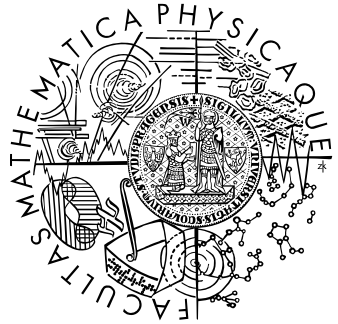
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Adrián Andráš

William S. Gosset a jeho význam pro rozvoj teorie a praxe statistiky

Katedra pravděpodobnosti a matematické statistiky
Vedoucí bakalářské práce: Mgr. Michal Kulich, Ph.D.

Studijní program: Matematika
Studijní obor: Obecná matematika

2009

Ďakujem Mgr. Michalovi Kulichovi, Ph.D. za odborné vedenie mojej bakalárskej práce, za cennú pomoc a čas, ktorý mi pri písaní venoval. Ďakujem tiež Mgr. Šárke Došlej, MSc. za lepšie porozumenie danej problematike a za pomoc pri numerických výpočtoch použitých v tejto práci. Veľká vďaka patrí aj PhDr. Milene Režnej a Veronike Stankovianskej, ktoré mi pomáhali s prekladom anglických odborných textov a tiež Naďke Langovej za užitočné rady pri estetickom spracovaní textu.

Prehlasujem, že som svoju bakalársku prácu napísal samostatne a výhradne s použitím citovaných prameňov. Súhlasím so zapožičaním práce a jej zverejňovaním.

V Prahe dňa 7. 8. 2009 Adrián Andráš

OBSAH

Názov práce: William S. Gosset a jeho význam pro rozvoj teorie a praxe statistiky
Autor: Adrián Andráš
Katedra: Katedra pravděpodobnosti a matematické statistiky
Vedúci bakalárskej práce: Mgr. Michal Kulich, Ph.D.
e-mail vedúceho: kulich@karlin.mff.cuni.cz

Abstrakt: Táto práca je písaná na počesť stého výročia Gossetovho významného článku „Pravdepodobná chyba priemeru“. Úvod je venovaný životopisu Williama Sealyho Gosseta, no prevažná časť sa už venuje jeho článku. Približuje nám ho, hovorí o tom, ako Gossetovo (Studentovo) t-rozdelenie vzniklo, akých chýb sa Gosset dopustil pri výpočtoch a akej filozofie sa držal. Taktiež (stručne) rozoberá jeho druhý významný článok „Pravdepodobná chyba korelačného koeficientu“, ktorý sa venuje rozdeleniu výberového korelačného koeficientu za predpokladu náhodného výberu z dvojrozmerného normálneho rozdelenia. V závere sú spomenutí významní štatistici a matematici, ktorí venovali Gossetovmu článku značnú pozornosť, prípadne čerpali z jeho článkov niektoré závery. Posledná kapitola porovnáva t-rozdelenie v súčasnej štatistike s tou spred 100 rokov.

Kľúčové slová: Gosset, Studentovo *t* – rozdelenie, pravdepodobná chyba, výberový korelačný koeficient

Title: William S. Gosset and his contribution to the advancement of statistical theory and practice
Author: Adrián Andráš
Department: Department of Probability and Mathematical Statistics
Supervisor: Mgr. Michal Kulich, Ph.D.
Supervisor's e-mail address: kulich@karlin.mff.cuni.cz

Abstract: This thesis is written in honour of the centenary of Gosset’s seminal paper “The Probable Error of a Mean.” While the introduction is devoted to William Sealy Gosset’s biography, the rest focuses on Gosset’s (Student’s) article. The work makes the original text accessible as it explains the origins of Student’s *t* – distribution and Gosset’s errors in calculation in addition to the concepts he adopted. Rather briefly, the paper equally deals with another Gosset’s influential article “The Probable Error of a Correlation Coefficient,” addressing the distribution of the sample correlation coefficient in the case of a random sample from a bivariate normal distribution. A number of distinguished mathematicians and statisticians, either paying considerate attention to Gosset’s work or basing their results on his findings, are mentioned. The conclusive chapter is concerned with the comparison of the *t* – distribution as perceived in contemporary statistics as opposed to the statistics of a century ago.

Keywords: Gosset, Student’s *t* – distribution, the probable error, a sample correlation coefficient

1 William Sealy Gosset

1.1 Životopis

William Sealy Gosset (Student)

(1876 – 1937)

William Sealy Gosset sa narodil 13. júna 1876 v Caterbury, Veľká Británia (VB) a zomrel 16. októbra 1937, Beaconsfield, VB. Študoval na Winchester College a New College na Oxforde, VB. Počas štúdia bol ocenený niekoľkými titulmi za jeho vedomosti z matematiky („First in the Mathematical Moderations examination“) a z chémie („First-Class Degree in Chemistry“).

Po skončení školy ho zamestnal Arthur Guiness do svojej spoločnosti v Dubline, Írsku v roku 1899, ktorá sa zaoberala výrobou piva. Gosset si postupom času v práci začal uvedomovať dôležitosť štatistických metód pri postupoch, ktoré sa vo firme prevádzali a obzvlášť tiež problém malých výberov. To Gosseta donútilo, aby sa začal hlbšie venovať práve týmto problémom (zhrnul to v nepublikovanom zázname „Aplikácia ‘Zákonu chyby’ na prácu pivovaru“ datovanom 3. novembra 1904). Keď nenašiel potrebné informácie v literatúre, dohodol si stretnutie s Karlom Pearsonom (významným štatistikom) v júli 1905. A neskôr v rokoch 1906-1907 s povolením Guinessa, Gosset strávil dva semestre akademického roku v Pearsonovom biometrickom laboratóriu na univerzite v Londýne (University College London). Problémy, ktorými sa zaoberal, zhrnul vo vedeckom časopise *Biometrika*, ktorého spoluzakladateľom bol Pearson. Počas 25 ročného obdobia celkovo Gosset publikoval v *Biometrike* 14 článkov, ktoré napísal pod pseudonymom *Student*.

Podrobnejší Studentov životopis je možné nájsť v knihe [6].

1.2 Prečo práve Student?

Touto otázkou sa zaoberal článok [5] str. 189, v ktorom sa vysvetľuje Gossetovo tvrdenie, že jeho matematické a filozofické závery nebolo možné použiť na praktické využitie u konkurencie, no nakoniec ich mohol publikovať, ale len pod pseudonymom, aby sa zabránilo ťažkostiam so zvyškom personálu.

No taktiež sa stretávame s tvrdením, že anonymita bola vyžadovaná priamo Guinnessom, aby sa konkurencia nedozvedela, že je veľmi užitočné zamestnávať štatistikov. Toto tvrdenie je zase úplne protichodné Hotellingovmu (matematickému štatistikovi a významnému ekonomickému teoretikovi), podľa ktorého anonymita bola potrebná kvôli pivovarníkom „vo vnútri“ a nie tých „zvonka“.

Pseudonym Student bolo vymyslené Christopherom Diggesom La Toucheom, ktorý bol výkonným riaditeľom pivovaru.

1.3 Studentova štatistická filozofia

Student (odteraz mu budeme takto stále hovoriť, lebo pseudonym sa zachoval v štatistike až do súčasnosti) bol ovplyvnený spočiatku *Bayesovou filozofiou* aj preto, že čítal články a knihy od Karla Pearsona.

Pre priblíženie Bayesovej filozofie a základných princípov *bayesovských metód* uvedieme jednoduchý príklad, ktorý je možné nájsť v Andělovej knihe [2].

Máme veľkú sériu výrobkov. Nech N je počet kusov v tejto sérii a M značí počet vadných výrobkov. Číslo M nie je známe. Pravdepodobnosť, že vybraný kus bude chybný, je rovná $p = M/N$ (tá tiež nie je známa).

Uskutočníme tzv. *výber s vrátením*. To znamená, že preskúmaný výrobok sa znova vráti do série a v ďalších ťahoch môže byť znova vytiahnutý. Tým sa docieli, že jednotlivé ťahy budú na sebe nezávislé.

Pravdepodobnosť, že z n vytiahnutých výrobkov bude práve m vadných je

$$r(m \mid p) = \binom{n}{m} p^m (1-p)^{n-m}, \quad 0 \le m \le n. \tag{1.1}$$

Ak vo výbere bolo nájdených práve *m* vadných výrobkov, je najlepším nestranným odhadom parametru *p* veličina $p^* = m/n$.

Odôvodnenie: Nech séria výrobkov pozostáva z náhodných rovnako rozdelených veličín $X_1,\ldots,X_n$, kde $X_i \sim Alt(p)$.

Nech $P(X_i = 1) = p$ je pravdepodobnosť, že vybraný výrobok bude chybný a $P(X_i = 0) = 1 - p$ je pravdepodobnosť, že vybraný výrobok nebude chybný.

Ak $m = \sum_{i=1}^{n} X_i$ označuje počet vadných výrobkov vo výbere a vieme, že X_i sú rovnako rozdelené, potom

$$Ep^* = \frac{E\sum_{i=1}^{n} X_i}{n} = \frac{\sum_{i=1}^{n} EX_i}{n} = \frac{n \cdot EX_1}{n} = \frac{np}{n} = p,$$

z čoho plynie nestrannosť p^*.

Rozptyl p^* je rovný $\operatorname{var} p^* = p(1-p)/n$.

Ak sa stane, že výroba takých rovnako veľkých sérií prebieha už dlhšiu dobu a vedie sa dostatočne podrobná evidencia, je možné zistiť po čase presne, koľko bolo vyrobených vadných výrobkov v každej sérii, a teda sú známe aj podiely $p_1,\ldots,p_K$ vadných výrobkov v každej zo skôr vyrobených *K* sérií. Keďže podiely nie sú obvykle totožné, ľahko prijmeme názor, že podiel *p* vadných výrobkov v sérii je náhodná veličina, o ktorej rozdelení nás informuje náhodný výber $p_1,\ldots,p_K$.

Keďže náhodná veličina *p* môže nadobúdať len hodnoty $0, 1/N, \ldots, N/N$, je diskrétna. Ak je *N* veľké, je možné distribučnú funkciu tejto veličiny dostatočne presne aproximovať nejakou absolútne spojitou distribučnou funkciou.

V našom prípade budeme predpokladať následovnú aproximáciu náhodnej veličiny *p* B-rozdelením s hustotou

$$q(p) = \frac{1}{\mathrm{B}(a,b)}\, p^{a-1}(1-p)^{b-1}, \qquad 0 < p < 1, \tag{1.2}$$

kde $a>0$, $b>0$ sú známe parametre napr. dostatočne presne určené z hodnôt $p_1,\ldots,p_K$, ktoré považujeme za výber z rozdelenia s hustotou (1.2). Hustota *q* je marginálna hustota veličiny *p* a nazýva sa *apriórna hustota.* Predpokladá sa totiž, že

je známa skôr, než sa začneme zaoberať danou skupinou výrobkov. Pravdepodobnosti (1.1) potom tvoria podmienenú hustotu náhodnej veličiny *m* pri danej hodnote *p*.

Aposteriórnou hustotou nazývame výraz $\pi(p \mid m)$, ktorý je podľa Bayesovej vety rovný

$$\pi(p \mid m) = \frac{\frac{1}{\mathrm{B}(a,b)}\, p^{a-1}(1-p)^{b-1}\binom{n}{m} p^m (1-p)^{n-m}}{\int_0^1 \frac{1}{\mathrm{B}(a,b)}\, p^{a-1}(1-p)^{b-1}\binom{n}{m} p^m (1-p)^{n-m}\, \mathrm{d}p} =$$

$$= \frac{1}{\mathrm{B}(a+m, b+n-m)}\, p^{a+m-1}(1-p)^{b+n-m-1}. \qquad (1.3)$$

Na základe aposteriórnej hustoty $\pi(p \mid m)$ je možné zostrojiť intervalový odhad pre veličinu *p* následovne.

Nech čísla *D* a *H* $(0 \le D < H \le 1)$ splňujú podmienky

$$\int_D^H \pi(p \mid m)\, \mathrm{d}p = 1-\alpha, \qquad \int_0^D \pi(p \mid m)\, \mathrm{d}p = \alpha/2, \qquad \int_H^1 \pi(p \mid m)\, \mathrm{d}p = \alpha/2,$$

kde $\alpha \in (0,1)$. Potom platí

$$P(D < p < H \mid m) = 1-\alpha. \qquad (1.4)$$

Interval *(D, H)* je *bayesovský interval spoľahlivosti* pre *p.*

Ak sú čísla *D* a *H* zvolené tak, aby spĺňali podmienku (1.4), môžeme pristúpiť k testovaniu hypotéz pre parameter *p.*

Predpokladáme test hypotézy $\mathrm{H}_0 : p \in A$, kde A je ľubovolný interval na (0, 1). Ak celý interval *A* padne mimo (D, H), hypotézu H_0 zamietame. V tomto prípade je $P(A \mid m) \le \alpha$.

Nedá sa však založiť test len na hodnote $P(A \mid m)$ bez porovnania s intervalom (*D, H*). Pretože, ak by interval A bol príliš krátky, tak $P(A \mid m)$ by bola veľmi malá a to by spôsobilo zamietnutie H_0, napriek tomu, že poloha *A* by vzhľadom k hustote (1.3) nebola nijako „podozrivá“.

Ak sa vrátime k nášmu príkladu, parameter *p* (binomického rozdelenia) nebol považovaný za konštantu, ale za náhodnú veličinu. O tejto možnosti prvýkrát hovoril Thomas Bayes, a preto sa jeho metódy nazývajú *bayesovské.*

Najväčším kladom bayesovských postupov je ten, že sa neobmedzujú len na samotný výber, ale berú do úvahy aj doplňujúce informácie. Táto výhoda sa však stráca, ak doplňujúce informácie nemáme k dispozícii.

Konkrétne využitie Bayesovej filozofie Student sčasti uplatnil napr. vo svojom článku [9], ku ktorému sa neskôr dostaneme.

V ďalšej kapitole sa však pozrieme najskôr na Studentov významnejší článok [10].

2 Pravdepodobná chyba priemeru

2.1 Úvod

V celom článku sa predpokladá náhodný výber n $(n \geq 2)$ jedincov $Y_1,\ldots,Y_n$ z populácie s normálnym rozdelením, čiže $Y_i \sim N(\mu,\sigma^2)$, kde μ značí strednú hodnotu náhodnej veličiny Y_i a σ označuje jej smerodajnú odchýlku.

Ďalej vieme, že *výberový priemer* $\bar{Y}$ má normálne rozdelenie (kde $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$) so strednou hodnotou μ a smerodajnou odchýlkou $\frac{\sigma}{\sqrt{n}}$, čiže $\bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

Označíme *s* ako výberovú smerodajnú odchýlku náhodného výberu $X_1,\ldots,X_n$, kde

$$X_i = Y_i - \mu \qquad \text{a} \qquad s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} \quad \text{pre } i = 1,\ldots,n.$$

Cieľom Studentovho článku, bolo určiť rozdelenie náhodnej veličiny

$$z = \frac{\bar{Y} - \mu}{s}$$

a pre túto veličinu zostaviť štatistické tabuľky distribučnej funkcie $F(z)$.

2.2 Rozdelenie výberovej smerodajnej odchýlky

Ak uvažujeme výberovú smerodajnú odchýlku *s* náhodného výberu $X_1,\ldots,X_n$ a náhodnú veličinu X_i ako v úvode (kap. 2.1), potom

$$s^2 = \frac{\sum_{i=1}^{n} X_i^2}{n} - \left(\frac{\sum_{i=1}^{n} X_i}{n}\right)^2 = \frac{\sum_{i=1}^{n} X_i^2}{n} - \frac{\sum_{i=1}^{n} X_i^2}{n^2} - \frac{2\sum_{\substack{i=1 \\ i\neq}}^{n}\sum_{j=1}^{n} X_i X_j}{n^2}.$$

Strednú hodnotu veličiny s^2 spočítame ako

$$Es^2 = \frac{E\sum_{i=1}^{n} X_i^2}{n} - E\left(\frac{\sum_{i=1}^{n} X_i}{n}\right)^2 = \frac{E\sum_{i=1}^{n} X_i^2}{n} - \frac{E\sum_{i=1}^{n} X_i^2}{n^2} - \frac{2\sum_{\substack{i=1 \\ i\neq j}}^{n}\sum_{j=1}^{n} EX_iX_j}{n^2} =$$

$$= \frac{\sum_{i=1}^{n} EX_i^2}{n} - \frac{\sum_{i=1}^{n} EX_i^2}{n^2} - \frac{2\sum_{\substack{i=1 \\ i\neq j}}^{n}\sum_{j=1}^{n} EX_iX_j}{n^2} =$$

$$= \frac{nEX_i^2}{n} - \frac{nEX_i^2}{n^2} - \frac{2\sum_{\substack{i=1 \\ i\neq j}}^{n}\sum_{j=1}^{n} EX_iX_j}{n^2} = EX_i^2 - \frac{EX_i^2}{n} - \frac{2\sum_{\substack{i=1 \\ i\neq j}}^{n}\sum_{j=1}^{n} EX_iX_j}{n^2}. \quad (2.1)$$

Keďže

$$EX_i = EY_i - E\mu = \mu - \mu = 0,$$

$$\operatorname{var} X_i = EX_i^2 - \left(EX_i\right)^2 = EX_i^2 = E\left(Y_i - \underbrace{\mu}_{=EY_i}\right)^2 = \sigma^2,$$

dostávame, že

$$X_i \sim N(0, \sigma^2). \quad (2.2)$$

Vieme, že $X_1, \ldots, X_n$ sú nekorelované náhodné veličiny (lebo sa jedná o náhodný výber z normálneho rozdelenia a odčítanie μ tomu nijak neublíži). Potom môžeme písať, že

$$\operatorname{corr}(X_i, X_j) = 0, \text{ čiže } \operatorname{cov}(X_i, X_j) = 0 = EX_iX_j - \underbrace{\underbrace{EX_i}_{=0} EX_j}_{=0}, \text{ a teda } EX_iX_j = 0.$$

Teda posledný člen $\frac{2\sum_{\substack{i=1 \\ i\neq j}}^{n}\sum_{j=1}^{n} EX_iX_j}{n^2}$ v (2.1) vypadne a po zavedení $\mu_k = EX^k$, dostaneme

$$Es^2 = \mu_2 \frac{(n-1)}{n},$$

čo je v podstate *1. moment* (označíme ho M_1') veličiny s^2.

Student sa pokúsil spočítať aj ďalšie momenty veličiny s^2.

Výpočet *2. momentu* (M_2') s^2:

$$s^4 = \left\{ \frac{\sum_{i=1}^{n} X_i^2}{n} - \left(\frac{\sum_{i=1}^{n} X_i}{n} \right)^2 \right\}^2 = \left(\frac{\sum_{i=1}^{n} X_i^2}{n} \right)^2 - \frac{2\sum_{i=1}^{n} X_i^2}{n} \left(\frac{\sum_{i=1}^{n} X_i}{n} \right)^2 + \left(\frac{\sum_{i=1}^{n} X_i}{n} \right)^4 = \tag{2.3}$$

$$= \frac{\sum_{i=1}^{n} X_i^4}{n^2} + \frac{2\sum_{\substack{i=1 \\ i\neq}}^{n} \sum_{j=1}^{n} X_i^2 X_j^2}{n^2} - \frac{2\sum_{i=1}^{n} X_i^4}{n^3} - \frac{4\sum_{\substack{i=1 \\ i\neq}}^{n} \sum_{j=1}^{n} X_i^2 X_j^2}{n^3} + \frac{\sum_{i=1}^{n} X_i^4}{n^4} + \frac{6\sum_{\substack{i=1 \\ i\neq}}^{n} \sum_{j=1}^{n} X_i^2 X_j^2}{n^4}$$

$+$ ďalšie výrazy, pri ktorých je X_i umocnené na nepárnú mocninu.

Vieme, že platí $X_i \sim N(0, \sigma^2)$ a pre $k \in \mathbf{N}$,

$$E(X_i - EX_i)^k \overset{EX_i=0}{=} EX_i^k = \int_{-\infty}^{+\infty} x^k f(x)\,\mathrm{d}x = \int_{-\infty}^{+\infty} x^k \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}\,\mathrm{d}x = \tag{2.4}$$

$$= \begin{cases} 0 & \text{pre } k \text{ nepárne,} \\ (k-1)(k-3)\cdot\ldots\cdot 1\cdot\sigma^k & \text{pre } k \text{ párne.} \end{cases}$$

Keďže sú $X_1, \ldots, X_n$ nekorelované náhodné veličiny, dostávame pre $l \in \mathbf{N}$ a i ≠ j

$$EX_i^k X_j^l = \begin{cases} 0 \text{ (plynie z (2.4))} & \text{pre } k \text{ alebo } l \text{ nepárné,} \\ EX_i^k EX_j^l & \text{pre } k \text{ aj } l \text{ párne,} \end{cases} \tag{2.5}$$

Teda všetky výrazy, v ktorých sa vyskytuje X_i umocnené na nepárnu mocninu pri výpočte Es^4 v (2.3) zmiznú a keďže $\sum_{i=1}^{n} X_i^4$ má n členov a $\sum_{\substack{i=1 \\ i\neq}}^{n} \sum_{j=1}^{n} X_i^2 X_j^2$ má

$\frac{n(n-1)}{2}$ členov, ostane nám

$$M_2' = Es^4 = \frac{\mu_4}{n} + \mu_2^2 \frac{(n-1)}{n} - \frac{2\mu_4}{n^2} - 2\mu_2^2 \frac{(n-1)}{n^2} + \frac{\mu_4}{n^3} + 3\mu_2^2 \frac{(n-1)}{n^4} = \tag{2.6}$$

$$= \frac{\mu_4}{n^3}\left\{n^2 - 2n + 1\right\} + \frac{\mu_2^2}{n^3}(n-1)\left\{n^2 - 2n + 3\right\}.$$

Z (2.4) plynie, že

$$\mu_2 = EX^2 = \sigma^2,$$
$$\mu_4 = EX^4 = 3\sigma^4,$$

čiže

$$\mu_4 = 3\mu_2^2.$$

Teda po dosadení do (2.6) dostaneme

$$M_2' = \mu_2^2 \frac{(n-1)}{n^3}\{3n-3+n^2-2n+3\} = \mu_2^2 \frac{(n-1)(n+1)}{n^2}.$$

Analogicky sa spočíta *3. a 4. moment* veličiny s^2 a dostaneme

$$M_3' = \mu_2^3 \frac{(n-1)(n+1)(n+3)}{n^3},$$

$$M_4' = \mu_2^4 \frac{(n-1)(n+1)(n+3)(n+5)}{n^4}.$$

Čiže sa črtá pravidlo formovania jednotlivých momentov s^2.

Označme teraz M_R ako *R – tý centrálny moment* veličiny s^2. Potom

$$M_2 = E(s^2 - Es^2)^2 = \underbrace{E(s^2)^2}_{M_2'} - (E(s^2))^2 = \mu_2^2 \frac{(n-1)(n+1)}{n^2} - \left(\mu_2 \frac{(n-1)}{n}\right)^2$$

$$= 2 \cdot \frac{\mu_2^2 (n-1)}{n^2}$$

a analogicky sa spočítajú ďalšie momenty (podrobnejší výpočet je možné nájsť v Studentovom článku [10]).

$$M_3 = E(s^2 - Es^2)^3 = 8\mu_2^3 \frac{(n-1)}{n^3},$$

$$M_4 = E(s^2 - Es^2)^4 = 12\mu_2^4 \frac{(n-1)(n+3)}{n^4}.$$

Ďalej Student zistil, že ak spočíta šikmosť náhodnej veličiny s^2

$$\beta_1 = \left[\frac{E(s^2 - Es^2)^3}{\left(\sqrt{E(s^2 - Es^2)^2}\right)^3}\right]^2 = \frac{M_3^2}{M_2^3} = \frac{8}{n-1},$$

a následne špicatosť

$$\beta_2 = \frac{E(s^2 - Es^2)^4}{\left(\sqrt{E(s^2 - Es^2)^2}\right)^4} = \frac{M_4}{M_2^2} = \frac{3(n+3)}{n-1},$$

vyjde mu

$$2\beta_2 - 3\beta_1 - 6 = 2 \cdot \frac{3(n+3)}{n-1} - 3 \cdot \frac{8}{n-1} - 6 = 0$$

a to odpovedá *Pearsonovému rozdeleniu typu III*, ktoré v súčasnosti poznáme pod názvom gama rozdelenie.

Teda hustota náhodnej veličiny s^2 bude vyzerať ako

$$f(x) = Cx^p e^{-\gamma x},$$

kde

$$\gamma = 2\frac{M_2}{M_3} = \frac{4\mu_2^2(n-1)n^3}{8n^2\mu_2^3(n-1)} = \frac{n}{2\mu_2},$$
$$p = \frac{4}{\beta_1} - 1 = \frac{n-1}{2} - 1 = \frac{n-3}{2},$$

z čoho plynie hustota

$$f(x) = Cx^{\frac{n-3}{2}} e^{-\frac{nx}{2\mu_2}}, \quad \text{kde} \quad x > 0.$$

Poznámka: Presné zdôvodnenie *Pearsonového rozdelenia typu III.* je možné nájsť v Pearsonovom článku [7].

Ak si označíme

$$I = C\int_0^{+\infty} x^{\frac{n-3}{2}} e^{-\frac{nx}{2\mu_2}} \mathrm{d}x,$$

potom

$$1 = \frac{C\int_0^{+\infty} x^{\frac{n-3}{2}} e^{-\frac{nx}{2\mu_2}} \mathrm{d}x}{I},$$

z čoho plynie, že $\frac{C\,x^{\frac{n-3}{2}} e^{-\frac{nx}{2\mu_2}}}{I}$ je hustota s^2.

Potom <u>1. moment</u> náhodnej veličiny s^2 je (keďže gama funkcia nebola dobre známa, Student použil metódu per partes)

$$Es^2 = \frac{C\int_0^{+\infty} \overbrace{x^{\frac{n-1}{2}}}\cdot x^{\frac{n-3}{2}} e^{-\frac{nx}{2\mu_2}} \mathrm{d}x}{I} \overset{\text{per partes}}{=} \left| \begin{array}{ll} u = x^{\frac{n-1}{2}} & u' = \left(\frac{n-1}{2}\right) x^{\frac{n-3}{2}} \\ v' = e^{-\frac{nx}{2\mu_2}} & v = \frac{e^{-\frac{nx}{2\mu_2}}}{-\frac{n}{2\mu_2}} \end{array} \right| =$$

$$= \frac{C}{I}\left\{\left[-\frac{2\mu_2}{n}x^{\frac{n-1}{2}}e^{-\frac{nx}{2\mu_2}}\right]_{x=0}^{+\infty} + \frac{2\mu_2}{n}\left(\frac{n-1}{2}\right)\underbrace{\int_0^{+\infty} x^{\frac{n-3}{2}}e^{-\frac{nx}{2\mu_2}}\mathrm{d}x}_{=\frac{I}{C}}\right\} =$$

$$= \left(-\frac{2\mu_2}{n}\cdot\lim_{x\to\infty}\frac{x^{\frac{n-1}{2}}}{e^{\frac{nx}{2\mu_2}}} - 0\right) + \frac{C}{I}\mu_2\frac{(n-1)}{n}\frac{I}{C} \overset{(\#)}{=} 0 + \frac{C}{I}\mu_2\frac{(n-1)}{n}\frac{I}{C} = \mu_2\frac{(n-1)}{n}.$$

Analogicky sa spočítajú druhy, tretí, štvrtý moment, ... a dostaneme

$$M_2' = \mu_2^2\frac{(n-1)(n+1)}{n^2},$$

$$M_3' = \mu_2^3\frac{(n-1)(n+1)(n+3)}{n^3},$$

$$M_4' = \mu_2^4\frac{(n-1)(n+1)(n+3)(n+5)}{n^4}.$$

Teda rozdelenie s^2 má všetky momenty rovnaké ako Pearsonovo rozdelenie typu III a Student prehlásil, že rozdelenie s^2 je teda práve Pearsonovým rozdelením typu III. dané hustotou $f(x) = Cx^{\frac{n-3}{2}}e^{-\frac{nx}{2\mu_2}}$. (V súčasnosti je známe, že neplatí tvrdenie: Rovnosť momentov dvoch rozdelení implikuje aj rovnosť rozdelení. Viď príklad v Andělovej knihe [3] str. 16. Teda Studentov postup tu nie je úplne ošetrený.)

Keďže už poznáme rozdelenie náhodnej veličiny s^2, nie je problém konečne nájsť hľadané rozdelenie náhodnej veličiny *s*. K tomu použijeme *vetu o transformácií náhodnej veličiny.*

Veta (O transformácií náhodnej veličiny) *Nech X má spojitú distribučnú funkciu F. Nech* $F'(x) = f(x)$ *existuje všade až na konečne mnoho bodov. Nech je t rýdzo monotónna funkcia, ktorá má všade nenulovú deriváciu. Položme* $Y = t(X)$. *Označme τ inverznú funkciu k t. Potom Y má hustotu*

$$g(y) = f[\tau(y)]|\tau'(y)|.$$

*Dôkaz. V*iď Andělovu knihu [3] str. 48.

V našom prípade môžeme označiť

(#) Na výpočet limity použijeme $(n-1)/2$ krát L'Hospitalovo pravidlo pre prípad "$\frac{\infty}{\infty}$"

$$X = s^2, \qquad \text{kde } \phi(s^2) = Cx^{\frac{n-3}{2}} e^{-\frac{nx}{2\mu_2}} \text{ je hustota X,}$$
$$Y = s = \sqrt{\mathrm{X}}, \quad \text{kde } \psi(\mathrm{s}) \text{ je hustota Y a } x = y^2.$$

Najskôr však potrebujeme overiť predpoklady vety, aby sme zistili, či vetu je možné aplikovať na náš prípad. Čiže máme zobrazenie

$$t : X \to \sqrt{X}\ .$$

Deriváciu *t* spočítame ako

$$t'(x) = \frac{1}{2\sqrt{x}} \neq 0, \quad \forall x \in \Re^+ \setminus \{0\}.$$

A keďže

$$t'(x) = \frac{1}{2\sqrt{x}} > 0, \quad \forall x \in \Re^+ \setminus \{0\},$$

dostávame rýdzo-monotónnosť t.

Keďže všetky predpoklady sú splnené, môžeme použiť vetu nasledujúcim spôsobom:

$$\psi(y) = \phi[\tau(y)]\,|\tau'(y)|,$$

kde

$$\tau(y) = x = y^2,$$

čiže po dosadení do $\tau(y)$ a $\tau'(y)$ dostaneme

$$\psi(y) = \phi(y^2) \cdot 2y\,.$$

Keď nahradíme *y* premennou s, môžeme písať

$$\psi(s) = 2 \cdot C \cdot s \cdot \left(s^2\right)^{\frac{n-3}{2}} e^{-\frac{ns^2}{2\mu_2}} = 2 \cdot C \cdot s^{n-2} e^{-\frac{ns^2}{2\mu_2}}, \quad \text{kde } s > 0,$$

čiže sme nakoniec dostali hustotu *s*.

Teda ak $A = 2C$, potom

$$f(s) = A \cdot s^{n-2} e^{-\frac{ns^2}{2\mu_2}} \tag{2.7}$$

je hustota výberovej smerodajnej odchýlky *s* výberu z populácie, ktorého náhodné veličiny majú normálne rozdelenie so smerodajnou odchýlkou σ ($X_i \sim N(0, \sigma^2)$).

Chceme teraz spočítať konštantu *A*. Tú vypočítame z hustoty smerodajnej odchýlky *s* a hustota musí spĺňať

$$(\text{Plocha pod krivkou } f(\mathrm{s})) = A \int_0^{+\infty} s^{n-2} e^{-\frac{ns^2}{2\mu_2}} \mathrm{d}s\,. \tag{2.8}$$

Nech

$$I_p = \int_0^{+\infty} s^p e^{-\frac{ns^2}{2\mu_2}} \mathrm{d}s$$

a z (2.2) plynie, že

$$\mu_2 = \sigma^2\,.$$

Potom

$$\begin{aligned} I_p &= \frac{\sigma^2}{n} \int_0^{+\infty} s^{p-1} \frac{\mathrm{d}}{\mathrm{d}x}(-e^{-\frac{ns^2}{2\sigma^2}})\mathrm{d}s = \\ &\overset{\text{per partes}}{=} \frac{\sigma^2}{n}\left[s^{p-1}(-e^{-\frac{ns^2}{2\sigma^2}}) \right]_{s=0}^{+\infty} + \frac{\sigma^2}{n}(p-1) \int_0^{+\infty} s^{p-2} e^{-\frac{ns^2}{2\sigma^2}} \mathrm{d}s = \\ &= \frac{\sigma^2}{n}(p-1)\, I_{p-2}. \end{aligned}$$

Ak budeme ďalej pokračovať týmto rekurentným postupom, dostaneme

$$\begin{aligned} I_{n-2} &= \left(\frac{\sigma^2}{n}\right)^{\frac{n-2}{2}} (n-3)(n-5)\ldots 3\cdot 1\, I_0 \qquad \text{pre } n \text{ párne,} \\ &= \left(\frac{\sigma^2}{n}\right)^{\frac{n-3}{2}} (n-3)(n-5)\ldots 4\cdot 2\, I_1 \qquad \text{pre } n \text{ nepárne,} \end{aligned} \tag{2.9}$$

kde

$$I_0 = \int_0^{+\infty} e^{-\frac{ns^2}{2\sigma^2}} \mathrm{d}s = \sqrt{\frac{\pi}{2n}}\sigma, \tag{2.10}$$

$$I_1 = \int_0^{+\infty} s\, e^{-\frac{ns^2}{2\sigma^2}} \mathrm{d}s = \frac{\sigma^2}{n}. \tag{2.11}$$

Potom z (2.8), (2.9), (2.10) a (2.11) plynie

$$\begin{aligned} A &= \frac{(\text{Plocha pod krivkou } f(\mathrm{s}))}{(n-3)(n-5)\ldots 3\cdot 1\cdot \sqrt{\frac{\pi}{2}}\left(\frac{\sigma^2}{n}\right)^{\frac{n-1}{2}}} \qquad \text{pre } n \text{ párne,} \\ &= \frac{(\text{Plocha pod krivkou } f(\mathrm{s}))}{(n-3)(n-5)\ldots 4\cdot 2\cdot \left(\frac{\sigma^2}{n}\right)^{\frac{n-1}{2}}} \qquad \text{pre } n \text{ nepárne.} \end{aligned}$$

Dosadením do (2.7) dostaneme

$$\boxed{\begin{aligned} f(s) &= \frac{N}{(n-3)(n-5)\ldots 3\cdot 1}\cdot\sqrt{\frac{\pi}{2}}\left(\frac{n}{\sigma^2}\right)^{\frac{n-1}{2}} s^{n-2}e^{-\frac{ns^2}{2\sigma^2}} \quad \text{pre } n \text{ párne,} \\ &= \frac{N}{(n-3)(n-5)\ldots 4\cdot 2}\cdot\left(\frac{n}{\sigma^2}\right)^{\frac{n-1}{2}} s^{n-2}e^{-\frac{ns^2}{2\sigma^2}} \quad \text{pre } n \text{ nepárne,} \end{aligned}} \tag{2.12}$$

kde N značí plochu pod krivkou $f(s)$. (V súčasnosti je známe, že $N=1$, pretože vo vzorci (2.8) sa A volí tak, aby plocha pod krivkou $f(s)$ bola rovná 1).
(2.12) je už konečný tvar hustoty s výberu z normálneho rozdelenia.

2.3 Korelácia medzi $\overline{X}$ - EX a s výberu z normálneho rozdelenia

Predpokladajme opäť náhodný výber $X_1,\ldots,X_n$ z normálneho rozdelenia $X_i \sim N(0,\sigma^2)$. Viď kap. 2.1.

Najskôr Student ukázal, že náhodné veličiny $\overline{X} - EX \overset{EX=0}{=} \overline{X}$ a s sú nekorelované, argumentoval to tým, že je rovnako pravdepodobná pozícia výberového priemeru ($\overline{X}$) naľavo aj napravo od počiatku 0. (t.j. že platí $\mathrm{P}(\overline{X}>0) = \mathrm{P}(\overline{X}<0) = \frac{1}{2}$ resp. že sa jedná o symetrické rozdelenie.) V dnešnej dobe je známe, že Studentov argument nebol správny a neplatí. Vieme však, že $\overline{X}$ a s^2 sú nezávislé (viď Andělovu knihu [3] str. 70) a z toho plynie nezávislosť $\overline{X}$ a s resp. aj nekorelovanosť $\overline{X}$ a s .

Student ešte zisťoval koreláciu medzi $\left(\overline{X}\right)^2$ a s^2, aby si bol istý výpočtami ďalej.
Označme

$$u^2 := \left(\overline{X}\right)^2, \text{ kde } X = Y_i - \mu \quad \text{pre } i = 1,\ldots,n.$$

Ďalej zavedieme premenné m_1 a M_1 nasledujúcim spôsobom ako

$$m_1 = Eu^2 = E\left(\frac{\sum_{i=1}^{n} X_i}{n}\right)^2 = \frac{E\sum_{i=1}^{n} X_i^2}{n^2} + \underbrace{\frac{2\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n} EX_iX_j}{n^2}}_{z(2.5)=0} = \frac{\mu_2}{n},$$

$$M_1 = Es^2 = \mu_2\frac{(n-1)}{n}.$$

Spočítame

$$u^2s^2 = \left(\frac{\sum_{i=1}^{n} X_i}{n}\right)^2\left[\frac{\sum_{i=1}^{n} X_i^2}{n} - \left(\frac{\sum_{i=1}^{n} X_i}{n}\right)^2\right] =$$

$$= \left(\frac{\sum_{i=1}^{n} X_i^2}{n}\right)^2 + 2\frac{\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n} X_iX_j \cdot \sum_{i=1}^{n} X_i^2}{n^2} - \frac{\sum_{i=1}^{n} X_i^4}{n^4} - \frac{6\cdot\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n} X_iX_j}{n^4} - (\text{zvyšné}$$

členy, ktoré pri strednej hodnote zmiznú),

potom

$$Eu^2s^2 = E\left(\frac{\sum_{i=1}^{n} X_i^2}{n}\right)^2 + 2E\frac{\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n} X_iX_j \cdot \sum_{i=1}^{n} X_i^2}{n^2} - E\frac{\sum_{i=1}^{n} X_i^4}{n^4} - E\frac{6\cdot\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n} X_iX_j}{n^4} =$$

$$= \frac{\mu_4}{n^2} + \mu_2^2\frac{(n-1)}{n^2} - \frac{\mu_4}{n^3} - 3\mu_2^2\frac{(n-1)}{n^3}.$$

Keďže pre koreláciu medzi náhodnými veličinami u^2 a s^2 platí vzťah

$$\operatorname{corr}(u^2, s^2) = \frac{\operatorname{cov}(u^2, s^2)}{\sqrt{\operatorname{var} u^2}\sqrt{\operatorname{var} s^2}} = \frac{Eu^2s^2 - Eu^2Es^2}{\sqrt{\operatorname{var} u^2}\sqrt{\operatorname{var} s^2}} = \frac{Eu^2s^2 - m_1M_1}{\sqrt{\operatorname{var} u^2}\sqrt{\operatorname{var} s^2}},$$

dostávame, že

$$\operatorname{corr}(u^2, s^2)\sqrt{\operatorname{var} u^2}\sqrt{\operatorname{var} s^2} + m_1M_1 = Eu^2s^2. \qquad (2.13)$$

Po dosadení za m_1M_1 a Eu^2s^2 vo vzorci (2.13) s použitím (2.4) dostaneme

$$\operatorname{corr}(u^2, s^2)\sqrt{\operatorname{var} u^2}\sqrt{\operatorname{var} s^2} + \frac{\mu_2}{n}\mu_2\frac{(n-1)}{n} = \frac{\mu_4}{n^2} + \mu_2^2\frac{(n-1)}{n^2} - \frac{\mu_4}{n^3} - 3\mu_2^2\frac{(n-1)}{n^3},$$

$$\operatorname{corr}(u^2, s^2)\sqrt{\operatorname{var} u^2}\sqrt{\operatorname{var} s^2} + \mu_2^2\frac{(n-1)}{n^2} = \frac{3\mu_2^2}{n^2} + \mu_2^2\frac{(n-1)}{n^2} - \frac{3\mu_2^2}{n^3} - 3\mu_2^2\frac{(n-1)}{n^3},$$

$$\operatorname{corr}(u^2, s^2)\sqrt{\operatorname{var} u^2}\sqrt{\operatorname{var} s^2} + \mu_2^2 \frac{(n-1)}{n^2} = \mu_2^2 \frac{(n-1)}{n^2}.$$

Teda

$$\operatorname{corr}(u^2, s^2)\sqrt{\operatorname{var} u^2}\sqrt{\operatorname{var} s^2} = 0, \quad \text{čiže} \quad \operatorname{corr}(u^2, s^2) = 0$$

a vidíme, že veličiny u^2 a s^2 sú nekorelované.

Poznámka Student vo svojej práci [10] použil značenie $R_{u^2 s^2}$ pre $\operatorname{corr}(u^2, s^2)$,

σ_{u^2} pre $\sqrt{\operatorname{var} u^2}$,

σ_{s^2} pre $\sqrt{\operatorname{var} s^2}$.

2.4 Hľadanie hustoty nezávislej na σ

Vzorec $f(s) = \frac{C}{\sigma^{n-1}} s^{n-2} e^{-\frac{ns^2}{2\sigma^2}}$,

kde

$$C = \frac{N}{(n-3)(n-5)\ldots 3\cdot 1} \cdot \sqrt{\frac{2}{\pi}} \cdot n^{\frac{n-1}{2}} \qquad \text{pre } n \text{ párne,}$$

$$= \frac{N}{(n-3)(n-5)\ldots 4\cdot 2} \cdot n^{\frac{n-1}{2}} \qquad \text{pre } n \text{ nepárne,}$$

označuje hustotu smerodajnej odchýlky s náhodného výberu n jedincov populácie z normálneho rozdelenia ($N(0, \sigma^2)$).

Rovnica $f(x) = \frac{\sqrt{n}N}{\sqrt{2\pi}\sigma} e^{-\frac{nx^2}{2\sigma^2}}$ označuje hustotu aritmetického priemeru $\overline{X}$ vyššie zmieneného náhodného výberu. (Vzťah bol získaný z Airyho knihy [1].)

Nech

$$z = \frac{\overline{X}}{s},$$

kde

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu) = \overline{Y} - \mu$$

a μ označuje strednú hodnotu náhodnej veličiny Y_i.

Chceme zistiť rozdelenie náhodnej veličiny *z*. Na to použijeme *vetu o transformácii náhodného vektoru.*

Veta (O transformácii náhodného vektoru) *Nech náhodný vektor* $\boldsymbol{X} = (X_1, \ldots, X_n)'$ *má hustotu p vzhľadom k Lebesgueovej miere v* $\Re^n$. *Nech* $\boldsymbol{t}$ *je zobrazenie z* $\Re^n$ *do* $\Re^n$, *ktoré je regulárne a prosté na takej otvorenej množine G, pre ktorú platí* $\int_G p(x)\mathrm{d}x = 1$. *Označme* $\boldsymbol{\tau}$ *inverzné zobrazenie k* $\boldsymbol{t} : G \to \boldsymbol{t}(G)$. *Potom náhodný vektor* $\boldsymbol{Y} = \boldsymbol{t}(\boldsymbol{X})$ *má hustotu vzhľadom k Lebesgueovej miere a táto hustota je rovná*

$$q(\boldsymbol{y}) = \begin{cases} p[\boldsymbol{\tau}(\boldsymbol{y})]\left|D_{\tau}(\boldsymbol{y})\right| & \text{pre } \boldsymbol{y} \in \boldsymbol{t}(G), \\ 0 & \text{pre } \boldsymbol{y} \notin \boldsymbol{t}(G). \end{cases}$$

*Dôkaz. V*iď Andělova kniha [3] str. 49.

Vrátime sa späť k nášmu problému, a tým je zistiť rozdelenie $z = \dfrac{\overline{X}}{s}$, kde $\overline{X}$ budeme značiť ako *x*.

Pri počítaní združenej hustoty $f(s,x)$ sa Student dopustil nepresnosti. Pretože podľa jeho výpočtov $f(s,x) = f(s)f(x)$, lenže to by platilo, pokiaľ by náhodné veličiny $\overline{X}$ a *s* boli nezávislé. No Student dokázal len nekorelovanosť $\overline{X}$ a *s* a potom nekorelovanosť $\overline{X}^2$ a s^2 (viď kap. 2.3), a to neimplikuje nezávislosť náhodných veličín $\overline{X}$ a *s* (viď [1], str. 34). Teda podľa Studenta združená hustota náhodných veličín *s* a *x* je rovná

$$f(s,x) = \frac{C}{\sigma^{n-1}} s^{n-2} e^{-\frac{ns^2}{2\sigma^2}} \cdot \frac{\sqrt{n}N}{\sqrt{2\pi}\sigma} e^{-\frac{nx^2}{2\sigma^2}} = \frac{CN}{\sigma^n}\sqrt{\frac{n}{2\pi}} \cdot s^{n-2} e^{-\frac{n(s^2+x^2)}{2\sigma^2}}.$$

Ideme použiť vetu o transformácii náhodného vektoru.

Nech

$$\boldsymbol{t}: \begin{pmatrix} s \\ x \end{pmatrix} \to \begin{pmatrix} z \\ w \end{pmatrix} = \begin{pmatrix} \frac{x}{s} \\ s \end{pmatrix}.$$

Keďže

$$z = \frac{x}{s}, \quad x = z \cdot w,$$
$$w = s,$$

pre inverzné zobrazenie $\boldsymbol{\tau}$ platí, že

$$\boldsymbol{\tau}: \begin{pmatrix} z \\ w \end{pmatrix} \to \begin{pmatrix} w \\ z.w \end{pmatrix},$$

kde $s \in (0,\infty)$, $x \in \Re$, teda aj $w \in (0,\infty)$, $z \in \Re$.

Najskôr overíme predpoklady vety.

Keďže

$$D_t = \det\begin{pmatrix} -\frac{x}{s^2} & \frac{1}{s} \\ 1 & 0 \end{pmatrix} = -\frac{1}{s} \neq 0, \quad s \neq 0,$$

$$x \in \Re,\ s \in (0,\infty) \ \Rightarrow\ G \subset \Re \times (0,\infty) \quad \text{je otvorená}$$

a

$$\left.\begin{array}{ll} \frac{\partial t_1}{\partial s} = -\frac{x}{s^2} & \frac{\partial t_1}{\partial x} = \frac{1}{s} \\ \frac{\partial t_2}{\partial s} = 1 & \frac{\partial t_2}{\partial x} = 0 \end{array}\right\} \text{ spojité na } G,$$

dostávame, že zobrazenie $\boldsymbol{t}$ je regulárne na množine $G \subset \Re \times (0,\infty)$.

Teraz sa pozrieme na podmienku prostoty.

Keďže

$$z = \frac{x}{s} \quad \text{a} \quad w = s,$$

potom

$\forall (s,x) \in G,\ \forall (\tilde{s},\tilde{x}) \in G$ platí, že ak $[(s,x) \neq (\tilde{s},\tilde{x})]$, potom $[\boldsymbol{t}(s,x) \neq \boldsymbol{t}(\tilde{s},\tilde{x})]$,

teda zobrazenie $\boldsymbol{t}$ je prosté na množine $G \subset \Re \times (0,\infty)$.

Keďže predpoklady sú splnené, použijeme danú vetu.

Združená hustota náhodných veličín Z a W je

$$g_{ZW}(z,w) = f(\boldsymbol{\tau}(z,w))\left|D_{\tau}\right|, \tag{2.14}$$

kde

$$D_{\tau} = \begin{pmatrix} \frac{\partial \tau_1}{\partial z} & \frac{\partial \tau_1}{\partial w} \\ \frac{\partial \tau_2}{\partial z} & \frac{\partial \tau_2}{\partial w} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ w & z \end{pmatrix} = -w,$$

čiže po dosadení do (2.14) dostaneme

$$g_{ZW}(z,w) = f(w,zw)\left|-w\right| = \frac{CN}{\sigma^n}\sqrt{\frac{n}{2\pi}} \cdot w^{n-2} e^{-\frac{n(zw^2+w^2)}{2\sigma^2}} w$$
$$= \frac{CN}{\sigma^n}\sqrt{\frac{n}{2\pi}} \cdot w^{n-1} e^{-\frac{nw^2(z^2+1)}{2\sigma^2}}$$

a hustotu $g_Z(z)$ s použitím substitúcie

$$y = \frac{nw^2(z^2+1)}{2\sigma^2},$$
$$\mathrm{d}y = \frac{n(z^2+1)}{\sigma^2} w\,\mathrm{d}w,$$
$$w = \sqrt{\frac{2\sigma^2 y}{n(z^2+1)}}$$

a gama funkcie

$$\Gamma(a) \overset{def.}{=} \int_0^{+\infty} x^{a-1} e^{-x}\,\mathrm{d}x, \quad a>0 \quad \text{a}\ \Gamma(a+1) = a!,\ \Gamma\left(\frac{1}{2}\right) \overset{def.}{=} \sqrt{\pi},$$

spočítame vzťahom

$$g_Z(z) = \frac{CN}{\sigma^n}\sqrt{\frac{n}{2\pi}} \cdot \int_0^{+\infty} w^{n-1} e^{-\frac{nw^2(z^2+1)}{2\sigma^2}}\,\mathrm{d}w =$$
$$= \frac{CN}{\sigma^n}\sqrt{\frac{n}{2\pi}} \cdot \frac{\sigma^2}{n(z^2+1)} \int_0^{+\infty} \left[\frac{2\sigma^2}{n(z^2+1)}\right]^{\frac{n-2}{2}} y^{\frac{n-1}{2}} e^{-y}\mathrm{d}y$$
$$= \frac{C}{\sqrt{2\pi}} N \cdot \frac{\sigma^{2+n-2}}{\sigma^n} \frac{\sqrt{n}}{n \cdot n^{\frac{n-2}{2}}} \cdot 2^{\frac{n-2}{2}} (z^2+1)^{-\left[\frac{n-2}{2}+1\right]} \Gamma\left(\frac{n-2}{2}+1\right) =$$
$$= \frac{CN}{\sqrt{2\pi}} \cdot 2^{\frac{n-2}{2}} \cdot \frac{\Gamma\left(\frac{n}{2}\right)}{n^{\frac{n-1}{2}}} \cdot (z^2+1)^{-\frac{n}{2}}, \qquad z \in \Re. \tag{2.15}$$

C môžeme prepísať cez gama funkciu (integračným cvičením zo vzťahu (2.8)) ako

$$C = \frac{N \cdot n^{\frac{n-1}{2}}}{2^{\frac{n-3}{2}}} \cdot \frac{1}{\Gamma\left(\frac{n-1}{2}\right)}.$$

Po dosadení do (2.15), dostaneme

$$g_Z(z) = \frac{N^2 n^{\frac{n-1}{2}}}{\sqrt{2\pi} \cdot 2^{\frac{n-3}{2}} \Gamma\left(\frac{n-1}{2}\right)} \frac{2^{\frac{n-2}{2}} \Gamma\left(\frac{n}{2}\right)}{n^{\frac{n-1}{2}}} (1+z^2)^{-\frac{n}{2}} =$$

$$\overset{N=1}{=} \frac{1}{2}\frac{n-2}{n-3}\frac{n-4}{n-5}\cdots\frac{5}{4}\cdot\frac{3}{2}(1+z^2)^{-\frac{n}{2}} \quad \text{pre } n \text{ nepárne,}$$

$$\overset{N=1}{=} \frac{1}{\pi}\frac{n-2}{n-3}\frac{n-4}{n-5}\cdots\frac{4}{3}\cdot\frac{2}{1}(1+z^2)^{-\frac{n}{2}} \quad \text{pre } n \text{ párne.}$$

Pretože táto rovnica nám dáva nezávislosť na σ, dostali sme rozdelenie náhodnej veličiny $z = \frac{\overline{X}}{s}$ pre akýkoľvek náhodný výber $(X_1, \ldots, X_n)$ z normálneho rozdelenia.

2.5 Vlastnosti hustoty *z*

Student chcel vytvoriť štatistickú tabuľku, z ktorej je možné vyčítať pravdepodobnosť, že výberový priemer $\overline{X}$ vydelený s $\left(z = \frac{\overline{X}}{s}\right)$, bude ležať medzi $-\infty$ a z $\left(\text{čiže } P(-\infty < \frac{\overline{X}}{s} < z) = F(z)\right)$. Teda potreboval vypočítať distribučnú funkciu $F(z)$.

Tá sa spočíta ako

$$F(z) = K \cdot \int_{-\infty}^{z} \frac{1}{\left(1+x^2\right)^{\frac{n}{2}}} \mathrm{d}x,$$

kde

$$K := \frac{1}{2}\frac{n-2}{n-3}\frac{n-4}{n-5}\cdots\frac{5}{4}\cdot\frac{3}{2} \quad \text{pre } n \text{ nepárne,}$$

$$:= \frac{1}{\pi}\frac{n-2}{n-3}\frac{n-4}{n-5}\cdots\frac{4}{3}\cdot\frac{2}{1} \quad \text{pre } n \text{ párne.}$$

S použitím substitúcie

$$x = \tan\theta, \quad \text{kde} \quad \theta \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right],$$

$$\mathrm{d}x = \frac{1}{\cos^2\theta}\mathrm{d}\theta,$$

kde medze spočítame ako

$$\theta = \arctan z,$$

$$\lim_{z\to-\infty} \arctan z = -\frac{\pi}{2},$$

$$\text{čiže} \quad \theta \in \left[-\frac{\pi}{2}, \arctan z\right],$$

dostaneme

$$\begin{aligned} F(z) &= K \cdot \int_{-\frac{\pi}{2}}^{\arctan z} \frac{1}{\left(1+\tan^2\theta\right)^{\frac{n}{2}}} \cdot \frac{1}{\cos^2\theta}\mathrm{d}\theta = \\ &= K \cdot \int_{-\frac{\pi}{2}}^{\arctan z} \frac{1}{\left[\frac{1}{\cos^2\theta}\right]^{\frac{n}{2}}} \cdot \frac{1}{\cos^2\theta}\mathrm{d}\theta = \\ &= K \cdot \int_{-\frac{\pi}{2}}^{\arctan z} \left[\cos^2\theta\right]^{\frac{n}{2}-1}\mathrm{d}\theta = \\ &= K. \int_{-\frac{\pi}{2}}^{\arctan z} \cos^{n-2}\theta\,\mathrm{d}\theta \end{aligned}$$

a to je možné dopočítať už pre akékoľvek $z \in \Re$.

Na základe toho Student zostavil *tabuľky distribučných funkcii* $F(z)$, kde

$$F(z) = \frac{n-2}{n-3}\frac{n-4}{n-5}\cdots\begin{pmatrix} \frac{3}{2}\cdot\frac{1}{2} & \text{pre } n \text{ nepárne} \\ \frac{2}{1}\cdot\frac{1}{\pi} & \text{pre } n \text{ párne} \end{pmatrix} \int_{-\frac{\pi}{2}}^{\arctan z} \cos^{n-2}\theta\,\mathrm{d}\theta. \qquad (2.16)$$

Pre porovnanie si vzal distribučnú funkciu $F(x)$ normálneho rozdelenia so smerodajnou odchýlkou $s = \sqrt{\frac{1}{7}}$ a $n = 10$:

$$F(x) = \frac{\sqrt{7}}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{7x^2}{2}}\mathrm{d}x\,.$$

Čiže tabuľka vyzerá následovne.

2.5.1 Tabuľky distribučnej funkcie $F(z)$ *získané výpočtom zo vzťahu* (2.16) *pre hodnoty* $n = 10$ *v porovnaní s distribučnou funkciou* $F(x)$ *normálneho rozdelenia, kde* $n = 10$ a $s = \sqrt{\frac{1}{7}}$.

z	$n = 10$	$F(x)$
0,1	0,61462	0,60411
0,2	0,71846	0,70159
0,3	0,80423	0,78641
0,4	0,86970	0,85520
0,5	0,91609	0,90691
0,6	0,94732	0,94375
0,7	0,96747	0,96799
0,8	0,98007	0,98253
0,9	0,98780	0,99137
1,0	0,99252	0,99820
1,1	0,99539	0,99926
1,2	0,99713	0,99971
1,3	0,99819	0,99986
1,4	0,99885	0,99989
1,5	0,99926	0,99999

Poznámka: Celú tabuľku je možné nájsť v Studentovom článku [10] str. 21.

2.6 Použitie tabuľky v praxi

Student urobil experiment na porovnanie účinnosti dvoch liečiv. Hodnoty daných liečiv získal z tabuliek A. R. Cushnyho a A. R. Peeblesa z časopisu *Journal of Physiology z roku 1904.* V nich bol zaznamenaný spánok 10 pacientov pred použitím hypnotických látok a po podaní (1) Dextro – hyoscyaminu hydromrobidu, (2) Laevo – hyoscyaminu hydrobromidu. Tabuľka 2.6.1 nám udáva počet hodín spánku získaných navyše u pacientov, ktorí užívali liek (1) a (2):

2.6.1 Tabuľka udávajúca počet hodín spánku získaných navyše užívaním lieku hyoscyaminu hydromrobidu získaná z [10].

Pacient	**1 (Dextro-)**	**2 (Laevo-)**	**Rozdiel (2-1)**
1.	+0,7	+1,9	+1,2
2.	-1,6	+0,8	+2,4
3.	-0,2	+1,1	+1,3
4.	-1,2	+0,1	+1,3
5.	-1,0	-0,1	0
6.	+3,4	+4,4	+1,0
7.	+3,7	+5,5	+1,8
8.	+0,8	+1,6	+0,8
9.	0	+4,6	+4,6
10.	+2,0	+3,4	+1,4
aritmetický priemer $\overline{X}$	+0,75	+2,33	+1.58
smerodatná odchýlka *s*	+1,70	+1,90	+1,17

Najskôr nás zaujíma pravdepodobnosť, že liek (1) v priemere dáva nárast spánku.

Keďže $z = \frac{\overline{X}}{s} = \frac{+0,75}{1,70} = 0,44$, hľadáme v tabuľkách 2.5.1 hodnotu distribučnej funkcie $F(z)$ pre $z = 0,44$ a $n = 10$. Tá nám vyšla niekde medzi 0,8697 a 0,9161. Student presne spočítal, že je to 0,8873. Inak povedané, šance sú 0,887 k 0,113. (Čiže pravdepodobnosť, že liek (1) zaznamená nárast spánku je približne 0,8873) . Teda, je veľmi pravdepodobné, že liek (1) zvýši dĺžku spánku.

Analogicky spočítame pravdepodobnosť, že liek (2) v priemere spôsobuje nárast spánku. Keďže $z = \frac{\overline{X}}{s} = \frac{+2,33}{1,90} = 1,23$ a opäť $n = 10$, z tabuliek 2.5.1 dostávame pravdepodobnosť rovnú 0,9974. (t.j. šance sú približne 400 k 1)

Z toho sa zdá byť pravdivé, že liek (2) je účinnejší ako liek (1). Ideme túto hypotézu otestovať tak, že odčítame dané hodnoty lieku (1) od lieku (2). (viď tabuľka 2.6.1 stĺpec „rozdiel (2-1)“).

Spočítame $z = \frac{\overline{X}}{s} = \frac{+1{,}58}{1{,}17} = 1{,}35$ a pre $n = 10$ z tabuľky 2.5.1 dostávame hodnotu 0,9985. Čiže pravdepodobnosť, že liek (2) je účinnejší ako liek (1) je okolo 0,9985. (t.j. šanca je približne 666 k 1) Pri takej vysokej pravdepodobnosti možme už predpokladať, že liek (2) je určite účinnejší ako liek (1).

Ďalšie zaujímavé experimenty, ktoré Student uskutočnil je možné nájsť v jeho článku [10].

3 Objavené nepresnosti v Studentovom článku

Prvou Studentovou nepresnosťou pri uvažovaní bolo, že na ukázanie gama rozdelenia náhodnej veličiny s^2, spočítal prvé štyri momenty veličiny s^2 a tie mu vyšli rovnako ako pri gama rozdelení, teda usúdil, že môžeme predpokladať gama rozdelenie náhodnej veličiny s^2. Hoci to tvrdenie všeobecne pre všetky rozdelenia neplatí (viď kap. 2.2 str. 14).

Ďalšieho omylu sa dopustil, keď vo vete o transformácii namiesto nezávislosti $\overline{X}$ a s, ukázal nekorelovanosť $\overline{X}$ a s a následne aj nekorelovanosť $\left(\overline{X}\right)^2$ a s^2. Čo v podstate neimplikuje nezávislosť náhodných veličín $\overline{X}$ a s.

Student počítal s testovacou štatistikou

$$z = \frac{\overline{X} - \mu}{s},$$

kde

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

V dnešnej dobe je už zrejmá testovacia štatistika

$$t = \sqrt{n}\,\frac{\overline{X} - \mu}{S},$$

kde

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Na ňu však prišiel v roku 1925 štatistik Ronald A. Fisher a to tak, že použil transformáciu $t = z\sqrt{n-1}$, ktorú si neskôr osvojil Student.

Ďalším významným článkom, ktorým sa Student zaoberal a ktorému sa budeme v krátkosti venovať je *Pravdepodobná chyba korelačného koeficientu* [9].

4 Pravdepodobná chyba korelačného koeficientu

Student sa pokúsil týmto článkom zaoberať problémom významu korelačných koeficientov odvodených z malých výberov a taktiež chcel motivovať matematikov, ktorí majú čas aj schopnosti sa daným problémom tiež zaoberať.

Predpokladá sa náhodný výber $(X_1,Y_1)^T,\ldots,(X_n,Y_n)^T$ z populácie (kde $(X_i,Y_i)^T$ sú stĺpcové vektory) s dvojrozmerným normálnym rozdelením $N_2\left((\mu_X,\mu_Y)^T,\begin{pmatrix}\sigma_X^2 & \operatorname{cov}(X,Y)\\ \operatorname{cov}(X,Y) & \sigma_Y^2\end{pmatrix}\right)$.

Definujme *výberový korelačný koeficient R* vzorcom

$$R=\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}}.$$

(Viď [3] str. 93).

Ďalej definujme *korelačný koeficient r* ako

$$r=\frac{\operatorname{cov}(X_i,Y_i)}{\sqrt{(\operatorname{var}X_i)(\operatorname{var}Y_i)}}.$$

My budeme zisťovať pravdepodobnosť, že *r* leží medzi danými dvoma medzami. (Tzv. *interval spoľahlivosti* $\mathrm{P}(D\le r\le H)$, kde *D* je dolná hranica a *H* je horná hranica).

Aby sme vyriešili tento problém, musíme vedieť dve veci: (1) rozdelenie hodnôt *R* získaných z výberu populácie (ktorá nám dáva *r*) a (2) *apriórnu pravdepodobnosť* (viď kap. 1.3), že *r* leží medzi danými dvoma medzami.

Student uvažoval, že rozdelenie *R* by mohlo mať len jednu z dvoch hustôt. A to buď *R* ma rovnomerné rozdelenie $R[-1,1]$, potom hustota *R* je

$$f(r)=\frac{1}{2}, \qquad \text{ak } r\in(-1,1),$$
$$=0, \qquad \text{inak.}$$

Alebo hustota náhodnej veličiny *R* môže byť rovná

$$f(r) = C \cdot \left(1 - r^2\right), \qquad \text{ak r} \in (-1,1),$$
$$= 0, \qquad \text{inak.}$$

V súlade s jeho pokusmi a skúsenosťami stanovil apriórne rozdelenie ako

$$f(r) = \frac{3}{4} \cdot \left(1 - r^2\right), \qquad \text{ak r} \in (-1,1),$$
$$= 0, \qquad \text{inak.}$$

Pre veľké výbery štatistici Pearson a Filon ukázali, že rozdelenie *R* odpovedá asymptoticky normálnemu rozdeleniu. No malými výbermi sa nik nezaoberal.

Student sa snažil tento problém vyriešiť a spravil nasledujúci experiment. Vzal 3000 hodnôt výšky postavy a dĺžky prostredného prsta na ľavej ruke kriminálnikov a urobil náhodný výber po štyroch z tejto populácie. Čím dostal 750 hodnôt *R* z populácie, ktorej skutočný korelačný koeficient *r* bol 0,66.

Keď vzal výšky postáv z nejakého výberu a dĺžky prostredných prstov ľudí iného výberu, dostal 750 hodnôt *R* z populácie, ktorej skutočný korelačný koeficient *r* bol rovný 0 (pretože výška jedného jedinca a dĺžka prostredného prsta iného jedinca, sú nezávislé náhodné veličiny, čiže korelácia je nulová).

Na základe získaných údajov z experimentov, s ktorými presnejšie pracuje v jeho článku [9] a matematickou úvahou prišiel na to, že ak je korelácia medzi dvoma normálne rozdelenými náhodnými veličinami nulová, hustota

$$f(r) = r_0 \left(1 - r^2\right)^{\frac{n-4}{2}}, \quad \text{kde } n \geq 3,$$

nám dáva presné rozdelenie výberového korelačného koeficientu *R* získaného z náhodného výberu *n* jednotlivcov. r_0 spočítame z rovnice

$$r_0 \int_{-1}^{1} \left(1 - r^2\right)^{\frac{n-4}{2}} \mathrm{d}r = N,$$

kde *N* je ako vo vzťahu (2.12) v kapitole 2. 2. V súčasnosti je známe, že *N* je rovné 1, pretože r_0 sa volí tak, aby to bola normalizujúca konštanta, čiže aby platilo, že hustota

$$f(r) = r_0 \int_{-1}^{1} \left(1 - r^2\right)^{\frac{n-4}{2}} \mathrm{d}r = 1. \qquad (4.1)$$

Spočítame teraz integrál zo vzťahu (4.1). Pred tým si ale potrebujeme zadefinovať pojem *beta funkcia*.

Pre $a > 0,\ b > 0$ definujeme beta funkciu $B(a, b)$ vzorcom

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\,\mathrm{d}x\,.$$

Medzi gama funkciou a beta funkciou platí vzťah

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}. \tag{4.2}$$

Vrátime sa späť k integrálu $\int_{-1}^{1}\left(1-r^2\right)^{\frac{n-4}{2}}\,\mathrm{d}r$:

$$\int_{-1}^{1}\left(1-r^2\right)^{\frac{n-4}{2}}\,\mathrm{d}r = 2\int_0^1\left(1-r^2\right)^{\frac{n-4}{2}}\,\mathrm{d}r\,.$$

Použijeme substitúciu

$$\begin{aligned} y &= r^2,\\ \sqrt{y} &= r,\\ \frac{1}{2}\frac{1}{\sqrt{y}}\,\mathrm{d}y &= \mathrm{d}r, \end{aligned}$$

a dostaneme

$$I = \int_{-1}^{1}\left(1-r^2\right)^{\frac{n-4}{2}}\,\mathrm{d}r = \int_0^1 y^{-1/2}(1-y)^{\frac{n-4}{2}}\,\mathrm{d}y = \int_0^1 y^{(1/2)-1}(1-y)^{\frac{n-2}{2}-1}\,\mathrm{d}y = B\left(1/2, \frac{n-2}{2}\right),$$

kde $n \geq 3$.

Čiže nakoniec

$$I = B\left(1/2, \frac{n-2}{2}\right) = \frac{\Gamma(1/2)\Gamma\left(\frac{n-2}{2}\right)}{\Gamma\left(\frac{n-2+1}{2}\right)} = \sqrt{\pi}\,\frac{\Gamma\left(\frac{n-2}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}, \quad n \geq 3\,.$$

Teda dostávame

$$r_0 = \frac{N}{\int_{-1}^{1}\left(1-r^2\right)^{\frac{n-4}{2}}\,\mathrm{d}r} = \frac{1}{\frac{\sqrt{\pi}\,\Gamma\left(\frac{n-2}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}} = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{n-2}{2}\right)}, \text{ kde } n \geq 3\,.$$

Takže výsledná hustota výberového korelačného koeficientu R pri výbere o rozsahu $n \geq 3$ z dvojrozmerného normálneho rozdelenia s nulovým korelačným koeficientom r je rovná

$$f(r) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{n-2}{2}\right)}\left(1-r^2\right)^{\frac{n-4}{2}}.$$

V ďalšej kapitole sa budeme venovať štatistikom, ktorí na Studentovu tvorbu neskôr naväzovali.

5 Štatistici, ktorých ovplyvnila Studentova tvorba

V období pred Fisherom nikto nevenoval veľkú pozornosť Studentovým článkom. Prvým významných štatistikom, ktorého ovplyvnili Studentove práce, bol už vyššie zmienený Ronald A. Fisher. Ten nielen oživil Studentove články, ale tiež v nich opravil chyby a nedostatky. Keď sa hovorí o „Studentovom t-rozdelení“, máme v podstate namysli skôr Fisherove myšlienky ako Studentove. Napríklad t-rozdelenie o $n - 1$ stupňov voľnosti, ako rozdelenie založené na normálnom rozdelení, no tiež jeho aplikácia v regresnej analýze, boli výplodom Fisherovho myslenia. Aplikáciu „Studentovho t-rozdelenia“ Fisher zahrnul vo svojom článku [4]. Student sa vo svojich prácach venoval len jednovýberovými problémami, no Fisher v r. 1922 ukázal aplikáciu t- rozdelenia aj na dvojvýberové problémy, kde v oboch výberoch predpokladal rovnaký rozptyl náhodných veličín.

t-rozdelenie bolo neskôr v roku 1931 Haroldom Hotellingom (matematickým štatistikom a významným ekonomickým teoretikom) zovšeobecnené aj na testovanie vektoru stredných hodnôt s mnohorozmerným normálnym rozdelením.

Studentove práce sa dostali do povedomia aj štatistikom z ruskej ríše. Významným štatistikom, ktorého zaujala Studentova tvorba, bol Evgenii Evgenevič Sluckij. Jeho meno sa stalo významné v oblasti ekonometrie a matematickej štatistiky, hlavne kvôli štúdiu stochasticky závislých náhodných procesov. Sluckij venoval dlhé roky štúdiu publikácií Karla Pearsona, čím sa dostal aj k prácam Studenta. Neskôr napísal knihu [8], ktorej názov nám už sám o sebe napovedá , že sa v nej zaoberal sčasti (a dokonca z väčšej časti) teóriou korelácie, špeciálne pravdepodobnej chybe. To ho počas písania priviedlo k Studentovmu článku [9], z ktorého potom vo svojej knihe citoval niektoré výsledky a závery.

Ďalšími osobnosťami z Ruska, ktorí venovali pozornosť hlavne druhému Studentovmu článku [9], boli Alexander Alexandrovič Čuprov (štatistik) a Andrej Andrejevič Markov (matematik). Tí sa zaoberali prevažne šiestimi typmi Pearsonových kriviek (tie je možné nájsť v Pearsonovom článku [7]), čo v podstate do istej miery súviselo aj s t-rozdelením Studentovej štatistiky.

6 Návrat k súčasnosti

6.1 Studentovo t-rozdelenie

V súčasnej štatistike sa Studentovo t-rozdelene zavádza nasledovne:

Veta (súčasnosť) *Nech* $X_1,\ldots,X_n$ *je náhodný výber z* $\mathrm{N}(\mu,\sigma^2)$, *kde* $n\geq 2$ *a* $\sigma^2>0$, *nech*

$$S=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}\quad a\quad \overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i\,,$$

potom

$$T=\sqrt{n}\,\frac{\overline{X}-\mu}{S},$$

kde $\overline{X}$ *a* S *sú nezávislé náhodné veličín, má t – rozdelenie o n – 1 stupňov voľnosti s hustotou*

$$f(t)=\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)\sqrt{\pi(n-1)}}\left(1+\frac{t^2}{n-1}\right)^{-\frac{n}{2}}. \tag{6.1}$$

Dôkaz. Viď Andělova kniha [3] str. 74.

Pre porovnanie uvedieme, ako by asi sformuloval vetu Student pred 100 rokmi, na základe zmienených faktov v jeho článku.

Veta (Student pred 100 rokmi) *Nech* $X_1,\ldots,X_n$ *je náhodný výber z* $\mathrm{N}(0,\sigma^2)$, *kde* $n\geq 2$, *(* X_i *je definované ako v kapitole 2.2), nech*

$$S=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}\quad a\quad \overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i\,,$$

potom

$$T=\frac{\overline{X}}{S},$$

kde $\overline{X}$ *a* S *sú nekorelované náhodné veličiny, má hustotu*

$$f(t) = \frac{N^2 n^{\frac{n-1}{2}}}{\sqrt{2\pi}\cdot 2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)} \frac{2^{\frac{n-2}{2}}\Gamma\left(\frac{n}{2}\right)}{n^{\frac{n-1}{2}}}(1+t^2)^{-\frac{n}{2}} = \overset{N=1}{=} \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{n-1}{2}\right)}(1+t^2)^{-\frac{n}{2}}. \tag{6.2}$$

Všimnime si, že súčasný vzorec sa veľmi nelíši od toho spred 100 rokov. V podstate sa odlišuje od toho Studentovho len o premennú $n-1$. Už vyššie sme spomenuli, že túto nepresnosť opravil Fisher v roku 1925 a dospel tým k hustote, s ktorou sa počíta v štatistike dodnes.

6.2 Výberový korelačný koeficient

Student vo svojom článku [9] prišiel na to, že hustota výberového korelačného koeficientu *R* pri náhodnom výbere o rozsahu $n \geq 3$ z dvojrozmerného normálneho rozdelenia s nulovým korelačným koeficientom *r* by mala byť daná vzorcom

$$f(r) = r_0\left(1-r^2\right)^{\frac{n-4}{2}}, \tag{6.3}$$

kde

$$r_0 = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{n-2}{2}\right)}.$$

V súčasnej štatistike (viď Andělovu knihu [3]) je jasné, že hustota *R* je daná rovnicou

$$f(r) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-2}{2}\right)\sqrt{\pi}}\left(1-r^2\right)^{\frac{n-4}{2}}, \tag{6.4}$$

kde

$$-1 < r < 1 \text{ a } n \geq 3.$$

Čiže sa Student vo výpočte nemýlil.

7 Záver

Student vo svojich prácach urobil pár vážnejších chýb, s ktorými jeho počítanie nebolo celkom korektné, no aj tak mu patrí obrovský obdiv. Svojimi výpočtami a závermi ovplyvnil množstvo štatistikov a matematikov, ktorí nakoniec Studentove postupy vylepšili a opravili natoľko, že našli využitie v súčasnej štatistike. Taktiež je potrebné spomenúť meno Arthur Guiness, ktorý Studentovi umožnil stráviť dva semestre akademického roku v Pearsonovom biometrickom laboratóriu na univerzite v Londýne, kde sa Student veľa veci naučil. Ďalšou dôležitou osobou bol Ronald A. Fisher, bez ktorého by Studentove práce ostali bez povšimnutia.

No aj napriek vážnejším chybám sa v podstate Studentove výsledky až tak nelíšia od tých súčasných. Stačí porovnať hustotu (6.1) so (6.2) a hustotu (6.3) so (6.4), kde jasne vidieť len veľmi malé rozdiely. Ak však vezmeme do úvahy fakt, že pred 100 rokmi neexistovali žiadne počítače a ani štatistika nebola natoľko vyvinutá ako dnes a že Student bol v podstate zamestnanec pivovaru, určite prijmeme názor, že Studentovi patrí obrovská pocta a Studentovo t-rozdelenie nesie právom jeho meno.

Literatúra

[1] Airy, G. B.: *On the Algebraical and Numerical Theory of Errors of Observations and the Combination of Observations,* Macmillan and Co., Cambridge and London 1861.

[2] Anděl, J.: *Matematická statistika*, SNTL, Praha 1985.

[3] Anděl, J.: *Základy matematické statistiky*, Matfyzpress, Praha 2005.

[4] Fisher, R. A.: *Statistical Methods for Research Workers*, Oliver & Boyd., Edinburgh and London 1925.

[5] Hotteling, H.: *British Statistics and Statisticians Today*, Journal of the American Statistical Association **25** (1930), 186 – 190.

[6] Pearson, E. S., Plackett R. L., Barnard G. A.: *'Student', A Statistical Biography of William Sealy Gosset*, Oxford University Press, Oxford 1990.

[7] Pearson, K.: *Contributions to the mathematical theory of evolution, II: Skew variation in homogeneous material*, Philosophical Transactions of the Royal Society of London, A **186** (1895), 343 – 414.

[8] Slutsky, E. E.: *The Theory of Correlation and Elements of the Study of Curves of Distribution,* Izvestiia Kievskago Kommercheskago Instituta, Kiev 1912.

[9] Student: *Probable Error of a Correlation Coefficient*, Biometrika **6** (1908), 302 – 310.

[10] Student: *The Probable Error of a Mean*, Biometrika **6** (1908), 1 – 25.