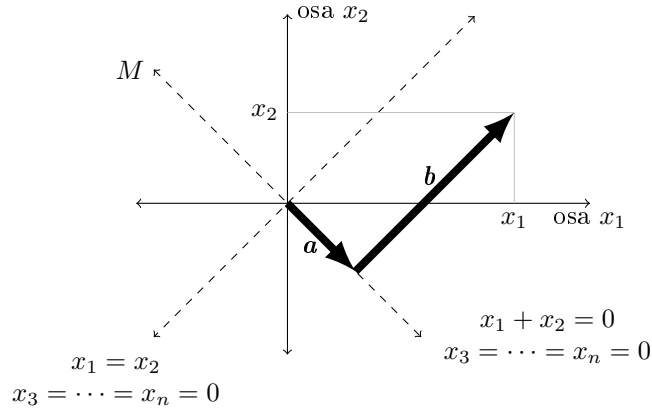
jehož délka je skutečně (3.12). Fisherova argumentace je však jiná.

Přímky (3.7), (3.8) jsou osami kvadrantů roviny  $x_1, x_2$ ; postup jeho úvah si tedy vykládáme tak, že nejprve rozloží vektor  $\mathbf{x} \equiv (x_1, x_2, 0, \dots, 0)^T$  na součet dvou vektorů  $\mathbf{a}$ ,  $\mathbf{b}$ , které jsou rovnoběžné s těmito osami a jejichž délky jsou

$$\frac{|x_1 - x_2|}{\sqrt{2}}, \quad \frac{|x_1 + x_2|}{\sqrt{2}}$$

(viz obr. 3.3). Vektor  $\mathbf{a}$  leží v podprostoru  $M$ , vektor  $\mathbf{b}$  s tímto podprostorem

<sup>8</sup>Má být zřejmě „distribution“.



**Obrázek 3.3:** Ilustrace k rotaci os  $x_1, x_2$  o  $45^\circ$ . Platí  $\|\mathbf{a}\| = |x_1 - x_2|/\sqrt{2}$ ,  $\|\mathbf{b}\| = |x_1 + x_2|/\sqrt{2}$ . Podprostor  $M$  protíná náčrtku v ose druhého a čtvrtého kvadrantu.

svírá úhel velikosti  $\alpha$ , pro který platí

$$\sin \alpha = \frac{|(1, 1, 0, \dots, 0)^T \circ (1, \dots, 1)^T|}{\|(1, 1, 0, \dots, 0)^T\| \cdot \|(1, \dots, 1)^T\|} = \sqrt{\frac{2}{n}},$$

tj.

$$\cos \alpha = \sqrt{1 - \sin^2 \alpha} = \sqrt{\frac{n-2}{n}}.$$

Dále je třeba si představit další vektor, řekněme  $\mathbf{c}$ , směřující mimo rovinu  $x_1, x_2$  z koncového bodu vektoru  $\mathbf{x}$  do bodu, který je z množiny  $L'$  nejbližší počátku soustavy souřadnic. Tento vektor zřejmě musí být tvaru  $(0, 0, x_3, \dots, x_n)$ , je tedy kolmý na rovinu  $x_1, x_2$ . Délka vektoru  $\mathbf{a} + \mathbf{b} + \mathbf{c}$  je pak hledaná vzdálenost. Je-li nyní úhel  $\beta$  sevřený vektory  $\mathbf{b}$  a  $\mathbf{b} + \mathbf{c}$  stejný jako úhel  $\alpha$ , platí

$$\|\mathbf{b} + \mathbf{c}\|^2 = \frac{\|\mathbf{b}\|^2}{\cos^2 \beta} = \frac{n}{n-2} \cdot \frac{(x_1 + x_2)^2}{2},$$

neboť vektor  $\mathbf{b} + \mathbf{c}$  je přeponou pravoúhlého trojúhelníku, jehož odvěsny jsou vektory  $\mathbf{b}$  a  $\mathbf{c}$ . Z Pythagorovy věty

$$\|\mathbf{a} + \mathbf{b} + \mathbf{c}\|^2 = \|\mathbf{a}\|^2 + \|\mathbf{b} + \mathbf{c}\|^2$$

pak snadno dojdeme k výsledku (3.12).

Je třeba ovšem dokázat rovnost úhlů  $\alpha$  a  $\beta$ . Ta plyne ze vzájemné polohy roviny určené vektory  $\mathbf{b}, \mathbf{c}$  a podprostoru  $M$ ; můžeme totiž ukázat, že platí

$$M \perp [\mathbf{b}, \mathbf{c}], \quad (3.13)$$

$$[\mathbf{b}, \mathbf{c}] \cap M = [\mathbf{b} + \mathbf{c}], \quad (3.14)$$

z čehož vyplývá, že pravoúhlý průmět  $\mathbf{b}_M$  vektoru  $\mathbf{b}$  do podprostoru  $M$  musí být rovnoběžný s vektorem  $\mathbf{b} + \mathbf{c}$  a lze psát

$$\alpha \equiv |\angle \mathbf{b}, M| = |\angle \mathbf{b}, \mathbf{b}_M| = |\angle \mathbf{b}, \mathbf{b} + \mathbf{c}| \equiv \beta;$$

ke třetí rovnosti je třeba poznamenat, že oba úhly musí být menší než  $90^\circ$ , vektory  $\mathbf{b}_M$  a  $\mathbf{b} + \mathbf{c}$  musí mít tedy i stejnou orientaci.

Dokažme tedy ještě tvrzení (3.13) a (3.14). Druhé z nich plyne z toho, že vektory  $\mathbf{a}$  a  $\mathbf{a} + \mathbf{b} + \mathbf{c}$  leží oba v podprostoru  $M$ , leží v něm tedy i vektor  $\mathbf{b} + \mathbf{c}$ . Dokázat druhé tvrzení však není tak snadné, ačkoli je intuitivně snadno přijatelné – alespoň pokud se člověk spokojí s představou v trojrozměrném prostoru.

Uvědomme si tedy, že vektor  $\mathbf{a} + \mathbf{b} + \mathbf{c}$  musí být ze své definice kolmý na *zaměření* množiny  $L'$ , tj. na podprostor  $L$  tvořený všemi takovými vektory  $\mathbf{u}$ , pro které platí

$$\exists \mathbf{v}, \mathbf{w} \in L' : \mathbf{u} = \mathbf{v} - \mathbf{w}.$$

V našem případě se zřejmě jedná o všechny vektory tvaru  $(0, 0, x_3, \dots, x_n)^T$ . Oba vektory  $\mathbf{a}$ ,  $\mathbf{b}$  jsou na tento podprostor také kolmé, musí na něj být tedy kolmý i vektor  $\mathbf{c}$ . Protože jsou tyto tři vektory lineárně nezávislé a dimenze podprostoru  $L$  je  $n - 3$ , znamená to, že platí

$$[\mathbf{a}, \mathbf{b}, \mathbf{c}] = L^\perp.$$

Vektor  $(1, \dots, 1)^T$ , který je kolmý na podprostor  $M$  (a tedy i na  $L$ ) musí být tím pádem lineární kombinací těchto tří navzájem kolmých vektorů. Na vektor  $\mathbf{a} \in M$  je však také kolmý, musí tedy ležet v rovině  $[\mathbf{b}, \mathbf{c}]$ . A protože je jediným generátorem podprostoru  $M^\perp$ , platí  $M^\perp \subseteq [\mathbf{b}, \mathbf{c}]$ , z čehož vyplývá  $M \perp [\mathbf{b}, \mathbf{c}]$  a proto i  $M \sqcup [\mathbf{b}, \mathbf{c}]$  (viz podkapitulu 2.2.2).

Fisher tedy rychle a se značnou suverenitou dochází pomocí  $n$ -rozměrné geometrie ke správným výsledkům, domníváme se však, že podstatnou část argumentace vynechává. Fakta, která uvádí, nejsou vždy postačující podmínkou pro závěry, které z nich vyvozuje.

### 3.2.5 Sdružené rozdělení průměrné odchylky a výběrové směrodatné odchylky

Hlavním cílem článku [13] je porovnat přesnost odhadu směrodatné odchylky pomocí statistik odvozených od průměrné absolutní odchylky od průměru ( $\sigma_1$ ) a od výběrové směrodatné odchylky ( $\sigma_2$ ) vypočtené z náhodného výběru z rozdělení  $N(m; \sigma^2)$  o rozsahu  $n$ :

$$\sigma_1 \equiv \sqrt{\frac{\pi}{2}} \sum_{i=1}^n |x_i - \bar{x}|, \quad (3.15)$$

$$\sigma_2 \equiv \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}}. \quad (3.16)$$

V hlavní části se Fisher omezuje na případ, kdy  $n = 4$ , a odvozuje hustotu rozdělení veličiny  $\sigma_1$  podmíněné danou hodnotou  $\sigma_2$ . Vychází z představy, že při pevné hodnotě  $\sigma_2$  tvoří množina možných pozorovaných hodnot sféru  $K$  se

středem v bodě  $M \equiv (\bar{x}, \bar{x}, \bar{x}, \bar{x})^T$  a poloměrem  $2\sigma_2$ , která leží v trojrozměrné nadrovině  $N$  dané rovnicí

$$\sum_{i=1}^n (x_i - \bar{x}) = 0;$$

je to tedy „obyčejná“ dvojrozměrná sféra, tj. povrch trojrozměrné koule. Naproti tomu při pevné hodnotě  $\sigma_1$  tvoří množina možných pozorování v nadrovině  $N$  povrch mnohostěnu; jeho průnikem se sférou  $K$  je křivka, která reprezentuje pozorování vyhovující daným hodnotám  $\sigma_1$  a  $\sigma_2$ . Z její délky lze odvodit hledanou hustotu podmíněného rozdělení díky tomu, že na povrchu této koule je hustota náhodného vektoru konstantní (viz předchozí podkapitoly).

Fisher se nejprve zabývá vlivem absolutních hodnot ve vzorci (3.15) v různých částech nadroviny  $N$ :

Within this space the values of  $(x - \bar{x})$  will be positive or negative according as the representative point lies on the other four planes, through  $M$ , drawn parallel to the faces of a regular tetrahedron.

([13], str. 195)

Jedná se o to, že hodnoty  $(x_i - \bar{x})$  jsou pozitivní či negativní podle toho, v jaké části nadroviny  $N$  rozdělené nadrovinami

$$x_1 = \bar{x}, \quad x_2 = \bar{x}, \quad x_3 = \bar{x}, \quad x_4 = \bar{x}$$

daná čtveřice pozorování leží. Tyto čtyři nadroviny protínají podprostor  $M$  ve čtyřech rovinách, které jsou vůči sobě orientovány stejně jako stěny pravidelného čtyřstěnu (nejprostším důkazem tohoto tvrzení je požadavek symetrie, lze jej však ověřit i analyticky); protínají se ovšem všechny v jednom bodě  $M$ . Jakékoli dvě z nich tedy svírají úhel o velikosti  $\arccos \frac{1}{3}$  a jakékoli dvě jejich sousední průsečnice svírají úhel o velikosti  $\pi/3$ .

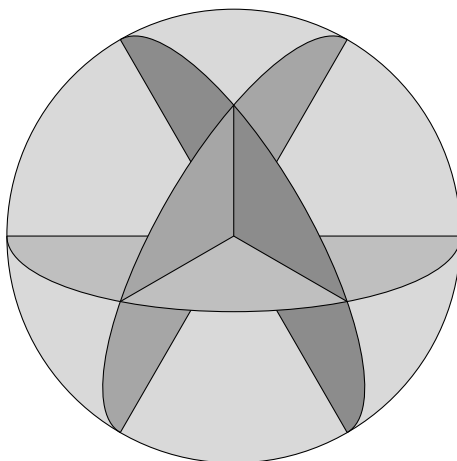
Tyto roviny rozdělují sféru  $K$  na čtrnáct oblastí, kde každá z nich odpovídá určité kombinaci znamének výrazů  $(x_i - \bar{x})$  – vyloučeny jsou případy všech znamének záporných a všech znamének kladných.<sup>9</sup> Šest těchto oblastí je tvaru čtverce – ty odpovídají případům dvou pozitivních a dvou negativních hodnot; zbylé mají tvar trojúhelníku a odpovídají případům jedné hodnoty pozitivní a tří negativních, resp. naopak (viz obr. 3.4). Čtvercové oblasti celkem pokrývají část sféry  $K$  o relativním obsahu

$$6 \cdot \left( \frac{1}{2} - \frac{\arccos \frac{1}{3}}{\pi} \right) = 3 - \frac{6 \arccos \frac{1}{3}}{\pi} = 0,6490,$$

zatímco na trojúhelníkové oblasti připadá podíl

$$8 \cdot \left( \frac{3 \arccos \frac{1}{3}}{4\pi} - \frac{1}{4} \right) = \frac{6 \arccos \frac{1}{3}}{\pi} - 2 = 0,3510;$$

<sup>9</sup>To je mimochodem zajímavá kombinatorická úvaha. Pokud poněkud zjednodušíme formulaci, znamená to, že nadrovina  $n$ -rozměrného prostoru popsána rovnicí  $x_1 + \dots + x_n = 0$  je  $n$  nadrovinami  $x_1 = 0, \dots, x_n = 0$  rozdělena na  $2^n - 2$  částí.



**Obrázek 3.4:** Čtyři roviny, které se protínají ve středu koule a jsou rovnoběžné se stěnami pravidelného čtyřstěnu, dělí povrch koule na čtrnáct oblastí; šest z nich má tvar čtverce, zbývající mají tvar trojúhelníku. Na obrázku je povrch koule průhledný, rovin nikoli. Jedna z rovin je rovnoběžná s nákresnou.

těchto komplikovaně vyhlížejících výsledků je ve skutečnosti překvapivě snadné dosáhnout, neboť relativní obsahy čtvercových a trojúhelníkových oblastí musí vyhovovat soustavě

$$\begin{aligned} 6 \square + 8 \triangle &= 1, \\ 1 \square + 2 \triangle &= \frac{\arccos \frac{1}{3}}{2\pi}. \end{aligned}$$

Pokud nyní fixujeme hodnotu  $\sigma_1$  a zvolíme určitou kombinaci pozitivních a negativních odchylek, dostaneme ze vzorce (3.15) rovnici nadroviny, jejímž průnikem s nadrovinou  $N$  je rovina; je-li tato rovina ve vhodné vzdálenosti od bodu  $M$ , je jejím průnikem s příslušnou částí sféry  $K$  kružnice či její část. Osou této kružnice je v nadrovině  $N$  vždy osa příslušné části sféry (to je v článku uvedeno bez důkazu pro případ čtvercových částí, je však patrné, že se to předpokládá i u částí trojúhelníkových). Fisher vypočte úhlovou velikost  $\theta$  poloměru těchto kružnic:

$$\theta = \arccos \left( \sqrt{\frac{2}{\pi}} \cdot \frac{\sigma_1}{\sigma_2} \right)$$

v případě čtvercových částí, resp.

$$\theta = \arccos \left( \sqrt{\frac{8}{3\pi}} \cdot \frac{\sigma_1}{\sigma_2} \right)$$

v případě částí trojúhelníkových, a pokračuje:

The frequency distribution of  $\sigma_1$ , for a given value of  $\sigma_2$ , is thus reduced to the frequency of occurrence of different values of  $\theta$ , in two types of spherical figures. For the quadrangles the greatest possible value of  $\theta$  is  $45^\circ$ , while the least distance from the perimeter to the

centre is  $\sin^{-1} \sqrt{\frac{1}{3}}$ . From these 6 regions we have

$$\begin{aligned} \text{From } 0 \text{ to } \sin^{-1} \frac{1}{\sqrt{3}} & \quad \text{frequency } 3 \sin \theta \, d\theta \\ \text{From } \sin^{-1} \frac{1}{\sqrt{3}} \text{ to } 45^\circ & \quad \text{frequency } 3 \sin \theta \left( 1 - \frac{4}{\pi} \cos^{-1} \frac{1}{\sqrt{2} \tan \theta} \right) d\theta \end{aligned}$$

([13], str. 196)

(symbolem  $\sin^{-1}$  míní arcsin). Co to znamená? Hustota náhodného vektoru je na sféře  $K$  konstantní. Pravděpodobnost, že tento vektor bude ležet na jejím povrchu v dané oblasti (při podmínce, že se nachází kdekoli na jejím povrchu), lze tedy určit jako podíl obsahu  $dS$  této oblasti vůči obsahu  $S$  celé sféry. Při změně úhlu  $\theta$  o hodnotu  $d\theta$  je touto oblastí pás šířky  $dS = 2\sigma_2 d\theta$  (nezapomeňme, že  $2\sigma_2$  je poloměr koule) a délky, která je součtem délek všech výše popsaných kružnic či jejich částí. Jak lze ověřit, na jedné čtvercové oblasti je délka části této křivky

$$\begin{aligned} l &= 4\pi\sigma_2 \sin \theta & \text{pro } 0 \leq \theta \leq \arcsin \frac{1}{\sqrt{3}}, \\ l &= 4\pi\sigma_2 \sin \theta \left( 1 - \frac{4}{\pi} \arccos \frac{1}{\sqrt{2} \tan \theta} \right) & \text{pro } \arcsin \frac{1}{\sqrt{3}} \leq \theta \leq 45^\circ; \end{aligned}$$

druhý řádek se týká situace, kdy je úhel  $\theta$  příliš velký, takže do čtvercové oblasti se vejde jen část kružnice v rozích. Hledaná pravděpodobnost je tedy

$$6 \cdot \frac{dS}{S} = \frac{6 \cdot l \cdot 2\sigma_2 d\theta}{4\pi (2\sigma_2)^2};$$

dosazením za  $l$  dostaneme Fisherův výsledek.

Analogický výpočet poté autor provede pro trojúhelníkové oblasti a nakonec dosadí za  $\theta$  a  $d\theta$  odpovídající funkce parametru  $\sigma_1$ . Hledaná podmíněná hustota je součtem takto získaných funkcí (bez diferenciálu), Fisher ji však uvádí stále odděleně:

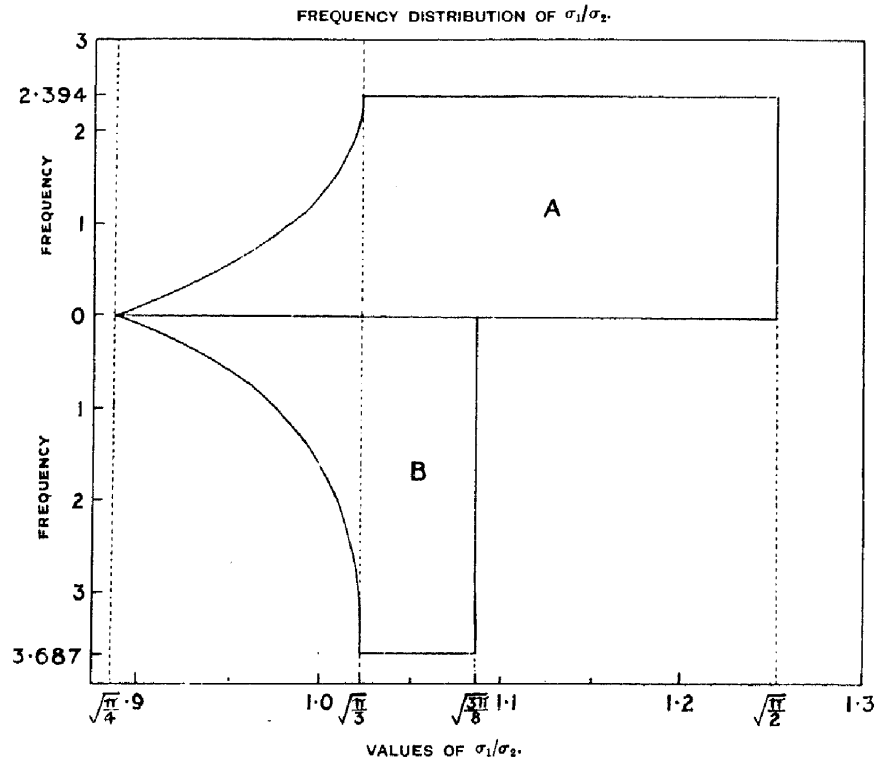
$$\begin{aligned} \frac{3}{\sigma_2} \sqrt{\frac{2}{\pi}} \left( 1 - \frac{4}{\pi} \arccos \frac{\sigma_1}{\sqrt{\pi\sigma_2^2 - 2\sigma_1^2}} \right) & \quad \text{pro } \sigma_2 \sqrt{\frac{\pi}{4}} \leq \sigma_1 \leq \sigma_2 \sqrt{\frac{\pi}{3}}, \\ \frac{3}{\sigma_2} \sqrt{\frac{2}{\pi}} & \quad \text{pro } \sigma_2 \sqrt{\frac{\pi}{3}} \leq \sigma_1 \leq \sigma_2 \sqrt{\frac{\pi}{2}} \end{aligned}$$

pro čtvercové oblasti a

$$\begin{aligned} \frac{8}{\sigma_2} \sqrt{\frac{2}{3\pi}} \left( 1 - \frac{3}{\pi} \arccos \frac{\sigma_1}{\sqrt{3\pi\sigma_2^2 - 8\sigma_1^2}} \right) & \quad \text{pro } \sigma_2 \sqrt{\frac{\pi}{4}} \leq \sigma_1 \leq \sigma_2 \sqrt{\frac{\pi}{3}}, \\ \frac{8}{\sigma_2} \sqrt{\frac{2}{3\pi}} & \quad \text{pro } \sigma_2 \sqrt{\frac{\pi}{3}} \leq \sigma_1 \leq \sigma_2 \sqrt{\frac{3\pi}{8}} \end{aligned}$$

pro oblasti trojúhelníkové (viz obr. 3.5, kde je odfiltrován vliv  $\sigma_2$ ).

Z historického hlediska je významné, že toto rozdělení nezávisí na skutečné hodnotě rozptylu  $\sigma^2$ ; to znamená, že při známé hodnotě statistiky  $\sigma_2$  již znalost statistiky  $\sigma_1$  nemůže o hodnotě parametru  $\sigma^2$  přinést žádnou novou informaci. Tento postřeh dovedl Fishera později k zavedení pojmu *suficientní statistiky*.



**Obrázek 3.5:** Hustota podílu  $\sigma_1/\sigma_2$  (viz (3.15), (3.16)) pro případ  $n = 4$ . Horní část křivky zahrnuje kombinace dvou pozitivních a dvou negativních odchylek od průměru, dolní část zbývající možnosti. Převzato z práce [13].

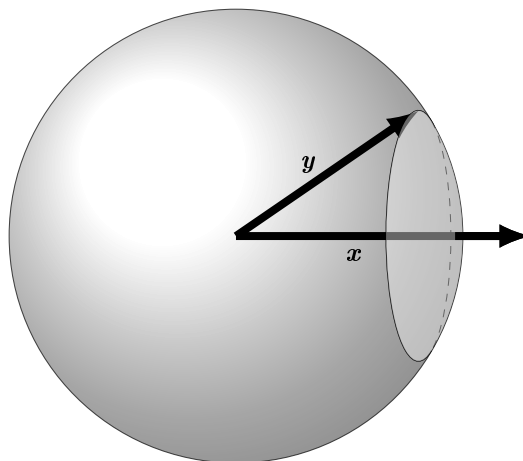
### 3.2.6 Test významnosti

Základ Fisherových úvah ohledně testu významnosti objasňuje v jeho životopise [5] jeho dcera J.F.Box následujícím způsobem. Předpokládejme jednoduchý lineární model

$$\mathbf{Y} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \cdot \beta + \begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \end{pmatrix} \equiv \mathbf{x}\beta + \mathbf{Z}, \quad (3.17)$$

kde  $\mathbf{Z} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_3)$ . Chceme-li zvažovat nulovou hypotézu  $H_0 : \beta = 0$ , je zřejmé, že ji zamítneme tehdy, když bude úhel mezi získanou realizací  $\mathbf{y}$  a vektorem  $\mathbf{x}$  malý. Jaké kritérium musíme pro své rozhodnutí zvolit, abychom se v případě platnosti hypotézy  $H_0$  zmýlili nanejvýš v 5% případech, tj. abychom nepřekročili obvykle požadovanou pravděpodobnost chyby prvního druhu? Nulová hypotéza vlastně znamená, že hustota závisí pouze na vzdálenosti od počátku soustavy souřadnic, takže nezávisí na směru. Vhodné kritérium můžeme tedy získat tak, že necháme získaný vektor  $\mathbf{y}$  rotovat kolem vektoru  $\mathbf{x}$ , čímž opišeme na sféře se středem v počátku a poloměrem  $\|\mathbf{y}\|$  hranici vrchlíku (viz obr. 3.6). Je-li relativní obsah tohoto vrchlíku vůči obsahu celé sféry menší než 5%, zamítneme nulovou hypotézu.

Po zobecnění na více rozměrů představuje tato myšlenka podstatu klasického jednovýběrového  $t$ -testu; jeho výsledná hladina významnosti je tedy vlastně podíl obsahu příslušného „hypervrchlíku“ vůči obsahu celé  $n$ -rozměrné koule. Tato



**Obrázek 3.6:** Získáme-li za platnosti modelu (3.17) realizaci  $\mathbf{y}$ , zamítneme nulovou hypotézu  $H_0 : \beta = 0$  tehdy, když bude úhel sevřený vektory  $\mathbf{y}$  a  $\mathbf{x}$  dostatečně malý. Hladinou významnosti tohoto testu je podíl  $S_v/S_k$ , kde  $S_k$  je obsah sféry se středem v počátku soustavy souřadnic a poloměrem  $\|\mathbf{y}\|$  a  $S_v$  je obsah vrchlíku, který na této sféře vznikne, necháme-li rotovat vektor  $\mathbf{y}$  kolem vektoru  $\mathbf{x}$ .

hodnota je funkcí úhlu sevřeného vektory  $\mathbf{y}$  a  $\mathbf{x}$ . V obecnějším případě, kdy je model vícerozměrný, je analogická úvaha základem  $F$ -testu jako hlavního výsledku analýzy rozptylu.

### 3.3 Další osudy geometrického přístupu

R. A. Fisher sice v některých svých člancích geometrii použil, v jeho nejvýznamnějších pracích však zmíněna není. Důvodem je patrně to, že Gosset, který byl jeho blízkým spolupracovníkem, nebyl schopen porozumět  $n$ -rozměrné geometrii (viz [5]); ze skromnosti to sice připisoval svým nedostatečným matematickým schopnostem, je však třeba přiznat, že Fisher si s vysvětlováním opravdu nedával příliš mnoho práce. Snad právě v reakci na Gossetovu prosbu „please don’t let too much be clear or obvious“, narážející na Fisherovy stručné argumenty typu „it is obvious, that...“, použil Fisher v článku [15], shrnujícím hlavní poznatky týkající se možného použití „Studentova“ rozdělení, algebraický přístup. Sám to komentoval těmito slovy:

It is perhaps worth while to give, at length, an algebraical method of proof, since analogous cases have hitherto been demonstrated only geometrically, by means of a construction in Euclidian hyperspace, and the validity of such methods of proof may not be universally admitted. ([15], str. 48)

Tak byl položen základ tradici, která v matematické statistice převládla, a geometrický přístup se od té doby využívá v literatuře jen sporadicky.

Poměrně podrobný přehled tohoto vývoje uvádí D. G. Herr v článku [17]. Rozzebírá zde přístup vybraných sedmi autorů, kteří příležitostně použili geometrický přístup k lineárnímu modelu. Nicméně pouze přístup R. A. Fishera a přístup autorů článku [8] jsou zde charakterizovány jako čistě geometrické v tom smyslu, že

jejich autoři z popsaných geometrických představ vyvozují výsledné vzorce zcela bezprostředně. Ostatní práce Herr popisuje spíše jako analyticko-geometrické, čímž míní to, že v nich jsou geometrické představy reprezentovány analytickými vzorci.

Autoři publikace [27] nějaký hlavní zdroj inspirace neuvádějí a v úvodu zmiňují pouze Fishera:

Unfortunately Fisher found it difficult to explain his geometric proofs to his colleagues, and in the main they decided the geometry was too difficult. For this reason, and easy computing formulae were needed in the early days, the tradition arose of expressing the geometric results in algebraic form... The book aims to retrieve Fisher's lost insight, and to present it in a dish palatable to all and even delicious to many.

Michael Wichura, autor námi často citované učebnice [31], připisuje hlavní zásluhu na jeho seznámení s geometrickým přístupem (a tím i pochopení matematické teorie stojící v pozadí analýzy rozptylu a regresních modelů) Willamu Kruskalovi, který mu někdy v polovině 60. let dal přečíst předběžnou verzi své učebnice věnované tomuto tématu.

Podobnou zkušenost zmiňuje i Morris Eaton v článku [9]. Role Williama Kruskala (1919–2005) je zřejmě v tomto směru zásadní. Ostatně jeho práce [18], [19] a [20] patří v literatuře zabývající se geometrickým přístupem k nejcitovanějším. K vydání výše zmíněné učebnice však nikdy nedošlo. Kruskala samotného k tomuto tématu podle [33] přivedl L. J. Savage.

V české literatuře je nám známa pouze jediná publikace věnovaná tomuto tématu, a tou je článek Pázmanův [23]. Ten se však zabývá podstatně specializovanějšími technikami než tato práce a obrací se k pokročilejším čtenářům. V učebnicích [3] a [35] jsou příležitostně použity geometrické argumenty, jedná se však o výjimky.

Ačkoli je výše uvedený výčet pouze ilustrativní a neklade si žádné nároky na úplnost, považujeme za zřejmé, že geometrický přístup k lineárnímu modelu je v záplavě statistické literatury poměrně okrajovým jevem. Nepočítaje nevydanou Kruskalovu publikaci, existují v této oblasti dosud zřejmě pouhé dvě učebnice (konkrétně knihy [27] a [31]).<sup>10</sup> Přitom málokterý z autorů zabývajících se tímto tématem opomene zdůraznit jeho eleganci a srozumitelnost, díky nimž, jak zdůrazňuje Box v životopise [5], „the results can be immediately seen rather than laboriously derived“. Herr navrhuje v článku [17] čtyři různé teorie, které tuto skutečnost vysvětlují:

- tradice;
- Fisherův způsob použití geometrie vyvolal dojem, že takový přístup může být srozumitelný pouze výjimečně nadaným jedincům;<sup>11</sup>
- nedostatečné matematické vzdělání některých statistiků;

<sup>10</sup>Několik skrovných, leč výstižných poznámek upozorňujících na užitečnost geometrické interpretace je také v učebnici [7].

<sup>11</sup>Autor Fisherův stručný styl komentuje těmito výstižnými slovy: „If you see it, it's beautifully elegant; if you don't, there is very little there to improve your vision. It seems the kind of discussion that inspires the reader to honor the genius that produced it, but does not to inspire him to try to emulate the approach.“ ([17], str. 45)

- maticová algebra je obecnější než geometrie (tuto možnost Herr vzápětí zavrhuje).

Přidejme k tomuto výčtu ještě jednu možnost, která by mohla přispívat k udržování současného stavu:

- oproti ostatním matematickým oborům má statistika výrazně silný vztah k aplikacím. Je tedy možné, že matematikové, kteří si ji volí jako svoji specializaci, jsou svou povahou praktičtější založení a méně dbají o jakousi pochybnou a subjektivní „eleganci“; dosažené výsledky jsou pro ně důležitější než způsoby jejich získání a necítí potřebu hledat způsoby nové.

Ať už jsou však příčiny této situace jakékoli, jsme přesvědčeni, že geometrický přístup by si zasloužil ve statistické literatuře širší prostor; myšlenky prezentované v naší práci jsou snad pro tento názor dostatečnou oporou.

## Shrnutí

První část předkládané práce nabízí alternativní koncept výkladu teorie lineárního modelu, založený na geometrii vektorových prostorů.

Základním východiskem je představa náhodného vektoru jako geometrického objektu umístěného ve vektorovém prostoru dimenze  $n$ , kde  $n$  je obvykle počet měření. Lineární model pak specifikuje polohu neznámé střední hodnoty náhodného vektoru v tom smyslu, že ji omezuje na nějaký vektorový podprostor. Je-li varianční matice tvaru  $\sigma^2 \mathbf{I}_n$  (což je případ, na který se – spolu s předpokladem normality – omezujeme téměř v celé práci), je nejlepším nestranným lineárním odhadem střední hodnoty pravoúhlý průmět náhodného vektoru do příslušného podprostoru.

Dále se ukazuje, že vzorce definující rozdělení  $\chi_n^2$ ,  $t_n$  a  $F_n$  lze interpretovat jako funkce jednoho či více pravoúhlých průmětů náhodného vektoru do navzájem kolmých podprostorů, přičemž tyto funkce nezávisí na zvolené soustavě souřadnic. To umožňuje vyslovit užitečná tvrzení pravděpodobnostního charakteru ohledně neznámých parametrů modelu a eventuální platnosti submodelu, neboť navzájem kolmé podprostory a pravoúhlé průměty hrají v další analýze lineárního modelu dominantní roli. Odhad střední hodnoty, odhad rozptylu a test submodelu jsou témata, v nichž efektivita geometrického přístupu vyniká nejnápadněji. Ostatní kapitoly této části vyžadují o něco hlubší zamyšlení a znalosti.

Druhá část publikace se zabývá jednak teoretickými základy, na kterých je geometrický přístup položen, a jednak způsoby, jak lze jeho použití zobecnit. Rovněž jsou zde pomocí geometrie odvozeny hustoty rozdělení  $\chi_n^2$ ,  $t_n$  a  $F_n$ ; tento způsob byl inspirován některými postupy R. A. Fishera a považujeme jej za další mimořádně efektivní ukázkou využití geometrie ve statistice.

Značný prostor je v druhé části věnován studiu pojmu kolmosti; ačkoli řada zde odvozených tvrzení nemá přímý vztah ke zbytku práce, považujeme tyto kapitoly za její nepostradatelnou součást. Odvoláváme se totiž často na geometrickou intuici, ta však nemůže být spolehlivým vodítkem, není-li patřičně kultivována pečlivými formálními důkazy.

Ve třetí části podáváme stručný popis historie spolupráce W.S. Gosseta a R. A. Fishera, kterou považujeme za počátek moderní matematické statistiky. Dále je na příkladu několika článků analyzován způsob Fisherova geometrického myšlení. Práci uzavírá krátká úvaha nad dalším vývojem geometrického přístupu. Zde uvedený přehled literatury je ovšem pouze orientační; další úsilí by mělo být zaměřeno právě tímto směrem.

# Přehled použité literatury

- [1] Anděl, J.: *Matematická statistika*. SNTL, Praha, 1985.
- [2] Anděl, J.: *Statistické metody*. Matfyzpress, Praha, 1998.
- [3] Anděl, J.: *Základy matematické statistiky*. Matfyzpress, Praha, 2005.
- [4] Bečvář, J.: *Lineární algebra*. Matfyzpress, Praha, 2005.
- [5] Box, J.F.: *R. A. Fisher. The Life of a Scientist*. John Wiley & Sons, New York, 1978.
- [6] Box, J.F.: *Gosset, Fisher and the t Distribution*. *The American Statistician* 35 (1981), 61–66.
- [7] Box, E. P. G., Hunter, W. G., Hunter, J. S.: *Statistics for Experimenters*. John Wiley & Sons, New York, 1978.
- [8] Durbin, J., Kendall, M. G.: *The Geometry of Estimation*. *Biometrika* 38 (1951), 150–158.
- [9] Eaton, M. L.: *William H. Kruskal and the Development of Coordinate-Free Methods*. *Statistical Science* 22 (2007), 264–265.
- [10] Eisenhart, C.: *On the Transition From “Student’s”  $z$  to “Student’s”  $t$* . *The American Statistician* 33 (1979), 6–10.
- [11] Fisher, R. A.: *On an Absolute Criterion for Fitting Frequency Curves*. *Messenger of Mathematics* 41 (1912), 507–521.
- [12] Fisher, R. A.: *Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population*. *Biometrika* 10 (1915), 507–521.
- [13] Fisher, R. A.: *A Mathematical Examination of the Methods of Determining the Accuracy of an Observation by the Mean Error, and by the Mean Square Error*. *Monthly Notices of the Royal Astronomical Society* 80 (1920), 758–770.
- [14] Fisher, R. A.: *Note on Dr Burnside’s Recent Paper on Error of Observation*. *Proceedings of the Cambridge Philosophical Society* 21 (1923), 655–658.
- [15] Fisher, R. A.: *Applications of “Student’s” Distribution*. *Metron* 5 (1925), 90–104.
- [16] Fisher, R. A.: *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh and London, 1925 [odkaz v textu je na 6. vydání z roku 1936].

- [17] Herr, D. G.: *On the History of the Use of Geometry in the General Linear Model*. The American Statistician 34 (1980), 43–47.
- [18] Kruskal, W. H.: *The Coordinate-free Approach to Gauss-Markov Estimation and Its Application to Missing and Extra Observations*. 4th Berkeley Symposium on Mathematical Statistics and Probability 1 (1961), 435–451.
- [19] Kruskal, W. H.: *When Are Gauss-Markov and Least Squares Estimators Identical? A Coordinate-free Approach*. Annals of Mathematical Statistics 39 (1968), 70–75.
- [20] Kruskal, W. H.: *The Geometry of Generalized Inverses*. Journal of the Royal Statistical Society, Ser. B 37 (1975), 272–283.
- [21] Kubát, V., Trkiovská, D.: *Analytická geometrie v afinních a eukleidovských prostorech*. Matfyzpress, Praha, 2011.
- [22] McMullen, L.: “Student” as a Man. Biometrika 30 (1939), 205–210.
- [23] Pázman, A.: *Geometrické metody v matematickej štatistike*. Pokroky matematiky, fyziky a astronomie 33 (1988), 314–326.
- [24] Pearson, E. S.: “Student”. *A Statistical Biography of William Sealy Gosset*. Clarendon Press, Oxford, 1990.
- [25] Rao, C. R., Rao, M. B.: *Matrix Algebra and Its Applications to Statistics and Econometrics*. World Scientific Publishing, Singapore, 1998.
- [26] Rényi, A.: *Teorie pravděpodobnosti*. Academia, Praha, 1972.
- [27] Saville, D. J., Wood, G. R.: *Statistical Methods: The Geometric Approach*. Springer-Verlag, New York, 1991.
- [28] “Student”: *The Probable Error of a Mean*. Biometrika 6 (1908), 1–25.
- [29] “Student”: *Probable Error of a Correlation Coefficient*. Biometrika 6 (1908), 302–310.
- [30] Tjur, T.: *Analysis of Variance Models in Orthogonal Designs*. International Statistical Review 52 (1984), pp. 33–65.
- [31] Wichura, M. J.: *The Coordinate-Free Approach to Linear Models*. Cambridge University Press, New York, 2006.
- [32] Zabell, J. S.: *On Student’s 1908 Article “The Probable Error of a Mean”*. Journal of the American Statistical Association 103 (2008), 1–7.
- [33] Zabell, S.: *A conversation with William Kruskal*. Statistical Science 9 (1994), 285–303.
- [34] Zvára, K., Štěpán, J.: *Pravděpodobnost a matematická statistika*. Matfyzpress, Praha, 2006.
- [35] Zvára, K.: *Regrese*. Matfyzpress, Praha, 2008.