MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Haojun Zhan

Párové testy a jejich interpretace

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. Ing. Marek Omelka, Ph.D.
Studijní program: Finanční matematika

Praha 2024

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného titulu.

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

Rád bych poděkoval vedoucímu své bakalářské práce, doc. Ing. Marku Omelkovi, Ph.D., za jeho odborné vedení, cenné rady a návrhy na zlepšení. Zároveň děkuji mé rodině za jejich podporu a pochopení při studiu.

Název práce: Párové testy a jejich interpretace

Autor: Haojun Zhan

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. Ing. Marek Omelka, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Tato bakalářská práce se zaměřuje na párové testy a jejich interpretaci, přičemž klade důraz na pochopení jejich významu v rámci statistické analýzy. Nejdříve zavedeme značení a zformulujeme základní předpoklady pro párové testy. V práci představíme celkem tři typy párových testů. Jedná se o párový t-test, párový znaménkový test a párový Wilcoxonův test. Párový t-test budeme porovnávat s dvouvýběrovým t-testem. Dále u párového znaménkového testu se budeme podrobně zabývat interpretací testu. Zde bude naším cílem zjistit, kdy se medián rozdílu dvou náhodných veličin rovná rozdílu mediánů těchto náhodných veličin. Nakonec u párového Wilcoxonova testu budeme hledat podobné způsoby interpretace testu jako u párového znaménkového testu.

Klíčová slova: párový t-test, párový Wilcoxonův test, párový znaménkový test

Title: Paired tests and their interpretation

Author: Haojun Zhan

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. Ing. Marek Omelka, Ph.D., Department of Probability and Mathematical Statistics

Abstract: This bachelor's thesis focuses on paired tests and their interpretation, with an emphasis on understanding their significance in statistical analysis. First, we will introduce notation and formulate the basic assumptions for paired tests. In this work, we will present a total of three types of paired tests. These are paired sample t-test, paired sample sign test, and Wilcoxon signed-rank test. We will compare the paired t-test with the two-sample t-test. Next, with the paired sample sign test, we will examine the interpretation of the test in detail. Our goal here will be to find out when the median of the difference of two random variables is equal to the difference of the medians of these random variables. Finally, with the Wilcoxon signed-rank test, we will look for similar ways of interpreting the test as with the paired sample sign test.

Keywords: paired sample sign test, Wilcoxon signed-rank test, paired sample t-test

Obsah

Značení

$\mathcal{F}$	pravděpodobnostní model pro pozorovaná data
$\mathbb{R}$	množina reálných čísel
v^T	transpozice vektoru v
P	pravděpodobnost
X	náhodná veličina
$X \sim \mathcal{L}$	X má rozdělení $\mathcal{L}$
$\mathsf{E}\, X$, μ_X	střední hodnota náhodné veličiny X
$\operatorname{var} X$, σ_X^2	rozptyl náhodné veličiny X
$\operatorname{cov}(X, Y)$	kovariance náhodných veličin X, Y
f_X	hustota náhodné veličiny X
f_{XY}	sdružená hustota náhodného vektoru $(X, Y)^T$
F_X	distribuční funkce náhodné veličiny X
F_{XY}	sdružená distribuční funkce náhodného vektoru $(X, Y)^T$
F_X^{-1}	kvantilová funkce náhodné veličiny X
$F_X^{-1}(1/2)$, m_X	medián náhodné veličiny X
$\hat{x}$	modus náhodné veličiny X
$\mathcal{N}(0, 1)$	normované normální rozdělení
$\overline{X}_n$	výběrový průměr náhodného výběru $X_1, \ldots, X_n$
S_X^2	výběrový rozptyl náhodného výběru $X_1, \ldots, X_n$
S_X	výběrová směrodatná odchylka náhodného výběru $X_1, \ldots, X_n$
$S_{X,Y}$	výběrová kovariance dvojic $(X_1, Y_1)^T, \ldots, (X_n, Y_n)^T$
H_0	nulová hypotéza
H_1	alternativa
u_α	α-kvantil rozdělení $\mathcal{N}$(0,1)
$t_n(\alpha)$	α-kvantil t-rozdělení s n stupni volnosti
$\xrightarrow[n\to\infty]{\mathcal{D}}$	konvergence v distribuci
$\xrightarrow[n\to\infty]{\mathsf{P}}$	konvergence v pravděpodonosti
$\mathcal{L}^2$	množina reálných náhodných veličin, pro které $\mathsf{E}\, X^2 < \infty$ a $\operatorname{var} X > 0$

Úvod

Při testování hypotézy někdy potřebudeme porovnávat dvě závislé skupiny, v takovém případě použijeme párové testy. Tyto statistické testy nám umožňují posoudit, zda existuje nějaký statistický rozdíl mezi dvěma měřeními provedenými na stejných subjektech.

Tato práce se zabývá párovými testy a jejich interpretacemi. V této práci představíme tři statistické testy, které jsou v praxi hojně využívány. Při tvorbě této práce budeme převážně vycházet z knih Anděl (2011), Anděl (2019) a z článku Hutson a Yu (2023). Budeme se opírat o informace, metodologie a výsledky uvedené v těchto odborných textech jako základ pro naši práci.

Základní předpoklady

V celé této práci budeme uvažovat náhodný výběr

$$\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \dots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$$

z libovolného dvourozměrného rozdělení se sdruženou distribuční funkcí F_{XY} pro náhodný vektor $(X,Y)^T$, kde rozsah výběru $n \geq 2$. Náhodné veličiny X a Y nejsou na sobě nezávislé. Nechť náhodná veličina X má distribuční funkci F_X a náhodná veličina Y má distribuční funkci F_Y. Pro zjednodušení výkladu klíčových bodů budeme předpokládat, že náhodný vektor $(X,Y)^T$ má spojité rozdělení a střední hodnoty $\mathsf{E}\, X$ a $\mathsf{E}\, Y$ jsou konečné.

Označme

$$Z_i := X_i - Y_i \ (i = 1, \dots, n)$$

jako rozdíly pozorování. Potom $Z_1, \dots, Z_n$ jsou stějně nezávislé rozdělené a pochází z rozdělení s distribuční funkcí F_Z pro náhodnou veličinu Z, která má střední hodnotu μ_Z, rozptyl σ_Z^2 a medián m_Z. Pozorování $Z_1, \dots, Z_n$ můžeme interpretovat jako rozdíly stavů na počátku a na konci nějakého vyšetření. Tato transformace ze dvourozměrného problému na jednorozměrný problém nám umožňuje použít jednovýběrové testy pro analýzu rozdílů mezi párovými pozorováními, což usnadňuje statistickou analýzu a interpretaci výsledků.

U jednotlivých testů budeme uvádět předpoklady, nulovou hypotézu a alternativu, testovou statistiku, kritický obor a p-hodnotu. Jako první krok při testování hypotéz párovými testy je ověřit, zda získaná data vyhovují předpokladům. U některých testů je potřeba předpokládat například normalitu dat, konečnost druhých momentů náhodné veličiny nebo symetrii hustoty kolem nějakého bodu. Nulovou hypotézu lze chápat jako výchozí tvrzení, které snažíme testovat pomocí statistické analýzy. Proti nulové hypotéze sestavujeme alternativu, kterou se snažíme prokázat. Kritický obor je množina hodnot pro testovou statistiku, která vede k zamítnutí nulové hypotézy na základě předem stanovené hladiny významnosti testu. P-hodnotu rozumíme nejmenší hladinu testu, na které bychom ještě nulovou hypotézu H_0 zamítli.

Členění práce

Práce je rozdělena do tří kapitol. V první kapitole je představen párový t-test, kde zformulujeme předpoklady a především se zaměřujeme na porovnání testových statistik párového t-testu a dvouvýběrového t-testu.

Druhá kapitola je věnována párovému znaménkovému testu. V rámci této kapitoly budeme postupně rozvolňovat běžné předpoklady, a tím získáme nové způsoby interpretace testu. Nejvíce pozornosti pak věnujeme otázce, kdy se medián rozdílu dvou náhodných veličin rovná rozdílu mediánů těchto náhodných veličin.

V poslední kapitole se budeme zabývat párovým Wilcoxonovým testem, kdy opět budeme rozvolňovat běžné předpoklady pro získání nových poznatků.

1. Párový t-test

Jedním z častých nástrojů používaných k porovnání dvou skupin jsou párové t-testy. Tato statistická metoda nám umožňuje posoudit, zda existuje nějaký statistický rozdíl mezi dvěma měřeními provedenými na stejných subjektech.

Nejdříve se podíváme na předpoklady párového t-testu, které jsou popsány v Anděl (2011). Poté prozkoumáme vztahy párového t-testu s dvouvýběrovým t-testem.

1.1 Formulace úlohy

Uvažujeme model

$$\mathcal{F}_t = \{Z_i = X_i - Y_i \sim \mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\},$$

kde parametry μ, σ^2 jsou neznámé, nebo

$$\mathcal{F}_z = \{Z_i = X_i - Y_i \in \mathcal{L}^2\},$$

kde $\mathcal{L}^2$ je množina rozdělení s konečným druhým momentem a nenulovým rozptylem. Model $\mathcal{F}_t$ označuje množinu náhodných veličin z normálního rozdělení s parametry μ, σ^2 a je předpokladem pro přesný párový t-test. Model $\mathcal{F}_z$ je předpokladem pro asymptotický párový t-test. Zřejmě platí $\mathcal{F}_t \subseteq \mathcal{F}_z$.

Označme aritmetické průměry $\overline{X}_n$, $\overline{Y}_n$ a výběrové rozptyly S_X^2, S_Y^2 obou výběrů:

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2, S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2.$$

Párový t-test je založen na principu porovnání průměrů dvou vzorků párových dat, tedy na testování rozdílu středních hodnot $\mathsf{E}\, X_i = \mu_X$ a $\mathsf{E}\, Y_i = \mu_Y$. Testujeme nulovou hypotézu

$$H_0 : \mu_X - \mu_Y = d_0$$

proti alternativě

$$H_1 : \mu_X - \mu_Y \neq d_0,$$

kde $d_0 \in \mathbb{R}$ je předem daná konstanta. Tedy testujeme, zda střední hodnoty dvou skupin jsou stejné, nebo se vzájemně liší o hodnotu d_0. Nejčastěji volíme $d_0 = 0$. Hypotéza H_0 pak tvrdí, že střední hodnoty dvou skupin jsou stejné, a naopak alternativa H_1 tvrdí, že střední hodnoty dvou skupin nejsou stejné. Testová statistika párového t-testu je definována ve tvaru

$$T_n = \sqrt{n}\,\frac{\overline{Z}_n - d_0}{S_Z},$$

kde $\overline{Z}_n = \overline{X}_n - \overline{Y}_n$ je aritmetický průměr rozdílů $Z_i = X_i - Y_i$ a S_Z je výběrová směrodatná odchylka rozdílů $Z_i = X_i - Y_i$. Výběrový rozptyl S_Z^2 je definován jako

$$S_Z^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \overline{Z}_n)^2.$$

Dá se ukázat, že za předpokladu platnosti nulové hypotézy H_0 v modelu $\mathcal{F}_t$ platí $T_n \sim t_{n-1}$ (viz např. Dupač a Hušková, 2005, věta 5.9. (c)), kde t_{n-1} značí t-rozdělení s $n-1$ stupni volnosti. Nulovou hypotézu H_0 pro přesný test zamítneme na hladině α, právě když

$$|T_n| \geq t_{n-1}(1-\alpha/2), \tag{1.1}$$

kde $t_{n-1}(1-\alpha/2)$ je $(1-\alpha/2)$-tý kvantil t-rozdělení s $n-1$ stupni volnosti. Rovněž se dá ukázat, že za platnosti nulové hypotézy H_0 v modelu $\mathcal{F}_z$ platí

$$T_n \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0,1).$$

Nulovou hypotézu H_0 pro asymptotický test zamítneme na hladině α, právě když

$$|T_n| \geq u_{1-\alpha/2},$$

kde $u_{1-\alpha/2}$ je $(1-\alpha/2)$-tý kvantil normovaného normálního rozdělení. V praxi se obvykle volí zamítací kritérium (1.1). Pro velká n (tj. $n \to \infty$) se kvantil t-rozdělení asymptoticky blíží ke kvantilu normovaného normálního rozdělení, proto pro zamítnutí nulové hypotézy můžeme použít zamítací kritérium pro přesný párový t-test.

V modelu $\mathcal{F}_t$ budou vypočtené p-hodnoty a intervaly spolehlivosti přesné, ale v modelu $\mathcal{F}_z$ pouze asymptotické. P-hodnota párového t-testu je definována následujícím vzorcem

$$p = 2(1 - F_n(|T_n|)),$$

kde T_n je pozorovaná hodnota testové statistiky a F_n je distribuční funkce rozdělení t_{n-1}. Oboustranný interval spolehlivosti pro rozdíl $\mu_X - \mu_Y$ na hladině $1-\alpha$ je

$$\left(\overline{Z}_n - \frac{S_Z}{\sqrt{n}}\, t_{n-1}(1-\alpha/2), \overline{Z}_n + \frac{S_Z}{\sqrt{n}}\, t_{n-1}(1-\alpha/2)\right),$$

viz Dupač a Hušková (2005, věta 5.10. (a)).

1.2 Vztahy t-testů

V této sekci budeme porovnávat párový t-test s dvouvýběrovým t-testem. Zajímá nás, za jakých předpokladů jsou testy ekvivalentní, tj. oba testy vydávají s pravděpodobností 1 totéž rozhodnutí (buď zamítneme, nebo nezamítneme nulovou hypotézu). Nejprve musíme zformulovat hypotézu dvouvýběrového t-testu, a poté definujeme jeho testovou statistiku.

1.2.1 Dvouvýběrový t-test

Uvažujme nezávislé náhodné výběry $X_1, \ldots, X_n$ a $Y_1, \ldots, Y_m$ s příslušnými distribučními funkcemi F_X a F_Y. Modelem dvouvýběrového t-testu je

$$\mathcal{F}_t^* = \{F_X, F_Y \text{ mají konečné rozptyly}\},$$

Testované parametry jsou střední hodnoty $\mu_X = \mathsf{E}\, X_i$, $\mu_Y = \mathsf{E}\, Y_j$. Testujeme nulovou hypotézu

$$H_0 : \mu_X - \mu_Y = d_0$$

proti alternativě

$$H_1 : \mu_X - \mu_Y \neq d_0,$$

kde $d_0 \in \mathbb{R}$ je předem daná konstanta.

Testová statistika dvouvýběrového t-testu je tvaru

$$T_{n,m} = \frac{\overline{X}_n - \overline{Y}_m - d_0}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}.$$

Dá se ukázat, že za platnosti nulové hypotézy H_0 v modelu $\mathcal{F}_t^*$ má testová statistika $T_{n,m}$ asymptoticky rozdělení $\mathcal{N}(0,1)$. Nulovou hypotézu H_0 zamítneme na hladině α, právě když

$$|T_{n,m}| \geq u_{1-\alpha/2},$$

kde $u_{1-\alpha/2}$ je $(1-\alpha/2)$-tý kvantil normovaného normálního rozdělení. V případě, že budeme navíc předpokládat, že oba náhodné výběry pocházejí z normálního rozdělení se stejným rozptylem $\sigma^2 > 0$, pak má testová statistika $T_{n,m}$ za platnosti nulové hypotézy $H_0 : \mu_X - \mu_Y = d_0$ přesně t-rozdělení s $n+m-2$ stupni volnosti, viz Anděl (2011, věta 4.26).

P-hodnota je

$$p = 2(1 - G(|T_{n,m}|)),$$

kde $T_{n,m}$ je pozorovaná hodnota testové statistiky a G je distribuční funkce rozdělení t_{n+m-2}.

1.2.2 Porovnání testových statistik

Z sekce 1.2.1 víme, že testová statistika dvouvýběrového t-testu je

$$T_{n,m} = \frac{\overline{X}_n - \overline{Y}_m - d_0}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}.$$

Nyní předpokládejme, že rozsahy obou výběrů jsou stejné, tj. $n = m$. Pak platí

$$T_{n,n} = \frac{\overline{X}_n - \overline{Y}_n - d_0}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{n}}} = \sqrt{n}\,\frac{\overline{X}_n - \overline{Y}_n - d_0}{\sqrt{S_X^2 + S_Y^2}}.$$

Z sekce 1.1 víme, že testová statistika párového t-testu je

$$T_n = \sqrt{n}\,\frac{\overline{Z}_n - d_0}{S_Z} = \sqrt{n}\,\frac{\overline{X}_n - \overline{Y}_n - d_0}{\sqrt{S_Z^2}}.$$

Výběrový rozptyl S_Z^2 vyjádříme pomocí výběrových rozptylů S_X^2, S_Y^2 a výběrové kovariance $S_{X,Y}$.

$$\begin{aligned}
S_Z^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \overline{Z}_n)^2 = \frac{1}{n-1}\sum_{i=1}^{n}[X_i - Y_i - (\overline{X}_n - \overline{Y}_n)]^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 - \frac{2}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)(Y_i - \overline{Y}_n) + \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2 \\
&= S_X^2 - 2\,S_{X,Y} + S_Y^2
\end{aligned}$$

Pak platí
$$T_n = \sqrt{n}\,\frac{\overline{X}_n - \overline{Y}_n - d_0}{\sqrt{S_X^2 - 2\,S_{X,Y} + S_Y^2}}.$$

Porovnáním testových statistik T_n, $T_{n,n}$ zjistíme, že se liší pouze jejich jmenovatelé, kdy pod odmocninou v T_n oproti $T_{n,n}$ se objevuje navíc výraz $-2\,S_{X,Y}$, kde $S_{X,Y}$ je výběrová kovariance. Mají-li náhodné veličiny X_i a Y_i konečné druhé momenty, pak výběrová kovariance $S_{X,Y}$ je nestranný a kozistentní odhad kovariance $\mathrm{cov}(X_i, Y_i)$.

Nyní budeme analyzovat hodnoty výběrové kovariance $S_{X,Y}$, která může nabývat jakýchkoli reálných hodnot. Zajímá nás, co se stane, když chybně použijeme dvouvýběrový t-test na situaci, kdy jsme měli použit párový t-test. Budeme porovnávat testové statistiky T_n pro párový t-test a $T_{n,n}$ pro dvouvýběrový t-test. Tím zjistíme, jak se změní p-hodnota, přičemž bychom neměli zapomínat, že zatímco T_n porovnáváme s kvantily t-rozdělení s $n-1$ stupni volnosti, tak $T_{n,n}$ s kvantily t-rozdělení s $2(n-1)$ stupni volnosti.

Testové statistiky jsou stejné, pokud je výběrovová kovariance nulová. Uvědomme si, že u dvouvýběrového t-testu předpokládáme dva nezávislé náhodné výběry, zatímco u párového t-testu uvažujeme dva náhodné výběry, které nejsou nezávislé.

Je-li hodnota výběrové kovariance mezi X_i a Y_i kladná, pak platí $T_n > T_{n,n}$. V takovém případě chybné použití dvouvýběrového t-testu místo párového t-testu vede k menší hodnotě testové statistiky a vyšší p-hodnotě. Při použití dvouvýběrového t-testu místo párového t-testu může dojít k podcenění statistické významnosti rozdílů mezi dvěma skupinami a může se stát, že nezamítneme nulovou hypotézu, i když bychom ji měli zamítnout. V takovém případě se tedy dopouštíme chyby druhého druhu.

Podobně, je-li hodnota výběrové kovariance mezi X_i a Y_i záporná, pak platí $T_n < T_{n,n}$. V takovém případě chybné použití dvouvýběrového t-testu místo párového t-testu vede k vyšší hodnotě testové statistiky a nižší p-hodnotě. Při použití dvouvýběrového t-testu místo párového t-testu může dojít k nadhodnocení statistické významnosti rozdílů mezi dvěma skupinami a může se stát, že zamítneme nulovou hypotézu i v případě, že by neměla být zamítnuta. V takovém případě se tedy dopouštíme chyby prvního druhu.

2. Párový znaménkový test

Párový znaménkový test se opět zaměřuje na porovnání rozdílů mezi dvěma párovými měřeními. V tomto kontextu se používají znaménka rozdílů, což jsou indikátory směru změny pro každý pár. Jeho základní myšlenkou je minimalizovat vliv variabilnosti mezi subjekty tím, že se zaměřuje na rozdíly v rámci jednotlivých párových jednotek. To nám umožňuje identifikovat, zda došlo ke statisticky významné změně mezi dvěma měřeními.

V rámci této kapitoly budeme analyzovat předpoklady párového znaménkového testu převzaté z knihy Anděl (2011) a dále budeme interpretovat párový znaménkový test s využitím Hutson a Yu (2023).

2.1 Formulace úlohy

Uvažujeme model

$$\mathcal{F} = \{Z_i = X_i - Y_i \text{ mají spojitá rozdělení}\}.$$

Párový znaménkový test zkoumá medián rozdílu náhodný veličin X, Y. Testujeme nulovou hypotézu

$$H_0 : m_Z = 0$$

proti alternativě

$$H_1 : m_Z \neq 0,$$

kde $m_Z = F_Z^{-1}(1/2)$ značí medián rozdílu $Z = X - Y$.

Testová statistika párového znaménkového testu je definována pomocí následující veličiny

$$Y_n = \sum_{i=1}^{n} \mathbb{I}_{(0,\infty)}(Z_i),$$

což můžeme interpretovat jako počet rozdílů $Z_i = X_i - Y_i$ s kladným znaménkem. Za platnosti hypotézy $H_0 : m_Z = 0$ lze ukázat, že

$$Y_n = \sum_{i=1}^{n} \mathbb{I}_{(0,\infty)}(Z_i) \sim Bi\left(n, \frac{1}{2}\right),$$

kde $Bi(n,1/2)$ značí binomické rozdělení s parametry n a $1/2$ (viz Anděl, 2011, str. 231). Rovněž se dá odvodit:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\mathbb{I}_{(0,\infty)}(Z_i) - \frac{1}{2}\right] \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}\left(0, \frac{1}{4}\right),$$

$$Y_n' := \frac{2}{\sqrt{n}}\left(Y_n - \frac{n}{2}\right) = \frac{2}{\sqrt{n}}\left(\sum_{i=1}^{n} \mathbb{I}_{(0,\infty)}(Z_i) - \frac{n}{2}\right) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0,1),$$

viz Anděl (2011, str. 231.).

Pro přesný test nulovou hypotézu H_0 zamítneme právě, když

$$Y_n \leq k_1 \text{ nebo } Y_n \geq k_2,$$

kde k_1 je největší celé číslo splňující

$$\mathsf{P}[Y_n \leq k_1] = \sum_{j=0}^{k_1} \binom{n}{j} \left(\frac{1}{2}\right)^n \leq \frac{\alpha}{2}$$

a k_2 je nejmenší celé číslo splňující

$$\mathsf{P}[Y_n \geq k_2] = \sum_{j=k_2}^{n} \binom{n}{j} \left(\frac{1}{2}\right)^n \leq \frac{\alpha}{2}.$$

To znamená, že hypotézu budeme zamítat pro malé nebo příliš velké hodnoty testové statistiky Y_n. Ze symetrie hustoty binomického rozdělení plyne, že $k_1 + k_2 = n$. Pro asymptotický test nulovou hypotézu H_0 zamítneme, právě když

$$|Y_n'| \geq u_{1-\alpha/2},$$

kde $u_{1-\alpha/2}$ je $(1-\alpha/2)$-tý kvantil normovaného normálního rozdělení.

P-hodnota přesného párového znaménkového testu je dána

$$\begin{aligned} p = 2 \min\{\mathsf{P}(Bi(n, 1/2) \leq y_n),\, \mathsf{P}(Bi(n, 1/2) \geq y_n)\} = \\ = 2 \min\{1 - G_0(y_n - 1), G_0(y_n)\}\,, \end{aligned}$$

kde G_0 je distribuční funkce $Bi(n,1/2)$ a y_n je napozorovaná hodnota Y_n. P-hodnota asymptotického testu je

$$p = 2(1 - \Phi\left(|y_n'|\right))\,,$$

kde y_n' je napozorovaná hodnota testové statistiky Y_n' a Φ je distribuční funkce normovaného normálního rozdělení.

2.2 Způsoby interpretace testu

V této sekci uvedeme několik způsobů interpretace párového znaménkového testu v závislosti na různých formulacích nulových hypotéz a alternativ.

2.2.1 Test mediánu

Nejdříve budeme uvažovat nulovou hypotézu $H_0 : m_Z = 0$ a alternativu $H_1 : m_Z \neq 0$. Hypotéza H_0 tvrdí, že medián rozdílu je nulový, což znamená, že X_i je s poloviční pravděpodobností větší než Y_i a s poloviční pravděpodobností menší než Y_i, matematicky můžeme zapsat jako

$$\mathsf{P}[X_i > Y_i] = \mathsf{P}[X_i < Y_i] = \frac{1}{2}.$$

Alternativa H_1 naopak tvrdí, že medián rozdílu není nulový, což znamená, že hustota náhodné veličiny Z není symetrická kolem 0.

Nyní zobecníme nulovou hypotézu a alternativu na tvar

$$H_0^* : m_Z = m_0,\ H_1^* : m_Z \neq m_0,$$

kde $m_0 \in \mathbb{R}$. V tom případě testujeme, zda platí

$$\mathsf{P}[X_i > Y_i + m_0] = \mathsf{P}[X_i < Y_i + m_0] = \frac{1}{2}.$$

Tato rovnost vyjadřuje, že pravděpodobnost, že hodnota X_i je větší než hodnota Y_i posunutá o hodnotu m_0, je stejná jako pravděpodobnost, že hodnota X_i je menší než hodnota Y_i posunutá o hodnotu m_0, a obě prevděpodobnosti jsou rovny 1/2.

2.2.2 Test rovnosti dvou mediánů

Interpretace statistických testů, zejména párových testů, může být někdy složitá a vyžaduje pečlivou pozornost k detailům. Jedním z klíčových poznatků při interpretaci párových testů je uvědomění si, že medián rozdílu náhodných veličin X, Y se nemusí obecně rovnat rozdílu mediánů těchto náhodných veličin. Ve skutečnosti, aby platila rovnost mediánů, je potřeba, aby rozdělení rozdílu v rámci jednoho páru mělo symetrickou hustotu.

Definujme náš problém. Předpokládejme, že náhodný vektor $(X, Y)^T$ má spojitou sdruženou distribuční funkci F_{XY}, a zároveň X a Y jsou stejně rozdělené náhodné veličiny, tj. platí

$$F_X(x) = F_Y(x) \text{ pro každé } x \in \mathbb{R}, \tag{2.1}$$

$m_X = m_Y$ a $\mathsf{E}\ X = \mathsf{E}\ Y$. Uvažujme např. X_i jako nějaká hodnota před lékařským zákrokem a Y_i jako nějaká hodnota po zákroku. Pak za nulové hypotézy dává smysl předpokládat, že zákrok neměl žádný efekt na pacienta, tj. platí (2.1).

Zformulujeme novou nulovou hypotézu a alternativu

$$H_0^{**} : m_X = m_Y,\ H_1^{**} : m_X \neq m_Y.$$

Zajímá nás, kdy je nulová hypotéza $H_0 : m_Z = 0$ ekvivalentní s hypotézou $H_0^{**} : m_X = m_Y$.

Nejdříve si zavedeme pojem sdružené symetrie náhodného vektoru $(X, Y)^T$ kolem $(0, 0)^T$.

Definice 1 (Hutson a Yu, 2023, str. 36)**.** *Řekneme, že náhodný vektor $(X, Y)^T$ je sdruženě symetrický*[1] *kolem $(0, 0)^T$, pokud platí*

$$\begin{pmatrix} X \\ Y \end{pmatrix} \overset{\mathcal{D}}{=} \begin{pmatrix} -X \\ -Y \end{pmatrix},$$

kde $\overset{\mathcal{D}}{=}$ značí rovnost v distribuci.

Věta 1. *Nechť $(X, Y)^T$ je dvourozměrný náhodný vektor se sdruženou hustotou f_{XY}. Je-li $(X, Y)^T$ sdruženě symetrický kolem $(0, 0)^T$, potom má náhodná veličina daná rozdílem $Z = X - Y$ symetrickou hustotu f_Z kolem 0.*

[1] Angl. *joint symmetry of the pair*

Důkaz.

Předpokládejme, že máme sdruženě symetrický dvourozměrný náhodný vektor $(X, Y)^T$ kolem $(0, 0)^T$ se sdruženou hustotou f_{XY}.

Distribuční funkce náhodné veličiny Z je $F_Z(z) = \mathsf{P}(Z \leq z)$. Ze symetrie rozdělení $(X, Y)^T$ platí $f_{XY}(x,y) = f_{XY}(-x, -y)$ pro každé x, $y \in \mathbb{R}$.

$$
\begin{aligned}
F_Z(z) = \mathsf{P}(Z \leq z) = \mathsf{P}(X - Y \leq z) &= \iint_{\{(x,y): x-y \leq z\}} f_{XY}(x,y)\, dx\, dy \\
&= \iint_{\{(x,y): x-y \leq z\}} f_{XY}(-x, -y)\, dx\, dy
\end{aligned}
$$

Nyní provedeme substituci $u = -x$, $v = -y$,

$$
\begin{aligned}
\iint_{\{(x,y): x-y \leq z\}} f_{XY}(-x, -y)\, dx\, dy &= \iint_{\{(u,v): u-v \geq -z\}} f_{XY}(u,v)\, du\, dv \\
&= \mathsf{P}(Z \geq -z) \\
&= 1 - F_Z(-z).
\end{aligned}
$$

Celkem máme $F_Z(z) = 1 - F_Z(-z)$ pro každé $z \in \mathbb{R}$.

Víme, že pro spojité náhodné veličiny platí $f(x) = F'(x)$ pro každé $x \in \mathbb{R}$. Zderivujeme obě strany rovnosti $F_Z(z) = 1 - F_Z(-z)$, získáme $f_Z(z) = f_Z(-z)$ pro každé $z \in \mathbb{R}$. Z toho plyne, že náhodná veličina Z má symetrickou hustotu kolem nuly.

□

Uvažujme případ, kdy máme sdruženě symetrický náhodný vektor $(X, Y)^T$ kolem bodu $(0, 0)^T$. V následující větě ukážeme, že za tohoto předpokladu je nulová hypotéza $H_0 : m_Z = 0$ ekvivalentní hypotéze $H_0^{**} : m_X = m_Y$.

Věta 2 (Hutson a Yu, 2023, str. 36)**.** *Nechť F_{XY} je distribuční funkce sdruženě symetrického náhodného vektoru $(X, Y)^T$ kolem bodu $(0, 0)^T$, nechť platí (2.1) a nechť $Z = X - Y$. Potom rozdíl marginálních mediánů $m_X - m_Y = 0$ a medián $m_Z = 0$.*

Důkaz. Je-li náhodný vektor $(X, Y)^T$ sdruženě symetrický kolem bodu $(0, 0)^T$, pak dle věty 1 má náhodná veličina Z symetrickou hustotu kolem 0, tedy dostáváme $m_Z = 0$. Jelikož platí (2.1), potom máme také $m_X - m_Y = 0$.

□

2.2.3 Vliv asymetrie na ekvivalenci hypotéz

Nyní budeme uvažovat, co se stane, pokud budeme předpokládat, že náhodný vektor $(X, Y)^T$ není sdruženě symetrický kolem $(0, 0)^T$ ve smyslu definice 1. Opět budeme předpokládat, že náhodný vektor $(X, Y)^T$ splňuje podmínku (2.1). Zde budeme používat klasické definice střední hodnoty, mediánu a modu.

Definice 2 (Anděl, 2011, str. 17)**.** *Nechť Z je spojitá náhodná veličina s hustotou f_Z. Řekneme, že $\hat{z}$ je modus Z, pokud platí*

$$f(\hat{z}) \geq f(z) \text{ pro každé } z \in \mathbb{R}.$$

Poznámka. Modus je hodnota, ve kterém nabývá hustota náhodné veličiny Z maxima. Modus nemusí být určen jednoznačně (např. hustota rovnoměrného rozdělení na intervalu $(0,1)$) a nemusí vždy existovat (např. hustota beta rozdělení s parametry 1/2, 1/2, která je neomezená).

Definice 3 (Birgé, 1997, definice 1)**.** *Řekneme, že funkce f_Z je unimodální hustota náhodné veličiny Z, pokud existuje $c \in \mathbb{R}$ takové, že $f_Z(z)$ je neklesající pro $z \leq c$ a $f_Z(z)$ je nerostoucí pro $z \geq c$. Navíc, v takovém případě je každé takové c modem Z.*

Poznámka. Je-li hustota f unimodální, potom je modus určen jednoznačně.

Věta 3 (van Zwet, 1979, str. 2)**.** *Nechť náhodná veličina Z je spojitá s hustotou f_Z a distribuční funkcí F_Z. Nechť f_Z je unimodální s modem $\hat{z}$ ve smyslu definice 3. Nechť střední hodnota μ_Z je konečná. Pokud platí*

$$F_Z(m_Z + z) + F_Z(m_Z - z) \geq 1 \text{ pro } z \in \mathbb{R}, \tag{2.2}$$

potom

$$\mu_Z \leq m_Z \leq \hat{z}.$$

Jestliže navíc $m_Z \neq \hat{z}$, potom $\mu_Z < m_Z < \hat{z}$.

Důkaz.

Jako první vypočítáme rozdíl mediánu a střední hodnoty náhodné veličiny Z,

$$m_Z - \mu_Z = m_Z - \mathsf{E}\ Z = \mathsf{E}\,(m_Z - Z) = \int_{-\infty}^{\infty} (m_Z - z) f_Z(z) dz.$$

Tento integrál budeme integrovat po částech.

$$\int_{-\infty}^{\infty} (m_Z - z) f_Z(z) dz = \int_{-\infty}^{m_Z} (m_Z - z) f_Z(z) dz + \int_{m_Z}^{\infty} (m_Z - z) f_Z(z) dz$$

Dle van Zwet (1979, str. 1) z konečnosti střední hodnoty Z platí

$$\lim_{z \to -\infty} z F_Z(z) = 0, \ \lim_{z \to \infty} z(1 - F_Z(z)) = 0.$$

Každý výraz spočítáme zvlášť s využitím metody per partes a nakonec aplikujeme substituce,

$$\begin{aligned}
\int_{-\infty}^{m_Z} (m_Z - z) f_Z(z) dz &= \begin{vmatrix} u = m_Z - z & v' = f_Z(z) \\ u' = -1 & v = F_Z(z) \end{vmatrix} \\
&= [(m_Z - z) F_Z(z)]_{-\infty}^{m_Z} + \int_{-\infty}^{m_Z} F_Z(z) dz \\
&= \begin{vmatrix} z = m_Z - x \\ dz = -dx \end{vmatrix} = - \int_{\infty}^{0} F_Z(m_Z - x) dx \\
&= \int_{0}^{\infty} F_Z(m_Z - x) dx.
\end{aligned}$$

Podobně,

$$\begin{aligned}
\int_{m_Z}^{\infty} (m_Z - z) f_Z(z) dz &= \begin{vmatrix} u = m_Z - z & v' = f_Z(z) \\ u' = -1 & v = -(1 - F_Z(z)) \end{vmatrix} \\
&= -\left[(m_Z - z)(1 - F_Z(z)\right]_{m_Z}^{\infty} - \int_{m_Z}^{\infty} \left(1 - F_Z(z)\right) dz \\
&= \begin{vmatrix} z = m_Z + x \\ dz = dx \end{vmatrix} = - \int_{0}^{\infty} (1 - F_Z(m_Z - x)) dx.
\end{aligned}$$

Dáme dohromady

$$m_Z - \mu_Z = \int_0^\infty F_Z(m_Z - z)dz - \int_0^\infty (1 - F_Z(m_Z - z))dz$$

a dostaneme

$$m_Z - \mu_Z = \int_0^\infty (F_Z(m_Z + z) + F_Z(m_Z - z) - 1)\, dz. \tag{2.3}$$

Ze vztahu (2.3) ukážeme nerovnost $\mu_Z \leq m_Z$. Jelikož platí (2.2), potom celá pravá strana je nezáporná, tudíž levá strana je také nezáporná, tj. platí $m_Z - \mu_Z \geq 0$. Z toho dostáváme $\mu_Z \leq m_Z$. Uvědomme si, je-li hustota f_Z unimodální a symetrická kolem m_Z, potom platí $\mu_Z = m_Z = \hat{z}$ a

$$F_Z(m_Z + z) + F_Z(m_Z - z) = 1. \tag{2.4}$$

Nyní dokážeme vztah mezi mediánem m_Z a modem $\hat{z}$. Pro spor předpokládejme, že $\hat{z} < m_Z$. Protože f_Z je klesající na intervalu $[\hat{z}, \infty)$, tak platí

$$f_Z(m_Z + w) - f_Z(m_Z - w) < 0 \tag{2.5}$$

pro všechny dostatečně malé $w > 0$, aby $m_Z - w \geq \hat{z}$. Zintegrujeme (2.5) podle $w \in (0, z)$, kdy $F_Z(m_Z) = 1/2$.

$$\begin{aligned}
\int_0^z (f_Z(m_Z + w) - f_Z(m_Z - w))\, dw &< 0 \\
F_Z(m_Z + z) - F_Z(m_Z) + F_Z(m_Z - z) - F_Z(m_Z) &< 0 \\
F_Z(m_Z + z) - \frac{1}{2} + F_Z(m_Z - z) - \frac{1}{2} &< 0 \\
F_Z(m_Z + z) + F_Z(m_Z - z) - 1 &< 0
\end{aligned}$$

pro dostatečně malé $z > 0$, což je spor s (2.2). Tedy platí $m_Z \leq \hat{z}$.

Jestliže kromě (2.2) navíc platí $m_Z \neq \hat{z}$, pak je zřejmě, že $m_Z < \hat{z}$ a hustota f_Z není symetrická kolem m_Z. Pak výraz $(F_Z(m_Z + z) + F_Z(m_Z - z) - 1)$ v (2.3) musí být kladný pro nějaké $z > 0$, tedy máme $\mu_Z \neq m_Z$. Z toho dostáváme také $\mu_Z < m_Z$, viz van Zwet (1979, str. 1).

□

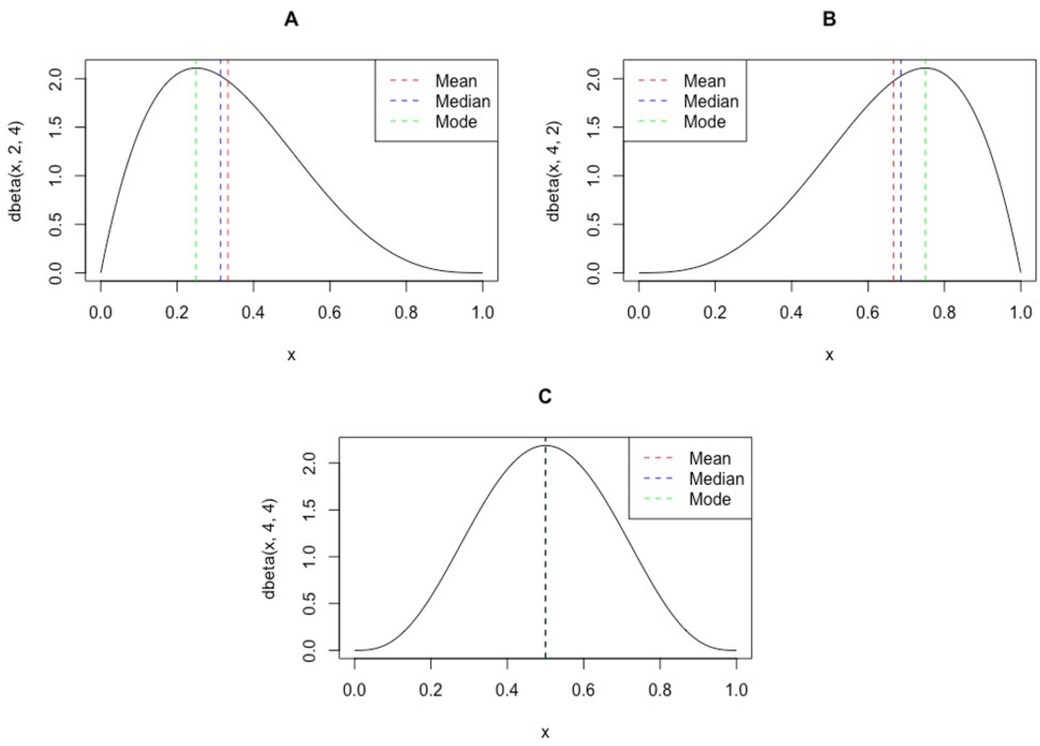
Poznámka. Lze také ukázat, že pokud budeme místo (2.2) předpokládat

$$F_Z(m_Z + z) + F_Z(m_Z - z) \leq 1 \text{ pro } z > 0, \tag{2.6}$$

získáme $\mu_Z \geq m_Z \geq \hat{z}$ (viz Abadir, 2005, str. 477).

Poznámka. Obrázek 2.1 nám poskytuje jednoduchý vizuální pohled na větu 3. Nechť f_Z je unimodální hustota náhodné veličiny Z. Lze ukázat, že pokud $m_Z > \hat{z}$ a platí (2.6), potom $\mu_Z > m_Z$, viz graf A v obrázku 2.1. Podobně, pokud $m < \hat{z}$ platí (2.2), potom $\mu_Z < m_Z$, viz graf B v obrázku 2.1. Pro rozdělení se symetrickou hustotou kolem nějakého bodu platí $\mu_Z = m_Z = \hat{z}$ za předpokladů (2.1) a (2.4), viz graf C v obrázku 2.1.

Následující věta nám dává postačující podmínku, kdy hypotéza $H_0 : m_Z = 0$ není ekvivalentní hypotéze $H_0^{**} : m_X = m_Y$.



Obrázek 2.1: Vztahy mezi střední hodnotou, mediánem a modem

Věta 4 (Hutson a Yu, 2023, str. 37)**.** *Nechť platí* (2.1)*. Nechť náhodná veličina* $Z = X - Y$ *je spojitá s hustotou* f_Z *a distribuční funkcí* F_Z*. Nechť* f_Z *je unimodální s modem* $\hat{z}$ *ve smyslu definice 3, pro který platí* $\hat{z} \neq m_Z$*. Nechť střední hodnota* μ_Z *je konečná. Nechť platí buď* (2.2)*, nebo* (2.6)*. Potom platí* $m_X - m_Y \neq m_Z$*.*

Důkaz. Nejprve si všimneme, že pokud platí (2.1), potom platí $m_X - m_Y = 0$ a $\mu_Z = \mu_X - \mu_Y = 0$. Bez újmy na obecnosti předpokládáme, že platí (2.2). Jelikož $\hat{z} \neq m_Z$, pak dle věty 3 máme $\mu_Z < m_Z$. Protože však $\mu_Z = 0$, tak máme $m_Z > 0$. Tedy celkem máme $m_X - m_Y \neq m_Z$.

V případě platnosti (2.6) bychom postupovali naprosto analogicky.

□

Numerická ilustrace

Nyní ilustrujeme tento poznatek na následujícím příkladě z článku (Hutson a Yu, 2023, str. 37). Nejdříve si připomeneme definici momentové vytvořující funkce.

Definice 4 (Dupač a Hušková, 2005, str. 24)**.** *Nechť* X *je reálná náhodná veličina. Potom momentovou vytvořující funkci náhodné veličiny* X *definujeme jako*

$$M_X(t) = \mathsf{E}\left[e^{tX}\right],$$

pro taková t*, pro která je* $\mathsf{E}\left[e^{tX}\right] < \infty$*.*

Nechť U a V jsou nezávislé náhodné veličiny taková, že

$$U \sim Exp(1) \text{ a } V \sim Exp(1),$$

kde $Exp(1)$ značí exponenciální rozdělení s parametrem 1. Pak náhodná veličina U má momentovou vytvořující funkci

$$M_U(t) = \mathsf{E}\left[e^{tU}\right] = \int_0^\infty e^{tu} e^{-u} du = \int_0^\infty e^{(t-1)u} du = \left[\frac{e^{(t-1)u}}{t-1}\right]_0^\infty = \frac{1}{1-t}$$

pro $t < 1$. Podobně, náhodná veličina V má momenotovou vytvořující funkci

$$M_V(t) = \mathsf{E}\left[e^{tV}\right] = \frac{1}{1-t} \text{ pro } t < 1.$$

Definujme novou náhodnou veličinu

$$W = U\phi + V\sqrt{1-\phi^2},$$

kde $\phi \in [0,1]$. Vypočítáme její momentovou vytvořující funkci

$$M_W(t) = \mathsf{E}\left[e^{tW}\right] = \mathsf{E}\left[e^{t\left(U\phi + V\sqrt{1-\phi^2}\right)}\right].$$

Protože náhodné veličiny U a V jsou nezávislé, platí:

$$M_W(t) = \mathsf{E}\left[e^{tU\phi}\right] \cdot \mathsf{E}\left[e^{tV\sqrt{1-\phi^2}}\right] = M_U(t\phi) \cdot M_V\left(t\sqrt{1-\phi^2}\right)$$

pro $t < \min\{1/\phi, 1/\sqrt{1-\phi^2}\}$. To dle Yanev (2020, str. 1) odpovídá momentové vytvořující funkci náhodné veličiny W, která má hypoexponenciální rozdělení[2] s parametry $\lambda_1 = 1/\phi$ a $\lambda_2 = 1/\sqrt{1-\phi^2}$ a s distribuční funkcí

$$F_W(w) = 1 - \frac{\lambda_2}{\lambda_2 - \lambda_1} e^{-\lambda_1 w} + \frac{\lambda_1}{\lambda_2 - \lambda_1} e^{-\lambda_2 w}$$

pro $w > 0$.

Věta 5 (Metoda inverzní transformace)**.** *Nechť X má rovnoměrné rozdělení na intervalu $[0,1]$ a nechť F je libovolná spojitá distribuční funkce. Potom má náhodná veličina $Y = F^{-1}(X)$ rozdělení s distribuční funkcí F.*

Důkaz. Viz Devroye (1986, věta 2.1.).

□

Poznámka. Má-li nějaká náhodná veličina distribuční funkci

$$F(y) = 1 - e^{-\lambda y} \text{ pro } y \geq 0, \tag{2.7}$$

potom inverzní funkce (kvantilová funkce) k funkci F je

$$F^{-1}(x) = -\frac{1}{\lambda}\log(1-x) \text{ pro } 0 < x < 1.$$

Pokud náhodná veličina $X \sim R(0,1)$, potom má náhodná veličina

$$Y = -\frac{1}{\lambda}\log(1-X)$$

exponenciální rozdělení s distribuční funkcí (2.7).

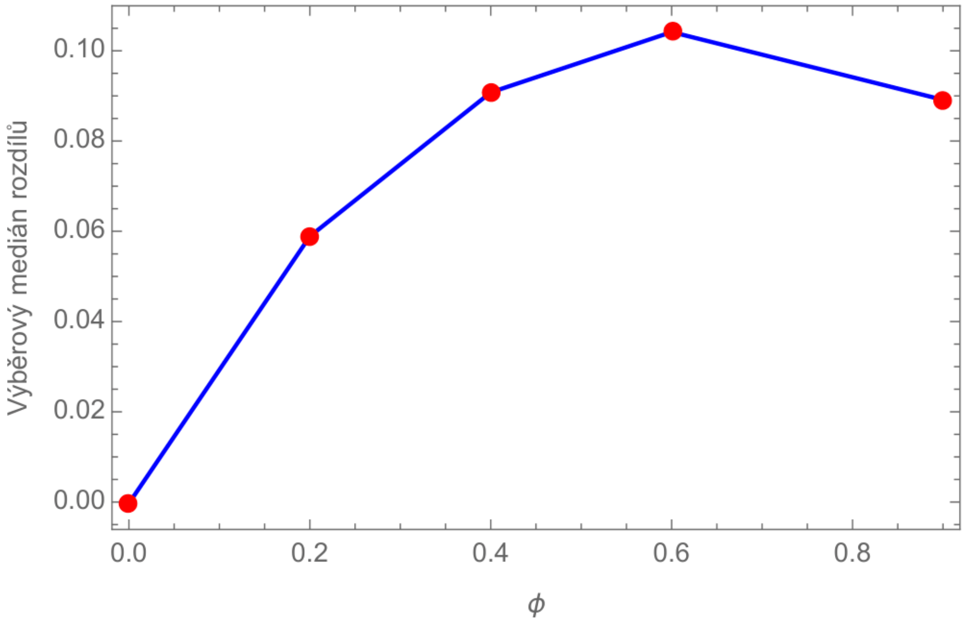
[2] Angl. *hypoexponential distribution* (Yanev, 2020)

Pomocí metody inverzní transformace definujeme náhodné veličiny

$$X = U \text{ a } Y = F_V^{-1}(F_W(W)),$$

kde $U \sim Exp(1)$, $V \sim Exp(1)$ a W má hypoexponenciální rozdělení s parametry $\lambda_1 = 1/\phi$ a $\lambda_2 = 1/\sqrt{1-\phi^2}$. Dle Anděl (2011, věta 1.4) platí, že $F_W(W)$ má rovnoměrné rozdělení na intervalu $(0,1)$. Pak z věty 5 nám plyne, že náhodná veličiny $Y \sim Exp(1)$. Celkem máme, že X a Y mají exponenciální rozdělení s parametrem 1, ale nejsou nezávislé pro $\phi > 0$.

Pomocí softwaru Wolfram Mathematica vytvoříme náhodný výběr o rozsahu 10 000 párových pozorování dvou nezávislých stejně rozdělených náhodných veličin U a V s exponenciálním rozdělením s parametrem 1. Vypočítáme výběrový medián rozdílů $Z_i = X_i - Y_i$ v závislosti na $\phi = 0, 0.2, 0.4, 0.6, 0.9$.



Obrázek 2.2: Výběrový medián rozdílů Z_i v závislosti na ϕ

Z grafu 2.2 vidíme, že medián Z se obecně odchyluje od nuly s výjimkou $\phi = 0$, kdy jsou náhodné veličiny X a Y nezávislé. Tento poznatek můžeme vypozorovat z grafu, že výběrový medián při $\phi = 0$ se nachází u 0.

Definice 5 (Rohatgi a Ehsanes Saleh, 2001, sekce 4.3, definice 7)**.** *Řekneme, že náhodné veličiny X a Y jsou zaměnitelné* [3]*, pokud platí $(X,Y)^T \overset{\mathcal{D}}{=} (Y,X)^T$, kde $\overset{\mathcal{D}}{=}$ značí rovnost v distribuci.*

Poznámka. Ve větě 2 by se dala místo předpokladu sdružené symetrie náhodného vektoru $(X,Y)^T$ kolem $(0,0)^T$ ve smyslu definice 1 uvažovat také zaměnitelnost náhodných veličin X a Y ve smyslu definice 5, jelikož dle Rohatgi a Ehsanes Saleh (2001, sekce 4.3, věta 5) zaměnitelnost náhodných veličin X a Y implikuje symetrii rozdělení náhodné veličiny definované jejich rozdílem $Z = X - Y$.

Pro $\phi = 0$ jsou náhodné veličiny X a Y nezávislé, a zároveň stejně rozdělené s exponenciálním rozdělením s parametrem 1. Pak v takovém případě, náhodné veličiny X a Y jsou zaměnitelné ve smyslu definice 5, Tedy dle poznámky výše dostáváme tvrzení věty 2. Celkem pro $\phi = 0$ platí, že $m_X - m_Y = m_Z$.

[3] Angl. exchangeable

3. Párový Wilcoxonův test

Dalším typem párových testů je párový Wilcoxonův test. Pro práci s párovými daty nemusíme znát konkrétní typ rozdělení, ze kterého pocházejí analyzovaná data.

U párového znaménkového testu jsme uvažovali pouze to, zda jsou jednotlivé rozdíly pozorování $Z_i = X_i - Y_i$ větší nebo menší než 0 (nebo v obecném případě m_0), nebrali jsme v úvahu velikosti těchto rozdílů od 0. Tedy párový znaménkový test může být málo eficientní a může docházet ke ztrátě síly testu, jelikož jeho testová statistika bere v úvahu pouze znaménka, a nikoli velikosti rozdílů mezi páry pozorovaní. Na druhou stranu u párového Wilcoxonova testu budeme hodnotit pořadí náhodných veličin v náhodném výběru, které jsou seřazeny od nejmenší po největší. Tento testu oproti párového znaménkového testu vyžaduje silnější předpoklad, který si zformulujeme v následující sekci. Budeme vycházet z knihy Anděl (2011).

3.1 Formulace úlohy

V této sekci budeme uvažovat model

$$\mathcal{F} = \{Z_i = X_i - Y_i \text{ má spojité rozdělení s hustotou } f_Z \text{ splňující}$$
$$\exists\, \delta \in \mathbb{R} : f_Z(\delta - z) = f_Z(\delta + z) \ \ \forall\, z \in \mathbb{R}\}.$$

Tedy předpokládáme, že hustota f_Z náhodné veličiny $Z_i = X_i - Y_i$ byla symetrická kolem bodu δ. Budeme brát na vědomí, že předpoklad o symetrii hustoty f_Z se týká rozdílů Z_i, nikoli původních pozorování X_i a Y_i.

Testujeme nulovou hypotézu

$$H_0 : \delta_Z = \delta_0$$

proti alternativě

$$H_1 : \delta_Z \neq \delta_0,$$

kde $\delta_0 \in \mathbb{R}$ je předem daná konstanta.

Testová statistika párového Wilcoxonova testu je

$$W_S = \sum_{i \in \mathcal{I}} R_i,$$

kde $\mathcal{I} \subset \{1, \ldots, n\}$ je množina všech indexů takových, že $Z_i^* \overset{\text{df}}{=} X_i - Y_i - \delta_0$ má kladné znaménko pro $i \in \mathcal{I}$, a R_i je pořadí náhodné veličiny $|Z_i^*|$ mezi všemi $|Z_1^*|, \ldots, |Z_n^*|$. Testová statistika W_n může nabývat hodnot z množiny $\{0, 1, \ldots,\ n(n+1)/2\}$.

Poznámka. Hodnotu testové statistiky vypočítáme následujícím způsobem:

1. Spočítáme odchylky $Z_i = X_i - Y_i - \delta_0$ a určíme množinu indexů $\mathcal{I}$.

2. Spočteme $|Z_1^*|, \ldots, |Z_n^*|$.

3. Seřadíme všechny $|Z_i^*|$ od nejmenší do největší a získáme uspořádaný náhodný výběr
$$0 < |Z^*|_{(1)} < |Z^*|_{(2)} < \cdots < |Z^*|_{(n)} .$$
4. Určíme pořadí R_i náhodné veličiny $|Z_i^*|$ mezi všemi
$$|Z^*|_{(1)} , |Z^*|_{(2)} , \ldots , |Z^*|_{(n)} .$$
Platí $|Z_i^*| = |Z^*|_{(R_i)}$.
5. Sečteme pořadí R_i pro $i \in \mathcal{I}$.

Za platnosti hypotézy $H_0 : \delta_Z = \delta_0$ v modelu $\mathcal{F}$ se dá ukázat následující vlastnosti testové statistiky W_S:

- střední hodnota:
$$\mathsf{E}\, W_S = \frac{1}{4}n(n+1),$$
- rozptyl:
$$\operatorname{var} W_S = \frac{1}{24}n(n+1)(2n+1),$$
- asymptotická normalita:
$$U_n := \frac{W_S - \mathsf{E}\, W_S}{\sqrt{\operatorname{var} W_S}} \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0,1),$$

viz Anděl (2011), str. 234). Odtud můžeme určit kritické hodnoty asymptotického testu. Nulovou hypotézu H_0 zamítneme, právě když
$$|U_n| = \frac{\left|W_S - \frac{n(n+1)}{4}\right|}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \geq u_{1-\alpha/2}.$$

3.2 Symetrie hustoty Z kolem nuly

V této sekci se budeme zabývat případem, kdy je hustota náhodné veličiny dané rozdílem $Z = X - Y$ symetrická kolem 0, což je předpokladem párového Wilcoxonova testu.

Jednou z postačujících podmínek pro symetrickou hustotu náhodné veličin Z kolem 0 je například sdružená symetrie náhodného vektoru $(X,Y)^T$ kolem $(0,0)^T$ ve smyslu definice 1, viz věta 2.

Jiná postačující podmínka je zaměnitelnost náhodných veličin X, Y ve smyslu definice 5, viz poznámka za touto definicí.

3.3 Způsoby interpretace testu

Často se v praxi setkáváme s tím, že výsledky testu jsou špatně interpretovány. Jedním z hlavních důvodů je způsob formulace hypotéz a alternativ. V této sekci uvedeme různé způsoby interpretace párového Wilcoxonova testu.

3.3.1 Test o středu symetrie hustoty rozdílu Z

Jak jsme již zmínili, že testujeme hypotézu $H_0 : \delta_Z = \delta_0$ proti alternativě $H_1 : \delta_Z \neq \delta_0$. Jinými slovy testujeme, zda střed symetrie je roven předepsané hodnotě δ_0. Tato hypotéza říká, že hustota náhodné veličiny dané rozdílem Z_i mezi dvěma měřeními má symetrickou hustotu kolem nějaké specifické hodnoty δ_0. Alternativní hypotéza tvrdí, že hustota náhodné veličiny dané rozdílem Z_i mezi dvěma měřeními nemá symetrickou hustotu kolem hodnoty δ_0.

Vzpomeňme si na důkaz věty 3. Uvažovali jsme hustotu náhodné veličiny Z, která byla unimodální s modem $\hat{z}$ ve smyslu definice 3 a symetrická kolem m_Z. Pak platí $\mu_Z = m_Z = \hat{z}$ a (2.4). Můžeme zformulovat následující hypotézu a alternativu

$$\begin{aligned} H_0^* &: \forall z \in \mathbb{R} : F_Z(\delta_0 + z) + F_Z(\delta_0 - z) = 1, \\ H_1^* &: \exists z \in \mathbb{R} : F_Z(\delta_0 + z) + F_Z(\delta_0 - z) \neq 1, \end{aligned}$$

což je v rámci modelu $\mathcal{F}$ ekvivalentní hypotéze a alternativě $H_0 : \delta_Z = \delta_0$, $H_0 : \delta_Z \neq \delta_0$.

Pokud $\delta_0 = 0$, potom test o středu symetrie hustoty kolem nuly můžeme také testovat pomocí následující hypotézy a alternativy:

$$\begin{aligned} H_0^{**} &: \forall\, x,y \in \mathbb{R} : F_{XY}(x, y) = F_{XY}(y, x) \\ H_1^{**} &: \exists\, x,y \in \mathbb{R} : F_{XY}(x, y) \neq F_{XY}(y, x). \end{aligned}$$

Hypotéza $H_0^{**} : \forall\, x,y \in \mathbb{R} : F_{XY}(x,y) = F_{XY}(y,x)$ bude implikovat hypotézu $H_0^* : \forall z \in \mathbb{R} : F_Z(z) + F_Z(-z) = 1$, nicméně obrácená implikace obecně neplatí.

3.3.2 Test mediánu rozdílu Z

Zformulujeme novou hypotézu a alternativu

$$H_0^{***} : m_Z = \delta_0,\ H_1^{***} : m_Z \neq \delta_0,$$

kde $\delta_0 \in \mathbb{R}$ je předem daná konstanta. Všimneme si, že střed symetrie δ v modelu $\mathcal{F}$ je roven mediánu m_Z. Tedy párový Wilcoxonův test můžeme zároveň interpretovat jako test mediánu rozdílu Z, tj. za platnosti modelu $\mathcal{F}$ je hypotéza $H_0 : \delta_Z = \delta_0$ ekvivalentní hypotéze $H_0^{***} :: m_Z = \delta_0$.

3.3.3 Test rovnosti mediánů X a Y

Zde budeme uvažovat následující nulovou hypotézu a alternativu

$$H_0^{****} : m_X - m_Y = 0,\ H_1^{****} : m_X - m_Y \neq 0.$$

Stejně jako u párového znaménkového testu zde nás opět zajímá, za jakých podmínek může být hypotéza $H_0^{***} : m_Z = 0$ ekvivalentní hypotéze $H_0^{****} : m_X - m_Y = 0$. Jinými slovy zkoumáme, kdy platí rovnost $m_Z = m_X - m_Y$. Zde můžeme využít stejné úvahy, které jsme dělali u párového znaménkového testu v sekci 2.2.2.

V předchozí sekci 3.3.2 jsme zjistili, že v modelu $\mathcal{F}$ je hypotéza $H_0 : \delta_Z = 0$ ekvivalentní hypotéze $H_0^{***} : m_Z = 0$. Pak rovností mediánů dvou náhodných veličin dosáhneme předpokladem 2.1. Tedy celkem platí, že v modelu $\mathcal{F}$ se medián rozdílu $Z = X - Y$ rovná mediánů rozdílu těchto náhodných veličin pouze v případě, pokud jsou náhodné veličiny X a Y stejně rozdělené.

3.3.4 Test rovnosti středních hodnot X a Y

Obecně se v modelu $\mathcal{F}$ konečnost střední hodnoty nepředpokládá. V případě, že bychom vyžadovali, aby $\mathsf{E}\ X_i$ a $\mathsf{E}\ Y_i$ byly konečné ($X, Y \in \mathcal{L}^1$), předpoklad dává $\delta_Z = \mathsf{E}\ Z_i = \mathsf{E}\ X_i - \mathsf{E}\ Y_i$. Tedy můžeme alternativně testovat hypotézu o rozdílu středních hodnot, tj. testovali bychom nulovou hypotézu

$$H_0^{*****} : \mu_X - \mu_Y = 0$$

proti alternativě

$$H_1^{*****} : \mu_X - \mu_Y \neq 0.$$

Párový t-test také testuje rozdíl středních hodnot nepotřebujeme k tomu symetrie Z_i. Na druhou stranu je však párový t-test citlivý na odlehlá pozorování.

Závěr

V této práci jsme se zabývali třemi párovými testy, tj. párovým t-testem, párovým znaménkovým testem a na závěr párovým Wilcoxonovým testem. V úvodu jsme si definovali párový problém a připomněli jsme si některé základní matematické pojmy.

V kapitole 1 jsme se zabývali párovým t-testem. Zformulovali jsme předpoklady testu. Párový t-test můžeme interpretovat jako test rovnosti středních hodnot dvou závislých pozorování. Dále jsme porovnali testovou statistiku párového t-testu s testovou statistikou dvouvýběrového t-testu. Zjistili jsme, že můžeme párový t-test mylně zaměnit s dvouvýběrovým t-testem pouze v případě, že výběrová kovariance dvou náhodných výběrů je nulová.

V kapitole 2 jsme pro párový znaménkový test zformulovali základní předpoklady. Hledali jsme různé způsoby, jak ho interpretovat. Jako první a zároveň nejjednodušší je test mediánu rozdílu dvou náhodných veličin. S dalšími předpoklady, jako je sdružená symetrie náhodného vektoru kolem nuly ve smyslu definice 1, jsme ho mohli interpretovat jako test rovnosti dvou marginálních mediánů. Podrobně jsme dokázali větu 3, pomocí které jsme zjistili, že bez dodatečného předpokladu sdružené symetrie náhodného vektoru kolem nuly bychom takovou interpretaci nemohli provést. Toto jsme také numericky ilustrovali na konkrétním příkladu.

V kapitole 3 jsme se zabývali párovým Wilcoxonovým testem. Opět jsme zformulovali předpoklady a hledali různé možnosti interpretace testu, přičemž jsme používali poznatky z kapitoly 2.

Seznam použité literatury

ABADIR, K. M. (2005). The mean-median-mode inequality: Counterexamples. *Econometric Theory*, **21**(2), 477–482.

ANDĚL, J. (2011). *Základy matematické statistiky.* Třetí vydání. Matfyzpress, Praha. ISBN 978-80-7378-162-0.

ANDĚL, J. (2019). *Statistické metody.* Páté vydání. Matfyzpress, Praha. ISBN 978-80-7378-381-5.

BIRGÉ, L. (1997). Estimation of unimodal densities without smoothness assumptions. *The Annals of Statistics*, **22**(3), 970–981.

DEVROYE, L. (1986). *Non-Uniform Random Variate Generation.* Springer-Verlag, USA. ISBN 0-387-96305-7.

DUPAČ, V. a HUŠKOVÁ, M. (2005). *Pravděpodobnost a matematická statistika.* První vydání. Karolinum, Praha. ISBN 80-246-0009-9.

HUTSON, A. D. a YU, H. (2023). The sign test, paired data, and asymmetric dependence: a cautionary tale. *Amer. Statist.*, **77**(1), 35–40.

ROHATGI, V. K. a EHSANES SALEH, A. K. M. (2001). *An Introduction to Probability and Statistics.* Druhé vydání. John Wiley and Sons, Kanada. ISBN 978-0-471-34846-7.

VAN ZWET, W. R. (1979). Mean, median, mode. II. *Statist. Neerlandica*, **33**(1), 1–5.

YANEV, G. P. (2020). Exponential and Hypoexponential Distributions: Some Characterizations. *Mathematics*, **8**(12), 1–10.