Matematická statististika

Doc. Ing. Jiří Likeš, CSc. Ing. Josef Machek, CSc.

Druhé vydání

Kniha je věnována základům matematické statistiky v rozsahu učiva na vysokých školách technických. Je určena hlavně pro posluchače vysokých škol technických. Přivítají ji také absolventi těchto škol a technici z praxe, kteří se chtějí s touto problematikou seznámit. *Autoři, 1983*

Dobrých knih není nikdy dost. Při hledání vhodných česky psaných učebnic základů matematické statistiky pro studenty technické univerzity jsme ani po 35 letech nenašli lépe napsanou knihu, kterou bychom mohli studentům doporučit, než je tato učebnice autorů Jiřího Likeše a Josefa Machka. Svojí nadčasovostí, důkladným a vlídným výkladem je stále aktuální i v době sofistikovaného statistického software a výkonných počítačů všech velikostí. Ba naopak, připomíná nám, že statistické programy jsou pouze nástrojem, který může dobře používat pouze dobrý řemeslník, který zná a ovládá teoretické základy. Domníváme se, že si tato kniha zaslouží reedici a pozornost studentů a učitelů i v době 4. průmyslové revoluce. *Reeditoři, 2019*

Redakce teoretické literatury –
hlavni redaktorka RNDr. Blanka Kulinová, CSC.
Lektorovali doc. RNDr. Jozef Nagy. CSc.. prof. RNDr. Jiří Anděl. DrSc.

Odpovědná redaktorka
RNDr. Jarmila Novotná. CSc.

Reedice –
prof. RNDr. Jaromír Antoch, CSc., prof. RNDr. Gejza Dohnal, CSc.
Ing. Eliška Cézová, Ph.D.

Obsah

Úvod

Hlavním úkolem matematické statistiky je rozbor dat vykazujících náhodná kolísání, ať už jde o data získaná pokusem pečlivě připraveným a provedeným pod stálou kontrolou experimentálních podmínek v laboratoři, či o data získaná sledováním výrobků určitého druhu v provozu. Takovými náhodnými fluktuacemi se vyznačují výsledky většiny experimentů jak laboratorních, tak provozních, uskutečňovaných při fyzikálním, chemickém i technickém výzkumu. Velmi často je povaha celého experimentu taková, že experimentální data jsou ve své podstatě realizacemi náhodných veličin.

Tak například při radiometrickém zjišťování tzv. objemové hmotnosti stavebních materiálů (nebo zemin) se určuje objemová hmotnost podle části záření emitovaného radioizotopem, která pronikne měřeným materiálem. Množství impulsů vyslaných zářičem během časového intervalu dané délky je – jak je dobře známo – náhodná veličina, takže i počty impulsů registrované detektorem po průchodu měřenou hmotou jsou náhodné.

Podobně doba života či doba do poruchy strojírenského výrobku je náhodná veličina; ve velké sérii výrobků téhož druhu vyrobených stejným postupem za stejných podmínek budou výrobky s různými hodnotami doby života. Doby do opotřebení nebo do poruchy změřené na omezeném počtu vzorků v provozu jsou tedy ve své podstatě pozorovanými hodnotami náhodné veličiny. Zkoušky životnosti se samozřejmě nekonají proto, aby se získaly údaje pro deset či dvacet zkoušených vzorků, nýbrž proto, aby se získaly informace o vlastnostech celé série výrobků nebo o daném druhu výrobku, o vhodnosti použitého technologického postupu ve srovnání s jiným apod. Při takovémto rozšířeni platnosti výsledků pokusu právě vzniká problém matematicko-statistického rozboru dat: Jakýkoliv ukazatel, např. průměr, vypočítaný z dané skupiny výsledků, je sám zatížen náhodnou chybou; kdyby se pokus opakoval s jinou skupinou vzorků, dostaly by se jiné výsledky tedy i jiná hodnota počítaného ukazatele. Cílem matematicko-statistického rozboru je využití počtu pravděpodobnosti k ohodnoceni přesnosti a spolehlivosti získaných výsledků, např. ke stanovení hranic, které chyba výsledků s vysokou pravděpodobností nepřekročí, k výpočtu rizika, že chyba bude větší než určitá přípustná mez, k výpočtu rizika, že rozhodnutí učiněné na základě výsledků podobného experimentu bude chybné, atd. Právě tak patří k úlohám matematické statistiky použiti počtu pravděpodobnosti ke stanovení rozsahu pokusu, tj. počtu pozorováni potřebného k tomu, aby shora zmíněná rizika

chyb byla udržena na přijatelné úrovni.

V následujících kapitolách jsou vyloženy základní postupy, kterých se užívá při matematicko-statistické analýze experimentálních dat k dosažení naznačených cílů. V předcházejících řádcích se nejednou vyskytla slova jako „pozoruji se náhodné veličiny“, „stanovení rizika chybných závěrů“ apod. Z toho plyne, že nepostradatelným nástrojem matematicko-statistického rozboru experimentálních dat je počet pravděpodobnosti a k pochopení podstaty statistických metod je nezbytná znalost základů počtu pravděpodobnosti přibližně v rozsahu práce „Počet pravděpodobnosti“ [24] V dalším textu se řady výsledků a pojmů z počtu pravděpodobnosti používá už bez podrobného vysvětlování.

Část I

Náhodný výběr

Kapitola 1

Statistický model experimentu a náhodný výběr

1.1 Úvod.

Zkušenost ukazuje, že výsledky experimentu či měření určitého druhu při mnohonásobném opakování pokusu vykazují zvláštní pravidelnost, která se projevuje v relativním zastoupení různých možných výsledků. Často lze tuto pravidelnost vyjádřit matematickou formulí, vyrovnat (nebo aproximovat) četnosti různých výsledků některým ze známých rozdělení pravděpodobnosti. Tuto skutečnost osvětlí nejlépe numerické příklady.

1.2 Příklady

1.2.1 Počty mikroorganismů v zorném poli mikroskopu.

Běžným typem experimentu v biochemické laboratoři je počítání mikroorganismů v jednotlivých políčkách čtvercové sítě, kterou je rozděleno zorné pole mikroskopu. Tabulka 1.1 obsahuje výsledky zjištěné na $n = 118$ takových polích (data podle [29]). V tabulce značí

x ..... počet mikroorganismů,

n_x ..... počet políček, ve kterých bylo nalezeno právě x mikroorganismů (tzv. absolutní četnost hodnoty x),

$f_x = \frac{n_x}{n}$ .. tzv. relativní četnost hodnoty x,

p_x ..... hodnotu výrazu $(2,96)^x \exp(-2,96)/x!$ pro $x = 0, 1, \ldots, 5$ a

$\sum_{t=6}^{\infty}(2,96)^t \exp(-2,96)/t!$ pro $x \geq 6$.

x	n_x	f_x	p_x
0	5	0,0424	0,0518
1	19	0,1610	0,1534
2	26	0,2203	0,2270
3	26	0,2203	0,2240
4	21	0,1780	0,1657
5	13	0,1102	0,0981
≥ 6	8	0,0678	0,0800
	118	1,0000	1,0000

Tab. 1.1: Výsledky počítání mikroorganismů na $n = 118$ políčkách

Je vidět, že pozorované relativní četnosti f_x souhlasí dobře s hodnotami p_x, ve kterých poznáváme (viz [24], odst. 15.1) pravděpodobnosti hodnot $x = 0.1\ldots$. v Poissonově rozdělení s parametrem $\lambda = 2,96$. V [29] je uvedeno ještě několik podobných tabulek pro jiné druhy mikroorganismů a ve všech případech je shoda pozorovaných relativních četností f_x s pravděpodobnostmi Poissonova rozdělení při vhodně zvoleném λ velmi dobrá. To napovídá, že počty mikroorganismů na jednotlivých polích čtvercové sítě se chovají jako náhodné veličiny s Poissonovým rozdělením; rozsáhlé experimenty s daty tohoto druhu takovou hypotézu podporují (nadto existuje pro ni i teoretické zdůvodnění).

1.2.2 Doba do poruchy navigačních přístrojů.

V tab. 1.2 jsou uvedeny výsledky sledování provozu $n = 125$ navigačních přístrojů. Zjišťována byla doba bezporuchového provozu. Možné výsledky v hodinách (nezáporná čísla) byly rozděleny do intervalů délky $h = 50$ hodin; $t_i = 50i$ je horní hranice i-tého intervalu, n_i značí počet přístrojů, které pracovaly bez poruchy po dobu delší než $50(i-1)$ hodin, ale ne delší než $50i$ hodin. Ve čtvrtém sloupci jsou uvedeny počty přístrojů, které pracovaly bez poruchy po dobu delší než t_i, ve sloupci pátém podíly R_i/n a v šestém hodnoty funkce

$$P(t_i) = \exp(-t_i/186,82).$$

Pořadové číslo intervalu i	Horní hranice intervalu t_i	Četnost intervalu n_i	Četnost výsledků větších t_i R_i	n_i/n	$P(t_i)$
1	50	32	93	0,744	0,765
2	100	25	68	0,544	0,586
3	150	16	52	0,416	0,448
4	200	6	46	0,368	0,343
5	250	10	36	0,288	0,262
6	300	8	28	0,224	0,201
7	350	7	21	0,168	0,154
8	400	7	14	0,112	0,118
9	450	3	11	0,088	0,090
10	500	2	9	0,072	0,069
11	1000	9	0	0	0,005

Tab. 1.2: Doby bezporuchového provozu $n = 125$ navigačních přístrojů.

Srovnáním posledních dvou sloupců zjistíme, že relativní četnosti přístrojů pracujících bez poruchy po dobu delší než $t_i = 50, 100, \ldots$ jsou vcelku dobře aproximovány hodnotami funkce $P(t_i)$. Ale $P(t_i)$ není nic jiného než pravděpodobnost, že náhodná veličina s exponenciálním rozdělením $\mathrm{Exp}(0, \delta)$ (viz [24], čl. 20) překročí t_i.To ukazuje, že na dobu bezporuchového chodu zařízení daného druhu lze pohlížet jako na náhodnou veličinu X s exponenciálním rozdělením s parametrem $\delta = 186,82$. Stejný úkaz byl pozorován mnohokrát při měření dob života různých výrobků, zvlášť u elektronických zařízení: Relativní četnosti dob života delších než t přiléhaly dobře k funkci typu $P(t) = \exp(-t/\delta)$ s různými hodnotami δ závislými na složitosti zařízení, provozních podmínkách atd. U jiných druhů výrobků zase lépe vyhovovala jiná funkce, např. $P(t) = \exp\big(-(t/\delta)^c\big)$.

1.3 Statistický model

Uvedené příklady naznačují, že výsledky určitých druhů experimentu se chovají jako náhodné veličiny se zcela určitými typy rozdělení pravděpodobnosti. Rozdělení pravděpodobnosti náhodné veličiny, kterou v pokusu pozorujeme, se nazývají často *statistický model* příslušného experimentu. Rozhodnutí,

z jakého modelu budeme při zpracování výsledků vycházet, tj. jaké rozdělení budeme u pozorovaných náhodných veličin uvažovat, je prvním krokem při rozboru pozorování. Někdy lze určit model na základě fyzikálních a podobných úvah o znalosti a podstatě pokusu. Častěji však je nutno opírat se o výsledky podobných pokusů nashromážděné z dřívějška.

1.4 Náhodný výběr

Opakujeme-li n-krát nezávisle na sobě pokus (měření, pozorování), jehož výsledek je náhodná veličina X s distribuční funkcí $F(x)$, pozorujeme vlastně náhodný vektor $\mathbf{X} = (X_1, \ldots, X_n)'$, jehož složky jsou vzájemně nezávislé náhodné veličiny s touž distribuční funkci $F(x)$. Pro takovýto vektor vzájemně nezávislých náhodných veličin se stejným $F(x)$ (čili vektor vzájemně nezávislých pozorování se stejným rozdělením) se vžilo pojmenování *náhodný výběr* z rozdělení $F(x)$ (ve starší literatuře se často setkáváme také s termínem „výběr ze základního souboru s rozdělením $F(x)$") Náhodný výběr tedy dostaneme opakováním měření (či pozorování) stejného druhu za stejných podmínek nezávisle na sobě, tj. tak, aby výsledek žádného měření neovlivňoval výsledek kteréhokoliv z ostatních. Počet opakování n se nazývá *rozsah* náhodného výběru.

1.5 Příklady

1.5.1

Skupina n opakování chemické analýzy určitého materiálu nezávisle na sobě, stejnou technikou a za nezměněných podmínek je náhodný výběr; nejčastěji budeme s takovou skupinou zacházet jako s náhodným výběrem z normálního rozdělení.

1.5.2

Pozorování hustoty provozu (vyjádřené např. počtem projíždějících vozidel za hodinu) na jednom místě ve stejnou dobu (např. v ranní dopravní špičce) konané m pracovních dnů stejného ročního období poskytne náhodný výběr rozsahu n; zpravidla lze předpokládat, že jde o náhodný výběr z Poissonova

rozdělení. Všimněme si zde všech vyjmenovaných podmínek: jedno roční období, stejná denní doba, pracovní dny; spojit např. pozorování z různých ročních období by znamenalo porušit podmínku stálosti podmínek a způsobilo by, že pozorování by neměla stejné rozdělení pravděpodobnosti, neboť charakter provozu bývá v zimě jiný než v létě. Měli bychom pak co dělat s několika náhodnými výběry z rozdělení s různými hodnotami parametru.

1.5.3

Pozorování stupně opotřebení n výrobků po předepsané době používáni bude tvořit náhodný výběr rozsahu n z rozděleni, jehož tvar bude závislý na druhu výrobku.

1.6 Náhodný výběr z vícerozměrného rozdělení

Náhodný výběr z vícerozměrného rozděleni. Při mnohých pokusech je výsledek jedné realizace dán dvojicí, trojicí, obecně vektorem p čísel. To znamená, že při každém opakování pozorujeme p-rozměmý náhodný vektor. Pak mluvíme o náhodném výběru z dvourozměrného, třírozměrného atd. rozdělení.

1.7 Příklady

1.7.1

Při chemické analýze může být výsledkem stanoveni obsahu p různých prvků ve vzorku. Analýza n vzorků téže látky pak dá náhodný výběr rozsahu n z p-rozměmého rozděleni (např. z p-rozměmého normálního rozdělení).

1.7.2

Při sledování provozu na určitém úseku silnice může být pozorována veličina X – hustota provozu vyjádřená počtem vozidel za hodinu – a Y – průměrná rychlost vozidel. Sledováním po dobu 11 dnů dostaneme náhodný výběr rozsahu n z dvourozměrného rozděleni. Uvědomme si, že dvojice $(X_1, Y_1), \ldots,$

(X_n, Y_n) jsou vzájemně nezávislé, avšak jednotlivé souřadnice uvnitř dvojice mohou mít mezi sebou změnou závislost; s rostoucí hustotou provozu bude klesat rychlost.

1.8 Nezávislé výběry z několika rozdělení

Často se setkáváme se složitějšími experimenty než n opakovaných měření, např. s experimenty určenými ke srovnání několika měřicích technik, několika druhů materiálu, několika technologií apod. Např. měření pevnosti n_1 vzorků materiálu A, n_2 vzorků materiálu B a n_3 vzorků materiálu C představují tři náhodné výběry ze tří rozdělení. Příslušná tři rozdělení mohou mít stejný tvar, např. všechna tři mohou být rozdělení gama, ale mohou se lišit ve svých parametrech. Úkolem statistického rozboru v takových případech je nejčastěji právě zjištění, zda výsledky opravňují např. k závěru, že materiál A má vyšší pevnost než B apod.

1.9 Poznámka

Jakmile je pokus hotov, stává se náhodný výběr vektorem daných čísel; místo náhodných veličin máme co dělat už jen s určitou realizací náhodného vektoru $\mathbf{X}$. Podstata statistického rozboru spočívá v tom, že při vyvozováni jakýchkoliv závěrů z této realizace máme stále na paměti příslušný model, počítáme s náhodným kolísáním pozorovaných veličin a na základě zvoleného modelu odhadujeme hranice a rizika chyb všech výpovědi s danou realizaci provedených.

Kapitola 2

Statistiky a jejich rozdělení

2.1 Statistiky

Z dat získaných pokusem nebo pozorováním, tj. z náhodných výběrů, se vypočítávají hodnoty různých ukazatelů; do zpráv o výsledcích se uvádí průměrná hodnota výsledků (aritmetický průměr napozorovaných hodnot), maximální a minimální napozorovaná hodnota apod. Výpočet těchto ukazatelů má několikerý význam. Předně pomocí podobných ukazatelů lze naměřené či napozorované hodnoty shrnout stručně a přehledně. Za druhé, vhodně vybrané ukazatele umožňují určité závěry o rozdělení pravděpodobnosti náhodných veličin, které se pozorují, umožňují úsudky o charakteristikách těchto rozdělení (viz [24], čl. 10). A některé z těchto charakteristik rozdělení bývají často konečným cílem pokusu.

Např. při n opakovaných stanoveních koncentrace nějaké látky v roztoku jde konec konců o stanovení koncentrace této látky v celé dávce; opakování má za účel jen zmenšení chyby výsledku a vyloučení případných hrubých omylů. V důsledku působeni náhodných chyb jsou jednotlivá stanovení náhodnými veličinami, n opakování $(X_1, \ldots, X_n)'$ představuje náhodný výběr z nějakého rozděleni pravděpodobnosti (nejčastěji normálního). Jestliže metoda stanoveni je taková, že střední hodnota veličin X_i je rovna skutečné hodnotě koncentrace (v takovém případě se říká, že metoda není zatížena systematickou chybou), pak vlastně jde o úlohu zjištění střední hodnoty určitého rozdělení pravděpodobnosti, tj. jedné ze základních charakteristik náhodné veličiny (rozdělení pravděpodobnosti). V části II uvidíme, že při X_i s normálním rozdělením je aritmetický průměr $\overline{X} = (X_1 + \cdots + X_n)/n$ veličin

X_i nejlepším řešením této úlohy.

V článku 10 práce [24] je zavedena řada charakteristik rozdělení pravděpodobnosti – charakteristiky polohy, variability, šikmosti aj. Všechny mají svoje protějšky v ukazatelích vypočítaných z náhodného výběru.

Každý statistický ukazatel je funkce veličin $X_1, \ldots, X_n$. Abychom ho odlišili od jemu odpovídající charakteristiky rozdělení pravděpodobnosti veličin X_i, říkáme mu *statistika* (je to vlastně výběrová charakteristika). Statistikou je každá funkce (měřitelná, obecně vícerozměrná) veličin $X_1, \ldots, X_n$, k jejímuž určení není třeba znalosti správných hodnot parametrů příslušného rozdělení. Tak např. součet všech pozorování, minimální pozorování, maximální pozorování, rozdíl mezi maximálním a minimálním pozorováním, součet čtverců odchylek od daného čísla jsou statistiky.

2.2 Nejčastěji používané statistiky

Uvedeme nejběžnější statistiky, které jsou takřka nevyhnutelnou součástí každého referátu o jakémkoliv provedeném pokusu či statistickém sledování a které mají své přímé protějšky v charakteristikách náhodných veličin:

2.2.1

Výběrový průměr

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{2.2.1}$$

tj. součet pozorování, dělený rozsahem výběru.

S rostoucím rozsahem výběru n konverguje $\overline{X}$ podle pravděpodobnosti ke střední hodnotě μ rozdělení, z kterého výběr pochází (viz [24], odst. 25.6), a je tedy empirickým protějškem střední hodnoty.

Má-li rozdělení, z něhož výběr pochází, střední hodnotu μ a konečný rozptyl σ^2, má statistika (2.2.1) střední hodnotu a rozptyl

$$E(\overline{X}) = \mu, \quad \mathrm{var}(\overline{X}) = \frac{\sigma^2}{n}, \tag{2.2.2}$$

neboť

$$E(\overline{X}) = \frac{1}{n} \sum_{i=1}^{n} E(X_i) = \frac{n\mu}{n}, \quad \mathrm{var}(\overline{X}) = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{var}(X_i) = \frac{n\sigma^2}{n^2}.$$

2.2.2

Výběrový rozptyl

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2}{n(n-1)}, \quad n \geq 2. \tag{2.2.3}$$

Tato statistika vyjadřuje míru variability experimentálních výsledků a je protějškem rozptylu $\mathrm{var}(X_i) = \sigma^2$, ke kterému konverguje podle pravděpodobnosti při rostoucím n.

Má-li rozdělení, z něhož výběr pochází, rozptyl σ^2, je střední hodnota statistiky (2.2.3) rovna

$$E(S^2) = \sigma^2, \tag{2.2.4}$$

neboť

$$\begin{aligned}(n-1)E(S^2) = E(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2) = \sum_{i=1}^{n} E(X_i^2) - nE(\overline{X})^2 = \\ = \sum_{i=1}^{n}\Big(\mathrm{var}(X_i) + E^2(X_i)\Big) - n\Big(\mathrm{var}(\overline{X}) + E^2(\overline{X})\Big) = \\ = n\sigma^2 + n\mu^2 - n\frac{\sigma^2}{n} - n\mu^2 = (n-1)\sigma^2.\end{aligned}$$

Kdybychom ve jmenovateli statistiky (2.2.3) uvažovali n místo $n-1$, nerovnala by se střední hodnota takovéto statistiky σ^2, ale $(n-1)\sigma^2/n$.

Rozptyl statistiky (2.2.3) je roven (viz např. [16], str. 153)

$$\mathrm{var}(S^2) = \frac{\mu_4}{n} - \frac{n-3}{n(n-1)}\sigma^4, \quad n \geq 3, \tag{2.2.5}$$

kde μ_4 je čtvrtý centrální moment rozdělení, z něhož výběr pochází.

2.2.3

Výběrová směrodatná odchylka S je druhá odmocnina z výběrového rozptylu

$$S = \left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\right)^{1/2}. \tag{2.2.6}$$

Statistika S je analogií směrodatné odchylky náhodné veličiny X_i. Protože

$$\mathrm{var}(S) = E(S^2) - E^2(S) = \sigma^2 - E^2(S) \geq 0,$$

platí pro výběr z libovolného rozdělení

$$E(S) \leq \sigma. \tag{2.2.7}$$

Mimoto se často používá statistik:

2.2.4

Výběrový r-tý obecný moment

$$M'_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r, \quad r = 1, 2, \ldots; \tag{2.2.8}$$

speciálně $M'_1 = \overline{X}$.

Jestliže existuje $\mu'_r = E(X_i^r)$, konverguje M'_r při rostoucím n podle pravděpodobnosti k μ'_r.

Střední hodnota a rozptyl statistiky (2.2.8) jsou rovny

$$E(M'_r) = \mu'_r, \quad \mathrm{var}(M'_r) = \frac{1}{n}\Big(\mu'_{2r} - (\mu'_r)^2\Big), \tag{2.2.9}$$

neboť

$$\mathrm{var}(M'_r) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{var}(X_i^r) = \frac{1}{n^2}\sum_{i=1}^{n}\Big(E(X_i^{2r} - (\mu'_r)^2\Big).$$

2.2.5

Výběrový r-tý centrální moment

$$M_r = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^r, \quad r = 2, 3, \ldots; \tag{2.2.10}$$

tj. aritmetický průměr r-tých mocnin odchylek pozorování ve výběru od výběrového průměru.

Speciálně, $M_2 = (n-1)S^2/n$, takže výběrový rozptyl S^2 se neshoduje s druhým výběrovým centrálním momentem M_2. Důvodem pro tuto nedůslednost v terminologii je právě vztah (2.2.4); druhý výběrový centrální moment má střední hodnotu $(n-1)\sigma^2/n$, takže soustavně podceňuje rozptyl σ^2 rozdělení pozorované veličiny.

2.2.6

Výběrový koeficient šikmosti a špičatosti

$$A_3 = \frac{M_3}{M_2^{3/2}}, \quad A_4 = \frac{M_4}{M_2^2} - 3. \tag{2.2.11}$$

V dalších kapitolách se setkáme ještě s jinými statistikami, které budou mít význam pro výběry ze speciálních rozdělení nebo pro řešení zvláštních úloh.

2.3 Příklad.

Při zkouškách únavy kovů bylo u osmi zkoušek při napětí 560 MPa dosaženo následujících hodnot počtu kmitů do lomu (v tisících; data viz [7]):

$$3\,322, \quad 14\,713, \quad 763, \quad 46\,296, \quad 2\,845, \quad 9\,411, \quad 1\,532, \quad 24\,023.$$

Vypočtěme hodnoty výběrového průměru, výběrového rozptylu a výběrové směrodatné odchylky pro tyto údaje a pro jejich logaritmy.

Pro původní hodnoty dostáváme

$$\begin{aligned} \overline{x} &= \frac{102\,905.10^3}{8} = 12\,863,125.10^3; \\ s_X^2 &= \frac{8.3\,047\,522\,337 - 102\,905^2}{56}.10^6 = \frac{13\,790\,739\,671}{56}.10^6 = \\ &= 246\,263\,208,410\,714.10^6; \\ s_X &= 15\,692,776.10^3. \end{aligned}$$

$\overline{x}$ i s_X jsou ve stejných rozměrech jako původní údaje (v našem případě počty kmitů do lomu).

Pro logaritmy $y_i = \ln x_i, \quad i = 1, \ldots, 8$, je

$$\begin{aligned} \overline{x} &= \frac{124,870\,970}{8} = 15,609; \\ s_X^2 &= \frac{8.1\,963,224\,527 - 124,870\,970^2}{56} = \frac{113,037\,068}{56} = 2,018\,519; \\ s_X &= 1,421. \end{aligned}$$

2.4 Rozdělení statistik.

Jak jsme již uvedli v odst. 2.1, každý výběrový ukazatel, tj. každá statistika, slouží při matematicko-statistickém rozboru dat k získání informace o hodnotě odpovídající charakteristiky rozdělení pravděpodobnosti pozorovaných náhodných veličin čili. jak se říká, k odhadu odpovídající charakteristiky rozdělení pravděpodobnosti pozorovaných náhodných veličin. Proto je důležité moci určit, s jakými pravděpodobnostmi se vyskytují různé odchylky statistik od jim odpovídajících charakteristik pozorovaných náhodných veličin (někdy se též říká od odpovídajících charakteristik tzv. základního souboru).

Můžeme např. vznést otázku, s jakou pravděpodobností se $\overline{X}$ odchýlí od $E(X_i) = \mu$ nejvýše s danou hodnotu Δ, s jakou pravděpodobností bude relativní odchylka statistiky S^2 od skutečné hodnoty $\sigma^2 = \mathrm{var}(X_i)$ menší než dané kladné číslo δ, tj. jaká je pravděpodobnost

$$P\left(\frac{|S^2 - \sigma^2|}{\sigma^2} < \delta\right).$$

Podobně se můžeme ptát, pro jakou hodnotu je Δ je

$$P\big(|\overline{X} - E(X_i)| < \Delta\big) = 0,99$$

nebo při kolika pozorováních (tj. při jakém rozsahu výběru n) je zaručeno např.

$$P(0,9\,\sigma^2 < S^2 < 1,1\,\sigma^2) \geq 0,95$$

apod.

Každá statistika je totiž funkci náhodného výběru, tj. funkci náhodných veličin $X_i, ..., X_n$. Nabývá tedy od jednoho náhodného výběru k druhému různých hodnot, je sama náhodnou veličinou, které přísluší určité rozdělení pravděpodobnosti, závis1é na rozsahu výběru a na rozdělení veličin $X_i, ..., X_n$. Toto rozdělení pravděpodobnosti, které je základem pro řešení naznačených otázek, se nazývá *rozdělení* příslušné *statistiky*. Mluvíme tedy např. o rozdělení výběrového průměru $\overline{X}$ o rozdělení výběrového rozptylu S^2, o rozdělení statistiky M_2 atd.

Odvození rozdělení různých statistik pro výběry z rozděleni vyskytujících se často v aplikacích je tedy jedním z prvních úkolů matematické statistiky. Užívá se při něm především matematických prostředků uvedených v [24], zejména teorie charakteristických funkcí (odst. 10.6) a transformací náhodných veličin (čl. 13).

2.5 Rozdělení statistik $\overline{X}$ a $\sum_{i=1}^{n} X_i$.

V některých jednoduchých příkladech lze stanovit rozděleni statistik přímo na základě výsledků uvedených v [24] To se týká především rozdělení statistik $\overline{X}$ a $\sum_{i=1}^{n} X_i = n\overline{X}$:.

2.5.1

V náhodném výběru z normálního rozděleni $\mathrm{N}(\mu, \sigma^2)$ má statistika $\overline{X}$ rozdělení $\mathrm{N}(\mu, \sigma^2/n)$ a statistika $\sum_{i=1}^{n} X_i$ má rozděleni $\mathrm{N}(n\mu, n\sigma^2)$ (viz [24], odst. 18.6).

V náhodném výběru z libovolného rozdělení, majícího střední hodnotu a konečný rozptyl σ^2, lze pro dostatečně velké rozsahy výběrů n aproximovat rozděleni statistik $\overline{X}$ a $\sum_{i=1}^{n} X_i$ uvedenými normálními rozděleními (viz [24], odst. 26.3).

2.5.2

V náhodném výběru z alternativního rozděleni $\mathrm{A}(\pi)$ má statistika $\sum_{i=1}^{n} X_i$ binomické rozdělení $\mathrm{Bi}(n, \pi)$ (viz [24], odst. 14.5).

Pro dostatečně velká n lze rozdělení statistiky $\sum_{i=1}^{n} X_i$ aproximovat rozdělením $\mathrm{N}\big(n\pi, n\pi(1-\pi)\big)$ (viz [24], odst. 26.7) a rozdělení statistiky $\overline{X}$ aproximovat rozdělením $\mathrm{N}\big(\pi, \pi(1-\pi)/n\big)$.

Připomeňme, že statistika $\sum_{i=1}^{n} X_i$ značí absolutní a $\overline{X}$ relativní četnost úspěchů (zdařilých pokusů) v n nezávislých opakováních téhož pokusu.

2.5.3

V náhodném výběru z Poissonova rozděleni $\mathrm{Po}(\lambda)$ má statistika $\sum_{i=1}^{n} X_i$ rozděleni $\mathrm{Po}(n\lambda)$ (viz [24], odst. 15.4).

Pro dostatečně velká n lze rozdělení statistiky $\sum_{i=1}^{n} X_i$ aproximovat rozdělením $\mathrm{N}(n\lambda, n\lambda)$ (viz [24], odst. 26. 7) a rozdělení statistiky $\overline{X}$ aproximovat rozdělením $\mathrm{N}(\lambda, \lambda/n)$

2.5.4

V náhodném výběru z rozděleni $\Gamma(m, \sigma)$ (tj. z rozděleni gama) má statistika $\sum_{i=1}^{n} X_i$ rozděleni $\Gamma(nm, \sigma)$ a statistika $\overline{X}$ má rozděleni $\Gamma(nm, \sigma/n)$ (viz

[24],odst. 22:31)

Pro dostatečně velká n lze aproximovat rozdělení statistiky $\sum_{i=1}^{n} X_i$ rozdělením $\mathrm{N}(nm,\, n\sigma^2)$ a rozdělení statistiky $\overline{X}$ rozdělením $\mathrm{N}(m,\, \sigma^2/n)$.

2.6 Příklady.

2.6.1

Nechť $\mathbf{X} = (X_1, \ldots, X_n)'$ je náhodný výběr z rozdělení $\mathrm{Po}(\lambda)$. Podle 2.5.3 má statistika $T = \sum_{i=1}^{n} X_i$ rozdělení $\mathrm{Po}(n\lambda)$. Odtud statistika $\overline{X} = T/n$ má rozdělení

$$P(\overline{X} = w) = P(T = nw) = \frac{(n\lambda)^{nw}}{(nw)!} \exp(-n\lambda), \quad w = 0, \frac{1}{n}, \frac{2}{n}, \ldots .$$

Rozdělení X		Rozdělení $\overline{X}$	
x	$P(X = x)$	w	$P(\overline{X} = w)$
0	0,367 9	0	0,006 7
1	0,367 9	0,2	0,033 7
2	0,183 9	0,4	0,084 2
3	0,061 3	0,6	0,140 4
4	0,015 3	0,8	0,175 5
5	0,003 1	1,0	0,175 5
6	0,000 5	1,2	0,146 2
7	0,000 1	1,4	0,104 4
		1,6	0,065 3
		1,8	0,036 3
		2,0	0,018 1
		2,2	0,008 2
		2,4	0,003 4
		2,6	0,001 3
		2,8	0,000 5
		3,0	0,000 2

Tab. 2.1: Rozdělení pravděpodobnosti veličiny X s Poissonovým rozdělením s parametrem $\lambda = 1$ a rozdělení statistiky $\overline{X}$ ve výběru rozsahu $n = 5$ z tohoto rozdělení.

Pro případ $\lambda = 1,\ n = 5$ je rozdělení pozorované veličiny X (tj. rozdělení Po(1)) a rozdělení statistiky $\overline{X}$ tabelováno v tab. 2.1 (potřebné hodnoty byly vypočteny z tabulek [23]). Z tabulky 2.1 je vidět, že zatímco pro veličinu X pravděpodobnost hodnot rovných 2 nebo větších je 0,1839 + 0,0613 + ... = 0,2642 a pravděpodobnost hodnoty 0 je rovna 0,3679, tj. pravděpodobnost náhodného jevu $|X-\lambda| \geq 1$ je rovna 0,3679 + 0,2642 = 0,6321, pravděpodobnost jevu $|\overline{X} - \lambda| \geq 1$ je jen 0,0067 + 0,0181 + 0,0082 + 0,0034 + ... = 0,0386. Podobně pravděpodobnost jevu $|\overline{X} - \lambda| < 0,5 = 0,742$, zatímco $P(|X - \lambda| < 0,5) = 0,3679$. Rozdělení statistiky $\overline{X}$ ukazuje, jaký je stupeň koncentrace statistiky $\overline{X}$ kolem skutečné hodnoty λ.

2.6.2

Nechť $\mathbf{X} = (X_1, \ldots, X_n)'$ je náhodný výběr z rozdělení s hustotou pravděpodobnosti

$$f(x) = \frac{x^2}{2\delta^2} \exp\left(-\frac{x}{\delta}\right), \quad x > 0.$$

To je hustota rozdělení gama (viz [24], odst. 22.1) s $m = 3$. Podle 2.5.4 má výběrový průměr $\overline{X}$ hustotu

$$g(x) = \frac{n^{3n} x^{3n-1}}{(3n-1)!\delta^{3n}} \exp\left(-\frac{nx}{\delta}\right), \quad x > 0.$$

Na obr. 2.1 je graficky znázorněna hustota pravděpodobnosti veličiny X (tj. rozdělení $\Gamma(3,\ \delta)$) a hustota pravděpodobnosti statistiky $\overline{X}$ pro případ $\delta = 2$ a $n = 5$.

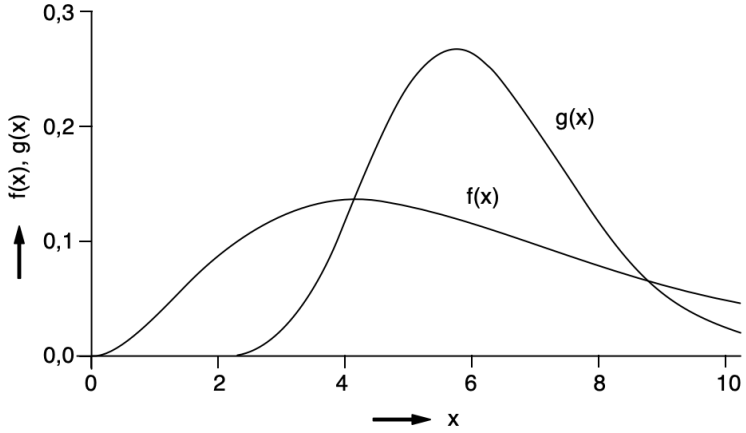
2.6.3

Předpokládejme, že pravděpodobnost proražení izolátoru určitého druhu je při daném napětí rovna π. Za účelem zjištění této pravděpodobnosti se provádí n nezávislých zkoušek. Zavedeme-li náhodné veličiny X_i, kde

$X_i = 1$, když při i-té zkoušce dojde k průrazu,
$= 0$, když při i-té zkoušce nedojde k průrazu,

pak $\mathbf{X} = (X_1, \ldots, X_n)'$ je za uvedených podmínek náhodný výběr z alternativního rozdělení $\mathrm{A}(\pi)$. Statistika

$$T = \sum_{i=1}^{n} X_i \quad \text{(celkový počet průrazů v } n \text{ nezávislých zkouškách)}$$



Obr. 2.1: Hustota pravděpodobnosti $f(x)$ rozdělení $\Gamma(3,2)$ a hustota pravděpodobnosti $g(x)$ statistiky $\overline{X}$ ve výběru rozsahu $n = 5$ z tohoto rozdělení.

má podle 2.5.2 rozdělení $\mathrm{Bi}(n,\pi)$ a statistika

$$\overline{X} = \frac{T}{n} \qquad \text{(relativní četnost průrazů)}$$

má rozdělení

$$P(\overline{X} = w) = P(T = nw) = \binom{n}{nw}\pi^{nw}(1-\pi)^{n-nw}, \qquad w = 0, \frac{1}{n}, \frac{2}{n}, \ldots, 1.$$

2.7 Úlohy

2.7.1

Nechť $X_1,\ X_2,\ X_3,\ X_4$ jsou nezávislá stanovení koncentrace určitého prvku v roztoku. Předpokládá se, že směrodatná odchylka výsledků analytické metody je $\sigma = 0,2\%$ výsledků. Navrhněte statistický popis (model) tohoto experimentu, najděte rozděleni pravděpodobnosti výsledku a vypočítejte pravděpodobnost, že chyba výsledku bude menší než 0,15%.

$\left[\begin{array}{l} \mathbf{X} = (X_1, \ldots, X_4)' \text{ je náhodný výběr rozsahu } n = 4 \text{ z rozdělení} \\ \mathrm{N}(\mu;\, 0,04), \text{ kde } \mu \text{ je správná hodnota. Konečný výsledek je} \\ \text{statistika } \overline{X} = \frac{1}{4}\sum_{i=1}^{4} X_i \text{mající rozdělení } \mathrm{N}(\mu;\, 0,01). \\ \text{Pravděpodobnost } P\big(|\overline{X} - \mu| < 0,15\big) = 0,8664. \end{array}\right]$

2.7.2

Nechť $X_1, \dots X_{20}$ jsou výsledky laboratorních zkoušek doby života výrobku, o kterém je známo, že jeho doba života má exponenciální rozdělení

$$f(x) = \frac{1}{\delta} \exp\left(-\frac{x}{\delta}\right), \qquad x > 0.$$

Účelem zkoušek je zjistit střední dobu života; za konečný výsledek se považuje aritmetický průměr naměřených hodnot. Uveďte statistickou formulaci úlohy a vypočtěte pravděpodobnost, že celá zkouška dá výsledek aspoň o 20% větší, než je skutečná střední hodnota doby života.

$$\left[\begin{array}{l} \mathbf{X} = (X_1, \dots, X_{20})' \text{ je náhodný výběr z rozdělení } \mathrm{E}(0;\, \delta), \\ \text{tj. z rozdělení } \Gamma(1;\, \delta). \text{ Statistika } \overline{X} = \frac{1}{20} \sum_{i=1}^{20} X_i \text{ má} \\ \text{podle 2.5.4 rozdělení } \Gamma(20;\, \delta/20). \\ \text{Odtud } P(\overline{X} > 1,2\delta) = \int_{1,2\delta}^{\infty} f(w)\,\mathrm{d}w = \int_{24}^{\infty} \mathrm{e}^{\cdot t} \frac{t^{19}}{19!}\,\mathrm{d}t = 0,18. \end{array}\right]$$

Kapitola 3

Rozdělení statistik při výběrech z normálního rozdělení

3.1 Úvod.

3.2 Rozdělení statistik $\overline{X}$ a S^2.

Mějme náhodný výběr $\mathbf{X} = (X_1, \ldots, X_n)',\ n \geq 2$ z rozdělení $\mathrm{N}(\mu, \sigma^2)$. Hledejme rozdělení statistik $\overline{X}$ a S^2. Za tím účelem uvažujme náhodné veličiny

$$Y_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i,$$

$$Y_j = \frac{1}{\sqrt{\big(j(j-1)\big)}} \left(\sum_{i=1}^{j-1} X_i - (j-1)X_j \right), \quad j = 2, \ldots, n., \tag{3.2.1}$$

Označíme-li $\mathbf{Y} = (Y_1, \ldots, Y_n)'$, je $\mathbf{Y} = \mathbf{AX}$, kde $\mathbf{A}$ je matice

$$\mathbf{A} = \begin{pmatrix} \frac{1}{\sqrt{n}}, & \frac{1}{\sqrt{n}}, & \frac{1}{\sqrt{n}}, & \cdots, & \frac{1}{\sqrt{n}}, & \frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{2}}, & \frac{-1}{\sqrt{2}}, & 0, & \cdots, & 0, & 0 \\ \frac{1}{\sqrt{6}}, & \frac{1}{\sqrt{6}}, & \frac{-2}{\sqrt{6}}, & \cdots, & 0, & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{1}{\sqrt{[n(n-1)]}}, & \frac{1}{\sqrt{[n(n-1)]}}, & \frac{1}{\sqrt{[n(n-1)]}}, & \cdots, & \frac{1}{\sqrt{[n(n-1)]}}, & \frac{-(n-1)}{\sqrt{[n(n-1)]}} \end{pmatrix} \tag{3.2.2}$$

Protože **A** je ortogonální matice, platí (viz [24], příkl. 24.8 a úloha 24.9.4) $\sum_{j=1}^{n} Y_j^2 = \sum_{i=1}^{n} X_i^2$ a veličiny (3.2.1) jsou vzájemně nezávislé, Y_1 má rozdělení $\mathrm{N}(\mu\sqrt{n},\, \sigma^2)$ a $Y_2, , Y_n$ mají rozdělení $\mathrm{N}(0,\, \sigma^2)$. Odtud vyplývá, že statistiky $\overline{X} = Y_1/\sqrt{n}$ a $\sum_{j=2}^{n} Y_j^2 = \sum_{j=1}^{n} Y_j^2 - n\overline{X}^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2 = (n-1)S^2$ jsou nezávislé. Tudíž i statistiky $\overline{X}$ a S^2 jsou nezávislé, $\overline{X}$ má rozdělení $\mathrm{N}(\mu,\, \sigma^2/n)$ a veličina $(n-1)S^2/\sigma^2$ má rozdělení $\chi^2(n-1)$, neboť

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{j=2}^{n} Y_j^2 = \sum_{j=2}^{n} U_j^2,$$

kde veličiny $U_2, \ldots, U_n$ jsou vzájemně nezávislé a každá má rozdělení $\mathrm{N}(0,\, 1)$. Tudíž veličiny $U_2^2, \ldots, U_n^2$ jsou vzájemně nezávislé, každá má rozdělení $\chi^2(1)$ a jejich součet má (viz [24], čl 22) rozdělení $\chi^2(n-1)$.

Střední hodnota

$$E\left(\frac{(n-1)S^2}{\sigma^2}\right) = 2^r\,\frac{\Gamma\left(\frac{n-1}{2}+r\right)}{\Gamma\left(\frac{n-1}{2}\right)}, \qquad r > -\frac{n-1}{2}. \tag{3.2.3}$$

Tento vztah vyplývá ze vztahu (22.2.3) práce [24], neboť rozdělení $\chi^2(n-1)$ je speciální případ rozdělení $\Gamma(m,\, \delta)$ pro $m = (n-1)/2$ a $\delta = 2$ (viz [24], odst. 22.5).

Z (3.2.3) pak dostáváme pro střední hodnotu a rozptyl statistiky (2.2.6)

$$E(S) = \left(\frac{2}{n-1}\right)^{\frac{1}{2}}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\sigma = \frac{1}{c_{n-1}}\sigma, \tag{3.2.4}$$

$$\mathrm{var}(S) = \left(1 - \frac{2}{n-1}\frac{\Gamma^2\left(\frac{n}{2}\right)}{\Gamma^2\left(\frac{n-1}{2}\right)}\right)\sigma^2 = \frac{c_{n-1}^2 - 1}{c_{n-1}^2}\sigma^2.$$

V [23] jsou tabelovány hodnoty c_{n-1} a c_{n-1}^2 pro $n - 1 = 1\,(1)\,100$.

3.3 Rozdělení t (Studentovo)

Uvažujme dvě nezávislé náhodné veličiny U a χ^2, přičemž U má rozdělení $N(0,1)$ a χ^2 má rozdělení $\chi^2(\nu)$. Hledejme rozdělení veličiny

$$T = \frac{U}{\sqrt{\chi^2/\nu}}. \tag{3.3.1}$$

Sdružená hustota pravděpodobnosti $f(u,x)$ veličin U a χ^2 je rovna

$$f(u,x)=\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{u^2}{2}}\frac{1}{2^{\nu/2}\Gamma(\nu/2)}x^{(\nu/2)-1}\mathrm{e}^{-\frac{x}{2}},\quad -\infty<u<\infty,\qquad x>0.$$

Použijeme-li postupu odst. 13.8 v [24], je sdružená hustota pravděpodobnosti veličin $T=U/\sqrt{\frac{\chi^2}{\nu}}$ a $Z=\chi^2$ dána výrazem

$$g(t,z)=\frac{1}{2^{(\nu+1)/2}\Gamma(\nu/2)\sqrt{\pi}}z^{(\nu/2)-1}\mathrm{e}^{-\frac{z}{2}\left(1+\frac{t^2}{\nu}\right)}\sqrt{\frac{z}{\nu}},\quad -\infty<t<\infty,\quad z>0.$$

Odtud hustota pravděpodobnosti

$$g_\nu(t)=\int_0^\infty g(t,z)dz$$

veličiny (3.3.1) je rovna

$$g_\nu(t)=\frac{1}{B\left(\frac{1}{2},\,\frac{\nu}{2}\right)\sqrt{\nu}}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},\quad -\infty<t<\infty \tag{3.3.2}$$

Rozdělení závisí na jediném parametru ν, který je roven počtu stupňů volnosti rozdělení veličiny χ^2 ve jmenovateli veličiny (3.3.1). Rozdělení s hustotou pravděpodobnosti (3.3.2) se nazývá *rozdělení t* (nebo *Studentovo rozdělení) o ν stupních volnosti.* Toto rozdělení označíme symbolem $\mathrm{t}(\nu)$.

Hustota pravděpodobnosti $g_\nu(t)$ je symetrická podle bodu $t=0$ a její tvar závisí na parametru ν (viz obr. 3.1). Jelikož

$$\lim_{\nu\to\infty}\frac{1}{B\left(\frac{1}{2},\,\frac{\nu}{2}\right)\sqrt{\nu}}=\frac{1}{\sqrt{2\pi}}\lim_{\nu\to\infty}\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\frac{\nu}{2}}}=\frac{1}{\sqrt{2\pi}},$$

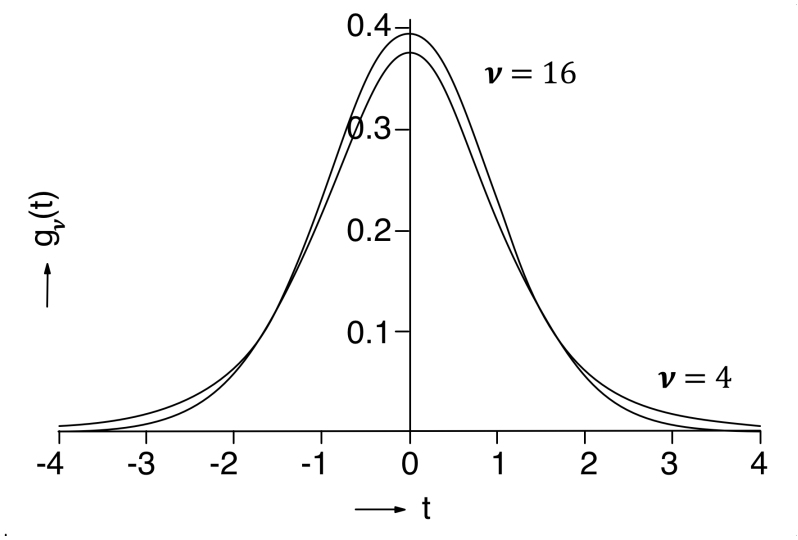
je

$$\lim_{\nu\to\infty}g_\nu(t)=\frac{1}{\sqrt{(2\pi)}}e^{-\frac{t^2}{2}},\qquad -\infty<t<\infty,$$

tzn. že pro rostoucí ν konverguje $g_\nu(t)$ k hustotě pravděpodobnosti rozdělení $\mathrm{N}(0,\,1)$.

Pro distribuční funkci $G_\nu(t)$ v důsledku symetrie $g_\nu(t)$ platí

$$G_\nu(t)=1-G_\nu(-t),\qquad -\infty<t<\infty. \tag{3.3.3}$$



Obr. 3.1: Hustota pravděpodobnosti rozdělení t(ν).

$100P\%$ kvantily $t_p(\nu)$ rozdělení t(ν), tj. hodnoty, pro něž platí

$$G_\nu(t_P(\nu)) = P, \quad 0 < P < 1, \quad \nu = 1, 2, \ldots, \tag{3.3.4}$$

jsou tabelovány, např. v [23] pro $\nu = 1\,(1)\,150\,(5)\,250\,(10)\,300\,(20)\,500\,(50)\,1000$ a $P = 0,9;\ 0,95;\ 0,975;\ 0,99;\ 0,995;\ 0,997\,5;\ 0,999;\ 0,999\,5$.

Vzhledem k symetrii rozdělení t(ν) podle bodu $t = 0$ platí

$$t_P(\nu) = -t_{1-P}(\nu), \quad 0 < P < 1, \quad \nu = 1, 2, \ldots. \tag{3.3.5}$$

Vraťme se nyní ke statistikám $\overline{X}$ a S^2 ve výběru z rozdělení N(ν, σ^2). Z odstavce 3.2 vyplývá, že veličina $U = (\overline{X} - \nu)\sqrt{n}/\sigma$ má rozdělení N(0, 1), veličina $(n-1)S^2/\sigma^2$ má rozdělení $\chi^2(n-1)$ a že tyto dvě veličiny jsou nezávislé. Tudíž veličina

$$T = \frac{(\overline{X} - \nu)\sqrt{n}}{\sigma\sqrt{\frac{(n-1)S^2}{\sigma^2(n-1)}}} = \frac{\overline{X} - \mu}{S}\sqrt{n} \tag{3.3.6}$$

má rozdělení t($n-1$).

3.4 Rozdělení F

Mějme dvě nezávislé náhodné veličiny χ_1^2 a χ_2^2, přičemž χ_1^2 má rozdělení $\chi^2(\nu_1)$ a χ_2^2 má rozdělení $\chi^2(\nu_2)$. Stanovme rozdělení veličiny

$$F = \frac{\chi_1^2/\nu_1}{\chi_2^2/\nu_2}. \tag{3.4.1}$$

Uvažujme nejprve náhodnou veličinu $Y = \chi_1^2/\chi_2^2$. Tato veličina má (viz [24], vztah (13.7.4)) hustotu pravděpodobnosti

$$g(y) = \int_0^\infty x f_{\nu_1}(xy) f_{\nu_2}(x)\,\mathrm{d}x = \frac{1}{B\left(\frac{\nu_1}{2},\,\frac{\nu_2}{2}\right)} y^{\frac{1}{2}\nu_1 - 1}(1+y)^{-\frac{\nu_1+\nu_2}{2}}, \qquad y > 0.$$

Veličina $F = \nu_2 Y/\nu_1$ má pak hustotu pravděpodobnosti

$$\begin{aligned} h_{\nu_1,\nu_2}(x) &= \frac{1}{B\left(\frac{\nu_1}{2},\,\frac{\nu_2}{2}\right)} \left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} x^{\frac{\nu_1}{2}-1} \left(1 + \frac{\nu_1}{\nu_2} x\right)^{-\frac{\nu_1+\nu_2}{2}}, \qquad x > 0, \\ &= 0, \qquad x \le 0. \end{aligned} \tag{3.4.2}$$

Rozdělení závisí na dvou parametrech ν_1 a ν_2 (počty stupňů volnosti) a nazývá se *rozdělení F o ν_1 a ν_2 stupních volnosti.* Toto rozdělení označíme $\mathrm{F}(\nu_1, \nu_2)$. Rozdělení je asymetrické (viz obr. 3.4) a jeho modus $\hat{x} = \nu_2(v_1 - 2)/\big(\nu_1(\nu_2+2)\big)$ pro $\nu_1 > 2$.

Uvažujme veličina (3.4.1). Pak veličinu F^{-1} lze vyjádřit ve tvaru

$$F^{-1} = \frac{\chi_2^2/\nu_2}{\chi_2^2/\nu_1}$$

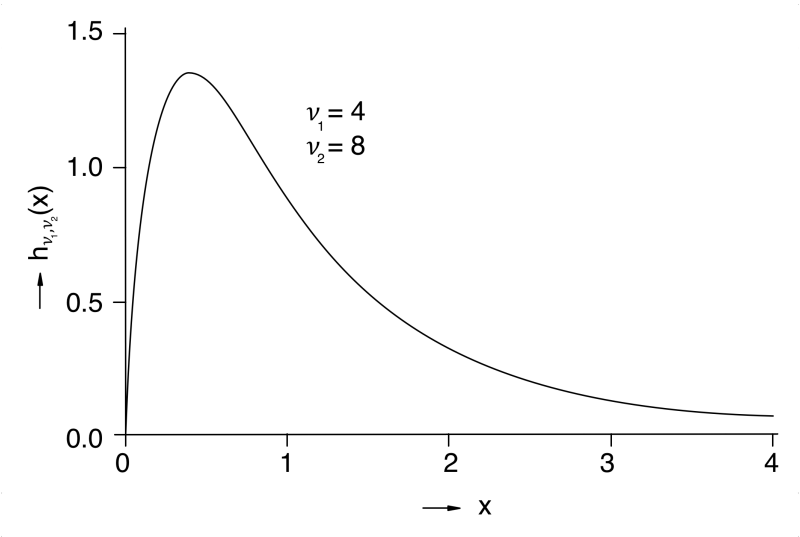
a tato veličina má rozdělení $\mathrm{F}(\nu_2, \nu_1)$. Platí tedy:

Má-li náhodná veličina F rozdělení $F(\nu_1, \nu_2)$, má veličina F^{-1} rozdělení $F(\nu_2, \nu_1)$, tj. opět rozdělení F, ale s opačným pořadím počtu stupňů volnosti.

$100P\%$ kvantily $F_P(\nu_1, \nu_2)$ rozdělení $\mathrm{F}(\nu_1, \nu_2)$, tj. hodnoty, pro které platí

$$P\big(F \le F_P(\nu_1,\nu_2)\big) = H_{\nu_1,\nu_2}\big(F_P(\nu_1,\nu_2)\big) = P, \qquad 0 < P < 1, \quad \nu_1, \nu_2 = 1, 2, \ldots, \tag{3.4.3}$$

jsou tabelovány, např. v [23] pro $\nu_1 = 1\,(1)\,30\,(5)\,50\,(10)\,100,\ 120,\ 150,\ 180,\ 200\,(100)\,1000,\ \infty,\ \ \nu_2 = 1\,(1)\,30\,(2)\,50\,(5)\,100\,(20)\,200\,(100)\,500,\ 1000,\ \infty$ a



Obr. 3.2: Hustota pravděpodobnosti rozdělení $F(\nu_1, \nu_2)$.

$P = 0,9;\ 0,95;\ 0,975;\ 0,99;\ 0,995;\ 0,997\,5;\ 0,999;\ 0,999\,5$. V (3.4.3) značí $H_{\nu_1,\nu_2}(x)$ distribuční funkci veličiny (3.4.1).

Pro každá $\nu_1, \nu_2 = 1, 2, \ldots$ platí

$$F_P(\nu_1, \nu_2) = \frac{1}{F_{1-P}(\nu_2, \nu_1)}, \qquad 0 < P < 1, \tag{3.4.4}$$

neboť

$$1 - P = P\big(F \geq F_P(\nu_1, \nu_2)\big) = \\ = P\big(F^{-1} \leq 1/F_P(\nu_1, \nu_2)\big) = P\big(F^{-1} \leq F_{1-P}(\nu_2, \nu_1)\big).$$

Poslední vztah vyplývá z toho, že veličina F^{-1} má rozdělení $\mathrm{F}(\nu_2, \nu_1)$.

3.5 Dva nezávislé výběry.

Mějme náhodný výběr $\mathbf{X} = (X_1, \ldots, X_{n_1})'$ z rozdělení $\mathrm{N}(\mu_1, \sigma_1^2)$ a náhodný výběr $\mathbf{Y} = (Y_1, \ldots, Y_{n_2})'$ z rozdělení $N(\mu_2, \sigma_2^2)$. Nechť výběry $\mathbf{X}$ a $\mathbf{Y}$ jsou

nezávislé. Označme

$$\begin{aligned} \overline{X} &= \frac{1}{n_1}\sum_{i=1}^{n_1} X_i, \quad S_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i-\overline{X})^2, \\ \overline{Y} &= \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i, \quad S_2^2 = \frac{1}{n_2-1}\sum_{i=1}^{n_2}(Y_i-\overline{Y})^2, \end{aligned} \tag{3.5.1}$$

výběrové průměry a výběrové rozptyly těchto dvou výběrů (S_1^2 má význam pro $n_1 \geq 2$ a S_2^2 pro $n \geq 2$).

Z odstavce 3.2 a z nezávislosti náhodných výběrů vyplývá, že statistiky $\overline{X}$, S_1^2, $\overline{Y}$ a S_2^2 jsou vzájemně nezávislé, statistika $\overline{X}$ má rozdělení $\mathrm{N}\big(\mu_1, \sigma_1^2/n_1\big)$, statistika $\overline{Y}$ má rozdělení $\mathrm{N}\big(\mu_2, \sigma_2^2/n_2\big)$, veličina $(n_1-1)S_1^2/\sigma_1^2$ má rozdělení $\chi^2(n_1-1)$ a veličina $(n_2-1)S_2^2/\sigma_2^2$ má rozdělení $\chi^2(n_2-1)$. V důsledku toho má veličina

$$U = \frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}} \tag{3.5.2}$$

rozdělení $\mathrm{N}(0,1)$.

Je-li $\sigma_1^2 = \sigma_2^2 = \sigma^2$, má veličina

$$\begin{aligned} \frac{(n_1-1)S_1^2}{\sigma^2} + \frac{(n_2-1)S_2^2}{\sigma^2} &= \frac{1}{\sigma^2}\Big[\sum_{i=1}^{n_1}(X_i-\overline{X})^2 + \sum_{i=1}^{n_2}(Y_i-\overline{Y})^2\Big] = \\ &= \frac{(n_1+n_2-2)S^2}{\sigma^2} \end{aligned} \tag{3.5.3}$$

rozdělení $\chi^2(n_1+n_2-2)$, neboť jsou-li χ_1^2 a χ_2^2 dvě nezávislé náhodné veličiny, přičemž veličina χ_1^2 má rozdělení $\chi^2(\nu_1)$, tj. rozdělení $\Gamma(\nu_1/2,\, 2)$, a veličina χ_2^2 má rozdělení $\chi^2(\nu_2)$, tj. rozdělení $\Gamma(\nu_2/2,\, 2)$, má jejich součet rozdělení $\Gamma\big((\nu_1+\nu_2)/2, 2\big)$, tj. rozdělení $\chi^2(\nu_1+\nu_2)$ - viz [24], odst. 22.3 a 22.5; v našem případě je $\nu_1 = n_1-1$ a $\nu_2 = n_2-1$.

Protože veličiny (3.5.2) a (3.5.3) jsou nezávislé, má v případě $\sigma_1^2 = \sigma_2^2 = \sigma^2$ veličina

$$T = \frac{U}{\sqrt{\frac{S^2}{\sigma^2}}} = \frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{S\sqrt{1/n_1+1/n_2}} \tag{3.5.4}$$

rozdělení $\mathrm{t}(n_1+n_2-2)$.

Dále veličina

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2}{S_2^2}\frac{\sigma_2^2}{\sigma_1^2} \tag{3.5.5}$$

má rozdělení $\mathrm{F}(n_1 - 1, n_2 - 1)$.

3.6 Příklady.

3.6.1

Stanovme rozdělení náhodné veličiny

$$Y = \frac{\nu_1 F}{\nu_2 + \nu_1 F}, \tag{3.6.1}$$

má-li veličina F rozdělení $\mathrm{F}(\nu_1, \nu_2)$.

Z (3.4.2) a ze vztahů (13.2.1) a (23.1.1) práce [24] vyplývá, že veličina (3.6.1) má rozdělení $\mathrm{Be}(\nu_1/2, \nu_2/2)$.

Lze tedy distribuční funkci $H_{\nu_1,\nu_2}(x)$ rozdělení $\mathrm{F}(\nu_1, \nu_2)$ vyjádřit pomocí neúplné funkce beta (viz [24], odst. 23.2)

$$H_{\nu_1,\nu_2}(x) = I_a\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right), \tag{3.6.2}$$

kde

$$a = \frac{\nu_1 x}{\nu_2 + \nu_1 x}. \tag{3.6.3}$$

3.6.2

Mějme náhodný výběr $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ z logaritmicko-normálního rozdělení $\mathrm{LN}(\mu, \sigma^2)$ (viz [24], odst. 19.1). Označíme-li

$$X_i = \ln Y_i, \qquad i = 1, \ldots, n,$$

představuje $\mathbf{X} = (X_1, \ldots, X_n)' = (\ln Y_1, \ldots, \ln Y_n)'$ náhodný výběr z rozdělení $\mathrm{N}(\mu, \sigma^2)$.

Tudíž statistiky

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} \ln Y_i \tag{3.6.4}$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\ln Y_i - \overline{X})^2 = \frac{1}{n(n-1)}\Big(n\sum_{i=1}^{n}(\ln Y_i)^2 - \big(\sum_{i=1}^{n}\ln Y_i\big)^2\Big)$$

jsou nezávislé, $\overline{X}$ má rozdělení $\mathrm{N}(\mu,\ \sigma^2/n)$ a veličina $(n-1)S^2/\sigma^2$ má rozdělení $\chi^2(n-1)$.

3.7 Statistiky ve výběrech z dvourozměrného normálního rozdělení.

V práci [24], odst. 24.5, jsme uvažovali dvourozměrné normální rozdělení, které závisí na pěti parametrech $\mu_1,\ \mu_2,\ \sigma_1^2,\ \sigma_2^2,\ \rho$.

Náhodným výběrem rozsahu n z tohoto rozdělení rozumíme n dvourozměrných náhodných veličiny $(X_i, Y_i)',\ i = 1, \ldots, n$; přičemž tyto dvourozměrné náhodné veličiny jsou vzájemně nezávislé a každá má totéž dvourozměrné normální rozdělení.

Označme

$$D_i = X_i - Y_i, \quad i = 1, \ldots, n. \tag{3.7.1}$$

Veličiny $D_1, \ldots, D_n$ jsou vzájemně nezávislé, všechny mají normální rozdělení $\mathrm{N}(\Delta, \sigma_D^2)$, kde

$$\Delta = E(D_i) = \mu_1 - \mu_2, \quad \sigma_D^2 = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2, \quad i = 1, \ldots, n. \tag{3.7.2}$$

Uvažujme statistiky

$$\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i = \overline{X} - \overline{Y}, \tag{3.7.3}$$

$$S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2 = \frac{1}{n(n-1)}\Big(n\sum_{i=1}^{n} D_i^2 - \big(\sum_{i=1}^{n} D_i\big)^2\Big), \quad n \geq 2.$$

Protože $\mathbf{D} = (D_1, \ldots, D_n)'$ tvoří náhodný výběr z rozdělení $\mathrm{N}(\Delta, \sigma_D^2)$, vyplývá z odst. 3.2, že statistiky $\overline{D}$ a S_D^2 jsou nezávislé, $\overline{D}$ má rozdělení $\mathrm{N}(\Delta, \sigma_D^2/n)$ a veličina $(n-1)S_D^2$ má rozdělení $\chi^2(n-1)$. Tudíž veličina

$$T = \frac{\overline{D} - \Delta}{S_D}\sqrt{n} \tag{3.7.4}$$

má rozdělení $\mathrm{t}(n-1)$.

Dále uvažujme statistiku

$$\begin{aligned} r &= \frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\Big(\sum_{i=1}^{n}(X_i-\overline{X})^2\Big)\Big(\sum_{i=1}^{n}(Y_i-\overline{Y})^2\Big)}} = \qquad (3.7.5)\\ &= \frac{n\sum_{i=1}^{n}X_iY_i-(\sum_{i=1}^{n}X_i)(\sum_{i=1}^{n}Y_i)}{\sqrt{\Big(n\sum_{i=1}^{n}X_i^2-(\sum_{i=1}^{n}X_i)^2\Big)\Big(n\sum_{i=1}^{n}Y_i^2-(\sum_{i=1}^{n}Y_i)^2\Big)}}, \quad n\geq 2. \end{aligned}$$

Tato statistika se nazývá *výběrový koeficient korelace.* Její hustota pravděpodobnosti závisí jen na parametru ρ a má tvar (viz [6], str. 398).

$$\begin{aligned} g(r) &= \frac{(1-\rho^2)^{\frac{n-1}{2}}}{\Gamma\big(\frac{n-1}{2}\big)\Gamma\big(\frac{n-2}{2}\big)\sqrt{\pi}}(1-r^2)^{\frac{n-4}{2}}. \qquad (3.7.6)\\ &\quad \Big(\Gamma^2\Big(\frac{n-1}{2}\Big)+\sum_{j=1}^{\infty}\frac{(2\rho r)^j}{j!}\Gamma^2\Big(\frac{n+j-1}{2}\Big)\Big), \quad -1<r<1,\\ &= 0, \qquad |r|\geq 1. \end{aligned}$$

V případě $\rho=0$ se $g(r)$ zjednoduší na tvar

$$\begin{aligned} g(r) &= \frac{1}{B\big(\frac{1}{2},\frac{n-2}{2}\big)}(1-r^2)^{\frac{(n-4)}{2}}, \quad -1<r<1, \qquad (3.7.7)\\ &= 0, \qquad |r|\geq 1. \end{aligned}$$

Použijeme-li vztahu (13.2.1) práce [24], zjistíme, že statistika

$$T=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}, \quad n\geq 3, \qquad (3.7.8)$$

má v případě $\rho=0$ hustotu pravděpodobnosti

$$h(t)=\frac{1}{B\big(\frac{1}{2},\frac{n-2}{2}\big)\sqrt{n-2}}\Big(1+\frac{t^2}{n-2}\Big)^{-\frac{n-1}{2}}, \quad -\infty<t<\infty,$$

tj. – vzhledem k (3.3.2) – má statistika (3.7.8) v případě $\rho=0$ rozdělení $t(n-2)$.

Pro případ obecného $\rho,\ -1<\rho<1$, se často používá transformace

$$Z=\frac{1}{2}\ln\frac{1+r}{1-r}, \qquad (3.7.9)$$

navržená R. A. Fisherem. Již pro malá n (řekněme $n \geq 10$, není-li $|\rho|$ příliš blízké 1) má veličina Z přibližně normální rozdělení se střední hodnotou

$$E(Z) = \frac{1}{2} \ln \frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)} = \zeta + \frac{\rho}{2(n-1)} \tag{3.7.10}$$

a rozptylem

$$\mathrm{var}(Z) = \frac{1}{n-3}, \tag{3.7.11}$$

takže náhodná veličina

$$U = (n-3)^{\frac{1}{2}} \Big(Z - \zeta - \frac{\rho}{2(n-1)} \Big) \tag{3.7.12}$$

má pro takováto n přibližně rozdělení $\mathrm{N}(0,1)$.

Uvažujme ještě dvourozměrné veličiny $(D_i, B_i)',\ i = 1, \ldots, n$, kde D_i jsou veličiny (3.7.1) a B_i jsou dány vztahy

$$B_i = X_i + Y_i, \qquad i = 1, \ldots, n. \tag{3.7.13}$$

Dvourozměrné veličiny $(D_1, B_1)', \ldots, (D_n, B_n)'$ jsou vzájemně nezávislé, každá má dvourozměrné normální rozdělení (viz [24], odst. 24.5) se středními hodnotami $E(D_i)$ a $E(B_i)$, rozptyly $\mathrm{var}(D_i)$ a $\mathrm{var}(B_i)$ a koeficientem korelace $\rho(D_i, B_i),\ i = 1, \ldots, n$.

Přitom $E(D_i)$ a $\mathrm{var}(D_i)$ jsou dány výrazy (3.7.2) a

$$E(B_i) = \mu_1 + \mu_2, \quad \mathrm{var}(B_i) = \sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2, \qquad i = 1, \ldots, n. \tag{3.7.14}$$

Dále kovariance

$$\mathrm{cov}(D_i, B_i) = E\big((X_i - Y_i)(X_i + Y_i)\big) - (\mu_1 - \mu_2)(\mu_1 + \mu_2) =$$

$$= E(X_i^2) - E(Y_i^2) - (\mu_1^2 - \mu_2^2) = \sigma_1^2 - \sigma_2^2, \qquad i = 1, \ldots, n,$$

takže

$$\rho(D_i, B_i) = \frac{\sigma_1^2 - \sigma_2^2}{\sqrt{(\sigma_1^2 + \sigma_2^2)^2 - 4\rho^2\sigma_1^2\sigma_2^2}}, \qquad i = 1, \ldots, n. \tag{3.7.15}$$

Je-li $\sigma_1^2 = \sigma_2^2 = \sigma^2$, je $\rho(D_i, B_i) = 0, i = 1, \ldots, n$, takže v tomto případě jsou (viz [24], odst. 24.3) veličiny D_i a B_i nezávislé, D_i má rozdělení $\mathrm{N}\big(\mu_1 - \mu_2, 2\sigma^2(1-\rho)\big)$ a B_i má rozdělení $\mathrm{N}\big(\mu_1 + \mu_2, 2\sigma^2(1+\rho)\big),\ i = 1, \ldots, n$.

3.8 Úlohy.

3.8.1

Stanovte střední hodnoty a rozptyly veličin (3.3.1) a (3.4.1).

$$\left[\begin{array}{l} E(T) = 0 \text{ pro } \nu \geq 2, \quad \mathrm{var}(T) = \frac{\nu}{\nu-2} \text{ pro } \nu \geq 3, \\ E(F) = \frac{\nu_2}{\nu_2-2} \text{ pro } \nu_2 \geq 3, \quad \mathrm{var}(F) = \frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)} \text{ pro } \nu_2 \geq 5. \end{array}\right]$$

3.8.2

Ukažte, že čtverec náhodné veličiny (3.3.1) má rozdělení $\mathrm{F}(1,\nu)$. Vyjádřete $100P\%$ kvantil rozdělení $\mathrm{F}(1,\nu)$ pomocí kvantilu rozdělení $\mathrm{t}(\nu)$.

$$\left[F_P(1,\nu) = t^2_{(1+P)/2}(\nu), \quad 0,5 \leq P < 1, \quad \nu = 1,2,\ldots.\right]$$

3.8.3

Čemu je roven medián rozdělení $t(\nu)$ a rozdělení $F(\nu,\nu)$?

$$\left[\begin{array}{l} t_{0,5}(\nu) = 0, \quad \nu = 1,2,\ldots; \\ F_{0,5}(\nu,\nu) = 1, \quad \nu = 1,2,\ldots \text{ (vyplývá z (3.4.4).} \end{array}\right]$$

3.8.4

Vyjádřete $100P\%$ kvantil statistiky (3.7.5) v případě $\rho = 0$ pomocí kvantilu rozdělení t.

$$\left[r_P = \frac{t_P}{(t_P^2+n-2)^{\frac{1}{2}}}, \quad t_P = t_P(n-2), \quad 0 < P < 1.\right]$$

Kapitola 4

Uspořádaný výběr

4.1 Pořádkové statistiky.

Mějme náhodný výběr $\mathbf{X} = (X_1, \ldots, X_n)'$ z daného rozdělení. Nechť $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ je týž náhodný výběr uspořádaný vzestupně podle velikosti. Pak $X_{(i)}$ se nazývá *i-tá pořádková statistika,* $i = 1, \ldots, n$. *Vektor* $\mathbf{X}^* = (X_{(1)}, \ldots, X_{(n)})'$ se nazývá *uspořádaný výběr* z daného rozdělení.

Statistiky $X_{(i)}$ hrají např. zásadní roli v testech životnosti, jestliže test se provádí tak, že se začne v témže okamžiku zkoušet životnost n výrobků, ale zkouška se skončí po dožití prvních r výrobků, $1 \leq r < n$. Výsledky těchto zkoušek jsou hodnoty statistik $X_{(1)}, \ldots, X_{(r)}$.

Jindy nás zajímá statistika $X_{(1)}$ (nejmenší pozorování), např. při studiu pevnosti nebo únavy materiálu, či statistika $X_{(n)}$ (největší pozorování), např. při sledování maximální hladiny nádrže či při analýze vlivu maximální tíhy sněhu na stavební konstrukce nebo při studiu hrubých chyb n opakovaných měření (zda $X_{(n)}$ či $X_{(1)}$ nejsou tzv. odlehlá pozorování).

Některé funkce statistik $X_{(i)}$, např. výběrový medián či výběrové rozpětí (viz dále), mohou sloužit jako charakteristiky polohy či variability náhodného výběru a pomocí nich usuzujeme na odpovídající charakteristiky rozdělení, z něhož náš výběr pochází.

Nejprve nalezneme rozdělení statistik $X_{(1)}$ a $X_{(n)}$ a poté rozdělení statistiky $X_{(i)}$ pro obecné i, $1 \leq i \leq n$. Pak budeme uvažovat vícerozměrné statistiky $(X_{(i_1)}, \ldots, X_{(i_k)})'$, $i \leq i_1 < i_2 < \ldots < i_k \leq n$, $2 \leq k \leq n$.

4.2 Rozdělení statistik $X_{(1)}$ a $X_{(n)}$.

Mějme náhodný výběr $\mathbf{X} = (X_1, \ldots, X_n)'$ z rozdělení, které má distribuční funkci $F(x)$.

Uvažujme statistiku $X_{(n)}$ a označme $F_{(n)}(x)$ distribuční funkci této statistiky. Zřejmě

$$F_{(n)}(x) = P(X_{(n)} \leq x) = P(X_i \leq x, i = 1, \ldots, n) = \prod_{i=1}^{n} P(X_i \leq x), \quad (4.2.1)$$

takže

$$F_{(n)}(x) = \big(F(x)\big)^n, \quad -\infty < x < \infty.$$

Je-li rozdělení, z něhož výběr pochází, rozdělení spojitého typu s distribuční funkcí $F(x)$ a hustotou pravděpodobnosti $f(x) = \mathrm{d}F(x)/\mathrm{d}x$, má statistika $X_{(n)}$ hustotu pravděpodobnosti

$$f_{(n)}(x) = \frac{\mathrm{d}F_{(n)}(x)}{\mathrm{d}x} = nf(x)\big(F(x)\big)^{n-1}, \quad -\infty < x < \infty. \quad (4.2.2)$$

Obdobně označme $F_{(1)}(x)$ distribuční funkci statistiky $X_{(1)}$. Jelikož

$$1 - F_{(1)}(x) = P(X_{(1)} > x) = P(X_i > x, i = 1, \ldots, n) = \prod_{i=1}^{n} P(X_i > x),$$

je

$$F_{(1)}(x) = 1 - \big(1 - F(x)\big)^n, \quad -\infty < x < \infty. \quad (4.2.3)$$

Pro výběr ze spojitého rozdělení má statistika $X_{(1)}$ hustotu

$$f_{(1)}(x) = nf(x)\big(1 - F(x)\big)^{n-1}, \quad -\infty < x < \infty. \quad (4.2.4)$$

4.3 Rozdělení statistiky $X_{(i)}$.

Dosud jsme uvažovali statistiky $X_{(i)}$ pro $i = 1$ a $i = n$. Pro libovolné $1 \leq i \leq n$ je distribuční funkce $F_{(i)}(x)$ statistiky $X_{(i)}$ v bodě x rovna pravděpodobnosti, že aspoň i z n náhodných veličin $X_1, \ldots, X_n$ nabude hodnoty menší nebo rovné x. Tato pravděpodobnost je však rovna součtu pravděpodobností, že právě t z n veličin $X_1, \ldots, X_n$ nabude hodnoty menší nebo rovné x a $n-t$ jich nabude hodnoty větší než x, pro $t = i,\ i+1, \ldots, n$.

Je tedy

$$F_{(i)}(x) = \sum_{t=i}^{n} \binom{n}{t} \big(F(x)\big)^t \big(1 - F(x)\big)^{n-t}, \quad -\infty < x < \infty. \tag{4.3.1}$$

Je ihned vidět, že (4.2.1) a (4.2.3) jsou speciální případy (4.3.1) pro $i = n$ a $i = 1$.

V případě výběru ze spojitého rozdělení má statistika $X_{(i)}$ hustotu

$$f_{(i)}(x) = \frac{\mathrm{d}F_{(i)}(x)}{\mathrm{d}x} = \sum_{t=i}^{n} n\binom{n-1}{t-1} \big(F(x)\big)^{t-1} f(x) \big(1 - F(x)\big)^{n-t} - \\ - \sum_{t=i}^{n-1} n\binom{n-1}{t} \big(F(x)\big)^{t} f(x) \big(1 - F(x)\big)^{n-t-1},$$

tj.

$$\begin{aligned} f_{(i)}(x) &= n\binom{n-1}{i-1} \big(F(x)\big)^{i-1} f(x) \big(1 - F(x)\big)^{n-i} = \qquad (4.3.2) \\ &= \frac{1}{B(i, n-i+1)} \big(F(x)\big)^{i-1} f(x) \big(1 - F(x)\big)^{n-i}, \quad -\infty < x < \infty. \end{aligned}$$

Pro výběry ze spojitých rozdělení lze tedy distribuční funkci $F_{(i)}(x)$ statistiky $X_{(i)}$ vyjádřit pomocí neúplné funkce beta (viz [24], odst. 23.2)

$$\begin{aligned} F_{(i)}(x) &= \int_{-\infty}^{x} f_{(i)}(x)\,\mathrm{d}x = \frac{1}{B(i, n-i+1)} \int_0^{F(x)} t^{i-1}(1-t)^{n-i}\,\mathrm{d}t = \\ &= I_{F(x)}(i, n-i+1), \quad 1 \le i \le n, \quad -\infty < x < \infty, \qquad (4.3.3) \end{aligned}$$

případně pomocí distribuční funkce rozdělení F (viz vztahy (3.6.2) a (3.6.3))

$$F_{(i)}(x) = H_{i-1,n-i+1}\left(\frac{(n-i+1)F(x)}{(i-1)\big(1-F(x)\big)}\right), \quad 1 \le i \le n, \quad -\infty < x < \infty. \tag{4.3.4}$$

Je-li spojité rozdělení symetrické podle bodu μ, tj. platí-li vztahy (viz [24], odst. 10.13)

$$f(x) = f(2\mu - x), \quad F(x) = 1 - F(2\mu - x), \quad -\infty < x < \infty, \tag{4.3.5}$$

pak platí

$$f_{(n-i+1)}(x) = f_{(i)}(2\mu - x), \qquad -\infty < x < \infty. \tag{4.3.6}$$

pro všechny $i = 1, \ldots, n$; o tom se ihned přesvědčíme, dosadíme-li do výrazu (4.3.2) pro $n - i + 1$ výrazy (4.3.5).

Pro výběry ze spojitého rozdělení jsou střední hodnota a rozptyl statistiky $X_{(i)}$ dány výrazy

$$E(X_{(i)} = \int_{-\infty}^{\infty} x f_{(i)}(x)\,\mathrm{d}x \tag{4.3.7}$$

a

$$\mathrm{var}(X_{(i)} = \int_{-\infty}^{\infty} \big(x - E(X_{(i)})\big)^2 f_{(i)}(x)\,\mathrm{d}x = \int_{-\infty}^{\infty} x^2 f_{(i)}(x)\,\mathrm{d}x - E^2(X_{(i)}). \tag{4.3.8}$$

Většinou střední hodnoty a rozptyly nelze jednoduše vyjádřit a musejí se tabelovat.

4.4 Příklady.

4.4.1

Stanovme hustotu pravděpodobnosti $f_{(i)}(x)$ pro případ výběru z exponenciálního rozdělení $\mathrm{E}(A,\, \delta)$. Dosazením do (4.3.2) dostáváme

$$\begin{aligned} f_{(i)}(x) &= \frac{n!}{(i-1)!(n-i)!}\frac{1}{\delta}\Big(1 - \mathrm{e}^{-\frac{x-A}{\delta}}\Big)^{i-1}\mathrm{e}^{-\frac{1}{\delta}(n-i+1)(x-1)}, \qquad x > A, \\ &= 0, \qquad\qquad x \le A. \end{aligned} \tag{4.4.1}$$

Pro $i = 1$ zjistíme, že statistika $X_{(1)}$ má rozdělení $\mathrm{E}(A,\, \delta/n)$.

4.4.2

Pro výběr ze spojitého rozdělení symetrického podle bodu μ vyplývá z (4.3.6)

$$\begin{aligned} E(X_{(n-i+1)} &= \int_{-\infty}^{\infty} x f_{(n-i+1)}(x)\,\mathrm{d}x = \int_{-\infty}^{\infty} x f_{(i)}(2\mu - x)\,\mathrm{d}x = \\ &= \int_{-\infty}^{\infty} (2\mu - y) f_{(i)}(y)\,\mathrm{d}y, \end{aligned}$$

takže

$$E(X_{(n-i+1)}) = 2\mu - E(X_{(i)}), \qquad i = 1, \ldots, n. \tag{4.4.2}$$

Odtud vyplývá, že

$$E\left(\frac{X_{(i)} + X_{(n-i+1)}}{2}\right) = \mu, \qquad i = 1, \ldots, n. \tag{4.4.3}$$

Připomeňme, že pro symetrické rozdělení jsou střední hodnota i medián tohoto rozdělení rovny μ.

Dále rozptyl

$$\begin{aligned} \mathrm{var}(X_{(n-i+1)} &= \int_{-\infty}^{\infty} \big(x - E(X_{(n-i+1)})\big)^2 f_{(n-i+1)}(x)\,\mathrm{d}x = \\ &= \int_{-\infty}^{\infty} \big(x - 2\mu + E(X_{(i)})\big)^2 f_{(i)}(2\mu - x)\,\mathrm{d}x = \\ &= \int_{-\infty}^{\infty} \big(y - E(X_{(i)})\big)^2 f_{(i)}(y)\,\mathrm{d}y, \end{aligned}$$

takže

$$\mathrm{var}(X_{(n-i+1)} = \mathrm{var}(X_{(i)}), \qquad i = 1, \ldots, n. \tag{4.4.4}$$

4.4.3

Výběrový medián je statistika

$$\begin{aligned} \tilde{X} &= X_{((n+1)/2)}, && n \text{ liché}, \\ &= \frac{1}{2}(X_{(n/2)} + X_{((n+2)/2)}, && n \text{ sudé}. \end{aligned} \tag{4.4.5}$$

Z (4.3.2) vyplývá, že pro n liché má $\tilde{X}$ hustotu pravděpodobnosti

$$f_{((n+1)/2)}(x) = \frac{n!}{\Big(\big(\frac{n+1}{2}\big)!\Big)^2}\big(F(x)\big)^{(n+1)/2} f(x)\big(1 - F(x)\big)^{(n+1)/2},$$

$$-\infty < x < \infty \tag{4.4.6}$$

Z příkladu 4.4.2 vyplývá, že pro symetrické rozdělení je

$$E(\tilde{X}) = \mu \tag{4.4.7}$$

pro n liché i pro n sudé.

4.4.4

Označme $x_{(n),P}100P\%$ kvantil statistiky $X_{(n)}$ (viz [24], odst. 10.8). V případě náhodného výběru ze spojitého rozdělení vyplývá z definice kvantilu [viz [24], vztah (10.8.2)] a z (4.2.1), že

$$P = F_{(n)}(x_{(n),P}) = \big(F(x_{(n),P})\big)^n,$$

takže

$$x_{(n),P} = X_Q, \quad Q = \sqrt[n]{P}, \quad 0 < P < 1, \tag{4.4.8}$$

tj. $100P\%$ kvantil statistiky $X_{(n)}$ je roven $100\sqrt[n]{P\%}$ kvantilu rozdělení, z něhož náhodný výběr pochází.

Obdobně, označíme-li $x_{(1),P}$ $100P\%$ kvantil statistiky $X_{(1)}$, vyplývá z rovnice (4.2.3), že

$$x_{(1),P} = x_Q, \quad Q = 1 - \sqrt[n]{(1-P)}, \quad 0 < P < 1. \tag{4.4.9}$$

Např. pro výběr rozsahu $n = 5$ z rozdělení $\mathrm{N}(10, 4)$ je 90% kvantil statistiky $X_{(5)}$ roven

$$x_{(5);0,9} = \mu + \sigma\, u_{0,979} = 10 + 2 \cdot 2,033\,52 \doteq 14,067,$$

neboť $\sqrt[5]{0,9} = 0,979$.

4.4.5

Označme $x_{(i),P}$ $100P\%$ kvantil statistiky $X_{(i)}, 1 \le i \le n$. V případě výběru ze spojitého rozdělení vyplývá z definice kvantilu a z (4.3.3), že

$$P = F_{(i)}(x_{(i),P}) = \frac{1}{B(i, n-i+1)} \int_0^{F(x_{(i),P})} t^{i-1}(1-t)^{n-i}\,\mathrm{d}t =$$

$$= \frac{1}{B(i, n-i+1)} \int_0^{w_p} t^{i-1}(1-t)^{n-i}\,\mathrm{d}t,$$

kde w_p značí $100P\%$ kvantil rozdělení $Be(i, n-i+1)$. Tudíž $x_{(i),P}$ se nalezne ze vztahu

$$F(x_{(i),P} = w_P, \quad 1 \le i \le n, \quad 0 < P < 1. \tag{4.4.10}$$

Například pro výběry z rozdělení $\mathrm{N}(\mu, \sigma^2)$ je

$$F(x_{(i),P} = \Phi\left(\frac{x_{(i),P} - \mu}{\sigma}\right) = w_P,$$

takže

$$x_{(i),P} = \mu + \sigma u_Q, \quad Q = w_P. \tag{4.4.11}$$

Z (3.4.3), (3.6.2) a (3.6.3) vyplývá, že

$$w_P = \frac{iF_P\big(2i, 2(n-i+1)\big)}{n-i+1+iF_P\big(2i, 2(n-i+1)\big)}, \quad 1 \le i \le n, \quad 0 < P < 1. \tag{4.4.12}$$

Označíme-li $w_P = w_P(i, n-i+1)$, vyplývá z (4.4.12) a (3.4.4), že

$$w_{1-P}(i, n-i+1) = 1 - w_P(n-i+1, i), \quad 1 \le i \le n, \quad 0 < P < 1. \tag{4.4.13}$$

Např. pro výběr rozsahu $n = 5$ z rozdělení $\mathrm{N}(10, 4)$ je 1% kvantil statistiky $X_{(2)}$ roven

$$x_{(2);0,01} = 10 + 2u_{0,033} = 10 - 2u_{0,967} = 10 - 2 \cdot 1,838\,424 \doteq 6,323,$$

neboť

$$Q = w_{0,01}(2,4) = 1 - w_{0,99}(4,2) = \\ = 1 - \frac{4\,F_{0,99}(8,4)}{2+4\,F_{0,99}(8,4)} = \frac{2}{2+4\cdot 14,799} = 0,033.$$

4.5 Rozdělení vícerozměrných pořádkových statistik.

Uvažujme nyní dvourozměrnou statistiku $(X_{(i_1)}, X_{(i_2)})'$, $1 \le i_1 < i_2 \le n$. *Sdružená distribuční funkce* $F_{(i_1,i_2)}(x,y) = P(X_{(i_1)} \le x,\ X_{(i_2)} \le y)$ statistik $X_{(i_1)}$ a $X_{(i_2)}$ je rovna pravděpodobnosti, že aspoň i_1 z n náhodných veličin $X_1, \ldots, X_n$ nabude hodnoty menší nebo rovné x a zároveň aspoň i_2 z těchto n náhodných veličin nabude hodnoty menší nebo rovné y. Pro $x < y$ je tato pravděpodobnost rovna součtu pravděpodobností, že právě t_1 veličin nabude hodnoty z intervalu $(-\infty, x\rangle$, právě t_2 veličin nabude hodnoty z intervalu $(x, y\rangle$ a právě $n - t_1 - t_2$ veličin nabude hodnoty z intervalu (y, ∞), pro $t_1 = i_1, \ldots, n$, $t_2 = T, \ldots, n - t_1$, kde $T = \max(0,\ i_2 - t_1)$.

Tudíž

$$F_{(i_1,i_2)}(x,y) =$$
$$= \sum_{t_1=i_1}^{n} \sum_{t_2=T}^{n-t_1} \frac{n!}{t_1!t_2!(n-t_1-t_2)!} \big(F(x)\big)^{t_1} \big(F(y)-F(x)\big)^{t_2} \big(1-F(y)\big)^{n-t_1-t_2},$$
$$-\infty < x < y < \infty. \quad (4.5.1)$$

Jednotlivé členy součtu (4.5.1) jsou pravděpodobnosti multinomického rozdělení (viz [24], odst. 17.1) pro $k = 3$.

Pro $x \geq y$ je

$$F_{(i_1,i_2)}(x,y) = P(X_{(i_2)} \leq y) = F_{(i_2)}(y). \quad (4.5.2)$$

V případě výběru ze spojitého rozdělení je sdružená hustota pravděpodobnosti

$$f_{(i_1,i_2)}(x,y) = \frac{\partial^2 F_{(i_1,i_2)}(x,y)}{\partial x \partial y} =$$
$$= \frac{n!}{(i_1-1)!(i_2-i_1-1)!(n-i_2)!} \big(F(x)\big)^{i_1-1} f(x) \big(F(y)-F(x)\big)^{i_2-i_1-1} \times$$
$$f(y)\big(1-F(y)\big)^{n-i_2}, \quad -\infty < x < y < \infty, \quad (4.5.3)$$
$$= 0, \qquad \text{jinak,}$$

pro všechna $1 \leq i_1 < i_2 \leq n$.

První z výrazů (4.5.3) dostaneme výpočtem parciálních derivací výrazu (4.5.1) a úpravou vzniklých součtů.

Je-li spojité rozdělení symetrické podle bodu μ, vyplývá z (4.5.3) a (4.3.5), že

$$f_{(n-i_2+1,n-i_1+1)}(x,y) = f_{(i_1,i_2)}(2\mu-y, 2\mu-x), \quad -\infty < x < y < \infty, \quad (4.5.4)$$

pro všechna $1 \leq i_1 < i_2 \leq n$.

Uvažujeme statistiku

$$Y = \sum_{i=1}^{n} a_i X_{(i)} + b, \quad (4.5.5)$$

kde $a_1, \ldots, a_n, b$ jsou známá reálná čísla. Pro stanovení střední hodnoty a rozptylu statistiky Y je zapotřebí znát střední hodnoty a rozptyly statistik $X_{(1)}, \ldots, X_{(n)}$ a jejich kovariance.

V případě výběru ze spojitého rozdělení je kovariance

$$\mathrm{cov}(X_{(i_1)}, X_{(i_2)}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \big(x - E(X_{(i_1)})\big)\big(y - E(X_{(i_2)}\big) f_{(i_1,i_2)}(x,y)\,\mathrm{d}x\mathrm{d}y =$$
$$= \int_{-\infty}^{\infty}\int_{-\infty}^{infty} xy f_{(i_1,i_2)}(x,y)\,\mathrm{d}x\mathrm{d}y - E(X_{(i_1)}E(X_{(i_2)}), \quad i_1, i_2 = 1,\ldots,n. \tag{4.5.6}$$

Je-li spojité rozdělení symetrické, pak z (4.4.2) a (4.5.4) vyplývá, že

$$\mathrm{cov}(X_{(n-i_2+1)}, X_{(n-i_1+1)}) = \mathrm{cov}(X_{(i_1)}, X_{(i_2)}), \quad 1 \le i_1 < i_2 \le n. \tag{4.5.7}$$

Statistiky $X_{(i)},\ 1 \le i \le n$ a $(X_{(i_1)}, X_{(i_2)})',\ 1 \le i_1 < i_2 \le n$, jsou speciální případy k-rozměrné statistiky $(X_{(i_1)},\ldots,X_{(i_k)})',\ 1 \le i_1 < \ldots < i_k \le n, 1 \le k \le n$. V případě výběru ze spojitého rozdělení hustota pravděpodobnosti (viz [8], str. 9)

$$f_{(i_1,\ldots,i_k)}(x_1,\ldots,x_k) =$$
$$= \frac{n!}{(i_1-1)!(i_2-i_1-1)!\ldots(i_k-i_{k-1}-1)!(n-i_k)!}\big(F(x_1)\big)^{i_1-1} f(x_1)\times$$
$$\big(F(x_2)-F(x_1)\big)^{i_2-1} f(x_2)\ldots\big(F(x_k)-F(x_{k-1})\big)^{i_k-i_{k-1}-1}\times$$
$$f(x_k)\big(1-F(x_k)\big)^{n-i_k}, \quad -\infty < x_1 < x_2 < \ldots < x_k < \infty, \tag{4.5.8}$$
$$= 0, \qquad \text{jinak.}$$

4.6 Rozdělení výběrového rozpětí.

Z (4.5.3) pro $i_1 = 1$ a $i_2 = n$ dostáváme hustotu pravděpodobnosti statistiky $(X_{(1),X_{(n)}})'$

$$\begin{aligned} f_{(1,n)}(x,y) &= n(n-1)f(x)\big(F(y)-F(x)\big)^{n-2} f(y), \quad -\infty < x < y < \infty, \\ &= 0, \qquad \text{jinak.} \end{aligned} \tag{4.6.1}$$

Statistika $R = X_{(n)-X_1}$, nazývaná *výběrové rozpětí*, má distribuční funkci

$$\begin{aligned} G(z) = P(X_{(n)} \le X_{(1)} + z) &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{x+z} f_{(1,n)}(x,y)\,\mathrm{d}y\right)\mathrm{d}x, \quad z > 0, \\ &= 0, \qquad z \le 0. \end{aligned}$$

Většinou distribuční funkci $G(z)$ nelze jednoduše vyjádřit a je nutné ji tabelovat. Hustota pravděpodobnosti $g(z)$ statistiky R je rovna

$$g(z) = \int_{-\infty}^{\infty} f_{(1,n)}(x, x+z)dx,$$

takže

$$\begin{aligned} g(z) &= n(n-1)\int_{-\infty}^{\infty} \big(F(x+z) - F(x)\big)^{n-2} f(x)f(x+z)\,\mathrm{d}x, \quad z > 0, \\ &= 0, \qquad\qquad z \le 0. \end{aligned} \tag{4.6.2}$$

Střední hodnotu a rozptyl statistiky R můžeme určit ze vztahů

$$E(R) = \int_0^{\infty} zg(z)\,\mathrm{d}z, \quad \mathrm{var}(R) = \int_0^{\infty} z^2 g(z)\,\mathrm{d}z - E^2(R),$$

kde $g(z)$ je výraz (4.6.2), nebo ze vztahů

$$E(R) = E(X_{(n)} - E(X_{(1)}), \tag{4.6.3}$$

$$\mathrm{var}(R) = \mathrm{var}(X_{(1)}) + \mathrm{var}(X_{(n)}) - 2\mathrm{cov}(X_{(1)}, X_{(n)}).$$

Pro symetrické rozdělení je

$$E(R) = 2\mu - 2E(X_{(1)}) = 2E(X_{(n)}) - 2\mu, \tag{4.6.4}$$

$$\mathrm{var}(R) = 2\mathrm{var}(X_{(1)}) - 2\mathrm{cov}(X_{(1)}, X_{(n)}) = 2\mathrm{var}(X_{(n)}) - 2\mathrm{cov}(X_{(1)}, X_{(n)}).$$

4.7 Příklady.

4.7.1

Mějme náhodný výběr $\mathbf{X} = (X_1, \ldots, X_n)'$ z rozdělení $\mathrm{N}(\mu, \sigma^2)$. Pak i-tou pořádkovou statistiku $X_{(i)}$ můžeme vyjádřit ve tvaru

$$X_{(i)} = \mu + \sigma U_{(i)}, \qquad i = 1, \ldots, n,$$

kde U_i je i-tá pořádková statistika ve výběru $\mathbf{U} = (U_1, \ldots, U_n)'$ z rozdělení $\mathrm{N}(0, 1)$.

Pro stanovení střední hodnoty a rozptylu veličiny

$$Y = \sum_{i=1}^{n} a_i X_{X_{(i)}} + b = \sigma \sum_{i=n} a_i U_{(i)} + \mu \sum_{i=1}^{n} a_i + b$$

je zapotřebí znát střední hodnoty a rozptyly veličin $U_{(1)}, \ldots, U_{(n)}$ a jejich kovariance. V [32] jsou tabelovány střední hodnoty $E(U_{(i)})$ a rozptyly $\mathrm{var}(U_{(i)})$ pro $n = 2\,(1)\,20,\ i = 1, 2, \ldots, n/2$ sudé a $i = 1, 2, \ldots, (n-1)/2$ pro n liché; pro ostatní i se použije vztahů

$$E(U_{(i)}) = -E(U_{(n-i+1)}), \quad \mathrm{var}(U_{(i)}) = \mathrm{var}(U_{(n-i+1)})$$

vyplývajících z (4.4.2) a (4.4.4). V [30] jsou tabelovány hodnoty $E(U_{(i)})$ pro $n = 2\,(1)\,50$.

Dále jsou v [32] tabelovány hodnoty $\mathrm{cov}(U_{(i_1)}, U_{(i_2)})$ pro $n = 2\,(1)\,20,\ i_1 = 1, 2, \ldots, n/2$ pro n sudé, $i_1 = 1, 2, \ldots, (n-1)/2$ pro n liché, a pro $i_2 = i_1 + 1, \ldots, n - i_1 + 1$. Pro ostatní i_1 a i_2 se použije vztahu (4.5.7).

Uvažujme např. výběrový medián

$$\tilde{X} = \frac{1}{2}(X_{(4)} + X_{(5)}) = \mu + \frac{1}{2}(U_{(4)} + U_{(5)})$$

ve výběru rozsahu $n = 8$ z rozdělení $\mathrm{N}(\mu, \sigma^2)$. Z tabulek v [32] zjistíme, že rozptyl

$$\mathrm{var}(\tilde{X}) = \frac{1}{4}\big(2\mathrm{var}(U_{(4)}) + 2\mathrm{cov}(U_{(4)}, U_{(5)})\big) = \frac{1}{2}(0,1872 + 0,1492) = 0,1682.$$

4.7.2

V případě výběru z rozdělení N(0, 1) má statistika $W = U_{(n)} - U_{(1)}$ hustotu pravděpodobnosti

$$\begin{aligned} g(w) &= \frac{n(n-1)}{2\pi} \int_{-\infty}^{\infty} \big(\Phi(u+w) - \Phi(u)\big)^{n-2} e^{-\frac{u^2+(u+w)^2}{2}}\,\mathrm{d}u, \qquad w > 0, \\ &= 0, \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad w \le 0, \end{aligned} \tag{4.7.1}$$

kde $\Phi(u)$ je distribuční funkce rozdělení N(0, 1) (viz [24], odst. 18.2).

Práce [18] obsahuje hodnoty distribuční funkce $P(W \le w)$ pro $n = 2\,(1)\,20$ a $w = 0,05\,(0,05)\,7,25$. V [23] jsou tabelovány hodnoty $E(W)$ a

$\mathrm{var}(W)$ pro $n = 2\,(1)\,20$ a $100P\%$ kvantily w_P statistiky W pro $n = 2\,(1)\,20$ a hodnoty P v rozmezí $0,000\,1$ až $0,999\,9$.

Je-li $R = X_{(n)} - X_{(1)}$ výběrové rozpětí ve výběru z rozdělení $\mathrm{N}(\mu,\, \sigma^2)$, je $R = \sigma_W$, kde W má hustotu pravděpodobnosti (4.7.1). Pak

$$E(R) = \sigma E(W), \quad \mathrm{var}(R) = \sigma^2 \mathrm{var}(W), \quad R_P = \sigma w_P,$$

kde R_P značí $100P\%$ kvantil statistiky R.

Tak např. pro výběr rozsahu $n = 8$ z rozdělení $N(\mu, 4)$ s libovolným μ je

$$E(R) = 2 \cdot 2,847\,201 = 5,694\,402, \quad \mathrm{var}(R) = 4 \cdot 0,672\,124 = 2,688\,496$$

a

$$R_{0,95} = 2 \cdot 4,286\,309 = 8,572\,618.$$

4.7.3

Mějme náhodný výběr $\mathbf{X} = (X_1, \ldots, X_n)'$ z rozdělení $\mathrm{E}(A,\, \delta)$ (viz [24], odst. 20.1). V příkladě 4.4 jsme nalezli hustotu pravděpodobnosti statistiky $X_{(i)},\ 1 \le i \le n$. Stanovme nyní hustotu pravděpodobnosti statistiky

$$Z_i = X_{(i)} - X_{(i-1)}, \quad 2 \le i \le n.$$

Vzhledem k (4.5.3) je

$$\begin{aligned} f_{(i-1,i)}(x, y) = & \\ &= \frac{n!}{(i-2)!(n-i)!} \frac{1}{\delta^2} \left(1 - \mathrm{e}^{-\frac{x-A}{\delta}}\right)^{i-2} \mathrm{e}^{-\frac{x-A}{\delta}} \mathrm{e}^{-(n-i+)\frac{y-A}{\delta}}, \\ & \qquad\qquad A < x < y < \infty, \\ &= 0, \qquad\qquad \text{jinak.} \end{aligned}$$

Hustota pravděpodobnosti $h_i(z)$ statistiky Z_i je pak rovna

$$h_i(z) = \int_A^\infty f_{(i-1,i)}(x, x+z)\, \mathrm{d}x, z > 0.$$

Po dosazení a substituci $t = e^{-\frac{x-A}{\delta}}$ dostaneme

$$\begin{aligned} h_i(z) &= \frac{n-i+1}{\delta} \mathrm{e}^{-(n-i+1)\frac{z}{\delta}}, \quad z > 0, \\ &= 0, \qquad\qquad\qquad\quad z \le 0, \end{aligned} \tag{4.7.2}$$

pro $i = 2, \ldots, n$; položíme-li $X_{(0)} = A$, platí (4.7.2) i pro $i = 1$ (viz příkl. 4.4.1).

Uvažujme ještě náhodné veličiny

$$Y_i = \frac{2}{\delta}(n - i + 1)(X_{(i)} - X_{(i-1)}), \quad i = 1, \ldots, n. \tag{4.7.3}$$

Z (4.7.2) vyplývá, že pro každé $i = 1, \ldots, n$ má Y_i hustotu

$$\begin{aligned} g_i(y) &= \frac{1}{2}\mathrm{e}^{-\frac{y}{2}}, \quad y > 0, \\ &= 0, \quad y \le 0; \end{aligned}$$

to je však hustota pravděpodobnosti rozdělení $\chi^2(2)$.

4.7.4

Ukažme, že náhodné veličiny (4.7.3) jsou vzájemně nezávislé. Z (4.5.8) vyplývá, že sdružená hustota pravděpodobnosti uspořádaného výběru $\mathbf{X}^* = (X_{(1)}, \ldots, X_{(n)})'$ je rovna

$$\begin{aligned} f_{(1,\ldots,n)}(x_1, \ldots, x_n) &= n! f(x_1) f(x_2) \ldots f(x_n), \quad -\infty < x_1 < \ldots < x_n < \infty, \\ &= 0, \qquad \text{jinak.} \end{aligned} \tag{4.7.4}$$

V případě výběru z rozdělení $\mathrm{E}(A, \delta)$ je

$$\begin{aligned} f_{1,\ldots,n)}(x_1, \ldots, x_n) &= \frac{n!}{n} e^{-\frac{1}{\delta}\sum_{i=1}^{n}(x_i - A)}, \quad -\infty < x_1 < \ldots < x_n < \infty, \\ &= 0, \qquad \text{jinak.} \end{aligned}$$

Protože platí

$$\begin{aligned} &\frac{1}{\delta}\sum_{i=1}^{n}(x_i - A) = \\ &= \frac{1}{\delta}\Big(n(x_1 - A) + (n-1)(x_2 - x_1) + \ldots + 2(x_{n-1} - x_{n-2}) + (x_n - x_{n-1})\Big) = \\ &= \frac{1}{2}(y_1 + y_2 + \ldots + y_n) \end{aligned}$$

a protože jakobián

$$J = \begin{vmatrix} \frac{\delta}{2n}, & -\frac{\delta}{2(n-1)}, & 0, & \dots, & 0 \\ 0, & \frac{\delta}{2(n-1)}, & -\frac{\delta}{2(n-2)}, & \dots, & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0, & 0, & 0, & \dots, & \frac{\delta}{2\cdot 1} \end{vmatrix} = \frac{\delta^2}{2^n n!},$$

má vektor $\mathbf{Y} = (Y_1, \dots, Y_n)'$ sdruženou hustotu pravděpodobnosti (viz [24], vztah (13.8.4))

$$\begin{aligned} g(y_1, \dots, y_n) &= \frac{1}{2^n} \exp\Big(-\frac{1}{2} \sum_{i=1}^{n} y_i \Big) = \prod_{i=1}^{n} g_i(y), \quad y_i > 0, \quad i = 1, \dots, n, \\ &= 0, \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{jinak.} \end{aligned}$$

tj. veličiny (4.7.3) jsou vzájemně nezávislé.

Odtud např. vyplývá, že statistiky $X_{(i_1)}$ a $X_{(i_2)} - X_{(i_1)}$ jsou nezávislé pro všechna $1 \le i_1 < i_2 \le n$.

4.8 Úlohy.

4.8.1

Uvažujte náhodný výběr $\mathbf{X} = (X_1, \dots, X_n)'$ z rozdělení, které má hustotu pravděpodobnosti $f(x) = 1,\ 0 < x < 1,\ f(x) = 0$, jinak. Stanovte hustoty pravděpodobnosti, střední hodnoty a rozptyly statistik $X_{(i)},\ 1 \le i \le n$.

$$\left[\begin{array}{l} f_{(i)}(x) = \frac{1}{B(i, n-i+1)} x^{i-1}(1-x)^{n-i}, \quad 0 < x < 1, \\ E(X_{(i)} = \frac{i}{n+1}, \quad \mathrm{var}(X_{(i)}) = \frac{i(n-i+1)}{(n+1)^2(n+2)}, \quad 1 \le i \le n. \end{array} \right]$$

4.8.2

Uvažujte $100P\%$ kvantil statistiky $X_{(n)}$ ve výběru rozsahu n z rozdělení $\mathrm{N}(\mu, \sigma^2)$ a $100(1-P)\%$ kvantil statistiky $X_{(1)}$ v témže výběru. Jaký vztah platí mezi těmito dvěma kvantily?

$$\Big[x_{(n),P} = 2\mu - x_{(1),1-P}, \quad 0 < P < 1. \Big]$$

4.8.3

Pomocí tabulek středních hodnot výběrového rozpětí W z rozdělení N(0; 1) (viz [23], tab 15 A) stanovte střední hodnotu $E(X_{(1)})$ pro výběr rozsahu 10 z rozdělení N(30; 4) a střední hodnotu $E(X_{(12)}$ pro výběr rozsahu 12 z rozdělení N$(-10; 2,25)$.

$$\Big[26,922;\ -7,556.\Big]$$

4.8.4

Uvažujte uspořádaný výběr $\mathbf{X}^* = (X_{(1)}, \ldots, X_{(n)})'$ z rozdělení E(A, δ). Jaké rozdělení má statistika

$$T = \frac{2}{\delta}\Big(\sum_{i=p+1}^{q} (X_{(i)} - X_{(p)} + (n-q)(X_{(q)} - X_{(p)})\Big), \quad 0 \le p < q \le n?$$

Čemu jsou rovny střední hodnoty a rozptyl veličiny $V = \delta T/2$? (Využijte veličin (4.7.3)!)

$$\Big[\chi^2(2q-2p), \quad E(V) = (q-p)\delta, \quad \mathrm{var}(V) = (q-p)\delta^2.\Big]$$

4.8.5

Stanovte střední hodnotu statistiky $U_{(2)}$ ve výběru rozsahu 2 z rozdělení N(0, 1). (Využijte vztahů (4.6.2) a (4.6.4)!)

$$\Big[\pi^{-\frac{1}{2}}.\Big]$$

Část II

Metody odhadu parametrů a jejich funkcí

Kapitola 5

Podstata úlohy odhadu: bodové a intervalové odhady

5.1 Příklad.

Obecnou úlohu odhadu parametrů v rozdělení pravděpodobnosti osvětlí nejlépe jednoduchý příklad. V laboratoři se měří doba života n výrobků nového typu (např. počet sepnutí, který vydrží nový druh kontaktů pro relé, než opálení jejich povrchu dosáhne určitého stupně, nebo počet otáček, které snese valivé ložisko, dokud nedojde k vydrolování povrchu kuliček nebo vnitřního povrchu pouzdra, apod.). Cílem zkoušek je zjistit, jaký podíl výrobků tohoto typu bude mít dobu života delší než dané číslo A (např. 800 000 sepnutí u kontaktů relé, 10^7 otáček u ložiska apod.). Označme $X_1, \ldots, X_n$ měřené doby života. Je třeba najít funkci n proměnných $T(x_1, \ldots, x_n)$, která po dosazení naměřených hodnot veličin $X_1, \ldots, X_n$ bude „dobrým přiblížením“ k hledanému podílu výrobků s dobou života větší než A.

Statistická podstata uvedené úlohy je tato: Doba života výrobku je náhodná veličina X s distribuční funkcí $F(x)$, i při přesném dodržení postupu a stejných provozních podmínkách (ty jsou při zkoušce v laboratoři zaručeny) budou doby života jednotlivých vzorků různé. Tvar distribuční funkce $F(x)$ závisí na druhu výrobku a na mechanismu, který vyvolává opotřebení (tření, tepelné namáhání atd.). V našem příkladě lze očekávat, že distribuční funkce $F(x)$ bude mít tvar

$$F(x) = 1 - \mathrm{e}^{-\left(\frac{x}{\delta}\right)^c}, \qquad x > 0,$$

tj. že X má Weibullovo rozdělení. Distribuční funkce $F(x)$ zde závisí na číslech δ a c, tzv. parametrech rozdělení. Tyto parametry jsou zase dány způsobem zpracování (technologií), podmínkami provozu atd. Z velkého počtu výrobků daného typu při dané technologii podíl těch, jejichž doba života překročí dané kladné číslo A, bude přibližně roven

$$P(X > A) = 1 - F(A) = \mathrm{e}^{-\left(\frac{A}{\delta}\right)^c}.$$

To znamená, že hledaná funkce je funkcí dvou parametrů δ a c závislých na konkrétních podmínkách (materiál, technologie, způsob užívání).

Měřené doby života $X_1, \ldots, X_n$ tedy jsou v daném příkladě náhodným výběrem z rozdělení s distribuční funkcí

$$F(x; \delta, c) = 1 - \mathrm{e}^{-\left(\frac{x}{\delta}\right)^c}, \qquad x > 0;$$

zápisem $F(x; \delta, c)$ jsme vyjádřili skutečnost, že distribuční funkce $F(x)$ závisí na dvou parametrech δ a c. Hledá se statistika $T(X_1, \ldots, X_n)$, která by „dobře aproximovala“ skutečnou hodnotu

$$\tau(\delta, c) = \mathrm{e}^{-\left(\frac{A}{\delta}\right)^c}.$$

Jakožto statistika – tj. funkce náhodného výběru – je $T = T(X_1, \ldots, X_n)$ náhodná veličina, která má také své rozdělení pravděpodobnosti; bylo by žádoucí zvolit funkci $T(X_1, \ldots, X_n)$ tak, aby s co největší pravděpodobností nabývala hodnot blízkých skutečné hodnotě $\tau(\delta, c)$. Protože uznáváme, že $T = T(X_1, \ldots, X_n)$ je náhodná veličina a může nabývat s různými pravděpodobnostmi hodnot blízkých správné hodnotě $\tau = \tau(\delta, c)$, nazýváme $T = T(X_1, \ldots, X_n)$ *odhadem* parametrické funkce $\tau(\delta, c)$.

Při důkladnějším rozboru měřených hodnot $X_1, \ldots, X_n$ se musíme věnovat ještě otázce ohodnocení chyb, kterých se můžeme dopustit, když nahradíme skutečnou hodnotu $\tau = \tau(\delta, c)$ hodnotou statistiky $T(X_1, \ldots, X_n)$. Zpravidla se tato otázka řeší jedním ze dvou následujících způsobů:

- údaj hodnotě statistiky T se doprovodí údajem hodnoty některé charakteristiky variability statistiky T, např. $\sqrt{\mathrm{var}(T)}$;
- jako odhad τ se neuvede jediná hodnota T, nýbrž dvojice hodnot, řekněme T_d, T_h, kde T_d a T_h jsou statistiky zvolené tak, aby pravděpodobnost, že T_d nabude hodnoty menší než správná hodnota τ a zároveň

T_h nabude hodnoty větší než správná hodnota τ, byla alespoň přibližně rovna zvolenému číslu blízkému 1, tj. aby

$$P(T_d < \tau < T_h) \doteq 1 - \alpha.$$

kde $1 - \alpha$ je zvolené číslo, např. $0,95$ nebo $0,99$.

Druhý z obou uvedených postupů lze také vyjádřit takto: T_d a T_h jsou statistiky zvolené tak, aby interval s náhodnými krajními body (T_d, T_h) pokryl správnou hodnotu τ s předem danou pravděpodobností.

5.2 Obecná formulace úlohy odhadu.

Shrnutím úvah z předchozího odstavce dospějeme k následující formulaci *úlohy odhadu.* Nechť $\mathbf{X} = (x_1, \ldots, X_n)'$ je náhodný výběr z rozdělení s distribuční funkcí $F(x; \theta_1, \ldots, \theta_k)$ závislou na k-rozměrném vektoru parametrů $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$. Skutečné hodnoty složek θ_i vektoru $\boldsymbol{\theta}$ jsou neznámy, je jen známo, že $\boldsymbol{\theta}$ je prvkem nějaké dané množiny Ω, kterou nazýváme *parametrickým prostorem.* Nechť na Ω je dána reálná funkce $\tau(\boldsymbol{\theta})$. Řešení úlohy odhadu funkce $\tau(\boldsymbol{\theta})$ spočívá v konstrukci funkce $T(\mathbf{x})$ na množině všech možných $\mathbf{x}$ takové, že rozdělení statistiky $T = T(\mathbf{X})$ se vyznačuje co nejvyšším stupněm koncentrace okolo správné hodnoty $\tau(\boldsymbol{\theta})$, a to pokud možno při všech hodnotách $\boldsymbol{\theta}$. Takovou statistiku $T(\mathbf{X})$ nazýváme *bodovým odhadem* funkce $\tau(\boldsymbol{\theta})$.

Požadavek co největšího soustředění rozdělení statistiky $T(\mathbf{X})$ kolem skutečné hodnoty $\tau(\boldsymbol{\theta})$ se nejčastěji vyjadřuje pomocí střední hodnoty $E\big(T(\mathbf{X})\big)$ a rozptylu $\operatorname{var}\big(T(\mathbf{X})\big)$ odhadu $T(\mathbf{X})$.

5.2.1 Nestranný odhad.

Jestliže platí

$$E\big(T(\mathbf{X})\big) = \tau(\boldsymbol{\theta}) \quad \text{pro všechna} \quad \boldsymbol{\theta} \in \Omega, \tag{5.2.1}$$

nazýváme $T(\mathbf{X})$ *nestranným (též nevychýleným) odhadem* funkce $\tau(\boldsymbol{\theta})$.

5.2.2 Nejlepší nestranný odhad.

Jestliže $T(\mathbf{X})$ je nestranný odhad funkce $\tau(\boldsymbol{\theta})$ a navíc při všech $\boldsymbol{\theta} \in \Omega$ platí pro jakýkoliv jiný nestranný odhad $T^*(\mathbf{X})$ funkce $\tau(\boldsymbol{\theta})$ nerovnost

$$\operatorname{var}\big(T(\mathbf{X})\big) \leq \operatorname{var}\big(\mathbf{T}^*(\mathbf{X})\big), \tag{5.2.2}$$

nazývá se $T(\mathbf{X})$ *nejlepší nestranný odhad* funkce $\tau(\boldsymbol{\theta})$.

5.2.3 Vychýlení odhadu.

Rozdíl

$$B(\boldsymbol{\theta}) = E\big(T(\mathbf{X})\big) - \tau(\boldsymbol{\theta}) \tag{5.2.3}$$

se nazývá *vychýlení* (nebo *jednostrannost*) odhadu $T(\mathbf{X})$.

Nejlepší nestranný odhad je většinou přijatelným řešením úlohy odhadu. Pro některé funkce $\tau(\boldsymbol{\theta})$ však vůbec žádný nestranný odhad neexistuje – příklady takových situací uvidíme v čl. 6 – nebo je jeho konstrukce natolik obtížná, že se pro daný účel nevyplatí. V takových případech je třeba posuzovat odhady přicházející pro danou funkci $\tau(\boldsymbol{\theta})$ v úvahu podle různých kritérií „ad hoc“, např. volit ten, který má pro hodnoty $\boldsymbol{\theta}$ nejčastěji se vyskytující přijatelné malé vychýlení a malý rozptyl apod. Někdy se odhady posuzují také podle tzv. střední kvadratické chyby.

5.2.4 Střední kvadratická chyba.

Střední kvadratickou chybou odhadu $T(\mathbf{X})$ rozumíme střední hodnotu čtverce odchylky odhadu od skutečné hodnoty odhadované funkce

$$K(\boldsymbol{\theta}) = E\Big(\big(T(\mathbf{X}) - \tau(\boldsymbol{\theta})\big)^{\mathbf{2}}\Big) = \mathbf{B^2}(\boldsymbol{\theta}) + \operatorname{var}\big(\mathbf{T}(\mathbf{X})\big), \tag{5.2.4}$$

neboť

$$K(\boldsymbol{\theta}) = E\Big(\big(T(\mathbf{X}) - \mathbf{E}\big(\mathbf{T}(\mathbf{X})\big) + \mathbf{B}(\boldsymbol{\theta})\big)^{\mathbf{2}}\Big) =$$

$$= \operatorname{var}\big(T(\mathbf{X})\big) + \mathbf{B^2}(\boldsymbol{\theta}) + \mathbf{2B}(\boldsymbol{\theta})\mathbf{E}\Big(\mathbf{T}(\mathbf{X}) - \mathbf{E}\big(\mathbf{T}(\mathbf{X})\big)\Big)$$

a střední hodnota v posledním výrazu je nulová.

Jindy volíme odhad, který má aspoň příznivé asymptotické vlastnosti, tj. který má dobré vlastnosti při velkých rozsazích výběru n. Zcela logickým požadavkem, který by měl dobrý odhad splňovat, je požadavek, aby se blížil skutečné hodnotě odhadované funkce tím těsněji, čím větší je počet pozorování. Tento požadavek je formalizován v následující definici.

5.2.5 Konzistentní odhad.

Odhad $T(\mathbf{X})$ funkce $\tau(\boldsymbol{\theta})$ se nazývá *konzistentní,* jestliže platí

$$\lim_{n\to\infty} P(|T(\mathbf{X}) - \tau(\boldsymbol{\theta})| < \varepsilon) = 1 \tag{5.2.5}$$

pro libovolné $\varepsilon > 0$ a pro všechna $\boldsymbol{\theta} \in \Omega$, tj. jestliže odhad s rostoucím počtem pozorování konverguje podle pravděpodobnosti (viz [24], odst. 25.2) ke správné hodnotě odhadované funkce $\tau(\boldsymbol{\theta})$.

Pro některé funkce $\tau(\boldsymbol{\theta})$ parametru ($\boldsymbol{\theta}$ nelze použít nestranného odhadu [nestranný odhad $T(\mathbf{X})$ třeba vůbec neexistuje nebo sice existuje, ale je příliš náročný na numerické výpočty]. Často však v takových případech lze najít vychýlený odhad, jehož vychýlení $B(\boldsymbol{\theta})$ klesá při rostoucím počtu pozorování n poměrně rychle k nule.

5.2.6 Asymptoticky nestranný odhad.

Odhad $T(\mathbf{X})$ je asymptoticky nestranný, jestliže

$$\lim_{n\to\infty} B(\boldsymbol{\theta}) = \lim_{n\to\infty} \Big(E\big(T(\mathbf{X})\big) - \tau(\boldsymbol{\theta})\Big) = \mathbf{0}. \tag{5.2.6}$$

Uvažujme odhad $T(\mathbf{X})$, který splňuje podmínku (5.2.6) a dále podmínku

$$\lim_{n\to\infty} \operatorname{var}\big(T(\mathbf{X})\big) = \mathbf{0}. \tag{5.2.7}$$

Vzhledem k platnosti (5.2.6) existuje pro každé $\varepsilon > 0$ přirozené číslo n_ε takové, že pro $n > n_\varepsilon$ je $-\frac{\varepsilon}{2} < B(\boldsymbol{\theta}) < \frac{\varepsilon}{2}$. Tudíž pro $n > n_\varepsilon$ platí

$$P\big(-\varepsilon < T(\mathbf{X}) - \tau(\boldsymbol{\theta}) < \varepsilon\big) = P\Big(-\varepsilon - B(\boldsymbol{\theta} < T(\mathbf{X}) - E\big(T(\mathbf{X})\big) < \varepsilon - B(\boldsymbol{\theta})\Big) \geq$$

$$\geq P\Big(-\frac{\varepsilon}{2} < T(\mathbf{X}) - E\big(T(\mathbf{X})\big) < \frac{\varepsilon}{2}\Big) \geq 1 - \frac{4\operatorname{var}\big(T(\mathbf{X})\big)}{\varepsilon^2};$$

poslední vztah vyplývá z Čebyševovy nerovnosti (viz [24], odst. 25.4). Je tedy

$$\lim_{n\to\infty} P\big(\big|T(\mathbf{X} - \tau(\boldsymbol{\theta}))\big| < \varepsilon\big) \geq \lim_{n\to\infty} \left(1 - \frac{4\operatorname{var}\big(T(\mathbf{X})\big)}{\varepsilon^2}\right) = 1,$$

takže platí (5.2.5).

Jestliže tedy $T(\mathbf{X})$ je asymptotický nestranný odhad parametrické funkce $\tau(\boldsymbol{\theta})$ a platí vztah (5.2.7), je $T(\mathbf{X})$ konzistentní odhad $\tau(\boldsymbol{\theta})$.

5.3 Příklady.

5.3.1

Mějme náhodný výběr $\mathbf{X} = (X_1, \ldots, X_n)'$ z rozdělení, které má konečný čtvrtý centrální moment μ_4. Uvažujme odhady

$$T_1 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = S^2, \quad T_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = M_2 = \frac{n-1}{n}S^2$$

rozptylu σ^2 tohoto rozdělení.

Statistika T_1 je nestranný odhad σ^2 a pro její rozptyl $\mathrm{var}(T_1) = \mathrm{var}(S^2)$ platí vzhledem k (2.2.5)

$$\lim_{n\to\infty} \mathrm{var}(T_1) = 0.$$

Tudíž T_1 je konzistentní odhad σ^2.

Statistika T_2 je asymptotický nestranný odhad σ^2 a pro jeho rozptyl $\mathrm{var}(T_2)$ platí

$$\lim_{n\to\infty} \mathrm{var}(T_2) = \lim_{n\to\infty} \mathrm{var}(T_1) = 0,$$

takže i T_2 je konzistentní odhad σ^2.

5.3.2

Mějme náhodný výběr $\mathbf{X} = (X_1, \ldots, X_n)'$ z rozdělení, které má konečný $2r$-tý moment μ'_{2r}. Statistika

$$T = \frac{1}{n}\sum_{i=1}^{n} X_i^r = M'_r$$

je nestranný odhad μ'_r a pro její rozptyl platí vzhledem k (2.2.9)

$$\lim_{n\to\infty} \mathrm{var}(T) = 0.$$

Je tedy M'_r konzistentním odhadem μ'_r.

Z Chinčinovy věty (viz [24], odst. 25.6) vyplývá, že M'_r konverguje podle pravděpodobnosti k μ'_r (tj. M'_r je konzistentní odhad μ'_r), je-li μ'_r konečné, bez ohledu na to, zda μ'_{2r} je či není konečné.

5.4 Úlohy.

5.4.1

V příkladě 5.3.1 nalezněte střední kvadratické chyby odhadu T_1 a T_2 pro případ, že náhodný výběr $\mathbf{X}$ pochází z rozdělení $\mathrm{N}(\mu, \sigma^2)$. Dále uvažujte odhad

$$T_3 = c\sum_{\imath=1}^{n}(X_i - \overline{X})^2$$

a stanovte číslo c tak, aby střední kvadratická chyba odhadu T_3 byla minimální.

$$\left[K_1(\sigma^2) = var(S^2) = \tfrac{2\sigma^4}{n-1}, \quad K_2(\sigma^2) = \tfrac{(2n-1)\sigma^4}{n^2} < K_1(\sigma^2), \quad c = \tfrac{1}{n+1}.\right]$$

5.4.2

V uspořádaném výběru $\mathbf{X}^* = (X_{(1)}, \ldots, X_{(8)}$ z rozdělení $\mathrm{N}(\mu, \sigma^2)$ uvažujte odhad

$$T = cR = c(X_{(8)} - X_{(1)}).$$

Z tabulek [23] (tab. 15A) nalezněte c tak, aby T byl nestranný odhad směrodatné odchylky σ, a určete rozptyl tohoto odhadu.

$$\left[T = 0,351\,222\,R; \quad \mathrm{var}(T) = 0,082\,911\,\sigma^2.\right]$$

5.4.3

Nechť $\mathbf{X}^* = (X_{(1)}, \ldots, X_{(n)})'$ je uspořádaný výběr z rozdělení, které má hustotu pravděpodobnosti $f(x) = 1/\theta, \quad 0 < x < \theta, f(x) = 0$ jinak, kde $\theta > 0$ je neznámý parametr. Uvažujte statistiky

$$T_i = c_i X_{(i)}, \qquad i = 1, \ldots, n.$$

Stanovte čísla c_i tak, aby T_i byly nestranné odhady parametru θ. Které T_i má pak nejmenší rozptyl? (Využijte výsledků úlohy 4.8.1, kde se uvažuje $\theta = 1$.)

$$\left[c_i = \tfrac{n+1}{i}, \quad i = 1, \ldots, n; \quad \mathrm{var}(T_n) = \tfrac{\theta^2}{n(n+2)}.\right]$$

Kapitola 6

Metody konstrukce bodových odhadů

6.1 Exponenciální třída rozdělení pravděpodobnosti.

Nejlepší nestranné odhady – pokud vůbec pro danou parametrickou funkci nějaký nestranný odhad existuje – lze nejlépe konstruovat, když pozorovaná veličina X má rozdělení patřící do určité speciální třídy, totiž do tzv. exponenciální třídy rozdělení.

Říkáme, že veličina X má *rozdělení z exponenciální třídy* (nebo *rozdělení exponenciálního typu*), jestliže její hustota pravděpodobnosti $f(x)$ (případně pravděpodobnostní funkce $p(x) = P(X = x)$, jde-li o rozdělení diskrétního typu) má tvar

$$f(x;\boldsymbol{\theta}) = \exp\Big(\sum_{j=1}^{k} Q_j(\boldsymbol{\theta})U_j(x) + R(\boldsymbol{\theta}) + V(x)\Big) \tag{6.1.1}$$

a splňuje podmínky

$$\big\{x \,\big|\, f(x;\boldsymbol{\theta}) > 0\big\} \text{ nezávisí na } \boldsymbol{\theta}; \tag{6.1.2}$$

parametrický prostor Ω obsahuje k-rozměrný interval, tj. body $\boldsymbol{\theta}$, pro které $f(x;\boldsymbol{\theta})$ je hustotou pravděpodobnosti (pravděpodobnostní funkcí), neleží v žádné nadploše $g(\boldsymbol{\theta}) = 0$. (6.1.3)

Většina rozdělení užitečných pro aplikace patří do exponenciální třídy.

6.2 Příklady.

6.2.1

Poissonovo rozdělení patří do exponenciální třídy, neboť jeho pravděpodobnostní funkci (zde místo θ označíme parametr rozdělení symbolem λ, jak jsme to činili v [24])

$$p(x;\lambda) = P(X = x) = \frac{\lambda^x}{x!}\mathrm{e}^{-\lambda}, \qquad x = 0, 1, \ldots,$$

lze zapsat ve tvaru

$$p(x;\lambda) = \exp(x \ln \lambda - \lambda - \ln x!),$$

tj. ve tvaru (6.1.1), kde $k = 1,\ Q_1(\lambda) = \lambda,\ U_1(x) = x;\ \{x|p(x;\lambda) > 0\} = \{0, 1, \ldots\}$, tedy nezávisí na λ a $\Omega = (0, \infty)$, tedy Ω obsahuje jednorozměrný interval.

6.2.2

Logaritmicko-normální rozdělení patří do exponenciální třídy, neboť jeho hustotu

$$f(x;\mu,\sigma^2) = \frac{1}{\sigma x\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2\sigma^2}(\ln x-\mu)^2}, \qquad x > 0,$$

lze zapsat ve tvaru

$$f(x;\mu,\sigma^2) = \exp\big(-\frac{1}{2\sigma^2}(\ln x)^2 + \frac{\mu}{\sigma^2}\ln x - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\ln\sigma^2 - \ln x - \ln\sqrt{2\pi}\big),$$

tj. ve tvaru (6.1.1) s

$$k = 2,\ \boldsymbol{\theta} = (\mu,\ \sigma^2),\ Q_1(\boldsymbol{\theta}) = -\frac{1}{2\sigma^2},\ Q_2(\boldsymbol{\theta}) = \frac{\mu}{\sigma^2},\ U_1 = (\ln x)^2,\ U_2 = \ln x,$$

$$R(\boldsymbol{\theta}) = -\frac{1}{2}\left(\frac{\mu^2}{\sigma^2} + \ln\sigma^2\right) - \ln\sqrt{2\pi},\ V(x) = -\ln x.$$

Parametrický prostor $\Omega = \{(\mu,\sigma^2)|-\infty < \mu < \infty,\ \sigma^2 > 0\}$ je polorovina, množina $\{x|f(x;\mu,\sigma^2) > 0\} = (0,\infty)$, tedy nezávisí na (μ,σ^2).

6.2.3

Rovnoměrné rozdělení na intervalu $(0, \theta)$ nepatří do exponenciální třídy, jeho hustota totiž je

$$f(x;\theta) = \frac{1}{\theta}, \qquad 0 < x < \theta,$$

takže množina $\{x | f(x;\theta) > 0\}$ závisí na θ, a nesplňuje tedy podmínku (6.1.2).

6.2.4

Cauchyovo rozdělení nepatří do exponenciální třídy, neboť jeho hustota

$$f(x;\theta) = \frac{1}{\pi}\Big(1 + (x-\theta)^2\Big)^{-1} = \exp\Big(- \ln \pi - \ln\big(1 + (x-\theta)^2\big)\Big)$$

nemá tvar (6.1.1).

6.3 Postačující statistiky.

Je-li $\mathbf{X} = (X_1, \ldots, X_n)'$ náhodný výběr z rozdělení exponenciálního typu (exponenciální třídy), pak sdružená hustota náhodného vektoru $\mathbf{X}$ je

$$f(x;\boldsymbol{\theta}) = \exp\Big(\sum_{j=1}^{k} Q_j(\boldsymbol{\theta}) \sum_{i=1}^{n} U_j(x_i) + nR(\boldsymbol{\theta}) + \sum_{i=1}^{n} V(x_i)\Big) = \tag{6.3.1}$$

$$= \exp\Big(\sum_{j=1}^{k} Q_j(\boldsymbol{\theta}) S_j(\mathbf{x}) + nR(\boldsymbol{\theta}) + V(\mathbf{x})\Big),$$

kde

$$S_j(\mathbf{x}) = S_j(x_1, \ldots, x_n) = \sum_{i=1}^{n} U_j(x_i), \quad j = 1, \ldots, k, \qquad V(\mathbf{x}) = \sum_{i=1}^{n} V(x_i). \tag{6.3.2}$$

Statistiky $S_1(\mathbf{X}), \ldots, S_k(\mathbf{X})$ dané výrazy (6.3.2) představují největší možnou redukci výsledků pozorování, nejúčelnější nahrazení všech n pozorování menším počtem údajů. Říkáme jim proto *minimální postačující statistiky.* Odhady s nejlepšími vlastnostmi pro funkce $\tau(\theta)$ parametrů rozdělení exponenciální třídy jsou vždy funkcemi statistik (6.3.2). Plyne to z následující věty, kterou uvádíme bez důkazu [27].

6.4 Věta.

Nechť $\mathbf{X}$ *je náhodný výběr z rozdělení exponenciálního typu (6.1.1) a nechť* $\tau(\boldsymbol{\theta})$ *je daná funkce parametru* $\boldsymbol{\theta}$*. Jestliže existuje nestranný odhad* $T^*(\mathbf{X})$ *pro funkci* $\tau(\boldsymbol{\theta})$*, pak existuje funkce* $T(s_1, \ldots, s_k)$ k *proměnných taková, že*

$$T = T(S_1(\mathbf{X}, \ldots, S_k(\mathbf{X})) \tag{6.4.1}$$

je také nestranný odhad $\tau(\boldsymbol{\theta})$ *a přitom*

$$\operatorname{var}(T) \leq \operatorname{var}(T^*). \tag{6.4.2}$$

Funkce (6.4.1) *je právě jedna.*

Z uvedené věty plyne: Jestliže pro funkci $\tau(\boldsymbol{\theta})$ parametru $\boldsymbol{\theta}$ rozdělení exponenciální třídy existují nestranné odhady, pak nejlepší z nich je funkcí statistik typu (6.3.2) a ten je určen jednoznačně.

6.5 Nalezení nejlepšího nestranného odhadu.

Nalezení funkce T statistik $S_1(\mathbf{X}), \ldots, S_k(\mathbf{X})$ může být obtížné; nepodaří-li se takovou funkci „uhodnout“, osvědčuje se často tento postup: Najít co nejjednodušší nestranný odhad funkce $\tau(\boldsymbol{\theta})$, řekněme $T_0(\mathbf{X})$, vypočítat střední hodnotu odhadu $T_0(\mathbf{X})$ podmíněnou danými hodnotami statistik $S_1(\mathbf{X} = s_1, \ldots, S_k(\mathbf{X}) = s_k$,

$$T(\mathbf{s}) = E\Big(T_0(\mathbf{X} \,|\, S_1(\mathbf{X}) = s_1, \ldots, S_k(\mathbf{X}) = s_k\Big),$$

kde $\mathbf{s} = (s_1, \ldots, s_k)$, tj.

$$T(\mathbf{s}) = \sum_x T_0(x) P\Big(\mathbf{X} = \mathbf{x} \,|\, S_1(\mathbf{X}) = s_1, \ldots, S_k(\mathbf{X}) = s_k\Big) \tag{6.5.1}$$

v diskrétním případě a

$$T(\mathbf{s}) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} T_0(\mathbf{x}) f\big(\mathbf{x} | S_1(\mathbf{X}) = s_1, \ldots, S_k(\mathbf{X}) = s_k\big)\, \mathrm{d}x_1 \ldots \mathrm{d}x_n \tag{6.5.2}$$

ve spojitém případě. Postup bude ilustrován v následujícím odstavci na příkladech.

6.6 Příklady.

6.6.1

Budiž $\mathbf{X} = (X_1, \ldots, X_n)'$ náhodný výběr z Poissonova rozdělení s parametrem λ. To je rozdělení z jednoparametrické exponenciální třídy (viz příkl. 6.2.1); statistika $S(\mathbf{X})$ je rovna

$$S(\mathbf{X}) = \sum_{i=1}^{n} X_i.$$

Rozdělení této statistiky $S = S(\mathbf{X})$ je (viz [24], odst. 15.4) Poissonovo s parametrem $n\lambda$, takže

$$P(S = s) = \frac{(n\lambda)^s}{s!}\mathrm{e}^{-n\lambda}, \qquad s = 0, 1, \ldots.$$

Pro libovolné přirozené číslo k je

$$E\big(S(S-1)(S-2)\ldots(S-k+1)\big) = \sum_{s=k}^{\infty} s(s-1)\ldots(s-k+1)\frac{(n\lambda)^s}{s!}\mathrm{e}^{-n\lambda} =$$

$$= \sum_{s=k}^{\infty} \frac{(n\lambda)^s}{(s-k)!}\mathrm{e}^{-n\lambda} = (n\lambda)^k \sum_{t=0}^{\infty} \frac{(n\lambda)^t}{t!}\mathrm{e}^{-n\lambda} = (n\lambda)^k.$$

Odtud plyne, že

$$T_k(\mathbf{X}) = \frac{1}{n^k}\Big(\sum_{i=1}^{n} X_i\Big)\Big(\sum_{i=1}^{n} X_i - 1\Big)\ldots\Big(\sum_{i=1}^{n} X_i - k + 1\Big) \tag{6.6.1}$$

je nejlepší nestranný odhad parametrické funkce $\tau_k(\lambda) = \lambda_k$.

6.6.2

Za stejných podmínek jako v příkl. 6.6.1 stanovme nejlepší nestranný odhad parametrické funkce $\tau(\lambda) = \mathrm{e}^{-\lambda}$. Zde lze použít postupu naznačeného v odst. 6.5; využije se skutečnosti, že

$$\tau(\lambda) = e^{-\lambda} = P(X_1 = 0),$$

a položí se

$$\begin{aligned} T_0(\mathbf{x}) &= 1 \quad \text{pro} \quad x_1 = 0, \\ &= 0 \quad \text{pro} \quad x_1 \neq 0, \end{aligned}$$

Pak je skutečně

$$E\big(T_0(\mathbf{X})\big) = P(X_1 = 0) = \mathrm{e}^{-\lambda}$$

a počítá se

$$E\big(T_0(\mathbf{X}) \,\big|\, S = s\big) = P\Big(X_1 = 0 \,\Big|\, \sum_{i=1}^{n} X_i = s\Big) = \frac{P(X_1 = 0,\ \sum_{i=2}^{n} X_i = s)}{P(\sum_{i=1}^{n} X_i = s}$$

$$= \frac{\mathrm{e}^{-\lambda}\mathrm{e}^{-(n-1)\lambda}[(n-1)\lambda]^s/s!}{\mathrm{e}^{-n\lambda}(n\lambda)^s/s!} = \Big(1 - \frac{1}{n}\Big)^s.$$

Nejlepší nestranný odhad $\tau(\lambda) = \mathrm{e}^{-\lambda}$ tedy je

$$T(\mathbf{X}) = \Big(1 - \frac{1}{n}\Big)^{\sum_{i=1}^{n} X_i}. \tag{6.6.2}$$

Předpokládejme např., že při výrobě tabulového skla počet kazů (bublin) v tabuli je náhodná veličina s rozdělením $\mathrm{Po}(\lambda)$. Podrobnou prohlídkou $n = 25$ tabulí byly zjištěny tyto počty bublin v jednotlivých tabulích:

0, 0, 0, 1, 0, 5, 0, 0, 1, 0, 0, 2, 0, 0, 0, 0, 3, 0, 0, 1, 0, 0, 3, 0, 1,

tj. 17 tabulí bez kazu, 4 tabule s jedním kazem, 1 tabule se dvěma, 2 tabule se třemi a 1 s pěti kazy.

Statistika $T = \sum_{i=1}^{25} X_i$ zde nabývá hodnoty $17 = 0 \cdot 17 + 1 \cdot 4 + 2 \cdot 1 + 3 \cdot 2 + 5 \cdot 1$. Statistika (6.6.1) pro $k = 1$, která je nejlepší nestranný odhad parametru (středního počtu kazů v jedné tabuli) λ, nabývá hodnoty $\frac{17}{25} = 0,68$ a statistika (6.6.2), která je nejlepší nestranný odhad $\tau(\lambda) = e^{-\lambda}$ (tj. nejlepší nestranný odhad podílu tabulí ve výrobě nemajících žádný kaz), nabývá hodnoty

$$\big(1 - \frac{1}{25}\big)^{17} = 0,499\,6.$$

6.6.3

Nechť $\mathbf{X} = (X_1, \ldots, X_n)'$ je náhodný výběr z alternativního rozdělení

$$p(x;\pi) = \pi^x(1-\pi)^{1-x}, \quad x = 0, 1.$$

Toto rozdělení má tvar (6.1.1), položí-li se $k = 1,\ Q_1(\pi) = \ln\left(\frac{\pi}{1-\pi}\right),\ R(\pi) = \ln(1-\pi),\ V(x) = 0,\ U(x) = x$. Statistikou, na které lze založit nestranné odhady všech funkcí $\tau(\pi)$ parametru π, je tedy

$$S(\mathbf{X}) = \sum_{i=1}^{n} X_i.$$

Statistika $S = S(\mathbf{X})$ má binomické rozdělení Bi$(n,\ \pi)$, takže

$$P(S = s) = \binom{n}{s}\pi^s(1-\pi)^{n-s}, \quad s = 0, 1, \ldots, n.$$

Hledejme nejlepší nestranné odhady funkcí

$$\tau_1(\pi) = \pi = E(X_i), \quad \tau_2(\pi) = \pi(1-\pi)\frac{1}{n} = \mathrm{var}\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big),$$

$$\tau_3(\pi) = \pi^n = P(X_i = 1,\ i = 1, \ldots, n), \quad \tau_4(\pi) = \pi^m, \quad \text{kde } m > n.$$

(1) Jak je známo ([24], odst. 14.4), $E(S) = n\pi$; tedy

$$\frac{S}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{6.6.3}$$

je nejlepší nestranný odhad $\tau_1(\pi) = \pi$.

(2) Ke konstrukci nejlepšího nestranného odhadu funkce $\tau_2(\pi)$ použijeme postupu z odst. 6.5. Součin $\pi(1-\pi)$ je totiž roven pravděpodobnosti, že dvě nezávislé náhodné veličiny $X_1,\ X_2$ s alternativním rozdělením nabudou hodnot $X_1 = 1,\ X_2 = 0$. Položíme tedy

$$\begin{aligned} T_0(\mathbf{x}) &= 1 \quad \text{pro } x_1 = 1, \quad x_2 = 0, \\ &= 0 \quad \text{ve všech ostatních případech,} \end{aligned}$$

a vypočteme

$$E\Big(T_0(\mathbf{X})\sum_{i=1}^n X_i = s\Big) = P\Big(T_0(\mathbf{X}) = 1 \,\Big|\, \sum_{i=1}^n X_i = s\Big) =$$

$$= P\Big(X_1 = 1,\ X_2 = 0 \,\Big|\, \sum_{i=3}^n X_i = s-1\Big) =$$

$$= \frac{\pi(1-\pi)\binom{n-2}{s-1}\pi^{s-1}(1-\pi)^{n-s-1}}{\binom{n}{s}\pi^s(1-\pi)^{n-s}} = \frac{\binom{n-2}{s-1}}{\binom{n}{s}} = \frac{s(n-s)}{n(n-1)}.$$

Nejlepším nestranným odhadem funkce $\tau_2(\pi) = \mathrm{var}\big(\sum_{i=1}^n X_i/n\big)$ tedy je

$$T_2(\mathbf{X}) = \frac{1}{n}\frac{\big(\sum_{i=1}^n X_i\big)\big(n - \sum_{i=1}^n X_i\big)}{n(n-1)}; \tag{6.6.4}$$

nicméně v praxi se často používá mírně jednostranného, ale asymptoticky nestranného odhadu

$$\frac{1}{n}\,\frac{\sum_{i=1}^n X_i}{n}\left(1 - \frac{\sum_{i=1}^n X_i}{n}\right)$$

(3) Funkce $\tau_3(\pi) = \pi^n$ je vlastně rovna pravděpodobnosti, že n vzájemně nezávislých náhodných veličin $X_1, \ldots, X_n$ s alternativním rozdělením nabude hodnoty 1.

Aplikací postupu odst. 6.5 s $T_0(\mathbf{x}) = \mathbf{1}$ pro $x_1 = \ldots = x_n = 1,\ T_0(\mathbf{x}) = 0$ ve všech ostatních případech dostaneme jako nejlepší nestranný odhad

$$\begin{aligned} T_3(s) &= E\big(T_0(\mathbf{X})\big|\, S = s\,\big) \\ &= P\big(X_1 = \ldots = X_n = 1 \,\big|\, \sum_{i=1}^n X_i = s\big) = 1, \quad s = n, \\ &= 0, \qquad\qquad\qquad\qquad\qquad\qquad\quad s \neq n; \end{aligned} \tag{6.6.5}$$

tzn. že bychom odhadli π^n jako 0 při $\sum_{i=1}^n X_i \neq n$ a jako 1 při $\sum_{i=1}^n X_i = n$. To by sice byl nestranný odhad s minimálním rozptylem, avšak zřejmě prakticky nepřijatelný.

Máme zde příklad situace, kdy nestranný odhad existuje, ale je absurdní, a kdy je lépe upustit od požadavku nestrannosti.

(4) Pro funkci $\tau_4(\pi) = \pi^m$ při $m > n$ nestranný odhad dokonce vůbec neexistuje. Kdyby totiž existoval, musela by - podle věty odst. 6.4 - existovat funkce $h(s)$ taková, že

$$E\big(h(S)\big) = \sum_{s=0}^{n} h(s)\binom{n}{s}\pi^s(1-\pi)^{n-s} = \pi^m \quad \text{pro všechna } \pi;$$

to však není možné, protože $E\big(h(S)\big)$ je polynom stupně nejvýše n.

6.6.4

Nechť $\mathbf{X} = (X_1, \ldots, X_n)'$ je náhodný výběr z normálního rozdělení s neznámou střední hodnotou μ a neznámým rozptylem σ^2. Hustota tohoto rozdělení je

$$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \\ = \exp\Big(-\frac{x^2}{2\sigma^2} + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\ln\sigma^2 - \frac{1}{2}\ln(2\pi)\Big);$$

má tedy tvar (6.1.1) s $U_1(x) = x^2$, $U_2(x) = x$. Statistikou pro odhad všech funkcí parametru $\boldsymbol{\theta} = (\mu,\, \sigma^2)$ tedy je dvojice

$$(S_1, S_2)' = \Big(\sum_{i=1}^{n} X_i^2,\ \sum_{i=1}^{n} X_i\Big)'.$$

Jelikož

$$E\Big(\sum_{i=1}^{n} X_i\Big) = n\mu, \quad E\Big(\sum_{i=1}^{n} X_i^2 - \frac{1}{n}(\sum_{i=1}^{n} X_i)^2\Big) = \sigma^2(n-1),$$

jsou nejlepší nestranné odhady funkcí $\tau_1(\mu,\sigma^2) = \mu$ a $\tau_2(\mu,\sigma^2) = \sigma^2$, tj. střední hodnoty a rozptylu, rovny

$$T_1(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}, \quad T_2(\mathbf{X}) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = S^2. \qquad (6.6.6)$$

6.7 Metoda maximální věrohodnosti.

V příkladech předcházejícího odstavce jsme viděli, že pro některé funkce parametrů neexistují nestranné odhady, pro jiné existují, ale jsou zcela absurdní, a konečně, že existují rozdělení, která nepatří do exponenciální třídy, pro kterou dovedeme nejlepší nestranné odhady aspoň někdy sestrojit. V takových situacích dává často dobré výsledky tzv. *metoda maximální věrohodnosti.*

Budiž $\mathbf{X} = (X_1, \ldots, X_n)'$ náhodný výběr z rozdělení diskrétního typu s pravděpodobnostní funkcí $p(x;\boldsymbol{\theta})$ nebo z rozdělení spojitého typu s hustotou pravděpodobnosti $f(x;\boldsymbol{\theta})$, kde $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$ je vektor neznámých parametrů. Představme si, že byl pozorován výběr $\mathbf{X} = \mathbf{x}$. Sdružená pravděpodobnostní funkce (příp. hustota pravděpodobnosti) náhodného výběru $\mathbf{X}$ v bodě $\mathbf{x}$ je

$$p(\mathbf{x};\boldsymbol{\theta}) = \prod_{i=1}^{n} p(x_i,\boldsymbol{\theta}), \quad \text{příp.} \quad f(\mathbf{X};\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i,\boldsymbol{\theta}),$$

neboť jednotlivá pozorování v náhodném výběru jsou vzájemně nezávislé náhodné veličiny. Vezměme dva různé body θ v parametrickém prostoru Ω, řekněme $\boldsymbol{\theta}'$ a $\boldsymbol{\theta}''$. Je-li $p(\mathbf{x};\boldsymbol{\theta}')$ o mnoho menší než $p(\mathbf{x};\boldsymbol{\theta}'')$ (příp. $f(\mathbf{x};\boldsymbol{\theta}')$ o mnoho menší než $f(\mathbf{x};\boldsymbol{\theta}'')$), znamená to, že daný výsledek pozorování $\mathbf{X} = \mathbf{x}$ (tj. daný výběr) má při $\boldsymbol{\theta} = \boldsymbol{\theta}'$ o mnoho menší pravděpodobnost než při $\boldsymbol{\theta} = \boldsymbol{\theta}''$ (příp. výsledky z okolí daného bodu $\mathbf{x}$ mají při $\boldsymbol{\theta} = \boldsymbol{\theta}'$ menší pravděpodobnost než při $\boldsymbol{\theta} = \boldsymbol{\theta}''$), a jsme tedy nakloněni považovat za správnou hodnotu parametru $\boldsymbol{\theta}$ spíše $\boldsymbol{\theta}''$ než $\boldsymbol{\theta}'$. V souladu s touto úvahou volíme za odhad parametru $\boldsymbol{\theta}$ ten vektor $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)'$ při kterém nabývá sdružená pravděpodobnostní funkce $p(\mathbf{x},\boldsymbol{\theta})$ [příp. sdružená hustota pravděpodobnosti $f(\mathbf{x};\boldsymbol{\theta})$] maximální hodnoty pro daný výběr $\mathbf{X} = \mathbf{x}$.

6.8 Funkce věrohodnosti.

Sdruženou pravděpodobnostní funkci (příp. sdruženou hustotu pravděpodobnosti) náhodného výběru $\mathbf{X}$, uvažovanou při daném $\mathbf{X} = \mathbf{x}$ jako funkci parametru $\boldsymbol{\theta}$, nazýváme *funkcí věrohodnosti.*

Funkci věrohodnosti značíme $L(\boldsymbol{\theta})$; je tedy

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i;\boldsymbol{\theta}), \quad \text{příp.} \quad L(\boldsymbol{\theta}) = \prod_{i=1}^{n} p(x_i;\boldsymbol{\theta}). \tag{6.8.1}$$

6.9 Maximálně věrohodný odhad.

Vektor statistik $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1(\mathbf{X}, \ldots, \hat{\theta}_k(\mathbf{X}))'$ definovaný výrazem

$$L(\hat{\boldsymbol{\theta}}) \geq L(\boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \Omega, \tag{6.9.1}$$

nazýváme maximálně věrohodným odhadem vektoru parametrů $\boldsymbol{\theta}$.

Zpravidla je výhodné místo vlastní funkce věrohodnosti $L(\boldsymbol{\theta})$ maximalizovat její logaritmus $\ln L(\boldsymbol{\theta})$. Jestliže pravděpodobnostní funkce (příp. hustota pravděpodobnosti) rozdělení, ze kterého výběr pochází, splňuje při každém $\mathbf{x}$ podmínky pro výpočet bodu $\hat{\boldsymbol{\theta}}$ pomocí standardních metod matematické analýzy (množina $\{\mathbf{x} \,\big|\, p(\mathbf{x};\boldsymbol{\theta}) > 0\}$, resp. $\{\mathbf{x} \,\big|\, f(\mathbf{x};\boldsymbol{\theta}) > 0\}$ nezávislá na $\boldsymbol{\theta}$, existence parciálních derivací $p(\mathbf{x};\boldsymbol{\theta})$, resp. $f(\mathbf{x};\boldsymbol{\theta})$ podle všech složek vektoru $\boldsymbol{\theta}$ při každém $\mathbf{x}$), stanoví se maximálně věrohodné odhady parametrů $\theta_j, \quad j = 1, \ldots, k$, řešením soustavy rovnic

$$\left.\frac{\partial \ln L(\boldsymbol{\theta}}{\partial \theta_j}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\mathbf{x})} = 0, \quad j = 1, \ldots, k. \tag{6.9.2}$$

Rovnice (6.9.2) se nazývají *věrohodnostní rovnice.* Má-li se odhadnout funkce $\tau(\boldsymbol{\theta})$ vektoru parametrů $\boldsymbol{\theta}$, bere se při užití metody maximální věrohodnosti za její odhad

$$\hat{\tau} = \tau(\hat{\theta}_1, \ldots, \hat{\theta}_k). \tag{6.9.3}$$

6.10 Příklady.

6.10.1

Nechť $\mathbf{X} = (X_1, \ldots, X_n)'$ je náhodný výběr z Rayleighova rozdělení (viz [24], odst. 22.6)

$$f(x, \sigma^2) = \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}}, \quad x > 0,$$

kde σ^2 je neznámý parametr. Funkce věrohodnosti je

$$L(\sigma^2) = \sigma^{-2n} \prod_{i=1}^{n} x_i \exp\Big(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2\Big).$$

Věrohodnostní rovnice je

$$\frac{\mathrm{d}\,\ln L(\sigma^2)}{\mathrm{d}\,\sigma^2} = -\frac{2n}{\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n} x_i^2 = 0$$

a maximálně věrohodné odhady parametru σ^2 a parametrické funkce σ jsou

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2, \qquad \hat{\sigma} = \Big(\frac{1}{n}\sum_{i=1}^{n} X_i^2\Big)^{\frac{1}{2}}.$$

6.10.2

Nechť $\mathbf{X} = (X_1, \ldots, X_n)'$ je náhodný výběr z rozdělení gama s hustotou

$$f(x; m, \delta) = \frac{x^{m-1}}{\delta^m \Gamma(m)} e^{-\frac{x}{\delta}}, \qquad x > 0,$$

kde m a δ jsou neznámé parametry. Věrohodnostní rovnice jsou

$$\frac{\sum_{i=1}^{n} x_i}{\hat{\delta}^2} - \frac{n\hat{m}}{\hat{\delta}} = 0,$$

$$\sum_{i=1}^{n} \ln x_i - n \ln \hat{\delta} - n \frac{\mathrm{d}\,\ln \Gamma(m)}{\mathrm{d}\,m}\bigg|_{m=\hat{m}} = 0.$$

Jednoduchou úpravou dostaneme

$$\frac{\sum_{i=1}^{n} \ln x_i}{n} - \ln\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) = -\ln \hat{m} + \frac{\mathrm{d}\,\ln \Gamma(m)}{\mathrm{d}\,m}\bigg|_{m=\hat{m}}$$

$$\hat{\delta} = \frac{\sum_{i=1}^{n} x_i}{n\hat{m}}.$$

První rovnici je ovšem nutno řešit numericky; příklad ukazuje, že maximálně věrohodný odhad nemá vždy explicitní vyjádření, nýbrž že je někdy dán soustavou rovnic, které vyžadují numerické řešení.

6.10.3

Nechť $\mathbf{X} = (X_1, \ldots, X_n)'$ je náhodný výběr negativního binomického rozdělení (viz [24], odst. 14.8)

$$p(x; m, \pi) = \frac{\Gamma(m+x)}{\Gamma(m)x!}\pi^m(1-\pi)^x, \qquad x = 0, 1, \ldots,$$

kde $m > 0$ a $\pi \in (0, 1)$ jsou neznámé parametry.

Po úpravách lze zapsat věrohodnostní rovnice ve tvaru

$$\frac{1}{n}\sum_{i=1}^{n}\Psi(\hat{m}+x_i) = \Psi(\hat{m}) + \ln\Big(1 + \frac{1}{n\hat{m}}\sum_{i=1}^{n}x_i\Big),$$

$$\hat{\pi} = \Big(1 + \frac{1}{n\hat{m}}\sum_{i=1}^{n}x_i\Big)^{-1},$$

kde

$$\Psi(t) = \frac{\mathrm{d}\ln\Gamma(t)}{\mathrm{d}t}.$$

Z první rovnice se vypočte postupnými aproximacemi $\hat{m}$ a $\hat{\pi}$ je pak už jednoduchou funkcí $\sum_{i=1}^{n} x_i$ a $\hat{m}$; rovnice pro určení $\hat{m}$ je zde však ještě o mnoho obtížněji řešitelná než rovnice pro $\hat{m}$ v příkl. 6.10.2.

6.11 Vlastnosti maximálně věrohodných odhadů.

Maximálně věrohodné odhady mají dvě důležité vlastnosti, pro které se jich používá tak často, přestože jejich výpočet je numericky náročný. První z těchto vlastností se týká výběrů jakéhokoliv rozsahu n a zní:

6.11.1 Věta

Jestliže rozdělení pozorované veličiny X patří do exponenciální třídy (6.1.1),*pak maximálně věrohodné odhady $\hat{\theta}, \ldots, \hat{\theta}_k$ parametrů $\theta_1, \ldots, \theta_k$ jsou funkcemi minimálních postačujících statistik* (6.3.2).

D ů k a z. Má-li X hustotu (pravděpodobnostní funkci) tvaru (6.1.1), pak sdružená hustota náhodného výběru $\mathbf{X} = (X_1, \ldots, X_n)'$ je typu (6.3.1), logaritmus funkce věrohodnosti má pak tvar

$$\ln L(\boldsymbol{\theta}) = \sum_{l=1}^{k} Q_l(\boldsymbol{\theta}) S_l(\mathbf{x}) + nR(\boldsymbol{\theta}) + V(\mathbf{x}), \tag{6.11.1}$$

věrohodnostní rovnice jsou

$$\sum_{l=1}^{k} S_l(\mathbf{x}) \frac{\partial Q_l(\boldsymbol{\theta})}{\partial \theta_j}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = 0, \qquad j = 1, \ldots, k, \tag{6.11.2}$$

a jejich řešení závisí na $\mathbf{x}$ jen prostřednictvím hodnot $S_1(\mathbf{x}), \ldots, S_k(\mathbf{x})$ minimálních postačujících statistik.

Z věty odst. 6.4 a věty odst. 6.10.1 pak přímo plyne důležitý důsledek.

6.11.2

Jestliže pozorovaná náhodná veličina X má rozdělení z exponenciální třídy (6.1.1) *a jestliže pro nějako funkci $T(\hat{\theta}_1, \ldots, \hat{\theta}_k)$ maximálně věrohodných odhadů parametrů $\theta_1, \ldots, \theta_k$ platí*

$$E[T(\hat{\theta}_1, \ldots, \hat{\theta}_k)] = \tau(\boldsymbol{\theta}) \quad \textit{pro všechna } \boldsymbol{\theta} \in \Omega, \tag{6.11.3}$$

pak $T(\hat{\theta}_1, \ldots, \hat{\theta}_k)$ je nejlepším nestranným odhadem funkce $\tau(\boldsymbol{\theta})$ vektoru parametrů $\boldsymbol{\theta}$.

Druhá významná vlastnost maximálně věrohodných odhadů je asymptotická, platí pro výběry velkého rozsahu n; umožňuje určit aspoň asymptotické rozdělení maximálně věrohodných odhadů v případech, kdy jejich přesné rozdělení nelze určit nebo kdy je toto přesné rozdělení prakticky neupotřebitelné pro svou velkou složitost. K vyslovení příslušné věty definujme dva důležité pojmy.

6.11.3 Informace.

Nechť rozdělení veličiny X má hustotu (pravděpodobnostní funkci) $f(x;\theta)$ závislou na parametru θ nabývajícím hodnot z nějakého intervalu Ω na přímce. Nechť $f(x;\theta)$ splňuje podmínky:

$$M = \{x \,\big|\, f(x;\theta) > 0\} \quad \text{nezávisí na } \theta; \tag{6.11.4}$$

$$\int_M \frac{\partial f(x;\theta)}{\partial\,\theta}\,\mathrm{d}x = \int_M \frac{\partial \ln f(x;\theta)}{\partial\,\theta} f(x;\theta)\,\mathrm{d}x = 0 \quad \text{pro všechna } \theta \in \Omega, \tag{6.11.5}$$

$$\sum_M \frac{\partial f(x;\theta)}{\partial\theta} = \sum_M \frac{\partial \ln f(x;\theta)}{\partial\,\theta} f(x;\theta) = 0 \quad \text{pro všechna } \theta \in \Omega;$$

$$J(\theta) = \int_M \left[\frac{\partial \ln f(x;\theta)}{\partial\,\theta}\right]^2 f(x;\theta)\,\mathrm{d}x \tag{6.11.6}$$

resp.

$$J(\theta) = \sum_M \left[\frac{\partial \ln f(x;\theta)}{\partial\,\theta}\right]^2 f(x;\theta),$$

je konečné kladné číslo pro každé $\theta \in \Omega$.

Pak funkci $J(\theta)$ parametru θ nazveme informací (podrobněji *Fisherovou mírou informace o parametru θ)* příslušnou k $f(x;\theta)$.

6.11.4 Informační matice.

Nechť rozdělení veličiny X má hustotu pravděpodobnosti (pravděpodobnostní funkci) $f(x;\boldsymbol{\theta})$ závislou na vektorovém parametru $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$ nabývajícím hodnot z intervalu Ω v k-rozměrném euklidovském prostoru. Nechť $f(x;\boldsymbol{\theta})$ splňuje podmínky:

$$M = \{x \,|\, f(x;\boldsymbol{\theta} > 0\} \quad \text{nezávisí na } \boldsymbol{\theta}; \tag{6.11.7}$$

$$\int_M \frac{\partial f(x;\boldsymbol{\theta})}{\partial\theta_j}\,\mathrm{d}x = \int_M \frac{\partial \ln f(x;\boldsymbol{\theta})}{\partial\theta_j} f(x;\boldsymbol{\theta})\,\mathrm{d}x = 0 \tag{6.11.8}$$

pro všechna $\boldsymbol{\theta} \in \Omega$ a pro všechna $j = 1, \ldots, k$, resp.

$$\sum_M \frac{\partial f(x;\boldsymbol{\theta})}{\partial\theta_j} = \sum_M \frac{\partial \ln f(x;\boldsymbol{\theta})}{\partial\theta_j} f(x;\boldsymbol{\theta}) = 0$$

pro všechna $\boldsymbol{\theta} \in \Omega$ a pro všechna $j = 1, \ldots, k$; matice

$$J(\boldsymbol{\theta}) = \left(\int_M \frac{\partial \ln f(x;\boldsymbol{\theta})}{\partial\theta_{j_1}} \frac{\partial \ln f(x;\boldsymbol{\theta})}{\theta_{j_2}} f(x;\boldsymbol{\theta})\,\mathrm{d}x\right), \quad j_1, j_2 = 1, \ldots, k, \tag{6.11.9}$$

má kladný a konečný determinant pro všechna $\boldsymbol{\theta} \in \Omega$.

Pak nazveme matici $\mathbf{J}(\boldsymbol{\theta})$ *informační maticí* příslušnou k $f(x;\boldsymbol{\theta})$.

O maximálně věrohodných odhadech parametrů rozdělení splňujících podmínky odst. 6.11.3, příp. 6.11.4 pak platí následující věta (důkaz viz [6, 20]).

6.11.5 Věta.

Nechť $\mathbf{X} = (X_1, \ldots, X_n)'$ *je náhodný výběr z rozdělení s hustotou (pravděpodobnostní funkcí)* $f(x;\theta)$ *závislou na parametru* $\theta \in \Omega \subset \mathbf{E^1}$ *splňující podmínky odst.* 6.11.3. *Nechť skutečná hodnota* θ_0 *parametru* θ *je vnitřním bodem* Ω *a nechť v okolí bodu* θ_0 *platí*

$$-\int_M \frac{\partial^2 \ln f(x,\theta)}{\partial\theta^2} f(x,\theta), \mathrm{d}x = \int_M \left[\frac{\ln f(x,\theta)}{\partial\theta}\right]^2 f(x,\theta)\,\mathrm{d}x = J(\theta), \tag{6.11.10}$$

resp.

$$-\sum_M \frac{\partial^2 \ln f(x,\theta)}{\theta^2} f(x,\theta) = \sum_M \left[\frac{\partial \ln f(x,\theta)}{\partial\theta}\right]^2 f(x,\theta) = J(\theta).$$

Pak náhodná veličina

$$n^{1/2}(\hat{\theta} - \theta_0), \tag{6.11.11}$$

kde $\hat{\theta}$ *je kořen věrohodnostní rovnice, konverguje v distribuci (viz [24], odst. 26.2) k normálnímu rozdělení* $N\big(0,\, 1/J(\theta_0)\big)$.

Je-li $\tau(\theta)$ *diferencovatelná funkce na* Ω, *pak náhodná veličina*

$$n^{1/2}\big(\tau(\hat{\theta} - \tau(\theta_0)\big) \tag{6.11.12}$$

konverguje v distribuci k normálnímu rozdělení $N\big(0,\, \big(\tau'(\theta_0)\big)^2/J(\theta_0)\big)$.

Prakticky to znamená, že při velkých hodnotách n lze aproximovat rozdělení odhadu $\hat{\theta}$ parametru θ rozdělením $\mathrm{N}\big(\theta,\, 1/\big(nJ(\theta)\big)\big)$ a rozdělení odhadu $\tau(\hat{\theta})$ pro funkci $\tau(\theta)$ parametru θ rozdělením $\mathrm{N}\big(\tau(\theta),\, (\tau'(\theta))^2/\big(nJ(\theta)\big)\big)$. Stručně říkáme, že maximálně věrohodný odhad $\hat{\theta}$ je *asymptoticky nestranný* a jeho *asymptotické rozdělení je normální s rozptylem* $1/\big(nJ(\theta)\big)$. Protože žádný nestranný odhad parametru θ nemůže mít (za podmínek odst. 6.11.3) rozptyl menší než $1/\big(nJ(\theta)\big)$ (viz [1, 27]), říká se, že maximálně věrohodný odhad je *asymptoticky eficientní.*

Pro maximálně věrohodné odhady vektorového parametru $\boldsymbol{\theta}$ platí obdobná věta, kterou také uvádíme bez důkazu (viz [20]).

6.11.6 Věta

Nechť $\mathbf{X} = (X_1, \ldots, X_n)'$ *je náhodný výběr z rozdělení* $f(x;\boldsymbol{\theta}), \boldsymbol{\theta} \in \Omega \subset \mathbf{E^k}$, *splňujícího podmínky odst. 6.11.4. Nechť skutečná hodnota parametru* $\boldsymbol{\theta}_0 =$

$(\theta_{01},\dots,\theta_{0k})'$ je vnitřním bodem Ω a nechť pro všechna $\boldsymbol{\theta}$ z určitého okolí bodu $\boldsymbol{\theta}_0$ platí

$$-\int_M \frac{\partial^2 \ln f(x;\boldsymbol{\theta})}{\partial\theta_{j_1}\partial\theta_{j_2}} f(x;\boldsymbol{\theta}),\mathrm{d}x = \int_M \frac{\partial \ln f(x;\boldsymbol{\theta})}{\partial\theta_{j_1}} \frac{\partial \ln f(x;\boldsymbol{\theta})}{\partial\theta_{j_2}} f(x;\boldsymbol{\theta}),\mathrm{d}x = \\ = J_{j_1 j_2}(\boldsymbol{\theta}), \quad j_1, j_2 = 1,\dots,k, \quad (6.11.13)$$

přičemž $\mathbf{J}(\boldsymbol{\theta}) = \big(J_{j_1 j_2}(\boldsymbol{\theta})\big)$ je pozitivně definitní matice. Pak náhodný vektor

$$(n^{1/2}(\hat\theta_1 - \theta_{01},\dots,n^{1/2}(\hat\theta_k - \theta_{0k}))', \quad (6.11.14)$$

kde $\hat\theta_1,\dots,\hat\theta_k$ jsou řešení věrohodnostních rovnic, konverguje v distribuci ke k-rozměrnému normálnímu rozdělení s nulovou střední hodnotou a s kovarianční maticí $(\mathbf{J}(\boldsymbol{\theta_0}))^{-1}$, kde $\mathbf{J}(\boldsymbol{\theta})$ je informační matice (6.11.9) příslušná k $f(x;\boldsymbol{\theta})$.

Je-li $\tau(\boldsymbol{\theta})$ funkce parametru $\boldsymbol{\theta}$ mající parciální derivace

$$\tau'_j(\boldsymbol{\theta}) = \frac{\partial\tau(\boldsymbol{\theta})}{\partial\theta_j}$$

pro všechna $j = 1,\dots,k$, pak náhodná veličina

$$n^{1/2}\big(\tau(\hat{\boldsymbol{\theta}}) - \tau(\boldsymbol{\theta}_0)\big), \quad (6.11.15)$$

kde $\hat{\boldsymbol{\theta}} = (\hat\theta_1,\dots\hat\theta_k)'$ konverguje v distribuci k rozdělení

$$\mathrm{N}_k\Big(0, \sum_{j_1=1}^{k}\sum_{j_2=1}^{k} \tau'_{j_1}(\boldsymbol{\theta}_0)\tau'_{j_2}(\boldsymbol{\theta}_0) J^{j_1 j_2}(\boldsymbol{\theta}_0)\Big),$$

kde $J^{j_1 j_2}(\boldsymbol{\theta})$ jsou prvky inverzní matice k matici $\mathbf{J}(\boldsymbol{\theta})$.

Obsah věty 6.11.6 lze shrnout jinými slovy takto: Rozdělení maximálně věrohodných odhadů $\hat\theta_1,\dots,\hat\theta_k$ parametrů $\theta_1,\dots,\theta_k$ je při velkých n přibližně k-rozměrné normální rozdělení se střední hodnotou $\boldsymbol{\theta} = (\theta_1,\dots,\theta_k)'$ (tj. rovnou vektoru skutečných hodnot parametrů) a s kovarianční maticí $\frac{1}{n}(\mathbf{J}(\boldsymbol{\theta}))^{-1}$. Rozdělení odhadu $\tau(\hat{\boldsymbol{\theta}})$ funkce $\tau(\boldsymbol{\theta})$ je při velkých n přibližně

$$\mathrm{N}\Big(\tau(\boldsymbol{\theta}, \frac{1}{n}\sum_{j_1=1}^{k}\sum_{j_2=1}^{k} \tau'_{j_1}(\boldsymbol{\theta})\tau'_{j_2}(\boldsymbol{\theta}) J^{j_1 j_2}(\boldsymbol{\theta})\Big).$$

Speciálně při $\tau(\boldsymbol{\theta}) = \theta$, tj. pro j-tou souřadnici vektoru parametrů, má maximálně věrohodný odhad $\hat\theta_j$ při velkém n přibližně rozdělení $\mathrm{N}\big(\theta_j, \frac{1}{2}J^{jj}(\boldsymbol{\theta})\big)$, $j = 1,\dots,k$.

6.12 Metoda momentů.

Metoda maximální věrohodnosti vyložená v předcházejícím odstavci poskytuje odhady s velmi příznivými vlastnostmi, je však často velmi náročná numericky. Není-li možnost užití výkonných počítačů, je někdy nutno uchýlit se k jednodušším metodám. I při užití metody maximální věrohodnosti je někdy třeba jednodušších metod k získání výchozího odhadu, se kterým by bylo možno zahájit iterační postup řešení rovnic věrohodnosti. Jedna z nejběžnějších takových metod je metoda momentů.

Nechť $\mathbf{X} = (X_1, \ldots, X_n)'$ je náhodný výběr z rozdělení s hustotou pravděpodobnosti (pravděpodobnostní funkcí) $f(x; \boldsymbol{\theta})$ závislou na vektorovém parametru $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$. Při metodě momentů získáme odhady parametrů $\theta_1, \ldots, \theta_k$, řekněme $\theta_1^*, \ldots, \theta_k^*$ tak, že porovnáváme výběrové momenty $M_1', \ldots, M_k'$ (viz odst. 2.2) s odpovídajícími momenty $\mu_1', \ldots, \mu_k'$ rozdělení $f(x; \boldsymbol{\theta})$. Momenty $\mu_1', \ldots, \mu_k'$ jsou totiž funkcemi parametru $\boldsymbol{\theta}$,

$$\mu_r' = E(X^r) = \int_{-\infty}^{\infty} x^r f(x; \boldsymbol{\theta}) \, \mathrm{d}x,$$

příp.

$$\mu_r' = \sum_x x^r f(x; \boldsymbol{\theta}).$$

Jestliže $mu_r' = \mu_r'(\boldsymbol{\theta})$ mají tu vlastnost, že při libovolných hodnotách $M_1', \ldots, M_r'$ má soustava rovnic

$$\mu_r'(\boldsymbol{\theta}) = M_r', \qquad r = 1, \ldots, k,$$

jediné řešení $(\theta_1 *, \ldots, \theta_k^*)'$, kde $\theta_j^* = \theta_j^*(M_1', \ldots, M_k')$ je funkce $M_1', \ldots, M_k'$, pak dává metoda momentů jednoznačně určené odhady parametrů. Rovnice (6.12.1) bývají často mnohem jednodušší než rovnice věrohodnosti (6.9.2).

Kde je to výhodné, lze porovnávat místo obecných momentů momenty centrální; pak se samozřejmě porovnávají centrální momenty výběru s příslušnými centrálními momenty pozorované náhodné veličiny.

V důsledku Chinčinovy věty (viz [24], odst. 25.6 a příkl. 5.3.2 této práce) jsou za dosti obecných podmínek odhady metodou momentů konzistentní. Podle Chinčinovy věty platí totiž, že M_r' konverguje podle pravděpodobnosti k μ_r': jsou-li θ_j^* spojitými funkcemi $M_1', \ldots, M_k'$, konverguje i θ_j^* podle pravděpodobnosti k θ_j, $j = 1, \ldots, k$. Jestliže rozdělení $f(x; \boldsymbol{\theta})$ má konečné momenty až do $2k$-tého řádu včetně, mají podle centrální limitní věty (viz [24], odst.

26.3) výběrové momenty $M'_1, \ldots, M'_k$ asymptoticky normální rozdělení se středními hodnotami $\mu'_1, \ldots, \mu'_k$, a jsou-li $\theta^*_1, \ldots, \theta^*_k$ spojité funkce výběrových momentů, mající parciální derivace podle $M'_1, \ldots, M'_k$, pak i rozdělení odhadů $\theta^*_1, \ldots, \theta^*_k$ bude asymptoticky normální se středními hodnotami $\theta_1, \ldots, \theta_k$ a rozptyly, které lze určit pomocí Taylorova rozvoje.

Závěrem ještě poznamenejme, že u rozdělení s větším počtem parametrů může být použití metody momentů poněku nespolehlivé, zvláště při menších rozsazích výběru. Je to proto, že při větším počtu parametrů je nutno používat výběrových momentů vyšších řádů (až do řádu rovného počtu odhadovaných parametrů) a momenty vyšších řádů jsou velmi citlivé i na malé změny jednotlivých hodnot, takže např. větší nepřesnost jednoho či dvou měření může podstatně ovlivnit výsledek odhadu.

6.13 Příklady.

6.13.1

Nechť $\mathbf{X} = (X_1, \ldots, X_n)'$ je náhodný výběr z rozdělení $\Gamma(m, \delta)$. Pro toto rozdělení je (viz [24], odst. 22.2)

$$\mu'_1(m, \delta) = m\delta, \qquad \mu_2(m, \delta) = m\delta^2.$$

Pro odhad parametrů m a δ metodou momentů máme jednoduchou soustavu rovnic

$$\overline{X} = m\delta, \qquad M_2 = m\delta^2.$$

Odtud odhady metodou momentů jsou

$$m^* = \frac{\overline{X}^2}{M_2}, \qquad \delta^* = \frac{M_2}{\overline{X}}. \tag{6.13.1}$$

Odhad metodou momentů je mnohem jednodušší než maximálně věrohodné odhady (viz příkl. 6.10.2).

6.13.2

Nechť $\mathbf{X} = (X_1, \ldots, X_n)'$ je náhodný výběr z negativního binomického rozdělení s parametry m a π. Pro toto rozdělení platí (viz [24], odst. 14.8)

$$\mu'_1(m, \pi) = \frac{m(1-\pi)}{\pi}, \qquad \mu_2 = \frac{m(1-\pi)}{\pi^2}.$$

Porovnáním těchto momentů s odpovídajícími výběrovými momenty dostaneme odhady pro m a π

$$m^* = \frac{\overline{X}^2}{M_2 - \overline{X}}, \qquad \pi^* = \frac{\overline{X}}{M_2}. \tag{6.13.2}$$

Odhady jsou opět nesrovnatelně jednodušší než maximálně věrohodné odhady (příkl. 6.10.3). Zároveň však příklad ukazuje, že metoda momentů někdy nevede k cíli: ve jmenovateli odhadu m^* je rozdíl $M_2 - \overline{X}$; ačkoliv pro veličinu s negativním binomickým rozdělením je vždy $\mu_2 > \mu_1'$, ve výběru (zvláště při menším rozsahu) se může stát, že $M_2 < \overline{X}$, a potom vyjde odhad m^* záporný, a tedy nepřijatelný. S rostoucím rozsahem výběru n pravděpodobnost takové situace klesá.

6.14 Úlohy.

6.14.1

Nechť $\mathbf{X} = (X_1, \ldots, X_n)'$ je náhodný výběr z tzv. *Fisherova logaritmického rozdělení*; to je rozdělení diskrétního typu s pravděpodobnostní funkcí

$$f(x;\theta) = \frac{\theta^x}{x}\,\frac{1}{-\ln(1-\theta)}, \qquad x = 1, 2, \ldots,$$

kde θ je neznámý parametr z intervalu $(0, 1)$.

a) Zapište $f(x;\theta)$ v exponenciálním tvaru (6.1.1) a najděte postačující statistiku.

b) Ukažte, že pro toto rozdělení se odhad metodou momentů shodu s odhadem metodou maximální věrohodnosti.

c) Vypočtěte informaci pro toto rozdělení a najděte asymptotické rozdělení maximálně věrohodného odhadu pro parametr θ.

[
a) $f(x;\theta) = \exp\Big(x\ln\theta - \ln\big(-\ln(1-\theta)\big) - \ln x\Big)$; postačující statistika je $S(X) = \sum_{i=1}^n X_i$.
b) Odhad metodou momentů se získá řešením rovnice (numerickým) $\overline{X} = \frac{\theta}{(1-\theta)[-\ln(1-\theta)]}$. Rovnice věrohodnosti (6.9.3) je stejná.
c) $J(\theta) = \frac{1}{-\ln(1-\theta)}\,\frac{1}{\theta(1-\theta)^2}\big(1 - \frac{\theta}{-\ln(1-\theta)}\big)$, $\hat\theta$ má při velkých n přibližně rozdělení $\mathrm{N}\big(\theta, \frac{1}{nJ(\theta)}\big)$.
]

6.14.2

Řešte otázky a), b), c) z úlohy 6.14.1 pro případ náhodného výběru z geometrického rozdělení s pravděpodobnostní funkcí

$$f(x;\pi) = \pi(1-\pi)^x, \qquad x = 0, 1, \ldots,$$

kde π je neznámý parametr z intervalu $(0, 1)$.

[a) $f(x;\pi) = \hat{x}p\big(x\ln(1-\pi) + \ln\pi\big)$; postačující statistika je $S(x) = \sum_{i=1}^n X_i$.
b) Odhad metodou momentů je $\pi^* = n\big(\sum_{i=1}^n X_i + n\big)^{-1}$ maximálně věrohodný odhad $\hat{\pi}$ je stejný.
c) $J(\pi) = \sum_{x=0}^{\infty}\big(\frac{-1}{1-\pi}\big)^2\big(x - \frac{1-\pi}{\pi}\big)^2$, $f(x;\pi) = \frac{1}{(1-\pi)^2}$, $\mathrm{var}(X) = \frac{1}{\pi^2(1-\pi)}$. Odhad $\hat{\pi}$ má při velkých n přibližně rozdělení $\mathrm{N}\big(\pi, \pi^2(1-\pi)/n\big)$.]

6.14.3

Nechť $\mathbf{X} = (X_1, \ldots, X_n)'$ je náhodný výběr z rozdělení $\mathrm{N}(\mu, \sigma^2)$ se známým parametrem μ. Stanovte nejlepší nestranný a maximálně věrohodný odhad parametru σ^2, jejich rozdělení a jejich rozptyly.

[Obojí je statistika $\sum_{i=1}^n \frac{(X_i-\mu)^2}{n}$, která má rozdělení $\Gamma(n/2, 2\sigma^2/n)$ a rozptyl $2\sigma^4/n$.]

Kapitola 7

Odhady parametrických funkcí některých důležitých rozdělení

7.1 Normální rozdělení.

Nechť $\mathbf{X} = (X_1, \ldots, X_n)'$ je náhodný výběr z rozdělení $\mathrm{N}(\mu, \sigma^2)$. V příkladě 6.6.4 jsme ukázali, že $S_1(X) = \sum_{i=1}^n X_i$ a $S_2(X) = \sum_{i=1}^n X_i^2$ jsou postačující statistiky pro rozdělení $\mathrm{N}(\mu, \sigma^2)$, takže nejlepší odhady parametrů μ a σ^2 i všech funkcí těchto parametrů najdeme jako funkce uvedených statistik.

Výběrový průměr $\overline{X}$ je nejlepším nestranným odhadem střední hodnoty μ a výběrový rozptyl S^2 je nejlepším nestranným odhadem rozptylu σ^2. Tím je vysvětleno časté používání průměrů $\overline{X}$ a statistiky S^2 při zpracování souborů opakovaných měření za stejných podmínek a vůbec při zpracování výsledků laboratorních pokusů stejné přesnosti.

Střední hodnota statistiky S je rovna (viz odst. 3.2) σ/c_{n-1}, tzn. že nejlepším nestranným odhadem směrodatné odchylky σ je statistika

$$S' = c_{n-1}S = \left(\frac{n-1}{2}\right)^{\frac{1}{2}} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} S. \tag{7.1.1}$$

Častěji se však používá mírně jednostranného odhadu S, který je podstatně jednodušší a jehož jednostrannost $B(\sigma) = E(S - \sigma)$ nemá praktický význam, pokud není rozsah výběru n příliš malý.

Nejlepším nestranným odhadem $100P\%$ kvantilu $x_P = \mu + \sigma u_P$ rozdělení $N(\mu, \sigma^2)$ je statistika

$$\overline{X} + u_p c_{n-1} S \tag{7.1.2}$$

kde u_p je $100P\%$ kvantil rozdělení $N(0,1)$; častěji se však použije odhadu jen přibližně nestranného

$$\overline{X} + u_p S. \tag{7.1.3}$$

Statistika (7.1.2) má rozptyl

$$\mathrm{var}(\overline{X}+u_p c_{n-1} S) = \mathrm{var}(\overline{X})+u_P^2 c_{n-1}^2 \mathrm{var}(S) = \Big(\frac{1}{n}+u_P^2(c_{n-1}^2-1)\Big)\sigma^2. \tag{7.1.4}$$

V [12], str. 89, jsou uvedeny hodnoty 50 analýz pro stanovení koncentrace sodného louhu. Z těchto hodnot se vypočte

$$\overline{x} = 42,27; \qquad s^2 = 2,192.$$

To jsou hodnoty nejlepších nestranných odhadů střední hodnoty μ a rozptylu σ^2 koncentrace. Stanovme ještě hodnotu nejlepšího nestranného odhadu 90% kvantilu $x_{0,9}$. Dosazením do (7.1.2) dostáváme $\overline{x} + u_{0,9}\, c_{49}\, s = 42,27 + 1,281\,552 \cdot 1,005\,115 \cdot 1,48 \doteq 44,177$.

Nejlepším nestranným odhadem rozptylu (7.1.4) je statistika, kterou dostaneme tak, že v (7.1.4) σ^2 nahradíme jeho nejlepším nestranným odhadem S^2. V našem příkladě nabývá hodnoty

$$\Big(\frac{1}{n} + u_{0,9}^2(c_{49}^2 - 1)\Big)s^2 = \Big(\frac{1}{50} + 1,281\,552^2 \cdot 0,010\,256\Big) \cdot 2,192 \doteq 0,081.$$

7.2 Logaritmicko-normální rozdělení.

Nechť $\mathbf{X} = (X_1, \ldots, X_n)'$ je náhodný výběr z logaritmicko-normálního rozdělení $\mathrm{LN}(\mu,\, \sigma^2)$ (viz [24], odst. 19.1). Hustota tohoto rozdělení má tvar (6.1.1) (viz příkl. 6.2.2) a dvojice $\sum_{i=1}^n \ln X_i,\ \sum_{i=1}^n (\ln X_i)^2$ je tedy postačující statistikou pro parametry μ a σ^2. Jelikož $\mathbf{Y} = (Y_1, \ldots, Y_n)' = (\ln X_1, \ldots, \ln X_n)'$ je náhodný výběr z $\mathrm{N}(\mu,\, \sigma^2)$, je ihned vidět (s užitím výsledků odst. 7.1), že

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^n \ln X_i, \quad S_Y^2 = \frac{1}{n-1}\sum_{i=1}^n (\ln X_i - \overline{Y})^2 \tag{7.2.1}$$

jsou nejlepší nestranné odhady parametrů μ a σ^2. Rozdělení těchto odhadů jso stejná jako rozdělení nejlepších nestranných odhadů pro parametry rozdělení $N(\mu, \sigma^2)$, tj. $\overline{Y}$ má rozdělení $N(\mu, \sigma^2/n)$ a $(n-1)S_Y^2/\sigma^2$ má rozdělení $\chi^2(n-1)$.

Jestliže pro daný účel není třeba používat nejlepších nestranných odhadů (stačí menší přesnost) nebo nechceme počítat logaritmy všech výsledků, lze použít jednodušších odhadů založených na metodě momentů. Veličina X s rozdělením $\mathrm{LN}(\mu,\,\sigma^2)$ má první dva momenty rovny (viz [24], odst. 19.3)

$$\mu_1' = E(X) = e^{\mu+\frac{\sigma^2}{2}}, \qquad \mu_2' = E(X^2) = e^{2\mu+2\sigma^2}.$$

Odtud rovnice pro odhad metodou momentů jsou

$$\mu^* + \frac{\sigma^{2*}}{2} = \ln \overline{X}, \qquad 2\mu^* + 2\sigma^{2*} = \ln M_2'$$

a jejich řešení

$$\sigma^{2*} = \ln M_2' - 2\ln \overline{X} = \ln \frac{M_2'}{\overline{X}^2}, \tag{7.2.2}$$

$$\mu^* = \ln \frac{\overline{X}^2}{\sqrt{M_2'}}$$

U logaritmicko-normálního rozdělení je často třeba odhadovat funkce parametrů μ a σ^2 tvaru

$$\tau(\mu,\sigma^2) = e^{a\mu+b\sigma^2}, \tag{7.2.3}$$

kde a a b jsou daná čísla; toho tvaru jsou např. momenty rozdělení $\mathrm{LN}(\mu,\,\sigma^2)$, medián a modus tohoto rozdělení (podrobněji o významu takových funkcí viz [24], čl. 19). Nejlepším nestranným odhadem funkce (7.2.3) je

$$T(Y,S_Y^2) = e^{a\overline{Y}}\varphi\left(\left(b-\frac{a^2}{2n}\right)S_Y^2\right), \tag{7.2.4}$$

kde

$$\varphi(v) = 1 + \sum_{k=1}^{\infty} \frac{v^k}{k!}\frac{(n-1)^k}{(n-1)(n+1)\dots(n+2k-3)}. \tag{7.2.5}$$

Speciálně nejlepším nestranným odhadem funkce

$$\tau(\mu,\sigma^2) = E(X) = e^{\mu+\frac{\sigma^2}{2}}$$

je statistika

$$\mathrm{e}^{\overline{Y}}\varphi\left(\frac{n-1}{2n}S_Y^2\right) \tag{7.2.6}$$

nejlepším nestranným odhadem funkce

$$\tau(\mu, \sigma^2) = \mathrm{var}(X) = e^{2\mu+2\sigma^2} - e^{2\mu+\sigma^2}$$

je statistika

$$e^{2\overline{Y}}\Big(\varphi\Big(\frac{2(n-1)}{n}S_Y^2\Big) - \varphi\Big(\frac{n-2}{n}S_Y^2\Big)\Big) \tag{7.2.7}$$

a nejlepším nestranným odhadem funkce

$$\tau(\mu, \sigma^2) = e^{\mu},$$

tj. mediánu rozdělení, je statistika

$$\mathrm{e}^{\overline{Y}}\varphi\big(-\frac{1}{2n}S_Y^2\big). \tag{7.2.8}$$

Předpokládejme, že údaje příkl. 2.3 představují náhodný výběr z rozdělení LN(μ, σ^2). Porovnejme hodnoty odhadů (7.2.1) a (7.2.2) parametrů μ a σ^2:

$$\begin{aligned} \overline{y} &= 15,609; \qquad \mu^* = \ln \frac{102\ 905^2 \cdot 10^3}{(8 \cdot 3\ 047\ 522\ 337)^{\frac{1}{2}}} \doteq 15,953; \\ s_Y^2 &= 2,019; \qquad \sigma^{2*} = \ln \frac{8 \cdot 3\,047\,522\,337}{102\,905^2} \doteq 0,834. \end{aligned}$$

Je vidět, že (v důsledku malého počtu pozorování a velké variability údajů) se odhady parametru σ^2 od sebe dosti liší.

Stanovme ještě hodnotu nejlepšího nestranného odhadu střední hodnoty $E(X)$ rozdělení. Z (7.2.6) dostáváme

$$e^{15,609}\varphi\left(\frac{113,037\,068}{8 \cdot 16}\right) \doteq 13\,526,64 \cdot 10^3.$$

7.3 Exponenciální rozdělení.

Nechť $\mathbf{X} = (X_1, \ldots, X_n)'$ je náhodný výběr z exponenciálního rozdělení (viz [24], odst. 20.1) s hustotou

$$f(x; \delta) = \frac{1}{\delta}e^{-\frac{x}{\delta}}, \qquad x > 0. \tag{7.3.1}$$

Toto rozdělení je jednoparametrické tvaru (6.1.1), minimální postačující statistikou je součet všech pozorování

$$S(X) = \sum_{i=1}^{n} X_i. \tag{7.3.2}$$

Statistika S má rozdělení $\Gamma(n, \delta)$ (viz [24], příkl. 22.4.2); tudíž

$$E(S^r) = \frac{\Gamma(n+r)}{\Gamma(n)}\delta^r, \qquad r > -n. \tag{7.3.3}$$

Odtud nejlepším nestranným odhadem funkce $\tau_r(\delta) = \delta^r$ je

$$T_r = \frac{\Gamma(n)}{\Gamma(n+r)}\Big(\sum_{i=1}^{n} X_i\Big)^r, \qquad k > -n. \tag{7.3.4}$$

Speciálně pro $r = 1$ dostáváme jako nejlepší nestranný odhad parametru δ (tj. střední hodnoty rozdělení) průměr všech pozorování,

$$T_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}, \tag{7.3.5}$$

a pro veličinu $\tau_{-1} = 1/\delta = \lambda$ (tj. *intenzitu poruch,* je-li X doba do poruchy nějakého výrobku)

$$T_{-1} = \frac{n-1}{\sum_{i=1}^{n} X_i} = \frac{n-1}{n\overline{X}}. \tag{7.3.6}$$

Rozptyl těchto odhadů je

$$\operatorname{var}(T_1) = \frac{\delta^2}{n}, \quad \operatorname{var}(T_{-1}) = \frac{1}{(n-2)\delta^2}. \tag{7.3.7}$$

Odhady těchto rozptylů nebo příslušných směrodatných odchylek snadno stanovíme užitím (7.3.4); např. nejlepší nestranný odhad funkce $\sqrt{\operatorname{var}(T_{-1})} = \Big(\delta\sqrt{n-2}\Big)^{-1}$ je

$$\frac{1}{\sqrt{n-2}}\,\frac{n-1}{\sum_{i=1}^{n} X_i},$$

nejlepší nestranný odhad funkce $\operatorname{var}(T_1) = \delta^2/n$ je

$$\frac{\big(\sum_{i=1}^{n} X_i\big)^2}{n^2(n+1)}$$

atd.

Při aplikacích v otázkách spolehlivosti se často žádá odhad tzv. *funkce spolehlivosti* (též tzv. *funkce přežití*) $R(t)$, tj. pravděpodobnosti, že zařízení bude pracovat bez poruch po dobu aspoň rovnou t. Má-li doba bezporuchového chodu rozdělení $\mathrm{E}(0,\delta)$, je

$$R(t) = P(X \geq t) = e^{-\frac{t}{\delta}}, \qquad t > 0. \tag{7.3.8}$$

Nejlepším nestranným odhadem této funkce je

$$\begin{aligned} T &= \Big(1 - \frac{t}{\sum_{i=1}^n X_i}\Big)^{n-1} \quad \text{pro} \quad \sum_{i=1}^n X_i \geq t, \\ &= 0 \qquad\qquad\qquad\quad \text{pro} \quad \sum_{i=1}^n X_i < t; \end{aligned} \tag{7.3.9}$$

získá se postupem popsaným v odst. 6.5, vyjde-li se od jednoduchého odhadu

$$\begin{aligned} T_0(\mathbf{X}) &= 1 \quad \text{pro} \quad X_1 \geq t, \\ &= 0 \quad \text{pro} \quad X_1 < t, \end{aligned}$$

a stanoví se podmíněná pravděpodobnost

$$P(X_1 \geq t \,\Big|\, \sum_{i=1}^n X_i = s) = T(s);$$

odhad (7.3.9) je $T = T\big(\sum_{i=1}^n X_i\big)$.

Při výběrech většího rozsahu n lze nahradit nejlepší nestranné odhady všech funkcí parametru δ maximálně věrohodnými odhady. Maximálně věrohodný odhad parametru δ se shoduje s nejlepším nestranným odhadem, tj. $\hat{\delta} = \overline{X}$, a maximálně věrohodný odhad jakékoliv funkce $\tau(\delta)$ parametru δ je $\hat{\tau} = \tau(\overline{X})$; tedy např. maximálně věrohodný odhad intenzity poruch je $\hat{\lambda} = 1/\overline{X}$ a maximálně věrohodný odhad funkce spolehlivosti $R(t)$ je $\hat{R}(t) = e^{-t/\overline{X}}$.

Při sledování doby do poruchy (v hodinách) určitého zařízení se získalo následujících osm údajů: 48, 16, 75, 29, 96, 67, 89, 22. Předpokládejme, že se jedná o náhodný výběr z rozdělení $\mathrm{E}(0,\delta)$. Stanovme nejlepší nestranný a maximálně věrohodný odhad funkce $R(t)$ pro $t = 100$.

Protože $\sum_{i=1}^{n} x_i = 442$, dostáváme ze (7.3.9)

$$T = \left(1 - \frac{100}{442}\right)^7 \doteq 0,166.$$

Dále

$$\hat{R}(100) = e^{-\frac{800}{442}} \doteq 0,164.$$

7.4 Weibullovo rozdělení.

Nechť $\mathbf{X} = (X_1, \ldots, X_n)'$ je náhodný výběr z Weibullova rozdělení (viz [24], odst. 2.11) s hustotou

$$f(x; c, \delta) = \frac{cx^{c-1}}{\delta^c} \mathrm{e}^{-(x/\delta)^c}, \quad x > 0, \tag{7.4.1}$$

kde c a δ jsou neznámé kladné parametry. Hustota (7.4.1) zjevně nemá tvar (6.1.1), nepatří tedy do exponenciální třídy hustot a odhady parametrů c a δ i odhady jejich funkcí je nutno založit na metodě momentů (odst. 6.12) nebo na metodě maximální věrohodnosti (odst. 6.7), příp. na grafických metodách (odst. 13.9) či na uspořádaném výběru (odst. 21.2) nebo na kombinaci těchto metod.

Metoda momentů vede k poměrně jednoduchému řešení: Střední hodnota a rozptyl rozdělení (7.4.1) jsou

$$\mu_1' = \Gamma\Big(1 + \frac{1}{c}\Big)\delta, \quad \mu_2 = \left(\Gamma\Big(1 + \frac{2}{c}\Big) - \Gamma^2\Big(1 + \frac{1}{c}\Big)\right)\delta^2 \tag{7.4.2}$$

(viz [24], odst. 21.3). Z rovnic

$$\mu_1' = \overline{X}, \quad \mu_2 = M_2 \tag{7.4.3}$$

vyloučíme δ a dostaneme pro c^* rovnici

$$\frac{M_2}{\overline{X}^2} = \frac{\Gamma\big(1 + \frac{2}{c^*}\big)}{\Gamma^2\big(1 + \frac{1}{c^*}\big)} - 1, \tag{7.4.4}$$

kterou je třeba řešit numericky. Hodnoty pravé strany jako funkce odhadu c^* uvádí tab. 7.1

Odhad parametru δ pak je

$$\delta^* = \frac{X}{\Gamma\left(1+\frac{1}{c^*}\right)}. \tag{7.4.5}$$

Rovnice pro výpočet maximálně věrohodných odhadů (podle odst. 6.9) po úpravách vedou k jedné rovnici pro $\hat{c}$, kterou je třeba řešit numericky,

$$\left[\frac{\sum_{i=1}^n x_i^{\hat{c}} \ln x_i}{\sum_{i=1}^n x_i^{\hat{c}}} - \frac{\sum_{i=1}^n \ln x_i}{n}\right]^{-1} = \hat{c}, \tag{7.4.6}$$

a po výpočtu $\hat{c}$ je odhad parametru δ roven

$$\hat{\delta} = \left(\frac{1}{n}\sum_{i=1}^n x_i^{\hat{c}}\right)^{\frac{1}{\hat{c}}}. \tag{7.4.7}$$

Jako první aproximace pro $\hat{c}$ lze použít přibližného řešení (7.4.4) nebo hodnoty získané grafickou analýzou podle odst. 13.9. Maximálně věrohodný odhad pravděpodobnosti

$$R(t) = P(X \geq t) = e^{-(t/\delta)^c}, \quad t > 0, \tag{7.4.8}$$

je roven

$$\hat{R}(t) = \exp\left(-\frac{nt^{\hat{c}}}{\sum_{i=1}^n X_i^{\hat{c}}}\right). \tag{7.4.9}$$

Bez užití počítače s možností programování je výpočet maximálně věrohodných odhadů velice pracný (po každé aproximaci k $\hat{c}$ je nutno vypočítat $\sum_{i=1}^n x_i^{\hat{c}}$, $\sum_{i=1}^n x_i^{\hat{c}} \ln x_i$), zato mají maximálně věrohodné odhady ohromnou výhodu, že veličiny $\left(\hat{c}/c\right)$, $\left(\hat{\delta}/\delta\right)^{\hat{c}}$ mají rozdělení nezávislé na skutečných hodnotách parametrů a pro různé hodnoty n jsou známy a tabelovány [35] kvantily těchto rozdělení, takže lze posoudit i při malých hodnotách n přesnost odhadů. Při užití odhadů metodou momentů je známo je asymptotické rozdělení.

Při zkouškách životnosti určitého elektronického prvku byly zjištěny následující doby života (ve dnech): 4, 13, 26, 36, 51, 75, 100, 111, 162, 174 (viz [4]). Z údajů vypočteme

$$\overline{X} = \frac{752}{10} = 75,2; \qquad M_2 = \frac{10 \cdot 89\ 224 - 752^2}{100} = 3\,267,36.$$

Podíl $M_2/\overline{X}^2 = 0,577\,78$ a interpolací v tab. 7.1 nalezneme $c^* = 1,33$; odtud $\delta^* = 75,2/\Gamma(1+1/1,33) \doteq 81,78$.

Použijeme-li této hodnoty c^* jako první aproximace v rovnici (7.4.6), dostaneme řešení $\hat{c} = 1,19$ a dosazením této hodnoty do (7.4.7) dostáváme $\hat{\delta} = 79,5$.

7.5 Úlohy.

7.5.1

Hodnoty 50 analýz uvažované v odst. 7.1 byly v pořadí, jak byly získány, rozděleny postupně do 10 skupin po 5 hodnotách a z každé skupiny byla vypočtena hodnota rozpětí R. Dostali jsme následující hodnoty $R : 3,8;\ 3,5;\ 3,8;\ 3,9;\ 3,9;\ 2,2;\ 0,6;\ 3,5;\ 6,2;\ 4,4$.

Nalezněte nestranný odhad směrodatné odchylky σ koncentrace založený na průměrném rozpětí a porovnejte tento odhad s nejlepším nestranným odhadem σ. (Využijte příkl. 4.7.2 a tabulek [23].)

$$\left[a_5\overline{R} = 1,539; \quad c_{49}S = 1,488\right].$$

7.5.2

Pro údaje příkladu z odst. 7.3 nalezněte hodnotu nejlepšího nestranného a maximálně věrohodného odhadu intenzity poruch $\lambda = 1/\delta$. Dále nalezněte nejlepší nestranné odhady rozptylů těchto dvou odhadů.

$$\left[\begin{array}{ll} T_{-1} = 0,016; & \dfrac{n-1}{\left(\sum_{i=1}^n X_i\right)^2} = 36\cdot 10^{-6}; \\ \hat{\lambda} = 0,018; & \dfrac{n^2}{(n-1)\left(\sum_{i=1}^n X_i\right)^2} = 47\cdot 10^{-6}. \end{array}\right]$$

c	$\Gamma\left(1+\frac{1}{c}\right)$	$\frac{\Gamma\left(1+\frac{2}{c}\right)}{\Gamma\left(1+\frac{1}{c}\right)}-1$
0,4	3,323 35	9,864 97
0,5	2,000 00	5,000 00
0,6	1,504 58	3,090 79
0,7	1,265 82	2,138 68
0,8	1,133 00	1,588 89
0,9	1,052 18	1,238 83
1,0	1,000 00	1,000 00
1,1	0,964 91	0,828 49
1,2	0,940 66	0,700 40
1,3	0,923 58	0,601 74
1,4	0,911 42	0,523 82
1,5	0,902 74	0,461 00
1,6	0,896 57	0,409 48
1,7	0,892 24	0,366 61
1,8	0,889 29	0,330 48
1,9	0,887 36	0,299 70
2,0	0,886 23	0,273 24
2,1	0,885 69	0,250 29
2,2	0,885 62	0,230 24
2,3	0,885 92	0,212 60
2,4	0,886 48	0,196 99
2,5	0,887 26	0,183 10
2,6	0,888 21	0,170 69
2,7	0,889 28	0,159 54
2,8	0,890 45	0,149 48
2,9	0,891 69	0,140 37
3,0	0,892 98	0,132 09

Tab. 7.1: Hodnoty pro určení parametrů Weibullova rozdělení metodou momentu.

Kapitola 8

Intervalové odhady

8.1 Intervaly spolehlivosti.

V odstavci 5.1 jsme zdůraznili, že odhad parametru (nebo funkce parametrů) je funkce realizací náhodných veličin, tedy sám je realizací náhodné veličiny a prakticky vždy se bude lišit více či méně od skutečné hodnoty odhadovaného parametru. Proto se často výsledky experimentu shrnují pomocí dvojic statistik, řekněme T_d a $T_h, T_d = T_d(\mathbf{X}), T_h = T_h(\mathbf{X})$, sestrojených tak, aby bylo možno s rozumným stupněm důvěry či s rozumnou zárukou očekávat, že skutečná hodnota parametru leží v intervalu (T_d, T_h). Nejčastěji se požadavek „rozumné záruky, že skutečná hodnota parametru (funkce parametru) bude pokryta intervalem $T_d, T_h)$“ formuluje takto: Zvolí se číslo $1 - \alpha$ blízké 1, např. $1 - \alpha = 0,95$ nebo $0,99$ a žádá se, aby při jakékoliv skutečné hodnotě parametru platilo

$$P\big(T_h \leq \tau(\theta)\big) \leq \frac{\alpha}{2}, \qquad P\big(T_d \geq \tau(\theta)\big) \leq \frac{\alpha}{2}, \tag{8.1.1}$$

tj. aby pro všechna $\theta \in \Omega$ platilo

$$P(T_d < \tau(\theta) < T_h) \geq 1 - \alpha. \tag{8.1.2}$$

Dvojice statistik (T_d, T_h) splňující (8.1.2) se nazývá *interval spolehlivosti* (také *konfidenční interval*) pro funkci $\tau(\theta)$ parametru θ. Číslo $1 - \alpha$, tj. $\inf_{\theta\in\Omega} P(T_d < \tau(\theta) < T_h)$, se nazývá *koeficient spolehlivosti* (také *konfidenční koeficient).*

Někdy se mluví např. o devadesátipětiprocentním konfidenčním intervalu (je-li $1 - \alpha = 0,95$) či devadesátidevítiprocentním konfidenčním intervalu (je-li $1 - \alpha = 0,99$). Riziko, že skutečná hodnota funkce $\tau(\theta)$ bude

přeceněna, tj. pravděpodobnost, že dolní hranice intervalu spolehlivosti T_d padne nad správnou hodnotu $\tau(\theta)$, je při splnění podmínek (8.1.1) rovno $\alpha/2$ a stejnou hodnotu má i riziko podcenění skutečné hodnoty $\tau(\theta)$, tj. pravděpodobnost, že bude získán náhodný výběr $\mathbf{X}$, pro který $T_h(\mathbf{X}) \leq \tau(\theta)$. Z dlouhé řady výběrů přibližně $100(1-\alpha)\%$ bude takových, že interval (T_d, T_h) pokryje skutečnou hodnotu funkce $\tau(\theta)$. To znamená, že např. při soustavném používání metody s $1-\alpha = 0,99$ jen asi v jednom případě ze sta dostaneme interval, který neobsahuje správnou hodnotu parametru.

Dolní hranice intervalu spolehlivosti se často nazývá *dolní konfidenční mez* a horní hranice *horní konfidenční mez*. Často se dolní konfidenční mez označuje stejným písmenem jako odhadovaný parametr a připojuje se pruh pod symbolem a horní konfidenční mez stejným písmenem s pruhem nad symbolem, tedy např. pro parametr μ v $\mathrm{N}(\mu, \sigma^2)$ se píše $\underline{\mu}, \overline{\mu}$.

Někdy se stává, že důležitý je jen horní odhad příslušného parametru, tj. že se hledá jen nejvyšší hodnota parametru, se kterou je třeba při daných výsledcích experimentu počítat, nebo naopak jen dolní odhad, tj. nejmenší hodnota, kterou lze s rozumným stupněm důvěry očekávat. Pak se používá jen horní, příp. dolní konfidenční meze a požadavek (8.1.2) se nahrazuje požadavkem

$$P\big(T_d < \tau(\theta)\big) \geq 1-\alpha \quad \text{resp.} \quad P\big(T_h > \tau(\theta)\big) \geq 1-\alpha. \tag{8.1.3}$$

Lze také říci, že v takových případech používáme intervalů tvaru

$$(\inf_{\theta\in\Omega} \tau(\theta),\ T_h), \quad \text{resp.} \quad (T_d,\ \sup_{\theta\in\Omega} \tau(\theta)),$$

takže užijeme jen jedné z rovnic (8.1.1) a číslo $\alpha/2$ nahradíme číslem α.

8.2 Obecný postup při kosntrukci intervalu spolehlivosti v případě jednoho neznámého parametru.

Pro výklad obecného postupu při konstrukci intervalu spolehlivosti uvažujme nejprve jednoduchý případ, kdy rozdělení závisí na jediném neznámém parametru θ. Nechť tedy $\mathbf{X} = (X_1, \ldots, X_n)'$ je náhodný výběr z rozdělení s hustotou (příp. pravděpodobnostní funkcí) $f(x;\theta)$ závislou na jediném reálném parametru θ. Budiž $T = T(\mathbf{X})$ nějaká statistika, na které lze založit odhad

parametru θ. Předpokládejme, že existuje funkce $h(t,\theta)$ dvou proměnných splňující tyto podmínky:

$$\text{Rozdělení náhodné veličiny } h(T,\theta) \text{ je nezávislé na } \theta; \tag{8.2.1}$$

$$\begin{gathered} h(t,\theta) \text{ je při každém pevném } t \text{ klesající funkcí } \theta \\ \text{a při každém pevné } \theta \textit{ neklesající funkcí } t. \end{gathered} \tag{8.2.2}$$

Pro zvolené číslo α z intervalu $0 < \alpha < 1$ označme

$$h_1 = \sup\Big\{h \,|\, P\big(h(T,\theta) < h\big) \le \frac{\alpha}{2}\Big\}, \tag{8.2.3}$$

$$h_2 = \inf\Big\{h \,|\, P\big(h(T,\theta) > h\big) \le \frac{\alpha}{2}\Big\}. \tag{8.2.4}$$

Z podmínky (8.2.1) plyne, že čísla h_1, h_2 nezávisí na hodnotě θ. Dále definujme pro libovolné t funkce $\underline{\theta}(t)$ a $\overline{\theta}(t)$ rovnicemi

$$h\big(t,\underline{\theta}(t)\big) = h_2, \tag{8.2.5}$$

$$h\big(t,\overline{\theta}(t)\big) = h_1. \tag{8.2.6}$$

Z podmínky (8.2.2) plyne, že tyto funkce jsou jednoznačně určeny. Z definice čísel h_1, h_2 a z definice funkcí $\underline{\theta}(T), \overline{\theta}(T)$ platí

$$P\big(\underline{\theta}(T) \ge \theta\big) \le \frac{\alpha}{2} \quad \text{pro libovolné } \theta, \tag{8.2.7}$$

$$P\big(\overline{\theta}(T) \le \theta\big) \le \frac{\alpha}{2} \quad \text{pro libovolné } \theta, \tag{8.2.8}$$

čili že pro všechna θ platí

$$P\big(\underline{\theta}(T) < \theta < \overline{\theta}(T)\big) \ge 1 - \alpha. \tag{8.2.9}$$

To znamená, že $\big(\,\underline{\theta}(T),\, \overline{\theta}(T)\,\big)$ je $100(1-\alpha)$-procentní interval spolehlivosti pro parametr θ.

Důkaz platnosti vztahů (8.2.7) a (8.2.8) je jednoduchý: Z podmínky (8.2.2) plyne, že pro libovolné dané θ platí

$$\begin{aligned} \overline{\theta}(t) \le \theta &\iff h(t,\theta) \le h_1, \\ \underline{\theta}(t) \ge \theta &\iff h(t,\theta) \ge h_2. \end{aligned}$$

Odtud

$$\begin{aligned} P\big(\overline{\theta}(T) \le \theta\big) &= P\big(h(T,\theta) \le h_1\big), \\ P\big(\underline{\theta}(T) \ge \theta\big) &= P\big(h(T,\theta) \ge h_2\big); \end{aligned}$$

pravděpodobnosti na pravé straně jsou nejvýše rovny $\alpha/2$ podle definice čísel h_1, h_2.

Jestliže náhodná veličina $h(T,\theta)$ má při každém θ rozdělení spojitého typu s ryze rostoucí distribuční funkcí, pak ve vztazích (8.2.7) až (8.2.9) platí znamení rovnosti a pravděpodobnost, že konfidenční interval $(\,\underline{\theta}(T),\, \overline{\theta}(T)\,)$ pokryje skutečnou hodnotu neznámého parametru, je právě rovna zvolenému koeficientu $1-\alpha$. Má-li hodnotu neznámého parametru, je právě rovna zvolenému koeficientu $1-\alpha$. Má-li $h(T,\theta)$ rozdělení diskrétního typu, pak skutečná hodnota pravděpodobnosti pokrytí neznámého parametru intervalem $(\,\underline{\theta}(T),\, \overline{\theta}(T)\,)$ je závislá na skutečné hodnotě parametru, je však vždy větší nebo rovna zvolenému koeficientu spolehlivosti.

Užití uvedeného postupu je ilustrováno v příkl. 8.4.1.

Za funkci $h(t,\theta)$ je někdy výhodné volit funkci

$$F(t;\theta) = P(T \le t), \tag{8.2.10}$$

která automaticky splňuje podmínku, že je při každém θ neklesající funkcí t. Konstrukce intervalu spolehlivost s užitím této funkce je ilustrována příkl. 8.4.3 a 8.4.4.

8.3 Několik neznámých parametrů.

Jestliže rozdělení závisí na vektorovém parametru $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$ a žádá se interval spolehlivosti pro některou složku tohoto vektoru nebo interval spolehlivosti pro některou funkci vektoru $\boldsymbol{\theta}$, řekněme $\tau(\boldsymbol{\theta})$, bývá často nemožné najít funkci závislou jen na odhadované složce vektoru $\boldsymbol{\theta}$ (resp. odhadované funkci $\tau(\boldsymbol{\theta})$) a na jedné reálné statistice $T(\mathbf{X})$. Někdy se v takových situacích podaří najít náhodnou veličinu $h(T,\tau; S_1(\mathbf{X}), \ldots, S_r(\mathbf{X}))$ jako funkci bodového odhadu T pro $\tau(\boldsymbol{\theta})$, dalších statistik $S_1(\mathbf{X}), \ldots, S_r(\mathbf{X})$ a odhadované funkce $\tau(\theta)$ tak, aby její rozdělení nezáviselo na „přebytečných“ složkách parametru $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$. Příkladem takové funkce je (8.4.9).

Jindy je možné najít statistiky $S_1(\mathbf{X}), \ldots, S_{k-1}(\mathbf{X})$ a funkci $h(T(\mathbf{X}),\tau)$ tak, že aspoň podmíněné rozdělení náhodné veličiny $h(T(\mathbf{X},\tau)$ při daných

hodnotách statistik $S_1, \ldots, S_{k-1}$ nezávisí na vektoru parametrů. Obecnou teorii zde nelze vykládat, je obsažena např. v [22].

Někdy bývá užitečné konstruovat simultánní (sdružené) intervaly spolehlivosti pro několik parametrů (nebo funkcí parametrů) současně, popř. i oblasti jiného tvaru než k-rozměrný interval, tzn. sestrojit funkce $\underline{\tau}_1(\mathbf{x}), \ldots, \underline{\tau}_k(\mathbf{x})$ a $\overline{\tau}_1(\mathbf{x}), \ldots, \overline{\tau}_k(\mathbf{x})$ tak, aby

$$P\Big(\underline{\tau}_j(\mathbf{X}) < \tau_j(\theta_1, \ldots, \theta_k) < \overline{\tau}_j(\mathbf{X}), \quad j = 1, \ldots, k\Big) \geq 1 - \alpha \tag{8.3.1}$$

nebo obecněji náhodné oblasti $M(\mathbf{x})$ v k-rozměrném prostoru tak, aby

$$P\Big(M(\mathbf{X}) \supset (\tau_1(\boldsymbol{\theta}), \ldots, \tau_k(\boldsymbol{\theta}))\Big) \geq 1 - \alpha. \tag{8.3.2}$$

Příslušnou teorii lze najít v [1].

8.4 Příklady.

Jako příklady užití obecných postupů z odst. 8.2 a 8.3 uvedeme intervaly spolehlivosti pro parametry a některé funkce parametrů rozdělení, jejichž bodovými odhady jsme se zabývali v čl. 7.

8.4.1 Exponenciální rozdělení.

Nechť $\mathbf{X} = (X_1, \ldots, X_n)'$ je náhodný výběr z exponenciálního rozdělení $\mathrm{E}(0, \delta)$. To je rozdělení s jediným parametrem δ, pro který máme jednorozměrnou minimální postačující statistiku $T = \sum_{i=1}^{n} X_i$. Lze tedy postupovat podle odst. 8.2 Náhodná veličina

$$h(T, \delta) = \frac{2T}{\delta} = \frac{2}{\delta} \sum_{i=1}^{n} X_i \tag{8.4.1}$$

má rozdělení $\chi^2(2n)$, jak se lze snadno přesvědčit užitím výsledku příkl. 22.4.2 z [24] a jednoduchou transformací. Funkce (8.4.1) zřejmě splňuje podmínky (8.2.1) a (8.2.2); pro libovolné δ platí

$$P\big(\, \chi^2_{\alpha/2}(2n) < \frac{2}{\delta} \sum_{i=1}^{n} X_i < \chi^2_{1-\alpha/2}(2n) \,\big) = 1 - \alpha.$$

Rovnice (8.2.5) a (8.2.6) přejdou v

$$\frac{2t}{\underline{\delta}} = \chi^2_{1-\alpha/2}(2n), \qquad \frac{2t}{\overline{\delta}} = \chi^2_{\alpha/2}(2n), \tag{8.4.2}$$

a interval spolehlivosti pro δ je tedy

$$\left(\frac{2\sum_{i=1}^n X_i}{\chi^2_{1-\alpha/2}},\ \frac{2\sum_{i=1}^n X_i}{\chi^2_{\alpha/2}} \right). \tag{8.4.3}$$

Pravděpodobnost $R(t) = P(X \geq t)$, kde t je dané kladné číslo, je dána výrazem (7.3.8) a je rostoucí funkcí parametru δ. Platí tedy

$$\exp\big(-\frac{t}{\underline{\delta}}\big) < \exp\big(-\frac{t}{\delta}\big) < \exp\big(-\frac{t}{\overline{\delta}}\big);$$

odtud plyne, že konfidenční interval pro tzv. funkci spolehlivosti při exponenciálním rozdělení doby života je

$$\left(\exp\left[-\frac{t\,\chi^2_{1-\alpha/2}(2n)}{2\sum_{i=1}^n X_i} \right],\ \exp\left[-\frac{t\,\chi^2_{\alpha/2}(2n)}{2\sum_{i=1}^n X_i} \right] \right). \tag{8.4.4}$$

Uvažujme údaje odst. 7.3 a stanovme 95% konfidenční interval pro $R(50)$. Z tabulek [24] nalezneme $\chi^2_{0,95}(16) = 26,296$. Pak

$$\frac{50 \cdot 26,296}{2 \cdot 442} \doteq 1,487; \qquad e^{-1,487} \doteq 0,226,$$

takže dostáváme jednostranný konfidenční interval $(0,226;\ 1)$ pro $R(50)$.

8.4.2 Normální rozdělení.

Nechť $\mathbf{X} = (X_1, \ldots, X_n)'$ je náhodný výběr z rozdělení $\mathrm{N}(\mu,\, \sigma^2)$. To je rozdělení s dvourozměrným parametrem; minimální postačující statistika je dvourozměrná, pro libovolnou funkci dvojice $(\mu,\, \sigma^2)$ nelze užít jednoduchého postupu z odst. 8.2. Je to však možné pro $\tau(\mu,\, \sigma^2) = \sigma^2$, tj. pro odhad rozptylu normálního rozdělení. Podle odst. 3.2 má náhodná veličina

$$h(S^2, \sigma^2) = \frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma^2} \tag{8.4.5}$$

rozdělení $\chi^2(n-1)$. Rovnice (8.2.5) a (8.2.6) přejdou v

$$\frac{(n-1)s^2}{\underline{\sigma}^2} = \chi^2_{1-\alpha/2}(n-1), \quad \frac{(n-1)s^2}{\overline{\sigma}^2} = \chi^2_{\alpha/2}(n-1), \tag{8.4.6}$$

odkud interval spolehlivosti pro σ^2 je

$$\left(\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}, \frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}\right). \tag{8.4.7}$$

Interval spolehlivosti pro směrodatnou odchylku σ je

$$\left(\frac{S\sqrt{n-1}}{\sqrt{\chi^2_{1-\alpha/2}(n-1)}}, \frac{S\sqrt{n-1}}{\sqrt{\chi^2_{\alpha/2}(n-1)}}\right). \tag{8.4.8}$$

K sestrojení intervalu spolehlivosti pro střední hodnotu μ je třeba použít funkce bodového odhadu $\overline{X}$ parametru μ a druhé statistiky S^2; podle (3.3.7) má funkce

$$h(\overline{X}, \mu; S^2) = \frac{\overline{X} - \mu}{S}\sqrt{n} \tag{8.4.9}$$

rozdělení $\mathrm{t}(n-1)$, tedy nezávisí na (μ, σ^2) a při každém $(\overline{X}, S^2)$ je $h(\overline{X}, \mu; S^2)$ klesající funkce μ. Rovnice (8.2.5) a (8.2.6) přejdou v

$$\frac{\overline{x} - \underline{\mu}}{s}\sqrt{n} = t_{1-\alpha/2}(n-1), \quad \frac{\overline{x} - \overline{\mu}}{s}\sqrt{n} = t_{\alpha/2}(n-1) = -t_{1-\alpha/2}(n-1), \tag{8.4.10}$$

odkud interval spolehlivosti pro μ je

$$\left(\overline{X} - t_{1-\alpha/2}(n-1)\frac{S}{\sqrt{n}}, \ \overline{X} + t_{1-\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right). \tag{8.4.11}$$

Stanovme 95% intervaly spolehlivosti $\underline{\mu} < \mu < \overline{\mu}$ a $\sigma < \overline{\sigma}$, máme-li data z odst. 7.1. Z tabulek [23] nalezneme $t_{0,975}(49) = 2,0096$; $\chi^2_{0,05}(49) = 33,930$. Odtud

$$\underline{\mu} = 42,27 - \frac{2,009\,6\sqrt{2,192}}{\sqrt{50}} \doteq 41,849; \quad \overline{\mu} = 42,27 - \frac{2,0096\sqrt{2,192}}{\sqrt{50}} \doteq 42,691;$$

$$\overline{\sigma} = \sqrt{\frac{49 \cdot 2,192}{33,930}} \doteq 1,779.$$

8.4.3 Alternativní rozdělení.

Nechť $\mathbf{X} = (X_1, \ldots, X_n)'$ je náhodný výběr z rozdělení $\mathrm{A}(\pi)$ (viz [24], odst. 14.2). Statistika $T = \sum_{i=1}^{n} X_i$ má binomické rozdělení $\mathrm{Bi}(n, \pi)$ (viz [24], odst.14.5). Funkce

$$h(t,\pi) = \sum_{i=0}^{t} \binom{n}{j} \pi^j (1-\pi)^{n-j}, \tag{8.4.12}$$

tj. distribuční funkce statistiky T v bodě t, splňuje podmínky (8.2.1) a (8.2.2); podle příkl. 23.3 v [24] je totiž

$$h(t,\pi) = 1 - I_\pi(t+1, n-t) = \frac{1}{\mathrm{B}(t+1, n-t)} \int_\pi^1 u^t (1-u)^{n-t-1}\,\mathrm{d}u, \tag{8.4.13}$$

což je při každém $t \in \langle 0, n)$ klesající funkce π; naopak z (8.4.12) je vidět, že $h(t,\pi)$ je při každém $\pi \in (0,1)$ neklesající funkce t. Hranice $\underline{\pi}(t)$, $\overline{\pi}(t)$ konfidenčního intervalu konstruované podle odst. 8.2 s užitím funkce $\overline{h}(t,\pi)$ definované v (8.4.12) jsou tedy dány rovnicemi

$$\sum_{j=0}^{t} \binom{n}{j} \overline{\pi}^j (1-\overline{\pi})^{n-j} = \frac{\alpha}{2}, \quad \sum_{j=0}^{t-1} \binom{n}{j} \underline{\pi}^j (1-\underline{\pi})^{n-j} = 1 - \frac{\alpha}{2}. \tag{8.4.14}$$

Nejpraktičtější postup při řešení rovnic (8.4.14) je nahradit levé strany distribuční funkcí rozdělení beta podle (8.4.13) a dále s užitím vztahu (3.6.2) distribuční funkci rozdělení beta vyjádřit pomocí distribuční funkce rozdělení F. Rovnice (8.4.14) přejdou na rovnice

$$\begin{aligned} 1 - P\left(F\big(2(t+1), 2(n-t)\big) \le \frac{(n-t)\overline{\pi}}{(t+1)(1-\overline{\pi})} \right) &= \frac{\alpha}{2}, \\ 1 - P\left(F\big(2t, 2(n-t+1)\big) \le \frac{(n-t+1)\underline{\pi}}{t(1-\underline{\pi})} \right) &= 1 - \frac{\alpha}{2}. \end{aligned} \tag{8.4.15}$$

Odtud plyne vyjádření $\underline{\pi}(t)$ a $\overline{\pi}(t)$ pomocí kvantilů rozdělení F. Z rovnice (8.4.15) máme

$$\begin{aligned} \frac{(n-t)\overline{\pi}}{(t+1)(1-\overline{\pi})} &= F_{1-\alpha/2}\big(2(t+1), 2(n-t)\big), \\ \frac{(n-t+1)\underline{\pi}}{t(1-\underline{\pi})} &= F_{\alpha/2}\big(2t, 2(n-t+1)\big) \end{aligned}$$

a odtud

$$\begin{aligned}\overline{\pi}(t) &= \frac{(t+1)F_{1-\alpha/2}\big(2(t+1),2(n-t)\big)}{n-t+(t+1)F_{1-\alpha/2}\big(2(t+1),2(n-t)\big)}, \quad t=0,1,\ldots,n-1,\\ \overline{\pi}(n) &= 1 \end{aligned} \tag{8.4.16}$$

a

$$\begin{aligned}\underline{\pi}(0) &= 0, \\ \underline{\pi}(t) &= \frac{t}{(n-t+1)F_{1-\alpha/2}\big(2(n-t+1),2t\big)}, \quad t=1,\ldots,n.\end{aligned} \tag{8.4.17}$$

Hodnoty $\underline{\pi}(t)$ z (8.4.17) jsou tabelovány v [23] pro řadu kombinací, $t, n-t$ a $\alpha/2$. Hodnoty $\overline{\pi}(t)$ není třeba tabelovat, neboť srovnáním (8.4.16) a (8.4.17) snadno zjistíme, že

$$\overline{\pi}(t) = 1 - \underline{\pi}(n-t),$$

takže horní hranice intervalu spolehlivosti se dostane snadno záměnou t za $n-t$.

Při vysokých hodnotách n a podílu t/n nepříliš blízkém 0 nebo 1 (tak, aby bylo přibližně $n(t/n)(1-t/n) > 9$) se aproximuje rozdělení statistiky T normálním rozdělením (viz [24], odst. 26.7) a rovnice (8.4.14) přejdou v

$$\Phi\left(\frac{t+\frac{1}{2}-n\overline{\pi}}{n\overline{\pi}(1-\overline{\pi})}\right)=\frac{\alpha}{2}, \quad \Phi\left(\frac{t-\frac{1}{2}-n\underline{\pi}}{n\underline{\pi}(1-\underline{\pi})}\right)=1-\frac{\alpha}{2}. \tag{8.4.18}$$

Jejich přibližné řešení (po zanedbání členů tvaru $\frac{c}{n}$, kde c je konstanta je

$$\begin{aligned}\overline{\pi}(t) &\doteq p + u_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}, \\ \underline{\pi}(t) &\doteq p - u_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}},\end{aligned} \tag{8.4.19}$$

kde $p = t/n$.

Uvažujme příklad, kdy z $n = 30$ nezávisle zkoušených izolátorů jich bylo $t = 4$ proražených. Stanovme 90% interval spolehlivosti pro rozdíl π proražených izolátorů tohoto typu.

Z tabulek [23] pro $t = 4,\ n-t = 26$ a $\alpha/2 = 0,05$ nalezneme $\underline{\pi}(4) = 0,047$ a $\overline{\pi}(4) = 1 - \underline{\pi}(26) = 0,280$. Použitím aproximativních vztahů (8.4.19) pro $p = \frac{4}{30}$ a $u_{0,95} = 1,644\,854$ dostáváme $\underline{\pi}(4) \doteq 0,031$ a $\overline{\pi}(4) \doteq 0,235$.

8.4.4 Poissonovo rozdělení.

Nechť $\mathbf{X} = (X_1, \ldots, X_n)'$ je náhodný výběr z rozdělení $\mathrm{Po}(\lambda)$. Statistika $T = \sum_{i=1}^{n} X_i$ má rozdělení $\mathrm{Po}(n\lambda)$. Zvolíme-li za $h(t, \lambda)$ funkci

$$h(t, \lambda) = \sum_{j=0}^{t} \frac{(n\lambda)^j}{j!} \mathrm{e}^{-n\lambda}, \tag{8.4.20}$$

získáme pro interval spolehlivosti pro λ rovnice

$$\sum_{j=0}^{t} \frac{(n\overline{\lambda})^j}{j!} \mathrm{e}^{-n\overline{\lambda}} = \frac{\alpha}{2}, \quad \sum_{j=0}^{t-1} \frac{(n\underline{\lambda})^j}{j!} \mathrm{e}^{-n\underline{\lambda}} = 1 - \frac{\alpha}{2}. \tag{8.4.21}$$

Nejvýhodnější postup řešení rovnic (8.4.21) je založen na vztahu

$$\sum_{j=0}^{t} \frac{(n\overline{\lambda})^j}{j!} \mathrm{e}^{-n\overline{\lambda}} = \int_{2n\lambda}^{\infty} \frac{e^{-\frac{x}{2}} x^{2(t+1)-1}}{2^{2(t+1)}(2t+1)!}\, \mathrm{d}x = P\Big(\chi^2\big(2(t+1)\big) > 2n\overline{\lambda}\Big), \tag{8.4.22}$$

viz [24], příkl. 22.4. Aplikací tohoto vztahu na levé strany rovnic (8.4.21) dostaneme rovnice

$$P\Big(\chi^2\big(2(t+1)\big) > 2n\overline{\lambda}\Big) = \frac{\alpha}{2}, \quad P\big(\chi^2(2t) > 2n\underline{\lambda}\big) = 1 - \frac{\alpha}{2}. \tag{8.4.23}$$

Odtud

$$\overline{\lambda}(t) = \frac{1}{2n} \chi^2_{1-\alpha/2}\big(2(t+1)\big), \quad t = 0, 1, \ldots, \tag{8.4.24}$$

a

$$\begin{aligned} \underline{\lambda}(t) &= \frac{1}{2n} \chi^2_{\alpha/2}(2t), \qquad t = 1, 2, \ldots, \\ \underline{\lambda}(0) &= 0. \end{aligned} \tag{8.4.25}$$

Kvantily rozdělení χ^2 najdeme v [23]. Při vyšších hodnotách t takových, že $\chi^2_{1-\alpha/2}(2(t+1))$ není už v tabulkách obsaženo, použije se aproximace rozdělení χ^2 normálním rozdělením (viz [24], úloha 26.8.2); pak bude přibližně

$$\overline{\lambda}(t) \doteq \frac{1}{n}\big(t + 1 + u_{1-\alpha/2}\sqrt{t+1}\big), \quad \underline{\lambda}(t) \doteq \frac{1}{n}\big(t - u_{1-\alpha/2}\sqrt{t}\big). \tag{8.4.26}$$

Uvažujme údaje příkl. 6.6.2. Konfidenční interval s koeficientem spolehlivosti $1-\alpha = 0,95$ pro střední hodnotu počtu kazů v jedné tabuli má krajní body

$$\begin{aligned} \underline{\lambda}(17) &= \frac{1}{50}\chi^2_{0,025}(34) = \frac{19,806}{50} \doteq 0,396; \\ \overline{\lambda}(17) &= \frac{1}{50}\chi^2_{0,975}(36) = \frac{54,437}{50} \doteq 1,089. \end{aligned}$$

Krajní body přibližného intervalu spolehlivosti jsou podle (8.4.26)

$$\underline{\lambda}(17) \doteq \frac{1}{25}(17 - 1,96\sqrt{17}) \doteq 0,357; \qquad \overline{\lambda}(17) \doteq \frac{1}{25}(18 + 1,96\sqrt{18}) \doteq 1,053.$$

8.4.5 Dvě normální rozdělení.

Nechť $\mathbf{X} = (X_1, \ldots, X_n)'$ je náhodný výběr z rozdělení $\mathrm{N}(\mu_1, \sigma_1^2)$ a $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ náhodný výběr z rozdělení $\mathrm{N}(\mu_2, \sigma_2^2)$. Nechť výběry $\mathbf{X}$ a $\mathbf{Y}$ jsou nezávislé.

Předpokládejme nejprve, že $\sigma_1^2 = \sigma_2^2 = \sigma^2$ a uvažujme parametrickou funkci $\tau = \tau(\mu_1, \mu_2, \sigma^2) = \mu_1 - \mu_2 = \Delta$. Podle odst. 3.5 má náhodná veličina

$$h(\overline{X}, \overline{Y}, \Delta, S^2) = \frac{\overline{X} - \overline{Y} - \Delta}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{8.4.27}$$

kde S^2 je dáno výrazem (3.5.3), rozdělení $\mathrm{t}(n_1 + n_2 - 2)$, tedy nezávislé na (μ_1, μ_2, σ^2), a při každém $(\overline{X}, \overline{Y}, S^2)$ je (8.4.27) klesající funkcí Δ. Rovnice (8.2.5) a (8.2.6) přejdou v

$$\frac{\overline{X} - \overline{Y} - \underline{\Delta}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = t_{1-\alpha/2}(n_1 + n_2 - 2), \quad \frac{\overline{X} - \overline{Y} - \overline{\Delta}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = -t_{1-\alpha/2}(n_1 + n_2 - 2), \tag{8.4.28}$$

odkud dostáváme interval spolehlivosti pro $\Delta = \mu_1 - \mu_2$

$$\left(\overline{X} - \overline{Y} - t_{1-\alpha/2}(n_1 + n_2 - 2)s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \right.$$
$$\left. \overline{X} - \overline{Y} + t_{1-\alpha/2}(n_1 + n_2 - 2)s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right). \tag{8.4.29}$$

V práci [12], str. 108, jsou uvedeny výsledky obsahu aktivního chloru (v g/l) v bělicím roztoku ze dvou nádrží. Z nich se vypočte:

$$\begin{aligned} \text{I. nádrž: } n_1 &= 25; \quad \overline{x} = 34,48; \quad s_1^2 = 1,748\,2; \\ \text{II. nádrž: } n_2 &= 10; \quad \overline{y} = 35,59; \quad s_2^2 = 1,712\,1. \end{aligned}$$

Předpokládejme, že se jedná o nezávislé výběry ze dvou normálních rozdělení, a stanovme 95% interval spolehlivosti pro rozdíl obsahu chloru v obou nádržích.

Z tabulek [23] nalezneme $t_{0,975}(33) = 2,034\,5$. Dále

$$s^2 = \frac{24 \cdot 1,748\,2 + 9 \cdot 1,712\,1}{33} \doteq 1,738\,4,$$

takže pro určení (8.4.29) dostáváme

$$2,034\,5\left(1,738\,4\left(\frac{1}{25} + \frac{1}{10}\right)\right)^{\frac{1}{2}} \doteq 1,003\ 7$$

a odtud interval spolehlivosti pro $\mu_1 - \mu_2$ je $(-2,114;\ -0,106)$.

Stanovme ještě interval spolehlivosti pro parametrickou funkci

$$\tau(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = \frac{\sigma_1^2}{\sigma_2^2}.$$

Uvažujme náhodnou veličinu

$$h(S_1^2, S_2^2, \sigma_1^2, \sigma_2^2) = \frac{S_1^2}{S_2^2}\frac{\sigma_2^2}{\sigma_1^2} = \frac{S_1^2}{S_2^2}\frac{1}{\tau}, \tag{8.4.30}$$

tj. veličinu (3.5.5), která má rozdělení $\mathrm{F}(n_1 - 1,\ n_2 - 1)$. Z rovnic

$$\frac{S_1^2}{S_2^2}\frac{1}{\underline{\tau}} = F_{1-\alpha/2}(n_1 - 1, n_2 - 1), \quad \frac{S_1^2}{S_2^2}\frac{1}{\overline{\tau}} = F_{\alpha/2}(n_1 - 1, n_2 - 1), \tag{8.4.31}$$

vyplývá interval spolehlivosti

$$\left(\frac{S_1^2/S_2^2}{F_{1-\alpha/2}(n_1 - 1, n_2 - 1)}, \quad \frac{S_1^2/S_2^2}{F_{\alpha/2}(n_1 - 1, n_2 - 1)},\right) \tag{8.4.32}$$

pro σ_1^2/σ_2^2.

8.4.6 Dvourozměrné normální rozdělení.

Nechť $(X_1, Y_1)', \ldots, (X_n, Y_n)'$ je náhodný výběr z dvourozměrného normálního rozdělení s parametry

$$\mu_1 = E(X), \quad \mu_2 = E(Y), \quad \sigma_1^2 = \mathrm{var}(X), \quad \sigma_2^2 = \mathrm{var}(Y), \quad \rho = \frac{\mathrm{cov}(X, Y)}{\sigma_1 \sigma_2}.$$

Uvažujme veličiny $D_i = X_i - Y_i, i = 1, \ldots, n$. Podle odst. 3.7 můžeme považovat $\mathbf{D} = (D_1, \ldots, D_n)'$ za náhodný výběr z rozdělení $\mathrm{N}(\Delta, \sigma_D^2)$, kde $\Delta = \mu_1 - \mu_2$ a $\sigma_D^2 = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2$.

Zcela analogickým postupem jako v příkl. 8.4.2 zjistíme, že interval spolehlivosti pro Δ je

$$\left(\overline{D} - t_{1-\alpha/2}(n-1)\frac{S_D}{\sqrt{n}}, \quad \underline{D} + t_{1-\alpha/2}(n-1)\frac{S_D}{\sqrt{n}}, \right) \tag{8.4.33}$$

kde $\overline{D}$ a S_D^2 jsou statistiky (3.7.3).

i	x_i	y_i	d_i	b_i
1	24,14	24,26	-0,12	48,40
2	36,99	37,31	-0,32	74,30
3	29,10	28,95	0,15	58,05
4	31,21	31,66	-0,45	62,87
5	48,82	49,33	-0,51	98,15
6	48,43	48,90	-0,47	97,33
7	38,11	38,37	-0,26	76,48
8	37,62	38,11	-0,49	75,73
9	57,21	57,44	-0,23	114,65
10	41,43	42,01	-0,58	83,44
11	38,94	39,28	-0,34	78,22
12	45,77	46,15	-0,38	91,92

Tab. 8.1: Hodnoty obsahu železa.

Tabulka 8.1 udává hodnoty obsahu železa u $n = 12$ vzorků železné ruda. Zde x_i jsou výsledky standardní analytické metody a y_i výsledky nově zaváděné analytické metody. Z těchto hodnot vypočteme

$$\overline{d} = \frac{-4}{12} \doteq -0,333; \qquad s_D^2 = \frac{12 \cdot 1,779\,8 - (-4)^2}{132} \doteq 0,040\,6$$

a z tabulek [23] nalezneme $t_{0,975}(11) = 2,201$. Je tedy $s_D/\sqrt{12} \doteq 0,058$, takže $(-0,461; -0,205)$ je 95% interval spolehlivosti pro $\Delta = \mu_1 - \mu_2$.

Ke konstrukci intervalu spolehlivosti pro koeficient korelace ρ lze využít statistiky

$$Z = \frac{1}{2} \ln \frac{1+r}{1-r},$$

kde r je výběrový koeficient korelace (3.7.5). Podle odst. 3.7 má veličina

$$h(Z, \rho) = (n-3)^{\frac{1}{2}} \Big(Z - \frac{1}{2} \ln \frac{1+\rho}{1-\rho} - \frac{\rho}{2(n-1)} \Big) \tag{8.4.34}$$

pro přiměřená n (viz odst. 3.7) přibližně rozdělení N(0, 1). Přitom pro každé r, a tedy i každé Z, je $h(Z, \rho)$ klesající funkcí ρ. Rovnice (8.2.5) a (8.2.6) přejdou v

$$h(Z, \underline{\rho}) = u_{1-\alpha/2}, \qquad h(Z, \overline{\rho}) = -u_{1-\alpha/2}. \tag{8.4.35}$$

Rovnice je zapotřebí řešit numericky; pro první aproximaci můžeme zanedbat v (8.4.34) výraz $\rho/\big(2(n-1)\big)$.

Byla zjišťována zralost viskózy X a tažnost Y umělého vlákna. Z $n = 46$ měření byl vypočten výběrový koeficient korelace $r = 0,638$. Stanovme 95% interval spolehlivosti pro ρ. Protože $Z = 0,754\,794$, budeme řešit rovnice

$$43^{\frac{1}{2}} \Big(0,754\,794 - \frac{1}{2} \ln \frac{1+\overline{\rho}}{1-\overline{\rho}} - \frac{\overline{\rho}}{90} \Big) = -1,959\,964,$$

$$43^{\frac{1}{2}} \Big(0,754\,794 - \frac{1}{2} \ln \frac{1+\underline{\rho}}{1-\underline{\rho}} - \frac{\underline{\rho}}{90} \Big) = 1,959\,964.$$

Jejich řešením nalezneme interval spolehlivosti $(0,423; 0,780)$ pro ρ.

8.5 Úlohy.

8.5.1

Uvažujte údaje odst. 7.3 a nalezněte 95% interval spolehlivosti pro parametrickou funkci $\lambda = \frac{1}{\delta}$.

$[(0,008;\ 0,033)]$

8.5.2

Uvažujte údaje příkl. 6.6.2 a nalezněte 95% interval spolehlivosti pro parametrickou funkci $\tau(\lambda) = e^{-\lambda}$.

$[(0,337; 0,673).]$

Část III

Ověřování statistických hypotéz

Kapitola 9

Úvodní poznámky

9.1 Podstata statistického rozhodování.

Statistický rozhodovací problém vzniká, když je třeba zvolit jedno rozhodnutí z určité třídy rozhodnutí v dané situaci možných a relativní výhodnost jednotlivých možných rozhodnutí závisí na hodnotách jednoho nebo několika neznámých parametrů, o kterých lze získat jen informace zatížené náhodnými chybami; lze jen pozorovat náhodné veličiny, jejichž rozdělení závisí na neznámých parametrech. Volba rozhodnutí na základě pozorování náhodných veličin je přirozeně spojena s rizikem, že zvolené rozhodnutí nebude v dané situaci nejlepší a v důsledku nesprávného rozhodnutí vznikne ztráta. Úkolem teorie statistického rozhodování je vypracovat pravidla pro volbu rozhodnutí v podmínkách nejistoty tak, aby střední hodnota ztrát byla co nejmenší.

9.2 Příklady.

9.2.1 Statistická přejímací kontrola.

Předpokládejme, že závod přebírá od subdodavatele dodávky po $N = 4\,000$ kusech polotovarů. V každé dodávce je určité procento $100\pi\%$ nevyhovujících kusů, číslo π kolísá od dodávky k dodávce. Pokud je π menší než určité číslo π_0, je výhodnější dodávku převzít, neboť potíže s tříděním by byly větší než potíže se zpracováním dodávky; jakmile π překročí π_0, je výhodnější celou dodávku třídit. Z každé dodávky se tedy vybere n (např. 200) kusů a podle počtu vadných se rozhodne, zda se dodávky použije bez dalších opatření či

zda se před použitím podrobí stoprocentní kontrole.

To je statistický rozhodovací problém: Třída možných rozhodnutí obsahuje dva prvky „přijmout dodávku“ a „přetřídit dodávku“. Parametr, na kterém závisí volba rozhodnutí, je π = podíl vadných kusů v dodávce. Je neznám, avšak pozoruje se náhodná veličina X =počet vadných kusů ve výběru, která má hypergeometrické rozdělení

$$P(X = x) = \frac{1}{\binom{N}{n}} \binom{N\pi}{x} \binom{N - N\pi}{n - x},$$
$$x = \max(0, N\pi - N + n), \ldots, \min(N\pi, n).$$

Rozhodovací pravidlo (rozhodovací funkce) přiřazuje každé z možných hodnot náhodné veličiny X jedno ze dvou rozhodnutí „přijmout dodávku“ nebo přetřídit dodávku. Každé z obou uvedených rozhodnutí má svoje důsledky: Přijme-li se dodávka obsahující $M = N\pi$ vadných kusů bez třídění, vzniknou náklady $aN\pi$, kde a je škoda způsobená vadným kusem ve výrobě (např. cena opravy hotového výrobku, který v důsledku zamontování vadné součástky nefunguje), a třídí-li se dodávka, vzniknou náklady rovné ceně třídění.

9.2.2 Volba optimální varianty technologického procesu.

Představme si, že se provozním pokusem srovnávají tři různé varianty technologického procesu v chemickém průmyslu. Předpokládejme, že rozhodující znak jakosti výrobku je náhodná veličina, která má při i-té variantě rozdělení $\mathrm{N}(\mu_i, \sigma_i^2)$, $i = 1, 2, 3$, a že na nejvhodnější se považuje varianta s nejvyšší hodnotou μ_i. V pokusu se připraví n várek každou ze zkoušených variant a získají se tři náhodné vektory $(X_{i1}, \ldots, X_{in})'$, tj. výsledky naměřené u n várek připravených i-tou technologií, $i = 1, 2, 3$. Na základě výsledků se má rozhodnout, která ze tří srovnávaných technologií je podle uvedeného kritéria nejlepší a má být zavedena ve velkém.

Jde o statistický rozhodovací problém se třemi možnými rozhodnutími typu „doporučit i-tou variantu“, $i = 1, 2, 3$. Označme s číslo varianty doporučené na základě výsledku pokusu; je-li např. doporučena varianta druhá, je $s = 2$. V důsledku rozhodnutí vznikne škoda úměrná rozdílu mezi μ_s a $\max_{1 \le i \le 3} \mu_i$, tedy škoda

$$a \max_{1 \le i \le 3} (\mu_i - \mu_s);$$

je-li zvolena podle výsledků experimentu práva varianta s maximální hodnotou μ_i, je škoda nulová. Je-li zvolena jiná varianta, vznikne určitá ztráta, zpravidla asi tím větší, čím větší bude rozdíl mezi $\max \mu_i$ a vybraným μ_s.

9.3 Rozhodovací funkce.

Jak je vidět z příkladů odst. 9.2, základní prvky úlohy statistického rozhodování jsou:

1. množina možných rozhodnutí, řekněme D, tzv. *prostor rozhodnutí*;
2. množina možných výsledků příslušného náhodného experimentu, tzv. *výběrový prostor*. který označíme $\mathfrak{X}$;
3. množina možných hodnot parametrů rozdělení pravděpodobnosti výsledků daného experimentu, tzv. *parametrický prostor,* který označíme Ω a jeho prvky θ;
4. funkce $W(d, \theta)$ na $D \times \Omega$ udávající ztrátu, která vznikne, přijme-li se rozhodnutí d, když správná hodnota parametru je θ, tzv. *ztrátová funkce.*

V příkladě 9.2.1 je prostor rozhodnutí $D = \{d_1, d_2\}$, kde d_1 značí rozhodnutí „přijmout dodávku“ a d_2 rozhodnutí „dodávku přetřídit“, výběrový prostor $\mathfrak{X} = \{\max(0, N\pi - N + n), \max(0, N\pi - N + n) + 1, \ldots \min(N\pi, n)\}$, parametrický prostor Ω je interval $\langle 0, 1\rangle$. Ztrátovou funkci lze zvolit různě, ale nejpřirozenější volba v popsané situaci by byla např.

$$\begin{aligned} W(d_1, \pi) &= aN\pi, \\ W(d_2, \pi) &= C, \end{aligned} \tag{9.3.1}$$

kde a značí škodu způsobenou vadným kusem a C náklad na třídění dodávky. Jiná přijatelná volba ztrátové funkce pro daný příklad je

$$\begin{aligned} W(d_1, \pi) &= 0, \quad \pi \le \pi_0, \\ &= 1, \quad \pi > \pi_0, \\ W(d_2, \pi) &= 1, \quad \pi \le \pi_0, \\ &= 0, \quad \pi > \pi_0. \end{aligned} \tag{9.3.2}$$

Při užití ztrátové funkce (9.3.2) nevyjadřujeme absolutní velikosti ztráty, nýbrž jen skutečnost, že při $\pi \leq \pi_0$ je rozhodnutí d_1 správné a d_2 nikoliv, zatímco při $\pi > \pi_0$ je správné rozhodnutí d_2.

Rozhodnutí d se vybírá na základě výsledku náhodného experimentu. Pravidlo, které přiřazuje každému z možných výsledků experimentu určité rozhodnutí, tj. funkci $d(x)$ zobrazující $\mathfrak{X}$ do D, je tzv. *rozhodovací funkce.* V příkladě 9.2.1 nejpřirozenější rozhodovací funkcí je

$$\begin{aligned} d(x) &= d_1, \quad x < c, \\ &= d_2, \quad x \geq c, \end{aligned}$$

tj. přijmout dodávku, když počet vadných kusů X ve výběru je menší než dané číslo c, a zamítnout ji, když $X \geq c$.

Jelikož rozhodnutí je pak funkcí výsledku náhodného experimentu, je i ztráta utrpěná v důsledku rozhodnutí náhodou veličinou; její střední hodnota

$$R(\boldsymbol{\theta}) = E\Big(W\big(d(X), \boldsymbol{\theta}\big)\Big) \tag{9.3.3}$$

je tzv. *riziková funkce.* Riziková funkce je základem pro srovnávání různých rozhodovacích pravidel. Velmi zhruba řečeno, cílem teorie rozhodovacích funkcí je konstrukce rozhodovacích pravidel s „příznivým průběhem“ rizikové funkce.

Obecná teorie rozhodovacích funkcí je předmětem mnoha speciálních pojednání (např. [3, 36]). Zde se omezíme jen na zvláštní případ, totiž na úlohy s dvěma rozhodnutími a s jednoduchou ztrátovou funkcí obdobného typu jako (9.3.2). Rozhodovací funkce pro úlohy tohoto druhu jsou ve statistice známy jako *testy statistických hypotéz.*

Kapitola 10

Testování statistických hypotéz

10.1 Úloha testování statistické hypotézy.

Jestliže v obecném statistickém rozhodovacím problému z čl. 9 množina možných rozhodnutí má jen dva prvky, $D = \{d_1, d_2\}$ a parametrický prostor Ω příslušného experimentu je rozdělen na dvě části, řekněme ω a $\overline{\omega} = \Omega - \omega$, kde při $\boldsymbol{\theta} \in \omega$ je výhodnější rozhodnutí d_1 než d_2 a při $\boldsymbol{\theta} \in \overline{\omega}$ je výhodnější rozhodnutí d_2 než d_1, nazýváme rozhodovací problém *úlohou ověření (či testu) statistické hypotézy.* Tvrzení „$\boldsymbol{\theta} \in \omega$“ označíme jako hypotézu H a opačné tvrzení „$\boldsymbol{\theta} \in \Omega - \omega$“ jako alternativní hypotézu A. Jestliže je hypotéza H správná, tj. jestliže ve skutečnosti je $\boldsymbol{\theta}$ prvkem ω, pak je na místě rozhodnutí d_1; volbu rozhodnutí d_1 označujeme tedy jako přijetí hypotézy H a volbu rozhodnutí d_2 jako zamítnutí H (čili přijetí alternativní hypotézy A).

Pravidlo, podle kterého se na základě výsledku náhodného experimentu rozhoduje, zda danou hypotézu přijmout či zamítnout, se nazývá *test hypotézy H.*

10.2 Příklady.

10.2.1 Přejímací kontrola dodávek výrobků (pokračování příkl. 9.2.1).

V příkladě 9.2.1 jde v podstatě o test hypotézy $H : \pi \leq \pi_0$ (podíl vadných výrobků v dodávce nepřekračuje dané číslo π_0) proti alternativní hypotéze $A : \pi > \pi_0$. Zamítnutí hypotézy H má za následek stoprocentní kontrolu

(třídění) celé dodávky. Parametrický prostor Ω (kterým je interval $\langle 0, 1\rangle$) je rozdělen na části $\omega = \langle 0, \pi_0\rangle$ a $\overline{\omega} = \Omega - \omega = (\pi_0, 1\rangle$. Test hypotézy H je: zamítnout $H : \pi \leq \pi_0$, když $x \geq c$, kde x je počet vadných výrobků ve výběru a c dané číslo (o jeho volbě bude řeč později).

10.2.2 Zkouška účinnosti nového způsobu ochrany proti korozi.

Představme si následující situaci: Chemik navrhl nový způsob povrchové úpravy materiálu od kterého očekává lepší ochranu proti korozi. K ověření, zda nový způsob skutečně poskytuje lepší ochranu proti korozi, provedl pokus: Připravil běžným standardním způsobem n_1 vzorků materiálu a nově navrženým způsobem n_2 vzorků a obě skupiny vystavil účinkům agresivního prostředí. Po určité době změří na všech vzorcích účinky koroze (např. procento povrchu zasažené korozí). Podle výsledku pokusu rozhodne, zda nově navržený způsob doporučí či nikoliv. Ze zkušenosti s podobnými pokusy ví, že zvolená míra poškození má normální rozdělení.

Matematicko-statistický popis daného problému zní: Pozorují se dva nezávislé náhodné výběry $\mathbf{X} = (X_1, \ldots, X_{n1})'$ a $\mathbf{Y} = (Y_1, \ldots, Y_{n2})'$; $\mathbf{X}$ je výběr z rozdělení $\mathrm{N}(\mu_1, \sigma_1^2)$ a $\mathbf{Y}$ je výběr z rozdělení $\mathrm{N}(\mu_2, \sigma_2^2)$. Přitom X_i je procento povrchu zachváceného korozí u i-tého standardního vzorku a Y_i procento povrchu zachváceného korozí u i-tého vzorku připraveného novým způsobem. Výběrový prostor $\mathfrak{X}$ tedy je $(n_1 + n_2)$-rozměrný euklidovský prostor. Parametrický prostor Ω je část 4-rozměrného euklidovského prostoru,

$$\Omega = \left\{\theta = (\mu_{\mathbf{1}}, \mu_{\mathbf{2}}, \sigma_{\mathbf{1}}^{\mathbf{2}}, \sigma_{\mathbf{2}}^{\mathbf{2}}) \,\middle|\, \sigma_{\mathbf{1}}^{\mathbf{2}} > \mathbf{0},\ \sigma_{\mathbf{2}}^{\mathbf{2}} > \mathbf{0}\right\}.$$

Rozhodnutí „doporučit nový způsob“ je na místě, když $\mu_2 < \mu_1$, rozhodnutí „nezavádět nový způsob“ je správné, když $\mu_1 \leq \mu_2$. Tím je dán rozklad Ω na dvě disjunktní části

$$\omega = \left\{(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) \,\middle|\, \mu_1 \leq \mu_2, \sigma_1^2 > 0, \sigma_2^2 > 0\right\},$$

$$\overline{\omega} = \left\{(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) \,\middle|\, \mu_1 > \mu_2, \sigma_1^2 > 0, \sigma_2^2 > 0\right\},$$

Jde o test hypotézy $H : \theta \in \omega$, tj. $\mu_1 \leq \mu_2$, proti alternativní hypotéze $A : \theta \in \overline{\omega}$, tj. $\mu_1 > \mu_2$. Postup při provedení testu je popsán v odst. 11.10.

10.3 Konstrukce testů statistických hypotéz.

Protože v úloze testování statistické hypotézy jsou možná jen dvě rozhodnutí (totiž „zamítnout hypotézu H“ a „nezamítnout hypotézu H“), je test (rozhodovací pravidlo) zcela popsán rozkladem výběrového prostoru $\mathfrak{X}$ na dvě disjunktní části: část, které je přiřazeno rozhodnutí „zamítnout H“, a část, které je přiřazeno rozhodnutí „nezamítat H“.

10.4 Kritický obor.

Množinu prvků $\mathbf{x}$ výběrového prostoru $\mathfrak{X}$, kterým je přiřazeno rozhodnutí „zamítnout hypotézu H“, nazveme *kritickým oborem* (úplněji *kritickým oborem pro testování hypotézy* H).

V příkladě 10.2.1 je výběrovým prostorem množina celých nezáporných čísel nejvýše rovných n, $\mathfrak{X} = \{0, 1, \ldots, n\}$. Náhodná veličina X sice může nabývat jen hodnot z intervalu $\langle \max(0, N\pi - N + n), \min(N\pi, n)\rangle$, ale parametr π není znám a může mít kteroukoliv hodnotu z intervalu $\langle 0, 1\rangle$; je tedy třeba považovat za možné výsledky experimentu všechna celá čísla z intervalu $\langle 0,\ n\rangle$. Kritický obor pro test $H : \pi \leq \pi_0$ je množina $W = \{c, c+1, \ldots, n\}$.

Kritický obor budeme v dalším textu této kapitoly značit písmenem W a pravděpodobnost, že experiment dá výsledek z W, když parametr má hodnotu $\boldsymbol{\theta}$, symbolem $P_w(\boldsymbol{\theta})$. Je-li $T = T(\mathbf{X})$ nějaká statistika, symbol $P\big(T(X) \geq c \,\big|\, \boldsymbol{\theta} = \boldsymbol{\theta}_0\big)$ bude značit pravděpodobnost jevu $\{T(\mathbf{X}) \geq c\}$ počítanou při hodnotě $\boldsymbol{\theta} = \boldsymbol{\theta_0}$; zde nejde o podmíněnou pravděpodobnost, protože parametr $\boldsymbol{\theta}$ nepovažujeme za náhodnou veličinu.

Vhodnost daného testu statistické hypotézy H (tj. daného kritického oboru pro test hypotézy H) můžeme posoudit, stanovíme-li pro každé $\boldsymbol{\theta} \in \Omega$ pravděpodobnost, že hypotéza H bude testem zamítnuta, když parametr má ve skutečnosti hodnotu $\boldsymbol{\theta}$. Tato pravděpodobnost by měla být co nejmenší, když H je správná (tj. když $\boldsymbol{\theta} \in \omega$), a co největší, když H neplatí (tj. když $\boldsymbol{\theta} \in \Omega - \omega$). Zavádíme tedy následující definici.

10.5 Silofunkce testu (kritického oboru).

Funkci $P_w(\boldsymbol{\theta})$ na Ω, která udává ke každému $\boldsymbol{\theta} \in \Omega$ pravděpodobnost zamítnutí hypotézy H, když parametr má hodnotu $\boldsymbol{\theta}$, tj.

$$P_{\mathbf{W}}(\boldsymbol{\theta}) = P(X \in W|\boldsymbol{\theta}), \tag{10.5.1}$$

nazveme *silofunkcí testu (kritického oboru)* W. Hodnotu silofunkce v určitém bodě $\boldsymbol{\theta} = \boldsymbol{\theta}'$ nazýváme *sílou testu vůči alternativě* $\boldsymbol{\theta} = \boldsymbol{\theta}'$.

Jestliže výsledek experimentu má rozdělení diskrétního typu s pravděpodobnostní funkcí $p(\mathbf{x};\boldsymbol{\theta})$; vypočítá se silofunkce jako

$$P_{\mathbf{W}}(\boldsymbol{\theta}) = \sum_{\mathbf{x}\in\mathbf{W}} p(\mathbf{x};\boldsymbol{\theta}), \tag{10.5.2}$$

jestliže X má rozdělení spojitého typu se sdruženou hustotou $f(\mathbf{x};\boldsymbol{\theta})$, vypočítá se silofunkce jako

$$P_W(\boldsymbol{\theta} = \int_W \ldots \int f(\mathbf{x};\boldsymbol{\theta})\,\mathrm{d}x_1 \ldots\, \mathrm{d}x_n. \tag{10.5.3}$$

Při testu statistické hypotézy $H : \theta \in \omega$ může dojít k chybnému rozhodnutí dvěma způsoby:

1. Hypotéza H je ve skutečnosti správná a na základě testu je zamítnuta (experiment dá výsledek z kritického oboru W),

2. hypotéza H neplatí, ale test nevede k jejímu zamítnutí (experiment dá výsledek, který nepatří do kritického oboru W, $\mathbf{X} \in \mathfrak{X} - W$).

Pro tyto dva druhy chyb je v matematické statistice zavedeno zvláštní označení.

10.6 Chyba prvního a druhého druhu.

Rozhodnutí „zamítnout hypotézu H“, když ve skutečnosti je H správná, se nazývá *chyba I. druhu.* Rozhodnutí „nezamítat hypotézu H“, když ve skutečnosti je správná alternativní hypotéza A, se nazývá *chyba II. druhu.*

10.7 Hladina významnosti.

Při volbě testu statistické hypotézy se snažíme především omezit riziko chyby I. druhu a kritický obor vybíráme tak, aby pravděpodobnost chyby I. druhu nepřekročila předem zvolené číslo α; zpravidla se používá hodnot α ne vyšších než $0,1$, např. $\alpha = 0,05$ nebo $0,01$. Hypotéza $H : \theta \in \omega$ je správná, když parametr θ patří do dané části ω parametrického prostoru Ω. Na kritický obor tedy klademe podmínku

$$P_W(\boldsymbol{\theta}) \leq \alpha \quad \text{pro všechna} \quad \boldsymbol{\theta} \in \omega. \tag{10.7.1}$$

Toto číslo α nazýváme *hladinou významnosti* a test (kritický obor) splňující podmínku (10.7.1) *testem na hladině významnosti* α.

Výklad o metodách a kritériích pro konstrukci kritických oborů na dané hladině významnosti lze najít v podrobnějších učebnicích matematické statistiky, např. [1, 20, 31], úplný rozbor úlohy v monografii [22]. Zde uvedeme jen jednoduchý postup použitelný při nejběžnějších typech hypotéz.

10.8 Běžné typy statistických hypotéz.

V aplikacích se nejčastěji setkáváme s hypotézami tohoto tvaru:

1. Nechť $\tau(\boldsymbol{\theta})$ je daná reálná funkce parametru $\boldsymbol{\theta}$. Hypotéza H_1 zní $\tau(\boldsymbol{\theta}) \leq \tau_0$ (kde τ_0 je dané číslo), alternativní hypotéza A_1 zní $\tau(\boldsymbol{\theta}) > \tau_0$.

2. Hypotéza H_2 zní $\tau(\boldsymbol{\theta}) \geq \tau_0$ (τ_0 dané číslo), alternativní hypotéza A_2 je $\tau(\boldsymbol{\theta}) < \tau_0$.

3. Hypotéza H_3 zní $\tau(\boldsymbol{\theta}) = \tau_0$ (τ_0 dané číslo), alternativní hypotéza A_3 je $\tau(\boldsymbol{\theta}) \neq \tau_0$.

V případě 1 a 2 mluvíme o *testu jednostranné hypotézy proti jednostranné alternativě*, v případě 3 mluvíme o *testu hypotézy H_3 proti oboustranné alternativě,* stručněji o *jednostranných* a *oboustranných testech.*

Kritické obory pro testování hypotéz tohoto typu lze zpravidla uvést na tento tvar: Nechť $T(\mathbf{X})$ je reálná statistika (reálná funkce výsledku experimentu), jejíž rozdělení závisí jen na hodnotě funkce $\tau(\boldsymbol{\theta})$ parametru $\boldsymbol{\theta}$. Kritický obor W_1 pro test hypotézy H_1 je

$$W_1 = \{\mathbf{x} \,|\, T(\mathbf{x}) \geq C_1\}, \tag{10.8.1}$$

kde C_1 je číslo zvolené tak, aby

$$P\big(T(\mathbf{X}) \geq C_1 \,|\, \tau(\boldsymbol{\theta}) = \tau_0\big) = \alpha. \tag{10.8.2}$$

Kritický obor W_2 pro test hypotézy H_2 je

$$W_2 = \{\mathbf{x} \,|\, T(\mathbf{x}) \leq C_2\} \tag{10.8.3}$$

kde C_2 je zvoleno tak, aby

$$P\big(T(\mathbf{X}) \leq C_2 \,|\, \tau(\boldsymbol{\theta}) = \tau_0\big) = \alpha. \tag{10.8.4}$$

Kritický obor pro test hypotézy H_3 je

$$W_3 = \{\mathbf{x} \,|\, T(\mathbf{x}) \leq C_3'\} \cup \{\mathbf{x} \,|\, T(\mathbf{x}) \geq C_3''\}, \tag{10.8.5}$$

kde C_3' a C_3'' jsou čísla zvolená tak, aby

$$P\big(T(\mathbf{X}) \leq C_3' \,|\, \tau(\boldsymbol{\theta}) = \tau_0\big) = P\big(T(\mathbf{X}) \geq C_3'' \,|\, \tau(\boldsymbol{\theta}) = \tau_0\big) = \frac{\alpha}{2}. \tag{10.8.6}$$

10.9 Konfidenční intervaly a testy statistických hypotéz.

Konstrukci kritických oborů typu $W_1,\ W_2,\ W_3$ pro hypotézy $H_1,\ H_2,\ H_3$ z předcházejícího odstavce, tj. výběr příslušné statistiky T a stanovení čísel $C_1,\ C_2,\ C_3',\ C_3''$, usnadní následující vztah mezi konfidenčními intervaly a testy statistických hypotéz.

Nechť $\big(\underline{\tau}(\mathbf{X}), \overline{\tau}(\boldsymbol{X})\big)$ je konfidenční interval pro $\tau(\boldsymbol{\theta})$ s koeficientem spolehlivosti $1 - 2\alpha$. Potom (viz odst. 8.1) pro libovolné $\boldsymbol{\theta} \in \Omega$ platí

$$P\big(\underline{\tau}(\mathbf{X}) \geq \tau(\boldsymbol{\theta})\big) = \alpha. \tag{10.9.1}$$

Odtud plyne: Je-li správná hypotéza H_1 z odst. 10.8, pak pravděpodobnost jevu $\underline{\tau}(\mathbf{X}) \geq \tau_{\mathbf{0}}$ je nejvýše rovna α.

$$P\big(\underline{\tau}(\mathbf{X}) \geq \tau_0 \,|\, \boldsymbol{\theta}\big) \leq \alpha \ \text{ pro všechna } \ \boldsymbol{\theta} \ \text{ taková, že } \ \tau(\boldsymbol{\theta}) \leq \tau_0, \tag{10.9.2}$$

a kritický obor W_1 lze zapsat ve tvaru

$$W_1 = \{\mathbf{x} \,|\, \underline{\tau}(\mathbf{x} \geq \tau_0\} \tag{10.9.3}$$

Je-li správná hypotéza H_2 z odst. 10.8, tj. je-li $\tau(\boldsymbol{\theta}) \geq \tau_0$, pak

$$P\big(\overline{\tau}(\mathbf{X}) \leq \tau_0\big) \leq \alpha \quad \text{pro všechna } \boldsymbol{\theta} \text{ s vlastností } \tau(\boldsymbol{\theta}) \geq \tau_0 \tag{10.9.4}$$

a kritický obor W_2 z (10.3.6) lze zapsat ve tvaru

$$W_2 = \{\mathbf{x} \,|\, \overline{\tau}(\mathbf{x} \leq \tau_0\}. \tag{10.9.5}$$

Pro konstrukci testu hypotézy H_3 vyjdeme z konfidenčního intervalu s koeficientem spolehlivosti $1 - \alpha$. Pro takový interval platí

$$P\big(\underline{\tau}(\mathbf{X}) \geq \tau(\boldsymbol{\theta})\big) = P\big(\overline{\tau}(\mathbf{X}) \leq \tau(\boldsymbol{\theta})\big) = \frac{\alpha}{2}. \tag{10.9.6}$$

Odtud plyne, že kritický obor W_3 definovaný vztahem

$$W_3 = \{\mathbf{x} \,|\, \underline{\tau}(\mathbf{x}) \geq \tau_0\} \cup \{\mathbf{x} \,|\, \overline{\tau}(\mathbf{x}) \leq \tau_0\} \tag{10.9.7}$$

je kritickým oborem na hladině α pro oboustranný test hypotézy H_3.

Všechny testy konkrétních hypotéz uvedené v čl. 11 jsou právě tohoto typu a doporučujeme čtenáři, aby si alespoň některé z nich odvodil z výsledků odst. 8.2.

Kapitola 11

Některé důležité testy

V tomto článku jsou uvedeny testy běžných hypotéz o parametrech nejdůležitějších rozdělení popsaných v [24]. Odstavec má dvojí účel: jednak ilustrovat základní pojmy z čl. 10, jednak dát návod k řešení nejčastěji se vyskytujících úloh testování statistických hypotéz.

11.1 Testy hypotéz o parametru alternativního rozdělení.

Budiž $\mathbf{X} = (X_1, \ldots, X_n)'$ náhodný výběr z alternativního rozdělení $\mathrm{A}(\pi)$. To znamená, že $X_i, i = 1, \ldots, n$, jsou „zakódované“ výsledky n nezávislých náhodných pokusů, když se „úspěch“ v i-tém pokusu zaznamená jako $X_i = 1$ a „neúspěch“ jako $X_i = 0$.

Kritický obor pro test jednostranné hypotézy $H : \pi \leq \pi_0$ (π_0 dané číslo) proti jednostranné alternativě A: $\pi > \pi_0$ je

$$W = \{\mathbf{x} \,|\, \sum_{i=1}^{n} x_i \geq c\}, \tag{11.1.1}$$

kde c je číslo určené tak, aby

$$P\big(\sum_{i=1}^{m} X_i \geq c \,|\, \pi = \pi_0\big) = \alpha. \tag{11.1.2}$$

Náhodná veličina $\sum_{i=1}^{n} X_i$, tj. celkový počet úspěchů v sérii n nezávislých pokusů, má rozdělení $\mathrm{Bi}(n, \pi)$ (viz [24], odst. 14.5). Číslo c by tedy mělo být

určeno tak, aby

$$\sum_{j=c}^{n} \binom{n}{j} \pi_0^j (1-\pi_0)^{n-j} = \alpha. \tag{11.1.3}$$

Číslo c splňující rovnici (11.1.3) však nelze pro libovolnou dvojici π_0, α najít. Volí se tedy buď takové c, při kterém

$$\sum_{j=c}^{n} \binom{n}{j} \pi_0^j (1-\pi_0)^{n-j} \le \alpha < \sum_{j=c-1}^{n} \binom{n}{j} \pi_0^j (1-\pi_0)^{n-j}, \tag{11.1.4}$$

nebo takové c, při kterém je

$$\sum_{j=c}^{n} \binom{n}{j} \pi_0^j (1-\pi_0)^{n-j}$$

nejbližší předepsané hladině významnosti α.

Ke stanovení čísla c lze užít tabulek rozdělení $\mathrm{Bi}(n, \pi)$ (např. [23]). Alternativní postup je tento: zamítnout hypotézu H, když

$$\frac{n - \sum_{i=1}^{n} x_i + 1}{\sum_{i=1}^{n} x_i} \frac{\pi_0}{1-\pi_0} \le F_\alpha\Big(2\sum_{i=1}^{n} x_i,\ 2(n - \sum_{i=1}^{n} x_i + 1)\Big), \tag{11.1.5}$$

kde $F_\alpha(\nu_1, \nu_2)$ je $100\alpha\%$ kvantil rozdělení F s ν_1 a ν_2 stupni volnosti (viz (3.4.3) a [23]). Ekvivalence (11.1.4) a (11.1.5) se dokáže užitím výsledků odst. 3.6.1 této práce a odst. 23.3 v [24].

Při velkých hodnotách n a $\sum_{i=1}^{n} x_i$ lze užít aproximace rozdělení veličiny $\frac{1}{n}\sum_{i=1}^{n} X_i$ normálním rozdělením podle odst. 26.3 v [24]; podmínka (11.1.4) se pak nahradí podmínkou

$$1 - \Phi\left(\frac{c/n - \pi_0}{\sqrt{\pi_0(1-\pi_0)}}\sqrt{n}\right) \doteq \alpha, \tag{11.1.6}$$

odkud

$$c \doteq \pi_0 + u_{1-\alpha}\sqrt{\frac{\pi_0(1-\pi_0)}{n}}. \tag{11.1.7}$$

11.2 Příklady.

11.2.1

Ve výběru $n = 50$ výrobků byl nalezeno 6 vadných. Není tento výsledek v příkrém rozporu s předpokladem, že daný technologický postup produkuje v průměru ne více než 5% vadných kusů?

Náhodné vybrání $n = 50$ kusů ke kontrole představuje $n = 50$ náhodných pokusů, v nichž jev $X_i = 1$ (i-tý vybraný kus je vadný) má neznámou pravděpodobnost π. Je třeba testovat hypotézu $H : \pi \leq \pi_0 = 0,05$. V našem případě bylo pozorováno $\sum_{i=1}^{50} x_i = 6$ čili

$$\frac{50 - 6 + 1}{6} \frac{0,05}{0,95} \doteq 0,395$$

a 5% kvantil rozdělení F(12, 90) je $F_{0,05}(12, 90) = 1/F_{0,95}(90, 12) \doteq 0,425$ (z tabulek [23]). Hypotézu $H : \pi \leq 0,05$ tedy zamítáme a usuzujeme, že průměrný podíl vadných kusů při daném výrobním procesu je vyšší než 5%.

11.2.2 Regulace výrobního procesu při kontrole posuzováním.

Dlouhodobým sledováním výrobního procesu je zjištěno, že při ustálených výrobních podmínkách vznikají přibližně 4% vadných kusů při výrobě jednoduchého výrobku (např. výlisky z umělé hmoty). Za účelem kontroly, zda se nezhoršily výrobní podmínky a průměrné procento vadných kusů nestouplo, odebírají se v pravidelných intervalech (např. každé dvě hodiny) náhodné výběry $n = 10$ výrobků. Jestliže počet vadných kusů ve výběru je roven nebo překročí číslo c, usuzuje se na zhoršení podmínek a podnikají se kroky k nápravě.

Popsaný postup je vlastně test hypotézy „pravděpodobnost π výroby vadného kusu $\leq 0,04 = \pi_0$“, prováděný každé dvě hodiny pomocí výběru rozsahu $n = 10$ z rozdělení A(π). Počet vadných kusů ve výběru, $T = \sum_{i=1}^{10} X_i$, kde $X_i = 1$, když i-tý vybraný kus je vadný, má binomické rozdělení Bi(10, π). Aby pravděpodobnost „planého poplachu“, tj. pravděpodobnost zamítnutí testované hypotézy $H : \pi \leq 0,04$, byla nejvýše rovna číslu α, musí c být voleno tak, aby

$$\sum_{j=c}^{10} \binom{10}{j} 0,04^j \cdot 0,96^{10-j} \leq \alpha < \sum_{j=c-1}^{10} \binom{10}{j} 0,04^j \cdot 0,96^{10-j}.$$

V tabulkách [23] nalezneme pro $\alpha = 0,05$

$$\sum_{j=3}^{10} \binom{10}{j} 0,04^j \cdot 0,96^{10-j} < 0,05 < \sum_{j=2}^{10} \binom{10}{j} 0,04^j \cdot 0,96^{10-j}.$$

odkud při $\alpha = 0,05$ je $c = 3$. Síla testu vůči alternativní hypotéze $\pi = 0,15$ (tj. pravděpodobnost odhalení takové změny výrobních podmínek, že pravděpodobnost výroby vadného kusu, čili průměrný podíl vadných kusů, je $0,15$ místo $0,04$), je rovna (opět z tabulek [23])

$$\sum_{j=3}^{10} \binom{10}{j} 0,15^j \cdot 0,85^{10-j} \doteq 0,179\,8.$$

11.3 Testy hypotéz o parametru Poissonova rozdělení.

Budiž $\mathbf{X} = (X_1, \ldots, X_n)'$ náhodný výběr z rozdělení $\mathrm{Po}(\lambda)$. Kritický obor pro test hypotézy $H : \lambda \leq \lambda_0$ (kde λ_0 je dané číslo) je

$$W = \{\mathbf{x} \,|\, \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{x_i} \geq \mathbf{c}\}, \tag{11.3.1}$$

kde c je určeno tak, aby

$$P\big(\sum_{i=1}^{n} X_i \geq c \,|\, \lambda = \lambda_0\big) \leq \alpha < P\big(\sum_{i=1}^{n} X_i \geq c - 1 \,|\, \lambda = \lambda_0\big). \tag{11.3.2}$$

Statistika $T = \sum_{i=1}^{n} X_i$ má rozdělení $\mathrm{Po}(n\lambda)$. Číslo c se tedy určí ze vztahu

$$\sum_{j=c}^{\infty} \frac{(n\lambda_0)^j}{j!} \mathrm{e}^{-n\lambda_0} \leq \alpha < \sum_{j=c-1}^{\infty} \frac{(n\lambda_0)^j}{j!} \mathrm{e}^{-n\lambda_0} \tag{11.3.3}$$

buď s užitím tabulek Poissonova rozdělení nebo pomocí počítače.

Jinak lze provést test tak, že zamítneme hypotézu H, jestliže

$$\chi^2_\alpha\Big(2\sum_{i=1}^{n} x_i\Big) \geq 2n\lambda_0, \tag{11.3.4}$$

kde $\chi^2_\alpha(\nu)$ značí $100\alpha\%$ kvantil rozdělení $\chi^2(\nu)$.

Ekvivalence nerovností (11.3.3) a (11.3.4) plyne ze vztahu mezi rozdělením χ^2, rozdělením gama a Poissonovým rozdělením, viz [24], odst. 22.4. Při vysokých hodnotách n a λ_0 (aspoň $n\lambda_0 > 30$) lze aproximovat Poissonovo rozdělení rozdělením normálním a podmínka (11.3.3) přejde v podmínku

$$\Phi\left(\frac{c - n\lambda_0}{\sqrt{n\lambda_0}}\right) \doteq 1 - \alpha, \tag{11.3.5}$$

odkud

$$c \doteq n\lambda_0 + u_{1-\alpha}\sqrt{n\lambda_0}. \tag{11.3.6}$$

Hypotéza $H : \lambda \leq \lambda_0$ se tedy zamítá, když

$$\sum_{i=1}^{n} x_i \geq n\lambda_0 + u_{1-\alpha}\sqrt{n\lambda_0} \quad \text{čili} \quad \overline{x} \geq \lambda_0 + u_{1-\alpha}\sqrt{\frac{\lambda_0}{n}}. \tag{11.3.7}$$

Síla testu proti alternativní hypotéze $\lambda = \lambda_1$ je rovna

$$P(\lambda_1) = \sum_{j=c}^{\infty} \frac{(n\lambda_1)^j}{j!} \mathrm{e}^{-n\lambda_1} = P\big(\chi^2(2c) \leq 2n\lambda_1\big) \tag{11.3.8}$$

a při $n\lambda_1 > 30$ přibližně

$$P(\lambda_1) \doteq 1 - \Phi\left(\frac{c - n\lambda_1}{\sqrt{n\lambda_1}}\right) = \Phi\left(\frac{\lambda_1 - \lambda_0}{\sqrt{n\lambda_1}}\sqrt{n} - u_{1-\alpha}\sqrt{\frac{\lambda_0}{\lambda_1}}\right). \tag{11.3.9}$$

11.4 Příklad.

Dlouhodobým sledováním je zjištěno, že při ustálených výrobních podmínkách má počet kazů na jeden běžný metr bavlněné tkaniny určité šíře Poissonovo rozdělení. Při prohlídce 15 metrů bylo nalezeno celkem 7 kazů. Je udržitelný předpoklad, že průměrný počet kazů na 1 metr nepřekračuje $\lambda_0 = 0,2$ (tj. že v průměru nepřipadá více než 1 kaz na každých 5 metrů)?

Vybraných 15 metrů představuje náhodný výběr rozsahu $n = 15$ z Poissonova rozdělení s neznámou střední hodnotou λ. Je třeba ověřit hypotézu $H : \lambda \leq \lambda_0 = 0,2$. Pozorovaná hodnota statistiky $T = \sum_{i=1}^{15} X_i$ je v dané úloze rovna 7. Z tabulek [23] zjišťujeme $\chi^2_{0,05}(14) = 6,75 > 30 \cdot 0,2 = 6$. Hypotézu $H :\leq 0,2$ tedy při testování na hladině významnosti $\alpha = 0,05$ zamítáme, tj. usuzujeme, že střední hodnota počtu kazů na 1 metr tkaniny je větší než $0,2$.

11.5 Test hypotézy o parametru exponenciálního rozdělení.

Nechť $\mathbf{X} = (X_1, \ldots, X_n)'$ je náhodný výběr z rozdělení $\mathrm{E}(0,\, \delta)$. Test hypotézy $H : \delta \leq \delta_0$ (δ_0 dané číslo) je založen na kritickém oboru tvaru

$$W = \{\mathbf{x} \,|\, \sum_{i=1}^{n} x_i \geq c\}, \tag{11.5.1}$$

kde c je určeno tak, aby

$$P\Big(\sum_{i=1}^{n} X_i \geq c \,\Big|\, \delta = \delta_0\Big) = \alpha. \tag{11.5.2}$$

Podle [24], odst. 22.4, má statistika $T = \sum_{i=1}^{n} X_i$ Erlangovo rozdělení a jednoduchou transformací lze ukázat, že náhodná veličina

$$\frac{2T}{\delta} = \frac{2}{\delta}\sum_{i=1}^{n} X_i \tag{11.5.3}$$

má rozdělení $\chi^2(2n)$. Odtud plyne, že požadavek (11.5.2), který lze zapsat jako

$$P\left(\frac{2}{\delta_0}\sum_{i=1}^{n} X_i \geq \frac{2c}{\delta_0} \,\middle|\, \delta = \delta_0\right) = \alpha \tag{11.5.4}$$

bude splněn při

$$c = \frac{1}{2}\delta_0 \chi^2_{1-\alpha}(2n). \tag{11.5.5}$$

Síla testu vůči alternativní hypotéze $\delta = \delta_1$ je rovna

$$\begin{aligned} P(\delta_1) &= P\Big(\sum_{i=1}^{n} X_i \geq \delta_0 \frac{1}{2}\chi^2_{1-\alpha}(2n) \,\Big|\, \delta = \delta_1\Big) = \\ &= P\left(\frac{2\sum_{i=1}^{n} X_i}{\delta_1} \geq \frac{\delta_0}{\delta_1}\chi^2_{1-\alpha}(2n) \,\middle|\, \delta = \delta_1\right) = \\ &= P\Big(\chi^2(2n) \geq \frac{\delta_0}{\delta_1}\chi^2_{1-\alpha}(2n)\Big). \end{aligned} \tag{11.5.6}$$

Test pravostranné hypotézy $H : \delta \geq \delta_0$ (δ_0 dané číslo) proti levostranné alternativě $A : \delta < \delta_0$ bude mít kritický obor

$$W = \{\mathbf{x} \,|\, \sum_{i=1}^{n} x_i \leq \frac{1}{2}\delta_0 \chi^2_\alpha(2n)\} \tag{11.5.7}$$

a jeho síla vůči alternativě $\delta = \delta_1$ je rovna

$$P(\delta_1) = P\Big(\chi^2(2n) \leq \frac{\delta_0}{\delta_1}\chi^2_\alpha(2n)\Big). \tag{11.5.8}$$

11.6 Příklad.

Předpokládejme, že o určitém výrobku je známo, že jeho doba života má rozdělení E(0, δ), kde parametr δ závisí na konkrétních podmínkách výroby a provozu. Zkouškami se ověřuje, zda aspoň 90% výrobků má dobu života delší než 200 hodin, tj. zda desetiprocentní kvantil $x_{0,1}$ rozdělení doby života není menší než 200 hodin. Jde tedy o test hypotézy

$$H : x_{0,1} \geq 200.$$

Kvantil $x_{0,1}$ rozdělení E(0, δ) je roven (viz [24], odst. 20.2) - $\delta \ln 0,9$, takže hypotéza vlastně zní

$$H : -\delta \ln 0,9 \geq 200 \equiv \delta \geq \delta_0 = -\frac{200}{\ln 0,9}.$$

Podle (11.5.7) se hypotéza H zamítne, jestliže

$$\sum_{i=1}^{n} x_i \leq \frac{1}{2}\delta_0 \chi^2_\alpha(2n),$$

kde $x_1, \ldots, x_n$ jsou pozorované doby života u n výrobků vybraných ke zkoušce. Je-li např. $n = 20$ a testuje-li se na hladině významnosti $\alpha = 0,05$, zamítne se H při

$$\sum_{i=1}^{20} x_i \leq \frac{1}{2}\Big(-\frac{200}{\ln 0,9}\Big)\chi^2_{0,05}(40) = 25\,160,28$$

čili při

$$\overline{x} \leq 1\,258,014.$$

Pravděpodobnost, že tento test odhalí sníženou životnost, když desetiprocentní kvantil bude ve skutečnosti jen 150 hodin, tj. $\delta = \delta_1 = -\frac{150}{\ln 0{,}9} \doteq 1\,423\,68$, je rovna

$$P(150) = P\Big(\chi^2(40) \leq \frac{200}{150}\chi^2_{0,05}(40)\Big) = P\big(\chi^2(40) \leq 34,35\big) \doteq 0,31.$$

11.7 Testy hypotéz o parametrech normálního rozdělení.

11.7.1 Test hypotézy o střední hodnotě.

Nechť $\mathbf{X} = (X_1, \ldots, X_n)'$ je náhodný výběr z rozdělení $\mathrm{N}(\mu, \sigma^2)$. Pro testování hypotézy $H : \mu \leq \mu_0$ proti alternativní hypotéze $A : \mu > \mu_0$ se nejčastěji užívá kritického oboru

$$W = \big\{\mathbf{x} \,\big|\, \overline{x} \geq \mu_0 + t_{1-\alpha}(n-1)\frac{s}{\sqrt{n}}\big\} = \big\{\mathbf{x}, \,\big|\, \frac{\overline{x} - \mu_0}{s}\sqrt{n} \geq t_{1-\alpha}(n-1)\big\}, \tag{11.7.1}$$

kde $t_{1-\alpha}(n-1)$ je $100(1-\alpha)\%$ kvantil rozdělení t o $n-1$ stupních volnosti a $\overline{X}$ a S statistiky definované v odst. 2.2.

K testování hypotézy $H : \mu \geq \mu_0$ proti $A : \mu < \mu_0$ se užije kritického oboru

$$W = \big\{\mathbf{x} \,\big|\, \overline{x} \leq \mu_0 - t_{1-\alpha}(n-1)\frac{s}{\sqrt{n}}\big\} \tag{11.7.2}$$

a k testování hypotézy $H : \mu = \mu_0$ proti dvoustranné alternativní hypotéze $A : \mu \neq \mu_0$ oboru

$$W = \big\{\mathbf{x} \,\big|\, |\overline{x} - \mu_0| \geq t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}}\big\}. \tag{11.7.3}$$

Je-li rozptyl σ^2 znám, užije se v (11.7.1) až (11.7.3) místo s skutečné směrodatné odchylky σ a kvantil rozdělení $\mathrm{t}(n-1)$ se nahradí kvantilem rozdělení $\mathrm{N}(0, 1)$.

11.7.2 Test hypotézy o rozptylu.

Nechť $\mathbf{X} = (X_1, \ldots, X_n)'$ je náhodný výběr z rozdělení $\mathrm{N}(\mu, \sigma^2)$. K testování hypotézy $H : \sigma^2 \leq \sigma_0^2$ (σ_0^2 dané číslo) proti alternativní hypotéze $A : \sigma^2 > \sigma_0^2$

se používá statistiky S^2 definované v odst. 2.2; kritický obor je

$$W = \Big\{\mathbf{x}\,\Big|\,\frac{\sum_{i=1}^n (x_i-\overline{x})^2}{\sigma_0^2} \geq \chi^2_{1-\alpha}(n-1)\Big\} = \Big\{\mathbf{x}\,\Big|\,\frac{s^2}{\sigma_0^2} \geq \frac{1}{n-1}\chi^2_{1-\alpha}(n-1)\Big\}. \tag{11.7.4}$$

Síla tohoto testu proti alternativě $\sigma^2 = \sigma_1^2$ je rovna

$$\begin{aligned} P(\sigma_1^2) &= P\left(\frac{\sum_{i=1}^n (X_i-\overline{X})^2}{\sigma_0^2} \geq \chi^2_{1-\alpha}(n-1)\,\Big|\,\sigma^2=\sigma_1^2\right) = \\ &= P\left(\frac{\sum_{i=1}^n (X_i-\overline{X})^2}{\sigma_1^2} \geq \frac{\sigma_0^2}{\sigma_1^2}\chi^2_{1-\alpha}(n-1)\,\Big|\,\sigma^2=\sigma_1^2\right) = \\ &= P\Big(\chi^2(n-1) \geq \frac{\sigma_0^2}{\sigma_1^2}\chi^2_{1-\alpha}(n-1)\Big). \end{aligned} \tag{11.7.5}$$

11.8 Příklady.

11.8.1

Pro kontrolu správnosti nastavení měřicího přístroje bylo provedeno 10 měření zkušebního etalonu se správnou hodnotou $\mu_0 = 15,20$. Byly získány tyto výsledky:

$$15,23;\ 15,21;\ 15,19;\ 15,16;\ 15,26;\ 15,22;\ 15,23;\ 15,26;\ 15,23;\ 15,29.$$

Lze považovat pozorované odchylky od správné hodnoty za náhodné chyby měření nebo je důvod k podezření na přítomnost systematické chyby? (Předpokládáme, že jde o měření, při kterém náhodné chyby mají normální rozdělení). Jelikož je žádoucí odhalit jak zápornou, tak kladnou systematickou chybu (odchýlení střední hodnoty výsledku měření od správné hodnoty), použijeme testu hypotézy $H : \mu = \mu_0$ proti oboustranné alternativní hypotéze $A : \mu \neq \mu_0$ podle (11.7.3). Je

$$\begin{aligned} x = 15,2 + \big[&0,03 + 0,01 + (-0,01) + (-0,04) + 0,06 + 0,02 + 0,03 + \\ &+ 0,06 + 0,03 + 0,09\big]/10 = 15,2 + 0,028 = 15,228; \end{aligned}$$

$$s^2 = \frac{1}{9}\Big[\sum_{i=1}^{9}(x_i - 15,2)^2 - 10(\overline{x} - 15,2)^2\Big] =$$
$$= \frac{1}{9}\big[0,000\,1(9+1+1+16+36+4+9+36+9+81) - 0,007\,84\big] =$$
$$= 0,001\,373;$$

$$\frac{s}{\sqrt{10}} \doteq 0,011\,72; \qquad t_{0,975}(9) = 2,262\,2;$$

$$|\overline{x} - \mu_0| = |15,228 - 15,2| = 0,028 > 0,026\,5 \doteq 2,262\,2 \cdot 0,011\,72.$$

Hypotézu $H : \mu = 15,2$ tedy zamítáme, tj. usuzujeme na přítomnost systematické chyby.

11.8.2

Pro bavlněnou přízi určitého druhu je předepsána horní mez variability pevnosti žádá se, aby směrodatná odchylka pevnosti vlákna nepřekročila hodnotu $\sigma_0 = 0,6\,\mathrm{kg}$, jinak vznikají potíže při tkaní. Pevnost má přibližně normální rozdělení $\mathrm{N}(\mu, \sigma^2)$. Při zkoušce $n = 16$ vzorků byly zjištěny tyto výsledky: 2,22; 3,54; 2,37; 1,66; 4,74; 4,82; 3,21; 5,44; 3,23; 4,79; 4,85; 4,05; 3,48; 3,89; 4,90; 5,37.

Je důvod k podezření na vyšší nestejnoměrnost, než je stanoveno?

Důvodné podezření na vyšší nestejnoměrnost vzniká, když test zamítne hypotézu $H : \sigma \le \sigma_0 = 0,6$. Pro uvedené výsledky je

$$\overline{x} = 3,91; \quad \sum_{i=1}^{1} 5(x_i - x)^2 = 20,212; \quad s^2 = 1,347\,5; \quad s \doteq 1,160\,8;$$

$$\frac{\sum_{i=1}^{15}(x_i - \overline{x})^2}{\sigma_0^2} = \frac{20,212}{0,36} \doteq 56,144\,4,$$

což je hodnota o mnoho větší než $\chi^2_{0,95}(15) = 24,996$ (z tabulek [23]). Hypotézu $H : \sigma \le 0,6$ tedy zamítáme na hladině významnosti $0,05$; výsledky potvrzují vyšší nestejnoměrnost, než žádá norma.

Konfidenční interval pro σ s koeficientem spolehlivosti $0,95$ je (dosazením $s = 1,160\,8,\ n = 15$ a $\frac{\alpha}{2} = 0,025$ do (8.4.8)) (0,837 5; 1,796 6).

11.9 Dva nezávislé náhodné výběry ze dvou normálních rozdělení.

Nechť $\mathbf{X} = (X_1, \ldots, X_{n1})'$ je náhodný výběr rozsahu n_1 z rozdělení $\mathrm{N}(\mu_1, \sigma_1^2)$ a $\mathbf{Y} = (Y_1, \ldots, Y_{n2})'$ náhodný výběr rozsahu n_2 z rozdělení $\mathrm{N}(\mu_2, \sigma_2^2)$. Nechť výběry $\mathbf{X}$ a $\mathbf{Y}$ jsou nezávislé. Uvažujme statistiky $\overline{X}$, $\overline{Y}$, S_1^2 a S_2^2 dané výrazy (3.5.1) a dále statistiku S^2 danou výrazem (3.5.3).

11.9.1 Testy hypotéz o středních hodnotách při stejných rozptylech.

Jsou-li rozptyly obou rozdělení stejné, $\sigma_1^2 = \sigma_2^2$ (nebo aspoň přibližně stejné), užívá se k testování hypotézy $H_1 : \mu_1 \leq \mu_2$ proti alternativní hypotéze $A : \mu_1 > \mu_2$ kritického oboru

$$W_1 = \left\{ (\mathbf{x}, \mathbf{y}) \;\middle|\; \frac{\overline{x} - \overline{y}}{s} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \geq t_{1-\alpha}(n_1 + n_2 - 2) \right\}, \tag{11.9.1}$$

k testování hypotézy $H_2 : \mu_2 \leq \mu_1$ proti $A_2 : \mu_2 > \mu_1$ obdobně

$$W_2 = \left\{ (\mathbf{x}, \mathbf{y}) \;\middle|\; \frac{\overline{y} - \overline{x}}{s} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \geq t_{1-\alpha/2}(n_1 + n_2 - 2) \right\} \tag{11.9.2}$$

k testování hypotézy $H_3 : \mu_1 = \mu_2$ proti oboustranné alternativní hypotéze $A_3 : \mu_1 \neq \mu_2$ kritického oboru

$$W_3 = \left\{ (\mathbf{x}, \mathbf{y}) \;\middle|\; \frac{|\overline{x} - \overline{y}|}{s} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \geq t_{1-\alpha}(n_1 + n_2 - 2) \right\}, \tag{11.9.3}$$

kde $t_{1-\alpha}(n_1+n_2-2)$ a $t_{1-\alpha/2}(n_1+n_2-2)$ jsou kvantily rozdělení $\mathrm{t}(n_1+n_2-2)$.

Z (3.5.4) vyplývá, že v případě $\mu_1 = \mu_2$ má statistika

$$\frac{(\overline{X} - \overline{Y}(n_1 n_2)^{\frac{1}{2}}}{S(n_1 + n_2)^{\frac{1}{2}}}$$

rozdělení $\mathrm{t}(n_1 + n_2 - 2)$.

Je-li tedy ve skutečnosti $\mu_1 = \mu_2$, pak pravděpodobnost zamítnutí hypotéz H_1, H_2 i H_3 je právě rovna zvolené hladině významnosti α.

11.9.2 Testy hypotéz o středních hodnotách při nestejných rozptylech.

Jsou-li důvody k domněnce, že rozptyly σ_1^2 a σ_2^2 se mezi sebou značně liší, nahradí se kritický obor (11.9.1) kritickým oborem

$$W = \left\{ (\mathbf{x}, \mathbf{y}) \,\middle|\, \frac{\overline{x} - \overline{y}}{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^{\frac{1}{2}}} \geq t_{1-\alpha}(\nu) \right\}, \tag{11.9.4}$$

kde ν je dáno vztahem (viz např. [16], str. 174)

$$\frac{1}{\nu} = \frac{c^2}{n_1 - 1} + \frac{(1-c)^2}{n_2 - 1}, \tag{11.9.5}$$

přičemž

$$c = \frac{s_1^2}{n_1} \left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^{-1}. \tag{11.9.6}$$

Kritické obory (11.9.2) a (11.9.3) se upraví obdobně.

11.9.3 Testy hypotéz o rozptylech.

K testování hypotézy $H : \sigma_1^2 \leq \sigma_2^2$ proti alternativní hypotéze $A : \sigma_1^2 > \sigma_2^2$ se užívá kritického oboru

$$W = \Big\{ (\mathbf{x}, \mathbf{y}) \,\Big|\, \frac{s_1^2}{s_2^2} \geq F_{1-\alpha}(n_1 - 1, n_2 - 1) \Big\}, \tag{11.9.7}$$

kde $F_{1-\alpha}(n_1 - 1, n_2 - 1)$ je $100(1-\alpha)\%$ kvantil rozdělení $F(n_1 - 1, n_2 - 1)$. Síla tohoto testu proti alternativě $\sigma_1^2 = \lambda \sigma_2^2$ je rovna

$$\begin{aligned} P(\lambda) &= P\left(\frac{S_1^2}{S_2^2} \geq F_{1-\alpha}(n_1 - 1, n_2 - 1) \,\middle|\, \sigma_1^2 = \lambda \sigma_2^2 \right) = \\ &= P\big(F(n_1 - 1, n_2 - 1) \geq \frac{1}{\lambda} F_{1-\alpha}(n_1 - 1, n_2 - 1) \big). \end{aligned} \tag{11.9.8}$$

K testování hypotézy $H : \sigma_1^2 = \sigma_2^2$ proti oboustranné alternativě $A : \sigma_1^2 \neq \sigma_2^2$ se užije kritického oboru

$$W = \Big\{ (\mathbf{x}, \mathbf{y}) \,\Big|\, \frac{\max(s_1^2, s_2^2)}{\min(s_1^2, s_2^2)} \geq F_{1-\alpha/2}(\nu_1, \nu_2) \Big\}, \tag{11.9.9}$$

kde ν_1 je počet stupňů volnosti příslušný k většímu z odhadů s_1^2, s_2^2 a ν_2 je počet stupňů volnosti příslušný k menšímu z nich, tj.

$$\nu_1 = n_1 - 1, \quad \nu_2 = n_2 - 1, \quad \text{jestliže} \;\; s_1^2 > s_2^2,$$

$$\nu_1 = n_2 - 1, \quad \nu_2 = n_1 - 1, \quad \text{jestliže} \;\; s_1^2 < s_2^2.$$

11.10 Příklady.

11.10.1

Bylo provedeno $n_1 = 5$ nezávislých stanovení obsahu SiO_2 v martinské strusce analyticky a $n_2 = 6$ nezávislých stanovení obsahu SiO_2 v téže strusce fotokolorimetricky s těmito výsledky (viz [28]):

analytická stanovení (x_i): $20,1;\ 19,6;\ 20,0;\ 19,9;\ 20,1$
fotokolorimetrická stanovení (y_i): $20,9;\ 20,1;\ 20,6;\ 20,5;\ 20,7;\ 20,5$.

Dosavadní zkušenosti s oběma metodami ukazují, že mezi rozptyly výsledků není podstatného rozdílu. Má se ověřit, zda u vzorků z téhož materiálu dávají obě metody v průměru stejné výsledky.

Pro uvedená data je

$$\overline{x} = 19,94; \quad \overline{y} = 20,55; \quad s_1^2 = \frac{0,172}{4} = 0,043; \quad s_2^2 = \frac{0,355}{5} = 0,071;$$

$$s^2 = \frac{0,172 + 0,355}{9} \doteq 0058\,6; \quad s \doteq 0,242;$$

$$\frac{|\overline{x} - \overline{y}|}{s}\sqrt{\frac{30}{5+6}} \doteq \frac{0,61}{0,242} \cdot 1,651\,4 \doteq 4,163 > 2,262\,2 = t_{0,975}(9).$$

Hypotézu, že střední hodnota výsledku kolorimetrického měření je rovna (při stejném analyzovaném materiálu) střední hodnotě analytického stanovení, zamítáme; mezi metodami patrně je systematický rozdíl.

11.10.2

Bylo provedeno po 18 zkouškách pevnosti v tahu na vzorcích dvou různých kabelů. Souhrnné výsledky jsou (viz [15]):

kabel druhu 1: $\overline{x} = 345,61; \quad s_1^2 = 11,90,$
kabel druhu 2: $\overline{y} = 340,50; \quad s_2^2 = 35,91.$

Ověřme,

a) zda lze považovat pozorovaný rozdíl ve variabilitě výsledků za náhodný (tj. zda je přijatelný předpoklad, že rozptyly pevnosti obou druhů kabelu jsou stejné),

b) zda střední hodnota pevnosti obou druhů kabelu je stejná.

Obě úlohy řešíme pomocí testu proti oboustranné alternativě; žádný z kabelů není označen jako experimentální, u kterého očekáváme menší variabilitu nebo větší průměrnou pevnost.

Při srovnání rozptylů postupujeme podle (11.9.9):

$$\frac{\max(s_1^2, s_2^2)}{\min(s_1^2, s_2^2)} = \frac{35,91}{11,90} = 3,018.$$

Pozorovaná hodnota je větší než $F_{0,975}(17,17) = 2,673\,3$. Hypotézu o rovnosti rozptylů tedy zamítáme. Při srovnání středních hodnot postupujeme podle 11.9.2:

$$\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} = \frac{11,90}{18} + \frac{35,91}{18} \doteq 2,656\,1; \quad c = 0,248\,9;$$

$$\nu = \left(\frac{0,061\,95}{17} + \frac{0,564\,15}{17} \right)^{-1} \doteq 27,15,$$

$$\frac{|\overline{x} - \overline{y}|}{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^{\frac{1}{2}}} = \frac{5,11}{1,629\,8} \doteq 3,135 > t_{0,975}(27) = 2,052;$$

i hypotézu o rovnosti středních hodnot zamítáme. Závěr zní: Kabel druhého typu má nižší pevnost a větší rozptyl pevnosti.

11.11 Dvourozměrné normální rozdělení.

Nechť $(X_1, Y_1)', \ldots, (X_n, Y_n)'$ je náhodný výběr z dvourozměrného normálního rozdělení s parametry $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho$.

11.11.1 Test hypotézy o rozdílu středních hodnot.

Obdobně jako v příkl. 8.4.6 uvažujme veličiny $D_i = X_i - Y_i,\ i = 1, \ldots, n$; pak $\mathbf{D} = (D_1, \ldots, D_n)'$ je náhodný výběr z rozdělení $\mathrm{N}(\Delta,\, \sigma_D^2)$, kde $\Delta = \mu_1 - \mu_2$ a σ_D^2 je dáno výrazem (3.7.2).

Pro testování hypotézy $H : \Delta \leq 0$ proti alternativní hypotéze $A : \Delta > 0$ užijeme kritického oboru

$$W = \Big\{(\mathbf{x}, \mathbf{y}) \,\Big|\, \overline{d} \geq t_{1-\alpha(n-1)} \frac{s_D}{\sqrt{n}}\Big\}, \tag{11.11.1}$$

kde $\overline{D}$ a S_D^2 jsou statistiky (3.7.3).

K testování hypotézy $H : \Delta \geq 0$ proti $A : \Delta < 0$ se užije kritického oboru

$$W = \Big\{(\mathbf{x}, \mathbf{y}) \,\Big|\, \overline{d} \leq -t_{1-\alpha}(n-1) \frac{s_D}{\sqrt{n}}\Big\} \tag{11.11.2}$$

a k testování hypotézy $H : \Delta = 0$ (tj. $H : \mu_1 = \mu_2$) proti oboustranné alternativní hypotéze $A : \Delta \neq 0$ kritického oboru

$$W = \Big\{(\mathbf{x}, \mathbf{y}) \,\Big|\, |\overline{d}| \geq t_{1-\alpha/2}(n-1) \frac{s_D}{\sqrt{n}}\Big\}. \tag{11.11.3}$$

11.11.2 Test hypotézy o koeficientu korelace.

Pro testování hypotézy $H : \rho = 0$ (tj. hypotézy nezávislosti veličin X a Y majících dvourozměrné normální rozdělení, viz [24], odst. 24.3) se používá výběrového koeficientu r definovaného v odst. 3.7; za platnosti hypotézy $H : \rho = 0$ má statistika (3.7.8) rozdělení $\mathrm{t}(n-2)$.

K testování hypotézy $H : \rho = 0$ proti alternativní hypotéze $A : \rho > 0$ použijeme kritického oboru

$$W = \Big\{(\mathbf{x}, \mathbf{y}) \,\Big|\, \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \geq t_{1-\alpha}(n-2)\Big\}, \tag{11.11.4}$$

k testování hypotézy $H : \rho = 0$ proti alternativní hypotéze $A : \rho < 0$ kritického oboru

$$W = \Big\{(\mathbf{x}, \mathbf{y}) \,\Big|\, \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \leq -t_{1-\alpha}(n-2)\Big\}, \tag{11.11.5}$$

a k testování hypotézy $H : \rho = 0$ proti oboustranné alternativní hypotéze $A : \rho \neq 0$ kritického oboru

$$W = \Big\{(\mathbf{x}, \mathbf{y}) \,\Big|\, \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}} \geq t_{1-\alpha}(n-2)\Big\}. \tag{11.11.6}$$

Pro testování hypotézy $H : \rho \leq \rho_0$ (ρ_0 dané číslo z intervalu $(-1, 1)$) proti alternativní hypotéze $A : \rho > 0$ použijeme pro $n \geq 10$, není-li $|\rho_0|$ příliš blízké jedné, kritického oboru

$$W = \big\{(\mathbf{x}, \mathbf{y}) \,\big|\, u \geq u_{1-\alpha}\big\}, \tag{11.11.7}$$

kde U je statistika

$$U = (n-3)^{\frac{1}{2}}\Big(\frac{1}{2}\ln\frac{1+r}{1-r} - \frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0} - \frac{\rho_0}{2(n-1)}\Big). \tag{11.11.8}$$

Obdobně k testování hypotézy $H : \rho \geq \rho_0$ proti alternativní hypotéze $A : \rho < \rho_0$ použijeme kritického oboru

$$W = \big\{(\mathbf{x}, \mathbf{y}) \,\big|\, u \leq -u_{1-\alpha}\big\}, \tag{11.11.9}$$

a k testování hypotézy $H : \rho = \rho_0$ proti oboustranné alternativě $A : \rho \neq \rho_0$ kritického oboru

$$W = \big\{(\mathbf{x}, \mathbf{y}) \,\big|\, |u| \geq u_{1-\alpha/2}\big\}. \tag{11.11.10}$$

Zde $u_{1-\alpha}$ a $u_{1-\alpha/2}$ jsou kvantily rozdělení $\mathrm{N}(0, 1)$.

11.11.3 Test hypotézy rovnosti rozptylů.

Uvažujme kromě veličin $D_i = X_i - Y_i$ ještě veličiny $B_i = X_i + Y_i$, $i = 1, \ldots, n$. V odstavci 3.7 jsme ukázali, že $(D_1, B_1)', \ldots, (D_n, B_n)'$ je náhodný výběr z dvourozměrného normálního rozdělení, jehož koeficient korelace $\rho(D, B)$ je dán výrazem (3.7.15). Je-li $\sigma_1^2 = \sigma_2^2$, je $\rho(D, B) = 0$.

Hypotéza $H : \sigma_1^2 = \sigma_2^2$ rovnosti rozptylů σ_1^2 a σ_2^2 v dvourozměrném normálním rozdělení $\mathrm{N}_2(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ náhodného vektoru $(X, Y)'$ je tedy

ekvivalentní hypotéze $H : \rho(D, B) = 0$, tj. hypotéze nulového koeficientu korelace v dvourozměrném normálním rozdělení náhodného vektoru $(D, B)'$.

Pro testování hypotézy $H : \sigma_1^2 = \sigma_2^2$ proti alternativní hypotéze $A : \sigma_1^2 > \sigma_2^2$ užijeme tedy kritického oboru (11.11.4), pro testování hypotézy H proti alternativní hypotéze $A : \sigma_1^2 < \sigma_2^2$ kritického oboru (11.11.5) a pro testování hypotézy H proti oboustranné alternativní hypotéze $A : \sigma_1^2 \neq \sigma_2^2$ kritického oboru (11.11.6), přičemž ve všech těchto kritických oborech je r statistika

$$r = \frac{n\sum_{i=1}^n D_i B_i - \left(\sum_{i=1}^n D_i\right)\left(\sum_{i=1}^n B_i\right)}{\sqrt{\left(n\sum_{i=1}^n D_i^2 - \left(\sum_{i=1}^n D_i\right)^2\right)\left(n\sum_{i=1}^n B_i^2 - \left(\sum_{i=1}^n B_i\right)^2\right)}}. \quad (11.11.11)$$

11.12 Příklady.

11.12.1

Byla zjišťována zralost viskózy X a tažnost umělého vlákna. Z $n = 46$ měření byl vypočten výběrový koeficient korelace $r = 0,638$. Testujme hypotézu $H : \rho \leq 0,5$ proti alternativní hypotéze $A : \rho > 0,5$ na hladině významnosti $\alpha = 0,05$.

Protože

$$u = 43^{\frac{1}{2}}\left(\frac{1}{2}\ln\frac{1,638}{0,362} - \frac{1}{2}\ln\frac{1,5}{0,5} - \frac{0,5}{90}\right) \doteq 1,311 < 1,645 = u_{0,95},$$

hypotézu H nezamítáme.

11.12.2

Uvažujme data tab. 8.1 a ověřme,

a) zda neexistuje systematická chyba mezi uvažovanými dvěma analytickými metodami,

b) zda lze předpokládat, že rozptyly obou metod jsou stejné.

V případě a) uvažujme $H : \Delta = 0$ a alternativní hypotézu $A : \Delta < 0$, v případě b) $H : \sigma_1^2 = \sigma_2^2$ a $A : \sigma_1^2 \neq \sigma_2^2$.

Protože

$$\overline{d} = -0,333 < -0,104 = -1,7959\sqrt{\frac{0,040\,6}{12}},$$

hypotézu $H : \Delta = 0$ zamítáme, tj. učiníme závěr, že první metoda dává systematicky nižší výsledky než metoda druhá.

Statistika (11.12.11) nabývá hodnoty

$$r = \frac{12 \cdot (-338,251) - (-4) \cdot 959,54}{\{[12 \cdot 1,779\,8 - (-4)^2][12 \cdot 80\,550,775 - 959 \cdot 54^2]\}^{\frac{1}{2}}} \doteq -0,445\,4.$$

Odtud

$$t = \frac{0,445\,4\sqrt{10}}{1 - 0,445\,4^2} \doteq 1,573 < 1,812\,5 = t_{0,975}(10),$$

takže hypotézu $H : \sigma_1^2 = \sigma_2^2$ nezamítáme.

Kapitola 12

Neparametrické testy

12.1 Parametrické a neparametrické testy.

Dosud jsme se zabývali testováním hypotéz o parametrech, případně parametrických funkcích jednotlivých rozdělení. Předpokládali jsme tedy, že máme náhodný výběr z rozdělení daného tvaru (např. z normálního rozdělení), které obsahuje jeden či více neznámých parametrů a na základě tohoto výběru jsme provedli test hypotézy. Takovéto testy se nazývají *parametrické testy.*

Neparametrické testy nevycházejí z předpokladu daného tvaru rozdělení, ale z obecnějších předpokladů. Často se pouze předpokládá, že rozdělení, z něhož náhodný výběr pochází, je rozdělení spojitého typu.

V řadě neparametrických testů se pozorování nahrazují jejich pořadími. Mějme náhodný výběr $\mathbf{X} = (X_1, \ldots, X_n)'$. Nechť X_1 je i_1-tá pořádková statistika, X_2 je i_2-tá pořádková statistika, $\ldots, X_n$ je i_n-tá pořádková statistika tohoto výběru. Pak R_1, tj. pořadí X_1, je rovno i_1; dále R_2, tj. pořadí X_2, je rovno $i_2, \ldots, R_n$, tj. pořadí X_n, je rovno i_n. Např. pro výběr uvažovaný v odst. 7.3 je:

$$R_1 = 4;\ R_2 = 1;\ R_3 = 6;\ R_4 = 3;\ R_5 = 8;\ R_6 = 5;\ R_7 = 7;\ R_8 = 2.$$

Někdy se může stát (nejčastěji vlivem zaokrouhlování výsledků měření), že dvě nebo více výběrových hodnot jsou stejné. Pak hovoříme o tzv. *shodách* a shodným hodnotám přiřazujeme jejich průměrné pořadí. Např. kdyby ve výběru v odst. 7.3 byla místo hodnoty 29 hodnota 22, dostali bychom:

$$R_1 = 4;\ R_2 = 1;\ R_3 = 6;\ R_4 = 2,5;\ R_5 = 8;\ R_6 = 5;\ R_7 = 7;\ R_8 = 2,5.$$

V tomto článku uvedeme základní neparametrické testy. Pro teoretické odvození těchto testů a jejich vlastností a pro jiné neparametrické testy odkazujeme na literaturu, např. [5, 14, 17, 21, 33].

12.2 Znaménkový test.

Mějme náhodný výběr $\mathbf{X} = (X_1, \ldots, X_n)'$ ze spojitého rozdělení, které má distribuční funkci $F(x)$. Pro medián $x_{0,5}$ tohoto rozdělení platí

$$P(X < x_{0,5}) = P(X > x_{0,5}) = 0,5. \tag{12.2.1}$$

Testujme hypotézu $H : x_{0,5} = x_0$ (x_0 je dané číslo). Označme

$$Z_i = X_i - x_0, \quad i = 1, \ldots, n, \tag{12.2.2}$$

a nechť Z je počet kladných rozdílů z těchto n rozdílů. Statistika Z nabývá hodnot $0, 1, \ldots, n$ a za platnosti hypotézy H má rozdělení Bi(n, 1/2).

Uvažujme alternativní hypotézu $A : x_{0,5} < x_0$ a hladinu významnosti $\alpha = 0,05$. Hypotézu H zamítáme, jestliže $Z \leq c$, kde c je číslo, pro které platí

$$\left(\frac{1}{2}\right)^n \sum_{t=0}^{c} \binom{n}{t} \leq \alpha < \left(\frac{1}{2}\right)^n \sum_{t=0}^{c+1} \binom{n}{t}. \tag{12.2.3}$$

V případě alternativní hypotézy $A : x_{0,5} > x_0$ hypotézu H zamítáme, jestliže $Z \geq n-c$, kde c je opět dáno vztahem (12.2.3) (vyplývá ze skutečnosti, že rozdělení Bi(n, 1/2) je symetrické).

V případě alternativní hypotézy $A : x_{0,5} \neq x_0$ se H zamítá, jestliže $Z \leq c$ nebo $Z \geq n - c$, kde c se určí ze vztahu (12.2.3), v němž se místo α uvažuje $\alpha/2$.

Hodnoty c lze určit z tabulek distribuční funkce rozdělení Bi(n, π) pro $\pi = 0,5$ (viz např. [23], tab. 24 pro $n \leq 50$). V [1], str. 333, jsou tabelovány hodnoty c pro alternativní hypotézu $A : x_{0,5} \neq x_0$ pro hladinu významnosti 0,05 a 0,01 (tj. hodnoty c splňující vztah (12.2.3) pro $\alpha = 0,025$ a $0,005$) a pro $n = 6$ (1) 80 (5) 100. V [23], tab. 29, jsou tabelovány hodnoty $c + 1$, kde c splňují vztahy (12.2.3) pro $\alpha \leq 0,05$ a $n = 5$ (1) 200.

Je-li (např. vlivem zaokrouhlování výsledků měření) z n rozdílů (12.2.2) r nulových, provede se test popsaným způsobem, přičemž se uvažuje rozsah výběru $n - r$ místo n.

Pro velká n lze rozdělení Bi(n, 1/2) aproximovat rozdělením N($n/2$, $n/4$) (viz [24], odst. 26.6). Platí-li hypotéza H, má pro velká n veličina

$$U = \frac{Z - n/2}{n/4} = \frac{2Z - n}{\sqrt{n}} \tag{12.2.4}$$

přibližně rozdělení N(0, 1), a lze jí tedy v případě velkých n užít pro testování uvedených hypotéz.

Mějme náhodný výběr $(X_1, Y_1)', \ldots, (X_n, Y_)'$ z dvourozměrného rozdělení. Uvažujme rozdíly $D_i = X_i - Y_i, i = 1, \ldots, n$. Pak veličiny $D_1, \ldots, D_n$ jsou vzájemně nezávislé a všechny mají stejnou distribuční funkci $F(d)$. Nechť $F(d)$ je spojitá distribuční funkce. Označme $d_{0,5}$ medián veličin $D_1, \ldots, D_n$ a testujme hypotézu $H : d_{0,5} = 0$.

Označíme-li Z počet kladných veličin z veličin $D_1, \ldots, D_n$, postupujeme při testování hypotézy $H : d_{0,5} = 0$ obdobně jako při testování hypotézy $H : x_{0,5} = x_0$.

12.3 Příklady.

12.3.1

Příklad 2.3 obsahuje údaje o počtu kmitů do lomu u $n = 8$ zkoušek únavy kovu. Testujme hypotézu $H : x_{0,5} = 8 \cdot 10^6$ proti alternativní hypotéze $A : x_{0,5} > 8 \cdot 10^6$ na hladině významnosti $\alpha = 0,05$.

Z tabulek pro $n = 8$ a $\alpha = 0,05$ nalezneme $c = 1$, takže $n - c = 7$. Protože $Z = 4 < 7$, hypotézu H nezamítáme.

12.3.2

Tabulka 8.1 udává hodnoty obsahu železa u $n = 12$ vzorků železné rudy; jsou to výsledky získané dvěma analytickými metodami. Označme D_i rozdíly výsledků u jednotlivých vzorků a testujme hypotézu $H : d_{0,5} = 0$ proti alternativní hypotéze $A : d_{0,5} < 0$ při $\alpha = 0,005$.

Z tabulek pro $n = 12$ a $\alpha = 0,005$ nalezneme $c = 1$. Protože $Z = 1 = c$, hypotézu H zamítáme, tj. učiníme závěr, že $P(X < Y) > P(X > Y)$.

12.4 Wilcoxonův test.

Mějme náhodný výběr $\mathbf{X} = (X_1, \ldots, X_n)'$ ze spojitého rozdělení symetrického podle mediánu $x_{0,5}$. Testujme opět hypotézu $H : x_{0,5} = x_0$ (x_0 dané číslo).

Uvažujme rozdíly (12.2.2). Jejich absolutní hodnoty

$$|Z_i| = |X_i - x_0|, \quad i = 1, \ldots, n, \tag{12.4.1}$$

seřaďme od nejmenší do největší a takto uspořádané posloupnosti přiřaďme pořadí $1,\ 2,\ \ldots,\ n$. Označme R_i^* pořadí $|Z_i|,\ i = 1, \ldots, n$. Položme

$$T = \sum_{i=1}^{n} a_i R_i^* \tag{12.4.2}$$

kde

$$\begin{aligned} a_i &= 1, \quad \text{jestliže} \ \ Z_i > 0, \\ &= 0, \quad \text{jestliže} \ \ Z_i < 0, \end{aligned} \tag{12.4.3}$$

Je tedy T součet pořadí pro kladná Z_i.

Za platnosti hypotézy H má statistika T symetrické rozdělení se střední hodnotou a rozptylem

$$E(T) = \frac{n(n+1)}{4}, \quad \operatorname{var}(T) = \frac{n(n+1)(2n+1)}{24} \tag{12.4.4}$$

V [23] jsou tabelovány hodnoty T_P splňující za platnosti hypotézy H vztahy

$$P(T \le T_P) \le P, \quad P(T \le T_P + 1) > P \tag{12.4.5}$$

pro $P \le 0,1$ a $n = 4\,(1)\,100$. Pro hodnoty T_{1-P} takové, že

$$P(T \ge T_{1-P} \le P, \quad P(T \ge T_{1-P} - 1) > P, \tag{12.4.6}$$

platí

$$T_{1-P} = \frac{n(n+1)}{2} - T_P. \tag{12.4.7}$$

Pro velká n má za platnosti hypotézy H statistika

$$U = \frac{T - n(n+1)/4}{[n(n+1)(2n+1)/24]^{\frac{1}{2}}} = \frac{4T - n(n+1)}{[n(n+1)(2n+1)]^{\frac{1}{2}}}\sqrt{\frac{3}{2}} \tag{12.4.8}$$

přibližně rozdělení N(0, 1), takže v případě velkých n můžeme stanovit přibližně hodnoty T_P pomocí kvantilů rozdělení N(0, 1).

Testujme hypotézu $H : x_{0,5} = x_0$. Uvažujme hladinu významnosti α. Pak

a) v případě alternativní hypotézy $A : x_{0,5} > x_0$ zamítáme H, jestliže $T \geq T_{1-\alpha}$;

b) v případě alternativní hypotézy $A : x_{0,5} < x_0$ zamítáme H, jestliže $T \leq T_\alpha$;

c) v případě oboustranné alternativní hypotézy $A : x_{0,5} \neq x_0$ zamítáme H, jestliže $T \leq T_{\alpha/2}$ nebo $T \geq T_{1-\alpha/2}$.

Obdobně jako v případě znaménkového testu můžeme Wilcoxonova testu užít i pro náhodný výběr z dvourozměrného rozdělení. Můžeme-li předpokládat, že rozdíly $D_i = X_i - Y_i, i = 1, \ldots, n$, jsou vzájemně nezávislé náhodné veličiny a všechny mají totéž spojité rozdělení symetrické podle mediánu $d_{0,5}$, pak lze při testování hypotézy $H : d_{0,5} = 0$ postupovat obdobně jako při testování hypotézy $H : x_{0,5} = x_0$.

Znaménkový i Wilcoxonův test v případě výběru z dvourozměrného rozdělení jsou neparametrické analogie parametrického testu uvažovaného v odst. 11.12.1 v případě výběru z dvourozměrného normálního rozdělení.

12.5 Příklady.

12.5.1

Aplikujme Wilcoxonův test na údaje uvažované v příkl. 12.3.1. Tabulka 12.1 uvádí hodnoty Z_i a pořadí R_i^*; z nich určíme hodnotu $T = 5+8+1+7 = 21$. Z tabulek nalezneme $T_{0,95} = 36 - 5 = 31$. Protože $21 < 31$, hypotézu $H : x_{0,5} = 8 \cdot 10^6$ nezamítáme.

12.5.2

Použijme Wilcoxonova testu pro údaje tab. 12.1. Testujme hypotézu $H : d_{0,5} = 0$ proti alternativní hypotéze $A : d_{0,5} < 0$ na hladině významnosti $\alpha = 0,005$.

Z tabulek pro $n = 12$ nalezneme hodnotu $T_{0,005} = 7$. Protože $T = 2 < 7$, hypotézu H zamítáme.

X_i	$Z_i = X_i - 800$	R_i^*	a_i
3322	-4678	2	0
14713	6713	5	1
763	-7237	6	0
46296	38296	8	1
2845	-5155	3	0
9411	1411	1	1
1532	-6468	4	0
24023	16023	7	1

Tab. 12.1: Pořadí R_i^*.

12.6 Mannův-Whitneyův test.

Mějme uspořádaný výběr $\mathbf{X}_1^* = (X_{1(1)}, \ldots, X_{1(n_1)})$ ze spojitého rozdělení s distribuční funkcí $F(x)$ a uspořádaný výběr $\mathbf{X}_2^* = (X_{2(1)}, \ldots, X_{2(n_2)})$ ze spojitého rozdělení s distribuční funkcí $G(x)$. Nechť výběr $\mathbf{X}_1^*$ a $\mathbf{X}_2^*$ jsou nezávislé.

Testujme hypotézu rovnosti těchto distribučních funkcí, tj. hypotézu H : $F(x) = G(x)$ pro všechna reálná x.

Mannova-Whitneyova statistika U je definována vztahem

$$U = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} h_{ij}, \tag{12.6.1}$$

kde

$$\begin{aligned} h_{ij} &= 1, \quad \text{jestliže } X_{1(i)} < X_{2(j)}, \\ &= 0, \quad \text{jestliže } X_{1(i)} > X_{2(j)}, \end{aligned} \tag{12.6.2}$$

tj. U je počet dvojic $(X_{1(i)}, X_{2(j)})$, pro které $X_{1(i)} < X_{2(j)}$, $i = 1, \ldots, n_1$, $j = 1, \ldots, n_2$,

Za platnosti hypotézy H má statistika (12.6.1) symetrické rozdělení se střední hodnotou a rozptylem

$$E(U) = \frac{1}{2} n_1 n_2, \quad \text{var}(U) = \frac{1}{12} n_1 n_2 (n_1 + n_2 + 1). \tag{12.6.3}$$

V [23] jsou tabelovány hodnoty U_P, pro něž za platnosti hypotézy H

$$P(U \leq U_P) \leq P, \quad P(U \leq U_P + 1) > P, \tag{12.6.4}$$

pro $P \leq 0,1$ a $1 \leq n_1 \leq n_2 \leq 40$.

Pro hodnoty U_{1-P} takové, že

$$P(U \geq U_{1-P} \leq P, \qquad P(U \geq U_{1-P} - 1) > P, \tag{12.6.5}$$

platí

$$U_{1-P} = n_1 n_2 - U_P. \tag{12.6.6}$$

Pro velká n_1 a n_2 lze za platnosti hypotézy H aproximovat rozdělení statistiky

$$\frac{U - n_1 n_2/2}{\big(n_1 n_2 (n_1 + n_2 + 1)/12\big)^{\frac{1}{2}}} = \frac{2U - n_1 n_2}{\big(n_1 n_2 (n_1 + n_2 + 1)\big)^{\frac{1}{2}}} \sqrt{3} \tag{12.6.7}$$

rozdělením N(0, 1).

Testujme hypotézu $H : F(x) = G(x)$ pro všechna reálná x. Uvažujme hladinu významnosti α. Pak

a) v případě alternativní hypotézy $A : F(x) \geq G(x)$, přičemž aspoň pro jedno x je $F(x) > G(x)$, zamítáme H, jestliže $U \geq U_{1-\alpha}$;

b) v případě alternativní hypotézy $A : F(x) \leq G(x)$, přičemž aspoň pro jedno x je $F(x) < G(x)$, zamítáme H, jestliže $U \leq U_{\alpha}$;

c) v případě, že alternativní hypotéza A je negací hypotézy H, zamítáme H, jestliže $U \leq U_{\alpha/2}$ nebo $U \geq U_{1-\alpha/2}$.

S Mannovým-Whitneyovým testem souvisí *Wilcoxonův dvouvýběrový test.* Spojme oba výběry, uspořádejme všech $n_1 + n_2$ výběrových hodnot podle velikosti a přiřaďme jim pořadí. Označme R_{1i} pořadí hodnoty X_{1i}, $i = 1, \ldots, n_1$. Pak Wilcoxonova dvouvýběrová statistika

$$S = \sum_{i=1}^{n_1} R_{1i}. \tag{12.6.8}$$

Protože

$$\sum_{j=1}^{n_2} h_{ij} = n_2 - (R_{1i} - i), \qquad i = 1, \ldots, n_1,$$

je

$$U = n_1 n_2 + \frac{1}{2} n_1 (n_1 + 1) - S. \tag{12.6.9}$$

Odtud a z (12.6.3) vyplývá, že

$$E(x) = \frac{1}{2}n_1(n_1 + n_2 + 1), \qquad \mathrm{var}(S) = \mathrm{var}(U). \tag{12.6.10}$$

Místo hodnot U_P daných výrazy (12.6.4) lze tabelovat hodnoty S_P splňující za platnosti hypotézy H vztahy

$$P(S \le S_P) \le P, \qquad P(S \le S_P + 1) > P. \tag{12.6.11}$$

Pro takováto S_P platí vzhledem k (12.6.9) a (12.6.6)

$$S_P = n_1 n_2 + \frac{1}{2}n_1(n_1 + 1) - U_{1-P} = \frac{1}{2}n_1)(n_1 + 1) + U_P. \tag{12.6.12}$$

Dále pro hodnoty S_{1-P} splňující vztahy

$$P(S \ge S_{1-P}) \le P, \qquad P(S \ge S_{1-P} - 1) > P \tag{12.6.13}$$

je

$$S_{1-P} = n_1 n_2 + \frac{1}{2}n_1(n_1 + 1) - U_P = n_1(n_1 + n_2 + 1) - S_P. \tag{12.6.14}$$

Pro velká n_1 a n_2 lze za platnosti hypotézy H aproximovat rozdělení statistiky

$$\frac{S - n_1(n_1 + n_2 + 1)/12}{\sqrt{n_1 n_2(n_1 + n_2 + 1)/12}} = \frac{2S - n_1(n_1 + n_2 + 1)}{\sqrt{n_1 n_2(n_1 + n_2 + 1)}}\sqrt{3} \tag{12.6.15}$$

rozdělením N(0, 1).

Pro testování hypotézy $H : F(x) = G(x)$ pro všechna reálná x odpovídají kritickým oborům $U \ge U_{1-\alpha}$, $U \le U_\alpha$ a $U \le U_{\alpha/2}$ nebo $U \ge U_{1-\alpha/2}$ pro statistiku U po řadě kritické obory $S \le S_\alpha$, $S \ge S_{1-\alpha}$ a $S \le S_{\alpha/2}$ nebo $S \ge S_{1-\alpha/2}$ pro statistiku S.

12.7 Příklad.

Při $n_1 = 5$ zkouškách pevnosti oceli A jsme dostali výsledky (uspořádané podle velikosti):

$$119,9;\ 124,6;\ 132,3;\ 135,6;\ 148,3.$$

Obdobně při $n_2 = 6$ zkouškách pevnosti oceli B jsme dostali výsledky:

$$109,9;\ 113,3;\ 113,9;\ 117,9;\ 124,0;\ 127,8.$$

Předpokládejme, že se jedná o dva nezávislé uspořádané výběry ze spojitých rozdělení, přičemž první výběr pochází z rozdělení s distribuční funkcí $F(x)$, druhý z rozdělení s distribuční funkcí $G(x)$.

Testujme hypotézu $H : F(x) = G(x)$ pro všechna reálná x, tj. hypotézu, že oba výběry pocházejí z téhož rozdělení, proti alternativní hypotéze $A : F(x) \leq G(x)$, přičemž aspoň pro jedno x je $F(x) < G(x)$. Uvažujme hladinu významnosti $\alpha = 0,05$.

Z tabulek pro $n_1 = 5$ a $n_2 = 6$ nalezneme $U_{0,05} = 5$. Protože $U = 2 + 1 + 0 + 0 + 0 = 3 < 5$, hypotézu H zamítáme.

Kdybychom použili Wilcoxonova dvouvýběrového testu, nalezli bychom podle (12.6.13) $S_{0,95} = 30 + 15 - 5 = 40$. Protože $S = 5 + 7 + 9 + 10 + 11 = 42 > 40$, hypotézu H zamítáme.

12.8 Spearmanův koeficient pořadové korelace.

Mějme náhodný výběr $(X_1, Y_1)', \ldots, (X_n, Y_n)'$ ze spojitého rozdělení dvourozměrné veličiny $(X, Y)'$. Testujme hypotézu H : Veličiny X a Y jsou nezávislé.

Uvažujme vektor $(X_1, \ldots, X_n)'$ a označme P_i pořadí X_i, $i = 1, \ldots, n$; obdobně uvažujme vektor $(Y_1, \ldots, Y_n)'$ a označme Q_i pořadí Y_i, $i = 1, \ldots, n$.

Dosazením P_i za X_i a Q_i za Y_i v (3.7.5) dostáváme *Spearmanův koeficient pořadové korelace* (pro $n \geq 2$)

$$r^{(S)} = \frac{n \sum_{i=1}^n P_i Q_i - \left(\sum_{i=1}^n P_i\right)\left(\sum_{i=1}^n Q_i\right)}{\sqrt{\left(n \sum_{i=1}^n P_i^2 - (\sum_{i=1}^n P_i)^2\right)\left(n \sum_{i=1}^n Q_i^2 - \left(\sum_{i=1}^n Q_i\right)^2\right)}} \qquad (12.8.1)$$

Protože

$$\sum_{i=1}^n P_i = \sum_{i=1}^n Q_i = \sum_{i=1}^n i = \frac{n(n+1)}{2},$$

$$\sum_{i=1}^n P_i^2 = \sum_{i=1}^n Q_i^2 = \sum_{i=1}^n i^2 = \frac{n(n+1)(2n+1)}{6}$$

a

$$2\sum_{i=1}^{n} P_i Q_i = -\sum_{i=1}^{n}(P_i - Q_i)^2 + \sum_{i=1}^{n} P_i^2 + \sum_{i=1}^{n} Q_i^2 = \\ = -\sum_{i=1}^{n}(P_i - Q_i)^2 + \frac{n(n+1)(2n+1)}{3},$$

dá se (12.8.1) upravit na tvar

$$r^{(S)} = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n}(P_i - Q_i)^2 = 1 - \frac{6}{n(n^2-1)}S. \tag{12.8.2}$$

V [23] jsou tabelovány hodnoty S_P splňující za platnosti hypotézy H vztahy

$$P(S \le S_P) \le P, \quad P(S \le S_P + 1) > P \tag{12.8.3}$$

pro $P \le 0,1$ a $n = 4\,(1)\,16$. Pro hodnoty S_{1-P} takové, že

$$P(S \ge S_{1-P}) \le P, \quad P(S \ge S_{1-P} - 1) > P, \tag{12.8.4}$$

platí

$$S_{1-P} = \frac{n(n^2-1)}{3} - S_P. \tag{12.8.5}$$

Označme

$$r_{1-P}^{(S)} = 1 - \frac{6}{n(n^2-1)}S_P, \tag{12.8.6}$$

kde pro S_P platí vztahy (12.8.3); ze vztahu (12.8.5) pak vyplývá, že

$$r_P^{(S)} = -r_{1-P}^{(S)}.$$

Pro větší n lze použít aproximace

$$r_{1-P}^{(S)} \doteq \frac{t_{1-P}}{\sqrt{t_{1-P}^2 + n - 2}}, \quad P < 0,5, \tag{12.8.7}$$

kde t_{1-P} je $100(1-P)\%$ kvantil rozdělení t o $n-2$ stupních volnosti.

x_i	y_i	P_i	Q_i	$(P_i - Q_i)^2$
21,2	18,3	7	1	36
18,8	19,2	2	4	4
22,5	20,9	10	6	16
21,6	19,8	8	5	9
23,2	22,7	11	11	0
19,5	21,3	4	7	9
18,2	18,9	1	3	4
19,7	18,7	5	2	9
22,3	21,4	9	8	1
20,9	22,5	6	10	16
19,2	21,6	3	9	36

Tab. 12.2: Pořadí P_i a Q_i pro Spearmanův koeficient korelace.

12.9 Příklad.

V náhodném výběru rozsahu $n = 11$ jsme dostali hodnoty uvedené v tab. 12.2. Testujme hypotézu H: Veličiny X a Y mající dvourozměrné rozdělení, z něhož byl vzat náš výběr, jsou nezávislé. Uvažujme kritický obor $r^{(S)} \geq r^{(S)}_{1-\alpha}$ (neboli $S \leq S_\alpha$) a hladinu významnosti $\alpha = 0,01$.

Tabulka 12.2 uvádí též pořadí P_i a Q_i a rozdíly $(P_i - Q_i)^2$. Z nich vypočteme $S = 140$ a $r^{(S)} = 1 - \frac{6 \cdot 140}{11 \cdot 120} \doteq 0,364$. V [23] nalezneme pro $n = 11$ hodnotu $S_{0,01} = 64$. Hypotézu H tedy nezamítáme.

12.10 Kendallův koeficient pořadové korelace.

Uvažujme stejné předpoklady i testovanou hypotézu H jako v odst. 12.8. Seřaďme opět pozorování $X_1, \ldots, X_n$ od nejmenšího do největšího a nahraďme je jejich pořadími. Totéž učiňme pro $Y_1, \ldots, Y_n$.

Uvažujme nyní posloupnost pořadí $1, 2, \ldots, n$ pro X a jim odpovídající pořadí $R_1,\ R_2,\ \ldots, R_n$ pro Y. Nyní pro každé i uvažujme pořadí $R_{i+1},\ \ldots,\ R_n$ a označme k_i počet z těchto pořadí, která jsou větší než $R_i,\ \ i = 1, \ldots, n-1$.

Označme $K = \sum_{i=1}^{n-1} k_i$. Pak *Kendallův koeficient pořadové korelace* je definován vztahem

$$r^{(K)} = \frac{4K}{n(n-1)} - 1. \tag{12.10.1}$$

V [23] jsou tabelovány hodnoty K_P takové, že za platnosti hypotézy H je

$$P(K \leq K_P) \leq P, \quad P(K \leq K_P + 1) > P, \tag{12.10.2}$$

pro $P \leq 0,1$ a $n = 4\,(1)\,100$. Hodnoty K_{1-P} takové, že

$$P(K \geq K_{1-P}) \leq P, \quad P(K \geq K_{1-P} - 1) > P, \tag{12.10.3}$$

splňují vztahy

$$K_{1-P} = \frac{n(n-1)}{2} - K_P. \tag{12.10.4}$$

Pro velká n lze za platnosti hypotézy H aproximovat rozdělení statistiky

$$\frac{K - n(n-1)/4}{\sqrt{n(n-1)(2n+5)/72}} = \frac{4K - n(n-1)}{\sqrt{n(n-1)(2n+5)}}\,\frac{3\sqrt{2}}{2} \tag{12.10.5}$$

rozdělením N(0, 1).

12.11 Příklad.

vypočteme hodnoty statistik K a $r^{(K)}$ pro údaje tab. 12.2. V tabulce 12.3 jsou seřazena pořadí pro X vzestupně od 1 do 11 a jim přiřazena odpovídající pořadí pro Y. Odtud vypočteme

$$K = 8 + 7 + 2 + 3 + 5 + 1 + 4 + 3 + 1 + 1 = 35$$

	1	2	3	4	5	6	7	8	9	10	11
R_i	3	4	9	7	2	10	1	5	8	6	11

Tab. 12.3: Pořadí R_i pro Kendallův koeficient korelace.

a $r^{(K)} = (4 \cdot 35/110 - 1 \doteq 0,273$. V [23] nalezneme pro $n = 11$ hodnotu $K_{0,01} = 12$, takže $K_{0,99} = 55 - 12 = 43$. Protože $K = 35 < 43$, hypotézu H, stejně jako v příkl. 12.9, nezamítáme.

Část IV

Ověřování shody empirických rozdělení s modelem

Kapitola 13

Grafické metody rozboru empirického rozdělení

13.1 Úvodní poznámka.

Všechny metody vyložené v předcházejících článcích (až na čl. 12) vycházely z předpokladu, že se pozorují náhodné veličiny s určitým známým rozdělením pravděpodobnosti a experiment slouží jen k poznání některých parametrů příslušného rozdělení. Jen zřídkakdy je tvar tohoto rozdělení přesně odvozen z fyzikálních či chemických nebo jiných zákonů a z podstaty experimentu. Zpravidla je tvar rozdělení pozorované veličiny určen na základě rozboru rozsáhlých nahromaděných dat, tj. na základě rozboru náhodného výběru $\mathbf{X} = (X_1, \ldots, X_n)'$ velkého rozsahu n z příslušného rozdělení. Při malém počtu pozorování n je naděje na přijatelně spolehlivé určení tvaru rozdělení, z kterého výběr pochází, nepatrná. v tomto článku se budeme zabývat jednoduchými grafickými metodami k orientačnímu posouzení tvaru rozdělení.

13.2 Skupinové (třídní) rozdělení; histogram.

Jak jsme uvedli v odst. 13.1, máme-li mít rozumnou naději na zjištění tvaru rozdělení náhodné veličiny X, potřebujeme zpravidla značný počet pozorování, tj. výběr značného rozsahu n. Záznam jednotlivých výsledků v pořadí, v jakém byly naměřeny, je nepřehledný a neumožňuje představu o tom, jaké asi může být rozdělení pozorované veličiny. Proto se pozorování ve výběru velkého rozsahu n zpravidla roztřídí do tzv. skupinového (třídního) rozdělení

četností.

Předpokládejme, že pozorovaná náhodná veličina X má rozdělení spojitého typu s distribuční funkcí $F(x)$ a že z tohoto rozdělení byl proveden výběr velkého rozsahu n. Rozdělme interval možných hodnot náhodné veličiny X na intervaly $(t_0, t_1\rangle, (t_1, t_2\rangle, \ldots, (t_{r-1}, t_r)$. Tyto intervaly nazýváme *třídními intervaly* nebo krátce *třídami.* Zjistíme počty pozorování s hodnotami z jednotlivých tříd a označíme n_i počet pozorování X_j splňujících nerovnost $t_{i-1} < X_j \leq t_i$.

Čísla $n_1, \ldots, n_r$ jsou tzv. *(absolutní) třídní četnosti;* přirozeně musí být $\sum_{i=1}^{r} n_i = n$.

Čísla

$$p_i = \frac{n_i}{n}, \qquad i = 1, \ldots, r, \tag{13.2.1}$$

jsou tzv. *relativní třídní četnosti;* splňují podmínku $\sum_{i=1}^{r} p_i = 1$. Relativní třídní četnosti $p_i, i = 1, \ldots, r$, jsou nestrannými odhady pravděpodobností

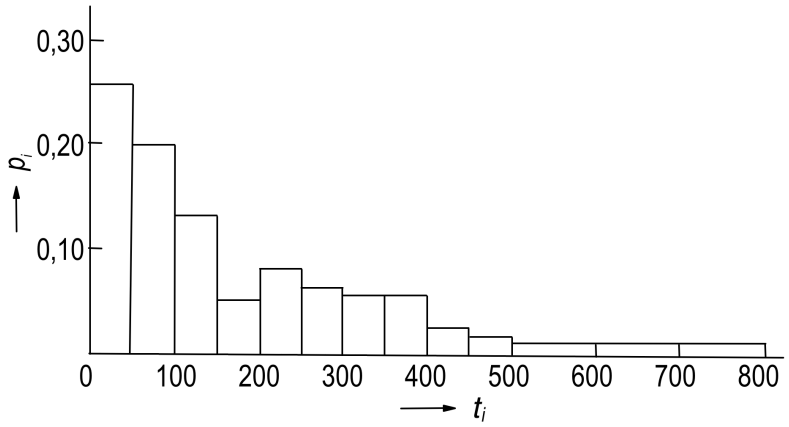
$$\pi_i = P(t_{i-1} < X \leq t_i) = F(t_i) - F(t_{i-1}) = \int_{t_{i-1}}^{t_i} f(x)\,\mathrm{d}x \doteq (t_i - t_{i-1}) f(\bar{t}_i), \tag{13.2.2}$$

kde $\bar{t}_i = (t_{i-1} + t_i)/2$ a $f(x)$ je hustota pravděpodobnosti příslušná k distribuční funkci $F(x)$.

Přibližnou představu o průběhu hustoty pravděpodobnosti lze získat z tzv. histogramu. *Histogram* je graf, ve kterém jsou na ose úseček vyznačeny třídní intervaly a nad každým intervalem sestrojen obdélník, jehož plocha je úměrná relativní četnosti $p_i,\ i = 1, \ldots, r$. Histogram je tedy po částech konstantní funkce nabývající pro x z intervalu $t_{i-1} < x \leq t_i$ hodnoty $h(x) = p_i/(t_i - t_{i-1})$. Srovnáním $h(x)$ s (13.2.2) zjistíme, že $h(\bar{t}_i)$ je odhadem hustoty $f(\bar{t}_i)$.

13.3 Příklad.

Tabulka 1.2 v příkl. 1.2.2 je příkladem třídního rozdělení četností. Na obr. 13.3 je znázorněn odpovídající histogram. Je vidět, že histogram má průběh dobře aproximovatelný funkcí tvaru $c\,\mathrm{e}^{-\frac{x}{\delta}}$, což naznačuje, že doba bezporuchového provozu daného typu zařízení bude mít exponenciální rozdělení nebo aspoň rozdělení blízké exponenciálnímu.



Obr. 13.1: Histogram rozdělení dob bezporuchového chodu navigačních přístrojů z tab. 1.2

13.4 Empirická distribuční funkce.

Podobně jako histogram poskytuje představu o možném průběhu hustoty pravděpodobnosti, empirická distribuční funkce je odhadem distribuční funkce pozorované veličiny. Distribuční funkce $F(x)$ náhodné veličiny X udává ke každému x pravděpodobnost $P(X \leq x)$.

Empirickou distribuční funkci (příslušnou k náhodnému výběru $\mathbf{X} = (X_1, \ldots, X_n)'$ na rozsahu n) definujeme jako funkci $F_n(x)$ přiřazující každému x relativní četnost pozorování menších nebo rovných x. Značí-li tedy N_x počet pozorování X_i splňujících nerovnost $X_i \leq x$, je

$$F_n(x) = \frac{N_x}{n}. \tag{13.4.1}$$

Označme – jako v čl. 4 – $X_{(i)}$ i-té pozorování ve výběru seřazeném podle velikosti. Empirická distribuční funkce $F_n(x)$ je schodovitá funkce rovná 0 pro všechna $x < X_{(1)}$ (protože ve výběru není žádné pozorování menší než $X_{(1)}$), rovná $1/n$ pro všechna x z intervalu $X_{(1)} \leq x < X_{(2)}$ atd., tj. konstantní v intervalech $\langle X_{(i)}, X_{(i+1)})$ se skokem velikosti $1/n$ v každém bodě $X_{(i)}$. To znamená, že lze psát

$$\begin{aligned} F_n(x) &= 0, \quad x < X_{(1)}, \\ &= \frac{i}{n}, \quad X_{(i)} \leq x < X_{(i+1)}, \quad i = 1, \ldots, n-1, \\ &= 1, \quad x \geq X_{(n)}. \end{aligned} \tag{13.4.2}$$

Při grafickém znázornění empirické distribuční funkce $F_n(x)$ se někdy místo celého průběhu funkce $F_n(x)$ zakreslují do grafu jen body se souřadnicemi

$$\left[X_{(i)};\ \frac{i-0,5}{n}\right], \quad i=1,\ldots,n, \tag{13.4.3}$$

tj. v každém bodě $X_{(i)}$, v němž $F_n(x)$ má diskontinuitu (mění skokem svou hodnotu), se zakresluje jako pořadnice průměr hodnot $F(X_{(i-1)}$ a $F(X_{(i)}$.

Při velkém počtu pozorování n se často do grafu zakreslují jen hodnoty empirické distribuční funkce ve vybraných bodech $t_1,\ldots,t_r$, např. ekvidistantních (ale ekvidistantnost není podmínkou). To znamená, že se vyjde z třídního rozdělení četností podle (13.2.1), vypočtou se tzv. *kumulativní četnosti* $N_i=\sum_{j=1}^{n} n_j, i=1,\ldots,r$, a položí se

$$F_n(t_i)=\frac{N_i}{n}, \quad i=1,\ldots,r; \tag{13.4.4}$$

do grafu se zakreslují jen body $[t_i;\ F_n(t_i)]$.

Všechny tři postupy jsou ilustrovány v následujících dvou příkladech.

13.5 Příklady.

13.5.1

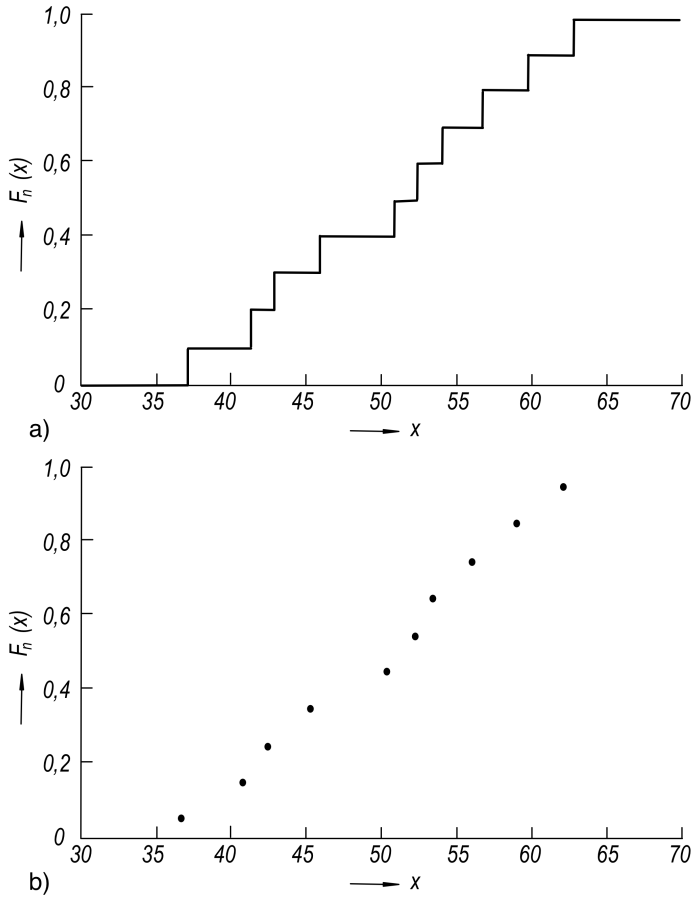
Tabulka 13.1 obsahuje podle velikosti seřazené hodnoty v náhodném výběru rozsahu $n=10$. Na obrázku 13.5.2 je zakreslena empirická distribuční funkce tohoto výběru podle (13.4.2) a podle (13.4.3).

37,3	41,4	42,8	45,9	51,0	52,6	54,0	56,7	59,8	62,8

Tab. 13.1: Hodnoty v náhodném výběru rozsahu $n=10$ seřazené podle velikosti.

13.5.2

Empirická distribuční funkce pro data z příkl. 13.3 je tabelována v tab. 13.2 a graficky znázorněna na obr. 13.6.



Obr. 13.2: Empirická distribuční funkce pro data z příkladu 13.5.1
a) podle (13.4.2), b) podle (13.4.3)

13.6 Transformovaná empirická distribuční funkce; pravděpodobnostní papír.

Empirická distribuční funkce, která by naznačovala tvar skutečné distribuční funkce pozorované náhodné veličiny X tak výrazně jako v příkl. 13.3, je úkazem spíše mimořádným než pravidelným. Většinou sledují body $[t_i;\ F_n(t_i)]$ vzestupnou esovitě prohnutou křivku a je obtížné rozlišit, zda tato křivka je blízká distribuční funkci rozdělení normálního nebo distribuční funkci rozdělení logaritmicko-normálního apod. Proto se často místo grafu empi-

rické distribuční funkce sestrojuje graf transformované empirické distribuční funkce, přičemž transformace je volena tak, aby distribuční funkce určitého tvaru měly po této transformaci lineární průběh. Jestliže analyzovaná data jsou náhodným výběrem z rozdělení daného typu, pak graf transformované distribuční funkce sleduje - až na náhodné odchylky - také přímku. Lineární průběh grafu je snadno rozeznatelný, a to je hlavní důvod k užití podobných transformovaných grafů.

i	Horní hranice třídy t_i	Třídní četnost n_i	Kumulovaná četnost N_i	Empirická distribuční funkce $F_n(t_i)$
1	50	32	32	0,256
2	100	25	57	0,456
3	150	16	73	0,584
4	200	6	79	0,632
5	250	10	89	0,712
6	300	8	97	0,776
7	350	7	104	0,832
8	400	7	111	0,888
9	450	3	114	0,912
10	500	2	116	0,928
11	550	5	121	0,968
12	600	3	124	0,992
13	650	0	124	0,992
14	700	1	125	1,000

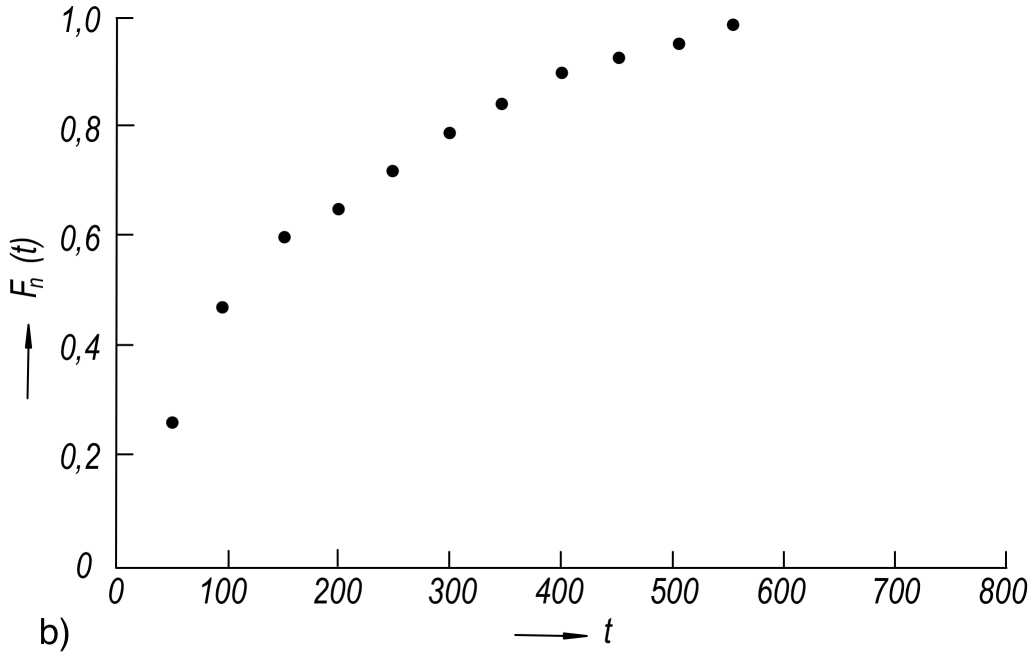
Tab. 13.2: Empirická distribuční funkce doby do poruch.

V následujících odstavcích uvedeme transformace linearizující průběh distribučních funkcí rozdělení, se kterými se nejčastěji v praxi setkáváme: normálního, logaritmicko-normálního, exponenciálního a Weibullova.

13.7 Normální rozdělení.

Jestliže náhodná veličina X má rozdělení $\mathrm{N}(\mu, \sigma^2)$, pak její distribuční funkce

$$F(x) = P(X \leq x) = \Phi\Big(\frac{x-\mu}{\sigma}\Big), \quad -\infty < x < \infty, \tag{13.7.1}$$



Obr. 13.3: Empirická distribuční funkce rozdělení dob bezporuchového provozu z tab. 13.2

kde

$$\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} \mathrm{e}^{-\frac{t^2}{2}}\, \mathrm{d}t. \tag{13.7.2}$$

Aplikujeme-li na obě strany rovnosti (13.7.1) funkci Φ^{-1} (tj. inverzní funkci k Φ), dostaneme

$$\Phi^{-1}\big(F(x)\big) = \frac{x}{\sigma} - \frac{\mu}{\sigma}; \tag{13.7.3}$$

v jiném značení

$$u_{F(x)} = \frac{x}{\sigma} - \frac{\mu}{\sigma}. \tag{13.7.4}$$

To znamená, že kvantil $u_{F(x)}$ je lineární funkcí x.

Skutečné hodnoty distribuční funkce $F(x)$ nejsou známy, ale jsou k dispozici jejich odhady, totiž $F_n(x)$, empirická distribuční funkce. Jestliže tedy pozorovaná náhodná veličina X má rozdělení $\mathrm{N}(\mu, \sigma^2)$, pak graf $[x; u_{F_n(x)}]$ má – až na náhodné odchylky – lineární průběh. Hodnoty $u_{F_{n(x)}}$ jsou seskupeny kolem přímky se směrnicí $1/\sigma$. Pro $x = \mu$ je $u_{F(x)} = 0$ (jak je zřejmé

z (13.7.4)); tudíž bod x_0, ve kterém přímka proložená „od oka“ grafem $u_{F_{n(x)}}$ protíná osu úseček, je hrubým odhadem parametru μ. Dále pro $x_1 = \mu + \sigma$ je $u_{F(x_1)} = 1$ a pro $x_{-1} = \mu - \sigma$ je $u_{F(x_{-1})} = -1$.

Odtud plyne: Je-li $\hat{x}$ (resp. $\hat{x}_{-1}$) bod, ve kterém přímka proložená „od oka“ body $[x; u_{F_{n(x)}}]$ protíná přímku $u = 1$ (resp. $u = -1$), pak $\hat{\sigma} = (\hat{x}_1 - \hat{x}_{-1})/2$ je hrubým odhadem směrodatné odchylky σ.

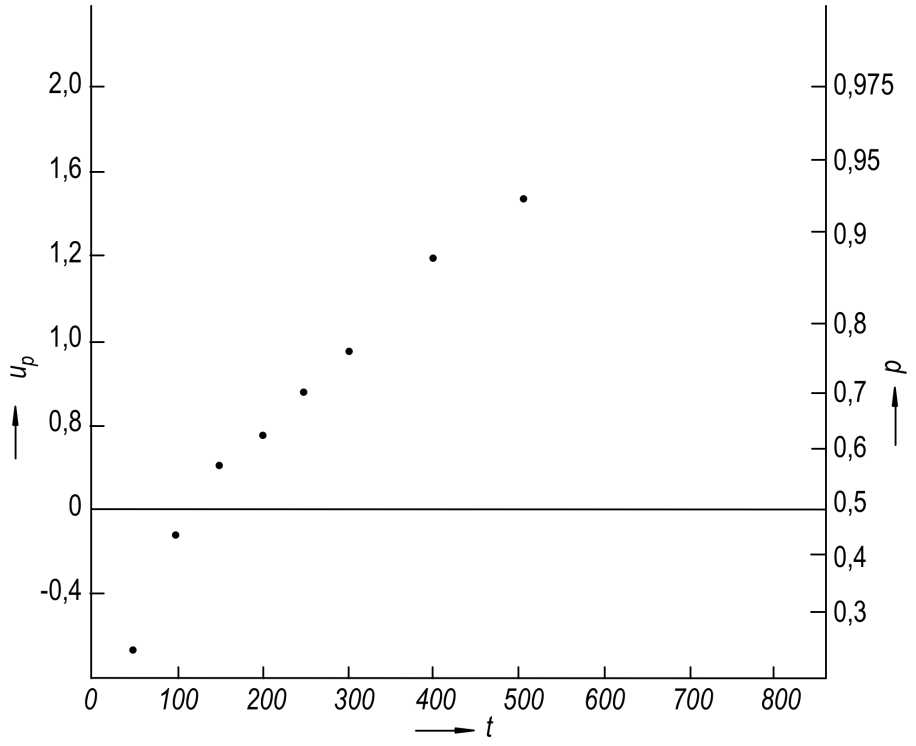
Pro ilustraci znázorněme graficky tímto způsobem empirickou distribuční funkci náhodného výběru z příkl. 13.5.2. S užitím tabulek [23] dostaneme hodnoty $t_i\, u_{F_{n(ti)}}$ uvedené v tab. 13.3.

i	t_i	$F_n(t_i)$	$u_{F_n}(t_i)$
1	50	0,256	-0,656
2	100	0,456	-0,111
3	150	0,584	0,212
4	200	0,632	0,337
5	250	0,712	0,559
6	300	0,776	0,759
7	400	0,888	1,216
8	500	0,928	1,461
9	600	0,992	2,409
10	700	1,000	∞

Tab. 13.3: Hodnoty $u_{F_n}(t_i)$.

Graf $[t_i; u_{F_{n(t_i)}}]$ je zakreslen na obr. 13.7. Seskupení bodů $[t_i; F_n(t_i)]$ ukazuje, že lepšího vyrovnání by se dosáhlo spíše nějakou konkávní funkcí než přímkou; normální rozdělení tedy asi nebude vhodným modelem pro dobu života (dobu do poruchy) u přístroje daného typu.

Pro časté používání popsaného způsobu analýzy je výhodné opatřit si speciální grafický papír (grafickou síť), který má vodorovnou osu dělenou lineárně jako obyčejný milimetrový papír a na svislé ose ve vzdálenosti u_p od počátku kótu P. Počátku tedy odpovídá kóta $0,5$ nebo 50%. Do této sítě se zakreslují přímo body $[X_{(i)}; i/n]$ nebo body $[t_i; F_n(t_i)]$. Popsaná síť se jmenuje *normální pravděpodobnostní papír.*



Obr. 13.4: Graf empirické distribuční funkce dat z příkl. 1.2.2 v normální pravděpodobnostní síti. Na levé straně lineární stupnice, na pravé straně pravděpodobnostní stupnice

13.8 Exponenciální rozdělení.

Distribuční funkce exponenciálního rozdělení je

$$F(x) = 1 - \mathrm{e}^{-\frac{x}{\delta}}, \qquad x > 0;$$

přirozený logaritmus funkce $\big(1 - F(x)\big)^{-1}$ zřejmě je

$$\ln \frac{1}{1 - F(x)} = \frac{x}{\delta} \tag{13.8.1}$$

čili lineární funkce proměnné x. Při použití obyčejného (dekadického) logaritmu

$$\log_{10} \frac{1}{1 - F(x)} = \frac{x}{2,302\,6\delta} \tag{13.8.2}$$

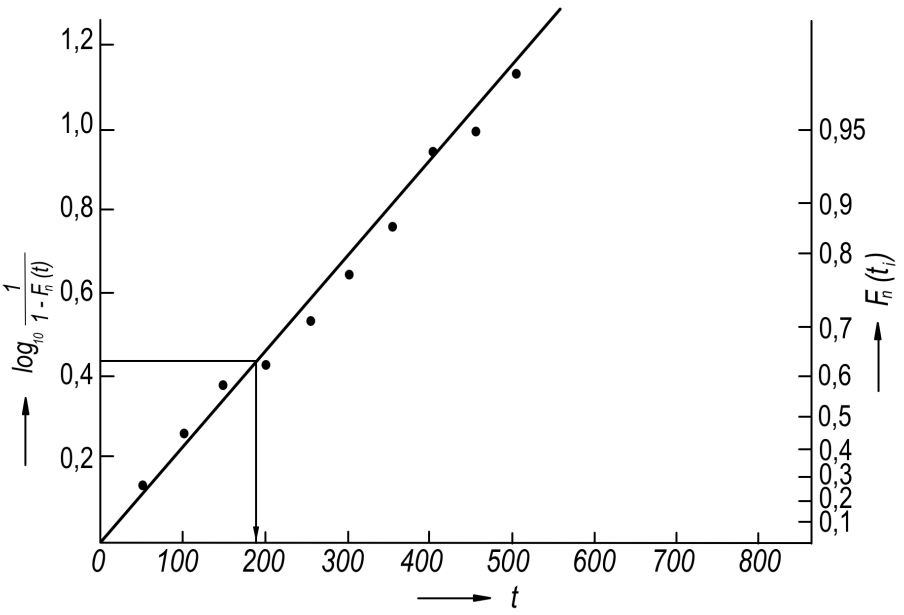
tj. opět lineární funkce procházející počátkem. To znamená, že transformovaný graf empirické distribuční funkce (označení log je zde i dále použito pro oba případy ln i $\log_{10}$)

$$\Big[X_{(i)};\ \log\frac{n}{n-i}\Big], \tag{13.8.3}$$

příp.

$$\Big[t_i;\ \log\frac{n}{n-N_i}\Big], \tag{13.8.4}$$

má – při X s exponenciálním rozdělením – lineární průběh až na náhodné odchylky; přirozeně lze s výhodou použít semilogaritmického papíru. Semilogaritmického papíru zpravidla použijeme, když nebude k dispozici kalkulátor a pak budeme chtít vůbec redukovat množství početních operací. Proto budeme místo $\log_{10}\big(1-F_n(x)\big)^{-1}$ vynášet graf funkce $\log_{10}\big(1-F_n(x)\big)$, který by měl – až na náhodné odchylky – sledovat přímku se směrnicí $-2,302\,6\,\delta$.



Obr. 13.5: Empirická distribuční funkce dat z tab. 1.2 transformovaná podle (13.8.4). Vlevo lineární stupnice pro přímé použití (13.8.4), vpravo stupnice $\log_{10}\big(1/(1-p)\big)$ pro zakreslení funkce $\big[t_i, F_n(t_i)\big]$. Grafickou metodou získaný odhad parametru δ je 190, v příkl. 14.3.1 vypočten numericky přesný odhad $186,82$

i	t_i	$n - N_i$	$1 - F_n(t_i)$	$\log_{10}\big(1 - F_n(t_i)\big)^{-1}$
0	0	125	1,000	0,0000
1	50	93	0,744	0,1284
2	100	68	0,544	0,2644
3	150	52	0,416	0,3809
4	200	46	0,368	0,4342
5	250	36	0,288	0,5406
6	300	28	0,224	0,6498
7	350	21	0,168	0,7747
8	400	14	0,112	0,9508
9	450	11	0,088	1,0555
10	500	9	0,072	1,1427

Tab. 13.4: Hodnoty $\log_{10}\big(1 - F_n(t_i)\big)^{-1}$.

Grafu lze použít i k hrubému odhadu parametru δ : v bodě $x = \delta$ je $\ln\big(1-F(x)\big)^{-1} = 1$, $\log_{10}\big(1-F(x)\big)^{-1} = 1/2,302\,6 \doteq 0,434\,3$; tzn. že bod x_0, ve kterém přímka proložená od oka grafem funkce $\log_{10}\big(1-F(x)\big)^{-1}$ protíná $0,434\,3$ (viz obr. 13.8), je odhadem parametru δ. Při použití semilogaritmické sítě a funkce $1-F_n(x)$ je odhadem parametru δ bod x_0, ve kterém graf funkce $1 - F_n(x)$ protíná antilog $(-0,434\,3) \doteq 0,367\,9$.

Popsaný postup ilustrujme opět na datech příkl. 13.5.2. Tabulka 13.4 obsahuje hodnoty funkce $1 - F_n(x)$ pro $x = t_1 = 50, x = t_2 = 100$ atd. Na obr. 13.8 je znázorněna funkce $1 - F_n(x)$ v semilogaritmické síti. Body $[t_i, F_n(t_i)]$ se dobře přimykají k přímce procházející bodem $[0, 1]$, což ukazuje na přijatelnost předpokladu, že X má exponenciální rozdělení. Na obrázku je tato přímka (proložená „od oka“) také vyznačena. Odhad parametru δ nalezený jako úsečka průsečíku této přímky s $0,367\,9$ je přibližně 190.

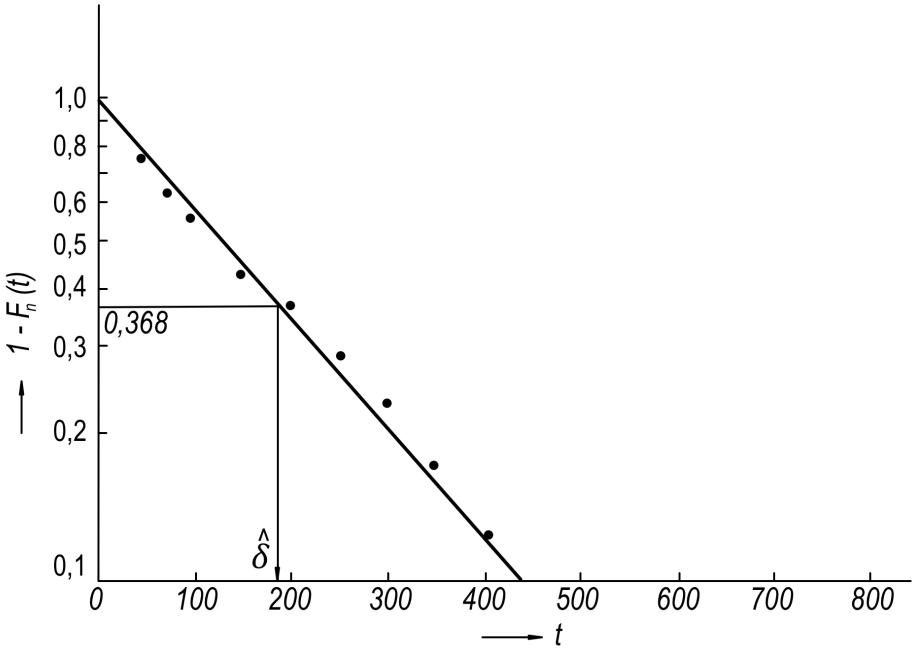
13.9 Weibullovo rozdělení.

Distribuční funkce Weibullova rozdělení je

$$F(x) = 1 - \mathrm{e}^{-(x/\delta)^c}, \qquad x > 0.$$

Odtud

$$\ln\Big(- \ln\big(1 - F(x)\big)\Big) = c \ln x - c \ln \delta, \tag{13.9.1}$$



Obr. 13.6: Graf 1 - $F_n(t_i)$ pro data příkl. 1.2.2 v semilogaritmické síti. V grafu je vyznačen odhad $\hat{\delta}$ rovný přibližně 190

což znamená, že funkce

$$\ln\ln\frac{1}{1-F(x)} \tag{13.9.2}$$

je lineární funkcí přirozeného logaritmu x. Má-li tedy X Weibullovo rozdělení, pak graf empirické distribuční funkce $F_n(x)$ transformované vztahem

$$\ln\ln\frac{1}{1-F_n(x)} = \ln\ln\frac{n}{n-N_x} \tag{13.9.3}$$

by měl sledovat - až na náhodné odchylky - přímku s rovnicí $y = c\ln x - c\ln\delta$.

Pro hrubé ověření, zda pozorovaná náhodná veličina X může mít Weibullovo rozdělení, tedy zakreslíme do grafu body

$$\left[\ln X_{(i)};\quad \ln\ln\frac{n}{n-i+0,5}\right] \tag{13.9.4}$$

při netříděných datech, příp.

$$\left[\ln t_i;\quad \ln\ln\frac{n}{n-N_i}\right] \tag{13.9.5}$$

při datech uspořádaných do skupinového rozdělení četností. Jsou-li body seskupeny přibližně kolem přímky, je přijatelný předpoklad, že X má Weibullovo rozdělení. Směrnice přímky proložené grafem „od oka“ je odhadem parametru c, úsek na svislé ose je odhadem $c \ln \delta$ a z přímky lze přímo vyčíst odhady kvantilů rozdělení. Pro rutinní použití podobných analýz je možno připravit si předem grafický papír s logaritmickou stupnicí na ose úseček a se stupnicí $\ln \ln(1/P)$ na ose pořadnic. To je tzv. *Weibullův pravděpodobnostní papír*. Lze si samozřejmě vypomoci použitím logaritmického papíru, do kterého se zakreslí body $\left[X_{(i)};\, log(1 - F_n(X_{(i)}))^{-1}\right]$; při odhadu parametrů z grafu je pak třeba dát pozor na vliv modulu pro převod přirozených logaritmů na dekadické.

Postup ilustrujme opět na datech z příkl. 13.5.2. Tabulka 13.5 obsahuje hodnoty $\log_{12} t_i$ a $\log_{10} \log_{10} \left(1 - F_n(t_i)\right)^{-1}$, na obr. 13.9 je znázorně graf transformované empirické distribuční funkce. Skutečnost, že body se dobře přimykají přímce, od které nevykazují systematické odchylky, ukazuje na to, že Weibullovo rozdělení je přijatelným modelem pro rozdělení X.

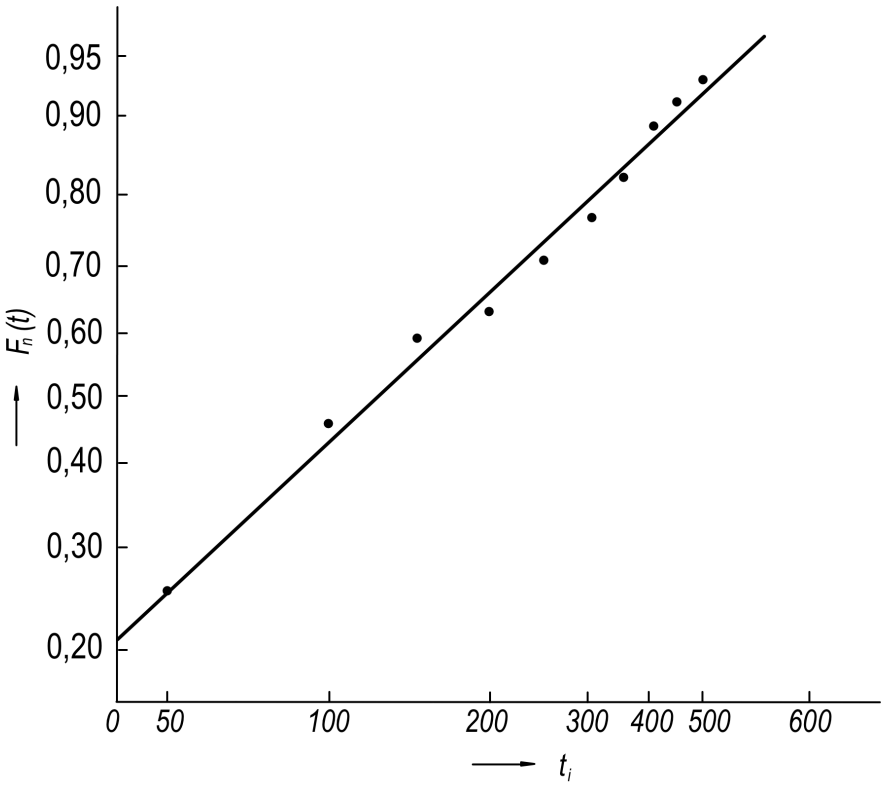
i	t_i	$\log_{10} t_i$	$\log_{10} \log_{10} \left(1 - F_n(t_i)\right)^{-1}$
1	50	1,699	-0,8913
2	100	2,000	-0,5777
3	150	2,176	-0,4192
4	200	2,301	-0,3624
5	250	2,398	-0,2671
6	300	2,477	-0,1873
7	350	2,544	-0,1109
8	400	2,602	-0,0219
9	450	2,653	0,0235
10	500	2,699	0,0579
11	800	2,903	∞

Tab. 13.5: Transformace dat z tabulky 13.2.

13.10 Logaritmicko-normální rozdělení.

Distribuční funkce logaritmicko-normálního rozdělení je rovna

$$F(x) = \Phi\left(\frac{\ln x - \mu}{\sigma}\right), \qquad x > 0,$$



Obr. 13.7: Empirická distribuční funkce rozdělení dob bezporuchového provozu navigačních přístrojů z tab. 1.2 ve Weibullově pravděpodobnostní síti

kde $\Phi(u)$ je distribuční funkce rozdělení $N(0,1)$. Aplikací inverzní funkce k Φ dostaneme

$$\Phi^{-1}\big(F(x)\big) = \frac{\ln x}{\sigma} - \frac{\mu}{\sigma} = \frac{2,3026\ \log_{10}\ x}{\sigma} - \frac{\mu}{\sigma}; \tag{13.10.1}$$

tzn. že $u_{F(x)}$ (kde u_P je $100\%P$ kvantil rozdělení N(0, 1)) je lineární funkcí logaritmu x. Má-li tedy X logaritmicko-normální rozdělení, pak obrazy bodů $[lnX_{(i)};\ u_{(i-0,5),n}]$ (při netříděných datech), resp. $[\ln t_i;\ u_{F_n(t_i)}]$ budou seskupeny – až na náhodné odchylky – kolem přímky.

S grafem se pracuje podobně jako s grafem z odst. 13.7. Síť s logaritmickou stupnicí na vodorovné ose a normální pravděpodobnostní stupnici na svislé ose tvoří tzv. *logaritmicko-normální pravděpodobnostní papír.*

13.11 Závěrečné poznámky k grafické analýze dat.

Metody grafické analýzy vyložené v odst. 13.6 až 13.10 neslouží vždy jen k posouzení tvaru rozdělení pozorované veličiny. Používá se jich někdy i v případech, kdy tvar rozdělení považujeme za známý a potřebujeme jen rychlý a jednoduchý, třeba hrubý, odhad parametrů. Tak např. lze použít grafické metody z odst. 13.9 k získání první aproximace pro maximálně věrohodné odhady parametrů Weibullova rozdělení (viz odst. 7.4); s odhady získanými z grafu podle odst. 13.9 se zahájí iterační postup řešení rovnice věrohodnosti (7.4.6).

Grafické metody odhadu postupy z předchozích odstavců jsou zvláště užitečné při odhadech z tzv. *cenzorovaných výběrů*, tj. z pozorování uspořádaných tak, že je známa jen část výsledků a o zbytku se ví jen, že pozorování nabyla hodnot větších než největší zaznamenaná hodnota. K takové situaci dochází zvláště často při zkouškách životnosti; do zkušebního provozu se dá n výrobků a odhady parametrů příslušného rozdělení se provádějí po určité době, řekněme t_0, ve které ještě některé výrobky neměly poruchu. To znamená, že je změřeno jen $m < n$ nejkratších dob života ve výběru $X_{(1)}, \ldots, X_{(m)}$, o zbývajících se ví jen tolik, že jsou větší než t_0. Lze sestrojit část grafu empirické distribuční funkce. Je-li znám tvar rozdělení doby života X, lze z grafu transformované empirické distribuční funkce odhadnout přibližné hodnoty parametrů. I když se použije jemnějších metod, např. metody maximální věrohodnosti rovnice v popsané situaci zpravidla vyžadují numerické řešení.

Grafy transformovaných distribučních funkcí lze upravit různými způsoby. Někteří statistici např. při netříděných datech raději zakreslují na osu úseček hodnoty $1/n, 2/n, \ldots$ a na osu pořadnic $X_{(1)}, X_{(2)}, \ldots$, jiní nahrazují jmenovatele n v empirické distribuční funkci jmenovatelem $n + 1$ atd. Výklad důvodů pro takové úpravy a diskuse jejich užitečnosti by vedla příliš daleko a není zde na ni místo.

13.12 Úloha.

Vyjádřete souřadnice pravděpodobnostního papíru pro Rayleighovo rozdělení (viz [24], odst. 22.6).

$$\Big[\,\big[x;\,\sqrt{-2\ln(1-F(x))}\big]\ \text{ nebo }\ \big[x^2;\,-2\ln(1-F(x))\big].\Big]$$

Kapitola 14

Testy dobré shody

14.1 Úloha testování dobré shody.

V odstavcích 13.6 až 13.10 se několikrát vyskytla fráze „jestliže X má distribuční funkci určitého tvaru, graf transformované distribuční funkce má – až na náhodné odchylky – lineární průběh“. Je přirozené položit si otázku, co jsou „náhodné odchylky“; kdy jsou odchylky empirické distribuční funkce tak malé, že je možno vysvětlit jejich přítomnost tím, že empirická distribuční funkce je sestrojena z náhodného výběru, a kdy jsou tak velké, že budí oprávněné pochybnosti o správnosti předpokladu „X má distribuční funkci tvaru $F(x)$“. Odchylky, které jeden experimentátor bude považovat za nepodstatné a přirozené, může jiný pokládat za výrazné. Proto je užitečné mít k dispozici nějakou míru statistické významnosti takových odchylek umožňující objektivnější posouzení souhlasu či nesouladu náhodného výběru $\mathbf{X} = (X_1, \ldots, X_n)'$ s předpokladem typu „X má rozdělení s distribuční funkcí tvaru $F(x)$“.

Tvrzení jako „X má normální rozdělení“ nebo „X má exponenciální rozdělení“ atd. jsou statistické hypotézy. Pravidlo, podle kterého se na základě opakovaných pozorování náhodné veličiny X (tj. na základě náhodného výběru z $F(x)$) rozhodně, zda daná hypotéza má být přijata či zamítnuta, je tzv. „test dobré shody.“

Obecný postup při testu dobré shody je tento: Z výběru $\mathbf{X} = (X_1, \ldots, X_n)'$ se vypočte vhodná míra nesouhlasu $\mathbf{X}$ s předpokladem „X má rozdělení s distribuční funkcí tvaru $F(x)$“ čili vhodná míra odchylky či vzdálenosti empirické distribuční funkce od tzv. „teoretické “ či „hypotetické distribuční

funkce“ (tj. distribuční funkce vypočítané za předpokladu, že má tvar daný hypotézou). Označme, na chvíli, tuto míru vzdálenosti empirické distribuční funkce od hypotetické jako D. Veličina D je funkcí náhodného výběru, je tedy sama náhodnou veličinou, která má svoje rozdělení pravděpodobnosti závislé na skutečné distribuční funkci náhodné veličiny X Označme $D_{1-\alpha}$ tu hodnotu, kterou D překročí s pravděpodobností α, když testovaná hypotéza je správná, tj. když X má skutečné rozdělení dané hypotézou. Zvolí se malé číslo α (tzv. hladina významnosti), např. $\alpha = 0,05$. Jestliže statistika D vypočtená z výběru nabude hodnoty větší než nebo rovné $D_{1-\alpha}$, pak nastal jev, který má za platnosti příslušné hypotézy jen malou pravděpodobnost, totiž α, a je tedy důvod k podezření, že hypotéza není správná, tj. že pozorovaná veličina X má jiné rozdělení než předpokládané.

V následujících odstavcích uvedeme některé nejběžnější typy statistiky D, vyjadřující souhlas či nesouhlas výsledků pozorování s hypotézou o tvaru rozdělení.

14.2 Test chí-kvadrát.

Při tzv. *testu* χ^2 dobré shody se vychází z třídního rozdělení náhodného výběru (viz odst. 13.2). Jako míry nesouhlasu mezi výsledky pozorování a hypotézou o tvaru distribuční funkce pozorovaného veličiny se používá statistiky

$$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - n\pi_i)^2}{n\pi_i} = \sum_{i=1}^{r} \frac{n_i^2}{n\pi_i} - n, \tag{14.2.1}$$

kde r je počet tříd, na které byl rozdělen interval možných hodnot pozorované veličiny, n_i absolutní třídní četnost i-tého třídního intervalu, tj. počet pozorování splňujících nerovnost $t_{i-1} < X_j \leq t_i$ a π_i pravděpodobnost, že pozorovaná náhodná veličina X nabude hodnoty z intervalu $(t_{i-1}, t_i\rangle$ počítaná za předpokladu, že X má dané rozdělení. Značí-li $F(x)$ distribuční funkci náhodné veličiny X, jak ji specifikuje ověřovaná hypotéza, a $F_n(x)$ empirickou distribuční funkci výběru, je tedy

$$n_i = n\big(F_n(t_i) - F_n(t_{i-1})\big), \tag{14.2.2}$$

$$\pi_i = F(t_i) - F(t_{i-1}); \tag{14.2.3}$$

$t_0, t_1, \ldots, t_r$ jsou hranice tříd, $t_0 = \inf\{t \,|\, F(t) > 0\}$, $t_r = \sup\{t \,|\, F(t) < 1\}$, tj. dolní a horní hranice intervalu možných hodnot X.

Vektor $(n_1, \ldots, n_r)'$ třídních četností je realizací r-rozměrné náhodné veličiny s multinomickým rozdělením (viz [24], odst. 17.1) s parametry n, $\pi_1, \ldots, \pi_r$; náhodný výběr rozsahu n z rozdělení $F(x)$ představuje totiž n vzájemně nezávislých náhodných pokusů, ve kterých jev $A_i = \{t_{i-1} < X \leq t_i\}$ má pravděpodobnost π_i danou (14.2.3) a n_i je počet výskytů jevu A_i v těchto n nezávislých pokusech. Podle [24], odst. 17.3, je střední hodnota četnosti jevu A_i rovna $n\pi_i$. Tuto střední hodnotu nazýváme (v souvislosti s testem dobré shody) *teoretickou* (nebo *hypotetickou) četností i-té třídy.*

Hypotéza o tvaru distribuční funkce $F(x)$ zpravidla neurčuje hypotetické četnosti $n\pi_i$ jednoznačně, nýbrž jen jako funkce jednoho či několika parametrů. Zní-li hypotéza např. „X má normální rozdělení“, pak

$$n\pi_i = n\Big(\Phi\Big(\frac{t_i - \mu}{\sigma}\Big) - \Phi\Big(\frac{t_{i-1} - \mu}{\sigma}\Big)\Big) = n\pi_i(\mu, \sigma^2)$$

Pro výpočet hypotetických četností ve statistice (14.2.1) je pak třeba potřebné parametry odhadnout také z pozorovaných četností n_i. Nejčastěji se k tomu používá metody maximální věrohodnosti (viz odst. 6.7). Pravděpodobnost, že při daných třídách $(t_{i-1}, t_i\rangle$, $i = 1, \ldots, r$, v náhodném výběru rozsahu n budou třídní četnosti rovny $n_1, \ldots, n_r$, je

$$P(n_1, \ldots, n_r) = \frac{n!}{n_1! \ldots n_r!} \pi_1^{n_1} \ldots \pi_r^{n_r}. \tag{14.2.4}$$

Jsou-li pravděpodobnosti tříd $\pi_1, \ldots, \pi_r$ funkcemi k parametrů, kde $k < r - 1$, $\pi_i = \pi_i(\Theta_1, \ldots, \Theta_k)$, pak funkce (14.2.4) považována při daných hodnotách pozorovaných četností $n_1, \ldots, n_r$ za funkci neznámých parametrů, je funkcí věrohodnosti pro odhad vektoru parametrů $(\Theta_1, \ldots, \Theta_k)'$.

Maximálně věrohodné odhady $\hat{\Theta}_1, \ldots, \hat{\Theta}_k$ parametrů $\Theta_1, \ldots, \Theta_k$ tedy získáme maximalizací logaritmu funkce věrohodnosti

$$\sum_{i=1}^{r} n_i \ln \pi_i(\Theta_1, \ldots, \Theta_k). \tag{14.2.5}$$

Jestliže pravděpodobnosti tříd $\pi_i(\Theta_1, \ldots, \Theta_k)$ splňují podmínku

$$\sum_{i=1}^{r} \pi_i(\Theta_1, \ldots, \Theta_k) = 1 \tag{14.2.6}$$

pro všechny možné hodnoty parametrů $\Theta_1, \ldots, \Theta_k$, mají spojité parciální derivace podle všech Θ a matice

$$\mathbf{D} = \left(\frac{\partial \pi_i}{\partial \Theta_j} \right), \quad i = 1, \ldots, r, \quad j = 1, \ldots, k, \tag{14.2.7}$$

má hodnost k, pak lze odhady $\hat{\Theta}_1, \ldots, \hat{\Theta}_k$ nalézt řešením soustavy rovnic

$$\sum_{i=1}^{r} \frac{n_i}{\pi_i(\hat{\Theta}_1, \ldots, \hat{\Theta}_k)} \frac{\partial \pi_i(\hat{\Theta}_1, \ldots, \hat{\Theta}_k}{\partial \hat{\Theta}_j} = 0, \quad j = 1, \ldots, k. \tag{14.2.8}$$

Statistika χ^2 dána vztahem (14.2.1) s $\pi_i = \hat{\pi}_i(\hat{\Theta}_1, \ldots, \hat{\Theta}_k)$, kde $\hat{\Theta}_1, \ldots, \hat{\Theta}_k$ jsou kořeny soustavy (14.2.8), má za platnosti ověřované hypotézy při $n \to \infty$ (tak, aby $n\hat{\pi}_i > 5$ pro všechna $i = 1, \ldots, r$) asymptoticky rozdělení χ^2 s počtem stupňů volnosti $v = r - 1 - k$. Důkaz tohoto výsledku a rozbor vlastností odhadů podle (14.2.8) lze nalézt v pracích [6, 31].

Test dobré shody založený na statistice (14.2.1) tedy zahrnuje tyto kroky:

1. Výpočet odhadů $\hat{\Theta}_1, \ldots, \hat{\Theta}_k$ neznámých parametrů rozdělení řešením soustavy rovnic (14.2.8).
2. Výpočet pravděpodobností $\pi_i(\hat{\Theta}_1, \ldots, \hat{\Theta}_k)$ dosazením $\hat{\Theta}_1, \ldots, \hat{\Theta}_k$ do (14.2.3).
3. Výpočet statistiky χ^2 podle (14.2.1).
4. Porovnání vypočtené hodnoty statistiky χ^2 s tabelovaným kvantilem $\chi^2_{1-\alpha}(r - 1 - k)$.

Soustava rovnic (14.2.8) má jen zřídka explicitní řešení, zpravidla je k přesnému řešení třeba numerických metod. V praxi se často spokojujeme jen s přibližným řešením. Užití popsaného testu je osvětleno v následujícím odstavci na příkladech.

14.3 Příklady.

14.3.1 Ověřování shody s exponenciálním rozdělením.

Předpokládejme, že hypotéza zní: X má rozdělení s distribuční funkcí

$$F(x) = 1 - e^{-x/\delta}, \quad x > 0, \tag{14.3.1}$$

kde δ je neznámý parametr. Byl proveden výběr dosti velkého rozsahu n a utvořeno třídní rozdělení četností. Pravděpodobnosti $\pi_i(\delta)$ jevů $\{t_{i-1} < X \leq t_i\}$ jsou

$$\pi_i = 1 - \mathrm{e}^{-t_i/\delta} - \left(1 - e^{-t_{i-1}/\delta}\right) = e^{-t_{i-1}/\delta} - \mathrm{e}^{-t_i/\delta}. \tag{14.3.2}$$

Předpokládejme, že intervaly $(t_{i-1}, t_i\rangle$ jsou ekvidistantní, $t_i = ih, \quad h > 0, \ i = 0, 1, \ldots, r-1, t_r = \infty$. Pak

$$\begin{aligned} \pi_i &= \mathrm{e}^{-(i-1)h/\delta} - \mathrm{e}^{-ih/\delta} = \mathrm{e}^{-ih/\delta}\big(\mathrm{e}^{h/\delta} - 1\big), \qquad i = 1, \ldots, r-1, \\ \pi_r &= \mathrm{e}^{-(r-1)h/\delta}. \end{aligned} \tag{14.3.3}$$

Soustava (14.2.8) obsahuje jedinou rovnici

$$\frac{n_1}{\pi_1}\frac{\mathrm{d}\pi_1}{\mathrm{d}\delta} + \sum_{i=2}^{r-1}\frac{n_i}{\pi_i}\frac{\mathrm{d}\pi_i}{\mathrm{d}\delta} + \frac{n_r}{\pi_r}\frac{\mathrm{d}\pi_r}{\mathrm{d}\delta} = 0, \tag{14.3.4}$$

která po úpravě s užitím (14.3.3) přejde na rovnici

$$e^{h/\delta}\Big(\sum_{i=1}^{r} i\, n_i - n\Big) = \sum_{i=1}^{r} i\, n_i - n_r, \tag{14.3.5}$$

odkud

$$\hat{\delta} = \{\frac{1}{h}\ln\frac{\sum_{i=1}^{r} i n_i - n_r}{\sum_{i=1}^{r} i n_i - n}\}^{-1}. \tag{14.3.6}$$

To je jeden z mála případů, kdy (14.2.8) má explicitní řešení. Pro ilustraci převedeme provedení testu na datech z příkl. 13.5.2 při použití $r = 11$ tříd s horními hranicemi $t_i = 50i$ pro $i = 1, \ldots, 10$, jedenáctá třída je $(500, \infty)$. Výpočet odhadu $\hat{\delta}$, pravděpodobností π_i a testové statistiky κ^2 je shrnut v tab. 14.

Odhad parametru δ je

$$\hat{\delta} = 50[ln(503-9) - ln(503-125)]^{-1} = \frac{50}{ln494 - ln378} \doteq 186,82.$$

Další výpočet je zřejmý z tabulky. Statistika κ^2 nabývá hodnoty $6,325$, počet stupňů volnosti je $\nu = 11-1-1 = 9$; statistika tedy zdaleka nedosahuje hodnoty $\kappa^2_{0,95}(9) = 16,919$, je dokonce menší než $\kappa^2_{0,5}(9) = 8,343$. Není tedy žádný důvod k podezření, že rozdělení se liší od exponenciálního.

i	t_i	n_i	$i\,n_i$	π_i	$n\,\pi_i$	$\frac{(n_i-n\pi_i)^2}{n\pi_i}$
1	50	32	32	0,2348	29,35	0,2393
2	100	25	50	0,1797	22,46	0,2873
3	150	16	48	0,1375	17,19	0,0824
4	200	6	24	0,1052	13,15	3,8876
5	250	10	50	0,0805	10,06	0,0004
6	300	8	48	0,0616	7,70	0,0117
7	350	7	49	0,0471	5,89	0,2092
8	400	7	56	0,0361	4,51	1,3747
9	450	3	27	0,0276	3,45	0,0587
10	500	2	20	0,0211	2,64	0,1552
11	∞	9	99	0,0688	8,60	0,0186
	–	125	503	1,0000	125,00	6,3250

Tab. 14.1: Test exponenciálního rozdělení z příkladu 13.3.

14.3.2 Test normality.

Nechť hypotéza zní: X má normální rozdělení. Předpokládejme, že náhodný výběr rozsahu n byl roztříděn do r tříd $(t_{i-1}, t_i\rangle$, kde $t_0 = -\infty, t_1$ je dané číslo, $t_i = t_1 + (i-1)h$ pro $i = 2, \ldots, r-1, t_r = \infty$. Pravděpodobnosti jednotlivých tříd jsou

$$\begin{aligned} \pi_i &= P(t_{i-1} < X \le t_i) = \Phi\left(\frac{t_i-\mu}{\sigma}\right) - \Phi\left(\frac{t_{i-1}-\mu}{\sigma}\right), \quad i = 1, \ldots, r-1, \\ \pi_r &= P(X > t_{r-1}) = 1 - \Phi\left(\frac{t_{r-1}-\mu}{\sigma}\right), \end{aligned} \tag{14.3.7}$$

kde $\Phi(u)$ je distribuční funkce rozdělení N(0, 1). Pravděpodobnosti π_i jsou závislé na dvou neznámých parametrech μ a σ^2. K získání aproximací odhadů pro parametry μ a σ nahraďme dolní hranici první třídy číslem $t_0 = t_1 - h$, horní hranici r-té třídy číslem, $t_r = t_{r-1} + h$ a použijme dalších aproximací

$$\pi_i = \frac{1}{\sqrt{2\pi}} \int_{\frac{t_{i-1}-\mu}{\sigma}}^{\frac{t_i-\mu}{\sigma}} \mathrm{e}^{-\frac{u^2}{2}}\,\mathrm{d}u \doteq \frac{h}{\sigma}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{(\overline{t_i}-\mu)^2}{2\sigma^2}}, \tag{14.3.8}$$

$$\frac{\pi_i}{\partial\mu} = -\frac{1}{\sigma\sqrt{2\pi}}\{\mathrm{e}^{-\frac{(t_i-\mu)^2}{2\sigma^2}} - \mathrm{e}^{-\frac{(t_{i-1}-\mu)^2}{2\sigma^2}}\} \doteq \frac{1}{\sigma\sqrt{2\pi}}\int_{\frac{t_{i-1}-\mu}{\sigma}}^{\frac{t_i-\mu}{\sigma}} ue^{-\frac{u^2}{2}}\,\mathrm{d}u \doteq$$

$$\doteq \frac{h}{\sigma\sqrt{2\pi}}\frac{\overline{t}_i-\mu}{\sigma}e^{-\frac{(\overline{t}_i-\mu)^2}{2\sigma^2}} = \frac{h}{\sigma}\frac{\overline{t}_i-\mu}{\sigma}\pi_i, \quad (14.3.9)$$

$$\begin{aligned}\frac{\partial\pi_i}{\partial\sigma} &= -\frac{t_i-\mu}{\sigma\sqrt{2\pi}}e^{-\frac{(t_i-\mu)^2}{2\sigma^2}} + \frac{t_{i-1}-\mu}{\sigma\sqrt{2\pi}}e^{-\frac{(t_{i-1}-\mu)^2}{2\sigma^2}} \doteq \quad (14.3.10)\\ &\doteq \frac{1}{\sqrt{2\pi}}\int_{\frac{t_{i-1}-\mu}{\sigma}}^{\frac{t_i-\mu}{\sigma}}(u^2-1)e^{-\frac{u^2}{2}}\,\mathrm{d}u \doteq\\ &\doteq \frac{h}{\sigma\sqrt{2\pi}}\left[\left(\frac{\overline{t}_i-\mu}{\sigma}\right)^2 e^{-\frac{(\overline{t}_i\mu)^2}{2\sigma^2}}\} - e^{-\frac{(\overline{t}_i-\mu)^2}{2\sigma^2}}\right] = \frac{h}{\sigma}\pi_i\left(\frac{\overline{t}_i-\mu}{\sigma}\right)^2 - \frac{h}{\sigma}\pi_i.\end{aligned}$$

Dosazením aproximací do (14.2.8) dostaneme pro odhady parametrů μ a σ^2 rovnice

$$\sum_{i=1}^{n} n_i(\overline{t}_i-\hat{\mu}) = 0, \frac{1}{\hat{\sigma}^2}\sum_{i=1}^{r} n_i(\overline{t}_i-\hat{\mu})^2 - \sum_{i=1}^{r} n_i = 0, \quad (14.3.11)$$

odkud

$$\begin{aligned}\hat{\mu} &= \frac{1}{n}\sum_{i=1}^{r} n_i\overline{t}_i, \quad (14.3.12)\\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{r} n_i(\overline{t}_i-\hat{\mu})^2 = \frac{1}{n}\Big(\sum_{i=1}^{r} n_i\overline{t}_i^2 - \frac{1}{n}(\sum_{i=1}^{r} n_i\overline{t}_i)^2\Big),\end{aligned}$$

kde $\overline{t}_i$ je střed i-té třídy.

Často se zůstává při této hrubé první aproximaci a nehledají se odhady splňující přesně soustavu (14.2.8); k přesnému řešení by byl nutný iterační postup, při kterém bychom už přešli ke skutečným hranicím $t_0 = -\infty, t_r = \infty$ a používali přesných výrazů pro derivace π_i (viz [1]).

V odhadech (14.3.12) poznáváme obvyklý průměr a druhý centrální moment náhodného výběru počítaný z třídního rozdělení.

Pro ilustraci postupu při testu normality pomocí statistiky χ^2 uveďme tento příklad: Údaje tab. 14.2 obsahují výsledky $n = 200$ zkoušek pevnosti v tahu lněné příze. Nejsou výsledky uvedené v této tabulce v příkrém rozporu s předpokladem, že pevnost tohoto druhu lněné příze má normální rozdělení?

Za předpokladu, že pevnost daného druhu příze má normální rozdělení, odhady parametrů tohoto rozdělení podle (14.3.12) jsou

$$\hat{\mu} = \frac{454,99}{200} = 2,275; \qquad \hat{\sigma}^2 = \frac{1\,079,11 - 1\,035,08}{200} = 0,22; \hat{\sigma} = 0,469.$$

Předposlední sloupec tab. 14.2 obsahuje normované hodnoty horních hranic intervalů u_i a poslední sloupec odhady hodnot distribuční funkce

$$\hat{F}(t_i) = \Phi\left(\frac{t_i - \hat{\mu}}{\hat{\sigma}}\right).$$

i	t_i	n_i	$\bar{t}_i$	$n_i\bar{t}_i$	$n_i\bar{t}_i^2$	$u_i = \frac{t_i-\hat{\mu}}{\hat{\sigma}}$	$\hat{F}(t_i)$
1	1,25	2	1,125	2,25	2,53	-2,1855	0,01444
2	1,50	6	1,375	8,25	11,34	-1,6525	0,04922
3	1,75	20	1,625	32,50	52,81	-1,1194	0,13148
4	2,00	24	1,875	45,00	84,38	-0,5864	0,27880
5	2,25	40	2,125	85,00	180,62	-0,0533	0,47875
6	2,50	52	2,375	123,50	293,31	0,4797	0,68428
7	2,75	32	2,625	84,00	220,50	1,0128	0,84442
8	3,00	13	2,875	37,38	107,45	1,5458	0,93892
9	3,25	5	3,125	15,62	48,83	2,0790	0,98119
10	3,50	3	3,375	10,12	34,17	2,6119	0,99550
11	3,75	1	3,625	3,62	13,14	3,1450	0,99917
12	4,00	2	3,875	7,75	30,03	3,6780	0,99988
		200		454,99	1079,11		

Tab. 14.2: Výsledky $n = 200$ zkoušek pevnosti lněné příze; test normality rozdělení.

Tabulka 14.3 uvádí hodnoty potřebné pro výpočet statistiky χ^2. Aby byl splněn požadavek, že hypotetické četnosti nemají být příliš malé, doporučuje se spojení tříd s malými hypotetickými četnostmi v jedinou třídu (zde třídy 10, 11 a 12). Při zcela přesném postupu by se po takové úpravě měly korigovat odhady parametrů tak, aby byly řešením „přesných“ rovnic věrohodnosti (14.2.8). většinou se od toho upouští a toleruje se skutečnost, že nebude naprosto přesně dodržena hladina významnosti.

i	$\hat{\pi}_i$	$n\hat{\pi}_i$		$\frac{(n_i - n\hat{\pi}_i)^2}{n\hat{\pi}_i}$
1	0,01444	2,888		0,2730
2	0,03478	6,956		0,1314
3	0,08226	16,452		0,7652
4	0,14732	29,464		1,0133
5	0,19995	39,990		0,0000
6	0,20553	41,106		2,8872
7	0,16014	32,028		0,0000
8	0,09450	18,900		1,8418
9	0,04227	8,454		1,4112
10	0,01431	2,862		
11	0,00367	0,734	}3,738	1,3688
12	0,00071	0,142		
	0,99988	199,976		9,6919

Tab. 14.3: Hodnoty potřebné pro výpočet statistiky χ^2.

Statistika χ^2 nabývá hodnoty $9,691\,9$. Má $\nu = 10 - 1 - 2 = 7$ stupňů volnosti. Na hladině významnosti $\alpha = 0,05$ by se hypotéza o normalitě zamítala, kdyby bylo $\chi^2_{0,095} > 14,067 = \chi^2_{0,95}(7)$. Pozorovaná hodnota je menší (leží mezi $\chi^2_{0,75}(7) = 9,037$ a $\chi^2_{0,8}(7) = 9,803$). Data nedávají důvod k zamítnutí hypotézy, že X má normální rozdělení.

14.4 Kolmogorovův test.

Při *Kolmogorovově testu dobré shody* se k ověření hypotézy: X má rozdělení s distribuční funkcí $F(x)$, kde $F(x)$ je dána distribuční funkce spojitého typu, používá statistiky

$$D = sup_x|F_n(x) - F(x)|. \tag{14.4.1}$$

Protože $F(x)$ jakožto distribuční funkce je neklesající a $F_n(x)$ je funkce po částech konstantní a mění svou hodnotu skokem v bodech $X_{(1)}, \ldots, X_{(n)}$, lze statistiku D zapsat také jako

$$D = \max_{1\le i\le n}\left\{\max\left\{|F(x_{(i)} - \frac{i}{n}|, |F(X_{(i)} - \frac{i-1}{n}|\right\}\right\} \tag{14.4.2}$$

kde $F(x_{(i)})$ je hodnota hypotetické distribuční funkce F v bodě $X_{(i)}$. Odtud plynou dvě důležité zásady pro korektní použití Kolmogorova testu:

1. Testu se má používat jen při udaném celém průběhu empirické distribuční funkce $F_n(x)$, tj. při udaných všech hodnotách $X_{(i)}$, nikoliv při pozorováních seskupených do třídního rozdělení četností. Při třídním rozdělení četností známe totiž jen hodnoty hypotetické distribuční funkce v r daných bodech $t_1, \ldots, t_r$ a hodnoty empirické distribuční funkce v týchž bodech, tj. $F(t_i)$ a $F_n(t_i), i = 1, \ldots, r$, a je jasné, že

$$\max_{1\le i\le r} \big\{|F_n(t_i) - F(t_i)|\big\} \le sup_x|F_n(x) - F(x)|. \tag{14.4.3}$$

   To znamená, že při použití Kolmogorovova testu na třídní rozdělení se snižuje účinnosti (síla) testu, neboť se nezkoumá celý průběh empirické distribuční funkce, nýbrž jen hodnoty ve vybraných bodech.

2. Přísně vztato, má se Kolmogorovova testu používat jen k ověřování hypotéz, které stanoví distribuční funkci $F(x)$ jednoznačně, tj. nejen co do tvaru, nýbrž i co do hodnot parametrů, tak, aby bylo možno vypočítat hodnoty $F(X_{(i)})$ za platnosti hypotézy; v definici statistiky D se nemluví o „odhadu $F(x)$ za platnosti hypotézy“, nýbrž o „hodnotě distribuční funkce $F(x)$“. Jakmile by hypotéza určovala jen tvar distribuční funkce a na místě $F(X_{(i)})$ v (14.4.1) nebo (14.4.2) by se užívalo $F(X_{(i)}; \hat{\Theta})$, kde $\hat{\Theta}$ by byl odhad počítaný z výběru, změnilo by se rozdělení statistiky D, a v důsledku toho i rizika chyb.

Rozdělení statistiky D je rozsáhle tabelováno. A. N. Kolmogorov odvodil asymptotické rozdělení statistiky $\sqrt{n}\,D$. Pro menší rozsahy výběrů, které nedovolují užít asymptotické teorie, jsou tabelovány hodnoty $D_{n,1-\alpha}$ s vlastností

$$P\big(D > D_{n,1-\alpha} \,|\, X \text{ má rozdělení } F(x)\big) = \alpha,$$

viz [18, 23].

Pro ověření hypotézy: X má rozdělení $\mathrm{N}(\mu, \sigma^2)$ bez udání parametrů (μ, σ^2) a hypotézy: X má rozdělení $\mathrm{E}(0, \delta)$ bez udání hodnoty parametru δ, jsou tabelovány (viz [25, 26]) hodnoty $D^*_{n,1-\alpha}$ takové, že

$$P\left(\sup_x \left|\Phi\Big(\frac{x-\overline{x}}{s}\Big) - F_n(x)\right| > D^*_{n,1-\alpha} \,\Big|\, X \text{ má } \mathrm{N}(\mu, \sigma^2)\right) = \alpha,$$

resp. hodnoty $D^{**}_{n,1-\alpha}$ takové, že

$$P\left(\sup_x \left|F_n(x) - (1 - e^{-\frac{x}{\bar{x}}})\right| > D^{**}_{n,1-\alpha} \,\middle|\, X \text{ má } \mathrm{E}(0,\, \delta)\right) = \alpha.$$

Kapitola 15

Testování nezávislosti kvalitativních znaků

15.1 Experimenty s kvalitativní odpovědí; kontingenční tabulka.

Jestliže výsledek experimentu není vyjádřen reálným číslem (nebo vektorem reálných čísel), nýbrž jen slovně popsán či zařazen do jedné z k vzájemně se vylučujících kategorií, říkáme, že jde o *experiment s kvalitativní odpovědí.* Tak např. spočívá-li experiment ve zkoušce funkční způsobilosti výrobku, může výsledek pozorování být vyjádřen zařazením do některé z kategorií „zkouška úspěšná“, „selhání v důsledku poruchy typu A_1“, „selhání v důsledku poruchy typu A_2“, ..., „selhání v důsledku poruchy typu A_{k-1}“.

V jednom experimentu můžeme současně sledovat dvě nebo i více kvalitativních odpovědí, podobně jako lze v jednom experimentu měřit či pozorovat dvě nebo více náhodných veličin. Tak např. při kontrole jakosti výrobků může být jednou odpovědí (se dvěma kategoriemi) přítomnost či nepřítomnost vady druhu A a druhou odpovědí přítomnost či nepřítomnost vady druhu B. Při psychologické zkoušce konané za účelem zjištění způsobilosti osoby k výkonu určité činnosti může zkoušená osoba dostat p různých úkolů a výsledek plnění jednotlivých úkolů může být hodnocen jako „vynikající“, „průměrný“ a „podprůměrný“ - pak by šlo o experiment s p kvalitativními odpověďmi, každá se třemi kategoriemi. Jestliže se u výrobku zjišťuje nikoliv přesným měřením, nýbrž jen pomocí měrek, zda splňuje či nesplňuje tolerance pro dva různé rozměry, máme co dělat s experimentem se dvěma

kvalitativními odpověďmi, jejichž kategorie jsou „rozměr A v tolerancích“, „rozměr A pod dolní tolerancí“, „rozměr A nad horní tolerancí“, a podobně pro rozměr B.

Představme si nyní n nezávislých opakování experimentu se dvěma kvalitativními odpověďmi (nebo se dvěma kvalitativními znaky); označme $A_1, \ldots, A_r$ možné kategorie znaku (odpovědi) A a $B_1, \ldots, B_S$ možné kategorie znaku (odpovědi) B. Výsledek celého složeného experimentu, tj. všech opakování, lze shrnout do tabulky, viz tab. 17.

Kategorie znaku A	Kategorie znaku B B_1	B_2	$\ldots$	B_s	Celkem
A_1	n_{11}	n_{12}	$\ldots$	n_{1s}	$n_{1\,.}$
A_1	n_{21}	n_{22}	$\ldots$	n_{2s}	$n_{2\,.}$
A_1	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_r	n_{r1}	n_{r2}	$\ldots$	n_{rs}	$n_{r\,.}$
Celkem	$n_{.\,1}$	$n_{.\,1}$	$\ldots$	$n_{.\,s}$	n

Tab. 15.1: Kontingenční tabulka.

V tabulce 15.1 značí n_{ij} počet experimentů, při kterých odpověď A byla z kategorie A_i a odpověď B z kategorie B_j, $n_{i\,.} = \sum_{j=1}^{s} n_{ij}$ je celkový počet opakování, při kterých se vyskytla i-tá kategorie znaku A, $n_{.\,j} = \sum_{i=1}^{r} n_{ij}$ je celkový počet opakování, při kterých se vyskytla j-tá kategorie znaku B.

Tabulka uvedeného typu se nazývá *kontingenční tabulka.* Cílem statistického rozboru takové kontingenční tabulky je zjištění, zda příslušné dvě odpovědi (či dva znaky) mezi sebou nějak souvisí či ne. Kontingenční tabulka může mít i jiný obsah, a pak přicházejí v úvahu i jiné hypotézy. Zde se však zabýváme jen hypotézou nezávislosti dvou znaků.

15.2 Hypotéza nezávislosti kvalitativních znaků.

Výskyt i-té kategorie znaku A a j-té kategorie znaku B při jedné realizaci příslušného experimentu je náhodný jev, který je průnikem dvou jevů, A_i a B_j, kde A_i značí jev „znak A patří do kategorie A_i“ a B_j jev „znak B patří do kategorie B_j“. Označme pravděpodobnost tohoto průniku π_{ij}. Statickým modelem pro složený experiment spočívající v n nezávislých opakováních

pak je multinomické rozdělení s parametry $n, \pi_{11}, \pi_{12}, \ldots, \pi_{rs}$ (viz [24], odst. 17.1). Každá z četností n_{ij} v tab. 15.1 je vlastně realizací náhodné veličiny s rozdělením Bi(n, π_{ij}).

Četnosti n_i výskytu jevů A_i a četnosti $n_{.j}$ výskytu jevů B_j jsou realizace náhodných veličin s rozděleními Bi(n, $\pi_{i.}$) a Bi(n, $\pi_{.j}$), kde $\pi_{i.} = \sum_{j=1}^{s} \pi_{i,j}$, $\pi_{.j} = \sum_{i=1}^{r} \pi_{ij}$.

V souladu s definicí nezávislosti jevů (viz [24], odst. 6.8) řekneme, že znaky (odpovědi) A a B jsou *nezávislé*, jestliže platí

$$\pi_{ij} = \pi_{i.}\pi_{ij}, \qquad i = 1, \ldots, r, \quad j = 1, \ldots, s. \tag{15.2.1}$$

15.3 Test nezávislosti kvalitativních znaků.

Z testu χ^2 dobré shody lze odvodit *asymptotický test nezávislosti dvou kvalitativních znaků (odpovědí).* Platí-li totiž hypotéza, že znaky (odpovědi) A a B jsou nezávislé, pak (15.2.1) vyjadřuje pravděpodobnosti jednotlivých tříd $A_i \cap B_j$ jako funkce $r - 1 + s - 1 = r + s - 2$ neznámých parametrů $\pi_1, \ldots, \pi_{r-1}, \ \pi_{.1}, \ldots, \pi_{.s-1}$ z podmínky $\sum_{i=1}^{r} \pi_{i.} = \sum_{j=1}^{s} \pi_{.j} = 1$ totiž plyne

$$\pi_{r.} = 1 - \sum_{i=1}^{r-1} \pi_{i.}, \qquad \pi_{.s} = 1 - \sum_{j=1}^{s-1} \pi_{.j}. \tag{15.3.1}$$

Metodou maximální věrohodnosti podle odst. 14.2 dostaneme za platnosti hypotézy (15.2.1) s přihlédnutím k (15.3.1) pro pravděpodobnosti π_{ij} odhady

$$\hat{\pi}_{ij} = \frac{n_{i.}}{n}\frac{n_{.j}}{n} \tag{15.3.2}$$

a odtud hypotetické četnosti kombinací $A_i \cap B_j$

$$n\hat{\pi}_{ij} = \frac{n_{i.}n_{.j}}{n}. \tag{15.3.3}$$

Podle odst. 14.2 má tedy za platnosti hypotézy (15.2.1) statistika

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{ij} - n_{i.}n_{.j}/n)^2}{n_{i.}n_{.j}/n} = n\left(\sum_{i=1}^{r}\sum_{j=1}^{s} \frac{n_{ij}^2}{n_{i.}n_{.j}} - 1\right) \tag{15.3.4}$$

při velkých hodnotách n (je vhodné n takové, aby pro všechna (i, j) bylo $n_{i.}n_{.j}/n > 5$) přibližné χ^2 rozdělení o $rs - 1 - (r + s - 2) = rs - s - r + 1 =$

$(r-1)(s-1)$ stupních volnosti. Hypotézu (15.2.1) tedy zamítáme na hladině významnosti α (tj. závislost znaků A a B pokládáme za prokázanou), když

$$\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij}-n_{i.}n_{.j}/n)^2}{\frac{n_{i.}n_{.j}}{n}} > \chi^2_{1-\alpha}\big((r-1)(s-1)\big).$$

15.4 Příklad.

U $n = 320$ součástek se zjišťovala jejich výška a vnější průměr, přičemž pro oba tyto rozměry se každá součástka označila buď jako dobrá (vyhovující předepsanému rozměru), nebo jako nevyhovující. Výsledky jsou uvedeny v tab. 15.2.

	B (průměr		
A (výška)	Dobrý	Nevyhovující	Celkem
Dobrá	239	60	299
Nevyhovující	14	7	21
Celkem	253	67	320

Tab. 15.2: Četnosti n_{ij} pro $n = 320$ součástek.

(viz [10]). Testujeme hypotézu nezávislosti znaků A a B na hladině významnosti $\alpha = 0,1$. Dosazením do (15.3.4) dostáváme

$$\chi^2 = 320\left(\frac{239^2}{253\cdot 299}+\frac{60^2}{67\cdot 299}+\frac{14^2}{253\cdot 21}+\frac{7^2}{67\cdot 21}-1\right) = 2,086.$$

Protože $2,086 < \chi^2_{0,9}(1) = 2,705\,5$, hypotézu o nezávislosti znaků A a B nezamítáme.

15.5 Úlohy.

15.5.1

Zjednodušte výraz (15.3.4) pro případ $r = s = 2$ a zkontrolujte hodnotu χ^2 v příkl. 15.4.

$$\left[\chi^2 = \frac{n(n_{11}n_{22}-n_{12}n_{21})^2}{n_{1.}n_{2.}n_{.1}n_{.2}}.\right]$$

15.5.2

Vypočtěte hypotetické četnosti $n\hat{\pi}_{ij}$ pro údaje příkl. 15.4.

$$\begin{bmatrix} 236,4; & 62,6; \\ 16,6; & 4,4. \end{bmatrix}$$

Kapitola 16

Některé speciální testy dobré shody

16.1 Úvod.

Testy dobré shody uvedené v kap. 15, tj. test χ^2 a Kolmogorovův test, jsou testy do jisté míry univerzální; lze jich použít – při splnění příslušných podmínek – k ověření shody empirického rozdělení s jakýmkoliv modelem (jsou to neparametrické testy). Za tuto univerzálnost se platí sníženou účinností testů při odhalování odchylek od testované hypotézy, tj. menší silou testu proti některým alternativním hypotézám. Tato okolnost vedla k návrhům speciálních testů založených na charakteristických vlastnostech předpokládaného modelu a na charakteristických vlastnostech modelů, které v dané situaci přicházejí v úvahu jako alternativy. Tak vznikly např. speciální testy pro ověření normality rozdělení, testy k ověření shody s exponenciálním rozdělením, shody s Poissonovým rozdělením apod. V tomto článku se stručně zmíníme o některých z nich.

16.2 Testování shody s Poissonovým rozdělením.

Nechť $\mathbf{X} = (X_1, \ldots, X_n)'$ je náhodný výběr z rozdělení diskrétního typu přiřazujícího kladné pravděpodobnosti $p(x)$ všem celým nezáporným hodnotám x, tj. $\mathbf{X}$ je n-tice vzájemně nezávislých pozorování náhodné veličiny

X, která nabývá hodnot $x = 0, 1, 2, \ldots$. K ověření hypotézy

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \qquad x = 0, 1, 2, \ldots,$$

tj. k ověření hypotézy: X má Poissonovo rozdělení, se často využívá následující vlastnosti Poissonova rozdělení:

$$E(X) = \mathrm{var}(X) = \lambda,$$

viz [24], odst. 15.3; test se zakládá na srovnání výběrového průměru $\overline{X}$ s výběrovým rozptylem S^2. Jestliže X má skutečně Poissonovo rozdělení, mělo by být – až na náhodné odchylky –

$$S^2 = \overline{X}, \qquad \text{tj.} \quad \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\overline{X}} = n - 1. \tag{16.2.1}$$

Předpokládejme, že $\mathbf{X} = (X_1, \ldots, X_n)'$ je náhodný výběr z Poissonova rozdělení $\mathrm{Po}(\lambda)$. Potom sdružené rozdělení $\mathbf{X}$ podmíněné danou hodnotou $\sum_{i=1}^n X_i = t$ je

$$P\big(X_1 = x_1, \ldots, X_n = x_n \,\big|\, \sum_{i=1}^n X_i = t\big) = \\ = \frac{\mathrm{e}^{-n\lambda}\lambda^{\sum_{i=1}^n x_i} / \prod_{i=1}^n x_i!}{\mathrm{e}^{-n\lambda}(n\lambda)^t/t!} = \frac{t!}{x_1! \ldots x_n!} \prod_{i=1}^n \left(\frac{1}{n}\right)^{x_i}$$

pro všechny vektory $(x_1, \ldots, x_n)'$ splňující podmínku $\sum_{i=1}^n x_i = t$. To znamená, že náhodný vektor $\mathbf{X}$ má za podmínky $\sum_{i=1}^n X_i = t$ multinomické rozdělení s parametry $t, 1/n, \ldots, 1/n$; i-tá souřadnice X_i náhodného vektoru má – při daném $\sum_{i=1}^n X_i = t$ – rozdělení $\mathrm{Bi}(t,\, 1/n)$. Podle odst. 14.2 má tedy při dosti velkých hodnotách t statistika

$$\frac{\sum_{i=1}^n (X_i - t/n)^2}{t/n} = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\overline{X}} \tag{16.2.2}$$

přibližně rozdělení $\chi^2(n-1)$.

Odtud plyne následující test hypotézy „X má Poissonovo rozdělení“: Zamítnout hypotézu, když výběr $\mathbf{X} = (X_1, \ldots, X_n)'$ splní některou z nerovností

$$\frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\overline{X}} \le \chi^2_{\alpha/2}(n-1), \quad \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\overline{X}} \ge \chi^2_{1-\alpha/2}(n-1). \tag{16.2.3}$$

Test je aplikovatelný, když $\overline{X} > 5$.

16.3 Příklad.

Při zkouškách přístroje pro medicínské aplikace ozařování radioizotopy byly u jednoho zdroje záření a jednoho čítače impulsů naměřeny tyto počty impulsů v $n = 50$ ekvidistantních intervalech:

4 020,	4 015,	3 940,	3 988,	4 011,	4 018,	3 985,	4 001,	4 024,	4 053,
4 041,	3 951,	3 986,	3 975,	4 008,	3 982,	4 002,	4 010,	3 992,	3 991,
3 990,	4 010,	4 007,	3 989,	4 024,	4 015,	4 009,	3 986,	4 025,	3 974,
4 002,	4 032,	3 990,	3 970,	3 990,	3 987,	3 899,	4 005,	3 992,	4 003,
3 965,	4 031,	3 980,	4 010,	4 015,	3 982,	3 965,	4 012,	4 007,	3 998.

Zde máme

$$\sum_{i=1}^{50} x_i = 199\,857; \qquad \overline{x} = 3\,997,14;$$

$$\sum_{i=1}^{50} (x_i - \overline{x})^2 = 33\,558,02; \quad \sum_{i=1}^{50} \frac{(x_i-\overline{x})^2}{\overline{x}} = 8,40.$$

Pozorovaná hodnota podílu $\sum_{i=1}^{50}(X_i - \overline{X})^2/\overline{X}$ je tedy o mnoho menší než $\chi^2_{0,025}(49) = 31,555$. Rozptyl $s^2 = 685,98$ je statisticky významně menší než $\overline{x} = 3\,997,34$. Prokázala se tedy významná odchylka od Poissonova rozdělení. V dané úloze může mít výsledek tento výklad: O radioaktivním záření je známo (teoreticky zdůvodněno i experimentálně potvrzeno), že počty emisí X v ekvidistantních intervalech mají při konstantní intenzitě Poissonovo rozdělení, tedy $\mathrm{var}(X) = E(x)$. Kdyby se intenzita záření v průběhu měření měnila, očekávali bychom $\mathrm{var}(X) > E(X)$. V dané sérii měření se však ukázalo $\mathrm{var}(X) < E(X)$. To může být způsobeno vlastnostmi registračního zařízení; některé čítače mají tzv. „mrtvou dobu“, po registraci impulsu se čítač na určitý čas zastaví a během tohoto času žádný impuls neregistruje. Je-li tato „mrtvá doba“ dlouhá v poměru ke střední hodnotě doby mezi impulsy, pak počty impulsů registrované v intervalech pevné délky mají menší rozptyl i střední hodnotu, přičemž rozptyl klesá více než střední hodnota. V našem příkladě lze tedy na základě výsledku usuzovat na existenci „mrtvé doby“ poměrně dlouhé ve srovnání se střední délkou intervalu mezi emisemi, což znamená, že čítač registruje jen část impulsů a skutečná intenzita záření je podceněna.

16.4 Test normality.

O normálním rozdělení je známo, že má nulové koeficienty šikmosti a špičatosti, tj. že

$$\alpha_3 = \frac{E\big((X-\mu)^3\big)}{\sigma^3} = 0, \quad \alpha_4 = \frac{E\big((X-\mu)^4\big)}{\sigma^4} - 3 = 0, \tag{16.4.1}$$

viz [24], odst. 18.4. Toho se někdy využívá k ověření hypotézy, že X má normální rozdělení. Z výběru se vypočtou odhady koeficientů α_3 a α_4,

$$A_3 = \frac{\sqrt{n}\sum_{i=1}^n (X_i - \overline{X})^3}{\big(\sum_{i=1}^n (X_i - \overline{X})^2\big)^{3/2}}, \qquad A_4 = \frac{n\sum_{i=1}^n \big(X_i - \overline{X}\big)^4}{\big(\sum_{i=1}^n (X_i - \overline{X})^2\big)^2} - 3, \tag{16.4.2}$$

a porovnají se s hodnotami (16.4.1), tj. ověří se významnost odchylky A_3 od 0 a A_4 od 0. Má-li X rozdělení $\mathrm{N}(\mu, \sigma^2)$, pak při velkých rozsazích výběru přibližné rozdělení A_3 je $\mathrm{N}\big(0, \mathrm{var}(A_3)\big)$ a přibližné rozdělení A_4 je $\mathrm{N}\big(E(A_4), \mathrm{var}(A_4)\big)$, kde (viz [1], str. 206)

$$\begin{aligned} E(A_3) &= 0, & \mathrm{var}(A_3) &= \frac{6(n-2)}{(n+1)(n+3)}, \\ E(A_4) &= -\frac{6}{n+1}, & \mathrm{var}(A_4) &= \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}. \end{aligned} \tag{16.4.3}$$

Hypotéza normality X se tedy zamítá, když

$$|A_3| \geq u_{1-\alpha/2}\sqrt{\mathrm{var}(A_3)} \quad \text{nebo} \quad |A_4 + \frac{6}{n+1}| \geq u_{1-\alpha/2}\sqrt{\mathrm{var}(A_4)}. \tag{16.4.4}$$

Jiný test hypotézy, že X má normální rozdělení, je založen na pořádkových statistikách; postup při jeho užití a potřebné tabulky lze nalézt v [13], zde je uvedena jen základní myšlenka.

Nechť $\mathbf{X} = (X_1, \ldots, X_n)'$ je náhodný výběr z rozdělení $\mathrm{N}(\mu, \sigma^2)$. Střední hodnota i-té pořádkové statistiky $X_{(i)}$ z tohoto výběru je

$$E(X_{(i)}) = \mu + a_i\sigma, \tag{16.4.5}$$

kde a_i jsou konstanty (viz dále odst. 21.2) tabelované např. v [30, 32]. Má-li tedy X rozdělení $\mathrm{N}(\mu, \sigma^2)$, pak body $(a_i, X_{(i)})$ budou seskupeny – až na náhodné odchylky – kolem přímky se směrnicí σ. Jinými slovy, regrese $X_{(i)}$

na a_i (viz kap. V) je lineární a regresní přímka má směrnici σ. Test normality rozdělení náhodné veličiny X spočívá v tom, že se metodou popsanou dále v odst. 21.5 odhadne σ a odhad se porovná s obvyklým odhadem založeným na $\sum_{i=1}^{n}(X_i - \overline{X})^2$. Rozsáhlými pokusy bylo zjištěno, že tento test je vůči většině odchylek od normality citlivější než ostatní známé testy dobré shody.

16.5 Testy exponenciálnosti.

Nechť $\mathbf{X} = (X_1, \ldots, X_n)'$ je náhodný výběr z rozdělení spojitého typu, které má hustotu pravděpodobnosti $f(x)$ kladnou pro všechna $x > 0$ a nulovou pro $x \leq 0$. K ověření hypotézy, že

$$f(x) = \frac{1}{\delta}\mathrm{e}^{-x/\delta}, \qquad x > 0,$$

tj. že X má rozdělení E(0, δ), se doporučuje v [13] test založený na statistice

$$\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\overline{X}^2} = n^2 \sum_{i=1}^{n} \left| \frac{X_i}{\sum_{i=1}^{n} X_i} - \frac{1}{n} \right|^2. \tag{16.5.1}$$

Ve prospěch této statistiky mluví dva důvody. Především, má-li X rozdělení E(0, δ), pak $E(X) = \delta$ a $\mathrm{var}(X) = \delta^2 = E^2(X)$. Ve výběrech z rozdělení E(0, δ) by tedy mělo být S^2 blízké $\overline{X}^2$ čili $\sum_{i=1}^{n}(X_i - \overline{X})^2/\overline{X}^2$ blízké $n - 1$. Za druhé, rozdělení náhodného vektoru

$$\mathbf{Z} = (Z_1, \ldots, Z_{n-1})', \tag{16.5.2}$$

kde

$$Z_i = X_1 + \ldots + X_i, i = 1, \ldots, n-1, \tag{16.5.3}$$

podmíněné danou hodnotou $Z_n = \sum_{i=1}^{n} X_i = z$, má hustotu

$$f(z) = \frac{(n-1)!}{z^{n-1}}, \qquad 0 < z_1 < \ldots < z_{n-1} < z. \tag{16.5.4}$$

To je však hustota uspořádaného výběru rozsahu $n - 1$ z rovnoměrného rozdělení na intervalu $(0, z)$. Je snadné se přesvědčit, že potom

$$E(Z_i - Z_{i-1}) = E(X_i) = \frac{z}{n}, \tag{16.5.5}$$

což znamená, že při výběru X rozsahu n z $\mathrm{E}(0,\,\delta)$ lze považovat $X_1, \ldots, X_n$ za délky intervalů, na které dělí pořádkové statistiky výběru rozsahu $n-1$ z rovnoměrného rozdělení na $(0,\,\sum_{i=1}^{n} X_i)$ interval $(0,\,\sum_{i=1}^{n} X_i)$. Tyto intervaly jsou – až na náhodné odchylky – rovny $\sum_{i=1}^{n} X_i/n$. Statistika (16.5.1) je právě rovna součtu čtverců odchylek $X_i/\sum_{i=1}^{n} X_i$ od hypotetických hodnot $1/n$.

V [13] jsou uvedeny tabulky kvantilů statistiky (16.5.1) a popsán též test obecnější hypotézy, že X má rozdělení $\mathrm{E}(A,\,\delta)$ s neznámým A.

Část V

Regresní analýza

Kapitola 17

Lineární regrese s jednou vysvětlující proměnnou

17.1 Regresní funkce.

Jedním z důležitých úkolů všech technických oborů je hledání a studium závislostí proměnných. Tak nás např. zajímá závislost pevnosti oceli na jejím chemickém složení, závislost doby do lomu na napětí a teplotě při hodnocení zkoušek pevnosti při tečení apod.

V technických a přírodních vědách se často pracuje s funkčními vztahy, kde závisle proměnná y je jednoznačně určena funkcí $r \geq 1$ nezávisle proměnných $x_1, \ldots, x_r$, tj.

$$y = \varphi(x_1, \ldots, x_r).$$

Pak pro daný vektor $\mathbf{x}' = (x_1, \ldots, x_r)$ dostaneme jedinou hodnotu závisle proměnné y.

Často však, v důsledku působení náhodných činitelů, má závisle proměnná povahu náhodné veličny mající určité rozdělení pravděpodobnosti.

Nezávisle proměnné mohou být buď nenáhodné (fixní) proměnné, nebo též náhodné veličiny.

V dalším budeme uvažovat případy, kdy závisle proměnná je náhodná veličina; označíme ji Y. Předpokládejme, že podmíněná střední hodnota náhodné veličiny Y pro dané hodnoty $x_1, \ldots, x_r$ je rovna

$$E(Y \mid x_1, \ldots, x_r) = \eta(x_1, \ldots, x_r; \beta_1, \ldots, \beta_p), \tag{17.1.1}$$

tj. je funkcí $x_1, \ldots, x_r$ a $p \geq 1$ parametrů $\beta_1, \ldots, \beta_p$.

Funkce (17.1.1) se nazývá *regresní funkce* a parametry $\beta_1, \ldots, \beta_p$ *regresní parametry* (nebo *regresní koeficienty).*

Příkladem regresní funkce je pro $r = 2$ a $p = 4$ funkce

$$E(Y \mid x_1, x_2) = \eta(x_1, x_2; \beta_1, \beta_2, \beta_3, \beta_4) = \beta_1 + \beta_2 x_1 + \beta_3 x_2 + \beta_4 x_1 x_2$$

nebo pro $r = 1$ a $p = 2$ funkce

$$E(Y|x_1) = \beta_1 + e^{\beta_2 x_2}; \tag{17.1.2}$$

první z těchto regresních funkcí je lineární v parametrech, druhá je nelineární v parametrech β_j.

Dále budeme uvažovat případy, kdy η má tvar

$$\eta = \beta_0 x_0 + \beta_1 x_1 + \ldots + \beta_k x_k, \tag{17.1.3}$$

přičemž $x_0, x_1, \ldots, x_k$ jsou nenáhodné (fixní) proměnné. Je tedy $r = p = k+1$. Velmi častým případem je případ $x_0 = 1$.

Máme-li n vektorů pozorování $(Y_i, x_{0i}, \ldots, x_{li}, \ldots, x_{ki})'$, $i = 1, \ldots, n$, můžeme z nich nalézt odhady regresních parametrů $\beta_0, \beta_1, \ldots, \beta_k$ a odhady regresní funkce η v daném bodě $\mathbf{x}' = (x_0, x_1, \ldots, x_k)$.

V pracích o regresní analýze se často proměnné $x_0, x_1, \ldots, x_k$ nazývají *vysvětlující proměnné* a veličina Y *vysvětlovaná proměnná.*

17.2 Regresní přímka.

Mějme jednu vysvětlující proměnnou x a nechť

$$E(Y \mid x) = \eta = \beta_0 + \beta_1 x. \tag{17.2.1}$$

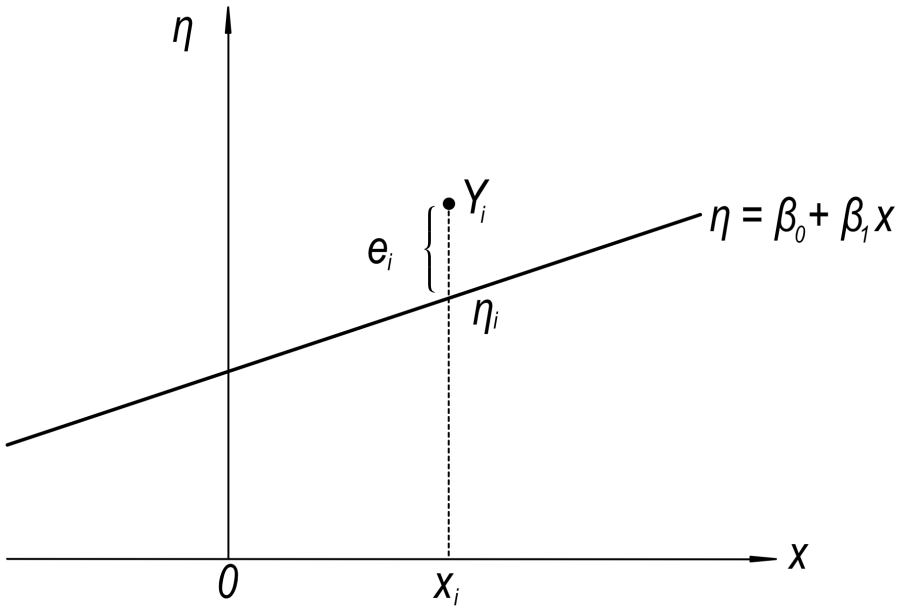
Tato funkce se nazývá *regresní přímka.* Pro fixní proměnnou x lze (17.2.1) přepsat též na tvar

$$Y = \beta_0 + \beta_1 x + e, \tag{17.2.2}$$

kde e je náhodná odchylka (též náhodná chyba).

Mějme $n > 2$ vektorů pozorování $(Y_i, x_i)'$, $i = 1, \ldots, n$, kde $x_1, \ldots, x_n$ jsou daná čísla, z nichž aspoň dvě jsou navzájem různá, a $Y_1, \ldots, Y_n$ jsou náhodné veličiny. Přitom vzhledem k (17.2.2) je (viz též obr. 11)

$$Y_i = \eta_i + e_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, \ldots, n. \tag{17.2.3}$$



Obr. 17.1: Regresní přímka

Odhady b_0 a b_1 parametrů β_0 a β_1 se naleznou *metodou nejmenších čtverců,* tj. z podmínky, aby výraz

$$\mathcal{S} = \sum_{i=1}^{n} (Y_i - \eta_i)^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$$

byl minimální (viz dále, odst. 18.2).

Ze vztahů

$$\frac{\partial \mathcal{S}}{\partial \beta_j} = 0, \qquad j = 0, 1,$$

dostáváme soustavu dvou rovnic

$$\begin{aligned} n b_0 + b_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} Y_i, \\ \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i Y_i, \end{aligned} \tag{17.2.4}$$

která se nazývá *soustava normálních rovnic.* Jejím řešením získáme odhady

b_0 a b_1 parametrů β_0 a β_1,

$$b_0 = \frac{1}{n}\Big(\sum_{i=1}^n Y_i - b_1 \sum_{i=1}^n x_i\Big) = \overline{Y} - b_1\overline{x}, \tag{17.2.5}$$

$$b_1 = \frac{\sum_{i=1}^n (x_i-\overline{x})Y_i}{\sum_{i=1}^n (x_i-\overline{x})^2} = \frac{n\sum_{i=1}^n x_iY_i - \big(\sum_{i=1}^n x_i\big)\big(\sum_{i=1}^n Y_i\big)}{n\sum_{i=1}^n x_i^2 - \big(\sum_{i=1}^n x_i\big)^2}.$$

Statistiky b_0 a b_1 jsou odhady regresních parametrů β_0 a β_1 získané metodou nejmenších čtverců. Je vidět, že b_0 i b_1 jsou lineární formy veličin $Y_1, \ldots, Y_n$; říká se též, že b_0 a b_1 jsou *lineární odhady* parametrů β_0 a β_1.

Odhadem hodnoty regresní funkce $\eta = \beta_0 + \beta_1 x$ pro dané x je statistika

$$\hat{Y} = b_0 + b_1 x = \overline{Y} + b_1(x - \overline{x}). \tag{17.2.6}$$

Odchylky

$$\hat{e}_i = Y_i - \hat{Y}_i = Y_i - b_0 - b_1x_i = Y_i - \overline{Y} - b_1(x_i - \overline{x}), \quad i = 1, \ldots, n, \tag{17.2.7}$$

se nazývají *rezidua* a statistika

$$S_R = \sum_{i=1}^n \hat{e}_i^2 = \sum_{i=1}^n \big(Y_i - \hat{Y}_i\big)^2 = \sum_{i=1}^n \big(Y_i - \overline{Y} - b_1(x_i - \overline{x})^2\big) \tag{17.2.8}$$

se nazývá *reziduální součet čtverců.*

Protože platí

$$S_R = \sum_{i=1}^n (Y_i - \overline{Y})^2 + b_1^2 \sum_{i=1}^n (x_i - \overline{x})^2 - 2b_1 \sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y}) =$$

$$= \sum_{i=1}^n Y_i^2 - nY^2 - b_1^2 \sum_{i=1}^n (x_i - \overline{x})^2 = \sum_{i=1}^n Y_i^2 - b_0 \sum_{i=1}^n Y_i - b_1\overline{x}n\overline{Y} - b_1 \sum_{i=1}^n (x_i - \overline{x})Y_i,$$

lze S_R vyjádřit ve tvaru

$$S_R = \sum_{i=1}^n Y_i^2 - b_0 \sum_{i=1}^n Y_i - b_1 \sum_{i=1}^n x_iY_i, \tag{17.2.9}$$

tj. od součtu čtverců veličin Y_i se odečte odhad b_0 násobený pravou stranou první rovnice a odhad b_1 násobený pravou stranou druhé rovnice soustavy (17.2.4).

17.3 Střední hodnoty a rozptylu odhadů.

Z (17.2.1) a (17.2.2) vyplývá, že $E(e) = 0$. Předpokládejme v (17.2.3), že

$$E(e_i) = 0, \quad \operatorname{var}(e_i) = E(e_i^2) = \sigma^2, \quad i = 1, \dots, n, \tag{17.3.1}$$

a

$$\operatorname{cov}(e_{i1}, e_{i2}) = E(e_{i1}, e_{i2}) = 0, \quad i_1,\ i_2 = 1, \dots, n, \quad i_1 \neq n_2, \tag{17.3.2}$$

tj. že náhodné chyby mají nulovou střední hodnotu, týž neznámý rozptyl $\sigma^2 > 0$ a že jsou nekorelované.

Stanovme za těchto předpokladů střední hodnoty a rozptyly statistik b_0 a b_1. Zřejmě

$$\Big(\sum_{i=1}^{n}(x_i - \overline{x})^2\Big)E(b_1) = \sum_{i=1}^{n}(x_i - \overline{x})E(Y_i) = \sum_{i=1}^{n}(x_i - \overline{x})(\beta_0 + \beta_1 x_i) =$$

$$= \beta_1 \sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Statistiku b_0 můžeme přepsat na tvar

$$b_0 = \sum_{i=1}^{n}\Big(\frac{1}{n} - \frac{(x_i - \overline{x})\overline{x}}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\Big)Y_i = \sum_{i=1}^{n} c_i Y_i.$$

Odtud

$$E(b_0) = \sum_{i=1}^{n} c_i E(Y_i) = \sum_{i=1}^{n} c_i(\beta_0 + \beta_1 x_i) = \beta_0.$$

Je tedy

$$E(b_j) = \beta_j, \quad j = 0, 1, \tag{17.3.3}$$

tzn. že statistika b_j je nestranný odhad parametru β_j, $j = 0, 1$.

Protože veličiny $e_1, \dots, e_n$ jsou nekorelované, jsou i veličiny $Y_1, \dots, Y_n$ nekorelované. Tudíž rozptyl (viz [24], vztah (11.7.12), a protože $\operatorname{var}(Y_i) = \operatorname{var}(e_i) = \sigma^2$, $i = 1, \dots, n$)

$$\operatorname{var}(b_0) = \sum_{i=1}^{n} c_i^2 \operatorname{var}(Y_i)\,\sigma^2 \sum_{i=1}^{n}\Big(\frac{1}{n} - \frac{(x_i - \overline{x})\overline{x}}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\Big)^2$$

a po rozvoji a úpravě dostáváme

$$var(b_0) = \Big(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\Big)\sigma^2 = \frac{\sum_{i=1}^n x_i^2}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}\sigma^2 = \sigma_{b_0}^2. \tag{17.3.4}$$

Dále

$$\mathrm{var}(b_1) = \frac{1}{\big(\sum_{i=1}^n (x_i - \overline{x})^2\big)^2} \sum_{i=1}^n (x_i - \overline{x})^2 \mathrm{var}(Y_i),$$

takže

$$\mathrm{var}(b_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{n\sigma^2}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} = \sigma_{b_1}^2. \tag{17.3.5}$$

Uvažujme ještě statistiku (17.2.6). Přepišme ji na tvar

$$\hat{Y} = \sum_{i=1}^n \Big(\frac{1}{n} + \frac{(x_i - \overline{x})(x - \overline{x})}{\sum_{i=1}^n (x_i - x)^2}\Big) Y_i = \sum_{i=1}^n c_i Y_i.$$

Střední hodnota

$$E(\hat{Y}) = \sum_{i=1}^n c_i(\beta_0 + \beta_1 x_i) = \beta_0 + \beta_1 \overline{x} + \beta_1(x - \overline{x}) = \beta_0 + \beta_1 x \tag{17.3.6}$$

a rozptyl

$$\mathrm{var}(\hat{Y}) = \sigma^2 \sum_{i=1}^n c_i^2 = \Big(\frac{1}{n} + \frac{n(x - \overline{x})^2}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}\Big)\sigma^2 = \sigma_{\hat{Y}}^2. \tag{17.3.7}$$

Je tedy $\hat{Y}$ lineární nestranný odhad parametrické funkce $\eta = \beta_0 + \beta_1 x$ pro dané x.

17.4 Odhad rozptylu σ^2.

Statistika S_R má střední hodnotu

$$E(S_R) = \sum_{i=1}^n E(Y_i)^2 - nE(\overline{Y}^2) - \Big(\sum_{i=1}^n (x_i - \overline{x})^2\Big) E(b_1^2) =$$

$$= \sum_{i=1}^{n}(\sigma^2 + \eta_i^2) - n\Big(\frac{\sigma^2}{n} + \overline{\eta}^2\Big) - \sigma^2 - \beta_1^2 \sum_{i=1}^{n}(x_i - \overline{x})^2,$$

kde

$$\eta_i = \beta_0 + \beta_1 x_i, \quad i = 1, \ldots, n, \quad \overline{\eta} = \frac{1}{n}\sum_{i=1}^{n}\eta_i = \beta_0 + \beta_1 \overline{x}.$$

Je tedy

$$E(S_R) = (n-2)\sigma^2 + \sum_{i=1}^{n}(\eta_i - \overline{\eta})^2 - \beta_1^2 \sum_{i=1}^{n}(x_i - \overline{x})^2,$$

tj.

$$E(S_R) = (n-2)\sigma^2. \tag{17.4.1}$$

Nestranným odhadem rozptylu $\sigma^2 = \mathrm{var}(e_i) = \mathrm{var}(Y_i), i = 1, \ldots, n$, je statistika

$$S^2 = \frac{S_R}{n-2} =$$
$$= \frac{1}{n(n-2)}\left(n\sum_{i=1}^{n}Y_i^2 - (\sum_{i=1}^{n}Y_i)^2 - \frac{\big(n\sum_{i=1}^{n}x_iY_i - (\sum_{i=1}^{n}x_i)(\sum_{i=1}^{n}Y_i)\big)^2}{n\sum_{i=1}^{x}{}_i^2 - (\sum_{i=1}^{n}x_i)^2}\right). \tag{17.4.2}$$

Pro kontrolu lze S^2 vypočítat pomocí vztahu (17.2.9).

Statistika

$$S_{b_j}^2 = S^2\left(\frac{1}{\sigma^2}\sigma_{b_j}^2\right), \quad j = 0, 1, \tag{17.4.3}$$

je pak nestranný odhad $\sigma_{b_j}^2, j = 0, 1$, a statistika

$$S_{\hat{Y}}^2 = S^2\left(\frac{1}{\sigma^2}\sigma_{\hat{Y}}^2\right) = S^2\left(\frac{1}{n} + \frac{n(x-\overline{x})^2}{n\sum_{i=1}^{n}x_i^2 - (\sum_{i=1}^{n}x_i)^2}\right) \tag{17.4.4}$$

je nestranný odhad $\sigma_{\hat{Y}}^2$.

17.5 Intervaly spolehlivosti a testy hypotéz.

Přidejme k předpokladům (17.3.1) a (17.3.2) ještě předpoklad, že $e_1, \ldots, e_n$ mají normální rozdělení. Předpokládáme tedy (viz [24], odst. 24.3), že náhodné

veličiny $e_1, \ldots, e_n$ jsou vzájemně nezávislé, všechny mají rozdělení $\mathrm{N}(0,\, \sigma^2)$. Tento předpoklad lze též formulovat takto: Náhodný vektor $\mathbf{Y} = (Y_1, \ldots, Y_n)'$, jehož složkami jsou veličiny (17.2.3), má n-rozměrné normální rozdělení (viz [24], odst.24.1) $\mathrm{N}_n(\boldsymbol{\eta},\, \sigma^2\mathbf{I}_n)$, kde

$$\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)'. \tag{17.5.1}$$

Protože statistiky b_0 a b_1 představují lineární formy veličin $Y_1, \ldots, Y_n$, má (viz [24], odst. 24.7) vektor $\mathbf{b} = (b_0, b_1)'$ dvourozměrné normální rozdělení, takže (viz [24], odst. 24.5) statistika b_j má rozdělení $\mathrm{N}(\beta_j,\, \sigma^2_{b_j})$, $j = 0, 1$. Statistika (17.2.6), která je také lineární formou veličiny $Y_1, \ldots, Y_n$, má rozdělení $\mathrm{N}(\beta_0 + \beta_1 x,\, \sigma^2_{\hat{Y}})$.

Dále náhodná veličina S_R/σ^2 má rozdělení $\chi^2(n-2)$ a $\mathbf{b}$ a S_R jsou nezávislé (viz dále odst. 18.6 a 18.7). Tudíž náhodné veličiny $(b_j - \beta_j)/\sigma_{b_j}$, $j = 0, 1$, mají rozdělení $\mathrm{N}(0,\, 1)$ a náhodné veličiny

$$T = \frac{b_j - \beta_j}{S_{b_j}}, \qquad j = 0, 1, \tag{17.5.2}$$

kde $S^2_{b_j}$ jsou statistiky (17.4.3), mají (viz odst. 3.3) Studentovo rozdělení $\mathrm{t}(n-2)$. Obdobně náhodná veličina

$$T = \frac{\hat{Y} - (\beta_0 + \beta_1 x)}{S_{\hat{Y}}}, \tag{17.5.3}$$

kde $S_{\hat{Y}}$ je statistika (17.4.4), má rozdělení $\mathrm{t}(n-2)$.

Pomocí těchto veličin můžeme zkonstruovat intervaly spolehlivosti pro parametry β_0 a β_1 či pro hodnotu regresní funkce $\eta = \beta_0 + \beta_1 x$ pro dané x nebo testovat hypotézy o těchto parametrických funkcích.

17.6 Příklad.

Při hodnocení zkoušek na únavu lze popsat závislost logaritmu počtu kmitů do lomu, $Y = \ln V$, na napětí x rovnicí

$$Y = \beta_0 + \beta_1 x + e. \tag{17.6.1}$$

Tabulka 17.1 udává hodnoty napětí x_i (MPA) a ϑ_i (počet kmitů), $i = 1, \ldots, 16$ (viz [7], str. 7). Z posledního řádku této tabulky dostáváme

$$\begin{aligned} b_1 &= \frac{16 \cdot 139\,447,176\,251 - 9\,560 \cdot 234,202\,085}{16 \cdot 5\,730\,000 - 9\,560^2} \doteq -0,027\,294; \\ b_0 &= \frac{1}{16}(234,202\,085 + 0,027\,294 \cdot 9\,560) \doteq 30,945\,795; \\ s^2 &= \frac{1}{16 \cdot 14}\Bigg(16 \cdot 3\,447,199\,511 - 234,202\,085^2 - \\ &\quad - \frac{(16 \cdot 139\,447,176\,251 - 9\,560 \cdot 234,202\,085)^2}{16 \cdot 5\,730\,000 - 9\,560^2}\Bigg) \doteq 0,406\,076. \end{aligned}$$

x_i	$\vartheta_i 10^{-3}$	$y_i = \ln \vartheta_i$	$x_i y_i$	x_i^2	y_i^2
560	2845	14,861 074	8 322,201 261	313 600	220,851 511
560	3322	15,016 078	8 409,003 467	313 600	225,482 587
560	9411	16,057 390	8 992,138 338	313 600	257,839 770
560	14713	16,504 242	9 242,375 610	313 600	272,390 009
580	2597	14,769 868	8 566,523 173	336 400	218,148 987
580	4429	15,303 684	8 876,136 981	336 400	234,202 758
580	5523	15,524 432	9 004,170 461	336 400	241,007 984
580	6868	15,742 384	9 130,582 482	336 400	247,822 641
600	554	13,224 920	7 934,951 988	360 000	174,898 508
600	1227	14,020 083	8 412,049 650	360 000	196,562 720
600	3446	15,052 725	9 031,634 850	360 000	226,584 522
600	3684	15,119 510	9 071,705 844	360 000	228,599 575
650	348	12,759 958	8 293,972 544	422 500	162,816 522
650	530	13,180 632	8 567,410 995	422 500	173,729 068
650	728	13,498 056	8 773,736 621	422 500	182,197 525
650	780	13,567 049	8 818,581 986	422 500	184,064 824
9 560		234,202 085	139 447,176 251	5 730 000	3 447,199 511

Tab. 17.1: Hodnoty potřebné pro výpočet statistik b_0, b_1 a S^2.

Testujme hypotézu $H : \beta_1 = 0$ proti alternativě $A : \beta_1 < 0$ na hladině významnosti $\alpha = 0,05$. Kritickým oborem je

$$W = \Big\{(\mathbf{x}, \mathbf{y}) \,|\, t = \frac{b_1}{s_{b_1}} \leq -t_{0,95}(14)\Big\},$$

kde

$$s_{b_1}^2 = \frac{16s^2}{16\sum_{i=1}^{16} x_i^2 - (\sum_{i=1}^{16} x_i)^2} = 2\,269 \cdot 10^{-8}.$$

Protože

$$t = \frac{-0,027\,294}{0,004\,763} \doteq -5,73 < -1,761\,3 = -t_{0,95}(14),$$

hypotézu $H : \beta_1 = 0$ zamítáme.

17.7 Použití pro některé jiné regresní funkce.

Výsledků předchozích odstavců můžeme využít pro regresní funkce

$$\eta = \beta_0 + \beta_1 f(t), \tag{17.7.1}$$

kde t je nenáhodná (fixní) proměnná a $f(t)$ je známá funkce této proměnné. Položíme-li $x = f(t)$, převedeme (17.7.1) na regresní přímku. Nejčastějšími případy jsou:

$$\eta = \beta_0 + \beta_1 \log t, \quad t > 0, \tag{17.7.2}$$

$$\eta = \beta_0 + \beta_1 t^r, \qquad \text{pro } r < 0 \text{ musí být } t \neq 0, \tag{17.7.3}$$

$$\eta = \beta_0 + \beta_1 \mathrm{e}^t. \tag{17.7.4}$$

Za předpokladu, že platí vztahy $Y_i = \eta_i + e_i,\ i = 1, \ldots, n$, se statistiky $b_0,\ b_1,\ S_R^2,\ S^2$ atd. získají postupy uvedenými v předchozích odstavcích, přičemž se všude dosadí $x_i = f(t_i),\ i = 1, \ldots, n$.

Uvažujme ještě regresní model

$$V_i = \zeta_i u_i = \delta\big(f(t_i)\big)^{\beta_1} u_i, \quad i = 1, \ldots, n, \tag{17.7.5}$$

kde V_i a u_i jsou náhodné veličiny nabývajících kladných hodnot a $\delta > 0,\ f(t_i) > 0,\ i = 1, \ldots, n$. Označíme-li

$$Y_i = \ln V_i, \quad \beta_0 = \ln \delta, \quad x_i = \ln f(t_i), \quad e_i = \ln u_i, \quad i = 1, \ldots, n, \tag{17.7.6}$$

převedeme (17.7.5) na (17.2.3) (místo přirozených logaritmů můžeme uvažovat i jiné, např. dekadické logaritmy).

Platí-li pro $e_i = \ln u_i$ předpoklady (17.3.1) a (17.3.2), jsou (17.2.5), v nichž se za Y_i a x_i dosadí výrazy (17.7.6), lineární nestranné odhady parametrů

$\beta_0 = \ln\delta$ a β_1 a nestranné odhady rozptylů $\sigma^2_{b_j}$ jsou dány výrazy (17.3.3). Mají-li navíc veličiny $e_1, \dots, e_n$ normální rozdělení (tj. mají-li veličiny $u_1, \dots, u_n$ logaritmicko-normální rozdělení, viz [24], odst. 19.1), mají veličiny (17.5.2) a (17.5.3) rozdělení $\mathrm{t}(n-2)$.

Příkladem funkce $f(t)$ v (17.7.5) je $f(t) = t$ nebo $f(t) = K^t$, kde K je dané kladné číslo, např. $K = 10$ nebo $K = 2,718\dots$.

17.8 Příklad.

Vztah (17.6.1) uvažovaný v příkl. 17.6 vznikl logaritmováním vztahu

$$V = \zeta u = \delta\, \mathrm{e}^{\beta_1 x} u, \tag{17.8.1}$$

přičemž jsme označili $Y = \ln V,\ \beta_0 = \ln\delta$ a $e = \ln u$.

Nalezněme nyní 95% interval spolehlivosti pro hodnotu $\zeta = \delta e^{\beta_1 x}$ při napětí $x = 630$. Z (17.2.6) a (17.4.4) dostáváme

$$\hat{Y} = 14,637\,630 - 0,027\,294(630 - 597,5) \doteq 13,750\,575;$$

$$s^2_{\hat{Y}} = 0,406\,077\left(\frac{1}{16} + \frac{16(630 - 597,5)^2}{16 \cdot 5\,730\,000 - 9\,560^2}\right) \doteq 0,049\,341$$

a z tabulek nalezneme $t_{0,975}(14) = 2,144\,8$. 95% interval spolehlivosti pro $\eta = \ln\zeta$ je

$$\left(\hat{Y} - t_{0,975}(14)S_{\hat{Y}},\ \hat{Y} + t_{0,975}(14)S_{\hat{Y}}\right),$$

tj. v našem případě $(13,274; 14,227)$. Odtud dostáváme 95% interval spolehlivosti pro $\zeta = e^{\eta}$

$$\left(582 \cdot 10^3,\ 1\,509 \cdot 10^3\right).$$

17.9 Úlohy.

17.9.1

Pro údaje příkl. 17.8 uvažujte regresní model $V = \delta\, t^{\beta_1} u$, kde jsme použili označení t místo x. Nalezněte nejlepší nestranné odhady b_0 a b_1 parametrů $\beta_0 = \ln\delta$ a β_1 a určete s^2. Nalezněte dvoustranný 95% interval spolehlivosti pro hodnotu $\zeta = \delta\, t^{\beta_1}$ v bodě $t = 630$ a porovnejte tento interval s intervalem příkl. 17.8.

$$\left[b_0 = 120,416;\ b_1 = -16,551;\ s^2 = 0,406;\ (213 \cdot 10^3, 3\,993 \cdot 10^3).\right]$$

17.9.2

Uvažujte regresní funkci $\eta = \beta_0 + \beta_1 t^{-1}, t \neq 0$. Stanovte b_0, b_1 a S_R.

$$\left[\begin{array}{l} b_1 = \frac{n\sum_{i=1}^n t_i^{-1}Y_i - (\sum_{i=1}^n t_i^{-1})(\sum_{i=1}^n Y_i)}{n\sum_{i=1}^n t_i^{-2} - \sum_{i=1}^n t_i^{-1}}, \\ \\ b_0 = \frac{1}{n}(\sum_{i=1}^n Y_i - b_1 \sum_{i=1}^n t_i^{-1}), \\ \\ S_R = \sum_{i=1}^n Y_i^2 - b_0 \sum_{i=1}^n Y_i - b_1 \sum_{i=1}^n t_i^{-1} Y_i. \end{array}\right]$$

Kapitola 18

Metoda nejmenších čtverců

18.1 Regresní model lineární v parametrech.

Mějme n náhodných veličin $Y_1, \ldots, Y_n$, které se dají vyjádřit jako

$$Y_i = \eta_i + e_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + e_i, \quad i = 1, \ldots, n, \tag{18.1.1}$$

tj. jako lineární funkce $k+1$ regresních parametrů $\beta_0, \beta_1, \ldots, \beta_k$. Veličiny $e_1, \ldots, e_n$ jsou náhodné chyby.

Označme

$$\mathbf{Y} = (Y_1, \ldots, Y_n)', \quad \mathbf{e} = (e_1, \ldots, e_n)' \tag{18.1.2}$$

a

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)', \quad \boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)'. \tag{18.1.3}$$

Pak (18.1.1) lze napsat ve tvaru

$$\mathbf{Y} = \boldsymbol{\eta} + \mathbf{e} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \tag{18.1.4}$$

kde

$$\mathbf{X} = \begin{pmatrix} x_{10}, & x_{11}, & \ldots, & x_{1k} \\ x_{20}, & x_{21}, & \ldots, & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n0}, & x_{n1}, & \ldots, & x_{nk} \end{pmatrix} \tag{18.1.5}$$

Nechť $n > k+1$, nechť x_{ij} jsou daná čísla (tj. proměnné $x_0, x_1, \ldots, x_k$ jsou nenáhodné) a nechť matice $\mathbf{X}$ má hodnost $h(\mathbf{X}) = k+1$.

Dále nechť náhodný vektor $\mathbf{e}$ má (viz [24], odst. 11.7) střední hodnotu $(0,\ldots,0)'$ a kovarianční matici $\sigma^2\mathbf{I}_n$, kde σ^2 je neznámý kladný parametr a $\mathbf{I}_n$ je jednotková matice typu (n,n). Tyto předpoklady o vektoru $\mathbf{e}$ lze též vyjádřit vztahy (17.3.1) a (17.3.2).

Všechny uvedené předpoklady popisují model, který se nazývá *regresní model lineární v parametrech.*

18.2 Soustava normálních rovnic.

Odhady $b_0,\ldots,b_k$ regresních parametrů (koeficientů) $\beta_0,\ldots,\beta_k$ metodou nejmenších čtverců jsou statistiky, pro které součet čtverců

$$\mathcal{S}=\sum_{i=1}^{n}\big(Y_i-\sum_{j=0}^{k}b_jx_{ij}\big)^2\le\sum_{i=1}^{n}\big(Y_i-\sum_{j=0}^{k}\beta_jx_{ij}\big)^2\quad\text{pro všechna}\quad\beta_0,\ldots,\beta_k. \tag{18.2.1}$$

Odhady b_j lze nalézt řešením soustavy rovnic

$$\frac{\partial\mathcal{S}}{\partial\beta_j}\Big|_{\beta_j=b_j}=0,\qquad j=0,\ldots,k, \tag{18.2.2}$$

tj. řešením soustavy rovnic

$$\begin{aligned} b_0S_{00}+b_1S_{01}+\ldots,+b_kS_{0k} &= S_{0y},\\ &\vdots\\ b_0S_{k0}+b_1S_{k1}+\ldots,+b_kS_{kk} &= S_{ky}, \end{aligned} \tag{18.2.3}$$

kde

$$S_{j_1j_2}=S_{j_2j_1}=\sum_{i=1}^{n}x_{ij_1}x_{ij_2},\qquad j_1,j_2=0,\ldots,k, \tag{18.2.4}$$

a

$$S_{jy}=\sum_{i=1}^{n}x_{ij}Y_i,\qquad j=0,\ldots,k. \tag{18.2.5}$$

Soustava (18.2.3) se nazývá *soustava normálních rovnic.* Lze ji též zapsat ve tvaru

$$\mathbf{X'Xb}=\mathbf{X'Y}, \tag{18.2.6}$$

kde

$$\mathbf{b} = (b_0, b_1, \ldots, b_k)'. \tag{18.2.7}$$

Podle předpokladu je hodnost matice $\mathbf{X}$ rovna $k+1$. Tudíž i matice $\mathbf{X}'\mathbf{X}$ má hodnost $k+1$ a inverzní matice $(\mathbf{X}'\mathbf{X})^{-1}$ existuje. Z (18.2.6) pak vyplývá, že

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{18.2.8}$$

Při aplikacích s konkrétními daty se odhady $b_j,\ j = 0, \ldots, k$, naleznou buď přímým řešením soustavy normálních rovnic (18.2.3), nebo (zejména při použití počítačů) podle vztahu (18.2.8).

18.3 Vlastnosti odhadů b_j.

Ukážeme, že statistiky b_j jsou:

- lineární odhady,
- nestranné odhady

parametrů $\beta_j,\ j = 0, \ldots, k$, a že kovarianční matice $\mathbf{\Sigma_b}$ vektoru (18.2.8) je rovna

$$\mathbf{\Sigma_b} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \tag{18.3.1}$$

Označme $t_{ji},\ j = 0, \ldots, k,\ i = 1, \ldots, n$, prvky matice

$$\mathbf{T}' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'. \tag{18.3.2}$$

Z (18.2.8) vyplývá, že

$$b_j = \sum_{i=1}^{n} t_{ji} Y_i, \qquad j = 0, \ldots, k, \tag{18.3.3}$$

takže statistika b_j je lineárním odhadem parametru $\beta_j,\ j = 0, \ldots, k$ (je lineární formou veličin $(Y_1, \ldots, Y_n)$. Přitom t_{ji} jsou známá čísla (neboť x_{ij} jsou známá čísla).

Protože střední hodnota vektoru $\mathbf{Y}$ je rovna $\boldsymbol{\eta}$, je střední hodnota vektoru $\mathbf{b}$ rovna

$$\mathbf{T}\boldsymbol{\eta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta},$$

takže platí

$$E(b_j) = \sum_{i=1}^{n} t_{ji}\eta_i = \beta_j, \qquad j = 0, \ldots, k, \tag{18.3.4}$$

tj. b_j je nestranný odhad parametru β_j.

Říkáme, že statistika b_j je *lineární nestranný odhad* parametru β_j, $j = 0, \ldots, k$.

Kovariance

$$\operatorname{cov}(b_{j_1}, b_{j_2}) = E\big((b_{j_1} - \beta_{j_1}(b_{j_2} - \beta_{j_2}\big) = \sum_{i=1}^{n}\sum_{l=1}^{n} t_{j_1 i}t_{j_2 l}E\big((Y_i - \eta_i)(Y_l - \eta_l)\big)$$

vzhledem k (18.3.3) a (18.3.4). Avšak [viz (18.1.1), (17.3.1) a (17.3.2)]

$$\begin{aligned} E\big((Y_i - \eta_i)(Y_l - \eta_l)\big) &= E(e_i e_l) = \sigma^2, \qquad i = l = 1, \ldots, n, \\ &= 0, \qquad\qquad\quad i, l = 1, \ldots, n, \quad i \neq l. \end{aligned}$$

Tudíž

$$\operatorname{cov}(b_{j_1}, b_{j_2}) = \sigma^2 \sum_{i=1}^{n} t_{j_1 i}t_{j_2 i}, \qquad j_1, j_2 = 0, \ldots, k; \tag{18.3.5}$$

speciálně pro $j_1 = j_2 = j$ je

$$\operatorname{var}(b_j) = \sigma^2 \sum_{i=1}^{n} t_{ji}^2, \qquad j = 0, \ldots, k. \tag{18.3.6}$$

Avšak (18.3.5) a (18.3.6) jsou prvky matice $\sigma^2\,\mathbf{T}\mathbf{T}' = \sigma^2\,(\mathbf{X}'\mathbf{X})^{-1}$ (neboť $\big((\mathbf{X}'\mathbf{X})^{-1}\big)' = (\mathbf{X}'\mathbf{X})^{-1}$, protože matice $\mathbf{X}'\mathbf{X}$ je symetrická). Platí tedy vztah (18.3.1).

Je vidět, že při řešení úloh regresní analýzy s konkrétními daty je výhodné vypočítat $\mathbf{b}$ pomocí vztahu (18.2.8), neboť v tomto vztahu se vyskytuje matice $(\mathbf{X}'\mathbf{X})^{-1}$, jejíž prvky násobené σ^2 představují rozptyly a kovariance odhadů $b_0, \ldots, b_k$.

18.4 Odhad parametrické funkce $\mathbf{c}'\boldsymbol{\beta}$.

Parametry β_j jsou speciální případy parametrické funkce

$$\gamma = \sum_{j=0}^{k} c_j\beta_j = \mathbf{c}'\boldsymbol{\beta}, \tag{18.4.1}$$

kde $\mathbf{c} = (c_0, \ldots, c_k)'$ je daný nenulový vektor. Jiným důležitým případem γ je hodnota regresní funkce $\boldsymbol{\eta} = \mathbf{x}'\boldsymbol{\beta}$ v daném bodě $\mathbf{x}' = (x_0, \ldots, x_k)$.

Uvažujme odhad

$$g = \sum_{j=0}^{n} c_j b_j = \mathbf{c}'\mathbf{b}, \tag{18.4.2}$$

který sestrojíme tak, že v (18.4.1) neznáme parametry $\beta_0, \ldots, \beta_k$ nahradíme jejich odhady $b_0, \ldots, b_k$ nalezenými metodou nejmenších čtverců.

Je ihned vidět, že (18.4.2) je lineární nestranný odhad γ, neboť g je lineární forma veličin $Y_1, \ldots, Y_n$ a

$$E(g) = \sum_{j=0}^{k} c_j E(b_j) = \sum_{j=0}^{k} c_j \beta_j = \gamma.$$

Pro rozptyl statistiky (18.4.2) platí (viz [24], vztah (11.7.10))

$$\operatorname{var}(g) = \sum_{j_1=0}^{k} \sum_{j_2=0}^{k} c_{j_1} c_{j_2} \operatorname{cov}(b_{j_1}, b_{j_2}) = \mathbf{c}'\boldsymbol{\Sigma}_{\mathbf{b}}\mathbf{c},$$

takže

$$\operatorname{var}(g) = \sigma^2 \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}. \tag{18.4.3}$$

Ukážeme, že ve třídě lineárních nestranných odhadů parametrické funkce (18.4.1) má statistika (18.4.2) nejmenší rozptyl neboli že statistika g je *nejlepší lineární nestranný odhad* parametrické funkce γ.

Uvažujme libovolný lineární odhad

$$G = \sum_{i=1}^{n} h_i Y_i = \mathbf{h}'\mathbf{Y}, \tag{18.4.4}$$

kde $h_1, \ldots, h_n$ jsou konstanty nezávislé na regresních parametrech $\beta_0, \ldots, beta_k$. Statistiku G můžeme vyjádřit ve tvaru

$$G = g + \sum_{i=1}^{n} d_i Y_i = \mathbf{c}'\mathbf{b} + \mathbf{d}'\mathbf{Y} = (\mathbf{c}'\mathbf{T} + \mathbf{d}')\mathbf{Y},$$

kde $\mathbf{d} = (d_1, \ldots, d_n)'$. Je tedy $\mathbf{h}' = \mathbf{c}'\mathbf{T} + \mathbf{d}'$.

Střední hodnota a rozptyl

$$E(G) = \sum_{i=1}^{n} h_i \eta_i = \mathbf{h}'\boldsymbol{\eta} = (\mathbf{c}'\mathbf{T} + \mathbf{d}')\mathbf{X}\boldsymbol{\beta} = (\mathbf{c}' + \mathbf{d}'\mathbf{X})\boldsymbol{\beta},$$

$$\begin{aligned}\mathrm{var}(G) &= \sigma^2 \sum_{i=1}^{n} h_i^2 = \sigma^2 \mathbf{h}'\mathbf{h} = \sigma^2(\mathbf{c}'\mathbf{T} + \mathbf{d}')(\mathbf{T}'\mathbf{c} + \mathbf{d}) \\ &= \sigma^2(\mathbf{c}'\mathbf{T}\mathbf{T}'\mathbf{c} + \mathbf{c}'\mathbf{T}\mathbf{d} + \mathbf{d}'T'\mathbf{c} + \mathbf{d}'\mathbf{d}).\end{aligned}$$

Aby G byl nestranný odhad γ, musí platit $\mathbf{d}'\mathbf{X} = \mathbf{0}' = (0, \ldots, 0)'$ (nulový vektor s $k+1$ prvky). Pak však $\mathbf{d}'\mathbf{T}'\mathbf{c} = \mathbf{d}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c} = 0$, rovněž $\mathbf{c}'\mathbf{T}\mathbf{d} = 0$, takže

$$\mathrm{var}(G) = \mathrm{var}(g) + \sigma^2 \sum_{i=1}^{n} d_i^2 \geq \mathrm{var}(g),$$

přičemž rovnost nastává jen pro případ $d_1 = \ldots d_n = 0$, tj. pro $G = g$.

Protože β_j je speciální případ $\gamma = \mathbf{c}'\boldsymbol{\beta}$ pro vektor $\mathbf{c}$, jehož j-tá složka je rovna 1 a všechny ostatní složky jsou nulové, vyplývá odtud, že statistika b_j je nejlepší lineární nestranný odhad parametru β_j, $j = 0, \ldots, k$. Za předpokladů uvedených v odst. 18.1 dává tedy metoda nejmenších čtverců odhady b_j mající tuto vlastnost.

18.5 Odhad parametru σ^2.

Z předchozího odstavce vyplývá, že nejlepším lineárním nestranným odhadem hodnoty $\eta_i = \sum_{j=0}^{k} \beta_j x_{ij}$ regresní funkce $\eta = \sum_{j=0}^{k} \beta_j x_j$ v bodě $\mathbf{x}_i'(x_{i0}, \ldots, x_{ik})$ je statistika

$$\hat{Y}_i = \sum_{j=0}^{k} b_j x_{ij}, \qquad i = 1, \ldots, n. \tag{18.5.1}$$

Označme $\hat{\mathbf{Y}} = (\hat{\mathbf{Y}}_\mathbf{1}, \ldots, \hat{\mathbf{Y}}_\mathbf{n})'$. Statistika

$$S_R = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = (\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}) = \mathbf{Y}'\mathbf{Y} - 2\hat{\mathbf{Y}}'\mathbf{Y} + \hat{\mathbf{Y}}'\hat{\mathbf{Y}}, \tag{18.5.2}$$

která se nazývá *reziduální součet čtverců,* charakterizuje variabilitu odchylek $Y_i - \hat{Y}_i$, $i = 1, \ldots, n$. Zdá se tedy rozumné využít pro odhad parametru σ^2

této statistiky. Soustavu (18.5.1) lze vyjádřit ve tvaru $\hat{\mathbf{Y}} = \mathbf{Xb}$. Odtud a z (18.2.6) vyplývá, že

$$\hat{\mathbf{Y}}'\mathbf{Y} = \mathbf{b}'\mathbf{X}'\mathbf{Y} = \mathbf{b}'\mathbf{X}'\mathbf{Xb} = \hat{\mathbf{Y}}'\hat{\mathbf{Y}},$$

takže S_R lze též vyjádřit ve tvaru

$$S_R = \mathbf{Y}'\mathbf{Y} - \hat{\mathbf{Y}}'\mathbf{Y} = \mathbf{Y}'\mathbf{Y} - \hat{\mathbf{Y}}'\hat{\mathbf{Y}}. \tag{18.5.3}$$

Střední hodnota této statistiky

$$\begin{aligned} E(S_R) &= \sum_{i=1}^{n} E(Y_i)^2 - \sum_{i=1}^{n} E(\hat{Y}_i^2) = \sum_{i=1}^{n}(\sigma^2 + \eta_i^2) - \sum_{i=1}^{n}\big(\mathrm{var}(\hat{Y}_i) + \eta_i^2\big) = \\ &= n\sigma^2 - \sum_{i=1}^{n}\mathrm{var}(\hat{Y}_i). \end{aligned}$$

Avšak

$$\begin{aligned} \mathrm{var}(\hat{Y}_i) &= E\big((\hat{Y}_i - \eta_i)^2\big) = E\Big(\big(\sum_{j=0}^{k}(b_j - \beta_j)x_{ij}\big)^2\Big) = \\ &= \sum_{j_1=0}^{k}\sum_{j_2=0}^{k} x_{ij_1}x_{ij_2}\mathrm{cov}(b_{j_1}, b_{j_2}), \qquad i = 1, \ldots, n. \end{aligned}$$

Označme matici $\mathbf{X}'\mathbf{X}$ jako $\mathbf{A} = (a_{j_1j_2}),\ j_1, j_2 = 0, 1, \ldots, k$. Dále označme $(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{A}^{-1} = (a^{j_1j_2},\ j_1, j_2 = 0, 1, \ldots, k$. Pak vzhledem k (18.3.1) je

$$\sum_{i=1}^{n}\mathrm{var}(\hat{Y}_i) = \sigma^2\sum_{j_1=0}^{k}\sum_{j_2=0}^{k} a_{j_1j_2}a^{j_1j_2} = (k+1)\sigma^2,$$

neboť $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}_{k+1}$, takže platí

$$\sum_{j_2=0}^{k} a_{j_1,j_2}a^{j_1j_2} = 1, \qquad j_1 = 0, 1, \ldots, k.$$

Je tedy

$$E(S_R) = (n - k - 1)\sigma^2, \tag{18.5.4}$$

takže nestranným odhadem parametru σ^2 je statistika

$$S^2 = \frac{1}{n-k-1}\, S_R. \tag{18.5.5}$$

Nestranné odhady rozptylů a kovariancí odhadů $b_0, \ldots, b_k$ dostaneme tak, že prvky matice $(\mathbf{X}'\mathbf{X})^{-1}$ násobíme statistikou (18.5.5).

Protože, jak již bylo uvedeno, platí $\hat{\mathbf{Y}}'\mathbf{Y} = \mathbf{b}'\mathbf{X}'\mathbf{Y} = \sum_{j=0}^{k} b_j S_{jy}$, kde S_{jy} jsou pravé strany normálních rovnic (18.2.3), lze též S_R vyjádřit ve tvaru

$$S_R = \sum_{i=1}^{n} Y_i^2 - \sum_{j=0}^{n} b_j S_{jy}. \tag{18.5.6}$$

18.6 Rozdělení statistik b a $g = \mathbf{c}'\mathbf{b}$.

K předpokladům odst. 18.1 přidejme ještě předpoklad normálního rozdělení veličin $e_1, \ldots, e_n$. Předpokládáme tedy, že vektor $\mathbf{e}$ má rozdělení $\mathrm{N}_n(\mathbf{0},\, \sigma^2\mathbf{I}_n)$ neboli, že vektor $\mathbf{Y}$ má rozdělení $\mathrm{N}_n(\mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n)$, přičemž matice $\mathbf{X}$ má hodnost $k+1$.

Protože odhady b_j jsou vzhledem k (18.2.8) lineární formy veličin $Y_1, \ldots, Y_n$, má vektor $\mathbf{b}$ za uvedených předpokladů (viz [24], odst. 24.7) rozdělení $\mathrm{N}_{k+1}\big(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\big)$.

Odtud dále vyplývá, že pro libovolný nenulový vektor $\mathbf{c} = (\mathbf{c_0}, \ldots, \mathbf{c_k})'$ má $g = \mathbf{c}'\mathbf{b}$ rozdělení $N(\gamma, \sigma_g^2)$, kde

$$\gamma = \mathbf{c}'\boldsymbol{\beta}, \qquad \sigma_g^2 = \mathbf{c}'\boldsymbol{\Sigma}_\mathbf{b}\mathbf{c}. \tag{18.6.1}$$

Tudíž statistika

$$U = \frac{g-\gamma}{\sigma_g} = \frac{\mathbf{c}'\mathbf{b} - \mathbf{c}'\boldsymbol{\beta}}{(\mathbf{c}'\boldsymbol{\Sigma}_\mathbf{b}\mathbf{c})^{1/2}} \tag{18.6.2}$$

má rozdělení N(0, 1).

18.7 Rozdělení veličiny S_R/σ^2.

Za předpokladu, že vektor $\mathbf{Y}$ má rozdělení $\mathrm{N}_n(\mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n)$ a že matice $\mathbf{X}$ má hodnost $k+1$, má veličina

$$\frac{1}{\sigma^2} S_R = \frac{(n-k-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \tag{18.7.1}$$

rozdělení $\chi^2(n-k-1)$. Důkaz viz [1], str. 100.

Vyjádřeme ještě S_R jako kvadratickou formu veličin $Y_1, \ldots, Y_n$. Uvažujme vektor

$$\mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\mathbf{b} = \big(\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\big)\mathbf{Y}. \tag{18.7.2}$$

Tento vektor má (viz [24], odst. 24.7) rozdělení $\mathrm{N}_n\Big(\mathbf{0},\, \sigma^2\big(\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\big)\Big)$, neboť

$$\big(\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\big)\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta} = \mathbf{0}$$

a

$$\big(\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\big)\sigma^2\mathbf{I}_n\big(\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\big) = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

Statistiku S_R lze vyjádřit ve tvaru

$$S_R = (\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}) = \mathbf{Y}'\big(\mathrm{l}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\big)\mathbf{Y}.$$

18.8 Nezávislost statistik b a S_R.

Za předpokladů uvedených v odst. 18.7 jsou statistiky **b** a S_R nezávislé.

Vzhledem k (18.2.8) je **b** lineární forma veličin $Y_1, \ldots, Y_n$ a vzhledem k (18.7.3) je S_R kvadratická forma veličin $Y_1, \ldots, Y_n$. Protože vektor **Y** má rozdělení $\mathrm{N}_n(\mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n)$ a protože platí

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}_n\big(\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\big) = \mathbf{0},$$

jsou (viz [1], str. 81, věta 16) statistiky **b** a S_R nezávislé.

18.9 Intervaly spolehlivosti a testy pro parametrické funkce $\gamma = \mathbf{c}'\boldsymbol{\beta}$.

Jak bylo uvedeno v předchozích odstavcích, za předpokladu, že vektor **Y** má rozdělení $\mathrm{N}_n(\mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n)$ a že matice **X** má hodnost $k+1$, jsou statistiky **b** a S_R nezávislé, **b** má rozdělení $\mathrm{N}_{k+1}\big(\boldsymbol{\beta},\, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\big)$ a veličina S_R/σ^2 má rozdělení $\chi^2(n-k-1)$.

Odtud a z odst. 18.6 vyplývá, že pro libovolnou parametrickou funkci $\gamma = \mathbf{c}'\boldsymbol{\beta}$, kde **c** je daný nenulový vektor, má veličina

$$T = \frac{g - \gamma}{S_g} = \frac{\mathbf{c}'\mathbf{b} - \mathbf{c}'\boldsymbol{\beta}}{S\big(\frac{1}{\sigma^2}\mathbf{c}'\boldsymbol{\Sigma}_{\mathbf{b}}\mathbf{c}\big)^{1/2}}, \tag{18.9.1}$$

kde S^2 je statistika (18.5.5), rozdělení t o $n-k-1$ stupních volnosti.

Veličina (18.9.1) vznikne úpravou veličiny

$$T = \frac{U}{\left(\frac{S_R}{\sigma^2}\,\frac{1}{n-k-1}\right)^{\frac{1}{2}}} = \frac{g-\gamma}{\sigma_g\left(\frac{S^2}{\sigma^2}\right)^{\frac{1}{2}}},$$

která má vzhledem k odst. 3.3 rozdělení $\mathrm{t}(n-k-1)$.

Veličiny (18.9.1) lze využít při konstrukci intervalu spolehlivosti pro γ; je vidět, že

$$\Big(g - t_{1-\alpha/2}(n-k-1)S_g;\ g + t_{1-\alpha/2}(n-k-1)S_g\Big) \tag{18.9.2}$$

je $100(1-\alpha)\%$ interval spolehlivosti pro γ.

Pro testování hypotézy $H : \gamma = \gamma_0$ (γ_0 dané číslo) proti oboustranné alternativě $A : \gamma \neq \gamma_0$ užijeme statistiky

$$T = \frac{g-\gamma_0}{S_g} \tag{18.9.3}$$

a H se zamítá na hladině významnosti α, jestliže $|T| \geq t_{1-\alpha/2}(n-k-1)$.

Tak např. pro testování hypotézy $H : \beta_j = 0$, tj. že model (18.1.1) neobsahuje vysvětlující proměnnou x_j, je

$$T = \frac{b_j}{S_{b_j}} = \frac{b_j}{S\sqrt{a^{jj}}}, \tag{18.9.4}$$

kde a^{jj} je prvek j-tého řádku a j-tého sloupce matice $\mathbf{A}^{-1} = (\mathbf{X}'\mathbf{X})^{-1}$.

18.10 Příklad.

Aplikujeme výsledky čl. 18 na regresní přímku $\eta = \beta_0 + \beta_1 x$ a porovnejme je s odpovídajícími vztahy čl. 17. Zřejmě

$$\mathbf{X} = \begin{pmatrix} 1, & x_1 \\ \vdots & \vdots \\ 1, & x_n \end{pmatrix}, \qquad \mathbf{X}'\mathbf{X} = \begin{pmatrix} n, & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i, & \sum_{i=1}^n x_i^2 \end{pmatrix}, \qquad \mathbf{X}'\mathbf{Y} = \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n x_i Y_i \end{pmatrix},$$

takže

$$\mathbf{b} = \frac{1}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \begin{pmatrix} \sum_{i=1} x_i^2, & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i, & n \end{pmatrix} \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n x_i Y_i \end{pmatrix},$$

a po výpočtu dostaneme pro b_0 a b_1 výrazy (17.2.5). Dále

$$\mathbf{\Sigma_b} = \frac{\sigma^2}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \begin{pmatrix} \sum_{i=1} x_i^2, & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i, & n \end{pmatrix}$$

takže $\mathrm{var}(b_0)$ je dána výrazem (17.3.4) a $\mathrm{var}(b_1)$ výrazem (17.3.5). Kovariance

$$\mathrm{cov}(b_0, b_1) = -\frac{\sigma^2 \sum_{i=1}^n x_i}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}.$$

Pro parametrickou funkci $\eta = \beta_0 + \beta_1 x$, tj. pro hodnotu regresní funkce v daném x, je

$$g = (1, x) \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = b_0 + b_1 x$$

a

$$\mathrm{var}(g) = \frac{\sigma^2}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} (1, x) \begin{pmatrix} \sum_{i=1} x_i^2, & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i, & n \end{pmatrix} \begin{pmatrix} 1 \\ x \end{pmatrix},$$

což po výpočtu vede k výrazu (11.3.7).

18.11 Úlohy.

18.11.1

Uvažujte regresní funkci $\eta = \beta_0$. Nalezněte $b_0,\ \mathrm{var}(b_0)$ a S^2.

$$\Big[b_0 = \overline{Y},\ \ \mathrm{var}(b_0) = \tfrac{\sigma^2}{n},\ \ S^2 = \tfrac{1}{n-1}\Big(\sum_{i=1}^n Y_i^2 - \tfrac{1}{n}(\sum_{i=1}^n Y_i)^2\Big).\Big]$$

18.11.2

Uvažujte regresní funkci $\eta = \beta_1 x$. Stanovte $b_1,\ \mathrm{var}(b_1)$ a S^2.

$$\Big[b_1 = \tfrac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2},\ \ \mathrm{var}(b_1) = \tfrac{\sigma^2}{\sum_{i=1}^n x_i^2},\ \ S^2 = \tfrac{1}{n-1}\Big|\sum_{i=1}^n Y_i^2 - \tfrac{(\sum_{i=1}^n x_i Y_i)^2}{\sum_{i=1}^n x_i^2}\Big|.\Big]$$

18.11.3

Čemu jsou pro model odst. 18.1 rovny výrazy

$$\sum_{i=1}^n (Y_i - \hat{Y}_i) x_{ij}, \qquad j = 0, \ldots, k?$$

Tyto výrazy mohou sloužit jako kontrola správnosti výpočtu odhadů $b_0, \ldots, b_k$ z normálních rovnic.

$\big[0;$ vyplývá to z (18.2.3).$\big]$

Kapitola 19

Lineární regrese s více vysvětlujícími proměnnými

19.1 Odhady regresních parametrů.

V článku 18 jsme uvažovali regresní model lineární v parametrech. Speciálním případem tohoto modelu je pro $k = 1$ a $x_0 = 1$ regresní přímka studovaná v čl. 17. Jejím zobecněním pro $k \geq 2$ je regresní funkce

$$\eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k. \tag{19.1.1}$$

Jestliže

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + e_i, \quad i = 1, \ldots, n > k + 1, \tag{19.1.2}$$

a platí-li předpoklady odst. 18.1, naleznou se odhady $b_0, \ldots, b_k$ řešením soustavy normálních rovnic

$$\begin{aligned}
b_0 n + b_1 \sum_{i=1}^{n} x_{i1} + b_2 \sum_{i=1}^{n} x_{i2} + \ldots + b_k \sum_{i=1}^{n} x_{ik} &= \sum_{i=1}^{n} Y_i, \\
b_0 \sum_{i=1}^{n} x_{i1} + b_1 \sum_{i=1}^{n} x_{i1}^2 + b_2 \sum_{i=1}^{n} x_{i1} x_{i2} + \ldots + b_k \sum_{i=1}^{n} x_{i1} x_{ik} &= \sum_{i=1}^{n} x_{i1} Y_i, \\
&\vdots \\
b_0 \sum_{i=1}^{n} x_{ik} + b_1 \sum_{i=1}^{n} x_{i1} x_{ik} + b_2 \sum_{i=1}^{n} x_{i2} x_{ik} + \ldots + b_k \sum_{i=1}^{n} x_{ik}^2 &= \sum_{i=1}^{n} x_{ik} Y_i.
\end{aligned} \tag{19.1.3}$$

Odečteme-li první rovnici násobenou výrazem $\sum_{i=1}^{n} x_{ij} = n\overline{x}_j$ od $(j+1)$-ní rovnice, $j = 1, \ldots, k$, dostáváme soustavu

$$\begin{aligned} b_1 \sum_{i=1}^{n} z_{i1}^2 + b_2 \sum_{i=1}^{n} z_{i1} z_{i2} + \ldots + b_k \sum_{i=1}^{n} z_{i1} z_{ik} &= \sum_{i=1}^{n} z_{i1} Y_i, \\ &\vdots \\ b_1 \sum_{i=1}^{n} z_{i1} z_{ik} + b_2 \sum_{i=1}^{n} z_{i2} z_{ik} + \ldots + b_k \sum_{i=1}^{n} z_{ik}^2 &= \sum_{i=1}^{n} z_{ik} Y_i, \end{aligned} \tag{19.1.4}$$

kde

$$z_{ij} = x_{ij} - \overline{x}_j = x_{ij} - \frac{1}{n} \sum_{i=1}^{n} x_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, k. \tag{19.1.5}$$

Tím snížíme počet rovnic soustavy z $k+1$ na k a dostaneme nižší numerické hodnoty součtů $|\sum_{i=1}^{n} z_{ij_1} z_{ij_2}|$, než byly hodnoty součtů $|\sum_{i=1}^{n} x_{ij_1} x_{ij_2}|$, $j_1, j_2 = 1, \ldots, k$.

Označme $\mathbf{Z}$ matici

$$\mathbf{Z} = \begin{pmatrix} z_{11}, & z_{12}, & \ldots, & z_{1k} \\ \ldots, & \ldots, & \ldots, & \ldots \\ z_{n1}, & z_{n2}, & \ldots, & z_{nk} \end{pmatrix}. \tag{19.1.6}$$

Pak se soustava (19.1.4) dá vyjádřit ve tvaru $\mathbf{Z}'\mathbf{Z}\mathbf{b}^* = \mathbf{Z}'\mathbf{Y}$, kde

$$\mathbf{b}^* = (b_1, \ldots, b_k)'. \tag{19.1.7}$$

Je-li aspoň $k+1$ z n bodů $\mathbf{x}_i' = (x_{i1}, \ldots, x_{ik})'$, $i = 1, \ldots, n$, různých, má matice $\mathbf{Z}$ hodnost k a

$$\mathbf{b}^* = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}. \tag{19.1.8}$$

Pro odhad b_0 vyplývá z první rovnice soustavy (19.1.3)

$$b_0 = \overline{Y} - b_1\overline{x}_1 - \ldots - b_k\overline{x}_k = \overline{Y} - \overline{\mathbf{x}}'\mathbf{b}^*, \tag{19.1.9}$$

kde $\overline{\mathbf{x}}' = (\overline{x}_1, \ldots, \overline{x}_k)$.

Protože tyto odhady $b_0, b_1, \ldots, b_k$ jsou ekvivalentní odhadům získaným řešením soustavy (19.1.3), jsou tato b_j vzhledem k odst. 18.3 a 18.4 nejlepší lineární nestranné odhady parametrů $\beta_0, \beta_1, \ldots, \beta_k$.

19.2 Rozptyly a kovariance odhadů b_j.

Ze vztahu (19.1.8) a z odst. 18.3 vyplývá, že kovarianční matice $\mathbf{\Sigma}_{\mathbf{b}^*}$ vektoru (19.1.7) je rovna

$$\mathbf{\Sigma}_{\mathbf{b}^*} = \sigma^2(\mathbf{Z}'\mathbf{Z})^{-1}. \tag{19.2.1}$$

Abychom mohli určit rozptyl statistiky $g = \sum_{j=0}^{k} c_j b_j$ pro libovolný nenulový vektor $\mathbf{c} = (c_0, c_1, \ldots, c_k)'$, potřebujeme ještě znát rozptyl $\mathrm{var}(\overline{Y})$ a $\mathrm{cov}(\overline{Y}, b_j),\ j = 1, \ldots, k$, neboť z (19.1.9) vyplývá, že

$$g = \sum_{j=0}^{k} c_j b_j = c_0\overline{Y} + \sum_{j=1}^{k}(c_j - c_0\overline{x}_j)b_j. \tag{19.2.2}$$

Rozptyl

$$\mathrm{var}(\overline{Y}) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{var}(Y_i) = \frac{\sigma^2}{n} \tag{19.2.3}$$

a kovariance

$$\mathrm{cov}(\overline{Y}, b_j) = E\Big(\big(b_0 + b_1\overline{x}_1 + \ldots + b_k\overline{x}_k - (\beta_0 + \beta_1\overline{x}_1 + \ldots + \beta_k\overline{x}_k)\big)(b_j - \beta_j)\Big) =$$

$$= E\Big(\big(b_0 - \beta_0 + \sum_{m=1}^{k}\overline{x}_m(b_m - \beta_m)\big)(b_j - \beta_j)\Big) = \mathrm{cov}(b_0, b_j) + \sum_{m=1}^{k}\overline{x}_m\mathrm{cov}(b_j, b_m).$$

Uvažujme nyní soustavu normálních rovnic (19.1.3) a označme pro tuto soustavu $\mathbf{X}'\mathbf{X} = \mathbf{A} = (a_{j_1 j_2})$ a $(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{A}^{-1} = (a^{j_1 j_2}),\ \ j_1 j_2 = 0, 1, \ldots, k$. Pak

$$\frac{1}{\sigma^2}\,\mathrm{cov}(\overline{Y}, b_j) = a^{0j} + \sum_{m=1}^{k}\overline{x}_m\, a^{jm} = a^{0j} + \frac{1}{n}\sum_{m=1}^{k} a_{0m}\, a^{jm}, \qquad j = 1, \ldots, k.$$

Protože matice $\mathbf{A}$ je symetrická a $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}_{k+1}$, platí

$$0 = \sum_{m=0}^{k} a_{0m}a^{jm} = a_{00}\, a^{j0} + \sum_{m=1}^{k} a_{0m}\, a^{jm} = na^{0j} + \sum_{m=1}^{k} a_{0m}\, a^{jm}, \qquad j = 1, \ldots, k.$$

Tudíž

$$\mathrm{cov}(\overline{Y}, b_j) = 0, \qquad j = 1, \ldots, k. \tag{19.2.4}$$

Pro rozptyl odhadu b_0 platí

$$\mathrm{var}(b_0) = \mathrm{var}(\overline{Y}) + \mathrm{var}(b_1\overline{x}_1 + \ldots + b_k\overline{x}_k) - 2\sum_{j=1}^{k} \overline{x}_j \,\mathrm{cov}(\overline{Y}, b_j) =$$

$$= \frac{\sigma^2}{n} + \sum_{j_1=1}^{k}\sum_{j_2=1}^{k} \overline{x}_{j_1}\,\overline{x}_{j_2}\,\mathrm{cov}(b_{j_1}, b_{j_2}).$$

Je tedy

$$\mathrm{var}(b_0) = \sigma^2\Big(\frac{1}{n} + \bar{\mathbf{x}}'(\mathbf{Z}'\mathbf{Z})^{-1}\bar{\mathbf{x}}\Big). \tag{19.2.5}$$

Pro dané $\mathbf{x}' = (x_1, \ldots, x_k)$ je statistika

$$\begin{aligned}\hat{Y} &= b_0 + b_1x_1 + \ldots + b_kx_k = \overline{Y} + b_1(x_1 - \overline{x}_1) + \ldots + b_k(x_k - \overline{x}_k) = \\ &= \overline{Y} + \mathbf{z}'\mathbf{b}, \end{aligned} \tag{19.2.6}$$

kde $\mathbf{z}' = (z_1, \ldots, z_k) = (x_1 - \overline{x}_1, \ldots, x_k - \overline{x}_k)$, nejlepší nestranný odhad hodnoty regresní funkce v bodě x. Její rozptyl

$$\mathrm{var}(\hat{Y}) = \sigma^2\Big(\frac{1}{n} + \mathbf{z}'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}\Big). \tag{19.2.7}$$

Rozptyl $\mathrm{var}(\hat{Y})$ se od rozptylu $\mathrm{var}(b_0)$ liší jen tím, že vektor $\overline{\mathbf{x}}$ se nahradí vektorem $\mathbf{z}$.

19.3 Odhad rozptylu σ^2.

Vzhledem k (18.5.6) a (19.1.3) lze reziduální součet čtverců S_R vyjádřit ve tvaru

$$\begin{aligned} S_R &= \sum_{i=1}^{n}(Y_i - b_0 - b_1x_{i1} - \ldots - b_kx_{ik})^2 = \\ &= \sum_{i=1}^{n} Y_i^2 - b_0\sum_{i=1}^{n} Y_i - b_1\sum_{i=1}^{n} x_{i1}Y_i - \ldots - b_k\sum_{i=1}^{n} x_{ik}Y_i. \end{aligned}$$

Použijeme-i výrazu (19.1.9) pro b_0, dostáváme

$$S_R = \sum_{i=1}^{n} Y_i^2 - \frac{1}{n}(\sum_{i=1}^{n} Y_i)^2 - b_1\sum_{i=1}^{n}(x_{i1} - \overline{x}_1)Y_i - \ldots - \sum_{i=1}^{n}(x_{ik} - \overline{x}_k)Y_i. \tag{19.3.1}$$

Dosadíme-li tento výraz do (18.5.5), dostaneme nestranný odhad σ^2.

19.4 Intervaly spolehlivosti a testy.

Nejlepším lineárním nestranným odhadem parametrické funkce $\gamma = \sum_{j=0}^{k} c_j\beta_j$ je statistika (19.2.2). Její rozptyl

$$\mathrm{var}(g) = \sigma^2\Big(\frac{c_0^2}{n} + \mathbf{d}'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{d}\Big), \tag{19.4.1}$$

kde $\mathbf{d} = (c_1 - c_0\overline{x}_1, \ldots, c_k - c_0\overline{x}_k)'$. Nestranným odhadem rozptylu $\mathrm{var}(g)$ je statistika

$$S_g^2 = \frac{S_R}{n-k-1}\frac{\mathrm{var}(g)}{\sigma^2}. \tag{19.4.2}$$

Mají-li veličiny $e_1, \ldots, e_n$ rozdělení $\mathrm{N}(0,\, \sigma^2)$, má statistika $T = \frac{g-\gamma}{S_g}$ rozdělení $\mathrm{t}(n-k-1)$. Pomocí této statistiky můžeme sestrojit intervaly spolehlivosti pro γ nebo testovat hypotézy o γ.

19.5 Příklady.

19.5.1

V případě regresní funkce

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{19.5.1}$$

se odhady b_1 a b_2 naleznou řešením soustavy rovnic

$$b_1 h_{11} + b_2 h_{12} = h_{1y},$$

$$b_1 h_{12} + b_2 h_{22} = h_{2y},$$

kde (viz (19.1.4) a (19.1.5))

$$\begin{aligned} h_{j_1 j_2} &= \sum_{i=1}^{n} z_{ij_1} z_{ij_2} = \sum_{i=1}^{n} x_{ij_1} x_{ij_2} - \frac{1}{n}(\sum_{i=1}^{n} x_{ij_1})(\sum_{i=1}^{n} x_{ij_2}), \qquad j_1, j_2 = 1, 2, \\ h_{j_y} &= \sum_{i=1}^{n} z_{ij} Y_i = \sum_{i=1}^{n} x_{ij} Y_i - \frac{1}{n}(\sum_{i=1}^{n} x_{ij})(\sum_{i=1}^{n} Y_i), \qquad j = 1, 2. \end{aligned} \tag{19.5.2}$$

Je tedy

$$\mathbf{H} = \begin{pmatrix} h_{11}, & h_{12} \\ h_{12}, & h_{22} \end{pmatrix} = \mathbf{Z}'\mathbf{Z}, \qquad \mathbf{Z}'\mathbf{Y} = \begin{pmatrix} h_{1y} \\ h_{2y} \end{pmatrix}$$

a z (19.1.8) vyplývá, že

$$b_1 = \frac{h_{22}h_{1y} - h_{12}h_{2y}}{h_{11}h_{22} - h_{12}^2}, \qquad b_2 = \frac{h_{11}h_{2y} - h_{12}h_{1y}}{h_{11}h_{22} - h_{12}^2}. \tag{19.5.3}$$

Kovarianční matice $\mathbf{\Sigma_b}$ vektoru $\mathbf{b}^* = (b_1, b_2)'$ je rovna

$$\mathbf{\Sigma_b} = \sigma^2\mathbf{H}^{-1} = \frac{\sigma^2}{h_{11}h_{22} - h_{12}^2}\begin{pmatrix} h_{22}, & -h_{12} \\ -h_{12}, & h_{11} \end{pmatrix}. \tag{19.5.4}$$

Statistika $b_0 = \overline{Y} - b_1\overline{x}_1 - b_2\overline{x}_2$ má vzhledem k (19.2.5) rozptyl

$$\mathrm{var}(b_0) = \sigma^2\Big(\frac{1}{n} + \bar{\mathbf{x}}'\mathbf{H}^{-1}\bar{\mathbf{x}}\Big) == \frac{\sigma^2}{n} + \frac{\sigma^2}{h_{11}h_{22} - h_{12}^2}(h_{22}\overline{x}_1^2 - 2h_{12}\overline{x}_1\overline{x}_2 + h_{11}\overline{x}_2^2) = \tag{19.5.5}$$

$$= \frac{\sigma^2}{n(h_{11}h_{22} - h_{12}^2)}\Big(\big(\sum_{i=1}^n x_{i1}^2\big)\big(\sum_{i=1}^n x_{i2}^2\big) - \big(\sum_{i=1}^n x_{i1}x_{i2}\big)^2\Big)$$

a statistika $\hat{Y} = b_0 + b_1x_1 + b_2x_2 = \overline{Y} + b_1(x_1 - \overline{x}_1) + b_2(x_2 - \overline{x}_2)$ má vzhledem k (19.2.7) rozptyl

$$\mathrm{var}(Y) = \sigma^2\Big(\frac{1}{n} + \mathbf{z}'\mathbf{H}^{-1}\mathbf{z}\Big) = \tag{19.5.6}$$

$$= \frac{\sigma^2}{n} + \frac{\sigma^2}{h_{11}h_{22} - h_{12}^2)}\big(h_{22}(x_1 - \overline{x}_1)^2 - 2h_{12}(x_1 - \overline{x}_1)(x_2 - \overline{x}_2) + h_{11}(x_2 - \overline{x}_2)^2\big).$$

Reziduální součet čtverců

$$S_R = \sum_{i=1}^n Y_i^2 - \frac{1}{n}\big(\sum_{i=1}^n Y_i\big)^2 - b_1h_{1y} - b_2h_{2y}. \tag{19.5.7}$$

19.5.2

Řešení normálních rovnic (19.1.4) se podstatně zjednoduší, je-li matice $\mathbf{Z}'\mathbf{Z} = \mathbf{H} = (h_{j_1j_2})$, $j_1, j_2 = 1, \ldots, k$, diagonální. V tomto případě

$$b_j = \frac{h_{jy}}{h_{jj}} = \frac{\sum_{i=1}^n z_{ij}Y_i}{\sum_{i=1}^n z_{ij}^2} = \frac{\sum_{i=1}^n (x_{ij} - \overline{x}_j)Y_i}{\sum_{i=1}^n (x_{ij} - \overline{x}_j)^2}, \qquad j = 1, \ldots, k, \tag{19.5.8}$$

a dále

$$\mathbf{\Sigma_b} = \sigma^2\begin{pmatrix} h_{11}^{-1}, & 0, & \ldots, & 0 \\ 0, & h_{22}^{-1}, & \ldots, & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0, & 0, & \ldots, & h_{kk}^{-1} \end{pmatrix} \tag{19.5.9}$$

a

$$S_R = \sum_{i=1}^{n} Y_i^2 - \frac{1}{n}\big(\sum_{i=1}^{n} Y_i\big)^2 - h_{11}b_1^2 - \ldots - h_{kk}b_k^2. \tag{19.5.10}$$

Této situace výhodné z hlediska numerického řešení regresní úlohy můžeme dosáhnout vhodnou volbou vysvětlujících proměnných $x_1, \ldots, x_k$.

Tak např., jestliže pro $k = 2$ uvažujeme hodnoty $-a_1, 0, a_1$ proměnné x_1 a hodnoty $-a_2, 0, a_2$ proměnné x_2 (a_1 a a_2 jsou daná kladná čísla) a jestliže n bodů (x_{i1}, x_{i2}) tvoří devět kombinací těchto hodnot:

$$(-a_1, -a_2), \quad (-a_1, 0), \ \ldots, \ (a_1, 0), \quad (a_1, a_2),$$

pak

$$\sum_{i=1}^{9} x_{i1} = \sum_{i=1}^{9} x_{i2} = \sum_{i=1}^{9} x_{i1}x_{i2} = 0$$

a

$$\mathbf{H} = \begin{pmatrix} 6a_1^2, & 0 \\ 0, & 6a_2^2 \end{pmatrix}$$

19.5.3

Obdobně jako v odst. 17.7 můžeme uvažovat regresní funkce, které lze vhodnou transformací převést na regresní funkce typu (19.1.1). Nechť

$$V_i = \zeta_i u_i = \delta f_{i1}^{\beta_1} f_{i2}^{\beta_2} \ldots f_{ik}^{\beta_k} u_i, \qquad i = 1, \ldots, n, \tag{19.5.11}$$

kde V_i a u_i jsou náhodné veličiny nabývající kladných hodnot, $\delta > 0$ a $\beta_1, \ldots, \beta_k$ jsou neznámé parametry. Dále $f_j = f_j(t_1, \ldots, t_r)$ jsou známé funkce $r \geq 1$ nenáhodných proměnných $t_1, \ldots, t_r$; přitom $f_{ij} = f_j(t_{i1}, \ldots, t_{ir}) > 0$ pro všechna $i = 1, \ldots, n, \ j = 1, \ldots, k$. Příklady ζ_i jsou funkce

$$\zeta_i = \delta t_{i1}^{\beta_1} t_{i2}^{\beta_2}, \qquad i = 1, \ldots, n.$$

tj. $f_1 = t_1, \ f_2 = t_2$, nebo

$$\zeta_i = \delta t_{i1}^{\beta_1} \exp(\beta_2 t_{i2} t_{i3}), \qquad i = 1, \ldots, n,$$

tj. $f_1 = t_1, \ f_2 = \exp(t_2\, t_3)$, apod.

Označíme-li

$$\begin{aligned} Y_i &= \ln V_i, \quad e_i = \ln u_i, \quad i = 1, \ldots, n, \\ x_{ij} &= \ln f_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, k, \\ \beta_0 &= \ln \delta, \end{aligned} \tag{19.5.12}$$

převedeme (19.5.11) na (19.1.2) (můžeme uvažovat i dekadické logaritmy). Platí-li pro $e_i = \ln u_i$ předpoklady (17.3.1) a (17.3.2), je nejlepším lineárním nestranným odhadem parametrické funkce $\gamma = \sum_{j=0} c_j \beta_j$ statistika (19.2.2) mající rozptyl (19.4.1); v obou těchto výrazech se za Y_i a x_{ij} dosadí výrazy (19.5.12).

Mají-li veličiny $e_1, \ldots, e_n$ rozdělení $\mathrm{N}(0, \sigma^2)$ (tj. mají-li veličiny $u_1, \ldots, u_n$ rozdělení $LN(0, \sigma^2)$), mají veličiny $\frac{g-\gamma}{S_g}$ rozdělení $\mathrm{t}(n-k-1)$.

19.5.4

Závislost doby do lomu V na teplotě t_1 a napětí t_2 při hodnocení zkoušek pevnosti při tečení lze vyjádřit vztahem

$$V = \zeta u = \delta exp\left(\beta_1 \frac{1}{t_1} + \beta_2 \frac{1}{t_1} \ln t_2\right) u. \tag{19.5.13}$$

Tabulka 19.1 udává hodnoty proměnných $t_1\,[K]$, $t_2\,[MPA]$ a $V\,[h]$ pro $n = 7$ zkoušek.

t_{i1}	t_{i2}	v_i
832	390	124
823	360	575
823	330	1 454
823	300	4 197
873	310	187
873	280	513
873	260	1298

Tab. 19.1: Hodnoty teploty (t_1), napětí (t_2) a doby do lomu (V).

Logaritmujeme-li (19.5.13) a označíme

$$Y_i = \ln V_i, \quad x_{i1} = t_{i1}^{-1}, \quad x_{i2} = t_{i1}^{-1} \ln t_{i2}, \quad e_i = \ln u_i, \quad i = 1, \ldots, n,$$

a dále $\beta_0 = \ln \delta$, můžeme pro určení b_0, b_1, b_2 a S_R použít vztahů příkl. 19.5.1.

Po výpočtech dostaneme

$$\sum_{i=1}^{7} x_{i1} \doteq 8,296\,693 \cdot 10^{-3}; \qquad \sum_{i=1}^{7} x_{i2} \doteq 4,777\,326 \cdot 10^{-2};$$
$$\sum_{i=1}^{7} x_{i1}^2 \doteq 9,841\,891 \cdot 10^{-6}; \qquad \sum_{i=1}^{7} x_{i1}x_{i2} \doteq 5,669\,798 \cdot 10^{-5};$$
$$\sum_{i=1}^{7} x_{i2}^2 \doteq 3,267\,970 \cdot 10^{-4}; \qquad \sum_{i=1}^{7} y_i \doteq 45,438\,815;$$
$$\sum_{i=1}^{7} x_{i1}y_i \doteq 0,053\,914; \qquad \sum_{i=1}^{7} x_{i2}y_i \doteq 0,309\,830;$$

takže

$$h_{11} \doteq 0,830\,319 \cdot 10^{-8}; \quad h_{12} \doteq 7,511\,260 \cdot 10^{-8}; \quad h_{22} \doteq 75,637\,557 \cdot 10^{-8};$$
$$h_{1y} \doteq 0,580\,2 \cdot 10^{-4}; \quad h_{2y} \doteq -2,786\,2 \cdot 10^{-4}.$$

Z (19.5.3) pak vyplývá, že

$$b_1 \doteq 10,152 \cdot 10^4, \quad b_2 \doteq -1,045 \cdot 10^4.$$

Dále

$$b_0 \doteq -42,516.$$

a protože $\sum_{i=1}^{7} y_i^2 \doteq 303,926\,861$, dostaneme z (19.5.7)

$$s^2 \doteq \frac{1}{4} \cdot 0,170 \doteq 0,042.$$

19.6 Úlohy.

19.6.1

Pro regresní funkci $\eta = \beta_6 + \beta_1 x_1 + \beta_2 x_2$ uvažujte $n = 5$ bodů $\mathbf{x}' = (x_{i1}, x_{i2}) : (-1;-1),\ (-1;1),\ (0;0),\ (1;-1),\ (1;1)$. Napište výrazy pro odhady $b_0,\ b_1,\ b_2$. Dále stanovte rozptyl $\mathrm{var}(\hat{Y})$ obecně a pro $\mathbf{x}' = (0, \frac{1}{2})$.

$$\left[\begin{array}{l} b_0 = \frac{1}{5}(Y_1 + Y_2 + Y_3 + Y_4 + Y_5), \quad b_1 = \frac{1}{4}(-Y_1 - Y_2 + Y_4 + Y_5), \\ b_2 = \frac{1}{4}(-Y_1 + Y_2 - Y_4 + Y_5); \quad \mathrm{var}(\hat{Y}) = \sigma^2\big(\frac{1}{5}\frac{1}{4}x_1^2 + \frac{1}{4}x_2^2\big); \quad \frac{21}{80}\sigma^2 \end{array}\right].$$

19.6.2

Tabulka 19.2 uvádí hodnoty proměnných $x_1,\ x_2,\ x_3,\ Y$ pro $n = 8$. Uvažujte regresní model (19.1.2), stanovte odhady $b_j,\ j = 0, 1, 2, 3$, a nalezněte 95% interval spolehlivosti pro parametr β_3.

x_{i1}	x_{i2}	x_{i3}	y_i
1	1	1	11
-1	2	0	7
0	3	2	10
4	-2	-3	11
2	1	0	10
4	3	-1	9
1	-1	1	15
2	0	2	16

Tab. 19.2: Hodnoty proměnných $x_1,\ x_2,\ x_3,\ Y$.

$$\Big[b_0 \doteq 10,181\,6;\ b_1 \doteq 0,979\,9;\ b_2 \doteq -1,268\,0;\ b_3 \doteq 1,842\,4;\ (1,212; 2,473).\Big]$$

Kapitola 20

Polynomická regrese

20.1 Regresní polynom.

Uvažujme regresní polynom stupně k

$$\eta = \beta_0 + \beta_1 x + \ldots + \beta_k x^k \tag{20.1.1}$$

ve vysvětlující proměnné x.

Nechť

$$Y_i = \beta_0 + \beta_1 x_i + \ldots + \beta_k x_i^k + e_i, \quad i = 1, \ldots, n > k+1, \tag{20.1.2}$$

a nechť platí předpoklady odst. 18.1 o veličinách $e_1, \ldots, e_n$. Nechť $x_1, \ldots, x_n$ jsou daná čísla a matice

$$\mathbf{X} = \begin{pmatrix} 1, & x_1 & \ldots, & x_1^k \\ 1, & x_2 & \ldots, & x_2^k \\ \vdots & \vdots & \ddots & \vdots \\ 1, & x_n & \ldots, & x_n^k \end{pmatrix} \tag{20.1.3}$$

má hodnost $k+1$.

Odhady $b_0, \ldots, b_k$ se naleznou řešením soustavy normálních rovnic

$$\begin{aligned} b_0 n + b_1 \sum_{i=1}^{n} x_i + \ldots + b_k \sum_{i=1}^{n} x_i^k &= \sum_{i=1}^{n} Y_i, \qquad (20.1.4) \\ b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 + \ldots + b_k \sum_{i=1}^{n} x_i^{k+1} &= \sum_{i=1}^{n} x_i Y_i, \\ &\vdots \\ b_0 \sum_{i=1}^{n} x_i^k + b_1 \sum_{i=1}^{n} x_i^{k+1} + \ldots + b_k \sum_{i=1}^{n} x_i^{2k} &= \sum_{i=1}^{n} x_i^k Y_i. \end{aligned}$$

Položíme-li

$$x_{ij} = x_i^j, \quad j = 1, \ldots, n, \quad j = 1, \ldots, k, \qquad (20.1.5)$$

a porovnáme-li soustavy rovnic (19.1.3) a (20.1.4), vidíme, že se jedná o stejné soustavy.

Lze tedy k určení odhadů $b_0, \ldots, b_k$, jejich rozptylů a odhadů těchto rozptylů a rovněž při konstrukci intervalů spolehlivosti pro parametrické funkce $\gamma = c'\beta$ či pro testy hypotéz $H : \gamma = \gamma_0$ použít postupů a výsledků čl. 19, kde se všude výrazy x_{ij} nahradí výrazy $x_i^j,\ i = 1, \ldots, n,\ j = 1, \ldots, k$.

20.2 Ekvidistantní hodnoty proměnné x.

Uvažujme případ, kdy hodnoty $x_1, \ldots, x_n$ jsou takové, že platí

$$\sum_{i=1}^{n} x_i^j = 0, \quad j = 1, 3, \ldots, 2k-1. \qquad (20.2.1)$$

Odhady $b_1, \ldots, b_k$ se naleznou řešením soustavy rovnic (19.1.4), kde platí

$$\begin{aligned} z_{ij_1} z_{ij_2} &= \sum_{i=1}^{n} x_i^{j_1+j_2} - \frac{1}{n}\Big(\sum_{i=1}^{n} x_i^{j_1}\Big)\Big(\sum_{i=1}^{n} x_i^{j_2}\Big) = h_{j_1 j_2}, \qquad (20.2.2) \\ &= 0, \qquad \text{pro všechna } j_1 + j_2 \text{ lichá}, \\ &= \sum_{i=1}^{n} x_i^{j_1+j_2} \qquad \text{pro } j_1 \text{ lichá nebo } j_2 \text{ lichá}, \end{aligned}$$

pro všechna $j_1, j_2 = 1, \ldots, k$, a dále

$$\begin{aligned} z_{ij}Y_i &= \sum_{i=1}^{n} x_i^j Y_i - \frac{1}{n}\Big(\sum_{i=1}^{n} x_i^j\Big)\Big(\sum_{i=1}^{n} Y_i\Big) = h_{jy}, \qquad (20.2.3)\\ &= \sum_{i=1}^{n} x_i^j Y_i \quad \text{pro } j \text{ lichá}, \end{aligned}$$

pro všechna $j = 1, \ldots, k$. Pro jiné kombinace j_1, j_2 není zjednodušení možné.

Soustava (19.1.4) se rozpadne na dvě podsoustavy, z nichž první obsahuje odhady b_j jen pro lichá j a druhá obsahuje b_j jen pro sudá j. Řešením těchto podsoustav nalezneme $b_2, b_4 \ldots$ a $b_1, b_3 \ldots$.

Z (19.1.9) vyplývá, že pro odhad b_0 platí

$$b_0 = \frac{1}{n}\Big(\sum_{i=1}^{n} Y_i - b_2\sum_{i=1}^{n} x_i^2 - b_4\sum_{i=1}^{n} x_i^4 - \ldots - b_q\sum_{i=1}^{n} x_i^q\Big), \qquad (20.2.4)$$

kde $q = k$ pro k sudé a $q = k-1$ pro k liché.

Tohoto postupu lze využít zejména v případě regresní funkce

$$\eta = \gamma_0 + \gamma_1 t + \ldots + \gamma_k t^k, \qquad (20.2.5)$$

jestliže hodnoty $t_1, t_2, \ldots, t_n$ vysvětlující proměnné t jsou takové, že

$$t_i - t_{i-1} = \Delta, \quad i = 2, \ldots, n,$$

kde *Delta* je dané kladné číslo. Položme

$$x_i = \frac{1}{\Delta}\left(t_i - \frac{1}{n}\sum_{i=1}^{n} t_i\right) = \frac{t_i - t_1}{\Delta} - \frac{n-1}{2}, \quad i = 1, \ldots, n; \qquad (20.2.6)$$

pak pro tato x_i platí vztahy (20.2.1).

Dosadíme-li $t = \Delta x + \bar{t}$ v (20.2.5), je

$$\begin{aligned} \eta &= \gamma_0 + \gamma_1(\Delta x + \bar{t})^2 + \gamma^2(\Delta x + \bar{t})^2 + \ldots + \gamma_k(\Delta x + \bar{t})^k = \qquad (20.2.7)\\ &= \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^k. \end{aligned}$$

Zjednodušeným výpočtem popsaným v tomto odstavci pak nalezneme odhady $b_0, \ldots, b_k$ a protože parametry $\gamma_0, \ldots, \gamma_k$ se dají vyjádřit jako lineární funkce parametrů β_0, β_k, určí se nejlepší nestranné odhady $g_0, \ldots, g_k$ parametrů $\gamma_0, \ldots, \gamma_k$ podle postupu odst. 19.2.

20.3 Příklad.

Demonstrujme postup odst. 20.2 na případě $k = 3$ a $n = 5$ hodnotách proměnné $t : 2,\ 5,\ 8,\ 11,\ 14$. Je tedy $\Delta = 3, \bar{t} = 8,\ x_i = -2,\ -1,\ 0,\ 1,\ 2$ a

$$\eta = \gamma_0 + \gamma_1(3x+8) + \gamma_2(3x+8)^2 + \gamma_3(3x+8)^3 = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3.$$

Protože pro tato x_i platí vztahy (20.3.1), rozpadne se soustava (19.1.4) na dvě podsoustavy

$$b_1 h_{11} + b_3 h_{13} = h_{1y},$$
$$b_1 h_{13} + b_3 h_{33} = h_{3y}$$

a

$$b_2 h_{22} = h_{2y},$$

z nichž obdržíme

$$b_1 = \frac{h_{33}h_{1y} - h_{13}h_{3y}}{h_{11}h_{33} - h_{13}^2} = \frac{\big(\sum_{i=1}^n x_i^6\big)\big(\sum_{i=1}^n x_i Y_i\big) - \big(\sum_{i=1}^n x_i^4\big)\big(\sum_{i=1}^n x_i^3 Y_i\big)}{\big(\sum_{i=1}^n x_i^2\big)\big(\sum_{i=1}^n x_i^6\big) - \big(\sum_{i=1}^n x_i^4\big)^2}, \tag{20.3.1}$$

$$b_3 = \frac{h_{11}h_{3y} - h_{13}h_{1y}}{h_{11}h_{33} - h_{13}^2} = \frac{\big(\sum_{i=1}^n x_i^2\big)\big(\sum_{i=1}^n x_i^3 Y_i\big) - \big(\sum_{i=1}^n x_i^4\big)\big(\sum_{i=1}^n x_i Y_i\big)}{\big(\sum_{i=1}^n x_i^2\big)\big(\sum_{i=1}^n x_i^6\big) - \big(\sum_{i=1}^n x_i^4\big)^2}, \tag{20.3.2}$$

$$b_2 = \frac{h_{2y}}{h_{22}} = \frac{n\sum_{i=1}^n x_i^2 Y_i - \big(\sum_{i=1}^n x_i^2\big)\big(\sum_{i=1}^n Y_i\big)}{n\sum_{i=1}^n x_i^4 - \big(\sum_{i=1}^n x_i^2\big)^2}. \tag{20.3.3}$$

Dále

$$b_0 = \frac{1}{n}\Big(\sum_{i=1}^n Y_i - b_2 \sum_{i=1}^n x_i^2\Big). \tag{20.3.4}$$

V našem případě je $\sum_{i=1}^5 x_i^2 = 10, \sum_{i=1}^5 x_i^4 = 34, \sum_{i=1}^5 x_i^6 = 130$ a po dosazení do výrazů (20.3.1) až (20.3.4) dostáváme

$$b_1 = \frac{1}{12}(Y_1 - 8Y_2 + 8Y_4 - Y_5), \tag{20.3.5}$$
$$b_2 = \frac{1}{14}(2Y_1 - Y_2 - 2Y_3 - Y_4 + 2Y_5),$$
$$b_3 = \frac{1}{12}(-Y_1 + 2Y_2 - 2Y_4 + Y_5),$$
$$b_0 = \frac{1}{5}\sum_{i=1}^5 Y_i - 2b_2 = \frac{1}{35}(-3Y_1 + 12Y_2 + 17Y_3 + 12Y_4 - 3Y_5).$$

Dosazením experimentálních hodnot $Y_1, \ldots, Y_5$ dostaneme hodnoty odhadů $b_0,\ b_1,\ b_2,\ b_3$.

Protože

$$\gamma_3 = \frac{1}{27}\beta_3, \quad \gamma_2 = \frac{1}{29}(\beta_2 - 8\beta_3), \quad \gamma_1 = \frac{1}{9}(3\beta_1 - 16\beta_2 + 64\beta_3), \quad (20.3.6)$$

$$\gamma_0 = \frac{1}{27}(27\beta_0 - 72\beta_1 + 192\beta_2 - 512\beta_3),$$

jsou nejlepšími nestrannými lineárními odhady parametrů $\gamma_0, \ldots, \gamma_3$ statistiky $g_0, \ldots, g_3$, které dostaneme tak, že v (20.3.6) nahradíme $\beta_0, \ldots, \beta_3$ jejich odhady $b_0, \ldots, b_3$ danými výrazy (20.3.5).

20.4 Úloha.

Uvažujte polynom $\eta = \beta_0 + \beta_1 x + \ldots + \beta_4 x^4$ a hodnoty $x_1, \ldots, x_n$ takové, že je splněna podmínka (20.2.1) pro $j = 1,\ 3,\ 5,\ 7$. Stanovte odhady $b_0, b_1, \ldots, b_4$.

$\left[\begin{array}{l} b_1 \text{ je dáno výrazem (20.3.1) a } b_3 \text{ výrazem (20.3.2)}; \quad b_2 = \frac{h_{44}h_{2y} - h_{24}h_{4y}}{h_{22}h_{44} - h_{24}^2}, \\ b_4 = \frac{h_{22}h_{4y} - h_{24}h_{2y}}{h_{22}h_{44} - h_{24}^2}, \ \text{ kde } h_{j_1 j_2} \text{ a } h_{jy} \text{ jsou výrazy (20.2.2) a (20.2.3)}, \\ b_0 = \frac{1}{n}\sum_{i=1}^n Y_i - b_2 \sum_{i=1}^n x_i^2 - b_4 \sum_{i=1}^n x_i^4 \end{array}\right]$

Kapitola 21

Aplikace metody nejmenších čtverců na uspořádané výběry

21.1 Zobecněná metoda nejmenších čtverců.

Dosud jsme uvažovali model (18.1.1) s pevnými proměnnými $x_0, x_1, \dots x_k$, v němž náhodné chyby $e_1, \dots, e_n$ mají nulové střední hodnoty, týž rozptyl $\sigma^2 > 0$, jsou nekorelované a matice (18.1.5) má hodnost $k+1$.

Zobecněme nyní tento model tak, že pozměníme pouze předpoklady o kovarianční matici vektoru $\mathbf{e} = (e_1, \dots, e_n)'$. Budeme předpokládat, že

$$\operatorname{cov}(e_{i1}, e_{i2}) = E(e_{i_1}, e_{i_2}) = w_{i_1 i_2}\sigma^2, \qquad i_1, i_2 = 1, \dots, n, \tag{21.1.1}$$

kde $\sigma^2 > 0$ je opět neznámý parametr a $w_{i_1 i_2}$ jsou známí čísla taková, že symetrická matice

$$\mathbf{W} = \begin{pmatrix} w_{11}, & w_{12}, & \dots, & w_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1}, & w_{n2}, & \dots, & w_{nn} \end{pmatrix} \tag{21.1.2}$$

je pozitivně definitní.

Pak vektor $\mathbf{Y} = (Y_1, \dots, Y_n)'$, jehož prvky jsou veličiny (18.1.1), má střední hodnotu $\boldsymbol{\eta}$ a kovarianční matici $\sigma^2\mathbf{W}$.

Odhady $b_0, b_1, \dots, b_k$ parametrů $\beta_0, \beta_1, \dots, \beta_k$ zobecněnou metodou nejmenších čtverců jsou statistiky, které minimalizují kvadratickou formu

$$(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{W}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}), \tag{21.1.3}$$

kde $\mathbf{W}^{-1}$ je inverzní matice k matici (21.1.2).

Lze ukázat (viz [1], str. 133 nebo [16], str. 426), že v tomto případě se dají normální rovnice zapsat ve tvaru

$$\mathbf{X}'\mathbf{W}^{-1}\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{W}^{-1}\mathbf{Y}, \tag{21.1.4}$$

že pro vektor $\mathbf{b} = (b_0, \ldots, b_k)'$ platí

$$\mathbf{b} = (\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{-1}\mathbf{Y} \tag{21.1.5}$$

a že kovarianční matice vektoru $\mathbf{b}$ je rovna

$$\mathbf{\Sigma}_{\mathbf{b}} = \sigma^2(\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})^{-1}. \tag{21.1.6}$$

Odhady b_j jsou opět nejlepší nestranné lineární odhady parametrů $\beta_j, j = 0, 1, \ldots, k$.

21.2 Nejlepší lineární nestranné odhady z pořádkových statistik.

Uvažujme náhodnou veličinu Z, jejíž rozdělení závisí na dvou neznámých parametrech μ a σ. Přitom nechť náhodná veličina $U = (Z-\mu)/\sigma$ má rozdělení, které nezávisí na žádném parametru.

Např. má-li veličina Z normální rozdělení N(μ, σ^2), má veličina $U = (Z-\mu)/\sigma$ rozdělení N(0, 1). Obdobně, má-li veličina Z exponenciální rozdělení E(μ, σ), má (viz [24], odst. 20.1) veličina $U = (Z-\mu)/\sigma$ rozdělení E(0, 1). Rozdělení N(0, 1), stejně jako rozdělení E(0, 1), nezávisí na žádném parametru.

Uvažujme uspořádaný výběr $(U_{(1)}, \ldots, U_{(n)})'$ z rozdělení náhodné veličiny U a označme

$$a_i = E(U_{(i)}, \quad i = 1, \ldots, n, \tag{21.2.1}$$

a

$$w_{i_1 i_2} = \mathrm{cov}(U_{(i_1)}, U_{(i_2)}), \quad i_1, i_2 = 1, \ldots, n; \tag{21.2.2}$$

pro výběry ze spojitých rozdělení se střední hodnotou $E(U_{(i)})$ určí podle vztahu (4.3.7) a kovariance $\mathrm{cov}(U_{(i_1)}, U_{(i_2)})$ podle vztahu (4.5.6).

Pro uspořádaný výběr $(Z_{(1)}, \ldots, Z_{(n)})'$ z rozdělení náhodné veličiny $Z = \mu + \sigma U$ pak platí

$$E(Z_{(i)}) = \mu + a_i\sigma, \quad i = 1, \ldots, n, \tag{21.2.3}$$

a

$$\operatorname{cov}(Z_{(i_1)}, Z_{i_2}) = w_{i_1 i_2}\sigma^2, \quad i_1, i_2 = 1, \ldots, n. \tag{21.2.4}$$

Označíme-li $Y_i = Z_{(i)}, i = 1, \ldots, n$, můžeme (21.2.3) přepsat ve tvaru

$$Y_i = \mu + a_i\sigma + e_i, \quad i = 1, \ldots, n; \tag{21.2.5}$$

to je však regresní model s dvěma regresními parametry $\beta_0 = \mu,\ \beta_1 = \sigma$, s maticí

$$\mathbf{X} = \begin{pmatrix} 1 & a_1 \\ \vdots & \vdots \\ 1, & a_n \end{pmatrix} \tag{21.2.6}$$

a s náhodnými chybami e_i splňujícími předpoklady odst. 21.1.

Označme μ^* a σ^* odhady parametrů μ a σ získané metodou nejmenších čtverců. Zapíšeme-li matici (21.2.6) jako

$$\mathbf{X} = (\mathbf{1}, \mathbf{a}),$$

kde

$$\mathbf{1} = (1, \ldots, 1)'; \quad \mathbf{a} = (a_1, \ldots, a_n)', \tag{21.2.7}$$

vyplývá z (21.1.4) soustavy normálních rovnic

$$\begin{pmatrix} \mathbf{1}' \\ \mathbf{a}' \end{pmatrix} \mathbf{W}^{-1}(\mathbf{1}, \mathbf{a}) \begin{pmatrix} \mu^* \\ \sigma^* \end{pmatrix} = \begin{pmatrix} \mathbf{1}' \\ \mathbf{a}' \end{pmatrix} \mathbf{W}^{-1}\mathbf{Y},$$

tj. soustava

$$\begin{aligned} \mathbf{1}'\mathbf{W}^{-1}\mathbf{1}\mu^* + \mathbf{1}'\mathbf{W}^{-1}\mathbf{a}\sigma^* &= \mathbf{1}'\mathbf{W}^{-1}\mathbf{Y}, \\ \mathbf{a}'\mathbf{W}^{-1}\mathbf{1}\mu^* + \mathbf{a}'\mathbf{W}^{-1}\mathbf{a}\sigma^* &= \mathbf{a}'\mathbf{W}^{-1}\mathbf{Y}. \end{aligned} \tag{21.2.8}$$

Označíme-li

$$\mathbf{\Gamma} = \frac{\mathbf{W}^{-1}(\mathbf{1}\mathbf{a}' - \mathbf{a}\mathbf{1}')\mathbf{W}^{-1}}{\Delta}, \tag{21.2.9}$$

kde Δ je determinant soustavy tj.

$$\Delta = |\mathbf{X}'\mathbf{W}^{-1}\mathbf{X}| = (\mathbf{1}'\mathbf{W}^{-1}\mathbf{1})(\mathbf{a}'\mathbf{W}^{-1}\mathbf{a}) - (\mathbf{1}'\mathbf{W}^{-1}\mathbf{a})^2, \tag{21.2.10}$$

dostáváme

$$\mu^* = -\mathbf{a}'\mathbf{\Gamma}\boldsymbol{Y} \tag{21.2.11}$$

a

$$\sigma^* = \mathbf{1}'\mathbf{\Gamma}\boldsymbol{Y}. \tag{21.2.12}$$

21.3 Rozptyly a kovariance odhadů.

Prvky matice $\mathbf{X'W^{-1}X}$ jsou koeficienty u μ^* a σ^* v rovnicích (21.2.8). Odtud a z (21.1.6) ihned vyplývá, že

$$\mathrm{var}(\mu^*) = \frac{\mathbf{a'W^{-1}a}}{\Delta}\sigma^2, \quad \mathrm{var}(\sigma^*) = \frac{\mathbf{1'W^{-1}1}}{\Delta}\sigma^2, \tag{21.3.1}$$

a

$$\mathrm{cov}(\mu^*, \sigma^*) = -\frac{\mathbf{1'W^{-1}a}}{\Delta}\sigma^2. \tag{21.3.2}$$

21.4 Příklad.

V případě uspořádaného výběru $\mathbf{U}^* = (U_{(1)}, \ldots, U_{(n)})'$ z exponenciálního rozdělení E(0, 1) jsou statistiky

$$T_j = 2(n-j+1)(U_{(j)} - U_{(j-1)}), \quad j = 1, \ldots, n, \quad U_{(0)} = 0,$$

vzájemně nezávislé a všechny mají rozdělení $\chi^2(2)$ (viz příkl. 4.7.3). Protože i-tá pořádková statistika $U_{(i)}$ se dá vyjádřit jako

$$U_{(i)} = \sum_{j=1}^{i} \big(U_{(j)} - U_{(j-1)}\big) = \frac{1}{2}\sum_{j=1}^{i} \frac{1}{n-j+1}\, T_j,$$

platí pro střední hodnoty a rozptyly

$$a_i \;=\; E(U_{(i)}) = \sum_{j=1}^{i} \frac{1}{n-j+1}, \qquad i = 1, \ldots, n, \tag{21.4.1}$$

$$w_{ii} \;=\; \mathrm{var}(U_{(i)}) = \sum_{j=1}^{i} \frac{1}{(n-j+1)^2}, \quad i = 1, \ldots, n. \tag{21.4.2}$$

Pro $1 \le i_1 < i_2 \le n$ platí

$$2\,\mathrm{cov}(U_{(i_1)}, U_{(i_2)}) = \mathrm{var}(U_{(i_1)}) + \mathrm{var}(U_{(i_2)}) - \mathrm{var}(U_{(i_2)} - U_{(i_1)}).$$

Avšak

$$\mathrm{var}(U_{(i_2)} - U_{(i_1)}) = \mathrm{var}\Big(\frac{1}{2}\sum_{j=i_1+1}^{i_2} \frac{1}{n-j+1} T_j\Big) = \sum_{j=i_1+1}^{i_2} \frac{1}{(n-j+1)^2},$$

takže

$$2\,\mathrm{cov}(U_{(i_1)}, U_{(i_2)}) = \sum_{j=1}^{i_1} \frac{1}{(n-j+1)^2} + \sum_{j=1}^{i_2} \frac{1}{(n-j+1)^2} - \sum_{j=i_1+1}^{i_2} \frac{1}{(n-j+1)^2}.$$

Odtud a z (21.4.2) vyplývá, že

$$w_{i_1 i_2} = \mathrm{cov}(U_{(i_1)}, U_{(i_2)}) = \mathrm{var}(U_{(i_1)}) = w_{i_1 i_1}, \quad 1 \le i_1 < i_2 \le n. \tag{21.4.3}$$

Matice (21.2.2) je tedy v případě výběru z rozdělení $E(0,1)$ rovna

$$\mathbf{W} = \begin{pmatrix} w_{11}, & w_{11}, & \dots, & w_{11} \\ w_{11}, & w_{22}, & \dots, & w_{22} \\ \vdots & \vdots & \ddots & \vdots \\ w_{11}, & w_{22}, & \dots, & w_{nn} \end{pmatrix} \tag{21.4.4}$$

Uvažujme případ $n = 2$. Pak

$$a_1 = \frac{1}{2}, \quad a_2 = \frac{3}{2}, \quad w_{11} = \frac{1}{4}, \quad w_{22} = \frac{5}{4}.$$

Odtud

$$\mathbf{W}^{-1} = \begin{pmatrix} \frac{1}{4}, & \frac{1}{4} \\ \frac{1}{4}, & \frac{1}{4} \end{pmatrix}^{-1} = \begin{pmatrix} 5, & -1 \\ -1, & 1 \end{pmatrix}, \qquad \mathbf{1a}' - \mathbf{a1}' = \begin{pmatrix} 0, & 1 \\ -1, & 0 \end{pmatrix}.$$

Dále

$$\mathbf{1}'\mathbf{W}^{-1}\mathbf{1} = 4, \quad \mathbf{a}'\mathbf{W}^{-1}\mathbf{a} = 2, \quad \mathbf{1}'\mathbf{W}^{-1}\mathbf{a} = 2, \quad \Delta = 4,$$

a tedy

$$\mathbf{\Gamma} = \begin{pmatrix} 0, & 1 \\ -1, & 0 \end{pmatrix}.$$

Pak nejlepší lineární nestranné odhady parametrů A a δ rozdělení $\mathrm{E}(A,\, \delta)$ rpo $n = 2$ jsou rovny

$$A^* = (-\frac{1}{2}, -\frac{3}{2}) \begin{pmatrix} 0, & 1 \\ -1, & 0 \end{pmatrix} \begin{pmatrix} Z_{(1)} \\ Z_{(2)} \end{pmatrix} = \frac{1}{2}(3Z_{(1)} - Z_{(2)}),$$

$$\delta^* = (1, 1) \begin{pmatrix} 0, & 1 \\ -1, & 0 \end{pmatrix} \begin{pmatrix} Z_{(1)} \\ Z_{(2)} \end{pmatrix} = Z_{(2)} - Z_{(1)};$$

jejich rozptyly a kovariance

$$\mathrm{var}(A^*) = \frac{1}{2}\delta^2, \quad \mathrm{var}(\delta^*) = \delta^2, \quad \mathrm{cov}(A^*, \delta^*) = -\frac{1}{2}\delta^2.$$

Lze ukázat (viz [32] a též [16], str. 204), že pro $n \geq 2$ jsou statistiky A^* a δ^* dány výrazy

$$A^* = \frac{n+1}{n} Z_{(1)} - \frac{1}{n(n-1)} \sum_{i=2}^{n} Z_{(i)} = Z_{(1)} - \frac{1}{n}\delta^*, \tag{21.4.5}$$

$$\delta^* = -Z_{(1)} + \frac{1}{n-2} \sum_{i=2}^{n} Z_{(i)} = \frac{1}{n-1} \sum_{i=2}^{n} (Z_{(i)} - Z_{(1)}) \tag{21.4.6}$$

a jejich rozptyly

$$\mathrm{var}(A^*) = \frac{\delta^2}{n(n-1)}, \qquad \mathrm{var}(\delta^*) = \frac{\delta^2}{n-1}. \tag{21.4.7}$$

21.5 Zjednodušení pro symetrická rozdělení.

V případě, že veličina Z má symetrické rozdělení podle bodu μ (tj. platí-li vztahy (4.3.5)), vyplývá ze vztahů (4.4.2) a (21.2.3), že

$$a_i = -a_{n-i+1}, \qquad i = 1, \ldots, n, \tag{21.5.1}$$

a ze vztahů (4.5.7) a (21.2.4) vyplývá, že

$$w_{i_1 i_2} = w_{n-i_2+1, n-i_1+1}, \qquad i_1, i_2 = 1, \ldots, n. \tag{21.5.2}$$

Uvažujme čtvercovou symetrickou matici

$$\mathbf{J} = \begin{pmatrix} 0, & 0, & \ldots, & 0, & 1 \\ 0, & 0, & \ldots, & 1, & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1, & 0, & \ldots, & 0, & 0 \end{pmatrix}. \tag{21.5.3}$$

Pak (21.5.1) lze zapsat ve tvaru

$$\mathbf{a} = -\mathbf{J}\mathbf{a} \tag{21.5.4}$$

a (21.5.2) ve tvaru $\mathbf{W} = \mathbf{JWJ}$ neboli (protože $\mathbf{J}^{-1} = \mathbf{J}$)

$$\mathbf{W}^{-1} = \mathbf{JW}^{-1}\mathbf{J}. \tag{21.5.5}$$

Nyní

$$\mathbf{1}'\mathbf{W}^{-1}\mathbf{a} = \mathbf{1}'\mathbf{JW}^{-1}\mathbf{J}(-\mathbf{Ja}) = -(\mathbf{1}'\mathbf{J})\mathbf{W}^{-1}(\mathbf{I}_n\mathbf{a}) = -\mathbf{1}'\mathbf{W}^{-1}\mathbf{a},$$

takže

$$\mathbf{1}'\mathbf{W}^{-1}\mathbf{a} = -\mathbf{1}'\mathbf{W}^{-1}\mathbf{a} = 0. \tag{21.5.6}$$

Odtud a z (21.3.2) vyplývá, že v případě výběru ze symetrického rozdělení jsou odhady μ^* a σ^* nekorelované.

Použijeme-li vztahu (21.5.6) v soustavě normálních rovnic (21.2.8), dostáváme

$$\mu^* = \frac{\mathbf{1}'\mathbf{W}^{-1}\mathbf{Y}}{\mathbf{1}'\mathbf{W}^{-1}\mathbf{1}}, \qquad \sigma^* = \frac{\mathbf{a}'\mathbf{W}^{-1}\mathbf{Y}}{\mathbf{a}'\mathbf{W}^{-1}\mathbf{a}}, \tag{21.5.7}$$

a protože v (21.2.10) je druhý člen nulový, vyplývá z (21.3.1), že

$$\mathrm{var}(\mu^*) = \frac{\sigma^2}{\mathbf{1}'\mathbf{W}^{-1}\mathbf{1}}, \qquad \mathrm{var}(\sigma^*) = \frac{\sigma^2}{\mathbf{a}'\mathbf{W}^{-1}\mathbf{a}}. \tag{21.5.8}$$

21.6 Příklad.

V případě výběru rozsahu $n = 4$ z rozdělení $\mathrm{N}(\mu, \sigma^2)$ nalezneme v [32] hodnoty

$$a_4 = -a_1 = 1,029\,375\,373; a_3 = -a_2 = 0,297\,011\,382;$$

$$w_{11} = w_{44} = 0,491\,715\,237; \quad w_{12} = w_{21} = w_{34} = w_{43} = 0,245\,592\,693;$$

$$w_{13} = w_{31} = w_{24} = w_{42} = 0,158\,008\,070; \quad w_{14} = w_{41} = 0,104\,684\,000;$$

$$w_{22} = w_{33} = 0,360\,455\,343; \quad w_{23} = w_{32} = 0,235\,943\,894.$$

Dosazením do (21.5.7) a (21.5.8) dostaneme

$$\begin{aligned} \mu^* &= 0,25\,(Z_{(1)} + Z_{(2)} + Z_{(3)} + Z_{(4)} = \overline{Z}, \\ \sigma^* &= 0,453\,9\,(Z_{(4)} - Z_{(1)}) + 0,110\,2\,(Z_{(3)} - Z_{(2)}), \end{aligned}$$

$$\mathrm{var}(\mu^*) = 0,250\,\sigma^2, \qquad \mathrm{var}(\sigma^*) = 0,180\,\sigma^2.$$

V [30] jsou tabelovány odhady μ^* a σ^* a jejich rozptyly pro $n = 2\,(1)\,10$ v případě výběrů z normálního rozdělení.

21.7 Cenzorované výběry.

Zatím jsme uvažovali případ úplných uspořádaných výběrů $(Z_{(1)}, \ldots, Z_{(n)})'$, tj. situaci, kdy známe všechna pozorování ve výběru. Často se však setkáváme s tzv. *cenzorovanými výběry,* kdy známe jen $r < n$ pozorování, např. pri testech životnosti jen r nejmenších pozorování, jindy zase $r < n$ největších pozorování (v těchto dvou případech hovoříme o *jednostranně cenzorovaných výběrech*) nebo r prostředních pozorování (neznáme s_1 nejmenších a s_2 největších pozorování, $r = n-(s_1+s_2)$; v tomto případě hovoříme o *oboustranně cenzorovaných výběrech*).

Uvažujme obecný případ, kdy máme r pořádkových statistik $Z_{(i_1)}, Z_{(i_2)}, \ldots, Z_{i_r}$, $1 \le i_1 < i_2 < \ldots < i_r \le n$, $2 \le r < n$, ve výběru z rozdělení náhodné veličiny Z uvažované v odst. 21.2. Známe hodnoty (21.2.1) a (21.2.2) pro tato $i_1, \ldots, i_r$. Pak můžeme označit

$$Y_k = Z_{(i_k)}, \quad a_k = a_{ik}, \quad w_{k_1 k_2} = w_{i_{k_1} i_{k_2}}, \quad k, k_1, k_2 = 1, \ldots, r, \qquad (21.7.1)$$

a přímo použít výsledků odst. 21.2.

V případě, že veličina Z má symetrické rozdělení podle bodu μ a že

$$i_k = n - i_{r-k+1} + 1, \quad k = 1, \ldots, r \qquad (21.7.2)$$

(tj. že $i_1 = n - i_r + 1$, $i_2 = n - i_{r-1} + 1, \ldots$), se použije výsledků odst. 21.5.

V [30] jsou tabelovány odhady μ^*, σ^* a jejich rozptyly pro cenzorované výběry rozsahu $n = 2(1)10$ z rozdělení $\mathrm{N}(\mu, \sigma^2)$.

21.8 Příklad.

Uvažujme výběr rozsahu $n = 4$ z rozdělení $\mathrm{N}(\mu, \sigma^2)$, v němž známe jen $r = 3$ největší pozorování. Hodnoty a_k a $w_{k_1 k_2}$, $k, k_1, k_2 = 1, 2, 3$, jsou uvedeny v příkl. 21.6. Dosazením do výrazů odst. 21.2 dostaneme

$$\begin{aligned} \mu^* &= 0,116 Z_{(2)} + 0,241 Z_{(3)} + 0,643 Z_{(4)}, \\ \sigma^* &= -0,697 Z_{(2)} - 0,127 Z_{(3)} + 0,824 Z_{(4)}, \end{aligned}$$

$$\mathrm{var}(\mu^*) = 0,287\,\sigma^2, \quad \mathrm{var}(\sigma^*) = 0,302\,\sigma^2.$$

21.9 Úloha.

V příkladě 21.8 uvažujte $s_1 = 1,\ s_2 = 1$ (tj. máte jen pozorování $Z_{(2)}$ a $Z_{(3)}$. Nalezněte $\mu^*,\ \sigma^*$ a jejich rozptyly. (Využijte výsledků odst. 21.5.)

$$\left[\begin{array}{ll} \mu^* = 0,5\,(Z_{(2)} + Z_{(3)}), & \mathrm{var}(\mu^*) = 0,298\,\sigma^2, \\ \sigma^* 1,683^4\,(Z_{(3)} - Z_{(2)}), & \mathrm{var}(\sigma^*) = 0,706\,\sigma^2. \end{array}\right]$$

Literatura

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

[2] Aplikovaná matematika I, II. Oborová encyklopedie SNTL. Praha, SNTL 1977 a 1978.

[3] *Blackwell, D., Girshick, M. A.*: Teorie her a statistického rozhodování. Praha, NČSAV 1964.

[4] *Bury, K. V.*: Statistical Models in Applied Science. New York, J. Wiley 1975.

[5] *Conover, W. J.*: Practical Nonparametric Statistics. New York, J. Wiley 1971.

[6] *Cramér, H.*: Mathematical Methods of Statistics. Princeton, Princeton University Press 1946.

[7] ČSN 42 0368. Zkoušky únavy kovů. Praha, Vydavatelství ÚNM 1974.

[8] *David, H. A.*: Order Statistics. New York, J. Wiley 1970.

[9] *Drapet, N. R., Smith, H.*: Applied Regression Analysis. New York, J. Wiley 1966.

[10] *Dunin-Barkovskij, I. V., Smirnov, N. V.*: Teorija verojatnostej i matematičeskaja statistika v technike. Moskva, GITTL 1955.

[11] *Fabian, V.*: Základní statistické metody. Praha, NČSAV 1963.

[12] *Felix, M., Bláha, K.*: Matematickostatistické metody v chemickém průmyslu. Praha, SNTL 1962.

[13] *Hahn, G. J., Shapiro, S. S.*: Statistical Models in Engineering. New York, J. Wiley 1967.

[14] *Hájek, J., Šidák, Z.*: Theory of Rank Tests. Praha, Academia 1967.

[15] *Hald, A.*: Statistical Theory with ENgineering Applications. New York, J. Wiley 1952.

[16] *Hátle, J., Likeš, J.*: Základy počtu pravděpodobnosti a matematické statistiky. Praha, SNTL/ALFA 1974.

[17] *Hollander, M., Wolfe, D. A.*: Nonparametric Statistical Methods. New York, J. Wiley 1973.

[18] *Janko, J.*: Statistické tabulky. Praha, NČSAV 1958.

[19] *Johnson, N. I., Kotz, S.*: Distribution in Statistics. Discrete Distributions. Continuous Univariate Distribution 1, 2. Boston. Houghton Mifflin Company, 1969 a 1970.

[20] *Kendall, M. G., Stuart, A.*: The Advanced Theory of Statistics. Volume II. London, Griffin 1967.

[21] *Kendall, M. G.*: Rank Correlation Methods. London, Griffin 1975.

[22] *Lehmann, E. L.*: Testing Statistical Hypothesis. New York, J. Wiley 1959.

[23] *Likeš, J., Laga, J.*: Základní statistické tabulky. Praha, SNTL, 1978.

[24] *Likeš, J., Machek, J.*: MVŠT. Počet pravděpodobnosti. Praha, SNTL, 1981.

[25] *Lilliefors, H. W.*: On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. J. Amer. Statist. Assoc., 62, 1967, s. 399.

[26] *Lilliefors, H. W.*: On the Kolmogorov-Smirnov Test for the Exponential Distribution with Mean Unknown. J. Amer. Statist. Assoc., 64, 1969, s. 387.

[27] *Machek, J.*: Teorie odhadu. Praha, SPN 1980.

[28] *Nalimov, V. V.*: Primenenie matematičeskoj statistiki pri analize veščestva. Moskva, Fizmatgizdat 1960.

[29] *Neymann, J.*: Lectures and Conferences on Mathematical Statistics. Washington, Department of Agriculture 1938.

[30] *Owen, D. B.*: Handbook of Statistical Tables. Massachusetts, Addison - Wesley 1962.

[31] *Rao, C. R.*: Lineární metody statistické indukce a jejich aplikace. Praha, Academia, 1972.

[32] *Sarhan, A. E., Greenberg, B. G.*: Contributions to Order Statistics. New York, J. Wiley 1962.

[33] *Siegel, S.*: Nonparametric Statistics for the Behavioral Sciences. New York, McGraw-Hill 1956.

[34] *Šor, J. B.*: Statistické metody analýzy a kontroly jakosti a spolehlivosti. Praha, SNTL 1965.

[35] *Thoman, D. R., Bain, L. J., Antle, C. E.*: Inferences on the Parameters of the Weibull Distribution. Technometrics, 11, 1969, s. 445 - 460.

[36] *Wald, A.*: Statistical Decision Functions. New York, J. Wiley 1950.

[37] *Wilks, S. S.*: Mathematicak Statistics. New York, J. Wiley 1962.

Rejstřík