UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA
KATEDRA MATEMATICKÉ ANALÝZY A APLIKACÍ MATEMATIKY

BAKALÁŘSKÁ PRÁCE

Statistické chyby v medicínském výzkumu

Vedoucí diplomové práce:
Mgr. Jana Vrbková
Rok odevzdání: 2010

Vypracovala:
Zuzana Tonhauserová
M – E, III. ročník

Prohlášení

Prohlašuji, že jsem diplomovou práci zpracovala samostatně pod vedením Mgr. Jany Vrbkové a s použitím uvedené literatury.

V Olomouci dne 13. dubna 2010

Poděkování

Na tomto místě bych chtěla poděkovat především své vedoucí bakalářské práce paní Mgr. Janě Vrbkové za odbornou pomoc, cenné rady a čas, který mi věnovala při tvorbě této práce. Také bych ráda poděkovala své rodině a přátelům, kteří mě po celou dobu studia podporovali.

Obsah

Úvod

Statistické metody testování hypotéz, odhadů parametrů, tvorby modelů apod. jsou důležitou součástí rozhodovacích procesů. Nikdy by však neměly být brány jako jediné východisko pro stanovení rozhodnutí. Extrapolace poznatků získaných z jednoho vzorku nebo více vzorků větší nekompletně prozkoumané populace je do jisté míry záležitostí důvěry. Aplikace statistických procedur je, nejen v medicíně, zatížena celým zástupem možných zdrojů chyb a omylů, mezi něž patří např. následující: užití stejné množiny dat pro formulaci hypotézy i pro její testování, výběr vzorků z nesprávné populace či chyby ve specifikaci populace pro níž má být odvozen nějaký závěr, chyby v získávání náhodného reprezentativního vzorku populace, užití nevhodných nebo neúčinných statistických metod, chyby v testování modelů apod. Podle Gooda a Hardina [1] je největším zdrojem chyb nechat statistické procedury rozhodovat za nás, podle hesla „Nepřemýšlej – použij počítač!“.

Cílem této práce je poskytnout zevrubný přehled často pozorovaných statistických chyb, nedostatků a úskalí, zejména v lékařské vědě, za účelem pomoci lékařům a dalším výzkumníkům vytvářet statisticky věrohodné výstupy jejich výzkumu. Aplikování těchto principů může vést ke zlepšení statistické kvality výzkumných prací publikovaných v lékařských časopisech. Soustřeďuji se na jednoduché, ale přitom důležité statistické problémy, a tak může být tato práce použita jako pomoc (návod) pro nestatistiky, jak provést statisticky věrohodný výzkum. Při tvorbě této práce mi jako základ posloužil zejména článek „Statistical errors in medical research – a review of common pitfalls.“ A. M. Strasaka a kol. [2].

V dnešní době je statistika všeobecně přijímána jako mocný nástroj ve vědeckém výzkumném procesu. V posledních čtyřech desetiletích byl zaznamenán významný vzestup v užívání statistických metod v odborných lékařských časopisech. Nicméně obecně platí, že standard užití statistických metod je nízký, proto jsou statistické analýzy prováděny často nesprávně a velké množství publikovaných lékařských výzkumných prací je zatíženo statistickými chybami a nedostatky. Problém je vážný, protože nevhodné užití statistické analýzy může vést k nesprávným závěrům a zkresleným výsledkům studií, v medicíně s dalekosáhlými život a zdraví ohrožujícími následky, a v neposlední řadě rovněž k plýtvání hodnotnými, často nenahraditelnými, zdroji.

Mnoho vydavatelů lékařské odborné literatury vynaložilo značné úsilí na to, aby zlepšili kvalitu statistiky přijetím směrnic (standardů, tzv. guidelines) užití statistických

metod pro autory a detailnějším recenzováním obdržených příspěvků z hlediska užití statistických metod. Přes všechna tato opatření se standardní používání statistiky v medicínských výzkumech v průběhu času jen málo posunulo k lepšímu. Poslední studie ukazují na to, že hlavní problémy přetrvávají [2].

Tato práce je rozsáhlá zejména z toho důvodu, že považuji za velmi důležité zmínit základní pojmy a principy nutné pro pochopení problematiky statistických chyb. Proto v první kapitole nejdříve seznámím čtenáře - „nestatistika“ s těmito důležitými pojmy a principy. Po přečtení této kapitoly by se měl čtenář lépe orientovat v následujících kapitolách. Ve druhé kapitole se pokusím čtenáři předložit srozumitelný přehled běžných statistických chyb, nástrah a nedostatků týkajících se různých stádií lékařského vědeckého výzkumného procesu, počínaje plánováním lékařských studií až po přípravu finální zprávy o výsledcích studie. Všechny diskutované problémy mají za cíl pomoci badatelům zaměřit se na to, co je důležité z pohledu statistiky a jak správně prezentovat statistické výsledky ve svých pracech. Místo výčtu závazných pravidel, která by se měla striktně dodržovat, se pokusím poskytnout rady a informace o obecných statistických problémech, které se týkají aspektů konceptu výzkumu (návrh/design studie), vlastní analýzy dat, dokumentace použitých statistických metod, prezentace studovaných dat a interpretace výsledků. V poslední kapitole uvádím, jak můžeme efektivně využít statistické metody, a přitom se vyhnout možným nástrahám.

1 Základní pojmy a principy

V této části se pokusím čtenáři přiblížit nejdůležitější pojmy a principy, které bychom měli znát pro pochopení problematiky statistických chyb a omylů v medicínském výzkumu. Nejdříve bych však chtěla zmínit, že statistiku lze rozdělit na statistiku deskriptivní (popisnou) a inferenční (induktivní).

- **Deskriptivní (popisná) statistika** – jak už z jejího názvu vyplývá, se zabývá popisem a účelnou sumarizací statistických dat a jejich uspořádáním. Nejdříve se vymezí soubor prvků, na nichž budeme uvažovaný jev zkoumat. Poté se z hlediska studovaného jevu všechny prvky vyšetří a výsledky se prezentují ve formě číselných charakteristik, tabulek nebo grafů.
- **Inferenční (induktivní) statistika –** tzv. indukce se používá v případě, chceme-li zobecnit naše poznatky (např. přenosem závěrů z vybraného vzorku na celou populaci). Induktivní statistika se zabývá metodami, jak určité závěry zobecnit s udáním stupně jejich spolehlivosti (např. intervalu spolehlivosti). Provedeme tedy výpočet například průměru nebo jiné statistiky z výběru. Indukcí se pak snažíme odhadovat skutečnou centrální hodnotu v populaci s určitým stupněm spolehlivosti.

1.1 Získávání dat

Je třeba zmínit dva důležité pojmy, které hrají hlavní roli v metodách induktivní statistiky, tj. populace a výběr.

- **Populace (základní soubor)** - je soubor, který zahrnuje všechny teoreticky možné objekty našeho zájmu. Je dána přesným vymezením svých prvků a může mít konečný nebo nekonečný rozsah.
- **Výběr (výběrový soubor, vzorek)** - je podmnožina základního souboru, tedy vybrané prvky z populace.
- **Statistická jednotka** - elementární jednotka neboli prvek statistického pozorování, na kterém můžeme zkoumat konkrétní projev určitého hromadného jevu.
- **Statistický znak (zkoumaná náhodná veličina, proměnná) -** je to složka statistické jednotky, která může být měřena nebo pozorována. Statistické znaky jsou tedy vlastnosti, které sledujeme na prvcích výběru či populace. Hodnoty statistického znaku se vyjadřují v jednotkách měřící stupnice, tzv. **škály**. Rozlišujeme čtyři typy škál:

a) **Nominální škála** - k označování *nominálních znaků* (tj. znaků měřených na nominální škále) používáme názvy (jména), čísla a další symboly. Tato škála je složena ze dvou či více vzájemně se vylučujících kategorií, které nemohou být seřazeny. Příkladem může být *barva* s kategoriemi {*červená, modrá, žlutá, zelená*}.

b) **Ordinální škála** – hodnoty *ordinálních znaků* jsou seskupeny, podobně jako hodnoty nominálních znaků, do neslučitelných kategorií, které jsou ale vzájemně uspořádány. Lze tedy určit, která kategorie má menší hodnotu než jiná, ale nelze zjistit o kolik jednotek. Například *úroveň vzdělání* s kategoriemi {*základní, střední, vysokoškolské*}.

c) **Intervalová škála** – umožňuje stanovit vzdálenost mezi hodnotami měřené veličiny. Tato stupnice má definovanou jednotku měření, ale nemá jednoznačně stanovenou nulovou pozici. Počátek může být stanoven například jako 0˚C (není absolutní nula, tj. nejnižší dosažitelná teplota) nebo stupnice výšky tónu či libovolná kalendářní stupnice.

d) **Poměrová škála** – začíná skutečným nulovým bodem, tzn. zachovává nejen rozdíly (intervaly) mezi hodnotami, ale taky podíly těchto hodnot. Např. hmotnost, objem, vzdálenost či teplota ve stupních Kelvina (0˚K je nejnižší měřitelná teplota – absolutní nula, tj. nepřítomnost jakéhokoli množství tepla).

Statistické znaky také dělíme na kvalitativní a kvantitativní.

Kvalitativní znak – k jeho měření používáme nominální škálu.

Kvantitativní znak - znak měřený na ordinální, intervalové či poměrové škále.

- **Náhodná veličina** – je to veličina X (proměnná), jejíž hodnota x je jednoznačně určena výsledkem náhodného pokusu. Jedná se o libovolnou reálnou funkci X, která je zobrazením z množiny elementárních jevů Kolmogorovova pravděpodobnostního prostoru do prostoru Borelovsky měřitelných podmnožin reálné osy.

Hodnoty náhodné veličiny jsou v konkrétním souboru dat představovány hodnotami statistického znaku.

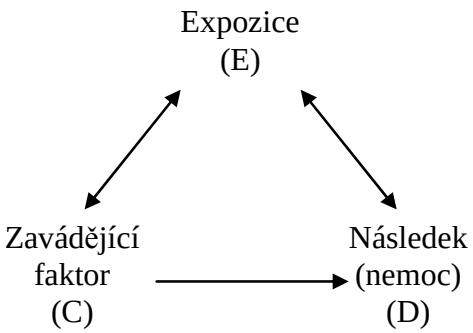
- **Náhodný výběr** – jedná se o výběr, při kterém se každá jednotka populace může s určitou pravděpodobností zahrnout do výběrového souboru, a skutečnost, zda bude či nebude vybrána, závisí čistě na náhodě. Náhodný výběr je náhodný vektor $X = (X_1, X_2,..., X_n)$, jehož složky jsou nezávislé náhodné veličiny, které mají stejné rozdělení pravděpodobností jako zkoumaná náhodná veličina X.

Ve vztahu k souboru dat jsou hodnoty náhodného výběru hodnotami statistického znaku (resp. znaků) měřených na jednotkách náhodně vybraných z populace.

Další pojem, se kterým se v textu setkáme, jsou tzv. zavádějící faktory (confounding factors).

- **Zavádějící faktor** – rušivý, matoucí faktor neboli třetí činitel. Při ignorování účinku zavádějícího faktoru může dojít k chybným odhadům velikosti účinku. Například, jak uvádí Marek Malý [3], konzumace alkoholu může mít za následek rakovinu plic, ale při popisu vztahu mezi konzumací alkoholu a rakovinou plic může být zavádějícím faktorem kouření cigaret (viz obrázek 1.1).

Obrázek 1.1: Působení zavádějícího faktoru



1.2 Základní statistické charakteristiky (míry)

Pro výpočty základních statistických charakteristik se v běžné popisné statistice pro označení používají hodnoty statistického znaku, které jsou reprezentované malými písmeny. V induktivní statistice se obdobné charakteristiky počítají pro náhodné výběry, u kterých jsou hodnoty statistického znaku nahrazeny hodnotami náhodných veličin příslušného náhodného výběru a označují se velkými písmeny.

1.2.1 Charakteristiky polohy

Jde o číselné hodnoty, pomocí nichž určujeme polohu míst, kolem kterých se data soustřeďují, tj. jejich „centrální hodnotu“.

- **Výběrový průměr** $\overline{X}_n$ - máme-li náhodný výběr rozsahu n, tj. X = (X_1, X_2,..., X_n), potom výběrový průměr $\overline{X}_n$ spočítáme následujícím způsobem:

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i . \qquad (1.2.1)$$

Aritmetický průměr $\bar{x}$ je analogií výběrového průměru v popisné statistice.

Příklad 1.2.1 Vypočtěte průměr následujících výsledků vyšetření: 20, 38, 49, 78, 80, 95.

Řešení: $\bar{x} = \frac{1}{6}(20 + 38 + 49 + 78 + 80 + 95) = 60.$

- **Modus** $\hat{x}$ - je hodnota znaku s největší četností. Pro diskrétní rozdělení pravděpodobnosti (znak nabývá nejvýše spočetně mnoha hodnot) je to hodnota znaku s nejvyšší pravděpodobností, tj. P($X = \hat{x}$) ≥ P($X = x_n$), n = 1,2,..., a pro spojitá rozdělení je to hodnota znaku, pro kterou platí f($\hat{x}$) ≥ f(x), -∞ < x < ∞, kde f označuje hustotu rozdělení pravděpodobnosti. Důležitý je pro kvalitativní, zejména nominální znaky. Má smysl tehdy, je-li počet pozorování vzájemně různých variant znaku X ve statistickém souboru podstatně menší než je rozsah n souboru.

Příklad 1.2.2 ([4], str. 81) Co je modus $\hat{x}$ v následujících výsledcích krevních skupin: A, 0, 0, B, B, AB, A, A, 0, 0, 0, AB, B, 0, B, A, 0, AB, 0, 0, B, 0, A?

Řešení: V tabulce 1.1 jsou přehledně zapsány výsledky.

Tabulka 1.1: Četnosti výskytu krevních skupin

krevní skupina	četnost výskytu
A	5
B	5
AB	3
0	10

Modus našich pozorování je tedy krevní skupina 0.

- **Kvantil** – nechť $\alpha \in (0, 1)$, α-kvantil náhodné veličiny X je takové reálné číslo x_α, pro které platí

$$P(X \le x_\alpha) \ge \alpha \quad \text{a současně} \quad P(X \ge x_\alpha) \ge 1 - \alpha. \qquad (1.2.2)$$

Kvantil se ve spojení s nějakým určitým rozdělením pravděpodobnosti používá často při testování hypotéz v kritériu, podle kterého se rozhodujeme, zda hypotézu zamítneme nebo ji nemůžeme zamítnout.

- **Medián $x_{0,5}$** - je nejčastěji užívaný kvantil s hodnotou α = 50%. Ve statistickém souboru je to prvek, který se po seřazení hodnot tohoto souboru (např. vzestupně) vyskytuje uprostřed (pro lichý počet pozorování prostřední hodnota, pro sudý počet pozorování průměr ze dvou prostředních pozorování). Není citlivý na odlehlé hodnoty a pro symetrické rozdělení se shoduje s průměrem.

  Příklad 1.2.3 Co je mediánem následujících výsledků šetření: 52, 13, 24, 79, 56, 83, 16?

  Řešení: Uspořádejme pozorování vzestupně: 13, 16, 24, 52, 56, 79, 83. Mediánem je 52.

Obecně pro daný p% kvantil určíme pořadové číslo jednotky n_p, pro které platí

$$n\frac{p}{100} < n_p < n\frac{p}{100} + 1, \tag{1.2.3}$$

kde n je počet prvků statistického souboru (výběru).

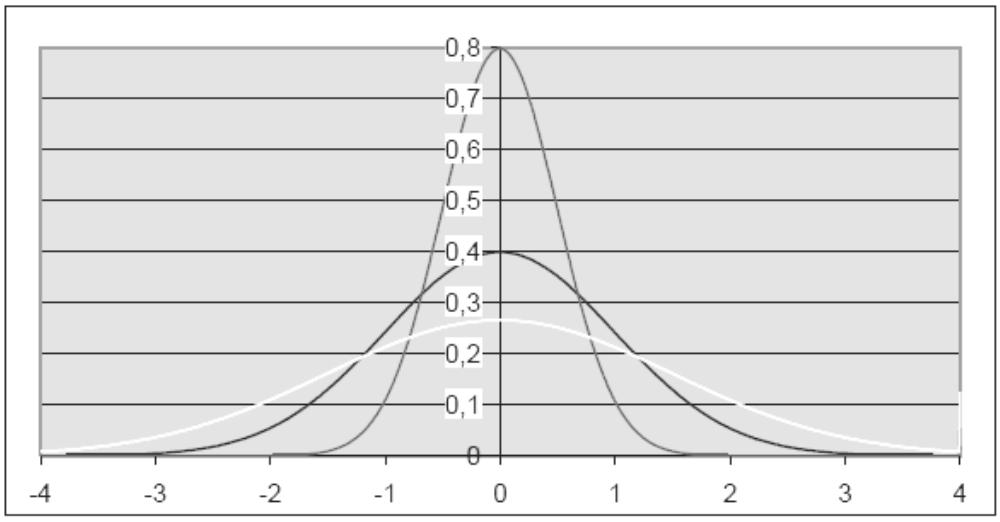
Pokud oddělujeme postupně hodnoty ve statistickém souboru po dvaceti pěti procentech, dostáváme tzv. **kvartily** (první kvartil je roven 25% kvantilu $x_{0,25}$, druhý 50% kvantilu, tj. mediánu a třetí kvartil je roven 75% kvantilu $x_{0,75}$). Jestliže hodnoty oddělujeme po deseti procentech, dostáváme **decily**, a podobně po jednom procentu, dostáváme **percentily**.

Pomocí kvartilů se počítá i **mezikvartilové rozpětí** (jde o míru variability) jako rozdíl mezi 3. kvartilem (75% kvantilem) a 1. kvartilem (25% kvantilem).

1.2.2 Charakteristiky variability (proměnlivosti, rozptýlení)

Výběr nelze přesně popsat jen pomocí charakteristik polohy, protože mnoho datových souborů má stejné nebo alespoň přibližně stejné hodnoty jednotlivých parametrů charakteristik polohy (průměr, modus, medián), ale už na první pohled se liší (jak je znázorněno na obrázku 1.2). Přestože mají data stejný průměr, modus, medián, liší se v soustředění hodnot kolem průměru. K vyjádření těchto odlišností nám slouží charakteristiky variability.

Obrázek 1.2: Odlišnost v soustředění hodnot kolem průměru (upraveno dle [5])



Mezi nejužívanější charakteristiky variability patří:

- **Variační rozpětí R** – je to rozdíl mezi největší a nejmenší hodnotou výběru (sledovaného znaku *X*), tedy

$$R = X_{\max} - X_{\min}. \tag{1.2.4}$$

Tato charakteristika není příliš spolehlivá, protože závisí na extrémních hodnotách, a také má nevýhodu v tom, že poskytuje pouze hrubý a předběžný odhad variability.

Příklad 1.2.4 ([4], str. 85) Sedm obyvatel malé obce A může mít stejný průměrný měsíční příjem jako sedm obyvatel malé obce B, ale rozdělení příjmů může být velmi odlišné, jak vidíme v tabulce 1.2.

Tabulka 1.2: Příjmy obyvatel v obcích A a B

obec A	obec B
4 000 Kč	8 000 Kč
6 000 Kč	8 000 Kč
8 000 Kč	9 000 Kč
10 000 Kč	10 000 Kč
12 000 Kč	11 000 Kč
14 000 Kč	12 000 Kč
16 000 Kč	12 000 Kč
$\hat{x}_A$ = 10 000 Kč	$\hat{x}_B$ = 10 000 Kč

V obci A je průměrný měsíční příjem 10 000 Kč, ale rozdíl mezi nejvyšší hodnotou (16 000 Kč) a nejnižší hodnotou (4 000 Kč) příjmu je 12 000 Kč. V obci B je také průměrný měsíční příjem 10 000 Kč, ale rozdíl je mnohem menší, pouze 4 000 Kč.

- **Výběrový rozptyl** S_n^2 – je základní mírou variability. Pomocí rozptylu měříme velikost čtverců odchylek jednotlivých hodnot výběru od průměru, tj.

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2. \tag{1.2.5}$$

Je důležité zmínit, že mnoho autorů statistické literatury uvádí název výběrový rozptyl pro S_n^2, přestože důsledně (v momentovém smyslu) je výběrovým rozptylem statistika

$$M_2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2. \tag{1.2.6}$$

Statistika S_n^2 se používá častěji, protože poskytuje lepší odhad (nestranný) skutečného rozptylu populace (σ^2), proto ji jako výběrový rozptyl uvádím zde také já.

- **Výběrová směrodatná odchylka S_n** – je definována jako kladná druhá odmocnina výběrového rozptylu, tedy

$$S_n = \sqrt{S_n^2}. \tag{1.2.7}$$

Fyzikálně je směrodatná odchylka vyjádřena ve stejných jednotkách jako měřené hodnoty, proto se uvádí v tabulkách sumarizujících naměřené hodnoty spolu s průměrem častěji než rozptyl.

1.3 Testování statistických hypotéz

Hypotéza znamená doslovně předpoklad či domněnku o pravděpodobnostním chování náhodné veličiny *X*. Úkolem teorie testování statistických hypotéz je konstrukce adekvátních matematických metod, pomocí nichž budeme posuzovat platnost či neplatnost zkoumané statistické hypotézy. Dříve než se budeme věnovat statistickým hypotézám, zmíním pojem **interval spolehlivosti** (konfidenční interval) pro náhodnou veličinu *X*. Je to interval, v němž se s pravděpodobností 1 - α realizace této náhodné veličiny nachází. Číslo 1 - α, $\alpha \in (0,1)$ se nazývá **spolehlivost odhadu** (koeficient spolehlivosti, konfidenční koeficient).

Pokud se hypotézy týkají hodnot parametrů rozdělení náhodné veličiny nebo parametrických funkcí, mluvíme o **parametrických hypotézách** a příslušné testy se rovněž nazývají parametrické (u těchto testů známe tvar rozdělení pravděpodobností

základního souboru). V ostatních případech mluvíme o **neparametrických hypotézách** a neparametrických testech (u nichž nemáme žádné informace o rozdělení pravděpodobností základního souboru). Dále dělíme hypotézy na jednoduché a složené. Pokud je hypotéza formulována tak, že jednoznačně určuje rozdělení náhodné veličiny, nazýváme ji **jednoduchou hypotézou**. Jestliže hypotéza rozdělení náhodné veličiny jednoznačně nespecifikuje, mluvíme o **složené hypotéze**. Dále se budeme zabývat pouze testováním pro případ jednoduché hypotézy.

Nechť rozdělení pravděpodobností náhodné veličiny X závisí na neznámém parametru θ, o kterém předem víme, že patří do parametrického prostoru $\Theta \subset R$ (tj. do množiny všech možných hodnot parametru).

Testovaná statistická hypotéza se nazývá **nulová hypotéza** (někdy také testovaná hypotéza) a značíme ji $\mathbf{H_0}$. Jestliže parametr základního souboru θ, odhadovaný na základě výběru, má hypotetickou hodnotu θ_0, lze zapsat nulovou hypotézu ve tvaru:

$$H_0: \theta = \theta_0 \qquad (1.3)$$

Nulová hypotéza tvrdí, že parametr základního souboru θ je roven hypotetické hodnotě θ_0. Proti hypotéze H_0 stavíme tzv. **alternativní hypotézu $\mathbf{H_1}$**, kterou přijímáme, jestliže jsme nulovou hypotézu H_0 zamítli jako nesprávnou. Alternativní hypotézu k hypotéze (1.3) bychom mohli vyjádřit ve formě:

1) oboustranné alternativy $H_1: \theta \neq \theta_0$,
2) jednostranné alternativy $H_1: \theta > \theta_0$ pravostranná alternativa,
   $H_1: \theta < \theta_0$ levostranná alternativa.

Své rozhodnutí o zamítnutí H_0 zakládáme na realizaci náhodného výběru $(X_1, X_2, ..., X_n)$ z rozdělení, které má náhodná veličina X, přesněji řečeno na realizaci určité statistiky $T = T(X_1, X_2, ..., X_n)$. Statistiku T nazýváme **testová statistika** nebo **testové kritérium**. Lze ji chápat jako míru nesouladu výsledků pokusu s nulovou hypotézou. Obor hodnot statistiky T (tj. výběrový prostor) rozdělíme na dvě disjunktní části – jednu z nich označíme W a budeme ji nazývat **kritickým oborem** nebo **oborem zamítnutí** pro test hypotézy H_0, druhou označíme V a nazveme oborem nezamítnutí (přijetí) pro test hypotézy H_0. V případě, že pro realizaci t statistiky T platí $T \in W$, nulovou hypotézu **zamítneme**, v opačném případě, kdy $T \in V$, tj. $T \notin W$, hypotézu H_0 **nelze zamítnout**. Při rozhodování o přijetí H_0 či H_1 provádíme testování na základě náhodného výběru, a proto se můžeme dopustit dvou chybných závěrů.

Následující tabulka znázorňuje, jakých chyb se při svém rozhodování o platnosti H_0 můžeme dopustit.

Tabulka 1.3: Chyby při rozhodování o H_0.

Rozhodnutí	**Skutečnost**	
	H_0 je pravdivá	H_0 není pravdivá
H_0 zamítneme	*chyba I. druhu*	*správné rozhodnutí*
H_0 nezamítneme	*správné rozhodnutí*	*chyba II. druhu*

Je přirozené požadovat, aby pravděpodobnosti obou těchto chyb byly co možná nejmenší.

Kritický obor W volíme tak, abychom omezili pravděpodobnost chyby I. druhu nějakým pevně zvoleným malým číslem α, kde α je z intervalu (0, 1) viz obrázky 1.3 a 1.4. V praxi se nejčastěji volí $\alpha = 0{,}05$ nebo $\alpha = 0{,}01$; jiné hodnoty se užívají řidčeji. Číslu α se říká **hladina významnosti testu** (pravděpodobnost chyby I. druhu), pro kterou platí

$$\alpha = P_{\theta_0}\left(T \in W \middle| H_0\right), \tag{1.4}$$

kde výraz na pravé straně označuje pravděpodobnostní míru, při níž statistika T patří do kritického oboru W za podmínky, že platí nulová hypotéza H_0 (statistika T má rozdělení s parametrem θ_0).

Pravděpodobnost, že do oboru přijetí nulové hypotézy H_0 padne hodnota testovacího kritéria, za podmínky, že platí H_1, tzn. pravděpodobnost chyby II. druhu, označované jako β, je

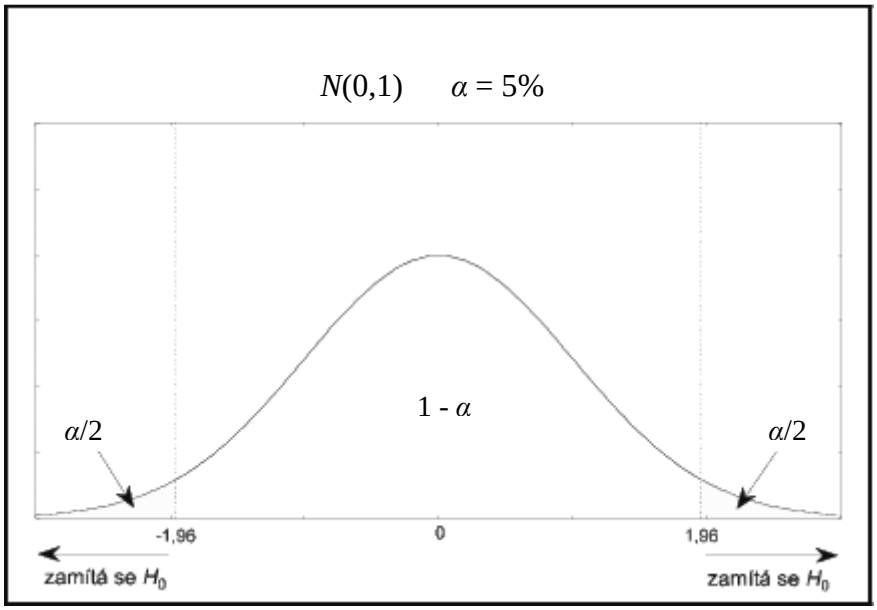
$$\beta = P_{\theta}\left(T \notin W \middle| H_1\right). \tag{1.5}$$

Doplněk pravděpodobnosti chyby II. druhu (β) do 1, tzn. 1 - β vyjadřuje pravděpodobnost opačného jevu, tj. správného zamítnutí testované hypotézy (schopnost testu odhalit neplatnost nulové hypotézy) a nazýváme ho **síla testu**.

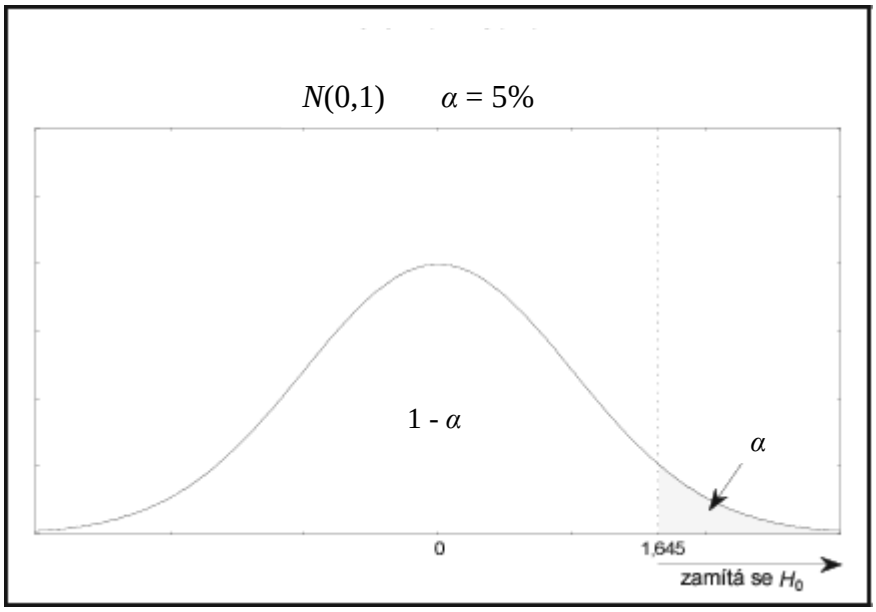
Při klasickém přístupu k testování hypotéz je nutné zvolit hladinu významnosti testu α ještě před pokusem, nezávisle na výsledku. Poté vypočteme na základě hodnot náhodného výběru hodnotu t testové statistiky T a určujeme, zda patří do kritického oboru W. Podle toho určíme, zda stanovenou nulovou hypotézu zamítáme nebo ji naopak nemůžeme zamítnout, tj. například, že rozdíl středních hodnot dvou nezávislých náhodných výběrů reprezentujících měření nějakého laboratorního parametru u kontrolní a léčené skupiny osob je statisticky významný či nikoliv. Alternativním postupem při rozhodnutí o platnosti či neplatnosti hypotézy je **dosažená hladina významnosti** testu neboli ***p*-hodnota** (anglicky *p*-value, significance value). Je to nejmenší hladina, při které

bychom ještě hypotézu H_0 zamítli (viz obrázek 1.5). *P*-hodnota tedy vyjadřuje nejmenší horní hranici pravděpodobnosti počítané za platnosti nulové hypotézy, že dostaneme právě naši realizaci *t* testové statistiky *T* nebo realizaci ještě více odporující nulové hypotéze. Hypotézu H_0 zamítáme na hladině α, právě když je *p*-hodnota menší než α, čím je tedy *p*-hodnota menší, tím méně důvěryhodná je nulová hypotéza.

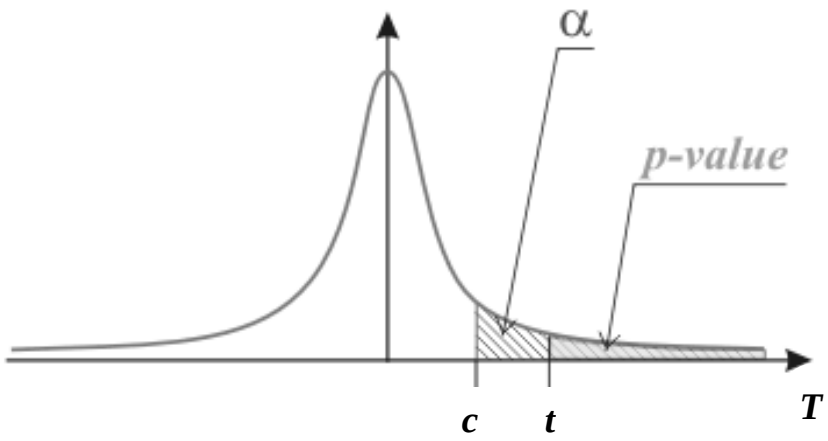
Obrázek 1.3: Kritický obor pro oboustranný test H_0 a hladinu významnosti 0,05 (upraveno dle [6]).



Obrázek 1.4: Kritický obor pro jednostranný test H_0 proti H_1: $\mu > \mu_0$ a hladinu významnosti 0,05 (upraveno dle [6]).



Obrázek 1.5: Dosažená hladina významnosti u jednostranného testu (upraveno dle [7]).*



* c = critical value, tj. (1-α) kvantil teoretického rozdělení testové statistiky,
t = hodnota testové statistiky T.

1.3.1 Obecný postup při testování statistických hypotéz

1) Formulace nulové hypotézy H_0 a alternativní hypotézy H_1, volba hladiny významnosti testu α.
2) Určení a výpočet testové statistiky $T = T(X_1, X_2,..., X_n)$.
3) Vymezení kritického oboru W pro test hypotézy H_0 proti alternativě na hladině testu α.
4) Rozhodnutí:
    - pokud $T \in W$, potom H_0 zamítneme ve prospěch alternativní hypotézy,
    - pokud $T \notin W$, potom H_0 nemůžeme zamítnout.

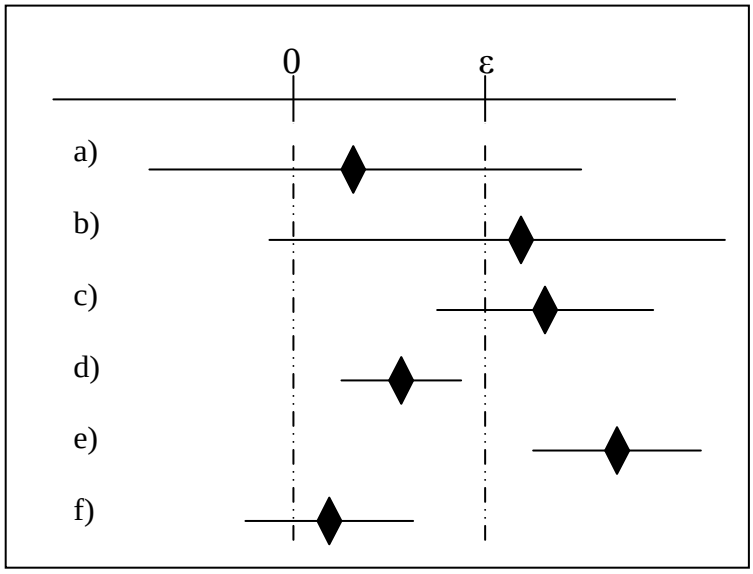
Alternativně můžeme zjistit *p*-hodnotu testu a pro nízkou *p*-hodnotu (je-li *p*-hodnota menší než předem stanovené α) zamítáme nulovou hypotézu ve prospěch alternativní hypotézy H_1 a pro vysokou *p*-hodnotu (je-li *p*-hodnota větší než předem stanovené α) nulovou hypotézu nelze zamítnout.

1.3.2 Rozdíl mezi statistickou a léčebnou (klinickou) významností

Vácha J. [8] definuje ve svém článku statisticky významný výsledek jako takový výsledek šetření (pozorování, pokusu), o kterém lze výpočtem zjistit, že nastává z náhodných příčin jen s jistou malou pravděpodobností. Obvykle se v praxi pokládá statisticky významný výsledek za skutečný efekt, a v důsledku toho i za výsledek klinicky důležitý, a naopak. Toto jednání, ale nemůžeme vždy považovat za správné. Zvárová [4] uvádí příklad, ve kterém při porovnávání krevního tlaku na levé a pravé ruce byl zjištěn průměrný rozdíl 1mm Hg. Zatímco je, díky velkému rozsahu výběru, tento rozdíl vysoce statisticky významný, není významný klinicky.

Mějme 95% interval spolehlivosti pro rozdíl středních hodnot. Předpokládejme, že překročením konstanty ε je rozdíl populačních průměrů klinicky významný. Pomocí obrázku 1.6 si ukážeme, jaké možnosti mohou nastat a na tabulce 1.4 si rozebereme jaký vliv má daný interval na statistickou a léčebnou významnost.

Obrázek 1.6: Intervaly spolehlivosti (upraveno dle [4])



Tabulka 1.4: Interpretace intervalů spolehlivosti [4]

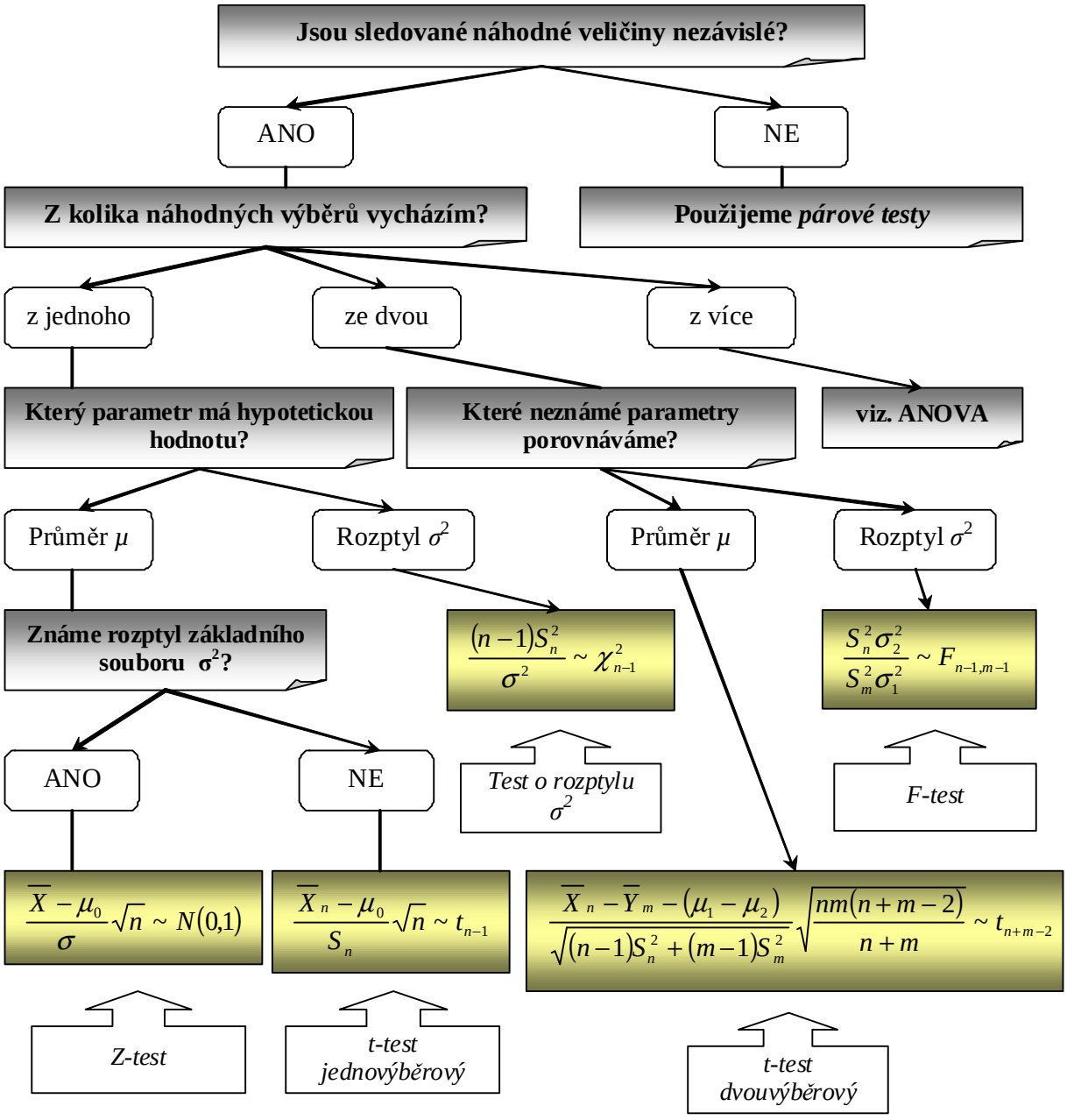
Varianta	Statistická významnost	Klinická významnost
a)	ne	možná
b)	ne	možná
c)	ano	možná
d)	ano	ne
e)	ano	ano
f)	ne	ne

Na obrázku vidíme, že např. v případě d) se interval spolehlivosti nachází mezi nulou a konstantou ε a zároveň konstantu ε nepřekračuje, tudíž rozdíl středních hodnot není klinicky významný, ale je statisticky významný. Narozdíl od případu e) kde se interval spolehlivosti nachází za konstantou ε, je tedy statisticky i klinicky významný.

1.3.3 Výběr správného testu

Předpokládejme, že chceme testovat parametrické hypotézy a zkoumané náhodné veličiny mají normální rozdělení. Následující obrázek 1.7 znázorňuje zjednodušený návod, jak při výběru parametrického testu postupovat a zvolit ten správný.

Obrázek 1.7: Výběr parametrického testu



1.3.4 Některé parametrické testy

Jak již bylo zmíněno na začátku kapitoly 1.3, parametrické testy jsou založeny na různých předpokladech. Jedním z nich je zpravidla to, že je určeno, z jakého rozdělení výběr pochází. Rozdělení pravděpodobnosti může být určeno úplně nebo je známo až na nějaké parametry. V této kapitole představím postupně několik nejčastěji používaných parametrických testů.

1.3.4.1 Parametrické testy jednovýběrové

Na základě jednoho výběrového souboru rozhodujeme, zda neznámý parametr základního souboru je nebo není roven určité předpokládané číselné hodnotě, či zda je (není) neznámý parametr větší (menší) než předpokládaná číselná hodnota.

Z-test

Mějme $(X_1, X_2, ..., X_n)$ náhodný výběr z rozdělení $X \sim N(\mu, \sigma^2)$ s neznámou střední hodnotou μ a známým rozptylem $\sigma^2 > 0$ a předem dané číslo μ_0. Na hladině významnosti α chceme testovat hypotézy znázorněné v tabulce 1.5. Test se provádí pomocí výběrové funkce:

$$T = \frac{\overline{X}_n - \mu_0}{\sigma}\sqrt{n}, \tag{1.3.4.1}$$

která má za platnosti $\mu = \mu_0$ rozdělení $N(0, 1)$, kde $\overline{X}_n$ je výběrový průměr, viz vztah (1.2.1).

Tabulka 1.5: Přehled kritických oborů a testovaných hypotéz

Nulová hypotéza	Alternativní hypotéza	Kritický obor
$H_0: \mu = \mu_0$	$H_1: \mu \neq \mu_0$	$W = \left(-\infty, -u_{1-\alpha/2}\right\rangle \cup \left\langle u_{1-\alpha/2}, \infty\right)$
$H_0: \mu \leq \mu_0$	$H_1: \mu > \mu_0$	$W = \left\langle u_{1-\alpha}, \infty\right)$
$H_0: \mu \geq \mu_0$	$H_1: \mu < \mu_0$	$W = (-\infty, -u_{1-\alpha}\rangle$

Pro zvolenou hladinu významnosti α určíme kvantily normovaného normálního rozdělení u_α, které najdeme ve statistických tabulkách.

Jednovýběrový t-test

Je-li $(X_1, X_2, ..., X_n)$ náhodný výběr z rozdělení $X \sim N(\mu, \sigma^2)$ s neznámou střední hodnotou μ a neznámým rozptylem σ^2 a μ_0 předem dané číslo, potom lze za testovou statistiku zvolit následující výběrovou funkci:

$$T = \frac{\overline{X}_n - \mu_0}{S_n}\sqrt{n}, \tag{1.3.4.2}$$

která má za platnosti $\mu = \mu_0$ rozdělení t_{n-1}, kde S_n^2 je výběrový rozptyl, viz vztah (1.2.5). Následující tabulka 1.6 znázorňuje testované hypotézy a odpovídající kritické obory.

Tabulka 1.6: Přehled kritických oborů a testovaných hypotéz

Nulová hypotéza	Alternativní hypotéza	Kritický obor
H_0: $\mu = \mu_0$	H_1: $\mu \neq \mu_0$	$W = \left(-\infty, -t_{n-1;1-\alpha/2}\right\rangle \cup \left\langle t_{n-1;1-\alpha/2}, \infty\right)$
H_0: $\mu \leq \mu_0$	H_1: $\mu > \mu_0$	$W = \left\langle t_{n-1;1-\alpha}, \infty\right)$
H_0: $\mu \geq \mu_0$	H_1: $\mu < \mu_0$	$W = (-\infty, -t_{n-1;1-\alpha}\rangle$

Při odvozování testového kritéria předpokládáme, že H_0 je správná, tj. do T budeme ve všech případech dosazovat $\mu = \mu_0$. Víme, že $\overline{X}$ je bodovým odhadem $\mu = E(X)$ a za platnosti H_0 má rozdělení $N(\mu_0, \sigma^2/n)$. Současně využijeme toho, že $(n - 1)S_n^2/\sigma^2$ má rozdělení χ^2_{n-1}. Do kritického oboru W pak patří takové hodnoty testového kriteria, které svědčí ve prospěch alternativy H_1. Ve statistických tabulkách najdeme pro zvolenou hladinu významnosti α a pro n - 1 stupňů volnosti kvantily Studentova rozdělení, pomocí nichž určíme kritický obor W (viz tabulka 1.6).

Mezi testy hypotéz o parametrech jednorozměrného normálního rozdělení patří také **test hypotézy o rozptylu σ^2**, kde testovým kritériem je následující výběrová funkce:

$$Z = \frac{(n-1)S_n^2}{\sigma^2}, \tag{1.3.4.3}$$

která má za platnosti $\sigma^2 = \sigma_0^2$ rozdělení χ^2_{n-1}. Tabulka 1.7 znázorňuje testované hypotézy a odpovídající kritické obory.

Tabulka 1.7: Přehled kritických oborů a testovaných hypotéz

Nulová hypotéza	Alternativní hypotéza	Kritický obor
H_0: $\sigma^2 = \sigma_0^2$	H_1: $\sigma^2 \neq \sigma_0^2$	$W = \left(0, \chi^2_{n-1;\alpha/2}\right\rangle \cup \left\langle \chi^2_{n-1;1-\alpha/2}, \infty\right)$
H_0: $\sigma^2 \leq \sigma_0^2$	H_1: $\sigma^2 > \sigma_0^2$	$W = \left\langle \chi^2_{n-1;1-\alpha}, \infty\right)$
H_0: $\sigma^2 \geq \sigma_0^2$	H_1: $\sigma^2 < \sigma_0^2$	$W = (0, \chi^2_{n-1;\alpha}\rangle$

V tabulce χ^2-rozdělení najdeme pro zvolenou hladinu významnosti α a n-1 stupňů volnosti kritické hodnoty, pomocí nichž určíme kritický obor.

1.3.4.2 Parametrické testy dvouvýběrové

Dvouvýběrové testy nám umožňují porovnávat neznámé hodnoty parametru mezi dvěma základními soubory.

1. Dvouvýběrový t-test

Nechť je (X_1, X_2,..., X_n) náhodný výběr z $N(\mu_1, \sigma^2)$ a (Y_1, Y_2,..., Y_m) náhodný výběr z $N(\mu_2, \sigma^2)$ a dále nechť tyto dva výběry jsou na sobě nezávislé.

V praxi ověříme předpoklad (nulovou hypotézu) o stejném rozptylu $\sigma_1^2 = \sigma_2^2 = \sigma^2$ pomocí tzv. F-testu, který uvedu dále.

Dvouvýběrový t-test se provádí pomocí výběrové funkce (viz [9], strana 91, Věta 21)

$$T = \frac{\overline{X}_n - \overline{Y}_m - (\mu_1 - \mu_2)}{\sqrt{(n-1)S_n^2 + (m-1)S_m^2}} \sqrt{\frac{nm(n+m-2)}{n+m}}, \qquad (1.3.4.4)$$

která má za platnosti nulové hypotézy rozdělení t_{n+m-2}.

Následující tabulka 1.8 znázorňuje testované hypotézy o středních hodnotách veličin X, Y a odpovídající kritické obory..

Tabulka 1.8: Přehled kritických oborů a testovaných hypotéz

Nulová hypotéza	Alternativní hypotéza	Kritický obor
H_0: $\mu_1 = \mu_2$	H_1: $\mu_1 \neq \mu_2$	$W = \left(-\infty, -t_{n+m-2;1-\alpha/2}\right\rangle \cup \left\langle t_{n+m-2;1-\alpha/2}, \infty\right)$
H_0: $\mu_1 \leq \mu_2$	H_1: $\mu_1 > \mu_2$	$W = \left\langle t_{n+m-2;1-\alpha}, \infty\right)$
H_0: $\mu_1 \geq \mu_2$	H_1: $\mu_1 < \mu_2$	$W = (-\infty, -t_{n+m-2;1-\alpha}\rangle$

Hodnotu testového kriteria T určujeme za předpokladu, že H_0 je správná, tj. dosazujeme $\mu_1 - \mu_2 = 0$. Obdobnou úvahou jako u jednovýběrového t-testu odvodíme kritické obory.

2. F-test shody rozptylů

Nechť je (X_1, X_2,..., X_n) náhodný výběr z $N(\mu_1, \sigma_1^2)$ a (Y_1, Y_2,..., Y_m) náhodný výběr z $N(\mu_2, \sigma_2^2)$ a dále nechť tyto dva výběry jsou na sobě nezávislé, $\sigma_1^2 \neq 0$, $\sigma_2^2 \neq 0$. Označme n, resp. m, rozsah náhodného výběru příslušného X, resp. Y a příslušné výběrové rozptyly symboly S_n^2 resp. S_m^2.

Test shody rozptylů provedeme pomocí výběrové funkce (viz [10] str.146, Věta 60)

$$V = \frac{S_n^2}{S_m^2}\frac{\sigma_2^2}{\sigma_1^2}, \tag{1.3.4.5}$$

která má rozdělení $F_{n\text{-}1,m\text{-}1}$.

Formulaci hypotéz a příslušné kritické obory znázorňuje tabulka 1.9.

Tabulka 1.9: Přehled kritických oborů a testovaných hypotéz

Nulová hypotéza	Alternativní hypotéza	Kritický obor
H_0: $\sigma_1^2 = \sigma_2^2$	H_1: $\sigma_1^2 \neq \sigma_2^2$	$W = \left(0, F_{n-1,m-1;\alpha/2}\right\rangle \cup \left\langle F_{n-1,m-1;1-\alpha/2}, \infty\right)$
H_0: $\sigma_1^2 \leq \sigma_2^2$	H_1: $\sigma_1^2 > \sigma_2^2$	$W = \left\langle F_{n-1,m-1;1-\alpha}, \infty\right)$
H_0: $\sigma_1^2 \geq \sigma_2^2$	H_1: $\sigma_1^2 < \sigma_2^2$	$W = (0, F_{n-1,m-1;\alpha}\rangle$

Označme $F_{\nu,\tau;\alpha}$ α-kvantil F-rozdělení o (ν, τ) stupních volnosti. Pomocí této hodnoty, kterou najdeme ve statistických tabulkách (např. [9], str. 327-329), určíme kritický obor. Předpokládáme, že je nulová hypotéza správná, tzn. při určování hodnoty testového kritéria dosazujeme $\frac{\sigma_2^2}{\sigma_1^2} = 1$.

V doposud zmíněných testech předpokládáme nezávislost zkoumaných náhodných veličin (jedná se o **nepárové testy**), někdy ale tento předpoklad není splněn. Jestliže jsou tedy zkoumané náhodné veličiny na sobě závislé použijeme pro testování **test párový**. Například, jestliže sledujeme účinnost nějakého léku na jedné skupině pacientů, je naše pozorování tvořeno dvojicí údajů, které byly naměřeny na stejných osobách (stavy pacientů před léčbou a po léčbě). Pro otestování můžeme použít **párový t-test**.

Označme *Y*, *Z* sledované statistické znaky. Ve zmíněné situaci je logické uvažovat jejich rozdíl $X = Y - Z$. Je-li možno předpokládat, že $X \sim N(\mu, \sigma^2)$, lze testovat například hypotézu

$$H_0: \mu = EY - EZ = 0 \quad \text{proti} \quad H_1: \mu \neq 0, \tag{1.3.4.6}$$

tj., že podání léku nemělo vliv na střední hodnotu sledovaného znaku (stav pacienta).

Uvedenou hypotézu ověříme jednovýběrovým t-testem viz vztah (1.3.4.2), který aplikujeme na rozdíly párových hodnot.

1.3.4.3 Parametrické testy *k* - výběrové

1. Testy o shodě středních hodnot (průměrů)

Nyní mějme k dispozici $k > 2$ nezávislých náhodných výběrů z normálních rozdělení se stejným rozptylem, tj.

$$Y_{11}, Y_{12},..., Y_{1n_1} \sim N(\mu_1, \sigma^2),$$
$$Y_{21}, Y_{22},..., Y_{2n_2} \sim N(\mu_2, \sigma^2),$$
$$...$$
$$Y_{k1}, Y_{k2},..., Y_{kn_k} \sim N(\mu_k, \sigma^2).$$

Na hladině významnosti α chceme testovat nulovou hypotézu, která má tvar H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ proti alternativě, která tvrdí, že alespoň dvě střední hodnoty se od sebe liší. K ověřování středních hodnot na základě více než dvou výběrů je při statistických analýzách nejpoužívanější metodou tzv. **ANOVA** (analysis of variance, analýza rozptylu), kterou vytvořil ve 30. letech 20. století R. A. Fisher.

Výše popsaný model se označuje jako **model analýzy rozptylu jednoduchého třídění**, kdy se předpokládá, že zkoumaný kvantitativní statistický znak je ovlivňován pouze jediným faktorem A, který se sleduje na několika jeho úrovních a je uspořádaný do tolika skupin (tříd), na kolika úrovních tento faktor sledujeme. V tabulce 1.10 si ukážeme princip třídění pozorovaných hodnot.

Tabulka 1.10: Princip třídění hodnot

Skupina	Pozorované hodnoty	Rozsah	Součet hodnot	Průměr
1	$Y_{11}, Y_{12},..., Y_{1n_1}$	n_1	$Y_{1.}$	$\overline{Y}_{1.}$
2	$Y_{21}, Y_{22},..., Y_{2n_2}$	n_2	$Y_{2.}$	$\overline{Y}_{2.}$
…	...	...	...	...
i	$Y_{i1}, Y_{i2},..., Y_{in_i}$	n_i	$Y_{i.}$	$\overline{Y}_{i.}$
…	...	...	...	...
k	$Y_{k1}, Y_{k2},..., Y_{kn_k}$	n_k	$Y_{k.}$	$\overline{Y}_{k.}$

Pro rozsahy výběrů platí, že jejich součet je roven n, tj. $\sum_{i=1}^{k} n_i = n$.

V případě jednoduchého třídění máme vlastně následující matematický model:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2), \quad i = 1,...,k, \quad j = 1,...,n_i, \tag{1.3.4.7}$$

kde Y_{ij} je *j*-té pozorování z *i*-té skupiny, α_i je efekt (účinek) faktoru A v *i*-tém výběru, μ je společná střední hodnota a ε_{ij} jsou nezávislé náhodné veličiny vyjadřující blíže nespecifikované náhodné chyby, jimiž je každé měření zatíženo. Za předpokladu platnosti nulové hypotézy, kterou lze přepsat ve tvaru H_0: $\alpha_1 = \alpha_2 = ... = \alpha_k = 0$ se model změní na

$$Y_{ij} = \mu + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2), \quad i = 1,...,k, \quad j = 1,...,n_i. \tag{1.3.4.8}$$

Výpočty pro analýzu rozptylu uspořádáme do tzv. tabulky analýzy rozptylu (viz tabulka 1.11)

Tabulka 1.11: Tabulka analýzy rozptylu

Zdroj variability	Součet čtverců	Stupně volnosti	Podíl	Testovací kritérium
skupiny	$S_A = S_T - S_e$	$f_A = k - 1$	S_A / f_A	$F_A = \dfrac{\frac{S_T - S_e}{k-1}}{\frac{S_e}{n-k}} = \dfrac{S_A}{S_e}\dfrac{n-k}{k-1}$
reziduální	$S_e = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_{i.}\right)^2$	$f_e = n - k$	S_e / f_e	
celkový	$S_T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_{..}\right)^2$	$f_T = n - 1$		

Odhadem středních hodnot μ_i jsou průměry

$$\bar{Y}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}, \tag{1.3.4.9}$$

a odhadem společné střední hodnoty je celkový průměr

$$\bar{Y}_{..} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij} = \frac{1}{n}\sum_{i=1}^{k} n_i \bar{Y}_{i.} \tag{1.3.4.10}$$

Za platnosti H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ má testová statistika F_A rozdělení $F_{k-1,n-k}$. Nulovou hypotézu tedy zamítneme na hladině α, jestliže $F_A \geq F_{k-1,n-k;1-\alpha}$.

V případě, že je sledovaný statistický znak ovlivněn dvěma faktory, hovoříme o tzv. **analýze rozptylu dvojného třídění**, u více jak dvou faktorů se jedná o **vícefaktorový model analýzy rozptylu** (postupy zpracování pro tyto případy můžete najít např. v publikaci J. Anděla [9], str. 157-175)

Položme si nyní otázku, proč se používá test ANOVA a ne několikeré porovnávání dvojic náhodných výběrů. Jaká je tedy pravděpodobnost chyby prvního druhu (tj.

zamítáme-li nulovou hypotézu, když ve skutečnosti platí) pro případ testu rovnosti středních hodnot více než dvou náhodných výběrů?

Jak jsme si již uvedli, máme k dispozici k nezávislých náhodných výběrů Y_{i1}, Y_{i2},..., Y_{in_i} z normálního rozdělení se střední hodnotou μ_i a stejným rozptylem σ^2, i=1,..., k a testujeme hypotézu H_0: $\mu_1 = \mu_2 = ... = \mu_k$.

Abychom dosáhli hladiny významnosti α je správné použít test ANOVA. Pokud bychom použili dvojice dvouvýběrových testů, pak bychom museli provést celkem k (k - 1)/2 testů, každý s hladinou významnosti (pravděpodobností chyby I. druhu) α:

$\Rightarrow$ hypotézu o rovnosti středních hodnot všech k výběrů v tomto případě zamítáme, jestliže zamítáme nulovou hypotézu o rovnosti středních hodnot alespoň v jednom dvouvýběrovém testu hypotézy H_0: $\mu_i = \mu_j$, $i \neq j$,

$\Rightarrow$ hladinu významnosti α tedy spočítáme jako šanci, že alespoň v jednom dvouvýběrovém testu uděláme chybu I. druhu, tj. 1 - pravděpodobnost opačného jevu (pravděpodobnost, že hypotéza platí a my ji v žádném dvouvýběrovém testu nezamítneme): $1 - (1 - \alpha)^{k(k-1)/2}$. V tabulce 1.12 vidíme, že s rostoucím počtem výběrů roste i pravděpodobnost chyby I. druhu.

Tabulka 1.12: Hladina významnosti α u násobných porovnání

počet výběrů (k)	α = 0.05	α = 0.01
2	0,05	0,01
3	0,142625	0,029701
4	0,264908	0,05852
5	0,401263	0,095618
10	0,90056	0,363815
∞	1	1

2. Mnohonásobná porovnávání

Pokud zamítneme na hladině α hypotézu o rovnosti středních hodnot, potom nám analýza rozptylu pouze říká, že střední hodnoty nejsou stejné. Je třeba na téže hladině provést další analýzu, abychom zjistili, které výběry se od sebe svými středními hodnotami liší. Rozhodujeme tedy o platnosti H_0: $\mu_i = \mu_j$ proti alternativě H_1: $\mu_i \neq \mu_j$ pro i-tý j-tý výběr (i, j = 1,…, k, $i \neq j$).

Pro řešení tohoto problému mnohonásobného porovnávání uvedu dvě nejužívanější metody - Tukeyho metodu a Scheffého metodu.

Tukeyho metoda

Tukeyho metoda se používá v případě tzv. **vyváženého třídění**, což znamená, že všechny výběry mají stejný rozsah $p = n_1 = \ldots = n_k$. Hypotézu o rovnosti středních hodnot μ_i a μ_j zamítneme na hladině významnosti α, jestliže platí

$$\left|\overline{Y}_{i.} - \overline{Y}_{j.}\right| > q_{k,n-k;1-\alpha}\frac{S}{\sqrt{p}}, \qquad (1.3.4.11)$$

přičemž hodnoty $q_{k,n-k;1-\alpha}$ jsou kvantily tzv. studentizovaného rozpětí, které najdeme ve statistických tabulkách (např. [5] str.131, 132).

Scheffého metoda

Scheffého metoda umožňuje porovnávat rozdíly mezi středními hodnotami jak u vyváženého, tak u nevyváženého třídění (tj. výběry mají různý rozsah). Hypotézu o rovnosti středních hodnot μ_i a μ_j zamítneme na hladině významnosti α, pokud platí

$$\left|\overline{Y}_{i.} - \overline{Y}_{j.}\right| > \sqrt{(k-1)S^2\left(n_i^{-1} + n_j^{-1}\right)F_{k-1,n-k;1-\alpha}}, \qquad (1.3.4.12)$$

kde $s^2 = \frac{S_e}{n-k}$ je nestranným odhadem parametru σ^2 a veličina $F_{k-1,n-k;1-\alpha}$ je kvantil F-rozdělení.

3. Testy rovnosti rozptylů

Jelikož se jedná o jeden z předpokladů užití metody ANOVA testujeme hypotézu, že všech k výběrů pochází z normálně rozdělených základních souborů majících stejný rozptyl σ^2, tzn. H_0: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$, proti alternativě, která tvrdí, že alespoň jedna rovnost neplatí. Pro řešení tohoto problému se používá **Bartlettův test** ([9], str. 155) nebo také **Leveneův test**.

1.3.5 Některé neparametrické testy

Neparametrické metody byly vypracovány pro testování výběrů poměrně malých rozsahů, které pocházejí z výrazně nenormálních základních souborů. Tyto metody obecně nepředpokládají konkrétní typ rozdělení, ale mohou klást určité požadavky (např. že distribuční funkce základního souboru je spojitá).

Neparametrické metody jsou prováděny na základě znalosti pořadí hodnot. Mějme dána různá reálná čísla $x_1,\ldots, x_N$. **Pořadím R_i** čísla x_i nazýváme počet těch čísel $x_1,\ldots, x_N$, která jsou menší nebo rovna číslu x_i. Tabulka 1.13 znázorňuje přiřazení pořadí vzájemně různým hodnotám.

Tabulka 1.13: Přiřazení pořadí vzájemně různým hodnotám

Vzestupně upořádané hodnoty x_i	-3	-1	0	4	12
Pořadí R_i	1	2	3	4	5

Někdy se ale stane, že hodnoty x_i nejsou různé, ale některé z nich jsou si rovny, potom vytvářejí tzv. **shody**. V tom případě se shodným hodnotám přiřazuje průměrné pořadí odpovídající takové skupině.

1.3.5.1 Neparametrické testy (pořadové)

Neparametrické testy, které pracují s pořadím hodnot náhodné veličiny v náhodném výběru, jsou tzv. **pořadové neparametrické testy**. Mezi nejznámější patří tzv. *jednovýběrový a dvouvýběrový Wilcoxonův test* (jsou neparametrickou obdobou jednovýběrového a dvouvýběrového t-testu) a *Kruskalův-Wallisův test*.

1. Jednovýběrový Wilcoxonův test

Nechť $X_1,\ldots, X_n$ je náhodný výběr ze spojitého rozdělení s distribuční funkcí $F(x)$. Chceme testovat nulovou hypotézu, že F je symetrická kolem nuly v tom smyslu, že platí

$$F(x) = 1 - F(-x), \qquad -\infty < x < \infty. \tag{1.3.5.1}$$

Seřaďme $X_1,\ldots, X_n$ do rostoucí posloupnosti podle jejich absolutní hodnoty, tj.

$$|X|^{(1)} < |X|^{(2)} < \ldots < |X|^{(n)}. \tag{1.3.5.2}$$

Předpokládejme, že R_i^+ je pořadí X_i při uspořádání (1.3.5.2). Nyní zaveďme veličiny

$$S^+ = \sum_{X_i \geq 0} R_i^+, \; S^- = \sum_{X_i < 0} R_i^+. \tag{1.3.5.3}$$

Přitom platí $S^+ + S^- = n(n+1)/2$, což můžeme užít pro kontrolu. Hypotéza se zamítá v případě, že je číslo $\min(S^+, S^-)$ menší nebo rovno tabelované kritické hodnotě, kterou nalezneme ve statistických tabulkách (např. [9], str. 334).

V případě velkého rozsahu n (uvádí se, že postačí již $n \geq 20$) lze použít toho, že za platnosti nulové hypotézy má veličina

$$U = \frac{S^+ - \frac{1}{4}n(n+1)}{\sqrt{\frac{1}{24}n(n+1)(2n+1)}} \qquad (1.3.5.4)$$

asymptoticky rozdělení $N(0, 1)$. V případě, že $|U| \geq u_{\alpha/2}$ nulovou hypotézu na hladině významnosti α zamítáme (podmínka platí u oboustranného testu).

2. Dvouvýběrový Wilcoxonův test

Nechť $X_1,\ldots, X_m$ *a* $Y_1,\ldots, Y_n$ jsou dva nezávislé náhodné výběry ze dvou spojitých rozdělení. Chceme testovat nulovou hypotézu, že distribuční funkce obou rozdělení jsou totožné.

Všech $m + n$ výběrových hodnot $X_1,\ldots, X_m, Y_1,\ldots, Y_n$ uspořádáme vzestupně podle velikosti. Zjistíme součet pořadí hodnot $X_1,\ldots, X_m$ a označíme ho T_1. Obdobně T_2 je součet pořadí hodnot $Y_1,\ldots, Y_n$. Vypočteme

$$U_1 = mn + \frac{m(m+1)}{2} - T_1, \; U_2 = mn + \frac{n(n+1)}{2} - T_2. \qquad (1.3.5.5)$$

Přitom platí $U_1 + U_2 = mn$. Pokud $\min(U_1, U_2)$ je menší nebo rovno kritické hodnotě uvedené ve statistických tabulkách (např. [9], str.335, 336, v těchto tabulkách je pro stručnost psáno n místo $\min(m, n)$ a m místo $\max(m, n)$). V případě, že jsou hodnoty m a n velké, vypočteme veličinu

$$U_0 = \frac{U_1 - \frac{1}{2}mn}{\sqrt{\frac{mn}{12}(m+n+1)}}, \qquad (1.3.5.6)$$

která má za platnosti nulové hypotézy asymptoticky rozdělení $N(0, 1)$. V případě, že $|U_0| \geq u_{1-\alpha/2}$ nulovou hypotézu zamítneme na hladině asymptoticky rovné α. Tím je popsán oboustranný test.

3. Kruskalův - Wallisův test

Kruskalův-Wallisův test je neparametrickou obdobou analýzy rozptylu jednoduchého třídění. Tento test slouží k ověření nulové hypotézy, že $k > 2$ nezávislých náhodných výběrů o rozsazích $n_1, n_2,\ldots, n_k$ pochází z jednoho základního souboru (tj. ze stejného rozdělení). Označme $n = n_1 + \ldots + n_k$. Předpokládejme, že každý tento výběr pochází z nějakého rozdělení se spojitou distribuční funkcí $F_i(x)$, $i = 1, 2,\ldots, k$. Chceme testovat hypotézu, že všechny výběry pocházejí ze stejného rozdělení, tedy

$$H_0: F_1(x) = F_2(x) = \ldots = F_k(x) \text{ pro } \forall x. \qquad (1.3.5.7)$$

Všech n prvků z k výběrů se seřadí do rostoucí posloupnosti a určí se pořadí každého prvku. Součet pořadí těch prvků, které patří do i-tého výběru ($i = 1, 2,\ldots, k$) označíme T_i, protože musí platit $T_1 + ...+ T_k = n(n + 1)/2$. Tento vztah můžeme použít pro kontrolu, zda jsou T_i vypočtena správně. Za platnosti nulové hypotézy má pak veličina

$$Q = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i} - 3(n+1) \qquad (1.3.5.8)$$

při $n_i \to \infty$ asymptoticky χ^2 rozdělení o $k - 1$ stupních volnosti. Nulovou hypotézu na hladině významnosti α zamítáme, je-li $Q \geq \chi^2_{k-1;\alpha}$.

V případě, že všech k výběrů má stejný rozsah, tzn. platí-li $n_1 = n_2 = \ldots = n_k = N$, můžeme Kruskalův - Wallisův test doplnit **Neményiho metodou mnohonásobného srovnávání** ([9], str. 231).

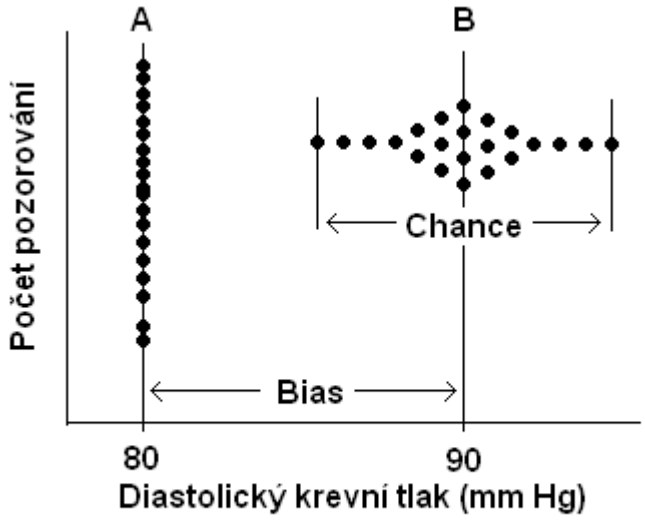
1.3.5.2 Testy dobré shody

Testy dobré shody umožňují srovnávat empirické (výběrové) rozdělení s jistým rozdělením teoretickým. K nejběžnějším testům dobré shody patří zejména **χ^2 – test dobré shody** a **Kolmogorov-Smirnovův test** ([5], str.120-123).

1.4 Techniky k zamezení vychýlení

Vychýlení (systematická chyba, bias) je výraz, který se v rámci klinického hodnocení, a ve statistice vůbec, vyskytuje velmi často. Vychýlení je cokoliv, co chybným způsobem ovlivňuje závěry a zkresluje porovnání léčebných skupin (ramen) klinického hodnocení. Může také způsobit úplné znehodnocení výsledků studie. Tato systematická chyba vzniká při sběru dat, jejich kontrole, analýze, interpretaci či publikaci a vede k závěrům systematicky se lišícím od skutečnosti. Proto je nutné eliminovat odstranitelné zdroje vychýlení, identifikovat všechny potenciální zdroje a ukázat, že jejich vliv nehraje důležitou roli. Existují různé techniky, pomocí kterých se dá předcházet závažným vychýlením. Mezi nejdůležitější patří randomizace a zaslepení. V publikaci *Clinical epidemiology and biostatistics* [11] je uveden příklad, na kterém lze ilustrovat vztah mezi biasem a náhodností (chance), viz obrázek 1.8, kdy byl měřen diastolický krevní tlak dvěma metodami: metodou (**A**), tj. intra-arteriální kanylou (přesná a objektivní metoda, poskytující nevychýlené výsledky) a metodou (**B**) sphygmomanometrem (tonometr – elektronický nebo mechanický přístroj na měření krevního tlaku s manžetou).

Obrázek 1.8: Vztah mezi biasem a náhodností



Na obrázku vidíme, že hodnoty naměřené metodou (**B**) jsou systematicky posunuty (bias) vlivem špatné velikosti manžety nebo sluchovým deficitem vyšetřující osoby, a dále jsou rovnoměrně rozptýleny vlevo i vpravo kolem předpokládané hodnoty (náhodná variabilita).

1.4.1 Zaslepení

Zaslepení (blinding nebo masking, oba termíny mají totožný význam) zajišťuje, aby byly po zahájení léčby odstraněny nežádoucí vlivy, které by působily různě v různých léčebných skupinách, narozdíl od randomizace, která dbá na to, aby před zahájením léčby mezi léčebnými skupinami nebyly žádné významné rozdíly. Cílem zaslepení je utajit znalost typu léčby.

V klinickém hodnocení hovoříme o zaslepení tehdy, je-li cílem experimentu srovnání dvou nebo více léčivých přípravků nebo aktivní látky a placeba, a kdy v zájmu objektivního zhodnocení jejich účinnosti a bezpečnosti je subjekt hodnocení (nebo i další účastníci studie) neschopen rozlišit, který ze srovnávaných přípravků je aplikován. Je tedy důležité dbát na to, aby podávané placebo bylo od zkoušeného léčiva nerozlišitelné co do vzhledu, způsobu aplikace, chuti, velikosti, zápachu či hmotnosti.

Existuje několik způsobů zaslepení:

- **Jednoduché zaslepení** (single blinding) - hovoříme o něm tehdy, je-li zaslepen studijní subjekt, a to buď subjekt hodnocení (pacient) nebo investigátor (zkoušející lékař).
- **Dvojité zaslepení** (double blinding) - je široce užívaný termín a používá se pro případ zaslepení, jak subjektu hodnocení, tak i investigátora.
- **Trojité zaslepení** (triple blinding) - používá se v případě, že je kromě subjektů hodnocení a investigátorů zaslepen rovněž personál spravující data studie.
- **Čtyřnásobné zaslepení** (quadruple blinding) – se používá, pokud jsou zaslepeni všichni členové týmu.

Zaslepení považujeme za zásadní prvek v designu klinických studií zvyšující objektivnost a validitu získaných dat, v některých případech je však samozřejmě nerealizovatelné (např. srovnání chirurgické léčby s léčbou jen medikamentózní).

1.4.2 Randomizace

Randomizace je proces, který se na základě předem určeného plánu, provádí formou náhodného (nebo pseudonáhodného) rozdělování subjektů hodnocení do dvou nebo více léčebných skupin srovnávaných v rámci klinického hodnocení. Zlatým standardem je dnes použití randomizace především při provádění srovnávacích klinických studií Fáze III (popis jednotlivých fází klinického hodnocení viz [12], [13]), tedy v projektech bezprostředně předcházejících registraci nového léčivého přípravku. Hlavním cílem randomizace je zamezit subjektivnímu a selektivnímu zařazování subjektů hodnocení do jednotlivých léčebných skupin (ramen studie). Dále také zajišťuje požadovaný počet subjektů ve srovnávaných ramenech, umožňuje rovnoměrné rozložení prognosticky významných faktorů (souběžná léčba, onemocnění atd.) i zavádějících faktorů (confounders).

Jak je uvedeno v publikaci [12] můžeme nejčastěji používané randomizační techniky podle vhodnosti jejich použití rozdělit na skupinu technik nepřípustných, méně vhodných a doporučených.

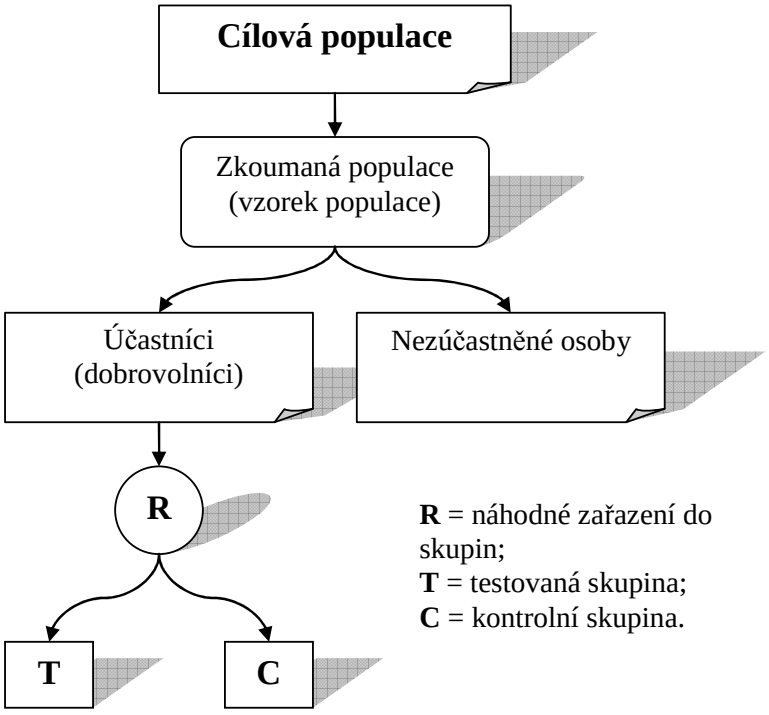
- **Nepřípustné randomizační techniky -** často používané v minulosti, patří mezi ně například randomizace podle pořadového čísla vstupu do studie, podle iniciál pacienta nebo také randomizace na základě data narození nebo data vstupu do studie. Tyto techniky jsou nevhodné především proto, že v případě nezaslepení pacienta ani zkoušejícího lékaře může experimentátor ovlivnit zařazení určitého subjektu do konkrétní léčebné skupiny, protože ještě před podpisem informovaného souhlasu ví, do jaké léčebné skupiny by byl pacient zařazen, a proto na základě této znalosti nemusí pacientovi účast v daném experimentu nabídnout. V případě zaslepených studií je nevýhodou těchto technik, že neumožňují vytvořit skupiny pacientů s homogenně rozloženými prognostickými faktory ani požadovaný poměr počtu subjektů v léčebných skupinách.
- **Méně vhodné randomizační techniky** – patří sem tzv. *kompletní randomizace*, kdy se rozhoduje o přiřazení pacienta do konkrétního ramene studie pouze náhodně (např. na základě hodu mincí). Tato technika se v současnosti příliš nepoužívá a to hlavně z důvodu rizika nevyváženého počtu pacientů v jednotlivých ramenech a také není zajištěna kontrola distribuce prognostických faktorů mezi rameny
- **Doporučované randomizační techniky –** tyto techniky se používají v současnosti, patří sem především *stratifikovaná permutační bloková randomizace*, která lépe vyhovuje požadavku nutnosti zajištění kontroly distribuce prognostických faktorů

v jednotlivých léčebných skupinách. Tyto faktory mohou u konkrétního subjektu zásadně ovlivnit účinnost a bezpečnost studijní léčby (např. věk, pohlaví, stadium onemocnění, amnestická data, souběžná léčba aj.). Pro zajištění srovnatelnosti porovnávaných ramen klinického hodnocení je proto zásadní možnost kontroly rovnoměrného rozložení prognostických faktorů. U stratifikované randomizace jsou popsány všechny teoreticky možné kombinace těchto faktorů u konkrétních jedinců, pro tyto kombinace se používá termín *strata* (tedy jedno stratum mohou představovat např. ženy mladší 50 let s klinickým stadiem I). Randomizace se poté provede v rámci všech těchto strat, což zajišťuje, že ve všech léčebných skupinách jsou rovnoměrně zastoupeni pacienti s časným a pokročilým stadiem, že mezi nimi budou srovnatelně zastoupeny všechny věkové skupiny a zároveň mezi nimi bude stejný podíl mužů a žen.

1.5 Uspořádání (design) klinického hodnocení

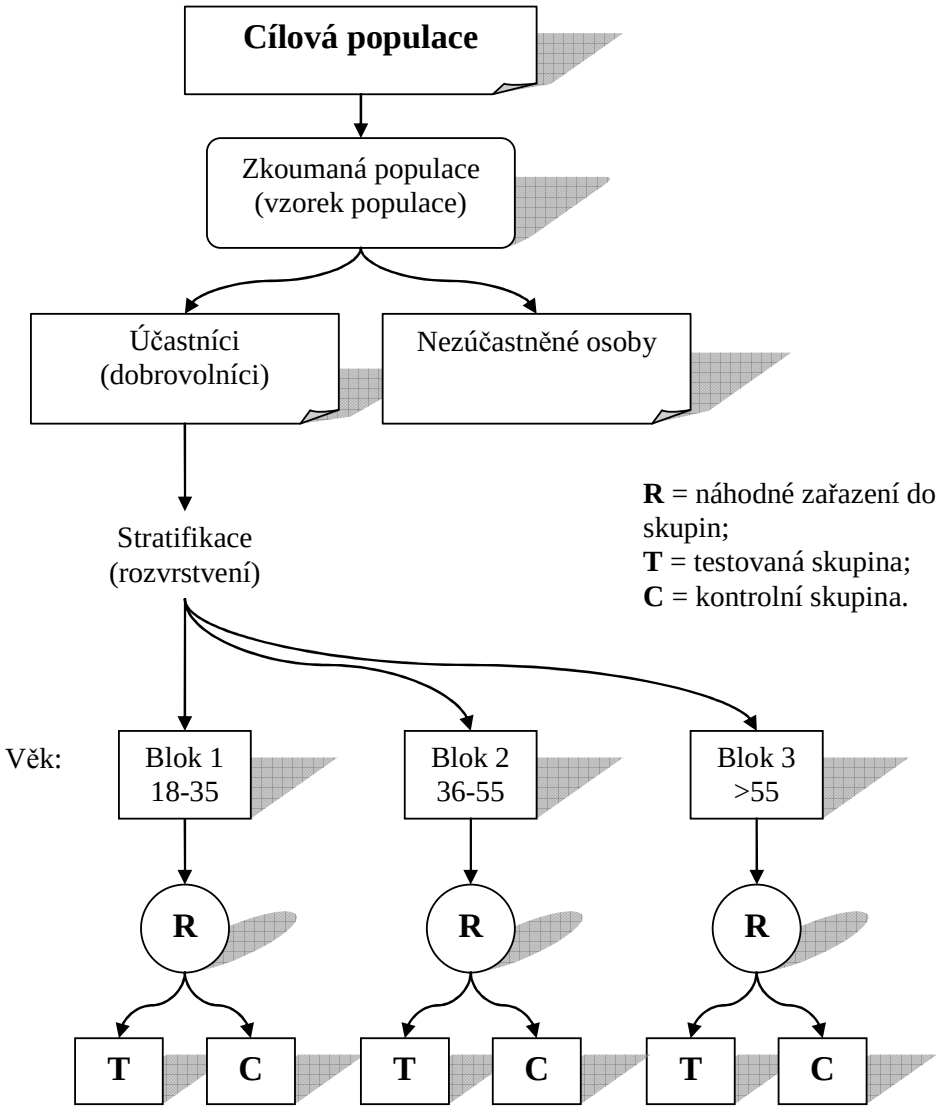
Následující obrázek 1.9 znázorňuje **kompletně náhodný návrh studie** (Completely random design), ve kterém se nekladou žádná omezení na randomizaci zařazení subjektů do skupin.

Obrázek 1.9: Kompletně náhodný návrh studie (upraveno dle [11])



Obrázek 1.10 popisuje **randomizovaný blokový návrh studie** (Randomized block design), ve kterém jsou subjekty nejdříve stratifikovány do bloků podle struktury zavádějících faktorů (např. věku nebo závažnosti onemocnění) nebo výchozí hodnoty závisle proměnné a poté jsou náhodně rozděleny na testovanou a kontrolní skupinu.

Obrázek 1.10: Randomizovaný blokový návrh studie (upraveno dle [11])



Uspořádání klinického hodnocení:

- **Paralelní uspořádání -** subjekty hodnocení jsou randomizovaně zařazeni do jednotlivých léčebných skupin s odlišným způsobem léčby, která se v průběhu celého klinického hodnocení nemění. Léčebné skupiny jsou nejméně dvě, v tomto případě je jedna testovaná (např. skupina léčená aktivní léčbou, kdy se zkoumá léčivý přípravek) a

druhá je kontrolní (skupina placebová). V případě existence více léčebných skupin může být příkladem klinické hodnocení s několika skupinami léčenými různými dávkami stejného léčivého přípravku.

- **„Cross-over“ uspořádání** (zkřížené uspořádání) - zde nejsou hodnocené subjekty randomizovány do jednotlivých léčebných skupin, ale do sekvencí léčebných period. Při tomto uspořádání podstoupí všechny subjekty hodnocení veškeré srovnávané léčby. V průběhu klinického hodnocení si subjekty v jednotlivých ramenech zkříženě mění léčbu (placebo, aktivní komparátor).
- **Faktoriální uspořádání** - umožňuje zkoumat účinek více léků současně. Při tomto uspořádání se v léčebných skupinách systematicky kombinují zkoumané léčebné postupy, což znamená, že některé skupiny dostávají více než jeden lék, tím je umožněno testování účinku více léků současně, bez zvýšení nároků na počet zařazených pacientů.

2 Důležité části statistického výzkumu

V této kapitole si nejdříve probereme hlavní problémy vznikající při aplikaci statistiky, které se vyskytují na různých úrovních vedení lékařského výzkumu, následovaným shrnutím důležitých statistických chyb, potíží a nástrah, tak, aby se jim dalo předejít.

2.1 Návrh/design studie

Nejdůležitější a přitom snad nejvíce opomíjenou částí jakéhokoliv výzkumu je plánování a design. Řádný a úplný plán studie je určitě základem pro úspěšný zdravotnický výzkum. Chyby a omyly objevující se ve fázi plánování výzkumu mohou mít obrovský negativní dopad na hodnotu a spolehlivost obdržených výsledků, protože přímo ovlivňují všechna další stádia výzkumu.

V první řadě je důležité, aby byly při plánování výzkumného projektu primárně přesně formulovány cíle studie a účel výzkumu, stanoveny všechny sledované výstupy (sledované parametry, výstupní proměnné) a rovněž konec studie jak ve výchozím, tak i v závěrečném protokolu výzkumné práce. Při vytváření přesné formulace cílů výzkumu je vhodné zahájit spolupráci lékaře a erudovaného statistika již na samém počátku plánování a návrhu výzkumného projektu. Statistik je nejprve obeznámen s nejdůležitějšími aspekty problému z lékařské stránky a lékař musí získat základní statistické znalosti. Také by měly být výslovně zmíněny a specifikovány zkoumané statistické a vědecké hypotézy, zvláště pokud nejsou evidentní či pokud se zkoumá více než jedna hypotéza. V případě, že neexistují předem specifikované zkoumané hypotézy, měl by být průzkumný charakter práce přiměřeně navržen. Zásadní chybou je vytvářet hypotézy až podle druhu nashromážděných dat a na těchto stejných datech je prověřovat.

Pro kvalitní statistickou výzkumnou práci je naprosto zásadní explicitně stanovit velikost efektu, provést odhad velikosti souboru (výběru) a provést vhodný výpočet statistické síly testu již ve fázi plánování, tak, aby bylo zajištěno, že studie bude provedena s dostatečnou statistickou silou k detekování léčebného efektu na základě pozorování, což je v přímé souvislosti s odstraněním nebo přinejmenším omezením možných chyb II. druhu. V protokolu studie by měla být vždy uvedena skutečná velikost výběru, spolu se záznamem o eventuálním odstoupení (vypadnutí) subjektu ze studie v jakékoliv její fázi včetně uvedení důvodu vypadnutí/odstoupení.

Kdykoliv je to možné, měly by se používat náhodné výběry, randomizace a zaslepování (viz kapitola 1.4). Je nutné si uvědomit, že všechny inferenční statistické metody jsou platné jen pro náhodné výběry a nemusí nezbytně fungovat (platit) pro data získaná jakýmkoli jiným způsobem. Úplný popis metody randomizace a získávání výběru by měl být povinnou součástí vytvářeného výzkumného protokolu.

Pokud se používají kontrolní skupiny, musí být zajištěna jejich počáteční rovnocennost a jejich vzájemná srovnatelnost, aby se zabezpečilo, že se nepracuje s částečně nebo přirozeně heterogenním materiálem, který by byl neporovnatelný. V takovémto případě není logicky možné diference po intervenci identifikovat jako účinky studované léčby, pokud nejsou adekvátně použity vícerozměrné techniky, které upravují studii vzhledem k těmto zavádějícím faktorům. Aplikace testů statistické významnosti pro porovnání homogenity studovaných skupin na počátku studie (což je běžná praxe) není ani přímo adekvátní ani doporučitelná. Pouhé prokázání, že mezi studovanými skupinami na počátku studie nejsou statisticky významné rozdíly, nemůže znamenat, že studijní skupiny jsou ekvivalentní, což platí zvláště pro malé skupiny, kde postrádáme dostatečnou statistickou sílu testu. Co je opravdu třeba, je stanovení homogenity skupin vzhledem k důležitým zavádějícím faktorům.

Shrnutí důležitých statistických chyb a omylů vztahujících se k plánování a designu studie:

- Cíle studie a primární výstupy nejsou uvedené nebo jsou nejasně stanovené.
- Chybí informace o skutečné velikosti výběru (rozsahu souboru); chybí evidence odstoupení od/ vyloučení ze studie včetně důvodů odstoupení/vyloučení.
- Úplně chybí nebo je špatně proveden odhad velikosti výběru na základě stanoveného efektu a/nebo chybí či je nekvalitně stanoven odhad síly plánovaného testu.
- Nejasně stanovená nebo zcela chybějící zkoumaná nulová hypotéza, zkoumané tvrzení.
- Chyby v užití randomizace a v jejím popisu; nejasně popsaná metoda randomizace.
- Chyby v zaslepení (pokud je možné).
- Chyby v popisu shody základních charakteristik a porovnatelnosti studijních skupin na začátku studie; používání nevhodných kontrolních skupin.
- Nevhodné testování shody základních charakteristik.

2.2 Analýza dat

Při provádění statistické analýzy dat a při aplikaci testů statistické významnosti nebo metod získávání odhadů by mělo být všem zřejmé, že každá metoda je založena na několika základních předpokladech, které musí být, alespoň asymptoticky, splněny, abychom získali správný a smysluplný výsledek. Bohužel dokonce tak jednoduché a základní postupy jako populární t-testy (viz kapitola 1.3) či χ^2-testy jsou v lékařském výzkumu špatně používány, protože předpoklady těchto testů nejsou dostatečně před jejich aplikací ověřeny. Navíc při používání t-testů a χ^2-testů se musí používat správné verze testů, protože existují ve více formách. Pokud jsou u χ^2-testů počty v buňkách menší než 5, neměly by být používány, protože za těchto okolností již nejsou aproximace χ^2 rozdělením spolehlivé. V případě, že je testovaný soubor malý, měla by se použít Yatesova korekce spojitosti, ještě lépe však exaktní testy, aby bylo dosaženo spolehlivých výsledků testů. Navíc velká rozmanitost dostupných statistických metod znamená, že výběr nejvhodnější a nejvýkonnější metody není vždy triviální, protože se musí vzít do úvahy spousta detailů.

Pokud studie obsahuje několik sledovaných výstupů, které vyžadují mnohonásobné testování, je důležité kontrolovat poměr falešně pozitivních výsledků, a možnost dosažení chyby I. druhu použitím odpovídajících korekcí pro násobná porovnávání.

Obzvláště důležité je poznání, že porovnání více než dvou skupin, vyžaduje použití metod parametrické nebo neparametrické analýzy rozptylu, protože opakovaným užitím dvouvýběrových testů se zvyšuje riziko nesprávně pozitivního výsledku testu, tj. zvyšuje se pravděpodobnost chyby prvního druhu. Protože se užití dvouvýběrových testů pro mnohonásobné porovnávání často vyskytuje jako důsledek špatně naplánované studie, lze se mu snadno vyhnout tím, že se postup zkonzultuje se statistiky již v plánovací fázi studie.

Měli bychom se vyhnout každé post-hoc (dodatečné) podskupinové analýze, která není specifikovaná v původním studijním protokolu, protože působí jako ˝koupě˝ statisticky významného výsledku („data fishing“).

Shrnutí hlavních statistických chyb a nedostatků v souvislosti s analýzou dat:

- Použití nevhodného statistického testu.
  - ⇒ Neslučitelnost statistického testu s typem zkoumaných dat.
  - ⇒ Nepárové testy pro párová pozorování nebo naopak.

⇒ Nevhodné použití parametrických metod (nesplnění či neověření předpokladů).

⇒ Použití nevhodného testu pro hypotézu, která je předmětem šetření.

- Nedodržení stanovené pravděpodobnosti chyby I. druhu.
  - ⇒ Chybování v zahrnutí korekcí pro násobná porovnávání (použití párových testů místo speciálních metod pro násobná porovnávání).
- Post-hoc analýza („fishing“).
- Typické chyby pro Studentův t-test.
  - ⇒ Neprokázání testových předpokladů.
  - ⇒ Rozdílné velikosti výběrů pro párový t-test.
  - ⇒ Nesprávné násobné párové porovnání více než dvou skupin.
  - ⇒ Použití nepárového t-testu pro párová data nebo naopak.
  - ⇒ Chybně stanovená alternativní hypotéza (jednostranná vs. oboustranná).
- Typické chyby pro χ^2-testy.
  - ⇒ Nepoužití korekce spojitosti v případě malých počtů pozorování.
  - ⇒ Užití χ^2-testu i v případě, že jsou pozorované četnosti velmi malé (pro test nezávislosti např. < 5).
  - ⇒ Nejasné stanovení testované nulové hypotézy.

2.3 Dokumentace

Při statistické analýze dat a při dokumentaci použitých statistických metod dochází k různým chybám a omylům. Proto je nezbytné, aby byly všechny použité statistické metody popsány jasně, správně a dostatečně podrobně tak, aby bylo umožněno kvalifikovanému čtenáři s přístupem k analyzovaným datům přepočítání všech výsledků.
Z tohoto důvodu je podkapitola zaměřená na statistickou analýzu, kde jsou uvedeny veškeré použité techniky a metody, povinná v každé lékařské výzkumné práci.

Běžně používané metody nemusí být podrobně vysvětleny, ale jakékoli nové aplikace a důvody pro použití metod by měly být uvedeny v souhrnu nebo odkazech. Při použití více než jednoho statistického testu, je třeba specifikovat, který test byl na daný soubor dat aplikován. Jednoduché tvrzení „v případě potřeby“ není v tomto případě postačující.

Pro statistické testy, které existují ve dvou verzích jako párové nebo nepárové (např. t-test, Wilcoxonův-test), je nutné určit, která verze testu (párová či nepárová) byla použita a musí být uvedeno, zda se jedná o oboustranný nebo jednostranný test.

Souhrn běžných chyb a nedostatků souvisejících s dokumentací používaných statistických metod:

- Nedostatek specifikace / definování všech používaných statistických testů jasně a správně.
  - ⇒ Neuvedení, zda se jedná o jednostranný nebo oboustranný test, tj. alternativní hypotézy.
  - ⇒ Neuvedení, zda je použitý test párový nebo nepárový.
- Špatné názvy statistických testů.
- Užití neobvyklých či nejasných metod bez jejich vysvětlení nebo odkazu na literaturu.
- Neschopnost určit, jaký test byl aplikován na daný soubor dat, pokud byl proveden více než jeden test; užití fráze „v případě potřeby“.

2.4 Prezentace

Dobrý výzkum si zaslouží být dobře prezentován a řádná prezentace je stejně tak součástí výzkumu jako shromažďování a analýza dat [2]. Proto, jestliže statisticky popisujeme nebo prezentujeme studovaná data, měli bychom si dát pozor, abychom použili vhodné statistické míry polohy a variability. Pokud se použije aritmetický průměr a směrodatná odchylka, mělo by být zřejmé, že data jsou alespoň přibližně normálně rozdělena a nejsou asymetrická (zešikmená). Jinak tyto míry nemohou být smysluplně použity pro popis dat. V každém případě, by měly být směrodatné odchylky uvedeny v závorkách [tj. průměr (SD)] spíše, než užívat zápis ± SD (viz příklad 2.1), neboť tato specifikace může být čtenářem zaměněna s 95% intervalem spolehlivosti. Pro zešikmené rozdělení hodnot, které je v biologickém a lékařském výzkumu poměrně časté, je vhodnější užít mediány, kvartily nebo rozpětí, s vědomím toho, že rozpětí je citlivé na odlehlé hodnoty a proto může být jako sumární statistika nevhodné. Pokud se následně pro statistickou analýzu dat využívají neparametrické testy, měli bychom se vyhnout středním

hodnotám a směrodatným odchylkám, neboť tyto parametry nejsou, podle definice, testovány neparametrickými testy, a tudíž nemají smysl pro popis hodnot v rámci šetření. Zde by měly být upřednostněny mediány, rozsahy hodnot (rozpětí) nebo mezikvartilové rozpětí. Také není dostačující prezentovat jen střední hodnotu bez udání míry variability dat.

Směrodatná odchylka průměru (SEM), která se běžně a nesprávně používá pro statistický popis hodnot (snad proto, aby se zdály být hodnoty méně variabilní), není popisná statistika (charakteristika dat), ale spíše inferenční statistika užívaná ve statistických odhadech. Směrodatná odchylka průměru je mírou variability střední hodnoty (spočítané jako aritmetický průměr) sady výsledků, zatímco směrodatná odchylka téže sady hodnot je mírou neurčitosti (rozptýlení) těchto výsledků. Není proto vhodné prezentovat průměr současně se směrodatnou odchylkou průměru jako míru disperze. Stejně tak je nezbytné, aby u všech znaků „±“ použitých v textu, v grafech, tabulkách nebo výpočtech bylo uvedeno (alespoň u prvního výskytu) co znamenají a stejně tak popsat význam chybových úseček v grafickém znázornění dat sloupcovými grafy.

Pro primární míry cílů a hlavní výsledky studie by měly být vždy, když je to možné, odhadnuty intervaly spolehlivosti, protože samotná hodnota pravděpodobnosti neposkytuje smysluplnou informaci o významu nebo velikosti efektu. Tedy správné používání technik statistického odhadu může čtenáři do značné míry rozšířit poskytnutou informaci obsaženou ve studii. Pokud používáme statistické odhady pro porovnávání skupin, potom by intervaly spolehlivosti měly být stanoveny spíše pro diference mezi skupinami, než pro každou skupinu zvlášť.

P-hodnoty by měly být uváděny přesně tak, jak byly obdrženy, než prostřednictvím zvolených prahových hodnot, jako např. „p = NS“, „$p < 0{,}05$“, nebo „$p > 0.05$“. Nicméně číselné informace by neměly být uváděny s nereálnou úrovní přesnosti, kterou nelze získat na základě daného rozsahu výběru dat pro studii.

Souhrn statistických chyb a nedostatků související s prezentací studijních dat:

- Nedostatečný grafický nebo numerický popis základních dat.
  - ⇒ Střední hodnota bez údaje o variabilitě dat.
  - ⇒ Uvádění SEM (Standard Error of Mean = směrodatná odchylka průměru) místo SD (Standard Deviation = směrodatná odchylka) u číselného popisu dat.

⇒ Použití střední hodnoty (SD) k popisu dat, která nejsou normálně rozdělena.

⇒ Chybějící definice „±“ u popisu variability nebo použití neoznačené (nedefinované) chybové úsečky v grafu.

- Nevhodné a nedostatečné vykazování výsledků.

⇒ Uvádění výsledků pouze jako *p*-hodnoty, neuvádění konfidenčních intervalů.

⇒ Intervaly spolehlivosti uváděné pro každou skupinu zvlášť a ne pro kontrast.

⇒ Tvrzení „p = NS“, „$p < 0{,}05$“ nebo jiné podobné vztahy místo uvádění přesné *p*-hodnoty.

⇒ Numerické informace uvedené s nereálnou úrovní přesnosti.

Příklad 2.1 Na následující tabulce vidíme příklad nevhodného užití zápisu „± SD“ při prezentaci výsledků [16].

Table 3. Laboratory-Test Results According to Outcome.*

VARIABLE	GROUP WITH EVENTS (N = 106)	EVENT-FREE GROUP (N = 2700)	P VALUE
Fibrinogen (g/liter)	3.28±0.74	3.00±0.71	0.01
Von Willebrand factor antigen (%)	137.5±48.8	124.6±49.1	0.05
t-PA antigen before venous occlusion (ng/ml)	11.9±4.7	10.0±4.2	0.02
C-reactive protein (mg/liter)	2.15±1.96	1.61±1.38	0.05

*Values are means ±SD, with adjustment for medical center, age, and sex. P values have also been adjusted for all the other risk factors for coronary disease on which we obtained data (see the Methods section). For statistical analyses of logarithmically transformed values, geometric means and approximate standard deviations are shown. Percentages have been calculated on the basis of the numbers with known values. The number of patients with definite coronary events included in the statistical analyses ranged from 89 to 106 for the various laboratory tests.

2.5 Interpretace

Pro konečnou fázi vědecko - výzkumného procesu, tj. pro interpretaci výsledků analýzy dat, je nezbytné, aby byly při samotné analýze dat použity nejvhodnější a nejsilnější statistické testy. Tím se zamezí zkresleným závěrům ze studie, které jsou nedostatečně podložené daty. Tvrdit, že efekt je významný, můžeme jen na základě použití

testu statistické významnosti. Pokud výsledky nevykazují statistickou významnost, je velmi důležité být opatrný ve vyvozování závěrů, přičemž ale nedostatek statistické významnosti vždy neznamená, že neexistuje žádný efekt nebo diference. Velikost výběru může být např. příliš malá na to, aby zajistila statistickou významnost, i když výsledky pořád ještě obsahují klinicky významné výsledky nebo poskytují hodnotný impuls pro další zkoumání.

Při použití malého rozsahu výběru a získání nevýznamných výsledků testů, by se měla nezbytně prověřit velikost chyb II. druhu, jakož i problematika násobného testování a rostoucího rizika falešně pozitivních výsledků testů. Při interpretaci výsledků by mělo být také dostatečně zváženo, zda je či není studie upravena vzhledem k potenciálním zavádějícím faktorům či biasu.

Souhrn důležitých statistických chyb a nástrah souvisejících s interpretací závěrů studie:

- Chybné interpretace výsledků.
  - ⇒ „Nevýznamný“ interpretovaný jako „bez efektu“ nebo „neodlišný“.
  - ⇒ Vytváření závěrů, které nelze doložit daty.
  - ⇒ Významnost uváděná bez potřebné analýzy dat a uvedení použitého statistického testu.
- Nedostatečná interpretace výsledků.
  - ⇒ Přehlížení chyby II. druhu při vykazování nevýznamného výsledku.
  - ⇒ Chybějící diskuse o problému významnosti testu mnohonásobného porovnávání, pokud se provádí.
  - ⇒ Nedostatky v diskusi o potenciálních zdrojích biasu a zavádějících faktorech.

Dostupnost různých statistických softwarových programů usnadňuje statisticky nekvalifikovaným lékařům provádění vlastní analýzy dat, což ale může na druhé straně vést k velkým problémům vyplývajícím z nedostatečné znalosti základních matematických konceptů a statistických principů. Lékaři a další výzkumní pracovníci musí být podporováni v tom, aby se stále dozvídali více o statistice, protože některé studie poukazují na závažný nedostatek znalostí statistiky již v rámci výuky a vzdělávání lékařů [2]. Konzultace se statistikem a výměna informací s kvalifikovanými odborníky často

probíhá až dlouho poté, co je studie naplánována, což téměř znemožňuje přímé odstranění nedostatků, k nimž došlo v předchozích krocích výzkumu.

V této souvislosti je pro všechny publikované články žádoucí, aby byly před jejich zveřejněním přezkoumány statistiky. Časopisy by měly zveřejnit svou politiku ohledně kontroly statistických výsledků. Hodnotitelům těchto metod by měla být alespoň nabídnuta možnost vidět přepracované rukopisy před konečnou publikací.

2.6 Publikace

Díky publikaci biomedicínského výzkumu, především v předních odborných časopisech, se po celém světě šíří rychlým tempem nové poznatky a výzkum se neustále rozvíjí. Jeho závěry stále výraznějším způsobem ovlivňují životy lidské společnosti. Články publikované v recenzovaných časopisech by měly zaručit nejen vědeckou kvalitu, ale i praktický význam publikovaných závěrů. V důsledku výskytu chyb, již zmíněných v předchozích kapitolách, tomu tak ale vždy není.

Publikované biomedicínské výzkumné práce by měly být rozděleny na následující části, podle nichž můžeme hodnotit kvalitu těchto prací [4]:

- **Souhrn** – má ukázat cíle výzkumu, díky kterému se čtenář může rozhodnout, zda číst celý článek či nikoliv.
- **Úvod** – měl by být krátký a čtenáři by měl zejména sdělit srozumitelně hlavní cíl studie a informovat o tom, zda byl stanoven dříve, než se nashromáždila data nebo v opačném případě zda byl formulován až po nashromáždění dat.
- **Metody** – je důležité, aby byly podrobně popsány použité metody, aby byli čtenáři informovaní o způsobu provedení studie.
- **Výsledky** – část publikace shrnující výsledky výzkumu by měla obsahovat závěry studie, které odpovídají stanoveným cílům studie (např. formou tabulek, grafů sumarizujících data a výsledků popisujících studii).
- **Diskuse a závěry**

3 Správné užití statistických metod

V celém textu jsem zatím uvedla, jakých chyb se v jednotlivých částech statistického výzkumu můžeme dopustit, a proto bych chtěla také nabídnout částečný přehled toho, jak můžeme statistické metody správně a efektivně využít a zároveň se vyhnout možným nástrahám, jak jej uvádí ve své publikaci Good a Hardin [1].

Částečný způsob bezchybného užití statistiky:

1. Před provedením laboratorního experimentu, klinického pokusu, nebo průzkumu a před analýzou existujícího souboru dat si nejprve stanovíme náš cíl a vytvoříme si příslušný plán našeho výzkumu.
2. Definujeme populaci, na kterou budeme aplikovat výsledky naší analýzy.
3. Důležité je zjištění všech možných zdrojů variability, jejich sledování a měření za účelem omezení vlivu zavádějících faktorů.
4. Zformulujeme hypotézy a všechny související alternativy. Vytvoříme seznam možných experimentálních poznatků spolu se závěry, ze kterých bychom mohli čerpat a postupy, které bychom mohli použít v případě, že by tento nebo jiný výsledek dokazoval, že tomu tak je. To vše musíme provést před dokončením jednotného formuláře pro sběr dat.
5. Podrobně popíšeme, jakým způsobem budeme vybírat reprezentativní vzorek z populace.
6. Použijeme odhady, pokud jsou nestranné, konzistentní, eficientní a zahrnující minimální ztrátu. Ke zlepšení výsledků se zaměříme na vhodné, klíčové a přípustné statistiky a použijeme intervalové odhady.
7. Musíme znát předpoklady (hypotézy), které tvoří základ pro testy, které používáme. Snažíme se použít testy, které jsou nejsilnější pro testovanou hypotézu.
8. Zahrneme do naší zprávy kompletní informace o tom, jak byla definována populace a jakým způsobem z ní byl vzorek vybrán. Pokud data chybí nebo plán odběru vzorků nebyl dodržen, musíme vysvětlit důvod, proč tomu tak je a vyjmenovat všechny rozdíly mezi daty, které prezentujeme ve vzorku a také data, která chyběla nebo byla vyloučena.

Závěr

Cílem této práce bylo popsat běžné statistické chyby, nástrahy a nedostatky týkající se různých stádií lékařského vědeckého výzkumného procesu. Při psaní této práce jsem se z velké části věnovala vysvětlení důležitých pojmů a principů, aby tato práce sloužila také jako pomoc pro „nestatistiky“ při vytváření statistického výzkumu, ale i přesto nelze popsat všechny metody, jako například analýzu síly testu, odhad velikosti výběru či testy používané v kontingenčních tabulkách, protože by tato práce výrazně překračovala doporučený rozsah bakalářské práce. Mohu alespoň čtenáře nasměrovat na publikaci, ve které jsou tyto metody popsány, jakou je např. *Matematická statistika* J. Anděla [9].

Pečlivé a přesné využití statistiky v lékařském výzkumu má velký význam. Správné či nesprávné užití statistiky v lékařské diagnostice a biomedicínském výzkumu může ovlivnit, zda jedinec bude žít nebo umře, zda je jeho zdraví chráněno nebo ohroženo, a zda lékařská věda dělá pokroky nebo naopak zaostává. Protože společnost závisí na dobré statistické praxi, všichni co statistiku v praxi využívají, bez ohledu na jejich vzdělání a povolání, mají společenskou povinnost vykonávat svoji práci profesionálním, kompetentním a etickým způsobem.

Věřím, že tato bakalářská práce pomůže nejen mně při dalším studiu, ale i lékařům a dalším výzkumníkům, aby z pohledu statistiky správně plánovali, analyzovali a prezentovali své výzkumy a výsledky ve svých publikacích.

Literatura

[1] Good, P. I., Hardin, J. W.: *Common Errors in Statistics: (and How to Avoid Them).* John Wiley & Sons, New Jersey 2006.

[2] Strasak, A.M., Zaman, Q., Pfeiffer, K.P., Göbel, G., Ulmer, H.: *Statistical errors in medical research – a review of common pitfalls*. [online], SWISS MED WKLY, 2007;137:44-49. Dostupné z WWW: <http://www.smw.ch/docs/pdf200x/2007/03/smw-11587.pdf>.

[3] Malý, M.: *Měření asociací v epidemiologických studiích*. [online], [cit. 2010-02-08]. Dostupné z WWW: <www.cittadella.cz/euromise/sites/File/Maly/MP301/maly01.pps>.

[4] Zvárová, J.: *Základy statistiky pro biomedicínské obory.* 1.vyd. Praha: Nakladatelství Univerzity Karlovy, 1998.

[5] *Základní statistické pojmy*. [online], [cit. 2010-02-11]. Dostupné z WWW: <http://www.kmt.zcu.cz/person/Kohout/info_soubory/letnisem/SS/stat10.pdf>.

[6] EuroMISE Centre [online]. 1999 [cit. 2010-02-12]. *7. Testování hypotéz* . Dostupné z WWW: <http://new.euromise.org/czech/tajne/ucebnice/html/html/node9.html>.

[7] AI ACCESS [online]. 2010 [cit. 2010-02-25]. *Hypothesis testing*. Dostupné z WWW: <http://www.aiaccess.net/English/Glossaries/GlosMod/e_gm_test_1.htm>.

[8] Vácha, J.: *Lékaři a věda statistická.* Vesmír [online]. 1995, č. 6, [cit. 2010-03-01]. Dostupné z WWW: <http://www.vesmir.cz/clanek/lekari-a-veda-statisticka>.

[9] Anděl, J.: *Matematická statistika*. 1. vyd. Praha: SNTL/ALFA, 1978.

[10] Kunderová, P.: *Základy pravděpodobnosti a matematické statistiky*. 1.vyd. Olomouc: Univerzita Palackého v Olomouci, Přírodovědecká fakulta, 2004.

[11] Knapp, R. G., Miller, M. C.: *Clinical epidemiology and biostatistics*. Baltimore: Williams & Wilkins, 1992.

[12] Svobodník, A., Coufal, O., Dušek, L.: *Základní pojmy v designu, analýze dat a interpretaci výsledků klinických hodnocení léčiv*. [online], [cit. 2010-03-12]. Klinická onkologie 2005;18(Suppl): 238-241. Dostupné z WWW: <www.linkos.cz/odbornici/vzdelavani/zvl1_05/02.pdf>.

[13] Suchý, D., Hora, M., Fínek, J.: *Vývoj a klinické hodnocení nových léčiv*. [online], [cit. 2010-02-22]. Ces Urol 2009;13(2):141-148. Dostupné z WWW: <www.czechurol.cz/dwnld/0902_141_148.pdf >.

[14] Svoboda, D.: *Statistika v klinickém hodnocení*. [online], [cit. 2010-02-23]. PharmTest, Hradec Králové, 395-398. Dostupné z WWW: <www.zdravcentra.cz/cps/rde/xbcr/zc/1876.pdf>.
[15] Svoboda, D.: *Statistika v klinickém hodnocení*. [online], [cit. 2010-02-23]. PharmTest, Hradec Králové, 617-618. Dostupné z WWW: <www.zdravcentra.sk/cps/rde/xbcr/zcsk/2005_6_FM_01.pdf>.
[16] Simon G. Thompson, M.A., et al.: *Hemostatic factors and the risk of myocardial infarction or sudden death*. The New England journal of Medicine [online]. 1995, [cit. 2010-03-08]. 332:635-641,. Dostupný z WWW: <http://content.nejm.org/cgi/content/full/332/10/635>.
[17] Novotný, J., Petruželka, L., Pecen, L.: *Metodické problémy klinických studií v onkologii*. [online], [cit. 2010-03-16]. Klinická onkologie 2005;18(Suppl): 252-254. Dostupné z WWW: <www.linkos.cz/odbornici/vzdelavani/zvl1_05/02.pdf>.
[18] Koutková, H.: *Pravděpodobnost a matematická statistika; Základy testování hypotéz; MODUL GA03 M04*. 1. vyd. Brno: Akademické nakladatelství CERM, 2007.
[19] Jindrová, A., Prášilová, M., Zeipelt, R.: *Statistika I*. 1. vyd. Praha: ČZU Provozně ekonomická fakulta, 2008.
[20] Martínek, J.: *Statistické metody v hodnocení léčby*. [online]. 2004, [cit. 2010-03-18]. Dostupné z WWW: <http://www.hpb.cz/index.php?pId=04-1-2-12>.
[21] VlastniServer.cz [online]. 2010 [cit. 2010-03-31]. *Základy pojmy statistiky. Popisná statistika: zpracování empirických dat; charakteristiky polohy, variability, šikmosti a špičatosti; charakteristiky vztahu dvou veličin* . Dostupné z WWW: <http://www.informatika.xcars.cz/zakladnipojmy.html>.
[22] Homola, V.: *Úvod do statistiky*. [online]. 2001, [cit. 2010-03-02]. Dostupné z WWW: <http://homel.vsb.cz/~hom50/SLBSTATS/UST/GS02.HTM>.
[23] *Minislovník statistiky*. [online], [cit. 2010-03-02]. Dostupné z WWW: <http://www.dc.fd.cvut.cz/material/msstat.doc>.
[24] Šedivá, B.: *Parametrické testy*. [online], [cit. 2010-02-22]. Dostupné z WWW: <home.zcu.cz/~sediva/stav/kap07.pdf>.
[25] Dohnal, L.: *Desatero pro porovnávání výsledků dvou metod*. [online], 2002 [cit. 2010-03-22]. Štatistické metódy pre klinickú epidemiológiu a laboratórnu prax;21-26. Dostupné z WWW: <http://www1.lf1.cuni.cz/~ldohna/publik/Kap_4_Desatero.pdf>.