rentní 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 zamítnuta neoprávněně, mluvíme o chybě I. druhu. Chyba II. druhu nastane, jestliže nulová hypotéza neplatí a zároveň není zamítnuta. 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 teorii 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. Na této adrese je také k dispozici program pro realizaci popsaných přístupů. Podrobněji objasňuje tuto moderní část statistiky Sprent (1998).

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 μ a její variabilitu směrodatnou odchylkou σ nebo rozptylem σ^2.

V této kapitole využijeme konceptů teorie odhadu a testování hypotéz a zaměříme se na popis metod pro hledání intervalů spolehlivosti a testy hypotéz pro průměr μ a rozptyl σ^2 na základě jedné množiny měření. V případě hodnocení parametru μ jsme se již s touto situací setkali v teoretické kapitole o statistickém usuzování. Uvedeme také metody pro porovnání parametrů μ a σ^2 ve dvou populacích. Poznáme test normality rozdělení, který pomáhá při rozhodování, zda je porušen předpoklad o normálním rozdělen 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.

V chemické laboratoři se přezkušuje přítomnost systematické chyby u nové měřicí laboratorní metody srovnáním se standardní metodou.

1. Jeden vzorek o neznámé koncentraci se změří dvacetkrát novou a dvacetkrát standardní metodou stejným způsobem, jako se provádí opakovací pokus po hodnocení reliability. Porovnají se průměry obou množin měření.
2. Novou i standardní metodou se změří dvacet různých vzorků (každý vzorek oběma metodami) a porovnají se získané hodnoty.
3. Také můžeme znát koncentraci vzorku a změřit ho dvacetkrát posuzovanou metodou. Průměr z naměřených hodnot srovnáme se známou koncentrací.

V tomto případě jsme v prvním úkolu získali data ze dvou nezávislých výběrů, druhý úkol vede ke dvěma závislým výběrům se spárovanými daty (měření vzorku novou metodu, měření vzorku standardní metodu). Poslední úkol je situací s jedním výběrem.

■ ■ ■

Druhý výzkumný problém pochází z oblasti pedagogického výzkumu.

1. Výzkumník se chce dozvědět, zda je při výuce matematiky na základní škole efektivnější položit otázky před, anebo po zavedení nového pojmu. Připraví dva textové segmenty, které zprostředkují příslušnou učební látku, jeden s motivujícími otázkami před uvedením nového pojmu a druhý s kontrolními otázkami po jeho uvedení. Každý z textů použije s jinou skupinou žáků. Následně výzkumník porovná mezi oběma skupinami výsledky dosažené ve znalostním testu.
2. Jiný výzkumník při řešení stejného problému postupuje trochu jinak. Připraví dva textové segmenty ke dvěma různým výukovým oblastem. Každý segment bude ve dvou verzích, jedna bude s otázkami před probráním látky a druhá s otázkami až po probrání látky. Sleduje se jedna skupina žáků. Každý žák studuje obě látky, jednu s otázkami před a druhou s otázkami po jejím probrání, přičemž pořadí výukové látky se volí náhodně. Výzkumník porovná u každého žáka výsledky znalostního testu pro obě látky, aby se ukázalo, která metoda vede k lepším výsledkům ve znalostním testu.

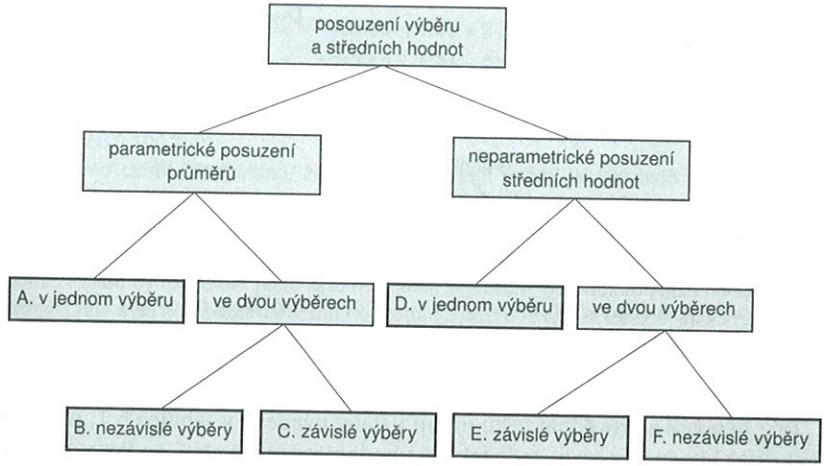
První experimentální situace se statisticky řeší jako problém dvou nezávislých výběrů, druhá situace patří do kategorie dvou závislých výběrů.

6.1 Hodnocení průměru v jednom výběru

Hodnotíme, zda odlišnost výběrového průměru od teoretické hodnoty μ_0 odpovídá náhodnému kolísání, nebo je projevem systematické odchylky. Máme k dispozici prostý náhodný výběr z dané populace. S takovou situací se setkáme např. v následujících situacích:

a) Firma vyrábějící marmelády zkoumá dodržení váhy produktu. Z každé série výrobků se vybírá náhodně několik balení a ověřuje se, zda získaná průměrná hodnota v dané sérii odpovídá nominální hodnotě.

Obr. 6.1 Uspořádání základních situací statistického usuzování



b) Zkušební laboratoř ověřuje práci vybrané nemocniční laboratoře. Opakovaně zadá do této laboratoře vzorek s danou koncentrací k analýze a přezkušuje, zda výsledky nahlášené laboratoří odpovídají v průměru této koncentraci.
c) Zkoumáme, zda antropometrické parametry v dané populaci mají předpokládanou úroveň, pomocí náhodného výběru z této populace. Odhadujeme také úroveň těchto parametrů.

Tuto situaci jsme podrobně probrali v předchozí kapitole o základech teorie statistického usuzování. Předpokládali jsme, že známe hodnotu směrodatné odchylky. V tomto odstavci splnění této podmínky nebudeme požadovat, což je pro statistickou praxi výhodnější. Předpokládáme normální rozdělení sledované proměnné. Jestliže tomu tak není, ale pozorování je více než 30, můžeme metodu s opatrností použít jako asymptoticky platnou v důsledku působení centrálního limitního teorému.

Využíváme poznatek, že hodnoty rozdílu průměru $\bar{x}$ od předpokládaného teoretického průměru μ_0 standardizovaná odhadem směrodatné chyby $s_{\bar{x}}$ mají za platnosti nulové hypotézy Studentovo t-rozdělení s $n - 1$ stupni volnosti (kap. 4.6.2).

Tab. 6.1 Schéma hodnocení průměru v jednom výběru

Jednostranný test	Dvoustranný test
$H_0: \mu = \mu_0$	$H_0: \mu = \mu_0$
$H_1: \mu > \mu_0$ (nebo $H_1: \mu < \mu_0$)	$H_1: \mu \neq \mu_0$
Testovací statistika: $t = \frac{\bar{x} - \mu_0}{s_{\bar{x}}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$	
Oblast zamítnutí: $t > t_\alpha$ (nebo $t < -t_\alpha$)	Oblast zamítnutí: $t < -t_{\alpha/2}$ a $t > t_{\alpha/2}$

Tvar hypotéz a provedení testu určují vztahy uvedené v tabulce 6.1, přičemž t_α je kritická hodnota t-testu, tedy taková t-hodnota, pro niž platí, že $P(t > t_\alpha) = \alpha$ za platnosti nulové hypotézy. Kritické hodnoty hledáme pro t-rozdělení o $n - 1$ stupních volnosti. Dvoustranný interval spolehlivosti pro hladinu spolehlivosti $1 - \alpha$ má tvar

$$\mu \in (\bar{x} - t_{\alpha/2} s_{\bar{x}},\ \bar{x} + t_{\alpha/2} s_{\bar{x}}).$$

Připomeňme, že interval spolehlivosti musíme odlišovat od predikčního intervalu pro sledovanou proměnnou X. Predikční interval určuje, kde se bude nacházet s pravděpodobností $(1 - \alpha)\,\%$ naměřený údaj. Predikční interval má tvar

$$X \in (\bar{x} - t_{\alpha/2} s,\ \bar{x} + t_{\alpha/2} s).$$

Abychom mohli použít predikční interval, musíme – na rozdíl od intervalu spolehlivosti – striktně požadovat normalitu rozdělení proměnné, protože se náš odhad týká krajních hodnot rozdělení, ne průměrných hodnot.

PŘÍKLAD 6.2

Hodnocení průměru v jednom výběru

Na základě dostupných dat chceme zjistit interval spolehlivosti pro průměrný roční příjem rodiny v dané oblasti. Pro náhodný výběr $n = 100$ rodin se získala průměrná hodnota ročního příjmu 105 510 Kč a směrodatná odchylka údajů 23 156 Kč. Vypočítáme 90% interval spolehlivosti pro průměr ročního příjmu v populaci:

$$s_{\bar{x}} = 23\,156/10 = 2\,315{,}6$$

$$t_{0{,}05}(99) = 1{,}66$$

$$\textit{dolní mez} = 105\,510 - 1{,}66 \times 2\,315 = 101\,667$$

$$\textit{horní mez} = 105\,510 + 1{,}66 \times 23\,115 = 109\,353$$

V tomto případě hledaná hodnota průměrného příjmu se nalézá s 90% spolehlivostí v intervalu (101 667; 109 353).

Chceme ověřit předpoklad, že průměrný příjem rodiny je 100 tisíc Kč. Testovací t-statistika má hodnotu (105 510 – 100 000)/(23 156/10) = 2,37 a dosažená p-hladina při dvoustranném testu má hodnotu 0,0193. Lze tedy uzavřít, že pokud by požadovaná hladina významnosti testu byla 0,01, nemáme dostatek evidence, abychom nulovou hypotézu zamítli. Pokud by rozdělení platů bylo normální, predikční interval pro jejich hodnotu, který zachytí platy s 95% spolehlivostí, má hodnotu: (105 510 – 1,95 × 23 156; 105 510 + 1,95 × 231 156). Takže predikční interval je tvořen mezemi: *dolní mez* = 60 355; *horní mez* = 150 664.

Uvedeme souhrnně základní pravidla pro použití t-testu s ohledem na splnění předpokladů. Jednovýběrový t-test je dosti robustní k odchylkám od normality:

- S výjimkou případu malých výběrů, předpoklad prostého náhodného výběru z posuzované populace je důležitější než předpoklad normálního rozdělení.
- Při výběrech o rozsahu menším než 15 použijeme t-test, pokud jsou data normálně rozdělená. Pokud data nejsou normálně rozdělená nebo obsahují odlehlou hodnotu, použijeme neparametrický test.
- Pro výběry o rozsahu ne menším než 15 použijeme t-test, pokud data nejsou výrazně zešikmená a neobsahují odlehlé hodnoty.
- Pro velké rozsahy (> 30) použijeme t-test i při zešikmených datech.

6.2 Porovnání průměrů ve dvou výběrech

Cílem statistického usuzování je v tomto případě porovnat odpovědi na dvě různá ošetření nebo intervence nebo porovnat dvě populace. Z každé populace máme zvláštní výběr. Porovnáváme průměry μ_1 a μ_2 sledované proměnné. Zajímá nás, zda rozdíl průměrů má určenou hodnotu Δ, nebo se od ní liší. Také chceme sestrojit příslušný interval spolehlivosti pro rozdíl průměrů proměnné v obou populacích. Uvedeme příklady této situace.

PŘÍKLAD 6.3

Porovnání průměrů ve dvou nezávislých výběrech

a) Provádí se mezipohlavní srovnání průměrné váhy novorozenců narozených v dané nemocnici za určité časové období. Jak je možné statisticky ověřit odlišnost průměrné váhy obou skupin novorozenců v této nemocnici, jestliže předpokládáme, že daný vzorek je reprezentativní pro všechny novorozence za delší časové období? Jak se spolehlivě odhadne rozdíl průměrných vah obou skupin při daném počtu pozorování?

b) Způsobuje zvýšený přívod kalcia v potravě snížení krevního tlaku? Provedl se randomizovaný pokus, v němž se do dvou skupin náhodně rozdělila skupina seniorů mužského pohlaví. Jedna ze skupin dostávala přídavek kalcia, druhá pouze placebový přípravek. Experiment se provedl jako dvojnásobně zaslepený. Získaly se údaje o změně systolického tlaku pro jedince z každé skupiny a dopočítaly se statistiky, které jsou popsány tabulkou 6.2. Skupina ošetřená kalciem zaznamenala průměrný pokles o 5,00 jednotek. Liší se tento pokles od průměrné změny −0,273 u skupiny ošetřené placebem?

Tab. 6.2 Příklad srovnávání průměrů – statistiky charakterizující změnu systolického tlaku pro dvě sledované skupiny

Skupina	Ošetření	n	$\bar{x}_1$	s
1	kalcium	10	5,00	8,74
2	placebo	10	−0,273	5,90

V obou popsaných případech se jedná o nezávislé skupiny měření. Odlišná je následující situace, jež vede ke dvěma závislým skupinám měření.

PŘÍKLAD 6.4

Porovnání průměrů ve dvou závislých výběrech – párová měření

Sledujeme jednu skupinu jedinců a provedeme u ní měření před a po aplikaci ošetření. Chceme porovnat rozdělení sledované proměnné před ošetřením a po něm. V tomto případě máme k dispozici pro každého jedince dvojici hodnot, párová měření: před intervencí a po ní. Statistický postup je odlišný od předchozích dvou situací. Při porovnávání pracujeme s rozdíly obou hodnot ve dvojici a použijeme t-test pro jednu skupinu. Tento postup podrobněji popíšeme v další části odstavce.

Předpokládejme obecně, že první výběr má rozsah n_1 a pro měřenou veličinu jsme vypočetli výběrový průměr $\bar{x}_1$. Tento výběr pochází z populace s průměrem μ_1. Druhý výběr má rozsah n_2, výběrový průměr $\bar{x}_2$. Pochází z populace

Tab. 6.3 Schéma porovnání průměrů ve dvou výběrech

Jednostranný test	Dvoustranný test
$H_0: (\mu_1 - \mu_2) = \Delta$	$H_0: (\mu_1 - \mu_2) = \Delta$
$H_1: (\mu_1 - \mu_2) > \Delta$	$H_1: (\mu_1 - \mu_2) \neq \Delta$
(nebo $H_1: (\mu_1 - \mu_2) < \Delta$)	

$$\bar{d} = \bar{x}_1 - \bar{x}_2$$

Testovací statistika:

$$t = \frac{\bar{d} - \Delta}{s_{\bar{d}}}$$

Oblast zamítnutí:	Oblast zamítnutí:
$t > t_\alpha$ (nebo $t < -t_\alpha$)	$t < -t_{\alpha/2}$ a $t > t_{\alpha/2}$

s průměrem μ_2. Tvar hypotéz a provedení testu pro uvedené situace určují vztahy uvedené v tabulce 6.3, v níž t_α je kritická hodnota t-testu. Je to taková t-hodnota, pro kterou platí, že $P(t > t_\alpha) = \alpha$ za platnosti nulové hypotézy. Kritické hodnoty vycházejí ze Studentova t-rozdělení. Testovací statistika obsahuje ve jmenovateli odhad směrodatné chyby čitatele $s_{\bar{d}}$.

Velikost rozdílu Δ volíme v závislosti na cílech výzkumu. Velmi často se spokojujeme s jednoduchým určením $\Delta = 0$, jež vyjadřuje rovnost obou průměrů μ_1 a μ_2. Dvoustranný interval spolehlivosti pro hladinu spolehlivosti $1 - \alpha$ má tvar

$$\Delta \in \bar{d} \pm t_{\alpha/2} s_{\bar{d}} \quad \text{nebo} \quad (\mu_1 - \mu_2) \in (\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} s_{\bar{d}}.$$

Velikost účinku ES (effect size) se obvykle měří odhadem Cohenova koeficientu Δ/σ, kde směrodatná odchylka σ měří variabilitu proměnné.

Rozlišujeme čtyři různé případy, jež vedou k odlišným volbám kritických hodnot a k odlišným výpočtům směrodatné chyby rozdílu průměrů $s_{\bar{d}}$. První tři se týkají situace dvou nezávislých výběrů, čtvrtá předpokládá párová data ze dvou závislých výběrů. V případě nezávislých výběrů jsme data získali pomocí dvou prostých náhodných výběrů z porovnávaných populací, v nichž má sledovaná proměnná normální rozdělení. Při větším počtu pozorování může být tento předpoklad nahrazen požadavkem náhodného přiřazení objektů do skupin. Předpoklady o rozdělení proměnných pro závislé výběry uvedeme v příslušném odstavci. Metody postupně popíšeme pomocí jednotného schématu.

Robustnost t-testu pro dva výběry vzhledem k odchylkám rozdělení dat od normality je značná. Jestliže rozsahy výběrů i tvary rozdělení jsou přibližně stejné, pak lze použít tabulky t-rozdělení i pro zešikmená dat již od rozsahu $n_1 = n_2 = 5$. Pokud jsou tvary rozdělení mezi výběry odlišné, je zapotřebí větších rozsahů, abychom mohli věřit v působení centrálního limitního teorému. Jako návod lze použít doporučení ze s. 207 s modifikací, že rozsah výběru se rovná součtu $n_1 + n_2$. Při plánování pokusů se dvěma výběry se snažíme, aby $n_1 = n_2$.

6.2.1 Metoda 1 – Dva velké nezávislé výběry

Jestliže rozsahy obou výběrů jsou velké (> 30) nebo oba výběry pocházejí z populací se známým rozptylem σ_1^2 and σ_2^2 a σ_1^2 and σ_2^2, pak dosadíme v tabulce 6.3 ve vztahu pro testovací statistiku za t symbol z a použijeme

$$s_{\bar{d}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

Pokud známe rozptyly, jejich hodnoty dosadíme místo výběrových rozptylů. Symbol z znamená, že použijeme tabulky kritických hodnot pro standardizované normální rozdělení. Mluvíme pak o z-testu průměrů. Takto postupujeme i v případě porovnání průměrů měření, která nemají normální rozdělení. Platnost procedury v tomto případě odvozujeme od působení centrálního limitního teorému a je podmíněna splněním požadavků na větší rozsah výběrů.

6.2.2 Metoda 2 – Dva nezávislé výběry se stejným rozptylem

Oba další případy rozlišují, zda rozptyly jsou v obou populacích stejné, nebo různé. Pokud přijmeme předpoklad stejných rozptylů sledované proměnné v populacích, počet stupňů volnosti je $n_1 + n_2 - 2$ a směrodatná chyba

$$s_{\bar{d}} = \sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, \qquad \text{kde} \quad s_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2},$$

přičemž s_1^2 a s_2^2 jsou vypočítané rozptyly v obou skupinách měření.

6.2.3 Metoda 3 – Dva nezávislé výběry, nestejné rozptyly

Tento případ je výpočetně nejkomplikovanější při hledání stupňů volnosti pro určení kritických hodnot:

$$st.v. = \left(\frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} \right)$$

Navíc uvedená formule je platná pouze přibližně. Výpočet směrodatné chyby odhadu je stejný jako v případě 1:

$$s_{\bar{d}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Uvedený vzorec se aproximativně s rostoucími rozsahy výběrů rovná výpočtu podle metody 1. Pokud jsou rozptyly stejné, pak dává podobné výsledky jako metoda 2.

Proč se objevují rozdíly ve výpočtech pro jednu problémovou situaci? Zaprvé při větším počtu pozorování se výpočty mohou zjednodušit (viz metoda 1). To je významné, jestliže nemáme k dispozici počítač. Zároveň test platí asymptoticky i pro měření, která nejsou normálně rozdělena. Třetí situace se týká případu, že v obou skupinách jsou statistické odlišné rozptyly. Při malém počtu pozorování tato skutečnost hraje roli při určení kritické meze, abychom opravdu dodrželi zvolenou hladinu významnosti našeho testovacího postupu.

PŘÍKLAD 6.5

Porovnání průměrů ve dvou nezávislých výběrech s nestejnými rozptyly

Pomocí experimentu chtěli výzkumníci ověřit, zda stupeň pohody u jedinců se zvýší, jestliže u nich indukujeme úsměv. Pokusu se účastnilo 40 dobrovolníků, kteří byli náhodně zařazeni buď do experimentální, nebo do kontrolní skupiny. Každá skupina měla velikost 20. Jedinci z obou skupin se dívali po určitou dobu na scénu s hrajícími si dětmi. Jedinci z druhé skupiny (experimentální skupina) měli mimikou napodobovat smích. Jedinci z první skupiny (kontrolní skupina) zachovávali naopak neutrální výraz. Po pokusu byl u všech jedinců změřen stupeň pohody pomocí odpovídajícího psychologického testu. Výsledkové skóry jsou zachyceny v tabulce 6.4. Krabicový graf (obr. 6.2) ukazuje rozdíly rozdělení výsledků ve skupinách. Také neindikuje žádné odlehlé hodnoty. Budeme zkoumat rozdílnosti pomocí numerických charakteristik. K tomu použijeme naše modelová data. Vycházíme z hodnot základních statistik, které jsou v tabulce 6.5.

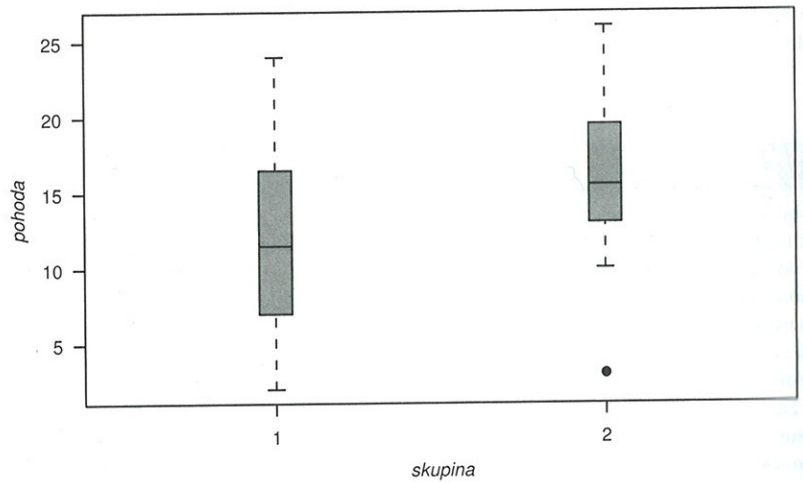
Tab. 6.4 Příklad porovnávání průměru ve dvou výběrech – stupeň pohody pokusných osob

Skupina 1	2	3	20	10	15	8	14	6	11	17	18	5	11	5	16	13	12	24	19	11
Skupina 2	15	10	14	19	16	26	23	12	20	13	17	3	13	17	23	14	20	15	13	17

Tab. 6.5 Příklad porovnávání průměru ve dvou výběrech – základní statistiky obou skupin

Skupina	*N*	Průměr	Směrodatná odchylka
experimentální	20	12	6,0087
kontrolní	20	16	5,1298

Obr. 6.2 Krabicový graf pro porovnání skupin



Obr. 6.3 Testy a intervaly spolehlivosti pro rozdíly teoretických průměrů

```
  Předpoklad          Diference  Směrod.    Směrod.   95% DM      95% HM
  o rozptylu  st.v.   průměrů    odchylka   chyba     průměru     průměru
  Stejné      38      -4         5,586638   1,76665   -7,576396   -0,4236038
  Nestejné    37,09   -4         7,9007     1,76665   -7,579287   -0,4207129
  Použité kritické hodnoty: T-alpha (rozptyly stejné) = 2,0244,
  T-alpha (rozptyly různé) = 2,0260

t-test, sekce pro stejné rozptyly

  Alternativní              Prav       Rozh.       Síla       Síla
  Hypotéza        t-hodn.   hlad.      (5%)        (a=0,05)   (a=0,01)
  Diference <> 0  -2,2642   0,029354   Odmítni Ho  0,597462   0,340933
  Diference < 0   -2,2642   0,014677   Odmítni Ho  0,718607   0,443265
  Diference > 0   -2,2642   0,985323   Přijmi Ho   0,000054   0,000003
  Diference: (Skupina=1)-(Skupina=2)

t-test, sekce pro různé rozptyly

  Alternativní              Prav       Rozh.       Síla       Síla
  Hypotéza        t-hodn.   hlad.      (5%)        (a=0,05)   (a=0,01)
  Difference <> 0 -2,2642   0,029500   Odmítni Ho  0,596913   0,340053
  Difference < 0  -2,2642   0,014750   Odmítni Ho  0,718268   0,442485
  Difference > 0  -2,2642   0,985250   Přijmi Ho   0,000055   0,000003
  Difference: (Skupina=1)-(Skupina=2)

Test rovnosti rozptylů jejich poměrem   F=1,3720 p=0,49730
Nelze zamítnout předpoklad stejných rozptylů.
```

Použitím procedury z programového systému NCSS 6.0 pro výpočet *t*-testu získáme výstupní sestavu s podrobnými výsledky (obr. 6.3). Podíváme se na dolní řádek tabulky s údaji o testu rovnosti rozptylů (výklad na s. 219). V závislosti na nich vybereme správné údaje z ostatních částí tabulky. Protože test nezamítl hypotézu rovnosti rozptylů, další údaje hledáme v sekci *t*-testu pro stejné rozptyly. Proto je hodnota směrodatné odchylky rozdílu průměrů $s_{\bar{d}} = 1{,}76$. Tuto hodnotu využijeme při výpočtu intervalu spolehlivosti, nebo ji dělíme hodnotu rozdílu průměrů $\bar{x}_1 - \bar{x}_2 = -4$, abychom získali testovací *t*-statistiku. V první části tabulky je uveden 95% interval spolehlivosti pro rozdíly průměrů pomocí dolní (DM) a horní (HM) meze. V další části najdeme výsledky testu rozdílu průměrů – test jednostranný i dvoustranný. Výpočty indikují statistické prokázání rozdílu mezi průměry skóre měření nálady v experimentální a kontrolní skupině. Použijeme testovací statistiku a hodnocení pro jednostranný test $t = -2{,}2642$ a $p = 0{,}014677$, protože předpokládáme určitý směr poruchy hypotézy. Nulovou hypotézu rovnosti průměrů lze zamítnout na hladině 5 %, ale ne na hladině 1 %.

Interval spolehlivosti s 95% hladinou spolehlivosti pro rozdíl průměrů má hodnotu (−7,576396; −0,4236038). Protože nepokrývá hodnotu nula, potvrzuje rozhodnutí o zamítnutí hypotézy rovnosti průměrů. Z jednostranných testů pochopitelně pouze jeden může indikovat statisticky významný rozdíl. Jako indikátor efektu intervence použijeme Cohenův index d. Pro naše výsledky $d = 4/5{,}59 = 0{,}71$.

Komentujme sloupce nadepsané „síla". Síla testu je pravděpodobnost zamítnutí hypotézy rovnosti průměrů za předpokladu, že si nejsou skutečně rovny. Požaduje se vysoká síla. Výpočty se opírají o odhady směrodatných odchylek ve skupinách a rozdíl výběrových průměrů, jako by to byly teoretické parametry. Sílu testu uvažujeme, pokud test nezamítl nulovou hypotézu. Vypočítaná hodnota síly testu indikuje, co bychom měli udělat, aby nulová hypotéza byla zamítnuta. Jestliže přijmeme nulovou hypotézu s velkou silou, nelze toho moc dělat. Oba průměry jsou si velmi blízko. Jestliže přijímáme hypotézu s malou silou, můžeme provést jednu z následujících akcí:

1. Zvýšíme hladinu α. Například budeme testovat na hladině $\alpha = 0{,}05$ místo na hladině $\alpha = 0{,}01$. Zvětšení této hladiny zvyšuje také sílu.
2. Zvýšíme rozsah výběru n. Tím zvýšíme sílu testu, pokud byla malá (< 0,80).
3. Snížíme velikost rozptylu měření. Uvažujeme možnost zlepšit plán výzkumu tak, aby naše měření byla přesnější a odstranily se vnější zdroje variability.

6.2.4 Párová data – dva závislé výběry

S problémem hodnocení párových dat se setkáváme ve čtyřech možných situacích:

1. Ve výzkumu typu pretest-posttest, kdy měříme objekt dvakrát, před pokusem a po něm. Každá hodnota x_i prvního výběru má odpovídající hodnotu y_i ve druhém výběru.
2. Použili jsme tzv. blokové schéma, kdy jsou párově vyrovnané objekty rozděleny podle hodnoty rušivého faktoru do bloků, je jim náhodně přiřazeno ošetření a pak jsou srovnávány jejich hodnoty závisle proměnné.
3. Na objektech máme párové části (oči, uši, paže, nohy) a provedeme na nich měření.
4. Dvě ošetření jsou aplikována u jedinců v náhodném pořadí.

V našem výkladu vycházíme z podobnosti s posuzováním průměrů ve dvou výběrech. Proto využijeme příslušné vzorce pro testovací statistiku a interval spolehlivosti z našeho schématu na s. 209. Směrodatnou chybu spočítáme podle vzorce

$$s_{\bar{d}} = \frac{1}{\sqrt{n}} \sqrt{\frac{\sum_i (d_i - \bar{d})^2}{n-1}},$$

v němž počítáme s rozdíly $d_i = x_i - y_i$. Jestliže n je počet párů, pak stupně volnosti (*st. v.*) pro hledání kritických hodnot nebo p-hodnot významnosti pomocí Studentova rozdělení jsou $n - 1$. O rozdílech předpokládáme, že je lze považovat za náhodný výběr z populace s normálním rozdělením, jehož rozptyl neznáme.

Je užitečné si uvědomit, že provedení t-testu v této situaci je v podstatě stejné jako použití t-testu pro jeden výběr pro rozdíly $d_i = x_i - y_i$. Obvykle testujeme hypotézu $H_0: \mu_d = 0$. V tomto případě má testovací statistika jednoduchý tvar

$$t = \frac{\bar{d}}{s_{\bar{d}}}.$$

Takto konstruovaný t-test pro párová data se nazývá obvykle **párový t-test**. Při větším rozsahu výběru použijeme asymptoticky platné kritické hodnoty normálního rozdělení. V tomto případě nemusíme striktně trvat na předpokladu, že rozdíly $d_i = x_i - y_i$ mají normální rozdělení. Interval spolehlivosti pro μ_d má tvar

$$\mu_d \in (\bar{d} - t_{\alpha/2} s_{\bar{d}}\,;\ \bar{d} + t_{\alpha/2} s_{\bar{d}}).$$

PŘÍKLAD 6.6

Porovnání výběrů pro párová data

Párový t-test demonstrujeme pro modelová data z tabulky 2.9 kapitoly 2.5. Hodnotíme u žáků rozdíly ve skoku do dálky při tréninku a při závodě. Použitím procedury pro výpočet párového t-testu z programového systému NCSS 6.0 dostaneme částečné výsledky

Obr. 6.4 Výstupní sestava párového t-testu

Proměná	Počet	Průměr	Směrod. odchylka	Směrod. chyba	95% DM průměru	95% HM průměru
SkokTrénink	26	3,582692	0,2880772	5,6496E-02	3,466336	3,699049
SkokZavod	26	3,578846	0,2953551	5,7929E-02	3,45955	3,698143
Rozdíl	26	3,846154E-03	0,1310901	2,5708E-02	-4,9102E-02	5,6794E-02

t kritická mez pro intervaly spolehlivosti = 2,0930

Párový t-test

Alternativní hypotézy	t-hodnota	Prav. úroveň	Rozh. (5%)	Síla (a=0,05)	Síla (a=0,01)
Skok-SkokZavod <>0	0,1496	0,882277	Přijmi Ho	0,052377	0,010735
Skok-SkokZavod <0	0,1496	0,558861	Přijmi Ho	0,036689	0,006790
Skok-SkokZavod >0	0,1496	0,441139	Přijmi Ho	0,066904	0,014460

(obr. 6.4). Výstupní sestava programu obsahuje mnoho údajů. Každý z nich má svůj smysl. Hodnota $\bar{d}$ je 0,0038. Hodnota $s_{\bar{d}}$ je 0,0257. V první části sestavy jsou uvedeny horní (HM) a dolní meze (DM) intervalu spolehlivosti na hladině spolehlivosti 95 % pro průměry v obou pokusech a průměrný rozdíl. Druhá část obsahuje výsledky pro jednostranné testy a dvoustranný párový t-test. Protože interval spolehlivosti pro rozdíl obsahuje nulu, nelze nulovou hypotézu o neexistenci rozdílu mezi výkonem při tréninku a výkonem v závodě zamítnout. To potvrzují i hodnoty t-testu. Sloupec „Síla" jsme komentovali v předchozím příkladu.

Poznamenejme, že párový t-test má smysl, pokud alternativní hypotézu lze vyjádřit modelem pro měření v jednotlivých párech i ve formě $y_i = x_i + d + e_i$, kde složka e_i má nulovou střední hodnotu. Podstatné je, že d je konstantní pro všechny páry. Někdy je však vhodnější model $y_i = a + bx_i + e_i$, pro který tento test není vhodný kromě případu, že $b = 1$.

Směrodatnou chybu $s_{\bar{d}}$ v párové t-statistice lze také vyjádřit pomocí vztahu

$$s_{\bar{d}} = \sqrt{1/n(s_x^2 + s_y^2 - 2rs_xs_y)},$$

v němž r je korelační koeficient (více o něm kap. 7.2). Z výrazu je patrné, že čím je vyšší korelace mezi oběma řadami měření, tím je větší testovací t-statistika. To má za následek nárůst síly testu. Vzorec dokumentuje, proč je výhodnější provést znáhodněný blokový experiment, a ne experiment s dvěma nezávislými skupinami, pokud máme představu o rušivé proměnné. Tuto úvahu uplatňujeme i v jiných souvislostech aplikace párového t-testu.

6.3 Hodnocení rozptylu

Metody statistického usuzování o velikosti rozptýlenosti měření kolem centrální hodnoty se nepoužívají tak často jako metody pro posouzení středních hodnot. Je však zřejmé, že ve výzkumu se nemusíme zajímat pouze o střední hodnoty sledovaných proměnných. Posouzení rozptýlenosti hraje roli při ověřování platnosti předpokladů t-testu při porovnávání průměrů dvou nezávislých populací a při volbě správné metody. Zkoumání velikosti rozptýlenosti dat však může být i vlastním cílem výzkumu.

a) Firma tvrdí, že dováží jablka, jejichž velikost má určitou průměrnou hodnotu a proměnlivost, která je určena směrodatnou odchylku 12 gramů. Tvrzení lze ověřit tím, že statisticky přezkoušíme pomocí náhodně vybraného vzorku jablek tento předpoklad.
b) Monitorujeme tepovou frekvenci v klidu a ověřujeme její stabilitu porovnáním zjištěného rozptylu tepových frekvencí s normou.

c) Porovnáváme kvalitu vyráběných výrobků na dvou strojích. Nezajímají nás pouze průměrné hodnoty sledovaného indikátoru kvality, ale i hodnoty jeho proměnlivosti. Ptáme se, zda je výrobní proces dostatečně stabilní.

Uvedeme způsoby statistického hodnocení, jež předpokládají normalitu dat. Budou zaměřeny na posouzení parametru rozptylu σ^2 normálního rozdělení pomocí jeho odhadu statistikou s^2. Na rozdíl od t-testů pro hodnocení průměru tyto metody nejsou bohužel robustní k porušení předpokladu normality ani rezistentní vůči odlehlým hodnotám (Conover, 1981). Tyto slabiny se nezlepšují ani s rostoucím počtem použitých měření. Proto se mají používat s největší opatrností po přezkoušení normality dat graficky a pomocí příslušného statistického testu. Jestliže data nemají normální rozdělení, volíme také jinou charakteristiku pro numerické posouzení jejich rozptýlenosti, protože empirický rozptyl s^2 není v tomto případě optimální mírou.

6.3.1 Hodnocení rozptylu v jednom výběru

Předpokládáme, že sledovaná proměnná má normální rozdělení. Analýzu rozptýlenosti dat založíme na velikosti rozptylu σ^2. Teprve v dalším kroku usuzujeme pomocí hodnoty rozptylu s^2 o velikosti směrodatné odchylky s.

Interval spolehlivosti

K nalezení intervalu spolehlivosti musíme nejdříve vypočítat bodový odhad rozptylu:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

Interval spolehlivosti pro σ^2 s hladinou spolehlivosti $1 - \alpha$ zjistíme pomocí kvantilů χ^2-rozdělení s $n - 1$ stupni volnosti. Interval spolehlivosti má tvar:

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

PŘÍKLAD 6.7

Intervalový odhad pro rozptyl dat v jednom výběru

Jestliže $n = 31$ a $s = 1000$, pak stupně volnosti jsou $n - 1 = 30$, takže pro interval spolehlivosti s hladinou 95 %, nalezneme kvantily $\chi^2_{0,025} = 46,988$ and $\chi^2_{0,975} = 16,789$. Dosazením těchto hodnot do vzorce dostaneme interval spolehlivosti ve tvaru

$$\frac{30 \times 1000^2}{46,988} \le \sigma^2 \le \frac{30 \times 1000^2}{16,789},$$

který lze zapsat také ve tvaru $0{,}638 \times 1000^2 \leq \sigma^2 \leq 1{,}786 \times 1000^2$. Jestliže chceme interval spolehlivosti pro směrodatnou odchylku, vypočteme ze všech tří hodnot druhou odmocninu a dostaneme intervaly $0{,}799 \times 1000 \leq \sigma \leq 1{,}337 \times 1000$ neboli $799 \leq \sigma \leq 1337$.

Test hypotézy

Předpokládejme hodnotu rozptylu proměnné σ_0^2. Tedy nulová hypotéza je

$$H_0 : \sigma^2 = \sigma_0^2$$

a alternativní hypotéza

$$H_1 : \sigma^2 \neq \sigma_0^2.$$

Za předpokladu normálního rozdělení proměnné má podíl

$$\chi^2 = \frac{(n-1)\, s^2}{\sigma_0^2}$$

rozdělení χ^2 s $n-1$ stupni volnosti. Pro dvoustranný test na hladině významnosti α najdeme kvantily $\chi^2_{1-\alpha/2}$ a $\chi^2_{\alpha/2}$, a nulovou hypotézu přijmeme, pokud vypočtená testovací statistika χ^2 leží mezi nimi.

PŘÍKLAD 6.8

Testování hypotéz o rozptylu dat v jednom výběru

Předpokládejme, že rozdělení dat je normální. Rozptyl má údajně hodnotu 64. Pro 17 údajů jsme zjistili rozptyl 100. Provedeme test s 2% hladinou významnosti. Situaci testování lze popsat takto:

$$H_0 : \sigma^2 = 64 \qquad n = 17$$
$$H_1 : \sigma^2 \neq 64 \qquad st.v. = n - 1 = 16$$
$$s^2 = 100$$
$$\sigma_0^2 = 64$$
$$\alpha = 0{,}02$$

Spočítáme

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{16 \times 100}{64} = 25.$$

Protože $\alpha/2 = 0{,}01$ a $st.v. = 16$, nalezneme kvantily $\chi^2_{0{,}99} = 5{,}812$ a $\chi^2_{0{,}01} = 32{,}006$. Protože vypočítaná hodnota leží v intervalu vytvořeném těmito kvantily, nemůžeme zamítnout nulovou hypotézu na hladině 0,02.

6.3.2 Porovnání rozptylů ve dvou nezávislých výběrech

Jednoduché řešení, jak vyhodnotit rozdíl rozptýlenosti ve skupinách měření, navrhl statistik H. Levene (Havránek, 1993). Doporučil nahradit každé měření x_i jeho absolutní odchylkou (nebo kvadratickou odchylkou) od odpovídajícího skupinového průměru a pak obě množiny takto vzniklých hodnot analyzovat t-testem pro dvě nezávislé skupiny. Průměr hodnot absolutních odchylek stejně jako průměr kvadratických odchylek totiž charakterizuje velikost rozptýlenosti původních měření. Přitom se využije asymptotická platnost t-testu pro hodnoty, které nemají normální rozdělení, při větším počtu měření.

Tradičnější je však postup založený na poměru rozptylů a využití F-rozdělení. Předpokladem F-testu pro porovnání rozptylů je normální rozdělení dat. F-test pro rozptyly je velmi citlivý na porušení předpokladu normality proměnných, jeho nesplnění může značně zvýšit pravděpodobnost chyby I. druhu. Hodnocení shody rozptýlenosti doplňujeme grafickým posouzením dat. Jestliže rozdělení jsou zešikmená, musíme použít jiný postup.

Pokud se rozptyly neliší, měl by být jejich poměr roven přibližně jedné. Z tohoto požadavku vychází konstrukce F-testu shody rozptylů. Opíráme se o poznatek, že chování rozptylu lze popsat pomocí χ^2 rozdělení (viz kap. 4.6.5), a o teoretický výsledek, že poměr dvou nezávislých χ^2-rozdělení vydělených jejich stupni volnosti má F-rozdělení. Test rovnosti rozptylů se realizuje tak, že provedeme dvě srovnání a hypotézu rovnosti zamítneme, když

$$\frac{s_1^2}{s_2^2} \geq F_{1-\alpha/2} \qquad \text{nebo} \qquad \frac{s_2^2}{s_1^2} \geq F_{1-\alpha/2},$$

kde $F_{1-\alpha/2}$ je kvantil F-rozdělení s n_1 a s n_2 stupni volnosti s hladinou $(1-\alpha/2)$. V případě jednostranné alternativy se použije jen jedna z těchto nerovností, přičemž místo $\alpha/2$ zvolíme α (pro alternativu $H_1 : \sigma_1^2 > \sigma_2^2$ užíváme první nerovnost).

V tabulkách jsou obvykle uvedeny přímo kritické meze pro jednostranné testy. Pak odpovídá kritická mez pro jednostranný test na hladině 0,01 pro alternativu $H_1 : \sigma_1^2 > \sigma_2^2$ kvantilu F-rozdělení s hladinou 0,99. Tento přístup zvolíme, jestliže využíváme tabulky v příloze B.

PŘÍKLAD 6.9

Porovnání rozptylů ve dvou nezávislých výběrech

Předpokládejme, že naměřené hodnoty pocházejí z normálního rozdělení. Testujeme jednostrannou hypotézu:

$$H_0: \sigma_1^2 \leq \sigma_2^2$$

$$H_1: \sigma_1^2 > \sigma_2^2$$

V prvním výběru jsme naměřili hodnoty 1,0; 2,0; 1,9; 4,0 a 0,5. Výběrový rozptyl má hodnotu $s_1^2 = 1{,}797$ a $n_1 = 5$. Druhý výběr přinesl hodnoty 0,6; 0,7; 0,8; 1,1; 0,4 a 0,6. Pro druhý výběrová rozptyl jsme spočítali hodnotu $s_2^2 = 0{,}056$ a $n_2 = 6$. Testovací statistika má tedy hodnotu:

$$F = \frac{s_1^2}{s_2^2} = \frac{1{,}797}{0{,}056} = 32{,}09$$

Nalezneme kritickou hodnotu $F_{0{,}01}(4,5) = 11{,}39$ v tabulce F-rozdělení. Jelikož testovací statistika je mnohem větší než kritická mez, lze nulovou hypotézu v této modelové situaci zamítnout.

6.4 Neparametrické posouzení středních hodnot a test normality dat

Postupy uvedené v předchozích odstavcích předpokládaly normální rozdělení proměnné v populaci. Protože v mnoha situacích je tato podmínka omezující, statistici vyvinuli alternativní postupy s méně přísnými předpoklady. Neparametrické testy lze použít za obecnějších podmínek, a proto jsou populární. Jejich teorii a aplikacím se věnuje rozsáhlá literatura (např. Sprent, 2001). Poznamenejme, že za neparametrický test lze při větších výběrech považovat i popsaný t-test, protože jeho platnost je zaručena, jak jsme uvedli, působením centrálního limitního teorému.

Popíšeme některé způsoby posuzování střední hodnoty rozdělení proměnné pro jeden nebo dva výběry, jimiž nahrazujeme t-testy o průměrech. Používáme je v těchto případech:

- data nejsou normálně rozdělena;
- data mají ordinální charakter;
- výběry jsou malé nebo existují velké rozdíly mezi rozsahy výběrů;
- chceme posílit validitu výsledků parametrických metod.

Můžeme je aplikovat zároveň s dříve popsanými t-testy a porovnávat jejich výsledky. Pokud jsou rozdílné, zkoumáme, proč tomu tak je. Někdy nás tento přístup upozorní na nekonzistenci dat.

Popíšeme podrobněji použití znaménkového testu a testů Wilcoxonova typu založených na pořadí měření. V předchozí kapitole jsme vysvětlili, jak se odvozují přesné varianty těchto testů. V této kapitole se soustředíme na jejich varianty využívající asymptoticky platné kritické meze nebo asymptotické výpočty příslušných p hodnot. Znaménkový test je jednodušší, protože se opírá pouze o četnosti měření, která leží pod nebo nad zvolenou hodnotu. Testy Wilcoxonova typu jsou odvozeny za předpokladu spojitého rozdělení zkoumané proměnné. Jestliže ve výběru existuje více stejných pozorování, musíme použít jejich modifikace. Tyto testy lze provést také tak, že na pořadí aplikujeme vzorce pro t-testy popsané v předchozím odstavci a použijeme vhodné kritické hodnoty. Lze ukázat, že takto sestrojené testovací procedury jsou ekvivalentní testům Wilcoxonova typu (Conover, 1981). Popsané procedury lze použít i pro sestrojení neparametrického intervalu spolehlivosti pro střední hodnotu.

Při porovnávání dvou nezávislých výběrů mají neparametrické testy platnost pro dvě rozdílné výběrové situace:

a) máme k dispozici dva nezávislé náhodné výběry ze dvou populací a zobecňujeme na příslušné populace;
b) objekty byly přiřazeny do nezávislých skupin randomizovaně, závěr testu má lokální platnost pro posouzení rozdílnosti dvou takto vzniklých skupin dat (viz kap. 9.4).

Na konci kapitoly popíšeme test hypotézy, že rozdělení proměnné je normální.

6.4.1 Znaménkový test střední hodnoty pro jeden výběr

Variantu tohoto neparametrického testu jsme probrali v modelovém příkladu v kapitole 5.4.1. Někdy se tento test nazývá mediánový test, protože si všímá směru odchylky měření od předpokládaného mediánu. Jestliže platí o tomto parametru, že má hodnotu $\tilde{\mu}_0$, pak četnosti měření, která se nacházejí na obou stranách $\tilde{\mu}_0$, by měly být přibližně stejné. Z této skutečnosti vychází znaménkový test hypotézy $H_0: \tilde{\mu} = \tilde{\mu}_0$ proti alternativě $H_1: \tilde{\mu} \neq \tilde{\mu}_0$.

Uvažujme rozdíly $d_i = x_i - \mu_0$ a spočítejme počet Z_+ hodnot d_i, jež mají kladné znaménko, resp. počet Z_- udávající, kolik hodnot d_i má záporné znaménko. Při výpočtech vynecháváme páry, kdy $d_i = 0$. Označme součet $Z_+ + Z_- = n$. Obě hodnoty Z mají za platnosti nulové hypotézy binomické rozdělení s parametry n

a 0,5. Test této hypotézy provádíme pomocí přesného testu, v němž využijeme tabulky binomického rozdělení pro hodnotu Z_+ (tab. VIII v příloze B), nebo provedeme asymptoticky platný test na základě aproximace pomocí normálního rozdělení náhodné proměnné. Testovací statistika má pak tvar

$$z = \frac{2Z_+ - n}{\sqrt{n}}.$$

Tento test lze uplatnit, pokud n je větší než 25. Hypotézu H_0 nezamítáme, jestliže testovací statistika z leží uvnitř intervalu $\pm z_{\alpha/2}$, kde $\pm z_{\alpha/2}$ je kritická hodnota standardizovaného normálního rozdělení. Jinak vyjádřeno – pokud absolutní hodnota testovací statistiky je větší než $\pm z_{\alpha/2}$, hypotézu H_0 zamítáme. Při jednostranné alternativní hypotéze srovnáme statistiku z s kritickou mezí $\pm z_\alpha$.

PŘÍKLAD 6.10

Provedení znaménkového testu střední hodnoty pro jeden výběr

Pro modelová data (s. 77) přezkoušíme hypotézu pro výkon ve skoku dalekém kategorie chlapci H_0: $\tilde{\mu}_0 = 3{,}8$ nejdříve asymptoticky platným testem. Posuzujeme výkony chlapců, které měly hodnoty (v metrech): 3,75; 3,75; 3,9; 3,65; 3,65; 3,65; 3,6; 4,05; 4,05; 3,5; 3,45; 4,2; 3,3. Zjistíme, že 4 měření ze 13 hodnot jsou větší než 3,8. Dosadíme do vzorce pro výpočet testovací statistiky a získáme hodnotu $z = (8 - 13)/\sqrt{13} = 1{,}388$, která není statisticky významná, pokud ji srovnáme s běžnou kritickou mezí standardizovaného normálního rozdělení (1,96). Nulovou hypotézu na základě tohoto testu tedy nemůžeme zamítnout.

Přesný test provedeme tak, že využijeme tabulku kumulativních binomických pravděpodobností z přílohy B. Pracujeme s částí tabulky, jež se týká binomického rozdělení $B(0{,}5; 13)$ – viz též tabulka 6.6. Pomocí tabulky zjistíme pravděpodobnost jevu $P(X < 5)$ pro rozdělení $B(0{,}5; 13)$. To je pravděpodobnost jevu, že 4 nebo méně měření jsou větší než 3,8. Nalezneme hodnotu 0,13342. Také tato hodnota nesvědčí v neprospěch nulové hypotézy.

Tab. 6.6 Část tabulky kumulativních pravděpodobností binomického rozdělení $B(0{,}5; 13)$

Hodnota x	0	1	2	3	4	5	6
Pravděpodobnost $P(X \le x)$	0,00012	0,00171	0,01123	0,04614	0,13342	0,29053	0,5
Hodnota x	7	8	9	10	11	12	13
Pravděpodobnost $P(X \le x)$	0,70947	0,86658	0,95386	0,98877	0,99829	0,99988	1

6.4.2 Wilcoxonův test střední hodnoty pro jeden výběr

Test, který popíšeme, navrhl spolu s testem pro dva nezávislé výběry F. Wilcoxon v roce 1945. Je silnější než znaménkový test. Předpokladem testu je spojité symetrické rozdělení dat. Postup vychází z absolutních hodnot rozdílů mezi měřeními a předpokládanou hodnotou mediánu $\tilde{\mu}_0$, které seřadíme podle velikosti. Sečteme zvlášť hodnoty pořadí rozdílů, u kterých se hodnoty měření nacházely pod předpokládaným mediánem dat, a zvlášť pořadí měření, která ležela nad ním. Získáme tak součty T_- a T_+. Pokud platí nulová hypotéza, měly by být součty T_- a T_+ přibližně stejné – mít hodnotu $n(n+1)/4$. Poznamenejme, že v případě skupin shodných absolutních hodnot rozdílů měření dosadíme za jejich pořadí průměrnou hodnotu jejich pořadí. Tuto úpravu popisujeme v příkladu 6.11.

Test nulové hypotézy H_0: $\tilde{\mu} = \tilde{\mu}_0$ proti alternativě H_1: $\tilde{\mu} \neq \tilde{\mu}_0$ provedeme srovnáním menší z obou hodnot T_+ a T_- s přesnými kritickými hodnotami (tab. VI v příloze B) pro zvolenou hladinu významnosti. Pokud je T menší než kritická mez, hypotézu H_0 můžeme zamítnout na zvolené hladině významnosti. Dosaženou p-hodnotu lze vypočítat přesně pomocí Fisherova principu permutování.

Při větším rozsahu výběru použijeme aproximaci pomocí normálního rozdělení. Testovací statistika se pak spočte pomocí standardizace veličiny T_+ její teoretickou střední hodnotou $E(T)$ a směrodatnou odchylkou σ_T za platnosti nulové hypotézy

$$z = \frac{T_+ - E(T)}{\sigma_T},$$

kde

$$E(T) = n(n+1)/4 \quad \text{a} \quad \sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}}.$$

Podobně jako u znaménkového testu, lze tuto aproximaci uplatnit, pokud n je větší než 25. Jestliže testovací statistika z leží v intervalu $\pm z_{\alpha/2}$ (kde $\pm z_{\alpha/2}$ je kritická hodnota standardizovaného normálního rozdělení), nezamítáme H_0. Jinak vyjádřeno – pokud absolutní hodnota testovací statistiky je větší než $\pm z_{\alpha/2}$, hypotézu H_0 zamítáme. Vzorec se modifikuje, pokud seřazené absolutní rozdíly měření obsahují více shodných hodnot.

Při větším počtu shod upravíme hodnotu σ_T. Někdy se používá při zohlednění shodných hodnot následující přístup. Označme výraz pod odmocninou ve výpočtu σ_T symbolem A. Místo A dosadíme výraz $A - B/48$, v němž se B vypočítá pomocí vzorce $B = \sum_{i=1}^{r} (t_i^2 - t_i)/12$, kde r je počet skupin shodných měření a t_i jsou postupně počty stejných měření ve skupinách i.

Rychlejší je tento postup. Označme R_i pořadí přiřazená absolutním odchylkám pozorování od mediánu. Upravené σ_T má hodnotu $\sigma_T = 0{,}5\sqrt{\sum R_i^2}$.

PŘÍKLAD 6.11

Wilcoxonův test střední hodnoty pro jeden výběr

Stejně jako v předchozím příkladu pro modelová data (s. 77) přezkoušíme hypotézu pro výkon ve skoku dalekém kategorie chlapci $H_0: \tilde{\mu}_0 = 3{,}8$, $H_1: \tilde{\mu}_0 \neq 3{,}8$ dvoustranným testem. S daty provedeme naznačené operace (viz tab. 6.7). Předposlední řádek tabulky obsahuje absolutní rozdíly seřazené podle velikosti. Druhý řádek indikuje znaménkem, zda bylo měření pod, nebo nad hodnotou 3,8 m.

Tab. 6.7 Příklad provádění Wilcoxonova testu střední hodnoty

Délka skoku x_i [m]	3,75	3,75	3,9	3,65	3,65	3,65	3,6	4,05	4,05	3,5	3,45	4,2	3,3
$x_i - \tilde{\mu}_0$	−0,05	−0,05	0,1	−0,15	−0,15	−0,15	−0,2	0,25	0,25	−0,3	−0,35	0,4	−0,5
Seřazené $\lvert x_i - \tilde{\mu}_0 \rvert$	0,05	0,05	0,1	0,15	0,15	0,15	0,2	0,25	0,25	0,3	0,35	0,4	0,5
Pořadí	1,5	1,5	3	5	5	5	7	8,5	8,5	10	11	12	13

Vychází $T_+ = 32$ a $T_- = 56$. Nejdříve provedeme srovnání $T_+ = 32$ s přesnou kritickou hodnotou pro hladinu významnosti 0,05 (tab. VI v příloze B). Protože T_+ není menší než kritická mez 13, nemůžeme hypotézu zamítnout. Vypočítáme také z-statistiku a získáme hodnotu

$$z = \frac{32 - 13 \times 14/4}{\sqrt{\dfrac{13 \times 14 \times 27}{24}}} = -0{,}943,$$

která není statisticky významná, když ji posoudíme kritickou hodnotou standardizovaného normálního rozdělení (1,96). Testovací statistiku jsme vypočítali bez úpravy na shody. V první skupině (1,5; 1,5) jsou dvě shody, v další skupině (5; 5; 5) jsou tři shody a ve třetí skupině (8,5; 8,5) jsou opět dvě shody, zřejmě $r = 3$. Opravný faktor B má hodnotu $[(4-2)+(9-3)+(4-2)]/12 = 10/12 = 0{,}83$. Oprava v tomto případě neovlivní významně hodnotu testovací statistiky.

6.4.3 Znaménkový a Wilcoxonův test pro dva závislé výběry

V této situaci použijeme neparametrické testy pro hodnoty $d_i = x_i - y_i$, kde x_i jsou měření v prvním výběru a y_i jsou odpovídající měření ve druhém výběru. Předpokládáme, že rozdíly mají symetrické rozdělení. Pro hodnoty d_i testujeme

hypotézu, že jejich teoretické rozdělení má nulový průměr nebo medián, znaménkovým nebo Wilcoxonovým jednovýběrovým testem. Pokud předem specifikujeme hodnotu účinku Δ, testujeme nulovou hypotézu, že střední hodnota rozdílů d_i má úroveň Δ. Test lze provést i sestrojením intervalu spolehlivosti pro medián nebo průměr pomocí hodnot d_i. Výzkumné plány, které vedou ke dvěma závislým výběrům, jsme popsali v kapitole o párovém t-testu.

PŘÍKLAD 6.12

Znaménkový test pro dva závislé výběry

Použijeme modelová data a srovnáme výkony ve skoku dalekém při tréninku a při závodě. Hypotézu o rovnosti průměrů u skupiny 26 jedinců otestujeme znaménkovým testem. Nejdříve spočítáme, u kolika jedinců byl výkon při závodě lepší než při tréninku, a dostáváme $Z' = 14$. Hodnotu dosadíme do vzorce pro testovací statistiku

$$z = \frac{28 - 26}{\sqrt{26}} = 0{,}39$$

a tento výsledek porovnáme s kritickou hodnotou 1,96 ($\alpha = 0{,}05$), kterou jsme vyhledali pomocí tabulky distribuční funkce standardizovaného rozdělení (tab. II v příloze B). Hodnota testovací statistiky $z = 0{,}39$ neindikuje porušení nulové hypotézy.

PŘÍKLAD 6.13

Wilcoxonův test pro dva závislé výběry

Ve skupině 12 studentů se sledovala srdeční frekvence při změně polohy z lehu do stoje. Získaly se tyto rozdíly frekvencí počtu tepů srdce za minutu:

−2; 4; 8; 25; −5; 16; 3; 1; 12; 17; 20; 9.

Jestliže předpokládáme symetrii rozdělení těchto rozdílů, můžeme použít Wilcoxonovu statistiku pro jeden výběr. Budeme testovat nulovou hypotézu, že rozdíl obou tepových frekvencí je v průměru 15 ($H_0: \Delta = 15$) proti alternativě $H_1: \Delta \neq 15$. Data seřadíme podle velikosti absolutních hodnot po odečtení hodnoty 15 od každého údaje:

2; −3; 5; −6; −7; 10; −11; −12; −14; −17; −20

Znaménkem opatřená pořadí vytvoří sekvenci:

1; 2; −3; 4; −5; −6; 7; −8; −9; −10; −11; −12

Součet kladných pořadí dá hodnotu $T_+ = 1 + 2 + 4 + 7 = 14$. Použijeme asymptotickou testovací statistiku

$$z = \frac{14 - (12 \times 13)/4}{\sqrt{\dfrac{12 \times 13 \times 25}{24}}} = -1{,}9612.$$

Transformací na hodnotu hladiny významnosti pomocí Excelu dostaneme $p = 0{,}0499$. Takže evidence proti nulové hypotéze významná na hladině 0,05 existuje, ale je poměrně slabá. Jestliže na stejné hodnoty použijeme znaménkový test, dostaneme

$$z = \frac{8-6}{\sqrt{12}} = 0{,}577.$$

Po transformaci do pravděpodobnostního měřítka dostaneme $p = 0{,}14$. Znaménkový test tedy neindikuje porušení nulové hypotézy. Výsledek ukazuje, jak se projevuje menší statistická síla znaménkového testu.

6.4.4 Konstrukce neparametrických intervalů spolehlivosti

Jak znaménkový test, tak Wilcoxonův test pro jeden výběr lze využít při konstrukci intervalů spolehlivosti pro hodnoty mediánu nebo průměru za předpokladu symetrického rozdělení. Způsob, jak se krajní meze intervalu hledají, spočívá v postupném provádění testů pro různé hodnoty parametru mediánu a určení, které extrémní hodnoty parametru příslušný test ještě nezamítne. Tento postup je numericky velmi náročný a bez příslušného programu sotva proveditelný. Ukážeme na příkladu, jejž popsal Sprent (2001), jak se hodnoty intervalového odhadu liší, jestliže použijeme znaménkový test, Wilcoxonův test a t-test.

PŘÍKLAD 6.14

Konstrukce neparametických intervalů spolehlivosti

Půjde o úlohu získat intervaly spolehlivosti pro medián diferencí systolického tlaku (mm Hg) změřeného před a po cvičení ve skupině 24 studentů, která byla náhodně získána ze skupiny 300 studentů. Uvádíme získané rozdíly naměřených hodnot seřazené podle velikosti:

−5; −5; 0; 2; 10; 15; 15; 15; 18; 20; 20; 20; 20; 22; 30; 30; 34; 40; 40; 40; 41; 47; 80; 85

K výpočtům byl využit program *StatXact*, jenž je vhodný pro realizaci různých neparametrických testů a určení jejich přesných hladin významnosti. Pomocí tohoto programu byly nalezeny 95% intervaly spolehlivosti na základě znaménkového testu, Wilcoxonova testu a t-testu (odpovídá běžně používanému intervalu spolehlivosti), které uvádí tabulka 6.8. Je patrné, že ve shodě s teorií dává znaménkový test nejširší interval. Ačkoli data jsou mírně zešikmená, všechny tři intervaly jsou centrovány kolem přibližně stejné hodnoty. Sprent

(1984, s. 127) také zkoumal vliv odlehlé hodnoty, způsobené chybou zápisu, která skutečně při přenosu dat nastala. Nejmenší diference −5 měla původně hodnotu −103, jež se spočetla ze špatně zapsaného údaje 15 z druhého měření (první měření bylo 118). Tři metody vypočetly pro data s touto chybnou hodnotou intervaly spolehlivosti, které uvádí druhý sloupec tabulky 6.8. Je patrná robustnost výsledků neparametrických metod. Odlehlá hodnota −103 má pochopitelně největší vliv na interval získaný pomocí t-testu.

Tab. 6.8 Příklad konstrukce intervalů spolehlivosti na základě neparametrických testů

Test	Interval spolehlivosti s opravenými daty	Interval spolehlivosti s odlehlou hodnotou
znaménkový test	(15; 40)	(15; 40)
Wilcoxonův test	(17,5; 35)	(17,5; 35)
***t*-test**	(16,87; 35,96)	(7,83; 36,83)

6.4.5 Mediánový test pro dva nezávislé výběry

Mediánový test pro dva výběry vychází ze skutečnosti, že pokud mají mít za platnosti nulové hypotézy oba výběry společný teoretický medián $\tilde{\mu}$, má se na jedné straně mediánu $\tilde{\mu}$ nacházet přibližně 50 % hodnot z každého výběru.

Nejdříve zjistíme společný výběrový medián $\tilde{x}$ ze všech dat. Pak odvodíme tabulku četností měření nad nebo pod mediánem v obou výběrech (tab. 6.9). Posuzujeme, zda relativní četnosti měření nad společným mediánem jsou pro oba výběry stejné. Test této hypotézy lze provést pomocí přesného testu nezávislosti v čtyřpolní tabulce podle Fishera (viz s. 314) nebo pomocí asymptoticky platného testu.

Tab. 6.9 Schéma provedení mediánového testu pro dva nezávislé výběry

	Výběr 1	Výběr 2	Součet
Počet hodnot pod $\tilde{x}$	a	b	$a+b$
Počet hodnot nad $\tilde{x}$	c	d	$c+d$
Součet	$a+c$	$b+d$	N

Pokud jsou všechny četnosti v políčkách větší než 5, je možné použít asymptoticky platný test pomocí aproximace testovací statistiky normálním rozdělením. Testovací z-statistika má tvar:

$$z = \frac{(ad - bc)\sqrt{n}}{\sqrt{(a+b)(b+d)(a+c)(c+d)}}$$

Hodnota $ad - bc$ je tím rozdílnější od nuly, čím více se liší relativní četnosti $c/(a + c)$ a $d/(b + d)$ popisující, kolik měření z první a druhé skupiny leží nad společným mediánem. (Podrobněji o schématu tohoto testu pojednáváme v kap. 8.3.1.) Jestliže testovací statistika z leží uvnitř intervalu $\pm z_{\alpha/2}$, nezamítáme H_0 ($z_{\alpha/2}$ je kritická hodnota standardizovaného normálního rozdělení). Jinak vyjádřeno – pokud absolutní hodnota testovací statistiky je větší než $z_{\alpha/2}$, hypotézu H_0 zamítáme. Tento test lze uplatnit, pokud n je větší než 25.

PŘÍKLAD 6.15

Mediánový test pro dva nezávislé výběry

Pro modelová data (s. 212) přezkoušíme hypotézu rovnosti mediánů skórů hodnotících stav jedinců v experimentální a kontrolní skupině jednostranným testem. Společný medián dat je 14. Vytvoříme čtyřpolní tabulku (tab. 6.10) a na jejím základě vypočítáme testovací statistiku

$$z = \frac{(144 - 42)\sqrt{37}}{\sqrt{18 \times 19 \times 19 \times 18}} = 1{,}814.$$

Tato hodnota je větší než kritická hodnota standardizovaného normálního rozdělení pro jednostranný test 1,65 (α = 5 %). Proto uzavíráme, že výběr indikuje neplatnost nulové hypotézy.

Tab. 6.10 Příklad provedení mediánového testu pro dva nezávislé výběry

	Kontrola	Intervence	Součet
Počet hodnot pod $\tilde{x}$	12	6	18
Počet hodnot nad $\tilde{x}$	7	12	19
Součet	19	18	37

6.4.6 Wilcoxonův test pro dva nezávislé výběry

Tento test vytěží z dat více informací než znaménkový test. Test vychází z pořadí údajů a ze skutečnosti, že větší naměřené hodnoty mají vyšší pořadí. Jestliže tedy v jedné skupině máme více větších pozorování, průměrná hodnota pořadí bude větší než ve druhé skupině. Náhodné chování pořadí za platnosti nulové hypotézy jsme zkoumali v modelovém příkladu v kapitole 5.4.2. Nulová hypotéza předpokládá, že rozdělení sledované proměnné v obou skupinách jsou totožná. Při výpočtu testovací statistiky postupujeme tak, že seřadíme všechny údaje společně podle velikosti a označíme, k jakému výběru patří. Určíme jejich pořadí. Pokud jsou údaje stejné, budou mít také stejná pořadí. Za pořadí pak ale bereme průměr z pořadí následných stejných měření. Spočítáme součet pořadí měření T_1 a T_2 pro obě skupiny 1 a 2. Platí, že $T_1 + T_2 = 0{,}5(n_1 + n_2)(n_1 + n_2 + 1)$. Tento vztah využíváme při kontrole výpočtu obou součtů. Předpokládáme, že menší součet má hodnotu T_1. Buď pro hodnotu T_1 zjistíme dosaženou p-hodnotu zcela přesně pomocí Fisherova principu permutování (viz s. 194), nebo najdeme kritickou hodnotu pro danou hladinu významnosti v tabulce 16 přílohy B. Pokud je hodnota T_1 menší než kritická mez, nulovou hypotézu můžeme zamítnout. Kritická hodnota závisí na velikosti obou výběrů. Výpočty pro přesné určení dosažené významnosti p musí být implementovány v programovém systému, kterým test provádíme, protože jsou numericky velmi náročné.

Pro test lze také použít asymptoticky platnou testovací z-statistiku, již vypočítáme tak, že veličinu T_1 standardizujeme její teoretickou střední hodnotou $E(T_1)$ a směrodatnou odchylkou σ_T za platnosti nulové hypotézy

$$z = \frac{T_1 - E(T_1)}{\sigma_T},$$

kde

$$E(T_1) = n_1(n_1 + n_2 + 1)/2 \quad \text{a} \quad \sigma_T = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}.$$

Jestliže testovací statistika z leží uvnitř intervalu $\pm z_{\alpha/2}$, nezamítáme H_0 ($z_{\alpha/2}$ je kritická hodnota standardizovaného normálního rozdělení). Jinak vyjádřeno – pokud je absolutní hodnota testovací statistiky větší než $z_{\alpha/2}$, hypotézu H_0 zamítáme. Tuto kritickou mez lze uplatnit, pokud je počet měření n v obou skupinách větší než 20. Při větším počtu shodných hodnot modifikujeme hodnotu jmenovatele testovací z-statistiky. Při hledání v tabulkách se častěji používají hodnoty $U_1 = T_1 - n_1(n_1 + 1)/2$, resp. $U_2 = T_2 - n_2(n_2 + 1)/2$. Jednodušší výpočet hodnot U vysvětlíme v dalším odstavci.

Při větším počtu shod upravíme hodnotu σ_T. Označme výraz pod odmocninou A. Místo A dosadíme výraz $A - B$, kde B se vypočítá pomocí vzorce

$$B = \frac{n_1 n_2 \sum_{i=1}^{r} (t_i^3 - t_i)}{12n(n-1)},$$

přičemž r znamená počet skupin se shodným měřením, t_i jsou postupně počty stejných měření a $n = n_1 + n_2$.

PŘÍKLAD 6.16

Wilcoxonův test pro dva nezávislé výběry

Pro modelová data (s. 212) přezkoušíme hypotézu rovnosti rozdělení testových skórů v kontrolní a experimentální skupině jednostranným testem. Hodnoty testových výsledků seřadíme podle velikosti bez ohledu na příslušnost ke skupině a přiřadíme jim pořadí (tab. 6.11). Dále spočítáme součet pořadí v experimentální skupině T_1 = 488,5 a dosadíme do vzorce pro z-statistiku

$$z = \frac{488{,}5 - 20 \times 41/2}{41{,}23} = 1{,}90.$$

Získaná hodnota je větší než kritická hodnota standardizovaného normálního rozdělení pro jednostranný test 1,65 (α = 5 %). Proto uzavíráme, že výběr indikuje neplatnost nulové hypotézy průměrné shody výkonu chlapců a dívek ve zkoumané populaci na hladině významnosti 5 %. Jestliže aplikujeme korekci na stejné hodnoty měření, dostaneme σ_T = 36,89 a z = 2,12.

Tab. 6.11 Příklad provedení Wilcoxonova testu pro dva nezávislé výběry

Skupina	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Měření	2	3	20	10	15	8	14	6	11	17	18	5	11	5	16	13	12	24	19	11
Pořadí	1	2,5	35	8,5	23	7	20	6	11	29	31	4,5	11	4,5	26	17	14	39	33	11
Skupina	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Měření	15	10	14	19	16	26	23	12	20	13	17	3	13	17	23	14	20	15	13	17
Pořadí	23	8,5	20	33	26	40	38	14	35	17	29	2,5	17	29	38	20	35	23	17	29

Krátce po Wilcoxonovi navrhli Mann a Whitney pro stejnou situaci vlastní test pomocí statistiky U, který vede k ekvivalentnímu výsledku, ale je možné ho jinak interpretovat. V jejich postupu se porovná každé měření z první skupiny s každým měřením ze druhé skupiny. Přitom se zjistí, kolikrát toto srovnání na

Tab. 6.12 Příklad skórování dat pro test podle Manna a Whitneye

Klinické příznaky	Laboratorní výsledky	Skór
žádné	negativní	1
zlepšení	negativní	2
zlepšení	pozitivní	3
zhoršení	pozitivní	4

Tab. 6.13 Příklad tabulky četností pro test podle Manna a Whitneye

	Výsledná kategorie				
	1	2	3	4	Celkem
Nový postup	6	3	1	0	10
Kontrola	1	0	0	9	10

ordinální stupnici dopadne ve prospěch druhé skupiny. Označme zkoumanou proměnnou v prvním výběru X a ve druhém výběru Y. Testem přezkušujeme nulovou hypotézu, že $P(X > Y) = 0{,}5$, proti alternativě, že $P(X > Y) \neq 0{,}5$.

Přístup popíšeme pro případ ordinálních dat kategoriálního typu. Moses (1984) na tomto příkladu ukázal, jak široké spektrum použití tento test má. Šlo o kontrolovaný znáhodněný experiment hodnotící novou terapii. Kontrolní skupinu tvořili jedinci ošetření placebem. Každý jedinec byl nakonec hodnocen skórovacím systémem, který je popsán tabulkou 6.12. Výsledek hodnocení obou skupin je uveden v tabulce četností jednotlivých hodnot skórů (tab. 6.13). Prohlídka dat vede k předběžnému závěru, že nový postup přináší lepší výsledek než placebo. Musíme se přesvědčit, zda tuto datovou konfiguraci nemůžeme přičíst náhodě.

Při zkoumání lze srovnat každého pacienta z experimentální skupiny (n_1 = 10) s každým pacientem z kontrolní skupiny (n_2 = 10). Tak dostaneme pro uvedené rozsahy celkem 100 srovnání. Pokud jsou intervence ekvivalentní, na každé straně by měl být přibližně stejný počet příznivých srovnání. Některá srovnání nemusí upřednostnit jeden z postupů. Jsou to ty případy z tabulky, které měly stejné hodnoty skórů. Pro skór 1 se jedná o 6 srovnání. Když takto vyhodnotíme celou tabulku, zjistíme, že 90 párů vedlo k upřednostnění výsledku nové intervence, 6 bylo rovnocenných a zbytek vyšel proti nové intervenci (4). Abychom spravedlivě rozsoudili nerozhodné případy, přiřadíme 3 z nich jako více zlepšené k intervenci a 3 jako více zlepšené k placebu. Dostaneme tak

odhad pravděpodobnosti 93/100, že jedinec náhodně vybraný z populace, z níž pochází experimentální skupina a kontrolní skupina, a ošetřený novou metodou bude mít lepší výsledek než při ošetření placebem. Získaný index (0,93) odpovídá na otázku: „Jaká je pravděpodobnost, že intervence bude mít pro pacienta lepší důsledek než ošetření placebem?"

Při rovnosti účinku intervence a placeba očekáváme průměrný počet srovnání ve prospěch intervence $50 = 10 \times 10/2$. Za platnosti nulové hypotézy má rozdíl (počet srovnání ve prospěch intervence zmenšený o padesát) směrodatnou odchylku danou jmenovatelem Wilcoxonovy z-statistiky. Testovací z-statistika podle Manna a Whitneye má tvar

$$z = \frac{U - n_1 n_2/2}{\sigma_T},$$

kde U je počet srovnání ve prospěch jedné ze skupin a n_i rozsahy skupin. Jestliže z našich dat (tab. 6.4, s. 212) vypočítáme tuto testovací statistiku, dostaneme

$$z = \frac{(93 - 50)}{\sqrt{\dfrac{10 \times 10 \times (10 + 10 + 1)}{12}}} = 3{,}25.$$

Z tabulky normálního rozdělení zjistíme, že výsledku odpovídá hladina $p = 0{,}0005$. Uzavíráme, že existuje dostatek evidence, že nová intervence působí příznivěji než placebo. Provedeme-li přesný výpočet hodnoty p pomocí Fisherova permutačního principu, dostaneme hodnotu $p = 0{,}00041$.

Alternativní přístup k hodnocení dat tohoto experimentu spočívá v tom, že spočítáme běžnou t-statistiku pro dva nezávislé výběry z naměřených skórů. Tato procedura má zjevný nedostatek spočívající v tom, že nejsou splněny předpoklady t-testu, protože data jsou ordinálního typu. Lze však ukázat, že v mnoha případech takto „nesprávně" použitého t-testu dostaneme velmi podobné výsledky jako při použití Wilcoxonova testu. Pro demonstraci aplikujeme t-testy na naše data pro skórovací systémy 1, 2, 3, 4 a 1, 2, 8 a 9 . Dostaneme hodnoty $p = 0{,}000014$ a opět $p = 0{,}000014$. Souhlas je tedy značný. To ilustruje skutečnost, že:

- změna skórovacího systému má malý efekt na t-statistiku;
- hodnocení pomocí t-testu je často kvalitativně stejné jako Wilcoxonovým testem.

Wilcoxonův test lze také aplikovat v situaci, kdy ordinální charakter nemá závisle proměnná, ale nezávisle proměnná, pokud závisle proměnná má binární charakter. Jestliže obě proměnné jsou ordinální kategoriálního typu a mají více kategorií, použijeme test založený na skórech nebo na koeficientu korelace podle Kendalla (kap. 8.4).

6.4.7 Kolmogorovův-Smirnovův test normality a Lillieforsův test

Oba tyto testy patří do kategorie testů „dobré shody". Zkoumáme pomocí nich průběh celé distribuční funkce. Test Kolmogorova-Smirnova je zcela obecný pro jakýkoli typ rozdělení. Lillieforsova modifikace se týká testu normálního rozdělení dat a používá se při ověřování tohoto předpokladu před aplikací dalších statistických postupů.

Oba testy hodnotí, zda data pocházejí z populace s určitou distribuční funkcí $F_0(x)$. Testuje se tedy hypotéza $H_0: F(x) = F_0(x)$ proti alternativě $H_1: F(x) \neq F_0(x)$ pro všechna x, kde $F(x)$ je distribuční funkce, která generuje naše data. Při provedení těchto testů srovnáváme přirozeným způsobem empirickou distribuční funkci $\hat{F}(x)$ s hodnotami teoretické distribuční funkce $F_0(x)$ pro všechna x. Počítáme testovací statistiku jako maximum jejich absolutního rozdílu $D = \max|\hat{F}(x) - F_0(x)|$, kterou porovnáme s kritickými hodnotami. Pro hladinu významnosti 0,05, resp. 0,01 má příslušná kritická mez asymptoticky platnou hodnotu $1{,}358/\sqrt{n}$, resp. $1{,}628/\sqrt{n}$.

Lilliefors upravil tuto hodnotu pro případ testu normality rozdělení dat, kdy nejdříve odhadujeme teoretické parametry μ a σ rozdělení pomocí $\bar{x}$ a s. Lilliefors doporučil použít asymptoticky platnou kritickou mez $0{,}89/\sqrt{n}$ pro hladinu významnosti 5 %, resp. mez $1{,}04/\sqrt{n}$ pro hladinu významnosti 1 %.

PŘÍKLAD 6.17

Ověření normality rozdělení dat

Zkoumáme, zda 9 měření sledované veličiny X {50; 68; 77; 81; 83; 84; 91; 92; 94} má normální rozdělení se střední hodnotou 85 a směrodatnou odchylkou 15, tedy $N(85; 225)$. Použijeme Kolmogorovův-Smirnovův test.

$$H_0: X \sim N(85; 225) \quad \text{proti alternativě} \quad H_1: X \sim \text{non}\, N(85; 225)$$

Měření jsme již seřadili podle velikosti, což je důležité pro sestrojení empirické distribuční funkce. Postup ukazuje tabulka 6.14. Sloupec s teoretickými hodnotami distribuční funkce $F_0(x)$ se spočte pomocí tabulky standardizovaného normálního rozdělení pro $(x - \mu)/\sigma = (x - 85)/15$. Například pro první řádek:

$$z = \frac{x - 85}{15} = \frac{50 - 85}{15} = -2{,}33$$

$$F_0(x) = P(z \leq -2{,}33) = 0{,}5 - 0{,}4901 = 0{,}0099$$

Asymptoticky platná kritická mez pro hladinu významnosti 0,05 je $1{,}358/3 = 0{,}453$. Raději však použijeme přesnou kritickou mez, protože pozorování je pouze 9. Z tabulky kritických hodnot pro test podle Kolmogorova-Smirnova (viz Hátle, Likeš, 1978, s. 330)

Tab. 6.14 Příklad postupu při provádění Kolmogorovova-Smirnovova testu normality

x_i	n_i	$Kumn_i$	$\hat{F}(x)$	z	$F_0(x)$	$\lvert D \rvert$
50	1	1	0,1111	−2,33	0,0099	0,1012
68	1	2	0,2222	−1,20	0,1151	0,1071
77	1	3	0,3333	−0,53	0,2981	0,0796
81	1	4	0,4444	−0,67	0,2514	0,0352
83	1	5	0,5556	−0,13	0,4483	0,1073
84	1	6	0,6667	−0,07	0,4721	0,1946
91	1	7	0,7778	0,40	0,6554	0,1224
92	1	8	0,8889	0,46	0,6772	0,2117
94	1	9	1,0000	0,60	0,7257	0,2743

nalezneme pro hladinu významnosti 5 % kritickou hodnotu 0,430. Protože největší hodnota $|D|$ není větší než tato hodnota, H_0 nemůžeme zamítnout.

Souhrn

Popsali jsme jednoduché situace analýzy dat a statistického usuzování, jež se často vyskytují v literatuře a vědecké praxi. Soustředili jsme se na hodnocení centrální tendence dat a jejich rozptýlenost. Důležité je rozpoznat, kdy je danou metodu zapotřebí použít. Zásadní krok ve statistické inferenci při hodnocení středních hodnot spočívá ve volbě testu nebo vhodné konstrukce intervalu spolehlivosti. Pro orientaci slouží obrázek 6.5, který nám pomůže nalézt správný postup. Naše rozhodování určuje povaha dat a uspořádání plánu výzkumu. Další rozhodnutí ovlivňuje měřítko posuzovaných znaků (ordinální-metrické) a normalita rozdělení. Před použitím metod statistického usuzování kriticky posuzujeme kvalitu dat. Data přezkoumáváme na přítomnost odlehlých hodnot a jiné nekonzistence pomocí grafických prostředků, jako je krabicový graf nebo histogram. Poznamenejme, že posouzení párových měření se provádí stejnou procedurou jako měření z jednoho výběru, protože se hodnotí rozdíly měření, jako by pocházely z jednoho výběru.

Přehled testů z této kapitoly a jejich různých variant podává Sachs (1974). Možnosti neparametrických metod podrobně rozebírá na praktických příkladech Sprent (2001).

Obr. 6.5 Rozhodovací schéma pro výběr testů středních hodnot