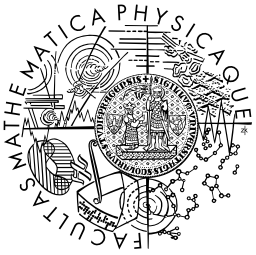


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Michaela Vařejková

Zobecněný Wilcoxonův test pro cenzorovaná data

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce:	RNDr. Matúš Maciak, Ph.D.
Studijní program:	Matematika
Studijní obor:	Obecná matematika

Praha 2019

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) 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 má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ..... Podpis autora

Ráda bych na tomto místě poděkovala vedoucímu práce, RNDr. Matúšovi Maciakovi, Ph.D., za zajímavé téma a všechny podnětné připomínky a rady.
Dále bych ráda poděkovala své rodině za podporu během celého studia.

Název práce: Zobecněný Wilcoxonův test pro cenzorovaná data

Autor: Michaela Vařejková

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: RNDr. Matúš Maciak, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Tato práce se zabývá zobecněným Wilcoxonovým testem a jeho použitím pro cenzorovaná data. V úvodní části je popsán standardní jednovýběrový a dvouvýběrový Wilcoxonův test a jejich základní vlastnosti, dále jsou popsána cenzorovaná data a metody cenzorování. Hlavní část práce se věnuje představení zobecněného Wilcoxonova testu a jeho vlastnostem. Nejprve je popsán test pro cenzorovaná, následně i pro dvojitě cenzorovaná data. Závěr práce je věnován praktické části, ve které jsou pomocí simulací ukázány statistické vlastnosti testu. První příklad porovnává zobecněný test se standardním dvouvýběrovým Wilcoxonovým testem, druhý příklad sleduje, jak míra cenzorování ovlivňuje sílu a hladinu zobecněného testu.

Klíčová slova: Wilcoxon test, nulová hypotéza, cenzorovaná data, náhodný výběr.

Title: Generalized Wilcoxon Test for Censored Data

Author: Michaela Vařejková

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Matúš Maciak, Ph.D., Department of Probability and Mathematical Statistics

Abstract: This paper deals with the generalized Wilcoxon test and its use for censored data. The introduction describes standard one-sample and two-samples Wilcoxon tests and their basic properties, censored data and methods of censoring. The main part of the paper is devoted to the introduction of the generalized Wilcoxon test and to its properties. First, a test for singly-censored data is described; the description of a test for doubly censored data follows. The paper concludes with a simulations part in which statistical properties of the test are demonstrated. The first example compares the generalized test with the standard two-samples Wilcoxon test. The second example shows how the censoring rate affects the power and significance level of the generalized test.

Keywords: Wilcoxon test, null hypothesis, censored data, random sample.

Obsah

Úvod

Tato práce se zabývá zobecněným Wilcoxonovým testem, který lze použít v případě, kdy analyzovaná data obsahují i data cenzorovaná. K cenzorování dat může docházet v mnoha různých odvětvích. Jedním z nich jsou například lékařské klinické studie, u kterých z časových či finančních důvodů nelze dopozorovat všechny jedince a získaná data jsou tedy nekompletní. Jedním z možných přístupů je nekompletní cenzorovaná data ze studie vynechat, ale jak si ukážeme ve třetí kapitole na provedených simulacích, připravili bychom se tím o značnou část informací. Z tohoto důvodu je velmi užitečné věnovat se metodám, které cenzorovaná data zpracovávají a používat statistické testy, které s těmito daty umějí pracovat.

Práce je rozdělena do tří kapitol. V první kapitole je představený jednovýběrový a dvouvýběrový Wilcoxonův test, ze kterého zobecnění vychází a jsou popsána cenzorovaná data a různé metody cenzorování. Ve druhé kapitole je představená zobecněná varianta Wilcoxonova testu, nejprve pro cenzorovaná, následně pro dvojitě cenzorovaná data. Poslední kapitola se věnuje simulační studii, na které jsou ukázané statistické vlastnosti testu.

Tato práce předpokládá základní znalost matematické statistiky a základních principů testování hypotéz.

1. Wilcoxonův test

V této části představíme nejprve jednovýběrový, posléze i dvouvýběrový Wilcoxonův test, jehož zobecnění je hlavní náplň práce. Vycházíme zde především z Omelka (2019).

Jednovýběrový i dvouvýběrový Wilcoxonův test jsou příklady tzv. neparametrických testů. Výhoda těchto testů spočívá v tom, že pro práci s nimi nemusíme znát konkrétní typ rozdělení, ze kterého pocházejí analyzovaná data. Jejich testové statistiky bývají často založeny na pořadí náhodných veličin v náhodném výběru, připomeneme tedy definici.

Definice 1. *Mějme $X_1,\ldots,X_n$ náhodný výběr ze spojitého rozdělení a $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ uspořádaný náhodný výběr, kde $X_{(k)}$ značí k-tou nejmenší hodnotu mezi $X_1,\ldots,X_n$.* Pořadím *náhodné veličiny X_i ve výběru $X_1,\ldots,X_n$ rozumíme přirozené číslo R_i takové, že $X_i = X_{(R_i)}$.*

Další výhodou neparametrických metod založených na pořadí je to, že nejsou příliš citlivé na odlehlá pozorování, například na rozdíl od testů, jejichž testové statistiky využívají výběrový průměr. Naopak nevýhodou je, že pokud bychom znali konkrétní rozdělení dat, může mít Wilcoxonův test ve srovnání s jinými parametrickými testy menší statistickou sílu.

1.1 Jednovýběrový Wilcoxonův test

Nechť $X_1,\ldots,X_n$ je náhodný výběr ze spojitého rozdělení s hustotou f, která je symetrická kolem bodu δ, tj. platí $f(\delta - x) = f(\delta + \mathrm{x})$ pro $x \in \mathbb{R}$. Právě z předpokladu symetrické hustoty X_i plyne, že střed symetrie δ je roven mediánu m. V případě, kdy existuje konečná střední hodnota μ, pak také $\mu = \delta$. Konečnost střední hodnoty však obecně nepředpokládáme. Jednovýběrový Wilcoxonův test je určen k testování hypotézy H_0: $\delta = \delta_0$ proti alternativě H_1: $\delta \neq \delta_0$, kde δ_0 je předem daná konstanta.

Definujme náhodné veličiny $Z_i := X_i - \delta_0$. Testová statistika jednovýběrového Wilcoxonova testu má pak tvar

$$W_s = \sum_{i \in I} R_i,$$

kde $\mathrm{I} = \{\ \mathrm{i} \in \{1,\ldots,n\} : Z_i \geq 0\ \}$ a R_i je pořadí $|Z_i|$ mezi $|Z_1|,\ldots,|Z_n|$.
Dá se ukázat, že za platnosti nulové hypotézy platí následující vztahy

$$\mathbb{E}W_s = \frac{n(n+1)}{4}, \quad \mathrm{var}(W_s) = \frac{n(n+1)(2n+1)}{24}$$

a testová statistika W_s je asymptoticky normální

$$\frac{W_s - \mathbb{E}W_s}{\sqrt{\mathrm{var}(W_s)}} \xrightarrow[n\to\infty]{d} N(0,1).$$

Díky tomu dostáváme asymptotické rozdělení statistiky za platnosti H_0:

$$U_n = \frac{W_s - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \overset{as.}{\sim} N(0,1).$$

Odtud získáme kritické hodnoty asymptotického testu a zamítáme nulovou hypotézu $H_0 \Leftrightarrow |U_n| \geq u_{1-\frac{\alpha}{2}}$.

1.2 Dvouvýběrový Wilcoxonův test

Mějme dva na sobě nezávislé náhodné výběry $X_1, \ldots, X_n$ a $Y_1, \ldots, Y_m$, po řadě s distribučními funkcemi F_X a F_Y. Naším cílem bude tyto dvě distribuční funkce porovnat. Nejprve budeme uvažovat, že F_X a F_Y jsou spojité a platí pro ně tzv. model posunutí v poloze, tj.

$$\exists \delta \in \mathbb{R} : F_X(x) = F_Y(x - \delta) \ \ \forall x \in \mathbb{R}.$$

Parametr, který testujeme je zde právě parametr posunutí δ. Uvažujme hypotézu $H_0 : \delta = 0$ proti alternativě $H_1 : \delta \neq 0$. Za platnosti nulové hypotézy jsou rozdělení náhodných výběrů totožná, rovnají se tedy jejich mediány a v případě, kdy existují střední hodnoty také $\mathbb{E}X = \mathbb{E}Y$. Dvouvýběrový Wilcoxonův test lze za platnosti modelu posunutí chápat jako test rovnosti mediánů a středních hodnot. Můžeme však uvažovat i obecnější model, který pouze předpokládá spojitost distribučních funkcí F_X a F_Y. Potom testujeme hypotézu $H_0 : F_X = F_Y$ proti alternativě $H_1 : F_X \neq F_Y$. V tomto případě však nemůžeme nic rozhodnout o mediánech ani středních hodnotách rozdělení, pouze můžeme určit zda se distribuční funkce rovnají či nikoli. Všimněme si, že za platnosti nulové hypotézy jsou oba výše popsané modely totožné.
Testovou statistiku dvouvýběrového Wilcoxonova testu můžeme zapsat ve tvaru

$$W_{n,m} = \sum_{i=1}^{n} R_i,$$

kde $R_1, \ldots, R_n$ jsou pořadí náhodných veličin $X_1, \ldots, X_n$ ve sdruženém náhodném výběru $X_1, \ldots, X_n, Y_1, \ldots, Y_m$.
Podobně jako u jednovýběrového testu se dá snadno ověřit, že za platnosti nulové hypotézy je

$$\mathbb{E}W_{n,m} = \frac{n(n+m+1)}{2}, \quad \mathrm{var}(W_{n,m}) = \frac{nm(n+m+1)}{12}$$

a platí

$$\frac{W_{n,m} - \mathbb{E}W_{n,m}}{\sqrt{\mathrm{var}(W_{n,m})}} \xrightarrow[n,m\to\infty]{d} N(0,1).$$

Můžeme tedy zapsat asymptotické rozdělení testové statistiky za H_0:

$$U_{n,m} = \frac{W_{n,m} - \frac{n(n+m+1)}{2}}{\sqrt{\frac{nm(n+m+1)}{12}}} \overset{as.}{\sim} N(0,1).$$

Nulovou hypotézu H_0 zamítáme $\Leftrightarrow |U_{n,m}| \geq u_{1-\frac{\alpha}{2}}$.

1.2.1 Mann-Whitneyho formulace

Dvouvýběrový Wilcoxonův test můžeme formulovat i pomocí tzv. Mann - Whitneyho statistiky, která nám počítá, kolikrát napozorované hodnoty z druhého výběru překročí hodnoty z prvního výběru, pokud bychom výběry spojili do jednoho sdruženého. Je definována jako

$$W^*_{n,m} = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbb{1}\left\{X_i < Y_j\right\}.$$

Mezi testovou statistikou klasického dvouvýběrového Wilcoxonova testu $W_{n,m}$ a statistikou $W^*_{n,m}$ existuje lineární vztah, testy za použití těchto statistik jsou tedy ekvivalentní. Jak přesně vypadá vztah mezi statistikami nám udává následující věta.

Věta 1. *Pro testové statistiky $W_{n,m}$ a $W^*_{n,m}$ platí vztah*

$$W_{n,m} + W^*_{n,m} = nm + \frac{n(n+1)}{2}.$$

Důkaz. Chceme ukázat, že

$$\sum_{i=1}^{n} R_i + \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbb{1}\left\{X_i < Y_j\right\} = nm + \frac{n(n+1)}{2}.$$

Z definice pořadí plyne, že

$$R_i = \sum_{j=1}^{n} \mathbb{1}\left\{X_j \le X_i\right\} + \sum_{j=1}^{m} \mathbb{1}\left\{Y_j \le X_i\right\}.$$

Dostáváme tedy

$$\begin{aligned}
&\sum_{i=1}^{n}\left(\sum_{j=1}^{n} \mathbb{1}\left\{X_j \le X_i\right\} + \sum_{j=1}^{m} \mathbb{1}\left\{Y_j \le X_i\right\}\right) + \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbb{1}\left\{X_i < Y_j\right\} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{1}\left\{X_j \le X_i\right\} + \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbb{1}\left\{Y_j \le X_i\right\} + \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbb{1}\left\{X_i < Y_j\right\} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{1}\left\{X_j \le X_i\right\} + \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbb{1}\left\{Y_j \le X_i \ \vee \ Y_j > X_i\right\} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{1}\left\{X_{(j)} \le X_{(i)}\right\} + \sum_{i=1}^{n}\sum_{j=1}^{m} 1 = \sum_{i=1}^{n} i + nm = \frac{n(n+1)}{2} + nm,
\end{aligned}$$

což jsme chtěli dokázat.

□

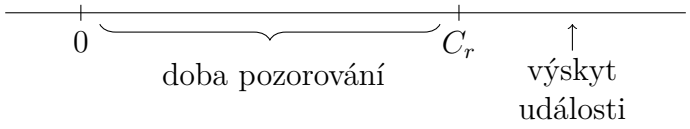
1.3 Cenzorovaná data

Dvouvýběrový Wilcoxonův test budeme zobecňovat pro případ, kdy naše data budou obsahovat i data cenzorovaná. Nejprve si tedy vysvětlíme, co to cenzorovaná data jsou, jak vznikají a jaké typy rozlišujeme.

Ve statistice často analyzujeme data, která nám udávají dobu čekání na výskyt nějaké události. S těmito daty se můžeme setkat v různých oborech, jako například v medicíně, průmyslu či ekonomii. V medicíně jsou běžná při lékařských studiích, kdy sledujeme čas do objevení příznaků jisté choroby, ve strojírenství můžeme zaznamenávat čas, kdy dojde k porouchání stroje. Někdy se ovšem může stát, že naše data budou obsahovat i jedince, u kterých nelze přesně určit, kdy k této požadované události došlo. Víme pouze, v jakém časovém intervalu to bylo. Tato data pak nazýváme cenzorovaná. Podle toho, jakým způsobem k cenzorování došlo rozlišujeme cenzorování zprava, cenzorování zleva a intervalové cenzorování. V této části vycházíme z Klein a Moeschberger (2003).

1.3.1 Cenzorování zprava

Cenzorování zprava vzniká v momentě, kdy nemůžeme pozorovat všechny jedince až do výskytu události. Víme pouze, jak dlouhou k události nedošlo, ale kdy přesně nastala už nejsme schopni určit. Podle metody cenzorování rozlišujeme cenzorování typu I a typu II. V následujícím odstavci si vysvětlíme oba typy, ale v práci se následně budeme zabývat pouze cenzorováním typu I, konkrétně jeho zobecněnou variantou.



Obrázek 1.1: Časová osa znázorňující cenzorování zprava.

Cenzorování typu I

Při klinických studiích se často stává, že z ekonomických, případně časových důvodů je potřeba studii ukončit ještě předtím, než pozorovaná událost nastane u všech jedinců. Označme čas ukončení studie C_r. V případě, kdy požadovaná událost nastala před tímto časem, známe přesný údaj, v opačném případě pouze čas cenzorovaný. Pokud by všichni pacienti vstoupili do studie zároveň, v jeden konkrétní společný čas, byl by i cenzorovaný čas u všech pacientů totožný. My ovšem budeme uvažovat, že pacienti mohou do studie vstupovat průběžně, cenzorované časy jednotlivých pacientů jsou tedy individuální. Tento typ cenzorování nazýváme zobecněné cenzorování typu I. Čas C_r však nemusí reprezentovat pouze čas ukončení studie. V průběhu našeho pozorování mohlo dojít k jiné mimořádné události, která nám znemožnila jedince dále pozorovat, pacient mohl například přestat docházet na pravidelné kontroly.

Obecně můžeme tato data reprezentovat dvojicí (T_i,δ_i), kde T_i je náhodná veličina a δ_i je tzv. indikátor cenzorování, pro který platí $\delta_i = 1$, pokud známe přesný časový údaj a k výskytu události došlo před časem C_r a $\delta_i = 0$, pokud došlo k cenzorování. Platí T_i=min(X_i,C_r), kde X_i je přesný čas i-tého jedince.

Cenzorování typu II

Uvažujme studii, do které jsme zahrnuli n jedinců. Zvolíme si pevně dané přirozené číslo r takové, že $r < n$. Studie poběží do té doby, dokud nedojde k selhání u prvních r jedinců, poté studii ukončíme a vyhodnotíme. Tento typ cenzorování se používá například při testování životnosti různých strojů, kdy by mohlo být velmi časově a finančně náročné čekat, než dojde k selhání u všech kusů.

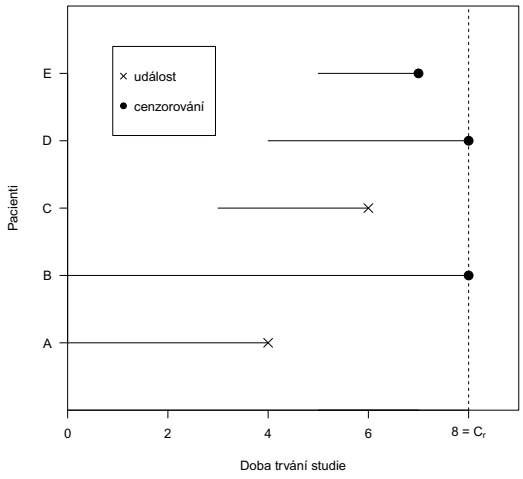
Příklad

Uveďme krátký příklad, na kterém si ilustrujeme cenzorování zprava. Budeme mít 5 pacientů, u kterých se postupně projeví příznaky jisté choroby. Těmto pacientům podáme lék, který příznaky okamžitě zmírní a budeme sledovat, za jak dlouho se dané příznaky opět objeví. Naše studie bude probíhat po dobu 8 týdnů, tedy $C_r = 8$. Na začátku studie budeme mít 2 pacienty A a B, v třetím týdnu vstoupí do studie pacient C, ve čtvrtém týdnu pacient D a jako poslední pacient E v pátém týdnu. Nejdříve k výskytu události, tj. k znovuobjevení příznaků, dojde u pacienta A ve čtvrtém týdnu, potom u pacienta C, a to v šestém týdnu. V sedmém týdnu se pacient E rozhodne studii opustit. U pacientů B a D k výskytu události po celou dobu trvání studie nedojde. Data, která jsme ze studie získali, budeme reprezentovat dvojicí (T_i,δ_i), kde T_i je naměřený čas a δ_i indikátor cenzorování. Jak vypadají napozorovaná data v našem konkrétním případě je uvedeno v Tabulce 1.1.

Pacienti	**A**	**B**	**C**	**D**	**E**
Data	(4,1)	(8,0)	(3,1)	(4,0)	(2,0)

Tabulka 1.1: Napozorovaná data uvedená ve tvaru (T_i, δ_i), $i = 1, \ldots 5$.

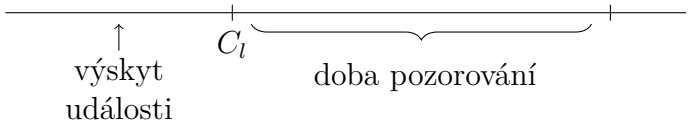
Pro lepší představu můžeme cenzorování znázornit i graficky.



Obrázek 1.2: Grafické znázornění nasbíraných dat, u kterých došlo k cenzorování typu I zprava.

1.3.2 Cenzorování zleva

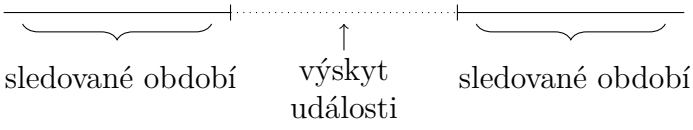
Cenzorování zleva je analogické cenzorování zprava jen s tím rozdílem, že daná událost u jedince nastala ještě předtím, než jsme ho začali pozorovat. Nevíme tedy kdy přesně k události došlo, víme jen, že to bylo před časem C_l (počáteční čas pozorování). Data můžeme opět reprezentovat pomocí dvojice (T_i,ω_i), kde ω_i je identifikátor cenzorování a pro náhodnou veličinu T_i platí T_i=max(X_i,C_l), kde X_i je přesný čas i-tého jedince.



Obrázek 1.3: Časová osa znázorňující cenzorování zleva.

1.3.3 Intervalové cenzorování

Intervalové cenzorování je zobecněním pravého a levého cenzorování. Vzniká v ten moment, kdy víme, že přesný údaj X_i spadá do nějakého intervalu $(L_i, R_i]$. V praxi se s tímto typem cenzorování setkáváme například při pravidelných kontrolách pacienta. Pokud u pacienta dojde k objevení příznaků, které u poslední kontroly ještě neměl, známe pouze časový interval. Kdy přesně k výskytu došlo už nejsme schopni určit.



Obrázek 1.4: Časová osa znázorňující intervalové cenzorování.

2. Zobecněný Wilcoxonův test pro cenzorovaná data

Dvouvýběrový Wilcoxonův test můžeme zobecnit pro případ, kdy naše data budou obsahovat právě data cenzorovaná. Toto zobecnění bylo představeno v Gehan (1965b). Uvažujme klinickou studii, ve které chceme porovnat dvě alternativní léčby a jejich schopnost prodloužit délku života pacientů. U každého pacienta budeme zaznamenávat jeho délku života po zahájení léčby, tedy po vstupu do studie. Pacienti vstupují do studie průběžně a jsou náhodně rozdělení do dvou skupin, jedné skupině bude aplikována léčba A, druhé skupině léčba B. Předpokládejme, že n pacientům byla přiřazena léčba A, m pacientům léčba B. Po uplynutí předem stanoveného času C_r studii ukončíme a vyhodnotíme. Mohly nastat 2 možnosti. Buď pacient již zemřel a známe přesný údaj o délce jeho života, nebo je pacient stále naživu a známe pouze cenzorovaný čas (doba, která uplynula od vstupu pacienta do studie až po čas C_r). Právě fakt, že pacienti mohou vstupovat do studie průběžně, nikoli v jeden společný okamžik, nám říká, že cenzorované časy se mohou lišit. Jedná se tedy o zobecněné cenzorování typu I zprava.

Napozorovaná data:

$$\left.\begin{array}{ll} X_1^{'},...,X_{r_n}^{'} & r_n \text{ cenzorovaných časů} \\ X_{r_n+1},...,X_n & n-r_n \text{ přesných časů} \end{array}\right\} \text{léčba A},$$

$$\left.\begin{array}{ll} Y_1^{'},...,Y_{r_m}^{'} & r_m \text{ cenzorovaných časů} \\ Y_{r_m+1},...,Y_m & m-r_m \text{ přesných časů} \end{array}\right\} \text{léčba B},$$

kde X_i,Y_j jsou přesné údaje o délce života pacientů, $X_i^{'}$,$Y_j^{'}$ časy cenzorované. Teoretické údaje o délce života jsou z distribučních funkcí $F_1(x), F_2(y)$, které mohou být diskrétní nebo spojité. Uvažujme nulovou hypotézu

$$H_0 : F_1(t) = F_2(t) \ \ (t \leq C_r)$$

a jednostrannou alternativu

$$H_1 : F_1(t) < F_2(t) \ \ (t \leq C_r),$$

nebo oboustrannou alternativu

$$H_2 : F_1(t) < F_2(t) \ \text{ nebo } \ F_1(t) > F_2(t) \ \ (t \leq C_r).$$

V případě klinické studie popsané výše se dá nulová hypotéza H_0 interpretovat tak, že léčby A a B jsou stejně efektivní, alternativa H_1 tak, že léčba A je efektivnější než B a alternativa H_2 že efektivnější je jedna z léčeb, nespecifikujeme jestli A nebo B.
Definujme náhodné veličiny U_{ij}:

$$U_{ij} = \begin{cases} -1 & X_i < Y_j \ \vee \ X_i \leq Y_j^{'} \\ 0 \quad X_i = Y_j \ \vee & (X_i^{'},Y_j^{'}) \ \vee \ X_i^{'} < Y_j \ \vee \ Y_j^{'} < X_i \\ 1 & X_i > Y_j \ \vee \ X_i^{'} \geq Y_j \end{cases}$$

Všimněme si, že relevantní jsou jen ta porovnání, kdy u obou pacientů známe přesný údaj nebo když pacient s cenzorovaným časem žil déle, než pacient s přesným časem.
Testovou statistiku zobecněného Wilcoxonova testu můžeme zapsat ve tvaru

$$W = \sum_{i=1}^{n}\sum_{j=1}^{m} U_{ij}.$$

V případě, kdyby žádná data nebyla cenzorovaná a nenapozorovali jsme žádné shody, je zobecněný test ekvivalentní klasickému dvouvýběrovému Wilcoxonově testu s testovou statistikou $W_{n,m}$. Jak přesně bude vypadat vztah nám udává následující věta.

Věta 2. *Platí vztah*

$$W + 2W_{n,m} = m(n + m + 1).$$

Důkaz. Testovou statistiku W můžeme ekvivalentně zapsat jako

$$W = \sum_{i=1}^{n}\sum_{j=1}^{m} U_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{m} \left(\mathbb{1}\left\{X_i > Y_j\right\} - \mathbb{1}\left\{X_i < Y_j\right\}\right).$$

Z definice pořadí dále plyne, že

$$W_{n,m} = \sum_{j=1}^{m}\left(\sum_{i=1}^{m}\mathbb{1}\left\{Y_i \le Y_j\right\} + \sum_{i=1}^{n}\mathbb{1}\left\{X_i \le Y_j\right\}\right).$$

Potom

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\mathbb{1}\left\{X_i > Y_j\right\} - \mathbb{1}\left\{X_i < Y_j\right\}\right) + 2\sum_{j=1}^{m}\left(\sum_{i=1}^{m}\mathbb{1}\left\{Y_i \le Y_j\right\} + \sum_{i=1}^{n}\mathbb{1}\left\{X_i \le Y_j\right\}\right)\\
&= \sum_{i=1}^{n}\sum_{j=1}^{m}\mathbb{1}\left\{X_i > Y_j\right\} - \sum_{i=1}^{n}\sum_{j=1}^{m}\mathbb{1}\left\{X_i \le Y_j\right\} + 2\sum_{j=1}^{m}\sum_{i=1}^{m}\mathbb{1}\left\{Y_i \le Y_j\right\}\\
&+ 2\sum_{j=1}^{m}\sum_{i=1}^{n}\mathbb{1}\left\{X_i \le Y_j\right\} = \sum_{i=1}^{n}\sum_{j=1}^{m}\mathbb{1}\left\{X_i > Y_j\right\} + \sum_{i=1}^{n}\sum_{j=1}^{m}\mathbb{1}\left\{X_i \le Y_j\right\}\\
&+ 2\sum_{j=1}^{m}\sum_{i=1}^{m}\mathbb{1}\left\{Y_{(i)} \le Y_{(j)}\right\} = \sum_{i=1}^{n}\sum_{j=1}^{m}\mathbb{1}\left\{X_i > Y_j \quad \vee \quad X_i \le Y_j\right\}\\
&+ 2\sum_{j=1}^{m}\sum_{i=1}^{m}\mathbb{1}\left\{Y_{(i)} \le Y_{(j)}\right\} = nm + 2\sum_{j=1}^{m} j = nm + m(m+1)\\
&= m(n + m + 1).
\end{aligned}$$

□

2.1 Statistické vlastnosti

V této podkapitole uvedeme, jak vypadají momenty testové statistiky W a ukážeme, že statistika je asymptoticky normální.

2.1.1 Střední hodnota a rozptyl

Předpokládejme, že platí nulová hypotéza H_0. Potom ze symetrie náhodné veličiny U_{ij} ihned plyne, že $\mathbb{E}(\mathrm{W}|H_0) = 0$.
Pro výpočet rozptylu budeme potřebovat napozorovaná data uspořádat do grafického schématu P, jehož části pak využijeme ve vzorci. Daný rozptyl pak bude tímto schématem podmíněn, budeme jej značit var($W|H_0$,P).

Algoritmus pro vytvoření schématu P:

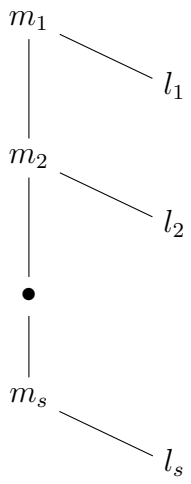
1. Spojíme všechna pozorování z obou výběrů do jednoho sdruženého výběru, který označíme $\boldsymbol{S} = (S_1, \ldots, S_{n+m})^\top$.

2. Z náhodného výběru $\boldsymbol{S}$ vybereme hodnoty necenzorovaných časů, seřadíme je podle velikosti a vytvoříme z nich vektor $\boldsymbol{Z} = (Z_1, \ldots, Z_s)^\top$. Pokud je v $\boldsymbol{S}$ více necenzorovaných pozorování se stejnou hodnotou, ve vektoru $\boldsymbol{Z}$ bude hodnota vždy jenom jednou.

3. Pro i = 1, ..., s označme
$$m_i = \sum_{j=1}^{n+m} \mathbb{1}\{Z_i = S_j\}.$$

4. Pro i = 1, ..., s označme l_i počet cenzorovaných časů , jejichž hodnoty jsou větší nebo rovny Z_i, ale zároveň jsou menší než Z_{i+1}.

Tímto způsobem můžeme reprezentovat libovolnou sadu napozorovaných dat.

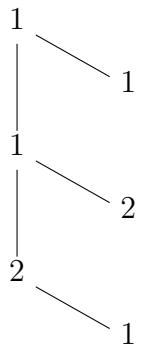


Obrázek 2.1: Grafické znázornění schámatu P.

V případě, kdy se objeví cenzorovaný čas ještě před prvním přesným časem, budeme uvažovat $m_1 = 0$.
Nyní uvedeme krátký příklad, na kterém si postup demonstrujeme.

Příklad
Mějme dva náhodné výběry $\boldsymbol{X} = (3{,}6{,}8+{,}10)^\top$, $\boldsymbol{Y} = (4+{,}7+{,}10{,}12+)^\top$, kde + značí cenzorování. Vytvoření schématu P:

1. Spojíme výběry do jednoho společného, $\boldsymbol{S} = (3{,}6{,}8+{,}10{,}4+{,}7+{,}10{,}12+)^\top$.
2. Seřadíme unikátní necenzorované časy, $\boldsymbol{Z} = (3{,}6{,}10)^\top$.
3. Hodnota 3 se ve výběru $\boldsymbol{S}$ vyskytuje pouze jednou, hodnota 6 také jednou a hodnota 10 dvakrát, tedy $m_1 = 1, m_2 = 1, m_3 = 2$.
4. Pouze cenzorovaný čas 4+ je větší než $Z_1(= 3)$ a zároveň menší než $Z_2(= 6)$, tedy $l_1 = 1$. Mezi Z_2 a $Z_3(= 10)$ leží časy 7+ a 8+, proto $l_2 = 2$. Poslední cenzorovaný čas 12+ je větší než Z_3, tedy opět $l_3 = 1$.



Obrázek 2.2: Schéma P pro výběry $\boldsymbol{X} = (3{,}6{,}8+{,}10)^\top$, $\boldsymbol{Y} = (4+{,}7+{,}10{,}12+)^\top$.

Teď už můžeme uvést vzorec pro výpočet rozptylu testové statistiky W.

$$\begin{aligned}\mathrm{var}(W|H_0{,}P) = \frac{nm}{(n+m)(n+m-1)} \Bigg\{ &\sum_{i=1}^{s} m_i M_{i-1}(M_{i-1}+1) + \sum_{i=1}^{s} l_i M_i(M_i+1) \\ &+ \sum_{i=1}^{s} m_i(n+m-M_i-L_{i-1})(n+m-3M_{i-1}-m_i-L_{i-1}-1) \Bigg\},\end{aligned}$$

kde

$$M_j = \sum_{i=1}^{j} m_i, \text{ pro } M_0 = 0,$$

$$\text{a } L_j = \sum_{i=1}^{j} l_i, \text{ pro } L_0 = 0.$$

Podrobné odvození je uvedeno v Gehan (1965b).

2.1.2 Asymptotická normalita

Ukážeme, že testová statistika W se střední hodnotou 0 a rozptylem $\mathrm{var}(W|H_0{,}P)$ má za platnosti nulové hypotézy asymptoticky normované normální rozdělení. Využijeme toho, že statistika W má tvar dvourozměrné U-statistiky a tato statistika je asymptoticky normální (Lehmann, 1951). Nejprve definujme dvourozměrnou U-statistiku.

Definice 2. *Nechť* $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ *jsou nezávislé náhodné výběry* $X_\alpha = (X_\alpha^{(1)}, X_\alpha^{(2)})$, $Y_\beta = (Y_\beta^{(1)}, Y_\beta^{(2)})$ *z rozdělení s distribučními funkcemi* $F_{X_\alpha}(x)$,$F_{Y_\beta}(y)$, *kde* $x_\alpha = (x_\alpha^{(1)}, x_\alpha^{(2)})$, $y_\beta = (y_\beta^{(1)}, y_\beta^{(2)})$. *Pro n,m* $\geq$ *1 a reálnou funkci t*(X_α,Y_β) *je statistika*

$$U = \frac{1}{nm} \sum_{\alpha=1}^{n} \sum_{\beta=1}^{m} t(X_\alpha, Y_\beta)$$

dvourozměrná U-statistika.

Víme, že pokud pro n $\to \infty$ existuje lim n/m, $\mathbb{E}$[t(X_α,Y_β)] je dobře definována a $\mathbb{E}[t(X_\alpha, Y_\beta)]^2 < \infty$, potom je U asymptoticky normální (Lehmann, 1951). Předpokládejme, že máme pravděpodobnostní rozdělení časů vstupu do studie n+m pacientů. Definujme $x_\alpha = (x_\alpha^{(1)}, x_\alpha^{(2)})$, α = 1, ..., n, kde $x_\alpha^{(1)} = x_i, x_i'$ (přesný čas,cenzorovaný čas) je z $F_{X_\alpha^{(1)}}(x_\alpha^{(1)})$ a $x_\alpha^{(2)}$ je indikátor cenzorování (má hodnotu 1, pokud známe přesný údaj a hodnotu 0, pokud došlo k cenzorování). Analogicky pro y_β. Nyní definujme

$$t(X_\alpha, Y_\beta) = \begin{cases} -1 & \text{pro} \quad x_\alpha^{(1)} < y_\beta^{(1)} \text{ a } (x_\alpha^{(2)}, y_\beta^{(2)}) = (1,1), \\ & \text{nebo} \quad x_\alpha^{(1)} \leq y_\beta^{(1)} \text{ a } (x_\alpha^{(2)}, y_\beta^{(2)}) = (1,0), \\ 1 & \text{pro} \quad x_\alpha^{(1)} > y_\beta^{(1)} \text{ a } (x_\alpha^{(2)}, y_\beta^{(2)}) = (1,1), \\ & \text{nebo} \quad x_\alpha^{(1)} \geq y_\beta^{(1)} \text{ a } (x_\alpha^{(2)}, y_\beta^{(2)}) = (0,1), \\ 0 & \text{jinak.} \end{cases}$$

Potom dostaneme, že statistika U je stejná jako $\frac{W}{nm}$. Protože $\mathbb{E}$[t(X_α,Y_β)] je dobře definována a $\mathbb{E}[t(X_\alpha, Y_\beta)]^2 < \infty$, je asymptotické rozdělení U normální. Platí tedy, že

$$\frac{W}{\sqrt{\text{var}(W|H_0)}} \overset{as.}{\sim} N(0,1).$$

V Gehan (1965b, str. 219) je dále ukázáno, že také

$$\frac{W}{\sqrt{\text{var}(W|H_0,P)}} \overset{as.}{\sim} N(0,1).$$

Označme

$$Z = \frac{W}{\sqrt{\text{var}(W|H_0,P)}}.$$

Test bude zamítat nulovou hypotézu H_0 ve prospěch alternativy $H_1 \Leftrightarrow \text{Z} < -u_{1-\alpha}$, případně ve prospěch oboustranné alternativy $H_2 \Leftrightarrow |Z| \geq u_{1-\frac{\alpha}{2}}$.

2.2 Zobecněný Wilcoxonův test pro dvojitě cenzorovaná data

Někdy se může stát, že naše data obsahují zároveň data cenzorovaná zprava i zleva. V takovém případě mluvíme o tzv. dvojitě cenzorovaných datech. Jednoduchou modifikací testové statistiky W můžeme odvodit test, který bude aplikovatelný i na tato data. Test byl popsán v Gehan (1965a).

Předpokládejme, že máme $n + m$ jedinců, které náhodně rozdělíme do dvou skupin A a B.

Napozorovaná data:

$$\left.\begin{array}{ll} X_1^{'},\ldots,X_{r_n}^{'} & r_n \text{ zprava cenzorovaných časů} \\ X_{r_n+1}^{''},\ldots,X_{r_n+s_n}^{''} & s_n \text{ zleva cenzorovaných časů} \\ X_{r_n+s_n+1},\ldots,X_n & n-r_n-s_n \text{ přesných časů} \end{array}\right\} \text{skupina A}$$

$$\left.\begin{array}{ll} Y_1^{'},\ldots,Y_{r_m}^{'} & r_m \text{ zprava cenzorovaných časů} \\ Y_{r_m+1}^{''},\ldots,Y_{r_m+s_m}^{''} & s_m \text{ zleva cenzorovaných časů} \\ Y_{r_m+s_m+1},\ldots,Y_m & m-r_m-s_m \text{ přesných časů} \end{array}\right\} \text{skupina B}$$

kde X_i,Y_j jsou přesné údaje, $X_i^{'}$,$Y_j^{'}$ údaje cenzorované zprava a $X_i^{''}$,$Y_j^{''}$ údaje cenzorované zleva. Pozorování jsou z distribučních funkcí $F_1(x), F_2(y)$, které mohou být diskrétní nebo spojité. Předpokládáme, že v obou výběrech se mohou vyskytovat všechny typy dat, jak údaje cenzorované zprava, tak zleva. Budeme testovat nulovou hypotézu

$$H_0 : F_1(t) = F_2(t) \ \ (t \leq C_r),$$

buď proti jednostranné alternativě

$$H_1 : F_1(t) < F_2(t) \ \ (t \leq C_r),$$

nebo oboustranné alternativě

$$H_2 : F_1(t) < F_2(t) \ \text{ nebo } \ F_1(t) > F_2(t) \ \ (t \leq C_r),$$

kde C_r je horní limit pro pozorování.
Definujme U_{ij}:

$$U_{ij} = \begin{cases} -1 & X_i < Y_j \ \vee \ X_i \leq Y_j^{'} \ \vee \ X_i^{''} \leq Y_j^{'}, Y_j; \\ 1 & X_i > Y_j \ \vee \ X_i^{'} \geq Y_j \ \vee \ X_i^{'}, X_i \geq Y_j^{''}; \\ 0 & \text{jinak.} \end{cases}$$

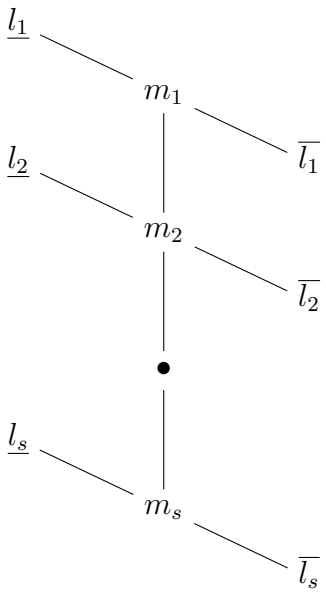
Testovou statistiku zobecněného Wilcoxonova testu pro dvojitě cenzorovaná data můžeme zapsat ve tvaru

$$W = \sum_{i=1}^{n} \sum_{j=1}^{m} U_{ij}.$$

Za předpokladu platnosti nulové hypotézy platí $\mathbb{E}(\mathrm{W}|H_0) = 0$ ze symetrie.
Pro výpočet rozptylu budeme opět potřebovat data uspořádat do schématu P, které je definováno obdobně jako pro případ jednoduše cenzorovaných dat. Mějme $\boldsymbol{S} = (S_1, \ldots, S_{n+m})^\top$ sdružený výběr a $\boldsymbol{Z} = (Z_1, \ldots, Z_s)^\top$ seřazené hodnoty necenzorovaných dat. Pro i = 1, ..., s označme

$$m_i = \sum_{j=1}^{n+m} \mathbb{1}\left\{Z_i = S_j\right\},$$

$\underline{l_i}$ počet zleva cenzorovaných časů , jejichž hodnoty jsou menší než Z_i, ale zároveň jsou větší než Z_{i-1} a $\overline{l_i}$ počet zprava cenzorovaných časů, jejichž hodnoty jsou větší než Z_i a zároveň jsou menší než Z_{i+1}.



Obrázek 2.3: Grafické znázornění schématu P pro dvojitě cenzorovaná data.

Vzorec pro výpočet rozptylu má tvar:

$$\begin{aligned}\mathrm{var}(W|H_0,P) = \frac{nm}{(n+m)(n+m-1)} \Bigg\{ &\sum_{i=1}^{s} m_i(M_{i-1}+\underline{L_i})(M_{i-1}+\underline{L_i}+1) \\ &+ \sum_{i=1}^{s} l_i(M_i+\underline{L_i})(M_i+\underline{L_i}+1) \\ + \sum_{i=1}^{s} m_i(M_p+\overline{L_p}-M_i-\overline{L}_{i-1})&(M_p+\overline{L_p}-3M_{i-1}-\overline{L}_{i-1}-2\underline{L_i}-m_i-1) \\ &+ \sum_{i=1}^{s} \underline{l_i}(M_p+\overline{L_p}-M_{i-1}-\overline{L}_{i-1})(M_p+\overline{L_p}-M_{i-1}-\overline{L}_{i-1}-1)\Bigg\},\end{aligned}$$

kde

$$\begin{aligned} M_j &= \sum_{i=1}^{j} m_i, \quad M_0 = 0, \\ \overline{L_j} &= \sum_{i=1}^{j} \overline{l_i}, \quad \overline{L_0} = 0, \\ \underline{L_j} &= \sum_{i=1}^{j} \underline{l_i}, \quad \underline{L_0} = 0. \end{aligned}$$

Také platí

$$Z = \frac{W}{\sqrt{\mathrm{var}(W|H_0,P)}} \overset{as.}{\sim} N(0,1).$$

Zamítáme tedy nulovou hypotézu H_0 ve prospěch alternativy $H_1 \Leftrightarrow \mathrm{Z} < -u_{1-\alpha}$, případně ve prospěch alternativy $H_2 \Leftrightarrow |Z| \geq u_{1-\frac{\alpha}{2}}$.

3. Simulační studie

V této kapitole si ukážeme krátkou simulační studii, na které představíme statistické vlastnosti testu. Nejprve zobecněný Wilcoxonův test pro cenzorovaná data porovnáme s klasickým dvouvýběrovým Wilcoxonovým testem pro případ, kdybychom cenzorovaná data vynechali a následně ukážeme, jak míra cenzorování ovlivňuje sílu a hladinu testu. K simulacím použijeme statistický software R.

3.1 Příklad 1

Budeme simulovat klinickou studii, do které je zapojeno $2n$ pacientů, n pacientům je aplikována léčba A, zbylým n léčba B. Studie bude probíhat po dobu 50 let, pacienti do ní můžou vstoupit kdykoliv během prvních 20 let od zahájení. Zaznamenávat budeme délku života pacientů od vstupu do studie, tedy po aplikování léčby. Po ukončení studii vyhodnotíme a u pacientů, kteří budou stále naživu, dojde k cenzorování. Naším cílem je zjistit, zda schopnost prodloužit délku života pacientů se u léčeb liší či nikoli.

Simulaci provedeme následovně. Označme $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ náhodný výběr odpovídající údajům z první skupiny, která reprezentuje léčbu A, $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ náhodný výběr ze skupiny druhé s léčbou B. Teoretickou délku života pacientů v případě první skupiny popisuje distribuční funkce $F_{\boldsymbol{X}}$, v případě druhé distribuční funkce $F_{\boldsymbol{Y}}$. V našem konkrétním případě budeme uvažovat $X_i \sim Exp(0.05)$, $Y_j \sim Exp(0.1)$, $i,j = 1, \ldots, n$. Každé hodnotě dále náhodně přiřadíme číslo vygenerované z rovnoměrného rozdělení na intervalu [0,20], které reprezentuje moment, kdy pacient vstoupil do studie. Pokud teoretická délka života pacienta přičtená k údaji o vstupu do studie přesáhne 50 let, nahradíme údaj o délce jeho života údajem pouze o tom, jak dlouho jsme pacienta sledovali a tento čas označíme za cenzorovaný.

Na tato data pak aplikujeme Wilcoxonův test a jeho zobecněnou verzi. Jelikož klasický Wilcoxonův test není definovaný pro cenzorovaná data, použijeme naivní řešení a tato data jednoduše vynecháme. Porovnávat budeme sílu obou testů. Budeme testovat $H_0 : F_X = F_Y$ proti $H_1 : F_X \neq F_Y$ a test provedeme na hladině $\alpha = 0.05$. Uložíme si příslušné p-hodnoty testů a celý postup zopakujeme 1000krát. Odhad síly pak spočteme jako relativní četnost případů, kde jsme hypotézu H_0 zamítli. Tuto simulaci provedeme pro různé rozsahy výběrů.
Data, která získáme z výše uvedené simulace jsou uvedena v Tabulce 3.1, pro grafické znázornění se můžeme podívat na Obrázek 3.1.

Vidíme, že zobecněný Wilcoxonův test nám dává dobré výsledky, stačí mít v každé skupině přibližně 50 pacientů a více jak v 80ti procentech případů test správně zamítne nulovou hypotézu. Oproti tomu vidíme, že použití klasického dvouvýběrového Wilcoxonova testu na tento typ studie je absolutně nevhodné. To je způsobeno nejenom tím, že vynecháním cenzorovaných dat si snížíme rozsah výběru, ale také se tím připravíme o částečné informace, které nám cenzorovaná data mohla poskytnout.

Rozsah	Wilcoxonův test	Zobecněný Wilcoxonův test
n=10	0.115	0.229
n=20	0.230	0.449
n=30	0.295	0.628
n=40	0.383	0.731
n=50	0.474	0.820
n=60	0.550	0.883
n=70	0.633	0.948
n=80	0.690	0.964
n=90	0.724	0.985
n=100	0.777	0.989

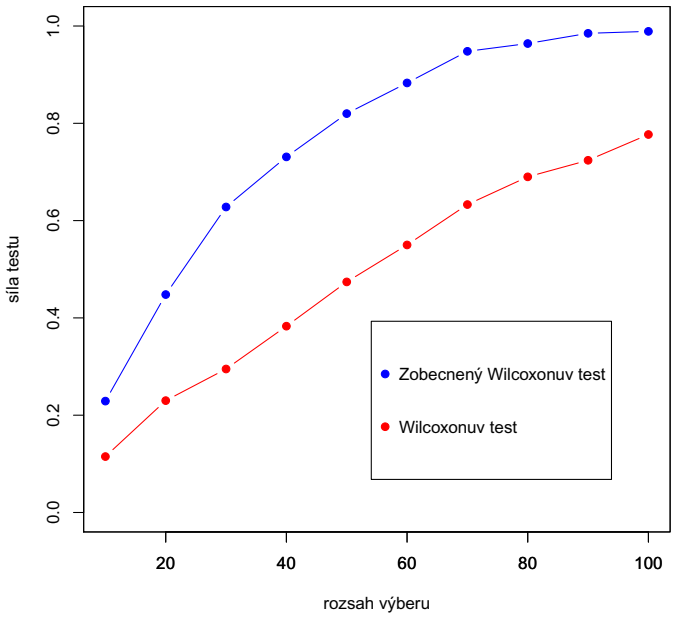
Tabulka 3.1: Odhad síly Wilcoxonova testu a zobecněného Wilcoxonova testu pro
cenzorovaná data v případě, kdy $X_i \sim Exp(0.05),\ Y_j \sim Exp(0.1)$.

3.2 Příklad 2

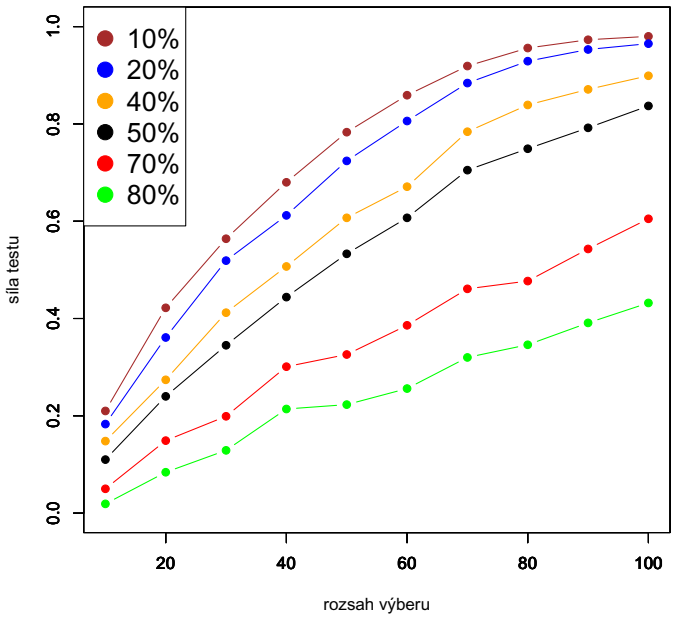
Na této simulaci si ukážeme, jak nám míra cenzorování ovlivní sílu a hladinu zobecněného testu. Abychom měli procento cenzorovaných dat přesně pod kontrolou, cenzorování zde budeme simulovat pomocí binomického rozdělení. Nejprve si nagenerujeme data z exponenciálního rozdělení, stejně jako v předchozím příkladě $X_i \sim Exp(0.05),\ Y_j \sim Exp(0.1)$. Následně si nagenerujeme náhodný výběr z binomického rozdělení, který nám bude říkat zda se jedná o přesný či cenzorovaný čas. Změnou pravděpodobnosti úspěchu u binomického rozdělení budeme regulovat procento cenzorovaných dat. Poté na data aplikujeme zobecněný Wilcoxonův test a stejně jako výše odhadneme sílu testu v závislosti na rozsahu výběru. Získané hodnoty jsou uvedeny v Tabulce 3.2, příslušný graf na Obrázku 3.2.

Rozsah	10%	20%	40%	50%	70%	80%
n=10	0.210	0.183	0.148	0.110	0.050	0.019
n=20	0.422	0.361	0.274	0.240	0.149	0.084
n=30	0.564	0.519	0.412	0.345	0.199	0.129
n=40	0.680	0.612	0.507	0.444	0.301	0.214
n=50	0.783	0.724	0.607	0.533	0.326	0.223
n=60	0.859	0.806	0.671	0.607	0.386	0.256
n=70	0.919	0.884	0.784	0.705	0.461	0.320
n=80	0.956	0.929	0.839	0.749	0.477	0.346
n=90	0.973	0.953	0.871	0.792	0.543	0.391
n=100	0.980	0.965	0.899	0.837	0.605	0.432

Tabulka 3.2: Odhad síly zobecněného Wilcoxonova testu pro různá procenta cenzorovaných dat a rozsahy výběrů, $X_i \sim Exp(0.05),\ Y_j \sim Exp(0.1)$.



Obrázek 3.1: Porovnání síly Wilcoxonova testu a zobecněného Wilcoxonova testu.



Obrázek 3.2: Síla zobecněného Wilcoxonova testu v závislosti na rozsahu výběru pro různá procenta cenzorovaných dat, $X_i \sim Exp(0.05)$, $Y_j \sim Exp(0.1)$.

Vidíme, že ačkoli pro vyšší procento cenzorovaných dat má zobecněný test nižší sílu, ve všech případech se zvyšujícím se rozsahem výběru síla roste. Tedy pro $n \to \infty$ síla asymptoticky konverguje k 1.

Ještě se můžeme podívat, jak cenzorování ovlivňuje hladinu testu. Simulace provedeme pro pevný rozsah výběru $n = 100$ a za platnosti nulové hypotézy, tedy $X_i \sim Exp(0.05)$ i $Y_j \sim Exp(0.05)$.

míra cenzorování (v %)	**hladina testu**
0	0.044
10	0.040
20	0.042
30	0.046
40	0.045
50	0.055
60	0.041
70	0.047
80	0.050
90	0.031

Tabulka 3.3: Hladina zobecněného Wilcoxonova testu pro různé procento cenzorovaných dat.

V Tabulce 3.3 můžeme vidět, že míra cenzorování nijak zásadně hladinu testu neovlivňuje a že v téměř všech případech zobecněný Wilcoxonův test dodržuje předepsanou hladinu $\alpha = 0.05$. V tomto konkrétním případě, kdy n=100, $X_i \sim Exp(0.05)$ a $Y_j \sim Exp(0.05)$ však dobře dodržuje hladinu i klasický dvouvýběrový Wilcoxonův test, jak můžeme vidět v Tabulce 3.4.

míra cenzorování (v %)	**hladina testu**
0	0.044
10	0.040
20	0.039
30	0.047
40	0.044
50	0.057
60	0.049
70	0.055
80	0.056
90	0.050

Tabulka 3.4: Hladina klasického Wilcoxonova testu pro různé procento cenzorovaných dat.

Závěr

Tato bakalářská práce se zabývala zobecněným Wilcoxonovým testem, který je určen pro cenzorovaná a dvojitě cenzorovaná data.

V první kapitole byl nejprve popsán jednovýběrový a dvouvýběrový Wilcoxonův test a jejich základní vlastnosti, v závěru kapitoly cenzorovaná data. K cenzorování může docházet z různých důvodů, podle toho se i liší různé typy cenzorovaných dat. V práci jsou popsány ty typy, se kterými se v praxi můžeme setkat nejčastěji.

V kapitole druhé byl představen už samotný zobecněný Wilcoxonův test. Motivací k tomuto zobecnění jsou lékařské studie, při kterých se snažíme porovnat dvě alternativní léčby a jejich účinnost. Údaje získané z těchto studií často obsahují právě data cenzorovaná. V práci je nejprve popsán test pro použití na jednoduše cenzorovaných datech, snadnou modifikací testové statistiky byl získán i test pro dvojitě cenzorovaná data.

Závěr práce se věnuje simulační studii, na které jsou předvedeny statistické vlastnosti testu. Nejprve je sledována síla zobecněného testu v porovnání se silou standardního Wilcoxonova testu, který použití cenzorovaných dat neuvažuje. Ukazuje se, že použitím zobecněného Wilcoxonova testu, který cenzorovaná data využívá, dosáhneme výrazně lepších výsledků. Ve druhé části pak sledujeme, jak je síla a hladina zobecněného testu ovlivněna mírou cenzorování.

Seznam použité literatury

Gehan, E. A. (1965a). A Generalized Two-Sample Wilcoxon Test for Doubly Censored Data. *Biometrika*, **52**, 650–653.

Gehan, E. A. (1965b). A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples. *Biometrika*, **52**, 203–223.

Klein, J. P. a Moeschberger, M. L. (2003). *Survival Analysis*. Springer.

Lehmann, E. L. (1951). Consistency and unbiasedness of certain nonparametric tests. *The Annals of Mathematical Statistics*, **22**, 165–179.

Omelka, M. (2019). Poznámky k přednášce NMSA331 Matematická statistika. Naposledy navštíveno 11. 7. 2019. URL `https://www.karlin.mff.cuni.cz/~omelka/Soubory/nmsa331/ms1.pdf`.