

## 5 Úvod do statistického usuzování

Při explorační a popisné analýze dat docházíme k závěrům, které se týkají pouze nashromážděných údajů. Naproti tomu cílem statistického usuzování (statistické inference) je odvodit na základě dat týkajících se výběru a jistých předpokladů o jejich rozdělení závěry o celé populaci nebo procesu. Naše schopnost takto zobecňovat závisí na plánu sběru dat a na chování numerických charakteristik, vypočítaných z dat.

Ačkoliv existuje mnoho různých postupů, jak provádět statistické usuzování v konkrétní situaci, všechny patří ke dvěma základním typům. Jedná se buď o metody pro **odhadování**, nebo postupy založené na **statistických testech**.

### PŘÍKLAD 5.1

#### Bodový odhad, intervalový odhad, test hypotézy

Odhadujeme průměrnou výšku chlapců v určité věkové kategorii. Ve studii se pro náhodně vybranou skupinu chlapců zjistil průměr 179 cm. To je „bodový odhad“ průměru v celé populaci. Víme ale, že kdybychom sestavili pro výzkum odlišnou skupinu (jiný výběr), dostali bychom nejspíš trochu jiný bodový odhad. Proto je lepší místo jediného čísla („bodu“) uvést interval, v němž se populační hodnota nachází s velkou spolehlivostí. To je „intervalový odhad“. Navíc chceme znát, zda se liší v průměrné výšce dvě specifikované subpopulace chlapců. To je problém testování hypotézy. Všechny tyto tři druhy otázek mají smysl.

V této kapitole demonstrujeme principy statistického usuzování na jednoduchých příkladech. Uvedeme základní metodu nalezení intervalu spolehlivosti pro průměr normálního rozdělení. Ukážeme také na průměru normálního rozdělení statistickou inferenci pomocí testu hypotézy. Postupy statistického testování hypotéz jsou užitečné tehdy, jestliže potřebujeme provést rozhodnutí o hodnotě parametru nebo obecně o tvaru rozdělení náhodné proměnné. Pomocí těchto postupů například dokážeme s velkou spolehlivostí rozhodnout, zda určitý parametr je větší, nebo menší než specifikovaná hodnota nebo zda se parametry v různých populacích liší.

**PŘÍKLAD 5.2****Statistické usuzování**

Výzkumníci v oblasti psychiatrie porovnávají skupinu duševně nemocných jedinců se skupinou zdravých jedinců na základě sledovaných 77 různých proměnných, popisujících děství a rodinné zázemí všech jedinců. Zjistila se „statistická rozdílnost“ mezi skupinami u dvou z těchto proměnných. Postupovali jsme při statistickém usuzování správně? Jak můžeme interpretovat tuto „statistickou významnost“?

Protože metody statistického usuzování vycházejí z výběrových rozdělení, vyžadují určitý pravděpodobnostní model dat. V následujících příkladech předpokládáme, že jsme provedli prostý náhodný výběr a data mají normální rozdělení. Opíráme se o poznatky z předcházející kapitoly, které se týkaly rozdělení výběrových statistik.

## 5.1 Základní koncepty statistického usuzování

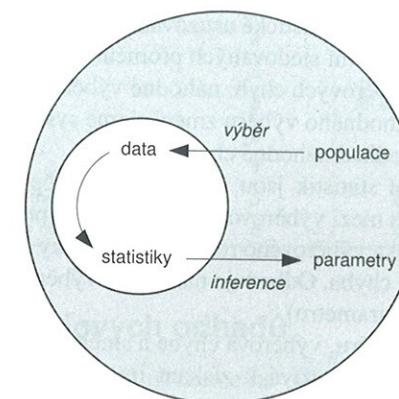
**Statistická inference (statistické usuzování)** znamená provedení zobecnění z náhodného výběru na populaci. Toto zobecnění se provádí s určitým stupněm jistoty, resp. spolehlivosti. Jak jsme už naznačili, rozlišujeme dvě hlavní formy statistického usuzování: odhadování a testování hypotéz. V obou typech statistického usuzování, jehož princip je schematicky zachycen na obrázku 5.1, počítáme z dat výběru určité statistiky, jež slouží jako základ tohoto usuzování. Odhadování vede k určení parametrů neznámého rozdělení. Testování hypotéz poskytuje jisté zdůvodnění pro úvahy, zda danou hypotézu o parametru nebo pravděpodobnostním rozdělení je možné zamítnout, nebo ne.

Účelem statistiky je získat porozumění výzkumným problémům pomocí dat. Přitom lze použít různé přístupy. Zmínili jsem se o explorační analýze dat (kap. 3.8). Jestliže klademe důraz na získání vhodných dat pomocí statistického šetření nebo experimentu, začínáme směřovat k statistické inferenci. Oba typy přístupů jsou důležité pro efektivní práci s daty. V tabulce 5.1 jsou schematicky ukázány rozdíly mezi oběma přístupy.

Oba přístupy se navzájem podporují. Statistické usuzování vyžaduje kvalitní data. Pomocí explorační analýzy odhalujeme odlehle hodnoty nebo datové konfigurace, které by mohly ovlivnit přesvědčivost statistické inference. Explorační analýza, především její grafické metody, jsou prvním krokem k validní inference. Hlavním předpokladem validity inference je získání dat pomocí dobré navrhované

Obr. 5.1

Z dat výběru počítáme statistiky, které podávají informaci o parametrech populace



Tab. 5.1 Porovnání explorační analýzy a statistické inference

Explorační analýza	Statistická inference
Účel je neomezený průzkum dat, hledání zajímavých konfigurací.	Cílem je odpovědět na specifickou otázku, kterou jsme položili před tím, než začal sběr dat.
Závěry platí pouze pro jedince a měření, jež jsme měli k dispozici.	Závěry platí pro větší skupinu jedinců (populaci) nebo širší třídu okolností.
Závěry jsou neformální, vycházíme z toho, co jsme nalezli v datech.	Závěry jsou formální, s upřesněním jejich spolehlivosti.

něho schématu výzkumu a sběru dat. Jestliže používáme metody statistického usuzování, postupujeme tak, jako by data představovala náhodný výběr nebo pocházela ze znáhodněného experimentu. Pokud tomu tak není, naše závěry lze snadno zpochybnit. Tomu nemůže zabránit ani složitá matematika, jež se používá při formálním odvození metod statistického usuzování. Jakkoli složitý matematický aparát nedokáže vylepšit základní pochybení při sběru dat. Proto je tak důležité porozumět základům metodologie výzkumné práce, plánování projektů, správné realizaci sběru dat a data zpracovat metodami statistické inference pouze tehdy, pokud jsme přesvědčeni, že si takovou analýzu zaslouží.

Shrneme v bodech principy, které stojí v základu statistického usuzování:

1. Statistické usuzování znamená zobecňování z výběrových statistik na parametry rozdělení.
2. Abychom mohli provést statistické usuzování, musíme mít nějakou teorii, jež popisuje náhodné chování sledovaných proměnných.
3. Existují dva typy výběrových chyb: náhodné výběrové chyby a systematické chyby. Získáním náhodného výběru zmenšujeme systematickou chybu a získaeme podklad pro odhad náhodné chyby výběru.
4. Výběrová rozdělení statistik jsou teoretická pravděpodobnostní rozdělení, která popisují vztah mezi výběrovou statistikou a populací.
5. Směrodatná odchylka výběrového rozdělení statistiky (odhadu parametru) se nazývá směrodatná chyba. Odhaduje náhodnou výběrovou chybu vypočítané statistiky (odhadu parametru).
6. Jak roste velikost výběru, výběrová chyba a směrodatná chyba se zmenšují.
7. Směrodatná chyba se používá k získání intervalového odhadu parametru i k testování hypotéz o parametrech rozdělení.

## 5.2 Spolehlivé odhadování

Parametr je číselná hodnota, jež platí pro celou populaci, kdežto **odhad parametru** získáváme pomocí výběru z populace. Parametr a jeho odhad jsou ve vztahu, ale nemůžeme je zaměnit. Výběrové charakteristiky jsou náhodné proměnné. Parametry se považují za konstantu (ačkoli v tzv. bayesovské teorii statistiky tomu je jinak). Parametry často neznáme, kdežto výběrovou charakteristiku můžeme pomocí získaných měření spočítat. Parametry a jejich bodové odhady (tj. výběrové statistiky) odlišujeme jiným značením – pro označení teoretických parametrů používáme často řecká písmena, odhady (statistiky) značíme písmeny běžné latinské abecedy. Tabulka 5.2 uvádí značení nejznámějších teoretických parametrů a jejich odhadů.

Populační parametry se snažíme odhadnout co nejlépe. Proto metody odhadu tvoří důležitou část statistické inference a statistické teorie. Odhad provádíme buď **jedinou hodnotou**, nebo **číselným intervallem**, v němž se nachází teoretická hodnota parametru se spolehlivostí  $S$ . V prvním případě mluvíme o bodovém odhadu, ve druhém případě o intervalovém odhadu.

Tab. 5.2 Značení teoretických a výběrových charakteristik

Parametr	Teoretický	Výslovnost	Odhad
průměr	$\mu$	mí	$\bar{x}$ nebo $M$
směrodatná odchylka	$\sigma$	sigma	$s$
medián	$\tilde{\mu}$	mí s vlnkou	$\tilde{x}$ nebo $Me$
modus	$\hat{\mu}$	mí se stříškou	$\hat{x}$ nebo $Mo$
korelační koeficient	$\rho$	ró	$r$
pravděpodobnost (někdy)	$\pi$	pí	$\hat{p}$ nebo $p$
směrnice regresní přímky (někdy)	$\beta$	beta	$b$
průsečík regresní přímky s osou $y$ (někdy)	$\alpha$	alfa	$a$

### 5.2.1 Kvalita bodových odhadů

Problematiku bodových odhadů ukážeme na odhadu parametru  $\mu$ . U symetrického rozdělení lze odhadovat teoretický průměr několika způsoby. Například výběrovým průměrem, mediánem nebo průměrem extrémních hodnot (minima a maxima). V čem se tyto charakteristiky liší v souvislosti s kvalitou odhadu teoretického průměru?

Obecně může být kvalita daného bodového odhadu teoretického parametru posuzována s ohledem na několik požadavků:

- **konzistence** – s rostoucím počtem pozorování se odhad blíží k teoretické hodnotě s pravděpodobností 1;
- **nestrannost** – jestliže při opakování výběrů kolísá odhad kolem teoretické hodnoty symetricky na obě strany, odhad je nestranný;
- **vydatnost** nebo **eficience** – rozptyl odhadů při opakování výběrů je malý;
- **rezistence** – odlehle hodnoty (způsobené hrubou chybou měření nebo špatným zápisem) nemají velký vliv na hodnotu odhadu.

U odhadů zjišťujeme, do jaké míry jsou uvedené vlastnosti splněné.

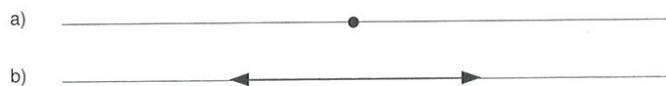
V případě odhadu průměru  $\mu$  je aritmetický průměr jeho nejvydatnějším odhadem, méně vydatným je výběrový medián a nejmenší hodnotu eficience (vydatnosti) má průměr z extrémních hodnot. Ten je také velmi citlivý k odlehlym hodnotám. Rezistentním odhadem parametru  $\mu$  je výběrový medián. Medián i průměr z extrémních hodnot jsou nestranným odhadem  $\mu$  pouze tehdy, jestliže rozdělení náhodné proměnné má symetrický tvar.

## 5.2.2 Interval spolehlivosti pro $\mu$

Jak jsme uvedli, místo bodového odhadu často užíváme odhad intervalový – sestrojujeme **interval spolehlivosti**. Princip postupu ukážeme na případu odhadu parametru  $\mu$ . Podobně bychom postupovali i v případě odhadu jiných parametrů. Vycházíme ze znalosti výběrového rozdělení aritmetického průměru  $\bar{x}$ . Předpokládáme, že měříme náhodnou proměnnou s normálním rozdělením.

V prvním kroku je důležité si uvědomit, že výběrový průměr je bodovým odhadem populačního průměru. Proto jeho hodnota tvoří střed, kolem něhož je interval spolehlivosti situován. Na obě strany od výběrového průměru se ve vzdálenosti určenémezí chyby nacházejí hranice intervalu spolehlivosti. Nalevo od něho je dolní hranice a napravo od něho je horní hranice intervalu spolehlivosti (obr. 5.2).

Obr. 5.2 Bodový (a) a intervalový (b) odhad



Když odečteme hodnotu dolní hranice od horní hranice intervalu spolehlivosti, dostaneme délku intervalu spolehlivosti, jež je celkovou mírou nepřesnosti vytvořeného odhadu. Krátké intervaly spolehlivosti jsou přesnější než dlouhé.

Délka intervalu spolehlivosti závisí na hladině spolehlivosti, se kterou ho určujeme. **Hladina spolehlivosti** je pravděpodobnost, s jakou se odhadovaný populační parametr ocitne v tomto intervalu při opakovém provádění výběru. Nejpoužívanější hladiny jsou 90 %, 95 % nebo 99 %, ale použít lze i jinou hladinu. Když např. pracujeme s 95% hladinou spolehlivosti, znamená to, že ze 100 vytvořených intervalů jich přibližně 95 pokryje hledanou hodnotu parametru. Interval spolehlivosti tedy není nic absolutně spolehlivého. Naopak má náhodně proměnlivý charakter, stejně jako příslušný bodový odhad. Každý výběr vede k trochu jiném intervalu spolehlivosti.

### Interval spolehlivosti pro $\mu$ při známém $\sigma$

Vycházíme ze znalosti rozdělení aritmetického průměru. Ten má jako náhodnou proměnnou normální rozdělení s průměrem  $\mu$ . Jeho směrodatná odchylka  $\sigma$

hodnotu směrodatné chyby průměru, z čehož plyne, že s pravděpodobností 95 % leží výběrový průměr v intervalu

$$\mu - 1,96\sigma_{\bar{x}} < \bar{X} < \mu + 1,96\sigma_{\bar{x}}.$$

Použitím jednoduchých algebraických úprav z této nerovnice plyne, že

$$P(\bar{X} - 1,96\sigma_{\bar{x}} < \mu < \bar{X} + 1,96\sigma_{\bar{x}}) = 0,95.$$

Proto je 95% interval spolehlivosti pro parametr  $\mu$

$$(\bar{x} - 1,96\sigma_{\bar{x}}; \bar{x} + 1,96\sigma_{\bar{x}}).$$

### PŘÍKLAD 5.3

#### Interval spolehlivosti pro střední hodnotu při známém rozptylu

Provádíme měření pomocí znalostního testu  $G$  na jisté škole. O hodnotách  $G$  je známo, že pro populaci dětí v daném věkovém pásmu jsou normálně rozděleny se střední hodnotou 100 a směrodatnou odchylkou 15. Předpokládáme, že proměnná  $G$  má na naší škole u dětí v daném věkovém pásmu stejnou rozptýlenost jako v celé populaci. Provedli jsme 9 měření a získali jsme průměr 112,8. Vypočítáme 95% interval spolehlivosti pro střední hodnotu znalostního parametru  $G$  dětí na škole:

$$(112,8 - 1,96 \times 15 / \sqrt{9}; 112,8 + 1,96 \times 15 / \sqrt{9}) = (112,8 - 1,96 \times 5; 112,8 + 1,96 \times 5) \\ = (103,0; 122,6)$$

### Interpretace intervalu spolehlivosti

Hladina spolehlivosti 95 % neznamená, že  $\mu$  leží uvnitř tohoto intervalu s touto pravděpodobností vzhledem k nějakému pravděpodobnostnímu rozdělení parametru  $\mu$ . Teoretický parametr nepředstavuje náhodnou proměnnou, a proto také nelze mluvit o pravděpodobnostech jeho hodnot. Zmínili jsme, že hladina spolehlivosti znamená pravděpodobnost pokrytí hodnoty  $\mu$  intervalem spolehlivosti při opakovém použití pokusu a celé procedury. Jedná se o dosti jemný rozdíl, který připomíná, že existuje jenom jeden průměr, ale mnoho intervalů spolehlivosti, jak opakově provádíme náhodné výběry. Průměr  $\mu$  je neznámý, ale má pro danou populaci určitou danou hodnotu. Interval spolehlivosti je sestrojen tak, aby pokryl parametr  $\mu$  s danou spolehlivostí.

Jestliže chceme jinou hladinu spolehlivosti, musíme změnit koeficient, jímž násobíme směrodatnou chybu odhadu. Nechť  $\alpha$  je pravděpodobnost, s níž výzkumník toleruje, že interval spolehlivosti nepokryje  $\mu$ . Pak  $(1 - \alpha)100\%$  interval spolehlivosti pro  $\mu$  má tvar

$$(\bar{x} - z_{1-\alpha/2}\sigma_{\bar{x}}; \bar{x} + z_{1-\alpha/2}\sigma_{\bar{x}}),$$

kde  $z_{1-\alpha/2}$  odpovídá  $(1 - \alpha/2)$  kvantilu standardizovaného normálního rozdělení.

Například abychom spočítali 90% interval spolehlivosti, volíme  $\alpha = 0,10$  a  $z_{1-0,05}$  má hodnotu 1,645. Proto má 90% interval spolehlivosti pro  $\mu$  z příkladu 5.3 tvar  $(104,6; 121,0)$ . Tento interval s 90% pravděpodobností pokryje správný populační průměr.

Jestliže počítáme 99% interval spolehlivosti, pak  $\alpha = 0,01$  a  $z_{1-0,005} = 2,58$ . Tedy 99% interval spolehlivosti pro  $\mu$  na základě dat příkladu 5.3 je dán hodnotami  $(99,9; 125,7)$ . Tento interval s 99% pravděpodobností pokryje správnou hodnotu průměru  $\mu$ . Obě pravděpodobnosti (90% a 99%) uvažujeme vzhledem k mnoha opakování provedení náhodného výběru.

Popsané vzorce se používají nejenom pro odhad průměru. Také u jiných bodových odhadů můžeme zjistit jejich směrodatnou chybu. Navíc má bodový odhad často asymptoticky normální rozdělení. V takovém případě použijeme obecný vzorec pro získání asymptoticky platného intervalu spolehlivosti ve tvaru:

$$\text{bodový odhad} \pm \text{koeficient spolehlivosti pro danou hladinu} \times \\ \times \text{směrodatná chyba odhadu}$$

Stejně jako směrodatná chyba průměru jsou směrodatné chyby jiných bodových odhadů nepřímo úměrné odmocnině z rozsahu výběru ( $n$ ).

### Interval spolehlivosti pro $\mu$ při neznámém $\sigma$

V případě, že neznáme směrodatnou odchylku  $\sigma$  náhodné proměnné, musíme k získání přesného intervalu spolehlivosti pro průměr  $\mu$  použít k určení koeficientu spolehlivosti tabulky Studentova  $t$ -rozdělení, s nímž jsme se seznámili v kapitole 4.6.2. Opět předpokládáme, že náhodná proměnná je normálně rozdělená. Ve vzorcích nahrazujeme parametr  $\sigma_{\bar{x}}$  výběrovou směrodatnou chybou  $s_{\bar{x}}$ . Interval spolehlivosti má tvar

$$(\bar{x} - t_{1-\alpha/2}s_{\bar{x}}; \bar{x} + t_{1-\alpha/2}s_{\bar{x}}),$$

kde ve vzorci použijeme kvantil  $t$ -rozdělení se stupni volnosti  $n - 1$  a hladinou  $1 - \alpha/2$ .

### PŘÍKLAD 5.4

#### Interval spolehlivosti pro střední hodnotu při neznámém rozptylu

Vycházíme ze zadání v předchozím příkladu 5.3. Provedli jsme měření u 9 náhodně vybraných žáků a získali jsme průměr 112,8 a směrodatnou odchylku 9. Vypočítáme 95% interval spolehlivosti pro střední hodnotu psychologického parametru  $G$  dětí na škole. Při hledání v tabulce  $t$ -rozdělení použijeme 8 stupňů volnosti a zjistíme, že 97,5% percentil  $t$ -rozdělení má hodnotu 2,306. Výsledný interval spolehlivosti je širší než v předchozím příkladu, kdy jsme předpokládali znalost rozptylu:

$$(112,8 - 2,306 \times 5; 112,8 + 2,306 \times 5) = (102,3; 124,3)$$

Projevuje se tak nejistota při odhadu směrodatné odchylky.

### 5.2.3 Potřebný počet pozorování

Uživatel statistiky nikdy neplánuje sběr dat, aniž by zároveň neuvažoval o použití statistické inference. Vhodný počet pozorování zajišťuje, že získáme odhady pomocí intervalů spolehlivosti s dostatečnou přesností a spolehlivostí. Za velikost chyby odhadování pomocí intervalu spolehlivosti, kterou označíme  $\Delta$ , budeme považovat polovinu délky tohoto intervalu. Jedná se o vzdálenost vypočteného průměru od meze intervalu spolehlivosti. Ze zápisu výpočtu intervalu spolehlivosti pro průměr náhodné proměnné s normálním rozdělením a se známou směrodatnou odchylkou plyne, že velikost chyby  $\Delta$  má hodnotu  $\Delta = z_{1-\alpha/2}\sigma_{\bar{x}}$ , resp.  $\Delta = z_{1-\alpha/2}\sigma/\sqrt{n}$ . Abychom získali intervalový odhad s požadovanou hodnotou přesnosti  $\Delta$  a spolehlivostí, dosadíme do vzorce příslušnou hodnotu  $z_{1-\alpha/2}$  pro zvolenou spolehlivost a vyřešíme rovnici vzhledem k rozsahu výběru  $n$ .

Interval spolehlivosti pro průměr bude mít specifikovanou velikost  $\Delta$ , jestliže zvolíme výběr o velikosti  $n$ , kde  $n$  se vypočte podle rovnice

$$n = \left( \frac{z_{1-\alpha/2}\sigma}{\Delta} \right)^2.$$

### PŘÍKLAD 5.5

#### Stanovení rozsahu výběru potřebného pro dosažení požadované spolehlivosti

Vraťme se k situaci z příkladů 5.3 a 5.4. Chceme získat odhad průměru rozdělení testu  $G$  ve zkoumané populaci s přesností 3 a víme, že směrodatná odchylka hodnot testu  $G$  v populaci je 15. Jaký musíme zvolit rozsah výběru  $n$ , jestliže průměr chceme odhadnout se spolehlivostí 95%? V tomto případě má koeficient z hodnotu 1,96 (97,5% kvantil normální rozdělení). Dosadíme příslušná čísla do vzorce a dostaneme

$$n = \left( \frac{1,96 \times 15}{3} \right)^2 \doteq 100.$$

Vypočtenou hodnotu zaokrouhlujeme na nejbližší vyšší celé číslo.

## 5.2.4 Výhody intervalů spolehlivosti

Mnoho těžkostí spojených s usuzováním pomocí statistiky je možné zmírnit nebo se jím vyhnout, jestliže budeme více používat intervaly spolehlivosti. Jejich výhody jsou následující:

- Šířka intervalu spolehlivosti charakterizuje přesnost odhadu parametru a indikuje potřebu dalších pozorování.
- Interval spolehlivosti poskytuje informaci o velikosti diference nebo odchylky od normy. Tato informace se vytrácí, pokud používáme jenom testy významnosti.
- Pokud provádíme rozhodnutí o platnosti nulové hypotézy ve vztahu k určitému parametru, interval spolehlivosti obsahuje dostatek informací k provedení tohoto rozhodnutí.
- Intervaly spolehlivosti se uplatňují v metaanalýze. Výsledky výzkumů lze popsat a porovnat pomocí intervalů spolehlivost mnohem lépe než pomocí  $p$ -hodnot nebo hvězdiček a jiných symbolů, které pouze vypovídají o tom, zda byla kritická mez testu překročena, nebo ne.
- Velmi často jsou intervaly spolehlivosti (jako v příkladu 5.4) aproximativně symetrické kolem bodového odhadu a meze odpovídají různým percentilům normálního rozdělení se směrodatnou chybou  $SE$ . V tomto případě se udává zkráceně pouze bodový odhad a směrodatná chyba odhadu  $SE$  pro přehled, jaké možné intervaly lze pomocí těchto hodnot sestrojit. Směrodatné chyby  $SE$  některých důležitých výběrových charakteristik uvádíme v tabulce 5.3.
- Přesná interpretace významu intervalu spolehlivosti je dána jeho vlastností při hypotetickém opakování celého pokusu. Rozdíl mezi spolehlivostí a pravděpodobnostním tvrzením, že parametr je s pravděpodobností  $1 - \alpha$  v nalezeném intervalu spolehlivosti, je dosti subtilní a v některých případech nedůležitý.
- Také pracujeme s intervaly pro predikci budoucích dat, jež vznikají v náhodném systému. Predikční intervaly a intervaly spolehlivosti musíme odlišovat. Například při predikci náhodné proměnné s přibližným průměrem  $m$  a směrodatnou odchylkou  $s$  lze použít interval  $(m - 2s, m + 2s)$  jako 95% predikční interval, protože se v něm bude nacházet přibližně 95% budoucích pozorování, pokud tato proměnná má normální rozdělení.

Tab. 5.3 Směrodatné chyby empirických charakteristik

Výběrová charakteristika	Směrodatná chyba $SE$
medián $Me$	$1,25s / \sqrt{n}$
směrodatná odchylka $s$	$s \sqrt{2/n}$
percentil $\bar{x} + zs$	$s \sqrt{(2+z)/(2n)}$
šíkmost $S_1$	$\sqrt{6/n}$
špičatost $S_2$	$\sqrt{24/n}$
$\hat{p}$	$\sqrt{(1-p)p/n}$

Poznámka: Vztahy kromě posledního platí při normálním rozdělení a větším počtu pozorování. U odhadu percentilu je z odpovídající percentil rozdělení  $N(0; 1)$ .

## 5.3 Testy významnosti

Procedury pro statistické testování se často používají při analýze dat. Na druhé straně je však tato oblast statistiky mnohdy špatně pochopena. Testování hypotéz je forma statistického usuzování, které hledá doporučení ve formě „ano“ nebo „ne“ na určitém způsobem formulované otázky. Například se můžeme zeptat, zda průměrný krevní tlak v určité subpopulaci je, anebo není nižší než průměrný krevní tlak v celé populaci nebo zda určitá forma výuky zlepšuje úspěšnost žáků v pedagogickém testu. V těchto případech se požaduje odpověď „ano“ nebo „ne“ a testuje se určité hypotetické tvrzení pomocí dat, jež máme k dispozici. Popíšeme podrobněji dvě situace z odlišných oblastí.

### PŘÍKLAD 5.6

#### Dvě situace uplatnění statistického testování hypotéz

V kapitole o teorii pravděpodobnosti (kap. 4.1) jsme popsali koeficienty pro hodnocení diagnostické kvality medicínského testu. Rozhodnutí pomocí diagnostického testu se velmi často provádí na základě jednoho měření. Výsledkem testu si lékař zodpovídá otázku: je tento pacient zdravý, nebo existuje podezření, že má určitou nemoc? Prakticky provádí statistický test „nulové hypotézy“, že pacient je zdravý. Tento „statistický test“ se dokonce opírá pouze o jedno měření. Čtenář se může pokusit po přečtení tohoto odstavce jako cvičení přeložit výrazy „specifita“ nebo „senzitivita“ medicínského testu do jazyka statistického testování hypotéz.

Hvězda místního basketbalového týmu tvrdí, že má v trestné střelbě úspěšnost 80 %. Požádali jsme ho: „Ukaž nám to.“ Hráč provede 20 pokusů a povede se mu jich 8. Ptáme

se: „Jestliže má úspěšnost 80 %, je možné, aby z 20 pokusů dal pouze 8 košů?“ Podíváme se do kumulativních tabulek binomického rozdělení a po chvíli pátrání hráči řekneme, že tato série nebo ještě horší série má za předpokladu jím uváděné úspěšnosti tak malou pravděpodobnost, že jeho tvrzení nemůže odpovídat pravdě. Dodejme, že jsme takto provedli statistický test parametru binomického rozdělení. Jako testovací statistiku (oporu našeho testu) jsme použili přímo počet úspěšných pokusů, takže jsme naštěstí nemuseli nic počítat. Při sobě jsme však měli vhodnou statistickou tabulkou.

Probereme dva obecné přístupy k testování hypotéz, nejdříve Fisherův přístup následně jeho modifikaci podle Neymana a Pearsona. V této kapitole budeme proces statistického testování demonstrovat na jednoduchém zkoumání hypotézy o parametru  $\mu$  normální rozdělení.

### 5.3.1 Kroky při testování hypotézy

Proceduru testování hypotézy lze schematicky rozložit do následujících kroků:

1. Formulace výzkumné otázky ve formě nulové a alternativní statistické hypotézy.
  2. Zvolení přijatelné úrovně chyby rozhodování.
  3. Vypočtení testovací statistiky.
  4. Doporučení.

Probereme podrobněji jednotlivé kroky

#### Krok 1: Určení statistické hypotézy

První krok při statistickém testování spočívá ve formulování nulové, resp. alternativní hypotézy. Pro tyto hypotézy používáme označení  $H_0$ , resp.  $H_1$ . Hypotéza se obvykle týkají nějakého parametru rozdělení náhodné proměnné, např. průměru  $\mu$ .

**Nulová hypotéza**  $H_0$  je tvrzení, které obvykle deklaruje „žádny rozdíl“ (tj. že kdykoli nalezený rozdíl lze přičíst přirozené variabilitě dat). To je hypotéza, kterou by výzkumník rád spíše zamítl. Tuto hypotézu také lze vymezit způsobem „rozdíl nedosahuje hodnoty  $\Delta$ “, což je někdy oprávněnější v důsledku úvah o „praktické významnosti“ intervence, ošetření apod.

**Alternativní hypotéza**  $H_1$  znamená situaci, kdy nulová hypotéza  $H_0$  neplatí. Obvykle se vyjadřuje jako „existence diference“ mezi skupinami nebo „existence závislosti“ mezi proměnnými. Nemusí jít o přesný logický opak nulové hypotézy, protože někdy máme důvod pracovat s tzv. jednostrannou alternativní hypotézou (jestliže nulová hypotéza říká, že neexistuje rozdíl mezi středními hodnotami).

pro dvě populace, pak jednostranná alternativní hypotéza může např. tvrdit, že druhá populace má střední hodnotu vyšší). Pokud jsme třeba ve vztahu k „praktické významnosti“ formulovali nulovou hypotézu ve tvaru „rozdíl nedosahuje hodnoty  $\Delta$ “, uvažujeme alternativní hypotézu ve formě „existuje diference větší než  $\Delta$ “.

Dokud nedokážeme opak, předpokládáme, že platí nulová hypotéza.

#### Krok 2: Určení hladiny chyby $\alpha$

Určíme **hladinu významnosti** alfa ( $\alpha$ ), což je pravděpodobnost, že se zamítne nulová hypotéza, ačkoliv ona platí. Tato hladina odpovídá míře ochoty výzkumníka smířit se s výskytom této chyby. Pochopitelně se hladina  $\alpha$  volí velmi malá, např. 0,05 nebo 0,01.

### Krok 3: Výpočet testovací statistiky

Z dat se vypočítá testovací statistika, která slouží jako základ pro provedení úvah o výsledném doporučení. Existuje mnoho testovacích statistik, výpočet závisí na povaze dat a hypotéze. Pro testování průměru, relativních četností a v mnoha dalších případech se používá jako testovací statistika standardizovaná vzdálenost odhadu od nulové hypotézy  $H_0$ . Testovací statistika má v těchto případech obecný tvar:

$$\text{testovací statistika} = \frac{\text{bodový odhad} - \text{hypotetická hodnota}}{\text{směrodatná chyba odhadu}}$$

#### Krok 4: Doporučení

Formulujeme závěr testování. Provádíme to dvěma způsoby. Srovnáme testovací statistiku s kritickoumezí nebo ji převedeme do pravděpodobnostní škály na tzv. hodnotu významnosti  $p$ . Oba způsoby popíšeme podobně.

Hodnota  $p$  odpovídá na otázku: „Jestliže nulová hypotéza platí, jaká je pravděpodobnost, že získáme právě vypočítanou hodnotu nebo ještě neobvyklejší hodnotu testovací statistiky?“ Hodnota  $p$  tedy kvantifikuje pravděpodobnost reálnice hodnoty testovací statistiky, pokud nulová hypotéza platí. Jestliže je malá, je zde doklad, že nulová hypotéza neplatí. Takže pravidlo pro volbu doporučení je jednoduché: Jestliže  $p$ -hodnota je menší než hladina  $\alpha$  nebo se jí rovná, data přinášejí evidenci pro zamítnutí  $H_0$ . Jestliže  $p$ -hodnota je větší než  $\alpha$ ,  $H_0$  se ponechává k dalšímu zkoumání.

Zopakujeme základní principy interpretace statistického testu pomocí dosažené  $p$ -hladiny:

## PŘEHLED STATISTICKÝCH METOD

1. Chceme potvrdit, že existuje nějaký rozdíl (alternativní hypotéza,  $H_1$ ).
  2. Postupujeme však tak, že zkoumáme předpoklad, že žádný rozdíl neexistuje (nulová hypotéza,  $H_0$ ), a hodnotíme sílu dokladů proti tomuto předpokladu.
  3. Konkrétně to znamená, že spočítáme pomocí dat pravděpodobnost  $p = P$  (stejná data nebo ještě neobvyklejší) za předpokladu, že platí nulová hypotéza.
  4. Jestliže hodnota  $p$  je malá, jedná se o evidenci proti nulové hypotéze.
  5. Jestliže hodnota  $p$  je větší, jedná se o evidenci připouštějící platnost nulovou hypotézu.

V souvislosti s tímto způsobem uvažování je nutné podotknout, že nezamítnutí nulové hypotézy neznamená její důkaz. Spíše to znamená, že nemáme dostatečnou evidence k jejímu zamítnutí. To je důležitý rozdíl.

Při dané hypotéze  $H_0$  a datech  $D$  počítáme  $p$ -hodnotu jako pravděpodobnost, že získáme data nejméně tak odporující (kontradiktorní) hypotéze  $H_0$ , jako jsou naše data  $D$ , za předpokladu, že platí hypotéza  $H_0$ .

Druhý způsob posouzení testovací statistiky je velmi názorný. Představme si spočívá v přímém srovnání testovací statistiky s **kritickou mezí**  $M_\alpha$ , která se určuje v závislosti na zvolené hladině (hodnotě) významnosti  $\alpha$ . Kritická meze určuje **kritickou oblast**, resp. **oblast zamítnutí**. Jestliže se hodnota testovací statistiky ocitne uvnitř kritické oblasti, znamená to, že existuje evidence pro zamítnutí nulové hypotézy. Tento způsob interpretace nahrazuje výpočet dosažené  $p$ -hodnoty významnosti odpovídající transformaci testovací statistiky do pravděpodobnostní škály. Jestliže testovací statistika je uvnitř kritické oblasti, pak dosažená  $p$ -hodnota je menší než příslušná hladina významnosti.

Testovací statistiku interpretujeme často jako míru nekompatibility dat  $D$  s nulovou hypotézou, resp. jako statistickou vzdálenost  $V(D, H_0)$  dat  $D$  od nulové hypotézy  $H_0$ . Pokud získaná konfigurace dat je ve shodě s nulovou hypotézou (např. vypočtený průměr se rovná hypotetickému průměru nebo dva vypočtené průměry si jsou skoro rovny), statistická vzdálenost  $V(D, H_0)$  je ohodnocená testovací statistikou je malá. Jestliže naopak získaná data jsou nepodobná datům vým konfiguracím očekávaným za platnosti nulové hypotézy, vypočtená statistická vzdálenost  $V(D, H_0)$  je veliká. Kritická vzdálenost je určena kritickoumezí  $M_\alpha$ . Jestliže  $V(D, H_0)$  je menší než kritická vzdálenost  $M_\alpha$  (kritickámez), získaná datová konfigurace indikuje neplatnost nulové hypotézy. Kritická vzdálenost  $M_\alpha$  se volí tak, že za platnosti nulové hypotézy je vzdálenost  $V(D, H_0)$  větší než  $M_\alpha$  s pravděpodobností pouze  $\alpha$ .

Proces konstrukce statistického testu pro nějaký zadaný obecný problém má význam.

- a) Sestrojení míry, jež charakterizuje, jak jsou data v současnosti s nulovou hypotézou se zohledňuje tvar  $H_1$ . Tuto míru počítáme z dat, proto se jedná o statistiku a říkáme **testovací statistika**. Hodnota testovací statistiky měří kompatibilitu dat  $D$  s modelem, je určen nulovou hypotézou  $H_0$ . Testovací statistiku lze považovat za míru nepodobnosti  $V(D, H_0)$  datové konfigurace  $D$  s konfigurací, již určuje nulová hypotéza  $H_0$ .

- b) Hledání výběrového rozdělení testovací statistiky za platnosti  $H_0$ . Toto rozdělení nám slouží k určení kritických mezd  $M_\alpha$ , resp. kritické oblasti pro testovací statistiku a k výpočtu pravděpodobnosti výskytu aktuální hodnoty testovací statistiky nebo hodnoty ještě extrémnější za platnosti nulové hypotézy. Co se považuje za extrémní hodnotu, závisí na alternativní hypotéze  $H_1$ .

Doporučujeme analyzovat každou testovací statistiku a vyjádřit, jak je v ní realizován princip určení míry nepodobnosti dat  $D$  a nulové hypotézy  $H_0$ .

Existuje mnoho statistických testů. V další části této kapitoly se věnujeme osvětlení konstrukce jednoduchého testu o průměru pro situaci jednoho výběru. Demonstrujeme na něm koncept jednostranného a dvoustranného testování.

### 5.3.2 Testování průměru jednostranným z-testem

Tomuto způsobu testování říkáme také testování při jednostranné alternativě. Co to znamená? Všimneme si situace, kdy nás zajímá test o průměrné hodnotě  $\mu$  za předpokladu, že známe  $\sigma$  a sledovaná náhodná proměnná má normální rozdělení. Testu se říká  $z$ -test, protože se opírá o standardizovanou hodnotu rozdílu mezi výběrovým průměrem a hodnotou  $\mu$ .

Abychom ilustrovali testovací proceduru, vrátíme se k příkladu 5.3 z odstavce o určování intervalu spolehlivosti. Předpokládáme, že hodnotíme polohu rozdělení zvoleného znalostního ukazatele  $G$  na určité škole. Z předchozích výzkumů je známo, že v celé populaci dětí v určitém věkovém pásmu jsou hodnoty  $G$  ukazatele normálně rozděleny se střední hodnotou 100 a směrodatnou odchylikou 15.

1. **Nulová a alternativní hypotéza:** Obě hypotézy mohou mít jednu ze tří forem: pravostranná hypotéza, levostranná hypotéza a dvoustranná hypotéza. Například se můžeme ptát, zda průměr ukazatele  $G$  na naší škole (pro dané věkové pásmo) je vyšší než teoretický průměr v celé populaci pro dané věkové pásmo. Pak se jedná o test s pravostrannou alternativní hypotézou. Jestliže  $\mu_0$  představuje očekávanou hodnotu za platnosti nulové hypotézy, pak nulová a alternativní hypotéza pro tuto podmíinku má tvar:

$$H_0: \mu \leq \mu_0 \quad \text{proti} \quad H_1: \mu > \mu_0$$

Také však můžeme zkoumat, zda průměr je nižší než očekávaný. Pak se jedná o levostrannou alternativu. Nulová a alternativní hypotéza pak mají tvar:

$H_0: \mu \geq \mu_0$  proti  $H_1: \mu < \mu_0$

Konečně za alternativu prostě považujeme průměr, jenž je odlišný (větší nebo menší) od nulové hypotézy, což vyjádříme takto:

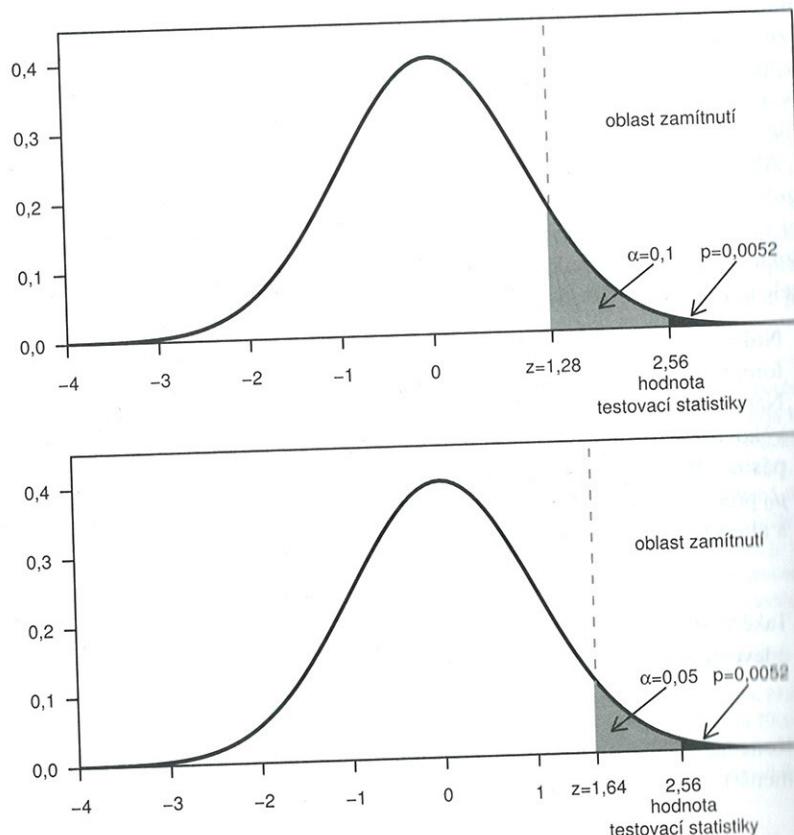
$$H_0: \mu = \mu_0 \quad \text{proti} \quad H_1: \mu \neq \mu_0$$

Test pro tento případ se nazývá dvoustranný test.  
Například chceme vědět, zda na naší škole průměrná hodnota znalostního ukazatele  $G$  je větší než hodnota 100. Jedná se o jednostrannou testovací situaci, v níž posuzujeme, zda průměr leží napravo od očekávané hodnoty. Nulovou a alternativní hypotézu zapíšeme ve tvaru podmínek:

$$H_0: \mu \leq 100 \quad \text{proti} \quad H_1: \mu > 100$$

Dále budeme předpokládat tuto formulaci obou hypotéz.

Obr. 5.3 Rozdělení testovací statistiky a plochy odpovídající statistické významnosti jednostranného  $z$ -testu



2. **Alfa ( $\alpha$ ):** Tuto hodnotu volí výzkumník. Zvolíme standardní hodnotu 0,05.
3. **Testová statistika:** Testová statistika se pro tento případ spočte jako  $z$ -hodnota:

$$z = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Předpokládáme, že výběr z naší školy měl rozsah 9, pro který jsme spočítali průměr  $\bar{x}_G = 112,9$ . Také předpokládáme, že směrodatná odchylka pro naší školu je stejná jako v celé populaci.  $SE$  má tedy hodnotu  $15/3 = 5$  a testovací statistika  $z = (112,8 - 100)/5 = 2,56$ .

4. **Závěr:** Testovací statistiku srovnáme s kritickoumezí. Při testování na hladině významnosti 0,1 (resp. 0,05) srovnáváme  $z$ -statistiku s kritickoumezí 1,28 (resp. 1,64). Obě dvě meze jsou znázorněny na obrázku 5.3 spolu s plochami, jejichž velikost odpovídá hladině významnosti. Pokud  $z$ -statistiku převádíme na  $p$ -hodnotu, hledáme plochu pod normální křivkou pro  $x$ -hodnoty nad číslem 2,58. Nalezneme hodnotu  $p = 0,0052$ . Jelikož  $p < 0,01$ , nulovou hypotézu lze zamítnout na hladině 0,01.

Uzavíráme, že existuje dostatečná evidence pro tvrzení, že děti z naší školy mají vyšší průměr ukazatele  $G$ , než je jeho průměrná hodnota v populaci. Poznámejme, že nás závěr se týká průměru ukazatele  $G$  a ne jeho konkrétní hodnoty u zvoleného dítěte.

Pro základní orientaci uvádíme tabulku s hodnotami  $P(Z > z)$  pro standardizované normální rozdělení (tab. 5.4).

### 5.3.3 Testování průměru dvoustranným $z$ -testem

Předchozí ilustrativní příklad se týkal jednostranné alternativy, která vyjadřovala, že nás zajímá jenom jeden směr odchylky od nulové hypotézy. Pokud nám záleží na odchylce od nulové hypotézy v obou směrech, nahoru i dolů od  $\mu_0$ , použijeme dvoustranný test.

Obecně lze říci, že dvoustranné testy se používají častěji než jednostranné testy. Obvykle je totiž naším cílem ohodnotit odchylky od nulové hypotézy bez ohledu na to, jaké byly naše alternativní hypotézy před pokusem. Proto jsou dvoustranné testy standardem. Vede to někdy k opomíjení správné formulace jednostranného testu v případech, kdy by to bylo užitečné. Použijeme data z předešlého příkladu a ukážeme, jak se na nich aplikuje dvoustranný test.

1. **Nulová a alternativní hypotéza:** Ověřujeme, zda 9 dětí z našeho příkladu má průměr znalostního ukazatele  $G$  statisticky odlišný od průměru v populaci (zda průměr je statisticky menší nebo větší než populační hodnota 100). To je

Tab. 5.4 Horní percentilové hodnoty pro standardizované normální rozdělení

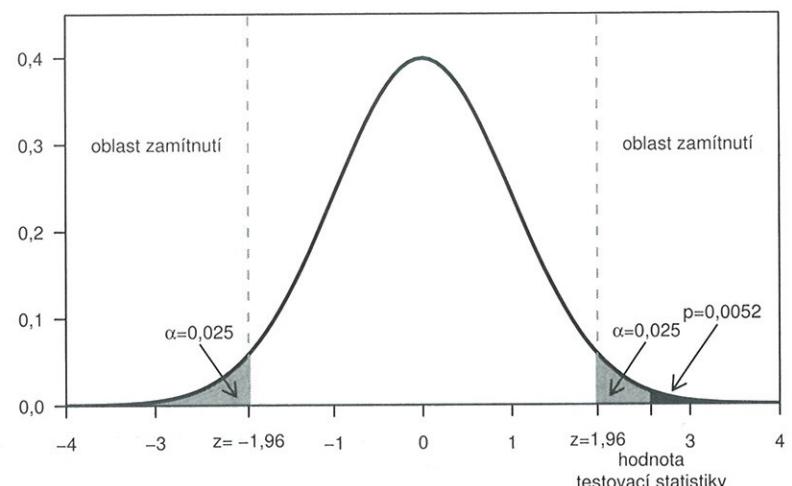
$q [\%]$	$z$						
50	0,000	14	1,080	2,3	1,995	0,7	2,457
45	0,126	13	1,126	2,2	2,014	0,6	2,512
40	0,253	12	1,175	2,1	2,034	0,5	2,576
35	0,385	11	1,227	2,0	2,054	0,4	2,652
30	0,524	10	1,282	1,9	2,075	0,3	2,748
25	0,674	9	1,341	1,8	2,097	0,2	2,878
24	0,706	8	1,405	1,7	2,120	0,1	3,090
23	0,739	7	1,476	1,6	2,144	0,09	3,121
22	0,772	6	1,555	1,5	2,170	0,08	3,156
21	0,806	5	1,645	1,4	2,197	0,07	3,195
20	0,842	4,5	1,695	1,3	2,226	0,06	3,239
19	0,878	4	1,751	1,2	2,257	0,05	3,291
18	0,915	3,5	1,812	1,1	2,290	0,04	3,353
17	0,954	3	1,881	1,0	2,326	0,03	3,432
16	0,994	2,5	1,960	0,9	2,366	0,02	3,540
15	1,036	2,4	1,977	0,8	2,409	0,01	3,719

Tabulka udává hodnoty  $z$ , pro které  $P(Z > z) = q \%$ , přičemž rozdelení  $Z$  je  $N(0, 1)$ .

ekvivalentní tvrzení, že průměr ukazatele  $G$  na naší škole se liší od hodnoty v populaci. Naše testová situace má tedy tvar:

$$H_0: \mu = 100 \quad \text{proti} \quad H_1: \mu \neq 100$$

2. **Úroveň alfa ( $\alpha$ ):** Opět použijeme  $\alpha = 0,05$ .
3. **Testová statistika:** Testovou statistiku vypočítáme podle stejného předpisu, tedy  $z = (112,8 - 100)/5 = 2,56$ .
4. **Závěr:** Testovací statistiku srovnáme s kritickou mezí. Při testování na hladině významnosti 0,05 srovnáme  $z$ -statistiku na rozdíl od předchozího případu s kritickoumezí 1,96. Tatomez spolu s plochami pod hustotou normálního rozdelení, jejichž součet odpovídá hladině významnosti pro dvoustranný test, je znázorněna na obrázku 5.4. Testovací statistika je v absolutní hodnotě větší než kritickámez, což znamená, že existuje evidence pro zamítnutí nulové hypotézy.

Obr. 5.4 Rozdělení testovací statistiky a plochy odpovídající statistické významnosti dvoustranného  $z$ -testu

Abychom vypočítali  $p$ -hodnotu dvoustranného testu, musíme uvažovat oba konci rozdělení statistiky. Proto vypočtenou plochu z předchozího příkladu pro jednostranný test vynásobíme dvěma. Plošky pod hustotou na obou stranách rozdělení mají totiž velikost 0,0052. Dohromady je jejich velikost  $0,0052 + 0,0052 = 0,0104$ . To reprezentuje  $p$ -hodnotu pro náš problém.

### 5.3.4 Chybné interpretace testů nulové hypotézy

Výsledky statistických testů se často špatně interpretují. Proto je užitečné na některé chyby předem upozornit, abychom se jim snáze vyhnuli.

- Jestliže test neindikuje zamítnutí nulové hypotézy, je nesprávné nulovou hypotézu přijmout jako definitivně pravdivou. Správně můžeme pouze prohlásit, že není dostatek dokladů (evidence) pro zamítnutí nulové hypotézy.
- Hodnota  $p$  se nesprávně interpretuje jako pravděpodobnost, že nulová hypotéza platí. Ve skutečnosti je  $p$  pravděpodobnost výskytu spočtené hodnoty testovací statistiky (a hodnot extrémnějších, viz obr. 5.4) za modelových předpokladů, že platí nulová hypotéza.
- Není pravda, že by hladina významnosti  $\alpha = 0,05$  měla jasné opodstatnění a byla jednou provždy daná. Tato hodnota je zvolena na základě konvence.

- Neplatí, že by menší  $p$ -hodnoty znamenaly silnější vědeckou evidenci než větší  $p$ -hodnoty. Ve skutečnosti  $p$ -hodnoty nevypovídají nic o síle evidence – jsou silně závislé na rozsahu výběru.
- Netvrďme, že data ukazují, že teorie platí/neplatí. Správnější je říci, že data podporují nebo nepodporují rozhodnutí o zamítnutí platnosti nulové hypotézy.
- „Statistická významnost“ neindikuje „vědeckou důležitost“ výsledků, ale týká se zamítnutí statistické nulové hypotézy. Teprve další diskuse vyjasní, do jaké míry jsou zjištěné výsledky „vědecky významné“ nebo „prakticky významné“.

### 5.3.5 Vztah testování hypotéz a intervalů spolehlivosti

Existuje formální vztah ekvivalence mezi intervalem spolehlivosti a testem významnosti parametru. Na interval spolehlivosti s hladinou spolehlivosti  $1 - \alpha$  se můžeme dívat jako na množinu možných hodnot parametru, které by na hladině významnosti  $\alpha$  nebyly považovány příslušným testem významnosti za nekonzistentní s posuzovanými daty. To znamená, že interval spolehlivosti lze použít při testování významnosti. Pokud hodnota hypotetického parametru neleží v intervalu spolehlivosti, hypotézu lze zamítnout na hladině významnosti  $\alpha$ .

Důkaz tohoto tvrzení není složitý. Kroky důkazu ukážeme pro interval spolehlivosti a dvoustranný  $z$ -test pro parametr  $\mu_0$ , jestliže známe směrodatnou odchylku  $\sigma$  náhodné proměnné a tedy i směrodatnou chybu průměru  $\sigma_{\bar{x}}$ . Interval spolehlivosti je určen kritickou hodnotou  $z^*$  pro danou hladinu spolehlivosti. Jestliže  $H_0$  zamítlame, pak platí:

$$\begin{aligned} \mu_0 \notin (\bar{x} - z^* \cdot \sigma_{\bar{x}}, \bar{x} + z^* \cdot \sigma_{\bar{x}}) &\Leftrightarrow \\ \mu_0 < \bar{x} - z^* \cdot \sigma_{\bar{x}} \text{ nebo } \mu_0 > \bar{x} + z^* \cdot \sigma_{\bar{x}} &\Leftrightarrow \\ -\mu_0 > z^* \cdot \sigma_{\bar{x}} - \bar{x} \text{ nebo } -\mu_0 < -z^* \cdot \sigma_{\bar{x}} - \bar{x} &\Leftrightarrow \\ \bar{x} - \mu_0 > z^* \cdot \sigma_{\bar{x}} \text{ nebo } \bar{x} - \mu_0 < -z^* \cdot \sigma_{\bar{x}} &\Leftrightarrow \\ \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}} > z^* \text{ nebo } \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}} < -z^* &\Leftrightarrow \\ \frac{|\bar{x} - \mu_0|}{\sigma_{\bar{x}}} > z^* & \end{aligned}$$

Používání intervalu spolehlivosti při testování má dvě přednosti. Intervalový odhad je vyjádřen v měřítku posuzovaného parametru. Také získáváme představu o přesnosti jeho odhadu a o tom, zda zvolený rozsah výběru  $n$  byl dostatečně veliký.

Jestliže odhadujeme intervalovým odhadem parametr ve dvou populacích, tak platí:

1. Pokud se odhadovací intervaly nepřekrývají, zamítneme hypotézu rovnosti parametrů, a to alespoň na hladině, která odpovídá spolehlivosti použitých odhadovacích intervalů.
2. Pokud se odhadovací intervaly překrývají, nemůžeme provést žádné rozhodnutí bez dalších výpočtů.

### 5.3.6 Test jako rozhodování

V teorii testování hypotéz vznikl koncept chyb I. a II. druhu o něco později než její základy. Tento příspěvek k hodnocení testovacích procedur navrhla dvojice statistiků Neyman a Pearson, kteří doplnili uvažování o statistickém testování pojmy, jež se rozvinuly v matematické teorii rozhodování v kontextu matematické ekonomie.

Průkopník statistických metod testování R. A. Fisher se opíral o koncept  $p$ -hodnot. Podstata Fisherova přístupu spočívá v tom, že zamítnutí nulové hypotézy má vycházet z malých  $p$ -hodnot, jež poukazují na neslučitelnost dat s nulovou hypotézou. Mnoho výzkumníků dokumentuje své výsledky právě tímto způsobem. Přitom srovnávají  $p$ -hodnoty s hraničními pravděpodobnostmi jako 0,05, 0,01 nebo dokonce 0,001.

Při popisu testovací procedury v předchozím odstavci jsme zatím uvažovali jenom falešné zamítnutí  $H_0$ . Je však také důležité uvažovat falešné ponechání nulové hypotézy v platnosti. Existují tedy dvě možnosti chyby:

1. **chyba I. druhu** – nulová hypotéza platí, ale zamítnete se;
2. **chyba II. druhu** – nulová hypotéza neplatí, ale nezamítnete se.

Při testování hypotéz proto mohou nastat čtyři možnosti, které popisuje tabulka 5.5. Jestliže přirovnáme chybu I. druhu k falešně pozitivnímu výsledku při medicínském testování, pak chyba II. druhu odpovídá falešně negativnímu výsledku v medicíně, kdy pacient je nemocen, ale test to neodhalí. Pravděpo-

Tab. 5.5 Schéma testování hypotéz

		Závěr testu	
		$H_0$ platí	$H_0$ neplatí
Skutečnost	$H_0$ platí	správný	chyba I. druhu
	$H_0$ neplatí	chyba II. druhu	správný

dobnost chyby I. druhu je podmíněná pravděpodobnost, že zamítneme nulovou hypotézu za předpokladu, že platí, a označujeme ji  $\alpha$ . Pravděpodobnost chyby II. druhu je podmíněná pravděpodobnost, že nezamítneme nulovou hypotézu za předpokladu, že neplatí, a označujeme ji  $\beta$ :

$$P(\text{chyba I. druhu} | H_0 \text{ platí}) = \alpha$$

$$P(\text{chyba II. druhu} | H_1 \text{ neplatí}) = \beta$$

Jak jsme již uvedli, hladinu  $\alpha$  obvykle volíme 0,05 nebo 0,01. Konvenční hodnoty pro  $\beta$  jsou 0,2 nebo 0,1.

Také můžeme někdy mluvit o opačných jevech k chybě I., resp. II. druhu, tzn. o podmíněné pravděpodobnosti, že neuděláme chybu I. druhu (spolehlivost testu) nebo že neuděláme chybu II. druhu. Síla testu odpovídá hodnotě ( $1 - \beta$ ). Jedná se tedy o podmíněnou pravděpodobnost, že správně odhalíme testem  $\beta$ . Neplatnost nulové hypotézy. To znamená:

$$P(\text{neuděláme chybu I. druhu} | H_0 \text{ platí}) = 1 - \alpha = \text{"spolehlivost"}$$

$$P(\text{neuděláme chybu II. druhu} | H_1 \text{ neplatí}) = 1 - \beta = \text{"síla testu"}$$

Konvenční hladina pro spolehlivost je 0,95 nebo 0,99 a konvenční hladina pro sílu je 0,8 nebo 0,9. Cílem při testování nulové hypotézy je omezit úrovně pravděpodobnosti chyb I. a II. druhu. Jinými slovy – usilujeme o maximalizaci spolehlivosti a síly testu.

### PŘÍKLAD 5.7

#### Analogie mezi testováním hypotéz a soudním procesem

Přiblížíme testování hypotéz metaforou. Odvoláme se na podobnost statistického usuzování (inference) se soudním procesem, kterou jsme naznačili v kapitole 3.8. Má padnout rozhodnutí, že obžalovaný spáchal/nespáchal zločin. Soudní systém se řídí zásadou, že obžalovaný je nevinen, pokud se nepodaří prokázat opak. Proto formulace hypotéz má tuto podobu:

$$H_0: \text{Obžalovaný je nevinný.}$$

$$H_1: \text{Obžalovaný je vinen.}$$

Různé možnosti vztahu mezi pravdou a rozhodnutím soudu ukazuje tabulka 5.6.

Uvědomujeme si, že chyba I. druhu má pro jedince nedozírné následky. Proto její možnost eliminujeme na nejmenší možnou míru. Soud musí jasně prokázat vinu obžalovaného. Jeho rozhodnutí také podléhají přezkoumání vyšších instancí. Odpovídá to volbě velmi malé hladiny významnosti. V mnoha případech však nevíme zcela přesně, která chyba je pro nás důležitější.

Tab. 5.6 Srovnání situace testování hypotéz a rozhodování soudu

		Závěr soudu	
		Obžalovaný je nevinen	Obžalovaný je vinen
Skutečnost	Obžalovaný je nevinen	správný	chyba I. druhu
	Obžalovaný je vinen	chyba II. druhu	správný

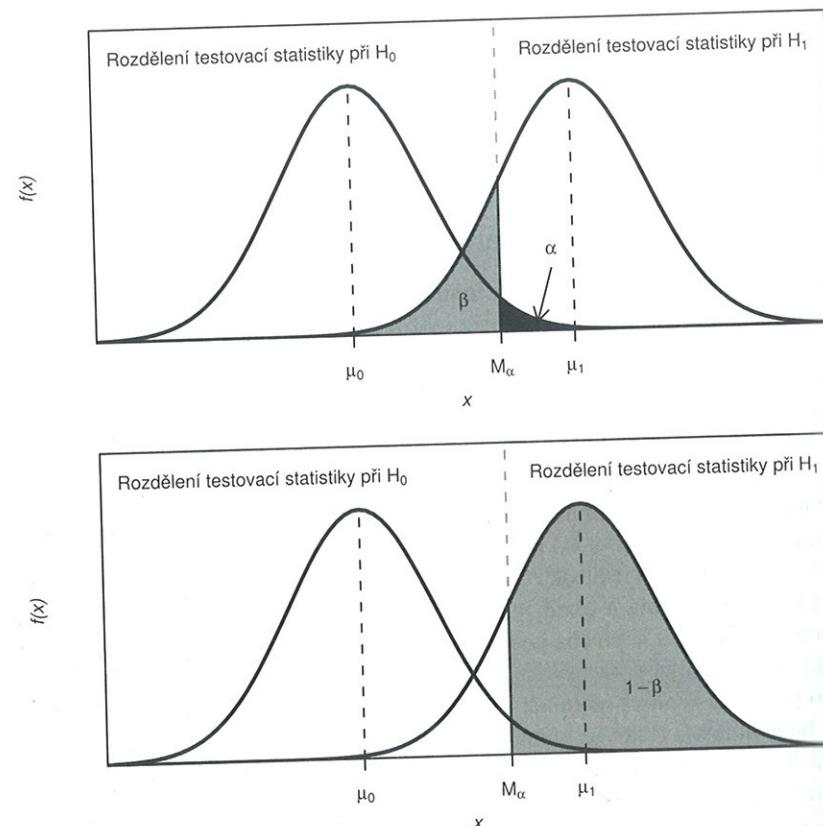
### 5.3.7 Vztah mezi silou testu, počtem pozorování a významnosti

Představme si, že testujeme  $z$ -testem nulovou hypotézu  $\mu_0$  proti zvolené alternativní hodnotě  $\mu_1$ .

Při rozhodování pomocí statistického testu použijeme koncept dvou typů chyb – chyby I. a II. druhu – který jsme zavedli v předchozích odstavcích. Vztah mezi nimi lze přiblížit obrázkem 5.5. Jsou na něm zobrazena rozdělení testovacích statistik za předpokladu platnosti nulové hypotézy (levá zvonovitá křivka) a za platnosti specifické alternativní hypotézy (pravá zvonovitá křivka). Kritická hodnota pro danou hladinu významnosti určuje odpovídající hladinu chyby II. druhu jako plošku nalevo od kritické hodnoty pro rozdělení při alternativní hypotéze, napravo od ní je ploška odpovídající hladině chyby I. druhu, již vymezuje opět kritická hodnota a okraj rozdělení testovací statistiky za platnosti nulové hypotézy. Síla testu je plocha pod křivkou odpovídající alternativní hypotéze napravo od kritické hodnoty. Aktuální hodnota testovací statistiky může být vyšší nebo nižší než kritická hodnota. Za platnosti nulové hypotézy, resp. její alternativy bude hodnota testovací statistiky vyšší než kritická mez s pravděpodobností, která se rovná hladině významnosti, resp. síle testu. Sílu testu bychom rádi maximalizovali. Při daném rozsahu výběru toho dosáhneme, když budeme pohybovat kritickou hodnotou směrem doleva k nižším hodnotám. Je však zřejmé, že tím zvýšujeme hladinu chyby I. druhu. To znamená, že při daném rozsahu výběru je síla testu přímo úměrná hladině významnosti. Hladina významnosti je nepřímo úměrná hladině chyby II. druhu.

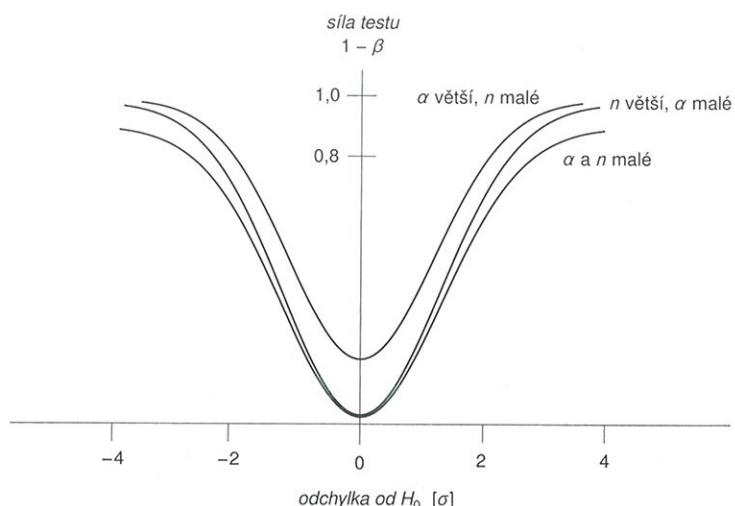
Při různé volbě alternativní hypotézy roste nebo se zmenšuje rozdíl  $\Delta$  mezi hodnotou  $\mu_1$  a hodnotou nulové hypotézy  $\mu_0$ . Z uvedeného lze odvodit, jak se budou měnit hladiny pravděpodobností chyb I. a II. druhu. Při rostoucím  $\Delta$  se bude snižovat hladina  $\beta$  chyby II. druhu a poroste síla testu. Jak vypadá průběh této změny v závislosti na velikosti  $\Delta$  pro dvě různé hladiny významnosti  $\alpha$  a dva rozsahy výběrů, ukazuje obrázek 5.6.

Obr. 5.5 Kritická hodnota a síla testu při dané hodnotě nulové hypotézy  $\mu_0$  a alternativní hypotézy  $\mu_1$



Z obrázku 5.6 je vidět, že s rostoucím rozsahem výběru při dané hladině významnosti a fixní hodnotě  $\Delta$  roste síla testu. Je-li rozsah výběru malý, neodhalíme ani velkou diferenci mezi nulovou hypotézou a aktuální hodnotou parametru. Naopak při dostatečně velkém rozsahu bude pravděpodobnost zamítnutí nulové hypotézy veliká i při malé hodnotě  $\Delta$ .

Obr. 5.6 Křivky síly z-testu pro různé podmínky při dvoustranném testu



### 5.3.8 Hodnocení velikosti účinku

Výsledek testování, ať ve Fisherově, nebo Neyman-Pearsonově tradici, dává důležité informace, ale jsou zde ještě jiné aspekty dat, které je rovněž nutné popsat. Například může být cílem kvantifikovat velikost nebo sílu účinku, jehož se dosáhlo zkoumanou intervencí – a právě tato informace není ve výsledku statistického testu obsažena. Jako by se student dozvěděl, že prošel zkouškou, ale neví, kolik bodů dosáhl. Hodnoty  $p$  o tom moc nevypovídají – jakkoli malý účinek intervence lze při dostatečném počtu měření prokázat, a naopak při malém počtu měření i velký efekt zůstane testem významnosti neodhalen. Přitom nám jde o prokázání nejenom statistické, ale i věcné, praktické, klinické významnosti.

Často proto vycházíme z odhadu koeficientu, který nazýváme **velikost účinku** (effect size). Ten je definován pro jednotlivé charakteristiky různě. Na tomto místě popíšeme pouze jeden případ.

Odhad účinku (effect size,  $ES$ ) ošetření nebo intervence, jenž závisí na velikosti průměru, může mít dvě formy:

- Uvedeme dosaženou diferenci průměrů (pro dvě skupiny, pro spárovaná měření, pro jednu skupinu):

$$ES = \bar{x}_1 - \bar{x}_2, \quad ES = \bar{d}, \quad ES = \bar{x}_1 - \mu$$

- Častěji používáme **Cohenův koeficient účinku  $d$** :

$$ES = d = \frac{\bar{x}_1 - \bar{x}_2}{s}, \quad ES = d = \frac{\bar{d}}{s}, \quad ES = d = \frac{\bar{x}_1 - \mu}{s},$$

kde  $s$  označuje směrodatnou odchylku měření (pozor – NE směrodatnou chybu odhadu). Jestliže jsme např. vypočítali průměry 8,1 a 3,1 a odhadli společnou směrodatnou odchylku náhodné proměnné v obou skupinách 2,3, bude mít Cohenův koeficient hodnotu  $d = (8,1 - 3,1)/2,3 = 2,2$ .

Praktický význam Cohenova návrhu spočívá v tom, že rozdíly se standardizují pomocí směrodatné odchylky. Tím se dosáhne toho, že se mohou srovnávat odchylky v působení intervencí a ošetření, které byly měřeny úplně jinými prostředky (psychologickými testy, laboratorními přístroji). Samozřejmě je tato volnost relativní, Cohen určil pro svůj index  $d$  konvenční hodnoty, jež usnadňují rozhodnutí, kdy můžeme mluvit o velkém efektu. Pokud je  $d$  větší než 0,8, je efekt velký; pro  $d$  z intervalu 0,5–0,8 je efekt střední; efekty podmezí 0,2 považujeme za malé. Je nutné si uvědomit, že posouzení, jaký účinek považovat za veliký, závisí na kontextu. Jak hranice navržené Cohenem, tak různé limity pro korelační koeficient nebo jiné míry účinku jsou podobně jako třeba hodnoty hladin významnosti sice do jisté míry zdůvodněné, ale nemají absolutní platnost.

Ačkoliv o to statistici a výzkumníci usilují, neexistuje ani nemůže existovat jeden koeficient, který by dobře ohodnotil a přehledně vyjádřil účinky (effect size) ve všech situacích. Jestliže posuzujeme sílu asociace mezi dvěma náhodnými proměnnými, používá se většinou jako míra účinku korelační koeficient  $r$  a jiná míra se nehledá. V určitých případech lze přecházet mezi různými mírami účinku pomocí přepočítávacích vzorců.

Podrobnosti k tomuto problému čtenář najde v přehledné práci Blahutové (2000).

### 5.3.9 Přesné a asymptotické testy

Skutečně dosaženou hladinu významnosti  $p$  pro testovací statistiku **zjistíme** přesně pouze v určitých případech. Takové testy nazýváme **přesné testy**. Je to možné, jestliže se např. jedná o  $t$ -test průměru, testy v regresní analýze a analýze

rozptylu nebo test korelačního koeficientu, pokud jsou splněny jejich předpoklady. Přesné testy jsou však spíše výjimkou. Proto je důležité mít ve výbavě také asymptotické testy, u kterých zjištěná  $p$ -hodnota je aspoň asymptoticky přesná s rostoucí velikostí výběru.

U **asymptoticky platných testů** využíváme toho, že se tvar rozdělení použitých testovacích statistik blíží k dobré známým rozdělením (normální nebo  $\chi^2$ ) v důsledku působení centrálního limitního teorému. To značně usnadňuje výpočet hledané  $p$ -hodnoty. Také u kritických mezí se ptáme, zda k nim příslušející hladina významnosti platí přesně, nebo pouze asymptoticky. Zůstává však riziko, že zjištěná  $p$ -hodnota není dostatečně validní, protože rozsah výběru není dostatečně velký. Pro malé výběry proto raději použijeme postup pro zjištění přesné  $p$ -hodnoty, jenž zohledňuje tuto skutečnost. Například při zkoumání hypotéz o relativních četnostech vychází příslušný výpočet přesné  $p$ -hodnoty z binomického rozdělení, jestliže výběr je malý, kdežto asymptotický přístup se opírá o normální rozdělení a zjistíme tak pouze přibližně platnou  $p$ -hodnotu.

V tomto a dalších případech se rozhodujeme podle problému, velikosti výběru a dostupnosti vhodných programů, jakou použijeme testovací statistiku a zda spočítáme přesnou  $p$ -hodnotu, nebo využijeme asymptotické vlastnosti chování testovací statistiky pro odhad získané  $p$ -hodnoty nebo pro nalezení kritické meze s asymptoticky platnou hladinou významnosti.

U testů, u kterých neznáme tak dobře rozdělení testovací statistiky jako u relativních četností, odhadujeme přesnou  $p$ -hodnotu různými metodami. Příslušné výpočty jsou mnohdy značně náročné a bez počítače prakticky neprověditelné. V některých případech se můžeme opřít o tabelované kritické hodnoty pro statistické testy na vybraných hladinách významnosti pro danou testovací statistiku a zvolené rozsahy výběrů. To se týká především neparametrických testů založených na pořadí nebo četnostech.

## 5.4 Neparametrické postupy statistického usuzování

Klasické postupy statistického usuzování obvykle předpokládají normální rozdělení náhodných proměnných. Tento předpoklad – přísně vzato – není skoro nikdy splněn. Proto statistici vyvinuli **procedury nezávislé na rozdělení**, resp. **neparametrické metody**, které tento předpoklad nepotřebují. Jejich výhoda spočívá v tom, že o tvaru rozdělení dat nemusíme mít větší vědomosti. Neparametrické postupy jsou tedy vhodné pro hodnocení ordinálních dat nebo dat naměřených v poměrovém nebo intervalovém měřítku, jež nemají normální rozdělení.

Existují tři hlavní třídy neparametrických procedur. V prvních dvou případech se u metrických dat uplatňuje změna měřítka dat na nižší úroveň.

1. V první skupině metod se zjišťuje, jakou četnost mají odchyly určitého znaménka (počítáme záporné nebo kladné odchyly). Například všem hodnotám ležícím nad určitou mezí se přiřadí kladné znaménko, ostatním záporné znaménko. Tím vlastně intenzivní data převedeme na kvalitativní dichotomická data. Tento typ neparametrických procedur reprezentuje znaménkový test.
2. Ve druhé skupině neparametrických technik se počítá s pořadími, která jsou číselným údajům přiřazena po jejich setřídění podle velikosti. Úloha se tak převádí na hodnocení dat v ordinálním měřítku. Těmito metodami se říká pořadové.

Neparametrické metody založené na četnostech odchylek nebo pořadích uplatňené na data intervalového a poměrového typu využívají jenom část dostupné informace. Proto jsou příslušné statistické testy méně citlivé k odchylkám od nulové hypotézy. Pro dané rozsahy výběru a alternativní hypotézu mají menší sílu. U některých neparametrických testů tato ztráta není tak značná a je vyvážena zvýšenou validitou a robustností vůči odlehlym hodnotám.

3. Třetí typ neparametrických metod reprezentují neparametrické postupy, které pracují s daty v jejich původní podobě. Přitom používáme stejné testovací statistiky jako v parametrických testech, ale jinak počítáme příslušné *p*-hodnoty – využíváme Fisherova permutačního principu. Jestliže *p*-hadinu počítáme pomocí tohoto principu, mluvíme o **permutačním testu**. Výhoda postupu spočívá v tom, že platnost *p*-hodnot nezávisí na způsobu výběru, ale na způsobu randomizovaného přiřazení experimentálních jednotek. Mění se pouze rozsah zobecnitelnosti. Tento přístup k výpočtu *p*-hodnot používáme i u postupů založených na četnostech a pořadích (viz kapitoly 5.4.2 a 9.4). Edgington (1987) říká: „Permutační testy jsou neparametrické, protože nevyžadují žádné předpoklady o populaci ani nepožadují podmínku náhodného výběru. Fyzický akt randomizace je náhodným procesem, jenž představuje základ platnosti těchto testů. Uvolnění od předpokladu náhodného výběru je nesmírně důležité především při experimentování v behaviorálních vědách, kde jenom málo dat splňuje podmínku náhodného výběru ... Neparametrické testy pro pořadí a binární data jsou permutační testy a tedy nevyžadují náhodný výběr.“

Poznamenejme, že ačkoliv nemusí být specifikována žádná třída rozdělení, neparametrické testy obvykle potřebují předpoklad symetrie rozdělení nebo jeho spojitosti, případně shodu rozdělení, která se při alternativní hypotéze liší pouze polohou nebo rozptýleností (Sprent, 1998, s. 9).

Určení přesných *p*-hadin významnosti u neparametrických metod je založeno na diskrétních výběrových rozděleních konstruovaných pomocí Fisherova principu permutování. Podrobněji tento postup ozějmíme na příkladu v kapitole 5.4.2.

**Neparametrický test** svým názvem napovídá, že se v něm neprovádí závěr o hodnotě nějakého parametru pravděpodobnostního rozdělení, avšak při splnění určitých předpokladů tomu tak být nemusí a neparametrické metody se někdy s výhodou používají i při hledání intervalů spolehlivosti některých základních parametrů.

Pokud neparametrické metody používají pořadí dat, při výpočtech někdy vznikají potíže, jestliže údaje ve výběru jsou shodné. Když tento případ nastane, je nutno příslušné postupy modifikovat a přihlédnout k tomu, že za pořadí dosazujeme necelé číslo (totiž průměr z pořadí, která přísluší dané sérii stejných hodnot).

Každému neparametrickému „přesnému testu“ lze obvykle přiřadit při větším rozsazích výběrů asymptotický test založený na normálním rozdělení nebo rozdělení  $\chi^2$ . První neparametrický test – tzv. znaménkový test – vycházel z binomického rozdělení, které popsal Jacob Bernoulli v roce 1713. Abraham de Moivre o tomto rozdělení v roce 1733 dokázal, že se blíží k normálnímu. Znaménkový test použil John Arbuthnot již v roce 1710.

Některé popsané vlastnosti neparametrických testů dále ozějmíme pomocí dvou modelových příkladů. První se týká znaménkového testu, druhý testu založeného na pořadí.

#### 5.4.1 Přesný znaménkový test hodnoty mediánu

Provedení tohoto neparametrického testu ukážeme na datech z medicínského výzkumu, v němž se monitorovalo po dobu 362 týdnů přežívání skupiny 10 pacientů s určitým typem rakoviny. Byly zjištěny tyto doby přežití v týdnech (poslední údaj označený hvězdičkou znamená cenzorované pozorování, tedy že pacient po této době stále ještě žil):

49, 58, 75, 110, 112, 132, 151, 276, 281, 362\*

Chceme testovat nulovou hypotézu, že medián *Me* času přežití v populaci pacientů je 200, proti alternativní hypotéze, že medián má jinou hodnotu.

$$H_0: Me = 200$$

$$H_1: Me \neq 200$$

Spočítáme hodnoty z výběru, které jsou větší než 200 (přiřadíme jim označení „+“). Jestliže teoretický medián je skutečně 200, měl by být počet pozorování

Tab. 5.7 Příklad znaménkového testu mediánu – druhý řádek tabulky uvádí pravděpodobnost, že v daném výběru bude $k$ hodnot větších, než je teoretický medián											
$k$	0	1	2	3	4	5	6	7	8	9	10
$p_k$	0,0010	0,0098	0,0439	0,1172	0,2051	0,2461	0,2051	0,1172	0,0439	0,0098	0,0010

pod hodnotu 200 přibližně roven počtu pozorování nad touto hodnotou. Počet pozorování  $k$  nad hodnotou 200 má za platnosti nulové hypotézy binomické rozdělení  $B(0,5; 10)$ . Pravděpodobnosti výskytu hodnot  $k$  uvádí tabulka 5.7. V našem souboru překročily 3 údaje hodnotu 200 (3 hodnoty mají znaménka +). Celková pravděpodobnost tohoto výsledku a pro nulovou hypotézu ještě nepříznivějších výsledků ( $k = 0, 1, 2, 3$ ) je  $0,1172 + 0,0439 + 0,0098 + 0,0010 = 0,1719$ , jestliže nulová hypotéza platí. Provádíme dvoustranný test. Protože pravděpodobnost, že získáme tři kladná znaménka nebo méně nebo 7 kladných znamének nebo více, je  $2 \times 0,1719 = 0,3438$ , nemáme jasnou evidenci proti nulové hypotéze. Tato situace může totiž nastat ve více než třetině pokusů, i když  $H_0$  platí. Abychom nulovou hypotézu mohli zamítнуть, musela by být vypočtená pravděpodobnost menší než zvolená hladina významnosti (např. menší než 0,05).

Je nutné si všimnout, že u příslušného parametrického  $z$ -testu jsou možné všechny hodnoty  $p$  v intervalu  $(0; 1)$ , kdežto u znaménkového testu tomu tak není. V tomto příkladu jsou totiž tři nejmenší možné hodnoty  $p$ :  $2 \times 0,0010 = 0,0020$ ;  $2 \times (0,0010 + 0,0098) = 0,0216$ ;  $2 \times (0,0010 + 0,0098 + 0,0439) = 0,1094$ . Je patrné, že naše statistika – počet znamének plus – má diskrétní rozdělení. To znamená, že nemůžeme určit kritickou oblast pro znaménkový test s přesnou hladinou významnosti 0,05. Musíme se rozhodnout mezi hladinami 0,0216 nebo 0,1094.

Pro výpočet hodnoty  $p$  lze také použít tabulku VIII distribuční funkce binomického rozdělení z přílohy B. Vyhledáme  $F(X < 4)$  pro rozdělení  $B(0,5; 10)$ . V řádce s  $n = 10$  pro  $p = 0,5$  a  $x = 3$  nalezneme 0,17188, po zaokrouhlení 0,1719.

## 5.4.2 Permutační testy

Testy založené na Fisherovu principu permutování pořadí hodnot ve výběru nebo na permutacích jistých funkcí pořadí hodnot ve výběru (včetně speciálního případu permutací původních měření) jsou základní výbavou z oblasti neparametrických metod. Patří mezi ně např. varianty statistických testů, které navrhl Frank Wilcoxon pro srovnání dvou skupin měření. Nazýváme je **permutační**

**testy** nebo **randomizační testy**. Tyto testy jsou nejenom snadno pochopitelné (podobně jako znaménkový test), ale splňují i teoretická kritéria pro validní statistickou inferenci v případech, pro něž standardní postupy statistické inference nebyly navrženy. Například je lze použít v randomizovaných komparativních experimentech, jestliže nemáme k dispozici náhodný výběr (základní populace se v tomto případě často kryje se skupinou sledovaných osob v experimentu). Pokud v tomto případě aplikujeme test založený na normálním rozdělení, dopouštíme se koncepční chyby, protože tato procedura vychází z teorie náhodných výběrů. Z tohoto pohledu má většina  $t$ -testů a  $F$ -testů použitych v komparačních studiích při srovnání dvou průměrů pochybnou validitu. Permutační testy předpoklad náhodného výběru nevyžadují. Jejich logiku ozřejmíme na modelovém příkladu z medicínského výzkumu, v němž půjde o dvoustranný test shody rozdělení hodnot v obou souborech.

### PŘÍKLAD 5.8

#### Permutační test

U čtyř náhodně vybraných pacientů z devíti je aplikována nová terapie. Po čtyřech týdnech je všech devět pacientů vyšetřeno. Na základě klinických testů jsou seřazeni podle stupně vážnosti stavu (nejpříznivější případ=1, nejhorší případ=9). Jaká je pravděpodobnost, jestliže ošetření nemá žádné příznivé působení, že pacienti s novou terapií mají pořadová čísla 1, 2, 3, 4?

Jestliže nové ošetření nemá žádný vliv, očekávali bychom, že pacienti s novou terapií budou mít svá pořadí náhodně rozdělena v sekvenci 1, 2, ..., 9. Pro skupinu devíti pacientů je možné vytvořit těchto kombinací 126, jak ukazuje tabulka 5.8 (podle Sprent, 2001). Jestliže nová terapie není účinná (neboli její účinnost je stejná jako účinnost placebového ošetření), má jakákoli kombinace z tabulky 5.8 stejnou pravděpodobnost. Pomocí tohoto předpokladu můžeme snadno spočítat pravděpodobnosti jednotlivých hodnot vhodné zvolené testovací statistiky vypočtené z pořadí. V tomto případě to provedeme pro součet pořadí v jedné skupině, protože takto zvolená statistika odráží nepodobnost obou skupin vzhledem k rozdílnosti střední hodnoty souboru.

Nejdříve zjistíme, kolik kombinací dá danou hodnotu součtu pořadí, a pak tuto hodnotu vynásobíme číslem 1/126. V tabulce 5.9 jsou uvedeny počty jednotlivých kombinací s danými hodnotami součtu pořadí. Uvažujeme test hypotézy

$$H_0: \text{nové ošetření nemá účinek}$$

proti dvoustranné alternativní hypotéze

$$H_1: \text{nové ošetření má účinek.}$$

Kombinace (1; 2; 3; 4) se může vyskytnout s pravděpodobností 1/126. Nejméně příznivá pro hodnocení účinku nové terapie by byla kombinace (6; 7; 8; 9), která se může vyskytnout také

Tab. 5.8

Příklad permutačního testu – možné kombinace výběru čtyř jedinců z devíti se součtem pořadí v závorce

1, 2, 3, 4 (10)	1, 2, 3, 5 (11)	1, 2, 3, 6 (12)	1, 2, 3, 7 (13)	1, 2, 3, 8 (14)	1, 2, 3, 9 (15)	1, 2, 4, 5 (12)
1, 2, 4, 6 (13)	1, 2, 4, 7 (14)	1, 2, 4, 8 (15)	1, 2, 4, 9 (16)	1, 2, 5, 6 (14)	1, 2, 5, 7 (15)	1, 2, 5, 8 (16)
1, 2, 5, 9 (17)	1, 2, 6, 7 (16)	1, 2, 6, 8 (17)	1, 2, 6, 9 (18)	1, 2, 7, 8 (18)	1, 2, 7, 9 (19)	1, 2, 8, 9 (20)
1, 3, 4, 5 (13)	1, 3, 4, 6 (14)	1, 3, 4, 7 (15)	1, 3, 4, 8 (16)	1, 3, 4, 9 (17)	1, 3, 5, 6 (15)	1, 3, 5, 7 (16)
1, 3, 5, 8 (17)	1, 3, 5, 9 (18)	1, 3, 6, 7 (17)	1, 3, 6, 8 (18)	1, 3, 6, 9 (19)	1, 3, 7, 8 (19)	1, 3, 7, 9 (20)
1, 3, 8, 9 (21)	1, 4, 5, 6 (16)	1, 4, 5, 7 (17)	1, 4, 5, 8 (18)	1, 4, 5, 9 (19)	1, 4, 6, 7 (18)	1, 4, 6, 8 (19)
1, 4, 6, 9 (20)	1, 4, 7, 8 (20)	1, 4, 7, 9 (21)	1, 4, 8, 9 (22)	1, 5, 6, 7 (19)	1, 5, 6, 8 (20)	1, 5, 6, 9 (21)
1, 5, 7, 8 (21)	1, 5, 7, 9 (22)	1, 5, 8, 9 (23)	1, 6, 7, 8 (22)	1, 6, 7, 9 (23)	1, 6, 8, 9 (24)	1, 7, 8, 9 (25)
2, 3, 4, 5 (14)	2, 3, 4, 6 (15)	2, 3, 4, 7 (16)	2, 3, 4, 8 (17)	2, 3, 4, 9 (18)	2, 3, 5, 6 (16)	2, 3, 5, 7 (17)
2, 3, 5, 8 (18)	2, 3, 5, 9 (19)	2, 3, 6, 7 (18)	2, 3, 6, 8 (19)	2, 3, 6, 9 (20)	2, 3, 7, 8 (20)	2, 3, 7, 9 (21)
2, 3, 8, 9 (22)	2, 4, 5, 6 (17)	2, 4, 5, 7 (18)	2, 4, 5, 8 (19)	2, 4, 5, 9 (20)	2, 4, 6, 7 (19)	2, 4, 6, 8 (20)
2, 4, 6, 9 (21)	2, 4, 7, 8 (21)	2, 4, 7, 9 (22)	2, 4, 8, 9 (23)	2, 5, 6, 7 (20)	2, 5, 6, 8 (21)	2, 5, 6, 9 (22)
2, 5, 7, 8 (22)	2, 5, 7, 9 (23)	2, 5, 8, 9 (24)	2, 6, 7, 8 (23)	2, 6, 7, 9 (24)	2, 6, 8, 9 (25)	2, 7, 8, 9 (26)
3, 4, 5, 6 (18)	3, 4, 5, 7 (19)	3, 4, 5, 8 (20)	3, 4, 5, 9 (21)	3, 4, 6, 7 (20)	3, 4, 6, 8 (21)	3, 4, 6, 9 (22)
3, 4, 7, 8 (22)	3, 4, 7, 9 (23)	3, 4, 8, 9 (24)	3, 5, 6, 7 (21)	3, 5, 6, 8 (22)	3, 5, 6, 9 (23)	3, 5, 7, 8 (23)
3, 5, 7, 9 (24)	3, 5, 8, 9 (25)	3, 6, 7, 8 (24)	3, 6, 7, 9 (25)	3, 6, 8, 9 (26)	3, 7, 8, 9 (27)	4, 5, 6, 7 (22)
4, 5, 6, 8 (23)	4, 5, 6, 9 (24)	4, 5, 7, 8 (24)	4, 5, 7, 9 (25)	4, 5, 8, 9 (26)	4, 6, 7, 8 (25)	4, 6, 7, 9 (26)
4, 6, 8, 9 (27)	4, 7, 8, 9 (28)	5, 6, 7, 8 (26)	5, 6, 7, 9 (27)	5, 6, 8, 9 (28)	5, 7, 8, 9 (29)	6, 7, 8, 9 (30)

Tab. 5.9

Počet různých kombinací vedoucích k určité hodnotě  $S$  součtu pořadí ve skupině

Součet pořadí $S$	10	11	12	13	14	15	16	17	18	19	20
Počet výskytů	1	1	2	3	5	6	8	9	11	11	12
Součet pořadí $S$	21	22	23	24	25	26	27	28	29	30	
Počet výskytů	11	11	9	8	6	5	3	2	1	1	

s pravděpodobností 1/126. Celková pravděpodobnost těchto dvou kombinací je  $2/126 = 0,0159$ , pokud platí nulová hypotéza. Takže můžeme říci, že nulovou hypotézu lze zamítнуť na přesné hladině významnosti 1,59 %.

Jak budeme postupovat, když dostaneme kombinaci (1; 2; 3; 5)? Opíráme se o součet pořadí, protože zřejmě malé nebo naopak velké hodnoty součtu jsou nepříznivé pro nulovou hypotézu. Abychom nalezli, jaký výsledek (součet pořadí) je konzistentní s  $p$ -hodnotou, která nepresahuje hodnotu 0,05, zvolíme oblasti na obou koncích rozdělení součtu pořadí, jejichž pravděpodobnost nebude větší než 0,025. Výpočty dostaneme

$$P(S \leq 11) = 0,0159,$$

$$P(S \leq 12) = 0,0317.$$

Druhá hodnota je již větší než 0,025. Proto za kritickou oblast testovací statistiky  $S$  pro realizaci dvoustranného testu zvolíme  $S \leq 11$  a k ní symetrickou část  $S \geq 29$ . Dosažená hladina významnosti je při takové volbě 0,03197.

Logika testu je jednoduchá. Výpočty nutné pro realizaci testu se provádí automaticky, pokud pracujeme s vhodným statistickým programovým systémem. Také existují tabulky, jež obsahují kritické meze pro hodnoty součtu pořadí při daném rozsahu obou skupin. Pro větší výběry pak použijeme approximaci rozdělení testovací statistiky normálním rozdělením.

Jestliže chceme určit kritickou oblast pro jednostranný test, všimáme si jenom jednoho konce setříděných hodnot součtu pořadí (tab. 5.9).

Uvedeme jednotlivé kroky Fisherova principu permutování pořadí, který lze uplatnit v mnoha situacích statistického testování hypotéz.

1. Vybereme testovací statistiku, jež dobře odlišuje nulovou hypotézu od alternativní.
2. Nahradíme původní pozorování  $\{x_{ij}; i = 1, 2, \dots, n_j; j = 1, 2, \dots, J\}$  jejich pořadími v sloučeném výběru  $\{R_k; k = 1, 2, \dots, n\}$ . Příklad: hodnoty 5, 2; 1 a 7 nahradíme pořadími 2; 1 a 3. Vypočítáme testovací statistiku  $S$  pro původní množinu pořadí. Tuto hodnotu označíme  $S'$ .
3. Získáme permutační rozdělení statistiky  $S$  opakováním přemísťováním pořadí a vypočítáváním testovací statistiky  $S$ . To, jakým způsobem provádíme přemísťování pořadí, závisí na plánu výzkumu, kterým jsme data získali. Protože hodnoty pořadí jsou (skoro) vždy čísla 1, 2, ..., použijeme podle možnosti tabelovanou distribuční funkci (např. tabulky pro přesný Wilcoxonův test), jež odpovídá permutačnímu rozdělení.
4. Odmítнемe nebo zamítнемe nulovou hypotézu, jestliže  $p$ -hodnota hodnoty testovací statistiky  $S'$  zjištěná na základě této distribuční je menší než zvolená hladina významnosti.

Princip permutování uplatňujeme i na původní hodnoty  $x_{ij}$  (viz kap. 9.4).

#### 5.4.3 Eficience neparametrických testů

Porovnání síly testů se provádí pomocí indexu **eficienze**  $E_n$ , který se vypočte jako poměr rozsahů výběru, jež jsou zapotřebí k dosažení stejné síly srovnávanými testy:

$$\text{eficience } E_n = \frac{\text{potřebný rozsah výběru pro parametrický test}}{\text{potřebný rozsah výběru pro neparametrický test}}$$

Tab. 5.10 Index asymptotické eficie pro porovnání vybraných testů

Statistický test v čitateli	Statistický test ve jmenovateli	Eficience $E$
znaménkový test	$t$ -test pro jednu skupinu	$2/\pi$
Wilcoxonův test	$t$ -test pro jednu skupinu	$3/\pi$
Wilcoxonův test	$t$ -test pro dvě nez. skupiny	$3/\pi$
Wilcoxonův test	znaménkový test	$3/2$
znaménkový test	$t$ -test pro dvě skupiny	$2/\pi$
Kruskal-Wallisův test	jednoduchá analýza rozptylu	$3/\pi$

Pojem **asymptotická eficience** označuje limitu eficie, když se  $n$  zvětšuje k velkým číslům. Například asymptotická eficience  $E = 0,95$ , kterou má Wilcoxonův test, znamená, že jestliže používáme Wilcoxonův test s rozsahy výběru 100, je to stejně, jako bychom používali v  $t$ -testu rozsahy 95.

Neparametrické testy se v případě *spojitých* měření používají, jestliže:

- parametrický test je málo robustní vůči odchylkám od jeho předpokladů (jak víme, robustnost znamená, že test neztratí všechny své dobré vlastnosti při porušení jeho předpokladů);
- tvar rozdělení nelze upravit transformacemi (např. logaritmická transformace zmenšuje zešikmení zleva);
- máme podezření, že data obsahují odlehlé hodnoty;
- chceme provést kontrolu parametrických testů;
- nám záleží na jednoduchosti.

Obvykle jsou neparametrické testy ve srovnání s parametrickými testy **konzervativní**. Tento pojem vyjadřuje, že testy dle setrvávají na rozhodnutí, že nulová hypotéza se nemá zamítat. Jinak řečeno, mají menší sílu testu. V tabulkce 5.10 uvádíme vybrané indexy asymptotické eficie pro porovnání některých testů, jestliže platí předpoklady pro použití parametrického testu.

## 5.5 Problém simultánního statistického usuzování

Teoretická tvrzení o optimálních vlastnostech statistického testu nebo intervalu spolehlivosti (tvrzení o hladině významnosti nebo spolehlivosti a o síle testu) byla odvozena pro situaci, kdy provedeme *jediný* test nebo sestrojíme jeden interval spolehlivosti, avšak při analýze dat obvykle současně provádíme *mnoho* statis-

tických inferenčních rozhodnutí. Často jsme v situaci, že počet testů nebo sestřených intervalů spolehlivosti nekontrolovatelně roste s tím, jak se zvětšuje počet porovnávaných skupin, počet proměnných a komplexnost použitých modelů závislostí proměnných. Stává se, že „lovíme“ v množině dat statisticky významné výsledky, kterými chceme zkrášlit naši zprávu o provedeném výzkumu. Lze však dokázat, že s rostoucím počtem provedených testů roste pravděpodobnost, že aspoň jeden výsledek bude statisticky významný, přestože ve skutečnosti platí všechny nulové hypotézy, jež jsme uvažovali a testovali. (To plyne jednoznačně z pravděpodobnostních úvah.)

Tohoto problému se jenom dotkneme (podrobnej viz Havránek, 1993). Rozlišujeme mezi mírou počtu chybných rozhodnutí v experimentu a v inferenci.

**Chybový index inference ve srovnání  $i$**  je dán poměrem

$$i = \frac{\text{počet chybných rozhodnutí}}{\text{počet rozhodnutí}}$$

a zajímá nás jeho střední hodnota. Inference odpovídá provedení rozhodnutí pomocí jednoho testu nebo sestření jednoho intervalu spolehlivosti. Chybné rozhodnutí odpovídá chybě I. druhu nebo tomu, že interval spolehlivosti nepokryje správnou hodnotu. Jestliže např. provádíme porovnávání k průměrů, lze vytvořit až  $k(k-1)/2$  párů průměrů a ty srovnat mezi sebou. Pokud testujeme na hladině významnosti  $\alpha$ , pak střední hodnota indexu  $i$  může být maximálně  $\alpha$ . To znamená, že při 50 srovnáních a použité 5% hladině významnosti uděláme v průměru asi 2 až 3 chybná rozhodnutí.

Statistici navrhli používat raději přísnější **index inference v experimentu**:

$$e = \frac{\text{počet chybných rozhodnutí}}{\text{počet pokusů}},$$

v němž uvažujeme počet chybných rozhodnutí v jednom uskutečněném experimentu nebo lépe v jednom výzkumném projektu. V tomto případě tvoří horní hranici pro průměrnou hodnotu indexu  $e$  číslo  $\alpha k(k-1)/2$ . Jedná se o horní hranici pravděpodobnosti, že uděláme alespoň jedno chybné rozhodnutí. Je patrné, že tato hranice bude s rostoucím počtem rozhodnutí větší než 1.

Pro udržení hladiny indexu inference v experimentu na přijatelné úrovni se používá Bonferronovo metoda. Nebudeme ji popisovat podrobně, ale je užitečné porozumět jejímu principu. V Bonferronově metodě snížujeme hladinu významnosti každého testu, který se ve výzkumné akci použije, aby se snížil index inference v experimentu tak, že bude určitě pod předem zvolenou úrovni, třeba 0,05. Jestliže např. chceme provést jeden dva testy na hladině významnosti 0,05, pak metoda požaduje vydělit 0,05 dvěma a výslednou hladinu 0,025 použít

v obou dvou testech. Obecně, jestliže chceme provést  $m$  rozhodnutí, volíme hladinu významnosti (platí také pro sestrojení intervalů spolehlivosti)  $\alpha/m$ , abychom dodrželi hladinu indexu  $i$  na hladině  $\alpha$ . V případě všech párových srovnání k průměrů tedy použijeme v jednotlivých testech hladinu  $(2\alpha)/(k(k - 1))$ .

Bonferroniho metoda se především používá, jestliže přibližně víme, kolik srovnání budeme provádět nebo plánujeme provést. Mluvíme o **plánovaném srovnávání**, jestliže

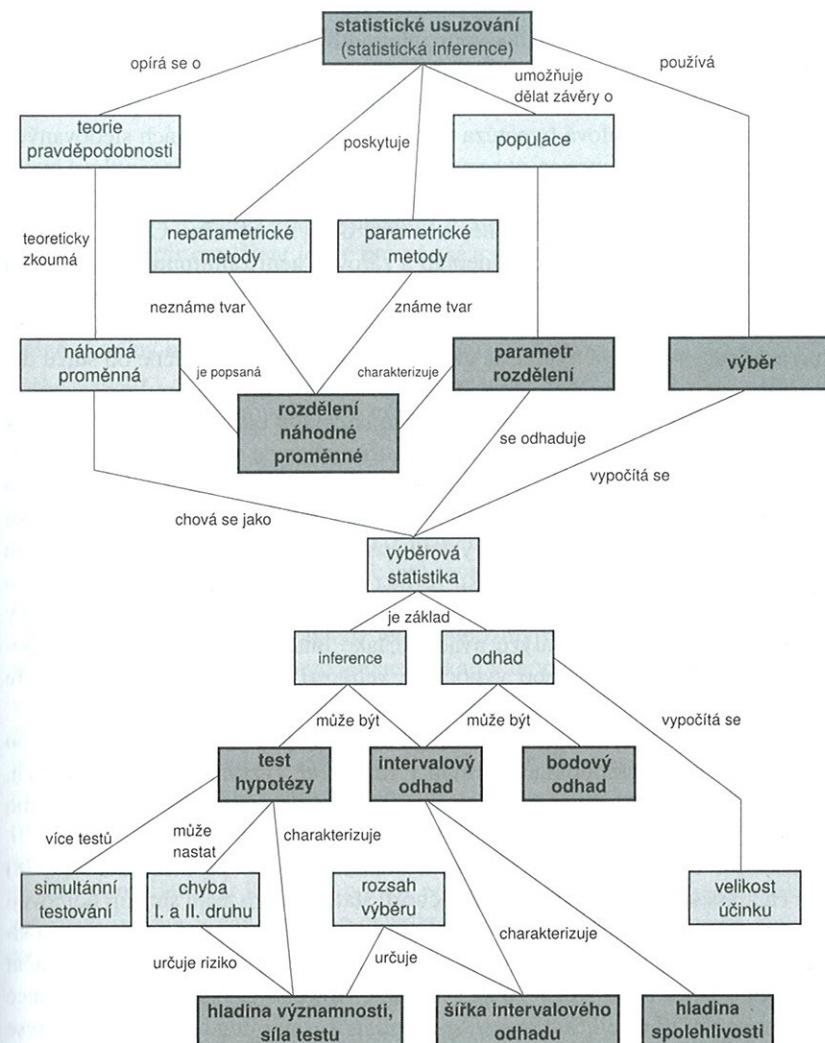
- chceme provést jenom podmnožinu všech možných testů;
- tuto podmnožinu specifikujeme předtím, než jsou data nasbírána.

Bonferroniho metoda a jiné příbuzné metody úprav hladin významnost nejsou vhodné v explorační analýze, kde chce výzkumník otestovat všechny možné hypotézy. Metody, které v této souvislosti umožní zachovat specifikovanou hladinu indexu  $e$ , nazýváme *post hoc* metody simultánní inference. *Post hoc* metody statistické inference dovolují výzkumníkovi zkoumat data a porovnávat jejich jednotlivé části, aniž by srovnání předem specifikoval, přičemž udržuje spolehlivost statistické inference. O některých těchto metodách se zmíníme v kapitole o analýze rozptylu.

## Souhrn

V kapitole jsme popsali základní pojmy statistického rozhodovaní a inference. Příslušná teorie vychází z pravděpodobnostních rozdělení výběrových statistik. Zdůraznili jsme vzájemnou vazbu počtu pozorování a parametrů jako síla testu, hladina významnosti a chyba I. a II. druhu. Kritika aplikací inferenční statistiky vedla k většímu uplatňování intervalů spolehlivosti a konceptu velikosti účinku. Odlišili jsme testy založené na normálním rozdělení zkoumaných náhodných proměnných a testy neparametrické, které tento předpoklad nevyžadují. Často lze využít asymptotických vlastností výběrových statistik, jež se projevují při větších rozsazích výběrů, k jednoduchému nalezení kritických hodnot testů nebo dokonce k nahradě neparametrického testu parametrickým. Návaznosti hlavních konceptů statistické inference schematicky znázorňuje obrázek 5.7.

Obr. 5.7 Vztahy mezi pojmy teorie statistického usuzování



Inferenční statistické postupy se zabývají testováním statistických hypotéz a odhady charakteristik rozdělení výzkumných proměnných. Zohledňují inherentní nejistotu dat v důsledku jejich náhodné proměnlivosti. Bodové odhady poskytují jedinou numerickou hodnotu, intervalové hodnoty udávají horní a dolní mez intervalu, kde se teoretická hodnota může vyskytovat s velkou spolehlivostí. Statistické testování hypotéz dovoluje provést jistá rozhodnutí o platnosti statistických hypotéz. Nulová hypotéza tvrdí něco o charakteristikách sledovaných proměnných. Cílem je vyvarovat se neoprávněného uznání porušení nulové hypotézy a odhalit skutečné porušení nulové hypotézy s velkou jistotou. Jestliže nulová hypotéza je zamítнутa neoprávněně, mluvíme o chybě I. druhu. Chyba II. druhu nastane, jestliže nulová hypotéza neplatí a zároveň není zamítнутa. Výzkumníci mají pod kontrolou hladinu chyby I. druhu tím, že určují hladinu významnosti. Četnost chyb druhého druhu závisí na zvolené hladině významnosti a na tom jak silně je porušena nulová hypotéza. Pokud se skutečnost liší značně od stavu daným nulovou hypotézou, pak získaná data povedou k zamítnutí nulové hypotézy častěji, než když se bude lišit méně. I malé odchylky od nulové hypotézy však odhalíme s velkou silou, jestliže budeme moci realizovat výzkum o velkých rozsazích výběrů. Při posuzování výsledků statistické analýzy musíme uvážit jejich význam v kontextu řešeného problému. Výsledky statistických testů významnosti doplňujeme úvahami o praktické významnosti. Přitom si pomáháme výpočtem různých koeficientů velikosti účinku (effect size). Někdy se požaduje předem specifikovat, jaké odchylky od nulové hypotézy budeme považovat za prakticky významné. Ve zprávě o analýze uvádíme také intervaly spolehlivosti pro posuzované charakteristiky nebo vypočtené velikosti účinku. Pomocí jejich šíře posuzujeme přesnost vypočtených bodových odhadů.

Parametrické testy zahrnují odhad parametrů, použití intervalových nebo poměrových dat a předpoklad normality rozdělení výzkumných proměnných. Neparametrické testy se používají s nominálními nebo ordinálními daty nebo v případech, kdy nelze předpokládat normální rozdělení.

Základní teorie statistických testů a odhadů vhodně probírají Rao (1978) nebo Havránek (1993). Diskusi o užitečnosti statistických testů shrnují nejnověji Wainer a Robinson (2003). Novější techniky statistického usuzování pomocí simulací na základě rozdělení výběrových dat (bootstrap, resampling, permutační testy) popisuje stručně v příloze ke své knize Howell (1992) na WWW stránce [www.uvm.edu/~dhowell/StatPages/Resampling/Resampling.html](http://www.uvm.edu/~dhowell/StatPages/Resampling/Resampling.html). Na této adrese je také k dispozici program pro realizaci popsaných přístupů. Podrobněji objasňuje tuto moderní část statistiky Sprent (1998).

## 6 Základní situace statistického usuzování

V prvních fázích analýzy dat využíváme grafické prostředky a číselné charakteristiky pro popis rozdělení jedné proměnné a porovnání několika skupin měření proměnné mezi sebou. Také výklad metod statistického usuzování začneme s metodami pro hodnocení rozdělení jedné proměnné. Bude nás zajímat posouzení jedné skupiny měření a srovnání mezi dvěma skupinami měření. Protože z hlediska statistiky půjde o výběry z určitých populací, mluvíme o problému jednoho nebo dvou výběrů. Porovnáváním více než dvou skupin měření se budeme zabývat v kapitole 9.

Dvě důležité vlastnosti dat jsou jejich střední hodnoty a rozptýlenost. Pokud je rozdělení proměnné normální, popisujeme střední hodnotu parametrem  $\mu$  a její variabilitu směrodatnou odchylkou  $\sigma$  nebo rozptylem  $\sigma^2$ .

V této kapitole využijeme konceptu teorie odhadu a testování hypotéz a zaměříme se na popis metod pro hledání intervalů spolehlivosti a testy hypotéz pro průměr  $\mu$  a rozptyl  $\sigma^2$  na základě jedné množiny měření. V případě hodnocení parametru  $\mu$  jsme se již s touto situací setkali v teoretické kapitole o statistickém usuzování. Uvedeme také metody pro porovnání parametrů  $\mu$  a  $\sigma^2$  ve dvou populacích. Poznáme test normality rozdělení, který pomáhá při rozhodování, zda je porušen předpoklad o normálním rozdělení dat. Popíšeme také některé neparametrické testy, jimiž nahrazujeme jednotlivé varianty parametrického  $t$ -testu pro hodnocení průměru. Podrobný rozhodovací algoritmus pro výběr jednotlivých metod uvedeme na konci kapitoly. Výklad o hodnocení průměrů a středních hodnot provedeme v logice a pořadí podle obrázku 6.1. Při posuzování průměrných hodnot v závislosti na charakteru dat použijeme parametrickou nebo neparametrickou proceduru.

### PŘÍKLAD 6.1

#### Závislé a nezávislé výběry

Na dvou problémech dokumentujeme rozdíl mezi dvěma závislými a dvěma nezávislými výběry. V obou případech vycházíme při posuzování rozdílů různých skupin měření z hodnot aritmetických průměrů.