UNIVERZITA J. E. PURKYNĚ

Fakulta přírodovědecká

STATISTICKÉ ZPRACOVÁNÍ VÝSLEDKŮ MĚŘENÍ

Josef HUMLÍČEK

BRNO 1984

V každém konkrétním měření je třeba zvládnout řadu problémů, většinou specifických pro danou úlohu. Správnost zjištěných hodnot záleží především na potlačení systematických chyb, způsobených měřicími přístroji nebo nevhodným postupem. Jádro tohoto problému vystihuje jednoduchý příklad: kratším metrem naměříme nesprávné (příliš velké) délky. Práce spojená s odstraněním možných systematických chyb představuje obvykle značnou část námahy vynaložené na celé měření.

Přes obrovskou různorodost mají měření výrazný společný rys. Je jím fakt, že opakování za stejných podmínek nedává přesně stejné výsledky. Jednak se uplatňují náhodné chyby měření, nebo se také samotné studované objekty projevují v náhodných jevech, které se řídí pouze pravděpodobnostními zákony (měření v mikrosvětě). Potřebné informace ze souboru naměřených dat, ve kterém jsou patrné náhodné vlivy, získáváme vhodným statistickým zpracováním.

Popisem náhodných jevů se zabývá teorie pravděpodobnosti, zpracováním pozorovaných náhodných výsledků další matematická disciplína - statistika. Statistika používá pojmů a výsledků teorie pravděpodobnosti; v kapitole I tohoto skripta jsou potřebné základy vyloženy. K pochopení statistických metod je třeba porozumět pravděpodobnostnímu popisu náhodných jevů a zvládnout použití náhodných proměnných. Užitečné je i podrobnější seznámení s několika typy často používaných rozdělení, které je rovněž soustředěno do kapitoly I.

Základní statistickou úlohou je zjišťování hodnot parametrů zkoumaného objektu z naměřených dat. Metody statistického odhadu parametrů jsou obsahem kapitoly II. Velká pozornost je věnována příkladům; je možné, že pro řadu čtenářů by mohly být právě tyto příklady vhodnými "vstupními body" do studované problematiky. Výklad statistických testů hypotéz v kapitole III je stručný, možnost posoudit rozdělení dat pomocí testů dobré shody by však měla být považována za důležitou.

K dalšímu studiu je k dispozici rozsáhlá literatura, z dostupných pramenů jsem vybral jen malou část. Zacházení s výsledky měření může být pobídkou k přemýšlení o základech matematiky náhody. Myslím, že by bylo chybou podceňovat elementární úvahy o pravděpodobnosti, které najdeme v několika dobrých populárních knihách.

Brno, červen 1983 Josef Humlíček

O b s a h

I. Základní pojmy teorie pravděpodobnosti

1. Pravděpodobnost jevů

V současné době je známo několik různých způsobů, jak definovat kvantitativně pravděpodobnost. Uvedeme tři možnosti, z nichž každá je svým způsobem výhodná a jejich srovnání je užitečné pro pochopení probl mů stojících v cestě zavedení univerzální definice.

Klasická definice

Pravděpodobnost P(X) určitého jevu X určujeme pomocí souboru tzv. elementárních událostí; označíme je $E_1,\ldots,E_n$. To jsou navzájem se vylu čující jevy (nastane-li jeden z nich, nemůže nastat žádný jiný), o kte rých předpokládáme, že jsou "stejně pravděpodobné", nebo "stejně možné" Pojem stejné pravděpodobnosti pokládáme za základní a nesnažíme se ho d finovat. Jestliže se událost X dá vyjádřit jako sjednocení některé m-ti různých elementárních událostí (t.j. jako jev, při kterém nastane E_{k_1} r bo E_{k_2} nebo ... nebo E_{k_m} se všemi $k_1,\ldots,k_m$ navzájem různými), položíme P(X)=m/n. Pro pravděpodobnost elementárních událostí máme tedy $P(E_i)=1/$ pro všechna i=1,...,n. Podstatná část klasické definice se dá vyjádřit následující formulací:

$$\text{pravděpodobnost} = \frac{\text{počet příznivých případů}}{\text{počet všech možných případů}} . \qquad (1)$$

Ihned je ovšem třeba doplnit, že všechny možné případy musí být stejně pravděpodobné. Vyhledání množiny elementárních událostí je obvykle založeno na symetrii objektů, které se daného jevu účastní (házení ideální kostkou, ruleta apod.). Není-li počet všech možných případů konečný, al zůstává možnost zdůvodnit stejnou pravděpodobnost některých podmnožin všech jevů, lze definici (1) v podstatě zachovat. Namísto počtu případů je nutné použít vhodnou míru velikosti oblastí, reprezentujících příznivé a všechny možné případy (délky, plochy atd). V učebnicích teorie pravděpodobnosti se v těchto okolnostech používá termínu geometrické pravděpodobnosti.

Statistická definice

Označme počet pokusů, ve kterých je sledován náhodný jev X, symbolem N. Jestliže v M případech jev X nastal (ve zbylých N-M nenastal), m žeme definovat pravděpodobnost X jako limitu relativní četnosti M/N při N jdoucím k nekonečnu:

$$P(X) = \lim_{N\to\infty} \frac{M}{N} . \qquad (2)$$

Pravděpodobnosti nemožného a jistého jevu jsou tedy po řadě 0 a 1; je-l X sjednocením konečného počtu vzájemně se vylučujících jevů $A_1,\ldots,A_k$, je zřejmě $P(X)=P(A_1)+\ldots+P(A_k)$. Tato definice vyjadřuje intuitivně zřejmou souvislost mezi pravděpodobností jevu a jeho četností při opakovaných pokusech. Ačkoliv nekonečnou řadu pokusů nelze realizovat, předpo-

kládáme, že s rostoucím počtem N se relativní četnost blíží k limitní (i když třeba neznámé)hodnotě (2).

Moderní definice

Pravděpodobnost je definována jako číselná míra na množině F všech možných jevů (ke každému jevu z F je přiřazeno číslo P), splňující následující axiomy:

(a) $P(X) \geq 0$ pro všechny jevy $X \in F$;

(b) $P(U) = 1$ pro jistý jev U(t.j. pro takový jev U, který nastává vždycky); (3)

(c) $P(A_1$ nebo A_2 nebo ...) $= P(A_1)+P(A_2)+\ldots$ pro libovolné vzájemně se vylučující jevy $A_1, A_2, \ldots$

Vlastnosti pravděpodobnosti z klasické a statistické definice jsou zachovány, chybí jen předpis pro konkrétní přiřazení numerických hodnot pravděpodobností jednotlivým jevům. To je přirozený důsledek požadavku, aby aparát teorie pravděpodobnosti mohl popisovat stejné množiny náhodných jevů, které se liší hodnotami pravděpodobností. Například při házení ideální kostkou je $P(1)=\ldots=P(6)=1/6$ (1,...,6 znamená výsledek hodu); odchylka od ideálního stavu vede k tomu, že se pravděpodobnosti liší od 1/6 a jejich hodnoty je třeba zjistit. Metody teorie pravděpodobnosti však fungují stejně v obou případech.

Pokusme se zformulovat hodnocení uvedených třech způsobů definice pravděpodobnosti. Pravděpodobnostní míra zavedená v moderní definici (3) reprezentuje podstatnou stránku společnou náhodným jevům; je vhodná pro logickou výstavbu matematické teorie. Klasický přístup (1) prokáže mnohdy cenné služby proto, že vyplňuje kostru obecných požadavků hodnotami pravděpodobností. I když velmi často potřebnou množinu stejně pravděpodobných elementárních událostí nenajdeme, nemá cenu klasickou definici odmítnout; tím bychom se ochudili o mnoho podstatných výsledků. Statistickou definici (2) považují někteří autoři za jedině správnou. V souvislosti se dvěma druhými alternativami se však přikloníme k chápání vztahu (2) jako prostředku k určení numerických hodnot pravděpodobnostní míry z moderní definice.

Podmíněná pravděpodobnost

Pravděpodobnost náhodného jevu A za předpokladu, že nastává jev B, se nazývá podmíněnou pravděpodobností. Značí se symbolem $P(A|B)$ a je definována pomocí pravděpodobnosti jevu AaB (t.j. jak A tak B současně) následujícím vztahem:

$$P(AaB) = P(B)P(A|B). \qquad (4)$$

$P(A|B)$ je definována jen tehdy, je-li $P(B) > 0$.

Nezávislost náhodných jevů

Dva náhodné jevy A_1, A_2 se nazývají nezávislé, jestliže

$P(A_1 a A_2) = P(A_1)P(A_2)$, neboli $P(A_1|A_2) = P(A_1)$, $P(A_2|A_1) = P(A_2)$. (

Pojem nezávislosti náhodných jevů je velmi důležitý a budeme se s ním často setkávat. Hořejší definice se však prakticky ke zjištění nezávislosti nepoužívá. Obvykle využijeme empirických poznatků k tomu, abychom rozhodli o správnosti tvrzení, že dva jevy spolu "nijak nesouvisí", a tu to nazávislost vyjádříme formálně vztahem (5).

Jednoduchá pravidla

Pro pravděpodobnost, že jev A nenastane (neboli nastane jev, který označíme buď ne A, nebo $\bar{A}$), vychází

$$P(\bar{A}) = 1-P(A). \qquad (6$$

Pravděpodobnost, že nastane alespoň jeden ze dvou jevů A,B je

$$P(A \text{ nebo } B) = P(A)+P(B)-P(AaB). \qquad (7$$

Pokud jev A musí nastat společně s právě jedním z k navzájem se vylučujících jevů $B_1,\ldots,B_k$, to jest A je jev (AaB_1)nebo(AaB_2)nebo...nebo(AaB_k), dostaneme s pomocí (3) a (4) tzv. vzorec pro úplnou pravděpodobnost

$$P(A) = \sum_{i=1}^{k} P(AaB_i) = \sum_{i=1}^{k} P(B_i)P(A|B_i). \qquad (8$$

Bayesův teorém

Z definice podmíněné pravděpodobnosti (4) plyne rovnost

$$P(AaB) = P(B)P(A|B) = P(A)P(B|A), \text{ neboli } P(B|A) = \frac{P(A|B)P(B)}{P(A)} \qquad (9$$

To je tzv. Bayesův teorém. Je-li jev B jedním ze vzájemně se vylučujících jevů B_i ze vztahu (8), vychází odtud Bayesův vzorec ve tvaru

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{i=1}^{k} P(B_i)P(A|B_i)}. \qquad (10)$$

Smysl posledního vztahu je následující: předpokládejme, že umíme najít pravděpodobnosti $P(B_1),\ldots,P(B_k)$ a $P(A|B_1),\ldots,P(A|B_k)$ pro náhodný jev A, který nastává spolu s právě jedním jevem $B_1,\ldots,B_k$. Budeme-li ze všech výsledků pokusů vybírat jen ty, ve kterých událost A nastala, dává Bayesův teorém (10) hodnoty pravděpodobností $P(B_i|A)$, které se obecně liší od výchozích pravděpodobností $P(B_i)$ náhodných jevů B_i sledovaných bez doplňující podmínky.

2. Náhodné proměnné

Studium náhodných jevů lze převést beze zbytku do řeči čísel prostřednictvím náhodné proměnné. To je proměnná veličina, jejíž hodnoty reprezentují všechny možné výsledky pokusu s náhodnými jevy; pravděpodobnosti jednotlivých výsledků jsou tak přiřazeny odpovídajícím hodnotám náhodné proměnné. Nejrůznější náhodné jevy, jejichž struktura je z hlediska uplatnění pravděpodobnostních zákonů stejná, jsou popisovány

jedinou náhodnou proměnnou. Pro zkoumání náhodných proměnných máme k dispozici rozvinutý aparát matematické analýzy.

Je-li množina hodnot náhodné proměnné η spočetná (lze je indexovat přirozenými čísly, například y_1, y_2,...), nazýváme η diskrétní náhodnou proměnnou a soubor pravděpodobností $P_\eta(y_1)$, $P_\eta(y_2)$,... diskrétní funkcí rozdělení. Výhodné je použití spojitých náhodných proměnných, které mohou nabývat libovolných reálných hodnot ze spojitých intervalů. Potom není možné přiřadit dané hodnotě nenulovou pravděpodobnost, protože pravděpodobnost spojená s intervalem hodnot by byla nekonečná. Přirozeným řešením tohoto problému je zavedení hustoty pravděpodobnosti; zadané hodnotě x spojité náhodné proměnné ξ přiřadíme hustotu

$$f_\xi(x) = \lim_{\Delta x \to 0} \frac{P(x \leq \xi < x + \Delta x)}{\Delta x} . \qquad (1)$$

Pravděpodobnost, že hodnoty ξ jsou z intervalu $\langle x, x + \Delta x)$ je tedy pro dostatečně malé Δx úměrná délce intervalu a koeficientem úměrnosti je hustota $f_\xi(x)$, neboli "pravděpodobnost na jednotkovou délku intervalu". Protože hustota je obecně funkcí x, dostaneme pravděpodobnost pro konečný interval $\langle x_1, x_2)$ integrací:

$$P(x_1 \leq \xi < x_2) = \int_{x_1}^{x_2} f_\xi(x)\,dx. \qquad (2)$$

S použitím Diracovy δ - funkce můžeme rozšířit zadání hustoty tak, že v sobě zahrnuje vlastnosti diskrétní i spojité náhodné proměnné. Například funkce

$$f_\xi(z) = P_1\,\delta(z-z_1) + P_2\,\delta(z-z_2) + (1-P_1-P_2)\,\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right) \qquad (3)$$

je hustotou náhodné proměnné ξ, která nabývá hodnot z_1, z_2 s pravděpodobnostmi P_1, P_2 a libovolných reálných hodnot různých od z_1, z_2 s hustotou danou třetím členem na pravé straně vzorce (3). Aby byl splněn požadavek pravděpodobnosti jistého jevu (viz (1.3)), musí platit

$$\sum_i P_\eta(y_i) = 1, \quad \int_{-\infty}^{\infty} f_\xi(x)\,dx = 1 \qquad (4)$$

pro libovolnou náhodnou proměnnou η (diskrétní) a ξ (spojitou, vně intervalů možných hodnot položíme hustotu rovnou nule). Z této normovací podmínky vychází faktor u exponenciální funkce ve vztahu (3).

Funkce náhodné proměnné

Je-li ξ náhodná proměnná s hustotou $f_\xi(x)$ a $h(x)$ zadaná funkce s hodnotami $y = h(x)$, můžeme přenést pravděpodobnostní míru z intervalů hodnot x do intervalů hodnot y, které jsou pak hodnotami nové náhodné proměnné (označme ji η). Je-li transformace $y = h(x)$ vzájemně jednoznačná, přechází interval $\langle x, x+dx)$ v $\langle y, y+dy)$ pro infinitezimální dx; přitom je $dy = |h'(x)|\,dx$. Pro hustotu $g_\eta(y)$ tedy s pomocí definice (1) dostaneme

$g_\eta(y)\,dy = f_\xi(x)\,dx$, a odtud

$$g_\eta(y) = \frac{f_\xi(x)}{|h'(x)|} = \frac{f_\xi[h^{-1}(y)]}{|h'[h^{-1}(y)]|} . \qquad (5)$$

$h'(x)$ značí derivaci a $h^{-1}(y)$ funkci inverzní: $x=h^{-1}(y)$, která podle předpokladu a jednoznačnosti transformace $y=h(x)$ existuje. V opačném případě se několik intervalů $\langle x, x+dx)$ transformuje do intervalu $\langle y, y+dy)$ a pravděpodobnosti je třeba sečíst přes všechny takové intervaly I:

$$g_\eta(y) = \sum_I \frac{f_\xi(x)}{|h'(x)|} . \qquad (6)$$

Vypočteme například hustotu pro $y=x^2$; v tomto případě je $x=\sqrt{y}$ pro $x>0$ a $x=-\sqrt{y}$ pro $x<0$. Odtud dostaneme

$$g_\eta(y) = \frac{f_\xi(-\sqrt{y}) + f_\xi(\sqrt{y})}{2\sqrt{y}} \quad \text{pro } y>0. \qquad (7)$$

Distribuční funkce

K úplnému zadání náhodné proměnné ξ se vedle diskrétní funkce rozdělení $P_\xi(x_i)$ nebo hustoty pravděpodobnosti $f_\xi(x)$ výborně hodí distribuční funkce $F_\xi(x)$ reálného argumentu $x\in(-\infty,\infty)$, definované vztahem

$$F_\xi(x) = P(\xi < x) \qquad (8)$$

pro diskrétní i spojité ξ. Souvislost s hustotou spojité náhodné proměnné je podle (2) následující:

$$F_\xi(x) = \int_{-\infty}^{x} f_\xi(t)\,dt, \quad f_\xi(x) = \frac{dF_\xi(x)}{dx} . \qquad (9)$$

Druhá z hořejších formulí platí jen tehdy, existuje-li v bodě x derivace. Funkce $F_\xi(x)$ má podle definice (8) skok velikosti $P(x_i)$ v každém bodě x_i, pro který je pravděpodobnost $P(x_i)$ nenulová. Distribuční funkce je neklesající, podle vztahu (4) jsou její krajní funkční hodnoty 0 a 1. Diskrétní funkci rozdělení, hustotě nebo distribuční funkci se stručně říká rozdělení náhodné proměnné.

Vícerozměrné náhodné proměnné

Některé náhodné jevy je třeba popisovat několika reálnými čísly. Uspořádaná n-tice reálných čísel $(x_1,\ldots,x_n)$, které je přiřazena pravděpodobnostní míra, tvoří hodnotu n-rozměrné náhodné proměnné. Často se používá také názvů náhodný vektor nebo soustava náhodných proměnných. Diskrétní náhodný vektor je určen pravděpodobnostmi $P(x_1,\ldots,x_n)$ toho, že jednotlivé složky nabudou diskrétních hodnot $x_{11},x_{12},\ldots,x_{n1},x_{n2},\ldots$. Hodnotám $(x_1,\ldots,x_n)$ spojitého náhodného vektoru ξ (podtržením symbolu zdůrazňujeme, že jde o veličinu s několika komponentami $\xi_1,\ldots,\xi_n$) je přiřazena hustota pravděpodobnosti

$$f_\xi(x_1,\ldots,x_n) = \lim_{\Delta x_1\to 0,\ldots,\Delta x_n\to 0} \frac{P(x_1\le\xi_1<x_1+\Delta x_1 \text{ a } \ldots \text{ a } x_n\le\xi_n<x_n+\Delta x_n)}{\Delta x_1\ldots\Delta x_n} . \qquad (10)$$

Distribuční funkce je definována zobecněním vztahu (8):

$$F_{\underline{\xi}}(x_1,\ldots,x_n) = P(\xi_1 < x_1 \text{ a } \ldots \text{ a } \xi_n < x_n). \tag{11}$$

Souvislost hustoty a distribuční funkce je analogií první z formulí (9):

$$F_{\underline{\xi}}(x_1,\ldots,x_n) = \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_n} f_{\underline{\xi}}(t_1,\ldots,t_n)\, dt_1 \ldots dt_n. \tag{12}$$

Pravděpodobnost nalezení funkčních hodnot v zadané oblasti Ω je

$$P(\underline{\xi} \in \Omega) = \int_{\Omega} \ldots \int f_{\underline{\xi}}(t_1,\ldots,t_n)\, dt_1 \ldots dt_n. \tag{13}$$

Marginální a podmíněné rozdělení

Projekce funkcí charakterizujících rozdělení pravděpodobností náhodného vektoru do "směrů" jeho komponent jsou označovány jako marginální rozdělení. Například

$$f_{\xi_1}(x_1) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f_{\underline{\xi}}(x_1,t_2,\ldots,t_n)\, dt_2 \ldots dt_n,$$

$$F_{\xi_1}(x_1) = F_{\underline{\xi}}(x_1,\infty,\ldots,\infty) \tag{14}$$

jsou marginální hustotou a marginální distribuční funkcí proměnné ξ_1. Tyto funkce určují pravděpodobnosti intervalů proměnné ξ_1 bez ohledu na hodnoty zbylých komponent $\underline{\xi}$.

Řezy rozdělovacích funkcí (jedna nebo několik komponent zadány pevně) se nazývají podmíněná rozdělení. Například z hustoty (10) dostaneme hustotu pravděpodobnosti komponent $\xi_2,\ldots,\xi_n$ za předpokladu, že ξ_1 nabývá pevné hodnoty $x_1^{(o)}$, vhodným normováním:

$$f_{\xi_2,\ldots,\xi_n}(x_2,\ldots,x_n \mid \xi_1 = x_1^{(o)}) = \frac{f_{\underline{\xi}}(x_1^{(o)},x_2,\ldots,x_n)}{f_{\xi_1}(x_1^{(o)})}, \tag{15}$$

kde f_{ξ_1} je marginální hustota (14). Normovací faktor vychází přímo z definice podmíněné pravděpodobnosti (1.4).

Nezávislost náhodných proměnných

Má-li náhodný vektor $\underline{\xi}$ s komponentami $\xi_1,\ldots,\xi_n$ distribuční funkci, která je součinem marginálních distribučních funkcí,

$$F_{\underline{\xi}}(x_1,\ldots,x_n) = F_{\xi_1}(x_1)\ldots F_{\xi_n}(x_n), \tag{16}$$

označujeme náhodné proměnné $\xi_1,\ldots,\xi_n$ jako nezávislé. Smysl tohoto pojmenování je stejný jako v případě nezávislosti náhodných jevů (§1, vzorec (1.5)) - pravděpodobnosti nebo hustoty jedné z proměnných nezáleží vůbec na ostatních. Hodnoty podmíněných pravděpodobností nebo hustot (jako např. ve vztahu (15)) jsou tytéž jako bez podmínek, což se dá stručně vyjádřit formulí (16).

Výsledky měření - hodnoty náhodných proměnných

Měřením získáváme ve velké většině případů soubory čísel. Náhodné

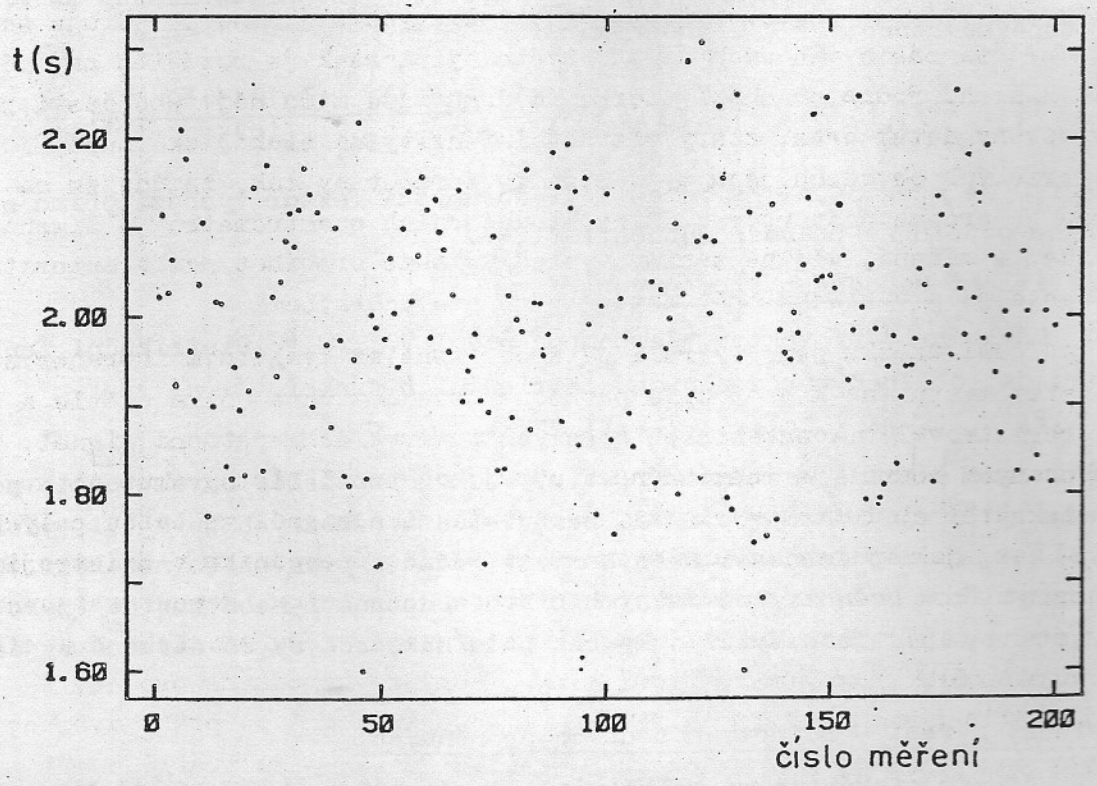
vlivy působící v procesu měření vedou k tomu, že je více možných výsledků. O tom se můžeme přesvědčit pouze opakováním celého měření v nezměněných podmínkách. Zkušenost nás učí, že se relativní četnosti možných výsledků s rostoucím počtem opakování blíží k pevným hodnotám - pravděpodobnostem. V opačném případě usuzujeme na změnu podmínek měření.

Naměřená čísla jsou hodnotami náhodných proměnných. Cílem je ovšem zjištění vlastností měřeného objektu, které jsou reprezentovány pevnými hodnotami parametrů. Tyto parametry, spolu s náhodnými vlivy, formují náhodné proměnné popisující výsledky měření. Vhodným zpracováním dat se snažíme potlačit vliv náhody a "určit", nebo lépe "odhadnout", hledané parametry. Statistický termín odhad je lepší, protože hodnoty parametrů vypočtené z naměřených dat tvoří opět náhodné proměnné a jejich souvislost se skutečnými parametry můžeme vyjádřit pouze pomocí pojmu pravděpodobnosti.

Na závěr ukážeme výsledky dvou typických měření.

Příklad měření časového intervalu

K měření byl vybrán časový interval známé délky - totiž doba, za kterou proběhne vteřinová ručička hodin dva vteřinové dílky ciferníku. Při průchodu ručky výchozí značkou byly "ručně" spuštěny, a po průběhu dvou dílků opět ručně zastaveny, digitální hodiny, které počítaly intervaly délky asi 0.58 ms (milisekundy). Počet tiků digitálních hodin byl



Obr. 1. Výsledky opakovaného měření času.

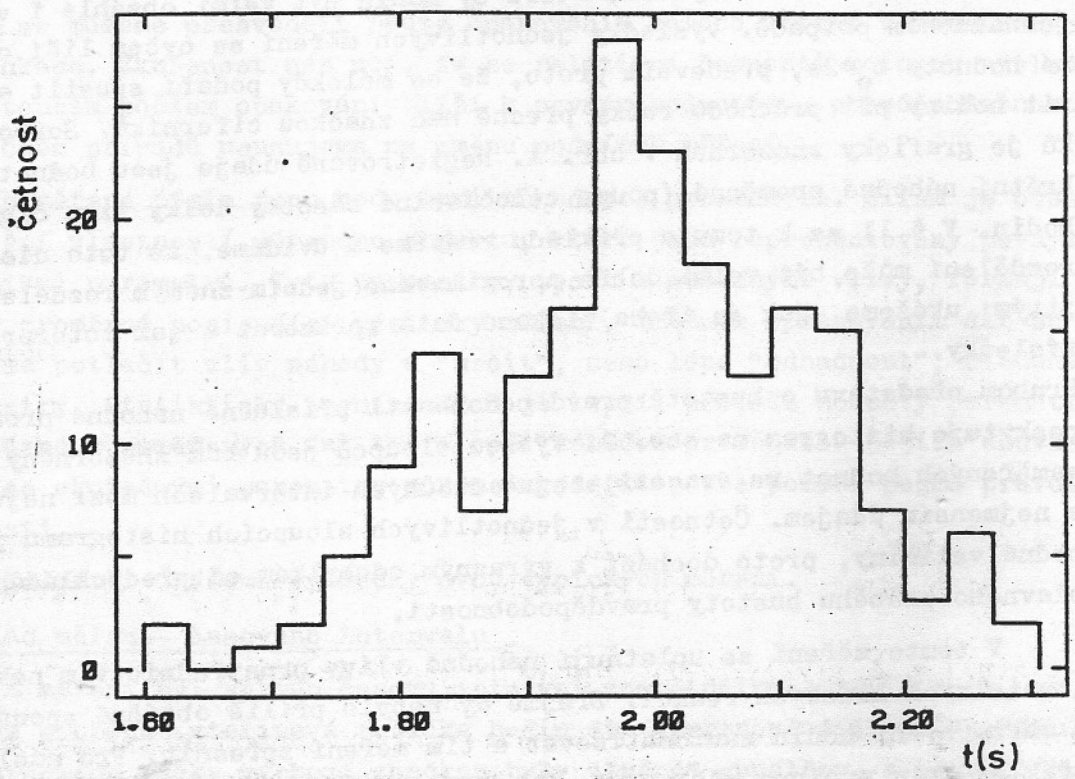
zaregistrován a měření opakováno celkem 200-krát. Vynecháme diskusi o možných systematických chybách, která by mohla být velmi obsáhlá i v tomto jednoduchém případě. Výsledky jednotlivých měření se ovšem liší od správné hodnoty t_0=2s, především proto, že se málokdy podaří spustit a zastavit hodiny při průchodu ručky přesně nad značkou ciferníku. Soubor výsledků je graficky znázorněn v obr. 1. Registrované údaje jsou hodnotami diskrétní náhodné proměnné (pouze celočíselné násobky délky tiku digitálních hodin. V § 11 se k tomuto příkladu vrátíme a uvidíme, že toto diskrétní rozdělení může být velmi dobře aproximováno jedním známým rozdělením spojitým; ukážeme, jak je třeba získaná data zpracovat a jak interpretovat výsledky.
Hrubou představu o hustotě pravděpodobnosti příslušné náhodné proměnné poskytuje histogram na obr. 2. Výškou sloupců jsou tam znázorněny počty naměřených hodnot ve dvaceti stejně dlouhých intervalech mezi největším a nejmenším údajem. Četnosti v jednotlivých sloupcích histogramu jsou náhodné veličiny, proto dochází k výrazným odchylkám od předpokládaného plavného průběhu hustoty pravděpodobnosti.

V tomto měření se uplatňují náhodné vlivy prostřednictvím nekontrolovatelných lidských reakcí. Zřejmě by nebylo příliš obtížné spouštění a zastavování hodin zautomatizovat a tím měření zpřesnit. Pro následující příklad bylo zvoleno měření, do jehož průběhu člověk nezasahuje.

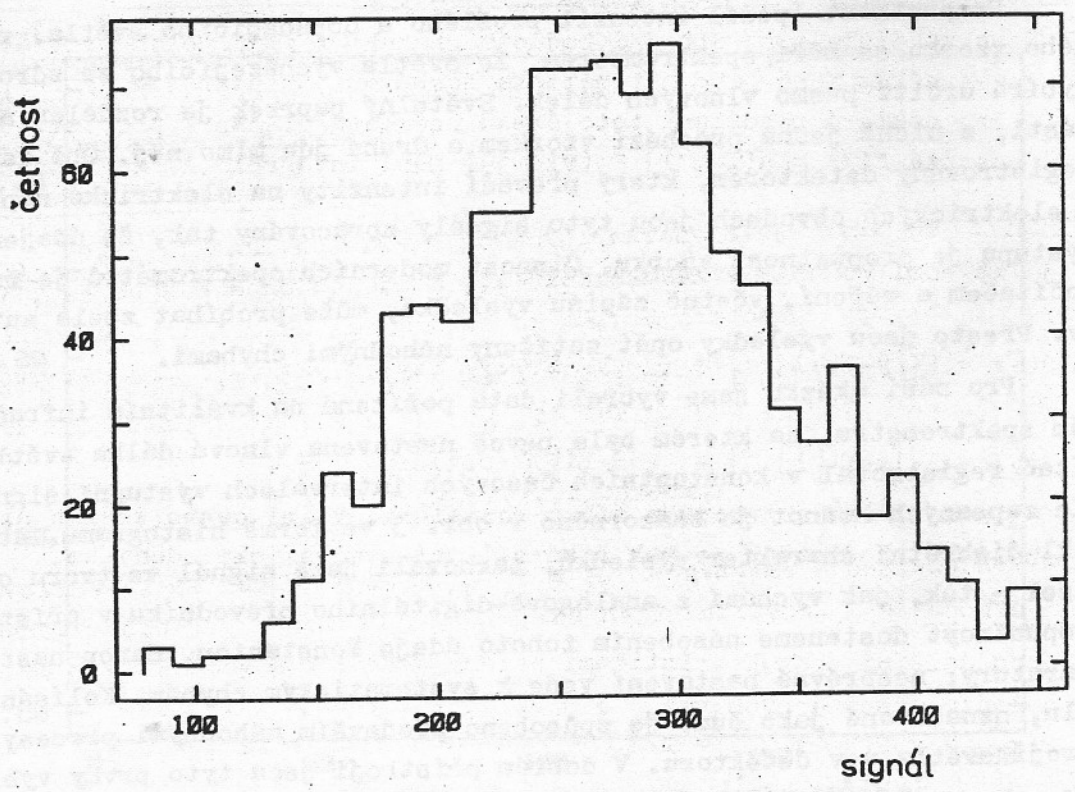
Příklad měření propustnosti spektrometrem

Propustnost (podíl intenzit prošlého a dopadajícího světla) zkoumaného vzorku se měří spektrometrem. Ze světla vycházejícího ze zdroje se vybírá určité pásmo vlnových délek. Světelný paprsek je rozdělen na dvě části, z nichž jedna prochází vzorkem a druhá jde mimo něj. Obě části jsou registrovány detektorem, který převádí intenzity na elektrické napětí. V elektrických obvodech jsou tyto signály zpracovány tak, že údajem na výstupu je propustnost vzorku. Činnost moderních spektrometrů je řízena počítačem a měření, včetně zápisu výsledků, může probíhat zcela automaticky. Přesto jsou výsledky opět zatíženy náhodnými chybami.

Pro naši ukázku jsme vybrali data pořízená na kvalitním infračerveném spektrometru, na kterém byla pevně nastavena vlnová délka světla a počítač registroval v konstantních časových intervalech výstupní signál. Tisíc zapsaných hodnot je znázorněno v obr. 3 ve formě histogramu. Aby vynikl diskrétní charakter výsledků, zachovali jsme signál ve tvaru celých čísel - tak, jak vychází z analogově-digitálního převodníku v přístroji. Propustnost dostaneme násobením tohoto údaje konstantou, danou nastavením aparatury; nesprávné nastavení vede k systematickým chybám. Kolísání signálu, označované jako šum, je způsobeno především náhodnými procesy ve zdroji světla a v detektoru. V dobrém přístroji jsou tyto prvky vybrány tak, aby poměr signál/šum byl co největší. Zlepšení dosažitelné s přijatelnou námahou obvykle není možné. Návod k optimálnímu zpracování dat,



2. Histogram sestavený z dat v obr. 1.



Obr. 3. Histogram výsledků 1000x opakovaného měření na infračerveném spektrometru.

která jsou k dispozici, poskytují statistické metody. Analýzou rozdělení náhodné veličiny z tohoto příkladu se budeme zabývat v §§ 20 a 21; ukážeme, že je stejného typu jako v hořejším příkladě ručního měření času.

3. Vlastnosti náhodných proměnných

Náhodná proměnná je úplně zadaná svojí distribuční funkcí, případně hustotou nebo diskrétní funkcí rozdělení. Velmi užitečné jsou následující číselné hodnoty, které vystihují některé podstatné vlastnosti rozdělení pravděpodobnosti.

Střední hodnota

náhodné veličiny η (diskrétní, rozdělení $P(y_i)$) a ξ (spojité, s hustotou $f_\xi(x)$) je definována vztahem

$$E(\eta) = \sum_i y_i P(y_i), \quad E(\xi) = \int_{-\infty}^{\infty} x f(x)\,dx , \tag{1}$$

jestliže tyto výrazy existují. Pro zadané funkce $g(\eta)$, $h(\xi)$ jsou střední hodnoty

$$E\left[g(\eta)\right] = \sum_i g(y_i)\cdot P(y_i), \quad E\left[h(\xi)\right] = \int_{-\infty}^{\infty} h(x) f(x)\,dx. \tag{2}$$

Je třeba si uvědomit, že střední hodnota není funkcí hodnot náhodné proměnné; symbolem $E(\xi)$ vyjadřujeme, že jde o střední hodnotu proměnné ξ. Je to lineární funkcionál, z definice (2) vychází

$$E\left[a d_1(\xi) + b d_2(\xi)\right] = a E\left[d_1(\xi)\right] + b E\left[d_2(\xi)\right] \tag{3}$$

pro libovolná čísla a,b. Střední hodnota charakterizuje polohu rozdělení. Označuje se někdy také jako matematické očekávání, expektance, střed rozdělení. Pro některá rozdělení toto číslo neexistuje, např. pro tzv. Cauchyovu hustotu $f(x) = \left[\pi(1+x^2)\right]^{-1}$ (viz §9) není integrál (1) definován.

Disperze

je střední hodnota kvadrátů odchylek od střední hodnoty náhodné veličiny:

$$D(\eta) = E\left\{\left[y-E(\eta)\right]^2\right\} = \sum_i \left[y_i - E(\eta)\right]^2 P(y_i),$$

$$D(\xi) = E\left\{\left[x-E(\xi)\right]^2\right\} = \int_{-\infty}^{\infty} \left[x-E(\xi)\right]^2 f(x)\,dx; \tag{4}$$

pokud existuje, charakterizuje šířku rozdělení. Velmi často se pro ni užívá symbolu σ^2, hodnotě σ se pak říká střední kvadratická odchylka.

Medián a moda

jsou dalšími charakteristikami polohy rozdělení. Medián $x_{1/2}$ je taková hodnota náhodné proměnné ξ s distribuční funkcí $F(x)$, pro kterou

$$F(x_{1/2}) = P(x < x_{1/2}) = 1/2. \tag{5}$$

Jinými slovy: pravděpodobnosti, že ξ nabude hodnot pod a nad mediánem jsou stejné a rovnají se jedné polovině. Pro symetrické rozdělení je medián

roven střední hodnotě. Moda x_m je hodnotou, pro kterou má hustota maximum $f(x_m)$, nebo nastává maximum pravděpodobnosti $P(x_m)$ v případě diskrétního rozdělení.

Momenty

Momentem řádu k náhodné veličiny ξ vzhledem k číslu c se nazývá střední hodnota $E\left[(\xi-c)^k\right]$. Momenty vzhledem k počátku (c=0) bývají označovány jako algebraické:

$$\nu_k = E(\xi^k), \tag{6}$$

momenty vzhledem ke střední hodnotě ($c = E(\xi)$) jsou centrální:

$$\mu_k = E\left\{\left[\xi - E(\xi)\right]^k\right\}. \tag{7}$$

Střední hodnota je tedy prvním algebraickým momentem (ν_1), disperze druhým centrálním momentem (μ_2).

Asymetrie a exces

Asymetrie γ_1 je definována jako

$$\gamma_1 = \sqrt{(\mu_3^2/\mu_2^3)} = \mu_3/\mu_2^{3/2}. \tag{8}$$

Pro symetrické rozdělení je μ_3 a tedy i asymetrie nulová. Pro nesymetrické rozdělení je γ_1 vhodnou mírou odchylky od symetrie. Exces

$$\gamma_2 = \mu_4/\mu_2^2 - 3 \tag{9}$$

je zvolen tak, aby pro normální rozdělení (§4) byl nulový. Umožňuje rychlé posouzení odlišnosti zadaného rozdělení od normálního (je mírou "špičatosti" ... $\gamma_2 > 0$ má rozdělení ostřejší, $\gamma_2 < 0$ rozdělení plošší než normální se stejnou disperzí).

Momenty náhodného vektoru

Pojmy střední hodnoty a momentů, zavedené ve vztazích (1),(2),(6) a (7), se dají snadno zobecnit pro vícerozměrnou náhodnou proměnnou. Kvůli jednoduchosti zápisu se omezíme na případ náhodného vektoru $\underline{\xi}$ se dvěma komponentami (ξ_1, ξ_2), který nabývá hodnot (x_1,x_2) s hustotou $f(x_1,x_2)$. Střední hodnota funkce $h(\xi_1,\xi_2)$ je definována formulí analogickou k (2):

$$E\left[h(\xi_1,\xi_2)\right] = \int_{-\infty}^{\infty}\int h(x_1,x_2)f(x_1,x_2)dx_1dx_2. \tag{10}$$

Zvolíme-li za funkci h mocniny ξ_1 a ξ_2, dostaneme z posledního vztahu momenty vektoru $\underline{\xi}$. Vypíšeme explicitně nejdůležitější z nich. Dvojice prvních algebraických momentů

$$E(\xi_1) = \int_{-\infty}^{\infty}\int x_1\, f(x_1,x_2)\, dx_1dx_2,\quad E(\xi_2) = \int_{-\infty}^{\infty}\int x_2\, f(x_1,x_2)\, dx_1dx_2 \tag{11}$$

bývá značena jako střed rozdělení vektoru $\underline{\xi}$. Dále jsou to druhé centrální momenty

$$D(\xi_1) = \sigma_1^2 = E\{[\xi_1 - E(\xi_1)]^2\}\,, \quad D(\xi_2) = \sigma_2^2 = E\{[\xi_2 - E(\xi_2)]^2\}, \tag{12}$$

což jsou disperze komponent ξ_1, ξ_2 (disperze marginálních rozdělení ze vztahů (2.14)). Konečně, smíšený druhý centrální moment

$$D(\xi_1, \xi_2) = E\{[\xi_1 - E(\xi_1)][\xi_2 - E(\xi_2)]\} = E(\xi_1\xi_2) - E(\xi_1)E(\xi_2) \tag{13}$$

se značí také jako kovariace ξ_1 a ξ_2, nebo korelační moment. Všimneme si, že z definicí (12) a (13) vychází $\sigma_1^2 = D(\xi_1, \xi_1)$.

Korelační koeficient

Koeficient korelace $\varrho(\xi_1, \xi_2)$ mezi ξ_1 a ξ_2 je definován vztahem

$$\varrho(\xi_1, \xi_2) = D(\xi_1, \xi_2)/\sqrt{D(\xi_1)D(\xi_2)} = D(\xi_1, \xi_2)/(\sigma_1\sigma_2); \tag{14}$$

snadno se ověří, že může nabývat pouze hodnot z intervalu $\langle -1, 1\rangle$. Jsou-li ξ_1, ξ_2 nezávislé (viz §2), je $E(\xi_1\xi_2) = E(\xi_1)E(\xi_2)$ a podle (13) vyjde $D(\xi_1, \xi_2) = 0$, tedy i $\varrho(\xi_1, \xi_2) = 0$. Je-li korelační koeficient nenulový, nemohou být příslušné náhodné proměnné nezávislé. Obrácené tvrzení však neplatí: existují závislé náhodné proměnné, které mají nulový korelační koeficient. Je-li $\varrho(\xi_1, \xi_2) = 0$, říkáme, že ξ_1 a ξ_2 jsou nekorelované, což je slabší vlastnost než nezávislost. Přesto je korelační koeficient užitečnou charakteristikou; podrobněji prozkoumáme jeho význam v případě normálně rozděleného náhodného vektoru v § 6.

Matice druhých momentů a korelačních koeficientů

Je-li počet n komponent náhodného vektoru $\underline{\xi}$ větší než 2, můžeme definovat smíšený druhý centrální moment a korelační koeficient pro každou dvojici ξ_i, ξ_j z příslušného dvojrozměrného marginálního rozdělení (rozdělení $\underline{\xi}$ integrované přes všechny složky kromě i-té a j-té). Čtvercovou matici s prvky

$$D_{ij} = D(\xi_i, \xi_j), \quad i,j=1,\ldots,n \tag{15}$$

nazýváme maticí druhých momentů, kovariančí nebo disperzní maticí, někdy také maticí chyb. V diagonále jsou disperze jednotlivých komponent: $D_{ii} = \sigma_i^2$, matice je symetrická ($D_{ij} = D_{ji}$). Čtvercová matice sestavená z korelačních koeficientů

$$\varrho_{ij} = \varrho(\xi_i, \xi_j) = D_{ij}/\sqrt{D_{ii}D_{jj}}\,, \quad i,j = 1,\ldots,n \tag{16}$$

se označuje jako korelační matice. Je opět symetrická, v diagonále jsou jedničky (korelační koeficient každé komponenty se sebou samou je roven jedné). Užitečnými čísly jsou tzv. globální korelační koeficienty ϱ_i pro každou komponentu ξ_i. Jsou to maximální hodnoty korelačního koeficientu $\varrho(\xi_i, \eta)$, když η probíhá všechny možné lineární kombinace všech komponent vektoru $\underline{\xi}$ kromě i-té. ϱ_i tedy udává míru korelace ξ_i se souborem zbylých komponent. Předpokládejme, že ke kovariační matici (15) existuje matice inverzní; její i-tý diagonální prvek označíme $(D^{-1})_{ii}$. Pro globální

korelační koeficient vyjde jednoduchá formule

$$\varrho_i = \sqrt{1 - \left[D_{ii}\ (D^{-1})_{ii}\right]^{-1}}\ . \tag{17}$$

Jestliže je kovariační matice singulární (neexistuje matice inverzní), je alespoň jedna ze složek ξ_j nějakou lineární kombinací ostatních, tedy úplně korelovaná s touto lineární kombinací ($\varrho_j = 1$).

Lineární funkce náhodných proměnných

Lineární kombinace $a_1\xi_1+\ldots+a_n\xi_n$ náhodných proměnných $\xi_1,\ldots,\xi_n$ ($a_1,\ldots,a_n$ jsou čísla) tvoří náhodnou proměnnou, pro niž snadno najdeme střední hodnotu a disperzi:

$$E\Big(\sum_{i=1}^{n} a_i \xi_i\Big) = \sum_{i=1}^{n} a_i\ E(\xi_i), \tag{18a}$$

$$D\Big(\sum_{i=1}^{n} a_i\xi_i\Big)=\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j D(\xi_i,\xi_j)=\sum_{i=1}^{n} a_i^2\ D(\xi_i)+2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} a_i a_j D(\xi_i,\xi_j). \tag{18b}$$

Podobně pro smíšený druhý centrální moment dvou lineárních kombinací vyjde vyjádření

$$D\Big(\sum_{i=1}^{n} a_i\xi_i,\ \sum_{i=1}^{n} b_i\xi_i\Big) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i b_j D(\xi_i,\xi_j). \tag{18c}$$

Disperze jsou kvadratickou formou koeficientů rozkladu a_i. Pokud jsou ξ_i vzájemně nekorelované (t.j. $\varrho_{ij}=0$ pro všechna $i\neq j$), zůstane v (18b) jen první člen na pravé straně:

$$D\Big(\sum_{i=1}^{n} a_i\xi_i\Big) = \sum_{i=1}^{n} a_i^2\ D(\xi_i). \tag{19}$$

Nalezení hustoty lineární kombinace z hustoty $f_{\xi}(x_1,\ldots,x_n)$ je trochu složitější. Elementární úvahou nebo využitím vztahu (2.5) zjistíme, že hustota konstantního násobku $a\xi$ náhodné proměnné ξ je násobkem hustoty $f_\xi(x)$:

$$g_{a\xi}(ax) = \frac{1}{a} f_\xi(x) = \frac{1}{a} f_\xi\Big(\frac{ax}{a}\Big). \tag{20}$$

Celý problém se tedy redukuje v podstatě na určení hustoty součtu dvou proměnných $\eta=\xi_1+\xi_2$. Distribuční funkce η je

$$F_\eta(y)= \iint\limits_{x_1+x_2<y} f_{\xi_1,\xi_2}(x_1,x_2)\,dx_1\,dx_2 = \int_{-\infty}^{\infty} dx_2 \int_{-\infty}^{y-x_2} f_{\xi_1,\xi_2}(x_1,x_2)\,dx_1. \tag{21}$$

Jsou-li ξ_1, ξ_2 nezávislé, je $f_{\xi_1,\xi_2}(x_1,x_2)=f_{\xi_1}(x_1)f_{\xi_2}(x_2)$; substitucí v (21) dostaneme

$$F_\zeta(y)=\int_{-\infty}^{\infty}dx_2\int_{-\infty}^{y}f_{\xi_2}(x_2)f_{\xi_1}(t-x_2)dt=\int_{-\infty}^{y}dt\left[\int_{-\infty}^{\infty}f_{\xi_2}(x_2)f_{\xi_1}(t-x_2)dx_2\right]. \qquad (22)$$

Podle vztahu (2.9) je tedy hustota součtu dvou nezávislých náhodných proměnných dána konvolucí

$$f_\zeta(y)=\int_{-\infty}^{\infty}f_{\xi_2}(u)f_{\xi_1}(y-u)du. \qquad (23)$$

Podíl nezávislých náhodných proměnných

Předpokládejme, že ξ_1, ξ_2 jsou nezávislé a mají hustoty $f_{\xi_1}(x_1)$, $f_{\xi_2}(x_2)$, resp. distribuční funkce $F_{\xi_1}(x_1)$, $F_{\xi_2}(x_2)$. Hustotu podílu $\eta=\xi_1/\xi_2$ najdeme pomocí distribuční funkce

$$F_\eta(y)=\iint_{x_1/x_2<y}f_{\xi_1}(x_1)f_{\xi_2}(x_2)dx_1dx_2=\int_{0}^{\infty}dx_2\int_{-\infty}^{yx_2}f_{\xi_1}(x_1)f_{\xi_2}(x_2)dx_1+$$

$$+\int_{-\infty}^{0}dx_2\int_{yx_2}^{\infty}f_{\xi_1}(x_1)f_{\xi_2}(x_2)dx_1=\int_{0}^{\infty}F_{\xi_1}(yx_2)f_{\xi_2}(x_2)dx_2+\int_{-\infty}^{0}\left[1-F_{\xi_1}(yx_2)\right]f_{\xi_2}(x_2)dx_2. \qquad (24)$$

Podle vztahu (2.9) dostaneme hledanou hustotu derivováním:

$$f_\eta(y)=\frac{d}{dy}F_\eta(y)=\int_{0}^{\infty}x_2f_{\xi_1}(yx_2)f_{\xi_2}(x_2)dx_2-\int_{-\infty}^{0}x_2f_{\xi_1}(yx_2)f_{\xi_2}(x_2)dx_2. \qquad (25)$$

Přibližná formule pro střední hodnotu a disperzi nelineární funkce

Střední hodnotu a disperzi funkce $h(\xi_1,\dots,\xi_n)$ náhodných proměnných můžeme aproximovat jednoduchými vztahy za předpokladu, že je průběh h v okolí středních hodnot $E(\xi_1),\dots,E(\xi_n)$ téměř lineární. V Taylorově rozvoji

$$h(\xi_1,\dots,\xi_m)=h\left[E(\xi_1),\dots,E(\xi_n)\right]+\sum_{i=1}^{n}\left[\xi_i-E(\xi_i)\right]\frac{\partial h}{\partial\xi_i}\Big|_E+\dots \qquad (26)$$

zachovávéme pouze uvedené dva členy; symbol E u derivací znamená, že jde o hodnoty v bodě $E(\xi_1),\dots,E(\xi_n)$. Střední hodnota druhého členu v (26) je nulová, proto

$$E\left[h(\xi_1,\dots,\xi_n)\right]\approx h\left[E(\xi_1),\dots,E(\xi_n)\right]. \qquad (27)$$

Pro disperzi funkce h dostáváme

$$D\left[h(\xi_1,\dots,\xi_n)\right]=E\left\{\left[h-E(h)\right]^2\right\}\approx E\left\{\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\xi_i-E(\xi_i)\right]\frac{\partial h}{\partial\xi_i}\Big|_E\cdot\left[\xi_j-E(\xi_j)\right]\frac{\partial h}{\partial\xi_j}\Big|_E\right\}=$$

$$=\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\partial h}{\partial\xi_i}\Big|_E\frac{\partial h}{\partial\xi_j}\Big|_E D(\xi_i,\xi_j). \qquad (28)$$

To je samozřejmě vztah (18b), jen na místě koeficientů a_i stojí derivace $\partial h/\partial \xi_i$. V rozvoji (26) jsme zachovali lineární kombinaci ξ_i a aditivní konstanty, které disperzi neovlivní. Úplně stejně odvodíme smíšený druhý moment dvou funkcí $h(\xi_1,\ldots,\xi_n)$ a $g(\xi_1,\ldots,\xi_n)$:

$$D(h,g)=E\left\{\left[h-E(h)\right]\left[g-E(g)\right]\right\} \approx \sum_{i=1}^{n}\sum_{j=1}^{n} \left.\frac{\partial h}{\partial \xi_i}\right|_E \left.\frac{\partial g}{\partial \xi_j}\right|_E D(\xi_i,\xi_j). \qquad (29)$$

Kvalita aproximací (27)-(29) záleží na tom, jak dobré je lineární přiblížení funkčních průběhů pomocí dvou členů Taylorova rozvoje v takové oblasti argumentů, která podstatně přispívá k disperzi. Velikost této oblasti záleží na tom, jak široká jsou rozdělení proměnných $\xi_1,\ldots,\xi_n$, tedy hlavně na jejich disperzích. Aproximace se v zásadě zlepšují při zmenšování druhých momentů $D(\xi_i,\xi_j)$.

Jsou-li ξ_i nekorelované, dostaneme z (28) a (29) jednodušší vztahy

$$D(h)\approx \sum_{i=1}^{n}\left(\left.\frac{\partial h}{\partial \xi_i}\right|_E\right)^2 D(\xi_i), \qquad D(h,g)\approx \sum_{i=1}^{n} \left.\frac{\partial h}{\partial \xi_i}\right|_E \left.\frac{\partial g}{\partial \xi_i}\right|_E D(\xi_i). \qquad (30)$$

První z relací (30), přepsané pro střední kvadratické odchylky $\sigma_h=\sqrt{D(h)}$, $\sigma_i=\sqrt{D(\xi_i)}$, se říká také (Gaussův) zákon pro přenos chyb:

$$\sigma_h \approx \sqrt{\left(\left.\frac{\partial h}{\partial \xi_1}\right|_E\right)^2 \sigma_1^2 + \ldots + \left(\left.\frac{\partial h}{\partial \xi_n}\right|_E\right)^2 \sigma_n^2}\ . \qquad (31)$$

Charakteristické funkce

Fourierova transformace hustoty nebo diskrétní funkce rozdělení se nazývá charakteristickou funkcí náhodné proměnné. Je to komplexní funkce reálné proměnné t:

$$\chi_\xi(t)=E\left[\exp(it\xi)\right] = \int_{-\infty}^{\infty} \exp(itx) f_\xi(x)\,dx \qquad (32)$$

pro spojitou proměnnou ξ s hustotou $f_\xi(x)$. Charakteristická funkce úplně popisuje náhodné proměnné; hustota je dána obrácenou transformací

$$f_\xi(x)= \frac{1}{2\pi}\int_{-\infty}^{\infty} \chi_\xi(t)\exp(-ixt)\,dt. \qquad (33)$$

Jsou-li a,b konstanty, platí

$$\chi_{a\xi+b}(t)=E\left\{\exp\left[it(a\xi+b)\right]\right\}=\exp(itb)\,\chi_\xi(at). \qquad (34)$$

Pro nezávislé ξ_1, ξ_2 dostaneme charakteristickou funkci součtu $\xi_1+\xi_2$ jako součin

$$\chi_{\xi_1+\xi_2}(t)=E\left\{\exp\left[it(\xi_1+\xi_2)\right]\right\} =E\left[\exp(it\xi_1)\right]\cdot E\left[\exp(it\xi_2)\right] =\chi_{\xi_1}(t)\cdot\chi_{\xi_2}(t). \qquad (35)$$

To je jeden z důvodů velké užitečnosti charakteristické funkce (namísto konvoluce hustot (23) máme jednoduchý součin charakteristických funkcí). Znalost $\chi_\xi(t)$ je užitečná i pro nalezení momentů (6) ze zápisu exponenty pomocí mocninné řady:

$$\chi_\xi(t)=E\left[\sum_{k=0}^{\infty}\frac{(it\xi)^k}{k!}\right]=\sum_{k=0}^{\infty}\frac{(it)^k}{k!}E(\xi^k)=\sum_{k=0}^{\infty}\frac{(i)^k}{k!}\nu_k t^k. \qquad (36)$$

Momenty ν_k jsou, až na faktor $i^k/k!$, rovny koeficientům u členů t^k v rozkladu $\chi(t)$ v mocninnou řadu.

4. Normální rozdělení

Spojitá náhodná proměnná, která nabývá libovolných reálných hodnot x s hustotou pravděpodobnosti

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right]. \qquad (1)$$

má tzv. normální, nebo Gaussovo-Laplaceovo, rozdělení. Kvůli stručnosti vyjadřování přestaneme v dalším textu odlišovat označení pro náhodnou proměnnou a její hodnoty, používané důsledně v §§ 2 a 3; budeme například říkat, že (1) je hustotou proměnné x.

Rozdělení (1) je zadáno dvěma reálnými parametry; μ může být libovolné, σ musí být kladné. μ je střední hodnota, σ střední kvadratická odchylka (σ^2 disperze):

$$E(x)=\mu,\quad D(x)=\sigma^2. \qquad (2)$$

Normální rozdělení je symetrické vzhledem ke střední hodnotě μ, která je zároveň mediánem i jedinou modou. Charakteristická funkce:

$$\chi(t)=\exp(i\mu t-t^2\sigma^2/2). \qquad (3)$$

Centrální momenty (3.7) lichého řádu jsou nulové, pro sudý řád vychází

$$\mu_{2k}=\frac{(2k)!}{2^k(k)!}\sigma^{2k},\quad k\geq 1. \qquad (4)$$

Asymetrie (3.8) i exces (3.9) jsou nulové: $\gamma_1=\gamma_2=0$.

Pro hustotu (1) se užívá značení $N(\mu,\sigma^2)$; její charakteristický "zvonový" průběh je pro tři různé disperze σ^2 nakreslen v obr. 4. Distribuční funkce je

$$F(x)=\Phi\left(\frac{x-\mu}{\sigma}\right),\quad \text{kde}\quad \Phi(z)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}\exp(-\frac{t^2}{2})dt. \qquad (5)$$

Funkci $\Phi(z)$ se říká integrál pravděpodobnosti nebo funkce chyb. F(x) pro tři různé disperze je nakreslena v obr. 5. Pomocí distribuční funkce (§2) můžeme vyjádřit pravděpodobmost, že hodnota x padne do zadaného intervalu:

To je jeden z důvodů velké užitečnosti charakteristické funkce (namísto konvoluce hustot (23) máme jednoduchý součin charakteristických funkcí). Znalost $\chi_\xi(t)$ je užitečná i pro nalezení momentů (6) ze zápisu exponenty pomocí mocninné řady:

$$\chi_\xi(t)=E\left[\sum_{k=0}^{\infty}\frac{(it\xi)^k}{k!}\right]=\sum_{k=0}^{\infty}\frac{(it)^k}{k!}E(\xi^k)=\sum_{k=0}^{\infty}\frac{(i)^k}{k!}\nu_k t^k. \qquad (36)$$

Momenty ν_k jsou, až na faktor $i^k/k!$, rovny koeficientům u členů t^k v rozkladu $\chi(t)$ v mocninnou řadu.

4. Normální rozdělení

Spojitá náhodná proměnná, která nabývá libovolných reálných hodnot x s hustotou pravděpodobnosti

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right]. \qquad (1)$$

má tzv. normální, nebo Gaussovo-Laplaceovo, rozdělení. Kvůli stručnosti vyjadřování přestaneme v dalším textu odlišovat označení pro náhodnou proměnnou a její hodnoty, používané důsledně v §§ 2 a 3; budeme například říkat, že (1) je hustotou proměnné x.

Rozdělení (1) je zadáno dvěma reálnými parametry; μ může být libovolné, σ musí být kladné. μ je střední hodnota, σ střední kvadratická odchylka (σ^2 disperze):

$$E(x)=\mu,\ D(x)=\sigma^2. \qquad (2)$$

Normální rozdělení je symetrické vzhledem ke střední hodnotě μ, která je zároveň mediánem i jedinou modou. Charakteristická funkce:

$$\chi(t)=\exp(i\mu t - t^2\sigma^2/2). \qquad (3)$$

Centrální momenty (3.7) lichého řádu jsou nulové, pro sudý řád vychází

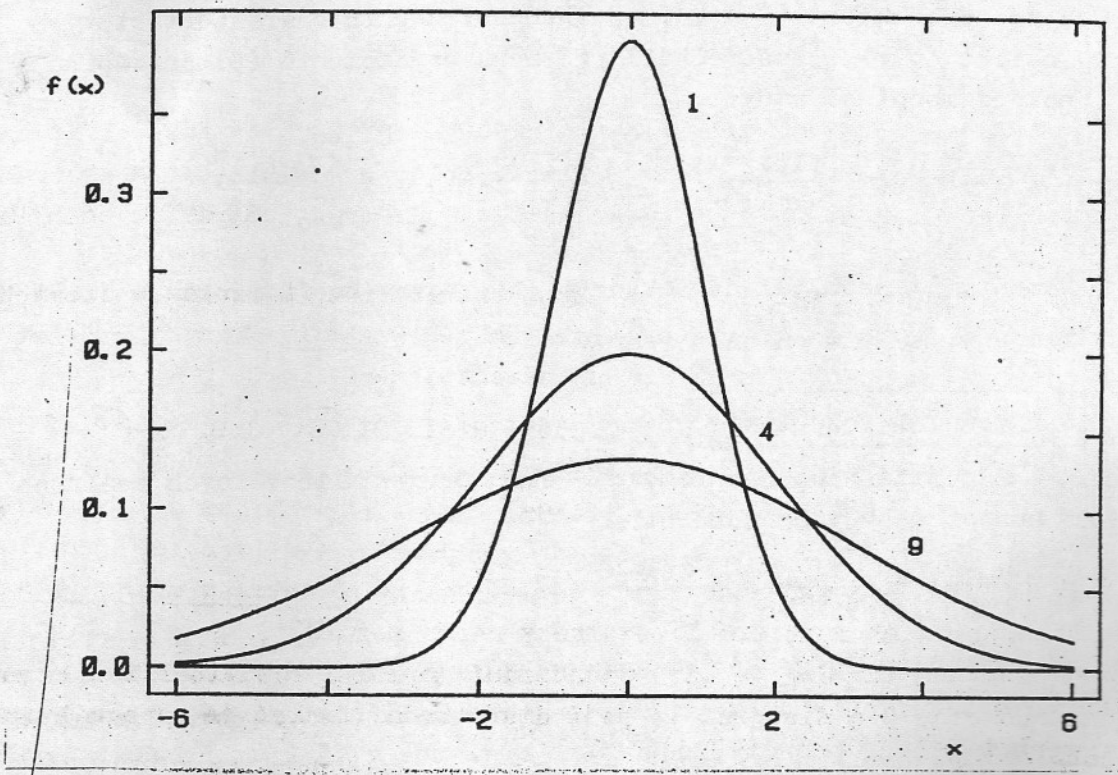
$$\mu_{2k}=\frac{(2k)!}{2^k(k)!}\sigma^{2k},\quad k\geq 1. \qquad (4)$$

Asymetrie (3.8) i exces (3.9) jsou nulové: $\gamma_1=\gamma_2=0$.

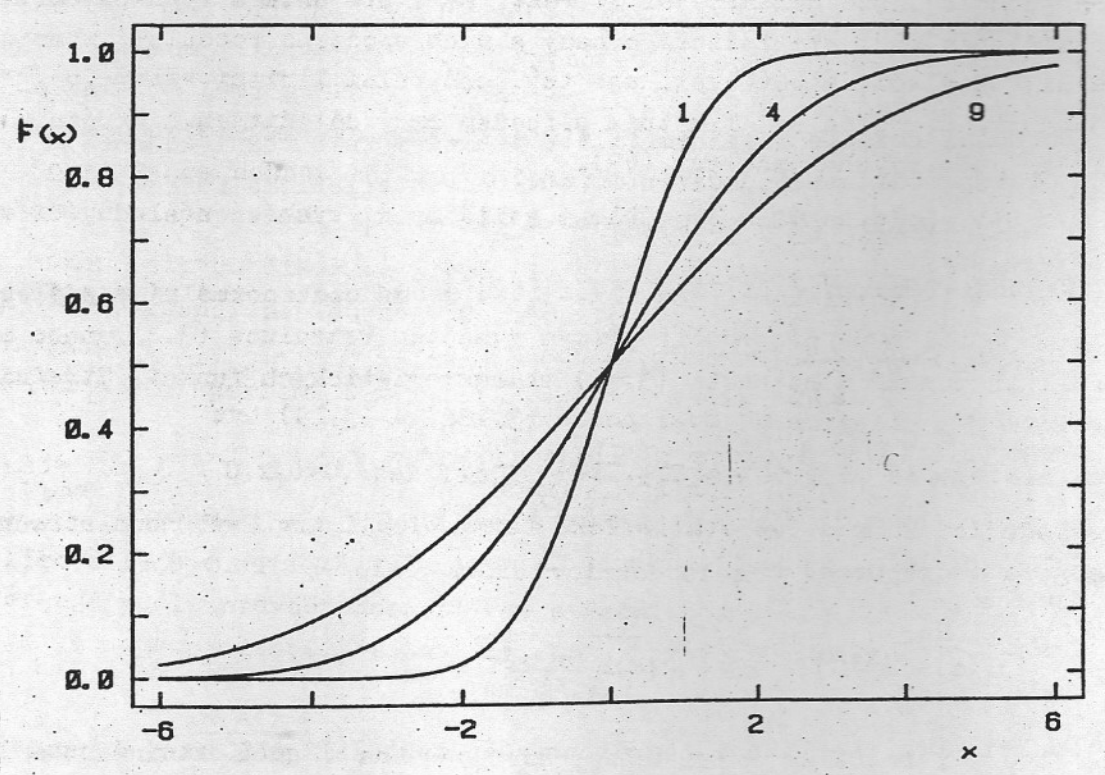
Pro hustotu (1) se užívá značení $N(\mu,\sigma^2)$; její charakteristický "zvonový" průběh je pro tři různé disperze σ^2 nakreslen v obr. 4. Distribuční funkce je

$$F(x)=\Phi\left(\frac{x-\mu}{\sigma}\right),\quad \text{kde}\quad \Phi(z)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}\exp(-\frac{t^2}{2})dt. \qquad (5)$$

Funkci $\Phi(z)$ se říká integrál pravděpodobnosti nebo funkce chyb. F(x) pro tři různé disperze je nakreslena v obr. 5. Pomocí distribuční funkce (§2) můžeme vyjádřit pravděpodobmost, že hodnota x padne do zadaného intervalu:



Obr. 4. Hustota normálního rozdělení se střední hodnotou $\mu=0$ a disperzemi $\sigma^2=1,4,9$.



Obr. 5. Distribuční funkce normálního rozdělení s $\mu=0$ a $\sigma^2=1,4,9$.

$P(x\in\langle x_1,x_2\rangle)=F(x_2)-F(x_1)$. Vyčíslením integrálu pravděpodobnosti (5) zjistíme, že

$$P(\mu-\sigma\leq x\leq\mu+\sigma)=0.683, \quad P(\mu-2\sigma\leq x\leq\mu+2\sigma)=0.954. \tag{6}$$

Střední kvadratické odchylce σ se v případě normálního rozdělení říká také standardní odchylka. Intervaly $\mu\pm\sigma$ a $\mu\pm2\sigma$ s pravděpodobnostním obsahem (6) se pak označují jako intervaly s jednou a dvěma standardními odchylkami.

Velmi potřebná funkce chyb (5) byla mnohokrát tabelována (v různých modifikacích). Užitečné jsou různé aproximace, které umožňují vypočítat dostatečně přesné hodnoty s minimální námahou, např.

$$\Phi(z)\approx 1-\frac{\exp(-z^2/2)}{\sqrt{2\pi}}t\,(0.3193815-0.3565638t+1.781478t^2-1.821256t^3+$$
$$+1.330274t^4), \quad t=1/(1+0.2316419z) \text{ pro } z\geq 0, \tag{7}$$

$$\Phi(z)=1-\Phi(|z|) \text{ pro } z<0;$$

chyba této aproximace je pro libovolné z menší než 10^{-7}.

Náhodná proměnná $(x-\mu)/\sigma$ má tzv. standardní normální rozdělení $N(0,1)$ se střední hodnotou 0 a disperzí 1; její distribuční funkcí je integrál pravděpodobnosti Φ.

Normální rozdělení má při zpracování výsledků měření podstatnou důležitost. Především v mnoha situacích velmi dobře vystihuje rozložení naměřených hodnot. Ještě důležitější je fakt, že i pro data s výrazně odlišným rozdělením mají statistické odhady z nich spočtené rozdělení zhruba normální; tuto souvislost vystihuje tzv. centrální limitní věta (§5). Navíc je normální rozdělení limitním případem řady důležitých diskrétních i spojitých modelových rozdělení (§9).

Pro nezávislé normálně rozdělené veličiny x_i vychází následující důležité výsledky:

a) Libovolná lineární kombinace $a_1x_1+\ldots+a_nx_n$ má opět normální rozdělení. O tom je možné se přesvědčit přímým výpočtem konvoluce (3.23) nebo mnohem lépe pomocí vlastnosti (3.35) charakteristických funkcí. Střední hodnota μ a disperze σ^2 musí podle (3.18a) a (3.19) být

$$\mu=a_1\mu_1+\ldots+a_n\mu_n, \quad \sigma^2=a_1^2\sigma_1^2+\ldots+a_n^2\sigma_n^2. \tag{8}$$

b) Následující funkce (ve statistické terminologii tzv. výběrová střední hodnota a disperze)

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i, \quad s^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

jsou nezávislé tehdy a jen tehdy, když všechna x_i mají stejné normální rozdělení (s týmiž μ,σ).

5. Zákon velkých čísel a centrální limitní věta

Zákon velkých čísel

Souvislost pravděpodobnosti p náhodného jevu X a četnosti M jeho výskytu v N-krát opakovaném pokusu je intuitivně zřejmá: čekáme, že se relativní četnost M/N bude s rostoucím N přibližovat k pravděpodobnosti p. Protože podíl M/N je náhodná veličina, je třeba pro očekávané přibližování k hodnotě p formulovat pravděpodobnostní tvrzení. Nejznámější je Bernoulliův teorém; pro každé $\varepsilon>0$ platí

$$\lim_{N\to\infty} P\left(\left|\frac{M}{N} - p\right|<\varepsilon\right) = 1 \,. \qquad (1)$$

Vyjádřeno slovy: ať zvolíme $\varepsilon>0$ jakkoli malé, s pravděpodobností libovolně blízkou k jedné jsou při dostatečně velkém počtu pokusů odchylky poměrných četností M/N od hodnoty p menší než ε. Formuli (1) se říká (slabý) zákon velkých čísel. Tvrzení

$$P\left(\lim_{N\to\infty} \frac{M}{N} = p\right) = 1 \qquad (2)$$

je silnější (z (2) plyne (1), ale ne naopak); objevil je Borel a říká se mu silný zákon velkých čísel.

Pro přívržence statistické definice pravděpodobnosti (§1) je zákon velkých čísel tautologií, protože pravděpodobnost určují právě z relativních četností při opakování pokusu. Pro zastánce názoru, že se pravděpodobnosti dají (alespoň někdy) vypočíst ze struktury jevů, je předpověď četností otázkou logické výstavby teorie a věty o konvergenci posloupnosti M/N jsou podstatnými výsledky. Slabý a silný zákon vyjadřují dva různé typy konvergence hodnot poměrných četností: tzv. konvergenci podle pravděpodobnosti, popsanou vztahem (1) a konvergenci téměř jistě (2)).

Jako zákon velkých čísel se kromě (1) a (2) označují také následující věty o konvergenci posloupnosti aritmetických průměrů náhodných proměnných. Jsou-li $x_1, x_2, \ldots$ nezávislé náhodné proměnné se stejnou střední hodnotou μ a disperzemi $D(x_1), D(x_2), \ldots$ takovými, že

$$\lim_{N\to\infty} \frac{1}{N^2} \sum_{i=1}^{N} D(x_i)=0, \text{ pak } \lim_{N} P\left(\left|\frac{1}{N} \sum_{i=1}^{N} x_i - \mu\right|<\varepsilon\right)=1 \qquad (3)$$

pro libovolné $\varepsilon>0$. To je slabý zákon velkých čísel - posloupnost průměrů konverguje podle pravděpodobnosti ke střední hodnotě. Silný zákon tvrdí, že pro náhodné proměnné, jejichž disperze splňují podmínku

$$\lim_{N\to\infty} \left[\sum_{i=1}^{N} \frac{D(x_i)}{i^2}\right] < \infty, \qquad (4)$$

konverguje průměr ke střední hodnotě téměř jistě:

$$P\left[\lim_{N\to\infty}\left(\frac{1}{N}\sum_{i=1}^{N}x_i\right)=\mu\right]=1. \qquad (5)$$

Obě věty (3) a (5) se dají zobecnit pro případ posloupnosti proměnných s různými středními hodnotami $E(x_1)$, $E(x_2)$,... , Jejich aritmetický průměr konverguje k limitě průměru středních hodnot $[\sum E(x_i)]/N$.

Centrální limitní věta

udává, jaké je v limitě rozdělení aritmetického průměru nezávislých náhodných proměnných $x_1, x_2, \ldots$, které mají stejnou distribuční funkci se střední hodnotou μ a disperzí σ^2. Podle (3.18a) a (3.19) je střední hodnota průměru $\bar{x}_N=(x_1+\ldots+x_N)/N$ rovna $E(\bar{x}_N)=\mu$ a jeho disperze $D(\bar{x}_N)=\sigma^2/N$; hustotu pravděpodobnosti $\bar{x}_N$ dostaneme z (3.23) komplikovaným způsobem, totiž N-1-krát opakovanou konvolucí.

S pomocí vlastnosti (3.35) charakteristické funkce však snadno zjistíme, že charakteristická funkce $\chi(\bar{x}_N)$ se pro $N\to\infty$ blíží k charakteristické funkci (4,3) normálního rozdělení. Hustota průměru je tedy

$$f(\bar{x}_N) \xrightarrow{N\to\infty} \frac{1}{\sqrt{2\pi\sigma^2/N}}\exp\left[\frac{-(\bar{x}_N-\mu)^2}{2\sigma^2/\sqrt{N}}\right]; \qquad (6)$$

rozdělení sčítanců x_i může být jakékoliv, stačí když má konečnou disperzi σ^2. To je tzv. Lindbergův-Lévyův teorém, nebo jedna z variant centrální limitní věty. Jiná varianta platí pro posloupnost nezávislých náhodných proměnných $x_1, x_2, \ldots$ se středními hodnotami $\mu_1, \mu_2, \ldots$ a disperzemi $\sigma_1^2, \sigma_2^2, \ldots$, které nemusí mít stejnou distribuční funkci. Rozdělení aritmetického průměru je v limitě $N\to\infty$ opět normální; veličina

$$\left(\sum_{i=1}^{N}x_i-\sum_{i=1}^{N}\mu_i\right)\Big/\sqrt{\sum_{i=1}^{N}\sigma_i^2} \qquad (7)$$

má asymptoticky rozdělení N(0,1). K platnosti tohoto tvrzení stačí, aby střední hodnoty μ_i a disperze σ_i^2 existovaly a nerostly příliš rychle s rostoucím i. Postačující je například splnění Ljapunovovy podmínky: existuje takové $a>0$, že

$$\lim_{N\to\infty}\left[\sum_{i=1}^{N}E(x_i-\mu_i)^{2+a}\right]\Big/\left(\sum_{i=1}^{N}\sigma_i^2\right)^{2+a}=0. \qquad (8)$$

Příklad: součet rovnoměrně rozdělených náhodných čísel

Konvergenci součtu nezávislých náhodných proměnných k normálnímu rozdělení budeme ilustrovat na příkladu rovnoměrně rozdělených (§9) veličin x_i s hustotou

$$f(x_i) = 1, \quad x_i \in \langle 0,1 \rangle . \tag{9}$$

Hustota aritmetického průměru $\overline{x}_N = s_N/N$, kde $s_N = x_1+\ldots+x_N$, se dá vyjádřit analyticky; je to po částech polynom stupně N-1:

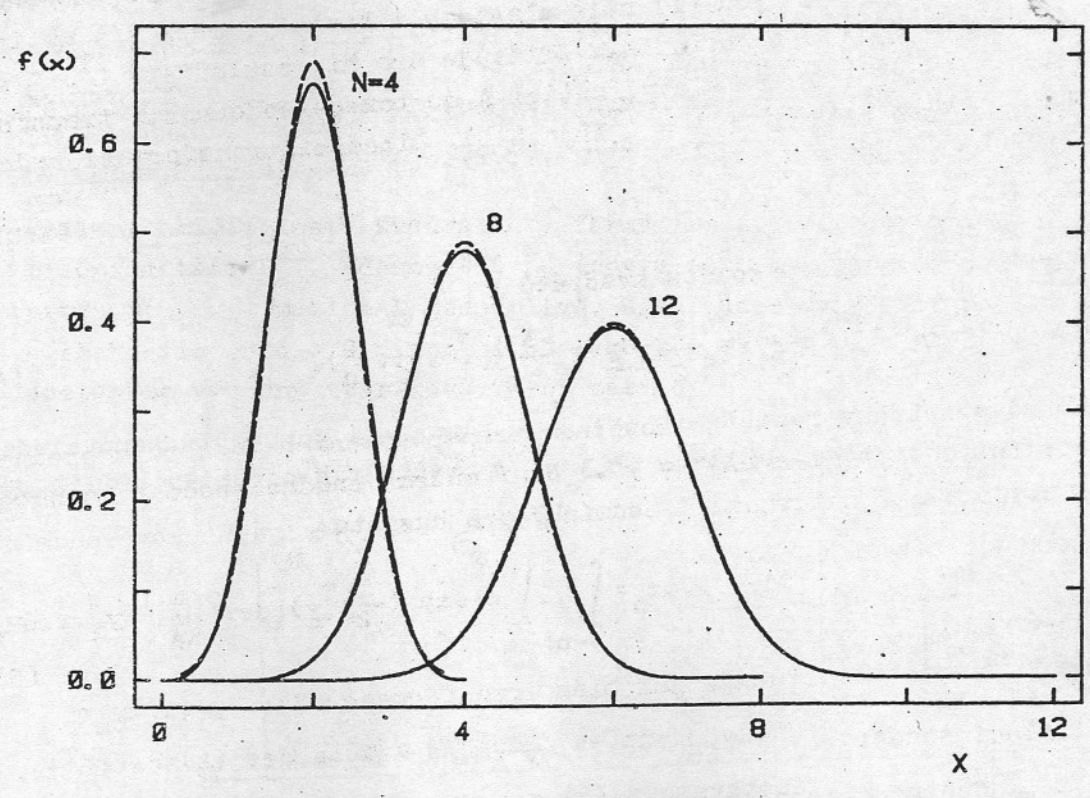
$$f(\overline{x}_N) = \frac{N^N}{(N-1)!} \sum_{i=0}^{k} (-1)^i \binom{N}{i} (\overline{x}_N - \frac{i}{N})^{N-1} , \tag{10}$$

$$\overline{x}_N \in \left\langle \frac{k}{N}, \frac{k+1}{N} \right\rangle , \quad k=0,\ldots,N-1.$$

Střední hodnotu a disperzi součtu s_N můžeme vypočíst mnohem snáze než hustotu; s pomocí (9.6) dostaneme

$$E(s_N) = NE(x_i) = N/2, \quad D(s_N) = ND(x_i) = N/12. \tag{11}$$

Srovnání hustoty součtu s_N a normální hustoty se stejnou střední hodnotou a disperzí je v obr. 6. Je vidět, že konvergence k normálnímu rozdělení je velmi rychlá, veličina $s_{12}-6$ má prakticky standardní normální rozdělení.



Obr. 6. Hustota pravděpodobnosti součtu N nezávislých rovnoměrně rozdělených náhodných proměnných (plná čára), normální hustota se stejnou střední hodnotou a disperzí (podle vztahu (11), čárkovaná čára).

6. Vícerozměrné normální rozdělení

Zobecnění normálního rozdělení (4.1) na případ n-rozměrného náhodného vektoru vychází z požadavku, aby hustota byla úměrná exponenciální funkci kvadratické formy jednotlivých složek:

$$f(x_1,\ldots,x_n) = c.\exp\left[-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}(x_i-\mu_i)(x_j-\mu_j)\right]. \qquad (1)$$

Konečný tvar této hustoty odvodíme podrobněji, protože se přitom ukáže řada užitečných souvislostí. Kvůli přehlednosti zápisu zavedeme následující konvence: čtvercovou matici budeme značit velkým písmenem s podloženou vlnovkou, sloupcové vektory podtržením, transpozici libovolné matice horním indexem T (například sloupcový vektor se složkami μ_i jako $\underset{\sim}{\mu}$, řádkový vektor $\underset{\sim}{\mu}^T = (\mu_1,\ldots,\mu_n)$). Argument exponenty v (1) tedy stručně zapíšeme jako

$$-\frac{1}{2}\,(\underline{x}-\underline{\mu})^T\,\underset{\sim}{A}\,(\underline{x}-\underline{\mu}), \qquad (2)$$

když matice $\underset{\sim}{A}$ má prvky $(\underset{\sim}{A})_{ij} = a_{ij}$. Hořejší výraz je skalár, protože součin matice $\underset{\sim}{A}$ se sloupcovým vektorem $\underline{x}-\underline{\mu}$ je sloupcový vektor, ze kterého vyjde násobením řádkovým vektorem $(\underline{x}-\underline{\mu})^T$ skalární hodnota dvojnásobné sumy v (1). Aby funkce typu (1) měla vlastnosti hustoty, musí být matice $\underset{\sim}{A}$ symetrická a pozitivně definitní. Existuje pro ni rozklad $\underset{\sim}{A} = \underset{\sim}{L}^T.\,\underset{\sim}{L}$ s regulární maticí $\underset{\sim}{L}$ (jinými slovy, matici $\underset{\sim}{A}$ dostaneme podobnostní transformací z jednotkové matice $\underset{\sim}{I}$: $\underset{\sim}{A} = \underset{\sim}{L}^T\underset{\sim}{I}\underset{\sim}{L} = \underset{\sim}{L}^T\underset{\sim}{L}$). Lineární transformací hodnot náhodného vektoru $\underline{x}$

$$\underline{y} = \underset{\sim}{L}(\underline{x}-\underline{\mu}) \qquad (3)$$

dostaneme (2) ve tvaru součtu kvadrátů

$$-\frac{1}{2}\,(\underline{x}-\underline{\mu})^T\,\underset{\sim}{L}^T\underset{\sim}{L}(\underline{x}-\underline{\mu}) = -\frac{1}{2}\,\underline{y}^T\underline{y} = -\frac{1}{2}\,(y_1^2+\ldots+y_n^2). \qquad (4)$$

(Transpozice součinu matic je součinem transponovaných faktorů v opačném pořadí: $(\underline{x}-\underline{\mu})^T\,\underset{\sim}{L}^T = \left[\underset{\sim}{L}(\underline{x}-\underline{\mu})\right]^T = \underline{y}^T$.) Nyní můžeme snadno spočíst hodnotu konstanty c v (1) z normovací podmínky pro hustotu:

$$\int_{-\infty}^{\infty}\!\!\ldots\int f(x_1,\ldots,x_n)\,dx_1\ldots dx_n = \int_{-\infty}^{\infty}\!\!\ldots\int c\cdot\exp\left(-\frac{1}{2}\underline{y}^T\underline{y}\right)\left[\det(\underset{\sim}{L})\right]^{-1}dy_1\ldots dy_n = 1. \qquad (5)$$

V posledním vztahu vystupuje jakobián transformace (3): $dy_1\ldots dy_n = \det(\underset{\sim}{L})\,dx_1\ldots dx_n$. Protože $\int_{-\infty}^{\infty}\exp(-t^2/2)\,dt = \sqrt{2\pi}$ a dále $\det(\underset{\sim}{A}) = \det(\underset{\sim}{L}^T)\cdot\det(\underset{\sim}{L}) = \left[\det(\underset{\sim}{L})\right]^2$, dostáváme z (5) pro konstantu c vztah

$$c = \frac{\det(\underset{\sim}{L})}{(\sqrt{2\pi})^n} = \frac{1}{\sqrt{(2\pi)^n\,\det(\underset{\sim}{A}^{-1})}}. \qquad (6)$$

Je třeba si všimnout faktu, že $\underline{y}$ ze vztahu (3) je náhodným vektorem s nezávislými komponentami, které mají standardní normální rozdělení (nulové střední hodnoty, jednotkové disperze a nulové korelační koeficienty). Zobecněním postupu z § 2 pro určení hustoty funkce náhodné proměnné na případ vektorů totiž zjistíme, že vztah (2.5) zůstane zachován, jen na místě $|h'(x)|$ se objeví jakobián transformace. Z hustoty (1) tedy dostaneme s využitím (6) hustotu vektoru $\underline{y}$

$$g(\underline{y}) = \frac{f(\underline{x})}{\det(\underset{\sim}{L})} = \frac{1}{\sqrt{(2\pi)^n}} \exp(-\frac{1}{2}\underline{y}^T\underline{y}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp(-\frac{y_i^2}{2}) \qquad (7)$$

ve tvaru součinu standardních normálních hustot jednotlivých komponent. Vektor středních hodnot $\underline{y}$ je nulový, což zapíšeme symbolicky jako $E(\underline{y})=\underline{0}$. Podle (3) můžeme do tohoto vztahu dosadit $\underline{y}=\underset{\sim}{L}(\underline{x}-\underline{\mu})$; protože podle (3.18a) je E lineární operátor a $\underset{\sim}{L}$ je regulární, dostaneme odtud $E(\underline{x}-\underline{\mu})=\underline{0}$ a tedy střední hodnoty vektoru $\underline{x}$:

$$E(\underline{x}) = \underline{\mu}. \qquad (8)$$

Matice druhých momentů vektoru $\underline{y}$ je jednotková, symbolicky $E(\underline{y}\underline{y}^T)=\underset{\sim}{I}$; součin sloupcového a řádkového vektoru $\underline{y}\underline{y}^T$ tvoří čtvercovou matici nxn a funkcionál E působí na každý její prvek zvlášť. Dosazením za $\underline{y}$ z (3) a využitím linearity E dostaneme

$$E\left[\underset{\sim}{L}(\underline{x}-\underline{\mu})(\underline{x}-\underline{\mu})^T\underset{\sim}{L}^T\right] = \underset{\sim}{L}E\left[(\underline{x}-\underline{\mu})(\underline{x}-\underline{\mu})^T\right]\underset{\sim}{L}^T = \underset{\sim}{I},$$

neboli pro matici $\underset{\sim}{D}$ druhých momentů vektoru $\underline{x}$ vztah

$$\underset{\sim}{D} = E\left[(\underline{x}-\underline{\mu})(\underline{x}-\underline{\mu})^T\right] = \underset{\sim}{L}^{-1}(\underset{\sim}{L}^T)^{-1} = (\underset{\sim}{L}^T\underset{\sim}{L})^{-1} = \underset{\sim}{A}^{-1}. \qquad (9)$$

Matice $\underset{\sim}{A}$ koeficientů kvadratické formy (2) je tedy rovna inverzní kovariační matici $\underset{\sim}{D}^{-1}$. Kdo nevěří maticovým zápisům, může se pokusit vypočíst prvky $D(x_i,x_j)$ matice $\underset{\sim}{D}$ jednotlivě. Kovariance vyjádříme pomocí korelačních koeficientů ρ a disperzí σ^2 jako $D(x_i,x_j) = \rho_{ij}\sigma_i\sigma_j$ (viz (3.16)); dostaneme výhodný tvar kovariační matice

$$\underset{\sim}{D} = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1n}\sigma_1\sigma_n \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & & \\ \vdots & & \ddots & \vdots \\ \rho_{1n}\sigma_1\sigma_n & \rho_{2n}\sigma_2\sigma_n & \cdots & \sigma_n^2 \end{bmatrix} \qquad (10)$$

Libovolné n-rozměrné normální rozdělení může být zadáno n-ticí středních hodnot $\underline{\mu}$ a $n(n+1)/2$ nezávislými prvky symetrické matice - buď $\underset{\sim}{A}$ nebo $\underset{\sim}{D}$. Vyjádření hustoty (1) s normalizační konstantou (6) pomocí matice $\underset{\sim}{D}$ je

$$f(x) = \frac{1}{\sqrt{(2\pi)^n \det(\underset{\sim}{D})}} \exp\left[-\frac{1}{2}(\underline{x}-\underline{\mu})^T \underset{\sim}{D}^{-1}(\underline{x}-\underline{\mu})\right]. \qquad (11)$$

Veličina

$$\zeta = (\underline{x}-\underline{\mu})^T \underset{\sim}{D}^{-1}(\underline{x}-\underline{\mu}) = (\underline{x}-\underline{\mu})^T \underset{\sim}{A}(\underline{x}-\underline{\mu}) \qquad (12)$$

se nazývá kovariační formou náhodného vektoru $\underline{x}$. Je to jednorozměrná náhodná proměnná s rozdělením χ^2 s n stupni volnosti (§ 8), neboť se dá napsat jako součet kvadrátů n-tice nezávislých proměnných se standardním normálním rozdělením... $\zeta = \underline{y}^T\underline{y}$. Hustota (11) je konstantní na plochách ζ = konst. a pravděpodobnostní obsah těchto elipsoidů (pravděpodobnost, že $\underline{x}$ padne dovnitř elipsoidu) je dána distribuční funkcí χ^2_n.

Z pozoruhodných vlastností rozdělení (11) uvedeme dvě:

a) Libovolná projekce na prostor menší dimenze (marginální rozdělení, §2) je opět normální s maticí druhých momentů sestavenou z prvků matice (10) odpovídajících zbylým proměnným. Například marginální rozdělení každé komponenty x_i je

$$f(x_i) = N(\mu_i, \sigma_i^2). \qquad (13)$$

b) Libovolný řez (podmíněné rozdělení, §2) je opět normální. Řez rovinou $x_i = x_i^{(o)}$, t.j. rozdělení s konstantní hodnotou $x_i^{(o)}$ složky x_i, má matici druhých momentů $\underset{\sim}{D}_{n-1}$, kterou dostaneme inverzí matice $\underset{\sim}{A}_{n-1}$, kovariační formy zbylých proměnných.

Dvojrozměrné normální rozdělení

Pro dvojrozměrné (n=2) rozdělení můžeme snadno vyjádřit explicitně prvky matice $\underset{\sim}{D}^{-1}$; hustota (11), zapsaná pomocí středních hodnot μ_1, μ_2, standardních odchylek σ_1, σ_2 a korelačního koeficientu ρ má tvar

$$f(x_1,x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right]\right\}. \qquad (14)$$

Můžeme si ji představit názorně jako zvonovitou plochu nad rovinou x_1, x_2, nebo při pohledu shora znázornit soustavu vrstevnic - čar s konstantní funkční hodnotou. Vrstevnicemi jsou elipsy

$$\frac{1}{1-\rho^2}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right] = \lambda = \text{konst.} \qquad (15)$$

Protože kovariační forma na levé straně (15) má známé rozdělení (χ^2 se dvěma stupni volnosti), můžeme vypočíst pravděpodobnost, že dvojice x_1, x_2 leží uvnitř elipsy (15):

$$P = F_{\chi^2_2}(\lambda) ;$$

$F_{\chi^2_2}$ je příslušná distribuční funkce. V obrázku 7 je nakresleno několik elips s různými pravděpodobnostními obsahy P=0.99, 0.954, 0.683, 0.5, 0.2, pro které podle tabulky v dodatku D1 vychází hodnoty λ po řadě 9.21, 6.158, 2.298, 1.386, 0.446; korelační koeficient je $\varrho = -3/4$.

Z hustoty (14) odvodíme podmíněné rozdělení x_1 za předpokladu, že x_2 nabývá pevné hodnoty (viz (2.15)):

$$g(x_1)=f(x_1|x_2) = \frac{1}{\sqrt{2\pi}\,\sigma_1\sqrt{1-\varrho^2}} \exp\left\{\frac{-1}{2(1-\varrho^2)\sigma_1^2}\left[x_1-\mu_1-\varrho\frac{\sigma_1}{\sigma_2}(x_2-\mu_2)\right]^2\right\}. \quad (16)$$

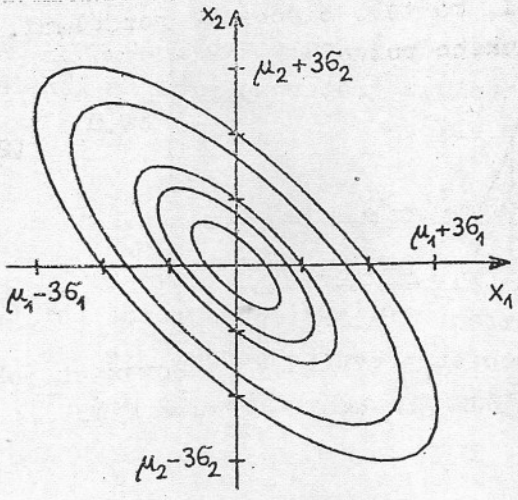
Je to normální rozdělení se střední hodnotou a disperzí

$$E(x_1|x_2) = \mu_1+\varrho\frac{\sigma_1}{\sigma_2}(x_2-\mu_2) , \quad D(x_1|x_2)= \sigma_1^2\,(1-\varrho^2). \quad (17)$$

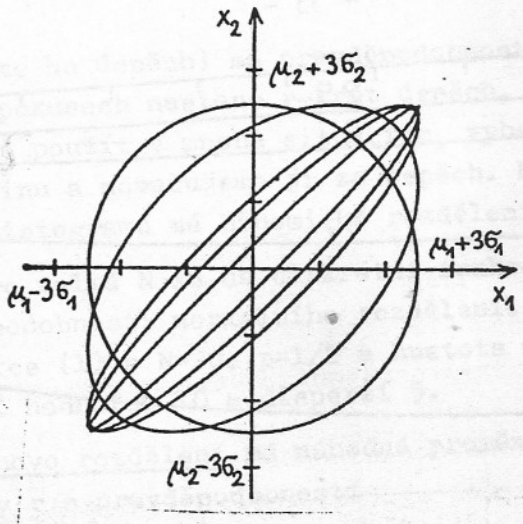
V tomto místě máme dobrou příležitost ilustrovat smysl pojmů závislost a korelace náhodných proměnných. Normálně rozdělené proměnné jsou nezávislé právě tehdy, když jsou nekorelované; je vidět, že hustota (14) je pro $\varrho = 0$ součinem hustot $N(\mu_1, \sigma_1^2)$ a $N(\mu_2,\sigma_2^2)$. Jinými slovy, rozdělení každé proměnné je nezávislé na tom, jakou hodnotu nabývá druhá z nich. Obecně (pro jiná rozdělení) je ovšem nekorelovanost slabší než nezávislost (§3). Je-li korelační koeficient různý od nuly, záleží podle (17) rozdělení x_1 na tom, jaké hodnoty nabývé x_2; při $|\varrho|\to 1$ se zužuje kolem střední hodnoty závislé na x_2. V limitním případě úplné korelace ($\varrho = 1$) nabývají náhodné proměnné x_1, x_2 hodnot, které spolu souvisí vztahem

$$(x_1-\mu_1)/\sigma_1 = (x_2-\mu_2)/\sigma_2 . \quad (18)$$

Míru závislosti ϱ obou proměnných můžeme znázornit ještě jinak. V obrázku 8 jsou nakresleny elipsy konstantní hustoty, které mají stejný pravděpodobnostní obsah P=0.954 a liší se hodnotou korelačního koeficientu. S rostoucím ϱ



Obr. 7. Elipsy (15) normálního rozdělení s korelačním koeficientem $\varrho = -3/4$. Pravděpodobnostní obsah je, po řadě od největší k nejmenší, roven 0.99, 0.954, 0.683, 0.5, 0.2.



Obr. 8. Elipsy s pravděpodobnostním obsahem 0.954 a různými korelačními koeficienty $\varrho = 0$ (kruh), 0.5, 0.9, 0.95, 0.99 (nejužší elipsa).

se oblast, do které bod x_1, x_2 padne s velkou pravděpodobností, zužuje. Pro $\varrho = 1$ zdegeneruje elipsa v úsečku a x_1 je lineární funkcí x_2 podle vztahu (18).

Pravděpodobnostní obsah eliptických oblastí počítáme jednoduše pomocí distribuční funkce χ^2 - rozdělení. Užitečným údajem je pravděpodobnostní obsah obdélníků $x_1 \in \langle \mu_1-k\sigma_1, \mu_1+k\sigma_1 \rangle$, $x_2 \in \langle \mu_2-k\sigma_2, \mu_2+k\sigma_2 \rangle$ pro zadané násobky k standardních odchylek; k jeho vyčíslení je třeba integrovat funkci chyb (4.5):

$$P = \int_{\mu_1-k\sigma_1}^{\mu_1+k\sigma_1} dx_1 \int_{\mu_2-k\sigma_2}^{\mu_2+k\sigma_2} f(x_1,x_2)\,dx_2 = \sqrt{\frac{2}{\pi}} \int_0^k \left[\Phi\left(\frac{k-\varrho t}{\sqrt{1-\varrho^2}}\right) - \Phi\left(\frac{-k-\varrho t}{\sqrt{1-\varrho^2}}\right) \right] \exp\left(-\frac{t^2}{2}\right) dt. \qquad (19)$$

Závislost P na korelačním koeficientu pro několik hodnot k je nakreslena v obr. 9.

7. Binomické a Poissonovo rozdělení

Binomické rozdělení. Diskrétní náhodná proměnná, která nabývá celé nezáporné hodnoty r s pravděpodobností

$$P(r) = \binom{N}{r} p^r (1-p)^{N-r}, \quad r = 0,1,\ldots,N, \qquad (1)$$

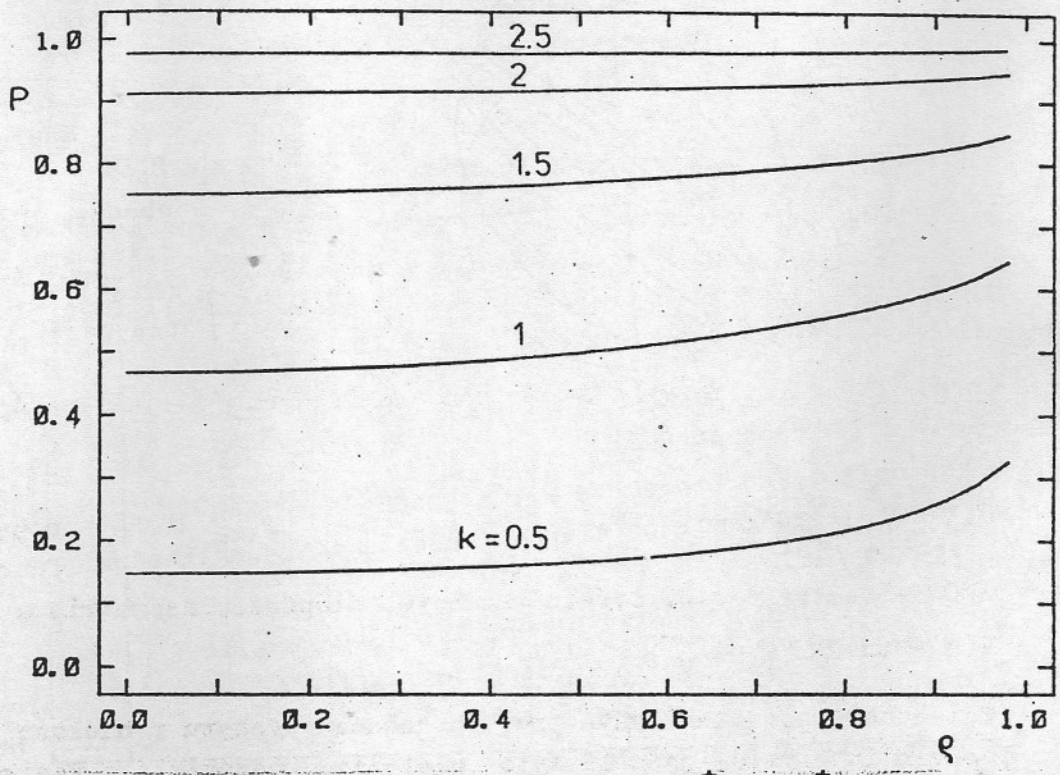
kde N je celé kladné, p reálné, $0 \leq p \leq 1$, má tzv. binomické rozdělení. Pravděpodobnosti (1) jsou členy binomického rozvoje

$$(p+q)^N = \sum_{r=0}^{N} \frac{N!}{r!(N-r)!} p^r q^{N-r} \qquad (2)$$

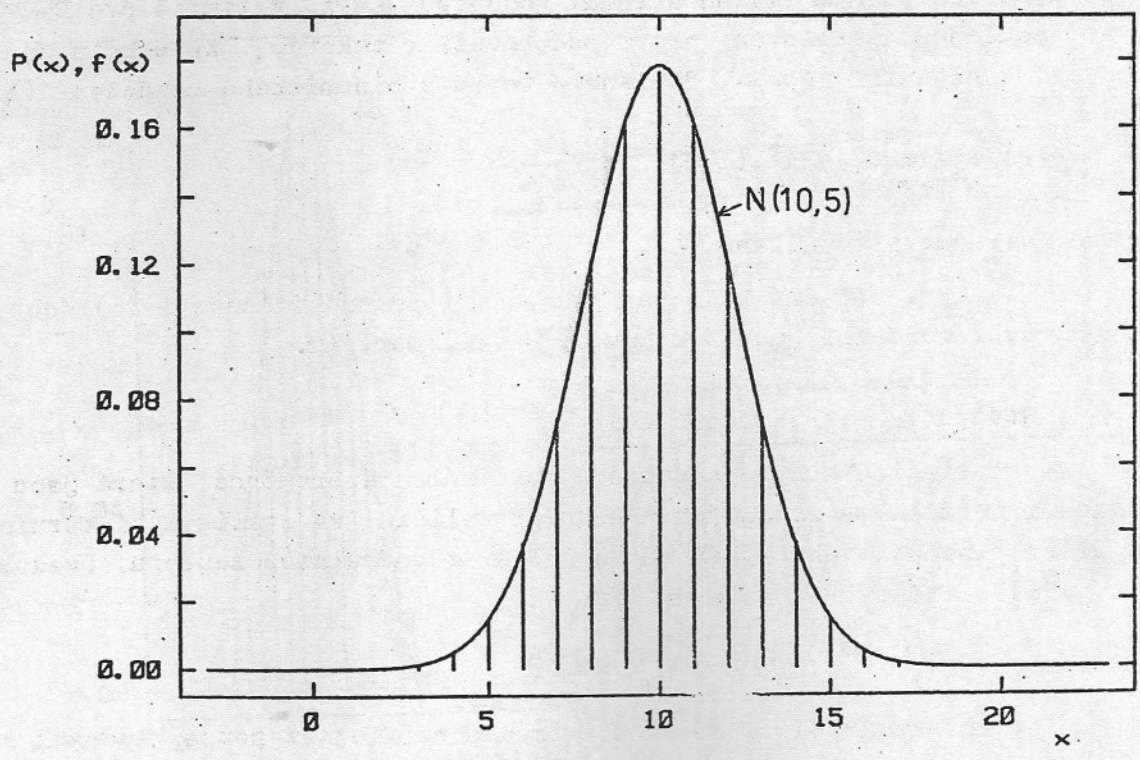
s $q = 1-p$. Střední hodnota, disperze, asymetrie a exces:

$$E(r)=Np, \quad D(r)=Np(1-p), \quad \gamma_1 = \frac{1-2p}{\sqrt{Np(1-p)}}, \quad \gamma_2 = \frac{1-6p(1-p)}{Np(1-p)}. \qquad (3)$$

Proměnná s binomickým rozdělením popisuje výsledky opakovaných pokusů s náhodným jevem, který má jen dva možné výsledky. Jeden z nich



Obr. 9. Pravděpodobnostní obsah obdélníků $\mu_1 \pm k\sigma_1, \mu_2 \pm k\sigma_2$ dvojrozměrného normálního rozdělení v závislosti na koeficientu korelace ρ.



Obr. 10. Binomické (N=10,p=1/2, svislé úsečky) a normální rozdělení.

(označíme ho úspěch) má pravděpodobnost p, druhý 1-p. Pravděpodobnost, že v N pokusech nastane r-krát úspěch, je dána formulí (1). Toto rozdělení se dá použít v mnoha situacích, vybereme-li z možných výsledků nějakou podmnožinu a považujeme ji za úspěch. Například počet událostí v jedné buňce histogramu má binomické rozdělení.

Pro velká N se dá diskrétní funkce (1) dobře aproximovat hustotou pravděpodobnosti normálního rozdělení. V obr. 10 je nakreslena rozdělovací funkce (1) s N=20, p=1/2 a hustota normálního rozdělení se stejnou střední hodnotou 10 a disperzí 5.

Poissonovo rozdělení má náhodná proměnná, která nabývá celé nezáporné hodnoty r s pravděpodobností

$$P(r) = \frac{\mu^r \exp(-\mu)}{r!}, \quad r=0,1,\ldots, \tag{4}$$

kde $\mu>0$ je reálné číslo. Střední hodnota, disperze, asymetrie a exces:

$$E(r)=D(r)=\mu, \quad \gamma_1=1/\sqrt{\mu}, \quad \gamma_2=1/\mu. \tag{5}$$

Poissonovo rozdělení dává pravděpodobnost výskytu r událostí v daném časovém intervalu, jsou-li tyto události nezávislé a vznikají s konstantní rychlostí. Například, z radiaktivního zdroje vylétají částice tak, že pravděpodobnost vyzáření jedné částice za infinitezimální čas δt je $\nu\delta t$. Pravděpodobnost vyzáření r částic za konečný interval délky t je dána rozdělením (4) se střední hodnotou $\mu=\nu t$. V limitě pro $N\to\infty$ a při současném zmenšování pravděpodobnosti p takovém, že součin Np zůstává konstantní, $Np=\mu$, dostaneme totiž z binomického rozdělení (1)

$$P(r) = \lim_{N\to\infty}\left[\binom{N}{r}\left(\frac{\mu}{N}\right)^r \cdot \left(1-\frac{\mu}{N}\right)^{N-r}\right] \tag{6}$$

právě Poissonovo rozdělení (4).

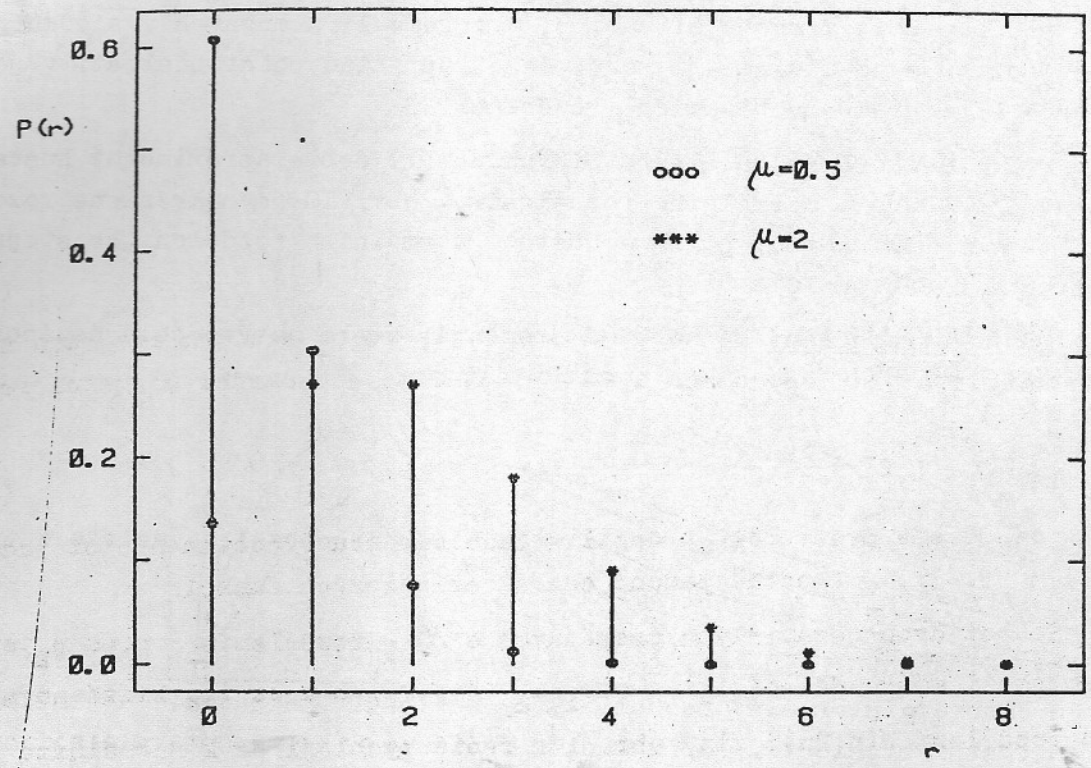
S rostoucí střední hodnotou μ se dají pravděpodobnosti (4) dobře aproximovat normální hustotou $N(\mu,\mu)$ - viz. obr. 12.

8. χ^2, Studentovo a F - rozdělení

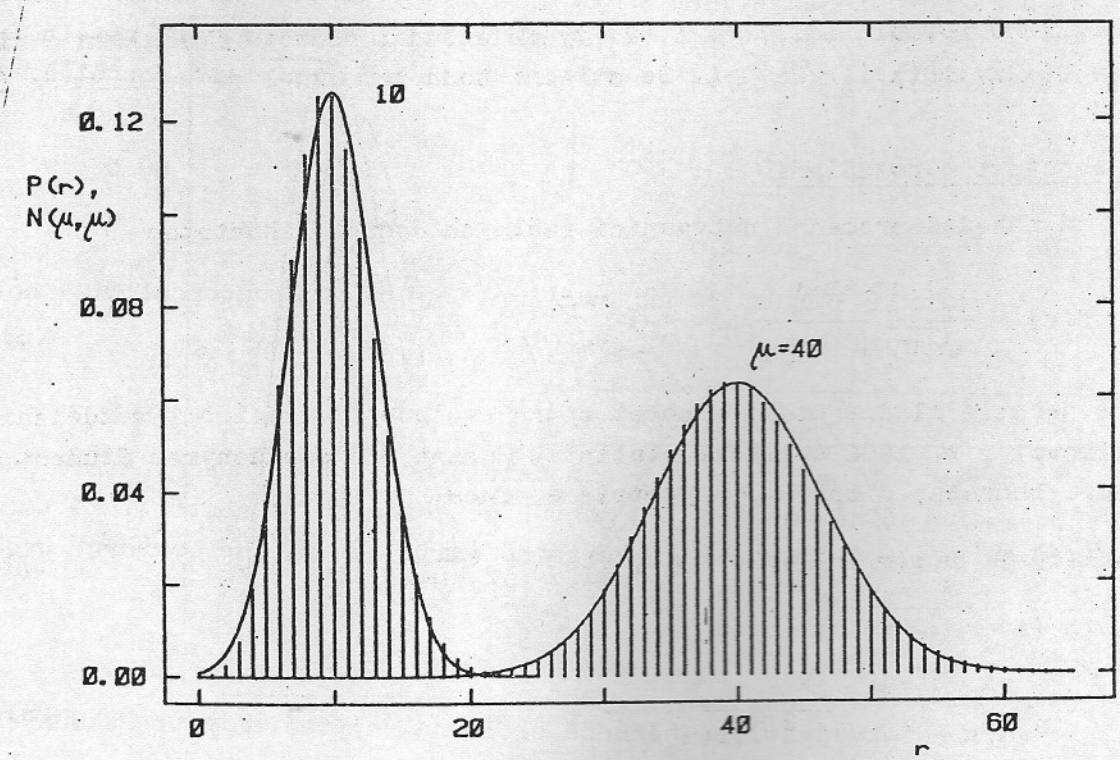
Ve statistice hrají podstatnou roli náhodné proměnné, které jsou funkcemi normálně rozdělených náhodných veličin; ve statistické terminologii se označují jako výběrová rozdělení z normálního souboru. Uvedeme tři nejdůležitější.

χ^2 - rozdělení

(čti chi-kvadrát) má náhodná proměnná nabývající pouze kladných reálných hodnot s hustotou



Obr. 11. Poissonovo rozdělení.



Obr. 12. Poissonovo rozdělení (svislé úsečky) a normální rozdělení $N(\mu,\mu)$.

$$f(x) = \frac{(x/2)^{(n-2)/2}\exp(-x/2)}{2\Gamma(n/2)}. \qquad (1)$$

n je celé kladné číslo, tzv. počet stupňů volnosti, funkce Γ je Eulerův integrál druhého druhu. Střední hodnota, disperze, asymetrie a exces:

$$E(x)=n,\ D(x)=2n,\ \gamma_1=2\sqrt{2/n}\ ,\ \gamma_2=12/n. \qquad (2)$$

Charakteristická funkce:

$$\chi(t) = (1-2it)^{-n/2}\ . \qquad (3)$$

Rozdělení (1) má náhodná veličina x, která je součtem kvadrátů nezávislých proměnných $x_1,\ldots,x_n$, z nichž každá má standardní normální rozdělení N(0,1):

$$x=x_1^2+\ldots+x_n^2\ . \qquad (4)$$

O tom se můžeme přesvědčit s dosti velkou námahou výpočtem hustot podle (2.7) a (3.23), elegantně pomocí charakteristických funkcí.

Součet dvou nezávislých proměnných s χ^2 - rozdělením s n_1 a n_2 stupni volnosti má rozdělení (1) s $n=n_1+n_2$. Pro velká n se (1) blíží normálnímu rozdělení N(n,2n), viz. obr. 14. Ještě rychleji se k normálnímu rozdělení blíží veličina $\sqrt{x}$, přičemž pro její hustotu platí

$$g(\sqrt{2x}) \approx N(\sqrt{2n-1}\ ,\ 1) \text{ pro } n \geq 30. \qquad (5)$$

Z přibližné formule (3.27) a z (2) vyjde střední hodnota $E(\sqrt{2x})\approx 2n$ a disperze $D(\sqrt{2x}) \approx D(x)(1/\sqrt{2n})^2=1$. Se střední hodnotou $\sqrt{2n-1}$ je aproximace (5) lepší.

Studentovo t - rozdělení

má náhodná proměnná nabývající reálných hodnot s hustotou

$$f(t) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\,\Gamma(n/2)}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}\ , \qquad (6)$$

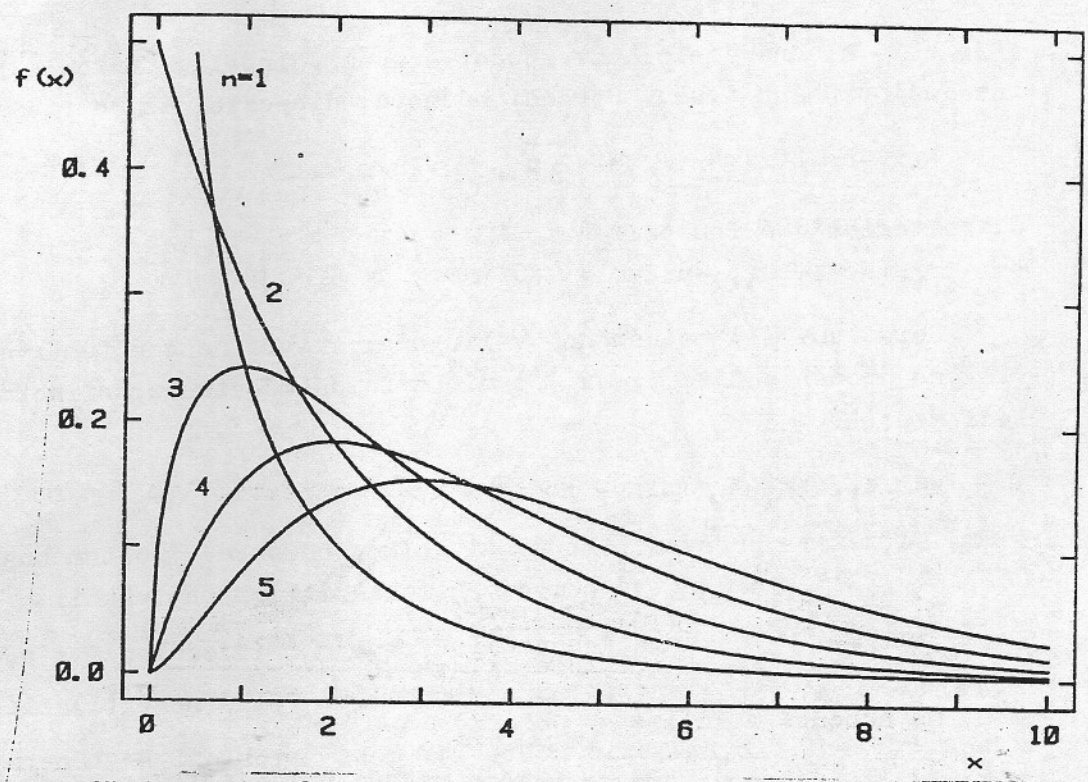
kde n je celé kladné číslo - počet stupňů volnosti. Práci o t-rozdělení publikoval v r. 1908 anglický statistik Gosset pod pseudonymem Student. Střední hodnota, disperze, asymetrie a exces:

$$E(t)=0,\ D(t)=\frac{n}{n-2} \text{ pro } n>2,\ \gamma_1=0,\ \gamma_2=\frac{6}{n-4} \text{ pro } n>4. \qquad (7)$$

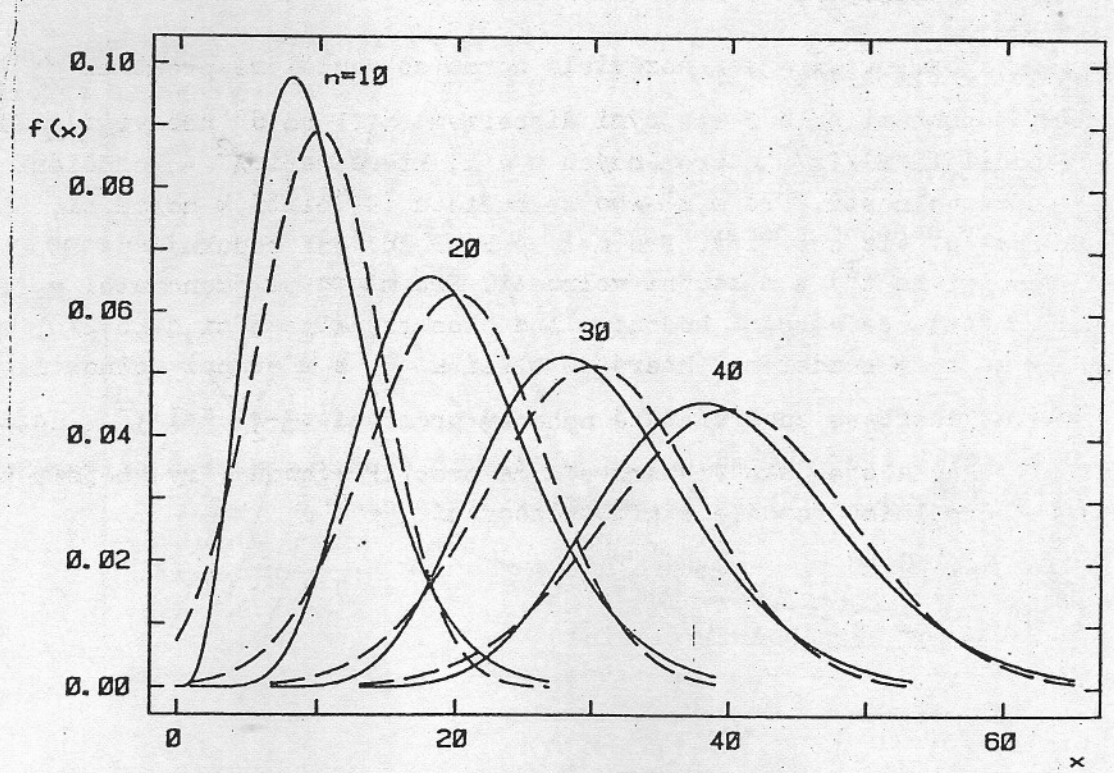
Hustotu (6) má náhodná veličina

$$t=\frac{x_o}{\sqrt{x/n}} = \frac{x_o}{\sqrt{(x_1^2+\ldots+x_n^2)/n}}\ , \qquad (8)$$

kde x_o a x jsou nezávislé, x_o má standardní normální rozdělení a x rozdě-



Obr. 13. Rozdělení χ^2.



Obr. 14. Rozdělení χ^2 (plná čára) a normální rozdělení se stejnou střední hodnotou a disperzí (čárkovaná čára).

lení χ^2 s n stupni volnosti (vztah (4)). Pro rozdělení s n=1 se používá názvu Cauchyovo (§9). S rostoucím n se hustota (6) přibližuje ke standardnímu normálnímu rozdělení. Obvykle se t-rozdělení nahrazuje normálním N(0,1) pro $n \geq 30$.

F - rozdělení

je zobecněním předchozích dvou. Označuje se často také jako Fisherovo-Snedecorovo nebo jen Snedecorovo, nebo jako rozdělení v^2. Hustota pravděpodobnosti je nenulová jen pro kladné hodnoty F:

$$f(F)=\left(\frac{m}{m'}\right)^{m/2} \frac{\Gamma\left(\frac{m+m'}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{m'}{2}\right)} F^{\frac{m-2}{2}} \cdot \left(1+\frac{m}{m'}F\right)^{-\frac{m+m'}{2}}, \quad F>0. \tag{9}$$

Zde jsou m, m' celá kladná čísla - počty stupňů volnosti. Střední hodnota a disperze:

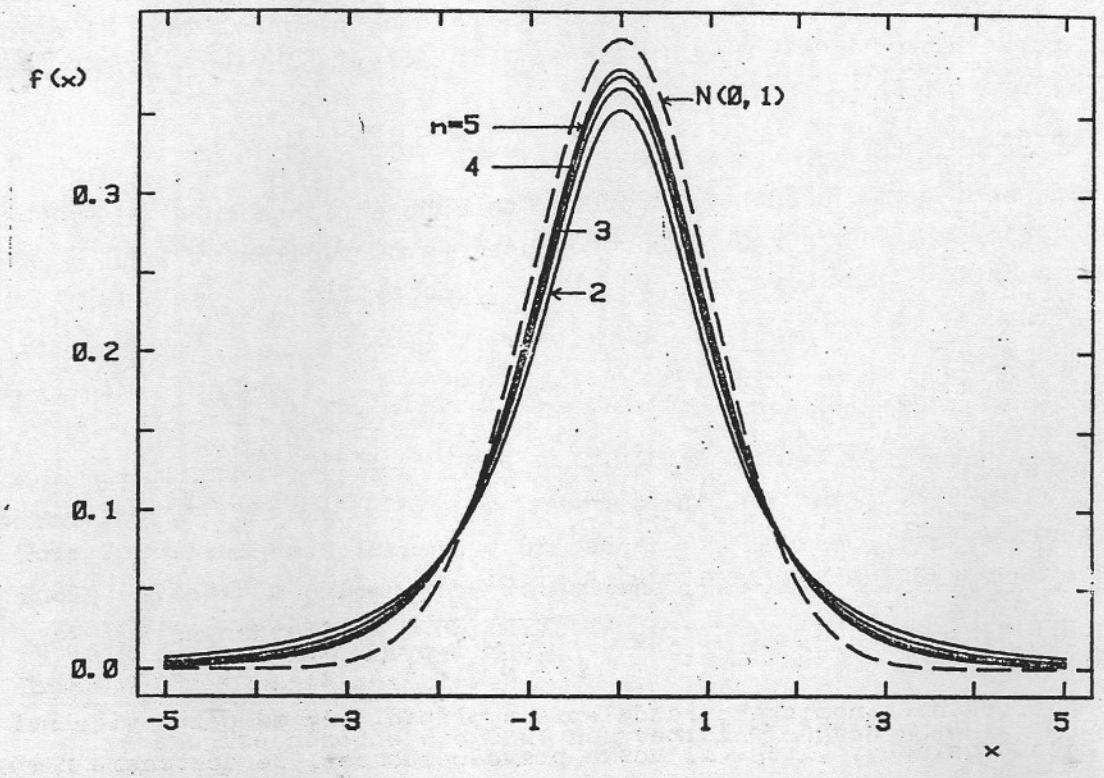
$$E(F)=\frac{m'}{m'-2} \text{ pro } m'>2, \quad D(F)=\frac{2m'^2(m+m'-2)}{m(m'-2)^2\cdot(m'-4)} \text{ pro } m'>4. \tag{10}$$

Hustotu (9) má náhodná veličina

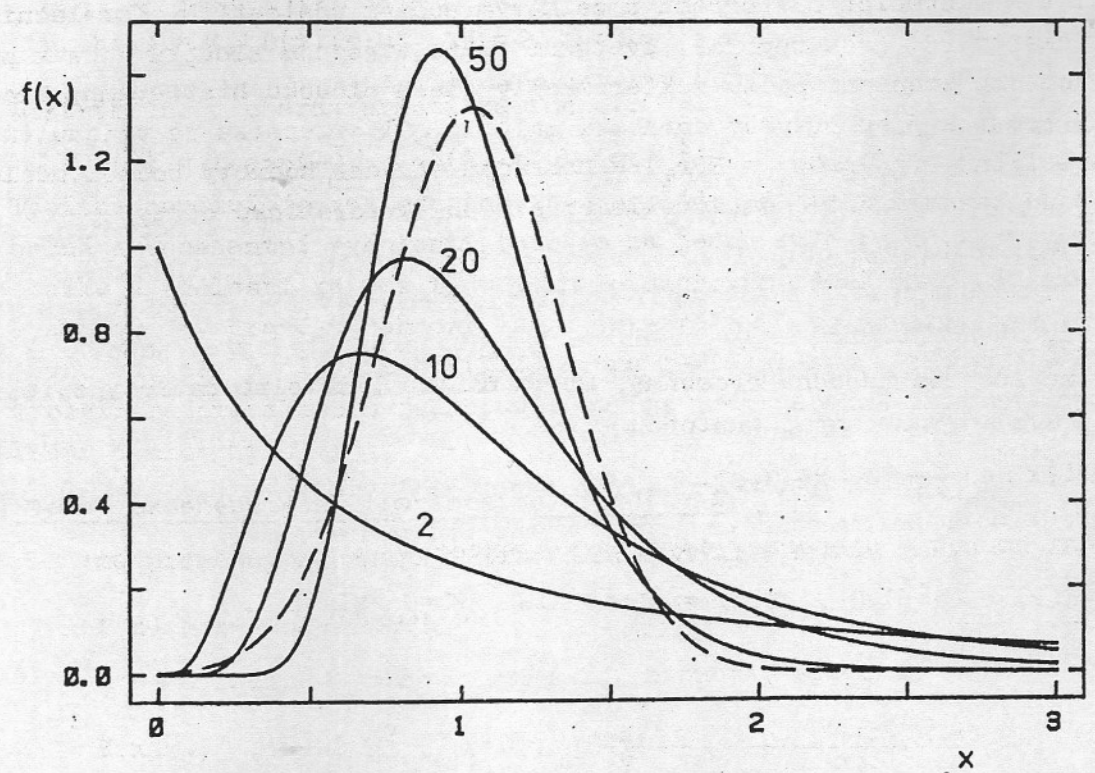
$$F=\frac{\frac{1}{m}(x_1^2+\ldots+x_m^2)}{\frac{1}{m'}(\tilde{x}_1^2+\ldots+\tilde{x}_{m'}^2)}, \tag{11}$$

kde $x_1,\ldots,x_m,\tilde{x}_1,\ldots,\tilde{x}_{m'}$ jsou nezávislé normálně rozdělené proměnné se středními hodnotami nula a stejnými disperzemi σ^2 (F na σ^2 nezávisí). Je to tedy podíl $(x/m)/(x'/m')$ proměnných x a x', které mají χ^2 - rozdělení s m a m' stupni volnosti. Pro $m,m' \to \infty$ se hustota (9) blíží k normální, ale poměrně pomalu (viz obr. 16). Pro m=1 se F - rozdělení redukuje na Studentovo (přesněji na t^2) s m' stupni volnosti. Pro $m' \to \infty$ má jmenovatel v (11) normální hustotu se střední hodnotou 1 a disperzí klesající jako $2/m'$; veličina mF má tedy rozdělení, které se blíží k χ^2 s m stupni volnosti.

Poměrně často se používá také náhodná proměnná $z=-\frac{1}{2}\ln F=\ln\sqrt{F}$. Její rozdělení se označuje jako Fisherovo a má proti F výhodu v rychlejším přiblížení k normálnímu rozdělení při zvětšování m a m'.



Obr. 15. t - rozdělení s různým počtem stupňů volnosti n a limitní normální rozdělení.



Obr. 16. F - rozdělení s různým počtem stupňů volnosti m=m' (plná čára) a normální rozdělení se stejnou střední hodnotou a disperzí jako má F při m=m'=50, t.j. N(1.04, 0.0925) (čárkovaná čára).

9. Další modelová rozdělení, souvislost některých rozdělení

Řadu základních rozdělení popsaných v předchozích paragrafech doplníme několika dalšími užitečnými typy.

Multinomické rozdělení

má k-rozměrná diskrétní náhodná proměnná nabývající celých nezáporných hodnot $r_1,\dots,r_k$ z rozmezí $0,1,\dots,N$ s pravděpodobnostmi

$$P(r_1,\dots,r_k) = \frac{N!}{r_1!\dots r_k!}\, p_1^{r_1}\dots p_k^{r_k} . \qquad (1)$$

Přitom jsou parametry $p_1,\dots,p_k$ nezáporná reálná čísla taková, že $p_1+\dots+p_k=1$. Střední hodnoty a disperze jsou

$$E(r_i) = Np_i , \quad D(r_i) = Np_i(1-p_i) , \qquad (2)$$

smíšené druhé momenty a korelační koeficienty

$$D(r_i,r_j) = -Np_ip_j , \quad \varrho_{ij} = -\sqrt{p_ip_j/(1-p_i)(1-p_j)} \quad \text{pro } i\neq j. \qquad (3)$$

Je to zobecnění binomického rozdělení na případ, kdy má pokus více než dva možné výsledky. Vztah (1) udává pravděpodobnost, že dostaneme r_i výsledků typu i v N nezávislých pokusech, když p_i je pravděpodobnost výsledku typu i v jednom pokusu. Multinomické rozdělení popisuje například četnosti v k sloupcích histogramu s celkovým počtem událostí N. Korelační koeficienty (3) jsou záporné, zvětšení počtu v jednom sloupku vede k pravděpodobnému zmenšení počtu v kterémkoliv jiném sloupku histogramu. Pro velký počet k jsou pravděpodobnosti malé: $p_i \ll 1$. Disperze ze vztahu (2) je přibližně $D(r_i)\approx Np_i = E(r_i)$ a náhradou střední hodnoty počtem událostí r_i dostaneme užitečnou aproximaci střední kvadratické odchylky

$$\sqrt{D(r_i)} = \sigma_i \approx \sqrt{r_i} . \qquad (4)$$

Rovnoměrné rozdělení

má spojitá náhodná proměnná, která nabývá libovolné hodnoty z intervalu $\langle a,b\rangle$ s konstantní hustotou

$$f(x) = \frac{1}{b-a} , \quad x\in\langle a,b\rangle . \qquad (5)$$

Střední hodnota, disperze, asymetrie a exces:

$$E(x) = (a+b)/2 , \quad D(x) = (b-a)^2/12 , \quad \gamma_1=0 , \quad \gamma_2=-1.2. \qquad (6)$$

Charakteristická funkce:

$$\chi(t) = \frac{\sinh[it(b-a)/2]}{it(b-a)} + \frac{it(b+a)}{2} . \qquad (7)$$

Rovnoměrné rozdělení může popisovat například chyby, vznikající zaokrouhlováním čísel.

Beta - rozdělení

má spojitá náhodná proměnná s hodnotami z $\langle 0,1\rangle$ s hustotou

$$f(x) = \frac{\Gamma(n+m)}{\Gamma(m)\Gamma(n)} x^{m-1} (1-x)^{n-1}, \quad x \in \langle 0,1\rangle, \tag{8}$$

kde n,m jsou parametry (celá kladná čísla). Střední hodnota a disperze:

$$E(x) = \frac{m}{m+n}, \quad D(x) = \frac{mn}{(m+n)^2(m+n+1)}, \tag{9}$$

asymetrie a exces:

$$\gamma_1 = \frac{2(n-m)\sqrt{m+n+1}}{\sqrt{mn}\,(m+n+2)}, \quad \gamma_2 = \frac{3(m+n+1)[2(m+n)^2+mn(m+n-6)]}{mn(m+n+2)(m+n+3)} - 3. \tag{10}$$

Toto rozdělení se uplatňuje v případech proměnných ohraničených shora i zdola. Zvláštním případem je rovnoměrné rozdělení (m=n=1). Několik hustot typu (8) je nakresleno v obr. 17.

Exponenciální rozdělení

má spojitá náhodná proměnná nabývající kladných hodnot s hustotou

$$f(x) = \frac{1}{\mu} \exp(-\frac{x}{\mu}), \quad x > 0, \tag{11}$$

kde $\mu > 0$ je reálný parametr. Střední hodnota, disperze, asymetrie a exces:

$$E(x) = \mu, \quad D(x) = \mu^2, \quad \gamma_1 = 2, \gamma_2 = 6. \tag{12}$$

Distribuční a charakteristická funkce:

$$F(x) = 1-\exp(-\frac{x}{\mu}), \quad \chi(t) = (1-i\mu t)^{-1}. \tag{13}$$

Typické použití je následující: předpokládejme, že události vznikají náhodně s konstantní rychlostí (počtem za jednotku času) ν. Pravděpodobnost vzniku N událostí za čas t je dána Poissonovým rozdělením (§7) se střední hodnotou νt. Pravděpodobnost, že v intervalu $\langle 0,t\rangle$ pozorujeme alespoň jednu událost je podle vztahu (7.4) rovna $1-P(0)=1-\exp(-\nu t)$. Čas t, během kterého zaregistrujeme alespoň jednu událost, je tedy náhodná proměnná s distribuční funkcí typu (13).

Dvojné exponenciální (Laplaceovo) rozdělení

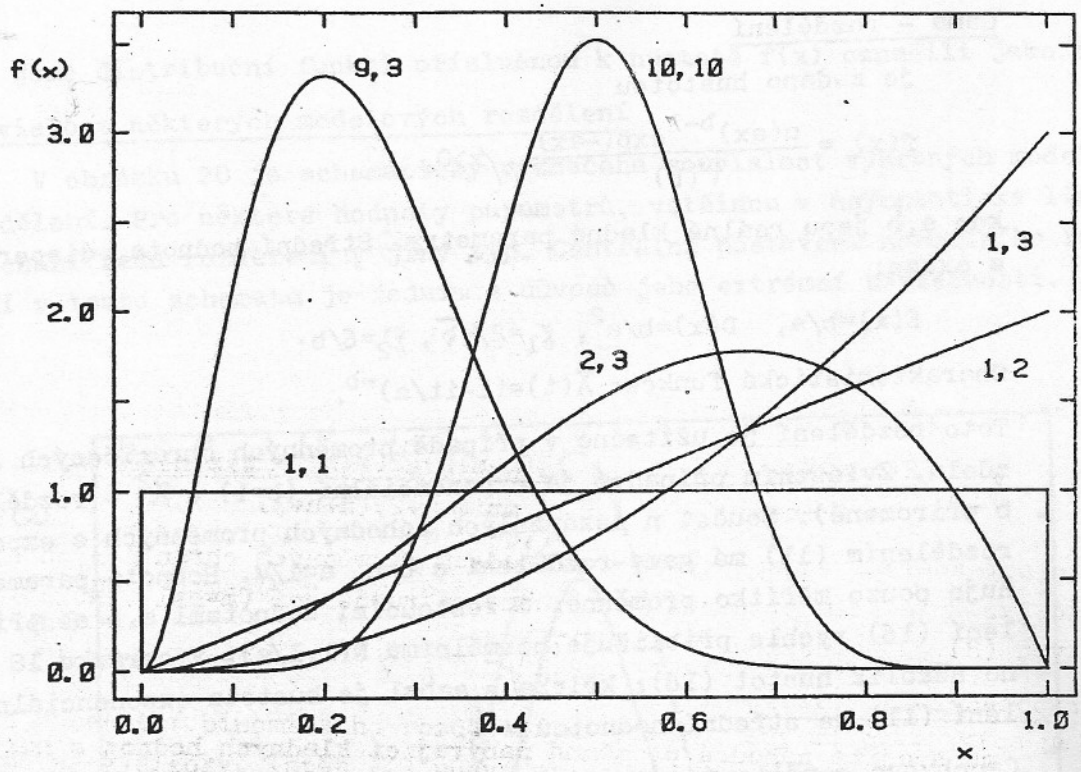
má proměnná nabývající libovolných hodnot s hustotou

$$f(x) = \frac{\lambda}{2} \exp(-\lambda|x-\mu|), \tag{14}$$

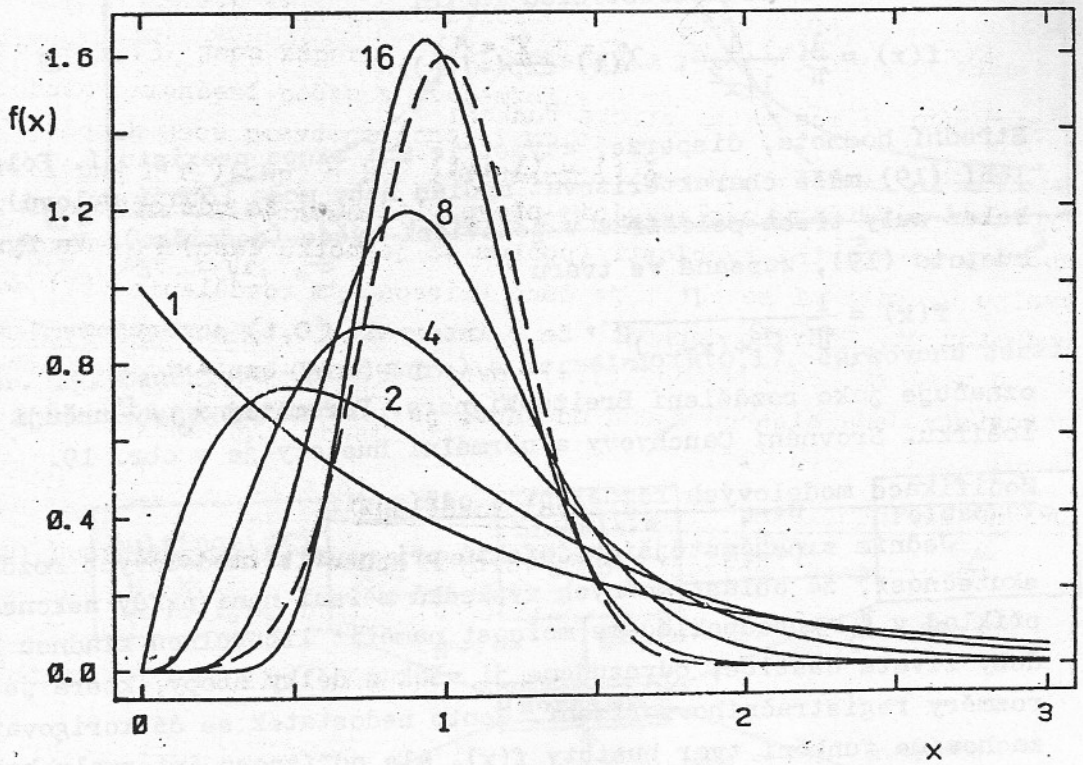
$\lambda > 0$, μ jsou reálné parametry. Střední hodnota, disperze, asymetrie a exces:

$$E(x)=\mu, \quad D(x)=2/\lambda^2, \quad \gamma_1=0, \quad \gamma_2=3. \tag{15}$$

Pro velké $|x|$ ubývá hustota (14) pomaleji než pro normální, ale rychleji než pro Cauchyovo rozdělení (19).



Obr. 17. Hustoty beta-rozdělení s různými parametry n,m.



Obr. 18. Hustoty gama-rozdělení s různými hodnotami a=b, plná čára; normální rozdělení N(1,1/16), čárkovaná čára.

Gama - rozdělení

je zadáno hustotou

$$f(x) = \frac{a(ax)^{b-1}\exp(-ax)}{\Gamma(b)}, \quad x>0, \tag{16}$$

kde a,b jsou reálné kladné parametry. Střední hodnota, disperze, asymetrie a exces:

$$E(x)=b/a, \quad D(x)=b/a^2, \quad \gamma_1=2/\sqrt{b}, \quad \gamma_2=6/b. \tag{17}$$

Charakteristická funkce: $\chi(t)=(1-it/a)^{-b}$. (18)

Toto rozdělení je užitečné v případě proměnných ohraničených shora nebo zdola. Zvláštním případem je exponenciální (b=1) a χ^2 - rozdělení (a=1/2, b přirozené). Součet n nezávislých náhodných proměnných s exponenciálním rozdělením (11) má gama-rozdělení s b=n, a=1/μ. Hodnota parametru a ovlivňuje pouze měřítko proměnné. S rostoucími hodnotami a,b se při a=b rozdělení (16) rychle přibližuje normálnímu N(1,1/a). V obrázku 18 je nakresleno několik hustot (16); křivka s a=b=1 je hustota exponenciálního rozdělení (11) se střední hodnotou μ=1.

Cauchyovo rozdělení

má spojitá náhodná proměnná, nabývající libovolných reálných hodnot s hustotou a charakteristickou funkcí

$$f(x) = \frac{1}{\pi}\frac{1}{1+x^2}, \quad \chi(t)=\exp(-|t|). \tag{19}$$

Střední hodnota, disperze, asymetrie ani exces neexistují. Polohu rozdělení (19) může charakterizovat medián nebo moda (obojí nulové), rozptyl kolem nuly třeba pološirka v poloviční výšce (jednička). Ve fyzice se hustota (19), zapsaná ve tvaru

$$f(x) = \frac{1}{\pi}\frac{\Gamma}{\Gamma^2+(x-x_0)^2}, \tag{20}$$

označuje jako rozdělení Breita-Wignera. Parametry x_0 a Γ určují modu a pološirku. Srovnání Cauchyovy a normální hustoty je v obr. 19.

Modifikace modelových rozdělení - odříznutí

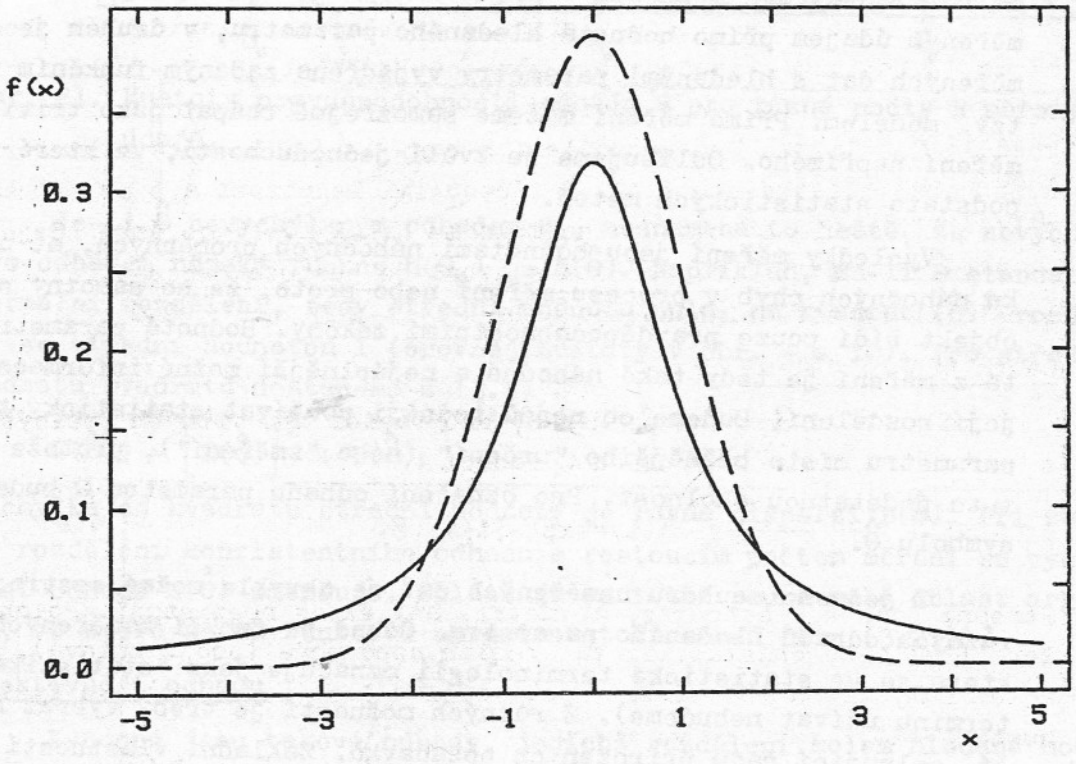
Jedním z nejčastějších defektů při použití modelových rozdělení je skutečnost, že oblast možných výsledků měření není nikdy nekonečná. Například v § 13 předpokládáme možnost naměřit libovolnou kladnou hodnotu doby života částice; odvozujeme ji však z délky stopy, která je omezena rozměry registračního zařízení. Tento nedostatek se dá korigovat tak, že zachováme funkční tvar hustoty f(x), ale odřízneme intervaly hodnot, které nemohou nastat. Pro rozdělení v intervalu (a,b) musíme původní f(x) normovat:

$$\tilde{f}(x) = \frac{f(x)}{F(b)-F(a)}, \quad x\in(a,b), \tag{21}$$

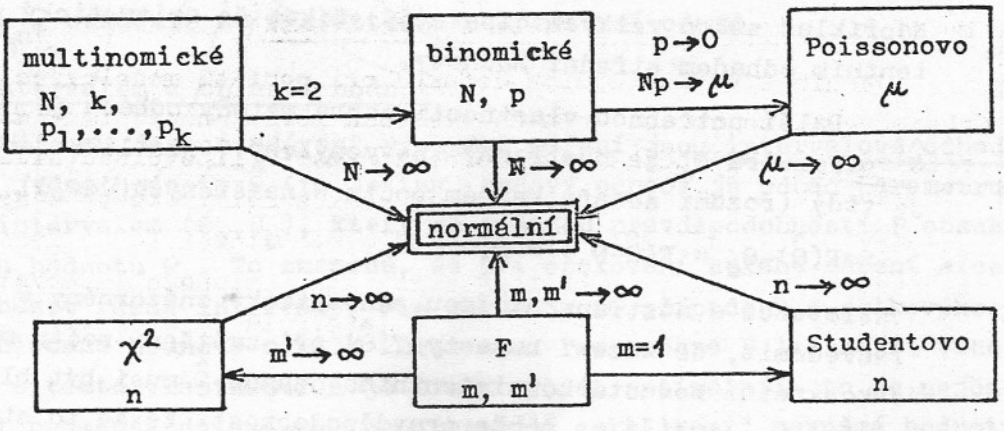
kde jsme distribuční funkci příslušnou k hustotě f(x) označili jako F(x).

Souvislost některých modelových rozdělení

V obrázku 20 je schematicky vyznačena souvislost vybraných modelových rozdělení. Pro některé hodnoty parametrů, většinou v asymptotické limitě, přechází řada rozdělení v jiný typ. Centrální postavení normálního rozdělení v tomto schematu je jedním z důvodů jeho extrémní užitečnosti.



Obr. 19. Cauchyovo (plná čára) a normální (N(0,1), čárkovaná čára) rozdělení.



Obr. 20. Souvislost modelových rozdělení.

II. Odhad parametrů

10. Metody statistického odhadu parametrů

Ve velké většině případů je cílem měření určit hodnoty neznámých veličin, které budeme označovat jako parametry. Někdy je cíl jiný, totiž posouzení správnosti jedné nebo několika hypotéz; v takové situaci se používají statistické metody testů hypotéz, kterými se budeme stručně zabývat v části III. V úloze určení hodnot parametrů z naměřených dat budeme rozlišovat dvě možnosti - přímé a nepřímé měření. V prvním případě je měřeným údajem přímo hodnota hledaného parametru, v druhém je souvislost měřených dat s hledanými parametry vyjádřena zadaným funkčním vztahem, tzv. modelem. Přímé měření můžeme samozřejmě chápat jako triviální případ měření nepřímého. Odlišujeme je kvůli jednoduchosti, ve které vynikne podstata statistických metod.

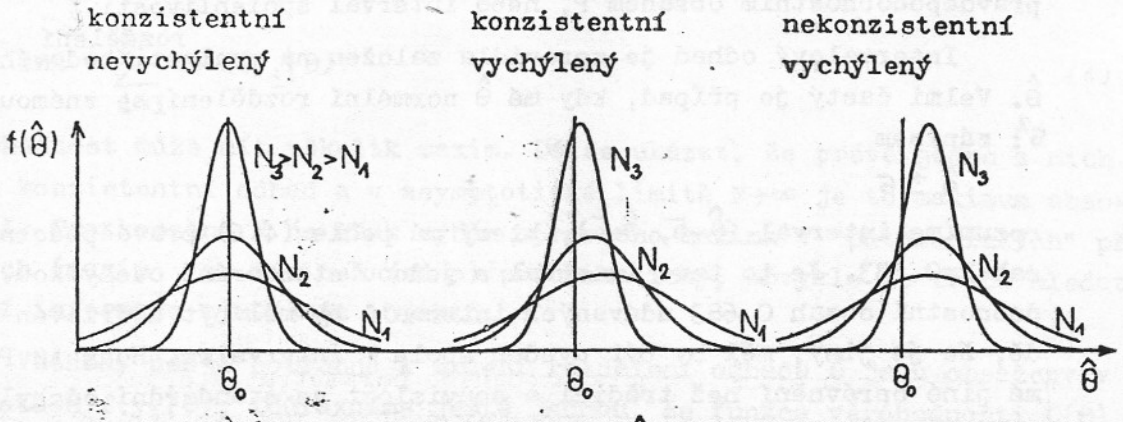
Výsledky měření jsou hodnotami náhodných proměnných, ať už v důsledku náhodných chyb v procesu měření nebo proto, že se samotný studovaný objekt řídí pouze pravděpodobnostními zákony. Hodnota parametru odhadnutá z měření je tedy také náhodná a nejúplnější možná informace o ní je její rozdělení. Budeme co nejdůsledněji používat statistický termín odhad parametru místo běžnějšího "určení" (nebo "změření"), protože vyjadřuje tuto podstatnou okolnost. Pro označení odhadu parametru Θ budeme užívat symbolu $\hat{\Theta}$.

Z jednoho souboru naměřených dat je obvykle možné sestrojit mnoho různých odhadů hledaného parametru. Odhad je funkcí naměřených hodnot, která se ve statistické terminologii označuje jako "statistika" (tohoto termínu užívat nebudeme). Z různých možností je třeba vybrat nejvhodnější, splňující řadu přirozených požadavků. Základní vlastností by měla být tzv. konzistence. Metoda odhadu se označuje jako konzistentní, konvergují-li odhady ke skutečné hodnotě parametru při zvětšování počtu měření. Konzistence odhadu zaručuje, že s pomocí dostatečně velkého počtu měření dokážeme "lokalizovat" neznámý parametr s libovolně velkou přesností. Například zákon velkých čísel (§5) říká, že aritmetický průměr je konzistentním odhadem střední hodnoty.

Další potřebnou vlastností dobré metody odhadu je nestrannost. Odhad $\hat{\Theta}$ parametru Θ_0 je nestranný (nevychýlený), jestliže jeho střední hodnota je vždy (rozumí se při každém počtu N naměřených údajů) rovna Θ_0:

$$E(\hat{\Theta})-\Theta_0 = E(\hat{\Theta}-\Theta_0) = 0. \qquad (1)$$

Konzistence a nestrannost jsou schematicky znázorněny v obr. 21. Je třeba si uvědomit, že zúžení hustoty $f(\hat{\Theta})$ při zvětšení počtu měření neznamená, že konkrétní hodnota konzistentního odhadu $\hat{\Theta}$ musí být blíž ke skutečné hodnotě Θ_0, zvětší se pouze pravděpodobnost, že se to stane.



Obr. 21. Hustoty pravděpodobnosti odhadu $\hat{\theta}$ pro různé počty N měřených údajů.

Je-li $\hat{\theta}$ nevychýleným odhadem θ_0, neznamená to ještě, že nevychýleným odhadem nějaké funkce $h(\theta_0)$ je $h(\hat{\theta})$. Například, má-li $\hat{\theta}$ standardní normální rozdělení, tedy střední hodnotu nula, má kvadrát $(\hat{\theta})^2$ rozdělení χ^2_1 se střední hodnotou 1 (srovnej hustoty v obr. 4 a 10). Pro střední hodnotu kvadrátu dostaneme z (3.4)

$$E(\hat{\theta}^2) = \left[E(\hat{\theta})\right]^2 + D(\hat{\theta}), \qquad (2)$$

odchylka od kvadrátu střední hodnoty je rovna disperzi $D(\hat{\theta})$. Při zužování rozdělení konzistentního odhadu s rostoucím počtem měření se vychýlenost odhadu $h(\hat{\theta})$ zmenšuje. Uplatňuje se totiž pouze malá oblast argumentů, ve které se dá funkce h aproximovat lineárně (viz (3.27)).

Efektivnost odhadu

Výhodné jsou takové odhady, jejichž rozdělení kolem hledané hodnoty je co nejužší. Vhodnou mírou šířky rozdělení $\hat{\theta}$ je disperze $D(\hat{\theta})$; k hodnocení efektivnosti používáme podíl $D_{min}/D(\hat{\theta})$, kde D_{min} je nejmenší možná disperze mezi všemi odhady. Obvykle se daří celkem snadno najít asymptotickou efektivnost v limitě $N\to\infty$ (N je počet změřených údajů). Je-li $D(\hat{\theta})=D_{min}$, označuje se $\hat{\theta}$ krátce jako efektivní odhad.

Odhad intervalem a oblastí hodnot

Ustálenou formou udávání výsledků měření jsou intervalové odhady. Namísto jedné hodnoty $\hat{\theta}$ (to je tzv. bodový odhad) je odhad parametru vyjádřen intervalem (θ_a,θ_b), který se zadanou pravděpodobností P obsahuje hledanou hodnotu θ_0. To znamená, že při opakování celého měření sice budou vycházet různé intervaly, ale zhruba v nP případech z celkového počtu n bude hledaná hodnota uvnitř intervalu. Pro zadané P lze najít více intervalů s touto vlastností a je třeba vybrat optimální - to je nejčastěji interval nejmenší délky (pro "nejpřesnější lokalizaci" neznámé hodnoty).

Takto vybranému intervalu se ve statistice říká konfidenční interval s pravděpodobnostním obsahem P, nebo interval spolehlivosti.

Intervalový odhad je zpravidla založen na znalosti rozdělení bodového odhadu $\hat{\theta}$. Velmi častý je případ, kdy má $\hat{\theta}$ normální rozdělení se známou disperzí σ^2; zápisem

$$\hat{\theta} \pm \sigma$$

rozumíme interval $(\hat{\theta}-\sigma, \hat{\theta}+\sigma)$, který má podle (4.6) pravděpodobnostní obsah P=0.683. Je to tzv. interval s jednou standardní odchylkou. Pravděpodobnostní obsah 0.683 udávaných intervalů by měl být dodržován a v případě, že je jiný, měl by být uveden spolu s intervalem. Hodnota P=0.683 nemá jiné oprávnění než tradici a souvislost se standardní odchylkou normálního rozdělení. Podobně interval $\hat{\theta}\pm 2\sigma$, podle (4.6) s P=0.954, se často uvádí jako výsledek měření - v případě, kdy chceme standardní pravděpodobnost 0.683 zvětšit. Mezi délkou intervalu a jeho pravděpodobnostním obsahem je třeba vybrat rozumný kompromis.

Odhadujeme-li několik parametrů současně, udáváme <u>oblast</u> hodnot, která se zadanou pravděpodobností obsahuje hledaný bod prostoru parametrů. V následujících odstavcích se budeme hledáním takových intervalů a oblastí několikrát zabývat.

Z běžných metod odhadu vybereme dvě nejdůležitější, které zpravidla dávají výsledky s požadovanými vlastnostmi (konzistence, efektivnost). Protože v tomto místě chceme vysvětlit podstatné myšlenky metod, budeme hovořit o jednom parametru; technické detaily postupu s větším počtem parametrů jsou v následujících odstavcích (zejména §§ 15-17).

<u>Metoda maximální věrohodnosti</u>

Předpokládejme, že nezávislé naměřené hodnoty $y_1,\ldots,y_N$ jsou náhodná čísla popsaná hustotami $f(y_i|\theta)$, závislými na hledaném parametru θ. Odhad je možné založit na principu maximální věrohodnoti - najít ho tak, aby s hodnotou $\hat{\theta}$ byla naměřená data pravděpodobnější než s jinými hodnotami θ. Hustota pravděpodobnosti N-tice nezávislých náhodných proměnných je rovna součinu jednotlivých hustot:

$$L(y_1,\ldots,y_N|\theta) = \prod_{i=1}^{N} f(y_i|\theta). \qquad (3)$$

Při dosazení naměřených hodnot y_i je L funkcí θ, pro kterou zavedl Fisher označení funkce věrohodnosti a použil ji k formulaci metody maximální věrohodnosti: pro hodnotu $\hat{\theta}$ má $L(\theta)$ maximum. Je nutné si uvědomit, že proměnná θ není náhodná; zacházíme s ní tak, že zkoušíme, jak velkou věrohodnost L mají její možné hodnoty a pro odhad vybíráme bod maxima. $\hat{\theta}$ už ovšem je náhodnou proměnnou, protože při opakování měření vyjde jiná N-tice y_i a tedy i jiná funkce $L(\theta)$.

Podmínku maxima L můžeme zapsat jako podmínku maxima logaritmu L

(L a lnL mají extrémy ve stejných bodech):

$$\ln L = \sum_{i=1}^{N} \ln f(y_i|\theta). \qquad (4)$$

Věrohodnost může mít několik maxim. Dá se ukázat, že právě jedno z nich dává konzistentní odhad a v asymptotické limitě $N\to\infty$ je to maximum absolutní. Pro konečné N je však výběr správného maxima v "patologických" případech (maxim je víc než jedno) problematický; obvykle je třeba hledat další informace o měřeném objektu.

Všechny údaje potřebné k určení rozdělení odhadu $\hat{\theta}$ jsou obsaženy v hustotách $f(y_i|\theta)$; zdůrazníme ještě jednou, že funkce věrohodnosti $L(\theta)$ není hustotou pravděpodobnosti odhadu $\hat{\theta}$. Prakticky se hustota $\hat{\theta}$ dá najít v některých jednoduchých a přitom důležitých případech. V následujících odstavcích uvidíme, že odhady mají typicky rozdělení normální nebo blízké k normálnímu. Obecně je hledání hustot odhadů značně obtížné, funkční závislost $\hat{\theta}$ na měřených datech y_i je dána pouze implicitně - podmínkou maxima věrohodnosti. Potěšitelné je zjednodušení pro $N\to\infty$; za velmi obecných podmínek mají odhady, díky platnosti centrální limitní věty, normální rozdělení. Pro disperzi $\hat{\theta}$ vychází v limitě jednoduchá formule

$$D(\hat{\theta}) = \left(-\frac{\partial^2 \ln L}{\partial \theta^2} \right)^{-1} \Bigg|_{\theta=\hat{\theta}} ; \qquad (5)$$

je to zároveň minimální možná hodnota disperze. Odhad metodou maximální věrohodnosti je asymptoticky efektivní.

Metoda nejmenších čtverců

Abychom mohli použít metodu maximální věrohodnosti, musíme znát rozdělení měřených hodnot v závislosti na odhadovaném parametru. V metodě nejmenších čtverců stačí znalost závislosti středních hodnot $E(y_i|\theta)$ a disperzí $D(y_i|\theta)$ na parametru θ. Odhad $\hat{\theta}$ hledáme, za předpokladu nezávislosti naměřených y_i, z podmínky minima součtu čtverců odchylek

$$S = \sum_{i=1}^{N} \frac{[y_i - E(y_i|\theta)]^2}{D(y_i|\theta)} . \qquad (6)$$

Vybíráme tedy takovou hodnotu, pro kterou jsou očekávané (modelové) střední hodnoty co nejblíže naměřeným y_i. Přitom počítáme s tím, že pro hledanou hodnotu θ_o budou odchylky $y_i - E(y_i|\theta_o)$ zpravidla tím větší, čím větší je disperze y_i. Proto jsou v sumě (6) kvadráty odchylek násobeny tzv. vahou $1/D(y_i|\theta)$. Čím větší je disperze i-tého bodu, tím menší je jeho váha a relativní příspěvek do součtu; podmínka minima S povoluje v tomto bodě větší odchylku. Naopak, modelová a naměřená hodnota s malou disperzí musí být blízké; velká váha v součtu čtverců ovlivňuje výběr odhadu v tomto směru.

Odhad $\hat{\theta}$ se nezmění, násobíme-li všechny členy v součtu (6) stejnou konstantou. To znamená, že není třeba znát všechny disperze $D(y_i|\theta)$, stačí jejich relativní velikosti. Jsou-li všechny $D(y_i|\theta)$ stejné a nezávislé na θ, neuplatní se v odhadu $\hat{\theta}$ z nejmenších čtverců vůbec; potom hledáme minimum sumy

$$S = \sum_{i=1}^{N} \left[y_i - E(y_i|\theta)\right]^2 = \sum_{i=1}^{N} \left[y_i - f_i(\theta)\right]^2 . \qquad (7)$$

Zde jsme zavedli nové označení $f_i(\theta)$ pro funkční závislost i-té hodnoty modelu měřených hodnot na parametru. Takový zápis je běžný v situaci, kdy měřené hodnoty y_i jsou součtem

$$y_i = f_i(\theta) + \varepsilon_i \qquad (8)$$

hodnot modelu a náhodné chyby ε_i s nulovou střední hodnotou. Často jsou měřené údaje získány při různých (známých) hodnotách nějakého parametru x, což zapíšeme symbolicky jako

$$y_i = f(x_i, \theta) + \varepsilon_i . \qquad (9)$$

Pozoruhodné vlastnosti má odhad metodou nejmenších čtverců v případě lineárního modelu, kdy $E(y_i|\theta)$ je lineární funkcí θ a $D(y_i|\theta)$ na θ nezávisí. Především jsou odhady z minima S lineárními funkcemi y_i, jsou nevychýlené při libovolném N a mají minimální disperzi ze všech možných nevychýlených lineárních odhadů (Gaussova-Markova věta). Tyto vlastnosti nezávisí na rozdělení dat, jsou dány pouze linearitou modelu.

Rozdělení dat

Mají-li měřené hodnoty y_i normální rozdělení (4.1) se středními hodnotami $f_i(\theta)$ a disperzemi σ_i^2 nezávislými na θ,

$$f(y_i|\theta) = \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\left\{ -\frac{\left[y_i - f_i(\theta)\right]^2}{2\sigma_i^2} \right\} , \qquad (10)$$

vyjde velmi jednoduchá souvislost logaritmu věrohodnosti (4) a součtu čtverců (6):

$$\ln L = \sum_{i=1}^{N} \left\{ -\frac{\left[y_i - f_i(\theta)\right]^2}{2\sigma_i^2} - \ln(\sqrt{2\pi}\,\sigma_i) \right\} = -\frac{S}{2} - \sum_{i=1}^{N} \ln(\sqrt{2\pi}\,\sigma_i) . \qquad (11)$$

Protože druhý člen na pravé straně (11) na θ nezávisí, maximum věrohodnosti L nastává pro tutéž hodnotu $\hat{\theta}$ jako minimum součtu čtverců S. Obě metody odhadu jsou v tomto případě ekvivalentní.

Data, která mají přibližně normální rozdělení, se prakticky vyskytují velmi často; jejich zpracování budeme věnovat největší pozornost. Je-li rozdělení jiné a přitom známé, je obvykle výhodné využít metodu maximální

věrohodnosti. S odhadem parametrů z dat s rozdělením jiným než normálním se setkáme v §§ 13-15. V případě neznámého rozdělení dat je zpravidla preferována metoda nejmenších čtverců, díky jejím optimálním vlastnostem pro lineární modely (nezávisle na rozdělení). Formulace odhadu je jednoduchá a názorná, což jistě přispívá k popularitě této metody; používá se velmi často pro nelineární metody, kdy už diskutované optimální vlastnosti nemá.

Volba metody odhadu by měla být adekvátní důležitosti řešeného problému a náročnosti experimentální práce. Bylo by nesmyslné znehodnotit výsledky obtížných měření na drahých aparaturách jednoduchou neefektivní metodou. Na druhé straně je v mnoha situacích hledání optimální metody nepřiměřeně náročné, mnohem výhodnější může být použití málo efektivní metody s tím, že potřebnou přesnost zajistíme třeba opakováním měření.

Poznámka o inverzní pravděpodobnosti

V předchozích úvahách jsme hledaný parametr θ_0 považovali za pevnou, i když neznámou, charakteristiku měřeného objektu, která se projeví v rozdělení odhadu $\hat{\theta}$. Pomocí symbolu podmíněné pravděpodobnosti (§1) označíme hustotu odhadu $f(\hat{\theta}|\theta_0)$. Fakt, že různé hodnoty θ_0 vedou k různým rozdělelením odhadu umožňuje formulaci pravděpodobnostních závěrů o souvislosti hodnoty $\hat{\theta}$ získané z konkrétního měření s hledaným θ_0.

O problému hledání θ_0 se dá hovořit úplně jiným způsobem: pozorovaná hodnota $\hat{\theta}$ specifikuje, které z možných hodnot θ_0 jsou více a které méně pravděpodobné. Tento pohled na problém odhadu je vyjádřen zavedením rozdělení $p(\theta_0|\hat{\theta})$, ve kterém je θ_0 proměnnou a $\hat{\theta}$ podmínkou (obráceně, než v hořejší hustotě $f(\hat{\theta}|\theta_0)$). Pravděpodobnosti p se označují jako inverzní. Použití pojmu inverzní pravděpodobnosti může být velmi přitažlivé; otázka "jaká je pravděpodobnost toho, že skutečná hodnota je θ_0, když z měření vychází $\hat{\theta}$?" se zdá být položena správně. Manipulace s $p(\theta_0|\hat{\theta})$ je založena na Bayesově teorému (1.10), přesněji řečeno na jistém způsobu jeho interpretace. Nebudeme se tímto problémem zabývat, odkážeme pouze na podrobnou a zajímavou diskusi v knize [14]. Přidržíme se běžného chápání odhadované veličiny jako neznámé konstanty a inverzní pravděpodobnost $p(\theta_0|\hat{\theta})$ používat nebudeme.

11. Příklad měření časového intervalu

Ukážeme, jak se dají prostředky teorie pravděpodobnosti a matematické statistiky použít v konkrétním případě - při zpracování dat získaných ručním měřením známého časového intervalu t_0=2s z příkladu v § 2. Celý postup založíme na předpokladu, že se rozdělení naměřených hodnot dá dobře aproximovat normální hustotou; k posouzení vhodnosti této aproximace se vrátíme na konci tohoto paragrafu. Vše co se dá říci o četnostech možných výsledků měření je tedy obsaženo ve dvou parametrech normálního rozdělení

$$f(t_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(t_i-t_o)^2}{2\sigma^2}\right]. \qquad (1)$$

Střední hodnota je rovna hledané veličině t_o a disperze σ^2 (nebo standardní odchylka σ) charakterizuje chyby měření.

Zapomeňme na chvíli, že střední hodnotu t_o známe; naším úkolem je odhadnout t_o a σ^2 z naměřených hodnot t_i, i=1,...,200. Ptáme se: která čísla t_o a σ^2 v normálním rozdělení nejlépe souhlasí s tím, co jsme ve dvou stech měření zaregistrovali? Nejlepší dosud známá odpověď na tuto otázku je ta, že je třeba parametry najít tak, aby s nimi byla právě tato naměřená data nejvěrohodnější (§ 10). Protože předpokládáme nezávislost jednotlivých t_i, je hustota pravděpodobnosti N-tice výsledků rovna součinu hustot (1)

$$L = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(t_i-t_o)^2}{2\sigma^2}\right]. \qquad (2)$$

Maximum věrohodnosti L najdeme nejlépe jako maximum funkce

$$\ln L(t_o, \sigma^2) = -\sum_{i=1}^{N} \frac{(t_i-t_o)^2}{2\sigma^2} - \frac{N}{2}(\ln\sigma^2 + \ln 2\pi). \qquad (3)$$

Z podmínek maxima $\partial(\ln L)/\partial t_o = 0$, $\partial(\ln L)/\partial\sigma^2 = 0$ dostaneme odhady

$$\hat{t}_o = \frac{1}{N}\sum_{i=1}^{N} t_i, \qquad \hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(t_i-\hat{t}_o)^2. \qquad (4)$$

Podmínka maxima L vzhledem k t_o je totožná s podmínkou minima součtu čtverců odchylek t_i-t_o.

Z kompletních dat (N=200) vychází v našem příkladě $\hat{t}_o$=1.99281s, $\sqrt{\hat{\sigma}^2}$ =0.1335s. To jsou ovšem hodnoty náhodných proměnných - při opakování celého experimentu budou vycházet různě. Rozdělení $\hat{t}_o$ je podle (4.8) a (1) normální se střední hodnotou t_o a disperzí $D(\hat{t}_o)=\sigma^2/N$. Rozdělení náhodné proměnné $N\hat{\sigma}^2/\sigma^2$ je χ^2 s N-1 stupněm volnosti (§ 8); $\hat{\sigma}^2$ ze vztahu (4) se totiž dá napsat jako součet N-1 kvadrátů nezávislých lineárních kombinací veličin t_i, z nichž každá má střední hodnotu nula a disperzi σ^2. Instruktivní je ověření tohoto faktu pro N=2. Odhad $\hat{\sigma}^2$ je vychýlený protože střední hodnota rozdělení χ^2_{N-1} je N-1 (viz(8.2)), odkud vyjde střední hodnota $E(\hat{\sigma}^2) = \sigma^2(N-1)/N \neq \sigma^2$. Podmínka konzistence odhadu $\hat{\sigma}^2$ ovšem splněna je (pro $N\to\infty$ je $E(\hat{\sigma}^2)\to\sigma^2$). Je zřejmé, že nevychýleným odhadem disperze σ^2 je veličina

$$\hat{\sigma}^{2\prime} = \frac{N}{N-1}\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(t_i-\hat{t}_o)^2 , \qquad (4a)$$

pro větší N je ovšem rozdíl mezi odhady (4) a (4a) nepodstatný.

Znalost rozdělení $\hat{t}_o$ a $\hat{\sigma}^2$ umožňuje zformulovat výsledek měření, totiž pravděpodobnostní závěry o souvislosti $\hat{t}_o$ s hledanou hodnotou t_o. Náhodná proměnná

$$\frac{(\hat{t}_o-t_o)\sqrt{\hat{\sigma}^2/N}}{\sqrt{N\hat{\sigma}^2/[(N-1)\sigma^2]}} = \frac{\hat{t}_o-t_o}{\sqrt{\hat{\sigma}^2/(N-1)}} \qquad (5)$$

má podle (8.8) Studentovo rozdělení s N-1 stupněm volnosti. Označíme-li

$$\hat{\delta} = \sqrt{\frac{\hat{\sigma}^2}{N-1}} = \sqrt{\frac{1}{N(N-1)}\sum_{i=1}^{N}(t_i-\hat{t}_o)^2} , \qquad (6)$$

můžeme vypočíst pravděpodobnost, že t_o leží v intervalu $(\hat{t}_o-k\hat{\delta}, \hat{t}_o+k\hat{\delta})$:

$$P\left[t_o\in(\hat{t}_o-k\hat{\delta},\hat{t}_o+k\hat{\delta})\right] = P\left(\left|\frac{\hat{t}_o-t_o}{\hat{\delta}}\right|<k\right) = F_{N-1}(k)-F_{N-1}(-k) = 2F_{N-1}(k)-1, \qquad (7)$$

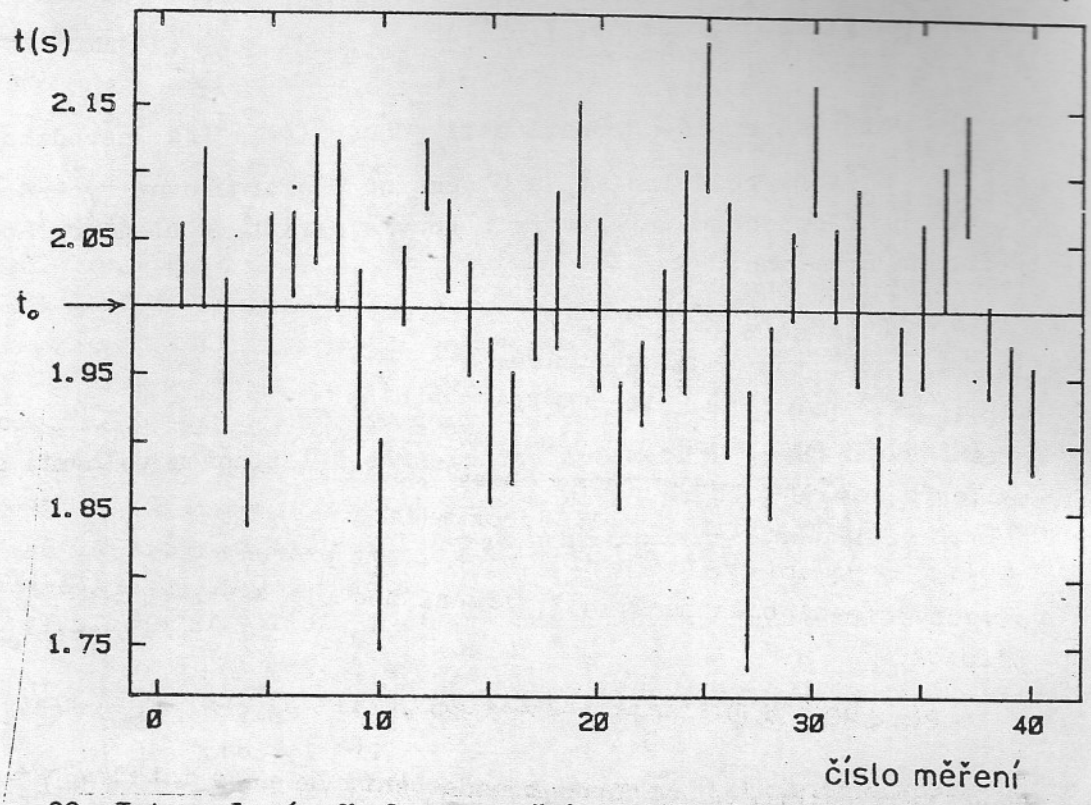
pomocí distribuční funkce F_{N-1} Studentova rozdělení s N-1 stupněm volnosti. Protože při N-1=199 je Studentovo rozdělení prakticky totožné s normálním, je pravděpodobnost (7) rovna 0.683 pro k=1 a 0.954 pro k=2 (srovnej s (4.6)). Měli bychom zachovat konvenci a udat jako výsledek měření intervalový odhad s pravděpodobnostním obsahem 0.683, kterým je $\hat{t}_o \pm \hat{\delta}$:

$$(1.9928 \pm 0.0095)s. \qquad (8)$$

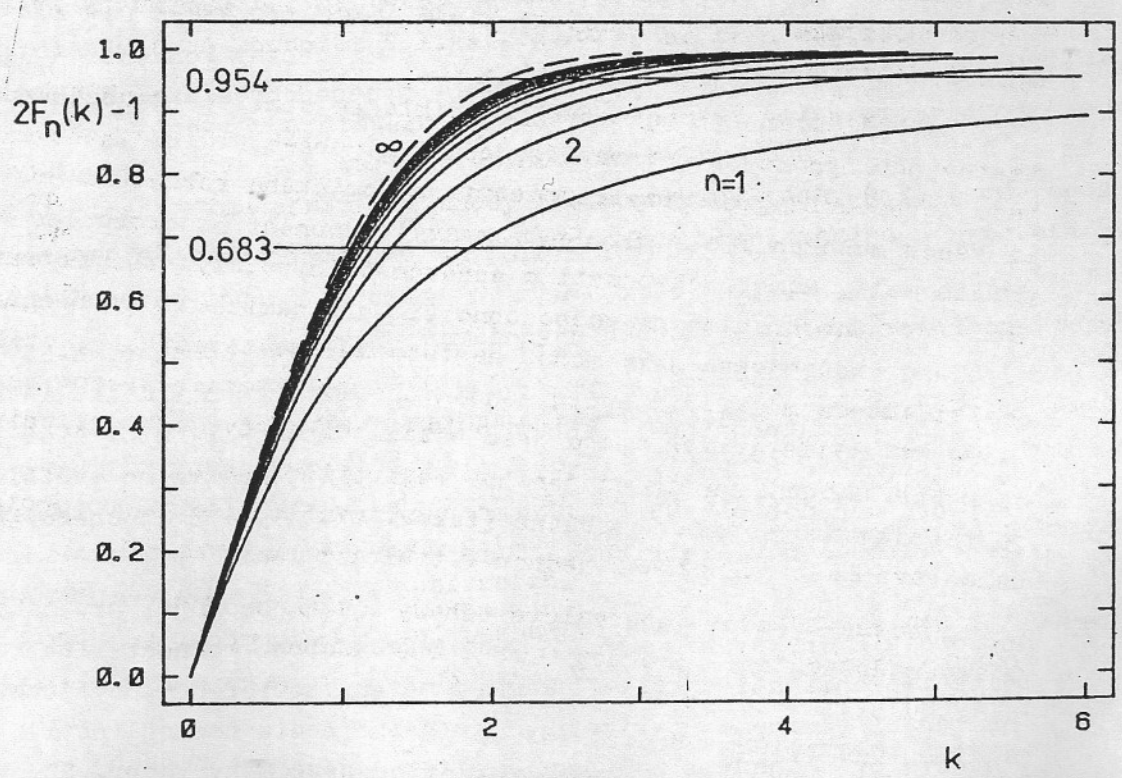
Shrneme smysl tohoto výsledku. S pravděpodobností 68.3% obsahují takto získané intervaly $\hat{t}_o \pm \hat{\delta}$ správnou hodnotu t_o (při mnohonásobném opakování bude v 68.3% případů t_o v mezích intervalového odhadu, ve zbylých 31.7% mimo ně). Přitom víme, jakým způsobem musíme změnit délku intervalu, aby se jeho pravděpodobnostní obsah změnil. Chceme-li "mít větší jistotu", že jsme správnou hodnotu v intervalu zachytili, musíme ho zvětšit. Vyhovuje-li pravděpodobnost 95.4%, vezmeme interval $\hat{t}_o \pm 2\hat{\delta}$; obecný návod je podle (7) obsažen v distribuční funkci t-rozdělení.

Odhad podle (8) správnou hodnotu t_o=2s obsahuje; hru náhody bychom pozorovali opakováním měření. Máme však dobrou možnost ilustrovat statistické zákonitosti tak, že budeme naměřený soubor dat považovat za opakované měření (řekněme čtyřicetkrát) menšího počtu hodnot (pěti). Výsledné intervalové odhady s N=5 jsou graficky znázorněny v obr. 22 svislými úsečkami, jejichž střed je v $\hat{t}_o$ a krajní body v $\hat{t}_o-\hat{\delta}$, $\hat{t}_o+\hat{\delta}$.

Z prvních čtyř pětic t_i například vyšly odhady



Obr. 22. Intervalové odhady t_o z pětic naměřených hodnot času.



Obr. 23. Distribuční funkce F(k) Studentova rozdělení. Nakresleny hodnoty $F_n(k)-F_n(-k)=2F_n(k)-1$ pro počty stupňů volnosti n=1,...,10 (plná čára) a limitní normální distribuční funkce (čárkovaná čá

$(2.030 \pm 0.032)s$, $(2.058 \pm 0.059)s$, $(1.963 \pm 0.057)s$, $(1.862 \pm 0.025)s$. (9)

V obrázku 22 vidíme názorně souvislost odhadu a správné hodnoty. Shodu pravděpodobnostních tvrzení o intervalových odhadech a pozorovanou skutečností posoudíme kvantitativně.

Ze čtyřiceti intervalů v obr. 22 jen dvacet obsahuje správnou hodnotu t_0. S pomocí tabulek v dodatku D2 zjistíme, že pro N-1=4 stupně volnosti je pravděpodobnost (7) zhruba P=0.62 (a ne 0.683 jako v případě velkého N!) pro interval $\hat{t}_0 \pm \hat{\delta}$, t.j. k=1. Počet n příznivých případů (interval obsahuje t_0) je náhodná veličina s binomickým rozdělením. Její střední hodnota je $40 \times 0.62 = 24.8$ a odmocnina z disperze $\sqrt{(40 \times 0.62 \times (1-0.62))} \approx 3.1$ (viz § 7). Kromě velmi pravděpodobných hodnot (25, 24, 26 atd.) se tedy při opakování celého pokusu občas objeví poněkud menší nebo větší počet příznivých případů. Použijeme-li aproximace binomického rozdělení normálním, dostaneme pomocí distribuční funkce (4.5) pravděpodobnost, že počet příznivých případů bude menší než střední hodnota alespoň o tolik co v našem případě:

$$P(n < 20.5) \approx \Phi\left(\frac{20.5-24.8}{3.1}\right) \approx 0.083 .$$

Není tedy celkem žádný důvod k podezření, že naše intervalové odhady nemají požadovaný význam. Pokud jde o počet příznivých případů v obr. 22, pozorovali jsme prostě výsledek, který se objeví zhruba jedenkrát v každých dvanácti opakováních.

Hořejší úvaha je jednoduchým příkladem statistického testu hypotézy. Testovanou hypotézou je pravděpodobnostní obsah P=0.62 intervalového odhadu, předpovídající prostřednictvím binomického rozdělení pravděpodobnosti všech možných výsledků. Je-li pravděpodobnost pozorovaného výsledku příliš malá, máme pochybnosti o správnosti hypotézy. Zřejmě není možné určit přesnou hranici pravděpodobnosti pro zamítnutí hypotézy, protože i málo pravděpodobné jevy mohou nastat. Musíme se smířit s omezenými možnostmi, které máme při studiu náhodných jevů. To samozřejmě neznamená, že výsledky statistických testů jsou bezcenné. Kdyby počet příznivých případů v obr. 22 byl například jen 15, měli bychom pádný důvod k domněnce, že je testovaná hypotéza nesprávná (takový případ, pokud je hypotéza správná, nenastává častěji než asi jednou v tisíci pokusech).

Chceme-li udat intervalové odhady s N=5 ve standardním tvaru, t.j. s pravděpodobností (7) rovnou 0.683, musíme zvolit k=1.142 (viz tabulku v dodatku D2). Pro P=0.954 je třeba $\hat{\delta}$ násobit faktorem k=2.858; z takto zvětšených intervalů v obr. 22 pak už jen tři (4., 12. a 33.) neobsahují hodnotu $t_0 = 2s$, což je v dobré shodě se střední hodnotou počtu příznivých případů $40 \times 0.954 \approx 38.2$. Stojí zato všimnout si odhadu ze čtvrté pětice (poslední v (9)); i toto se "náhodou" může stát (s pomocí tabulek v D2 zjistíme, že o něco méně jak v jednom ze sta pokusů). Důležitým faktem je nutnost

zvětšit interval $\hat{t}_o \pm \hat{\delta}$ faktorem k pro dosažení požadovaného pravděpodobnostního obsahu P, a to tím více, čím menší je počet stupňů volnosti N-1 a čím větší je P. To je vidět přehledně na distribučních funkcích Studentova rozdělení v obr. 23.

Rozdělení naměřených hodnot

Odhady (4) jsou založeny na předpokladu, že naměřené hodnoty mají normální rozdělení (1). To však může být nanejvýš aproximace: především víme, že t_i může nabývat pouze diskrétních hodnot, i když dosti jemně odstupňovaných. Dále je zřejmé, že výsledkem měření nebude nikdy záporné číslo, existuje jistě i horní hranice. Přesto je normální rozdělení v tomto případě velmi dobrou aproximací. Částečně je to vidět v histogramu v obr. 2. Je třeba si ovšem uvědomit, že četnosti v jednotlivých sloupcích jsou náhodné veličiny (s multinomickým rozdělením, § 9). Mohou, podle (9.4), silně kolísat kolem středních hodnot a tím ztěžovat srovnání s průběhem hustoty pravděpodobnosti. Lepší je použít tzv. empirické distribuční funkce

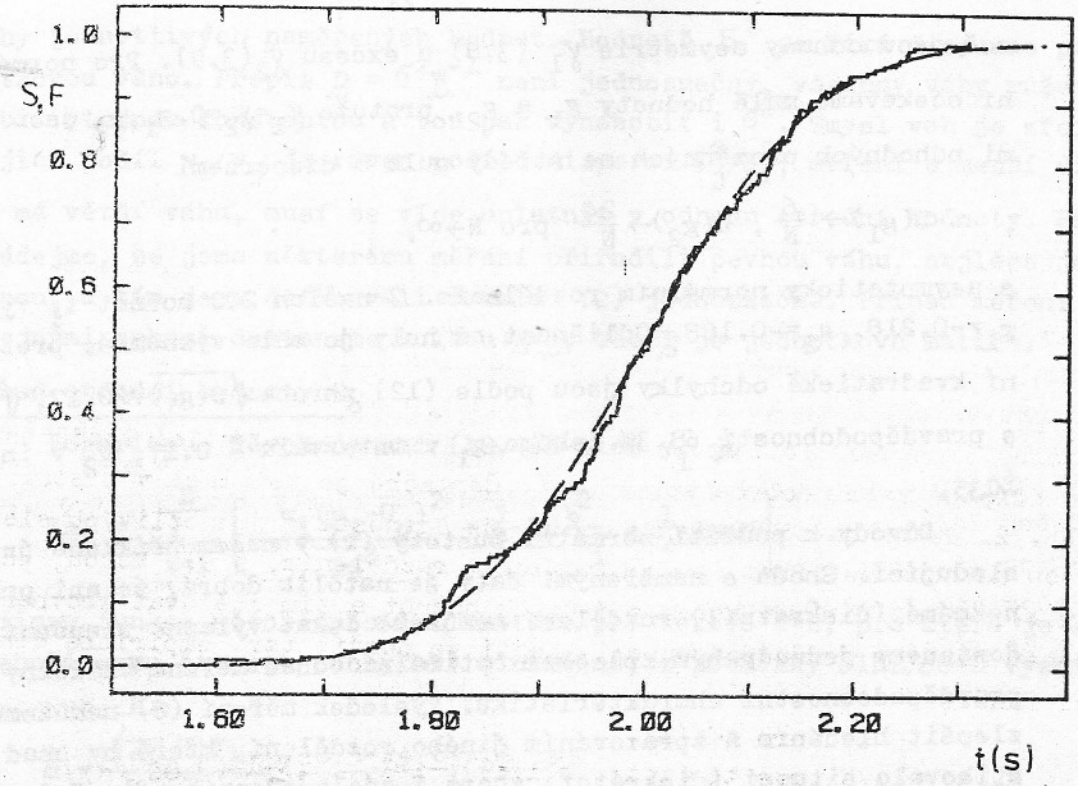
$$S_N(t) = \begin{cases} 0 & \text{pro } t<t(1), \\ i/N & \text{pro } t\in\langle t(i),\, t(i+1)),\ i=1,\dots,N-1, \\ 1 & \text{pro } t\geq t(N), \end{cases} \qquad (10)$$

kde $t(1),\dots,t(N)$ je N-tice výsledků měření $t_1,\dots,t_N$ uspořádaná podle velikosti od nejmenší k největší hodnotě. Funkce (10) má skok velikosti 1/N v každé naměřené hodnotě. V limitě $N\to\infty$ se blíží k distribuční funkci náhodné proměnné, která měření popisuje. Srovnání normální distribuční funkce se střední hodnotou a disperzí odhadnutou pomocí (4) s empirickou distribuční funkcí (10) je v obr. 24. Maxima odchylky obou závislostí se dá využít k přesnějšímu posouzení shody (Kolmogorovův test, § 21), zde se spokojíme s kvalitativním konstatováním souhlasu empirického a hypotetického rozdělení.

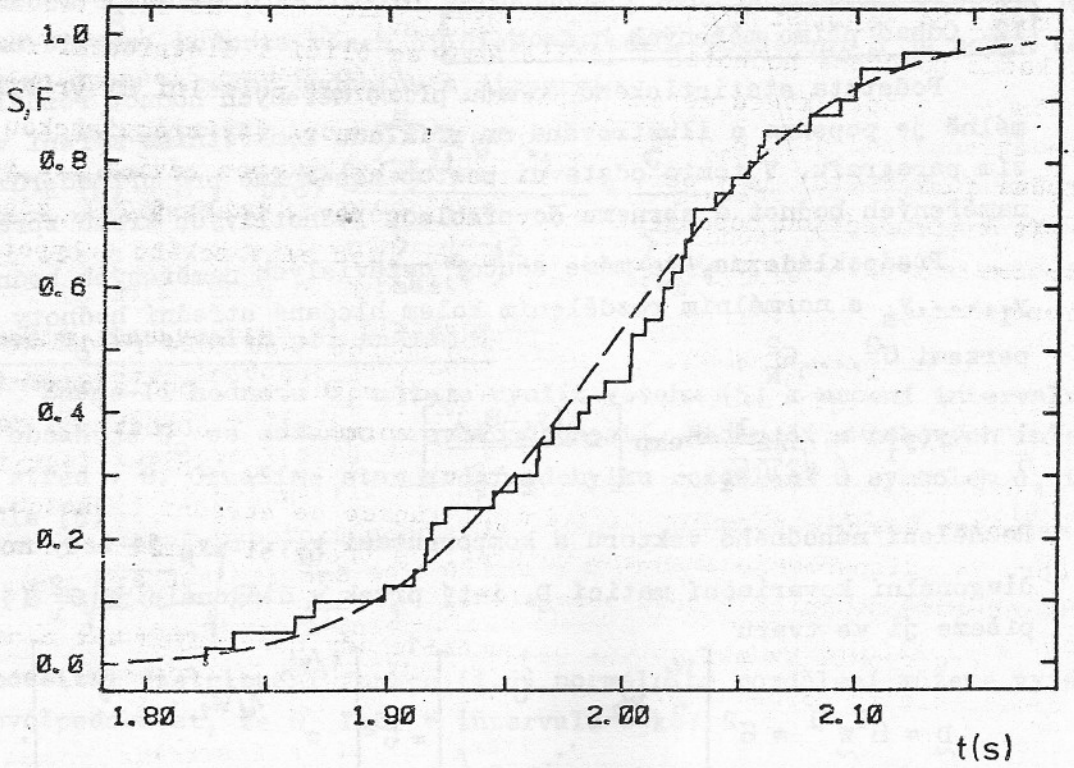
Aritmetické průměry $\hat{t}_o$ naměřených hodnot jsou také rozděleny zhruba normálně. I kdyby naměřená data měla jiné rozdělení, podle centrální limitní věty (§ 5) je průměr asymptoticky normální. V obrázku 25 je empirická distribuční funkce čtyřiceti průměrů z pětic po sobě následujících naměřených údajů a normální distribuční funkce se střední hodnotou jako v obr. 24, disperze je zmenšena pětkrát. Je třeba si všimnout různých měřítek časové stupnice v obou obrázcích.

Pro rychlé kvantitativní posouzení shody naměřených dat s normálním rozdělením je možné použít hodnot empirických koeficientů asymetrie a excesu

$$g_1 = \frac{\sum_{i=1}^{N}(t_i-t_o)^3/N}{\left[\sum_{i=1}^{N}(t_i-t_o)/N\right]^{3/2}}, \qquad g_2 = \frac{\sum_{i=1}^{N}(t_i-t_o)^4/N}{\left[\sum_{i=1}^{N}(t_i-t_o)^2/N\right]^2} - 3, \qquad (11)$$



Obr. 24. Empirická distribuční funkce z dvou set naměřených časů (plná čára) a hypotetická normální distribuční funkce (čárkovaná čára).



Obr. 25. Empirická distribuční funkce průměrů z 5-ti hodnot času (plná čára) a hypotetická normální distribuční funkce (čárkovaná čára).

což jsou odhady asymetrie γ_1 (3.8) a excesu γ_2(3.9). Pro normální rozdělení očekáváme malé hodnoty g_1 a g_2, protože $\gamma_1=\gamma_2=0$. g_1, g_2 jsou ale hodnotami náhodných proměnných se středy nula a disperzemi

$$D(g_1)\to\frac{6}{N},\quad D(g_2)\to\frac{24}{N}\quad \text{pro } N\to\infty, \tag{12}$$

s asymptoticky normálním rozdělením. Z našich 200 hodnot t_i vychází $g_1=-0.218$, $g_2=-0.108$. Odlišnost od nuly je málo významná, protože střední kvadratické odchylky jsou podle (12) zhruba $\sqrt{D(g_1)}\approx 0.17$, $\sqrt{D(g_2)}\approx 0.35$ s pravděpodobností 68.3% čekáme g_1 v intervalu $\pm$ 0.17, g_2 v intervalu ± 0.35.

Důvody k použití normální hustoty (1) v našem příkladě jsou tedy následující. Shoda s naměřenými daty je natolik dobrá, že ani pro skutečné neznámé (diskrétní!) rozdělení nemůžeme čekat výrazné zlepšení. Přitom dostaneme jednoduchým způsobem optimální odhad měřené veličiny a jeho pravděpodobnostní charakteristiku. Výsledek měření (8) nemůžeme podstatně zlepšit hledáním a zpracováním jiného rozdělení, které by snad lépe vystihovalo situaci (diskrétní, zhora i zdola ohraničené). Jednou větou: normální rozdělení je zde aproximací, která poskytuje vše co k dosažení cíle potřebujeme.

12. Odhad přímo měřených hodnot

Podstata statistického odhadu přímo měřených hodnot rozdělených normálně je popsána a ilustrována na příkladu konkrétního měření v předchozím paragrafu. V tomto odstavci postup zobecníme pro případ různých vah naměřených hodnot a shrneme do přehledu jednotlivých kroků zpracování dat.

Předpokládejme, že máme soubor nezávislých naměřených hodnot $y_1,\dots,y_n$ s normálním rozdělením kolem hledané střední hodnoty θ_o a s disperzemi $\sigma_1^2,\dots,\sigma_N^2$:

$$f(y_i)=\frac{1}{\sqrt{2\pi}\,\sigma_i}\exp\left[-\frac{(y_i-\theta_o)^2}{2\sigma_i^2}\right]. \tag{1}$$

Rozdělení náhodného vektoru s komponentami $y_1,\dots,y_N$ je tedy normální s diagonální kovariační maticí $\underset{\sim}{D}$, i-tý prvek v diagonále je σ_i^2 (§ 6); zapíšeme ji ve tvaru

$$\underset{\sim}{D}=\sigma^2\underset{\sim}{W}^{-1}=\sigma^2\begin{bmatrix} w_1 & & 0\\ & w_2 & \\ & \ddots & \\ 0 & & w_N\end{bmatrix}^{-1}=\sigma^2\begin{bmatrix} 1/w_1 & & 0\\ & 1/w_2 & \\ & \ddots & \\ 0 & & 1/w_N\end{bmatrix}. \tag{2}$$

$\underset{\sim}{W}$ je matice vah; jejími prvky jsou kladná čísla $w_i=\sigma^2/\sigma_i^2$ označovaná jako

váhy jednotlivých naměřených hodnot. Hodnotě σ^2 se říká disperze pro jednotkovou váhu. Přepis $\underset{\sim}{D} = \sigma^2 \underset{\sim}{W}^{-1}$ není jednoznačný, všechny váhy můžeme násobit stejnou konstantou a tou pak vynásobit i σ^2. Smysl vah je zřejmý: jejich podíl w_i/w_j je roven podílu disperzí σ_j^2/σ_i^2; měření s menší disperzí má větší váhu, musí se více uplatnit v odhadu střední hodnoty. Předpokládejme, že jsme některému měření přiřadili pevnou váhu, nejlépe jednotkovou, a tím jsme definovali rozklad (2) jednoznačně. Případ měření se stejnými vahami dostaneme volbou $\underset{\sim}{W}=\underset{\sim}{I}$, kde $\underset{\sim}{I}$ je jednotková matice.

Odhad střední hodnoty θ_o

Logaritmus věrohodnosti (10.4)N-tice y_i je

$$\ln L = -\sum_{i=1}^{N} \left[\frac{w_i (y_i-\theta_o)^2}{2\sigma^2} + \frac{1}{2} \ln \frac{\sigma^2}{w_i} + \frac{1}{2} \ln 2\pi \right]. \qquad (3)$$

Maximum funkce věrohodnosti L nastane pro takové $\theta=\hat{\theta}$, pro které je $-\ln L$ (neboli suma čtverců odchylek) minimální; z podmínky $\partial \ln L/\partial\theta=0$ vychází

$$\hat{\theta} = \frac{\sum_{i=1}^{N} w_i y_i}{\sum_{i=1}^{N} w_i} . \qquad (4)$$

Odhadem θ_o metodou maximální věrohodnosti nebo nejmenších čtverců je vážená střední hodnota všech dílčích výsledků. Rozdělení $\hat{\theta}$ je podle (4.8) normální se střední hodnotou a disperzí

$$E(\hat{\theta}) = \theta_o, \quad D(\hat{\theta}) = \frac{\sum_{i=1}^{N} w_i^2 \sigma^2/w_i}{(\sum_{i=1}^{N} w_i)^2} = \frac{\sigma^2}{\sum_{i=1}^{N} w_i} . \qquad (5)$$

Odhad θ_o intervalem při známém σ

Známe-li hodnotu σ, můžeme využít vztahu (5) k určení intervalu, který obsahuje θ_o se zadanou pravděpodobností. Nejkratší z takových intervalů má střed v $\hat{\theta}$. Označíme standardní odchylku rozdělení $\hat{\theta}$ symbolem δ, tedy podle (5)

$$\delta = \sqrt{D(\hat{\theta})} = \sqrt{\sigma / \sum_{i=1}^{N} w_i} . \qquad (6)$$

S použitím distribuční funkce (4.5) normálního rozdělení můžeme vyjádřit pravděpodobnost, že θ_o leží v intervalu $\hat{\theta} \pm k\delta$:

$$P\left[\theta_o \in (\hat{\theta}-k\delta,\ \hat{\theta}+k\delta)\right] = P\left(\left| \frac{\hat{\theta}-\theta_o}{\delta} \right| < k \right) = \Phi(k)-\Phi(-k) = 2\Phi(k)-1. \qquad (7)$$

K požadované pravděpodobnosti P tedy najdeme potřebný násobek k standardní odchylky δ. Pokud k intervalovému odhadu nepodáme jiné vysvětlení, měl

by mít pravděpodobnostní obsah P=0.683 a tedy k=1 (interval $\hat{\theta}\pm\delta$); interval $\hat{\theta}\pm 2\delta$ má P=0.954, viz (4.6).

Odhad neznámé hodnoty σ^2

Pokud disperzi pro jednotkovou váhu neznáme, můžeme ji odhadnout opět z podmínky maxima věrohodnosti L, čili $\partial \ln L/\partial(\sigma^2) = 0$. Ze vztahu (3) vyjde

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N} w_i(y_i-\hat{\theta})^2 . \qquad (8)$$

$\hat{\sigma}^2$ je ovšem náhodná veličina. Dá se dokázat, že $N\hat{\sigma}^2/\sigma^2$ má χ^2 rozdělení s N-1 stupněm volnosti; lze ji vyjádřit jako součet N-1 kvadrátů nezávislých lineárních kombinací jednotlivých y_i, z nichž každá má střední hodnotu nula a disperzi 1. Dvojice $y_i-\hat{\theta}$, $y_j-\hat{\theta}$ nezávislé nejsou, každé $y_1,\ldots,y_N$ vystupuje ve vztahu (4) pro $\hat{\theta}$. Odhad (8) je vychýlený, protože jeho střední hodnota není rovna odhadovanému σ^2:
$E(\hat{\sigma}^2) = (\sigma^2/N)E(N\hat{\sigma}^2/\sigma^2) = \sigma^2(N-1)/N < \sigma^2$. Nevychýlený odhad tedy může být

$$\hat{\sigma}^{2\prime} = \frac{N}{N-1}\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N} w_i\,(y_i-\hat{\theta})^2, \qquad (8a)$$

není to už ale odhad nejvěrohodnější. $(N-1)\hat{\sigma}^{2\prime}/\sigma^2$ má rozdělení χ^2_{N-1}, pro větší N je rozdíl mezi odhady (8) a (8a) nepodstatný. Znalost rozdělení proměnných $\hat{\sigma}^2$ resp. $\hat{\sigma}^{2\prime}$ umožňuje formulaci intervalových odhadů hodnoty σ^2 s předepsaným pravděpodobnostním obsahem (odhadnout "chybu $\sigma(\hat{\sigma})$ chyby $\hat{\sigma}$"). K tomu stačí využít distribuční funkce rozdělení χ^2_{N-1}.

Odhad $\hat{\theta}_o$ intervalem hodnot při neznámém σ

Je-li hodnota σ neznámá a odhadujeme ji z naměřených dat, je konstrukce intervalového odhadu pro θ_o o něco složitější než v předchozím případě (7). Využijeme faktu, že podíl

$$\frac{(\hat{\theta}-\theta_o)/\sqrt{D(\hat{\theta})}}{\sqrt{N\hat{\sigma}^2/[\sigma^2(N-1)]}} = \frac{\hat{\theta}-\theta_o}{\sqrt{N\hat{\sigma}^2/\left[(N-1)\sum_{i=1}^{N} w_i\right]}} \qquad (9)$$

má Studentovo rozdělení s N-1 stupněm volnosti (§ 8). S označením

$$\hat{\delta} = \sqrt{\frac{N\hat{\sigma}^2}{(N-1)\sum_{i=1}^{N} w_i}} = \sqrt{\frac{1}{(N-1)\sum_{i=1}^{N} w_i}\sum_{i=1}^{N} w_i(y_i-\hat{\theta})^2} \qquad (10)$$

dostaneme pravděpodobnost

$$P\left[\theta_o\in(\hat{\theta}-k\hat{\delta},\hat{\theta}+k\hat{\delta})\right] = P\left(\left|\frac{\hat{\theta}-\theta_o}{\hat{\delta}}\right|<k\right) = F_{N-1}(k)-F_{N-1}(-k) = 2F_{N-1}(k)-1 \qquad (11)$$

vyjádřenou pomocí distribuční funkce F_{N-1} Studentova rozdělení s N-1 stupněm volnosti. Přehled o souvislosti k a P poskytuje obr. 23. V limitě $N\to\infty$ je t-rozdělení normální, $\hat{\delta}$ ze vztahu (10) je standardní odchylkou odhadu $\hat{\theta}$ a interval $\hat{\theta}\pm\hat{\delta}$ (t.j. pro k=1) má pravděpodobnostní obsah P=0.683. Při zmenšování N je třeba pro zachování P zvětšovat k, a to výrazněji pro větší pravděpodobnosti (obr. 23 nebo tabulka v D2).

Kontrola rozdělení dat

Správnost předpokladu o normálním rozdělení (1) měřených hodnot se dá účinně posoudit pomocí statistických testů dobré shody (kapitola III). Zde se omezíme na orientační posouzení normality souboru dat pomocí empirických koeficientů asymetrie a excesu

$$g_1=\sqrt{N}\,\frac{\sum_{i=1}^{N} w_i(y_i-\hat{\theta})^3}{\left[\sum_{i=1}^{N} w_i(y_i-\hat{\theta})^2\right]^{3/2}}\,,\qquad g_2=N\,\frac{\sum_{i=1}^{N} w_i(y_i-\hat{\theta})^4}{\left[\sum_{i=1}^{N} w_i(y_i-\hat{\theta})^2\right]^{2}}-3\,. \tag{12}$$

Jsou to odhady asymetrie (3.8) a excesu (3.9), které jsou pro normální rozdělení nulové.
Disperze náhodných proměnných g_1, g_2 jsou

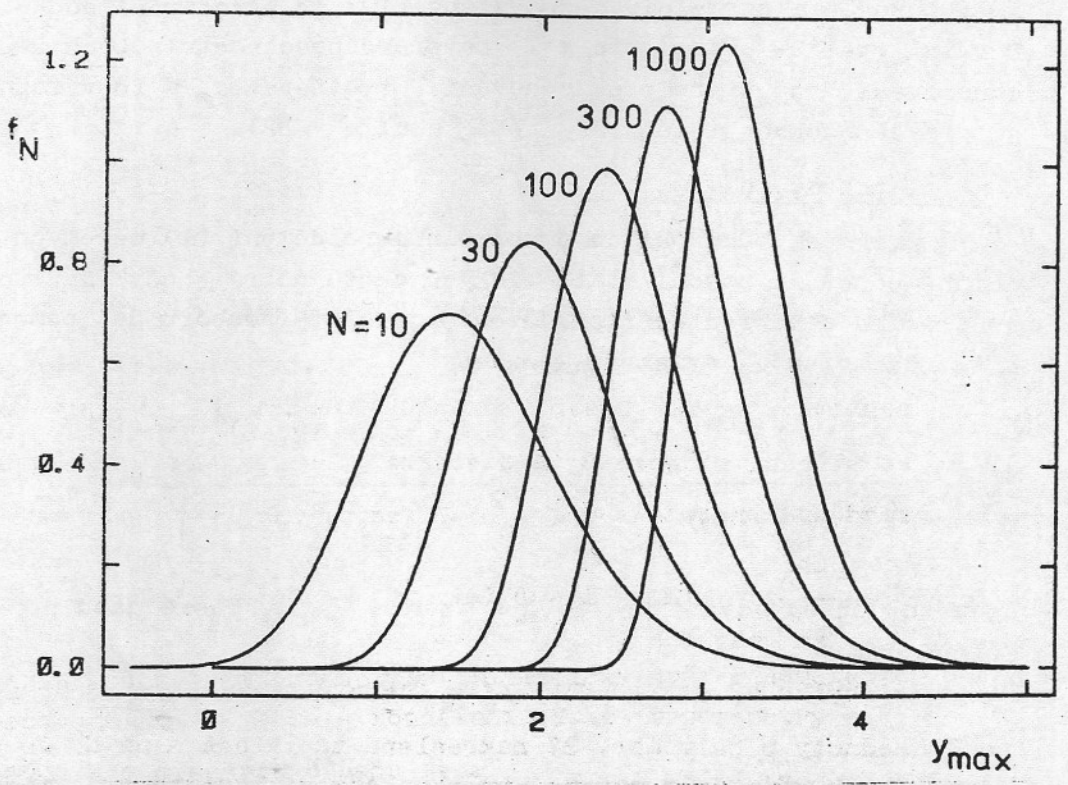
$$D(g_1)=\frac{6N(N-1)}{(N-2)(N+1)(N+3)}\,,\quad D(g_2)=\frac{24N(N-1)^2}{(N-3)(N-2)(N+3)(N+5)}, \tag{13}$$

rozdělení g_1 a g_2 jsou asymptoticky normální. Vyjdou-li hodnoty (12) daleko od nuly, máme podezření, že rozdělení dat normální není. Výsledek $|g_1|\gtrsim 2\sqrt{D(g_1)}$ nebo $|g_2|\gtrsim 2\sqrt{D(g_2)}$ většinou považujeme za významný nesouhlas s předpokladem, protože takový případ nastává při normálně rozdělených datech zřídka (méně jak v 5% případů).

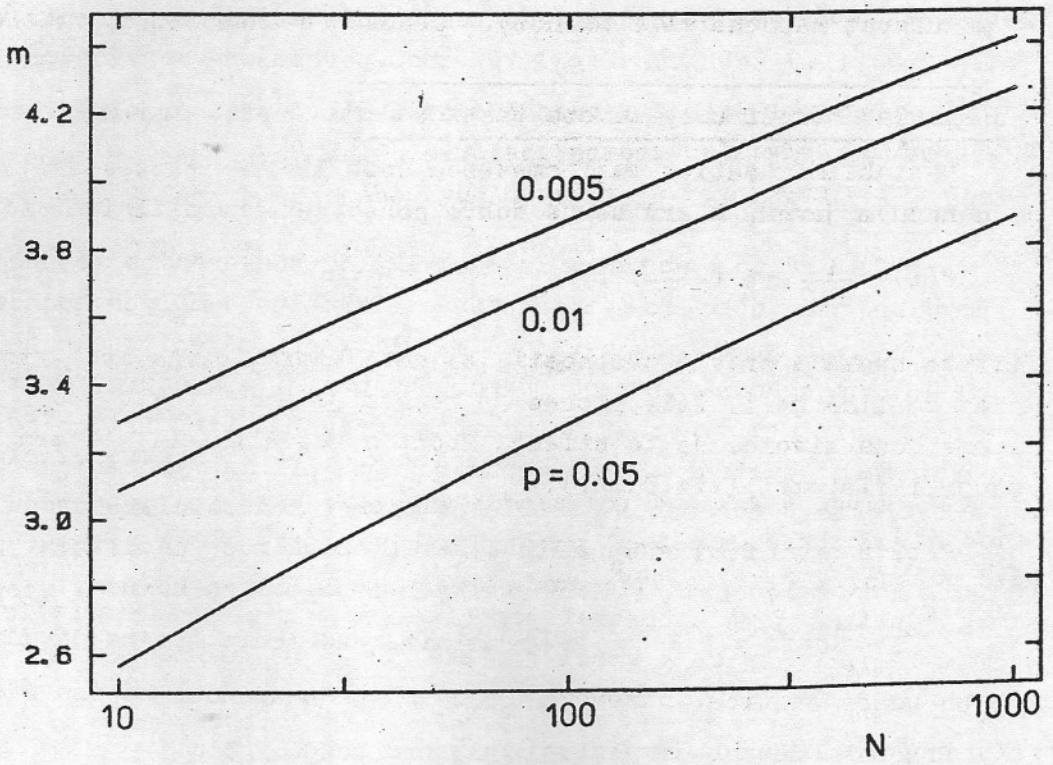
Nápadně vybočující hodnoty

Velmi často se stane, že odchylku od očekávaného rozdělení dat způsobuje jen jedna nebo několik málo nápadně velkých nebo malých hodnot. Jejich přítomnost bývá způsobena nežádoucími vlivy při měření, jako je chybný zápis údaje nebo náhodná krátkodobá porucha měřicí aparatury. Je možné nápadně vybočující data vynechat a tím měření "zachránit". Přitom je třeba postupovat velmi opatrně a s uvážením možných příčin vybočení, protože vynechání dat, které do souboru patří, je rovněž nežádoucí. Rozhodující je znalost konkrétního procesu měření, statistika může poskytnout pomocná kriteria pro vyloučení nečekaně velkých nebo malých hodnot.

Pro údaje $y_1,\ldots,y_N$ s normálním rozdělením (1) se dá snadno najít rozdělení maximální hodnoty y_{max}. Pro jednoduchost zápisu dvou následujících formulí položíme $\theta_o=0$ a $\sigma_i^2=1$; pro každé y platí



Obr. 26. Hustota největší z N-tice nezávislých hodnot se standardním normálním rozdělením.



Obr. 27. Řešení m rovnice (12.16) v závislosti na počtu N normálně rozdělených hodnot; p je pravděpodobnost, že $y_{max} > \Theta_0 + m\sigma$.

$$P(y_{max}<y) = P(y_1<y \text{ a } \dots \text{ a } y_N<y) = P(y_1<y)\dots P(y_N<y) = [\Phi(y)]^N \ . \quad (14)$$

To je distribuční funkce $F_N(y_{max})$ maximální hodnoty; vyjde tedy rovna N-té mocnině integrálu pravděpodobnosti (4.5). Odtud dostaneme, s pomocí (2.9), hustotu

$$f_N(y_{max}) = \frac{d}{dy_{max}} F_N(y_{max}) = N\Phi^{N-1}(y_{max})\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y_{max}^2}{2}\right) \ . \quad (15)$$

Hustoty (15) jsou nakresleny v obr. 26. S rostoucím N se rozdělení y_{max} posouvá k větším hodnotám a zužuje se. Například pro N=1000 je s velkou pravděpodobností $y_{max} \in (2.5, 4.5)$, pravděpodobnost $y_{max} > 4$ je malá. Vrátíme-li se ke střední hodnotě Θ_o a disperzi σ_i^2 v rozdělení (1), dostaneme ze (14) pravděpodobnost

$$p = P\left[\left(\frac{y_i-\Theta_o}{\sigma_i}\right)_{max} > m\right] = 1 - \Phi^N(m) \quad (16)$$

toho, že odchylka y_i od Θ_o překročí m-násobek standardní odchylky σ_i. Pro tři malé hodnoty p je v obr. 27 nakreslena závislost m na N. Je vidět, že v odůvodněných případech můžeme vynechat takové hodnoty y_i, které jsou o 3-až 4- násobek σ_i větší než odhad $\hat{\Theta}_o$ (počítaný ovšem bez vynechávaných hodnot); pravděpodobnost, že do souboru patří je velmi malá. Stejně se dají posuzovat nápadně malé hodnoty - menší o m- násobek σ_i než $\hat{\Theta}_o$.

13. Příklad měření doby života částice

Nestabilní částice mají omezenou dobu života - rozpadají se. Rozpad je náhodným jevem, který se dá dobře popsat exponenciálním rozdělením (§9):

$$f(t) = \frac{1}{\tau_o}\exp\left(-\frac{t}{\tau_o}\right) \ . \quad (1)$$

f(t) je hustota pravděpodobnosti, že doba které uplyne mezi vznikem a rozpadem částice je t. Celý proces je popsán jedinou konstantou τ_o, které říkáme doba života. Je to střední hodnota rozdělení (1), která zároveň určuje i disperzi (viz(9.12)):

$$E(t) = \tau_o, \quad D(t) = \tau_o^2 \ . \quad (2)$$

Život částic můžeme pozorovat pomocí stopy v registračním zařízení, stopa začíná v místě vzniku a končí v místě rozpadu. Dokážeme-li určit rychlost pohybu každé částice, můžeme z délky stopy vypočíst dobu mezi vznikem a rozpadem.

Předpokládejme, že jsme sledovali N částic a získali N-tici nezávislých hodnot $t_1, \dots t_N$. I když se nám podařilo potlačit náhodné chyby v

procesu měření na zanedbatelnou úroveň, jsou t_i náhodná čísla s rozdělení (1). Studujeme náhodný jev; cílem měření je určení konstanty τ_o, jejíž hodnota umožňuje předpovídat pravděpodobnosti prostřednictvím hustoty (1)

Optimální odhad $\hat{\tau}_o$ dostaneme z maxima věrohodnosti (10.3) naměřených nezávislých t_i:

$$L(\tau) = \prod_{i=1}^{N} \frac{1}{\tau} \exp\left(-\frac{t_i}{\tau}\right), \tag{3}$$

neboli z maxima funkce

$$\ln L = \frac{1}{\tau}\sum_{i=1}^{N} t_i + N\ln\tau \tag{4}$$

Maximum nastává pro hodnotu $\tau=\hat{\tau}_o$, pro kterou $\partial(\ln L)/\partial\tau = 0$:

$$\hat{\tau}_o = \frac{1}{N}\sum_{i=1}^{N} t_i . \tag{5}$$

Nejvěrohodnějším odhadem je aritmetický průměr naměřených časů. Náhodná proměnná $N\hat{\tau}_o$ má gama - rozdělení (9.16) s parametry b=N, $a=1/\tau_o$. Střední hodnota a disperze odhadu jsou

$$E(\hat{\tau}_o) = \tau_o, \quad D(\hat{\tau}_o) = \sum_{i=1}^{N} \frac{1}{N^2} D(t_i) = \sum_{i=0}^{N} \frac{\tau_o^2}{N^2} = \frac{\tau_o^2}{N}. \tag{6}$$

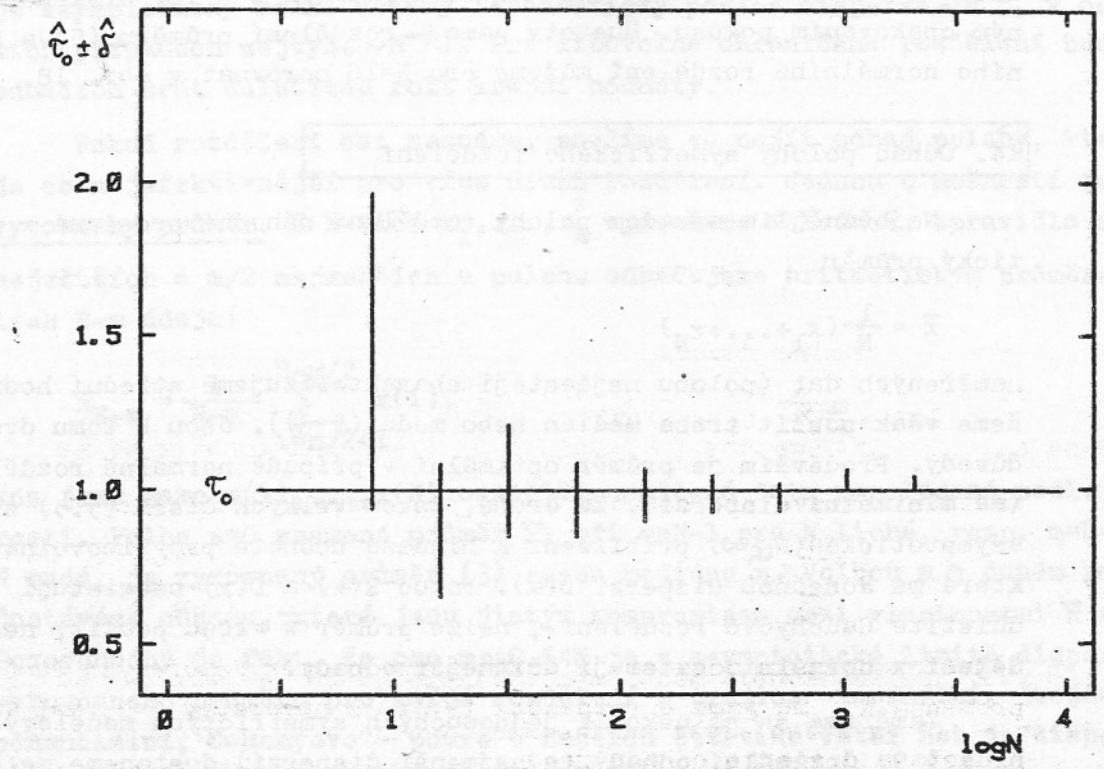
Odhad (5) je konzistentní a nevychýlený; jeho rozdělení je asymptoticky normální.

Skutečné měření dob života částic vyžaduje mohutné experimentální z řízení. My se spokojíme se simulací pomocí počítače. Vhodným programem g nerujeme N-tice pseudonáhodných čísel x_i s rovnoměrným rozdělením $g(x_i)=1$ pro $x_i \in (0,1)$. Transformací

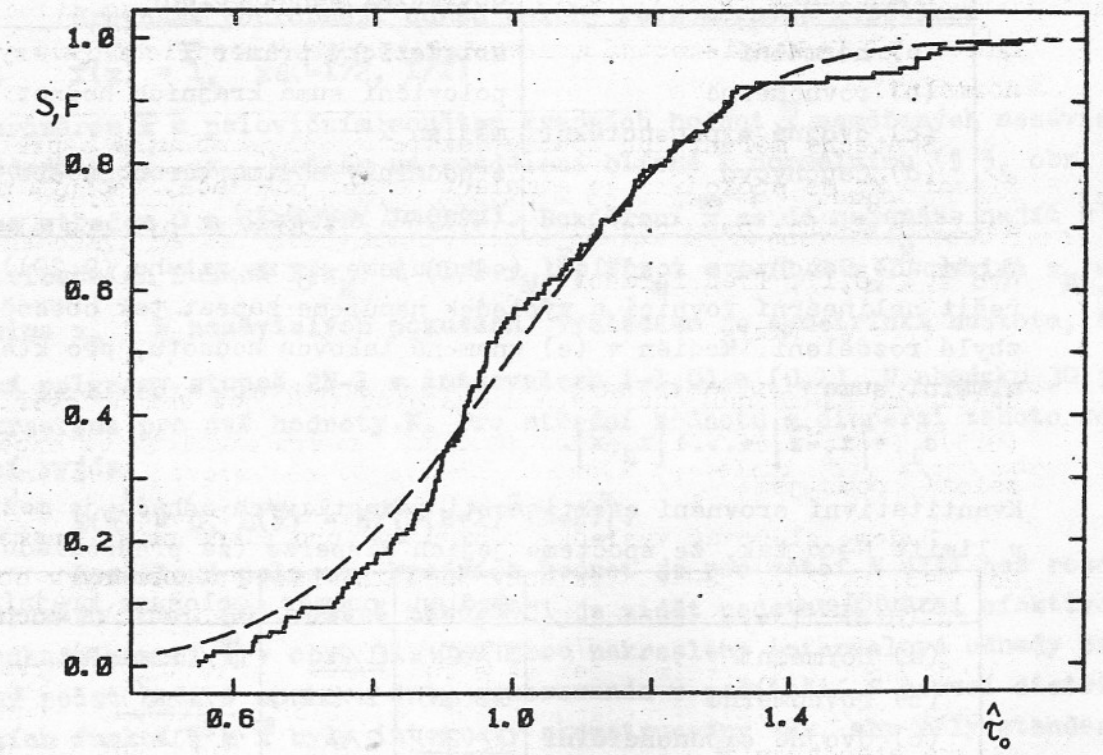
$$t_i = -\tau_o \ln x_i \tag{7}$$

dostaneme pseudonáhodné čísla s hustotou (1), což snadno ověříme pomocí (2.5). Protože na volbě jednotek času v našem simulovaném experimentu ne záleží, použijeme $\tau_o=1$.

Budeme sledovat výsledky odhadu (5) pro různé počty "naměřených" ho not N=8,16,32,...,2048. Výsledky jsou graficky znázorněny v obr. 28 ve tvaru intervalových odhadů $\hat{\tau}_o \pm \hat{\delta}$, kde střední kvadratická odchylka podl (6) je $\hat{\delta} \approx \hat{\tau}_o/\sqrt{N}$. Vidíme, jak se s rostoucím N zkracují intervaly (velmi pomalu, jako $1/\sqrt{N}$). Pravděpodobnost, že obsahují hledanou hodnotu τ_o je blízká standardním 0.683. Ačkoliv mají data t_i rozdělení (1) podstatně lišné od normálního, rozdělení průměru (5) se podle centrální limitní v ty normálnímu blíží, a to tím více, čím větší je N. I pro dosti malá N aproximace normálním rozdělením dobrá. Ukazuje to obr. 29, kde je empi



Obr. 28. Intervalové odhady τ_o s různým počtem N měřených hodnot.



Obr. 29. Empirická distribuční funkce odhadů doby života částice pro N=16, plná čára; normální distribuční funkce N(1,1/16), čárkovaná čára.

rická distribuční funkce (viz(11.10)) průměrů s N=16, získaná 128-násobným opakováním pokusu. Hustoty gama - rozdělení průměru (5) a aproximativního normálního rozdělení můžeme pro N=16 porovnat v obr. 18.

14. Odhad polohy symetrického rozdělení

Nejběžnějším odhadem polohy rozdělení náhodné proměnné x je aritmetický průměr

$$\bar{x} = \frac{1}{N}(x_1+\ldots+x_N) \qquad (1)$$

naměřených dat (polohu nejčastěji charakterizujeme střední hodnotou, můžeme však použít třeba medián nebo modu (§ 3)). Jsou k tomu dva podstatné důvody. Především je průměr optimální v případě normálně rozdělených dat (má minimální disperzi). Za druhé, zákon velkých čísel (5.5) zaručuje asymptotické ($N\to\infty$) přiblížení k hledané hodnotě pro libovolné rozdělení které má konečnou disperzi D(x). Pokud E(x) a D(x) neexistují (třeba pro důležité Cauchyovo rozdělení), nelze průměr $\bar{x}$ vůbec použít; není-li rozdělení x normální, existují účinnější odhady.

Zaměříme se na několik jednoduchých symetrických modelových rozdělení z § 9. Optimální odhady (s nejmenší disperzí) dostaneme metodou maximální věrohodnosti:

rozdělení	optimální odhad polohy
(a) normální	aritmetický průměr $\bar{x}$
(b) rovnoměrné	poloviční suma krajních hodnot $\bar{\bar{x}}$
(c) dvojné exponenciální	medián $\tilde{x}$
(d) Cauchyovo	z podmínky maxima věrohodnosti (řešení n lineární rovnice)

V případě Cauchyova rozdělení (odhadujeme x_0 ze vztahu (9.20)) je třeba řešit nelineární rovnici a výsledek nemůžeme zapsat tak obecně, jako pr zbylá rozdělení. Medián v (c) znamená takovou hodnotu, pro kterou je mi nimální suma

$$s_1 = \left|x_1-\tilde{x}\right| + \ldots + \left|x_N-\tilde{x}\right|. \qquad (2)$$

Kvantitativní srovnání efektivnosti jednotlivých odhadů je možné provés v limitě $N\to\infty$ tak, že spočteme jejich disperze (za předpokladu D(x)=1):

rozdělení	medián	průměr	polosuma krajních hodnot
(a) normální	$\pi/(2N)$	$1/N$	$\pi^2/(12\ \ln N)$
(b) rovnoměrné	$1/(4N)$	$1/(12N)$	$1/(2N^2)$
(c) dvojné exponenciální	$1/(2N)$	$2/N$	$\pi^2/12$
(d) Cauchyovo	$\pi^2/(4N)$	∞	∞

Podtržením jsou označeny nejmenší možné disperze, viz první tabulku v

tomto odstavci. Poloviční součet krajních hodnot pro rovnoměrné rozdělení je asymptoticky velmi účinný (kvadratický pokles disperze $\sim N^{-2}$, v ostatních případech nejvýše $\sim N^{-1}$). Pro libovolné ohraničené rozdělení budou v odhadech hrát důležitou roli krajní hodnoty.

Pokud rozdělení dat neznáme, snažíme se najít odhad polohy, který bude co nejefektivnější pro více druhů rozdělení. Jednou z možností je tzv. vyrovnaný průměr. Z N-tice $x_1,\ldots,x_N$ vynecháme m hodnot, zpravidla m/2 největších a m/2 nejmenších a polohu odhadujeme aritmetickým průměrem zbylých N-m údajů:

$$\bar{x}_{N-m} = \frac{1}{N-m} \sum_{i=m/2+1}^{N-m/2} x(i), \qquad (3)$$

kde jsme jako $x(1),\ldots,x(N)$ označili naměřená data uspořádaná podle velikosti. Volba m=0 znamená průměr $\bar{x}$; při m=N-1 pro N liché, resp. m=N-2 pro N sudé, je vyrovnaný průměr (3) roven mediánu $\tilde{x}$. Volbou m v daném rozmezí dostáváme odhady, které jsou jistým kompromisem mezi vlastnostmi $\bar{x}$ a $\tilde{x}$. Pozoruhodný je fakt, že pro $m \approx 0.54N$ je v asymptotické limitě disperze vyrovnaného průměru pro každé rozdělení z trojice - normální, dvojné exponenciální, Cauchyovo - pouze o necelou čtvrtinu větší než je disperze příslušného optimálního odhadu.

Příklad odhadu střední hodnoty rovnoměrného rozdělení

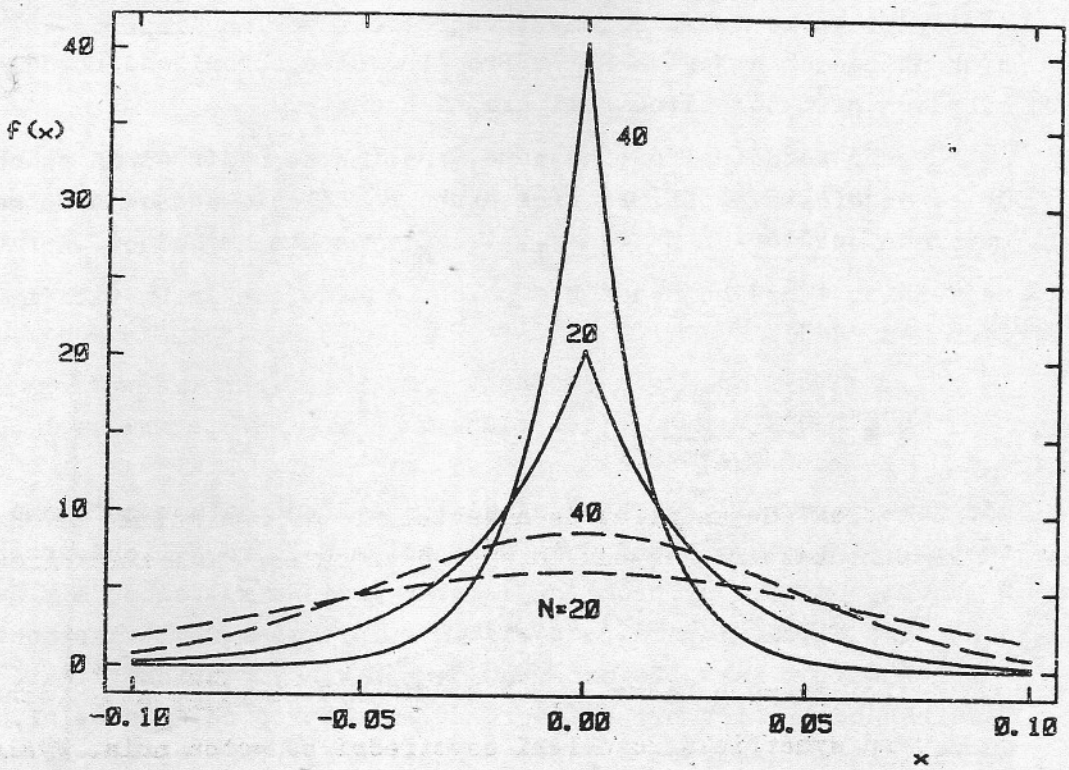
Srovnáme podrobněji odhad polohy rovnoměrného rozdělení

$$f(x) = 1, \quad x \in (-1/2,\ 1/2) \qquad (4)$$

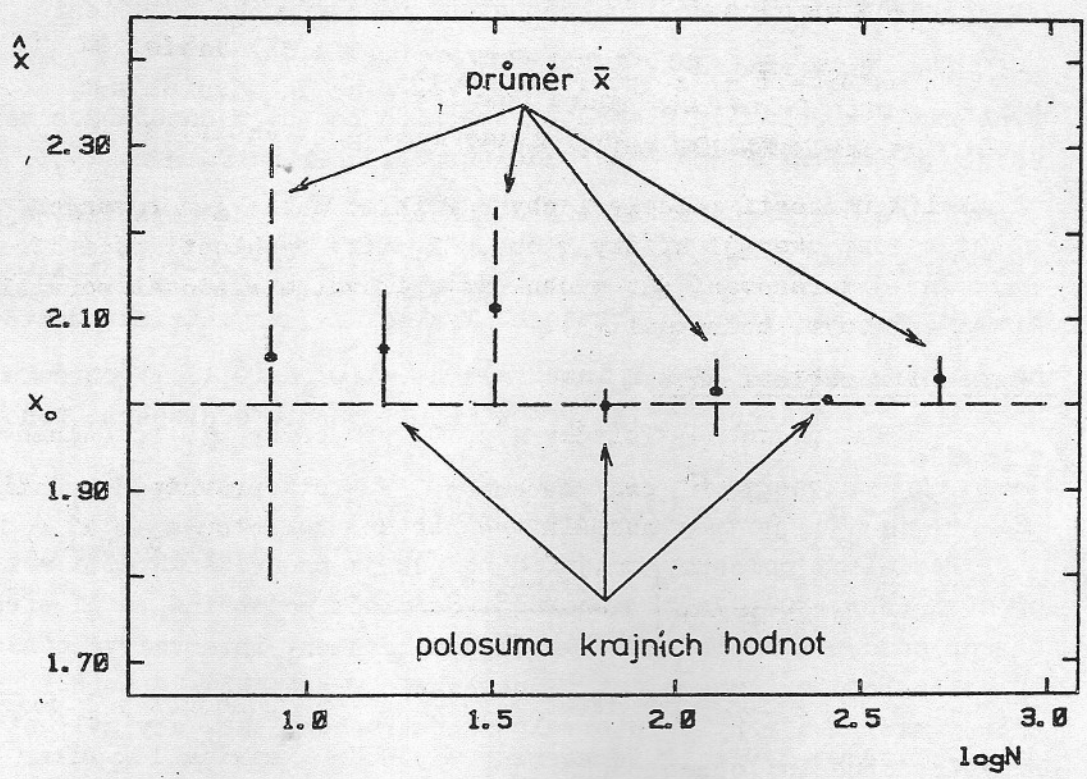
průměrem $\bar{x}$ a polovičním součtem krajních hodnot $\bar{\bar{x}}$ naměřených nezávislých údajů $x_1,\ldots,x_N$. Průměr má rozdělení blízké k normálnímu (§ 5, obr. 6) se středem 0 a disperzí 1/(12N). Rozdělení $\bar{\bar{x}}$ se dá nejsnáze najít z distribučních funkcí $F(x_M) = (1/2+x_M)^N$, $F(x_m) = 1-(1/2-x_m)^N$ maxima x_M a minima x_m v N nezávislých pokusech. Výsledkem je symetrická hustota, tvořená polynomy stupně 2N-1 v intervalech (-1,0) a (0,1). V obrázku 30 je nakreslena pro dvě hodnoty N. Pro střední hodnotu a disperzi tohoto rozdělení vyjde

$$E(\bar{\bar{x}}) = 0, \quad D(\bar{\bar{x}}) = N/\left[2(N+1)^2(N+2)\right]. \qquad (5)$$

Rozdělení polosumy krajních hodnot je pro větší N užší než rozdělení průměrů (obr. 30). Ještě názorněji je vidět podstatně větší efektivnost odhadu pomocí $\bar{\bar{x}}$ v obr. 31. Tam jsou nakresleny intervalové odhady pro různý počet bodů v souboru dat, generovaném v počítači. S pomocí distribučních funkcí $\bar{\bar{x}}$ a $\bar{x}$ byly intervaly zkonstruovány tak, aby měly standardní pravděpodobnostní obsah 68.3%.



Obr. 30. Hustoty pravděpodobnosti polovičního součtu (plná čára) a průměru (čárkovaná čára) z N hodnot s rovnoměrným rozdělením (14.4).



Obr. 31. Intervalové odhady polohy rovnoměrného rozdělení (soubor N hodnot generován v počítači)

15. Příklad odhadu dvou parametrů lineárního modelu

Metody odhadu více parametrů budeme ilustrovat na dvojrozměrném případě, který je dostatečně obecný a přitom velmi názorný. Předpokládejme, že pro různé hodnoty nezávisle proměnné x měříme hodnoty závisle proměnné

$$y = a_o + b_o x^2 \quad . \tag{1}$$

Závislost y(x) je určena dvojicí parametrů a_o, b_o, které vystupují v modelu (1) lineárně; jejich hodnoty hledáme nepřímo z naměřených dvojic x, y. O linearitě modelu rozhoduje závislost na parametrech, nikoliv na nezávisle proměnné.

Předpokládejme dále, že nezávisle proměnnou můžeme určit přesně (nebo se zanedbatelnou chybou) a výsledkem měření je N hodnot proměnné y:

$$y_i = a_o + b_o x_i^2 + \varepsilon_i, \quad i=1,\ldots,N, \tag{2}$$

pro N-tici pevných hodnot x_i. Náhodná chyba i-té hodnoty (označili jsme ji ε_i) má symetrické rozdělení se střední hodnotou nula. Výsledky měření, t.j. N-tice hodnot (x_i, y_i), byly simulovány na počítači za pomoci generátoru pseudonáhodných čísel. Zvolili jsme

$$a_o = 1, \; b_o = 2 \; , \tag{3}$$

ekvidistantní síť x_i z intervalu $\langle 0,1 \rangle$:

$$x_i = (i-1)/(N-1), \quad i=1,\ldots,N, \tag{4}$$

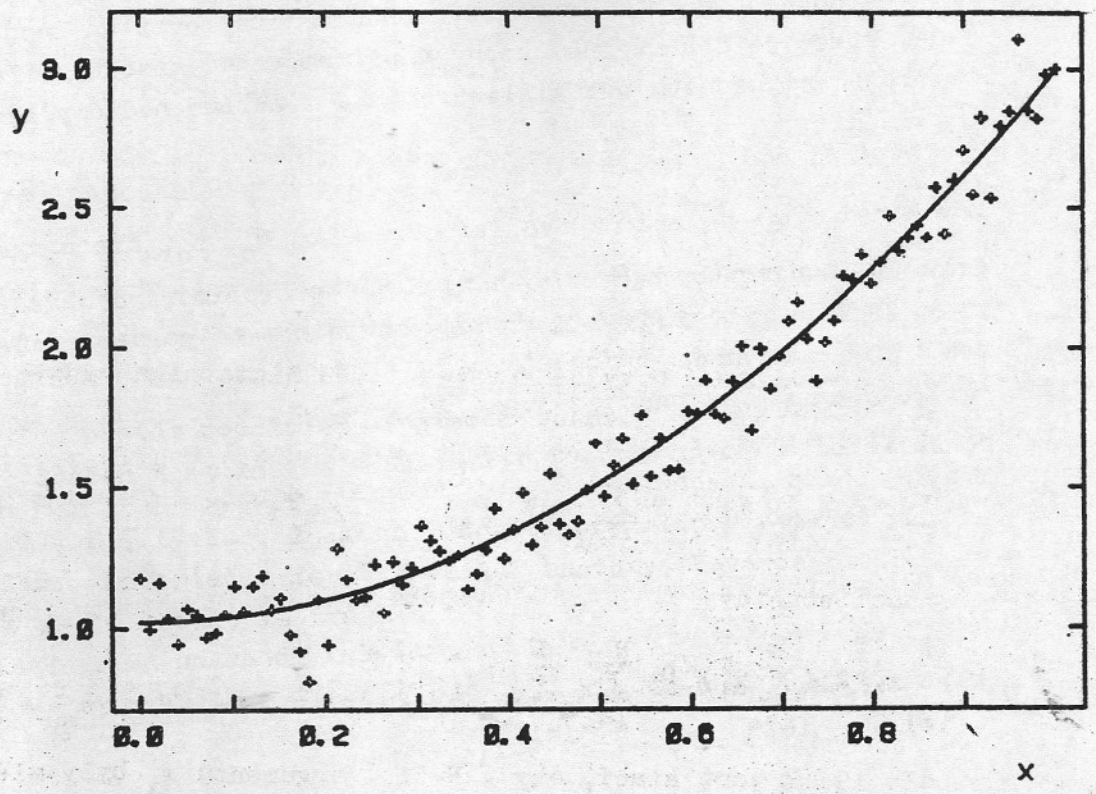
a několik možností rozdělení chyb. Příklad takto generovaných "experimentálních" dat ukazují křížky v obr. 32. Díky rychlosti počítače můžeme generování a zpracování dat mnohokrát opakovat a sledovat souvislost výsledků se správnými hodnotami (3).

Normální rozdělení chyb

Ukážeme podrobně výsledky výpočtů pro chyby ε_i rozdělené normálně se stejnou disperzí σ^2 pro všechna x_i. Hustota pravděpodobnosti jednotlivých hodnot y_i je tedy normální se střední hodnotou $a_o + b_o x_i^2$ a disperzí σ^2:

$$f(y_i) = \frac{1}{\sqrt{2\pi}\sigma^2} \exp\left[- \frac{(y_i - a_o - b_o x_i^2)^2}{2\sigma^2} \right]. \tag{5}$$

Předpokládáme samozřejmě nezávislost naměřených hodnot y_i; nejvěrohodnější odhad parametrů a_o, b_o dostaneme z podmínky maxima logaritmu funkce věrohodnosti



Obr. 32. Závislost (15.2) generovaná pro N=100 s normálně rozdělenou chybou se střední kvadratickou odchylkou σ=0.1 (křížky) a proložená závislost s odhadem parametrů (13a) (plná čára).

$$\ln L = \ln \prod_{i=1}^{N} f(y_i) = -\sum_{i=1}^{N} \frac{(y_i-a-bx_i^2)^2}{2\sigma^2} - \frac{N}{2}(\ln\sigma^2+\ln 2\pi). \qquad (6)$$

Maximum (6) vzhledem k a,b nastane právě tehdy, je-li minimální součet čtverců odchylek hodnot naměřených (y_i) a předpovězených modelem ($a+bx_i^2$):

$$S = \sum_{i=1}^{N} (y_i-a-bx_i^2)^2. \qquad (7)$$

Z podmínky $\partial S/\partial a = \partial S/\partial b = 0$ vyjde soustava dvou lineárních rovnic (říká se jim <u>normální rovnice</u>) pro hledané odhady $\hat{a}$, $\hat{b}$:

$$\hat{a}N + \hat{b}\sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} y_i, \quad \hat{a}\sum_{i=1}^{N} x_i^2 + \hat{b}\sum_{i=1}^{N} x_i^4 = \sum_{i=1}^{N} y_i x_i^2. \qquad (8)$$

Je-li determinant soustavy

$$d = N\sum_{i=1}^{N} x_i^4 - (\sum_{i=1}^{N} x_i^2)^2 = \sum_{i=1}^{N}\sum_{j=i+1}^{N} (x_i^2-x_j^2)^2 \qquad (9)$$

nenulový (podle (9) k tomu stačí, aby v N-tici argumentů x_i byly alespoň dva různé), existuje právě jedno řešení. Můžeme je napsat explicitně:

$$\hat{a} = \frac{1}{d}\sum_{i=1}^{N} y_i\left[\sum_{j=1}^{N} x_j^2(x_j^2-x_i^2)\right], \quad \hat{b} = \frac{1}{d}\sum_{i=1}^{N} y_i\left[\sum_{j=1}^{N}(x_i^2-x_j^2)\right]. \qquad (10)$$

Odhady jsou lineárními kombinacemi normálně rozdělených naměřených hodnot y_i - mají tedy také normální rozdělení. Přímým výpočtem můžeme najít střední hodnoty a druhé momenty; s pomocí (3.18) dostáváme

$$E(\hat{a}) = a_o, \quad E(\hat{b}) = b_o,$$

$$D(\hat{a}) = \sigma^2(\hat{a}) = \frac{\sigma^2}{d}\sum_{i=1}^{N} x_i^4, \quad D(\hat{b}) = \sigma^2(\hat{b}) = \sigma^2\frac{N}{d}, \quad D(\hat{a},\hat{b}) = \frac{\sigma^2}{d}(-\sum_{i=1}^{N} x_i^2). \qquad (11)$$

Využili jsme nezávislosti různých y_i ... $D(y_i,y_i) = \sigma^2\delta_{ij}$, kde δ_{ij} je Kroneckerovo delta (jednička pro i=j, jinak nula). Odhady $\hat{a}$, $\hat{b}$ jsou vždy korelované, protože $D(\hat{a},\hat{b}) \neq 0$. Korelační koeficient

$$\rho = \frac{D(\hat{a},\hat{b})}{\sigma(\hat{a})\sigma(\hat{b})} = -\sum_{i=1}^{N} x_i^2 \Big/ \sqrt{N\sum_{i=1}^{N} x_i^4} \qquad (12)$$

závisí pouze na hodnotách $x_1,\ldots,x_N$. Stojí zato si všimnout, že matice druhých momentů (11) je až na faktor σ^2 rovna inverzní matici soustavy normálních rovnic (8). Je-li disperze jednotlivých hodnot y_i známá, víme o odhadech $\hat{a},\hat{b}$ prostřednictvím (11) vše, co vědět můžeme.

Výpočet s daty z obr. 32 vedl k následujícím hodnotám:

$$\hat{a} = 1.022, \quad \hat{b} = 1.971, \tag{13a}$$

$$\sigma(\hat{a}) = 0.0150, \ \sigma(\hat{b}) = 0.0332, \ \rho = -0.743. \tag{13b}$$

Proložená závislost $\hat{a}+\hat{b}x^2$ je nakreslena v obr. 32.

Odhad eliptickou oblastí při známém σ

Při opakovaném pokusu budou body (a,b) vycházet náhodně kolem středu $(a_o,b_o) = (1,2)$ s normální hustotou s parametry (13b). Analogií intervalového odhadu jednoho parametru je zde odhad pomocí oblasti, která obsahuje hledané hodnoty s předepsanou pravděpodobností. Nejvýhodnější oblastí je vnitřek elipsy s konstantní hustotou (má při zadaném pravděpodobnostním obsahu minimální plochu, čímž nejlépe lokalizuje hledaný bod v rovině parametrů). Z § 6 víme, že konstrukce takové elipsy vyplývá ze znalosti rozdělení kovariační formy normální hustoty. Kovariační forma pro proměnné $\hat{a},\hat{b}$ se dá napsat ve tvaru

$$\zeta = \frac{1}{\sigma^2}\left[(\hat{a}-a_o)^2 N + 2(\hat{a}-a_o)(\hat{b}-b_o)\sum_{i=1}^{N} x_i^2 + (\hat{b}-b_o)^2\sum_{i=1}^{N} x_i^4\right] =$$

$$= \frac{1}{1-\rho^2}\left[\frac{(\hat{a}-a_o)^2}{\sigma^2(\hat{a})} - 2\rho\,\frac{(\hat{a}-a_o)(\hat{b}-b_o)}{\sigma(\hat{a})\sigma(\hat{b})} + \frac{(\hat{b}-b_o)^2}{\sigma^2(\hat{b})}\right]; \tag{14}$$

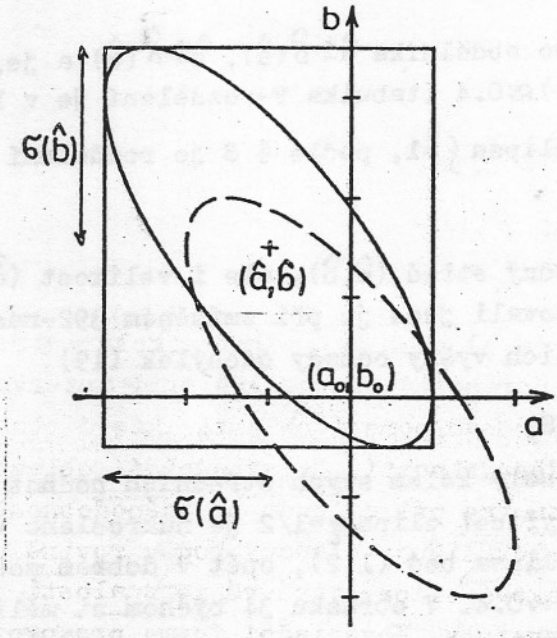
má rozdělení χ^2 se dvěma stupni volnosti. Vnitřek elipsy ζ = konst. $= \lambda$ má pravděpodobnostní obsah daný příslušnou distribuční funkcí:

$$P(\zeta < \lambda) = F_{\chi_2^2}(\lambda). \tag{15}$$

Elipsu $\zeta = \lambda$ můžeme interpretovat dvojím způsobem: buď má střed v (a_o,b_o) nebo v $(\hat{a},\hat{b})$ (obr. 33). V obou případech dává místa se stejnou hustotou pravděpodobnosti vzájemné polohy bodů (a_o,b_o) a $(\hat{a},\hat{b})$. První možnost bychom použili pro předpověď výskytu odhadů při známém (a_o,b_o), druhá se hodí pro řešení naší úlohy - elipsami kolem $(\hat{a},\hat{b})$ se snažíme zasáhnout hledaný bod (a_o,b_o). Volbou konstanty λ určujeme velikost elipsy a tím i její pravděpodobnostní obsah. Například $\lambda=1$ znamená podle (14) elipsu, která je vepsaná do obdélníka $\hat{a}\pm\sigma(\hat{a})$, $\hat{b}\pm\sigma(\hat{b})$ - viz obr. 33. Její pravděpodobnostní obsah je asi 0.4 (hodnota distribuční funkce (15) v bodě 1 ... tabulka v dodatku D2). Při 392-násobném opakování celého pokusu v počítači obsahovala elipsa $\zeta=1$ bod (1,2) ve 164 případech, což je ve velmi dobré shodě s očekávaným počtem $392\times0.4\approx157$.

Odhad disperze σ^2

Neznáme-li σ^2, můžeme ji odhadnout z naměřených dat. Z podmínky



Obr. 33. Elipsy $\zeta=1$ s kovariační formou ζ podle (14), $\varrho=-0.74$; obdélník $\hat{a}\pm\sigma(\hat{a})$, $\hat{b}\pm\sigma(\hat{b})$.

maxima věrohodnosti vzhledem k σ^2 (nejlépe s pomocí (6)...$\partial \ln L/\partial(\sigma^2)=0$) dostaneme odhad

$$\hat{\sigma}^2=\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{a}-\hat{b}x_i^2)^2 = S_o/N. \qquad (16)$$

Symbolem S_o jsme označili tzv. reziduální sumu čtverců, t.j. hodnotu S ze vztahu (7) v bodě minima. $N\hat{\sigma}^2/\sigma^2$ je náhodná veličina s χ^2 rozdělením s počtem stupňů volnosti rovným N-2 (dá se vyjádřit jako součet N-2 kvadrátů nezávislých proměnných s rozdělením N(0,1); jednotlivé sčítance v součtu (16) nezávislé nejsou!). Odhad (16) je vychýlený, protože jeho střední hodnota je $E(\hat{\sigma}^2) = \sigma^2(N-2)/N \neq \sigma^2$. Nevychýleným odhadem σ^2 je $\hat{\sigma}^{2\prime} = S_o/(N-2)$. V našem případě (N=100) je rozdíl mezi $\hat{\sigma}^2$ a $\hat{\sigma}^{2\prime}$ zanedbatelný; v pokusu, při kterém vyšly odhady (13a) bylo $\hat{\sigma}^{2\prime}\approx 0.0098$ v dobré shodě se skutečnou disperzí dat $\sigma^2 = 0.01$.

Odhad parametrů eliptickou oblastí při neznámé disperzi σ^2

Podíl

$$\eta = \frac{\frac{1}{2}\zeta}{\frac{1}{N-2}N\frac{\hat{\sigma}^2}{\sigma^2}} = \frac{\sigma^2\zeta/2}{S_o/(N-2)} \qquad (17)$$

na neznámé disperzi σ^2 nezávisí a má podle (8.11) F-rozdělení s m=2, m'=N-2 stupni volnosti. Vnitřek elipsy $\eta=\lambda$ má tedy pravděpodobnostní obsah roven hodnotě příslušné distribuční funkce v λ:

$$P(\eta<\lambda) = F_{2,N-2}(\lambda). \qquad (18)$$

Označíme

$$\hat{\delta}(\hat{a}) = \sqrt{\frac{S_o}{N-2}}\sqrt{\frac{1}{d}\sum_{i=1}^{N}x_i^4}\,, \quad \hat{\delta}(\hat{b}) = \sqrt{\frac{S_o}{N-2}}\sqrt{\frac{N}{d}} \qquad (19)$$

odhady standardních odchylek $\hat{a},\hat{b}$, které dostaneme ze vztahů (11) a (17). Rovnici elipsy $\eta=\lambda$ pak přepíšeme do tvaru

$$\frac{1}{2}\frac{1}{1-\varrho^2}\left\{\frac{(\hat{a}-a_o)^2}{[\hat{\delta}(\hat{a})]^2} - 2\varrho\frac{(\hat{a}-a_o)(\hat{b}-b_o)}{\hat{\delta}(\hat{a})\hat{\delta}(\hat{b})} + \frac{(\hat{b}-b_o)^2}{[\hat{\delta}(\hat{b})]^2}\right\} = \lambda. \qquad (20)$$

Pro $\lambda=1/2$ je to elipsa vepsaná do obdélníka $\hat{a}\pm\hat{\delta}(\hat{a})$, $\hat{b}\pm\hat{\delta}(\hat{b})$ a její pravděpodobnostní obsah je $F_{2,98}(0.5)\approx 0.4$ (tabulka F-rozdělení je v D3; vyšla stejná hodnota jako nahoře pro elipsu $\zeta=1$, podle § 8 je rozdělení veličiny $2F_{2,98}$ blízké k χ^2_2).

Elipsy (20) mají nejen náhodný střed $(\hat{a},\hat{b})$, ale i velikost ($\hat{\delta}(\hat{a})$, $\hat{\delta}(\hat{b})$ jsou náhodné veličiny). Sledovali jsme je při zmíněném 392-násobném opakování pokusu. Při prvním z nich vyšly odhady odchylek (19)

$$\hat{\delta}(\hat{a}) = 0.0148, \quad \hat{\delta}(\hat{b}) = 0.0328, \tag{21}$$

a v dalších podle očekávání kolísaly kolem svých středních hodnot $\sigma(\hat{a})$, $\sigma(\hat{b})$ ze vztahu (13b). Prvních čtyřicet elips $\eta=1/2$ je nakresleno v obr. 34. Celkem 169-krát obsahovala elipsa bod (1,2), opět v dobrém souhlasu s předpovězenou pravděpodobností ~ 0.4. V obrázku 34 bychom si měli všimnout toho, jak záporný koeficient korelace rozděluje odhady $(\hat{a},\hat{b})$ kolem úhlopříčky z levého horního do pravého dolního rohu. Pravděpodobnostní obsah obdélníka $\hat{a}\pm\hat{\delta}(\hat{a}), \hat{b}\pm\hat{\delta}(\hat{b})$ je samozřejmě o něco větší. Z obr. 9 v § 6 odečteme pro $\rho=0.74$ pravděpodobnost asi 0.55; pozorovaná relativní četnost příznivých případů $225/392\approx 0.57$ je s ní ve výborné shodě.

Intervalové odhady jednotlivých parametrů

Zatím jsme se soustředili na odhad obou parametrů současně. Můžeme najít také pravděpodobnostní tvrzení o každém parametru zvlášť. Především je zřejmé, že intervalový odhad $\hat{a}\pm\sigma(\hat{a})$ (chápaný tak, že sledujeme pouze souvislost a_o a $\hat{a}$ - bez ohledu na to, jak vychází $\hat{b}$) má pravděpodobnostní obsah 0.683; marginální rozdělení $\hat{a}$ je $N(a_o, \sigma^2(\hat{a}))$. Stejné tvrzení platí o odhadu $\hat{b}\pm\sigma(\hat{b})$.

Pokud disperzi σ^2 neznáme a nemůžeme spočíst $\sigma(\hat{a})$, $\sigma(\hat{b})$ ze vztahu (11), můžeme použít Studentova rozdělení pro sestrojení intervalů pomocí $\hat{\delta}(\hat{a})$, $\hat{\delta}(\hat{b})$ ze vztahu (19). Postup je přesně stejný jako v případě přímo měřené hodnoty (§12).

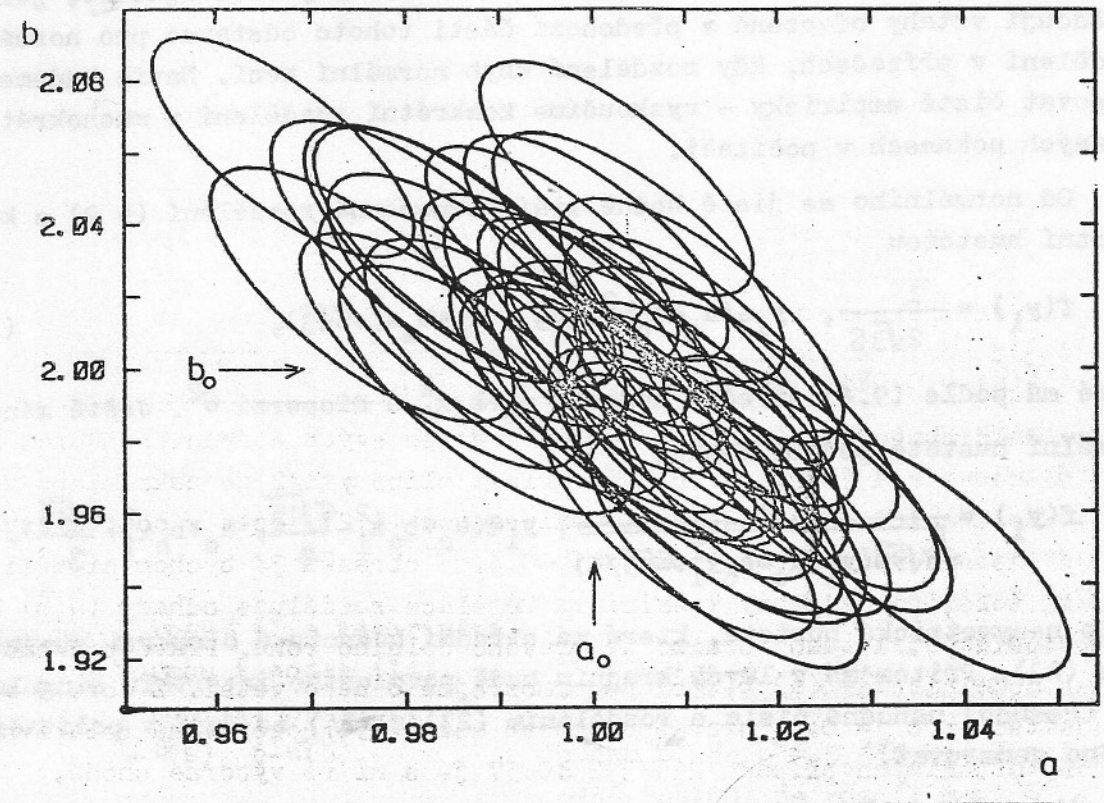
Je-li hodnota jednoho z parametrů známá, má odhad druhého opět normální rozdělení, s disperzí zmenšenou faktorem $1-\rho^2$ (viz podmíněné rozdělení (6.16), (6.17)). Například fixování hodnoty parametru $b=b_o$ vede ke zmenšení disperze $\hat{a}$ podle vztahů (11) a (12) na

$$D(\hat{a}|b=b_o) = D(\hat{a})(1-\rho^2) = \sigma^2/N. \tag{22}$$

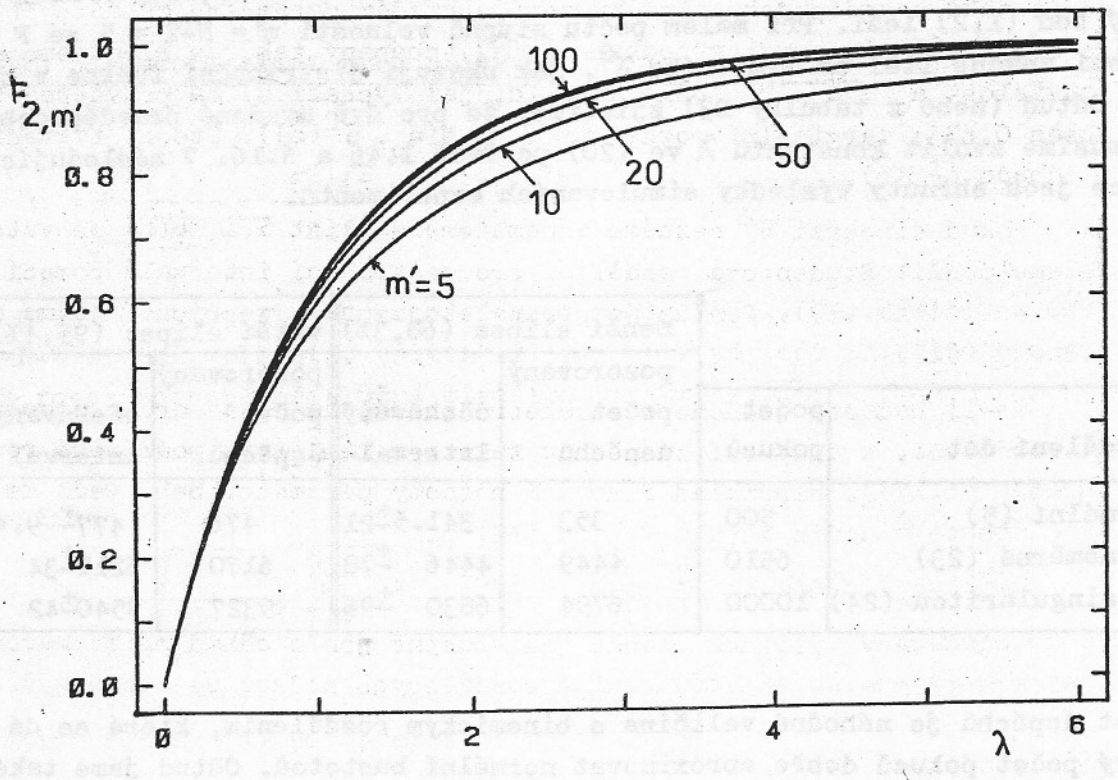
To je očekávaný výsledek, neboť jde vlastně jen o odhad přímo měřené veličiny z N-tice $a_i=y_i-b_o x_i^2$ normálně rozdělených hodnot se stejnou disperzí σ^2.

Jiné rozdělení dat

Posoudíme vliv různých rozdělení chyb ε_i ze vztahu (2) na odhad parametrů. Jde o velmi složitý problém, protože v zásadě každý typ rozdělení



Obr. 34. Odhady parametrů (a_o, b_o) eliptickou oblastí (15.20) s $\lambda = 1/2$.



Obr. 35. Distribuční funkce F - rozdělení s m=2 a různými hodnotami m'.

dává jiné formule pro nejvěrohodnější odhad a potřebné výpočty mohou být značně komplikované. Stanovíme si relativně skromný cíl: zjistit, jak se osvědčují vztahy odvozené v předchozí části tohoto odstavce pro normální rozdělení v případech, kdy rozdělení chyb normální není. Navíc budeme postupovat čistě empiricky - vyzkoušíme konkrétní rozdělení v mnohokrát opakovaných pokusech v počítači.

Od normálního se jistě hodně liší rovnoměrné rozdělení (§ 9) s konstantní hustotou

$$f(y_i) = \frac{1}{2\sqrt{36}}, \quad y_i \in (a_o+b_o x_i^2-6\sqrt{3},\ a_o+b_o x_i^2+6\sqrt{3}), \tag{23}$$

které má podle (9.6) střední hodnotu $a_o+b_o x_i^2$ a disperzi 6^2. Ještě méně se normální hustotě podobá funkce

$$f(y_i) = \frac{1}{\sqrt{6\sqrt{90}(y_i-a_o-b_o x_i^2+6\sqrt{45}/6)}}, \quad y_i \in (a_o+b_o x_i^2-\frac{6\sqrt{45}}{6},\ a_o+b_o x_i^2+\frac{6\sqrt{45}}{3}). \tag{24}$$

Je to nesymetrická hustota, která má střední hodnotu a disperzi stejnou jako (23). Přitom má v levém krajním bodě povoleného intervalu singularitu. (Pseudo) náhodná čísla s rozdělením (23) a (24) se dají v počítači snadno generovat.

Aby byl výpočet rychlejší, zvolíme menší počet bodů v závislosti (2): N=7. Budeme v každém pokusu konstruovat elipsy (20) s oblíbeným pravděpodobnostním obsahem 0.683 a 0.954 a sledovat počet případů, kdy v nich hledaný bod (1,2) leží. Při malém počtu stupňů volnosti $m' = N-2 = 5$ se F - rozdělení značně liší od limitního χ^2, jak ukazují distribuční funkce v obr. 35. Odtud (nebo z tabulky D3) zjistíme, že pro dvě uvedené pravděpodobnosti musíme zvolit konstantu λ ve (20) po řadě 1.46 a 6.16. V následující tabulce jsou shrnuty výsledky simulovaných experimentů.

		menší elipsa (68,3%)		větší elipsa (95,4%)	
rozdělení dat	počet pokusů	pozorovaný počet úspěchů	očekávaný interval	pozorovaný počet úspěchů	očekávaný interval
normální (5)	500	353	341.5 ± 21	476	477 ± 9.4
rovnoměrné (23)	6510	4449	4446 ± 78	6170	6211 ± 34
se singularitou (24)	10000	6794	6830 ± 96	9327	9540 ± 42

Počet úspěchů je náhodná veličina s binomickým rozdělením, které se dá pro velký počet pokusů dobře aproximovat normální hustotou. Odtud jsme také určili intervaly, do kterých by měly pozorované počty padnout s pravděpo-

dobností 95.4% ($\pm$ dvě standardní odchylky). Pro normálně rozdělená data v prvním řádku je vše v pořádku. Pozoruhodná je ovšem dobrá shoda pro obě další rozdělení. Jedině údaj pro větší elipsu v posledním řádku signalizuje závažný nesouhlas mezi pozorovaným a očekávaným údajem (liší se zhruba o 5 standardních odchylek). Na druhé straně je však většinou nepodstatné, že místo 0.954 je pravděpodobnostní obsah odhadu pouze 0.93. Celkově můžeme zhodnotit funkci postupu odvozeného pro normální rozdělení jako překvapivě dobrou.

Předpověď vlastností odhadů podle vztahů (10), (19) a (20) s testovacím rozdělením dat (23) a (24) by se dala provést přesně, byla by ale značně namáhavá. Použitelnost postupu můžeme vysvětlit v hrubých rysech pomocí centrální limitní věty - rozdělení odhadů (10) se blíží normálnímu i pro jiná rozdělení dat. Dobrá funkce studovaného postupu zpracování pro různá rozdělení je velmi vítaná. Neznamená to ale, že by nemělo cenu hledat jiné postupy respektující zvláštnosti rozdělení dat; zpracování pak může být efektivnější (viz příklad odhadu polohy rovnoměrného rozdělení v § 14).

16. Odhad parametrů lineárního modelu

Postup odhadu parametrů lineárního modelu, odvozený a ilustrovaný na příkladu v § 15, zobecníme pro případ různých vah naměřených hodnot a libovolného počtu parametrů. Budeme používat maticové zápisy podle konvencí použitých v § 6.

Předpokládejme, že hledáme hodnoty K parametrů $\theta_1,\ldots,\theta_K$ (uspořádáme je do sloupcového vektoru $\underline{\theta}$ nebo po transpozici do řádkového vektoru $\underline{\theta}^T=(\theta_1,\ldots,\theta_K)$). Měříme hodnoty $y_1,\ldots,y_N$, které jsou lineárními funkcemi $\underline{\theta}$:

$$\underline{y} = \underset{\sim}{A}\,\underline{\theta}\,, \tag{1}$$

kde $\underline{y}$ je vektor s N složkami, $\underline{y}^T = (y_1,\ldots,y_N)$, $\underset{\sim}{A}$ je matice koeficientů s N řádky a K sloupci (stručně NxK). Abychom mohli odhadnout všechny parametry, musí být $N>K$; o matici $\underset{\sim}{A}$ předpokládáme, že má hodnost K. Obvykle dostáváme hodnoty y_i měřením při různých (známých) hodnotách x_i jiné proměnné x, na které y také závisí:

$$y(x) = \theta_1 f_1(x)+\ldots+\theta_K f_K(x), \tag{2}$$

kde $f_1(x),\ldots,f_K(x)$ jsou zadané lineárně nezávislé funkce. Model (2) je lineární vzhledem k parametrům $\underline{\theta}$, matice $\underset{\sim}{A}$ má prvky

$$\underset{\sim}{A} = \begin{bmatrix} f_1(x_1) & & f_K(x_1) \\ & \ldots & \\ f_1(x_N) & & f_K(x_N) \end{bmatrix}. \tag{3}$$

Předpokládejme dále, že $\underline{y}$ má N-rozměrné normální rozdělení se středními hodnotami $E(\underline{y}) = \underset{\sim}{A}\underline{\theta}_o$ a diagonální maticí druhých momentů $\underset{\sim}{D} = \sigma^2 \underset{\sim}{W}^{-1}$ (složky $\underline{y}$ jsou nezávislé). Zde je, stejně jako v § 12, $\underset{\sim}{W}$ matice vah, σ^2 disperze pro jednotku váhy.

Odhad parametrů $\underline{\theta}_o$

Z principu maximální věrohodnosti (zobecněním postupu z § 15) dostaneme pro odhad $\hat{\underline{\theta}}$ hledaných parametrů $\underline{\theta}_o$ soustavu lineárních rovnic

$$(\underset{\sim}{A}^T \underset{\sim}{W} \underset{\sim}{A})\hat{\underline{\theta}} = \underset{\sim}{A}^T \underset{\sim}{W}\underline{y} . \tag{4}$$

Říká se jí soustava normálních rovnic. Podle předpokladu je matice

$$\underset{\sim}{H} = \underset{\sim}{A}^T \underset{\sim}{W} \underset{\sim}{A} \tag{5}$$

regulární - její hodnost je K. Soustava (4) má tedy právě jedno řešení

$$\hat{\underline{\theta}} = \underset{\sim}{H}^{-1} \underset{\sim}{A}^T \underset{\sim}{W}\underline{y} . \tag{6}$$

Odhad $\hat{\underline{\theta}}$ je lineární funkcí normálně rozděleného vektoru $\underline{y}$, má tedy také normální rozdělení. Střední hodnoty a matice druhých momentů jsou

$$E(\hat{\underline{\theta}}) = \underline{\theta}_o \ , \ \underset{\sim}{D}(\hat{\underline{\theta}}) = E\left[(\hat{\underline{\theta}}-\underline{\theta}_o)(\hat{\underline{\theta}}-\underline{\theta}_o)^T\right] = \sigma^2 \underset{\sim}{H}^{-1}. \tag{7}$$

Odhad $\hat{\underline{\theta}}_o$ elipsoidem při známém σ^2

Z § 6 víme, že kovariační forma

$$\varphi = (\hat{\underline{\theta}}-\underline{\theta}_o)^T \underset{\sim}{D}^{-1}(\hat{\underline{\theta}}-\underline{\theta}_o) = \frac{1}{\sigma^2}(\hat{\underline{\theta}}-\underline{\theta}_o)^T \underset{\sim}{H}(\hat{\underline{\theta}}-\underline{\theta}_o) \tag{8}$$

je náhodná proměnná s rozdělením χ^2 s K stupni volnosti. Pro kladnou konstantu λ znamená splnění nerovnosti $\varphi < \lambda$ fakt, že $\underline{\theta}_o$ leží uvnitř elipsoidu $\varphi = \lambda$, opsaného kolem středu $\hat{\underline{\theta}}$. Pravděpodobnost této události je

$$P(\varphi < \lambda) = F_{\chi^2_K}(\lambda), \tag{9}$$

kde $F_{\chi^2_K}$ je příslušná distribuční funkce, nakreslená pro několik hodnot K v obr. 36.

Odhad neznámé disperze σ^2

Odhad neznámé hodnoty σ^2 dostaneme opět z principu maximální věrohodnosti. Označíme-li $\hat{\underline{y}} = \underset{\sim}{A}\hat{\underline{\theta}}$ předpověď měřených hodnot z modelu (1) pro parametry získané odhadem (6), vyjde nejvěrohodnější odhad disperze

$$\hat{\sigma}^2 = \frac{1}{N}(\underline{y}-\hat{\underline{y}})^T \underset{\sim}{W}(\underline{y}-\hat{\underline{y}}) = \frac{1}{N}\sum_{i=1}^{N} w_i (y_i - \hat{y}_i)^2 = \frac{S_o}{N}. \tag{10}$$

Jako S_o jsme označili tzv. reziduální součet čtverců odchylek.

Dá se dokázat, že veličina $N\hat{\sigma}^2/\sigma^2$ má rozdělení χ^2_{N-K}; odtud můžeme najít intervalové odhady pro σ^2. Odhad (10) je vychýlený, protože jeho střední hodnota je

$E(\hat{\sigma}^2) = (\sigma^2/N)E(\chi^2_{N-K}) = \sigma^2(N-K)/N$. Nevychýleným odhadem disperze je například

$$\hat{\sigma}^{2\prime} = \frac{N}{N-K}\hat{\sigma}^2 = \frac{1}{N-K}\sum_{i=1}^{N} w_i(y_i-\hat{y}_i)^2 . \qquad (10a)$$

Odhad $\underline{\theta}_o$ elipsoidem při odhadované disperzi

Známého rozdělení kovariační formy (8) a odhadu $\hat{\sigma}^2$ (10) využijeme k odhadu $\underline{\theta}_o$ elipsoidem v prostoru parametrů; podíl

$$\eta = \frac{\zeta/K}{N\hat{\sigma}^2/[\sigma^2(N-K)]} = \frac{(\hat{\underline{\theta}}-\hat{\underline{\theta}}_o)^T \underset{\sim}{H}(\hat{\underline{\theta}}-\underline{\theta}_o)/K}{(\underline{y}-\hat{\underline{y}})^T\underset{\sim}{W}(\underline{y}-\hat{\underline{y}})/(N-K)} \qquad (11)$$

ma neznámé hodnotě σ^2 nezávisí a má F- rozdělení (§ 8) s m=K, m'=N-K stupni volnosti. Podmínku $\eta<\lambda$, kde λ je kladná konstanta, můžeme interpretovat tak, že bod $\underline{\theta}_o$ leží uvnitř elipsoidu $\eta=\lambda$ se středem v $\hat{\underline{\theta}}$. Pravděpodobnost této události je dána distribuční funkcí F - rozdělení:

$$P(\eta<\lambda) = F_{k,N-K}(\lambda). \qquad (12)$$

Volbou λ můžeme zajistit předepsaný pravděpodobnostní obsah elipsoidu (tabulka F - rozdělení je v D3).

Intervalové odhady jednotlivých složek $\underline{\theta}_o$

Znalost rozdělení vektoru $\hat{\underline{\theta}}$, případně odhadu disperze $\hat{\sigma}^2$, umožňuje konstrukci odhadů každé složky θ_{oj} vektoru $\underline{\theta}_o$ samostatně, t.j. bez ohledu na ostatní. Marginální rozdělení $\hat{\theta}_j$ je podle (6.13) normální se střední hodnotou θ_{oj} a disperzí danou j-tým diagonálním prvkem matice $\underset{\sim}{D}$:

$$\sigma_j^2 = (\underset{\sim}{D})_{jj} = \sigma^2(\underset{\sim}{H}^{-1})_{jj}, \quad j=1,\ldots,K. \qquad (13)$$

Známe-li σ^2, najdeme intervalové odhady $\hat{\theta}_j \pm k\sigma_j$ pomocí normální distribuční funkce ze vztahu (12.7).

Odhadujeme-li σ^2 z naměřených dat, vyjdeme z podílu

$$\frac{(\hat{\theta}_j-\theta_{oj})/\sigma_j}{\sqrt{N\hat{\sigma}^2/[(N-K)\sigma^2]}} = \frac{\hat{\theta}_j-\theta_{oj}}{\sqrt{(\underset{\sim}{H}^{-1})_{jj}\, S_o/(N-K)}} ; \qquad (14)$$

ten na σ^2 nezávisí a má Studentovo rozdělení s N-K stupni volnosti. Označíme

$$\hat{\delta}_j = \sqrt{(\underset{\sim}{H}^{-1})_{jj}\frac{S_o}{N-K}} \qquad (15)$$

a najdeme pravděpodobnostní obsah intervalu $\hat{\theta}_j \pm k\hat{\delta}_j$ pomocí distribuční

funkce F_{N-K} Studentova rozdělení (viz (12.11)). Pro dostatečně velký počet stupňů volnosti N-k (≥ 30) má interval $\hat{\theta}_j \pm \hat{\delta}_j$ pravděpodobnostní obsah 0.683, stejný, jako má interval $\hat{\theta}_j \pm \sigma_j$.

Souvislost se sumou čtverců odchylek

Podmínka maxima věrohodnosti je při normálně rozdělených datech ekvivalentní s podmínkou minima součtu čtverců - viz (10.11). V našem případě K- rozměrného vektoru parametrů $\underline{\theta}$ a diagonální matice vah $\underset{\sim}{W}$ je suma čtverců rovna

$$S = (\underline{y}-\underset{\sim}{A}\underline{\theta})^T \underset{\sim}{W}(\underline{y}-\underset{\sim}{A}\underline{\theta}) = \sum_{i=1}^{N} w_i (y_i - \sum_{j=1}^{K} a_{ij}\theta_j)^2. \qquad (16)$$

Snadno se můžeme přesvědčit, že podmínka minima S vzhledem k $\underline{\theta}$, t.j. $\partial S/\partial\theta_1 = \ldots = \partial S/\partial\theta_K = 0$, vede skutečně na soustavu normálních rovnic (4).

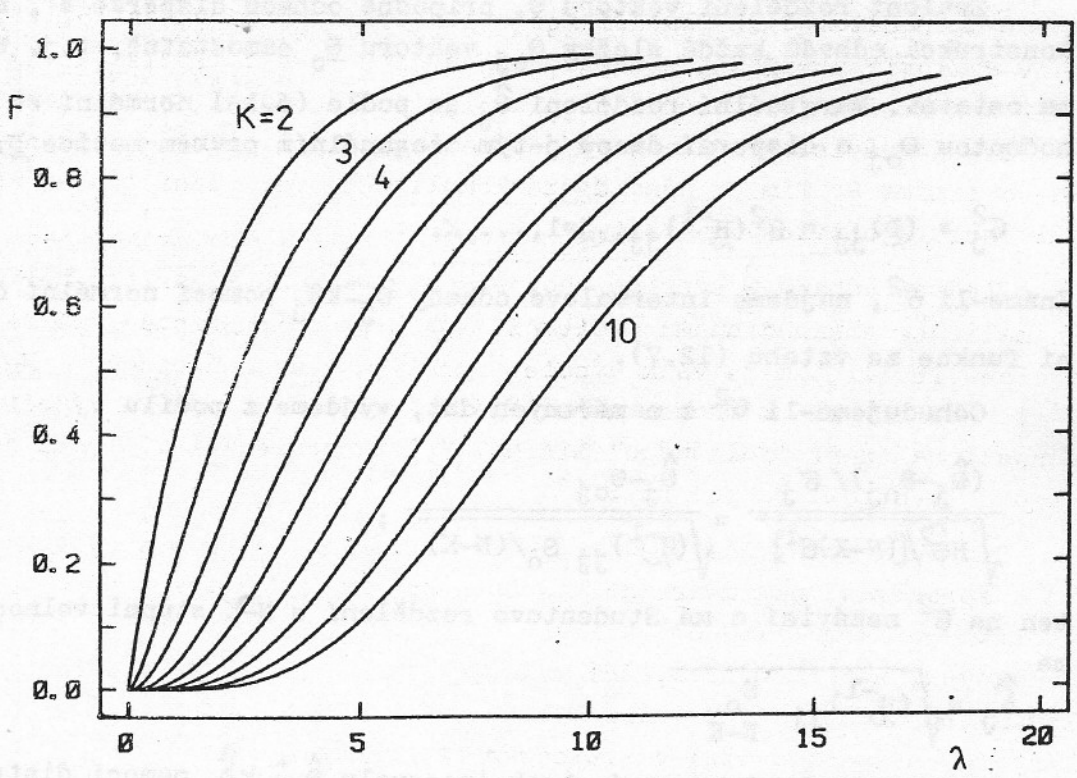
Pro součet čtverců (16) v minimu, t.j. v bodě $\hat{\underline{\theta}}$, dostaneme jednoduchý vztah

$$S_o = S(\hat{\underline{\theta}}) = \underline{y}^T \underset{\sim}{W}\underline{y} - 2\hat{\underline{\theta}}^T \underset{\sim}{A}^T \underset{\sim}{W}\underline{y} + \hat{\underline{\theta}}^T \underset{\sim}{H}\hat{\underline{\theta}} = \underline{y}^T \underset{\sim}{W}\underline{y} - \hat{\underline{\theta}}^T \underset{\sim}{A}^T \underset{\sim}{W}\underline{y}. \qquad (17)$$

Reziduální sumu čtverců tedy můžeme vypočíst tak, že od váženého součtu čtverců naměřených hodnot ($\underline{y}^T \underset{\sim}{W}\underline{y}$) odečteme skalární součin vektoru řešení a pravé strany normálních rovnic (4). Užitečné je také vyjádření S jako funkce posunutí $\underline{\Delta}$ z minima $\hat{\underline{\theta}}$:

$$S(\hat{\underline{\theta}}+\underline{\Delta}) = S_o + \underline{\Delta}^T \underset{\sim}{H}\underline{\Delta}. \qquad (18)$$

Matice $\underset{\sim}{H}$ kovariační formy (8) popisuje také funkci S v prostoru parametrů $\underline{\theta}$.



Obr. 36. Distribuční funkce χ^2 - rozdělení s K stupni volnosti.

17. Odhad parametrů nelineárního modelu

Měřené hodnoty $y_1,\dots,y_N$ mohou záviset na hledaných parametrech $\theta_1,\dots,\theta_K$ nelineárně:

$$y_i = h_i(\theta_1,\dots,\theta_K), i=1,\dots,N. \tag{1}$$

Předpokládejme, že y_i jsou nezávislé normálně rozdělené proměnné se středními hodnotami $E(y_i) = h_i(\theta_{o1},\dots,\theta_{oK})$ a diagonální kovariační maticí $\underset{\sim}{D} = \sigma^2 \underset{\sim}{W}^{-1}$ ($\underset{\sim}{W}$ je diagonální matice vah, σ^2 disperze pro jednotkovou váhu).

Funkce věrohodnosti pozorované N-tice $\underline{y}$ při hodnotách parametrů $\underline{\theta}$ je tedy

$$L = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2/w_i}} \exp\left\{ - \frac{\left[y_i - h_i(\theta_1,\dots,\theta_K)\right]^2}{2\sigma^2/w_i} \right\}. \tag{2}$$

Maximum L vzhledem k $\underline{\theta}$ nastane tehdy, je-li maximální

$$\ln L = - \sum_{i=1}^{N} \left\{ \frac{\left[y_i - h_i(\theta_1,\dots,\theta_K)\right]^2}{2\sigma^2/w_i} + \frac{1}{2} \ln(2\pi\sigma^2/w_i) \right\}; \tag{3}$$

k tomu stačí najít minimum váženého součtu čtverců odchylek pozorovaných a modelových hodnot v prostoru parametrů $\underline{\theta}$:

$$S = \sum_{i=1}^{N} w_i \left[y_i - h_i(\theta_1,\dots,\theta_K)\right]^2 = \left[\underline{y} - \underline{h}(\underline{\theta})\right]^T \underset{\sim}{W} \left[\underline{y} - \underline{h}(\underline{\theta})\right]. \tag{4}$$

Nalezení odhadu $\hat{\underline{\theta}}$ hledaného vektoru $\underline{\theta}_o$ je díky nelinearitě modelových funkcí h_i podstatně obtížnější než v lineárním případě (§§ 15,16), kde stačí sestavit a vyřešit soustavu lineárních rovnic. Podle (16.6) jsme mohli dokonce vyjádřit odhad jako explicitní lineární funkci naměřených dat. Zde je závislost $\hat{\underline{\theta}}$ na $\underline{y}$ vyjádřena pouze implicitně - podmínkou maxima L nebo minima S. Tím je dána druhá komplikace: rozdělení $\hat{\underline{\theta}}$ není normální a je zpravidla obtížné ho najít.

Hledání odhadu $\hat{\underline{\theta}}$ svěříme vhodnému numerickému algoritmu nelineární minimalizace a samočinnému počítači. I v lineárním případě řešíme normální rovnice (16.4); tam je však výpočet rychlý a vždy jednoznačný. Nelineární model vede obvykle k mnohem náročnějším (delším) strojovým výpočtům a k možnosti nalezení "falešného" minima. Při interpretaci výsledků je třeba s touto eventualitou počítat.

Druhou komplikaci, t.j. neznalost rozdělení odhadu $\hat{\underline{\theta}}$, obcházíme zpravidla aproximací nelineárního modelu lineárním. Jde v zásadě o použití přibližných formulí (3.28) a (3.29) pro druhé momenty funkcí náhodných proměnných, neboli o přibližné vyjádření "přenosu chyb měřených $\underline{y}$ do chyb hledaných parametrů". V okolí odhadu $\hat{\underline{\theta}}$ aproximujeme funkce (1) lineárními členy Taylorova rozvoje; v maticovém zápisu je

$$\underline{h}(\hat{\underline{\theta}}+\underline{\Delta}) \approx \underline{h}(\hat{\underline{\theta}}) + \underset{\sim}{A}\,\underline{\Delta} = \hat{\underline{y}} + \underset{\sim}{A}\,\underline{\Delta}, \tag{5}$$

kde prvky matice NxK koeficie rozvoje jsou d rivace $\underline{h}$ v bodě $\hat{\underline{\theta}}$:

$$(\underset{\sim}{A})_{ij} = (\partial h_i / \partial \theta_j)\Big|_{\underline{\theta} = \hat{\underline{\theta}}}. \tag{6}$$

Symbolem $\hat{\underline{y}}$ jsme označili hodno y modelových funkcí (1) v bodě $\hat{\underline{\theta}}$. V této aproximaci je suma čtverců (6) přibližně rovna

$$S \approx (\underline{y}-\hat{\underline{y}})^T \underset{\sim}{W} (\underline{y}-\hat{\underline{y}}) + \underline{\Delta}^T \underset{\sim}{A}^T \underset{\sim}{W} \underset{\sim}{A}\,\underline{\Delta} = S_o + \underline{\Delta}^T \underset{\sim}{H}\,\underline{\Delta}, \tag{7}$$

kde jsme zavedli označení

$$\underset{\sim}{H} = \underset{\sim}{A}^T \underset{\sim}{W} \underset{\sim}{A}, \text{ neboli } (\underset{\sim}{H})_{jm} = \sum_{i=1}^{N} w_i \frac{\partial h_i}{\partial \theta_j} \frac{\partial h_i}{\partial \theta_m}\Bigg|_{\underline{\theta}=\hat{\underline{\theta}}}. \tag{8}$$

Při odvození vztahu (7) jsme využili faktu, že vektor $(\underline{y}-\hat{\underline{y}})^T \underset{\sim}{W} \underset{\sim}{A}$ je úměrný gradientu

$$g_j = \frac{\partial S}{\partial \theta_j} = -2 \sum_{i=1}^{N} w_i \left[y_i - h_i(\underline{\theta}) \right] \frac{\partial h_i}{\partial \theta_j}, \quad j=1,\ldots,K \tag{9}$$

sumy čtverců (4) a je tedy v minimu S nulový.

Pokud je lineární aproximace (5) dobrá, můžeme použít všech výsledků z předchozího paragrafu. Rozdělení odhadu $\hat{\underline{\theta}}$ bude přibližně normální s kovariační maticí $\sigma^2 \underset{\sim}{H}^{-1}$ (viz (16.7)), kde prvky matice $\underset{\sim}{H}$ počítáme z (8). Použitelnost lineární aproximace záleží na průběhu funkcí h_i v tak velké oblasti prostoru parametrů $\underline{\theta}$, ve které je hustota pravděpodobnosti výsledku $\hat{\underline{\theta}}$ výrazně odlišná od nuly. Záleží tedy i na disperzi σ^2 naměřených hodnot $\underline{y}$ (viz diskusi o přibližných formulích v § 3). Představu o možnostech lineární aproximace poskytuje příklad v následujícím odstavci.

18. Příklad odhadu parametrů nelineárního modelu

Ukážeme použití metod předchozího paragrafu na příkladě odhadu tří parametrů Ω_o, Γ_o a α_o modelu

$$y(x; \Omega_o, \Gamma_o, \alpha_o) = \frac{1}{(x-\Omega_o)^2 + \Gamma_o^2} + \alpha_o. \tag{1}$$

Měříme N-tici hodnot $y_i = y(x_i) + \varepsilon_i$ pro známé $x_1,\ldots,x_N$, ε_i je náhodná chyba. Funkce (1) popisuje například tzv. Lorentzovský spektrální profil (závislost intenzity na frekvenci) čáry s centrální frekvencí Ω_o a položířkou Γ_o, přičtený ke konstantnímu pozadí α_o. Vzhledem k Ω_o, Γ_o je model (1) nelineární, α_o je lineární parametr.

Naměřená data byla simulována v počítači. Zvolili jsme

$$\Omega_o = 0,\ \Gamma_o = 1,\ \alpha_o = 0.5, \tag{2}$$

ekvidistantní síť s N=50 hodnotami x_i z intervalu $\langle -4,4\rangle$ a normálně rozdělené pseudonáhodné chyby ε se střední hodnotou 0 a standardní odchylkou

$$\sigma(\varepsilon) = 0.1 \tag{3}$$

stejnou pro všechny body x_i. V obrázku 37 jsou takto generovaná data znázorněna křížky.

Odhady $\hat{\Omega}$, $\hat{\Gamma}$ a $\hat{\alpha}$ jsme našli numerickou minimalizací součtu čtverců (17.4) s jednotkovými vahami (disperze jednotlivých y_i jsou stejné):

$$S(\Omega,\Gamma,\alpha) = \sum_{i=1}^{N} \left[y_i - y(x_i;\Omega,\Gamma,\alpha)\right]^2 . \tag{4}$$

Linearity modelu vzhledem k parametru α jsme nevyužili. Použitý minimalizační algoritmus hledá miminum funkce pomocí gradientu a matice druhých derivací (tzv. hessiánu). Označíme θ_1, θ_2, θ_3 po řadě parametry Ω, Γ, α; složky gradientu a hessiánu jsou

$$\frac{\partial S}{\partial \theta_j} = -2 \sum_{i=1}^{N} \left[y_i - y(x_i)\right] \frac{\partial y(x_i)}{\partial \theta_j} ,$$

$$\frac{\partial^2 S}{\partial \theta_j . \partial \theta_m} = 2 \sum_{i=1}^{N} \left\{ \frac{\partial y(x_i)}{\partial \theta_j} \frac{\partial y(x_i)}{\partial \theta_m} - \left[y_i - y(x_i)\right] \frac{\partial^2 y(x_i)}{\partial \theta_j \partial \theta_m} \right\}, \quad j,m = 1,\ldots,3. \tag{5}$$

V hessiánu zanedbáváme členy s druhými derivacemi (to je tzv. linearizace) a používáme vlastně (až na faktor 2) matici $\underset{\sim}{H}$ ze vztahu (17.8):

$$(\underset{\sim}{H})_{jm} = \frac{1}{2} \frac{\partial^2 S}{\partial \theta_j . \partial \theta_m} . \tag{6}$$

Vedlejším produktem minimalizace je tedy užitečná matice, která podle § 17 popisuje rozdělení odhadu $\hat{\underline{\theta}}$ parametrů v lineárním přiblížení ($\sigma^2(\varepsilon)\underset{\sim}{H}^{-1}$ je kovariační maticí normálního rozdělení $\hat{\underline{\theta}}$).

Suma čtverců (4) je minimální pro

$$\hat{\Omega} = 0.0470,\ \hat{\Gamma} = 0.999,\ \hat{\alpha} = 0.522; \tag{7}$$

modelová funkce (1) s těmito parametry je nakreslena plnou čarou v obr. 37. Součet čtverců v minimu je S_o=0.456. Další postup záleží na tom, je-li hodnota $\sigma(\varepsilon)$ známá nebo je třeba ji odhadnout (§ 16); ukážeme výsledky pro druhý případ. Počet stupňů volnosti je N-3=47, proto

$$\hat{\sigma}(\varepsilon) = \sqrt{S_o/47} \approx 0.0985,$$

v dobré shodě s hodnotou (3) použitou při generaci dat. Odhady standardních odchylek (16.15) jsou

$$\hat{\delta}_\Omega = 0.045, \quad \hat{\delta}_\Gamma = 0.024, \quad \hat{\delta}_\alpha = 0.017. \tag{8}$$

Počet stupňů volnosti je dostatečně velký k tomu, abychom namísto Studentova rozdělení v (16.14) mohli použít limitní normální rozdělení. Jako výsledek měření můžeme udat intervaly $\hat{\Omega} \pm \hat{\delta}_\Omega$, $\hat{\Gamma} \pm \hat{\delta}_\Gamma$, $\hat{\alpha} \pm \hat{\delta}_\alpha$ pro každý z parametrů zvlášť; jejich pravděpodobnostní obsah je asi 68%. Výhodnější může být odhad celé trojice elipsoidem (16.12) s pravděpodobnostním obsahem daným F - rozdělením. K tomu je třeba doplnit údaje o kovariační matici, nejlépe udáním korelačních koeficientů dvojic parametrů. V našem příkladě je

$$\varrho_{\Omega\Gamma} = 0.001, \quad \varrho_{\Omega\alpha} = 0.001, \quad \varrho_{\Gamma\alpha} = 0.551. \tag{9}$$

Předchozí úvaha o intervalových odhadech je založena na lineární aproximaci modelu (1) podle § 17. Posoudíme její použitelnost. Názorné je srovnání závislosti součtu čtverců odchylek (4) na parametrech pro modelovou funkci a její lineární aproximaci. Pro určitost budeme sledovat závislost na Γ. Numericky vypočteme funkci $S(\Gamma)$ pro pevně zadané Γ a parametry Ω, α nalezené tak, aby suma (4) byla minimální; s pomocí vztahu (17.7) najdeme parabolickou závislost $\tilde{S}(\Gamma)$ z lineární aproximace:

$$S(\Gamma) = \min_{\Omega,\alpha} S(\Omega,\Gamma,\alpha), \quad \tilde{S}(\Gamma) = S_0 + (\Gamma-\hat{\Gamma})^2 / (\underset{\sim}{H}^{-1})_{22} = S_0 + \hat{\sigma}^2(\varepsilon)(\Gamma-\hat{\Gamma})^2/\hat{\delta}_\Gamma^2 . \tag{10}$$

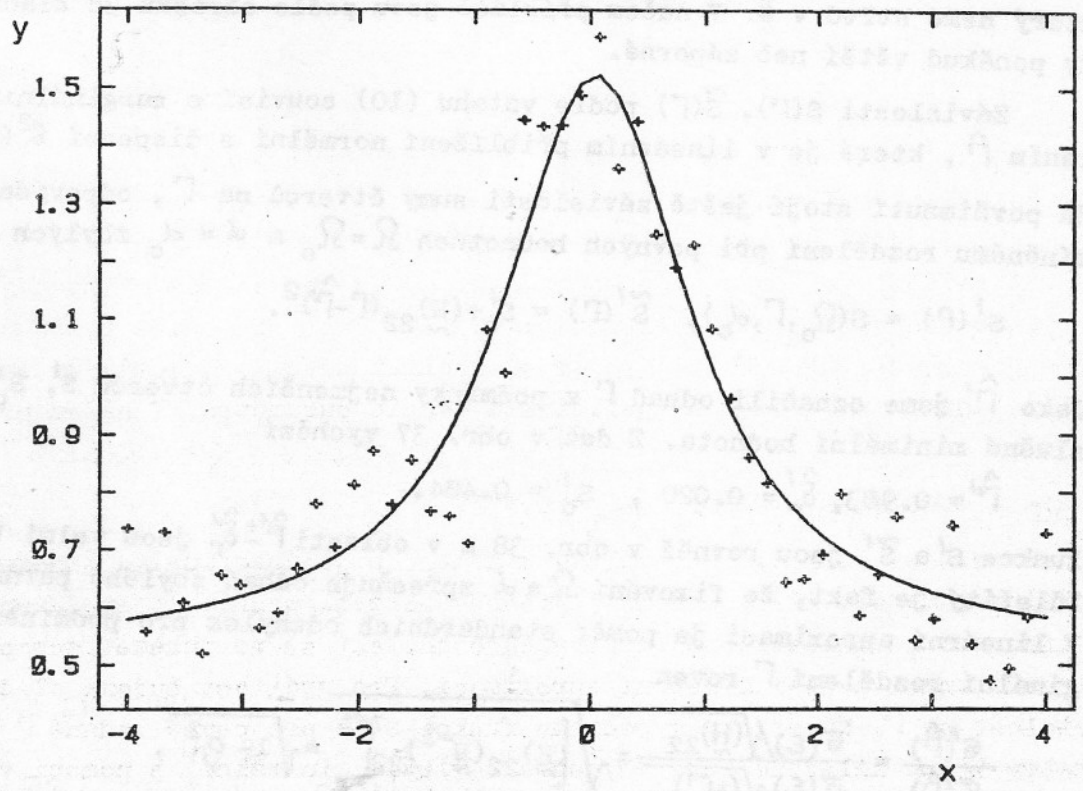
Při posunutí z minima o $\mp \hat{\delta}_\Gamma$ vzroste hodnota paraboly $\tilde{S}$ o $\hat{\sigma}^2(\varepsilon)$.

Obě funkce S, $\tilde{S}$ jsou nakresleny v obr. 38; je vidět, že se v intervalu $\hat{\Gamma} \pm \hat{\delta}_\Gamma$ liší velmi málo. Přestože závislost modelové funkce (1) na Γ je nelineární, odhady $\hat{\Gamma}$ jsou koncentrovány do dostatečně malého intervalu, v němž je lineární přiblížení vyhovující. Je třeba si uvědomit, že velikost intervalu $2\hat{\delta}_\Gamma$ je přímo úměrná střední kvadratické odchylce $\sigma(\varepsilon)$ naměřených hodnot. Budou-li chyby dat větší než v obr. 37, vliv nelinearity vzroste; naopak, pro menší chyby se bude dále zmenšovat.

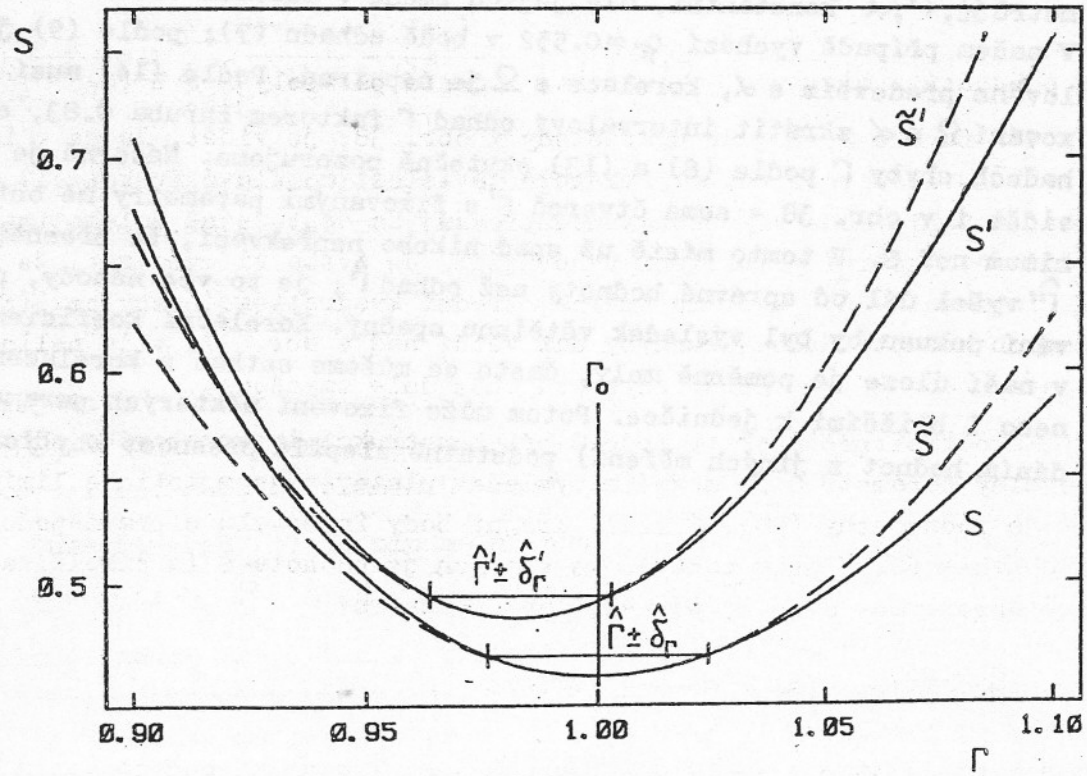
Započtení nelinearity modelu při konstrukci intervalového odhadu pro libovolný parametr Θ je možné a výsledek platný v asymptotické limitě $N \to \infty$ je jednoduchý ([14], § 9.3). Krajní body intervalu s pravděpodobnostním obsahem 68.3% jsou takové, ve kterých je hodnota S (a nikoli parabolické aproximace $\tilde{S}$) o $\hat{\sigma}^2(\varepsilon)$ větší než v minimu:

$$S(\hat{\theta} - \delta^{(-)}) = S(\hat{\theta} + \delta^{(+)}) = S_0 + \hat{\sigma}^2(\varepsilon). \tag{11}$$

Podobně vzrůst S o $k^2\hat{\sigma}^2(\varepsilon)$ definuje interval s pravděpodobnostním obsahem stejným, jako má v lineární aproximaci interval $\hat{\theta} \pm k\hat{\delta}$ (například 95.4% s $k=2$ a tedy s posunutím z minima o $4\hat{\sigma}^2(\varepsilon)$). Výsledkem je zpravidla interval,



Obr. 37. Experimentální data (křížky) a proložená závislost s parametry (18.8) (plná čára).



Obr. 38. Závislost součtu čtverců (18.10) a (18.12) na Γ, plná čára; parabolická aproximace, čárkovaná čára.

který nemá střed v $\hat{\theta}$. V našem příkladě jsou podle obrázku 38 kladné odchylky poněkud větší než záporné.

Závislosti $S(\Gamma)$, $\tilde{S}(\Gamma)$ podle vztahu (10) souvisí s marginálním rozdělením $\hat{\Gamma}$, které je v lineárním přiblížení normální s disperzí $\sigma^2(\varepsilon)(\underset{\sim}{H}^{-1})_{22}$. Za povšimnutí stojí ještě závislosti sumy čtverců na Γ, odpovídající podmíněnému rozdělení při pevných hodnotách $\Omega = \Omega_0$ a $\alpha = \alpha_0$ zbylých parametrů:

$$S'(\Gamma) = S(\Omega_0, \Gamma, \alpha_0), \quad \tilde{S}'(\Gamma) = S'_0 + (\underset{\sim}{H})_{22}(\Gamma - \hat{\Gamma})^2. \tag{12}$$

Jako $\hat{\Gamma}'$ jsme označili odhad Γ z podmínky nejmenších čtverců S', S'_0 je příslušná minimální hodnota. Z dat v obr. 37 vychází

$$\hat{\Gamma}' = 0.983, \ \hat{\delta}'_\Gamma = 0.020, \quad S'_0 = 0.484. \tag{13}$$

Funkce S' a $\tilde{S}'$ jsou rovněž v obr. 38 a v oblasti $\hat{\Gamma}' \pm \hat{\delta}'_\Gamma$ jsou velmi blízké. Důležitý je fakt, že fixování Ω a α zpřesňuje odhad zbylého parametru Γ. V lineární aproximaci je poměr standardních odchylek pro podmíněné a marginální rozdělení Γ roven

$$\frac{\sigma'(\hat{\Gamma})}{\sigma(\hat{\Gamma})} = \frac{\sigma(\varepsilon)\sqrt{(\underset{\sim}{H})_{22}}}{\sigma(\varepsilon)\sqrt{(\underset{\sim}{H}^{-1})_{22}}} = \sqrt{\left[(\underset{\sim}{H})_{22}(\underset{\sim}{H}^{-1})_{22}\right]^{-1}} = \sqrt{1-\rho_\Gamma^2}, \tag{14}$$

kde ρ_Γ je globální korelační koeficient parametru Γ ze vztahu (3.17). Matice $\underset{\sim}{H}$, $\underset{\sim}{H}^{-1}$, a tedy i veličiny z nich odvozené, sice nejsou v prostoru parametrů Ω, Γ, α konstantní, ale jejich změna v blízkém okolí odhadu je malá. V našem případě vychází $\rho_\Gamma \approx 0.552$ v bodě odhadu (7); podle (9) je Γ korelováno především s α, korelace s Ω je nepatrná. Podle (14) musí tedy fixování Ω a α zkrátit intervalový odhad Γ faktorem zhruba 0.83, což v odhadech chyby Γ podle (8) a (13) skutečně pozorujeme. Názorně je tento fakt vidět i v obr. 38 - suma čtverců S' s fixovanými parametry má ostřejší minimum než S. V tomto místě už snad nikoho nepřekvapí, že přesnější odhad $\hat{\Gamma}'$ vyšel dál od správné hodnoty než odhad $\hat{\Gamma}$; je to věc náhody, při opakování pokusu by byl výsledek většinou opačný. Korelační koeficient ~ 0.55 v naší úloze je poměrně malý, často se můžeme setkat s korelacemi ~ 0.99 nebo i bližšími k jedničce. Potom může fixování některých parametrů (zadáním hodnot z jiných měření) podstatně zlepšit přesnost zbylých odhadů.

III. Testy hypotéz

19. Statistické testy hypotéz

Podstatnou úlohou statistiky je odhad hledaných parametrů z naměřených hodnot. Naměřená data mohou být využita ještě jiným způsobem - k testu, který má rozhodnout o platnosti dané teorie nebo hypotézy, případně vybrat jednu z možných alternativ. Testované hypotéze velmi často odpovídá určitá hodnota některého parametru. Metody testu a odhadu mohou být v takovém případě podobné, formulace úlohy i výsledku jsou však podstatně odlišné. Pokud odhadujeme neznámý parametr θ_0 intervalem $\hat{\theta} \pm \sigma$, je výsledkem tvrzení o pravděpodobnostní souvislosti intervalu a neznámé hodnoty. Testujeme-li naopak hypotézu o zadané hodnotě θ_0, zformulujeme pravděpodobnostní tvrzení o možnostech správného nebo chybného přijetí či odmítnutí hypotézy na základě pozorovaných údajů.

Statistická hypotéza je soubor předpokladů, ze kterých plynou předpovědi rozdělení náhodných veličin. Pokud je předpověď rozdělení jednoznačná, označuje se hypotéza jako jednoduchá (například hypotéza, že rozdělení náhodné proměnné je normální se zadanou střední hodnotou a disperzí). V opačném případě jde o hypotézu složenou (normální rozdělení proměnné se střední hodnotou z nějakého intervalu).

Statistický test je založen na srovnání předpovědi plynoucí z hypotézy s pozorovanými daty. Je-li pozorovaný výsledek v rámci dané hypotézy málo pravděpodobný, soudíme na její neplatnost. Při testu hypotézy H_0 mohou nastat čtyři případy:

(a) H_0 platí, na základě testu ji přijímáme;

(b) H_0 neplatí, na základě testu ji zamítáme;

(c) H_0 platí, ale pomocí testu ji zamítáme; to je tzv. chyba prvního druhu;

(d) H_0 neplatí, ale pomocí testu ji přijímáme; to je chyba druhého druhu.

Shoda předpovědí plynoucích z hypotézy H_0 s pozorovanými fakty se testuje pomocí vhodné funkce t naměřených hodnot, které se ve statistické terminologii říká testovací statistika. Obor všech možných hodnot této náhodné proměnné rozdělíme na tzv. oblast přijetí H_0 a kritickou oblast. Kritickou oblast K vybíráme tak, že hodnoty t do ní padnou s malou pravděpodobností

$$\alpha = P(t \in K \mid H_0) \; ; \qquad (1)$$

nastane-li tento případ, hypotézu H_0 zamítneme. Říkáme, že pomocí testu zamítáme H_0 na hladině významnosti α (nebo s rizikem α). Přitom se můžeme podle (c) nahoře dopustit s pravděpodobností α chyby prvního druhu.

Existuje-li k H_0 jediná alternativní hypotéza H_1 (platí právě jedna z nich), můžeme najít pravděpodobnost chyby druhého druhu (d), čili neoprávněného přijetí H_0:

$$\beta = P(t \notin K | H_1). \tag{2}$$

Mírou možnosti oddělit H_0 a H_1 je tzv. síla (mohutnost) testu $1 - \beta$, která ovšem závisí na α. Tuto souvislost objasníme v následujícím příkladu.

Příklad testu zvětšení střední hodnoty normálního rozdělení

Uvažujme o následující situaci. Měřením intenzity zdroje záření (elektromagnetického nebo svazku částic) dostáváme náhodné výsledky x normálně rozdělené se střední hodnotou μ_0 a disperzí σ^2. Předpokládejme, že jsme znali μ_0 a potřebujeme rozhodnout, zda tato hodnota zůstala (hypotéza H_0) nebo se zvětšila na μ_1 (hypotéza H_1) po nějaké úpravě zdroje. Budeme postupovat tak, že změříme N-tici intenzit $x_1, \ldots, x_N$ a spočteme průměr (testovací statistiku)

$$t = \frac{1}{N} \sum_{i=1}^{N} x_i, \tag{3}$$

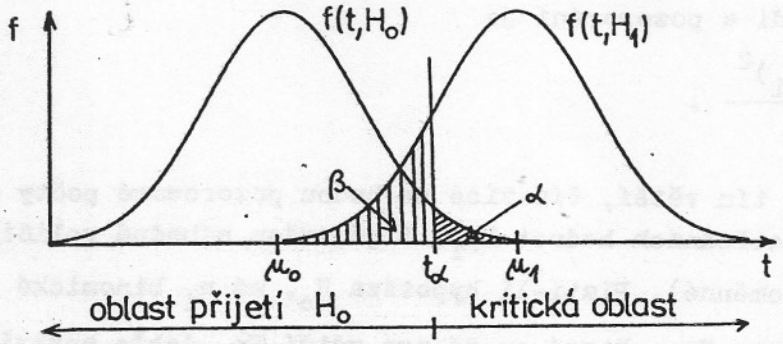
což je podle předpokladu o rozdělení x náhodná proměnná s rozdělením $N(\mu_0, \sigma^2/N)$ pokud platí H_0, respektive $N(\mu_1, \sigma^2/N)$ pokud platí alternativa H_1. V obrázku 39 jsou tyto dvě hustoty schematicky nakresleny spolu s plochami, reprezentujícími chyby α a β při zadané hranici t_α kritické oblasti. S použitím distribuční funkce (4.5) dostaneme

$$\alpha = P(t > t_\alpha | H_0) = 1 - \Phi\left(\frac{t_\alpha - \mu_0}{\sigma/\sqrt{N}}\right), \quad \beta = P(t < t_\alpha | H_1) = \Phi\left(\frac{t_\alpha - \mu_1}{\sigma/\sqrt{N}}\right). \tag{4}$$

S uvážením důsledků, které má přijetí jedné z hypotéz, je třeba rozhodnout o volbě kritické hodnoty t_α (tou jsou dány pravděpodobnosti chyb prvého i druhého druhu). S rostoucím počtem naměřených hodnot se rozdělení $f(t|H_0)$ i $f(t|H_1)$ zužují a v limitě $N \to \infty$ rozhodneme o platnosti jedné z hypotéz s libovolně malým rizikem chyby.

Dá se ukázat, že volba průměru (3) jako testovací statistiky je v tomto případě optimální - při zadaném α je síla $1 - \beta$ tohoto testu maximální. (V případě dvou jednoduchých hypotéz existuje nejsilnější test; obecný předpis pro jeho vyhledání udává tzv. Neymanova-Pearsonova věta). Můžeme použít mnoho jiných testů, ale žádný z nich nebude lepší než horší. Například lze testovat počet hodnot x_i, které jsou větší než μ_0; bude-li mnohem větší než N/2, platí pravděpodobně H_1. Z přesného rozboru tohoto testu zjistíme, že pro dané α dává větší β než test průměru (vztah (4)).

Přestože bychom mohli v dané situaci testovací veličinu t chápat



Obr. 39. Hustoty testovací statistiky t pro dvě hypotézy H_o a H_1.

jako odhad střední hodnoty proměnné x, je její použití v tomto příkladu podstatně jiné. Víme předem, že jsou pouze dvě možnosti (známé hodnoty μ_o nebo μ_1) a výsledkem měření je rozhodnutí, kterou z nich vybereme.

Testy dobré shody

Teorie testů hypotéz je velmi obsáhlou a propracovanou částí matematické statistiky. Dále se budeme zabývat pouze jedním druhem testů: prověrkou dané hypotézy H_o vzhledem ke všem jiným možným hypotézám (ne H_o). Takové srovnání H_o a alternativy, jaké jsme použili výše, pak nemá smysl; proti H_o stojí množina hypotéz, v níž jsou i takové, které vystihují data s libovolnou přesností. Chyba druhého druhu je v této situaci neznámá. Půjde tedy pouze o srovnání předpovědí plynoucích z H_o s naměřenými daty - tzv. kriteria dobré shody. Ve dvou následujících odstavcích jsou popsány dva z mnoha známých testů.

20. Pearsonův test dobré shody

Předpokládejme, že z N-tice naměřených hodnot $x_1,\ldots,x_N$ byl sestaven histogram s k sloupky (buňkami). V i-tém sloupku jsou hodnoty z intervalu $\langle m_i, M_i)$, jejich počet označíme n_i; zřejmě platí

$$\sum_{i=1}^{k} n_i = N. \qquad (1)$$

Počty "událostí" n_i v buňkách jsou náhodné veličiny s binomickým rozdělením (§ 7). Jsou určeny pravděpodobnostmi p_i toho, že naměřená hodnota padne do i-té buňky, neboli rozdělením měřené veličiny x:

$$p_i = P\left[x \in \langle m_i, M_i)\right] = F(M_i) - F(m_i), \quad i=1,\ldots,k. \qquad (2)$$

Zde je F(x) distribuční funkce náhodné proměnné x. Z hypotézy H_o, že se x řídí daným rozdělením, plyne kromě jiného i předpověď pravděpodobností různých počtů v buňkách histogramu. Vhodnou testovací veličinou pro srov-

nání shody předpovědi a pozorování je

$$T = \sum_{i=1}^{k} \frac{(n_i - Np_i)^2}{Np_i} \quad . \qquad (3)$$

Hodnota T bude tím větší, čím více se budou pozorované počty n_i lišit od očekávaných středních hodnot Np_i. T je ovšem náhodná veličina (n_i jsou náhodné proměnné). Platí-li hypotéza H_o, má n_i binomické rozdělení se střední hodnotou Np_i, které se dá pro větší Np_i dobře aproximovat normální hustotou (§ 7). Jednotlivé sčítance v (3) nejsou nezávislé, n_i splňují podmínku (1). Dá se ale ukázat, že veličina T je součtem k-1 kvadrátů nezávislých náhodných proměnmých, z nichž každá má přibližně standardní normální rozdělení N(0,1). Rozdělení T je tedy přibližně χ^2 s k-1 stupněm volnosti. Aproximace je tím lepší, čím větší jsou očekávané počty Np_i. Jako podmínka použitelnosti se obvykle uvádí $Np_i > 5$, nebo alespoň malý počet (ne více než 20%) intervalů s Np_i v rozmezí 1 až 5.

Znalosti rozdělení T využijeme k testu hypotézy H_o pomocí následující úvahy. Je-li pozorovaná hodnota T velká, mohly nastat dva případy:

H_o platí, velká hodnota vyšla náhodou;

H_o neplatí, velká hodnota vyšla proto, že rozdělení dat je jiné.

Rozhodneme se tedy, že H_o zamítneme, je-li hodnota T dostatečně málo pravděpodobná. Ve statistice se užívá ustáleného způsobu vyjadřování této souvislosti. Hypotézu H_o zamítáme na hladině významnosti α (nebo s rizikem α), jestliže vyšla hodnota $T \geq T_\alpha$, přičemž pravděpodobnost tohoto výsledku je α:

$$P(T \geq T_\alpha) = 1 - P(T < T_\alpha) = 1 - F_{\chi^2_{k-1}}(T_\alpha) = \alpha \quad . \qquad (4)$$

Distribuční funkce $F_{\chi^2_{k-1}}$ umožňuje najít k zadané pravděpodobnosti α kritickou hodnotu T_α. Podle rozdělení, kterým se řídí T, se tomuto testu říká Pearsonovo χ^2 - kriterium dobré shody.

Přesná hodnota α hladiny významnosti ve statistickém testu není zřejmě důležitá. Zacházíme se náhodnými jevy a pozorujeme občas i velmi málo pravděpodobné výsledky. Zkušenost však ukazuje, že pravděpodobnosti α pod $\sim$5% znamenají silný podnět k úvahám o tom, nemá-li být testovaná hypotéza nahrazena nějakou vhodnější.

Test při odhadovaných parametrech rozdělení

Zatím jsme předpokládali, že testované rozdělení nezávisí na žádném parametru určovaném z naměřených dat. Pokud ze souboru $x_1, \ldots, x_N$ naměřených hodnot nejprve odhadujeme parametry $\underline{\theta}$ distribuční funkce F(x), nemá

už proměnná (3) rozdělení χ^2_{k-1}. Dá se ukázat, že odhad r-tice parametrů z dat sdružených do histogramu vede ke zmenšení počtu stupňů volnosti v χ^2- - rozdělení proměnné T na k-r-1 (ztrácí se r stupňů volnosti). Je-li odhad založen na výchozích datech, bez sdružení do buněk histogramu, je rozdělení proměnné T někde mezi χ^2_{k-1} a χ^2_{k-r-1}. Pokud je k dostatečně velké proti r, je rozdíl mezi oběma krajními distribucemi malý a přesnější znalost rozdělení T není nutná.

Příklad použití χ^2- testu

Ukážeme použití Pearsonova testu na příkladě dat z obr. 3 - výsledků N=1000-krát opakovaného měření propustnosti infračerveným spektrometrem. Budeme testovat hypotézu H_o tvrdící, že data jsou rozdělená normálně se střední hodnotou $\hat{\mu}$ a disperzí $\hat{\sigma}^2$, odhadnutými metodou maximální věrohodnosti. Vychází

$$\hat{\mu} = 274, \quad \hat{\sigma}^2 = 4624 \quad (\hat{\sigma} \approx 68),$$

pravděpodobnosti (2) dostaneme s pomocí distribuční funkce (4.5) normálního rozdělení s těmito parametry. Pozorovaný a očekávaný počet událostí v histogramu z obr. 3 je spolu s příspěvkem každého sloupku do sumy (3) zapsán do následující tabulky.

i	1	2	3	4	5	6	7	8	9	10
n_i	3	1	2	2	6	11	24	20	43	44
Np_i	1.45	2.37	3.77	5.78	8.58	12.3	17.1	23.0	29.8	37.5
χ^2	1.67	0.79	0.83	2.47	0.78	0.14	2.77	0.39	5.81	1.14

i	11	12	13	14	15	16	17	18	19	20
n_i	42	55	55	72	72	73	69	75	63	50
Np_i	45.5	53.5	60.7	66.7	70.9	72.8	72.4	69.6	64.7	58.1
χ^2	0.27	0.04	0.54	0.42	0.02	0.00	0.16	0.43	0.04	1.13

i	21	22	23	24	25	26	27	28	29	30
n_i	46	31	27	36	18	23	14	10	3	10
Np_i	50.5	42.5	34.5	27.2	20.7	15.2	10.8	7.43	4.94	3.18
χ^2	0.41	3.10	1.65	2.88	0.34	4.02	0.95	0.89	0.76	14.6

Hodnota T podle (3) je 49.4. Protože jsme dva parametry odhadovali, je rozdělení náhodné proměnné T podle předchozího výkladu ohraničeno distribucemi χ^2_{29} a χ^2_{27}.

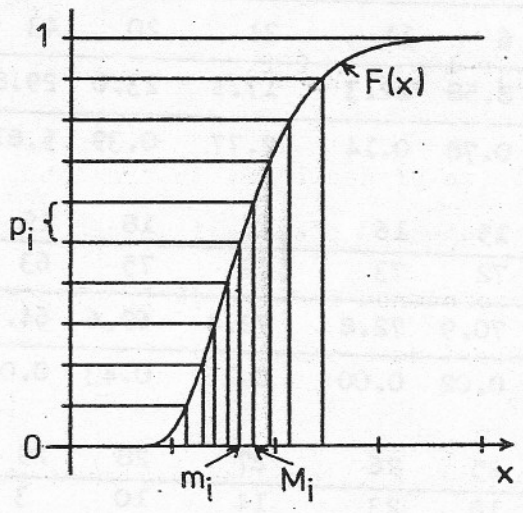
S pomocí tabulek v dodatku D1 můžeme formulovat výsledek testu - hypotézu

H_o zamítáme s malým rizikem $\alpha \approx 0.01$ (kritická hodnota $T_{o.01}$ je asi 49.6 pro 29 stupňů volnosti).

Podíváme-li se pozorně do hořejší tabulky, všimneme si nápadně velkého příspěvku poslední buňky do T. Kdyby v ní bylo jen o několik událostí méně, třeba místo deseti jen pět, vyšla by hodnota T zhruba 36; pak by hypotéza o normálním rozdělení byla zamítnuta χ^2- testem na hladině větší než 0.1 (viz D1). Při takovém výsledku býváme obvykle s hypotézou spokojeni. Pravděpodobně nejlepší vysvětlení těchto faktů je takové, že normální rozdělení skutečně dobře vystihuje registrovaná data. V průběhu 1000-krát opakovaného měření došlo asi k rušivému zásahu, při kterém bylo naměřeno několik příliš velkých hodnot (typické jsou náhodné impulzy v elektrické síti). V § 21 budeme testovat normalitu jedné menší části dat z obr. 3 (150 bodů) jiným testem; výsledek - dobrá shoda s hypotézou - hořejší závěry podporuje.

Volba buněk histogramu

Při sdružování naměřených dat do histogramu se ztrácí část informace (o rozdělení hodnot uvnitř jednotlivých buněk). To je nežádoucí jev, který ovlivňuje i kvalitu Pearsonova testu. Příliš jemné dělení intervalu vede zase k malému počtu událostí v buňkách a χ^2 - test nelze použít. Základní pravidlo pro optimální volbu buněk říká, že pravděpodobnosti (2) mají být stejné. Interval $\langle 0,1 \rangle$ funkčních hodnot distribuční funkce F(x) je tedy třeba rozdělit na k stejně velkých dílů a odečíst odpovídající argumenty jako hranice sousedních buněk (obr. 40). Optimální volba počtu buněk k vychází z požadavku maximální mohutnosti kriteria a je poměrně komplikovaná [14]. Spokojíme se s konstatováním, že optimální k roste s počtem dat N jako $N^{2/5}$ a pro N=200 je doporučená hodnota $k \approx 30$, pro N=500 zhruba $k \approx 43$. Histogram z obr. 3, který jsme použili v hořejším příkladu, nebyl z hlediska testu vybrán správně. Chtěli jsme však zachovat podobnost s hustotou pravděpodobnosti a proto byly zvoleny stejné velikosti buněk.



Obr. 40. Volba buněk histogramu se stejnou pravděpodobností.

21. Kolmogorovův test dobré shody

Při vytvoření histogramu z naměřených údajů se část informace ztrácí, proto je lepší testovat hypotézu o rozdělení dat bez sdružování do buněk ("třídních intervalů" ve statistické terminologii). Úspěšná kriteria jsou založena na sledování odchylek empirické a hypotetické distribuční funkce. Empirická distribuční funkce $S_N(x)$ souboru $x_1,\dots,x_N$ je po částech konstantní se skokem velikosti $1/N$ v každé naměřené hodnotě. Označíme-li $x(1),\dots,x(N)$ naměřené údaje uspořádané podle velikosti od nejmenšího k největšímu, je

$$S_N(x) = \begin{cases} 0 & \text{pro } x < x(1), \\ i/N & \text{pro } x \in \langle x(i),\ x(i+1)\rangle,\ i=1,\dots,N-1, \\ 1 & \text{pro } x \geq x(N). \end{cases} \tag{1}$$

Tuto empirickou distribuční funkci jsme už použili v § 11 pro kvalitativní srovnání s hypotetickou distribuční funkcí $F(x)$.

Vhodnou mírou odlišnosti $S_N(x)$ a $F(x)$ je maximum odchylky

$$D_N = \max_x \left| S_N(x) - F(x) \right|. \tag{2}$$

Hodnota D_N se celkem snadno spočte, protože maximum rozdílu $S_N(x)-F(x)$ nastává v některém z naměřených bodů $x(1),\dots,x(N)$. D_N je náhodná proměnná, která má za předpokladu platnosti hypotézy o rozdělení $F(x)$ asymptotickou distribuční funkci

$$F_K(z) = \lim_{N\to\infty} P\left[\sqrt{N}\, D_N > z\right] = 2 \sum_{r=1}^{\infty} (-1)^{r-1} \exp(-2r^2 z^2). \tag{3}$$

Průběh funkce $F_K(z)$ je v obr. 41; obvykle se předpokládá, že proměnná $\sqrt{N}\, D_N$ má asymptotické rozdělení (3) s dostatečnou přesností už při $N \geq 80$.

V Kolmogorovově testu posuzujeme pozorovanou hodnotu $\sqrt{N}\, D_N$. Vyjde-li příliš velká, zamítneme hypotézu o rozdělení dat podle $F(x)$: pro

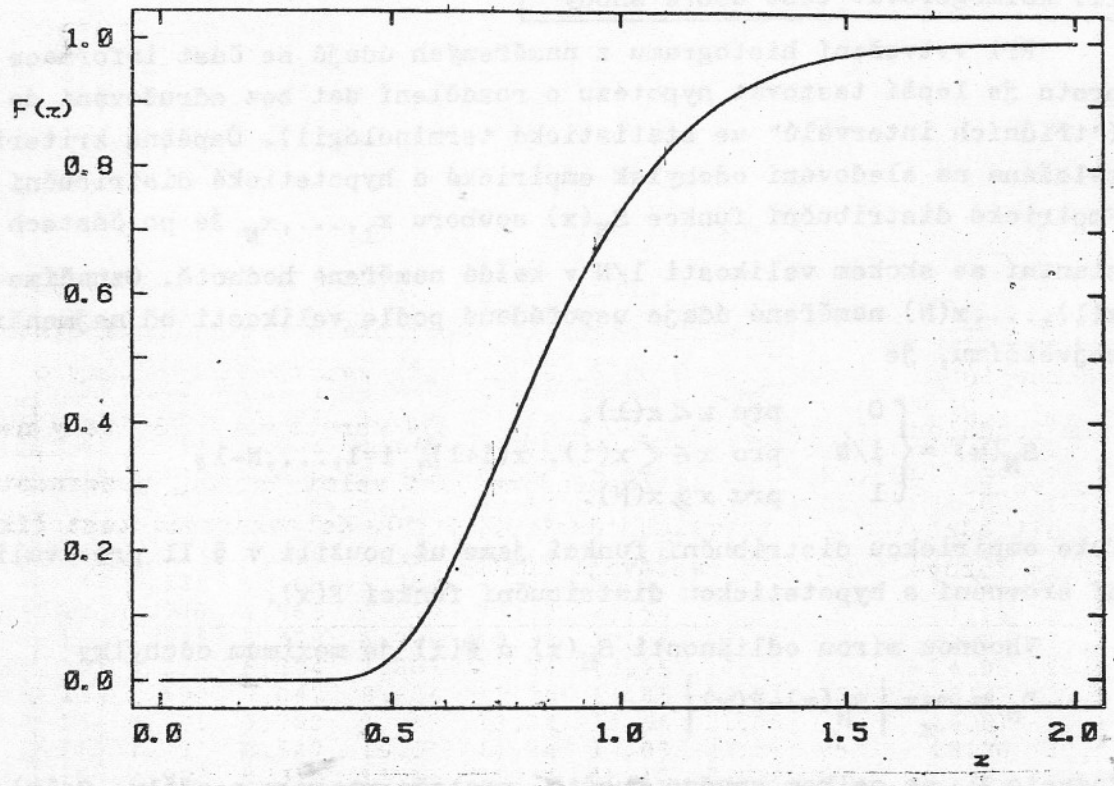
$$\sqrt{N}\, D_N > z_\alpha \ , \text{ kde } F_K(z_\alpha) = 1 - \alpha, \tag{4}$$

zamítáme hypotézu na úrovni α (s rizikem α). Kritické hodnoty jsou např.

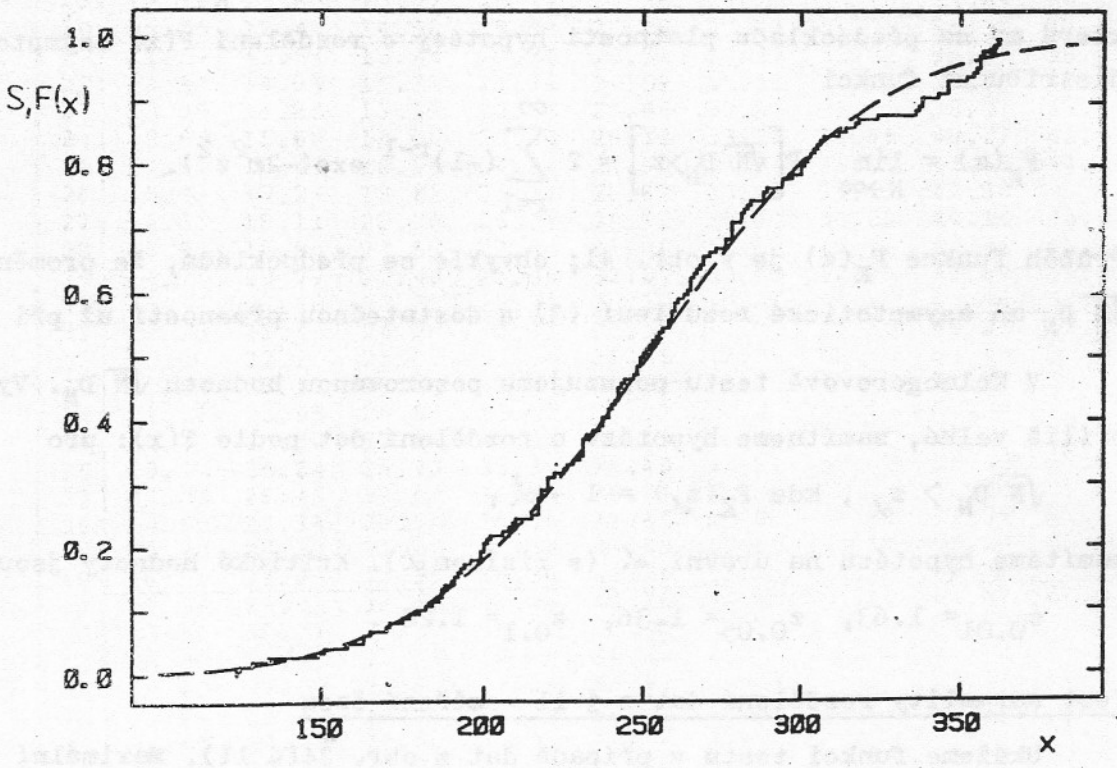
$$z_{0.01} = 1.63, \quad z_{0.05} = 1.36, \quad z_{0.1} = 1.22 \ . \tag{5}$$

Test normality rozdělení dat z § 11 - měření času

Ukážeme funkci testu v případě dat z obr. 24(§ 11). Maximální odchylka mezi empirickou distribuční funkcí a čárkovaně nakreslenou hypotetickou normální distribuční funkcí je $D_{200} = 0.0513$; testovaná hodnota $\sqrt{200}\, D_{200} = 0.725$ padá podle obrázku 41 do oblasti hodnot velmi pravděpodobných. Hladina významnosti pro $z_\alpha = 0.725$ je $\alpha = 1-F(z_\alpha) = 0.67$, riziko při



Obr. 41. Distribuční funkce (21.3) Kolmogorovova kriteria.



Obr. 42. Empirická distribuční funkce prvních 150-ti hodnot ze souboru dat z infračerveného spektrometru (§ 2, obr. 3), plná čára. Normální distribuční funkce, čárkovaná čára.

zamítnutí hypotézy daného normálního rozdělení je příliš velké. Připomeneme ještě, že není možné uvažovat tak, že shoda s hypotézou je tím lepší, čím menší je testované hodnota $\sqrt{N}\, D_N$. Jde o hodnotu náhodné proměnné, která podle obr. 41 padá s velkou pravděpodobností do intervalu mezi 0.5 a 1.5; pravděpodobnost výsledku $\sqrt{N}\, D_N < 0.25$ je prakticky nulová ($\sim 3 \times 10^{-8}$).

Test normality dat ze spektrometru

Z tisíce hodnot použitých v příkladu χ^2 - testu v § 20 jsme vybrali prvních 150 a vypočetli maximální odchylku empirické a hypotetické distribuční funkce (obr. 42): $D_{150} = 0.0538$. Riziko při zamítnutí hypotézy o normálním rozdělení je pro $z_\alpha = \sqrt{150}\, D_{150} = 0.659$ zhruba $\alpha = 0.78$, tedy nepřípustně velké. Shodu v obr. 42 posuzujeme jako velmi dobrou, pozornost může vzbudit odchylka v intervalu hodnot $300 \div 350$. Kolmogorovův test říká, že odchylky takové velikosti nastávají často.

Dodatky

Dl. χ^2- rozdělení

V tabulce jsou hodnoty x_P, pro které je pravděpodobnost $P(x < x_P) = F\chi^2_n(x_P) = P$, v závislosti na počtu stupňů volnosti n a pravděpodobnosti P.

n \ P:	0.050	0.100	0.200	0.500	0.683	0.900	0.954	0.980	0.990
1	0.004	0.016	0.064	0.455	1.001	2.705	3.981	5.411	6.632
2	0.103	0.211	0.446	1.386	2.298	4.605	6.158	7.824	9.210
3	0.352	0.584	1.005	2.366	3.529	6.251	8.000	9.837	11.34
4	0.711	1.064	1.649	3.357	4.722	7.779	9.689	11.67	13.28
5	1.145	1.610	2.343	4.351	5.891	9.236	11.29	13.39	15.09
6	1.635	2.204	3.070	5.348	7.042	10.64	12.82	15.03	16.81
7	2.167	2.833	3.822	6.346	8.180	12.02	14.31	16.62	18.48
8	2.733	3.490	4.594	7.344	9.308	13.36	15.76	18.17	20.09
9	3.325	4.168	5.380	8.343	10.43	14.68	17.18	19.68	21.67
10	3.940	4.865	6.179	9.342	11.54	15.99	18.58	21.16	23.21
11	4.575	5.578	6.989	10.34	12.65	17.28	19.95	22.62	24.72
12	5.226	6.304	7.807	11.34	13.75	18.55	21.31	24.05	26.22
13	5.892	7.042	8.634	12.34	14.85	19.81	22.66	25.47	27.69
14	6.571	7.790	9.467	13.34	15.94	21.06	23.99	26.87	29.14
15	7.261	8.547	10.31	14.34	17.03	22.31	25.30	28.26	30.58
16	7.962	9.312	11.15	15.34	18.12	23.54	26.61	29.63	32.00
17	8.672	10.09	12.00	16.34	19.20	24.77	27.91	31.00	33.41
18	9.390	10.86	12.86	17.34	20.29	25.99	29.20	32.35	34.81
19	10.12	11.65	13.72	18.34	21.36	27.20	30.48	33.69	36.19
20	10.85	12.44	14.58	19.34	22.44	28.41	31.75	35.02	37.57
21	11.59	13.24	15.44	20.34	23.52	29.62	33.02	36.34	38.93
22	12.34	14.04	16.31	21.34	24.59	30.81	34.28	37.66	40.29
23	13.09	14.85	17.19	22.34	25.66	32.01	35.53	38.97	41.64
24	13.85	15.66	18.06	23.34	26.73	33.20	36.78	40.27	42.98
25	14.61	16.47	18.94	24.34	27.80	34.38	38.03	41.57	44.31
26	15.38	17.29	19.82	25.34	28.87	35.56	39.26	42.86	45.64
27	16.15	18.11	20.70	26.34	29.94	36.74	40.50	44.14	46.96
28	16.93	18.94	21.59	27.34	31.00	37.92	41.73	45.42	48.28
29	17.71	19.77	22.48	28.34	32.07	39.09	42.95	46.69	49.59
30	18.49	20.60	23.36	29.34	33.13	40.26	44.17	47.96	50.89
31	19.28	21.43	24.26	30.34	34.19	41.42	45.39	49.23	52.19
32	20.07	22.27	25.15	31.34	35.25	42.58	46.60	50.49	53.49
33	20.87	23.11	26.04	32.34	36.31	43.75	47.81	51.74	54.78
34	21.66	23.95	26.94	33.34	37.37	44.90	49.02	53.00	56.06
35	22.47	24.80	27.84	34.34	38.43	46.06	50.23	54.24	57.34
36	23.27	25.64	28.73	35.34	39.48	47.21	51.43	55.49	58.62
37	24.07	26.49	29.64	36.34	40.54	48.36	52.63	56.73	59.89
38	24.88	27.34	30.54	37.34	41.60	49.51	53.82	57.97	61.16
39	25.70	28.20	31.44	38.34	42.65	50.66	55.02	59.20	62.43
40	26.51	29.05	32.34	39.34	43.70	51.81	56.21	60.44	63.69
41	27.33	29.91	33.25	40.34	44.76	52.95	57.39	61.67	64.95
42	28.14	30.77	34.16	41.34	45.81	54.09	58.58	62.89	66.21
43	28.96	31.63	35.07	42.34	46.86	55.23	59.76	64.12	67.46
44	29.79	32.49	35.97	43.34	47.91	56.37	60.95	65.34	68.71
45	30.61	33.35	36.88	44.34	48.96	57.51	62.13	66.56	69.96
46	31.44	34.22	37.80	45.34	50.02	58.64	63.30	67.77	71.20
47	32.27	35.08	38.71	46.34	51.06	59.77	64.48	68.99	72.44
48	33.10	35.95	39.62	47.34	52.11	60.91	65.65	70.20	73.68
49	33.93	36.82	40.53	48.33	53.16	62.04	66.82	71.41	74.92
50	34.76	37.69	41.45	49.33	54.21	63.17	67.99	72.61	76.15

D2. Studentovo rozdělení

V tabulce jsou hodnoty t_p, pro které je pravděpodobnost $P(|t| < t_p) = F_n(t_p) - F_n(-t_p) = P$, v závislosti na počtu stupňů volnosti n a pravděpodobnosti P.

n \ P	0.050	0.100	0.200	0.500	0.683	0.900	0.954	0.980	0.990
1	0.079	0.158	0.325	1.000	1.839	6.314	13.82	31.82	63.66
2	0.071	0.142	0.289	0.816	1.322	2.920	4.500	6.965	9.925
3	0.068	0.137	0.277	0.765	1.198	2.353	3.292	4.541	5.841
4	0.067	0.134	0.271	0.741	1.142	2.132	2.858	3.747	4.604
5	0.066	0.132	0.267	0.727	1.111	2.015	2.640	3.365	4.032
6	0.065	0.131	0.265	0.718	1.091	1.943	2.508	3.143	3.707
7	0.065	0.130	0.263	0.711	1.077	1.895	2.421	2.998	3.499
8	0.065	0.130	0.262	0.706	1.067	1.860	2.359	2.896	3.355
9	0.064	0.129	0.261	0.703	1.059	1.833	2.313	2.821	3.250
10	0.064	0.129	0.260	0.700	1.053	1.812	2.277	2.764	3.169
11	0.064	0.129	0.260	0.697	1.048	1.796	2.249	2.718	3.106
12	0.064	0.128	0.259	0.695	1.044	1.782	2.225	2.681	3.055
13	0.064	0.128	0.259	0.694	1.041	1.771	2.206	2.650	3.012
14	0.064	0.128	0.258	0.692	1.038	1.761	2.189	2.624	2.977
15	0.064	0.128	0.258	0.691	1.035	1.753	2.175	2.602	2.947
16	0.064	0.128	0.258	0.690	1.033	1.746	2.163	2.583	2.921
17	0.064	0.128	0.257	0.689	1.031	1.740	2.153	2.567	2.898
18	0.064	0.127	0.257	0.688	1.029	1.734	2.143	2.552	2.878
19	0.064	0.127	0.257	0.688	1.028	1.729	2.135	2.539	2.861
20	0.063	0.127	0.257	0.687	1.026	1.725	2.128	2.528	2.845
21	0.063	0.127	0.257	0.686	1.025	1.721	2.121	2.518	2.831
22	0.063	0.127	0.256	0.686	1.024	1.717	2.115	2.508	2.819
23	0.063	0.127	0.256	0.685	1.023	1.714	2.109	2.500	2.807
24	0.063	0.127	0.256	0.685	1.022	1.711	2.104	2.492	2.797
25	0.063	0.127	0.256	0.684	1.021	1.708	2.100	2.485	2.787
26	0.063	0.127	0.256	0.684	1.020	1.706	2.096	2.479	2.779
27	0.063	0.127	0.256	0.684	1.020	1.703	2.092	2.473	2.771
28	0.063	0.127	0.256	0.683	1.019	1.701	2.088	2.467	2.763
29	0.063	0.127	0.256	0.683	1.018	1.699	2.085	2.462	2.756
30	0.063	0.127	0.256	0.683	1.018	1.697	2.082	2.457	2.750
31	0.063	0.127	0.256	0.682	1.017	1.696	2.079	2.453	2.744
32	0.063	0.127	0.255	0.682	1.017	1.694	2.076	2.449	2.738
33	0.063	0.127	0.255	0.682	1.016	1.692	2.074	2.445	2.733
34	0.063	0.127	0.255	0.682	1.016	1.691	2.071	2.441	2.728
35	0.063	0.127	0.255	0.682	1.015	1.690	2.069	2.438	2.724
36	0.063	0.127	0.255	0.681	1.015	1.688	2.067	2.434	2.719
37	0.063	0.127	0.255	0.681	1.014	1.687	2.065	2.431	2.715
38	0.063	0.127	0.255	0.681	1.014	1.686	2.063	2.429	2.712
39	0.063	0.126	0.255	0.681	1.014	1.685	2.061	2.426	2.708
40	0.063	0.126	0.255	0.681	1.013	1.684	2.059	2.423	2.704
41	0.063	0.126	0.255	0.681	1.013	1.683	2.058	2.421	2.701
42	0.063	0.126	0.255	0.680	1.013	1.682	2.056	2.418	2.698
43	0.063	0.126	0.255	0.680	1.012	1.681	2.055	2.416	2.695
44	0.063	0.126	0.255	0.680	1.012	1.680	2.053	2.414	2.692
45	0.063	0.126	0.255	0.680	1.012	1.679	2.052	2.412	2.690
46	0.063	0.126	0.255	0.680	1.012	1.679	2.051	2.410	2.687
47	0.063	0.126	0.255	0.680	1.011	1.678	2.050	2.408	2.685
48	0.063	0.126	0.255	0.680	1.011	1.677	2.049	2.407	2.682
49	0.063	0.126	0.255	0.680	1.011	1.677	2.047	2.405	2.680
50	0.063	0.126	0.255	0.679	1.011	1.676	2.046	2.403	2.678

D3. F- rozdělení

V tabulce jsou hodnoty x_P, pro které je pravděpodobnost $P(x < x_P) = F_{m,m'}(x_P) = P$, v závislosti na počtech stupňů volnosti m, m' a pravděpodobnosti P.

m= 2

m' \ P	0.050	0.100	0.200	0.500	0.683	0.900	0.954	0.980	0.990
5	0.052	0.108	0.233	0.799	1.458	3.780	6.067	9.454	13.27
10	0.052	0.106	0.228	0.743	1.292	2.924	4.256	5.934	7.559
15	0.051	0.106	0.226	0.726	1.242	2.695	3.807	5.135	6.359
20	0.051	0.106	0.226	0.718	1.217	2.589	3.606	4.788	5.849
25	0.051	0.106	0.225	0.713	1.203	2.528	3.492	4.593	5.568
30	0.051	0.106	0.225	0.709	1.194	2.489	3.418	4.470	5.390
35	0.051	0.106	0.225	0.707	1.187	2.461	3.367	4.384	5.268
40	0.051	0.106	0.224	0.705	1.182	2.440	3.329	4.321	5.179
50	0.051	0.106	0.224	0.703	1.176	2.412	3.277	4.235	5.057
60	0.051	0.106	0.224	0.701	1.171	2.393	3.243	4.179	4.977
70	0.051	0.106	0.224	0.700	1.168	2.380	3.219	4.139	4.922
80	0.051	0.105	0.224	0.699	1.166	2.370	3.201	4.110	4.881
90	0.051	0.105	0.224	0.699	1.164	2.363	3.187	4.087	4.849
100	0.051	0.105	0.224	0.698	1.162	2.356	3.176	4.069	4.824

m= 3

m' \ P	0.050	0.100	0.200	0.500	0.683	0.900	0.954	0.980	0.990
5	0.111	0.188	0.337	0.907	1.523	3.619	5.659	8.669	12.06
10	0.114	0.191	0.336	0.845	1.337	2.728	3.835	5.218	6.551
15	0.115	0.192	0.335	0.826	1.281	2.490	3.388	4.447	5.416
20	0.115	0.193	0.335	0.816	1.254	2.380	3.187	4.113	4.937
25	0.116	0.193	0.335	0.811	1.238	2.317	3.074	3.927	4.675
30	0.116	0.193	0.335	0.807	1.227	2.276	3.001	3.809	4.509
35	0.116	0.194	0.335	0.804	1.220	2.247	2.950	3.727	4.395
40	0.116	0.194	0.335	0.802	1.214	2.226	2.913	3.667	4.312
50	0.117	0.194	0.335	0.800	1.207	2.197	2.862	3.585	4.199
60	0.117	0.194	0.335	0.798	1.201	2.177	2.828	3.532	4.125
70	0.117	0.194	0.335	0.796	1.198	2.164	2.804	3.494	4.074
80	0.117	0.194	0.335	0.795	1.195	2.153	2.787	3.466	4.036
90	0.117	0.194	0.335	0.795	1.193	2.146	2.773	3.445	4.006
100	0.117	0.194	0.335	0.794	1.191	2.139	2.762	3.428	3.983

m= 4

m' \ P	0.050	0.100	0.200	0.500	0.683	0.900	0.954	0.980	0.990
5	0.160	0.247	0.404	0.965	1.552	3.520	5.426	8.233	11.39
10	0.168	0.255	0.407	0.899	1.353	2.605	3.591	4.816	5.994
15	0.171	0.258	0.408	0.878	1.293	2.361	3.143	4.058	4.893
20	0.172	0.260	0.409	0.868	1.264	2.249	2.942	3.731	4.431
25	0.173	0.261	0.410	0.862	1.247	2.184	2.829	3.549	4.177
30	0.174	0.262	0.410	0.858	1.235	2.142	2.756	3.434	4.018
35	0.175	0.262	0.410	0.856	1.227	2.113	2.706	3.354	3.908
40	0.175	0.263	0.410	0.854	1.221	2.091	2.668	3.295	3.828
50	0.175	0.263	0.411	0.851	1.213	2.061	2.617	3.215	3.720
60	0.176	0.264	0.411	0.849	1.208	2.041	2.583	3.163	3.649
70	0.176	0.264	0.411	0.847	1.204	2.027	2.560	3.127	3.600
80	0.176	0.264	0.411	0.846	1.201	2.016	2.542	3.100	3.563
90	0.176	0.265	0.411	0.846	1.199	2.008	2.528	3.079	3.535
100	0.177	0.265	0.411	0.845	1.197	2.002	2.517	3.062	3.513

m= 5

m' \ P:	0.050	0.100	0.200	0.500	0.683	0.900	0.954	0.980	0.990
5	0.198	0.290	0.449	1.000	1.567	3.453	5.273	7.953	10.97
10	0.211	0.303	0.456	0.932	1.359	2.522	3.429	4.555	5.636
15	0.217	0.309	0.460	0.911	1.296	2.273	2.980	3.805	4.556
20	0.219	0.312	0.462	0.900	1.266	2.158	2.779	3.482	4.103
25	0.221	0.314	0.463	0.894	1.248	2.092	2.665	3.302	3.855
30	0.222	0.315	0.464	0.890	1.236	2.049	2.592	3.188	3.699
35	0.223	0.316	0.464	0.887	1.228	2.019	2.541	3.109	3.592
40	0.224	0.317	0.465	0.885	1.221	1.997	2.504	3.051	3.514
50	0.225	0.318	0.466	0.882	1.213	1.966	2.452	2.972	3.408
60	0.226	0.318	0.466	0.880	1.207	1.946	2.419	2.921	3.339
70	0.226	0.319	0.466	0.879	1.203	1.931	2.395	2.885	3.291
80	0.227	0.319	0.467	0.878	1.200	1.921	2.377	2.858	3.255
90	0.227	0.320	0.467	0.877	1.197	1.912	2.364	2.837	3.228
100	0.227	0.320	0.467	0.876	1.195	1.906	2.353	2.821	3.206

m= 6

m' \ P:	0.050	0.100	0.200	0.500	0.683	0.900	0.954	0.980	0.990
5	0.228	0.322	0.482	1.024	1.576	3.405	5.166	7.758	10.67
10	0.246	0.340	0.493	0.954	1.362	2.461	3.314	4.371	5.386
15	0.254	0.348	0.498	0.933	1.297	2.208	2.863	3.626	4.318
20	0.258	0.353	0.501	0.922	1.265	2.091	2.661	3.304	3.871
25	0.261	0.355	0.503	0.916	1.246	2.024	2.547	3.126	3.627
30	0.263	0.357	0.504	0.912	1.234	1.980	2.474	3.012	3.473
35	0.264	0.359	0.505	0.909	1.225	1.950	2.423	2.934	3.368
40	0.265	0.360	0.506	0.907	1.219	1.927	2.385	2.877	3.291
50	0.266	0.361	0.507	0.903	1.210	1.895	2.333	2.798	3.186
60	0.267	0.362	0.508	0.901	1.204	1.875	2.299	2.747	3.119
70	0.268	0.363	0.508	0.900	1.199	1.860	2.275	2.711	3.071
80	0.269	0.363	0.509	0.899	1.196	1.849	2.257	2.685	3.036
90	0.269	0.364	0.509	0.898	1.194	1.841	2.244	2.664	3.009
100	0.269	0.364	0.509	0.897	1.192	1.834	2.233	2.648	2.988

m= 7

m' \ P:	0.050	0.100	0.200	0.500	0.683	0.900	0.954	0.980	0.990
5	0.252	0.347	0.507	1.041	1.583	3.368	5.086	7.614	10.46
10	0.275	0.370	0.521	0.971	1.363	2.414	3.228	4.235	5.200
15	0.285	0.380	0.528	0.949	1.296	2.158	2.775	3.492	4.142
20	0.290	0.385	0.532	0.938	1.263	2.040	2.572	3.171	3.699
25	0.294	0.389	0.535	0.931	1.244	1.971	2.457	2.993	3.457
30	0.296	0.391	0.536	0.927	1.231	1.927	2.384	2.880	3.304
35	0.298	0.393	0.538	0.924	1.222	1.896	2.332	2.802	3.200
40	0.299	0.394	0.539	0.922	1.215	1.873	2.294	2.745	3.124
50	0.301	0.396	0.540	0.919	1.206	1.840	2.242	2.667	3.020
60	0.303	0.398	0.541	0.917	1.200	1.819	2.208	2.616	2.953
70	0.304	0.399	0.542	0.915	1.195	1.804	2.184	2.580	2.906
80	0.304	0.399	0.542	0.914	1.192	1.793	2.166	2.553	2.871
90	0.305	0.400	0.543	0.913	1.189	1.785	2.152	2.533	2.845
100	0.305	0.400	0.543	0.913	1.187	1.778	2.141	2.517	2.823

m= 8

m' \ P	0.050	0.100	0.200	0.500	0.683	0.900	0.954	0.980	0.990
5	0.271	0.367	0.526	1.055	1.587	3.339	5.025	7.503	10.29
10	0.299	0.394	0.544	0.983	1.363	2.377	3.161	4.129	5.057
15	0.311	0.406	0.552	0.960	1.295	2.119	2.706	3.387	4.004
20	0.317	0.412	0.557	0.950	1.261	1.999	2.502	3.067	3.564
25	0.322	0.417	0.560	0.943	1.241	1.929	2.387	2.890	3.324
30	0.325	0.420	0.562	0.939	1.228	1.884	2.312	2.777	3.173
35	0.327	0.422	0.564	0.936	1.219	1.852	2.260	2.699	3.069
40	0.329	0.423	0.565	0.934	1.212	1.829	2.222	2.641	2.993
50	0.331	0.426	0.567	0.930	1.202	1.796	2.170	2.563	2.890
60	0.333	0.428	0.568	0.928	1.196	1.775	2.135	2.512	2.823
70	0.334	0.429	0.569	0.927	1.191	1.760	2.111	2.476	2.777
80	0.335	0.430	0.569	0.926	1.188	1.748	2.093	2.450	2.742
90	0.336	0.430	0.570	0.925	1.185	1.739	2.079	2.429	2.715
100	0.336	0.431	0.570	0.924	1.183	1.732	2.068	2.413	2.694

m= 9

m' \ P	0.050	0.100	0.200	0.500	0.683	0.900	0.954	0.980	0.990
5	0.287	0.383	0.542	1.065	1.590	3.316	4.976	7.415	10.16
10	0.319	0.414	0.562	0.992	1.363	2.347	3.107	4.044	4.942
15	0.333	0.427	0.572	0.970	1.293	2.086	2.650	3.303	3.895
20	0.341	0.435	0.577	0.959	1.259	1.965	2.445	2.984	3.457
25	0.346	0.440	0.581	0.952	1.239	1.895	2.329	2.806	3.217
30	0.349	0.444	0.583	0.948	1.225	1.849	2.254	2.693	3.067
35	0.352	0.446	0.585	0.945	1.216	1.817	2.202	2.615	2.963
40	0.354	0.448	0.587	0.943	1.208	1.793	2.164	2.558	2.888
50	0.357	0.451	0.589	0.940	1.198	1.760	2.111	2.479	2.785
60	0.359	0.453	0.590	0.937	1.192	1.738	2.076	2.428	2.718
70	0.360	0.454	0.591	0.936	1.187	1.723	2.051	2.392	2.672
80	0.361	0.455	0.592	0.935	1.183	1.711	2.033	2.366	2.637
90	0.362	0.456	0.593	0.934	1.181	1.702	2.019	2.345	2.611
100	0.363	0.457	0.593	0.933	1.178	1.695	2.008	2.329	2.590

m=10

m' \ P	0.050	0.100	0.200	0.500	0.683	0.900	0.954	0.980	0.990
5	0.301	0.397	0.555	1.073	1.592	3.297	4.936	7.344	10.05
10	0.336	0.431	0.578	1.000	1.363	2.323	3.062	3.975	4.849
15	0.351	0.446	0.588	0.977	1.291	2.059	2.604	3.235	3.805
20	0.360	0.454	0.594	0.966	1.257	1.937	2.398	2.915	3.368
25	0.366	0.460	0.599	0.960	1.236	1.866	2.281	2.737	3.129
30	0.370	0.464	0.601	0.955	1.222	1.819	2.206	2.624	2.979
35	0.373	0.467	0.603	0.952	1.212	1.787	2.154	2.546	2.876
40	0.376	0.469	0.605	0.950	1.205	1.763	2.115	2.488	2.801
50	0.379	0.472	0.607	0.947	1.195	1.729	2.061	2.410	2.698
60	0.382	0.475	0.609	0.945	1.188	1.707	2.026	2.359	2.632
70	0.383	0.476	0.610	0.943	1.183	1.691	2.002	2.323	2.585
80	0.385	0.477	0.611	0.942	1.180	1.680	1.983	2.296	2.551
90	0.386	0.478	0.612	0.941	1.177	1.670	1.969	2.275	2.524
100	0.386	0.479	0.613	0.940	1.174	1.663	1.958	2.259	2.503

m=12

m' \ P	0.050	0.100	0.200	0.500	0.683	0.900	0.954	0.980	0.990
5	0.322	0.418	0.575	1.085	1.596	3.268	4.874	7.235	9.888
10	0.363	0.457	0.601	1.012	1.361	2.284	2.994	3.868	4.706
15	0.382	0.475	0.614	0.989	1.288	2.017	2.532	3.128	3.666
20	0.393	0.486	0.622	0.977	1.253	1.892	2.325	2.808	3.231
25	0.400	0.492	0.627	0.971	1.231	1.820	2.207	2.629	2.993
30	0.405	0.497	0.630	0.966	1.217	1.773	2.130	2.516	2.843
35	0.409	0.501	0.633	0.963	1.207	1.739	2.077	2.437	2.740
40	0.412	0.503	0.635	0.961	1.199	1.715	2.038	2.380	2.665
50	0.416	0.508	0.638	0.958	1.189	1.680	1.984	2.301	2.562
60	0.419	0.510	0.640	0.956	1.181	1.657	1.948	2.249	2.496
70	0.422	0.512	0.641	0.954	1.176	1.641	1.923	2.213	2.450
80	0.423	0.514	0.642	0.953	1.173	1.629	1.904	2.186	2.415
90	0.425	0.515	0.643	0.952	1.170	1.620	1.890	2.165	2.389
100	0.426	0.516	0.644	0.951	1.167	1.612	1.878	2.149	2.368

m=16

m' \ P	0.050	0.100	0.200	0.500	0.683	0.900	0.954	0.980	0.990
5	0.351	0.446	0.600	1.101	1.599	3.230	4.795	7.095	9.680
10	0.401	0.493	0.633	1.026	1.359	2.233	2.904	3.730	4.520
15	0.425	0.515	0.649	1.003	1.283	1.961	2.438	2.988	3.485
20	0.439	0.529	0.658	0.992	1.246	1.833	2.227	2.666	3.051
25	0.449	0.538	0.665	0.985	1.223	1.758	2.107	2.487	2.813
30	0.456	0.544	0.669	0.980	1.208	1.709	2.029	2.372	2.663
35	0.461	0.549	0.673	0.977	1.198	1.674	1.974	2.293	2.560
40	0.465	0.552	0.675	0.975	1.190	1.649	1.934	2.234	2.484
50	0.471	0.558	0.679	0.972	1.178	1.613	1.878	2.154	2.382
60	0.475	0.561	0.682	0.969	1.171	1.589	1.842	2.102	2.315
70	0.478	0.564	0.684	0.968	1.165	1.572	1.816	2.065	2.268
80	0.480	0.566	0.685	0.967	1.161	1.559	1.796	2.038	2.233
90	0.482	0.568	0.687	0.966	1.158	1.550	1.781	2.017	2.206
100	0.483	0.569	0.688	0.965	1.156	1.542	1.769	2.000	2.185

m=20

m' \ P	0.050	0.100	0.200	0.500	0.683	0.900	0.954	0.980	0.990
5	0.369	0.463	0.617	1.111	1.601	3.207	4.747	7.009	9.553
10	0.426	0.516	0.653	1.035	1.357	2.201	2.847	3.644	4.405
15	0.454	0.542	0.671	1.011	1.279	1.924	2.378	2.900	3.372
20	0.471	0.557	0.682	1.000	1.241	1.794	2.164	2.577	2.938
25	0.482	0.568	0.690	0.993	1.218	1.718	2.042	2.396	2.699
30	0.490	0.575	0.695	0.989	1.202	1.667	1.963	2.281	2.549
35	0.497	0.581	0.699	0.986	1.191	1.632	1.908	2.200	2.445
40	0.502	0.585	0.702	0.983	1.183	1.605	1.867	2.141	2.369
50	0.509	0.592	0.707	0.980	1.171	1.568	1.810	2.060	2.265
60	0.514	0.596	0.710	0.978	1.163	1.543	1.772	2.007	2.198
70	0.518	0.600	0.713	0.976	1.157	1.526	1.745	1.969	2.150
80	0.520	0.602	0.715	0.975	1.153	1.513	1.725	1.941	2.115
90	0.523	0.604	0.716	0.974	1.149	1.503	1.710	1.920	2.088
100	0.525	0.606	0.717	0.973	1.147	1.494	1.698	1.902	2.067