

4 Počet pravděpodobnosti jako základ statistického usuzování

K analýze dat přistupujeme z několika hledisek. Dosud jsme probrali explorační popisnou analýzu dat, jimiž dokážeme přehledně shrnout informace, jež se týkají právě těch objektů, které jsme pozorovali nebo změřili. Jestliže jsme však data získali na základě dobře navrženého výzkumného plánu, můžeme provádět zobecňující úsudky o chování sledovaných proměnných a jejich parametrech v celé uvažované populaci. Metody takového statistického usuzování se opírají o počet pravděpodobnosti. Proto jsou základy počtu pravděpodobnosti tématem této kapitoly.

Metody statistického testování a odhadovaní vyžadují data získaná náhodným výběrem nebo metodou znáhodněného experimentu. Statistické usuzování spočívá na kladení otázek typu: „Jak často tato metoda dá správnou odpověď, pokud ji použijí mnohokrát?“ Pokud si nemůžeme představit, že proces sběru dat lze opakovat (např. tím, že z populace vybereme jiný výběr), statistické usuzování nemá smysl. Jestliže však využijeme při získání dat náhodu, můžeme použít teorii pravděpodobnosti pro zodpovězení otázky: „Jak často nastane určitý jev, pokud experiment nebo výběr provedeme mnohokrát?“

Látka probíraná v této kapitole je klíčová pro porozumění většině postupů, jež uvedeme v dalších částech knihy. Nesmírně důležitý je pojem náhodné proměnné a rozdělení náhodné proměnné. V závěru této kapitoly se dostaneme k další problematice, která je pro statistickou analýzu rozhodující – k pravděpodobnostnímu rozvádění statistik vypočítaných z dat, a uvedeme základní teoretická pravděpodobnostní rozdělení, jež jsou vhodná pro popis variability statistik. Tato část tvoří základ pro rozvinutí principů teorie statistického usuzování, kterými se budeme zabývat v příští kapitole.

Aplikace počtu pravděpodobnosti a příslušné teorie pronikly do četných vědních oborů i oblastí praktické činnosti. Historici dovozují, že vývoj počtu pravděpodobnosti neprobíhal nijak jednoduše. Jak Řekové, tak první křesťané neměli důvod se zabývat kvantifikací náhody. Řekové si vymáhali působení náhody, ale věřili, že není správné matematicky spojovat to, co se stalo,

a to, co by se mělo stát, protože by šlo o překrývání „pozemského plánu“ a „nebeského plánu“. Navíc u Řeků hrál roli jejich antiempiricismus. Znalost se nemohla získat experimentováním, ale pouze logickou cestou. Tyto dva momenty jim bránily zabývat se problémem náhody v souvislosti s predikcí jevů. Pro první křesťany zase něco jako náhoda nemohlo vůbec existovat. Každá událost byla přímým svědectvím božského působení.

Nejdříve se prvky teorie pravděpodobnosti uplatňovaly při výpočtech sázek na hazardních hrách. Mezi první, kdo se zabýval pravděpodobnostními problémy, patřil Blaise Pascal (1623–1662). Hazardní hry v době, o níž mluvíme, měly za sebou historii dlouhou nejméně dva tisíce let, protože již Řekové a Římané byli vášnivými hráči. NejpopulárnějšíhraPascalovy doby se nazývala „hazard“, jejíž pojmenování pochází z arabského *al zhar*, což znamená „kostka“. Pascal si dal za úkol zodpovědět řadu otázek, které mu položil jeho přítel António Gombard rytíř de Mere. Například – proč je výhodné vsadit na čtyři šestky při čtyřech hodech a proč není výhodné vsadit při dvojnásobném hodu na dvě šestky v 24 pokusech? Vedl o tomto problému korespondenci s jiným významným vědcem té doby Pierre de Fermatem.

Nejstarší knihou o pravděpodobnosti bylo dílo holandského matematika Christiaena Huygense *De Ratiociniis in Ludo Aleae* (Výpočty v hrách náhody) z roku 1657. Po padesát letech sloužil jako standardní učební text o pravděpodobnosti. Právě jejho autora považují mnozí historici za zakladatele teorie pravděpodobnosti. Pierre Simon Laplace (1749–1827), autor přehledného pojednání o teorii pravděpodobnosti, prohlásil: „Je obdivuhodné, že počtu pravděpodobnosti, jenž vznikl při úvahách o hazardních hrách, bylo určeno stát se nejdůležitější složkou lidského vědění.“

Rozvoj teorie pravděpodobnosti si vyžádal kromě přemýšlení fyziků i využití mnoha pokusů statistika K. Pearsona, který v roce 1900 hodil 24 000krát mincí, aby se přesvědčil, zda relativní četnost jevu, že padne „orel“, konverguje k číslu 0,5. Jeho pokus vedl k četnosti „orla“ 12012.

4.1 Základní pojmy a výpočty

Vysvětlíme stručně pouze základy počtu pravděpodobnosti, které budeme potřebovat v této knize. Musíme si přitom uvědomit, že matematická teorie pravděpodobnosti nemůže objasnit podstatu náhodnosti a pravděpodobnosti. Je použitelným formálním popisem situací, v nichž se náhodnost, resp. nejistota projevuje; umožňuje o nich uvažovat.

4.1.1 Náhodné jevy, pravděpodobnost

Náhodnost vede k tomu, že jevy, které nás zajímají, se za daných podmínek mohou nebo nemusí vyskytnout. Například při házení mincí sledujeme jev, padne „orel“. V daném hodu můžeme predikovat jeho výsledek pouze vyjádřením pravděpodobnosti možností, jež mohou nastat. Pravděpodobnost, že padne „orel“, vyjadřujeme číslem, které má určitý význam. Slova pravděpodobný nebo nepravděpodobný, jež se vyskytují v běžné řeči, vyjadřují nejistotu kvalitativní. Vztah tohoto vyjádření k matematickému pojmu pravděpodobnost je dán kontextem.

Určitý fenomén považujeme za náhodný, jestliže jeho výskyt je nejistý, ale zároveň pozorujeme v dlouhé řadě situací určitou pravidelnost v rozdělení jeho výskytu.

PŘÍKLAD 4.1

Použití teorie pravděpodobnosti pro modelování jevů

Počet pravděpodobnosti lze využít pro modelování nejrůznějších situací. Uvedeme jednoduchý modelový příklad z oblasti sportu, který lze řešit pomocí pravděpodobnostního počtu. Jeden z přístupů k řešení popíšeme na konci tohoto odstavce (s. 122).

Jana má ve svém repertoáru dva druhy tenisového podání, tvrdý a měkký servis. Její tvrdý servis je v poli v 50 % podání a v 75 % pak uhraje míč. Měkký servis Jana nezkaží v 75 %, ale míč pak uhraje jenom v 50 %. Jakou má Jana zvolit strategii při svém podání, pokud lze předpokládat, že během utkání se tyto charakteristiky nezmění? Má hrát obě podání tvrdě nebo měkce? Nebo má začít tvrdým podáním a po chybě podávat měkce? Je snad pro ni lepší zahrát první podání měkce a druhé podání tvrdě? Jak by se měla rozhodovat, aby v průměru dosahovala nejlepších výsledků, jestliže předpokládáme, že uvedené relativní četnosti jsou platné bez ohledu na průběh utkání?

Existuje mnoho různých definic pravděpodobnosti: definice axiomatická; definice pravděpodobnosti jako kvantitativní míry jistoty; klasická definice, jež pojem pravděpodobnosti převádí na pojem stejné možnosti. Uvedeme **statistickou definici pravděpodobnosti**.

Mluvíme o **náhodném pokusu**, jestliže při pokusu lze dostat různé možné výsledky a přitom:

- | nelze předem určit, který z těchto výsledků získáme;
 - | pokus lze libovolně často opakovat, aniž se jednotlivá opakování vzájemně ovlivňují.

Množina všech možných výsledků náhodného pokusu tvoří prostor náhodných výsledků (E). Například při hodu mincí tvoří prostor jevů v jednom hodu „panna“ nebo „orel“.

Vymezená množina výsledků je **náhodný jev**. Všechny možné náhodné jevy tvoří **pole jevů**. Jev, jenž se skládá pouze z jednoho výsledku, se nazývá **elementární jev**. Jev, který nastává, jestliže dostaneme více možných výsledků, se nazývá **jev složený**.

Jev je výsledek náhodného pokusu. Pro pole náhodných jevů lze použít vztahy teorie množin.

Nymborem $A \cup B$ označujeme jev, že nastane jev A nebo nastane jev B nebo nastanou oba dva. Současný výskyt jevu A a jevu B označujeme symbolem $A \cap B$. Případ, že $A \cap B$ je prázdná množina, znamená vzájemně se vylučující jevy.

PŘEHLED STATISTICKÝCH METOD

Také říkáme, že tyto jevy jsou disjunktní. Jev doplňkový \bar{A} (nebo také opačný) k jevu A je jev, který nastane, když nenastane v pokusu jev A .

Pravděpodobnost náhodného jevu A je číslo $P(A)$, k němuž se blíží relativní četnost jevu A , jestliže pokus dostatečně často opakujeme. Jestliže jsme provedli n pokusů a v m z nich nastal jev A , pak názorně vyjádřena:

$$\lim_{n \rightarrow \infty} \frac{m}{n} = P(A)$$

Tuto hodnotu pravděpodobnosti považujeme v teorii pravděpodobnosti za danou. Výrok „Pravděpodobnost jevu A je rovna hodnotě p “ znamená, že $P(A) = p$.

Pravděpodobnost náhodného jevu je tedy číslo mezi 0 a 1, které popisuje relativní četnost, s jakou se jev vyskytne ve velmi dlouhé řadě opakování situace, kdy tento jev může nastat. Pravděpodobnosti popisují pouze to, co se stane v dlouhé řadě pokusů. Krátké série náhodných jevů, jako házení mincí nebo střelba na koš, často nevypadají náhodně, protože neukazují pravidelnost, jež se ve skutečnosti může prosadit jenom při mnoha opakováních.

Pravděpodobnost má tyto základní vlastnosti:

1. Pravděpodobnost jevu, který je jistý, se rovná 1.
2. Pravděpodobnost jevu nemožného je rovna 0.
3. Lze-li náhodný jev rozložit na několik vzájemně se vylučujících (disjunktivních) jevů, pak se jeho pravděpodobnost rovná součtu pravděpodobností těchto jevů.

Pro výpočet pravděpodobnosti jevu A často používáme pravidlo, které je východiskem definice pravděpodobnosti na základě stejné možnosti: Jestliže náhodný pokus může vést k r různým elementárním jevům, jež jsou stejně pravděpodobní, pak pravděpodobnost jevu A je

$$P(A) = \frac{\text{počet elementárních jevů, které vedou k } A}{r}.$$

PŘÍKLAD 4.2

Elementární a složený náhodný jev a jejich pravděpodobnosti

Při házení kostkou platí, že prostor náhodných výsledků E je $(1; 2; 3; 4; 5; 6)$. Příkladem elementárního jevu je jev, že padne číslo 5. Počet všech elementárních jevů je 6. Jev A , že padne sudé číslo, je jevem složeným – tvoří jej 3 elementární jevy. Proto je pravděpodobnost padnutí sudého čísla $P(A) = 3/6 = 1/2$. Relativní četnost tohoto jevu se tedy se vztahuje k počtem hodů blíží k číslu 0,5.

Často používáme pravděpodobnosti spojení a průniku jevů A a B nebo pravděpodobnost doplňku jevu. Pravidlo 3 lze napsat obecněji pomocí rovnice

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Rozšíření tohoto pravidla na tři jevy má tvar

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C).$$

Ze tří základních vlastností pravděpodobnosti plynou již všechny vlastnosti další. Uvedeme ty nejvýznamnější:

1. Pro libovolný jev A platí: $0 \leq P(A) \leq 1$.
2. Je-li jev \bar{A} doplňkový k jevu A , pak $P(\bar{A}) = 1 - P(A)$.
3. Je-li jev A částí jevu B , pak $P(A) \leq P(B)$.

PŘÍKLAD 4.3

Výpočet pravděpodobnosti různých jevů

Statistické šetření ukázalo u 1000 dotázaných občanů volební preference, které uvádí tabuľka 4.1. Jestliže z této skupiny náhodně vybereme jedince, jaká bude pravděpodobnost jevu:

- a) bude se jednat o ženu, která nepreferuje ODS;
- b) osoba bude ženského pohlaví nebo chce volit „ostatní“.

Obě úlohy nejsou složité, ale musíme si promyslet přesně obsah otázky.

V první úloze je odpověď dána zlomkem $(530 - 220)/1000$.

Abychom vyřešili druhou otázkou, musíme si uvědomit, že se jedná o výpočet pravděpodobnosti sjednocení jevu (žena) \cup (ostatní). Z toho plyne

$$\begin{aligned} P((žena) \cup (ostatní)) &= P(žena) + P(ostatní) - P((žena) \cap (ostatní)) \\ &= 530/1000 + 303/1000 - 157/1000. \end{aligned}$$

Tab. 4.1

Modelová data – výsledky průzkumu volebních preferencí

Preferovaná politická strana	Ženy	Muži	Celkem
ČSSD	153	130	283
ODS	220	194	414
Ostatní	157	146	303
Celkem	530	470	1000

4.1.2 Podmíněná pravděpodobnost, Bayesova formule

Často závisí pravděpodobnost výskytu určitého jevu na tom, zda nastal či nenastal nějaký jiný jev. Takovým pravděpodobnostem říkáme podmíněné a značíme je $P(A|B)$, což čteme: pravděpodobnost jevu A za předpokladu, že nastal jev B . Pro podmíněné pravděpodobnosti lze dokázat všechna základní pravidla, která jsme uváděli u nepodmíněné pravděpodobnosti, tedy zejména $0 \leq P(A|B) \leq 1$. Platí dvě ekvivalentní rovnice, pokud $P(B)$ nemá nulovou hodnotu:

$$P(A \cap B) = P(A|B)P(B) \quad \text{nebo} \quad P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Rovná-li se podmíněná pravděpodobnost pravděpodobnosti nepodmíněné, tedy když $P(A|B) = P(A)$, říkáme, že jevy A a B jsou **statisticky nezávislé**. Výskyt jevu B nemá v tomto případě vliv na pravděpodobnost výskytu jevu A v dané situaci. Jsou-li jevy A a B statisticky nezávislé, pak $P(A \cap B) = P(A) \cdot P(B)$. Platí: Jsou-li jevy A, B statisticky nezávislé, pak jsou statisticky nezávislé i dvojice jevů \bar{A}, B ; A, \bar{B} ; \bar{A}, \bar{B} .

Jestliže jev B nastává vždy s některým jevem A_1, \dots, A_n , přičnž A_i jsou jen disjunktní, pak

$$P(B) = \sum_{i=1}^n P(A_i)P(B|A_i).$$

Tento vztah nazýváme vzorec pro úplnou pravděpodobnost. Vzorec je často používá **Bavesova formule**. S

Při pravděpodobnostních úvahách se často používá Bayesova formule. Při vypočítání podmíněné pravděpodobnosti $P(A|B)$ za předpokladu, že známe pravděpodobnosti $P(B|A)$ a $P(A)$. Pomáhá nám např. při výpočtech, které provádíme při hodnocení diagnostických testů binárního typu v medicínské a psychologické diagnostice. Uvedeme její jednoduchou podobu:

$$P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A) + P(\bar{A})P(B|\bar{A})}$$

V čitateli této formule je pravděpodobnost, že současně nastane jev A a jev B , ve imenovateli je vzorec pro úplnou pravděpodobnost jevu B .

PŘÍKLAD 4.4

Aplikace Bayesova přístupu pro studium vlastností diagnostických testů

Významné použití nachází Bayesova formule při hodnocení diagnostických testů, jež mohou nabývat pouze dvou hodnot (pozitivní, negativní). Takové hodnocení se provádí i u uměle chotomizovaných kvantitativních testových výsledků (např. normální výsledek, výsledek normální mezí testu). Popíšeme stručně tuto situaci. Pacient muže, nebo nemusí mít danou chorobu (D_+ , D_-). Provedený diagnostický test muže, nebo nemusí tuto chorobu indikovat.

($T+$, $T-$). Záleží to na jeho specifitě a senzitivitě. **Senzitivita diagnostického testu** Se je podmíněná pravděpodobnost $P(T+|D+)$ toho, že výsledek testu bude pozitivní, když pacient má chorobu. **Specificita diagnostického testu** Sp je podmíněná pravděpodobnost $P(T-|D-)$, že za předpokladu, že pacient nemá danou chorobu, test bude negativní. **Prediktivní hodnota pozitivního testu** $P+$ je podmíněná pravděpodobnost $P(D+|T+)$, že pacient má chorobu, pokud byl test pozitivní. **Prediktivní hodnota negativního testu** $P-$ je podmíněná pravděpodobnost $P(D-|T-)$, že pacient nemá danou chorobu, když test byl negativní. **Prevalence** $P(D+)$ je pravděpodobnost choroby v populaci. Uvedené pravděpodobnosti se odhadují pomocí statistické evidence výsledků v medicínských databázích a zvlášť zaměřeného výzkumu diagnostické věrohodnosti diagnostického testu. Podle výsledků testu se sestavuje čtyřpolní tabulka s četnostmi (tabulka 4.2). Například četnost a je počet výsledků nemocných jedinců, kteří měli pozitivní test. Pro odhad uvedených charakteristik se četnosti tabulkou použijí takto:

$$Se = P(T+|D+) = a/(a+b)$$

$$Sp = P(T-|D-) = d/(c+d)$$

$$P_+ = P(D_+|T_+) = a/(a+c)$$

$$P_{-} = P(D_{-}|T_{-}) = b/(b+d)$$

To však lze provést pouze v případě, že získáváme výsledky pro jedince vybraného zcela náhodným způsobem. Častější je případ, kdy jsou k dispozici předem dané skupiny jedinců s diagnózou nebo bez ní a provedeme u obou skupin posuzovaný test. Odhad senzitivity a specificity je pořádku, ale odhad pravděpodobnosti $P+$ a $P-$ pomocí četností z tabulky je zkreslený. Musíme nejdříve získat informaci o výskytu uvažované nemoci v populaci. Proto najdeme prevalenci $P(D+)$ u různých subpopulací. Podle Bayesovy formule následně určíme prediktivní hodnotu pozitivního testu

$$P+ = \frac{SeP(D+)}{SeP(D+) + (1 - Sp)(1 - P(D+))}$$

ratio prediktivní hodnotu negativního testu

$$P- = \frac{Sp(1 - P(D+))}{Sp(1 - P(D+)) + (1 - Se)P(D+)}$$

Vy výzorech použijeme prevalenci $P(D+)$ podle toho, z které subpopulace jedinec pochází.

Tab. 4.2 Čtyřpolní tabulka s četnostmi

Skutečná diagnóza	Výsledek testu	
	T+	T-
D+	a	b
D-	c	d

4.1.3 Šance

Často používáme výraz, že šance vítězství fotbalového mužstva v daném zápase je 1 : 4 nebo 2 : 1. V prvním případě považujeme vítězství našeho klubu za málo pravděpodobné, ve druhém případě se domníváme, že pravděpodobnost vítězství je dvojnásobně větší než pravděpodobnost prohry $P(P)$. Tedy šance na vítězství mého klubu se rovná $P(V)/P(P) = 2 : 1$. Protože vítězství V a prohra P jsou vzájemně se vylučující jevy, můžeme šanci na vítězství zapsat takto:

$$\text{šance na vítězství} = \frac{P(V)}{1 - P(V)}$$

Formálně je šance ve prospěch nějakého jevu A definována poměrem

$$\text{šance ve prospěch } A = \frac{P(A)}{P(\bar{A})} = \frac{P(A)}{1 - P(A)},$$

kde $P(A)$ je pravděpodobnost jevu A.

Jestliže počítáme pravděpodobnosti pomocí vypočítaných relativních četností, pak platí, že jmenovatele ve zlomcích pro relativní četnost se vyruší. Proto lze vypočítat šanci pouze pomocí četností jevů A a \bar{A} , které se realizovaly v daném sledování:

$$\text{šance ve prospěch } A = \frac{\text{počet výskytů jevu } A}{\text{počet případů, kdy jev } A \text{ nenastal}}.$$

Jestliže známe šanci ve prospěch A, lze samozřejmě zpětně vypočítat pravděpodobnost jevu A v dané situaci.

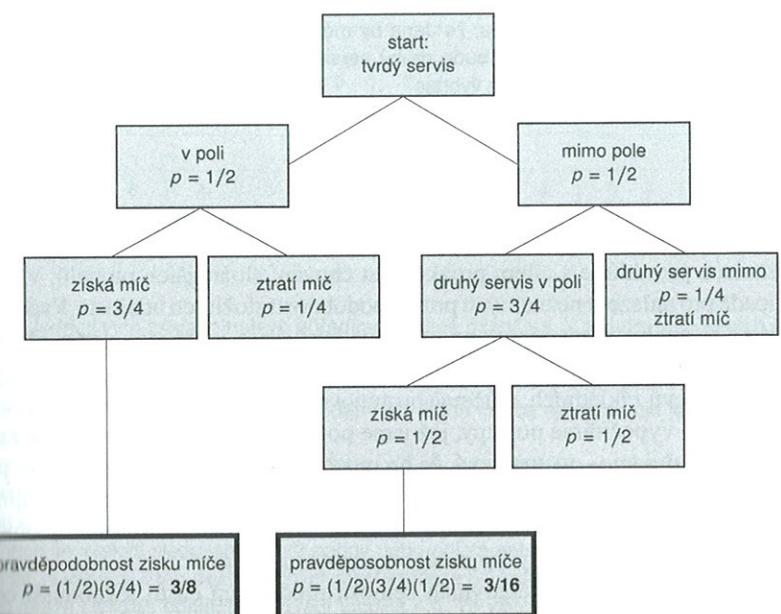
Řešení příkladu 4.1 ze stránky 117

Naším úkolem je navrhnut pro tenistku Janu, jak má využívat své podání. Hledané řešení by mělo vést v průměru k největšímu počtu vyhraných míčů. Tvrď servis se jí podaří v 50 %, míč pak vyhraje v 75 %. Měkký servis se jí podaří sice v 75 %, ale pak ho vyhraje pouze v 50 %. Jaké má zvolit pořadí obou typů servisů, aby dosáhla v průměru nejlepšího výsledku?

Úkol lze řešit sestrojením stromového grafu pro znázornění různých možností průběhu hry. Pomocí něj vypočítáme pravděpodobnosti výhry, které pak srovnáme.

Uvažujeme jako první možnost, že Jana začne tvrdým servisem a pokud se jí nepodaří servírovat měkkým servisem. První servis může nebo nemusí být v poli s pravděpodobností 1/2, tedy první servis je v poli, pravděpodobnost výhry míče je 3/4 a ztráty je 1/4. Jestliže první míč

Obr. 4.1 Stromový graf jedné varianty strategie podání (první podání tvrdé, druhé měkké)



Tab. 4.3 Příklad pravděpodobnostního modelu z oblasti sportu

První podání	Druhé podání	Pravděpodobnost výhry
tvrdé	měkké	9/16 = 0,563
tvrdé	tvrdé	9/16 = 0,563
měkké	tvrdé	15/32 = 0,469
měkké	měkké	15/32 = 0,469

Jana servíruje podruhé měkkce, je pravděpodobnost podařeného servisu 3/4 a druhá pravděpodobnost výhry míče se rovná 1/2. Jestliže i druhý servis je v autu, Jana má výhru. Celková pravděpodobnost výhry, když Jana začne tvrdým servisem a v případě porážení pokračuje měkkým servisem, se spočte jako součet pravděpodobností výhry ze dvou nezávislých větví grafu, jež vedou k výhře $3/8 + 3/16 = 9/16$. Poznamenejme,

PŘEHLED STATISTICKÝCH METOD

že ostatní pravděpodobnosti zisku míče v grafu jsou podmíněné pravděpodobností, kde podmínka je daná předpokladem, že nastaly jevy, které uvažovanému ziskovému uzlu předcházely.

Podobný způsobem odvodíme pravděpodobnosti výhry pro další tři možnosti. Výsledky znázorníme tabulkou 4.3. Vidíme, že Jana by měla nejdříve podávat tvrdě. Jestliže se jí podání nepodaří, je jedno, zda bude druhý servis podávat měkce nebo tvrdě. Můžeme jenom doufat, že ho dá do pole a vyhraje.

4.1.4 Využití simulace pro odhad pravděpodobnosti

Simulace provádíme s cílem prozkoumat chování složitějších modelů, v našem případě pro nalezení neznámých pravděpodobností složitých událostí. Vycházíme přitom ze znalosti pravděpodobností základních jevů, které jsou pro celý proces směrodatné. Pokud určíme tyto pravděpodobnosti a mechanismus, jak nové jevy vznikají z jevů základních, můžeme postupovat dvěma cestami. Buď nové pravděpodobnosti vypočítáme postupy, jež jsme popsali v předchozím příkladu, anebo proces simulujeme – to znamená, že ho mnohokrát opakujeme a jednotlivé pravděpodobnosti odhadujeme pomocí relativních četností jevů, které nás zajímají. Tento metodám se také říká metody Monte Carlo. Při modelování nám obvykle pomáhá počítač, ale někdy lze model realizovat manuálně.

Simulace a počítání relativních četností představují často jednodušší cestu, jak odhadnout neznámé pravděpodobnosti. Naše simulace povede k validním výsledkům, pokud jsme model dobrě sestavili. Simulace v kontextu statistiky lze definovat jako využívání tabulek náhodných čísel nebo generování náhodných čísel počítačem s cílem napodobit reálné procesy.

PŘÍKLAD 4.6

Odhad pravděpodobnosti pomocí simulace

Házíme minci a chceme zjistit pravděpodobnost, že v deseti hodech třikrát za sebou padne orel nebo třikrát za sebou padne panna. Popíšeme kroky sestavování modelu a realizaci simulace pomocí tabulky náhodných čísel z přílohy B.

Krok 1. Zadání pravděpodobností. Náš model má dvě části:

- v každém hodu je pravděpodobnost „orla“ 0,5;
- jednotlivé hody jsou na sobě nezávislými pokusy – to znamená, že výsledek jednoho hodu neovlivňuje pravděpodobnosti výsledku jiného hodu.

Tab. 4.4 Příklad využití simulace pro odhad pravděpodobnosti

1	9	2	2	3	9	5	0	3	4
P	P	O	O	P	P	P	O	P	O

0	5	7	5	6	2	8	7	1	3
O	P	P	P	O	O	O	P	P	P

9	6	4	0	9	1	2	5	3	1
P	O	O	O	P	P	O	P	P	P

Krok 2. Jednotlivým základním jevů přiřadíme číselné označení. V tabulce náhodných čísel (tab. I přílohy B) budou jednotlivé číslice zastupovat výsledek hodu mincí. Každá číslice v tabulce má pravděpodobnost 0,1 a jejich následné uspořádání je v tabulce nezávislé. Jedna číslice znamená jeden hod. Sudá číslice reprezentuje jev „orel“ (O), lichá „panna“ (P).

Krok 3. Simulace mnoha opakování daného pokusu. Deset číslic v tabulce náhodných čísel za sebou představuje jeden pokus (deset hodů). Zaznamenáme mnoho skupin po deseti číslicích a zjistíme v každé z nich jevy „O“ a „P“. Určíme, v kolika případech z nich se vyskytl sledovaný jev (3 panny nebo orli za sebou neboli 3 sudá nebo lichá čísla za sebou). Z tabulky náhodných čísel v příloze B zjistíme např. pro první tři skupiny po deseti číslicích údaje, které uvádí tabulka 4.4. Ve všech těchto třech simulovaných opakování pokusu deseti hodů mincí se realizoval jev tří za sebou stejných hodnot. Jestliže vyhledáme 25 skupin po deseti číslicích, získáme počet realizovaných jevů $m = 25$. Tedy relativní četnost jako odhad hledané pravděpodobnosti má hodnotu $m/n = 20/25 = 0.80$. Pro přesnější odhad musíme opakovat celý pokus mnohem vícekrát pomocí počítače, který umí generovat náhodná čísla. Pak dospejeme k číslu přibližně 0,826.

Příklad ilustroval mnoho pravděpodobnostních problémů z praxe. Ty se vyznačují tím, že se provádějí nezávislé pokusy, v nichž sledovaný jev má stejnou pravděpodobnost. Velmi důležitou roli zde hraje nezávislost jednotlivých pokusů (v našem příkladu nezávislost jednotlivých hodů mincí). Nezávislost pokusů lze ověřit pozorováním mnoha realizací pokusu.

4.2 Náhodná proměnná, rozdělení náhodné proměnné

Předpis, který přiřazuje každému výsledku náhodného pokusu určité číslo, se nazývá **náhodná proměnná**. Náhodná proměnná je tedy funkce, jež zobrazuje prostor výsledků do reálných čísel. Náhodné proměnné budeme značit velkými písmeny, např. A , X , Z , jejich jednotlivé realizace budou čísla (konkrétní výsledky), nebo malými písmeny (obecně).

Je třeba si uvědomit, že výsledkem pokusu nemusí být vždy nějaké číslo; vždy mu však můžeme nějaké číslo přiřadit. V praxi nás toto přiřazení zajímá často méně než pravděpodobnosti, s kterými náhodná proměnná nabývá určité hodnoty nebo je obsažena v určitých intervalech hodnot. Tyto pravděpodobnosti nazýváme **pravděpodobnostní rozdělení** (nebo jenom rozdělení) náhodné proměnné. Je určeno pravděpodobnostmi uvažovaných náhodných jevů z jevového pole.

Na prostoru výsledků lze definovat více náhodných proměnných. Funkce náhodných proměnných jsou zřejmě opět náhodné proměnné.

PŘÍKLAD 4.

Náhodná proměnná a její funkce

Při házení hrací kostkou je výsledkem pokusu jedna ze šesti jejích stran. Odpovídající náhodná proměnná X může nabývat hodnoty 1, 2, 3, 4, 5, 6. Rozdělení náhodné proměnné X je dáno pravděpodobnostmi jednotlivých stran ($1/6$). Definujeme funkci Y tak, že $Y(x) = 0$, jestliže číslo x je sudé, a $Y(x) = 1$, jestliže číslo x je liché. Funkce $Y(x)$ je opět náhodnou proměnnou.

Náhodné proměnné dělíme na diskrétní a spojité (viz kap. 2.2.3). Diskrétní proměnné nabývají navzájem izolované hodnoty – příkladem je počet bodů při hod kostkou. Někdy pozorujeme diskrétní náhodné proměnné, které mohou teoreticky nabývat nekonečně mnoha hodnot – přesněji řečeno tolika, kolik je přirozených čísel. Takovou proměnnou jsou např. počty nehod za rok v dané oblasti.

Spojité náhodné proměnné se v praxi vyskytují velice často. Například všechna běžná měření délky, váhy nebo i času modelujeme jako spojité náhodné proměnné. Do této kategorie náhodných proměnných patří i výsledky mnoha psychologických, znalostních nebo motorických testů. Když klademe důraz na nepřesnosti spojené se zápisem nebo zaokrouhlováním, musíme však v přísném slova smyslu tyto proměnné považovat za diskrétní. Závisí na problému a na cílech, zda danou proměnnou budeme považovat za spojitu, nebo diskrétní.

Tab. 4.5 Příklad rozdělení diskrétní náhodné proměnné – pravděpodobnost výsledku házení kostkou

Uvedli jsme, že pravděpodobnostní rozdělení náhodné proměnné (nebo zkráceně rozdělení náhodné proměnné) jsou pravděpodobnosti, s nimiž nabývá daných číselných hodnot. Diskrétní náhodná proměnná X nabývá hodnot x_1, x_2, \dots, x_m s pravděpodobnostmi p_1, p_2, \dots, p_m , přičemž platí, že součet všech p_i se rovná jedné. Pro naši kostku lze zapsat tyto údaje způsobem, který uvádí tabulka 4.5. Tímto způsobem je popsáno pravděpodobnostní rozdělení diskrétní náhodné proměnné, popisující výsledek hodu kostkou. Obecně se někdy funkci, která přiřazuje diskrétním hodnotám náhodné proměnné příslušné pravděpodobnosti, říká pravděpodobnostní funkce p_x . Také říkáme, že náhodná proměnná se řídí daným zákonem rozdělení.

PŘÍKLAD 4.8

Rozdělení diskrétní náhodné proměnné

Domácnost je skupina lidí, kteří spolu žijí. Jestliže vybereme náhodně domácnost z dané množiny domácností, lze její velikost – počet jejich členů X – považovat za diskrétní náhodnou proměnnou. V tabulce 4.6 uvádíme příklad rozdělení této proměnné (hodnoty větší než 7 jsme zanedbalí). Tyto pravděpodobnosti vyjadřují relativní četnosti velikostí domácností. Udávají pravděpodobnosti, že náhodně vybraná domácnost bude mít určitou velikost. Pravděpodobnost ievu, že domácnost bude mít více než dva členy, má hodnotu

$$\begin{aligned} P(X > 2) &= P(X = 3) + P(X = 4) + P(X = 5) + P(X = 6) + P(X = 7) = \\ &= 0,171 + 0,154 + 0,067 + 0,022 + 0,014 = \\ &= 0,428 \end{aligned}$$

Tab. 4.6 Příklad rozdělení náhodné proměnné - počet osob v domácnosti

Počet osob v domácnosti X	1	2	3	4	5	6	7
Pravděpodobnost p_x	0,251	0,321	0,171	0,154	0,067	0,022	0,014

4.3 Parametry rozdělení náhodné proměnné

Objasníme stručně pojem parametr rozdělení náhodné proměnné.

Pravděpodobnostní chování náhodné proměnné je dokonale popsáno zákonem rozdělení. Někdy však postačuje uvést jako charakteristiku chování náhodné proměnné jenom určité číselné hodnoty, kterým říkáme parametry rozdělení. Poznamenejme, že pro náhodnou proměnnou lze navrhnut mnoho různých parametrů. Obvykle jsou voleny tak, aby z určitého hlediska popsaly její náhodné parametry. Pro zjednodušení uvažujme chování. Zde zavedeme pouze parametry μ , σ a σ^2 . Pro zjednodušení uvažujme případ diskrétní proměnné.

Očekávaná hodnota diskrétní náhodné proměnné X , kterou označujeme $E(X)$ nebo μ , je součet součinů jednotlivých hodnot x_i a příslušných pravděpodobností p_i

$$E(X) = \mu = \sum_{i=1}^m x_i p_i.$$

Pro naš příklad 4.7, kdy se náhodná proměnná X rovná číslu, které padlo při hodu kostkou, snadno vypočítáme, že $E(X) = 3,5$. Takový počet bodů zjevně nelze kostkou realizovat, přesto tato hodnota představuje očekávanou hodnotu v žádném hodu dosažených bodů z mnoha hodů kostky.

Očekávanou hodnotu čtverce odchylek náhodné proměnné od očekávané hodnoty této náhodné proměnné nazýváme **rozptyl**. Rozptyl popisuje stupeň rozptýlenosti hodnot náhodné proměnné od její očekávané hodnoty. Někdy se podle synonyma *variance* označuje $Var(X)$, často též σ^2 , a vypočítá se podle vzorce

$$Var(X) = \sigma^2 = E(X - E(X))^2 = \sum_{i=1}^m (x_i - E(X))^2 p_i = \sum_{i=1}^m (x_i - \mu)^2 p_i.$$

Pro výsledek hodu kostkou X z našeho příkladu 4.7 má $Var(X)$ hodnotu 2,92.

Velice často se používá jiný parametr rozptýlenosti náhodné proměnné, **rodatná odchylka** σ , která se spočte jako druhá odmocnina z rozptylu.

PŘÍKLAD 4.9

Výpočet parametrů rozdělení diskrétní náhodné proměnné

Počet osob žijících v jedné domácnosti je náhodná proměnná. Spočítáme její očekávanou hodnotu pro rozdělení pravděpodobností z příkladu 4.8

$$\mu = 1 \times 0,251 + 2 \times 0,321 + 3 \times 0,171 + 4 \times 0,154 + 5 \times 0,067 + 6 \times 0,022 + 7 \times 0,014 = 2,92$$

V tabulce 4.7 uvádíme výpočet rozptylu této náhodné proměnné.

Tab. 4.7 Příklad postupu výpočtu rozptylu (diskrétní) náhodné proměnné

Hodnota náhodné proměnné x_i	Pravděpodobnost p_i	$p_i(x_i - \mu)^2$	
1	0,251	$0,251(1 - 2,587)^2$	= 0,6322
2	0,321	$0,321(2 - 2,587)^2$	= 0,1106
3	0,171	$0,171(3 - 2,587)^2$	= 0,0292
4	0,154	$0,154(4 - 2,587)^2$	= 0,3075
5	0,067	$0,067(5 - 2,587)^2$	= 0,3901
6	0,022	$0,022(6 - 2,587)^2$	= 0,2563
7	0,014	$0,014(7 - 2,587)^2$	= 0,2728
		$Var(X)$	= 1,995

Parametry rozdělení náhodné proměnné nejsou obvykle známé. Házení kostkou představuje vzácnou výjimku. Na druhé straně je pro posouzení chování náhodné proměnné důležité mít o těchto parametrech informace. V takovém případě o nich usuzujeme pomocí dat, která jsme získali v rámci experimentu nebo jiného výzkumného plánu.

Představme si, že jsme provedli n nezávislých pokusů, přičemž každá hodnota x_i se v těchto pokusech realizovala n_i -krát. Jestliže spočteme aritmetický průměr všech hodnot, pak ho můžeme vyjádřit takto:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \sum_{i=1}^m x_i \frac{n_i}{n} = \sum_{i=1}^m x_i \hat{p}_i$$

Hodnoty \hat{p}_i označují relativní četnosti různých hodnot x_i náhodné proměnné X v sérii n pokusů. Z definice pravděpodobnosti plyne, že relativní četnosti \hat{p}_i se v limitě (s rostoucím n) blíží k pravděpodobnostem p_i . Proto se také vypočtený průměr bude s rostoucím n stále méně lišit od očekávaná hodnoty $E(X)$. Říkáme, že výběrový průměr \bar{x} konverguje v pravděpodobnosti k $E(X)$. Lze nahlédnout, že totéž platí v případě výběrového rozptylu a směrodatné odchylky. Obecně lze vypočítat očekávanou hodnotu $E(g(X))$ jakékoli funkce g náhodné proměnné pomocí vztahu:

$$E(g(X)) \doteq \frac{\sum g(x_i)}{n}$$

Ještě získáme z náhodné proměnné X novou náhodnou proměnnou Y lineární transformací $Y = a + bX$, pak pro očekávanou hodnotu nové proměnné platí

Tab. 4.8 Některé vlastnosti parametrů rozdělení náhodné proměnné

Náhodná proměnná Y	Průměr $E(Y)$	Rozptyl $Var(Y)$
a	a	0
bX	$bE(X)$	$b^2 Var(X)$
$X + a$	$E(X) + a$	$Var(X)$

$E(Y) = a + bE(X)$. Rozptyl náhodné proměnné po lineární transformaci má hodnotu $Var(a + bX) = b^2 Var(X)$.

PŘÍKLAD 4.10

Změna průměru a rozptylu při lineární transformaci

Pokud provádíme transformaci dat $Y = a + bX$, kde a a b mají hodnotu 7 a 5, pak $E(5x + 7) = 5E(x) + 7$. Jestliže je např. průměr $X \mu_x = 3$, pak průměr nové proměnné $5x + 7$ je $E(5x + 7) = 5\mu_x + 7 = 5 \cdot 3 + 7 = 22$. Pro rozptyl bude platit $Var(5x + 7) = 5^2 Var(x)$. Jestliže rozptyl X je $\sigma_x^2 = 20$, pak $Var(5x + 7) = 5^2 Var(x) = 25 \times 20 = 500$.

Uvedeme ještě některé jednoduché případy a výsledky příslušných výpočtů v podobě vzorců (tab. 4.8). Pro dvě náhodné proměnné platí, že očekávaná hodnota z jejich součtu se rovná součtu očekávaných hodnot jednotlivých náhodných proměnných $E(X + Y) = E(X) + E(Y)$.

Další vztah předpokládá uplatnění důležitého konceptu **nezávislosti náhodných proměnných**, kterým vyjadřujeme, že realizace jedné náhodné proměnné neovlivňuje chování druhé náhodné proměnné (např. hodnota měření u jedné osoby neovlivňuje hodnotu měření u druhé osoby). Podrobněji pojem **nezávislosti** vysvětlíme v kapitole o korelační analýze, která bude zaměřena na hodnocení vztahů náhodných proměnných (kap. 7.2.2).

Pro dvě nezávislé náhodné proměnné platí, že rozptyl z jejich součtu, resp. rozdílu se rovná součtu rozptylů jednotlivých náhodných proměnných:

$$Var(X + Y) = Var(X) + Var(Y), \quad Var(X - Y) = Var(X) + Var(Y)$$

Vzorce pro výpočet očekávané hodnoty a rozptylu součtu náhodných proměnných platí také pro konečně mnoho sčítanců. Pomocí tohoto poznatku lze odvodit tvrzení o očekávané hodnotě aritmetickém průměru součtu n stejných náhodných proměnných, jeho rozptylu a směrodatné odchylce průměru:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

$$E(\bar{X}) = E(X)$$

$$Var(\bar{X}) = \frac{Var(X)}{n} \Rightarrow \sigma_{\bar{X}} = \sqrt{\frac{Var(X)}{n}} = \frac{\sigma}{\sqrt{n}}$$

Tyto vztahy se uplatňují v nejrůznějších souvislostech. Poslední z nich vyjadřuje, že směrodatnou odchylku průměru neboli směrodatnou chybu průměru můžeme snadno vypočítat pomocí směrodatné odchylky původní náhodné proměnné.

4.4 Distribuční funkce

Z teoretického hlediska nejúplnejší popis pravděpodobnostního chování diskrétní nebo spojité náhodné proměnné X představuje **distribuční funkce** $F(x)$. Distribuční funkce je pravděpodobnost, že náhodná proměnná X nabude určité hodnoty x nebo hodnoty menší, tedy

$$F(x) = P(X \leq x).$$

Distribuční funkce je definována pro všechna reálná čísla x , má tedy smysl pro $-\infty < x < +\infty$. Uvedeme několik jejích důležitých vlastností:

1. $0 \leq F(x) \leq 1$;
2. když $x \rightarrow -\infty$, pak $F(x) = 0$;
3. když $x \rightarrow +\infty$, pak $F(x) = 1$;
4. $F(x)$ je funkce neklesající, tedy když $x_i < x_j$, pak $F(x_i) \leq F(x_j)$;
5. $F(x)$ nemusí být spojitá.

Jestliže je $F(x)$ spojitá funkce, pak příslušná náhodná proměnná je spojitá.

Pro počítání pravděpodobností platí vzorce:

$$P(x_1 < X \leq x_2) = F(x_2) - F(x_1)$$

$$P(X > x) = 1 - F(x)$$

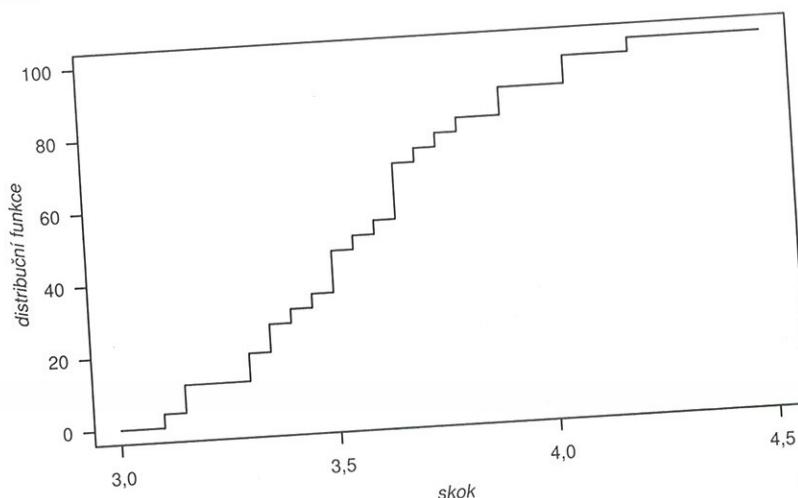
Výběrovým ekvivalentem teoretické distribuční funkce $F(x)$ je výběrová distribuční funkce $\hat{F}(x)$, jež popisuje rozdělení hodnot výběru (obr. 4.2). Empirická distribuční funkce $\hat{F}(x)$ je definována v bodě x relativním počtem měření, která jsou menší nebo rovna x . Tedy

$$\hat{F}(x) = \frac{\text{počet } x_i \leq x}{n}.$$

Jak jsme již uvedli u diskrétní náhodné proměnné, je její chování popsáno pravděpodobnostní funkcí $p(x) = P(X = x)$. Známe-li pravděpodobnostní

PŘEHLED STATISTICKÝCH METOD

Obr. 4.2 Grafické zobrazení empirické distribuční funkce vyjádřené v procentech pro měření skoku do délky pro data ze s. 77



funkci, umíme dopočítat distribuční funkci, a naopak. I pro spojitou náhodnou promennou existuje ekvivalent pravděpodobnostní funkce. Má-li $F(x)$ pro všechna x derivaci, nazýváme tuto derivaci **hustotou pravděpodobnosti** neboli **frekvenční funkcí** $f(x)$ náhodné promenné X . Můžeme ji interpretovat jako přibližnou pravděpodobnost, že hodnota náhodné promenné bude ležet v intervalu jednotkové délky kolem hodnoty x . Hustota pravděpodobnosti $f(x)$ vykazuje podobné vlastnosti jako $p(x) = P(X = x)$ u diskrétních náhodných promenných

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f(x) dx$$

$$\int_{-\infty}^{+\infty} f(x) dx = 1$$

$f(x) \geq 0$ pro každé x .

Distribuční funkce spojité náhodné promenné, pokud existuje její hustota, se spočte pomocí integrace:

$$F(x) = \int_{-\infty}^x f(z) dz$$

Pro diskrétní náhodnou spočítáme její distribuční funkci jednoduše pomocí pravděpodobnostní funkce jako součet hodnot pravděpodobnostní funkce:

$$F(x) = \sum_{x_i \leq x} p_i,$$

což znamená, že distribuční funkce má v bodě x hodnotu součtu těch pravděpodobností jednotlivých hodnot x_i , které jsou menší nebo rovné x .

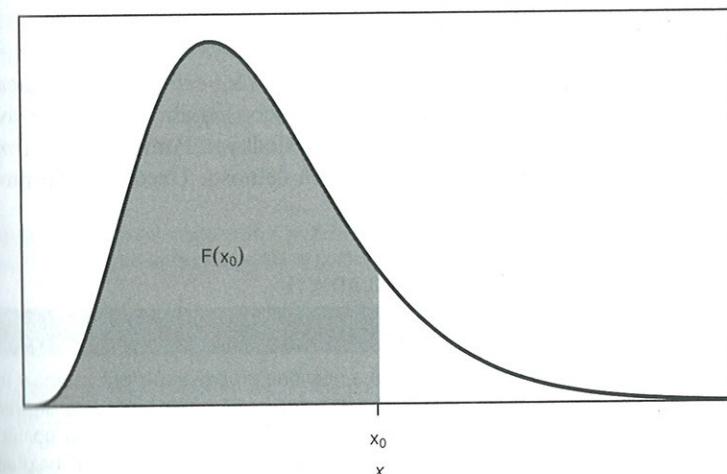
Kvantil x_p s hladinou p spojité náhodné promenné X s distribuční funkcí F je definován rovnicí

$$F(x_p) = p.$$

Kvantil $x_{0,5}$, pro který platí, že $F(x_{0,5}) = 0,5$, je **medián** (označení $\tilde{\mu}$).

Obrázek 4.3 ukazuje hustotu rozdělení a plochu, jež odpovídá hodnotě distribuční funkce v bodě x_0 .

Obr. 4.3 Hustota pravděpodobnosti $f(x)$ a hodnota distribuční funkce $F(x)$ v bodě x_0



PŘEHLED STATISTICKÝCH METOD

Hustota se využívá pro výpočet očekávané hodnoty, rozptylu nebo očekávané hodnoty jakékoli funkce $g(x)$ náhodné proměnné. Postupuje se podobně jako u diskrétní náhodné proměnné:

$$E(X) = \int_{-\infty}^{\infty} xf(x) dx \quad \text{Var}(X) = \int_{-\infty}^{\infty} (x - E(X))^2 f(x) dx$$

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x) dx$$

Některí autoři používají místo termínu rozdělení slovo **rozložení** nebo **distribuce**. Tedy rozdelení náhodných proměnných, rozložení náhodných proměnných, distribuce náhodných proměnných jsou termíny vzájemně zaměnitelné.

4.5 Základní pravděpodobnostní rozdelení

Pomocí distribuční funkce, frekvenční funkce nebo pravděpodobnostní funkce jsou popsána různá **pravděpodobnostní rozdelení náhodných proměnných**. Seznámíme se s nejdůležitějšími typy rozdělení. Nejvíce pozornosti budeme věnovat normálnímu a binomickému rozdelení. Normální rozdelení se týká spojité náhodné proměnné, binomické popisuje náhodné chování diskrétní proměnné.

4.5.1 Binomické rozdelení

Pomocí binomického rozdelení modelujeme chování četnosti prvků, které mají určitou vlastnost, v prostém náhodném výběru nebo variabilitu počtu nezávislých experimentů, jež skončily specifikovaným výsledkem. Pomocí tohoto rozdelení lze také popsat náhodné chování relativních četností. Uvedeme dva problémy vedoucí k použití binomického rozdelení.

PŘÍKLAD 4.11

Situace popsané binomickým rozdelením

- Univerzitní student Roman někdy zaspí a nestihne přednášku, která začíná v 9 hodin. Pravděpodobnost, že zaspí, je 0,4. V semestru je 12 přednášek. Lze si představit, že těchto 12 termínů představuje pro Romana 12 nezávislých pokusů dostat se na přednášku včas. Úkolem je nalézt pravděpodobnost, že v semestru přijde na přednášku včas. Roman pozdě v důsledku zaspání v polovině nebo více případů.

- Znalostní test sestává z otázek s několika volitelnými odpověďmi. Předpokládáme, že student má pravděpodobnost p , že správně odpoví na otázku náhodně zvolenou ze sestavy otázek (dobrý student má tuto pravděpodobnost vyšší, horší student nižší). Správnost odpovědi na specifickou otázku nezávisí na ostatních otázkách. Test obsahuje n otázek. Počet otázek x , které student správně zodpoví, je tedy četnost nezávislých pokusů, jež skončily „správnou odpověď“. Julie je dobrá studentka s $p = 0,75$. Máme odhadnout pravděpodobnost, že Julie zodpoví z 20 testových otázek více než 75 % otázek správně.

Zkoumané náhodné proměnné v obou úlohách (proměnnými jsou počet správných odpovědi na 20 otázek a počet „pozdních příchodů“ z 12 možných pokusů) modelujeme pomocí binomického rozdělení. Každý „pokus“ v úlohách představuje tzv. Bernoulliův pokus s dvěma možnými výsledky.

Předpoklady pro vznik náhodné proměnné X s binomickým rozdělením jsou tyto:

- provádíme n pozorování nebo pokusů;
- pozorování nebo pokusy jsou nezávislé – znalost výsledku v jednom pozorování nebo pokusu nám nic neříká o výsledku jiného pozorování;
- výsledky pozorování nebo pokusu mohou být jenom dva, nazveme je „úspěch“ a „neúspěch“;
- pravděpodobnost p každého „úspěchu“ je stejná pro všechna pozorování nebo pokusy.

Pravděpodobnostní rozdelení počtu „úspěchů“ za popsaných předpokladů nazýváme binomické rozdelení s parametry n a p . Označme počet úspěchů k . Pravděpodobnost $P(X = k)$ této hodnoty je dána vzorcem

$$P(X = k) = p(k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Připomeňme, že kombinační číslo v předchozím vzorci („ n nad k “) je počtem k -členných kombinací z n -členného souboru a počítá se podle vzorce

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

Binomické rozdelení je příklad diskrétního pravděpodobnostního rozdelení náhodné proměnné X , která může nabývat pouze $n + 1$ hodnot, přičemž n je jeden z parametrů příslušné funkce rozdelení. Při matematickém sestrojení binomického rozdelení vycházíme z tzv. Bernoulliova pokusu, jenž spočívá v tom, že v daném náhodném pokusu mohou nastat jenom dva stavy A a Ā

PŘEHLED STATISTICKÝCH METOD

s pravděpodobnostmi p a $1-p$. Takový pokus modelujeme tzv. binární náhodnou proměnnou Y , kde $P(Y=1) = p$ a $P(Y=0) = 1-p$. Náhodná proměnná Y má průměr μ a směrodatnou odchylku σ :

$$E(Y) = \mu_Y = 1 \times p_A + 0 \times (1 - p_A) = p_A$$

$$E(Y - p_A)^2 = \sigma^2 = p_A(1 - p_A)^2 + (1 - p_A)(p_A)^2 = (1 - p_A)p_A$$

Binomická náhodná proměnná určená parametry n a p vznikne jako součet n nezávislých binárních proměnných Y_i s hodnotami 0 nebo 1, které mají všechny stejné rozdělení určené parametrem p :

$$X = \sum_{i=1}^n Y_i$$

Z toho, co jsme uvedli, plyne pro očekávanou hodnotu a rozptyl součtu n nezávislých náhodných proměnných Y_i :

$$E(X) = np \quad \text{Var}(X) = np(1-p)$$

Binomické rozdělení s parametry p a n , kde p je pravděpodobnost jevu a n počet pokusů, označujeme $B(p; n)$. Proměnná X/n odpovídá relativní četnosti sledovaného jevu v n pokusech. Pro velké hodnoty n se rozdělení této náhodné proměnné blíží asymptoticky k normálním rozdělení, což rozvedeme podrobněji v kapitole 4.5.5. Tabulkou kumulativních pravděpodobností $P(X \leq x)$ binomického rozdělení pro $n < 14$ a zvolené hodnoty p nalezneme v tabulce VIII v příloze B.

PŘÍKLAD 4.12

Výpočet pravděpodobnosti na základě binomického rozdělení

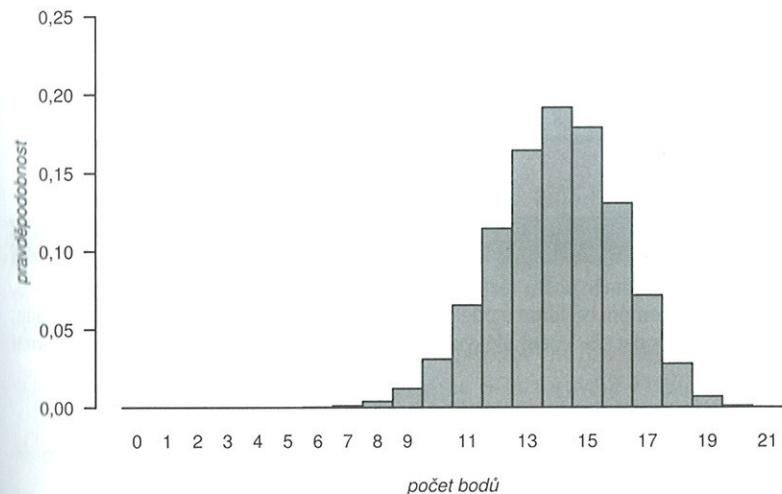
Hráč košíkové Petr promění z každých 10 trestních hodů průměrně 7. Jaká je pravděpodobnost, že v zápasu z dvaceti trestních hodů promění nejméně 15? Budeme předpokládat, že hody probíhají jako zcela nezávislé pokusy. Pak lze počet proměněných hodů Petra při 20 pokusech modelovat binomickým rozdělením $B(0,7; 20)$. Hledaná pravděpodobnost má hodnotu

$$P(X > 14) = p(15) + p(16) + p(17) + p(18) + p(19) + p(20) = \sum_{k=15}^{20} \binom{20}{k} (0,7)^k (0,3)^{20-k} = 0,41$$

Rozdělení pravděpodobností pro tento příklad je znázorněno graficky na obrázku 4.4. Odpovídající výpočty se snadno provedou v tabulkovém procesoru Excel pomocí funkce BINOMDIST. Konkrétně hledanou pravděpodobnost spočítáme takto:

$$P(X > 14) = 1 - \text{BINOMDIST}(14; 20; 0,7; \text{PRAVDA})$$

Obr. 4.4 Binomické rozdělení počtu bodů z 20 pokusů



4.5.2 Poissonovo rozdělení

Rozdělení, které je pojmenováno po francouzském matematikovi S. D. Poissonovi (1811–1840), popisuje mnoho náhodných procesů, např.:

- počet telefonních hovorů za den;
- počet nehod za danou časovou jednotku;
- počet přijímaných pacientů při noční službě na chirurgickém oddělení;
- počet tiskových chyb na jedné stránce.

Poissonovo rozdělení mají náhodné proměnné, které popisují četnosti jevů s těmito vlastnostmi:

- to, že jev v daném časovém intervalu nebo prostoru nastane (nebo nenastane), nezávisí na tom, co se stalo jindy nebo jinde;
- pro každý časový okamžik je pravděpodobnost jevu v malém časovém intervalu stejná (to se týká i jevů v malých oblastech prostoru);
- neexistuje případ, že by nastaly dva jevy přesně v jednom časovém okamžiku nebo místo prostoru.

PŘEHLED STATISTICKÝCH METOD

Střední hodnotu počtu jevů X za časovou jednotku nebo v prostorové jednotce označujeme λ . Rozdělení četnosti tohoto rozdělení je dáno je pravděpodobnostní funkcí

$$P(X = x) = p(x) = \frac{\lambda^x e^{-\lambda}}{x!},$$

kde x označuje četnost jevů ($x = 0, 1, 2, \dots$). Náhodná proměnná s tímto rozdělením má stejný průměr $E(X) = \mu = \lambda$ i rozptyl $Var(X) = \sigma^2 = \lambda$.

PŘÍKLAD 4.13

Náhodný jev popsán Poissonovým rozdělením

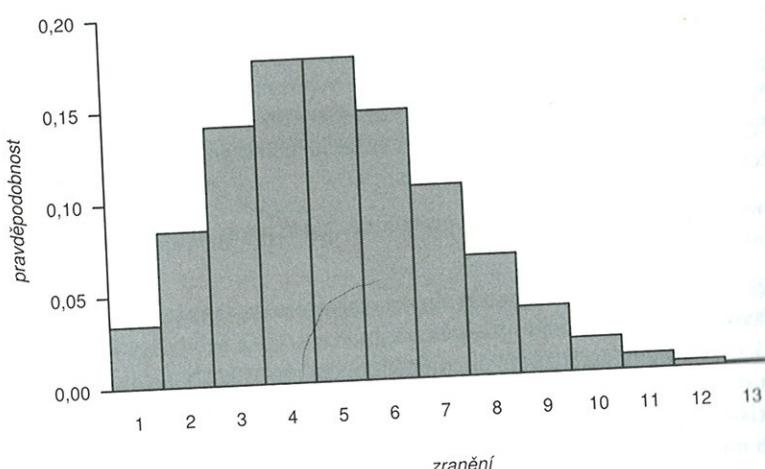
Předpokládejme, že průměr počtu těžkých zranění v jednom ročníku hokejové ligy je 5. Rozdělení četnosti zranění budeme modelovat pomocí Poissonova rozdělení. Máme zjistit, jaká je pravděpodobnost, že počet těžkých zranění bude více než 4. Rozdělení četnosti zranění X má tvar

$$p(x) = \frac{5^x e^{-5}}{x!}.$$

Pomocí tohoto vzorce vypočítáme pravděpodobnosti četnosti zranění $X = 0; 1; 2; 3; 4$:

$$p(0) = 0,00674, p(1) = 0,03370, p(2) = 0,08425, p(3) = 0,14042, p(4) = 0,17552.$$

Obr. 4.5 Poissonovo rozdělení pravděpodobnosti četnosti zranění



Tyto pravděpodobnosti sečteme a dostaneme hodnotu 0,44. Hledaná pravděpodobnost, že počet zranění přesahne 4, má tedy hodnotu $1 - 0,44 = 0,56$. Rozdělení pravděpodobností pro tento příklad je znázorněno graficky na obrázku 4.5. Odpovídající výpočty se snadno provedou v tabulkovém procesoru Excel pomocí funkce POISSON.

Například výše uvedenou pravděpodobnost spočítáme pomocí funkce POISSON takto:

$$P(X > 4) = 1 - \text{POISSON}(4; 5; \text{PRAVDA})$$

S rostoucí hodnotou λ se tvar tohoto rozdělení blíží k normálnímu rozdělení. Jestliže náhodná proměnná má binomické rozdělení $B(p; n)$, pak tvar jejího rozdělení se blíží k Poissonovu s parametrem np , pokud n je velké a p se blíží k nule. Aproximativně můžeme tedy rozdělení $B(p; n)$ s velkým n a malou hodnotou p nahradit Poissonovým rozdělením.

Součet nezávislých proměnných s Poissonovým rozdělením je opět rozdělen podle tohoto rozdělení. Jestliže máme n pozorování Poissonova rozdělení s průměrem λ , pak součet pozorování je možné považovat za pozorování s Poissonovým rozdělením a parametrem $n\lambda$.

4.5.3 Normální rozdělení

Normální rozdělení – nazývané též *Gaussovo rozdělení* (a v anglicky psané literatuře *rozdělení zvonovitého tvaru – bell curve*) je asi nejvíce používané rozdělení pro modelování náhodného chování proměnných v empirických vědách. Je tomu tak minimálně ze čtyř příčin:

- 1. mnoho sledovaných proměnných můžeme approximativně (tzn. s uspokojivým přiblížením) modelovat pomocí tohoto rozdělení;
- 2. některé jiné proměnné lze převést jednoduchou transformací na proměnnou, jež má normální rozdělení;
- 3. existuje mnoho statistických procedur, které byly v důsledku předchozích dvou bodů odvozeny pro toto rozdělení;
- 4. protože platí centrální limitní teorém (viz 4.5.5), lze často approximativně použít procedury, jež byly na základě normálního rozdělení navrženy, také při statistickém hodnocení proměnných, které se tímto rozdělením vůbec neffidí.

Matematik Abraham de Moivre v roce 1733 popsal pomocí normální křivky limitní chování binomického rozdělení, když se snažil approximovat výpočty jednotlivých pravděpodobností binomického rozdělení pro velká n . Rozdělení, jež navrhl de Moivre pro tento účel, se nakonec

PŘEHLED STATISTICKÝCH METOD

ukázalo být důležitější než samo binomické rozdělení. V roce 1812 odvodil nezávisle normální rozdělení francouzský matematik Pierre Laplace (1749–1827). Jak Laplace, tak Karl Friedrich Gauss (1772–1855) interpretovali toto rozdělení jako zákon chyb a používali ho pro interpretaci astronomických a geodetických měření, výsledků hazardních her a přesnosti dělostřelecké střelby.

Vzorec pro hustotu $f(x)$ normálního rozdělení $f(x)$ je dán vztahem

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2},$$

kde μ a σ jsou parametry, které ovlivňují tvar křivky této funkce. Když využijeme středoškolské znalosti a funkci analyzujeme, zjistíme, že parametr μ určuje, kde má křivka maximum. Parametr σ naproti tomu určuje, jak jsou po obou stranách od hodnoty μ vzdáleny inflexní body, tedy jak je křivka roztažena do šířky. Je patrné, že křivka je zvonovitého tvaru, symetrická kolem μ . Také je patrné, že je vždy kladná a nenulová pro všechny hodnoty x z oboru reálných čísel. Když vypočítáme pomocí integrace očekávanou hodnotu a směrodatnou odchylku, zjistíme že mají hodnoty právě μ a σ . Na obrázku 4.6 je grafické znázornění této funkce pro dané parametry μ a σ .

Zkráceně označujeme normální rozdělení se střední hodnotou μ a směrodatnou odchylkou σ symbolem $N(\mu; \sigma^2)$ s uvedením parametrů střední hodnoty a rozptylu (pozor: ne směrodatné odchylky σ). Jestliže náhodná proměnná X má takové rozdělení, vyjadřujeme to symbolicky zápisem

$$X \sim N(\mu; \sigma^2).$$

Například normální rozdělení se střední hodnotou 0 a směrodatnou odchylkou 1 označujeme $N(0; 1)$. Normální rozdělení se střední hodnotou 15 a směrodatnou odchylkou 3 označujeme $N(15; 9)$, protože jeho rozptyl je 9.

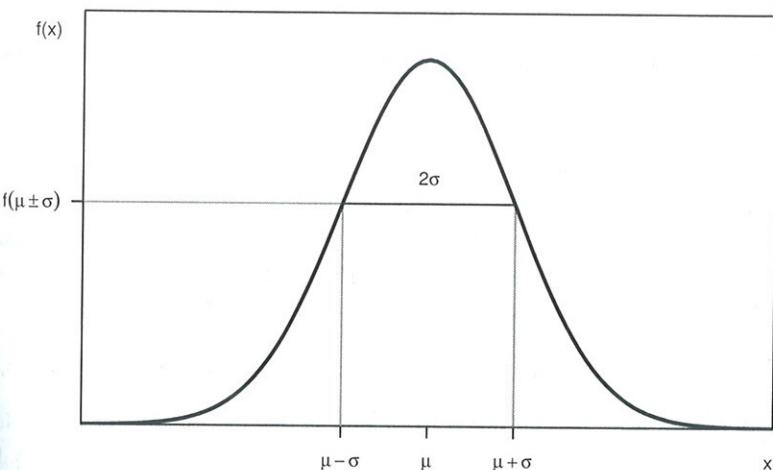
Přesnější interpretaci parametru rozptylenosti σ přibližují vztahy, které uvádějí pravděpodobnosti různých intervalů kolem středu rozdělení. Pro každé $N(\mu; \sigma^2)$ platí:

- interval $\mu \pm \sigma$ obsahuje 68,3 % populace,
- interval $\mu \pm 2\sigma$ obsahuje 95,5 % populace,
- interval $\mu \pm 3\sigma$ obsahuje 99,7 % populace;

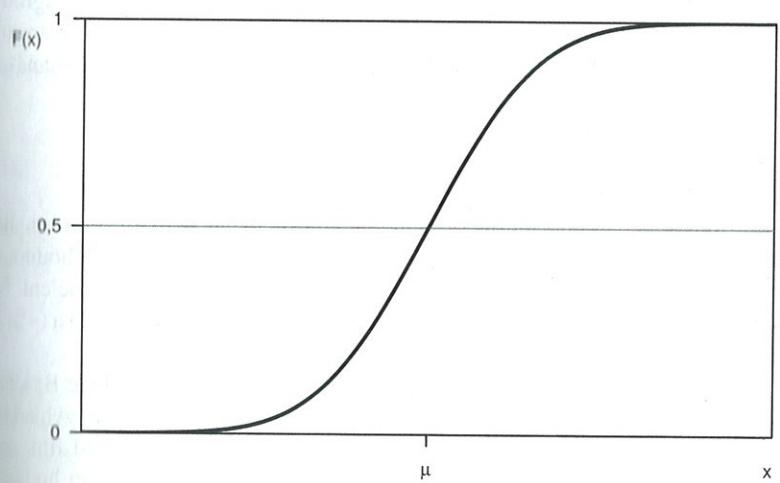
a obráceně:

- 95,0 % populace je obsaženo v intervalu $\mu \pm 1,96\sigma$;
- 99,0 % populace je obsaženo v intervalu $\mu \pm 2,58\sigma$;
- 99,9 % populace je obsaženo v intervalu $\mu \pm 3,29\sigma$.

Obr. 4.6 Křivka hustoty pravděpodobnosti normálního rozdělení daná parametry μ a σ



Obr. 4.7 Graf distribuční funkce $F(x)$ pro normální rozdělení z obr. 4.6



PŘEHLED STATISTICKÝCH METOD

Například 68 % pozorovaní normální náhodné proměnné se střední hodnotou 100 a směrodatnou odchylkou 15 bude ležet v intervalu 100 ± 15 (tzn. mezi hodnotami 85 a 115), 95 % hodnot bude ležet v intervalu 100 ± 30 (tzn. mezi hodnotami 70 a 130), a skoro všechna pozorování budou ležet uvnitř intervalu 100 ± 45 (tzn. mezi hodnotami 55 a 145). Tyto vztahy platí přibližně i v případě, že za teoretické parametry μ a σ dosadíme empirické hodnoty \bar{x} a s vypočítané pomocí dat za předpokladu, že měření jsou přibližně normálně rozdelená. Přesné určování kvantilů pro rozdělení $N(\mu; \sigma^2)$ se provádí pomocí vhodného počítačového programu nebo tabulek standardizovaného rozdělení $N(0; 1)$. Tento postup vysvětlíme v dalším odstavci.

Uvedeme ještě jeden významný teoretický poznatek, který se týká součtu nezávislých normálně rozdelených proměnných. Pokud sledujeme dvě nezávislé náhodné proměnné X a Y , jež jsou normálně rozdelené tak, že $X_1 \sim N(\mu_1; \sigma_1^2)$ a $X_2 \sim N(\mu_2; \sigma_2^2)$, má rozdělení jejich součtu $X_1 + X_2$ tvar

$$X_1 + X_2 \sim N(\mu_1 + \mu_2; \sigma_1^2 + \sigma_2^2).$$

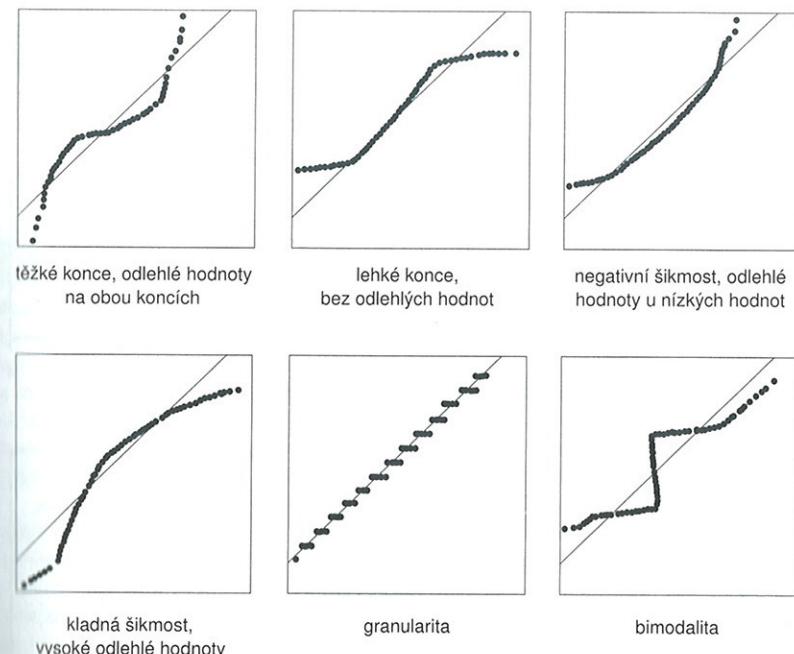
Předpoklad, že data mají normální rozdělení, se ověřuje mnoha způsoby. Mezi grafické prostředky patří histogram. Často se také používá diagram zvaný Q-Q graf. Body grafu vznikají tak, že na horizontální osu vynášíme pozorované hodnoty x_i a na vertikální osu hodnoty z_i . Hodnoty z_i odpovídají x_i , ale jsou vypočtené jako kvantily normálního rozdělení s průměrem \bar{x} a směrodatnou odchylkou s původními daty. Hladina kvantilu z_i odpovídá hladině, již má vzhledem k empirické distribuční funkci hodnota x_i . Jestliže data jsou normálně rozdelená, graf má přímkový charakter. Na obrázku 4.8 je znázorněno 6 Q-Q diagramů pro různé typy rozdělení jako reference pro interpretaci. Podobně se sestrojuje a interpretuje P-P graf, v němž se pro jednotlivá x_i na osy X , resp. Y nanáší příslušné hodnoty empirické, resp. normální distribuční funkce.

4.5.4 Standardizované normální rozdělení

Zvláštní místo ve třídě normálních rozdělení zaujímá standardizované normální rozdělení. Je to rozdělení $N(0; 1)$, tedy normální rozdělení se střední hodnotou nula a jednotkovým rozptylem. Někdy se toto rozdělení nazývá Z-rozdělení. Na obrázku 4.9 je znázorněn jeho tvar a pravděpodobnosti intervalů $(-1; 1)$ a $(-2; 2)$ v procentuálním vyjádření.

Užitečnost tohoto rozdělení spočívá v tom, že tabulku II v příloze B, která obsahuje vybrané hodnoty jeho distribuční funkce, lze použít pro vyhledání odpovídajících hodnot pro všechna ostatní normální rozdělení. Před tím, než tuto tabulku použijeme, musíme naše údaje standardizovat. Abychom hodnotu

Obr. 4.8 Možné tvary diagnostického Q-Q diagramu

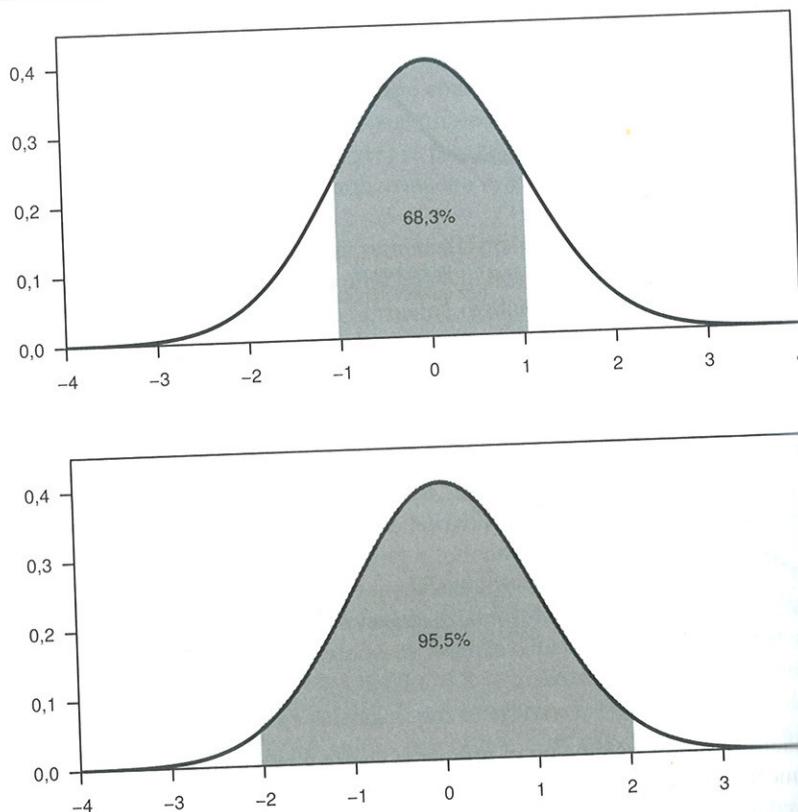


standardizovali, odečteme od ní průměr (výběrový nebo teoretický) a rozdíl vydělíme směrodatnou odchylkou (výběrovou nebo teoretickou). Když tuto operaci provedeme pro určitou řadu měření a použijeme přitom příslušné výběrové charakteristiky, získáme nové údaje, tzv. standardizované skóry nebo z -skóry, pro něž platí, že jejich průměr se rovná 0 a směrodatná odchylka jedná. V dalším výkladu provedeme provádět standardizaci teoretickými parametry μ a σ . Pro hodnotu z se používá standardizační transformace podle vzorce

$$z = \frac{x - \mu}{\sigma}.$$

Na příklad pro hodnotu $x = 115$ z normálního rozdělení se střední hodnotou 100 a směrodatnou odchylkou 15 je z -hodnota +1. Pro číslo $x = 85$ je z -hodnota -1, což také znamená, že hodnota 85 leží jednu směrodatnou odchylku pod průměrem.

Obr. 4.9 Standardizované normální rozdělení s hodnotami pravděpodobnosti (v procentech) dvou intervalů $(-1; 1)$ a $(-2; 2)$



Tabulky rozdělení $N(0; 1)$ se používají především ve dvou úlohách:

- při výpočtu kvantilu o dané hladině pro rozdělení $N(\mu; \sigma^2)$;
- při určení hodnot distribuční funkce rozdělení $N(\mu; \sigma^2)$.

Tabulky standardizovaného normálního rozdělení (viz tab. II v příloze B) obvykle uvádějí pro tabulované hodnoty z příslušnou hodnotu hustoty $f(z)$ a distribuční funkce $F(z)$. První hodnotu používáme zřídka. Hodnoty funkce $F(z)$ naopak velice často. V programových systémech jsou příslušné výpočty automaticky.

matizované – v programu Excel použijeme pro nalezení hodnot $F(z)$ funkci NORMDIST. Přesto je užitečné se naučit s těmito tabulkami pracovat. Malý problém spočívá v tom, že tabulky $N(0; 1)$ rozdělení jsou pro stručnost počítány jenom pro kladné hodnoty z . Jak se však dopočítají hodnoty distribuční funkce $F(z)$ pro záporná z ? Abychom to dokázali, využijeme symetrii rozdělení kolem 0. Pro dané $z < 0$ najděme nejdříve hodnotu $F(-z)$ a pak vypočítáme hledanou pravděpodobnost pomocí rovnice $F(z) = 1 - F(-z)$. Tento postup demonstруjeme při hledání hodnot distribuční funkce rozdělení $N(\mu; \sigma^2)$.

PŘÍKLAD 4.14

Výpočty kvantilů a hladin kvantilů normálního rozdělení

Ptáme se například, kolik procent hodnot leží pod číslem 115, jestliže sledujeme proměnnou s normálním rozdělením se střední hodnotou 100 a směrodatnou odchylkou 15? Chceme tedy určit hodnotu distribuční funkce $N(100; 15^2)$ pro $x = 115$. Postupujeme tak, že hodnotu x standardizujeme a pak najdeme vypočítaný z -skóre v tabulce rozdělení $N(0; 1)$ a k němu přířazenou hodnotu distribuční funkce: Vyjde nám $z = 1$ a v tabulce najdeme odpovídající hodnotu $F(1) = 0,8413$. Zjistili jsme, že pod hodnotou 115 leží přibližně 84 % měření dané proměnné.

Jak však tutéž úlohu vyřešíme pro $x = 85$, kdy $z = -1$? Protože pro $z < 0$ hodnoty distribuční funkce tabelovány nejsou, využijeme symetrii normálního rozdělení. Nejdříve zjistíme v tabulkách hodnotu distribuční funkce pro absolutní hodnotu z a pro záporné z odečteme nalezenou hodnotu od jedné. Tak dostaneme hledanou hodnotu distribuční funkce $F(z)$. Pro $z = -1$ získáme tímto postupem hodnotu 0,1587. Můžeme tedy tvrdit, že pod hodnotou 85 v případě rozdělení $N(100; 15^2)$ leží přibližně 16 % naměřených hodnot.

Opačný výpočet provádíme, když chceme zjistit hodnotu kvantilu x_p o dané hladině p pro obecné rozdělení $N(\mu; \sigma^2)$. V tomto případě nejdříve najdeme v tabulce II přílohy B kvantil z_p standardizovaného rozdělení pro danou hladinu a pak provedeme transformaci

$$x_p = z_p \sigma + \mu.$$

Jestliže hladina p je menší než číslo 0,5, musíme opět využít symetrii normálního rozdělení. Vyhledáme tedy kvantil x_{1-p} pro hodnotu $1-p$ (která je větší než 0,5). Hledaný kvantil pak má hodnotu $x_p = -x_{1-p}$.

Pokračování příkladu 4.14

Příklad chceme nalézt kvantil rozdělení $N(100; 15^2)$ o hladině 25 %. Jedná se tedy o první teoretický kvartil tohoto rozdělení. Nejdříve vyhledáme kvantil $z_{0,75}$, který má hodnotu 0,68. Našich úvah plyně, že hledaný 25% kvantil rozdělení $N(0; 1)$ má hodnotu $-0,68$. Pak

PŘEHLED STATISTICKÝCH METOD

provedeme transformaci $x_p = z_p \sigma + \mu = -0,68 \times 15 + 100 = 89,8$. Nalezli jsme hodnotu, pod níž leží přibližně 25 % hodnot. Nad číslem 89,8 se nachází 75 % hodnot rozdělení $N(100; 15^2)$.

Poslední příklad vyřešíme použitím funkce NORMINV programu Excel. Pracujeme s rozdělením $N(100; 15^2)$ a hledáme kvantil s hladinou 0,25:

$$89,8 = \text{NORMINV}(0,25; 100; 15)$$

Otázku, kolik procent leží pro stejné rozdělení pod hodnotou 115, zodpovíme využitím funkce NORMDIST:

$$0,8413 = \text{NORMDIST}(115; 100; 15)$$

4.5.5 Centrální limitní teorém

Mimořádné postavení normálního rozdělení spočívá kromě jiného v tom, že součet mnoha nezávislých libovolně rozdelených náhodných proměnných je přibližně normálně rozdelen, a to tím lépe, čím je sčítanců více. Toto tvrzení o asymptotickém (s rostoucím počtem sčítanců) chování součtu náhodných proměnných, které přesně vyjadřuje **centrální limitní teorém**, je základem pro skutečnost, že mnoho rozdělení výběrových statistik lze approximovat (přibližně popsat) při větším rozsahu výběru normálním rozdělením.

Jako první tento teorém formuloval Pierre Laplace v roce 1810. Neformálně lze ho vyjádřit takto: proměnná (výška, váha, reziduální hodnota apod.), pokud vznikla jako součet velkého počtu efektů nezávisle působících přičin, má přibližně normální rozdělení. Approximace je tím lepší, čím je počet přispívajících faktorů větší. Teorém se nazývá *limitní*, protože říká, co nastává v limitě, když počet přispívajících prvků se blíží k nekonečnu. Slovo *centrální* označuje jeho základní význam v počtu pravděpodobnosti. Uvedeme jednu z formulací tohoto teorému.

Centrální limitní teorém: Pokud mají prvky X_i posloupnosti nezávislých náhodných proměnných stejně rozdělení se střední hodnotou μ a směrodatnou odchylkou σ , pak rozdělení náhodné proměnné Z_n

$$Z_n = \frac{1}{n} \sum_{i=1}^n X_i$$

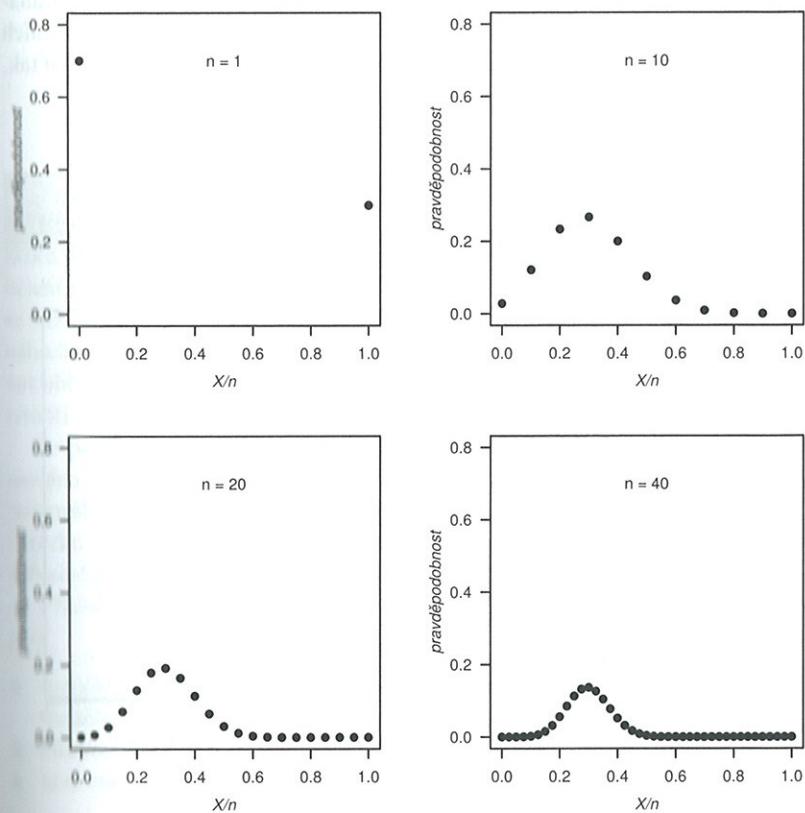
se s rostoucím n blíží k normálnímu rozdělení se střední hodnotou μ a směrodatnou odchylkou σ/\sqrt{n} .

Z této věty plyne, že aritmetický průměr jako náhodná proměnná je za velmi malých omezení asymptoticky normálně rozdelen. Jeho náhodné chování můžeme approximovat pomocí normálního rozdělení.

PŘÍKLAD 4.16

Případ asymptoticky normálního chování náhodné proměnné

Náhodná binární proměnná Y , která s pravděpodobností p nabývá hodnotu 1 a s pravděpodobností $1-p$ hodnotu 0, je extrémní případem náhodné proměnné, jež není normálně rozdelená. Rozdělení náhodné proměnné X , které vznikne jako součet n nezávislých proměnných Y , se řídí binomickým rozdělením. Jestliže vydělíme náhodnou proměnnou Z (což je součet n nezávislých proměnných Y) hodnotou n , získáme novou náhodnou proměnnou Z (jako aritmetický průměr). Pro chování tohoto průměru lze uplatnit centrální limitní teorém. Proto bude mít tento průměr s rostoucím n přibližně normální rozdělení se střední hodnotou $E(Z) = p$ a rozptylem $Var(Z) = p(1-p)/n$.

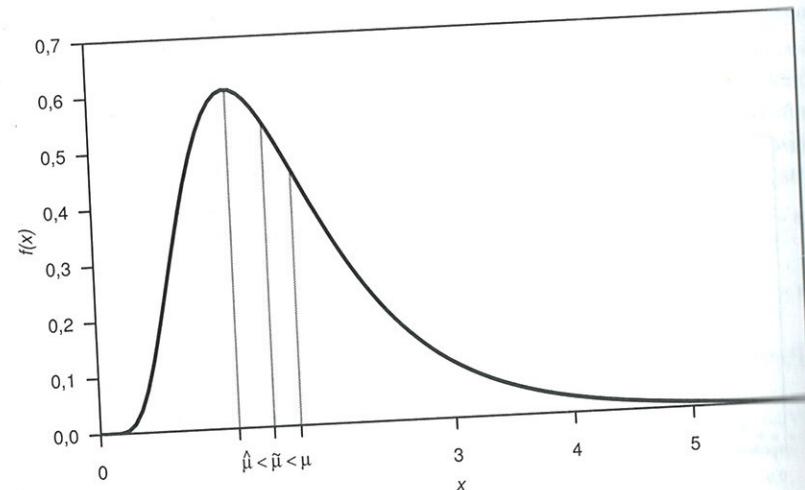
Obr. 4.10 Tvary binomického rozdělení $B(0,3; n)$ pro $n = 1, 10, 20$ a 40 

Obrázek 4.10 demonstruje, jak se mění tvar rozdělení s rostoucím n a konvergenci k normálnímu rozdělení. Budeme uvažovat transformovanou binomickou proměnnou X/n , která převádí její hodnoty do intervalu $(0; 1)$. Odpovídá to výpočtu průměru ze součtu n jednotlivých hodnot. Jednotlivé obrázky znázorňují funkční hodnoty pravděpodobnostní funkce $B(0,3; n)$ pro $n = 1, 10, 20$ a 40 , přičemž na osu X nanášíme hodnotu náhodné proměnné transformované do relativních četností X/n .

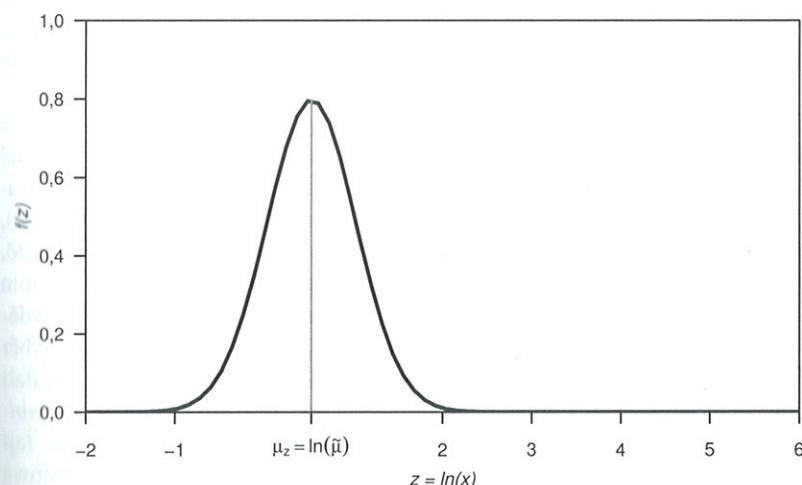
4.5.6 Log-normální rozdělení

Normální rozdělení není vhodné pro modelování *asymetricky* rozdělených náhodných proměnných. Jestliže zpracováváme měření takových proměnných pomocí technik založených na normálním rozdělení, musíme zvážit, zda je to vhodné. Mnoho technik je však k odchylkám od normality robustní (tzn. ponechávají si své dobré vlastnosti i při porušení předpokladu o normálním rozdělení analyzované náhodné proměnné). Nemusíme se zříci ani použití méně robustních technik, jestliže se nám podaří pomocí jednoduché transformace data upravit tak, aby jejich empirické rozdělení mělo přibližně normální tvar.

Obr. 4.11 Křivka log-normálního rozdělení s vyznačeným průměrem, mediánem a modem



Obr. 4.12 Graf hustoty f náhodné proměnné $Z = \ln(X)$ pro logaritmicko-normálně rozdělenou proměnnou X



Hustota mnoha empirických rozdělení má tvar, který je zobrazen na obrázku 4.11. Takové proměnné nenabývají záporných hodnot. Křivka jejich rozdělení je zleva nejdříve velmi strmá, po nabytí maxima se stává plošší a pomalu se blíží k ose X . Když zkoumáme vzájemnou polohu charakteristik polohy takové náhodné proměnné, zjistíme, že modus je menší než medián a průměr je vyšší než obě tyto hodnoty. (U symetrických rozdělení, jako je normální rozdělení, tyto tři charakteristiky mají samozřejmě stejné hodnoty.)

Je prokázané, že zejména empirická rozdělení mohou být dobře approximována log-normálním rozdělením. Říkáme, že náhodná proměnná X má **log-normální rozdělení**, jestliže proměnná $Z = \ln(X)$ má normální rozdělení. Na obrázku 4.12 je znázorněna hustota proměnné Z , jež vnikla logaritmováním náhodné proměnné X .

Příklad proměnných, které mají přibližně logaritmicko-normální rozdělení,

- rozdělení různě měřených časů nebo rozdělení věku obyvatelstva v populaci, rozdělení příjmů;
- citlivost lidí i zvířat na účinek farmak;
- koncentrace různých látek v krvi (bilirubin, kalcium).

4.6 Pojem výběrového rozdělení

Výzkumník vypočítává z dat různé statistiky a provádí jejich interpretaci pomocí metod statistického usuzování, o nichž pojednáme v další kapitole. V tomto odstavci popíšeme náhodné chování některých důležitých statistik. Jinými slovy, budeme se zabývat jejich pravděpodobnostním rozdělením. Tyto poznatky nám pomohou při výkladu postupů statistického usuzování.

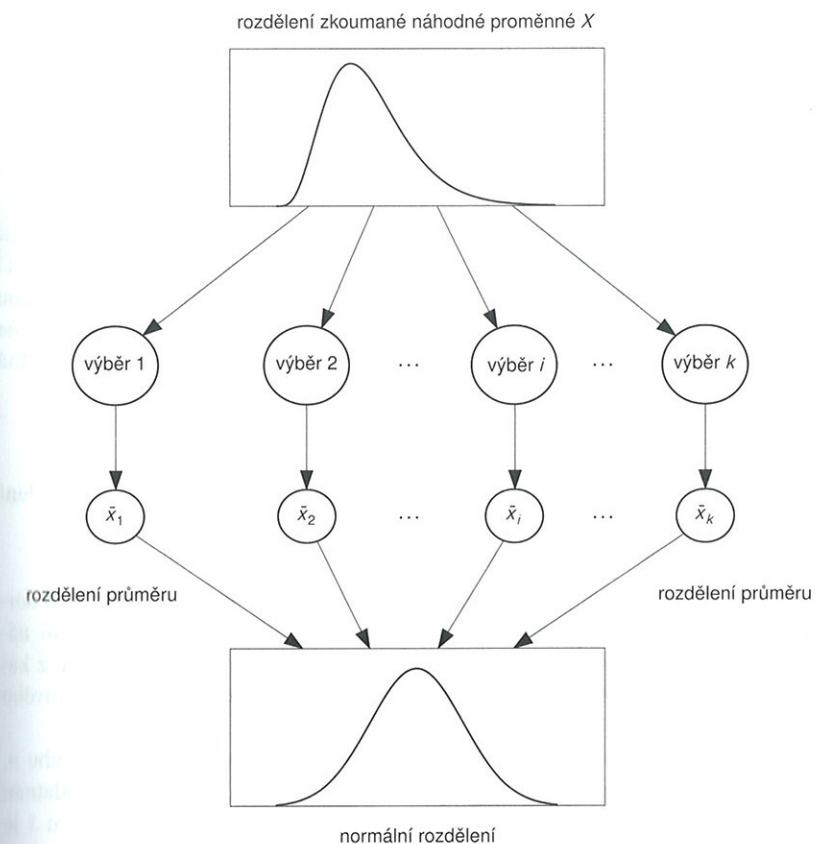
Výběrové rozdělení statistiky je pravděpodobnostní rozdělení hodnot, které statistika nabývá ve všech možných výběrech o daném rozsahu ze specifikované populace. Kdybychom provedli totální výběr z populace a spočítali pro sledovanou proměnnou popisné statistiky (např. průměr nebo směrodatnou odchylku), zjistili bychom parametry jejího statistického chování zcela přesně. V případě, že máme k dispozici pouze výběr z populace, vypočítáme pomocí něho jenom odhad parametrů rozdělení. Tyto odhady jsou statistikami a budou se pravděpodobně lišit od skutečných hodnot parametrů. Jestliže provedeme další výběr z téže populace, nové statistiky se budou lišit také od těch, jež jsme spočítali z prvního výběru. Tyto diferenční jsou známé pod názvem **výběrová variabilita**. Tako pojatou statistiku lze jistě považovat za náhodnou proměnnou. Její rozdělení nazýváme **výběrové rozdělení**. Je užitečné charakterizovat výběrové rozdělení běžných statistik.

Výběrové rozdělení statistiky je definováno množinou hodnot, které statistika může nabývat, když provedeme nezávisle všechny možné výběry o stejném rozsahu a ze stejné populace, a dále pravděpodobnostním rozdělením těchto hodnot. Pravděpodobnostní rozdělení hodnot statistiky charakterizuje, jak je statistika náhodně proměnlivá. Jestliže získáme určitý počet výběrů o stejném rozsahu těchto hodnot a approximujeme její výběrové rozdělení. V této kapitole máme vždy myslí prostý náhodný výběr z nekonečné nebo velmi rozsáhlé populace.

4.6.1 Výběrové rozdělení aritmetického průměru při známém σ

Zdůraznili jsme, že nejenom vlastní data jsou proměnlivá, také vypočítané statistiky jsou od výběru k výběru náhodně proměnlivé. Proměnlivost vypočítaných charakteristik zachycujeme často jedním parametrem, který je odvozen od směrodatné odchylky. Neměří rozptýlenost původní náhodné proměnné, ale rozptýlenost vypočítaných statistik. Tento parametr variability nazýváme **střední chybu odhadu** nebo **směrodatnou chybou**.

Obr. 4.13 Vznik rozdělení aritmetického průměru



Základní princip stojící za směrodatnou chybou odhadu ozřejmíme na variabilitě aritmetického průměru. Představme si, že parametr μ sledované náhodné proměnné neodhadujeme pomocí dat jednoho výběru, nýbrž k výběry (obr. 4.13). Přitom se vychází ze stejně velikých výběrů, jež se provedou ze stejné základní populace. Pro každý výběr můžeme spočítat průměr \bar{x}_i . Tyto průměry mají určité rozdělení, výběrové rozdělení průměrů.

Výsledné rozdělení má nějaký teoretický průměr $\mu_{\bar{x}}$ a směrodatnou odchylku $\sigma_{\bar{x}}$. Průměr $\mu_{\bar{x}}$ je roven průměru μ původní náhodné proměnné. Jakou hodnotu

má však směrodatná odchylka $\sigma_{\bar{x}}$? Lze ji přímočaře odhadnout pomocí získaných průměrů \bar{x}_i , když vypočteme jejich směrodatnou odchylku

$$s_{\bar{x}} = \sqrt{\sum (\bar{x}_i - \bar{x})^2 / (k - 1)},$$

kde \bar{x} je průměr vypočítaný z jednotlivých výběrových průměrů. Vztah mezi teoretickým parametrem σ rozdělení proměnné a rozdělením průměrů popisuje vzorec

$$\sigma_{\bar{x}} = \sigma / \sqrt{n},$$

který ukazuje, že rozptýlenost průměrů se s rostoucím n zmenšuje. Parametr $\sigma_{\bar{x}}$ nazýváme **směrodatná chyba odhadu průměru** nebo prostě směrodatná chyba průměru. Také se používá výraz střední chyba odhadu. Směrodatnou chybu průměru můžeme odhadnout pomocí tohoto vzorce, když do něho za σ dosadíme vypočítanou směrodatnou odchylku s z jednoho výběru. Výběrová směrodatná chyba průměru má tedy hodnotu

$$s_{\bar{x}} = s / \sqrt{n}.$$

Jestliže sledujeme normálně rozdělenou náhodnou proměnnou, pak pro rozdělení průměru platí

$$\bar{X} \sim N(\mu; \sigma^2/n).$$

Centrální limitní teorém zajišťuje, že výběrové rozdělení průměru se blíží k normálnímu rozdělení i v případě, že nepředpokládáme normalitu rozdělení náhodné proměnné v původním souboru. Původní formulaci tohoto teorému z kapitoly 4.5.5 uvedeme v jednodušší podobě a využijeme přitom pojmu výběrového rozdělení.

Centrální limitní teorém: Získáme náhodný výběr z populace o rozsahu n , kde náhodná proměnná X je rozdělena se střední hodnotou μ a směrodatnou odchylkou σ . Pokud n je veliké, pak výběrové rozdělení hodnot průměru \bar{X} je blízké normálnímu rozdělení $N(\mu; \sigma^2/n)$. Symbolicky vyjádřeno

$$\bar{X} \sim N(\mu; \sigma^2/n) \quad (\text{asymptoticky}).$$

Výraz *asymptoticky* znamená, že vztah platí tím lépe, čím je rozsah výběru větší.

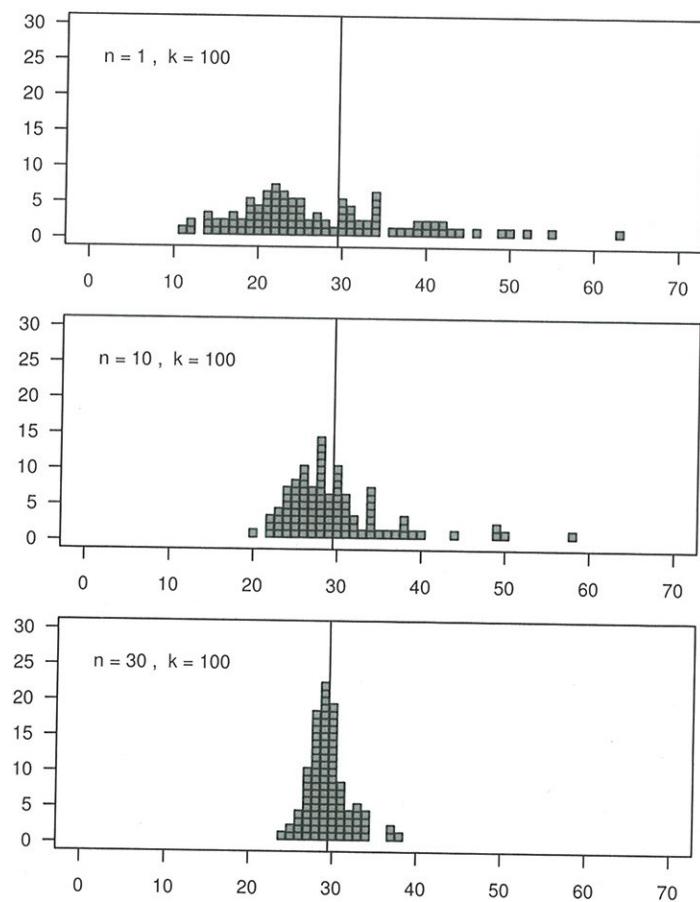
Tento vztah pak můžeme pro standardizovanou hodnotu průměru $(\bar{x} - \mu)/\sigma_{\bar{x}}$ přepsat výrazem

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0; 1) \quad (\text{asymptoticky}).$$

Jak je approximace dobrá, závisí na původním rozdělení. Approximace platí pouze měrně dobře pro rozsahy n větší než 30.

Podobné odvození a vzorce používáme pro charakteristiku variability mnoha dalších statistik. Statistika z v tomto případě udává, jak je výběrový aritmetický průměr vzdálen do hodnoty μ – měřeno v jednotkách, které určuje střední chyba. Další odstavce rozšiřují uvedené poznatky o rozdělení rozdílů průměrů, relativních četností a rozptylů.

Obr. 4.14 Tři experimenty s výběry o různých rozsazích ze stejné populace – histogramy ukazují četnosti hodnot průměrů spočtených pro každý výběr rozsahu n



PŘÍKLAD 4.17

Zkoumání vlastností výběrového průměru metodou simulace

Ilustrujeme rozdělení průměrů a závislost jeho chování na velikosti výběru pomocí simulacích experimentů, v nichž pozorujeme náhodnou proměnnou s průměrem 29,5 a směrodatnou odchylkou 13,6, která má mírně zešikmené rozdělení. Na obrázku 4.14 jsou pomocí histogramů zobrazeny výsledky tří experimentů.

Druhá simulace ukazuje výběr $k = 100$ hodnot (velikost výběru $n = 1$).

Druhá simulace ukazuje $k = 100$ průměrů spočítaných vždy pro výběr $n = 10$ hodnot.

Třetí simulace ukazuje $k = 100$ průměrů spočítaných vždy pro výběr $n = 30$ hodnot.

Všimněme si, že rozptýlenost zobrazovaných hodnot postupně klesá. Nejvíce je pro rozdělení původních dat. Nejmenší rozptýlenost mají průměry spočítané z 30 pozorování. Všimněme si také dvou dalších skutečností. Rozdělení se v důsledku působení centrálního limitního teorému stále více podobá normálnímu rozdělení. Navíc se stále silněji centruje kolem průměru 29,5, což ukazuje na působení zákona velkých čísel, jenž říká, že průměr spočtený ze stále více hodnot konverguje k teoretické průměrné hodnotě.

4.6.2 Výběrové rozdělení aritmetického průměru při neznámém σ

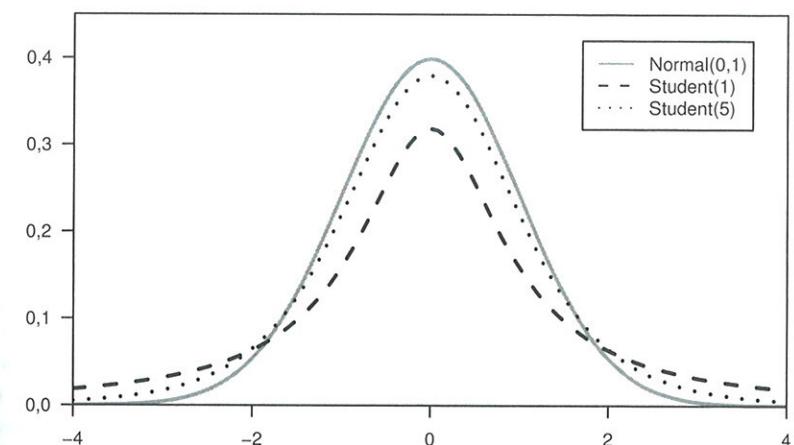
V předchozím odstavci jsme popsali rozdělení průměru, když známe teoretickou směrodatnou odchylku uvažované náhodné proměnné. Pokud neznáme směrodatnou odchylku σ , lze místo ní dosadit vypočítanou výběrovou směrodatnou odchylku s . Použití tohoto odhadu má však určité důsledky.

Tímto problémem se zabýval William S. Gosset (1876–1937), který pracoval jako sládek ve slavném pivovaru Guinness. Gosset se při své práci jako kontrolor kvality setkal s problémem, že neznal teoretickou směrodatnou odchylku měření, jež prováděl při kontrole kvality surovin k výrobě piva. Navíc jeho pokusy neobsahovaly mnoho pozorování. Proto si Gosset položil otázku, jaké je výběrové rozdělení statistiky $(\bar{x} - \mu)/s$. Guinness mu dovolil řešení publikovat pod pseudonymem Student v roce 1907. Proto se t -rozdělení, které Gosset popsal, nazývá někdy Studentovo t -rozdělení.

Gosset ukázal, že výběrové rozdělení aritmetického průměru se může po standardizaci střední hodnotou μ a výběrovou střední chybou s_x (kdy získáme t -statistiku) reprezentovat t -rozdělením, jehož tvar závisí na **stupních volnosti** ($st.v. = n - 1$). Symbolicky tento poznatek vyjádříme vztahem:

$$\frac{\bar{X} - \mu}{s_{\bar{x}}} = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t(st.v.) \quad (st.v. = n - 1)$$

Obr. 4.15 Srovnání tvaru Studentova t -rozdělení a normálního rozdělení



Tvrzení o tvaru rozdělení t -statistiky platí v tom případě, že rozdělení náhodné proměnné je normální. Poznamenejme, že t -statistika má stejnou interpretaci jako standardizovaná z -statistika z předchozho odstavce. Vyjadřuje, jak je aritmetický průměr vzdálen od μ v jednotkách určených střední chybou.

Studentovo t -rozdělení je podobné normálnímu standardizovanému rozdělení v tom, že je symetrické kolem nuly, dále je unimodální (má jeden vrchol) a má vonovitý tvar. Na rozdíl od něho má však trochu výraznější oba okraje.

Studentovo t -rozdělení je ve skutečnosti celá třída rozdělení, podobně jako u normálních rozdělení. Jeho určujícím parametrem jsou však pouze stupně volnosti. Stupně volnosti ovlivňují váhu (vyjádřenou pravděpodobnostmi) jeho okrajů. Pro menší stupně volnosti jsou váhy okrajů větší než pro velké stupně volnosti (obr. 4.15). S rostoucím n se tvar t -rozdělení approximativně blíží k normálnímu rozdělení. Protože existuje t -rozdělení nekonečně mnoho, v tabulce IV písmohy B uvádíme hodnoty jejich kvantilů pouze pro vybrané případy.

PŘÍKЛAD 4.18

Aplikace t -rozdělení pro studium chování průměru při neznámém rozptylu

Statistický úřad zjistil, že v daném městě je rozdělení věku živitelů rodin normální se střední hodnotou 41,3 roky. Prostý náhodný výběr 20 živitelů rodin vedl k průměru 44,1 roků se

směrodatnou odchylkou 9,6 roků. Jaká je pravděpodobnost, že výběr o tomto rozsahu bude mít průměr 44,1 nebo větší za předpokladu, že v celém městě je průměr 41,3 roku?

V této situaci máme $\mu = 41,3$, $n = 20$, $\bar{x} = 44,1$ a $s = 9,6$. Rozdělení veku je normální, takže můžeme použít naše tvrzení i přesto, že rozsah výběru je poměrně malý. Protože σ neznáme, platí $t = (\bar{x} - \mu)/(s/\sqrt{n})$.

Vypočítáme $P(\bar{X} > 44,1)$: $t = (44,1 - 41,3)/(9,6/\sqrt{20})$ s $(20 - 1 = 19)$ stupni volnosti, $t = 1,3044$. Pomocí podrobné tabulky t -rozdělení nebo pomocí vhodného programu (v programu Excel použijeme funkci TDIST) zjistíme, že k této hodnotě přísluší přibližně pravděpodobnost 0,1038. Proto lze tvrdit, že výběr o rozsahu 20 bude mít průměr 44,1 roku nebo více s pravděpodobností 0,1038.

4.6.3 Výběrové rozdělení relativní četnosti

Jestliže sledujeme výskyt náhodného jevu A v n opakovaných nezávislých pokusech, odhadujeme jeho pravděpodobnost $P(A)$ relativní četností m/n , kde m byla pozorovaná četnost výskytu jevu A . Tato relativní četnost se liší od pravděpodobnosti $P(A)$ vlivem náhodné odchylky. Lze si ji proto představit jako náhodnou proměnnou. Zajímá nás, jak popsat její náhodné chování při pevném počtu pokusů.

S podobným problémem se setkáváme, když cílem nějakého statistického řešení je určit relativní část populace, jež má danou vlastnost A . Například jaká část $P(A)$ populace mužů nad 50 let má určitou chorobu? Jaká část dospělé populace $P(A)$ by volila určitého prezidentského kandidáta, pokud by volby byly přímé? Provedeme z populace prostý náhodný výběr o rozsahu n a zjistíme v něm relativní četnost vlastnosti A . Chceme vědět, jak se mohou zjištěné relativní četnosti v rámci jejich náhodného kolísání lišit od populační hodnoty.

V obou příkladech, které jsme popsali, můžeme přijmout stejný rámec pravděpodobnostních úvah. Na položenou otázku lze přesně odpovědět použitím výpočtů, jež vycházejí z binomického rozdělení (viz kap. 4.5.1). V mnoha případech však vystačíme s asymptotickým přiblížením pomocí normálního rozdělení, jehož princip jsme popsali v kapitole 4.5.5.

Jehož princip jsme předchozího článku vysvětlovali.

Jestliže máme větší výběr o rozsahu n z rozsáhlé populace s teoretickou relativní četností sledované vlastnosti p a vypočítáme výběrovou relativní četnost \hat{p} , pak tato statistika na základě platnosti centrálního limitního teorému má přibližně normální rozdělení s průměrem p a směrodatnou odchylkou $\sqrt{p(1-p)/n}$. Symbolicky toto tvrzení vyjádříme vztahem

$\hat{P} \sim N(p, p(1 - p)/n)$ (asymptoticky).

Poznamenejme, že za větší populaci považujeme takovou, která má počet prvků větší než $10n$. Uvedena aproximace platí, jestliže je splněna podmínka pro rozsah výběru, že $np > 10$ a $n(1-p) > 10$.

PŘÍKLAD 4.19

Aproximace binárního rozdělení normálním rozdělením

Podle celostátního šetření má 40 % z 50 tisíc domácností v dané oblasti barevnou televizi. Jaká je pravděpodobnost, že v prostém náhodném výběru domácností o rozsahu 100 bude vlastnit 45 nebo více domácností barevnou televizi?

V tomto případě $p = 0,4$, $n = 100$, $\hat{p} = 0,45$. Předpoklady použití uvedené aproximace jsou splněny, protože 50 tisíc domácností představuje dost rozsáhlou populaci vzhledem k výběru o $n = 100(50000 > 10 \times 100)$; rovněž $np = 100 \times 0,4 > 10$ a $n(1-p) = 100 \times 0,6 > 10$.

3. Proto můžeme tvrdit

$$\hat{P} \sim N(0.40; 0.40 \times (1 - 0.40)/100)$$

$$z = (0.45 - 0.40) / \sqrt{0.40(1 - 0.40)/100} = 1.02$$

Pomocí tabulky normální rozdělení nebo vhodného programu (v Excelu použijeme funkci NORMDIST) zjistíme $P(\hat{P} > 0,45) = 0,1537$. Proto s přibližnou pravděpodobností 0,15 ve výběru o rozsahu 100 domácností bude 45 domácností nebo více vlastnit barevnou televizi. Můžeme tvrdit, že výběrová relativní četnost 0,45 nebo více v tomto případě není málo očekávaným jevem.

4.6.4 Výběrové rozdělení rozdílů dvou průměrů a dvou relativních četností

Pokud se dvě náhodné proměnné vzájemně neovlivňují, nazýváme je nezávislé. Příkladem nezávislých náhodných proměnných jsou průměry sledované proměnné nebo relativní četnosti určité vlastnosti prvků výběru získané pomocí náhodného výběru ze dvou různých populací. V kapitole 4.3 jsme uvedli, jak spočítat průměr a směrodatná odchylka nebo rozptyl takových nezávislých náhodných proměnných. Tyto poznatky aplikujeme na případ rozdílů průměrů a relativních četností.

Označme získané průměry \bar{x}_1, \bar{x}_2 , teoretické průměry a rozptyly proměnných v populacích μ_1, μ_2 a σ_1^2, σ_2^2 , označme také získané, resp. teoretické relativní frekvence \hat{p}_1, \hat{p}_2 resp. p_1, p_2 . Pro průměry a rozptyly rozdilů výběrových průměrů, resp. výběrových relativních četností platí:

$$\begin{aligned} E(\bar{X}_1 - \bar{X}_2) &= \mu_1 - \mu_2, & \text{Var}(\bar{X}_1 - \bar{X}_2) &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}, \quad \text{resp.} \\ E(\hat{P}_1 - \hat{P}_2) &= p_1 - p_2, & \text{Var}(\hat{P}_1 - \hat{P}_2) &= \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}, \end{aligned}$$

kde n_1, n_2 jsou rozsahy obou výběrů.

Jestliže navíc předpokládáme větší rozsahy výběrů a působení centrálního limitního teorému, pak usuzujeme, že oba rozdíly mají přibližně normální rozdělení. Rozdělení rozdílů průměrů a relativních četností lze tedy vyjádřit vztahy

$$\begin{aligned} \bar{X}_1 - \bar{X}_2 &\sim N(\mu_1 - \mu_2; \sigma_1^2/n_1 + \sigma_2^2/n_2), \\ \hat{P}_1 - \hat{P}_2 &\sim N(p_1 - p_2; p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2). \end{aligned}$$

Pokud jsou rozsahy výběrů velké a neznámé rozptyly proměnných σ_1^2, σ_2^2 , lze za ně dosadit výběrové rozptyly. Při menším počtu pozorování, ale normálním rozdělení sledovaných proměnných uplatňujeme podobnou modifikaci jako pro rozdělení standardizovaného průměru výběrovou směrodatnou odchylkou $s_{\bar{x}}$, které popisujeme Studentovým t -rozdělením.

PŘÍKLAD 4.20

Výběrové rozdělení dvou četností

Ve velkém městě se sleduje průměrná známka žáků na vysvědčení v osmé třídě. V předchozích šetření bylo zjištěno, že 10% jak chlapců, tak 10% dívek má horší průměrnou známku než 3,0. V náhodném výběru 80 dívek, resp. 100 chlapců byly zjištěny hodnoty 15%, resp. 12% žáků s průměrem známk horším než 3,0.

Máme nalézt pravděpodobnost, že rozdíl relativních četností dívek a chlapců ve výběru bude 3% nebo větší za předpokladu, že platí údaje z předchozího šetření.

$$\begin{aligned} \hat{p}_1 &= 0,15; \quad \hat{p}_2 = 0,12, \\ p_1 &= 0,10; \quad p_2 = 0,10. \end{aligned}$$

$$\hat{P}_1 - \hat{P}_2 \sim N(0,10 - 0,10; 0,10 \times 0,90/80 + 0,10 \times 0,90/100) = N(0; 0,002025),$$

přičemž σ tohoto rozdělení je $\sqrt{0,002025} = 0,045$.

Máme nalézt pravděpodobnost $P(\hat{P}_1 - \hat{P}_2 > 0,03)$.

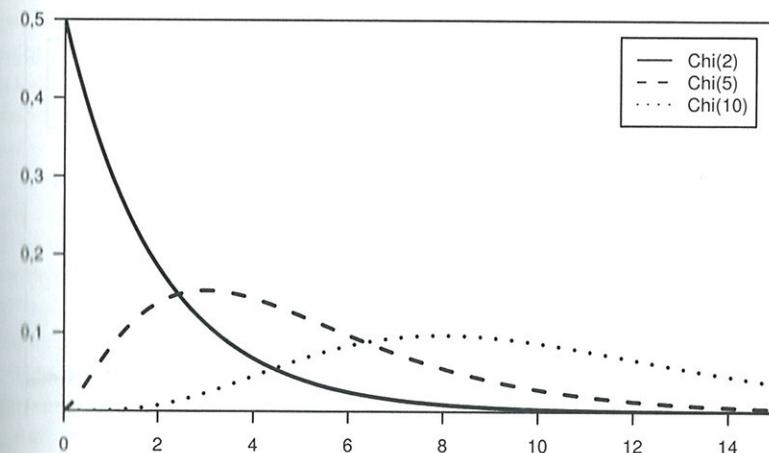
Standardizovaná hodnota $z = (0,03 - 0,0)/0,045 = 0,666$. Pomocí tabulek normálního rozdělení (tab. II přílohy B) určíme $F(z) = F(0,666) = 0,747$. Hledaná pravděpodobnost tedy hodnotu $1 - 0,747 = 0,253$.

4.6.5 Výběrové rozdělení rozptylu

Při zkoumání variability rozptylu a v mnoha dalších situacích se v pravděpodobnostních úvahách používá tzv. χ^2 -rozdělení (vyslovujeme *chi-kvadrát rozdělení*). Rozdělení χ^2 poprvé odvodil K. Pearson kolem roku 1900, kdy navrhl χ^2 -test dobré shody pro kategoriální data. Toto rozdělení má pouze jeden parametr, který nazýváme stupně volnosti. Pravděpodobnostní rozdělení χ^2 má statistika vytvořená z výběrového rozptylu, pokud náhodná proměnná, pro kterou se rozptyl počítá, má normální rozdělení. Pak platí, že hodnoty $(n-1)s^2/\sigma^2$ mají $\chi^2(n-1)$ rozdělení, tj. rozdělení χ^2 s $n-1$ stupňů volnosti. Parametr σ^2 je rozptyl náhodné proměnné a n počet pozorování. Na obrázku 4.16 jsou zobrazeny tři různé tvary tohoto rozdělení v závislosti na stupnících volnosti.

Pro lepší představu o χ^2 -rozdělení, jež má k stupňů volnosti, dodejme, že toto rozdělení v podstatě vzniká jako součet k nezávislých standardizovaných normálních proměnných, které jsou umocněny na druhou.

Obr. 4.16 Tvary χ^2 -rozdělení



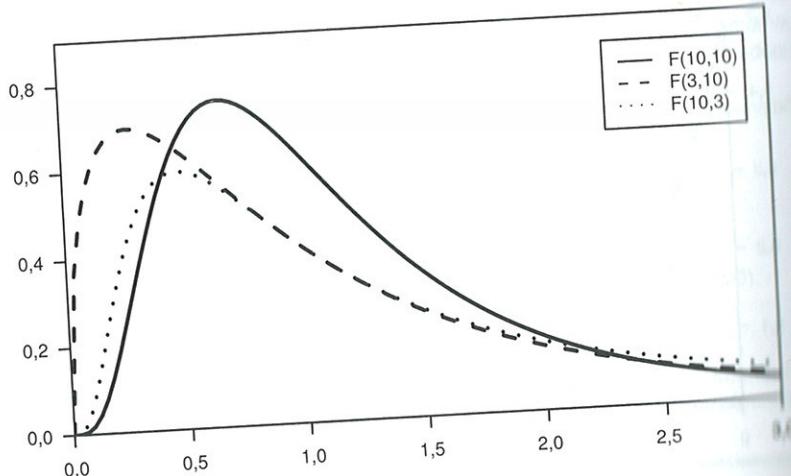
PŘÍKLAD 4.21

Použití rozdělení χ^2 pro studium chování výběrového rozptylu

Předpokládáme, že zkoumaná náhodná proměnná má rozptyl 4. Jaká je pravděpodobnost, že výběrový rozptyl, který spočítáme z 37 měření, bude větší než 5? Počítáme pravděpodobnost $P(s^2 > 5)$. Ta se rovná pravděpodobnosti $P(36s^2/4 > 36 \times 5/4) = P(\chi^2 > 45)$. Pomocí vhodného programu zjistíme, jaká hladina odpovídá kvantilu 45 rozdělení χ^2 o 36 stupních volnosti. V Excelu použijeme k potřebnému výpočtu funkce CHIDIST nebo CHIINV. Získáme hodnotu 0,85. To znamená, že přibližně pouze v 15 % případů lze očekávat, že hodnota výběrového rozptylu spočítaného z 37 měření bude větší než 5, jestliže skutečný rozptyl je 4.

4.6.6 Výběrové rozdělení poměru rozptylů

Také F -rozdělení, pojmenované po statistikovi R. A. Fisherovi, našlo uplatnění při popisu náhodného chování testovacích statistik. Základem je snaha popsat variabilitu poměru výběrových rozptylů, jež se vypočítaly z dat dvou nezávislých náhodných výběrů. Podobně jako u χ^2 -rozdělení musí být data normálně

Obr. 4.17 Tvary F -rozdělení

rozdělená. Navíc se předpokládá, že data pocházejí z populací, které mají stejný teoretický rozptyl. Pak platí, že hodnoty s_1^2/s_2^2 mají rozdělení F se stupni volnosti $n_1 - 1$ a $n_2 - 1$, kde n_1 a n_2 jsou rozsahy výběrů. F -rozdělení tedy závisí na dvou parametrech. Na obrázku 4.17 jsou zobrazeny tři různé tvary F -rozdělení v závislosti na stupních volnosti. Teoreticky je F -rozdělení s k_1 a k_2 stupni volnosti odvozeno jako podíl dvou nezávislých χ^2 -rozdělení s k_1 a k_2 stupni volnosti.

PŘÍKLAD 4.22

Aplikace F -rozdělení pro studium výběrového poměru rozptylů

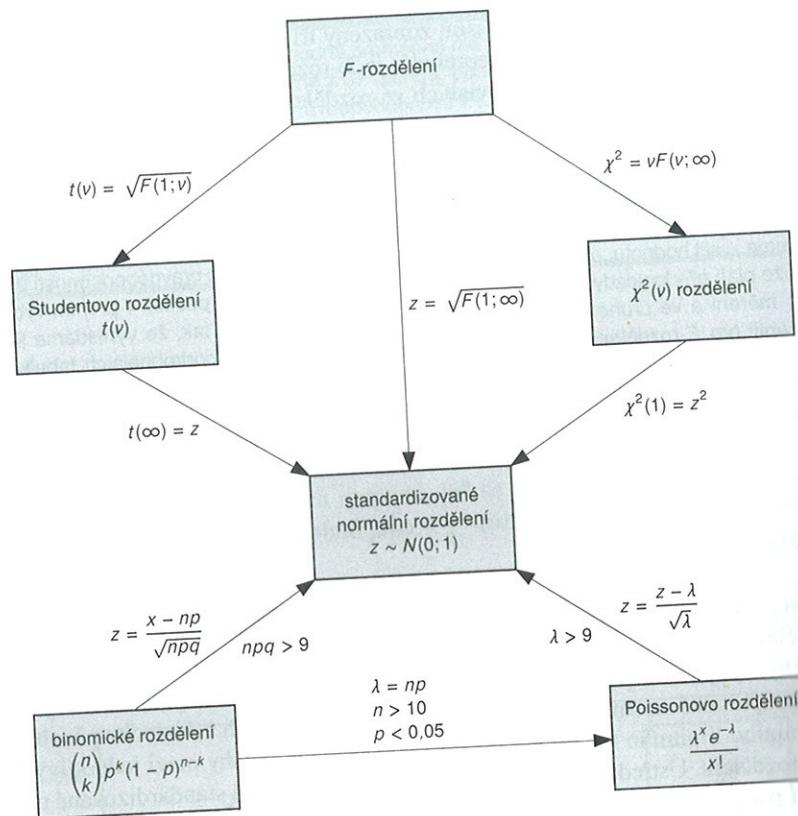
Chceme zjistit hodnotu, pod niž se bude nacházet poměr rozptylů s pravděpodobností 0,95, jestliže platí předpoklady pro konstrukci F -rozdělení. Rozptyly se počítají v prvním výběru z 21 měření a ve druhém výběru z 16 měření. Úlohu vyřešíme tak, že vyhledáme 95. percentil pro F -rozdělení s 20 a 15 stupni volnosti. Hledáme v podrobnějších tabulkách F -rozdělení (Likeš, Hebká, 1978). V Excelu použijeme k potřebnému výpočtu funkci FDIST a nalezneme hodnotu 2,328.

Souhrn

Vysvětlili jsme základní pravidla počítání s pravděpodobnostmi a popsali typy rozdělení, které nejčastěji používáme při analýze dat pomocí statistických metod. Také byla objasněna podstata pojmu výběrové rozdělení statistiky. Poukázali jsme na asymptotické vlastnosti průměru náhodných proměnných v důsledku působení centrálního limitního teoremu. Obrázek 4.18 ukazuje vztahy mezi jednotlivými typy rozdělení. Ústřední postavení má normální rozdělení a standardizované normální rozdělení. Ostatní se k němu přiblížují s rostoucím počtem pozorování nebo stupňů volnosti. Počet stupňů volnosti označujeme na obrázku symbolem v . Vztah mezi rozdělením χ^2 a $N(0; 1)$ vyplývá z definice χ^2 -rozdělení. Obrázek ukazuje, že rozdělení $N(0; 1)$, Studentovo t a χ^2 jsou speciálními případy F -rozdělení. Některá další rozdělení mají vztah mezi sebou nezávisle na normálním rozdělení. Z obrázku je patrné, že v důsledku podobnosti rozdělení lze jistě splnit určitých předpokladů dělat pravděpodobnostní úsudky o hodnotách výběrových statistik na základě standardizovaného normálního rozdělení, jestliže máme parametry μ , λ nebo p . Přehledně je tento poznatek vyjádřen v tabulce 4.9. Vybrané approximace rozdělení pravděpodobnosti uvádí tabulka 4.10.

Více o teorii pravděpodobnosti nalezneme v příslušných publikacích (např. Hebká, 1978 nebo Zvára, 1997).

Obr. 4.18 Vztahy mezi typy rozdělení



Tab. 4.9 Situace využití standardizovaného normálního rozdělení

Parametr	Výběrová statistika	Náhodná proměnná	Podmínka	Rozdělení
μ	\bar{x}	\bar{X}	normální rozdělení, také σ^2 známé	$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0; 1)$
μ	\bar{x}	\bar{X}	libovolné rozdělení, σ^2 známé, n velké	$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0; 1)$ (asympt.)
μ	\bar{x}	\bar{X}	libovolné rozdělení, σ^2 neznámé, n velké	$\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim N(0; 1)$ (asympt.)
λ	x	X	Poissonovo rozdělení, λ velké	$\frac{X - \lambda}{\sqrt{\lambda}} \sim N(0; 1)$ (asympt.)
p	\hat{p}	\hat{P}	binomické rozdělení, n velké	$\frac{\hat{P} - p}{\sqrt{p(1-p)/n}} \sim N(0; 1)$ (asympt.)

Tab. 4.10 Vybrané approximace rozdělení pravděpodobnosti

Aproximované rozdělení	Podmínky	Aproximující rozdělení
binomické rozdělení $X \sim B(p; n)$	$np > 10$ $n(p-1) > 10$	$P(X = x) \doteq P(x - 0,5 < W < x + 0,5)$, kde $W \sim N(np; npq)$ a $q = 1 - p$
Poissonovo rozdělení s parametrem λ	$\lambda > 25$	$P(X = x) \doteq P(x - 0,5 < W < x + 0,5)$, kde $W \sim N(\lambda; \lambda)$
binomické rozdělení $X \sim B(p; n)$	velké n malé p	Poissonovo rozdělení s parametrem $\lambda = np$