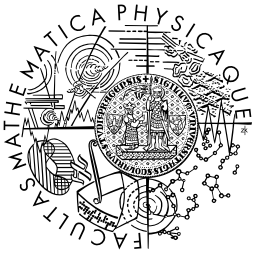

MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Kateřina Krejčová

Testy ekvivalence

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. RNDr. Zdeněk Hlávka, Ph.D.
Studijní program: Obecná matematika
Studijní obor: MOMP

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

Ráda bych poděkovala vedoucímu bakalářské práce doc. RNDr. Zdeňku Hlávkovi, Ph.D. za jeho cenné rady a věnovaný čas. Dále děkuji své rodině a přátelům za jejich trpělivost a podporu během studia.

Název práce: Testy ekvivalence

Autor: Kateřina Krejčová

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. RNDr. Zdeněk Hlávka, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: V této bakalářské práci se zabýváme testy ekvivalence využívanými v případech, kdy potřebujeme dokázat platnost tvrzení, které standardně dáváme do nulové hypotézy. Uvedeme postup řešení těchto testů nazvaný princip inkluze intervalu spolehlivosti. Poté se budeme věnovat dvěma vybraným testům ekvivalence pro párová data a ukážeme využití těchto testů na porovnání kvality nově vyvíjených algoritmů založených na umělé inteligenci se současně používanými metodami. Prvním testem ekvivalence je modifikace párového t-testu, druhým pak modifikovaný asymptotický McNemarův test. Součástí práce je u obou testů ekvivalence odvození plánování rozsahu výběru. Jejich použití poté demonstrujeme na reálných případech testování kvality algoritmů umělé inteligence používaných ve zdravotnictví.

Klíčová slova: hypotéza, testová statistika, ekvivalence

Title: Equivalence testing

Author: Kateřina Krejčová

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. RNDr. Zdeněk Hlávka, Ph.D., Department of Probability and Mathematical Statistics

Abstract: In this thesis, we focus on equivalence tests used in situation where we need to prove the validity of statements usually framed as null hypothesis. We present a procedure used for solving these tests called the confidence interval inclusion principle. We then discuss two selected equivalence tests for paired data. Using these tests, we demonstrate the comparison of quality of newly developed artificial intelligence-based algorithms with currently used methods. The first equivalence test is a modification of the paired t-test, and the second is a modified asymptotic McNemar test. For both tests, we present sample size calculation. Afterwards we show usage of these tests in quality testing of algorithms based on artificial intelligence which are applied in healthcare.

Keywords: hypothesis, test statistic, equivalence

Obsah

Úvod

V dnešní době jsou ve spoustě různých odvětví stále rozšířenější algoritmy založené na umělé inteligenci, jejichž cílem je nahradit současné metody nebo uspořit pracovní síly. Jeden z možných příkladů jejich použití je diagnóza rakoviny kůže u pacientů, kde uměle inteligentní algoritmus vyhodnotí, zda se jedná o zhoubný melanom či nikoliv. Než však tyto algoritmy budeme moci používat, je potřeba vyhodnotit, zda fungují stejně dobře nebo případně i lépe než aktuálně používané metody.

V klasických statistických oboustranných testech nulová hypotéza říká, že dvě metody jsou shodné, a alternativní hypotéza tvrdí, že jsou rozdílné. Těmito testy můžeme dokázat, že dvě metody, v našem případě algoritmus vytvořený umělou inteligencí a současná metoda, jsou dostatečně rozdílné. Neprokázání rozdílnosti nám však nedokazuje, že jsou shodné. Tento fakt nás přivádí k problému testování ekvivalence, kterým se v této práci zabýváme.

V první kapitole si připomeneme základní pojmy k testování hypotéz a zavedeme značení, které budeme dále v práci používat. Uvedeme také jednoduchý příklad na jednovýběrový t-test, na němž si dále znázorníme provedené modifikace u testování ekvivalence.

Ve druhé kapitole se podíváme na problém testování ekvivalence. Nejprve si ukážeme jeden z možných postupů řešení nazvaný princip inkluze intervalu spolehlivosti. Aplikaci tohoto postupu si znázorníme na zmíněném příkladu jednovýběrového t-testu. Dále se zaměříme na odvození dvou vybraných testů ekvivalence, jimiž budou modifikovaný párový t-test a modifikovaný McNemarův test pro ekvivalenci. V poslední části této kapitoly se budeme věnovat testům noninferiority jako „jednostranné verzi“ testů ekvivalence.

Ve třetí kapitole se budeme zabývat plánováním rozsahu výběru a odvodíme odhad u námi vybraného modifikovaného párového t-testu a u modifikovaného McNemarova testu.

Nakonec se ve čtvrté kapitole dostaneme k praktickému využití testů ekvivalence, konkrétně se podíváme na jejich možné použití u výše uvedeného problému testování kvality algoritmů založených na umělé inteligenci. Prvním příkladem bude porovnání radiologů a algoritmu vytvořeném umělou inteligencí při měření průměru aorty u CT angiografie před implementací aortální chlopně (TAVI). Zde použijeme modifikovaný párový t-test a ukážeme výpočet odhadu potřebného rozsahu výběru. Ve druhém příkladu se budeme věnovat porovnání úspěšnosti lékařů a uměle inteligentních algoritmů při diagnóze rakoviny kůže, kde použijeme modifikovanou verzi McNemarova testu.

1. Testování hypotéz

V této kapitole si připomeneme základní pojmy a zavedeme značení pro testování hypotéz. Následně uvedeme příklad testu hypotézy pro střední hodnotu. Vycházíme především z knihy Anděl (2007) a skript Kulich a Omelka (2022).

Mějme náhodný vektor $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ s rozdělením, které závisí na parametru $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^\top \in \Theta$. Nechť Θ_0 a $\Theta_1 = \Theta \setminus \Theta_0$ jsou neprázdné podmnožiny množiny Θ. Tvrzení $H_0 : \boldsymbol{\theta} \in \Theta_0$ nazýváme *nulová hypotéza* a naopak tvrzení $H_1 : \boldsymbol{\theta} \in \Theta_1$ nazýváme *alternativní hypotéza*. Říkáme, že *testujeme* hypotézu

$$H_0 : \boldsymbol{\theta} \in \Theta_0 \quad \text{proti alternativě} \quad H_1 : \boldsymbol{\theta} \in \Theta_1.$$

Mějme nyní jednorozměrný parametr θ. Jestliže $\Theta_0 = \{\theta_0\}$ je jednobodová množina, potom hypotézu nazýváme *jednoduchou* a test hypotézy

$$H_0 : \theta = \theta_0 \quad \text{proti} \quad H_1 : \theta \neq \theta_0$$

nazýváme *oboustranný test*. Pokud zvolíme $\Theta_0 = (-\infty, \theta_0\rangle$, resp. $\Theta_0 = \langle\theta_0, \infty)$, tedy testujeme hypotézu

$$H_0 : \theta \leq \theta_0 \quad \text{proti} \quad H_1 : \theta > \theta_0,$$

resp.

$$H_0 : \theta \geq \theta_0 \quad \text{proti} \quad H_1 : \theta < \theta_0,$$

potom výše uvedené testy nazýváme *jednostranné testy*.

Při rozhodování o platnosti hypotézy H_0 postupujeme následujícím způsobem. Zvolíme funkci $S_n(\boldsymbol{X})$, jejíž rozdělení za platnosti H_0 známe, a množinu $\mathcal{C}$. Funkci S_n nazýváme *testová statistika*, množinu $\mathcal{C}$ *kritický obor*. Pokud $S_n(\boldsymbol{X}) \in \mathcal{C}$, zamítáme hypotézu H_0 ve prospěch alternativy H_1, naopak pokud $S_n(\boldsymbol{X}) \notin \mathcal{C}$, potom hypotézu H_0 nezamítáme ve prospěch alternativy H_1. Ve druhém z uvedených případů se však nejedná o potvrzení hypotézy.

Během rozhodování o platnosti hypotézy H_0 se můžeme dopustit dvou chyb, které jsou uvedené v následující definici.

Definice 1. *Jestliže test zamítl platnou hypotézu, říkáme, že nastala* chyba I. druhu. *Jestliže test nezamítl neplatnou hypotézu, říkáme, že nastala* chyba II. druhu.

Celkově mohou nastat čtyři situace:

	Nezamítáme H_0	Zamítáme H_0
H_0 platí	Správné rozhodnutí	*chyba I. druhu*
H_0 neplatí	*chyba II. druhu*	Správné rozhodnutí

Chybám I. a II. druhu se zpravidla nemůžeme vyhnout. Chyba I. druhu je závažnější, a proto se pravděpodobnost jejího výskytu snažíme kontrolovat volbou kritického oboru. Zvolíme číslo $\alpha \in (0, 1)$, obvykle volíme $\alpha = 0{,}05$. Zkonstruujeme kritický obor $\mathcal{C}$ tak, aby platilo

$$\mathsf{P}_{\boldsymbol{\theta}}(S_n(\boldsymbol{X}) \in \mathcal{C}) \leq \alpha, \quad \forall \boldsymbol{\theta} \in \Theta_0.$$

Číslo $\alpha_0 = \sup_{\boldsymbol{\theta}\in\Theta_0} \mathsf{P}_{\boldsymbol{\theta}}(S_n(\boldsymbol{X}) \in \mathcal{C})$ nazýváme *hladina testu* a funkci $\beta(\boldsymbol{\theta}) = \mathsf{P}_{\boldsymbol{\theta}}(S_n(\boldsymbol{X}) \in \mathcal{C})$, $\boldsymbol{\theta} \in \Theta$, nazýváme *silofunkce testu*. Číslo $\beta(\boldsymbol{\theta})$, kde $\boldsymbol{\theta} \in \Theta_1$, nazýváme *síla* testu proti alternativě $\boldsymbol{\theta}$.

Formálně definujeme test jako funkci $\psi : \mathbb{R}^n \to \{0, 1\}$, kde máme rozhodovací pravidlo takové, že zamítáme H_0 právě tehdy, když $\psi(\mathbf{X}) = 1$.

Definice 2. Randomizovaný test *definujeme jako funkci* $\phi : \mathbb{R}^n \to [0, 1]$. *Silofunkci* ϕ *definujeme jako* $\beta(\theta) = \mathsf{E}_\theta\, \phi(\boldsymbol{X})$ *a hladinu jako* $\sup_{\theta\in\Theta_0} \mathsf{E}_\theta\, \phi(\boldsymbol{X})$.

Definice 3. Stejnoměrně nejsilnější test ϕ *pro*

$$H_0 : \theta \in \Theta_0 \quad \textit{proti} \quad H_1 : \theta \in \Theta_1$$

na hladině α *je test, který splňuje* $\sup_{\theta\in\Theta_0} \beta(\theta) \le \alpha$ *a zároveň pro každý jiný test* ϕ^* *splňující* $\sup_{\theta\in\Theta_0} \beta^*(\theta) \le \alpha$, *kde* β^* *je silofunkce* ϕ^*, *platí* $\beta(\theta) \ge \beta^*(\theta)$ *pro všechny* $\theta \in \Theta_1$.

Značení. Symbolem u_α budeme značit α-kvantil normovaného normálního rozdělení a symbolem $t_n(\alpha)$ α-kvantil Studentova rozdělení s n stupni volnosti. Distribuční funkci normovaného normálního rozdělení budeme značit $\Phi(\cdot)$.

Příklad 1 (Zadání příkladu převzato z knihy Anděl, 1998, str. 77-78)**.** Automat plní krabice pracím práškem. V každé krabici mají být 2 kg prášku. Z produkce bylo náhodně odebráno 6 krabic a jejich obsah přesně zvážen. Byly zjištěny tyto odchylky od požadované hmotnosti (v dkg):

$$\text{-5, 1, -1, -8, 7, -6.}$$

Je třeba ověřit, zda nedošlo k systematické odchylce nastavení automatu.

Označme $X_1, \ldots, X_6$ jednotlivé odchylky nastavení automatu, o kterých předpokládáme, že tvoří náhodný výběr z rozdělení $\mathrm{N}(\mu, \sigma^2)$. Budeme testovat hypotézu, že nedošlo k systematické odchylce nastavení automatu, proti alternativě, že k odchylce došlo, pomocí jednovýběrového t-testu. Neboli testujeme hypotézu

$$H_0 : \mu = 0 \quad \text{proti} \quad H_1 : \mu \neq 0.$$

Použijeme testovou statistiku

$$T_n = \sqrt{n}\,\frac{\overline{X}_n}{S_n},$$

kde

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

Víme, že testová statistika T_n má za platnosti H_0 Studentovo rozdělení s $n - 1$ stupni volnosti (viz Anděl, 2007, Věta 4.23). Tedy na hladině α zamítáme H_0 právě tehdy, když $|T_n| \ge t_{n-1}(1 - \alpha/2)$.

Vidíme, že pro $\alpha = 0{,}05$ platí $|T_6| \doteq 0{,}889 < 2{,}571 \doteq t_5(0{,}975)$, tedy H_0 na hladině α nezamítáme.

Interval spolehlivosti pro μ s koeficientem spolehlivosti $(1 - \alpha)$ je

$$\left(\overline{X}_n - t_{n-1}(1-\alpha/2)\,\frac{S_n}{\sqrt{n}},\ \overline{X}_n + t_{n-1}(1-\alpha/2)\,\frac{S_n}{\sqrt{n}}\right) \doteq (-7{,}787,\ 3{,}787).$$

H_0 nezamítáme a nedokážeme tedy říct, zda dochází k systematické odchylce nastavení automatu.

2. Testy ekvivalence

V této kapitole vycházíme z knihy Wellek (2010). U klasického testování hypotéz jsme formulovali oboustranný test pro jednobodovou množinu $\Theta_0 = \{\theta_0\}$ jako

$$H_0 : \theta = \theta_0 \quad \text{proti} \quad H_1 : \theta \neq \theta_0.$$

Nastane jedna ze dvou možností. Hypotézu buď zamítneme ve prospěch alternativy, nebo hypotézu nebudeme moci zamítnout ve prospěch alternativy. Ve druhém ze zmíněných případů se však nejedná o potvrzení platnosti hypotézy.

Uvažujeme nyní situaci, kdy bychom chtěli prokázat rovnost. Pokud bychom test hypotézy formulovali jako

$$H_0 : \theta \neq \theta_0 \quad \text{proti} \quad H_1 : \theta = \theta_0,$$

potom by měl hladinu α pouze v případě, že by síla testu také nepřesáhla hodnotu α (viz Wellek, 2010, str. 11). To znamená, že by pravděpodobnost zamítnutí neplatné hypotézy také nepřesáhla hodnotu α.

Z tohoto důvodu testujeme hypotézu

$$H_{0E} : \theta \leq \theta_0 - \epsilon_1 \text{ nebo } \theta \geq \theta_0 + \epsilon_2 \quad \text{proti} \quad H_{1E} : \theta_0 - \epsilon_1 < \theta < \theta_0 + \epsilon_2, \quad (2.1)$$

kde $\epsilon_1 > 0,\ \epsilon_2 > 0$ jsou stanovené konstanty. Hodnoty ϵ_1, ϵ_2 jsou zvolené před analýzou dat a určují interval ekvivalence tvaru $(\theta_0 - \epsilon_1,\ \theta_0 + \epsilon_2)$ okolo zvoleného parametru θ_0 takový, že pokud do něj naměřená hodnota náleží, dá se v dané situaci považovat za ekvivalentní se zvoleným parametrem. Tyto hodnoty jsou určeny experty z příslušného oboru, nikoliv statistikem. Nulová hypotéza zde říká, že θ a θ_0 nejsou ekvivalentní, alternativní hypotéza naopak říká, že jsou ekvivalentní. Pomocí testů ekvivalence prokazujeme, že je měřený parametr dostatečně blízko skutečné hodnotě parametru, neboli že jeho odchylka je prakticky irelevantní.

Příklad 2. Uvažujme obdobné zadání jako u příkladu 1.

Máme předem určené, že k systematické odchylce nedošlo, pokud se hmotnost prášku v krabici liší od požadované hmotnosti o méně než 4 dkg.

Na rozdíl od příkladu 1 se tedy nyní pokusíme dokázat, že k systematické odchylce nedošlo.

Označme $X_1, \ldots, X_6$ jednotlivé odchylky nastavení automatu, o kterých předpokládáme, že tvoří náhodný výběr z rozdělení $\mathrm{N}(\mu, \sigma^2)$. Budeme testovat hypotézu, že došlo k systematické odchylce nastavení automatu, proti alternativě, že k odchylce nedošlo. Ze zadání víme, že $\epsilon_1 = \epsilon_2 = 4$. Chceme tedy testovat hypotézu

$$H_{0E} : \mu \leq -4 \text{ nebo } \mu \geq 4 \quad \text{proti} \quad H_{1E} : -4 < \mu < 4. \quad (2.2)$$

V následující sekci popíšeme postup řešení a ukážeme jeho aplikaci na tomto příkladu.

2.1 Princip inkluze intervalu spolehlivosti

Jedním z přístupů, pomocí kterých testujeme ekvivalenci, je princip inkluze intervalu spolehlivosti. Dalším z možných přístupů je „*power approach*“, viz např. Schuirmann (1987). V této práci se budeme zabývat principem inkluze intervalu spolehlivosti.

Řešíme test hypotézy (2.1). Testování provedeme pomocí dvou následujících jednostranných testů

$$H_{0L} : \theta \leq \theta_0 - \epsilon_1 \quad \text{proti} \quad H_{1L} : \theta > \theta_0 - \epsilon_1, \tag{2.3}$$

$$H_{0P} : \theta \geq \theta_0 + \epsilon_2 \quad \text{proti} \quad H_{1P} : \theta < \theta_0 + \epsilon_2. \tag{2.4}$$

Test hypotézy (2.1) rozhodne ve prospěch ekvivalence právě tehdy, když oba testy hypotéz (2.3) a (2.4) zamítnou svoji nulovou hypotézu. Tomuto postupu se říká TOST procedura, název je odvozen z anglického „*two one-sided tests*“ („dva jednostranné testy“).

Výše uvedený postup je ekvivalentní následujícímu. U testování hypotézy (2.3) označme $\underline{\theta}(\mathbf{X}; \alpha)$ dolní mez $(1 - \alpha)100\%$ jednostranného intervalu spolehlivosti pro θ a obdobně u testu hypotézy (2.4) označme $\overline{\theta}(\mathbf{X}; \alpha)$ horní mez $(1 - \alpha)100\%$ jednostranného intervalu spolehlivosti pro θ. Pokud platí $\underline{\theta}(\mathbf{X}; \alpha) > \theta_0 - \epsilon_1$ a zároveň $\overline{\theta}(\mathbf{X}; \alpha) < \theta_0 + \epsilon_2$, respektive pomocí intervalů zapsáno, pokud platí $\Big(\underline{\theta}(\mathbf{X}; \alpha),\, \overline{\theta}(\mathbf{X}; \alpha)\Big) \subset (\theta_0 - \epsilon_1,\, \theta_0 + \epsilon_2)$, zamítáme nulovou hypotézu testu (2.1) ve prospěch alternativy H_{1E}. Odtud pochází název princip inkluze intervalu spolehlivosti.

Pozorování 1. *Interval* $\Big(\underline{\theta}(\boldsymbol{X}; \alpha),\, \overline{\theta}(\boldsymbol{X}; \alpha)\Big)$ *má pravděpodobnost pokrytí* $1 - 2\alpha$.

Věta 2. *Test ekvivalence založený na principu inkluze intervalu spolehlivosti má hladinu nejvýše* α.

Důkaz. Chceme ukázat, že pravděpodobnost chyby I. druhu, tedy zamítnutí platné hypotézy, je nejvýše α. Nechť pro θ platí H_0. Řešme postupně tři možné případy.

1. Nechť $\theta \in (-\infty,\, \theta_0 - \epsilon_1]$. Potom platí

$$\begin{aligned}
\mathsf{P}_\theta(\text{zamítáme } H_0) &= \mathsf{P}_\theta\Big(\Big(\underline{\theta}(\mathbf{X}; \alpha),\, \overline{\theta}(\mathbf{X}; \alpha)\Big) \subset \Big(\theta_0 - \epsilon_1,\, \theta_0 + \epsilon_2\Big)\Big) \\
&\leq \mathsf{P}_\theta\Big(\theta_0 - \epsilon_1 < \underline{\theta}(\mathbf{X}; \alpha)\Big) \\
&= \mathsf{P}_\theta\Big(\theta \leq \theta_0 - \epsilon_1 < \underline{\theta}(\mathbf{X}; \alpha)\Big) \\
&\leq \mathsf{P}_\theta\Big(\theta < \underline{\theta}(\mathbf{X}; \alpha)\Big) \leq \alpha,
\end{aligned}$$

   kde třetí nerovnost plyne z definice $\underline{\theta}(\mathbf{X}; \alpha)$.

2. Nechť $\theta \in [\theta_0 + \epsilon_2,\, \infty)$. Obdobným postupem ukážeme, že platí

$$\mathsf{P}_\theta(\text{zamítáme } H_0) \leq \mathsf{P}_\theta\Big(\theta \geq \overline{\theta}(\mathbf{X}; \alpha)\Big) \leq \alpha.$$

3. Nechť $\theta \in (\theta_0 - \epsilon_1,\, \theta_0 + \epsilon_2)$. Zde je zamítnutí nulové hypotézy správným řešením.

Test ekvivalence ze znění věty má tedy hladinu nejvýše α.

□

Příklad 2 (pokračování)**.** Připomínáme, že ověřujeme, zda-li nedošlo k systematické odchylce nastavení automatu, který plní krabice pracím práškem. Dospěli jsme k testu hypotézy (2.2). Mějme $\alpha = 0{,}05$. Řešení příkladu provedeme nejdřív pomocí TOST procedury a poté pomocí principu inkluze intervalu spolehlivosti.

TOST procedura. Řešení příkladu rozdělíme na několik kroků.

V prvním kroku provedeme test hypotézy

$$H_{0L} : \mu \leq -4 \quad \text{proti} \quad H_{1L} : \mu > -4. \tag{2.5}$$

Obdobným postupem z příkladu 1 vypočítáme

$$T_6 = \sqrt{n}\,\frac{\overline{X}_n + \epsilon_1}{S_n} = \sqrt{6}\,\frac{-2+4}{\sqrt{30{,}4}} \doteq 0{,}889.$$

Na hladině α zamítáme H_{0L} právě tehdy, když $T_n \geq t_{n-1}(1-\alpha)$. Vidíme, že platí $T_6 \doteq 0{,}889 < 2{,}015 \doteq t_5(0{,}95)$, tedy H_{0L} nezamítáme ve prospěch H_{1L}. Interval spolehlivosti pro μ s koeficientem spolehlivosti $(1-\alpha)$ je

$$\left(\overline{X}_n - t_{n-1}(1-\alpha)\,\frac{S_n}{\sqrt{n}},\, \infty\right) \doteq (-6{,}536,\, \infty). \tag{2.6}$$

Ve druhém kroku provedeme test hypotézy

$$H_{0P} : \mu \geq 4 \quad \text{proti} \quad H_{1P} : \mu < 4.$$

Obdobně jako v prvním kroku spočítáme

$$T_6 = \sqrt{n}\,\frac{\overline{X}_n - \epsilon_2}{S_n} = \sqrt{6}\,\frac{-2-4}{\sqrt{30{,}4}} \doteq -2{,}666.$$

Na hladině α zamítáme H_{0P} právě tehdy, když $T_n \leq -t_{n-1}(1-\alpha)$. Vidíme, že platí $T_6 \doteq -2{,}666 < -2{,}015 \doteq -t_5(0{,}95)$, tedy H_{0P} zamítáme ve prospěch H_{1P}. Interval spolehlivosti pro μ s koeficientem spolehlivosti $(1-\alpha)$ je

$$\left(-\infty,\, \overline{X}_n + t_{n-1}(1-\alpha)\,\frac{S_n}{\sqrt{n}}\right) \doteq (-\infty,\, 2{,}536). \tag{2.7}$$

Jelikož test hypotézy (2.5) nezamítl svoji nulovou hypotézu, potom i test hypotézy (2.2) nezamítá svoji nulovou hypotézu ve prospěch alternativy.

Princip inkluze intervalu spolehlivosti. Z (2.6) máme $\underline{\theta}(\mathbf{X};\alpha) \doteq -6{,}536$ a z (2.7) víme $\overline{\theta}(\mathbf{X};\alpha) \doteq 2{,}536$. Odtud již vidíme, že platí

$$\left(\underline{\theta}(\mathbf{X};\alpha),\, \overline{\theta}(\mathbf{X};\alpha)\right) \doteq (-6{,}536,\, 2{,}536) \not\subset (-4,\, 4) = (-\epsilon_1,\, \epsilon_2).$$

Nemůžeme tedy potvrdit, že k systematické odchylce nastavení automatu nedochází.

2.2 Vybrané testy ekvivalence

V kapitole 4 se budeme zabývat využitím testů ekvivalence na testování kvality algoritmů umělé inteligence. Z tohoto důvodu se v této sekci podíváme na dva konkrétní testy ekvivalence pro párová data, které budeme využívat na zmíněné testování kvality, na modifikovaný párový t-test a McNemarův test.

2.2.1 Párový t-test

Mějme náhodný výběr

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

náhodných vektorů s dvourozměrnou distribuční funkcí takový, že

$$D_i = X_i - Y_i \sim \mathrm{N}(\delta, \sigma_D^2), \quad \forall i \in \{1, \ldots, n\},$$

kde $\delta = \mathsf{E}\,(D_i)$, $\sigma_D^2 = \mathrm{var}(D_i)$.

Uvedeme dvě možná provedení testu. Vycházíme z knihy Wellek (2010) a diplomové práce Rychterová (2019).

TOST procedura. U párových t-testů je převládající volba testovaného parametru $\delta = \mu_X - \mu_Y$, kde $\mu_X = \mathsf{E}\, X_i$ a $\mu_Y = \mathsf{E}\, Y_i$. Dle teorie popsané v sekci 2.1 víme, že testování ekvivalence

$$H_{0E} : \delta \leq -\epsilon_1 \text{ nebo } \delta \geq \epsilon_2 \quad \text{proti} \quad H_{1E} : -\epsilon_1 < \delta < \epsilon_2,$$

kde $\epsilon_1 > 0$, $\epsilon_2 > 0$, můžeme provést pomocí dvou jednostranných testů

$$H_{0L} : \delta \leq -\epsilon_1 \quad \text{proti} \quad H_{1L} : \delta > -\epsilon_1, \tag{2.8}$$

$$H_{0P} : \delta \geq \epsilon_2 \quad \text{proti} \quad H_{1P} : \delta < \epsilon_2. \tag{2.9}$$

Hypotézu H_{0E} zamítáme, pokud zamítáme obě hypotézy H_{0L}, H_{0P}.

Volba testovaného parametru jako $\delta = \mu_X - \mu_Y$ nemusí být ve všech případech ideální volbou. Dvě rozdělení $\mathrm{N}(\mu_1, \sigma^2)$ a $\mathrm{N}(\mu_2, \sigma^2)$ budou prakticky nerozeznatelná pro dostatečně velké σ a naopak pro σ jdoucí k nule budou plochy pod jednotlivými hustotami téměř disjunktní. Z tohoto důvodu budeme u druhé metody za testovaný parametr volit $\theta = \delta/\sigma_D$.

Poznámka 1. Ukážeme spojitost se znaménkovým testem, kde využijeme právě volbu testovaného parametru $\theta = \delta/\sigma_D$. Označme $p_+ = \mathsf{P}(D_i > 0)$, $p_0 = \mathsf{P}(D_i = 0)$, $p_- = \mathsf{P}(D_i < 0)$. Jelikož $D_i \sim \mathrm{N}(\delta, \sigma_D^2)$, potom platí $p_0 = 0$ a $p_- = 1 - p_+$. Dále platí

$$p_+ = \mathsf{P}(D_i > 0) = \mathsf{P}\Big(\frac{D_i - \delta}{\sigma_D} > \frac{-\delta}{\sigma_D}\Big) = 1 - \Phi\Big(\frac{-\delta}{\sigma_D}\Big) = \Phi\Big(\frac{\delta}{\sigma_D}\Big) = \Phi(\theta),$$

kde jsme využili faktu $\Phi(x) = 1 - \Phi(-x)$. Tedy testování ekvivalence u znaménkového testu

$$H_{0E} : p_+ \leq 1/2 - \epsilon_1 \text{ nebo } p_+ \geq 1/2 + \epsilon_2 \quad \text{proti} \quad H_{1E} : 1/2 - \epsilon_1 < p_+ < 1/2 + \epsilon_2,$$

kde $\epsilon_1 > 0$, $\epsilon_2 > 0$, je ekvivalentní s testováním

$$H_{0E} : \theta \leq -u_{1/2+\epsilon_1} \text{ nebo } \theta \geq u_{1/2+\epsilon_2} \quad \text{proti} \quad H_{1E} : -u_{1/2+\epsilon_1} < \theta < u_{1/2+\epsilon_2}.$$

Stejnoměrně nejsilnější test. Profesor Wellek ve své knize (Wellek, 2010) popisuje odvození stejnoměrně nejsilnějších testů ekvivalence. Nyní se podíváme na toto odvození pro párový t-test.

Nejprve definujme definice a tvrzení, které budeme potřebovat dále při odvozování testu.

Definice 4. *Nechť X a Z jsou nezávislé náhodné veličiny takové, že $X \sim \mathrm{N}(0,1)$ a $Z \sim \chi^2_n$. Potom náhodná veličina*

$$T = \frac{X + \nu}{\sqrt{Z/n}}$$

má necentrální t-rozdělení s n stupni volnosti a parametrem necentrality ν. *Distribuční funkci tohoto rozdělení budeme značit $\mathcal{T}_n(\cdot \mid \nu)$, hustotu budeme značit $g_{n;\nu}(\cdot)$.*

Pozorování 3. *Pro $\nu = 0$ je necentrální t-rozdělení s n stupni volnosti a parametrem necentrality ν totožné se Studentovým rozdělením s n stupni volnosti.*

Definice 5 (Wellek, 2010, Definition A.1.1)**.** *Nechť $(p_\theta(\cdot))_{\theta\in\Theta}$ je rodina reálných funkcí s definičním oborem $\mathcal{X}$. Nechť $\mathcal{X}$ a Θ jsou lineárně uspořádané množiny. Dále pro libovolné $n = 1, 2, \ldots$ označme $\mathcal{X}^{(n)}$, resp. $\Theta^{(n)}$, uspořádanou množinu n prvků $x_1, \ldots, x_n \in \mathcal{X}$, kde $x_1 < x_2 < \cdots < x_n$, resp. $\theta_1, \ldots, \theta_n \in \Theta$, kde $\theta_1 < \theta_2 < \cdots < \theta_n$. Označme*

$$\Delta_n \begin{pmatrix} x_1, \ldots, x_n \\ \theta_1, \ldots, \theta_n \end{pmatrix} = \det \begin{pmatrix} p_{\theta_1}(x_1) & \ldots & p_{\theta_1}(x_n) \\ \vdots & \ddots & \vdots \\ p_{\theta_n}(x_1) & \ldots & p_{\theta_n}(x_n) \end{pmatrix}$$

pro libovolné $(x_1, \ldots, x_n) \in \mathcal{X}^{(n)}$, $(\theta_1, \ldots, \theta_n) \in \Theta^{(n)}$. Pokud pro všechna $n \in \{1, \ldots, r\}$, $r \in \mathbb{N}$, platí

$$\Delta_n \begin{pmatrix} x_1, \ldots, x_n \\ \theta_1, \ldots, \theta_n \end{pmatrix} > 0 \quad \forall\Big((x_1, \ldots, x_n), (\theta_1, \ldots, \theta_n)\Big) \in \mathcal{X}^{(n)} \times \Theta^{(n)}, \tag{2.10}$$

potom rodinu $(p_\theta(\cdot))_{\theta\in\Theta}$ nazýváme striktně totálně pozitivní řádu r. *Dále budeme používat značení STP_r. Pokud (2.10) platí pro všechna $n \in \mathbb{N}$, potom rodinu $(p_\theta(\cdot))_{\theta\in\Theta}$ nazýváme* striktně totálně pozitivní řádu ∞, *značíme STP_∞.*

Lemma 4. *Mějme libovolné $n \in \mathbb{N}$ a $\nu \in \mathbb{R}$. Potom $(g_{n;\nu}(\cdot))_{\nu\in\mathbb{R}}$ je STP_∞.*

Důkaz. Viz Wellek (2010, Lemma A.1.3).

□

Tvrzení 5. *Nechť $\alpha \in (0, 1)$, $\mathcal{X}$ je nedegenerovaný interval v $\mathbb{R}$, $\mathcal{B}_\mathcal{X}$ je borelovská σ-algebra na $\mathcal{X}$, μ je σ-konečná míra na $\mathcal{B}_\mathcal{X}$, jejíž nosič obsahuje alespoň dva různé body. Nechť $(p_\theta(\cdot))_{\theta\in\Theta}$ je STP_3 rodina hustot vůči μ na $\mathcal{X}$, kde Θ je nedegenerovaný interval v $\mathbb{R}$. Předpokládejme dále, že funkce $(x, \theta) \mapsto p_\theta(x)$ je spojitá v obou argumentech. Nechť $\theta_1, \theta_2 \in \Theta$, $\theta_1 < \theta_2$, jsou dány.*

1. Potom pro testování hypotézy

$$H_0 : \theta \in \Theta \setminus (\theta_1, \theta_2) \quad \textit{proti} \quad H_1 : \theta \in (\theta_1, \theta_2), \tag{2.11}$$

existuje stejnoměrně nejsilnější test $\phi : \mathcal{X} \to [0, 1]$ *s hladinou* α *tvaru*

$$\phi(x) = \begin{cases} 1, & \textit{pokud} \quad x \in (C_1, C_2), \\ \gamma_i, & \textit{pokud} \quad x = C_i, i = 1{,}2, \\ 0, & \textit{pokud} \quad x \in \mathcal{X} \setminus [C_1, C_2], \end{cases}$$

kde $C_i \in \mathcal{X}$, $i = 1, 2$, $C_1 < C_2$ *a*

$$E_{\theta_i}\phi(X) = \int \phi \, p_{\theta_i} \, d\mu = \alpha, \quad i = 1, 2.$$

2. Tento test ϕ *je stejnoměrně nejsilnějším testem pro* (2.11) *na hladině* α.

Důkaz. Viz Wellek (2010, Theorem A.1.5). □

Z důvodů popsaných výše budeme za testovaný parametr volit $\theta = \delta/\sigma_D$ a testujeme ekvivalenci

$$H_{0E} : \theta \leq \theta_1 \text{ nebo } \theta \geq \theta_2 \quad \text{proti} \quad H_{1E} : \theta_1 < \theta < \theta_2, \tag{2.12}$$

kde $\theta_1 < \theta_2$ jsou předem stanovené konstanty. Volíme testovou statistiku

$$T_n = \sqrt{n}\,\frac{\overline{D}_n}{S_n},$$

kde

$$\overline{D}_n = \frac{1}{n}\sum_{i=1}^{n} D_i, \quad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D}_n)^2.$$

Testová statistika T_n má necentrální t-rozdělení s $n-1$ stupni volnosti a parametrem necentrality $\sqrt{n}\,\delta/\sigma_D$, jelikož platí

$$T_n = \sqrt{n}\,\frac{\overline{D}_n}{S_n} = \frac{\sqrt{n}\,\dfrac{\overline{D}_n - \delta}{\sigma_D} + \sqrt{n}\,\dfrac{\delta}{\sigma_D}}{\sqrt{\dfrac{(n-1)S_n^2/\sigma_D^2}{n-1}}},$$

kde

$$\sqrt{n}\,\frac{\overline{D}_n - \delta}{\sigma_D} \sim \mathrm{N}(0, 1), \quad \frac{(n-1)S_n^2}{\sigma_D^2} \sim \chi^2_{n-1}$$

jsou nezávislé díky nezávislosti $\overline{D}_n$ a S_n^2.

Označíme $\tilde{\theta} = \sqrt{n}\,\delta/\sigma_D$. Potom (2.12) je ekvivalentní s testem ekvivalence

$$\tilde{H}_{0E} : \tilde{\theta} \leq \sqrt{n}\,\theta_1 \text{ nebo } \tilde{\theta} \geq \sqrt{n}\,\theta_2 \quad \text{proti} \quad \tilde{H}_{1E} : \sqrt{n}\,\theta_1 < \tilde{\theta} < \sqrt{n}\,\theta_2. \tag{2.13}$$

Z lemmatu 4 víme, že rodina $(g_{n;\tilde{\theta}}(\cdot))_{\tilde{\theta}\in\mathbb{R}}$, kde $g_{n;\tilde{\theta}}(\cdot)$ značí hustotu necentrálního t-rozdělení s $n-1$ stupni volnosti a parametrem necentrality $\tilde{\theta}$, je STP_∞,

tedy i STP_3. Z Wellek (2010, str. 93) víme, že $g_{n;\tilde{\theta}}(t)$ je spojitá v obou argumentech pro libovolné $n \in \mathbb{N}$.

Ověřili jsme předpoklady tvrzení 5, ze kterého dostaneme stejnoměrně nejsilnější test na hladině α pro (2.13) daný rozhodovacím pravidlem, které říká, že zamítáme nulovou hypotézu $\tilde{H}_{0E}$ právě tehdy, když

$$C_1 < T_n < C_2,$$

kde C_1, C_2 splňují

$$\begin{aligned} \mathsf{P}_{\sqrt{n}\,\theta_1}(C_1 < T_n < C_2) &= \mathcal{T}_{n-1}(C_2 \mid \sqrt{n}\,\theta_1) - \mathcal{T}_{n-1}(C_1 \mid \sqrt{n}\,\theta_1) = \alpha, \\ \mathsf{P}_{\sqrt{n}\,\theta_2}(C_1 < T_n < C_2) &= \mathcal{T}_{n-1}(C_2 \mid \sqrt{n}\,\theta_2) - \mathcal{T}_{n-1}(C_1 \mid \sqrt{n}\,\theta_2) = \alpha, \\ & -\infty < C_1 < C_2 < \infty. \end{aligned}$$

Hodnoty C_1, C_2 lze nalézt pomocí algoritmu uvedeného v knize Wellek (2010, str. 42--43).

Dále se podíváme na speciální případ, kdy je interval ekvivalence volen symetricky okolo nuly. Využijeme následující definici a tvrzení.

Definice 6. *Nechť X a Y jsou nezávislé náhodné veličiny takové, že X má necentrální χ^2-rozdělení s n stupni volnosti a parametrem necentrality ν (viz definice 7) a Y má χ^2-rozdělení s m stupni volnosti. Potom náhodná veličina*

$$F = \frac{X/n}{Y/m}$$

má necentrální F-rozdělení s n, m stupni volnosti a parametrem necentrality ν. *Pro libovolné $\alpha \in (0,\, 1)$ budeme α-kvantil tohoto rozdělení značit $F_{n,m;\,\alpha}(\nu)$.*

Pozorování 6. *Pro $\nu = 0$ je necentrální F-rozdělení s n, m stupni volnosti a parametrem necentrality ν totožné se F-rozdělením s n, m stupni volnosti.*

Tvrzení 7 (Johnson a kol., 1995, str. 516--517)**.** *Nechť má náhodná veličina T necentrální t-rozdělení s n stupni volnosti a parametrem necentrality ν. Potom T^2 má necentrální F-rozdělení s $1, n$ stupni volnosti a parametrem necentrality ν^2.*

Tvrzení 8. *Nechť jsou splněny předpoklady tvrzení 5. Nechť krajní body intervalu ekvivalence pro θ jsou zvoleny $\theta_1 = -\epsilon$, $\theta_2 = \epsilon$, kde $\epsilon > 0$ je libovolné pevné. Navíc předpokládejme, že za platnosti $\theta = \epsilon$ je rozdělení náhodné veličiny X stejné jako rozdělení $-X$ za platnosti $\theta = -\epsilon$. Potom stejnoměrně nejsilnější test $\phi : \mathcal{X} \to [0,\, 1]$ s hladinou α pro*

$$H_0 : \theta \in \Theta \setminus (-\epsilon,\, \epsilon) \quad \textit{proti} \quad H_1 : \theta \in (-\epsilon,\, \epsilon),$$

je daný předpisem

$$\phi(x) = \begin{cases} 1, & \textit{pokud} \quad |x| < C, \\ \gamma, & \textit{pokud} \quad |x| = C, \\ 0, & \textit{pokud} \quad |x| > C, \end{cases}$$

kde

$$C = \max\Big\{x \in [0, \infty) : \mathsf{P}_\epsilon(|X| < x) \le \alpha\Big\}$$

a

$$\gamma = \begin{cases} \dfrac{\alpha - \mathsf{P}_\epsilon(|X| < C)}{\mathsf{P}_\epsilon(|X| = C)}, & \textit{pokud} \quad \mathsf{P}_\epsilon(|X| = C) > 0, \\ 0, & \textit{pokud} \quad \mathsf{P}_\epsilon(|X| = C) = 0. \end{cases}$$

Důkaz. Viz Wellek (2010, Lemma A.1.6).

□

Nechť dále platí $\theta_1 = -\epsilon$, $\theta_2 = \epsilon$. Označme $\tilde{\epsilon} = \sqrt{n}\,\epsilon$. Potom (2.13) můžeme přepsat do tvaru

$$\tilde{H}_{0E} : \tilde{\theta} \leq -\tilde{\epsilon} \text{ nebo } \tilde{\theta} \geq \tilde{\epsilon} \quad \text{proti} \quad \tilde{H}_{1E} : -\tilde{\epsilon} < \tilde{\theta} < \tilde{\epsilon}. \tag{2.14}$$

Víme, že testová statistika T_n za podmínky $\tilde{\theta} = \tilde{\epsilon}$ má necentrální t-rozdělení s $n-1$ stupni volnosti a parametrem necentrality $\tilde{\epsilon}$. Díky symetrii normovaného normálního rozdělení má $-T_n$ necentrální t-rozdělení s $n-1$ stupni volnosti a parametrem necentrality $\tilde{\theta} = -\sqrt{n}\,\delta/\sigma_D$. Vidíme tedy, že za podmínky $\tilde{\theta} = -\tilde{\epsilon}$ má $-T_n$ stejné rozdělení jako T_n za podmínky $\tilde{\theta} = \tilde{\epsilon}$. Ověřili jsme předpoklady tvrzení 8, ze kterého dostaneme stejnoměrně nejsilnější test s hladinou α pro (2.14) s rozhodovacím pravidlem, že zamítáme nulovou hypotézu $\tilde{H}_{0E}$ právě tehdy, když

$$|T_n| < C,$$

kde $0 < C < \infty$ a

$$\alpha = \mathsf{P}_{\tilde{\epsilon}}(|T_n| < C) = \mathsf{P}_{\tilde{\epsilon}}(T_n^2 < C^2). \tag{2.15}$$

Jelikož T_n má necentrální t-rozdělení s $n-1$ stupni volnosti a parametrem necentrality $\tilde{\epsilon}$, potom z tvrzení 7 víme, že T_n^2 má necentrální F-rozdělení s $1, n-1$ stupni volnosti a parametrem necentrality $\tilde{\epsilon}^2$. Z tohoto faktu a z (2.15) určíme

$$C^2 = F_{1,\,n-1;\,\alpha}(\tilde{\epsilon}^2),$$
$$C = \sqrt{F_{1,\,n-1;\,\alpha}(\tilde{\epsilon}^2)}.$$

Tedy u testu ekvivalence (2.14) zamítáme nulovou hypotézu právě tehdy, když

$$|T_n| < \sqrt{F_{1,\,n-1;\,\alpha}(\tilde{\epsilon}^2)} = \sqrt{F_{1,\,n-1;\,\alpha}(n\epsilon^2)}.$$

V tabulce 2.1 uvádíme hodnoty $\sqrt{F_{1,\,n-1;\,0{,}05}(n\epsilon^2)}$ pro vybrané n a ϵ.

	ϵ				
n	0,1	0,2	0,3	0,4	0,5
10	0,0678	0,0787	0,1011	0,1431	0,2219
20	0,0702	0,0948	0,1556	0,3038	0,6136
30	0,0735	0,1151	0,2398	0,5744	1,0872
40	0,0771	0,1401	0,3610	0,8858	1,5034
50	0,0809	0,1704	0,5168	1,1783	1,8710
60	0,0850	0,2069	0,6912	1,4452	2,2039
70	0,0893	0,2503	0,8675	1,6909	2,5103
80	0,0938	0,3013	1,0371	1,9198	2,7958
90	0,0986	0,3600	1,1982	2,1349	3,0642
100	0,1036	0,4258	1,3511	2,3385	3,3183

Tabulka 2.1: Hodnoty $\sqrt{F_{1,\,n-1;\,0{,}05}(n\epsilon^2)}$.

2.2.2 McNemarův test

Při testování kvality algoritmů umělé inteligence narazíme na situaci, kdy u každé ze dvou porovnávaných metod mohou nastat pouze dvě možnosti, například úspěch či neúspěch metody nebo přítomnost či nepřítomnost sledovaného znaku. Z tohoto důvodu využijeme modifikovaný McNemarův test. Vycházíme z knihy Wellek (2010) a ze skript Kulich a Omelka (2022).

Mějme náhodný výběr

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

náhodných vektorů s dvourozměrnou distribuční funkcí, kde X_i a Y_i nabývají hodnot 0 a 1 pro všechna $i \in \{1, \ldots, n\}$. Označme 1 úspěch, resp. přítomnost testovaného znaku, a 0 neúspěch, resp. nepřítomnost znaku. Pro všechna $j, k \in \{0, 1\}$, $i \in \{1, \ldots, n\}$ zavedeme značení

$$\begin{aligned} p_{jk} &= \mathsf{P}(X_i = j,\, Y_i = k), \\ n_{jk} &= \sum_{i=1}^{n} \mathbb{I}\{X_i = j,\, Y_i = k\}, \end{aligned}$$

kde definujeme

$$\mathbb{I}\{X_i = j,\, Y_i = k\} = \begin{cases} 1, & \text{pokud } X_i = j,\, Y_i = k, \\ 0, & \text{jinak,} \end{cases}$$

přičemž zjevně platí

$$n = n_{00} + n_{01} + n_{10} + n_{11}.$$

Tyto hodnoty můžeme sestavit do následující kontingenční tabulky.

	Metoda B		
Metoda A	0	1	Σ
0	n_{00}	n_{01}	$n_{00} + n_{01}$
	(p_{00})	(p_{01})	$(p_{00} + p_{01})$
1	n_{10}	n_{11}	$n_{10} + n_{11}$
	(p_{10})	(p_{11})	$(p_{10} + p_{11})$
Σ	$n_{00} + n_{10}$	$n_{01} + n_{11}$	n
	$(p_{00} + p_{10})$	$(p_{01} + p_{11})$	(1)

Tabulka 2.2: Četnosti (pravděpodobnosti) u McNemarova testu.

Standardní McNemarův test je popsán například v knize profesora Anděla (Anděl, 2007, sekce 13.6). Dále odvodíme asymptotický test ekvivalence založený na McNemarově testu s pomocí knihy profesora Welleka (Wellek, 2010).

Naším cílem je odvodit test ekvivalence, kterým bychom mohli prokázat rovnost p_{10} a p_{01}. Testovaný parametr tedy zvolíme jako $\delta = p_{10} - p_{01}$ a testujeme hypotézu

$$H_{0E} : \delta \leq -\epsilon_1 \text{ nebo } \delta \geq \epsilon_2 \quad \text{proti} \quad H_{1E} : -\epsilon_1 < \delta < \epsilon_2, \tag{2.16}$$

kde $\epsilon_1 \in (0,\, 1)$, $\epsilon_2 \in (0,\, 1)$. Tři extrémní případy, kdy $p_{10} = p_{01} = 0$ nebo $p_{10} = 1$ nebo $p_{01} = 1$, vyřadíme a nebudeme dále uvažovat.

Označme

$$\mathbf{p} = (p_{00},\, p_{01},\, p_{10},\, p_{11})^\top,$$
$$\mathbf{N} = (n_{00},\, n_{01},\, n_{10},\, n_{11})^\top.$$

Jelikož je $\mathbf{N}$ vektor četností v kontingenční tabulce, tak víme, že platí

$$\mathbf{N} \sim \mathrm{Mult}_4(n,\, \mathbf{p}).$$

Pravděpodobnosti p_{jk}, $j,k \in \{0,\, 1\}$ odhadneme pomocí $\hat{p}_{jk} = n_{jk}/n$. Dále označíme $\hat{\delta}_n = \hat{p}_{10} - \hat{p}_{01} = (n_{10} - n_{01})/n$ a

$$\hat{\mathbf{p}}_n = (\hat{p}_{00},\, \hat{p}_{01},\, \hat{p}_{10},\, \hat{p}_{11})^\top.$$

Ze skript Kulich a Omelka (2022, Věta 8.3 (i)) víme, že platí

$$\sqrt{n}(\hat{\mathbf{p}}_n - \mathbf{p}) = \frac{1}{\sqrt{n}}(\mathbf{N} - n\mathbf{p}) \xrightarrow[n\to\infty]{D} \mathrm{N}_4(\mathbf{0}, \mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top),$$

kde $\mathrm{diag}(\mathbf{p})$ je diagonální matice, která má na diagonále prvky vektoru $\mathbf{p} = (p_{00},\, p_{01},\, p_{10},\, p_{11})^\top$.

Pomocí Δ-metody (viz Kulich a Omelka, 2022, Tvrzení 1.7), kde volíme funkci $g(t_1, t_2, t_3, t_4) = t_3 - t_2$, dostaneme

$$\sqrt{n}\Big((\hat{p}_{10} - \hat{p}_{01}) - (p_{10} - p_{01})\Big) \xrightarrow[n\to\infty]{D} \mathrm{N}\Big(0, (p_{10} + p_{01}) - (p_{10} - p_{01})^2\Big),$$

neboli

$$\sqrt{n}\frac{\hat{\delta}_n - \delta}{\sqrt{\eta - \delta^2}} \xrightarrow[n\to\infty]{D} \mathrm{N}(0, 1), \tag{2.17}$$

kde $\eta = p_{10} + p_{01}$.

Nyní uvedeme několik potřebných definic a tvrzení.

Definice 7. *Nechť $X_1, \ldots, X_n$, $n \in \mathbb{N}$, jsou nezávislé náhodné veličiny takové, že $X_i \sim \mathrm{N}(\mu_i, 1)$, $i \in \{1, \ldots, n\}$. Potom náhodná veličina*

$$Y = \sum_{i=1}^{n} X_i^2$$

má necentrální χ^2-rozdělení s n stupni volnosti a parametrem necentrality

$$\nu = \sum_{i=1}^{n} \mu_i^2.$$

Pro libovolné $\alpha \in (0,\, 1)$ budeme α-kvantil tohoto rozdělení značit $\chi^2_{n;\,\alpha}(\nu)$.

Pozorování 9. *Pokud položíme $\mu_i = 0$ pro všechna $i \in \{1, \ldots, n\}$, potom je necentrální χ^2-rozdělení s n stupni volnosti a parametrem necentrality ν totožné s χ^2-rozdělením s n stupni volnosti.*

Značení. Pro všechna $s > 0$ položme

$$c_\alpha^{\theta_1,\theta_2}(s) = C_\alpha\Big(s\,\frac{\theta_2-\theta_1}{2}\Big),$$

kde

$$C_\alpha(\psi) = \sqrt{\chi^2_{1;\,\alpha}(\psi^2)}.$$

Tvrzení 10. *Nechť pro všechna $n \in \mathbb{N}$ je*

$$X^{(n)} = \left(\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \ldots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}\right)$$

vektor 2-rozměrných náhodných veličin na pravděpodobnostním prostoru $(\Omega, \mathcal{A}, P)$ a nechť

$$\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \ldots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix} \sim \begin{pmatrix} F_1 \\ F_2 \end{pmatrix},$$

kde F_1, F_2 jsou 1-rozměrné distribuční funkce.

Nechť $\mathcal{F}$ je třída všech vektorů $\boldsymbol{F} = (F_1,\, F_2)$ distribučních funkcí splňujících níže uvedenou nulovou hypotézu a nechť pro všechna $n \in \mathbb{N}$ je testová statistika T_n dána tak, že platí

$$\sqrt{n}\frac{T_n - \theta(\boldsymbol{F})}{\sigma(\boldsymbol{F})} \xrightarrow[n\to\infty]{D} N(0,1) \quad \forall \boldsymbol{F} \in \mathcal{F}$$

pro vhodné funkcionály $\theta(\cdot) : \mathcal{F} \to \mathbb{R}$ a $\sigma(\cdot) : \mathcal{F} \to \mathbb{R}_+$.

Dále předpokládejme, že $\hat\sigma_n$ je konzistentní odhad $\sigma(\boldsymbol{F})$ pro $\forall \boldsymbol{F} \in \mathcal{F}$. Pokud definujeme

$$\phi_n(X^{(n)}) = \begin{cases} 1, & \textit{pokud} \quad \sqrt{n}\frac{|T_n-\theta_0|}{\hat\sigma_n} < c_\alpha^{\theta_1,\theta_2}(\frac{\sqrt{n}}{\hat\sigma_n}), \\ 0, & \textit{pokud} \quad \sqrt{n}\frac{|T_n-\theta_0|}{\hat\sigma_n} \geq c_\alpha^{\theta_1,\theta_2}(\frac{\sqrt{n}}{\hat\sigma_n}), \end{cases}$$

kde $\theta_0 = (\theta_1 + \theta_2)/2$, potom ϕ_n je asymptotický test na hladině $\alpha \in (0,\, 1)$ pro

$$H_0 : \theta(\boldsymbol{F}) \in \Theta\backslash(\theta_1,\, \theta_2) \quad \textit{proti} \quad H_1 : \theta(\boldsymbol{F}) \in (\theta_1,\, \theta_2).$$

Důkaz. Důkaz obecnější verze tohoto tvrzení lze najít v knize Wellek (2010, Theorem A.3.4).

□

Víme, že $\sqrt{\hat\eta_n - \hat\delta_n^2}$, kde $\hat\eta_n = \hat p_{10} + \hat p_{01}$, je konzistentní odhad $\sqrt{\eta - \delta^2}$. Tedy z (2.17) a z tvrzení 10 dostaneme, že pokud položíme

$$\phi_n = \begin{cases} 1, & \text{pokud} \quad \sqrt{n}\dfrac{\big|\hat\delta_n - (\epsilon_2-\epsilon_1)/2\big|}{\sqrt{\hat\eta_n - \hat\delta_n^2}} < c_\alpha^{-\epsilon_1,\epsilon_2}\Big(\dfrac{\sqrt{n}}{\sqrt{\hat\eta_n - \hat\delta_n^2}}\Big), \\ 0, & \text{pokud} \quad \sqrt{n}\dfrac{\big|\hat\delta_n - (\epsilon_2-\epsilon_1)/2\big|}{\sqrt{\hat\eta_n - \hat\delta_n^2}} \geq c_\alpha^{-\epsilon_1,\epsilon_2}\Big(\dfrac{\sqrt{n}}{\sqrt{\hat\eta_n - \hat\delta_n^2}}\Big), \end{cases}$$

potom ϕ_n je asymptotický test na hladině $\alpha \in (0,\, 1)$ pro test ekvivalence (2.16).

Zamítáme tedy nulovou hypotézu H_{0E} právě tehdy, když

$$\sqrt{n}\frac{\left|\hat{\delta}_n - (\epsilon_2 - \epsilon_1)/2\right|}{\sqrt{\hat{\eta}_n - \hat{\delta}_n^2}} < C_\alpha\Big(\frac{\sqrt{n}(\epsilon_1 + \epsilon_2)/2}{\sqrt{\hat{\eta}_n - \hat{\delta}_n^2}}\Big) = \sqrt{\chi^2_{1;\,\alpha}\Big(\frac{n(\epsilon_1 + \epsilon_2)^2/4}{\hat{\eta}_n - \hat{\delta}_n^2}\Big)},$$

neboli když

$$\sqrt{n}\frac{|n_{10} - n_{01} - n(\epsilon_2 - \epsilon_1)/2|}{\sqrt{n(n_{10} + n_{01}) - (n_{10} - n_{01})^2}} < \sqrt{\chi^2_{1;\,\alpha}\Big(\frac{n^3(\epsilon_1 + \epsilon_2)^2/4}{n(n_{10} + n_{01}) - (n_{10} - n_{01})^2}\Big)}.$$

Pro symetrický interval ekvivalence, kde $\epsilon = \epsilon_1 = \epsilon_2$, dostaneme pravidlo, že zamítáme H_{0E} právě tehdy, když

$$\frac{\sqrt{n}\ |n_{10} - n_{01}|}{\sqrt{n(n_{10} + n_{01}) - (n_{10} - n_{01})^2}} < \sqrt{\chi^2_{1;\,\alpha}\Big(\frac{n^3\,\epsilon^2}{n(n_{10} + n_{01}) - (n_{10} - n_{01})^2}\Big)}. \quad (2.18)$$

V tabulce 2.3 uvádíme hodnoty $\chi^2_{1;\,0,05}(\theta)$ a $\sqrt{\chi^2_{1;\,0,05}(\theta)}$ pro vybrané hodnoty θ.

θ	$\chi^2_{1;\,0,05}(\theta)$	$\sqrt{\chi^2_{1;\,0,05}(\theta)}$
5	0,3747	0,6122
10	2,3026	1,5174
15	4,9646	2,2281
20	7,9935	2,8273
25	11,2570	3,3551
30	14,6871	3,8324
35	18,2434	4,2712
40	21,8996	4,6797
45	25,6375	5,0634
50	29,4438	5,4262

Tabulka 2.3: Hodnoty $\chi^2_{1;\,0,05}(\theta)$ a $\sqrt{\chi^2_{1;\,0,05}(\theta)}$.

2.3 Testy noninferiority

Existují případy, kde nás zajímá, zda-li jsou naměřená data buď ekvivalentní dané konstantě ve smyslu uvedeném výše nebo je jejich hodnota vyšší, resp. nižší, než daná konstanta. Tyto situace si přiblížíme v této sekci. Opět vycházíme z knihy Wellek (2010).

Příklad 3. Při vývoji modifikace aktuálně používaného léku, která by snížila výrobní náklady o 40%, je potřeba zjistit, zda účinnost nové varianty není výrazně horší než u momentálně používaného léku. Zjišťujeme tedy, zda je nová verze léku ekvivalentní nebo lepší než jeho současná verze, abychom ji mohli uvést na trh.

Zápisem pomocí testu hypotézy dojdeme k jednostrannému testu, tedy testujeme hypotézu

$$H_{0I_L} : \theta \leq \theta_0 - \epsilon \quad \text{proti} \quad H_{1I_L} : \theta > \theta_0 - \epsilon, \tag{2.19}$$

resp. v případech, kde nižší hodnoty považujeme za „lepší“, testujeme hypotézu

$$H_{0I_P} : \theta \geq \theta_0 + \epsilon \quad \text{proti} \quad H_{1I_P} : \theta < \theta_0 + \epsilon, \tag{2.20}$$

kde $\epsilon > 0$ je předem stanovená konstanta, která je zvolena obdobně jako u testů ekvivalence. Hypotéza je zde stanovena jako „inferiorita“, alternativa naopak jako „noninferiorita“. Testy hypotéz (2.19) a (2.20) tak nazýváme testy noninferiority. Vidíme, že rozdíl mezi klasickým jednostranným testem a testem noninferiority spočívá v posunutí hranice doleva, resp. doprava, o hodnotu ϵ.

Podívejme se nyní blíže na vztah mezi testováním ekvivalence a noninferiority. U testů ekvivalence jsme zjišťovali, zda jsou naměřená data dostatečně podobná dané hodnotě, abychom je považovali za ekvivalentní. Naopak u testů noninferiority není naším cílem ukázat tuto „oboustrannou“ ekvivalenci, ale testovat, zda jsou námi naměřená data „jednostranně ekvivalentní“ předem dané hodnotě θ_0.

Volbou $\epsilon_1 = \epsilon$ a $\epsilon_2 = \infty$, nebo u množiny omezené zprava volbou $\epsilon_2 = \sup \Theta$, u testu ekvivalence (2.1) získáme test noninferiority (2.19). Obdobně můžeme postupovat u testu noninferiority (2.20).

Vidíme, že je mezi testy ekvivalence a noninferiority vztah, jelikož testy noninferiority můžeme získat modifikací testů ekvivalence. Z tohoto důvodu můžeme s testy noninferiority pracovat jako s „jednostrannou verzí“ testů ekvivalence, a tedy u TOST procedury je potřeba provádět pouze jeden ze dvou příslušných jednostranných testů.

Příklad 4. Uvažujme obdobné zadání jako u příkladu 1. Výrobce pracího prášku nechce na jeho výrobě prodělávat a rád by tedy ověřil, zda nedošlo k systematické odchylce nastavení automatu a ten neplní krabice větším množstvím prášku. Hranice ekvivalence je opět stanovena na 4 dkg.

Označme $X_1, \ldots, X_6$ jednotlivé odchylky nastavení automatu, o kterých předpokládáme, že tvoří náhodný výběr z rozdělení $\mathrm{N}(\mu, \sigma^2)$. Ze zadání víme $\epsilon = 4$. Chceme tedy testovat hypotézu

$$H_{0I_P} : \mu \geq 4 \quad \text{proti} \quad H_{1I_P} : \mu < 4.$$

Z druhého kroku u příkladu 2 víme, že platí $T_6 \doteq -2{,}666 < -2{,}015 \doteq -t_5(0{,}95)$, a zamítáme H_{0I_P} ve prospěch H_{1I_P}. Rozhodnutí testu si ještě znázorníme pomocí intervalů. Interval spolehlivosti pro μ s koeficientem spolehlivosti $(1-\alpha)$ je

$$\left(-\infty,\, \overline{X}_n + t_{n-1}(1-\alpha)\,\frac{S_n}{\sqrt{n}}\right) \doteq (-\infty,\, 2{,}536).$$

Odtud získáme $\overline{\theta}(\mathbf{X};\alpha) \doteq 2{,}536$ a vidíme, že platí

$$\left(-\infty,\, \overline{\theta}(\mathbf{X};\alpha)\right) \doteq (-\infty,\, 2{,}536) \subset (-\infty,\, 4) = (-\infty,\, \epsilon).$$

Potvrdili jsme, že nedochází k systematické odchylce takové, že by automat plnil krabice větším množstvím pracího prášku.

2.3.1 McNemarův test

Uvažujme stejnou situaci a značení jako v sekci 2.2.2. Nyní odvodíme asymptotický test noninferiority založený na McNemarově testu.

Testovaný parametr volíme jako $\delta = p_{10} - p_{01}$ a testujeme hypotézu

$$H_{0I_L} : \delta \leq -\epsilon \quad \text{proti} \quad H_{1I_L} : \delta > -\epsilon, \tag{2.21}$$

kde $\epsilon \in (0,\, 1)$.

Z (2.17) a z Cramer-Sluckého věty (viz Kulich a Omelka, 2022, Tvrzení 1.3) dostáváme

$$\sqrt{n}\frac{\hat{\delta}_n - \delta}{\sqrt{\hat{\eta}_n - \hat{\delta}_n^2}} \xrightarrow[n\to\infty]{D} N(0,1),$$

jelikož platí

$$\sqrt{\hat{\eta}_n - \hat{\delta}_n^2} \xrightarrow[n\to\infty]{\mathsf{P}} \sqrt{\eta - \delta^2}.$$

Mějme testovou statistiku

$$T_n = \sqrt{n}\frac{\hat{\delta}_n + \epsilon}{\sqrt{\hat{\eta}_n - \hat{\delta}_n^2}}.$$

Pro $\delta = -\epsilon$ je její rozdělení asymptoticky N(0, 1). Odtud dostáváme pravidlo, které říká, že zamítáme hypotézu H_{0I_L} právě tehdy, když $T_n \geq u_{1-\alpha}$, neboli když

$$\frac{\sqrt{n}(n_{10} - n_{10} + n\epsilon)}{\sqrt{n(n_{10} + n_{01}) - (n_{10} - n_{01})^2}} > u_{1-\alpha}.$$

Poznámka 2. Pro test hypotézy

$$H_{0I_P} : \delta \geq \epsilon \quad \text{proti} \quad H_{1I_P} : \delta < \epsilon \tag{2.22}$$

dostaneme pravidlo, které říká, že zamítáme hypotézu H_{0I_P} právě tehdy, když

$$\frac{\sqrt{n}(n_{10} - n_{10} - n\epsilon)}{\sqrt{n(n_{10} + n_{01}) - (n_{10} - n_{01})^2}} \leq -u_{1-\alpha}. \tag{2.23}$$

3. Plánování rozsahu výběru

V následující sekci se podíváme na plánování rozsahu výběru u testování ekvivalence. Vycházíme především z knih Chow a kol. (2017), resp. Julious (2023) a ze skript Kulich a Omelka (2022).

V této kapitole se zaměříme na plánování rozsahu výběru u modifikovaných párových t-testů a McNemarova testu popsaných výše.

Pravděpodobnost chyby I. druhu budeme značit α, pravděpodobnost chyby II. druhu β. Někteří autoři doporučují u plánování rozsahu výběru u testů ekvivalence za pravděpodobnost chyby I. druhu dosadit poloviční hodnotu než bychom použili u oboustranného testu hypotézy (např. viz Julious, 2023, Sekce 2.6.2). Jiní autoři naopak doporučují ponechat hodnotu pravděpodobnosti chyby I. druhu stejnou jako u oboustranného testu hypotézy (např. viz Wellek, 2010, Sekce 2.7). V této práci se budeme řídit doporučením z knihy profesora Welleka (Wellek, 2010). Hodnota síly testu, která je rovna doplňku pravděpodobnosti chyby II. druhu do hodnoty jedna, se standardně volí jako 0,90, minimální volená hodnota bývá 0,80.

Nejprve uvedeme potřebné pojmy a tvrzení. Připomeňme, že distribuční funkci necentrálního t-rozdělení s n stupni volnosti a parametrem necentrality ν značíme $\mathcal{T}_n(\cdot \mid \nu)$.

Tvrzení 11. *Nechť ν_1 a ν_2 jsou dva parametry necentrality. Potom pro libovolné $t \in \mathbb{R}$ platí*

$$\nu_1 \leq \nu_2 \Rightarrow \mathcal{T}_n(t \mid \nu_1) \geq \mathcal{T}_n(t \mid \nu_2).$$

Důkaz. Nechť ν_1 a ν_2 jsou hodnoty parametru necentrality takové, že $\nu_1 \leq \nu_2$. Potom platí

$$\mathcal{T}_n(t \mid \nu_1) = \mathsf{P}\Big(\frac{X+\nu_1}{\sqrt{Z/n}} \leq t\Big) = \mathsf{P}\Big(X - t\sqrt{Z/n} \leq -\nu_1\Big) \geq$$

$$\geq \mathsf{P}\Big(X - t\sqrt{Z/n} \leq -\nu_2\Big) = \mathsf{P}\Big(\frac{X+\nu_2}{\sqrt{Z/n}} \leq t\Big) = \mathcal{T}_n(t \mid \nu_2),$$

kde $X \sim \mathrm{N}(0,1)$ a $Z \sim \chi^2_n$.

□

Poznámka 3 (Chow a kol., 2017, str. 41). Pro dostatečně velká n platí, že $t_n(\alpha)$ můžeme aproximovat pomocí u_α a $\mathcal{T}_n(t_n(\alpha) \mid \nu)$ pomocí $\Phi(u_\alpha - \nu)$.

3.1 Párový t-test

Uvažujme náhodný výběr a značení zavedené v sekci 2.2.1. Testujeme

$$H_{0E} : \delta \leq \delta_0 - \epsilon_1 \text{ nebo } \delta \geq \delta_0 + \epsilon_2 \quad \text{proti} \quad H_{1E} : \delta_0 - \epsilon_1 < \delta < \delta_0 + \epsilon_2. \tag{3.1}$$

Budeme postupovat pomocí TOST procedury. Používáme značení $\overline{D}_n$ a S_n, kde

$$\overline{D}_n = \frac{1}{n}\sum_{i=1}^{n} D_i, \quad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D}_n)^2,$$

a značení β_1, β_2 pro pravděpodobnost chyby II. druhu u jednotlivých jednostranných testů hypotéz H_{0L}, H_{0P} z TOST procedury. Z Kulich a Omelka (2022, Věta 2.8 (ii)) víme, že $\overline{D}_n$ a S_n^2 jsou nezávislé.

V prvním kroku spočítáme pravděpodobnost chyby II. druhu u testu hypotézy

$$H_{0L} : \delta \leq \delta_0 - \epsilon_1 \quad \text{proti} \quad H_{1L} : \delta > \delta_0 - \epsilon_1.$$

Máme testovou statistiku

$$T_n = \sqrt{n}\,\frac{\overline{D}_n - (\delta_0 - \epsilon_1)}{S_n}.$$

Víme, že pro $\delta = \delta_0 - \epsilon_1$ platí $T_n \sim t_{n-1}$, jelikož platí

$$T_n = \sqrt{n}\,\frac{\overline{D_n} - (\delta_0 - \epsilon_1)}{S_n} = \frac{\sqrt{n}\,\dfrac{\overline{D}_n - (\delta_0 - \epsilon_1)}{\sigma_D}}{\sqrt{\dfrac{(n-1)S_n^2/\sigma_D^2}{n-1}}},$$

kde

$$\sqrt{n}\,\frac{\overline{D}_n - (\delta_0 - \epsilon_1)}{\sigma_D} \sim \mathrm{N}(0,1), \quad \frac{(n-1)S_n^2}{\sigma_D^2} \sim \chi^2_{n-1},$$

jsou nezávislé díky nezávislosti $\overline{D}_n$ a S_n^2. Kritický obor volíme tvaru $\mathcal{C}(\alpha) = [t_{n-1}(1-\alpha), \infty)$.

Vezměme nyní libovolné δ. Potom testová statistika

$$\begin{aligned} T_n &= \sqrt{n}\,\frac{\overline{D}_n - (\delta_0 - \epsilon_1)}{S_n} \\ &= \frac{\sqrt{n}\,\dfrac{\overline{D}_n - \delta}{\sigma_D} + \sqrt{n}\,\dfrac{\delta - \delta_0 + \epsilon_1}{\sigma_D}}{\sqrt{\dfrac{(n-1)S_n^2/\sigma_D^2}{n-1}}} \end{aligned}$$

má necentrální t-rozdělení s $n-1$ stupni volnosti a parametrem necentrality

$$\sqrt{n}\,\frac{\delta - \delta_0 + \epsilon_1}{\sigma_D},$$

jelikož platí, že

$$\sqrt{n}\,\frac{\overline{D}_n - \delta}{\sigma_D} \sim \mathrm{N}(0,1), \quad \frac{(n-1)S_n^2}{\sigma_D^2} \sim \chi^2_{n-1}$$

jsou nezávislé díky nezávislosti $\overline{D}_n$ a S_n^2.

Spočítejme silofunkci tohoto testu. Platí

$$\begin{aligned} \beta(\delta) &= \mathsf{P}_\delta\Big(T_n \in \mathcal{C}(\alpha)\Big) = \mathsf{P}_\delta\Big(T_n \geq t_{n-1}(1-\alpha)\Big) \\ &= 1 - \mathcal{T}_{n-1}\Big(t_{n-1}(1-\alpha) \,\Big|\, \sqrt{n}\,\frac{\delta - \delta_0 + \epsilon_1}{\sigma_D}\Big). \end{aligned}$$

Pokud neplatí hypotéza H_{0L}, existuje δ takové, že $\delta > \delta_0 - \epsilon_1$. Potom pravděpodobnost chyby II. druhu je

$$\beta_1 = \mathcal{T}_{n-1}\Big(t_{n-1}(1-\alpha)\,\Big|\,\sqrt{n}\,\frac{\delta-\delta_0+\epsilon_1}{\sigma_D}\Big).$$

Druhým krokem je výpočet pravděpodobnosti chyby II. druhu u testu hypotézy

$$H_{0P} : \delta \geq \delta_0 + \epsilon_2 \quad \text{proti} \quad H_{1P} : \delta < \delta_0 + \epsilon_2.$$

Máme testovou statistiku

$$T_n = \sqrt{n}\,\frac{\overline{D}_n - (\delta_0+\epsilon_2)}{S_n}.$$

Obdobně jako v prvním kroku víme, že pro $\delta = \delta_0 + \epsilon_2$ platí $T_n \sim t_{n-1}$, kritický obor volíme tvaru $\mathcal{C}(\alpha) = (-\infty, -t_{n-1}(1-\alpha)]$. Pro libovolné δ má testová statistika T_n necentrální t-rozdělení s $n-1$ stupni volnosti a parametrem necentrality

$$\sqrt{n}\,\frac{\delta-\delta_0-\epsilon_2}{\sigma_D}.$$

Ze symetrie normovaného normálního rozdělení dostaneme, že

$$-T_n = \frac{\sqrt{n}\,\dfrac{-\overline{D}_n+\delta}{\sigma_D} + \sqrt{n}\,\dfrac{-\delta+\delta_0+\epsilon_2}{\sigma_D}}{\sqrt{\dfrac{(n-1)S_n^2/\sigma_D^2}{n-1}}}$$

má necentrální t-rozdělení s $n-1$ stupni volnosti a parametrem necentrality

$$\sqrt{n}\,\frac{-\delta+\delta_0+\epsilon_2}{\sigma_D}.$$

Silofunkce tohoto testu je

$$\begin{aligned}\beta(\delta) &= \mathsf{P}_\delta\Big(T_n \in \mathcal{C}(\alpha)\Big) = \mathsf{P}_\delta\Big(T_n \leq -t_{n-1}(1-\alpha)\Big) = \mathsf{P}_\delta\Big(-T_n \geq t_{n-1}(1-\alpha)\Big)\\ &= 1 - \mathcal{T}_{n-1}\Big(t_{n-1}(1-\alpha)\,\Big|\,\sqrt{n}\,\frac{-\delta+\delta_0+\epsilon_2}{\sigma_D}\Big).\end{aligned}$$

Pokud neplatí hypotéza H_{0P}, potom existuje δ takové, že $\delta < \delta_0 + \epsilon_2$. Pravděpodobnost chyby II. druhu potom je

$$\beta_2 = \mathcal{T}_{n-1}\Big(t_{n-1}(1-\alpha)\,\Big|\,\sqrt{n}\,\frac{\epsilon-(\mu_1-\mu_0)}{\sigma_D}\Big).$$

Vyjádřili jsme pravděpodobnost chyby II. druhu u obou jednostranných testů hypotéz z TOST procedury testu ekvivalence (3.1). Pro pravděpodobnost chyby II. druhu u tohoto testu ekvivalence pro $\delta \in (\delta_0-\epsilon_1,\, \delta_0+\epsilon_2)$ platí

$$\begin{aligned}\beta &= \beta_1 + \beta_2\\ &= \mathcal{T}_{n-1}\Big(t_{n-1}(1-\alpha)\,\Big|\,\sqrt{n}\,\frac{\delta-\delta_0+\epsilon_1}{\sigma_D}\Big) + \mathcal{T}_{n-1}\Big(t_{n-1}(1-\alpha)\,\Big|\,\sqrt{n}\,\frac{-\delta+\delta_0+\epsilon_2}{\sigma_D}\Big)\\ &= \mathcal{T}_{n-1}\Big(t_{n-1}(1-\alpha)\,\Big|\,\sqrt{n}\,\frac{\delta+\epsilon-\tilde{\epsilon}}{\sigma_D}\Big) + \mathcal{T}_{n-1}\Big(t_{n-1}(1-\alpha)\,\Big|\,\sqrt{n}\,\frac{-\delta+\epsilon+\tilde{\epsilon}}{\sigma_D}\Big)\\ &= \mathcal{T}_{n-1}\Big(t_{n-1}(1-\alpha)\,\Big|\,\sqrt{n}\,\frac{\epsilon+|\delta-\tilde{\epsilon}|}{\sigma_D}\Big) + \mathcal{T}_{n-1}\Big(t_{n-1}(1-\alpha)\,\Big|\,\sqrt{n}\,\frac{\epsilon-|\delta-\tilde{\epsilon}|}{\sigma_D}\Big)\\ &\leq 2\,\mathcal{T}_{n-1}\Big(t_{n-1}(1-\alpha)\,\Big|\,\sqrt{n}\,\frac{\epsilon-|\delta-\tilde{\epsilon}|}{\sigma_D}\Big),\end{aligned}$$

kde nerovnost plyne z tvrzení 11, a kde

$$\epsilon = \frac{\epsilon_1 + \epsilon_2}{2}, \quad \tilde{\epsilon} = \frac{\epsilon_2 - \epsilon_1}{2} + \delta_0.$$

Vyřešením rovnice

$$\frac{\beta}{2} = \mathcal{T}_{n-1}\Big(t_{n-1}(1-\alpha) \,\Big|\, \sqrt{n}\,\frac{\epsilon - |\delta - \tilde{\epsilon}|}{\sigma_D}\Big) \tag{3.2}$$

$$= \mathcal{T}_{n-1}\left(t_{n-1}(1-\alpha) \,\middle|\, \sqrt{n}\,\frac{\frac{\epsilon_1+\epsilon_2}{2} - \left|\delta - \delta_0 - \frac{\epsilon_2-\epsilon_1}{2}\right|}{\sigma_D}\right) \tag{3.3}$$

obdržíme odhad rozsahu výběru potřebný k dosažení požadované síly $1-\beta$. K získání řešení můžeme využít tabulku 3.1 vhodnou substitucí za θ, ve které vidíme odhad rozsahu výběru pro danou sílu $1-\beta$ a zvolenou hodnotu θ.

	$1-\beta$			$1-\beta$	
θ	0,8	0,9	θ	0,8	0,9
0,05	3427	4331	0,55	30	38
0,10	858	1084	0,60	26	32
0,15	382	483	0,65	22	28
0,20	216	272	0,70	19	24
0,25	139	175	0,75	17	21
0,30	97	122	0,80	15	19
0,35	72	90	0,85	14	17
0,40	55	70	0,90	13	15
0,45	44	55	0,95	11	14
0,50	36	45	1,00	11	13

Tabulka 3.1: Nejmenší n splňující $\mathcal{T}_{n-1}(t_{n-1}(0{,}95) \,|\, \sqrt{n}\,\theta) \leq \beta/2$.

Použitím poznámky 3 vidíme, že pro rovnici (3.2) platí

$$\frac{\beta}{2} \approx \Phi\Big(u_{1-\alpha} - \sqrt{n}\,\frac{\epsilon - |\delta - \tilde{\epsilon}|}{\sigma_D}\Big),$$

$$-u_{1-\beta/2} = u_{\beta/2} \approx u_{1-\alpha} - \sqrt{n}\,\frac{\epsilon - |\delta - \tilde{\epsilon}|}{\sigma_D}.$$

Odtud dostáváme odhad rozsahu výběru

$$n \approx \frac{\sigma_D^2\,(u_{1-\alpha} + u_{1-\beta/2})^2}{\Big(\frac{\epsilon_1+\epsilon_2}{2} - \Big|\delta - \delta_0 - \frac{\epsilon_2-\epsilon_1}{2}\Big|\Big)^2}. \tag{3.4}$$

Rozsah výběru u testu hypotézy (3.1) tedy závisí na volbě hodnot pravděpodobností chyb I. a II. druhu, hodnotách určujících meze intervalu ekvivalence a na střední hodnotě a rozptylu měřených dat. Hodnoty δ a σ_D^2 většinou určíme na základě předchozích studií nebo pomocí pilotní studie.

3.2 McNemarův test

Uvažujme náhodný výběr a značení specifikované v sekci 2.2.2. Jelikož test, který jsme odvodili, je asymptotický, tak i zde bude odhad rozsahu výběru platit asymptoticky. Testujeme

$$H_{0E} : \delta \leq -\epsilon_1 \text{ nebo } \delta \geq \epsilon_2 \quad \text{proti} \quad H_{1E} : -\epsilon_1 < \delta < \epsilon_2. \tag{3.5}$$

Stejně jako u párového t-testu budeme postupovat pomocí TOST procedury a používat značení β_1, β_2 pro pravděpodobnost chyby II. druhu u jednotlivých jednostranných testů hypotéz H_{0L}, H_{0P} z TOST procedury.

Spočítejme nejprve pravděpodobnost chyby II. druhu u testu hypotézy

$$H_{0L} : \delta \leq -\epsilon_1 \quad \text{proti} \quad H_{1L} : \delta > -\epsilon_1.$$

Vezměme testovou statistiku

$$T_n = \sqrt{n}\,\frac{\hat{\delta}_n + \epsilon_1}{\sqrt{\eta - \delta^2}}.$$

Z (2.17) víme, že pro $\delta = p_{10} - p_{01} = -\epsilon_1$ má T_n asymptotické rozdělení N(0, 1), tedy hypotézu zamítáme, pokud $T_n \geq u_{1-\alpha}$.

Vezměme nyní libovolné p_{10}, p_{01}. Potom silofunkce testu je

$$\beta(p_{10}, p_{01}) = \mathsf{P}_{p_{10},p_{01}}\Big(T_n \geq u_{1-\alpha}\Big) = \mathsf{P}_{p_{10},p_{01}}\Big(\sqrt{n}\,\frac{\hat{\delta}_n - \delta}{\sqrt{\eta - \delta^2}} \geq u_{1-\alpha} - \sqrt{n}\,\frac{\delta + \epsilon_1}{\sqrt{\eta - \delta^2}}\Big).$$

Jelikož platí

$$\sqrt{n}\frac{\hat{\delta}_n - \delta}{\sqrt{\eta - \delta^2}} \xrightarrow[n\to\infty]{D} \mathrm{N}(0, 1),$$

můžeme silofunkci odhadnout výrazem

$$1 - \Phi\Big(u_{1-\alpha} - \sqrt{n}\,\frac{\delta + \epsilon_1}{\sqrt{\eta - \delta^2}}\Big).$$

Odtud pro $\delta = p_{10} - p_{01} > -\epsilon_1$ dostaneme odhad chyby II. druhu

$$\beta_1 \approx \Phi\Big(u_{1-\alpha} - \sqrt{n}\,\frac{\delta + \epsilon_1}{\sqrt{\eta - \delta^2}}\Big).$$

Podobným postupem dostaneme u testu hypotézy

$$H_{0P} : \delta \geq \epsilon_2 \quad \text{proti} \quad H_{1P} : \delta < \epsilon_2$$

pro $\delta = p_{10} - p_{01} < \epsilon_2$ odhad chyby II. druhu

$$\beta_2 \approx \Phi\Big(u_{1-\alpha} - \sqrt{n}\,\frac{-\delta + \epsilon_2}{\sqrt{\eta - \delta^2}}\Big).$$

Odhad chyby II. druhu testu ekvivalence (3.5) pro $\delta \in (-\epsilon_1, \epsilon_2)$ je

$$\begin{aligned}
\beta &= \beta_1 + \beta_2 \\
&\approx \Phi\Big(u_{1-\alpha} - \sqrt{n}\,\frac{\delta + \epsilon_1}{\sqrt{\eta - \delta^2}}\Big) + \Phi\Big(u_{1-\alpha} - \sqrt{n}\,\frac{-\delta + \epsilon_2}{\sqrt{\eta - \delta^2}}\Big) \\
&= \Phi\Big(u_{1-\alpha} - \sqrt{n}\,\frac{\delta + \epsilon - \tilde{\epsilon}}{\sqrt{\eta - \delta^2}}\Big) + \Phi\Big(u_{1-\alpha} - \sqrt{n}\,\frac{-\delta + \epsilon + \tilde{\epsilon}}{\sqrt{\eta - \delta^2}}\Big) \\
&= \Phi\Big(u_{1-\alpha} - \sqrt{n}\,\frac{\epsilon - |\delta - \tilde{\epsilon}|}{\sqrt{\eta - \delta^2}}\Big) + \Phi\Big(u_{1-\alpha} - \sqrt{n}\,\frac{\epsilon + |\delta - \tilde{\epsilon}|}{\sqrt{\eta - \delta^2}}\Big) \\
&\leq 2\Phi\Big(u_{1-\alpha} - \sqrt{n}\,\frac{\epsilon - |\delta - \tilde{\epsilon}|}{\sqrt{\eta - \delta^2}}\Big),
\end{aligned}$$

kde nerovnost plyne z faktu, že Φ je rostoucí funkce, a kde

$$\epsilon = \frac{\epsilon_1 + \epsilon_2}{2}, \quad \tilde{\epsilon} = \frac{\epsilon_2 - \epsilon_1}{2}.$$

Vyřešením následující rovnice dostaneme odhad rozsahu výběru, který je potřeba k dosažení síly $1 - \beta$. Řešme

$$\begin{aligned}
\beta &= 2\Phi\Big(u_{1-\alpha} - \sqrt{n}\,\frac{\epsilon - |\delta - \tilde{\epsilon}|}{\sqrt{\eta - \delta^2}}\Big), \\
u_{1-\beta/2} &= -u_{1-\alpha} + \sqrt{n}\,\frac{\epsilon - |\delta - \tilde{\epsilon}|}{\sqrt{\eta - \delta^2}}, \\
n &= \frac{(\eta - \delta^2)(u_{1-\alpha} + u_{1-\beta/2})^2}{(\epsilon - |\delta - \tilde{\epsilon}|)^2}.
\end{aligned}$$

Dostáváme odhad rozsahu výběru

$$n \approx \frac{\big(p_{10} + p_{01} - (p_{10} - p_{01})^2\big)(u_{1-\alpha} + u_{1-\beta/2})^2}{\Big(\frac{\epsilon_1 + \epsilon_2}{2} - \Big|p_{10} - p_{01} - \frac{\epsilon_2 - \epsilon_1}{2}\Big|\Big)^2}. \tag{3.6}$$

Obdobně jako u párového t-testu uvedeného dříve vidíme, že i zde rozsah výběru u testu hypotézy (3.5) závisí na volbě hodnot pravděpodobností chyb I. a II. druhu, hodnotách určujících meze intervalu ekvivalence a na pravděpodobnostech p_{10}, p_{01}. Hodnoty p_{10}, p_{01} většinou získáme na základě předchozích studií nebo pomocí pilotní studie.

Poznámka 4. Z výpočtu v této kapitole snadno dopočítáme odhad rozsahu výběru pro test noninferiority (2.21). Dostaneme

$$n \approx \frac{\big(p_{10} + p_{01} - (p_{10} - p_{01})^2\big)(u_{1-\alpha} + u_{1-\beta})^2}{(p_{10} - p_{01} + \epsilon)^2}.$$

Pro test noninferiority (2.22) dostaneme odhad

$$n \approx \frac{\big(p_{10} + p_{01} - (p_{10} - p_{01})^2\big)(u_{1-\alpha} + u_{1-\beta})^2}{(p_{01} - p_{10} + \epsilon)^2}.$$

4. Praktické využití

S rozvojem umělé inteligence (AI) jsou vyvíjeny různé algoritmy založené na AI, pomocí kterých se v daném oboru snažíme dosahovat podobných nebo lepších výsledků než současně používané metody. Abychom tyto algoritmy mohli používat, nesmí dosahovat odlišných nebo horších výsledků než současné metody. Na testování kvality těchto algoritmů můžeme využít právě párové testy ekvivalence nebo noninferiority.

V této kapitole uvedeme dva příklady testování kvality uměle inteligentních algoritmů ve zdravotnictví s použitím modifikovaného párového t-testu a McNemarova testu.

4.1 Párový t-test

Praktické použití párového t-testu vidíme v odborném článku Boninsegna a kol. (2024), který se věnuje porovnání výsledků radiologů a výsledků umělé inteligence při měření průměrů cév u CT angiografie před implementací aortální chlopně (TAVI).

Ve studii dva radiologové pracující společně a umělá inteligence měří rozměr aorty na devíti orientačních místech u 50 pacientů. Autoři studie na základě těchto výsledků ukazují, že lze využít algoritmus umělé inteligence pro analýzu CT angiografie, což může vést ke snížení variability a zrychlení analýzy. Na data, která splňovala předpoklad normálního rozdělení, použili autoři modifikovaný párový t-test. Pro data, která předpoklad normality nesplňovala, použili modifikovaný Mann-Whitneyho U test. Statistické testy byly provedeny pomocí TOST procedury, kde testovaný parametr byl zvolen jako rozdíl středních hodnot mezi měřením radiologů a umělé inteligence. Pro testování ekvivalence byly předem určeny dva intervaly ekvivalence, přísnější volba $\epsilon_1 = \epsilon_2 = 1\,mm$ a volnější volba $\epsilon_1 = \epsilon_2 = 2\,mm$, a hladina testu zvolena jako $\alpha = 0{,}05$.

V naší práci se zaměříme na orientační body měření z daného článku, ve kterých data splňují předpoklad normálního rozdělení a byl u nich použit párový t-test. Jedná se o místa nazvaná kořen aorty (*Sinus of Valsalva*), střed vzestupné aorty (*Mid ascending aorta*), střední oblouk (*Mid arch*) a střed sestupné aorty (*Mid descending aorta*).

Uvažujme konkrétní orientační bod a značení zavedené v sekci 2.2.1, kde $n = 50$, X_i značí naměřená data od radiologů a Y_i data od umělé inteligence, v obou případech uvedená v milimetrech. Připomeňme, že $\delta = \mathsf{E}(X_i) - \mathsf{E}(Y_i)$. Testujeme hypotézy

$$H_{0E_1} : \delta \leq -1 \text{ nebo } \delta \geq 1 \quad \text{proti} \quad H_{1E_1} : -1 < \delta < 1$$

a

$$H_{0E_2} : \delta \leq -2 \text{ nebo } \delta \geq 2 \quad \text{proti} \quad H_{1E_2} : -2 < \delta < 2$$

pomocí TOST procedury.

V článku Boninsegna a kol. (2024, Table 4) uvedli autoři p-hodnoty testů ekvivalence, kde výsledná p-hodnota testu ekvivalence provedeného pomocí TOST procedury se uvádí jako vyšší z dvojice příslušných jednostranných testů. Na výše

uvedených čtyřech orientačních bodech vyšla p-hodnota u volnější volby intervalu ekvivalence vždy menší než α, tedy H_{0E_2} zamítáme ve všech případech. Naopak u přísnější volby vyšla p-hodnota nižší než α pouze u středního oblouku, tedy H_{0E_1} zamítáme jen u tohoto případu.

Na základě výsledků uvedených v článku se nyní podíváme, jaký bychom potřebovali použít rozsah výběru, abychom dosáhli testu ekvivalence s hladinou $\alpha = 0{,}05$ a sílou testu 0,80.

Poznámka 5. Pro jednostranný test hypotézy (2.8) z TOST procedury označme realizovanou hodnotu testové statistiky

$$t_x = \sqrt{n}\,\frac{\overline{D}_n + \epsilon_1}{S_n} = \sqrt{n}\,\frac{\overline{X}_n - \overline{Y}_n + \epsilon_1}{S_n}.$$

Pro p-hodnotu platí

$$p(x) = \inf\{\alpha \in (0,\,1) : t_x \geq t_{n-1}(1-\alpha)\} = 1 - \mathcal{T}_{n-1}(t_x),$$

kde $\mathcal{T}_{n-1}(\cdot)$ značí distribuční funkci Studentova rozdělení s $n-1$ stupni volnosti. Tedy $t_x = t_{n-1}(1-p(x))$ a platí

$$S_n = \sqrt{n}\,\frac{\overline{X}_n - \overline{Y}_n + \epsilon_1}{t_{n-1}(1-p(x))}.$$

Obdobně pro test hypotézy (2.9) platí

$$S_n = \sqrt{n}\,\frac{\overline{X}_n - \overline{Y}_n - \epsilon_2}{t_{n-1}(p(x))}.$$

V orientačním bodě nazvaný kořen aorty máme hodnoty $n = 50$, $\delta_0 = 0$, $\overline{X}_{50} = 33{,}9$, $\overline{Y}_{50} = 33{,}6$. Pro $\epsilon_1 = \epsilon_2 = 1$ je p-hodnota rovna 0,1735. Z poznámky 5 dopočítáme $S_{50} \doteq 5{,}21$. Dosazením hodnot do rovnice (3.3), kde volíme $\delta = 33{,}9 - 33{,}6 = 0{,}3$ a $\sigma_D = 5{,}21$, a následným vyřešením dostaneme odhad rozsahu výběru $n = 476$. Použitím (3.4) dostaneme odhad $n \approx 475$.

Řešení pro ostatní případy shrneme do tabulky 4.1, kde $\epsilon = \epsilon_1 = \epsilon_2$.

	Metoda 1[a]		Metoda 2[b]	
	$\epsilon = 1$	$\epsilon = 2$	$\epsilon = 1$	$\epsilon = 2$
Kořen aorty	476	83	475	82
Střed vzestupné aorty	310	76	309	75
Střední oblouk	151	35	150	34
Střed sestupné aorty	158	37	157	35

Pozn: [a] Výpočet pomocí rovnice (3.3).

Pozn: [b] Výpočet pomocí odhadu (3.4).

Tabulka 4.1: Výpočet rozsahu výběru u příkladu měření aorty.

4.2 McNemarův test

Článek Menzies a kol. (2023) se věnuje porovnání úspěšnosti lékařů a umělé inteligence při diagnóze a průběhu léčení rakoviny kůže. Lékaři se zkušenostmi s diagnózou rakoviny kůže jsou rozděleni na specialisty se speciální kvalifikací v oboru a začínající lékaře bez speciální akreditace v oboru. Výsledky obou skupin lékařů se porovnávají s jedním ze dvou AI algoritmů: *7-class AI algorithm* nebo *International Skin Imaging Collaboration (dále jen ISIC) AI algorithm*. V naší práci ukážeme testování úspěšnosti diagnózy mezi specialisty a oběma algoritmy AI, ostatní testování by se provedlo obdobně.

Studie ve článku se zabývá diagnózou pigmentových lézí u pacientů s různými typy pleti podle Fitzpatrickovy škály (I-III). Každou pigmentovou lézi klasifikují do jedné ze sedmi daných kategorií vždy jeden specialista a oba umělé algoritmy. Symbolem P_s označíme pravděpodobnost správného rozhodnutí specialistů a symbolem P_a u AI algoritmů.

K testování naměřených dat použili autoři v článku asymptotický modifikovaný McNemarův test pro testování ekvivalence a noninferiority pomocí TOST procedury. Jednotlivé jednostranné testy z této TOST procedury jsou shodné s testy noninferiority popsanými v sekci 2.3.1.

Máme předem dané hodnoty hladiny testů $\alpha = 0{,}05$, intervalu ekvivalence $\epsilon_1 = \epsilon_2 = 0{,}1$, resp. $\epsilon = 0{,}1$, a máme celkem $n = 172$ měřených pigmentových lézí. Testovaný parametr volíme jako $\delta = \mathsf{P}_a - \mathsf{P}_s$ a dále uvažujeme značení zavedené v sekci 2.2.2. Úspěšné určení diagnózy značíme 1, neúspěšné 0.

Specialisté vs. 7-class AI algorithm. V tabulce 4.2 jsou uvedena naměřená data.

	Specialisté		
7-class AI	0	1	Σ
0	21	24	45
1	26	101	127
Σ	47	125	172

Tabulka 4.2: Specialisté vs. 7-class AI algorithm.

Hodláme prokázat, že *7-class AI algorithm* je v diagnóze pigmentových lézí stejně úspěšný jako specialisté. Testujeme ekvivalenci

$$H_{0E} : \delta \leq -0{,}1 \text{ nebo } \delta \geq 0{,}1 \quad \text{proti} \quad H_{1E} : -0{,}1 < \delta < 0{,}1.$$

Jelikož platí $\delta = \mathsf{P}_a - \mathsf{P}_s = (p_{10} + p_{11}) - (p_{01} - p_{11}) = p_{10} - p_{01}$, můžeme použít test odvozený v sekci 2.2.2, který říká, že hypotézu H_{0E} zamítáme právě tehdy, když platí (2.18). Dosazením do vzorce dostaneme

$$\frac{\sqrt{n}\ |n_{10} - n_{01}|}{\sqrt{n(n_{10} + n_{01}) - (n_{10} - n_{01})^2}} \doteq 0{,}283 <$$

$$< 0{,}794 \doteq \sqrt{\chi^2_{1;\,0{,}05}(5{,}920)} \doteq \sqrt{\chi^2_{1;\,\alpha}\Big(\frac{n^3\,\epsilon^2}{n(n_{10} + n_{01}) - (n_{10} - n_{01})^2}\Big)},$$

tedy zamítáme H_{0E} a prokázali jsme, že v diagnóze pigmentových lézí je *7-class AI algorithm* stejně úspěšný jako specialisté.

Specialisté vs. ISIC AI algorithm. Nyní se podíváme na porovnání diagnózy pigmentových lézí specialistů s druhým testovaným algoritmem. V tabulce 4.3 jsou uvedena naměřená data.

	Specialisté		
ISIC AI	0	1	Σ
0	27	40	67
1	20	85	105
Σ	47	125	172

Tabulka 4.3: Specialisté vs. ISIC AI algorithm.

Hodláme prokázat, že specialisté nejsou v určování diagnózy pigmentových lézí horší než *ISIC AI algorithm*. Testujeme hypotézu

$$H_{0I_P} : \delta \geq 0{,}1 \quad \text{proti} \quad H_{1I_P} : \delta < 0{,}1.$$

Ze sekce 2.3.1 víme rozhodovací pravidlo, která nám říká, že zamítáme H_{0I_P} právě tehdy, když platí (2.23). Dosazením do vzorce dostaneme

$$\frac{\sqrt{n}(n_{10} - n_{10} - n\epsilon)}{\sqrt{n(n_{10} + n_{01}) - (n_{10} - n_{01})^2}} \doteq -4{,}898 < -1{,}6449 \doteq -u_{1-\alpha},$$

tedy hypotézu H_{0I_P} zamítáme a prokázali jsme, že v určování diagnózy pigmentových lézí není *ISIC AI algorithm* lepší než specialisté.

Poznámka 6. Pokud bychom zde testovali ekvivalenci mezi diagnózou specialistů a *ISIC AI algorithm* obdobně jako u příkladu s *7-class AI algorithm*, po dosazení do vzorce (2.18) bychom dostali $2{,}633 \not< 0{,}638$ a hypotézu H_{0E} bychom nezamítali.

Ze vzorce (3.6) získáme odhad rozsahu výběru potřebný pro dosažení testu ekvivalence na hladině 0,05 se sílou testu 0,8. Dostáváme

$$n \approx \frac{\left(p_{10} + p_{01} - (p_{10} - p_{01})^2\right)(u_{1-\alpha} + u_{1-\beta/2})^2}{\left(\frac{\epsilon_1 + \epsilon_2}{2} - \left|p_{10} - p_{01} - \frac{\epsilon_2 - \epsilon_1}{2}\right|\right)^2} \doteq 10\,836.$$

V obou případech jsme použitím námi odvozených testů dospěli ke stejným výsledkům jako autoři v článku Menzies a kol. (2023).

Závěr

V této práci jsme se zabývali testy ekvivalence. Tyto testy využíváme v případech, kdy potřebujeme prokázat tvrzení, která se standardně dávají do nulové hypotézy. V první části práce jsme zavedli základní pojmy testování hypotéz a následně jsme se seznámili s testy ekvivalence a jedním možným postupem řešení nazvaným princip inkluze intervalu spolehlivosti.

Největším přínosem této práce je kapitola 3, ve které jsme se zabývali plánováním rozsahu výběru, hlavně u modifikovaného McNemarova testu.

Snahou bylo demonstrovat využití testů ekvivalence na testování kvality algoritmů založených na umělé inteligenci vůči předchozím používaným metodám. Z tohoto důvodu jsme se zaměřili na párová data a na modifikaci dvou testů právě pro tato data.

Prvním testem ekvivalence byla modifikace párového t-testu. Pro testovaný parametr δ jsme prezentovali řešení pomocí TOST procedury a zmínili nevýhodu volby tohoto parametru. Pro testovaný parametr δ/σ_D jsme ukázali nalezení stejnoměrně nejsilnějšího testu. Následně jsme pro volbu δ nalezli odhad rozsahu výběru potřebného k dosažení požadované síly testu a uvedli praktický příklad využití tohoto testu pro porovnání výsledků měření aorty provedené uměle inteligentním algoritmem a radiology.

Druhým vybraným testem ekvivalence byl modifikovaný McNemarův test. Předvedli jsme odvození tohoto asymptotického testu pro ekvivalenci i noninferioritu a následně jsme se zabývali plánováním rozsahu výběru testu. Na závěr jsme uvedli praktický příklad použití modifikovaného McNemarova testu pro ekvivalenci a noninferioritu na porovnání úspěšnosti výsledků specialistů a uměle inteligentních algoritmů při stanovení diagnóze rakoviny kůže.

Seznam použité literatury

Anděl, J. (1998). *Statistické metody*. Druhé přepracované vydání. Matfyzpress, Praha. ISBN 80-85863-27-8.

Anděl, J. (2007). *Základy matematické statistiky*. Druhé opravené vydání. Matfyzpress, Praha. ISBN 80-7378-001-1.

Boninsegna, E. a kol. (2024). CT angiography prior to endovascular procedures: can artificial intelligence improve reporting? *Physical and Engineering Sciences in Medicine*. doi: 10.1007/s13246-024-01393-1. URL `https://doi.org/10.1007/s13246-024-01393-1`.

Chow, S.-C., Shao, J., Wang, H. a Lokhnygina, Y. (2017). *Sample size calculations in clinical research*. Third Edition. Chapman and Hall/CRC, New York. ISBN 978-1-138-74098-3.

Johnson, N. L., Kotz, S. a Balakrishnan, N. (1995). *Continuous Univariate Distributions, Volume 2*. Second Edition. John Wiley & Sons. ISBN 0-471-58494-0.

Julious, S. A. (2023). *Sample sizes for clinical trials*. Second Edition. Chapman and Hall/CRC, New York. ISBN 978-1-138-58789-2.

Kulich, M. a Omelka, M. (2022). NMSA331 Matematická statistika 1 Poznámky k přednášce. URL `https://www.karlin.mff.cuni.cz/~komarek/vyuka/2022_23/nmsa331/ms1.pdf`. Naposledy upraveno dne 13. srpna 2022.

Menzies, S. W. a kol. (2023). Comparison of humans versus mobile phone-powered artificial intelligence for the diagnosis and management of pigmented skin cancer in secondary care: a multicentre, prospective, diagnostic, clinical trial. *The Lancet. Digital health*, **5**(10), 679–691. doi: 10.1016/s2589-7500(23)00130-9. URL `https://doi.org/10.1016/s2589-7500(23)00130-9`.

Rychterová, N. (2019). Testování ekvivalence a noninferiority. Diplomová práce, Univerzita Karlova.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, **15**(6), 657–680.

Wellek, S. (2010). *Testing Statistical Hypotheses of Equivalence and Noninferiority*. Second Edition. Chapman and Hall/CRC, New York. ISBN 978-1-4398-0818-4.

Seznam tabulek

Seznam použitých zkratek

$\mathsf{E}\ X$	střední hodnota náhodné veličiny X
$\operatorname{var} X$	rozptyl náhodné veličiny X
α	hladina testu
β	pravděpodobnost chyby II. druhu
β_1, β_2	pravděpodobnosti chyb II. druhu u jednotlivých jednostranných testů z TOST procedury
ϵ_1, ϵ_2	konstanty určující interval ekvivalence
ϵ	konstanta určující interval noninferiority
$\Phi(\cdot)$	distribuční funkce normovaného normálního rozdělení
u_α	α-kvantil normovaného normálního rozdělení
$\mathcal{T}_n(\cdot)$	distribuční funkce t-rozdělení s n stupni volnosti
$t_n(\alpha)$	α-kvantil Studentova rozdělení s n stupni volnosti
$\mathcal{T}_n(\cdot \mid \nu)$	distribuční funkce necentrálního t-rozdělení s n stupni volnosti a parametrem necentrality ν
$g_{n;\,\nu}(\cdot)$	hustota necentrálního t-rozdělení s n stupni volnosti a parametrem necentrality ν
$\chi^2_{n;\,\alpha}(\nu)$	α-kvantil necentrálního χ^2-rozdělení s n stupni volnosti a parametrem necentrality ν
$F_{n,\,m;\,\alpha}(\nu)$	α-kvantil necentrálního F-rozdělení s n stupni volnosti a parametrem necentrality ν
TOST	procedura používající dva jednostranné testy
$\underline{\theta}(\mathbf{X};\alpha)$	dolní mez $(1-\alpha)100\%$ jednostranného intervalu spolehlivosti pro θ
$\overline{\theta}(\mathbf{X};\alpha)$	horní mez $(1-\alpha)100\%$ jednostranného intervalu spolehlivosti pro θ
STP_r	striktně totálně pozitivní řádu r
AI	umělá inteligence