Kapitola 1

Základy teorie pravděpodobnosti

1.1 Náhodné jevy, pravděpodobnost

1.1.1 Náhoda, náhodný jev

„Život je jen náhoda“, jak se zpívá v jedné oblíbené písni [1]. S trochou zobecnění můžeme tuto větu považovat za jakousi „definici“ náhody. Pojem *náhody* patří do filosofických kategorií a dlouhá léta se filosofové přou o to, co to vlastně ta *náhoda* je. Je to to, co nelze poznat, nebo je to to, co poznat lze, ale my to neumíme, nebo její poznání je tak složité a nákladné, že se o něj raději ani nepokoušíme? v této souvislosti vzpomínám ještě na jednu „definici“, kterou vyslovil kriminální rada Vacátko v jednom též oblíbeném televizním seriálu: „Náhoda je Bůh, pánové“[2]. Jisté je, že náhoda je všude kolem nás a ovlivňuje náš život od samého počátku.

Pro naše potřeby budeme vycházet spíše z pojmu *náhodného pokusu*. Ten je definovaný zcela jasně a srozumitelně: **náhodný pokus** je jakákoli naše činnost, která skončí nějakým výsledkem z předem známé množiny všech možných výsledků, ale jejíž konkrétní výsledek nemůžeme z nějakých důvodů před zahájením pokusu s jistotou určit.

Příklad 1.1.1 *Takovým pokusem je třeba sledování počasí, konkrétně teplota. Dnes nelze s jistotou říci, jaká bude teplota zítra ráno v 8.00. Víme, že „výsledky“ tohoto pokusu mohou být nějaká reálná čísla (teplota ve stupních), ale konkrétní výsledek se dozvíme až po „uskutečnění“ pokusu, tedy zítra ráno v osm hodin na základě měření.*

Příklad 1.1.2 *Při cestě na poštovní úřad se můžeme pouze domnívat, kolik lidí bude ve frontě u přepážky pro výdej balíků. Jisté je, že to bude nějaké přirozené číslo, možná i nula. Teprve pozorováním na místě můžeme určit výsledek.*

Příklad 1.1.3 *Před koupí nového mobilního telefonu jsme vystaveni otázce: bude fungovat, nebo ne? Máme pouze dva možné výsledky: ano či ne. Přesto konkrétní výsledek neznáme, dokud přístroj nevyzkoušíme.*

Takto bychom mohli pokračovat donekonečna[3]. Příkladů je kolem nás nespočet. A co je na tom zajímavé? Když budeme stejný pokus opakovat za „přibližně stejných“ podmínek, nejistota výsledku zůstává. Ani velký počet opakování nám nedává jistotu konečného výsledku. To, co můžeme

[1] Jan Werich, Jiří Voskovec, Jaroslav Ježek, 1932

[2] Hříšní lidé města Pražského, režie Jiří Sequens, podle povídek Jiřího Marka,1968

[3] Donekonečna samozřejmě ne – to je nadsázka. Ale hodně dlouho.

mnohonásobným opakováním stejného pokusu získat, je pouze míra našeho očekávání, že nastane některý z často se opakujících výsledků. No a této *míře očekávání* budeme říkat *pravděpodobnost*.

Protějškem k *náhodnému* pokusu je pokus *deterministický*. v takovém případě je výsledek dopředu vždy přesně znám. v praktickém životě však tento typ pokusu existuje pouze v teoretické rovině, kdy abstrahujeme od řady vlivů a uvažujeme pouze jakési *ideální* podmínky.

Tedy vrátíme-li se na začátek tohoto úvodu a použijeme-li ještě jednu lidovou moudrost, můžeme konstatovat, že jediná jistota v životě je to, že všechno je nejisté. Proto je užitečné rozvíjet takovéto teorie, jako je teorie pravděpodobnosti, které nám pomohou zorientovat se v té nejistotě kolem nás a ukáží nám, co je jak pravděpodobné.

Čtenář může namítnout, že přestože je nejistota všude kolem nás, přesto v tomto světě žijeme, aniž bychom k tomu potřebovali teorii pravděpodobnosti. To je pravda, ale pouze z části. Každý z nás v životě pracuje s jakousi „subjektivní“ pravděpodobností, která nám dovoluje činit rozhodnutí aniž bychom si to uvědomovali. Problémy spojené s aplikací této subjektivní pravděpodobnosti pociťuje každý, kdo se má rozhodnout mezi několika možnostmi. Snažíme se „upřesnit“ tuto pravděpodobnost (rozuměj: míru očekávání výsledku) studiem, cvičením, získáváním dodatečných informací a podobně. Subjektivní pravděpodobnost má jednu velkou nevýhodu: je subjektivní. To znamená, že když se sejdou dva lidé, zpravidla mají na stejnou věc dva různé názory – dvě různé hodnoty subjektivní pravděpodobnosti možných výsledků.

Teorie pravěpodobnosti je matematická disciplína, která pracuje s „objektivní“ pravděpodobností. Objektivní pravděpodobnost je vyjádřena matematickou funkcí, vycházející z matematické logiky a objektivně uznatelnou všemi, kterých se týká.

Příklad 1.1.4 *Představme si známý příklad házení hrací kostkou. Subjektivní pocit, že „v následujícím hodu už mi určitě padne šestka“, vyjadřující vysokou subjektivní pravděpodobnost toho, že „musím vyhrát“ zde stojí proti objektivní zkušenosti, že šestka je pouze jedním ze šesti stejně možných výsledků a objektivně není důvod se domnívat, že by se zrovna v našem případě měla stát výjimka.*

Příklad 1.1.5 *Ze šesti přístrojů, mezi nimiž jsou dva vadné, vybereme náhodně jeden. Jaká je pravděpodobnost, že vybereme bezvadný přístroj?*

Řešení: Pokud vybíráme náhodně, můžeme předpokládat, že každý přístroj má stejnou možnost být vybrán a tedy objektivně v jedné třetině možných

výsledků (výběru konkrétního přístroje) vybereme vadný, ve dvou třetinách bezvadný přístroj. Očekávaná pravděpodobnost tedy bude rovna dvěma třetinám.

Předpoklad existence objektivní pravděpodobnosti představuje ovšem určité zjednodušení, abstrakci. v předchozím příkladu předpokládáme, že každý přístroj má „stejnou šanci“ být vybrán. To samo o sobě je idealizace, zvláště v případě, že přístroje nejsou na jednom místě, nejsou vzhledově úplně totožné, existují určité vjemy, které nás při výběru ovlivňují. Nemluvě o situaci, kdy vybíráme z většího počtu umístěného například v nějakém kontejneru. Potom ty kusy, které jsou na špatně přístupném místě mají reálně menší šanci být vybrány než ty, které jsou dostupnější. Nicméně, jisté zobecnění a zjednodušení je nezbytné k tomu, abychom mohli vytvářet matematické modely reality a na základě těchto modelů provádět další úvahy.

Poznámka: v těchto skriptech je uvedena řada různých příkladů. Všechny tyto příklady popisují jakési „modelové“ situace. To znamená, že není důležité, zda se v nich hovoří o hrací kostce nebo o vadných přístrojích. Například výše uvedené dva příklady by stejně dobře mohly znít takto:

Příklad 1.1.6 *Představme si šest stejných přístrojů, mezi nimiž je jeden vadný. Subjektivní pocit, že „při náhodném výběru jednoho z nich vyberu určitě ten vadný“, vyjadřující vysokou subjektivní pravděpodobnost plynoucí ze znalosti Murphyho zákonů*[4] *zde stojí proti objektivní zkušenosti, že výběr vadného přístroje je pouze jedním ze šesti stejně možných výsledků.*

Příklad 1.1.7 *Jaká je pravděpodobnost, že při hodu hrací kostkou nám padne číslo dělitelné třemi?*

Tyto příklady, ač jsou formulovány jinak než předchozí dva, mají úplně stejné řešení. Proto se nepozastavujte nad tím, když v některých příkladech „taháme koule z urny“, ale spíše se pokuste najít paralelu s reálným životem, formulaci téže úlohy v termínech mnohem bližších reálnému životu. Pokud se vám to podaří, začneme chápat oč tu běží.

1.1.2 Speciální definice pravděpodobnosti

Jak bylo uvedeno v předchozím odstavci, východiskem pro naše další úvahy bude vždy nějaký **náhodný pokus**. Ten vždy skončí nějakým konkrétním

[4] http://www.cska.cz/murphy/index.htm

výsledkem z předem známé množiny Ω. Tyto výsledky jsou navzájem *neslučitelné*, což znamená, že vždy může nastat pouze jeden z nich, nikoli dva zároveň. Takovéto výsledky nazýváme **elementární výsledky**.

Příklad 1.1.8

- *Při měření teploty chladící kapaliny je množinou elementárních výsledků množina reálných čísel.*
- *Při sledování počtu poruch zařízení za určitou dobu je množinou elementárních výsledků množina nezáporných celých čísel.*
- *Výsledkem posuzování kvality výrobku na výstupní kontrole je jedna ze tří jakostních kategorií.*
- *Pozorovaný výsledek testu odolnosti povrchu proti otěru může být buď obstál nebo neobstál.*

Kromě těchto elementárních výsledků jsou předmětem našeho studia často výsledky komplikovanější, složené z několika elementárních výsledků. Tyto výsledky už nemusejí být neslučitelné a mohou se vyskytnout současně. Zpravidla je můžeme vyjádřit jako sjednocení elementárních jevů. v uvedených příkladech to mohou být následující výsledky:

Příklad 1.1.9

- *Při měření teploty chladící kapaliny sledujeme, zda je teplota v určitém povoleném rozmezí. To je vyjádřeno intervalem, obsahujícím nekonečně mnoho reálných čísel.*
- *Při sledování počtu poruch zařízení za určitou dobu nás zajímá, zda tento počet nepřekročí povolený limit, řekněme k. Tento výsledek obsahuje čísla $\{0, 1, 2, \ldots, k\}$.*

To nás vede k následující definici:

Náhodné jevy (dále jen **jevy**) jsou množiny, jejichž prvky jsou elementární jevy z množiny Ω. Systém všech jevů budeme nazývat **jevové pole** $\mathcal{F}$. Jev Ω se nazývá **jistý jev**, $\emptyset$ prázdná množina elementárních jevů se nazývá **nemožný jev**.

Při práci s jevy se užívají množinové operace s následujícím významem: pro jevy $A, B \in \mathcal{F}$ lze definovat

- sjednocení $A \cup B$ (*nastane jev A nebo jev B*),

- průnik $A \cap B$ (*jevy A a B nastanou současně*)[5],
- doplněk A^c (*nenastane jev A*),
- rozdíl $A - B = AB^c$ (*nastane jev A, ale nenastane jev B*),
- inkluzi $A \subset B$ (*když nastane jev A, nastane i jev B*)
- disjunkci $A \cap B = \emptyset$ (*jevy A a B nemohou nastat současně - jsou* ***neslučitelné***)

V této souvislosti jsou užitečná takzvaná **de Morganova pravidla**, která platí i pro sjednocení či průnik většího počtu jevů:

$$(A \cup B)^c = A^c \cap B^c, \qquad (A \cap B)^c = A^c \cup B^c.$$

Klasická definice pravděpodobnosti. Uvažujme náhodný pokus s konečnou množinou elementárních jevů $\Omega = \{\omega_1, ..., \omega_n\}$ a nějaký jev $A \subset \Omega$. Za předpokladu, že všechny elementární výsledky pokusu jsou stejně možné, pravděpodobnost jevu A je definována jako podíl

$$P(A) = \frac{n_A}{n},$$

kde n_A je počet elementárních jevů, při nichž nastane jev A a n je počet všech elementárních jevů.

Při výpočtu pravděpodobnosti podle klasické definice je třeba znát počty n_A a n. K tomu se často používá kombinatorických úvah. Tak například $C_k(n) = \binom{n}{k}$ je počet **kombinací k-té třídy z n prvků**. Předpokládáme, že pro $k \leq n$. $C_k(n)$ odpovídá počtu k-tic, které lze vybrat z n prvků bez ohledu na pořadí. Pokud budeme rozlišovat i pořadí prvků ve vybíraných k-ticích, pužijeme **variace k-té třídy z n prvků**$V_k(n) = n(n-1)...(n-k\ +1)$. Specielně pro $k = n$ je $P_e(n) = V_n(n) = n!$ počet **permutací n prvků**, což představuje počet všech možných pořadí při uspořádání n prvků.

Příklad 1.1.10 *Deset aut zaparkuje náhodně vedle sebe. Jaká je pravděpodobnost, že zvolená tři auta budou spolu sousedit?*

[5]Často se můžete setkat se zápisem, v němž je průnik dvou jevů zapsán jako součin (bez symbolu $\cap$), tedy $A \cap B = AB$.

Řešení: Označme A sledovaný jev. Celkový počet možných rozmístění tří aut na 10 míst je roven $V_3(10) = 720$. Počet rozmístění, která vyhovují jevu A je osm možných pozic krát počet permutací tří aut v jedné pozici, tedy celkem $P(A) = \frac{8.3!}{720} = 0,0667$.

Geometrická pravděpodobnost. Nejjednodušší příklad definice pravděpodobnosti na nespočetné množině elementárních jevů pro $\Omega \subset R^k$ (R^k je k-rozměrný Eukleidovský prostor) a jev $A \subset \Omega$ vychází ze znalosti k-rozměrného objemu $V_k(A)$ a klade

$$P(A) = \frac{V_k(A)}{V_k(\Omega)}.$$

Všimněte si, že při této definici pravděpodobnosti a $k \geq 1$ mají jednotlivé elementární jevy pravděpodobnost 0, jsou to totiž body prostoru R^k, které mají nulový objem.

Příklad 1.1.11 *Na rovinu R^2 pokrytou systémem ekvidistantních rovnoběžných přímek vzdálených od sebe o $d > 0$ je náhodně vhozen*

***a**) kruh o průměru r,*

***b**) úsečka o délce $r < d$.*

Stanovte pravděpodobnost jevu $A =$ [útvar protne jednu z přímek].

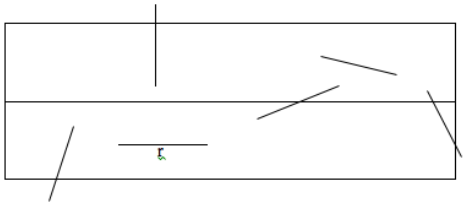
Řešení:

1. Poloha kruhu je určena (x, y), jev A nezávisí na souřadnici x. Stačí zvolit $\Omega = \langle 0, d/2\rangle \subset R^1$ interval vyjadřující vzdálenost středu od nejbližší přímky, v tomto modelu $A = \langle 0, r/2\rangle$ a tedy $P(A) = \frac{r}{d}$.

2. Zde je nutno navíc uvažovat orientaci φ úsečky z intervalu $\langle 0, \pi\rangle$, tedy $\Omega = \langle 0, \pi\rangle \, x \, \langle 0, d/2\rangle \subset R^2$. Jev A vyjádříme z podmínky protnutí jako $A = \left\{(\varphi, y) \in \Omega; y \leq \frac{r}{2} sin\varphi\right\}$. Protože $\int_0^\pi = \frac{r}{2} sin\varphi d\varphi = r$, je podle definice $P(A) = \frac{2r}{d\pi}$.

Předchozí příklad b) je známý Buffonův problém házení jehlou, formulovaný v roce 1777. Je možná četnostní interpretace výsledku. Na obdélníkovou oblast o stranách 1 a d je náhodně vhozeno n úseček délky r, n_p je počet úseček, které protnou přímku půlící obdélník, viz Obr. 1.1:

Nahradíme-li pravděpodobnost poměrnou četností, dostáváme vztah $\frac{n_p}{n} \approx \frac{2r}{d\pi}$, tedy $\frac{\pi}{2} n_p \approx \frac{rn}{d}$. Zde n_p je počet průsečíků na jednotku délky přímky, výraz



Obrázek 1.1: Buffonova úloha

na pravé straně je celková délka úseček v jednotce plochy. Poslední vzorec lze obecněji použít pro odhad délky vláken s náhodnou orientací (např. hustoty dislokačních čar na metalografickém snímku) pomocí počtu průsečíků na testovací přímce.

Statistická definice pravděpodobnosti. V řadě případů nelze použít ani jednu z předchozích definic pravděpodobnosti. V takových případech provedeme statistický pokus, při kterém mnohokrát opakujeme náhodný pokus za (přibližně) stejných podmínek a pravděpodobnost odhadneme poměrnou četností výskytu jevu A v této sérii pokusů. Tento způsob určení pravděpodobnosti je sice nejméně přesný, při různých sériích pokusů dostaneme dokonce různé výsledky a nelze rozhodnout, který z nich je ten správný, ale na druhou stranu, je to nejčastěji používaný způsob odhadu pravděpodobnosti v reálných aplikacích.

1.1.3 Axiomatická definice pravděpodobnosti

Obecná teorie pravděpodobnosti, která zahrnuje výše uvedené výklady pravděpodobnosti, vychází z následujících předpokladů:

1. Je dána neprázdná množina Ω, **prostor elementárních jevů**.
2. Je dáno **jevové pole** $\mathcal{F}$ podmnožin Ω splňující podmínky
   (a) nemožný jev $\emptyset \in \mathcal{F}$,
   (b) jestliže $A \in \mathcal{F}$, potom $A^c \in \mathcal{F}$,
   (c) jestliže $A_i \in \mathcal{F}, i \in N$, potom $U_{i \in N} A_i \in \mathcal{F}$
3. Každému jevu $A \in \mathcal{F}$ je přiřazena **pravděpodobnost** $P(A)$ s vlastnostmi
   (a) $P(A) \geq 0$ pro každé $A \in \mathcal{F}$
   (b) $P(\Omega) = 1$
   (c) $P(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty}(A_n)$ pro každou posloupnost $\{A_n\}$ po dvou neslučitelných jevů.

Trojice $(\Omega, \mathcal{F}, P)$ se nazývá **pravděpodobnostní prostor**. Pravděpodobnostní prostor je vlastně matematickým modelem náhodného pokusu. Uvedený soubor axiomů se nazývá podle jejich autora **Kolmogorovova axiomatická definice pravděpodobnosti**[6]

Je-li Ω konečná nebo spočetná množina, potom obyčejně $\mathcal{F}$ je systém všech podmnožin Ω. V případě nespočetného Ω je $\mathcal{F}$ nějaký dostatečně bohatý systém jevů splňující podmínku b) z předchozího odstavce, který nemusí obsahovat příliš komplikované podmnožiny Ω. Reálné funkci P na $\mathcal{F}$ se říká **pravděpodobnostní míra**, zkráceně pravděpodobnost.

Z výše uvedených axiomů lze odvodit následující vlastnosti pravděpodobnosti:

(**1**) $P(\emptyset) = 0$

(**2**) $0 \leq P(A) \leq 1$ pro každé $A \in \mathcal{F}$,

(**3**) jestliže $A \subset B$, potom $P(A) \leq P(B)$,

[6] A.N.Kolmogorov (1903 – 1987), sovětský matematik, zakladatel moderní teorie pravděpodobnosti.

(**4**) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$,

(**5**) $P(A \cap B) = P(A) + P(B)$ pro A, B neslučitelné,

(**6**) $P(A_c) = 1 - P(A)$,

(**7**) jestliže $A \subset B$, potom $P(B - A) = P(B) - P(A)$.

Vzorec pro pravděpodobnost sjednocení jevů lze indukcí rozšířit na libovolný počet sčítanců

$$P(\bigcup_{i=1}^{n} A_i) = \sum_{1\le i\le n} P(A_i) - \sum_{1\le i<j\le n} P(A_iA_j) + \sum_{1\le i<j<k\ \le n} P(A_iA_jA_k) - \ldots + (-1)^{n-1}P(A_1A_2\ldots A_n)$$

Příklad 1.1.12 *Při vyslání n zpráv určených po jedné n příjemcům dojde k chaosu a zprávy jsou přijaty náhodně. Jaká je pravděpodobnost jevu A = [aspoň jeden příjemce dostane svou zprávu], vyšetřete i limitní chování pro n rostoucí nade všechny meze.*

Řešení: Elementární jevy jsou všechny permutace, jejichž počet je $n!$. Označme $A_i = [i\text{-tý příjemce dostane svou zprávu}]$. Zřejmě je $A = \bigcup_{i=1}^{n} A_i$. Jevy A_i nejsou neslučitelné a tak nelze použít axiom 3c) z definice pravděpodobnosti. Užitím výše uvedeného vzorce dostáváme

$$P(A) = P(\bigcup_{i=1}^{n} A_i) = n\frac{(n-1)!}{n!} - \binom{n}{2}\frac{(n-2)!}{n!} + \binom{n}{3}\frac{(n-3)!}{n!} - \ldots + (-1)^{n-1}\frac{1}{n!} = 1 - \frac{1}{2!} + \frac{1}{3!} - \ldots + (-1)^{n-1}\frac{1}{n!}$$

dále je

$$\lim_{n\to\infty} P(\bigcup_{i=1}^{n} A_i) \;=\; \sum_{i=1}^{\infty}\frac{(-1)^{i-1}}{i!} \;=\; 1 - \sum_{i=0}^{\infty}\frac{(-1)^{i}}{i!} \;=\; 1 - e^{-1} \;=\; 0,628.$$

1.1.4 Podmíněná pravděpodobnost a stochastická závislost jevů

V některých případech jsou dva výsledky náhodného pokusu takové, že míra očekávání jednoho z nich, řekněme A, se změní, víme-li že nastal jev B. Tato pozměněná míra očekávání se nazývá **podmíněná pravděpodobnost** a o jevech A a B říkáme, že jsou **stochasticky závislé**. Podmíněná pravděpodobnost se spočte podle následující definice:
Je-li $A, B \in \mathcal{F}$, $P(B) > 0$, potom podmíněná pravděpodobnost $P(A|B)$ jevu A za podmínky, že nastal jev B je rovna

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

Z této definice dostáváme další zajímavý výsledek. Všimněte si, že z axiomů definice pravděpodobnosti nelze odvodit vztah pro pravděpodobnost průniku dvou jevů, ačkoli se často používá. Z definice podmíněné pravděpodobnosti vyplývá, že

$$P(A \cap B) = P(B).P(A|B).$$

Pokud jsou výsledky pokusu A a B takové, že míra očekávání jednoho z nich, řekněme A, se *nezmění* získáním informace o tom, že nastal jev B, potom říkáme, že jevy A a jsou **stochasticky nezávislé**. V tomto případě je $P(A|B) = P(A)$

Věta 1.1 (Kritérium nezávislosti) *Náhodné jevy $A, B \in \mathcal{F}$ jsou stochasticky nezávislé právě když platí $P(A \cap B) = P(A).P(B)$.*

Důkaz: Důkaz plyne bezprostředně ze vzorce pro podmíněnou pravděpodobnost a z toho, že stochastická nezávislost je ekvivalentní vlastnosti $P(A|B) = P(A)$.

Všimněte si, že jsou-li stochasticky nezávislé jevy A, B, potom jsou nezávislé i jevy A a B^c. Je totiž

$$\begin{aligned} P(AB^c) = P(A - B) = P(A - AB) = P(A) - P(AB) \\ = P(A) - P(A)P(B) = P(A)(1 - P(B)) = P(A)P(B^c). \end{aligned}$$

Pravděpodobnost, že padne rub při hodu mincí (jev A), považujeme za stejnou jako pravděpodobnost, že padne rub za podmínky, že házíme levou rukou (jev B), neboť výsledek pokusu nezávisí na tom, kterou rukou se hází. Naproti tomu v následujícím příkladu tato rovnost splněna nebyla a z výsledku vyplývá, že se lze domnívat, že nezaměstnanost mírně závisí na pohlaví.

Příklad 1.1.13 *Při vyšetřování nezaměstnanosti v populaci pracovních sil byly zjištěny poměrné četnosti v tabulce. Můžeme je interpretovat jako pravděpodobnosti, že náhodně vybraný jedinec patří k dané skupině. Vypočtěte $P(N)$, $P(N|M)$, $P(N|Z)$ při označení jevů podle tabulky.*

	M (muž)	*Z (žena)*	*celkem*
Z (zaměstnaný)	*0,519*	*0,409*	*0,928*
N (nezaměstnaný)	*0,039*	*0,033*	*0,072*
celkem	*0,558*	*0,442*	*1,000*

Řešení: Je $P(N) = 0,072$, $P(N|M) = \frac{0,039}{0,558} = 0,07$, $P(N|Z) = 0,075$

Příklad 1.1.14 *Nezávisle na sobě jsou vyrobeny tři ventilátory, přitom pravděpodobnost, že výrobek je vadný (V) resp. dobrý (D), je vždy 0,1 resp. 0,9. Jaké jsou pravděpodobnosti možných výsledků tohoto pokusu?*

Řešení: Lze uvažovat osm možných výsledků ω_i tohoto pokusu (variace třetí třídy ze dvou prvků s opakováním). Označme $A_i = [i\text{-tý ventilátor je vadný}]$, potom např. elementární jev $\omega_3 = [DVD]$, lze vyjádřit jako $\omega_3 = A_1^c A_2 A_3^c$. Vzhledem k nezávislosti výrobků je potom $P(\omega_3) = 0,9^2 0,1$.

Příklad 1.1.15 *Uvažujme dva náhodné jevy A a B, jejichž pravděpodobnosti jsou $P(A) = P(B) = 0,5$. Spočtěte pravděpodobnosti $P(A \cup B), P(A - B)$ za předpokladu, že*

a) A a B jsou nezávislé jevy,

b) A a B jsou stochasticky závislé a je $P(A|B) = 0,8$.

Řešení: V obou případech je $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ a $P(A - B) = P(A) - P(A \cap B)$. V případě a) plyne z nezávislosti $P(A \cap B) = P(A).P(B) = 0,25$. V případě b) je $P(A \cap B) = P(B).P(A|B) = 0,4$. Tedy odpověď je

a) $P(A \cup B) = 0,75,\ P(A - B) = 0,25$

b) $P(A \cup B) = 0,6,\ P(A - B) = 0,1$

Příklad 1.1.16 *Dva hráči hází střídavě mincí, vyhraje ten, komu padne první rub mince. Určete pravděpodobnost jevu $A =$* [vyhraje hráč, který začíná].

Řešení: Elementární jevy v tomto pokusu jsou posloupnosti rubů (R) a líců (L) tvaru $\omega_1 = \{R\}, \omega_2 = \{LR\}, \omega_3 = \{LLR\}, \ldots, \omega_n = \{L \ldots LR\}, \ldots$ Na každý z těchto jevů se však můžeme dívat z hlediska konkrétního hodu mincí,

jehož elementární výsledky jsou pouze L nebo R, každý s pravděpodobností $\frac{1}{2}$ (za předpokladu symetrické homogenní mince). Potom výsledky sledovaného pokusu (celé hry) lze chápat jako průniky nezávislých jevů L a R a tedy platí, že $P(\omega_i) = p_i = \left(\frac{1}{2}\right)^i = 2^{-i}$. Všimněte si, že $\sum_{i=1}^{\infty} p_i = 1$ a jev ω_∞ = [hra nikdy neskončí] má pravděpodobnost $P(\omega_\infty) = 0$. Jev A obsahuje pouze ty elementární výsledky, jejichž délka je lichá (rub padne při $1., 3., 5., \ldots$ hodu). Tedy je $A = \bigcup_{i=1}^{\infty} \omega_{2i-1}$ a $P(A) = \sum_{i=1}^{\infty} \frac{1}{2_{2i-1}} = 2\sum_{i=1}^{\infty} 4^{-i} = \frac{2}{3}$. Takováto hra je nespravedlivá, protože hráči nemají stejnou pravděpodobnost výhry.

Vraťme se k příkladu 1.1.12 a pokusme se řešit ho pomocí obratu z předchozího příkladu. Lze psát $A = \bigcup_{i=1}^{n} A_i$, kde A_i = [i-tý příjemce dostal svou zprávu], tedy $A^c = \bigcap_{i=1}^{n} A_i^c$.

Pravděpodobnost tohoto průniku ovšem nelze jednoduše vyjádřit, protože jevy A_i, tedy ani jevy A_i^c, nejsou nezávislé. Je totiž $P(A_i) = \frac{1}{n}, P(A_i)P(A_j) = \frac{1}{n^2}$, ale $P(A_iA_j) = \frac{(n-2)!}{n!} = \frac{1}{n(n-1)}$.

Věta 1.2 *Nechť $A_1, \ldots, A_n \in \mathcal{F}, P(A_1 \cap A_2 \cap \ldots \cap A_n) > 0, n \geq 2$. Potom* $P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \ldots P(A_n|A_1 \cap \ldots \cap A_{n-1})$

Důkaz: Důkaz se provede matematickou indukcí: Pro $n = 2$ je z definice podmíněné pravděpodobnosti $P(A_1 \cap A_2) = P(A_1)P(A_2|A_1)$. Dále je $P(A_1 \cap \ldots \cap A_{n+1}) = P(A_1 \cap \ldots \cap A_n)P(A_{n+1}|A_1 \cap \ldots \cap A_n)$, dosazením za $P(A_1 \cap \ldots \cap A_n)$ z indukčního předpokladu je důkaz ukončen.

Příklad 1.1.17 *Ze sedmi dodaných televizorů tři potřebují seřízení. Náhodně vybereme tři kusy, jaká je pravděpodobnost jevu A = [žádný vybraný televizor nepotřebuje seřízení].*

Řešení: Úlohu lze řešit buď jako klasický pravděpodobnostní pokus, kde počet elementárních jevů je $C_3(7) = \binom{7}{3} = 35$,

$$P(A) = \frac{\binom{3}{0}\binom{4}{3}}{35} = \frac{4}{35}$$

nebo pomocí věty z předchozího odstavce: buď A_i = [i-tý vybraný televizor nepotřebuje seřízení], $A = A_1A_2A_3$,

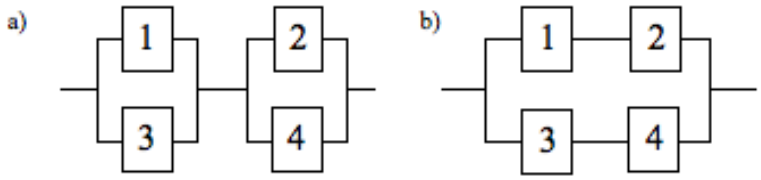
$$P(A) = P(A_1)P(A_2|A_1)P(A_3|A_1A_2) = \frac{4}{7}\frac{3}{6}\frac{2}{5} = \frac{4}{35}$$

Příklad 1.1.18 *Kolikrát je třeba hodit hrací kostkou, aby pravděpodobnost jevu A, že padne aspoň jedna šestka, byla větší než 0,9?*

Řešení: Buď $A_i = [\text{v } i\text{-tém hodu padne šestka}]$, potom $B_n = \bigcup_{i=1}^{n} A_i$ je jev, že šestka padne v některém z prvních n hodů a B_n^c jev, že šestka v prvních n hodech nepadne.

Podle de Morganových pravidel je $B_n^c = (\bigcup_{i=1}^{n} A_i)^c = \bigcap_{i=1}^{n} A_i^c$ což je průnik nezávislých jevů, tedy $P(B_n^c) = \prod_{i=1}^{n} P(A_i^c) = \left(\frac{5}{6}\right)^n$. Z nerovnosti $P(B_n) = 1 - \left(\frac{5}{6}\right)^n > 0,9$ dostáváme $n > \frac{ln0,1}{ln\left(\frac{5}{6}\right)} = 12,63$. Odpověď tedy zní, je třeba hodit alespoň třináckrát.

Příklad 1.1.19 *Vyšetřujme dva elektrické obvody se čtyřmi očíslovanými prvky, viz obr. 1.2. Předpokládáme, že jevy A_i = [dojde k poruše i-tého prvku] jsou nezávislé, $P(A_1) = P(A_2) = 0,1, P(A_3) = P(A_4) = 0,2$. Jaká je pravděpodobnost jevu B = [dojde k přerušení proudu v obvodu]?*



Obrázek 1.2: Zálohování po prvcích (a), zálohování celé větve (b).

Řešení: Je třeba kombinovat užití vzorců pro pravděpodobnost průniku a sjednocení jevů. V prvním obvodu a) je $B = A_1A_3 \bigcup A_2A_4, P(B) = 0,02 + 0,02 - 0,0004 = 0,0396$. Ve druhém obvodu b) je $B = (A_1 \bigcap A_2)(A_3 \bigcap A_4)$, jevy ve sjednocení nejsou neslučitelné, tedy $P(B) = (0,1+0,1-0,01)(0,2+0,2-0,04) = 0,0684$. Výsledek má následující interpretaci v teorii spolehlivosti: jsou-li 1,2 hlavní prvky a 3,4 záložní prvky v obvodu, potom protože $P(B)$ je nižší v obvodu a) potvrzuje se, že zálohování jednotlivých prvků je účinnější než zálohování celé hlavní větve v obvodu b).

Pro pevné $B \in \mathcal{F}$ je funkcí $P(.|B)$ definována na $\mathcal{F}$ opět pravděpodobnostní míra[7], přičemž základní prostor Ω se redukuje na B. Specielně buď $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}, B \subset \Omega, P(B) > 0$. V pojetí klasické definice pravděpodobnosti definujeme podmíněnou pravděpodobnost $P(\omega_i|B)$ přirozeně jako $P(\omega_i|B) =$

[7] tečka zde zastupuje místo pro proměnnou

0 pro $\omega_i \notin B$, $P(\omega_i|B) = \frac{P(\omega_i)}{P(B)}$ pro $\omega_i \in B$. V četnostní interpretaci označme n_i resp. n_B počet výskytů jevu ω_i resp. B v n opakováních pokusu. Potom je přibližně

$$P(\omega_i|B) \approx \frac{n_i}{n_B} = \frac{\frac{n_i}{n}}{\frac{n_B}{n}} \approx \frac{P(\omega_i)}{P(B)}$$

Potom, je-li $A \subset \Omega$, je

$$P(A|B) = \sum_{\omega_i \in A} P(\omega_i|B) = \sum_{\omega_i \in A \cap B} \frac{P(\omega_i)}{P(B)}.$$

Rozšíříme nyní definici nezávislosti na větší počet jevů. **Jevy systému $B = B_n, n \in M \subset \Omega$ se nazývají vzájemně nezávislé**, jestliže pro každou konečnou podmnožinu $n_1, \ldots, n_k \subset M$ indexové množiny M přirozených čísel platí

$$P(B_{n_1} B_{n_2} \ldots B_{n_k}) = P(B_{n_1}) P(B_{n_2}) \ldots P(B_{n_k}))$$

K nezávislosti systému jevů podle této definice nestačí nezávislost po dvou. Je-li např. $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}, P(\omega_i) = \frac{1}{4}, i = 1, \ldots, 4$. Jevy $A = \{\omega_1, \omega_2\}, B = \{\omega_1, \omega_3\}, C = \{\omega_1, \omega_4\}$ jsou po dvou nezávislé podle definice I.4.1, protože $P(A) = P(B) = P(C) = \frac{1}{2}$ a $P(AB) = P(AC) = P(BC) = \frac{1}{4}$. Naproti tomu $P(ABC) = \frac{1}{4}$ zatímco $P(A)P(B)P(C) = \frac{1}{8}$, tedy jevy nejsou vzájemně nezávislé.

1.1.5 Věta o úplné pravděpodobnosti a Bayesova věta

Vzájemně neslučitelné jevy $A_i \in \mathcal{F}, i = 1, \ldots, n$, tvoří **úplný systém jevů**, jestliže $P(A_i) > 0$ pro každé $i = 1, \ldots, n$ a $P(\bigcup_{i=1}^n A_i) = 1$. Někdy se též hovoří o **úplném pokrytí** množiny Ω.

Věta 1.3 (o úplné pravděpodobnosti) *Nechť $\{H_i\}_{i=1}^n$ je úplný systém jevů, $A \in \mathcal{F}$. Potom*

$$P(A) = \sum_{i=1}^{n} P(A|H_i) P(H_i)$$

Důkaz: $P(A) = P(A \cap \Omega) = P(A \cap (\cup_{i=1}^n H_i)) = P(\cup_{i=1}^n (A \cap H_i)) = \sum_{i=1}^n P(A \cap H_i) = \sum_{i=1}^n P(A|H_i)P(H_i)$

Příklad 1.1.20 *Jaká je pravděpodobnost, že vybraná součástka, která patří s pravděpodobností 0,2; 0,3 resp. 0,5 do první, druhé, resp. třetí skupiny, vydrží zátěž, jestliže součástka z první, druhé resp. třetí skupiny vydrží zátěž s pravděpodobností 0,95; 0,9 resp. 0,85.*

Řešení: Označme jev $A = [\text{součástka vydrží zátěž}]$, $H_i = [\text{součástka vybrána z } i\text{-té skupiny}]$. Ze zadání plyne $P(H_1) = 0,2, P(H_2) = 0,3, P(H_3) = 0,5$, dále $P(A|H_1) = 0,95, P(A|H_2) = 0,9, P(A|H_3) = 0,85$. Podle věty o úplné pravděpodobnosti je tedy $P(A) = 0,885$. Výsledná pravděpodobnost je váženým průměrem podmíněných pravděpodobností, přičemž váhy tvoří pravděpodobnosti jevů z úplného systému.

Věta 1.4 (Bayesova) *Nechť $\{H_i\}_{i=1}^n$ je úplný systém jevů, $A \in \mathcal{F}, P(A) > 0$. Potom*

$$P(H_i|A) = \frac{P(A|H_i P(H_i)}{\sum_i P(A|H_i)P(H_i)}$$

Důkaz: Je $P(H_i|A) = \frac{P(A \cap H_i)}{P(A)}$. Vyjádříme-li pravděpodobnost průniku v čitateli podle definice podmíněné pravděpodobnosti a dosadíme-li pravděpodobnost jevu A ve jmenovateli podle věty o úplné pravděpodobnosti, dostáváme tvrzení věty (tzv. *Bayesův vzorec*).

V aplikacích Bayesova vzorce z věty 1.4 mají jevy H_i význam hypotéz, které se navzájem vylučují a právě jedna z nich je správná. $P(H_i)$ jsou jejich pravděpodobnosti před provedením doplňujícího pokusu či testu, říká se jim *apriorní pravděpodobnosti*. $P(.|H_i)$ je pravděpodobnostní míra výsledků testu za platnosti hypotézy H_i, která je často známá. Vzorec umožňuje vypočítat podmíněné pravděpodobnosti hypotéz po provedení testu v němž nastal jev A, takzvané *aposteriorní pravděpodobnosti*.

Příklad 1.1.21 *Zamýšlíte koupit v autobazaru vůz jisté značky, je ovšem známo, že 30 procent takových nabízených vozů má vadné převodovky. Abyste získali více informací, najmete si mechanika, který je po projížďce schopen odhadnout stav vozu a jen s pravděpodobností 0,1 se zmýlí. Jaká je pravděpodobnost, že vůz, který zamýšlíte koupit, má vadnou převodovku*

***a**) předtím, než si najmete mechanika?*
***b**) jestliže mechanik předpoví, že je dobrý?*

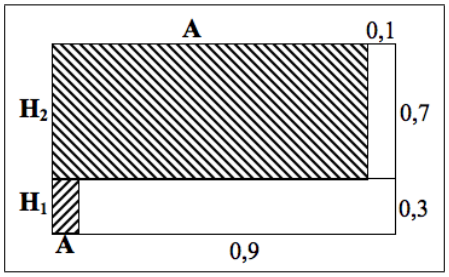
Řešení: Označme $H_1 = [\text{vůz je vadný}]$ a $H_2 = [\text{vůz je dobrý}]$ dvě hypotézy. Jedna z nich určitě nastane a navzájem se vylučují – tvoří tedy úplné pokrytí. Dále označme výsledek mechanikova testu $A = [\text{doporučuji vůz koupit}]$. Odpověď na otázku a) je dána podílem vadných vozů $P(H_1) = 0,3$,

což je jediná informace před najmutím mechanika. Výpočet aposteriorní pravděpodobnosti b) lze provést podle Bayesova vzorce

$$P(H_1|A) = \frac{P(A|H_1).P(H_1)}{P(A|H_1).P(H_1) + P(A|H_2).P(H_2)} = \frac{0,1.0,3}{0,1.0,3 + 0,9.0,7} = \frac{0,03}{0,63} = 0,045.$$

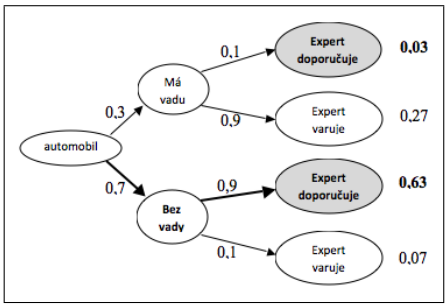
Tedy pravděpodobnost koupě vadného vozu podle doporučení experta se sníží na necelých pět procent. Odhaduje-li mechanik, že vůz je dobrý, šance na nákup dobrého vozu se zvýšila ze 70 % na 95,5 %.

Jedno z možných grafických znázornění výsledku b) je na obr. 1.3. Celková pravděpodobnost (plocha čtverce) je rozdělena vodorovnou příčkou v poměru pravděpodobnosti hypotéz, dále svislými příčkami v poměru podmíněných pravděpodobností výsledků testu. $P(H_1|A)$ hledáme v pravděpodobnostním prostoru redukovaném na vyšrafovanou plochu (t.j. za podmínky A) jako podíl plochy odpovídající H_1 ku celkové ploše.



Obrázek 1.3: Úloha o nákupu automobilu

Jiný způsob grafického znázornění výpočtu pomocí stromového grafu je na obr. 1.4. Zde pravděpodobnost nákupu vadného vozu s doporučením mechanika odpovídá první větvi shora – označme ji α. Případy, kdy mechanik doporučí koupi, jsou zahrnuty v první a třetí větvi – jejich součet označme β. Potom hledaná pravděpodobnost je rovna podílu $\frac{\alpha}{\beta}$.



Obrázek 1.4: Úloha o nákupu automobilu

1.2 Náhodná veličina

1.2.1 Rozdělení náhodné veličiny

Pod pojmem **(reálná) náhodná veličina** budeme rozumět reálnou funkci $X : \Omega \to R$, definovanou na prostoru elementárních jevů Ω a takovou, že pro každé reálné číslo $x \in R$ je $A = \{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}$.

Pokud je obor hodnot náhodné veličiny podmnožinou komplexních čísel, budeme hovořit o *komplexní* náhodné veličině, v případě, že hodnoty leží ve vektorovém prostoru R^n, budeme hovořit o náhodném vektoru. V této části se omezíme pouze na reálné náhodné veličiny.

Nadále budeme používat zkráceného zápisu $A = \{X \leq x\}$. Nenechte se však tímto zkráceným zápisem zmást: vynecháním ω v zápisu jeho vliv na hodnotu náhodné veličiny X nemizí a je třeba s ním vždy počítat!

Podmínka, že A je prvkem jevového pole $\mathcal{F}$ – jinými slovy že A je náhodný jev v uvažovaném náhodném pokusu – je potřeba k tomu, abychom mohli jevu $A = \{X \leq x\}$ přiřadit jeho pravděpodobnost. Připomeňme, že pravděpodobnost je definována právě na množině $\mathcal{F}$.

Je-li $\mathcal{F}$ systém všech podmnožin Ω, potom každá reálná funkce X na Ω je náhodná veličina. V obecném případě se může stát, že některé příliš komplikované funkce nemusí splňovat podmínku z definice náhodné veličiny. S takovými funkcemi se ale v praktických aplikacích nesetkáváme.

Funkce $F(x)$ definovaná vztahem

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

se nazývá **distribuční funkce** náhodné veličiny X. (Tento zápis je zjednodušen vynecháním složené závorky u jevu $\{X \leq x\}$.)

Věta 1.5 *Nechť $F(x)$ je distribuční funkce náhodné veličiny X. Potom platí*

(1) $0 \leq F(x) \leq 1$ *pro každé* $x \in R$,

(2) $\lim_{x\to-\infty} F(x) = 0, \lim x \to \infty F(x) = 1$,

(3) $\lim_{h\to 0_+} F(x+h) = F(x)$, *tj. F je zprava spojitá*[8],

(4) je-li $x_1 \leq x_2$, *potom* $F(x_1) \leq F(x_2)$, *tj. F je neklesající.*

[8] V některých učebnicích se můžete setkat s poněkud jinou definicí distribuční funkce: $F(x) = P(X < x)$ (s ostrou nerovností). Takováto distribuční funkce potom bude spojitá zleva.

Tyto vlastnosti jsou též postačující pro to, aby daná funkce $F(x)$ byla distribuční funkcí nějaké náhodné veličiny. Často se používá další vlastnost: pro reálná čísla $a \leq b$ platí

$$P(a < X \leq b) = P(\{X \leq b\} \cap \{X > a\}) = P(\{X \leq b\} - \{X \leq a\}) \\ = P(X \leq b) - P(x \leq a) = F(b) - F(a).$$

Distribuční funkce $F(x)$ se nazývá **diskrétní**, existuje-li konečná nebo spočetná posloupnost bodů $\{x_i\}$ a posloupnost kladných čísel p_i splňujících $\sum_i p_i = 1$ takové, že

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

pro libovolné reálné číslo $x \in R$. Diskrétní distribuční funkce má schodovitý tvar, se skoky velikosti p_i v bodech x_i. Má-li náhodná veličina X diskrétní distribuční funkci, tj. $p_i = P(X = x_i)$, říkáme, že X má **diskrétní rozdělení pravděpodobnosti**, stručně **diskrétní rozdělení**.

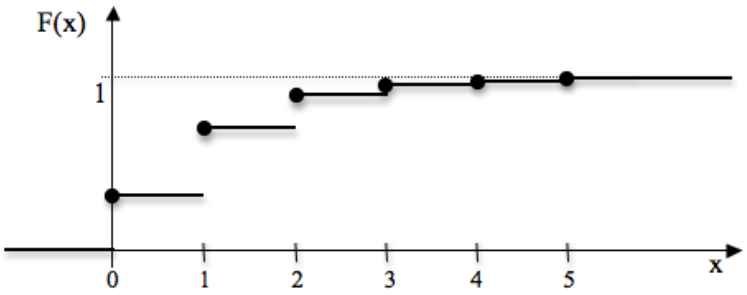
Vraťme se k náhodnému experimentu, jehož reprezentace je pravděpodobnostní prostor $(\Omega, \mathcal{F}, P)$. Jestliže je množina elementárních výsledků Ω konečná nebo spočetná, to znamená, že všechny možné elementární výsledky lze seřadit do posloupnosti $\Omega = \{\omega_i\}_{i=1}^{\infty}$, potom i náhodná veličina X definovaná pro tento experiment má nejvýše spočetnou množinu hodnot $X(\omega_i) = x_i$ tvořících posloupnost $\{x_i\}_{i=1}^{\infty}$. Pravděpodobnostní rozdělení je v tomto případě určeno jednoznačně pravděpodobnostní mírou P, tedy pravděpodobnostmi $\{p_i\}_{i=1}^{\infty}$, kde $p_i = P(X = x_i)$.

Typický průběh diskrétní distribuční funkce je na obr. 2.1. Tato funkce je vždy po částech konstantní („schodovitá“), s hodnotami vždy na levé straně „schodu“ (spojitost zprava).

Příklad 1.2.1 *Nezávisle na sobě je vyrobena série pěti ventilátorů, přičemž pravděpodobnost toho, že vyrobený ventilátor bude vadný, je* $0,1$. *Najděte rozdělení pravděpodobnosti náhodné veličiny, popisující počet vadných v sérii.*

Řešení: Označme $A_k =$ [v sérii 5 výrobků bude k vadných]. Pravděpodobnost toho, že k výrobků bude vadných a $(5-k)$ výrobků bude bez vady je – vzhledem k nezávislosti – rovna součinu $0,1^k 0,9^{5-k}$. Kromě toho, k vadných výrobků může být mezi 5 vyrobenými rozděleno celkem $\binom{5}{k}$ způsoby. Tedy $P(A_k) = \binom{5}{k} 0,1^k.0,9^{(5-k)}$. V následující tabulce jsou uspořádány hodnoty k a p_k, $k = 1, \ldots, 5$ a na obr.2.1 je graf příslušné distribuční funkce.

i	0	1	2	3	4	5
p_i	0,3277	0,4096	0,2048	0,0512	0,0064	0,0003



Obrázek 1.5: Distribuční funkce diskrétní náhodné veličiny z příkladu 1.2.1.

Distribuční funkce $F(x)$ se nazývá **absolutně spojitá**, jestliže existuje spojitá nezáporná funkce $f(x)$ nazývaná **hustota pravděpodobnosti**, stručně **hustota**, taková, že

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

pro každé $x \in R$. Má-li náhodná veličina X absolutně spojitou distribuční funkci, říkáme, že X má **spojité rozdělení pravděpodobnosti**, stručně **spojité rozdělení**.

Hustota pravděpodobnosti $f(x)$ musí splňovat rovnost $\int_R f(x)dx = 1$. Existuje-li derivace F' distribuční funkce bodě x, potom je $F'(x) = f(x)$. Pro $a, b \in R, a < b$, platí

$$P(a < X \leq b) = F(b) - F(a) = \int_{a}^{b} f(t)dt$$

Tedy pravděpodobnost toho, že spojitá náhodná veličina bude mít hodnoty v nějakém intervalu $\langle a, b\rangle$ je tedy plocha pod křivkou hustoty nad intervalem $\langle a, b\rangle$.

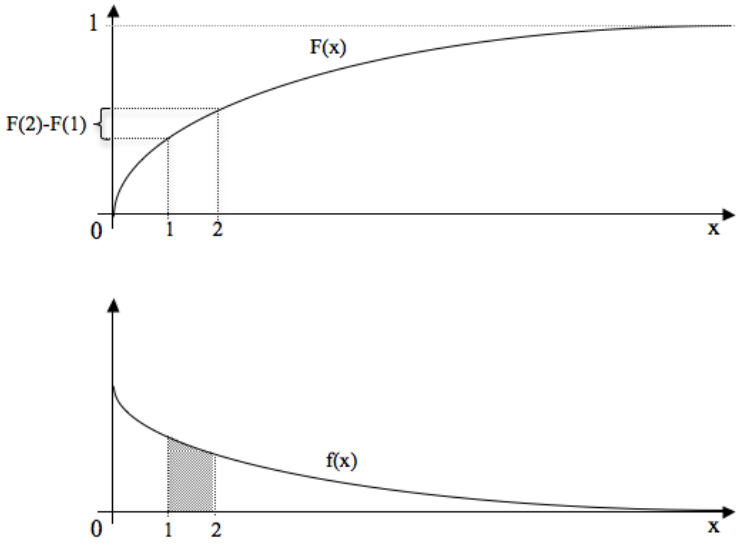
Jaká je vlastně interpretace hustoty pravděpodobnosti? Především $f(x)$ není pravděpodobnost jevu $\{X = x\}$! Ta je v případě náhodné veličiny se spojitým rozdělením vždy rovna nule. To plyne z $P(X = x) = \int_x^x f(t)dt = 0$ pro libovolné $x \in R$. Uvažujme interval $\langle x, x + dx)$, to jest interval reálných čísel o délce dx. Potom pro velmi malá dx je přibližně

$$P\left(X \in \langle x, x + dx)\right) = f(x)dx$$

.

Příklad 1.2.2 *Ukažte, že funkce* $F(x) = 1 - e^{-\lambda x}, x \geq 0, F(x) = 0, x < 0$, *kde* $\lambda > 0$ *je reálný parametr, je distribuční funkcí nějaké náhodné veličiny* Y *se spojitým rozdělením. Spočtěte* $P(1 < Y \leq 2)$.

Řešení: Derivace $\frac{dF(x)}{dx} = F'(x) = \lambda e^{-\lambda x}$ existuje a je pro $x \geq 0$ kladná, tedy $F(x)$ je rostoucí funkcí pro $x \geq 0$. Navíc $\lim_{x\to 0+} F(x) = 0$ a $\lim_{x\to\infty} = 1$. Z toho plynou vlastnosti (1)–(4) z věty 1.5 a tedy $F(x)$ je podle poznámky za větou 1.5 distribuční funkcí nějaké náhodné veličiny Y. Odpovídající hustota je $f(x) = F'(x) = \lambda e^{-\lambda x}, x \geq 0, f(x) = 0, x < 0$ zřejmě spojitá funkce.
Dále je $P(1 < Y \leq 2) = \lambda \int_1^2 e^{-\lambda x} dx = 1 - e^{-2\lambda} - 1 + e^{-\lambda} = 0,117$.



Obrázek 1.6: Distribuční funkce (nahoře) a hustota (dole) náhodné veličiny z příkladu 1.2.2.

Poznámka: V dalším výkladu se budeme zabývat především dvěma zavedenými třídami náhodných veličin, tzn. se spojitým a diskrétním rozdělením. Samozřejmě mohou existovat i náhodné veličiny, které mají smíšený charakter. Například při měření doby do poruchy složitého zařízení je třeba v některých případech uvažovat kladnou pravděpodobnost, že se zařízení vůbec nepodaří uvést do chodu. Tomu může odpovídat distribuční funkce která je nulová pro záporná x, absolutně spojitá a rostoucí pro $x > 0$ a se skokem v bodě $x = 0$.

1.2.2 Základní charakteristiky náhodných veličin

Náhodná veličina je funkce náhodných jevů. V jistém smyslu si ji můžeme představovat jako číselnou reprezentaci výsledků náhodného pokusu. Práce s náhodnými veličinami se však výrazně odlišuje od práce s matematickými funkcemi jako je sinus, exponenciela nebo mocnina. Na rozdíl od matematických funkcí nelze například nakreslit graf náhodné veličiny. Nelze stanovit její průběh nebo limitu. Lze pouze stanovit její **pravděpodobnostní charakteristiky**.

Pravděpodobnostní vlastnosti náhodné veličiny jsou plně popsány její distribuční funkcí. Pomocí distribuční funkce lze určit rozdělení pravděpodobnosti $\{p_i\}_{i=1}^{\infty}$ v případě diskrétní náhodné veličiny nebo hustotu $f(x)$ v případě spojité náhodné veličiny. To platí i obráceně: známe-li rozdělení pravděpodobnosti $\{p_i\}_{i=1}^{\infty}$ nebo hustotu $f(x)$, můžeme najít distribuční funkci $F(x)$.

Kromě těchto charakteristik se často používají i další, mezi něž patří především **momenty** a **kvantily**.

***K*-tý obecný moment** EX^k náhodné veličiny X s diskrétním resp. spojitým rozdělením je hodnota

$$E(X^k) = \sum_{i=1}^{\infty} x_i^k p_i, \text{ resp. } E(X^k) = \int_{-\infty}^{\infty} x^k f(x) dx.$$

Často budeme psát stručněji pouze EX^k (bez závorek). Významnou úlohu hraje první obecný moment $E(X)$, který se nazývá **střední hodnota** náhodné veličiny X (někdy se můžete setkat i s názvem **očekávaná hodnota**). Pomocí něho jsou definovány takzvané **centrální momenty**.

***K*-tý centrální moment** $\mu_k(X)$ náhodné veličiny X (s diskrétním nebo spojitým rozdělením) je hodnota

$$\mu_k(X) = E(X - E(X))^k.$$

Opět se poněkud liší výpočet centrálních momentů pro diskrétní resp. spojitou náhodnou veličinu:

$$E(X-EX)^n = \sum_{i=1}^{\infty} (x_i - EX)^k p_i, \text{ resp. } E(X-EX)^k = \int_{-\infty}^{\infty} (x-EX)^k f(x) dx.$$

Nejčastěji používaným centrálním momentem je druhý centrální moment, který se nazývá **rozptyl** náhodné veličiny X a ozančuje se speciálním symbolem $Var(X)$. Druhá odmocnina rozptylu je nazývána **směrodatnou odchylkou** náhodné veličiny X a budeme ji obvykle označovat $\sigma(X)$.

Znalost všech momentů (obecných nebo centrálních) pro $k = 1, 2, \ldots$ je opět ekvivalentní znalosti distribuční funkce a podává nám plnou informaci o náhodné veličině.

Příklad 1.2.3 *Náhodná veličina X může nabývat nezáporných celočíselných hodnot $k = 0, 1, \ldots$ s pravděpodobnostmi $p_k = P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$. Spočtěte střední hodnotu a rozptyl náhodné veličiny X.*

Řešení: Protože veličina X nabývá pouze spočetně mnoha hodnot, jedná se o veličinu diskrétní. Pro její střední hodnotu tedy platí
$EX = \sum_{k=0}^{\infty} k e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\sum_{k=1}^{\infty} k\ \frac{\lambda^k}{k!} = \lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!} = \lambda.$
Podobně pro rozptyl
$Var(X) = E(X-EX)^2 = \sum_{k=0}^{\infty}(k-\lambda)^2 e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\sum_{k=0}^{\infty}(k^2-2k\lambda+\lambda^2)\frac{\lambda^k}{k!}.$
Po krátkém počítání a úpravách se dostaneme k výsledku $Var(X) = \lambda$.
Rozdělení této náhodné veličiny se nazývá Poissonovo a budeme se jím ještě zabývat později.

Při výpočtu momentů náhodné veličiny se používá funkcionál E, který je lineární, to znamená, že pro libovolné náhodné veličiny X a Y a konstantu $k \in R$ platí:

- $E(X + Y) = E(X) + E(Y)$ (*aditivita*)
- $E(kX) = kE(X)$ (*homogenita*).

Tyto vlastnosti jsou zřejmé, když si uvědomíme, že funkcionál E je vyjádřen buď jako součet v diskrétním případě, nebo ve formě integrálu ve spojitém případě. Z druhé vlastnosti přímo plyne, že $E(k) = k$.

Speciálně lze uvedené vlastnosti interpretovat jako vlastnosti střední hodnoty. Uvědomíme-li si, že EX je číslo (nikoli funkce) a s využitím linearity střední hodnoty lze pro rozptyl snadno odvodit tyto užitečné vlastnosti:

- $Var(X) = E(X - EX)^2 = E(X^2 - 2XEX + (EX)^2) =$
  $EX^2 - 2(EX)^2 + (EX)^2 = EX^2 - (EX)^2,$
- $Var(kX + q) = k^2 Var(X)$ pro libovolné konstanty k, q.

První vlastnost se často využívá při výpočtu rozptylu. Všimněte si navíc, že tato vlastnost vlastně vyjadřuje vztah mezi druhým centrálním momentem a prvními dvěma obecnými momenty náhodné veličiny. Toto zjištění lze zobecnit a můžeme tvrdit, že libovolný k-tý centrální moment lze vyjádřit pouze pomocí prvních k obecných momentů.

Příklad 1.2.4 *Vyjádřete čtvrtý centrální moment náhodné veličiny X pomocí jejích prvních čtyř obecných momentů.*

Řešení: $\mu_4(X) = E(X-EX)^4 = E(X^4-4X^3EX+8X^2(EX)^2-4X(EX)^3+(EX)^4) = EX^4 - 4EX^3EX + 8EX^2(EX)^2 - 3(EX)^4$

Příklad 1.2.5 *Náhodná veličina X má hustotu rozdělení pravděpodobnosti $f(x) = \lambda e^{-\lambda x}, x \geq 0, f(x) = 0$ jinak. Spočtěte její obecné momenty.*

Řešení: X je spojitá náhodná veličina (její rozdělení pravděpodobnosti je dáno hustotou) a proto pro výpočet použijeme integrál
$EX^n = \lambda \int_0^\infty x^n e^{-\lambda x} dx$, který substitucí $\lambda x = t, \lambda dx = dt$ převedeme na integrál $\frac{1}{\lambda^n} \int_0^\infty t^n e^{-t} dt$. Integrál v tomto výrazu je znám jako takzvaná *Gama funkce* definovaná vztahem $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt, x > 0$. Pro funkci $\Gamma(x)$ platí několik vztahů, mezi jinými též $\Gamma(x+1) = x\Gamma(x), \Gamma(1) = 1, \Gamma(\frac{1}{2}) = \sqrt{\pi}$. V našem případě tedy dostáváme $EX^n = \frac{\Gamma(n+1)}{\lambda^n} = \frac{n!}{\lambda^n}$.

Následující věta popisuje vztah mezi střední hodnotou a rozptylem náhodné veličiny.

Věta 1.6 (Čebyševova nerovnost) *Nechť náhodná veličina X má konečný druhý moment. Potom pro libovolné $\epsilon > 0$ platí*

$$P(|X - EX| \geq \epsilon) \leq \frac{Var(X)}{\epsilon^2}.$$

Často se pracuje s takzvanými **normovanými momenty**. Jsou to vlastně obecné momenty **normované náhodné veličiny** $U = \frac{X - E(X)}{\sigma(X)}$. Tedy k-tý normovaný moment náhodné veličiny X je

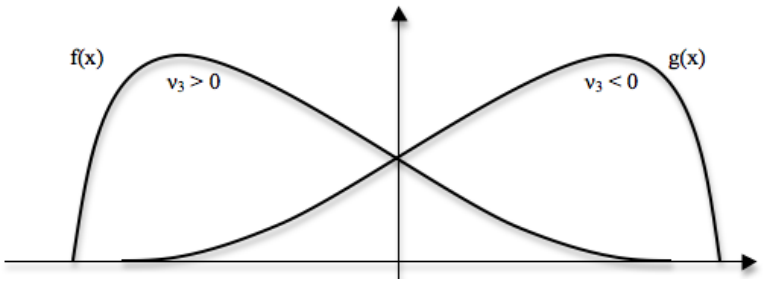
$$\nu_k(X) = E(U^k).$$

Výpočet normovaných momentů pro diskrétní resp. spojitou náhodnou veličinu je následující:

$$E\left(\frac{X - EX}{\sigma(X)}\right)^k = \sum_{i=1}^{\infty} \left[\frac{x_i - EX}{\sigma(X)}\right]^k p_i, \text{ resp. } \int_{-\infty}^{\infty} \left(\frac{x - EX}{\sigma(X)}\right)^k f(x) dx.$$

Z normovaných momentů náhodné veličiny jsou důležité ν_3 a ν_4, které popisují tvar jejího rozdělení. Normovaný moment ν_3 se nazývá **koeficient šikmosti** a je mírou symetrie rozdělení. Koeficient šikmosti je roven nule například pro náhodné veličiny, jejichž rozdělení pravděpodobnosti je symetrické kolem střední hodnoty, je kladný pro jednovrcholové hustoty šikmé zprava (obr.1.7), naopak záporný pro jednovrcholové hustoty šikmé zleva.

Normovaný moment ν_4 je nazýván **koeficientem špičatosti** nebo také **kurtoze** a je mírou toho, jak rychle klesá pravděpodobnost extrémních hodnot (směrem k $-\infty$ nebo do $+\infty$).



Obrázek 1.7: Hustoty nesymetrických rozdělení a jejich koeficienty šikmosti.

Při studiu chování náhodné veličiny si zpravidla klademe otázku: *jaká je pravděpodobnost α, že sledovaná náhodná veličina X nepřekročí předem danou hodnotu x*? Často je však kladena i opačná otázka: *jakou hodnotu x nepřekročí sledovaná náhodná veličina s předem danou pravděpodobností α*? Odpověď nám dávají **kvantily rozdělení náhodné veličiny**, které tvoří další důležitou skupinou charakteristik náhodné veličiny.

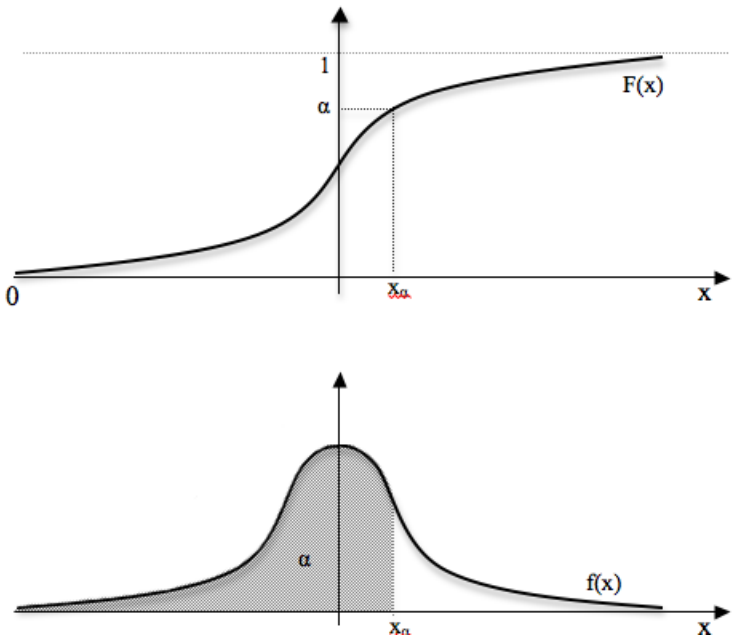
Mějme nějaké $0 < \alpha < 1$. Potom α-**kvantilem** náhodné veličiny X nazýváme takovou největší hodnotu x_α, pro kterou je

$$P(X \leq x_\alpha) \leq \alpha.$$

Má-li náhodná veličina X absolutně spojitou distribuční funkci $F(x)$, která je rostoucí pro ta x, pro která $0 < F(x) < 1$, potom existuje x_α takové, že $P(X \leq x_\alpha) = \alpha$ a je $x_\alpha = F^{-1}(\alpha)$, kde F^{-1} je inverzní funkce k F. Vyjádřeno pomocí hustoty f je $\int_{-\infty}^{x_\alpha} f(x)dx = \alpha$, viz obr.1.8.

Pravděpodobnost se často vyjadřuje v procentech. Potom budeme hovořit o $100\alpha\%$-kvantilu náhodné veličiny. Podobně jako v případě momentů i zde platí tvrzení, že známe-li všechny α-kvantily náhodné veličiny X pro všechna $\alpha \in \langle 0, 1\rangle$, potom máme úplnou informaci o jejím pravděpodobnostním chování.

Mezi všemi kvantily má významné postavení 50%-kvantil, který budeme označovat $x_{0,5}$ a budeme jej nazývat **medián** $Me(X)$ náhodné veličiny X. Medián se také někdy nazývá *prostřední* hodnota z hlediska pravděpodobnosti, neboť pravděpodobnost, že náhodná veličina X nabyde hodnoty menší než $Me(X)$ je rovna 0,5 což je stejná hodnota jako pravděpodobnost, že X bude mít hodnotu větší než $Me(X)$. Vedle střední hodnoty je to další takzvaná *charakteristika polohy* náhodné veličiny X.



Obrázek 1.8: Vztah mezi pravděpodobností α a α-kvantilem x_α.

Příklad 1.2.6 *Určete medián náhodné veličiny s hustotou* $f(x) = 0$ *pro* $x < 0$ *a* $f(x) = \lambda e^{-\lambda x}$ *pro* $x \geq 0$.

Řešení: Medián je taková hodnota x, pro kterou má platit, že $\int_{-\infty}^{x} f(t)dt = \int_{x}^{\infty} f(t)dt = 0,5$. Tedy v našem případě $\int_0^x \lambda e^{-\lambda t}dt = 1 - e^{-\lambda x} = 0,5$ a odtud dostaneme $Me(X) = \frac{ln2}{\lambda}$.

Při analýze dat se často používají takzvané **kvartily**. To jsou 25%, 50% a 75% kvantily. Přitom 25%-kvantil se nazývá **dolní kvartil**, 50%-kvantil je již zmíněný medián a 75%-kvantil je **horní kvartil**. Spolu s minimem a maximem se těmto charakteristikám říká *pět Tukeyho charakteristik* podle zakladatele takzvané „průzkumové analýzy dat“, amerického statistika Johna Tukeye.

Ve statistice se dále pracuje s takzvaným *horním* a *dolním* 5%-kvantilem. To jsou 5%- a 95%-kvantily. Podobně se můžete setkat i s pojmem *horní* nebo *dolní decil* – tedy 10%- nebo 90%-kvantil.

Kromě již zmíněných pravděpodobnostních charakteristik náhodné veličiny se používají i další, které nepatří ani do jedné z uvedených skupin. Takovou

charakteristikou je například **Modus** $Mo(X)$ náhodné veličiny X. Je to hodnota x, v níž má spojitá, resp. diskrétní náhodná veličina lokální extrém hustoty resp. posloupnosti pravděpodobností p_i. Rozdělení s jediným lokálním extrémem se nazývají **jednovrcholová**. V diskrétním případě jednovrcholového rozdělení lze modus interpretovat jako nejpravděpodobnější hodnotu náhodné veličiny X. Existuje však řada případů, kdy modus nelze jednoznačně určit.

V teorii spolehlivosti, zabývající se pravděpodobností bezporuchového chodu zařízení, se vedle distribuční funkce používá také **funkce spolehlivosti**[9] $R(t)$, definovaná následujícím způsobem: Je-li $F(t)$ distribuční funkce náhodné veličiny T, popisující dobu do poruchy zařízení, potom funkce spolehlivosti je rovna $R(t) = 1 - F(t)$. Tedy zatímco $F(t)$ je pravděpodobnost toho, že se zařízení porouchá do doby t, $R(t)$ je pravděpodobnost toho, že zařízení „přežije“ dobu t.

Další charakteristikou, používanou ve spolehlivosti, je **intenzita poruch**[10] $h(t)$, která je definována jako $h(t) = \frac{f(t)}{1-F(t)}$. Její interpretace je následující: pro malé hodnoty dt součin $h(t)dt$ udává přibližně podmíněnou pravděpodobnost toho, že se zařízení neporouchá v nejbližším časovém intervalu délky dt, pokud se neporouchalo do doby t.

[9] nebo též **funkce přežití**.

[10] V literatuře se můžete setkat i s názvem **riziková funkce**, anglicky **hazard function**, především v souvislosti s analýzou doby přežívání.

1.3 Pravděpodobnostní modely

Aplikace matematických metod při popisu reálného světa používá matematické modely. Stejně je tomu i v případě teorie pravděpodobnosti. V reálném světě kolem nás existuje řada standardních situací, které mají cosi společného. Za cenu jistého zjednodušení a zanedbání některých nepodstatných okolností jsme schopni tyto situace popsat pomocí matematických symbolů a vytvořit matematický model. Ten potom slouží ke studiu chování, závislostí, předpovídání budoucích jevů, a odhadů různých parametrů. Podle stupně zobecnění dostáváme více či méně složité modely. Neexistuje ideální model[11]

Zde uvedené pravděpodobnostní modely popisují řadu standardních situací, se kterými se můžeme setkat. Tak například házení mincí. Samozřejmě si lze představit, jak z dlouhé chvíle házíme mincí a sledujeme co nám padne. Lze si představit i hazardní hru, založenou na házení mincí. Ale mnohem praktičtější varianta tohoto standardního modelu je například zkouška jističe. Při každém pokusu buď sepne nebo nesepne. Zkoušíme jej tak dlouho, dokud nenastane první porucha. Nebo vybírání černých a bílých koulí z urny. To je oblíbená zábava ve středoškolských učebnicích pravděpodobnosti. Dokonce i ve sbírkách úloh z teorie pravděpodobnosti pro vysoké školy. Proč se máme zabývat takovou naprosto absurdní zábavou? Pokud si ale namísto černých a bílých koulí v urně představíme výrobky, například pístky do kompresoru, teré buď odpovídají požadavkům odběratele nebo vykazují nějakou vadu (rozměry mimo toleranci nebo nepřípustná drsnost povrchu) a jsou dodány v kontejneru, ze kterého náhodně vybereme určitý počet a zkoumáme jeho kvalitu[12], jedná se o stejný model o jehož praktičnosti už nepochybujeme.

Klasickým případem je model normálního rozdělení, ke kterému v minulosti došlo několik matematiků v různých dobách nezávisle na sobě a v různých souvislostech. Ať už zkoumali výšku branců pro královskou armádu nebo polohu nebeských těles či pohyb molekul, vždy se dostali k témuž modelu.

Každý pravděpodobnostní model představuje určitý typ náhodného pokusu, se kterým je spojena sledovaná náhodná veličina a nějaké standardní rozdělení pravděpodobnosti. Tato rozdělení zpravidla závisí na různém počtu parametrů. Pochopení těchto modelů a významu jejich parametrů je velmi důležité pro aplikaci pravděpodobnostních metod v praktických úlohách.

Následující přehled pravděpodobnostních modelů není zdaleka vyčerpávající. Uvádíme zde pouze ty nejzákladnější, nejčastěji používané modely .

[11] Jak napsal zakladatel kybernetiky, Norbert Wiener: „nejlepším modelem kočky je kočka“ a dodal: „a nejlépe ta samá“.

[12] Ve statistice tomu říkáme „statistická přejímka“.

1.3.1 Diskrétní modely

Model diskrétního rovnoměrného rozdělení $U(n)$**.** Uvažujme náhodný pokus, který může skončit konečným počtem n stejně možných výsledků $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$. Nemáme důvod domnívat se, že některý výsledek má větší naději nastat než jiný. v takovém případě budou mít všechny stejnou pravděpodobnost $p = \frac{1}{n}$. Počet možných výsledků n je parametrem tohoto rozdělení. V některých případech má smysl definovat náhodnou veličinu $X(\omega_k) = x_k$. Potom

$$\begin{array}{ll} \textit{Pravděpodobnostní funkce}: & p_k = P(X = x_k) = \frac{1}{n}, \quad k = 1, 2, \ldots, n \\ \textit{Základní charakteristiky}: & E(X) = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} \text{ (aritmetický průměr)} \\ & Var(X) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 \end{array}$$

Příklad 1.3.1 *Představme si deset přístrojů, například mobilních telefonů. Vybereme náhodně jeden z nich.*

v takovémto případě můžeme definovat náhodnou veličinu různým způsobem. Pokud veličina $X =$ [výrobní číslo telefonu], potom zřejmě nemá smysl počítat střední hodnotu ani rozptyl. Jiná situace je, uvažujeme-li náhodnou veličinu $Y =$ [cena přístroje]. v tomto případě střední neboli očekávaná hodnota je průměrná cena spočtená z cen všech deseti přístrojů. To má svůj význm například chceme-li dopředu (před výběrem) znát alespoň orientačně výslednou cenu vybraného přístroje.

Můžete namítnout, co je to za náhodnou veličinu, když ceny přístrojů jsou předem jednoznačně dané. Nahodilost je zde opět třeba chápat z hlediska toho, kdo provádí experiment: ten dopředu neví, který z přístrojů vybere a jaká bude jeho cena. Střední hodnota navíc vypovídá o ceně vybraného přístroje více v případě, že se ceny jednotlivých přístrojů příliš neliší. Jestliže se ceny pohybují od několika set korun do desítek tisíc korun, potom nám zřejmě střední hodnota mnoho informace nepřinese. Proto je třeba údaj o střední hodnotě doplnit údajem o rozptylu nebo alespoň směrodatné odchylce.

Alternativní model $Alt(p)$**.** Náhodný pokus v tomto modelu může skončit pouze dvěmi výsledky. Někdy se takovému pokusu také říká *Bernoulliovský*[13]. Náhodnou veličinu zde definujeme takto: jednomu výsledku – v tomto modelu jej budeme označovat jako „úspěch“ – přiřadíme hodnotu $X = 1$ a

[13]Podle nejstaršího ze savné rodiny švýcarských matematiků, Jacoba Bernoulli (27.12.1654 – 16.8.1705).

pravděpodobnost $P(X = 1) = p$; druhý výsledek – neúspěch – bude mít číselnou hodnotu 0 a nastane s pravděpodobností $P(X = 0) = 1 - p$. Hodnota p je parametrem tohoto modelu, který odpovídá přibližně poměru počtu úspěchů k počtu neúspěchů při velkém počtu opakování alternativního pokusu. Pokud náhodná veličina bude odpovídat tomuto modelu, budeme to vyjádřovat stručně $X \approx Alt(p)$.

Pravděpodobnostní funkce : $p_0 = P(X = 0) = 1 - p, \quad p_1 = P(X = 1) = p$
Základní charakteristiky : $E(X) = 0.(1 - p) + 1.p = p$
$Var(X) = p - p^2 = p.(1 - p)$

Příklad 1.3.2 *Ochrana parního turbogenerátoru neodpojí přívod páry a tím neodstaví turbínu v případě poruchy, spočívající v překročení povolené úrovně vibrací rotoru s pravděpodobností* $p = 7,6.10^{-4}$.

Jedná se o alternativní model, v němž „úspěchem“ je nezastavení turbogenerátoru v případě poruchy. Střední hodnota, parametr p zde říká, že lze očekávat, že při dlouhém provozu za přibližně stejných podmínek tento případ nastane přibližně v jednom z 1316 případů poruchy ($= \frac{1}{p}$). Pokud by k této poruše docházelo v průměru čtyřikrát do roka, potom by selhání bezpečnostní funkce ochran selhalo přibližně jednou za 329 let.

Binomický model $Bin(n, p)$**.** Budeme-li opakovat Bernoulliovské pokusy n-krát nezávisle na sobě, bude nás zajímat počet „úspěchů“ při těchto n pokusech. Když označíme výsledky jednotlivých Bernoulliovských pokusů jako $X_i \approx Alt(p), i = 1, \ldots, n$ a položíme $Y = \sum_{i=1}^{n} X_i$, potom Y představuje počet výskytů jevu $\{X = 1\}$ v n opakováních. Y může nabývat hodnot $0, 1, \ldots, n$.

Pravděpodobnostní funkce : $p_k = P(Y = k) = \binom{n}{k} p^k.(1-p)^{(n-k)}, \; k = 0, \ldots, n$
Základní charakteristiky : $E(X) = n.p$
$Var(X) = n.p.(1 - p)$

Odvození pravděpodobnostní funkce bylo naznačeno v řešení příkladu 1.2.1. Tvar této funkce připomíná *binomickou větu* $(p+q)^n = \sum_{k=0}^{n} \binom{n}{k} p^k.q^{(n-k)}$ pro libovolná reálná čísle p, q a přirozené n. Dosadíme-li za $q = 1 - p$, dostáváme dkaz toho, že $\{p_k\}$ opravdu tvoří pravděpodobnostní rozdělení, neboli že $\sum_{k=0}^{n} p_k = 1$.

Binomický model lze také interpretovat jako počet vybraných prvků s určitou vlastností při náhodném výběru z celkového počtu n prvků s vracením. **Náhodným výběrem** rozumíme takový výběr z konečného souboru prvků, při kterém má každý prvek stejnou pravděpodobnost být vybrán (rovnoměrné rozdělení pravděpodobnosti výběru).

Příklad 1.3.3 *V urně je pět bílých a tři černé koule*[14]*. Postupně vybíráme tři koule s vracením (to znamená, že po každém výběry kouli vrátíme zpět do urny). Jaká je pravděpodobnost toho, že ve výběru bude nejvýše jedna černá koule?*

Řešení: Každý výběr koule představuje jeden alternativní pokus s pravděpodobností $p = \frac{3}{8}$ pro vytažení černé koule. Vzhledem k tomu, že koule vracíme zpět, tato pravděpodobnost se v průběhu experimentu (= vytahování tří koulí) nemění. Tím dostáváme binomický model s $n = 3$ a $p = \frac{3}{8}$ s množinou elementárních výsledků $\Omega = \{0, 1, 2, 3\}$ a podle binomického rozdělení pravděpodobnosti je $P(\text{ve výběru bude}k \text{ černých koulí}) = P(Y = k) = \binom{3}{k} \frac{3}{8}^k \frac{5}{8}^{(3-k)}$, $k = 0, \ldots, 3$. Jev že ve výběru nebude více než jedna černá sestává ze dvou elementárních výsledků $\{0, 1\}$. Proto hledaná pravděpodobnost je $P(Y = 0) + P(Y = 1) = \frac{5}{8}^3 + 3\frac{3}{8}\frac{5}{8}^2 = \frac{175}{256}$

Geometrický model $Geom(p)$ Provádějme Bernoulliovský pokus se dvěma možnými výsledky, které nazveme *úspěch* a *neúspěch*; tak dlouho, dokud nenastane první *úspěch*Nechť pravděpodobnost úspěchu je p. Počet nezávislých opakování pokusu předcházejících prvnímu úspěchu je náhodná veličina X s **geometrickým rozdělením**. *Pravděpodobnostní funkce* : $p_k = P(X = k) = (1-p)^k.p,\ k = 0, \ldots, n$

Základní charakteristiky : $E(X) = \frac{1-p}{p}$

$Var(X) = \frac{1-p}{p^2}$

V geometrickém modelu lze definovat také veličinu Y jako počet pokusů, potřebných k prvnímu úspěchu. Charakteristiky veličiny Y je následující *Pravděpodobnostní funkce* : $p_m^Y = P(Y = m) = (1-p)^(m-1).p,\ m = 1, \ldots, n$

Základní charakteristiky : $E(Y) = \frac{1}{p}$

$Var(Y) = \frac{1-p}{p^2}$

Veličina X se také někdy označuje jako „diskrétní doba do poruchy“, vyjadřující počet *zátěžových cyklů* , které zařízení vydrží, než se porouchá, přičemž pravěpodobnost poruchy při každém cyklu je stejná a je rovna p.

Příklad 1.3.4 *Po elektrickém jističi se požaduje, aby s pravděpodobností 0,98 vydržel více než 500 sepnutí. Jaká může být maximální pravděpodobnost selhání při jednom sepnutí?*

Řešení: Budeme předpokládat, že jednotlivá sepnutí jsou nezávislé jevy a pravděpodobnost selhání se nemění v čase. Potom můžeme tuto situaci popsat geometrickým modelem, přičemž pravděpodobnost přežití 500 sepnutí je

[14]... už je to tady!

rovna $P(X \geq 500) = \sum_{k=501}^{\infty}(1-p)^k.p$. S využitím vlastnosti pravděpodobnostního rozdělení a vzorce pro částečný součet geometrické řady dostáváme $P(X \geq 500) = 1 - p\sum_{k=0}^{500}(1-p)^k = 1 - p\frac{1-(1-p)^{500}}{1-(1-p)} = (1-p)^{500}$. Podle zadání má být $(1-p)^{500} = 0,98$. Odtud $p = 1 - \exp\left(\frac{\ln(0,98)}{500}\right) = 4,0405.10^{-5}$. To znamená, že má-li být zajištěna požadovaná spolehlivost, jistič se může porouchat v průmětu jednou za 25 tisíc cyklů, což je přibližně rovno střední hodnotě geometrického rozdělení s parametrem p (přesně je to 24.749 cyklů).

Hypergeometrický model $Hyp(N, M, n)$ Mezi N prvky je M s určitou vlastností. z těchto prvků provádíme náhodný výběr n prvků bez vracení. Náhodná veličina X představující počet prvků s určitou vlastností ve výběru má takzvané **hypergeometrické rozdělení**:

$$\begin{aligned} &\textit{Pravděpodobnostní funkce}: p_k = P(X=k) = \frac{C_k(M)C_{n-k}(N-M)}{C_n(N)} = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} \\ &\qquad\qquad \text{kde } max(0, M+n-N) \leq k \ \leq min(n, M). \\ &\textit{Základní charakteristiky}: \quad E(X) = n\frac{M}{N} \\ &\qquad\qquad Var(X) = \frac{nM(N-M)(N-n)}{N^2(N-1)} \end{aligned}$$

Je-li N velké a n je oproti němu hodně malé (jedná se o malý výběr z velké populace, uvádí se: $\frac{n}{N} < 0,1$), je hypergeometrické rozložení velmi blízké binomickému (s parametrem $p = \frac{M}{N}$), které se snadněji počítá.

Příklad 1.3.5 *Ve Sportce tipuje sázející 6 čísel ze 49 možných, o ktreých předpokládá, že budou při losování Sportky tažena. Výhru v V. pořadí získá, pokud se mu podaří uhodnout libovolná tři čísla ze šesti tažených. Jaká je pravděpodobnost výhry v V. pořadí?*

Řešení: V tomto příkladě je $N = 49$ (počet všech čísel), $M = 6$ (počet vyhrávajících čísel), $n = 6$ (počet vytažených čísel) a $k = 3$ (počet vyhrávajících čísel, která sázející uhodne). Čísla nemohou být tažena dvakrát, což odpovídá modelu výběru bez vracení. Slosování Sportky se obvykle provádí prostřednictvím elektromechanického osudí, čímž by měla být zajištěna náhodnost výběru. Hledanou pravděpodobnost lze potom vyjádřit pomocí hypergeometrického rozdělení:

$$P(\text{V. pořadí}) = \frac{\binom{6}{3}\binom{43}{3}}{\binom{49}{6}} = 0,0177$$

Šance na výhru v V. pořadí Sportky je tedy necelá dvě procenta.

Poissonův model $Poiss(\lambda T)$. Uvažujme události nastávající zcela náhodně v čase (např. poruchy přístroje, příchody telefonních volání do call-centra, rozpady atomů radioaktivního prvku apod.). Předpoklad náhodnosti lze matematicky formulovat následovně: Nezávisle na tom, ke kolika událostem došlo v časovém intervalu $\langle 0, t\rangle$, pravděpodobnost toho, že během krátkého intervalu $(t, t+h\rangle$, kde h je velmi malé číslo, dojde k právě jedné události je dána přibližně výrazem λh[15]. Pravděpodobnost toho, že v intervalu $(t, t+h\rangle$ dojde k více než jedné události považujeme téměř za nulovou[16]. Za těchto předpokladů uvažujeme náhodnou veličinu N_T vyjadřující počet událostí za dobu T. Pro N_T platí

$$\begin{aligned} &\textit{Pravděpodobnostní funkce}: p_k = P(N_T = k) = e^{-\lambda T}\frac{(\lambda T)^2}{k!}, k \ = 0, 1, 2, \dots \\ &\textit{Základní charakteristiky}: \quad E(X) = \lambda T \\ &\qquad\qquad\qquad\qquad\quad Var(X) = \lambda \end{aligned}$$

Toto rozdělení se nazývá **Poissonovo**[17], jeho jediný parametr λ lze interpretovat jako průměrný počet událostí v tomto modelu za jednotku času, λT je potom průměrný počet událostí za dobu T. Velký význam má Poissonovo rozdělení v teorii hromadné obsluhy, kde popisuje takové náhodné jevy jako jsou příchody zákazníků.

Poissonovo rozdělení bývá označováno jako rozdělení *řídkých jevů*, neboť se podle něj řídí četnosti jevů, které mají velmi malou pravděpodobnost výskytu. Interpretace parametru T jako času není podstatná. Často se Poissonovo rozdělení používá i pto rozdělení výskytu zrn nebo částic v objemu nějaké látky či materiálu. Potom T má význam objemu.

Poissonovo rozdělení se někdy též používá k aproximaci binomického rozdělení pro velký počet n pokusů, tzn. $n \to \infty$ a malou pravděpodobnost p výskytu sledovaného jevu v jednom pokusu, tzn. $p \to 0$. Potom pokládáme $\lambda = np$. Obvykle můžeme binomické rozdělení aproximovat Poissonovým tehdy, pokud $n > 30$ a $p \leq \frac{1}{10}$.

Příklad 1.3.6 *Jaká je pravděpodobnost, že do autoservisu přijedou*

a) alespoň dva zákazníci během pěti minut,

b) žádný zákazník během půl hodiny,

jestliže ve sledované době sem přijíždí průměrně čtyři zákazníci za hodinu?

---

[15] Přesněji $\lambda h + o(h)$, kde $o(h)$ je nekonečně malá veličina splňující $\lim_{h\to 0_+} \frac{o(h)}{h} = 0$.

[16] Přesněji za rovnu $o(h)$.

[17] Podle francouzského matematika, Laplaceova studenta Simeona Denise Poissona (1781-1840), který se kromě teorie pravděpodobnosti zabýval především fyzikou (elektřinou, magnetismem a optikou).

Řešení: Pokud zákazníci přijíždějí po jednom nezávisle na sobě v čase, můžeme na tento příklad použít Poissonův model. Ze zadání vyplývá, že je $\lambda = 4$, v případě a) $T = \frac{1}{12}$hodiny, v případě b) $T = \frac{1}{2}$hodiny. Tedy
a) $P(N_T \geq 2) = 1 - P(N_T < 2) = 1 - p_0 - p_1 = 1 - e^{-\frac{1}{3}}(1 + \frac{1}{3}) = 0,045$,
b) $P(N_T = 0) = 0,135$.

Příklad 1.3.7 *Do 10kg těsta bylo přidáno 1000ks hrozinek. Jeden kus pečiva se vyrobí z 10 dkg těsta. Jaká je pravděpodobnost, že v jednom kusu vyrobeného pečiva bude alespoň 10 hrozinek?*

Řešení: Je-li těsto s hrozinkami důkladně promícháno, budou se jednotlivé hrozinky v hmotnostním objemu těsta vyskytovat náhodně (rovnoměrně), přičemž známe průměrný počet hrozinek v 1 kg těsta - označme jej λ a je $\lambda = 100ks$. Počet hrozinem v pečivu je náhodná veličina a označme ji X. K výpočtu požadované pravděpodobnosti použijeme Poissonovo rozdělení s parametrem λT, kde $T = 0,1$ je hmotnost těsta potřebného k výrobě jednoho kusu pečiva. Odpověď na otázku jaká je $P(X \geq 10)$ dostaneme následujícím výpočtem:

$P(X \geq 10) = 1 - P(X < 10) = 1 - e^{-10} \sum_{k=0}^{9} \frac{10^k}{k!} = 0,5421.$

Příklad 1.3.8 *Vraťme se k jističům z příkladu 1.3.4. Jaká je pravděpodobnost, že z 1000 jističů nesepnou právě 3 kusy?*

Řešení: Použijeme-li binomický model (počet „úspěchů“ při n nezávislých pokusech), dostáváme

$P(X = 3) = \binom{1000}{3}(4,0405.10^{-5})^3(1 - 4,0405.10^{-5})^{997} = 1,0528.10^{-5}.$

Při použití aproximace Poissonovým rozdělením je $\lambda = 0,040405$ a pro náš příklad dostáváme o trochu lépe „spočítatelný“ výraz

$P(X = 3) \doteq \frac{4,0405^3}{3!} e^{-0,040405} 10^{-6} = 1,0559.10^{-5}.$

1.3.2 Spojité pravděpodobnostní modely

Model rovnoměrného rozdělení na intervalu $U(a, b)$**.** Náhodný pokus spočívá v náhodném výběru jednoho čísla z reálného intervalu $\langle a, b\rangle$. V tomto případě nemá smysl hovořit o tom, že každé číslo z intervalu $\langle a, b\rangle$ má stejnou pravděpodobnost být vybráno – to je splněno triviálně pro každý spojitý model (rozdělení pravděpodobnosti), neboť pro každou spojitou náhodnou veličinu X a každé $x \in R$ platí $P(X = x) = 0$. Rovnoměrnost v tomto modelu znamená to, že vezmeme-li libovolný podinterval $I_x(\triangle) = \langle x, x + \triangle\rangle$, potom

pravděpodobnost $P(X \in I_x(\triangle))$ bude stejná pro jakékoli $x \in \langle a, b-\triangle\rangle$ a bude rovna $\frac{\triangle}{b-a}$.

Hustota pravděpodobnosti : $f(x) = \frac{1}{b-a}$, pro $a \le x \le b$, $f(x) = 0$ jinde

Distribuční funkce $$F(x) = \begin{cases} 0 & \text{pro } x < a, \\ \frac{x-a}{b-a} & \text{pro } a \le x \le b, \\ 1 & \text{pro } x > b. \end{cases}$$

Základní charakteristiky : $E(X) = \frac{a+b}{2}$

$Var(X) = \frac{(b-a)^2}{12}$

Jednou z aplikací rovnoměrného rozdělení je vyšetřování zaokrouhlovacích chyb v numerických výpočtech. Při zaokrouhlení na k desetinných míst lze chybu považovat za náhodnou veličinu s rovnoměrným rozdělením na intervalu $< -5.10^{-k-1}, 5.10^{-k-1} >$.

Exponenciální model $Exp(\lambda)$**.** Uvažujme náhodný výskyt událostí v čase. Jak jsme uvedli dříve, počet takovýchto událostí v čase t je náhodná veličina $X(t)$, kterou lze modelovat Poissonovým modelem $Poiss(\lambda t)$, kde λ je průměrný počet událostí za jednotku času. Náhodná veličina Y odpovídající době mezi výskytem událostí v takovémto modelu má rozdělení pravděpodobnosti pro které platí $P(Y \le t) = 1 - P(Y > t) = 1 - P(X(t) = 0) = 1 - e^{-\lambda t}$. Rozdělení náhodné veličiny Y se nazývá **exponenciální rozdělení** s parametrem λ.

Hustota pravděpodobnosti: $$f(x) = \begin{cases} 0 & \text{pro } x < 0, \\ \lambda e^{-\lambda x} & \text{pro } x \ge 0, \end{cases}$$

Distribuční funkce $$F(x) = \begin{cases} 0 & \text{pro } x < 0, \\ 1 - e^{-\lambda x} & \text{pro } x \ge 0, \end{cases}$$

Základní charakteristiky : $E(X) = \frac{1}{\lambda}$

$Var(X) = \frac{1}{\lambda^2}$

V tomto modelu je střední doba mezi jednotlivými událostmi rovna $\frac{1}{\lambda}$, v některých aplikacích se exponenciální rozdělení uvádí s parametrem $\theta = \frac{1}{\lambda}$. Jedná se o stejné rozdělení pravděpodobnosti, pouze jeho hustota a distribuční funkce jsou zapsány v poněkud jiném tvaru:

Hustota pravděpodobnosti: $$f(x) = \begin{cases} 0 & \text{pro } x < 0, \\ \frac{1}{\theta} e^{-\frac{x}{\theta}} & \text{pro } x \ge 0, \end{cases}$$

Distribuční funkce $$F(x) = \begin{cases} 0 & \text{pro } x < 0, \\ 1 - e^{-\frac{x}{\theta}} & \text{pro } x \ge 0, \end{cases}$$

Základní charakteristiky : $E(X) = \theta$

$Var(X) = \theta^2$

Exponenciální rozdělení se používá především v modelech doby mezi událostmi jako jsou například poruchy zařízení, příchody zákazníků do obslužného systému, požadavky na zpracování náhodného signálu a podobně.

Zajímavou vlastností exponenciálního rozdělení je jeho „ztráta paměti“. Představme si elektronické zařízení (například síťovou kartu v počítači), jehož poruchy přicházejí náhodně v čase. Když je zařízení nové, můžeme spočítat například pravděpodobnost $p = P(Y > \tau)$, že přežije dobu τ. Pokud zařízení již pracovalo po nějakou (třeba i dlouhou) dobu T, pravděpodobnost že přežije dalších τ jednotek času je stejná jako na počátku jeho života, tedy p. Zařízení si „nepamatuje“ jak dlouho už pracuje. To lze vyjádřit v následujícím tvrzení:

Věta 1.7 *Má-li náhodná veličina X exponenciální rozdělení s nějakým paramterem λ, potom pro libovolná $T \geq 0, \tau > 0$ platí*

$$P(X > \tau) = P(X > T + \tau | X \geq T)$$

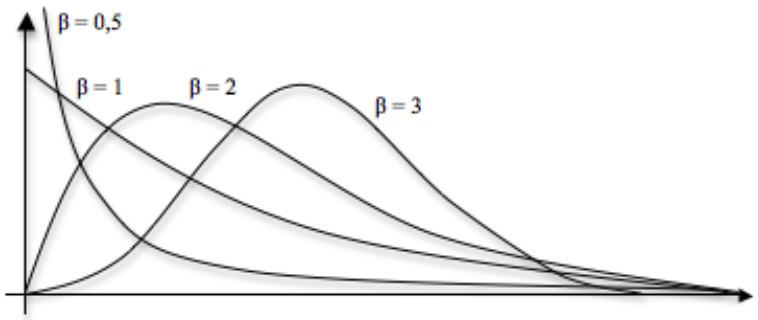
Důkaz: Důkaz zkuste provést sami jako cvičení.

Toto tvrzení platí i obráceně: pokud pro náhodnou veličinu X a libovolná $T \geq 0, \tau > 0$ platí uvedená rovnost, potom se tato veličina řídí exponenciálním rozdělením s nějakým parametrem λ.

Vlastnost „zapomínání“ exponenciálního rozdělení je velice užitečná pro výpočty, ale v praxi to znamená, že při modelování doby života technického zařízení nebereme do úvahy jeho „stárnutí“ či opotřebení. Proto se v řadě aplikací používá obecnější modelem pro délku života technických zařízení, který se nazývá Weibullův.

Weibullův model. Tento model se často používá jako model délky života technického zařízení.

Náhodná veličina X má **Weibullovo rozdělení** s parametry $\theta, \beta (X \approx W(\theta, \beta)$, je-li distribuční funkce rovna Grafy této hustoty pro pevné θ (parametr měřítka) a různé hodnoty β (parametr tvaru) jsou na obr. 1.9. Vzhledem k rozmanitosti tvaru se často užívá jako přibližný model v technických aplikacích. Pro $\beta > 1$ modeluje délku života zařízení, u něhož se pravděpodobnost poruchy s časem zvětšuje, v případě $\beta < 1$ modeluje životnost zařízení, u něhož se pravděpodobnost poruchy s časem zmenšuje. Je-li $\beta = 1$, dostáváme exponenciální model, jehož pravděpodobnost poruchy nezávisí na stáří (nemá paměť).



Obrázek 1.9: Hustoty Weibullova rozdělení pro různé hodnoty parametru β.

$$\textit{Hustota pravděpodobnosti: } f(x) = \begin{cases} 0 & \text{pro } x < 0, \\ \frac{\beta x^{\beta-1}}{\theta^\beta} e^{-\left(\frac{x}{\theta}\right)^\beta} & \text{pro } x \geq 0, \end{cases}$$

$$\textit{Distribuční funkce} \qquad F(x) = \begin{cases} 0 & \text{pro } x < 0, \\ 1 - e^{-\left(\frac{x}{\theta}\right)^\beta} & \text{pro } x \geq 0, \end{cases}$$

$$\textit{Základní charakteristiky}: \quad E(X) = \theta\Gamma(\tfrac{1}{\beta} + 1)$$

$$Var(X) = \theta^2\Gamma\left(\tfrac{2}{\beta} + 1\right) - \Gamma^2\left(\tfrac{1}{\beta} + 1\right)$$

n-tý obecný moment Weibullova rozdělení můžeme vyjádřit pomocí Gamma funkce[18]:

$$EX^n = \frac{\beta}{\theta^\beta} \int_0^\infty x^{\beta+n-1} e^{-\left(\frac{x}{\theta}\right)^\beta} dx = \theta^n \int_0^\infty z^{\frac{n}{\beta}} e^{-z} dz = \theta^n \Gamma\left(\frac{n}{\beta} + 1\right)$$

kde jsme užili substituci $z = \left(\frac{x}{\theta}\right)^\beta$

Uvažujeme-li množinu $\{X_i, i = 1, 2, \ldots, n\}$ nezávislých náhodných veličin, pak $\min_i\{x_i\}$ má za určitých podmínek pro velká n Weibullovo rozdělení.

Příklad 1.3.9 *Déka života Y oběžného kola turbiny je dána životností funkčně nejslabší lopatky. Nechť životnosti jednotlivých lopatek jsou nezávislé náhodné veličiny $X_i, i = 1, \ldots, n$, se stejným Weibullovým rozdělením s distribuční funkcí $F(x) = 1 - e^{(-\frac{x}{\theta})^\beta}, x \geq 0$. Najděte rozdělení pravděpodobnosti veličiny Y.*

[18]Funkce $\Gamma(x)$ je definována jako integrál $\int_0^\infty t^{x-1}e^{-t}dt$. Tento integrál nelze obecně (pro všechna x vyjádřit konečnou analytickou formulí a lze jej počítat pouze rozvojem v řadu. Pro $\Gamma(x)$ platí $\Gamma(x+1) = x\Gamma(x)$. Specielně pro přirozená n je $\Gamma(n) = (n-1)!$. Další užitečný vztah je $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

Řešení: Vypočteme distribuční funkci $G(y)$ náhodné doby života Y oběžného kola turbíny. Zřejmě je $Y = min(X_1, \ldots, X_n)$ a tedy

$$G(y) = P(Y \leq y) = P(min(X_1, \ldots, X_n) \leq y) = \\ 1 - P(min(X_1, \ldots X_n) > y) = 1 - P(\bigcap_{i=1}^{n} [X_i > y]).$$

Minimum z n čísel je totiž větší než y právě tehdy, když všechna čísla mají tuto vlastnost. Dále využijeme nezávislost a dostáváme

$$G(y) = 1 - \Pi_{i=1}^{n} P(X_i > y) = 1 - \Pi_{i=1}^{n}(1 - F(y)) = 1 - (1 - F(y))^n = 1 - e^{-n(\frac{x}{\theta})^\beta}.$$

Exponent lze zapsat ve tvaru $-n\left(\frac{x}{\theta}\right)^\beta = -\left(\frac{x}{\theta n^{-\frac{1}{\beta}}}\right)^\beta$. Vzhledem k jednoznačnosti vyjádření distribuční funkce daného rozdělení činíme závěr, že rozdělení životnosti oběžného kola je tedy opět Weibullovo, $Y \approx W(\theta n^{-\frac{1}{\beta}}, \beta)$ se stejným parametrem tvaru jako rozdělení životnosti lopatek.

Normální model $N(\mu, \sigma^2)$. Tento model bývá také označován jako model rozdělení chyb při měření.

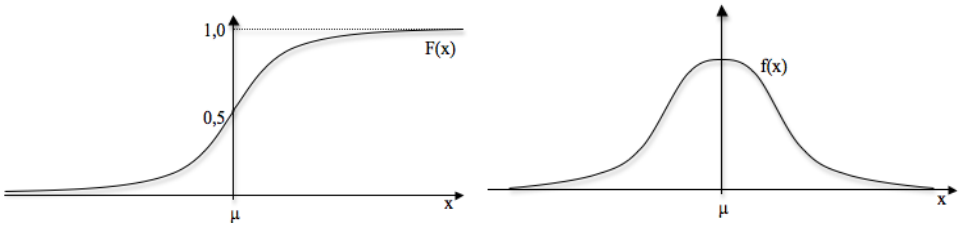
Model normálního rozdělení má bohatou historii. Postupně byl objevován a opět zapomínám, až se trvale dostal do poředí zájmu teorie pravděpodobnosti a především matematické statistiky. První, kdo popsal zvonovou křivku hustoty normálního rozdělení byl A. Moivre[19] v roce 1733. K normálnímu modelu se dostal zobecněním binomického modelu při házení mincí. V té době mu nikdo nevěnoval zvláštní pozornost a křivka i rovnice upadly v zapomenutí. Až na přelomu 18. a 19. století ji znovu „objevili“ Gauss[20] a Laplace[21] při

[19] Abraham de Moivre (1667–1754) byl francouzský matematik žijící větší část svého života v Anglii.

[20] Carl Friedrich Gauss (1777–1855)byl jeden z největších matematiků a fyziků všech dob. Zabýval se teorií čísel, matematickou analýzou, geometrií, geodézií, magnetismem, astronomií, optikou. Někdy bývá označován za „knížete matematiky“ nebo „největšího matematika od dob antiky“ – silně ovlivnil většinu oblastí svého oboru.

[21] Pierre Simon de Laplace (1749–1827) byl francouzský matematik, fyzik, astronom a politik; člen Francouzské akademie věd, královské společnosti v Londýně a Komise pro míry a váhy. Laplace je právem považován za jednoho z největších vědců vůbec. Zanechal monumentální dílo již svým rozsahem. Zabýval se matematickou analýzou, teorií pravděpodobnosti, nebeskou mechanikou, teorií potenciálu, zavedl pojem Laplaceovy transformace, užil tzv. Laplaceův operátor (v parciální diferenciální rovnici pro potenciál silového pole). Je autorem teorie o vzniku sluneční soustavy z rotující mlhoviny (Kantova-Laplaceova teorie) a mnoha dalších teorií a metod s mnoha aplikacemi

zkoumání astronomických měření. Byli postaveni před úlohu z mnoha měření, zatížených chybou, určit hodnotu, která se bude co nejvíce blížit skutečnosti. Odtud získal tento model přívlastek „model rozdělení chyb měření“ a odpovídající křivka hudtoty se někdy též nazývá „Gaussova“. Třetí, kdo tento model objevil a zároveň první, kdo jej nazval „normálním“, byl Quételet[22] v roce 1835. Normální křivku dostal v souvislosti s měřením obvodu prsou 5738 skotských vojáků a představou jakéhosi „normálního“, neboli průměrného jedince. Od té doby si model normálního rozdělení začal budovat svoji pevnou pozici ve všech oblastech vědy.



Obrázek 1.10: Distribuční funkce a hustota normálního rozdělení.

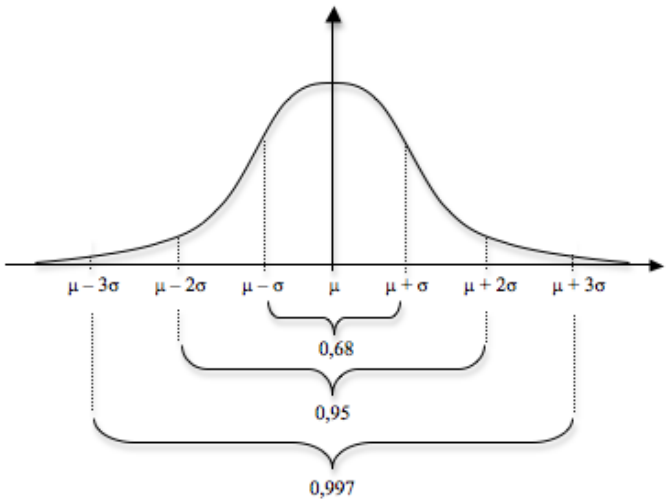
Hustota pravděpodobnosti: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in R$
Distribuční funkce $F(x) = \int_{-\infty}^{x} f(t)dt, \quad x \in R$
Základní charakteristiky : $E(X) = \mu$
$Var(X) = \sigma^2$

Poněkud nepříjemné v tomto modelu je to, že distribuční funkci, která je dána výše uvedeným integrálem, nelze vyjádřit konečnou analytickou formulí; její hodnoty se počítají pomocí distribuční funkce takzvaného *normovaného* normálního rozdělení (viz dále). V současné době to však není zásadní omezení, neboť řada programů (včetně tabulkového procesoru MS Excel) umějí distribuční funkci normálního rozdělení spočítat s dostatečnou přesností.

Parametry normálního rozdělení lze interpretovat jako *parametr polohy* $EY = \mu$ a *parametr měřítka* $VarY = \sigma^2$, vzhledem k symetrii rozdělení je μ též mediánem i modem. Význam směrodatné odchylky σ je ilustrován obr. 1.11, kde je znázorněna pravděpodobnost toho, že Y se liší od střední hodnoty μ v absolutní hodnotě o méně než $k\sigma, k = 1, 2, 3$.

K dalšímu výkladu potřebujeme následující tvrzení:

[22]Lambert Adolphe Jacques Qételet (1796–1874). Belgický vědec, jeden ze zakladatelů Královské statistické společnosti v Londýně



Obrázek 1.11: Vliv parametrů μ a σ normálního rozdělení na tvar křivky.

Věta 1.8 *Má-li náhodná veličina X normální rozdělení pravděpodobnosti se střední hodnotou μ a rozptylem σ^2 $(X \approx N(\mu, \sigma^2))$, potom pro libovolné konstanty $a, b \in R, a > 0$ má veličina $Y = \frac{X-b}{a}$ opět normální rozdělení se střední hodnotou $\mu - b$ a rozptylem $\left(\frac{\sigma}{a}\right)^2$, neboli patí $(Y \approx N\left(\mu - b, \left(\frac{\sigma}{a}\right)^2\right))$.*

Důkaz: Důkaz této věty vyplývá z věty 1.9 v následujícím odstavci.

Náhodná veličina Z má **normované normální rozdělení**, nebo též **standardní normální rozdělení**, $(Z \approx N(0,1))$, je-li její hustota rovna $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, z \in R$. Příslušná distribuční funkce se označuje obvykle symbolem Φ a lze ji vyjádřit jako

$$\Phi(z) = \frac{1}{2\pi} \int_{-\infty}^{z} e^{-\frac{t^2}{2}} dt.$$

Hodnoty této funkce se počítají rozvojem v řadu a integrací člen po členu, nebo jsou uvedeny v takzvaných „Statistických tabulkách“.

n-tý obecný moment pro liché $n = (2k-1)$ je vždy roven nule; tedy je $EZ^{(2k-1)} = 0, k \in N$. Pro n sudé, $n = 2k$ je $EZ^{2k} = \sqrt{\frac{2}{\pi}} \int_0^\infty z^{2k} e^{-\frac{z^2}{2}} dz$. Po substituci $\frac{z^2}{2} = t$ dostaneme[23]

$$EZ^{2k} = \sqrt{\frac{2}{\pi}} \int_0^\infty (2t)^{\frac{2k-1}{2}} e^{-t} dt = \frac{2^k}{\sqrt{(\pi)}} \Gamma(k+\frac{1}{2}) = (2k-1)!! = 1.3.5 \ldots (2k-1)$$

[23]Symbol $(2k-1)!!$ se používá k vyjádření „lichého“ faktoriálu, tedy součinu všech lichých čísel od 1 do $(2k-1)$.

Toto rozdělení se používá například tehdy, je-li třeba porovnat vlastnosti více náhodných veličin s různým normálním rozdělením. S takzvanou *normalizací* náhodné veličiny jsme se už setkali při definici *normovaných momentů* v odstavci 1.2. Jestliže má náhodná veličina X obecné normální rozdělení ($X \approx N(\mu, \sigma)$), vytvoříme normalizovanou náhodnou veličinu $Z = \frac{X-\mu}{\sigma}$. Tato veličina má normované normální rozdělení ($Z \approx N(0,1)$). Vztah mezi hustotou $f(x)$ veličiny X a hustotou $\phi(z)$ veličiny Z je následující:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{z^2}{2}} = \frac{1}{\sigma}\phi(z).$$

neboli

$$f(x) = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right).$$

Distribuční funkci $F(x)$ veličiny X lze vyjádřit podobně pomocí $\Phi(z)$, neboť platí

$$F(x) = \Phi(\frac{x-\mu}{\sigma}).$$

To lze snadno ověřit, dosadíme-li do integrálu pro $F(x)$ substituci $\frac{x-\mu}{\sigma} = z, dx = \sigma dz$.

Příklad 1.3.10 *Odhad vzdálenosti je zatížen systematickou chybou $a = -50$ (zkrácení v m) a náhodná chyb Y má normální rozdělení s nulovou střední hodnotou a se směrodatnou odchylkou $\sigma = 100m$. Jaká je pravděpodobnost, že*

a) odhad převýší skutečnou hodnotu,

b) odhad se nebude lišit od skutečné hodnoty o více než 50 m?

Řešení: Umístíme-li skutečnou hodnotu do počátku soustavy souřadnic, je odhadnutá hodnota X součtem systematické a náhodné chyby, tedy $X = Y + a$. Posunutím rozdělení Y o a dostáváme $X \approx N(-50, 100)$, označme distribuční funkci X jako $F(x)$. v příkladu a) počítáme

$P(X > 0) = 1 - P(X \geq 0) = 1 - F(0) = 1 - \Phi(\frac{50}{100}) = 0,309,$

v b) jde o $P(|X| < 50) = P([-50 < X] \cap [X < 50]) = P([X < 50] - [X < -50]) = F(50) - F(-50) = \Phi(1) - \Phi(0) = 0,341.$

1.3.3 Funkce náhodné veličiny

V řadě případů je třeba pracovat s náhodnou veličinou, která je vyjádřena jako funkce jiné náhodné veličiny. Nejjednodušším příkladem je *normalizovaná* náhodná veličina nebo náhodná veličina v příkladu 1.3.10.

Věta 1.9 *Mějme náhodnou veličinu X s distribuční funkcí $F(x), x \in R$. Označme $Y = aX + b$ její lineární transformaci pro libovolné konstanty $a, b \in R, a \neq 0$. Potom pro distribuční funkci $G(y)$ náhodné veličiny Y platí*

$$\begin{aligned} G(y) &= F(\frac{y-b}{a}) \quad pro \quad a > 0, \\ G(y) &= 1 - \lim_{x \to y^-} F(\frac{y-b}{a}) \quad pro \quad a < 0. \end{aligned}$$

Důkaz: Pro $a > 0$ je zřejmě

$$G(y) = P(Y \leq y) = P(aX + b \leq y) = P(X \leq \frac{y-b}{a}) = F(\frac{y-b}{a}).$$

Pokud je $a < 0$, nerovnost se po vynásobení výrazu v pravděpodobnosti číslem a obrací a dostáváme

$$G(y) = P(X \geq \frac{y-b}{a}) = 1 - P(X < \frac{y-b}{a}) = 1 - \lim_{x \to y^-} F(\frac{y-b}{a}).$$

Limita v posledním výrazu je důsledkem toho, že distribuční funkce nemusí být spojitá zleva (pro diskrétní rozdělení). Je-li F absolutně spojitá s hustotou $f(x)$ potom i G je absolutně spojitá s hustotou $g(y) = G'(y) = \frac{1}{|a|} f(\frac{y-b}{a})$ pro libovolné a.

Dále buď $Y = h(X)$ transformace obecnou reálnou funkcí h. Při označení z věty 1.9 je distribuční funkce $G(y) = P(Y \leq y) = P(h(X) \leq y)$. Dále specielně pro $F(x)$ diskrétní se skoky p_n v bodech x_n je

$$G(y) = \sum_{n: h(x_n) \leq y} p_n$$

a pro $F(x)$ absolutně spojitou s hudtotou $f(x)$

$$G(y) = \int_{x: h(x) \leq y} f(x) dx$$

Příklad 1.3.11 *Rozdělení náhodné veličiny X je dáno pravděpodobnostmi $P(X = 0) = \frac{1}{2}, P(X = 1) = \frac{1}{4}, P(X = 2) = \frac{1}{4}$. Najděte rozdělení náhodné veličiny $Y = (X - 1)^2$.*

Řešení: Náhodná veličina Y může nabývat pouze hodnot 0 nebo 1. Jejich pravděpodobnosti jsou $P(Y = 0) = P(X = 1) = \frac{1}{4}, P(Y = 1) = P(X = 0) + P(X = 2) = \frac{3}{4}$.

Příklad 1.3.12 *Najděte rozdělení spojité náhodné veličiny $Y = |X|$, kde X je náhodná veličina symetrická kolem 0.*

Řešení: Symetrie náhodné veličiny X kolem bodu 0 znamená, že její hustota $f(x)$ je sudá funkce, tj. $f(x) = f(-x)$ pro všechna $x \in R$. Potom je $F(x) = -F(-x) + C$ a z vlastností distribuční funkce plyne, že musí být $C = 1$. Pro $Y = |X|$ je potom distribuční funkce $G(y) = \int_{|x| \leq y} f(x)dx = F(y) - F(-y) = 2F(y) - 1,\ y \geq 0$.

Je-li funkce h ryze monotonní s derivací h', existuje inverzní funkce h^{-1} a potom je $G(y) = P(X \leq h^{-1}(y)) = F(h^{-1}(y))$ pro h klesající. Derivací podle proměnné x dostaneme vzorec pro **transformovanou hustotu** $g(y)$ **náhodné veličiny** $Y = h(X)$

$$g(y) = \frac{f(h^{-1}(y))}{h'(h^{-1}(y))}$$

Příklad 1.3.13 *Najděte hustotu $g(y)$ náhodné veličiny $Y = asin\Psi$ kde Ψ je náhodná fáze s rovnoměrným rozdělením $U(-\pi, \pi)$. Y je náhodná výchylka harmonického pohybu s konstantní amplitudou a.*

Řešení: Hustota Ψ je $f(x) = \frac{1}{2\pi}, -\pi < x < \pi$, inverzní transformace $h^{-1}(y) = arcsin\frac{y}{a}, -a < y < a$. Ve jmenovateli transformačního vzorce dostáváme $h'(h^{-1}(y)) = acosarcsin\frac{y}{a} = \sqrt{a^2 - y^2}$. Odsud $g(y) = \frac{1}{2\pi}\frac{1}{\sqrt{a^2-y^2}}, -a < y < a$. „Nejpravděpodobnější“ jsou výchylky v okolí amplitud, kde je hustota neomezená funkce.

Příklad 1.3.14 *Nechť $Z \approx N(0, 1)$. Spočtěte hustotu pravděpodobnosti náhodné veličiny $X = Z^2$.*

Řešení: Zde je $h(z) = z^2$ a opačně, $h^{-1}(z) = \sqrt(z)$ pro $z \geq 0$. Pro transformovanou hustotu dostáváme

$$h(x) = \frac{1}{\sqrt{2x\pi}} e^{-\frac{x}{2}}.$$

Rozdělení náhodné veličiny $X = Z^2$ se nazývá **chí-kvadrát rozdělení** $\chi^2(1)$. Toto rozdělení se často používá ve statistice jako rozdělení součtu n druhých mocnin náhodných veličin s rozdělením $N(0,1)$.

Nechť $X \approx N(\mu, \sigma)$. Uvažujme rozdělení pravděpodobnosti náhodné veličiny $Y = e^X$. Transformace v tomto případě má tvar $h(x) = e^x$, což je spojitá, prostá a diferencovatelná funkce, pro $y > 0$ je navíc $h^{-1}(y) = lny$. Dosazením do předchozího vzorce tedy dostaneme pro $y > 0$

$$g(y) = \frac{1}{y\sigma\sqrt{(2\pi)}} e^{-\frac{(lny-\mu)^2}{2\sigma^2}}.$$

Případ, kdy by $y \leq 0$ je nemožný (exponenciela má pouze kladné hodnoty) a proto položíme

$$g(y) = 0 \text{ pro } y < 0.$$

Náhodná veličina Y má **logaritmicko-normální rozdělení** $LN(\mu, \sigma^2)$.

Základní charakteristiky : $E(X) = e^{\mu+\frac{\sigma^2}{2}}$
$Var(X) = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)$

Obráceně: má-li náhodná veličina X rozdělení $LN(\mu, \sigma^2)$, potom náhodná veličina $Y = lnX$ má rozdělení $N(\mu, \sigma^2)$.

Logaritmicko-normální rozdělení se používá v teorii spolehlivosti, ve vodohospodářství při modelování průtoků vody v řekách, při popisu velikosti částic sypkých metriálů a pod.

Pro výpočet střední hodnoty funkce náhodné veličiny není třeba transformavat původní rozdělení. Jednodušší je použít přímého vzorce

$$Eh(X) = \sum_j h(x_j)p_j$$

pro diskrétní náhodnou veličinu a

$$Eh(X) = \int h(x)f(x)dx$$

pro spojité rozdělení s hustotou f.

Příklad 1.3.15 *Opotřebení Z náprav železničních vagonů se zvyšuje s druhou mocninou zatížení X, tedy $Z = X^2$. Předpokládejme, že vagon jede po velmi nerovné trati, takže zatížení kolísá s diskrétním rozdělením s pravděpodobnostmi $P(X = 8) = 0,25, P(X = 25) = 0,7, P(X = 80) = 0,05$. Vypočtěte střední opotřebení a porovnejte ho s opotřebením při konstantním zatížení 25.*

Řešení: Užitím prvního vzorce dostáváme $EZ = \sum_j x_j^2 p_j = 773,5$. Při konstantním zatížení je opotřebení $25^2 = 625$, tedy menší.

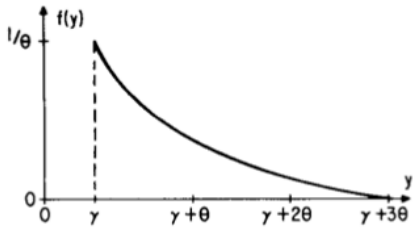
1.3.4 Posunutá, podmíněná a useknutá rozdělení

Příklad 1.3.16 *Náhodná veličina Y popisuje dobu potřebnou k vykonání určité technologické operace. Ta sestává ze dvou částí: první z nich trvá vždy stejnou dobu γ, druhá je náhodná, s exponenciálním rozdělením $Exp\left(\frac{1}{\theta}\right)$. Popište rozdělení pravděpodobnosti veličiny Y.*

Řešení: Hustota pravděpodobnosti veličiny Y je na obrázku 1.12. Tuto funkci lze popsat vztahem

$$f(y) = \begin{cases} 0 & \text{pro } y < \gamma, \\ \frac{1}{\theta} e^{-\frac{y-\gamma}{\theta}} & \text{pro } y \geq \gamma. \end{cases}$$

Jedná se o takzvané **posunuté rozdělení**. Jeho střední hodnota je posunutá



Obrázek 1.12: Graf hustoty posunutého exponenciálního rozdělení.

o γ, tedy $EX = \theta + \gamma$, zatímco rozptyl se posunutím nemění a je tedy roven $VarX = \theta^2$.

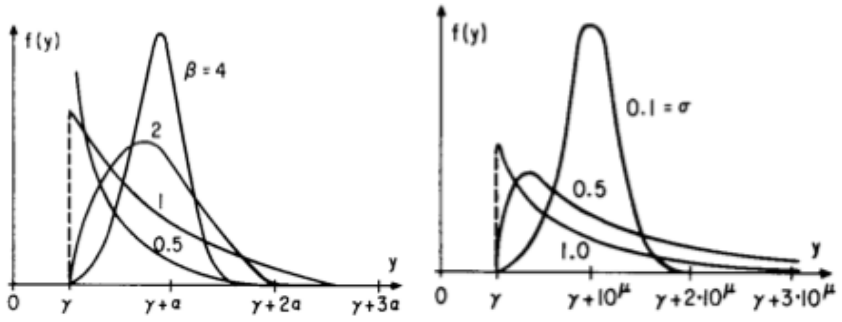
Příklad 1.3.17 *Najděte hustotu posunutého rozdělení ve Weibullově modelu a pro logaritmicko-normální pravděpodobnostní model.*

Řešení: Pro hustotu **posunutého Weibullova rozdělení** do bodu γ dostáváme vztah

$$f(y) = \begin{cases} 0 & \text{pro } y < \gamma, \\ \frac{\beta(y-\gamma)^{\beta-1}}{\theta^\beta} e^{-\left(\frac{(y-\gamma)}{\theta}\right)^\beta} & \text{pro } y \geq \gamma, \end{cases}$$

v případě **posunutého logaritmicko-normálního rozdělení** má posunutá hustota tvar

$$f(x) = \begin{cases} 0 & \text{pro } y < \gamma, \\ \frac{1}{(y-\gamma)\sigma\sqrt{(2\pi)}} e^{-\frac{(ln(y-\gamma)-\mu)^2}{2\sigma^2}} & \text{pro } y \geq \gamma, \end{cases}$$

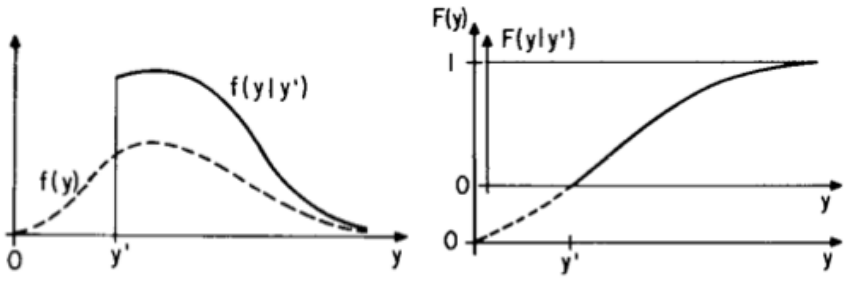


Obrázek 1.13: Posunuté Weibullovo (vlevo) a logaritmicko-normální (napravo) rozdělení.

Příklad 1.3.18 *Jeké je rozdělení délky života Y zařízení, když víme, že se dožilo doby y'?*

Řešení: Hledané rozdělení je takzvané **podmíněné rozdělení**. Označíme-li hustotu veličiny Y jako $f(y)$ a její distribuční funkci $F(y)$, potom hustota $f(y|y')$ a distribuční funkce $F(y|y')$ podmíněného rozdělení Y za podmínky $Y > y'$ jsou dány vztahy

$$f(y|y') = \begin{cases} 0 & \text{pro } y < y', \\ \frac{f(y)}{1-F(y')} & \text{pro } y \geq y', \end{cases} \qquad F(y|y') = \begin{cases} 0 & \text{pro } y < y', \\ \frac{F(y)-F(y')}{1-F(y')} & \text{pro } y \geq y'. \end{cases}$$



Obrázek 1.14: Hustota (vlevo) a distribuční funkce (napravo) podmíněného rozdělení.

Příklad 1.3.19 *Jaké je rozdělení zbývající doby života zařízení, když se dožilo doby y'?*

Řešení: V tomto případě se počátek měření času posouvá do okamžiku y', od kdy se zbývající doba života začíná měřit. **Rozdělení zbývající doby**

života je tedy něco jiného, než podmíněné rozdělení z předchozího příkladu. Hustotu $g(y)$ a distribuční funkci $G(y)$ zbývající doby života dostaneme posunutím podmíněné rozdělení do bodu y':

$$g(y) = \begin{cases} 0 & \text{pro } y < 0, \\ \frac{f(y+y')}{1-F(y')} & \text{pro } y \geq 0, \end{cases} \qquad G(y) = \begin{cases} 0 & \text{pro } y < 0, \\ \frac{F(y+y')-F(y')}{1-F(y')} & \text{pro } y \geq 0. \end{cases}$$

Příklad 1.3.20 *Automat dávkuje tekutinu, jejíž objem má být μ_0. v okolí μ_0 se veličina X, popisující objem dávky, chová jako náhodná veličina s normálním rozdělením $N(\mu_0, \sigma^2)$. Jaké je rozdělení pravděpodobnosti náhodné veličiny X?*

Řešení: Zřejmě lze předpokládat, že objem dávky X bude vždy pouze nezáporné číslo, může být i nulový. Problém je v tom, že normální rozdělení pravděpodobnosti dává kladnou pravděpodobnost i záporným hodnotám, které jsou v tomto případě nemožné.

Je-li $\mu_0 > 3\sigma$, je pravděpodobnost jevu $\{X < 0\}$ prakticky nulová a rozdělení X zpravidla považujeme za normální s parametry μ_0 a σ^2. Pokud je ale $0 < \mu_0 < 3\sigma$, potom by mohla být chyba v předpokladu normálního rozdělení příliš velká. Musíme tedy použít takzvané **useknuté normální rozdělení**. To je rozdělení, jehož hustota má pro $x > 0$ tvar zvonové křivky normálního rozdělení, pro $x \leq 0$ je nulová a přesto plocha pod ní zůstává rovna 1. Tvar této hustoty a jí odpovídající distribuční funkci je

$$g(x) = \begin{cases} 0 & \text{pro } x < 0, \\ \frac{f(x)}{1-F(0)} & \text{pro } x \geq 0, \end{cases} \qquad G(x) = \begin{cases} 0 & \text{pro } x < 0, \\ \frac{F(x)-F(0)}{1-F(0)} & \text{pro } x \geq 0, \end{cases}$$

kde $f(x)$, resp. $F(x)$ je hustota, resp. distribuční funkce rozdělení $N(\mu_0, \sigma^2)$. Všimněte si, že se vlastně jedná o podmíněné normální rozdělení za podmínky $\{X \geq 0\}$.

Pro střední hodnotu useknutého rozdělení potom zřejmě platí

$$EX = \int_{-\infty}^{\infty} x g(x) dx = \frac{\int_0^{\infty} x f(x) dx}{1 - F(0)}$$

a pro rozptyl

$$VarX = \frac{1}{(1 - F(0))^2} \left[\int_0^{\infty} x^2 f(x) dx - \left(\int_0^{\infty} x f(x) dx \right)^2 \right]$$

1.3.5 Rozdělení součtu nezávislých náhodných veličin

Uvažujme dvě nezávislé náhodné veličiny, X a Y, první s distribuční funkcí $F(x)$, druhou s distribuční funkcí $G(y)$. Bude nás zajímat rozdělení pravděpodobnosti jejich součtu, kterým je opět náhodná veličina $Z = X+Y$. Označme distribuční funkci Z jako $H(z)$.

V případě diskrétní náhodné veličiny vzhledem k nezávislosti X a Y zřejmě platí

$$H(z) = P(Z \leq z) = P(X + Y \leq z) = \sum \sum_{x+y \leq z} P(X = x)P(Y = y).$$

Pro spojité náhodné veličiny s hustotami $f(x)$ a $g(y)$ je

$$H(z) = \iint_{x+y \leq z} f(x)g(y)dxdy = \int_{-\infty}^{\infty} f(x) \int_{-\infty}^{z-x} g(y)dydx = \int_{-\infty}^{\infty} f(x)G(z-x)dx$$

Výsledné rozdělení pravděpodobnosti se nazývá *konvolucí* rozdělení X a Y.

Příklad 1.3.21 *Předpokládejme, že doba obsluhy zákazníka v systému hromadné obsluhy má exponenciální rozdělení pravděpodobnosti $Exp(\lambda)$ a obsluhy jednotlivých zákazníků jsou nezávislé. Jaké je rozdělení doby čekání zákazníka, který přijde do systému, v němž jsou už dva zákazníci?*

Řešení: Doba čekání příchozího zákazníka je součtem dob obsluhy dvou předchozích. Tedy hledáme rozdělení součtu dvou nezávislých náhodných veličin s exponenciálním rozdělením s parametrem λ.

$$H(z) = \iint_{x+y \leq z} \lambda^2 e^{-\lambda x} e^{-\lambda y} dxdy = \int_0^z \lambda e^{-\lambda x} \int_0^{z-x} \lambda e^{-\lambda(z-x)}) dydx = \int_0^z \lambda e^{-\lambda x}(1 - e^{-\lambda(z-x)})dx = 1 - e^{-\lambda z} - \lambda z\ e^{-\lambda z}$$

Rozdělení s touto distribuční funkcí se nazývá **Erlangovo rozdělení** s parametry 1 a λ. Obecně je Erlangovo rozdělení $Erl(n, \lambda)$ modelem součtu n nezávislých náhodných veličin s exponenciálním rozdělením $Exp(\lambda)$ a jeho charakteristiky jsou

Hustota pravděpodobnosti: $f(x) = \begin{cases} 0 & \text{pro } x < 0, \\ \lambda e^{-\lambda x} \frac{(\lambda x)^{n-1}}{(n-1)!} & \text{pro } x \geq 0, \end{cases}$

Distribuční funkce $F(x) = \begin{cases} 0 & \text{pro } x < 0, \\ 1 - e^{-\lambda x} \sum_{k=0}^{n-1} \frac{(\lambda x)^k}{k!} & \text{pro } x \geq 0, \end{cases}$

Základní charakteristiky : $E(X) = \frac{n}{\lambda}$

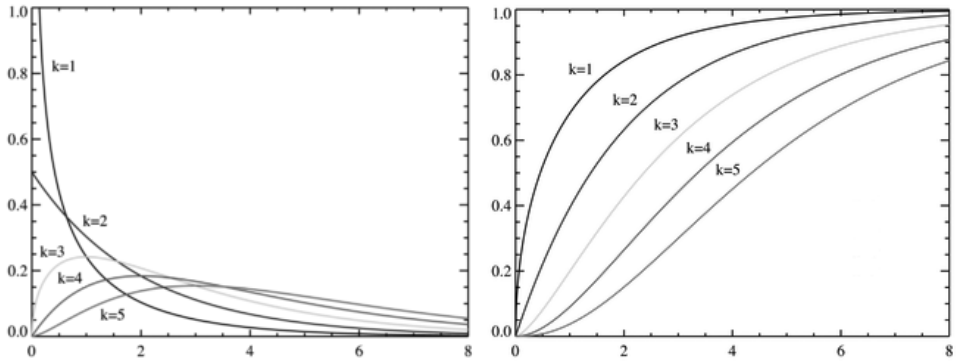
$Var(X) = \frac{n}{\lambda^2}$

Příklad 1.3.22 *Najděte rozdělení součtu druhých mocnin nezávislých náhodných veličin, které mají standardní normální rozdělení* $N(0,1)$.

Řešení: Rozdělení náhodné veličiny $X = Z^2$ jsme spočetli už v příkladu 1.3.14 kde jsme dostali hustotu $h(x) = \frac{1}{\sqrt{2x\pi}} e^{-\frac{x}{2}}$ pro $x \geq 0$. Pro posloupnost $\{Z_1, Z_2, \ldots, Z_n\}$ nezávislých náhodných veličin s rozdělením $N(0,1)$ má náhodná veličina $X_n^2 = \sum_{k=1}^{n} Z_k^2$ rozdělení pravděpodobnosti **chí-kvadrát** $\chi^2(n)$. Toto rozdělení se často používá v matematické statistice a ve spolehlivosti. Počet sčítanců n je parametr tohoto rozdělení (takzvané *stupně volnosti*).

Výpočet distribuční funkce a hustoty tohoto rozdělení je poměrně složitý a nebudeme se jím zde zabývat. Pro představu si uvedeme grafy hustoty a distribuční funkce.

Základní charakteristiky rozdělení $\chi^2(n)$ jsou: $E(X_n^2) = n, \quad VarX_n^2 = 2n.$



Obrázek 1.15: Hustota (vlevo) a distribuční funkce (napravo) rozdělení $\chi^2(k)$ pro různá k.

1.3.6 Momentová vytvořující funkce

Momentová vytvořující funkce $M_X(t)$ náhodné veličiny X je definována jako střední hodnota

$$M_X(t) = Ee^{tX}, t \in R$$

Věta 1.10 *Je-li $M_X(t)$ konečná v intervalu $< -b, b >$ pro nějaké $b > 0$, je*

$$EX^n = M_X^{(n)}(0), n = 0, 1, \ldots$$

kde $M_X^(n)$ značí n-tou derivaci funkce. Je-li $M_Y(t) = M_X(t)$ pro $|t| < b$, potom náhodné veličiny X a Y mají stejné rozdělení.

Naznačíme důkaz prvního tvrzení věty. Užitím rozvoje $e^{tX} = \sum_{k=0}^{\infty} \frac{(tX)^k}{k!}$ počítáme střední hodnotu $Ee^{tX} = \sum_{k=0}^{\infty} \frac{t^{k-n}}{(k-n)!} EX^k$, v bodě 0 je jediný nenulový člen této řady pro $k = n$ roven EX^n.

Příklad 1.3.23 *Spočtěte momentovou vytvořující funkci veličiny $Y = a + bX$, jestliže $X \approx N(0, 1)$.*

Řešení: Je-li $M_X(t)$ momentová vytvořující funkce X, potom

$$M_Y(t) = Ee^{(a+bX)t} = e^{at} Ee^{btX} = e^{at} M_X(tb).$$

Pro $X \approx N(0, 1)$ dostáváme

$$M_X(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{i} nftye^{tx} e^{-\frac{x^2}{2}} dx = e^{\frac{t^2}{2}}.$$

Z obou získaných poznatků plyne, že pro náhodnou veličinu $Y \approx N(\mu, \sigma)$ je $M_Y(t) = e^{\mu t} e^{\frac{\sigma^2 t^2}{2}}$.

Příklad 1.3.24 *Najděte obecné momenty náhodné veličiny $X \approx U(0, 1)$.*

Řešení: Momentová vytvořující funkce rovnoměrně rozdělené náhodné veličiny $X \approx U(0, 1)$ nesplňuje předpoklady věty 1.10: $M_X(t) = \int_0^1 e^{tx} dx = \frac{e^t - 1}{t}$. Přesto lze najít momenty X srovnáním s řadou v důkazu věty 1.10:

$$\frac{e^t - 1}{t} = \sum_{k=1}^{\infty} \frac{t^{k-1}}{k!} = \sum_{k=0}^{\infty} \frac{t^k}{k!} \frac{1}{k+1},$$

odsud $EX^k = \frac{1}{k+1}$

Věta 1.11 *Jsou-li X, Y nezávislé, je momentová vytvořující funkce součtu $Z = X + Y$ rovna součinu momentových vytvořujících funkcí složek, tj.*

$$M_Z(t) = M_{X+Y}(t) = M_X(t) M_Y(t)$$

Tato věta umožňuje stanovit rozdělení součtu nezávislých náhodných veličin mnohdy jednodušším způsobem než pomocí hustot pravděpodobnosti.

Příklad 1.3.25 *Dokažte, že rozdělení součtu dvou nezávislých náhodných veličin s normálním rozdělením je opět normální.*

Řešení: Buď $X \approx N(\mu_1, \sigma_1^2), Y \approx N(\mu_2, \sigma_2^2)$ nezávislé. Potom

$$M_{X+Y}(t) = M_X(t)M_Y(t) = e^{\mu_1 t + \frac{\sigma_1^2 t^2}{2}} e^{\mu_2 t + \frac{\sigma_2^2 t^2}{2}} = e^{(\mu_1+\mu_2)t + \frac{(\sigma_1^2+\sigma_2^2)t^2}{2}}$$

Podle druhé části věty 1.10 (vzájemná jednoznačnost přiřazení momentové vytvořující funkce rozdělení náhodné veličiny) musí mít náhodná veličina $Z = X + Y$ normální rozdělení $N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Příklad 1.3.26 *Nechť náhodné veličiny $X_i, i = 1, \ldots, n$ jsou nezávislé stejně rozdělené s alternativním rozdělením s parametrem p. Větu 1.11 lze rozšířit na konečný počet sčítanců. Je-li $Y = \sum_{i=1}^n X_i$, je $M_Y(t) = \Pi_{i=1}^n M_{X_i}(t)$. Podle binomického modelu je $Y \approx Bin(n, p)$*

Momentová vytvořující funkce X_i má tvar $M_{x_i} = Ee^{X_i t} = pe^t + (1 - p)$, tedy momentová vytvořující funkce binomického rozdělení je

$$M_Y(t) = (pe^t + 1 - p)^n$$

Podle věty 1.10 lze počítat např. $EY = M'_Y(0) = n(pe^0 + 1 - p)^{n-1}p = np$.

Jsou-li $X \approx Bin(n_1, p), Y \approx Bin(n_2, p)$ dvě nezávislé náhodné veličiny, je $M_{X+Y}(t) = (pe^t + 1 - p)^{n_1+n_2}$, tedy rozdělení součtu je opět binomické, $X + Y \approx Bin(n_1 + n_2, p)$. Tento výsledek je zřejmý z interpretace, přidáme-li k n_1 nezávislým opakováním pokusu se dvěma výsledky (zdar, nezdar) dalších n_2 nezávislých opakování, je počet výskytů zdarů v serii délky $n_1 + n_2$ roven součtu zdarů v obou seriích délky n_1 resp. n_2.

1.4 Náhodný vektor

Zobrazení $\mathbb{X} : \Omega \to R^n$, jehož jednotlivé složky $X_i, i = 1, 2, \ldots, n$ jsou náhodné veličiny, budeme nazývat **náhodným vektorem**. Náhodný vektor si tedy lze představit jako **vektor náhodných veličin**. v následujícím textu se budeme zabývat pouze případem $n = 2$.

1.4.1 Rozdělení náhodného vektoru

Sdružená distribuční funkce náhodného vektoru (X, Y) je definována předpisem

$$F(x, y) = P(X \leq x, Y \leq)$$

a má následující vlastnosti:

a) $0 \leq F(x, y) \leq 1$ pro každé $(x, y) \in R^2$,

b) $\lim_{\substack{x\to\infty \\ y\to\infty}} F(x, y) = 1$,

c) $\lim_{x\to-\infty} F(x, y) = \lim_{y\to-\infty} F(x, y) = 0$,

d) F je zprava spojitá v každé proměnné.

Uvažujme náhodný vektor $\mathbb{Z} = (X, Y)$. Jestliže X a Y jsou diskrétní náhodné veličiny, označme $\{x_i\}$ resp. $\{y_j\}$ konečné nebo spočetné posloupnosti všech hodnot X resp. Y. Dále označme $P(X = x_i, Y = y_j) = p_{ij}$. Potom posloupnost $\{p_{ij}\}$ tvoří **diskrétní sdružené rozdělení pravděpodobnosti** náhodného vektoru $\mathbb{Z}$. Musí samozřejmě platit $\sum_{i,j} p_{ij} = 1$.

Sdružená distribuční funkce $F(x, y)$ diskrétního náhodného vektoru $\mathbb{Z}$ se nazývá **diskrétní** a spočteme ji podle vztahu

$$F(x, y) = \sum_{\substack{i:x_i\leq x \\ j:y_j\leq y}} p_{ij}$$

Příklad 1.4.1 *Výrobky jsou produkovány ve třech jakostních kategoriích: třída I s pravděpodobností 0,3, v kategorii II s pravděpodobností 0,5 a v kategorii III s pravděpodobností 0,2. Jaké je rozdělení počtu výrobků I. a II. kategorie?*

Řešení: Náhodným pokusem zde je výroba n výrobků v různých jakostních kategoriích. Sledujeme veličiny X=[počet výrobků I. kategorie] a Y=[počet výrobků II. kategorie]. Počet výrobků II. kategorie je potom doplněk do celkového počtu n vyrobených. Podle klasické definice pravděpodobnosti lze odvodit následující vztah

$$P(X = k, Y = l) = p_{kl} = \frac{2!}{k!l!(2 - k \ - l)!}(0,3)^k(0,5)^l.$$

Obecně, může-li jeden pokus skončit některým ze tří možných výsledků $\{\omega_1, \omega_2, \omega_3\}$ s pravděpodobnostmi p_1, p_2 a $(1 - p_1 - p_2)$ a označíme-li X, Y počty výsledků $\{\omega_1, \omega_2\}$ v n nezávislých opakováních tohoto pokusu, potom rozdělení pravděpodobnosti náhodného vektoru (X, Y) se nazývá **multinomické** s parametry n, p_1, p_2 a platí

$$P(X = k, Y = l) = p_{kl} = \frac{2!}{k!l!(2 - k \ - l)!}p_1^k p_2^l.$$

Tento příklad lze zobecnit pro k možných výsledků s pravděpodobnostmi $p_i, i = 1, \dots k$. V n nezávislých opakováních pokusu buď X_i počet výskytů i-tého výsledku. Sdružené rozdělení pravděpodobnosti náhodného vektoru $(X_1, \dots, X_k)$ je **k-rozměrné multinomické rozdělením** definované předpisem

$$P(X_1 = x_1, \dots X_k = x_k) = \frac{n!}{x_1!x_2! \dots x_k!}p_1^{x_1} \dots p_k^{x_k}$$

pro x_i splňující $\sum_{i=1}^k x_i = n$. V případě $k = 2$ se jedná o binomické rozdělení.

Diskrétní rozdělení náhodného vektoru se často vyjadřuje tabulkou. Význam jednotlivých buněk v tabulce takovéhoto rozdělení je následující:

		Y			
		y_1	$\dots$	y_n	$\sum$
X	x_1	p_{11}	$\dots$	p_{1n}	p_1^X
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	x_m	p_{m1}	$\dots$	p_{mn}	p_m^X
	$\sum$	p_1^Y	$\dots$	p_n^Y	1

kde $p_{ij} = P(X = x_i, Y = y_j)$ a $p_i^X = P(X = x_i)$, resp. $p_j^Y = P(Y = y_j)$.

Příklad 1.4.2 *Podle statistického zkoumání zaměstnanosti mužů a žen v určitém regionu byly získány tyto výsledku: mezi muži je 12% nezaměstnaných, mezi ženami je nezaměstnaných 16%. Sestavte tabulku sdruženého rozdělení vektoru (X, Y), kde X reprezentuje pohlaví a Y zaměstnanost. Předpokládejme, že ve sledovaném regionu tvoří část mužské populace 48%.*

Řešení: Nechť jev $\{X = 0\}$ znamená, že náhodně vybraný člověk je muž, $\{X = 1\}$ že je žena. Podobně, jev $\{Y = 0\}$ bude znamenat nezaměstnanost, jev $\{Y = 1\}$ zaměstnanost. Zadané pravděpodobnosti lze interpretovat jako $P(Y = 0|X = 0) = 0,12$, $P(Y = 0|X = 1) = 0,16$ a $P(X = 0) = 0,48$. Potom je

$$\begin{aligned} p_{00} &= P(Y = 0|X = 0).P(X = 0) = 0,12.0,48 = 0,0576; \\ p_{01} &= P(Y = 1|X = 0).P(X = 0) = 0,88.0,48 = 0,4224; \\ p_{10} &= P(Y = 0|X = 1).P(X = 1) = 0,16.0,52 = 0,0832; \\ p_{11} &= P(Y = 1|X = 1).P(X = 1) = 0,84.0,52 = 0,4368; \end{aligned}$$

Hledaná tabulka má tedy tvar

		Y		
		0	1	$\sum$
X	0	0,0576	0,4224	0,48
	1	0,0832	0,4368	0,52
	$\sum$	0,1408	0,8592	1

Distribuční funkce $F(x, y)$ se nazývá **absolutně spojitá**, jestliže existuje nezáporná funkce $f(x, y)$ nazývá **sdružená hustota pravděpodobnosti**, stručně hustota, taková, že

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v)dudv.$$

Zřejmě musí platit $\int \int_{R^2} f(x, y)dxdy = 1$. Vztah mezi spojitou distribuční funkcí a její hustotou lze vyjádřit pomocí parciálních derivací

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}$$

pokud derivace F existuje. Říkáme, že náhodný vektor s absolutně spojitou distribuční funkcí má **spojité rozdělení**.

Pravděpodobnost $P(x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2)$ je pomocí distribuční funkce vyjádřena jako

$$F(x_2, y_2) - F(x_1, y_2) - F(x_2, y_1) + F(x_1, y_1)$$

pro každé $x_1 \leq x_2,\ y_1 \leq y_2$. Ve spojitém případě je

$$P(x_1 < X \leq x_2, y_1 < Y \leq y_2) = \int_{x_1}^{x_2} \int_{y_1}^{y_2} f(x, y)dxdy.$$

Obr.III.1: Graf hustoty dvourozměrného rozdělení, pravděpodobnost $P(x_1 < X < x_2, y_1 < Y < y_2)$.

Příklad 1.4.3 *Na obr.* **??** *je graf hustoty* $f(x,y) = \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}}$, $(x,y) \in R^2$ *dvourozměrného normálního rozdělení.*

Graf spojité hustoty (X,Y) tvoří plochu v R^3. Tento model může sloužit např. jako rozdělení náhodné chyby odhadu velikosti dvourozměrného objektu. V grafu je znázorněna pravděpodobnost $P(x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2)$, kterou lze interpretovat jako objem tělesa shora ohraničeného plochou hustoty nad obdélnikovou podstavou $\langle x_1, x_2) \times \langle y_1, y_2)$.

Nechť $F(x,y)$ je sdružená distribuční funkce náhodného vektoru (X,Y). Potom **marginální distribuční funkcí** jeho složky X nazveme distribuční funkci

$$F^X(x) = P(X \leq x) = \lim_{y \to \infty} F(x,y)$$

a podobně $F^Y(y) = P(Y \leq y) = \lim_{x \to \infty} F(x,y)$

Marginální rozdělení je tedy obyčejné rozdělení náhodné veličiny. V případě nutnosti rozlišení se užívá přívlastek sdružené rozdělení (pravděpodobnost, hustota) pro náhodný vektor a marginální rozdělení (pravděpodobnost, hustota) pro jeho složky. Jinak je možné tento přívlastek vynechat.

Je-li F diskrétní, jsou **marginální pravděpodobnosti** rovny

$$P(X = x_i) = p_i^X = \sum_j p_{ij}, \quad P(Y = y_j) = p_j^Y = \sum_i p_{ij},$$

v tabulce příkadu 1.4.2 jsou umístěny na dolním resp. pravém okraji (angl. margin).

U náhodného vektoru se spojitým rozdělením je **marginální hustota** $f_X(x)$ resp. $f_Y(y)$ složky X resp. Y rovna

$$f^X(x) = \int_{-\infty}^{\infty} f(x,y)dy, \quad f^Y(y) = \int_{-\infty}^{\infty} f(x,y)dx.$$

Příklad 1.4.4 *Uvažujme náhodný vektor* X, Y *s rovnoměrným rozdělením pravděpodobnosti nad jednotkovým kruhem. Spočtěte marginální rozdělení jeho složek.*

Řešení: Náhodný vektor (X,Y) nabývá hodnot z jednotkového kruhu $K = \{(x,y) \in R^2 : x^2 + y^2 = 1\}$, tedy bude se zřejmě jednat o spojitý náhodný vektor. Rovnoměrné rozdělení v tomto případě znamená konstantní hustotu $f(x,y) = c$, takovou, že musí platit $\iint_K f(x,y)dxdy = \iint_K cdxdy = c\pi = 1$.

Odtud dostáváme $f(x,y) = c = \frac{1}{\pi}$ pro všechna $(x,y) \in K$ a $f(x,y) = 0$ jinde.

Marginální hustotu pro X dostaneme integrací

$$f^X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{1}{\pi} dy = \frac{2}{\pi}\sqrt{1-x^2}, \quad x \in \langle -1, 1\rangle$$

Podobně je i

$$f^Y(y) = \int_{-\infty}^{\infty} f(x,y)dx = \int_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}} \frac{1}{\pi} dx = \frac{2}{\pi}\sqrt{1-y^2}, \quad y \in \langle -1, 1\rangle.$$

Příklad 1.4.5 *Najděte marginální rozdělení složek náhodného vektoru s dvourozměrným normálním rozdělením z příkladu 1.4.3.*

Řešení: V tomto případě lze psát $f^X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \frac{1}{2\pi} e^{-\frac{x^2}{2}} \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} dy$. Protože je $\int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} dy = \sqrt{2\pi}$, dostáváme pro sdruženou hustotu z 1.4.3 marginální hustotu $f^X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, x \in R$ a obdobně $f^Y(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}, y \in R$. Marginální rozdělení obou složek je tedy opět normální $N(0,1)$.

1.4.2 Nezávislost náhodných veličin

Říkáme, že náhodné veličiny X, Y jsou **stochasticky nezávislé**, jestliže jsou jevy $\{X \le x\}, \{Y \le y\}$ stochasticky nezávislé pro každé $(x,y) \in R^2$.

Stochastická nezávislost složek náhodného vektoru (X,Y) je ekvivalentní s každou z následujících vlastností:

a) Pro každé $(x,y) \in R^2$ lze sdruženou distribuční funkci $F(x,y)$ vyjádřit jako součin marginálních distribučních funkcí $F^X(x)$ a $F^Y(y)$, tedy platí

$$F(x,y) = F^X(x)F^Y(y).$$

b) Pro diskrétní rozdělení platí pro každé i, j

$$p_{ij} = p_i^X p_j^Y.$$

c) Pro každé $(x,y) \in R^2$ lze sdruženou hustotu $f(x,y)$ vyjádřit jako součin marginálních hustot $f^X(x)$ a $f^Y(y)$, tedy platí

$$f(x,y) = f^X(x)f^Y(y).$$

Příklad 1.4.6 *Existuje stochastická závislost mezi pohlavím a zaměstnaností v příkladu 1.4.2?*

Řešení: Náhodné veličiny v příkladu 1.4.2 nejsou nezávislé, neboť např. $p_0^X p_0^Y = 0,48.0,1408 = 0,067584 \neq 0,0576 = p_{00}$.

Příklad 1.4.7 *Jsou složky vektoru (X,Y) v příkladu 1.4.4 stochasticky závislé?*

Řešení: Z výpočtu marginálních hustot v příkladu 1.4.4 je na první pohled vidět, že jejich součin nemůže být roven konstantní funkci, kterou je sdružená hustota. Tedy složky X a Y tohoto vektoru jsou stochasticky závislé.

Nakonec tohot odstavce si uvedeme ještě jedno užitečné tvrzení:

Věta 1.12 *Jsou-lí náhodné veličiny X,Y stochasticky nezávislé a g,h jsou spojité reálné funkce, potom též $g(X), h(Y)$ jsou stochasticky nezávislé náhodné veličiny.*

1.4.3 Charakteristiky náhodného vektoru

Střední hodnota náhodného vektoru $E(X,Y)$ je definována jako **vektor středních hodnot** (EX, EY) jeho složek. Tyto střední hodnoty spočteme buď pomocí sdruženého rozdělení v diskrétním případě

$$\begin{aligned} EX &= \sum_i \sum_j x_i p_{ij} = \sum_i x_i \sum_j p_{ij} = \sum_i x_i p_i^X, \\ EY &= \sum_i \sum_j y_j p_{ij} = \sum_j y_j \sum_i p_{ij} = \sum_j y_j p_j^Y. \end{aligned}$$

Výrazy napravo odpovídají výpočtu střední hodnoty jednorozměrné veličiny X, resp. Y s rozdělením $\{p_i^X\}$, resp. $\{p_j^Y\}$.

Ve spojitém případě mají tyto rovnosti tvar

$$\begin{aligned} EX &= \iint_{R^2} x f(x,y)dxdy = \int_R x \int_R f(x,y)dydx = \int_R x f^X(x)dx, \\ EY &= \iint_{R^2} y f(x,y)dxdy = \int_R y \int_R f(x,y)dxdy = \int_R y f^Y(y)dy. \end{aligned}$$

Kovariance $cov(X,Y)$ náhodných veličin X,Y je definována jako

$$cov(X,Y) = E(X - EX)(Y - EY) = EXY - EXEY$$

Ve výrazu se objevuje takzvaný *smíšený moment* EXY, který je v diskrétním případě roven

$$EXY = \sum_i \sum_j x_i y_j p_{ij},$$

ve spojitém případě

$$EXY = \iint_{R^2} xyf(x,y)dxdy.$$

Všimněte si, že pokud jsou veličiny X a Y stochasticky nezávislé, potom je

$$EXY = \sum_i \sum_j x_i y_j p_i^X p_j^Y = \sum_i x_i p_i^X \sum_j y_j p_j^Y = EX.EY,$$

ve spojitém případě

$$EXY = \iint_{R^2} xyf^X(x)f^Y(y)dxdy = \int_R xf^X(x)dx \int_R yf^Y(y)dy = EX.EY.$$

Potom je

$$cov(X,Y) = EXY - EXEY = EXEY - EXEY = 0.$$

Tedy platí tvrzení: *pokud jsou X a Y stochasticky nezávislé náhodné veličiny, potom je jejich kovariance vždy rovna nule.*

Toto tvrzení neplatí opačně, jak je ukázáno v následujícím příkladu.

Příklad 1.4.8 *Spočtěte kovarianci složek X a Y náhodného vektoru z příkladu 1.4.4.*

Řešení:

$$\begin{aligned} EX &= \frac{1}{\pi}\int\int_K xdxdy = \frac{1}{\pi}\int_{-1}^{1}\int_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}} xdxdy = 0, \\ EY &= \frac{1}{\pi}\int\int_K ydxdy = \frac{1}{\pi}\int_{-1}^{1}\int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} ydydx = 0, \\ EXY &= \frac{1}{\pi}\int\int_K xydxdy = \frac{1}{\pi}\int_{-1}^{1} x\int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} ydydx = 0, \end{aligned}$$

neboť ve všech případech jsou v integrandu liché funkce a integruje se přes symetrickou oblast. Tedy $cov(X,Y) = 0$.

Pro tentýž vektor jsme ukázali v příkladu 1.4.7 že jeho složky jsou stochasticky závislé a přesto je jejich kovariance nulová.

Kovarianční matice náhodného vektoru (X, Y) má tvar

$$D = \begin{pmatrix} Var(X) & cov(X,Y) \\ cov(X,Y) & Var(Y) \end{pmatrix}$$

Na diagonále má tedy rozptyly složek, mimo diagonálu jejich kovarianci.

Korelační koeficient $\rho(X, Y)$ náhodných veličin X a Y je roven

$$\rho(X,Y) = \frac{cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

Je-li $\rho(X, Y) = 0$, potom se náhodné veličiny nazývají **nekorelované**.

Věta 1.13 *Platí* $-1 \leq \rho(X,Y) \leq 1$. *Dále* $\rho(X,Y) = \pm 1$ *právě tehdy, když pro nějaké* a, b *reálné konstanty,* $b \neq 0$, *platí s pravděpodobností* 1 *rovnost* $Y = a + bX$, *přitom znaménko* $\rho(X,Y)$ *se shoduje se znaménkem* b.

Věta 1.14 *Jsou-li náhodné veličiny* X, Y *stochasticky nezávislé, potom je* $\rho(X,Y) = 0$. *Naopak, je-li* $\rho(X,Y) \neq 0$, *potom jsou* X *a* Y *stochasticky závislé.*

Pokud je $\rho(X, Y) = 0$, veličiny X, Y jsou pouze nekorelované, to znamená že mezi nimi není lineární závislost, ale mohou být stochasticky závislé (jak je ukázáno v příkladu 1.4.8.

Korelační koeficient se používá v experimentálním výzkumu k vyjádření míry lineární závislosti mezi dvěma náhodnými veličinami.

Příklad 1.4.9 *Nechť* $X \approx Exp(1), Y = X^2$. *Vypočtěte kovarianční matici a korelační koeficient* $\rho(X,Y)$.

Řešení: Užitím vzorce pro obecné momenty exponenciálního rozdělení pro $\lambda = 1$ je $EX = 1, EX^2 = 2 = EY, EY^2 = EX^4 = 24$, tedy $VarX = 1, VarY = 20$. Pro výpočet EXY si stačí uvědomit, že $EXY = EX^3 = 6$, tedy kovariance $cov(X,Y) = 4$. Kovarianční matice je tedy

$$D = \begin{pmatrix} 1 & 4 \\ 4 & 20 \end{pmatrix}$$

a $\rho(X,Y) = \frac{2}{\sqrt{5}}$. Mezi X a Y je funkcionální závislost, tj. hodnoty jedné složky náhodného vektoru určují druhou složku (v praxi řídký případ). Korelační koeficient je kladný, neboť s rostoucím X roste Y, ale nedosahuje hodnoty 1, protože závislost není lineární.

Náhodný vektor (X,Y) má **dvourozměrné normální rozdělení** s vektorem středních hodnot (μ_1,μ_2) a kovarianční maticí

$$D = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix},$$

jestliže jeho hustota $f(x,y)$ má tvar

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{(1-\rho^2)}} e^{-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right)}$$

$(x,y) \in R^2$, kde $\rho = \frac{\sigma_{12}}{\sigma_1\sigma_2}$ je korelační koeficient složek. Pro $|\rho| = 1$, kdy X a Y jsou lineárně závislé, není hustota definována.

Hustotu dvourozměrného normálního rozdělení lze rozložit na součin jednorozměrných (marginálních) hustot právě tehdy, když $\rho = 0$, jak je vidět z tvaru exponentu. Lze vyslovit následující tvrzení:

Věta 1.15 *Nechť rozdělení pravděpodobnsti náhodného vektoru (X,Y) je dvourozměrné normální. Potom složky X a Y jsou stochasticky nezávislé právě tehdy, je-li jejich korelační koeficient $\rho(X,Y)$ roven nule.*

Příklad 1.4.10 *Napište hustotu náhodného vektoru (X,Y) s dvourozměrným normálním rozdělením, je-li $(\mu_1,\mu_2) = (26,-12)$ a*

$$D = \begin{pmatrix} 196 & -91 \\ -91 & 169 \end{pmatrix}$$

Řešení: Je $\rho = \frac{\sigma_{12}}{\sigma_1\sigma_2} = -0,5$, $\sqrt{1-\rho} = \frac{\sqrt{3}}{2}$ tedy

$$f(x,y) = \frac{1}{182\pi\sqrt{3}} e^{-\frac{2}{3}\left(\frac{(x-26)^2}{196} + \frac{(x-26)(y+12)}{182} + \frac{(y+12)^2}{169}\right)}$$

1.4.4 Funkce náhodného vektoru

Uvažujme lineární funkci $h(X,Y) = aX + bY, a, b$ jsou reálné konstanty, X, Y náhodné veličiny. Pro střední hodnotu $Eh(X,Y)$ je zřejmě (z linearity) $E(aX + bY) = aEX + bEY$. Dále pro rozptyl lze odvodit

$$\begin{aligned} Var(aX + bY) = E(aX + bY)^2 - (aEX + bEY)^2 = \\ a^2E(X - EX)^2 + b^2E(Y - EY)^2 + 2abE(X - EX)(Y - EY) = \\ a^2VarX + b^2VarY + 2ab.cov(X,Y) \end{aligned}$$

Specielně, jsou-li X, Y nekorelované, tj. $cov(X,Y) = 0$, potom $Var(aX + bY) = a^2VarX + b^2VarY$.

Příklad 1.4.11 *Chcete investovat 10000 Kč a rozhodujete se mezi akciemi a krátkodobými vkladovými certifikáty. Návratnosti z obou zdrojů jsou náhodné veličiny s diskrétním rozdělením pravděpodobností výnosů. Značí-li X resp. Y procentní výnos z akcií resp. certifikátů, je*

P(X = -5, Y = 8) = 0,1 P(X = 5, Y = 5) = 0,3
P(X = 5, Y = 8) = 0,3 P(X = 15, Y = 5) = 0,3

***a**) Najděte střední výnos μ a směrodatnou odchylku σ, (riziko) pro rozložení investic mezi akcie a certifikáty v poměru 100:0, 50:50, 0:100.*

***b**) Najděte optimální portfolio vzhledem k střednímu výnosu resp. vzhledem k míře rizika.*

Řešení: Výnos Z je dán vzorcem $Z = aX + bY$, kde $a + b = 100$ (1% z 10000). Hodnoty $\mu = EZ$ a $\sigma = \sqrt{VarZ}$ plynou ze výše uvedených vzorců, kam dosadíme po výpočtech $EX = 7, EY = 6,2, VarX = 36, VarY = 2,16, cov(X,Y) = -5,4$.

a) Pro poměr $a : b = 100 : 0$ je $\mu = 700, \sigma = 600$ pro $50 : 50$ je $\mu = 660, \sigma = 216,5$ pro $0 : 100$ konečně $\mu = 620, \sigma = 147$.

b) Protože $EX > EY$, nabývá EZ zřejmě maxima pro $a = 100, b = 0$. Optimalizace vzhledem k míře rizika představuje výpočet vázaného extrému $VarZ$ jako funkce a, b. Z Lagrangeovy funkce $L(a,b,\lambda) = a^2VarX + b^2VarY + 2ab.cov(X,Y) + \lambda(a + b - 100)$ odečtením rovnic $\frac{\partial L}{\partial a} = 2aVarX + 2b.cov(X,Y) + \lambda = 0, \frac{\partial L}{\partial b} = 2bVarY + 2a.cov(X,Y) + \lambda = 0$ dostáváme podíl

$$\frac{a}{b} = \frac{VarY - cov(X,Y)}{VarX - cov(X,Y)} = 0,1826$$

Z podmínky $a + b = 100$ tomu odpovídá poměr $a : b = 15,44 : 84,56$ (optimální vzhledem k riziku), pro nějž $\mu = 632,4, \sigma = 99,6$. Ukázalo se, že nelze docílit současně vysokého výnosu (velké μ) a nízkého rizika (malé σ), což bylo patrné již z výpočtů v a).

Uvažme obecnou funkci $h(x, y)$ dvou proměnných a náhodný vektor (X, Y). Potom $Z = h(X, Y)$ je náhodná veličina, pro jejíž distribuční funkci F_Z platí:

$$F_Z(z) = P(h(X, Y) \leq z)$$

což je v diskrétním případě rovno

$$F_Z(z) = \sum_{h(x_i,y_j)\leq z} p_{ij}$$

a pro spojité rozdělení (X, Y) s hustotou $f(x, y)$

$$F_Z(z) = \iint_{h(x,y)\leq z} f(x, y)dxdy.$$

Příklad 1.4.12 *Náhodná veličina $Z = 50X+50Y$ z III.4.2 má diskrétní distribuční funkci $F_Z(z)$ danou následujícími pravděpodobnostmi: $P(Z = 150) = 0,1, P(Z = 500) = 0,3, P(Z = 650) = 0,3, P(Z = 1000) = 0,3$.*

Pro výpočet střední hodnoty funkce náhodného vektoru není třeba počítat transformované sdružené rozdělení. Střední hodnotu lze počítat přímo podle vztahů

$$Eh(X, Y) = \sum_i \sum_j h(x_i, y_j)p_{ij}$$

pro diskrétní rozdělení a

$$Eh(X, Y) = \int \int_R^2 h(x, y)f(x, y)dxdy$$

pro spojité rozdělení.

Příklad 1.4.13 *Při měření dvou rozměrů součástky má chyba (X, Y) dvourozměrné normální rozdělení s nezávislými složkami, nulovým vektorem středních hodnot a stejnými rozptyly $\sigma_1^2 = \sigma_2^2 = \sigma^2$, tedy s hustotou*

$$f(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}, (x, y) \in R^2$$

Vypočtěte střední hodnotu a rozptyl celkové velikosti chyby dané náhodnou veličinou $Z = \sqrt{X^2 + Y^2}$.

Řešení: Počítáme i-tý obecný moment Z: je $Z^i = (X^2 + Y^2)^{\frac{i}{2}}$, tedy

$$EZ^i = \frac{1}{2\pi\sigma^2} \int \int_R^2 (x^2 + y^2)^{\frac{i}{2}} e^{\frac{x^2+y^2}{2\sigma^2}} dxdy.$$

Substitucí $x = rcos\beta, y = rsin\beta$, jejíž Jakobián je r, dostáváme

$$EZ^i = \frac{1}{2\pi\sigma^2} \int_0^{2\pi} d\beta \int_0^{\infty} r^{i+1} e^{-\frac{r^2}{2\sigma^2}} dr$$

a s použitím další substituce $t = \frac{r^2}{2\sigma^2}$

$$EZ^i = \sigma^i \int_0^{\infty} (2t)^{\frac{i}{2}} e^{-t} dt = \left(\sigma\sqrt{2}\right)^i \Gamma(\frac{i}{2} + 1).$$

Odsud specielně plyne $EZ = \sigma\sqrt{2}\Gamma(\frac{3}{2}) = 1,253\sigma$. Pomocí druhého obecného momentu $EZ^2 = 2\sigma^2\Gamma(2)$ je rozptyl $VarZ = EZ^2 - (EZ)^2 = 0,429\sigma^2$.

1.5 Limitní věty teorie pravděpodobnosti

Limitní věty teorie pravděpodobnosti se zabývají chováním posloupnosti náhodných veličin. Náhodné veličiny $X_1, \dots X_n$ jsou nezávislé, jsou-li jevy $\{X_1 \leq x_1\}, \dots, \{X_n \leq x_n\}$ nezávislé pro každé $(x_1, \dots, x_n) \in R^n$. V dalším textu budeme často mluvit o posloupnosti nezávislých, stejně rozdělených náhodných veličin[24]

Náhodné veličiny $X_1, X_2, \dots$ tvoří posloupnost **nezávislých, stejně rozdělených** náhodných veličin, jsou-li pro každé $n \in N$ veličiny $X_1, \dots, X_n$ nezávislé a mají všechny tutéž distribuční funkci $F(x)$.

1.5.1 Zákon velkých čísel

Věta 1.16 (Bernoulliho zákon velkých čísel) *Nechť $\{X_i\}_{i=1}^{\infty}$ je posloupnost nezávislých stejně rozdělených náhodných veličin s alternativním rozdělením $Alt(p)$. Položme $S_n = \sum_{i=1}^{n} X_i$. Potom pro každé $\epsilon > 0$ je*

$$\lim_{n\to\infty} P(|\frac{S_n}{n} - p| > \epsilon) = 0.$$

Důkaz: Pro každé n je $S_n \approx Bin(n,p), ES_n = np, VarS_n = np(1-p)$. Odtud $E\left(\frac{S_n}{n}\right) = p, Var\left(\frac{S_n}{n}\right) = \frac{p(1-p)}{n}$. Podle Čebyševovy nerovnosti (věta 1.6) je

$$P(|\frac{S_n}{n} - p| > \epsilon) \leq \frac{p(1-p)}{n\epsilon^2}$$

při $n \to \infty$ konverguje výraz na pravé straně k nule pro každé pevné $\epsilon > 0$.

Výraz $\frac{S_n}{n}$ v předchozí větě je poměrná četnost jevu $A = \{X_i = 1\}$ v n opakováních pokusu, je to náhodná veličina, neboť v různých seriích opakování nabývá různých hodnot. Setkali jsme se s ní již v úvodní kapitole. Zákon velkých čísel potvrzuje, že pro $n \to \infty$ veličina $\frac{S_n}{n}$ konverguje ke konstantě – k pravděpodobnosti jevu A. Pojem **konvergence posloupnosti náhodných veličin** lze definovat různým způsobem, ve větě 1.16 jde o **konvergenci v pravděpodobnosti**. Následující věta říká (bez předpokladu o konečnosti rozptylu), že aritmetický průměr konverguje ke střední hodnotě. To je zobecnění věty 1.16, neboť poměrná četnost je průměrem alternativních veličin a pravděpodobnost jevu A jejich střední hodnotou. Tak jsou potvrzeny úvahy, které jsme používali k četnostní interpretaci pravděpodobnosti.

[24] V literatuře se můžete setkat se zkratkou *iid* z anglického „independent and identically distributed“

Věta 1.17 *Nechť $\{X_i\}$ je posloupnost nezávislých stejně rozdělených náhodných veličin se střední hodnotou a. Potom pro každé $\epsilon > 0$ je*

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - a\right| > \epsilon\right) = 0.$$

Předchozí větu lze aplikovat k výpočtu určitého integrálu metodou *Monte Carlo* , tj. s použitím počítačových simulací pseudonáhodných čísel. Pro výpočet $\int_a^b g(x)dx$, kde g je reálná funkce jedné proměnné, nechť $\{X_i\}$ je posloupnost nezávislých stejně rozdělených náhodných veličin s rovnoměrným rozdělením $U(a,b)$, tj. s hustotou $f(x) = \frac{1}{b-a}$ pro $a \le x \le b$ a $f(x) = 0$ jinde. Platí $Eg(X_i) = \int_R g(x)f(x)dx = \frac{1}{b-a}\int_a^b g(x)dx$, podle věty 1.17 je pro každé $\epsilon > 0$

$$\lim_{n\to\infty} P\left(\left|\frac{b-a}{n}\sum_{i=1}^{n} g(X_i) - \int_a^b g(x)dx\right| > \epsilon\right) = 0$$

Tento výraz ukazuje, že hledaný integrál lze aproximovat uvedeným součtem a vede k následujícímu algoritmu:

1) proveďme n opakování simulace pseudonáhodného čísla $Y_i \approx U(0,1)$,

2) spočteme $X_i = (b-a)Y_i + a, X_i \approx U(a,b), i = 1,\ldots,n$,

3) potom $\int_a^b g(x)dx \approx \frac{b-a}{n}\sum_{i=1}^{n} g(x_i)$.

Přesnost metody Monte Carlo při výpočtu jednoduchého integrálu není vyšší než u běžných numerických metod. Více se v praxi uplatňuje při výpočtu složitých vícenásobných integrálů.

1.5.2 Centrální limitní věta

Centrální limitní věty v teorii pravděpodobnosti podtrhují výsadní postavení normálního rozdělení jako limitního pro součty či průměry náhodných veličin s obecným výchozím rozdělením. Těchto vět je celá řada a nejjednodušší z nich, takzvaná *Moivreova* věta přináší možnost aproximovat binomické rozdělení normálním.

Věta 1.18 *Za předpokladů věty* **??** *položme*

$$Z_n = \frac{S_n - np}{\sqrt{np(1-p)}}.$$

Potom $\lim_{n\to\infty} P(Z_n \le x) = \Phi(x), x \in R$.

Jinými slovy, distribuční funkce normovaných součtů Z_n (normováním náhodné veličiny rozumíme odečtení střední hodnoty a vydělení směrodatnou odchylkou) konverguje k distribuční funkci Φ normovaného normálního rozdělení $N(0,1)$.

Příklad 1.5.1 *Jaká je pravděpodobnost, že mezi* $n = 20000$ *nezávislými výrobky bude víc než* 100 *vadných, je-li pravděpodobnost výroby vadného* $p = 0,004$*?*

Řešení: Počet zmetků $S_n \approx Bin(n,p)$. Místo výpočtu pravděpodobnosti binomického rozdělení využijeme normální aproximaci. Při označení z věty 1.18 je

$$P(S_n > 100) = P\left(Z_n > \frac{100 - np}{\sqrt{np(1-p)}}\right) = P(Z_n > 2,24)$$

a tato pravděpodobnost je přibližně rovna $1 - \Phi(2,24) = 0,0125$. Poznamenejme, že přesnost tohoto přibližného nahrazení je poměrně velká.

Příklad 1.5.2 *Letecká společnost provozuje linku, na které se průměrně 5 procent cestujících nedostaví k letu. I když je všech* 67 *míst rezervováno, letadlo často létá s prázdnými místy. Společnost tedy plánuje, že bude přijímat rezervaci od* $n = 69$ *cestujících. Jaká je pravděpodobnost, že nebude místo pro všechny cestující s rezervací, kteří se k letu dostaví?*

Řešení: Cestující se dostaví k letu nezávisle na sobě s pravděpodobností $p = 0,95$. Celkový počet příchozích $S_n \approx Bin(n,p)$. Přesný výpočet dává $P(S_n > 67) = P(S_n = 69) + P(S_n = 68) = 0,95^{69} + 69.0,95^{68}0,05 = 0,1344$. Protože n není příliš vysoké, je při použití normální aproximace třeba provést **korekci pro nahrazení diskrétního rozdělení spojitým**. Protože S_n je diskrétní celočíselná veličina, jsou zřejmě jevy $\{S_n > 67\}$ a $\{S_n \geq 68\}$ totožné, tedy i jejich pravděpodobnosti by měly být stejné. Jenže při použití aproximace bez korekce by bylo $P(S_n > 67) = P(Z_n > 0,801) \approx 0,2118$ a současně $P(S_n \geq 68) = P(Z_n \geq 1,353) \approx 0,0885$. Proto volíme kompromis, spočívající v tom, že namísto dvou jevů $\{S_n > 67\}$ a $\{S_n \geq 68\}$ budeme uvažovat jev $\{S_n \geq 67,5\}$, pro který je $P(S_n \geq 67,5) = P(Z_n \geq 1,077) \approx 1 - \Phi(1,077) = 0,141$, což se již málo liší od přesné hodnoty $0,1344$.

Nyní uvádíme obecnější verzi centrální limitní věty.

Věta 1.19 *Nechť* $\{X_i\}$ *je posloupnost nezávislých stejně rozdělených náhodných veličin s konečnou střední hodnotou* μ *a rozptylem* σ^2*. Položme* $S_n = \sum_{i=1}^{n} X_i$*,* $Z_n = \frac{S_n - ES_n}{\sqrt{DS_n}} = \frac{S_n - n\mu}{\sigma\sqrt{(n)}}$*. Potom pro* $x \in R$ *je*

$$\lim_{n\to\infty} P(Z_n \leq x) = \Phi(x).$$

Tato věta říká, že distribuční funkce normovaných součtů konverguje k distribuční funkci $N(0,1)$ rozdělení dokonce i pro libovolné výchozí rozdělení s konečnou střední hodnotou a rozptylem.

Příklad 1.5.3 *Při výpočtu integrálu* $J = \int_0^1 \sqrt{x}dx$ *metodou Monte Carlo s* $n = 10000$ *určete pravděpodobnost, že chyba výsledku bude menší než* 10^{-3}.

Řešení: Při odhadu chyby potřebujeme skutečnou hodnotu $J = \frac{2}{3}$. Nechť $X_i \approx U(0,1), i = 1, \ldots, n$, jsou nezávislé. Je $E\sqrt{X_i} = J = \frac{2}{3}, EX_i = \frac{1}{2}$, tedy $Var\sqrt{X_i} = \frac{1}{18}$. Podle IV.1.5 má odhad metodou Monte Carlo tvar $J \approx \frac{1}{n}\sum_{i=1}^n \sqrt{X_i}$. Označme $S_n = \sum_{i=1}^n \sqrt{X_i}$, potom $ES_n = nJ, DS_n = \sum_{i=1}^n D\sqrt{X_i} = \frac{n}{18}$. Jsme připraveni použít větu IV.2.4:

$$P\left(\left|\frac{S_n}{n} - J\right| < 10^{-3}\right) = P\left(\left|\frac{S_n - nJ}{\sqrt{\frac{n}{18}}}\right| < 10^{-3}\sqrt{18n}\right) = P(|Z_n| < 0,424)$$
$$\approx 2\Phi(0,424) - 1 = 0,328$$

Pravděpodobnost, že odhad se bude lišit na třetím desetinném místě, je značná.

Příklad 1.5.4 *Kolikrát je třeba změřit fyzikální veličinu jejíž přesná hodnota je* m*, abychom mohli s pravděpodobností* $0,96$ *tvrdit, že průměr těchto měření se liší od* m *o méně než* 2*? Je známo, že směrodatná odchylka měřicí metody je* $\sigma = 4$.

Řešení: Přesná hodnota m představuje střední hodnotu měření $X_i, i = 1, \ldots, n, DX_i = \sigma^2 = 16$. Pro $S_n = \sum_{i=1}^n$ je $ES_n = nm, DS_n = 16n$, použitím věty IV.2.4 dostáváme

$$P\left(\left|\frac{S_n}{n} - m\right| < 2\right) = P\left(\left|\frac{S_n - nm}{4\sqrt{n}}\right| < \frac{\sqrt{n}}{2}\right) \approx 2\Phi\left(\frac{\sqrt{n}}{2}\right) - 1.$$

Poslední výraz má být roven alespoň $0,96$. Úpravou této nerovnosti je

$$2\Phi\left(\frac{\sqrt{n}}{2}\right) \geq 1,96 \text{ a tedy } \frac{\sqrt{n}}{2} \geq \Phi^{-1}(0,96) = 2,055,$$

tedy počet měření musí být alespoň 17. Příklad představuje úlohu plánování experimentu na základě požadavků na přesnost.

Kapitola 2

Základy matematické statistiky

2.1 Statistická indukce, náhodný výběr

(Upozornění: tento text ještě nebyl upraven do konečné podoby. Proto zde chybějí některé obrázky a objevují se zde odkazy na odstavce, které nejsou označeny odpovídajícím způsobem. Tyto nedostatky budou postupně odstraňovány. Do té doby prosím o trpělivost a omlouvám se za ztížené čtení. GDo)

2.1.1 Úloha statistické indukce

Řešení většiny praktických problémů v každodenním život je založeno na dostupných informacích o chování reálného světa. Tyto informace získáváme pozorováním nebo měřením. Přitom nás zajímají nejen kvantitativní znaky, jako je například hmotnost, napětí, rychlost, ale i znaky kvalitativní, například druh výrobní technologie, kvalitu, způsob reakce a další. Při získávání těchto informací ovšem narážíme na řadu problémů. Nejsme například schopni změřit spolehlivost celé produkce zářivek, pokud tuto spolehlivost vyjadřujeme v hodinách životnosti. Podobně nelze změřit kvalitu všech dodávaných komponent, jejichž denní spotřeba je několik tisíc a kontrola kvality je časově či finančně náročná. Dokonce řadu veličin, které v mnoha případech měříme zcela běžně a bez problémů, jako například hmotnost, nejsme s rostoucím nárokem na přesnost schopni „přesně“ změřit. Při měření totiž obvykle nelze absolutně vyloučit další působící vlivy, nepřesnosti přístrojů, vliv prostředí. Pokusy o maximální vyloučení těchto „vnějších“ vlivů bývají často velmi nákladné a ne vždy úspěšné. Přesto však měření sledovaných veličin a znaků tvoří základní úlohu, bez níž se neobejdeme a na jejímž výsledku závisí naše další závěry, rozhodování a činnost. Zde přichází ke slovu **matematická statistika** a její silná zbraň - **statistická indukce**, jejíž princip se pokusíme v následujícím textu vyložit.

Statistická indukce je induktivní způsob usuzování, při kterém činíme závěry o náhodném chování a vlastnostech celku - **základního souboru** - na základě pozorování jeho části - **výběru**. Chování náhodné veličiny, jakou je například hmotnost výlisku, lze popsat jejím rozdělením pravděpodobnosti. Abychom je poznali, museli bychom znát pravděpodobnosti všech hodnot, které může tato veličina teoreticky nabývat. My však můžeme v konečném čase prostřednictvím měření pozorovat pouze konečný počet hodnot z této nekonečné množiny. Podobně při měření životnosti zářivek, které probíhá tak, že měříme dobu, po kterou zářivka svítí až do jejího zničení. V tomto případě provedeme výběr z celkové produkce a na něm změříme sledovanou veličinu. Předpokládáme, že ostatní zářivky, které byly vyrobeny za přibližne stejných podmínek, budou mít i přibližně stejnou životnost.

Z předchozích příkladů je zřejmé, že statistickou indukci lze aplikovat pouze na náhodné jevy (veličiny), které mají **statistickou povahu**. Rozumíme tím možnost neomezeného nezávislého opakování jevu za „stejných“ podmínek.

Posloupnost n měření, která jsou na sobě nezávislá a probíhají pokud možno za stejných podmínek, můžeme interpretovat jako realizace n nezávislých, stejně rozdělených náhodných veličin $X_1, X_2, \ldots, X_n$. Mají-li všechny tyto veličiny pravděpodobnostní rozdělení s distribuční funkcí $F(x)$, řekneme, že se jedná o **náhodný výběr** $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ **o rozsahu** n z pravděpodobnostního rozdělení $F(x)$. Takto provedený výběr považujeme za reprezentativní. Pod pojmem *výběr* budeme nadále rozumět náhodný výběr.

Výsledky statistické indukce, a tedy všechny statistické metody, jsou založeny na naměřených hodnotách, na **datech**. Na jejich „kvalitě“ přímo závisí i kvalita našich závěrů, to znamená co nejvěrnější popis reálné situace. Data jsou náš výchozí „materiál“ a je zřejmé, že ze „špatných“ dat nelze „vyrobit“ dobré závěry, podobně jako ze špatného materiálu nelze sebelepším obráběním vyrobit kvalitní výrobek. Nejčastější požadavky kladené na výběr jsou **nezávislost** a **reprezentativnost** výběru.

2.1.2 Náhodný výběr

Kvalita dat je – kromě přesnosti – určena především způsobem výběru. Způsoby získávání dat ze základního souboru se zabývá část statistiky, nazývaná **výběrová šetření** (viz např. [Če]). Zde se zkoumají vlastnosti různých způsobů výběru. Nejčastěji provádíme náhodný výběr tak, aby každý prvek základního souboru měl stejnou pravděpodobnost, že bude vybrán, a aby každý prvek byl vybrán nezávisle na ostatních vybraných prvcích. V praxi se realizují i jiné způsoby výběru. Například tzv. **záměrný výběr**, při kterém vybíráme záměrně prvky, o nichž předpokládáme, že jsou pro daný soubor typickými. Provádíme jej tehdy, máme-li o základním souboru dostatek informací. Dalším způsobem je **oblastní výběr**. Ten spočívá v tom, že základní soubor nejprve rozdělíme do skupin (oblastí) a v každé z nich potom provádíme náhodný výběr. V tomto případě je třeba mít kriterium pro rozdělení do oblastí (např. regionální příslušnost, výrobní rezort apod.). **Systematický výběr** je způsob, při němž vybíráme prvky základního souboru podle předem zvoleného kriteria, nesouvisejícího s vyšetřovanými znaky (například vybíráme každou desátou vyrobenou součástku). V praxi se věnuje velká pozornost takzvanému **plánování experimentu**, při kterém se - kromě jiného - stanoví i způsob organizace měření, pořadí atd (podrobněji viz [Li]).

2.1.3 Četnosti, empirické rozdělení

Jak víme již z první části těchto skript, vlastnosti náhodných veličin lze odvozovat za předpokladu znalosti jejich rozdělení pravděpodobnosti. Při pozorování reálných procesů zpravidla teoretické rozdělení neznáme a k dispozici máme jen číselné hodnoty $x_1, x_2, \ldots, x_n$, které dostaneme pozorováním (měřením) náhodného výběru $X = X_1, X_2, \ldots, X_n$ z tohoto rozdělení. Naměřené hodnoty $x_1, x_2, \ldots, x_n$ budeme nazývat **pozorování** nebo **vstupní data**. Naší snahou obvykle bývá získat představu o rozdělení pravděpodobnosti sledované náhodné veličiny X. K tomu používáme **empirické rozdělení**, získané na základě pozorování náhodného výběru X.

Předpokládejme, že vstupní data $x_1, x_2, \ldots, x_n$ leží všechna v intervalu $\langle a, b\rangle$. Rozdělme tento interval body $a = a_0 < a_1 < a_2 < \cdots < a_k = b$ na k disjunktních podintervalů $(a_{i-1}, a_i\rangle$, $i = 1, 2, \ldots, k$, kterým budeme říkat **třídní intervaly** nebo jen **třídy**. Počet pozorování x_j, splňujících nerovnost $a_{i-1} < x_j \leq a_i$, nazveme **absolutní četností** nebo jen **četností** třídy i a budeme ji značit n_i. Počet všech pozorování náhodného výběru n nazveme **rozsahem výběru**. **Poměrné**, nebo též **relativní četnosti** $f_i = \frac{n_i}{n}$ potom lze považovat za odhad **empirického rozdělení** pravděpodobnosti jevu, že hodnota sledované náhodné veličiny bude v i-tém intervalu.

Tento postup lze použít jak pro diskrétní náhodné veličiny, tak i pro spojité (metrické) náhodné veličiny. V případě, že náhodná veličina může nabývat pouze konečného a malého počtu hodnot, není třeba vytvářet třídní intervaly. V takovém případě počítáme přímo četnosti výskytu jednotlivých hodnot náhodné veličiny v našem výběru.

Kromě četností pracujeme i s takzvanými *kumulativními* četnostmi. **Absolutní kumulativní četnost** N_i třídy i je rovna počtu pozorování x_j, pro která $a \leq x_j \leq a_i$, tedy je $N_i = \sum_{r=1}^{i} n_r$. Podobně **poměrná kumulativní četnost** je dána vztahem $F_i = \sum_{r=1}^{i} f_r$ a může sloužit jako hrubý odhad empirické distribuční funkce.

Příklad 2.1.1 *Uvažujme náhodný výběr, při kterém byly naměřeny následující hodnoty hmotnosti 40 výlisků (v gramech):*

1114	*1105*	*1121*	*1120*	*1118*	*1132*	*1121*	*1115*	*1128*	*1110*
1119	*1135*	*1128*	*1122*	*1116*	*1108*	*1111*	*1118*	*1118*	*1119*
1122	*1124*	*1112*	*1102*	*1121*	*1135*	*1123*	*1128*	*1116*	*1131*
1109	*1120*	*1128*	*1132*	*1117*	*1118*	*1122*	*1108*	*1125*	*1119*

Sestavte četnostní tabulku.

Řešení: Vytvoříme-li třídy po 6 gramech, dostaneme následující tabulku:

třídní interval			prosté četnosti		kumulativní četnosti	
levá mez	střed	pravá mez	absolutní	poměrná	absolutní	poměrná
1101	1103,5	1106	2	0.050	2	0.050
1107	1109,5	1112	6	0.150	8	0.200
1113	1115,5	1118	9	0.225	17	0.425
1119	1121,5	1124	13	0.325	30	0.750
1125	1127,5	1130	5	0.125	35	0.875
1131	1133,5	1136	5	0.125	40	1.000
součet			40	1.000	–	–

Pro grafické zobrazení četností používáme nejčastěji sloupkový graf, takzvaný **histogram**. Každý sloupek odpovídá jedné třídě (třídnímu intervalu) a výška sloupků odpovídá četnosti. Dalším, často používaným grafem pro poměrné (relativní) četnosti je **kruhový graf** (nebo též koláčový graf). Četnosti lze samozřejmě zobrazit i v jiných typech grafů.

Na levém obrázku vidíte histogram absolutních četností, na pravém je graf poměrných kumulativních četností:

V případě četnostní analýzy je důležitá volba počtu třídních intervalů. Příliš malý počet tříd povede k

2.1.4 Výběrové charakteristiky

V odstavci II.2 těchto skript se můžete seznámit s řadou základních charakteristik rozdělení pravděpodobnosti náhodné veličiny X. V našem případě, kdy máme k dispozici pouze pozorování náhodného výběru $\{X_1, X_2, \ldots, X_n\}$, nahrazujeme teoretické pravděpodobnostní charakteristiky jejich výběrovými protějšky - **statistickými ukazateli**, neboli **výběrovými charakteristikami**.

Výběrové charakteristiky se vyjadřují pomocí **statistik**. Statistiky jsou náhodné veličiny, které jsou vyjádřeny jako funkce výběru, to jest náhodných veličin $X_1, X_2, \ldots, X_n$. Takovou statistikou je například **úhrn** $\sum_{i=1}^{n} X_i$. Výběrové charakteristiky lze spočítat pouze na základě znalosti vstupních dat, bez znalosti parametrů skutečného rozdělení.

Výběrový k-tý obecný moment je statistika $m'_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, k = 1, 2, \ldots$. První obecný výběrový moment $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ je mírou polohy a nazývá se **výběrový průměr**.

Výběrový k-tý centrální momentje roven $m_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}^k), k = 1, 2, \ldots$. Pomocí druhého výběrového centrálního momentu m_2 je definován

výběrový rozptyl $s^2 = \frac{n}{n-1} m_2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$ a **výběrová směrodatná odchylka** $s = \sqrt{s^2}$. Obě tyto statistiky jsou mírami rozptýlenosti, podobně jako **variační koeficient** $V = s/\overline{X}$.

Pomocí druhého, třetího a čtrvtého centrálního momentu je definován **výběrový koeficient šikmosti** $\gamma_3 = \frac{m_3}{m_2^{3/2}}$ a **výběrový koeficient špičatosti** $\gamma_4 = \frac{m_4}{m_2} - 3$.

Při výpočtu některých statistik používáme takzvaný **uspořádaný výběr** $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$, který vznikne vzestupným uspořádáním prvků výběru $\{X_1, X_2, \ldots, X_n\}$ podle velikosti jeho realizací $\{x_1, x_2, \ldots, x_n\}$. Prvky $X_{(1)}$, $X_{(2)}, \ldots, X_{(n)}$ uspořádaného výběru nazýváme **pořádkovými statistikami**. Tak například první pořádková statistika $X_{(1)}$ je vlastně náhodnou veličinou z výběru, u níž byla naměřena nejmenší hodnota, neboli **výběrové minimum**. Podobně, $X_{(n)}$ je **výběrovým maximem**. Rozdíl $X_{(n)} - X_{(1)}$ nazýváme **výběrovým rozpětím**.

Pomocí pořádkových statistik definujeme **empirickou distribuční funkci** předpisem

$$F_n(x) = \begin{cases} 0, & x < X_{(1)}, \\ \frac{i}{n}, & X_{(i-1)} \le x < X_{(i)}, i = 1, \ldots, n, \\ 1, & x \le X_{(n)}. \end{cases}$$

100p% výběrový kvantil je definován pro $0 \le p \le 1$ následujícím způsobem:

$$\tilde{x}_{100p} = \begin{cases} X_{([np]+1)} & \text{pokud není } np \text{ celé číslo}, \\ \frac{1}{2}(X_{(np)} + X_{(np+1)}), & \text{pro } np \text{ celé}. \end{cases}$$

Tak například **dolní výběrový kvartil** $\tilde{x}_{25} = X_{([\frac{n}{4}]+1)}$, pokud je n nedělitelné 4, $\tilde{25} = \frac{1}{2}(X_{(\frac{n}{4})} + X_{\frac{n}{4}+1})$ pro n dělitelné 4. Podobně **výběrový medián** $\tilde{x}_{50} = X_{([\frac{n}{2}]+1)}$ pokud je n liché, $\tilde{50} = \frac{1}{2}(X_{(\frac{n}{2})} + X_{\frac{n}{2}+1})$ pro n sudé a **horní výběrový kvartil** $\tilde{75} = (X_{[\frac{3}{4}n)]+1}$, pokud je n nedělitelné 4, $\tilde{75} = \frac{1}{2}(X_{(\frac{3}{4}n)} + X_{\frac{3}{4}n+1})$ pro n dělitelné 4. Jako míra rozptýlenosti dat se používá také **výběrové mezikvartilové rozpětí** $\tilde{x}_{75} - \tilde{x}_{25}$.

2.1.5 Rozdělení některých statistik.

Jelikož jsou statistiky funkcemi náhodných veličin, jsou také náhodnými veličinami. Abychom s nimi mohli pracovat a poznat jejich vlastnosti, je

třeba znát jejich rozdělení pravděpodobnosti. Bohužel, explicitní odvození rozdělení statistik bývá složité a ne vždy možné. Toto rozdělení závisí na rozdělení výběru. Známe-li rozdělení výběru, můžeme při stanovení statistik postupujeme podle postupů popsaných v kapitole II.5. a III.4.

Příklad 2.1.2 *Nalezněte rozdělení výběrového průměru při výběru z Poissonova rozdělení s parametrem λ.*

Řešení: Momentová vytvořující funkce $M_X(t)$ Poissonova rozdělení je

$$M_X(t) = \sum_{x=0}^{\infty} e^{tx} e^{-\lambda} \frac{\lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} (\lambda e^t)^x \frac{1}{x!} = exp[\lambda(e^t - 1)]$$

. Momentová vytvořující funkce součtu $T = \sum_{j=1}^{n} X_j$ nezávislých veličin $X_1, X_2, \ldots, X_n$ je podle Věty III.3.6 rovna součinu jejich vytvořujících funkcí, a tedy

$$M_T(t) = \sum_{j=1}^{n} exp[\lambda(e^t - 1)] = exp[n\lambda(e^t - 1)]$$

. Odtud plyne, že statistika T má opět Poissonova rozdělení s parametrem $n\lambda$. Pro výběrový průměr $\overline{X} = \frac{1}{n}T$ potom dostáváme rozdělení

$$P(\overline{X} = \omega) = P(T = n\omega) = e^{-n\lambda} \frac{(n\lambda)^{n\omega}}{(n\omega)!}, \omega = 0, \frac{1}{n}, \frac{2}{n}, \ldots$$

.

Příklad 2.1.3 *Najděte střední hodnotu a rozptyl výběrového průměru $\overline{X}$*

Řešení: Jsou-li $X_1, \ldots, X_n$ nezávislé, stejně rozdělené náhodné veličiny se střední hodnotou μ a rozptylem σ^2, potom jejich výběrový průměr $\overline{X}$ má střední hodnotu

$$E(\overline{X}) = E(\frac{1}{n} \sum_{i=1}^{n} X_i) = \frac{1}{n} \sum_{i=1} nE(X_i) = \mu$$

a rozptyl

$$Var(\overline{X}) = Var(\frac{1}{n} \sum_{i=1}^{n} X_i) = \frac{1}{n^2} \sum_{i=1} nVar(X_i) = \sigma^2$$

Příklad 2.1.4 *Uvažujme výběr $\{X_1, \ldots, X_n\}$ z normálního rozdělení $N(\mu, \sigma^2)$. Jaké má rozdělení pravděpodobnosti výběrový průměr $\overline{X}$?*

Řešení: Podle příkladu III.3.7 má součet normálně rozdělených nezávislých veličin opět normální rozdělení. Tedy statistika $T = \sum_{j=1}^{n} X_j$ bude mít rozdělení $N(n\mu, n\sigma^2)$. Rozdělení výběrového průměru $\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j = \frac{1}{n}T$ bude potom $N(\mu, \frac{1}{n}\sigma^2)$.

Z centrální limitní věty (věta IV.2.4) plyne, že pro výběry z libovolného rozdělení s konečnou střední hodnotou μ, s konečným rozptylem σ^2 a s velkým rozsahem n, lze rozdělení výběrového průměru přibližně nahradit normálním rozdělením $N(\mu, \frac{1}{n}\sigma^2)$.

Pro praktické aplikace jsou významné výběry z rozdělení $N(\mu, \sigma^2)$. S tímto rozdělením jsou spojena některá další důležitá rozdělení pravděpodobnosti, s nimiž se v této části skript budeme setkávat.

Má-li náhodná veličina X rozdělení $N(0, 1)$, potom veličina X^2 bude mít tzv. rozdělení **chí-kvadrát s jedním stupněm volnosti**, které označujeme $\chi^2(1)$ a jehož hustotu lze snadno odvodit podle II.5.4. Uvažujme výběr $X_1, X_2, \ldots, X_n$ z rozdělení $N(0, 1)$. Potom rozdělení statistiky $\chi_n^2 = X_1^2 + X_2^2 + \cdots + X_n^2$ označujeme $\chi^2(n)$ a nazýváme jej **chí-kvadrát o** n **stupních volnosti**. Hustota tohoto rozdělení je kladná pouze pro kladné hodnoty argumentu a pro statistiku χ_n^2 platí $E(\chi_n^2) = n$, $Var(\chi_n^2) = 2n$.

Úpravou vztahu pro výběrový rozptyl s^2 dostaneme

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{\sigma^2}{n-1}\sum_{i=1}^{n}(\frac{X_i - \mu}{\sigma} - \frac{\overline{X} - \mu}{\sigma})^2$$

Označme $U_0 = \frac{\overline{X}-\mu}{\sigma}\sqrt{n}$ a $U_i = \frac{X_i-\mu}{\sigma}, i = 1, \ldots, n$. Tyto náhodné veličiny mají rozdělení $N(0, 1)$ a platí $U_0 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i$. Potom

$$S^2 = \frac{\sigma^2}{n-1}[\sum_{i=1}^{n} U_i^2 - 2\frac{U_0}{\sqrt{n}}\sum_{i=1}^{n} U_i + U_0^2] = \frac{\sigma^2}{n-1}[\sum_{n-1}^{n} U_i^2 - U_0^2].$$

V [An] je ukázáno, že tato statistika má rozdělení $\chi^2(n-1)$.

Jsou-li U a χ_n^2 dvě nezávislé náhodné veličiny s rozdělením $N(0, 1)$ a $\chi^2(n)$, potom rozdělení pravděpodobnosti statistiky

$$T = \sqrt{n}\frac{U}{\sqrt{\chi_n^2}}$$

se nazývá **rozdělení** t (nebo též **Studentovo**) **o** n **stupních volnosti**, které budeme značit $t(n)$. Hustota $g_n(t)$ tohoto rozdělení je symetrická kolem nuly. Tedy pro distribuční funkci $G_n(t)$ platí (podobně jako pro distribuční funkci

standardního normálního rozdělení) $G_n(t) = 1 - G_n(-t)$ a pro $100\alpha\%$-ní kvantily $t_n(\alpha)$, definované vztahem $G_n(t_n(\alpha)) = \alpha, 0 \leq \alpha \leq 1, n = 1, 2, \ldots,$ platí vztah $t_n(\alpha) = -t_n(1-\alpha)$.

Uvažujme statistiky $\overline{X}$ a s^2 výběru $\{X_1, X_2, \ldots, X_n\}$ z rozdělení $N(\mu, \sigma^2)$. Lze ukázat, že tyto dvě statistiky jsou nezávislé náhodné veličiny, veličina $U = \sqrt{n}\frac{\overline{X}-\mu}{\sigma}$ má rozdělení $N(0,1)$ a veličina $Z = \frac{(n-1)S^2}{\sigma^2}$ má rozdělení $\chi^2(n-1)$. Potom statistika $T = \sqrt{n-1}\frac{U}{\sqrt{Z}} = \sqrt{n}\frac{\overline{X}-\mu}{s}$ má rozdělení $t(n-1)$.

Označme

$$F = \frac{X_m^2/m}{X_n^2/n}$$

kde X_m^2 a X_n^2 jsou dvě nezávislé náhodné veličiny s rozdělením $\chi^2(m)$ a $\chi^2(n)$. Potom rozdělení pravděpodobnosti statistiky F nazýváme **rozdělením** F, (respektive **Fisher-Snedeckorovým**) **o** m **a** n **stupních volnosti** budeme jej označovat $F(m,n)$. Toto rozdělení je asymetrické, kladné pouze pro kladné hodnoty argumentu. Pro jeho $100\alpha\%$ kvantily $F_{m,n}(\alpha)$ platí pro každou dvojici $m, n = 1, 2, \ldots$ následující vztah

$$F_{m,n}(\alpha) = \frac{1}{F_{n,m}(1-\alpha)}, 0 < \alpha < 1$$

.

Předpokládejme, že byly provedeny dva nezávislé výběry z rozdělení $N(\mu_1, \sigma_1^2)$ a $N(\mu_2, \sigma_2^2)$ a rozsazích m a n. Označme S_1^2 a S_2^2 jejich výběrové rozptyly. Potom statistika

$$F = \frac{\frac{1}{m-1}\frac{(m-1)S_1^2}{\sigma^2}}{\frac{1}{n-1}\frac{(n-1)S_2^2}{\sigma_2^2}} = \frac{S_1^2}{S_2^2}$$

má rozdělení $F(m-1, n-1)$.

2.2 Odhady parametrů rozdělení náhodné veličiny

(Upozornění: tento text ještě nebyl upraven do konečné podoby. Proto zde chybějí některé obrázky a objevují se zde odkazy na odstavce, které nejsou označeny odpovídajícím způsobem. Tyto nedostatky budou postupně odstraňovány. Do té doby prosím o trpělivost a omlouvám se za ztížené čtení. GDo)

Při statistickém pozorování sledujeme určitou náhodnou veličinu, aniž bychom znali její pravděpodobnostní charakteristiky. Jediná informace, kterou o této veličině máme, je skryta v datech, v napozorovaných (naměřených) hodnotách. Proto musíme všechny její charakteristiky z těchto dat „odhadnout“. k tomuto účelu je zaměřena jedna část matematické statistiky, takzvaná teorie odhadu. Typy odhadů jsou v zásadě tři:

- odhady rozdělení,
- odhady parametrů za předpokladu nějakého rozdělení (takzvané „parametrické“ odhady),
- odhady charakteristik bez předpokladu rozdělení pravděpodobnosti (tyto odhady označujeme jako "neparametrické").

Odhady typu rozdělení provádíme zpravidla graficky, podle histogramu četností či z empirické distribuční funkce srovnáním se známými typy rozdělení. V některých případech odhadujeme typ rozdělení podle logického či fyzikálního modelu experimentu. Takto vytvořené odhady poté ověřujeme pomocí statistických testů (takzvané „testy dobré shody“).

Parametrické odhady a odhady pravděpodobnostních charakteristik rozdělujeme ještě podle jejich charakteru na

- bodové odhady a
- odhady intervalové

V prvním případě dostáváme sice přesné, ale zato zcela nespolehlivé hodnoty odhadující hledané charakteristiky. Ve druhém případě obdržíme nepřesně určenou hodnotu pomocí intervalu, zato ale víme, že skutečná hodnota se v tomto intervalu bude vyskytovat s předem danou, zpravidla dosti vysokou pravděpodobností.

V jistém smyslu se na odhady – nebo přesněji na odhadové statistiky – díváme jako na náhodné veličiny. Sloo "statistika"je zde použito ve smyslu "funkce náhodného výběru", tedy funkce náhodných veličin $X_1, X_2, \ldots, X_n$, což je opět náhodná veličina s pravděpodobnostním rozdělením, střední hodnotou, rozptylem a všemi dalšími charakteristikami.

2.2.1 Bodové odhady a jejich vlastnosti

V předchozí kapitole jsme popsali některé charakteristiky výběru, který si lze představit jako nezávislá pozorování jisté náhodné veličiny, získané pouze na základě naměřených hodnot. V některých případech však máme k dispozici více informací o výběru, například informaci o typu pravděpodobnostního rozdělení sledované náhodné veličiny. Potom vyvstává otázka identifikace tohoto rozdělení, tedy určení - **odhad** - jeho parametrů na základě pozorování náhodné veličiny, na základě výběru.

Předpokládejme, že máme k dispozici výběr $X = \{X_1, \ldots, X_n\}$ z rozdělení s distribuční funkcí $F(x, \theta)$, kde $\theta = (\theta_1, \ldots, \theta_k)$ je vektor reálných parametrů, které neznáme. Víme o nich pouze to, že leží v nějaké množině $\Theta \leq R^k$, v tzv. **parametrickém prostoru**. Nahradíme-li neznámé hodnoty parametrů $\theta_1, \ldots, \theta_k$ hodnotami statistik (reálných funkcí výběru) $T_1(X), \ldots, T_k(X)$, budeme $T(X) = (T_1(X), T_k(X))$ nazývat **odhadovou statistikou** pro parametry $(\theta_1, \ldots, \theta_k)$. Jestliže za X dosadíme hodnoty pozorování x, dostaneme tzv. **bodový odhad** $T(x) = (T_1(x), \ldots, T_k(x))$ těchto parametrů, získaný na základě pozorování $x = (x_1, \ldots, x_n)$.

Definice 2.1 *Bodový odhad se nazývá* ***nestranný****, jestliže pro střední hodnotu odhadové statistiky platí $E(T) = \theta$ pro všechna $\theta \in \Theta$. Neplatí-li tato rovnost, pak tento odhad nazýváme* ***vychýleným*** *a rozdíl $B(\theta) = E(T) - \theta$ je* ***vychýlení*** *odhadu T.*

Příklad 2.2.1 *Je druhý centrální výběrový moment nestranným odhadem rozptylu?*

Řešení: Druhý centrální výběrový moment je roven $M_2(x) = \frac{1}{n}\sum_{i=1}^{2}(x_i - \bar{x})^2$. Zzískáme jej na základě nezávislých pozorování $x_1, \ldots, x_n$ náhodné veličiny X. Spočteme-li střední hodnotu $E(M_2(X))$, dostaneme s použitím výsledku příkladu V.4.2

$$E(M_2(X)) = \frac{1}{n}E\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n}E(X_i - \mu + \mu - \bar{X})^2 =$$

$$\frac{1}{n}\sum_{i=1}^{n}[E(X_i - \mu)^2 - 2E(X_i - \mu)(\bar{X} - \mu) + E(\bar{X} - \mu)^2] =$$

$$\frac{1}{n}\sum_{i=1}^{n}(\sigma^2 - \frac{1}{n}\sigma^2) = (1 - \frac{1}{n})\sigma^2$$

Tedy tento odhad není nestranný a jeho vychýlení je $-\frac{1}{n}\sigma^2$.

Abychom dostali odhad nevychýlený, musíme M_2 vynásobit konstantou $\frac{n}{n-1}$. Tím dostáváme jiný odhad rozptylu, takzvaný **výběrový rozptyl** $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Tento odhad už je nevychýleným odhadem.

Definice 2.2 ***Střední kvadratická chyba*** *odhadu $T(X)$ je hodnota výrazu*

$$E(T(X) - \theta)^2 = D(T(X)) + B(\theta)^2$$

Příklad 2.2.2 *Spočtěte střední kvadratickou chybu statistiky s^2 jako odhadu rozptylu σ^2 náhodné veličiny X na základě výběru $X_1, \ldots, X_n$.*

Řešení: Jest

$$E(s^2 - \sigma^2)^2 = E(s^4) - 2\sigma^2 E(s^2) + \sigma^4 = E(s^4) - \sigma^4 = \frac{2n-1}{n^2}\sigma^4,$$

což je méně, než pro s^2, neboť platí nerovnost $\frac{2n-1}{n^2} < \frac{2}{n-1}$. Tedy každý z těchto dvou odhadů je „lepší“ v jiném smyslu.

Definice 2.3 *Nestranný odhad neznámého parametru θ, který má nejmenší rozptyl mezi všemi nestrannými odhady $T(X)$ parametru θ, se nazývá* ***nejlepším nestranným*** *odhadem parametru θ.*

Definice 2.4 *Platí-li pro skutečnou hodnotu θ a libovolné $\epsilon > 0$*

$$\lim_{n\to\infty} P(|T(X) - \theta| \geq \epsilon) = 0,$$

řekneme, že odhad $T(X)$ je ***konzistentním odhadem*** *parametru θ. Tento požadavek odpovídá představě, při které se vzrůstajícím počtem měření se stále více přibližujeme skutečné hodnotě θ.*

Věta 2.1 *Platí-li pro odhad $T(X)$*

$$\lim_{n\to\infty} B(\theta) = 0 a \lim_{n\to\infty} D(T(X)) = 0,$$

potom odhad $T(X)$ je konzistentním odhadem parametru θ.

Důkaz: plyne z aplikace Čebyšeovy nerovnosti (viz II.2.11) na pravděpodobnost v definici VI.1.7.

Příklad 2.2.3 *Ukažte, že pro náhodný výběr $X_1, \ldots, X_n$ z rozdělení se střední hodnotou μ a konečným rozptylem σ^2, je konzistentním odhadem střední hodnoty aritmetický průměr.*

Řešení: Toto tvrzení vyplývá přímo z příkladu V.4.2. Podle něho je $E(\bar{X}) = \mu$ a $D(\bar{X}) = \frac{\sigma^2}{n}$. Odtud

$$B(\mu) = E(\bar{X} - \mu) = E(\bar{X}) - \mu = 0$$

a

$$\lim_{n\to\infty} D(\bar{X}) = \lim_{n\to\infty} frac\sigma^2 n = 0$$

a tedy platí předpoklady Tvrzení VI.1.8.

2.2.2 Momentová metoda

Nejjednodušší metodou pro konstrukci bodových odhadů parametrů známého rozdělení náhodné veličiny je takzvaná **momentová metoda**. Při této metodě srovnáváme obecné nebo centrální momenty náhodné veličiny s předpokládaným rozdělením, vyjádřené pomocí neznámých parametrů, s výběrovými momenty spočtenými z naměřených dat. Pokud pozorovaná data odpovídají předpokládanému rozdělení, měly by se tyto momenty rovnat. Tím dostáváme soustavu rovnic, v níž jako neznámé vystupují hledané parametry. Zřejmě stačí tolik rovnic, kolik parametrů chceme odhadovat. Řešením takovéto soustavy jsou potom bodové odhady těchto parametrů, získané momentovou metodou.

Příklad 2.2.4 *Najděte bodové odhady parametrů normálního rozdělení.*

Řešení: Použijeme momentovou metodu. Máme odhadnout dva parametry, tedy potřebujeme dvě rovnice. k tomu použijeme první dva obecné momenty normálního rozdělení $N(\mu, \sigma^2)$:
$M_1 = E(X) = \mu$,
$M_2 = E(X^2) = Var(X) + (E(X))^2 = \sigma^2 + \mu^2$.
Tomu odpovídající výběrové momenty jsou
$m_1 = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$,
$m_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$.
Ze soustavy dvou rovnic

$M_1 = m_1$,
$M_2 = m_2$,
dostáváme bodové odhady parametrů μ a σ^2 momentovou metodou
$\tilde{\mu} = m_1 = \bar{x}$
$\tilde{\sigma}^2 = m_2 - m_1^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2$.

2.2.3 Maximálně věrohodné odhady.

Při hledání bodového odhadu zpravidla požadujeme, aby byl z určitého hlediska co nejlepší, tedy například nejlepší nestranný a zároveň konzistentní apod. To je ovšem v některých případech nemožné, neboť takový odhad ani nemusí existovat. Poměrně nejlepší výsledky při hledání bodových odhadů poskytuje **metoda maximální věrohodnosti**. Odhady získané touto metodou se nazývají **maximálně věrohodné odhady**.

Uvažujme náhodný výběr $X_1, \ldots, X_n$ ze spojitého rozdělení s hustotou $f(x, \theta)$, závisející na jednom neznámém parametru $\theta \in \Theta$. Potom sdružená hustota náhodného vektoru $(X_1, \ldots, X_n)$ je rovna

$$g(x, \theta) = f(x_1, \theta) f(x_2, \theta) \ldots f(x_n, \theta)$$

(Tento výraz je rozšířením multiplikativní vlastnosti z III.3.1 na n náhodných veličin). Funkci $g(x, \theta)$ budeme nyní chápat jako funkci proměnné θ při pevných hodnotách $x_1, \ldots, x_n$ a budeme jí říkat **věrohodnostní funkce**. Namísto této funkce je někdy výhodnější pracovat s jejím logaritmem a potom budeme mluvit o **logaritmické věrohodnostní funkci** $L(x, \theta) = ln(g, (x, \theta))$.

Uvažujme nyní prostý náhodný výběr $X_1, \ldots, X_n$ z diskrétního rozdělení s pravděpodobnostmi $p_i = P(X_i = x_i) = p(x_i, \theta)$, závisejícími na jednom neznámém parametru $theta \in \Theta$. Potom za věrohodnostní funkci budeme považovat funkci

$$g(x, \theta) = p(x_1, \theta) p(x_2, \theta) \ldots p(x_n, \theta)$$

jako funkci proměnné θ a logaritmická věrohodnostní funkce bude opět dána vztahem $L(x, \theta) = ln(g, (x, \theta))$.

Za maximálně věrohodný odhad parametru θ budeme považovat tu hodnotu $\theta \in \Theta$, při které je hodnota věrohodnostní funkce maximální (při dané realizaci $x_1, \ldots, x_n$ je „nejvěrohodnější“). Získáme ji řešením tzv. **věrohodnostní rovnice**

$$\frac{\partial g(x, \theta)}{\partial \theta} = 0.$$

Vzhledem k tomu, že funkce logaritmus je rostoucí funkce, lze maximum věrohodnostní funkce nalézt také maximalizací logaritmické věrohodnostní funkce, což bývá někdy výhodnější, tj. jako řešení **logaritmické věrohodnostní rovnice**

$$\frac{\partial L(x,\theta)}{\partial \theta} = 0.$$

Výše popsaný postup lze zobecnit na případ, kdy je neznámých parametrů více, tedy na případ odhadu vektorového parametru $\theta = \theta_1, \ldots, \theta_k$.

Potom budeme hledat extrém funkce $g(x,\theta)$, resp. $L(x,\theta)$. To vede k řešení **soustavy věrohodnostních rovnic**

$$\frac{\partial g(x,\theta)}{\partial \theta_i} = 0, i = 1, 2, \ldots, k$$

Příklad 2.2.5 *Metodou maximální věrohodnosti určete odhad neznámého parametru p binomického rozdělení, je-li n známé.*

Řešení: Předpokládejme, že máme k dispozici m nezávislých pozorování $x = x_1, \ldots, x_m$ náhodné veličiny s tímto rozdělením. Věrohodnostní funkce má podle II.3.2 a VI.2.2 tvar

$$g(x,n,p) = \sum_{i=1}^{m} \binom{n}{x_i} p^{x_i}(1-p)^{n-x_i}.$$

Zřejmě v tomto případě bude výhodnější pracovat s logaritmickou věrohodnostní funkcí

$$L(x,n,p) = \sum_{i=1}^{m} \left(ln\binom{n}{x_i} + x_i ln(p) + (n-x_i) ln(1-p) \right).$$

řešit logaritmickou věrohodnostní rovnicí

$$\frac{\partial L(x,n,p)}{\partial p} = \sum_{i=1}^{m} \left(\frac{x_i}{p} - \frac{n-x_i}{1-p} \right) = \frac{1}{p} \sum_{i=1} m x_i - \frac{1}{1-p} \sum_{i=1}^{m} (n - x_i) =$$
$$\frac{1}{p(1-p)} \left[(1-p) \sum_{i=1}^{m} x_i - pmn - p \sum_{i=1}^{m} x_i \right] = 0. \quad (2.1)$$

Za předpokladu $p \neq 0$ a $p \neq 1$ (pro $p = 0$ nebo $p = 1$ nabývá věrohodnostní funkce nulovou hodnotu, což je její minimum) dostaneme rovnici

$$(1-p)m\bar{x} - pmn - pm\bar{x} = 0,$$

kde $\bar{x}$ je aritmetický průměr čísel $x_1, \ldots, x_m$. Poslední rovnost po vydělení číslem m bude

$$p\bar{x} = (1-p)(n-\bar{x})$$

a její řešení má tvar $\hat{p} = \frac{1}{n}\bar{x}$, což je hledaný maximálně věrohodný odhad parametru p.

2.2.4 Intervaly spolehlivosti

Odhadová statistika $T(X)$ je náhodná veličina. Bodový odhad $\hat{\theta}$ parametru θ je získán na základě konkrétních pozorování $x_1, \ldots, x_n$ jako hodnota $T(x)$ odhadové statistiky. Proto lze očekávat, že takto získaná hodnota bude pouze „blízko“ skutečné hodnotě parametru θ. Jak „blízko“, to z bodového odhadu nelze zjistit. Proto se často používají tzv. **intervaly spolehlivosti**.

Nechť $\underline{\theta}, \bar{\theta}$, jsou dvě statistiky takové, že pro interval $(\underline{\theta}, \bar{\theta})$ platí

$$P(\theta \in (\underline{\theta}, \bar{\theta})) = 1 - \alpha,$$

tj. pravděpodobnost, že skutečná hodnota parametru leží v tomto intervalu je rovna $(1-\alpha)$, kde $\alpha \in (0,1)$, nazveme **intervelem spolehlivosti** pro parametr α s **koeficientem spolehlivosti** $(1-\alpha)$. Interval spolehlivosti se často také nazývá $100(1-\alpha)\%$ **intervalem spohlehlivosti** nebo též **konfidenčním intervalem**.

Příklad 2.2.6 *Najděte interval spolehlivosti pro střední hodnotu μ, máme-li k dispozici náhodný výběr z normálního rozdělení $N(\mu, \sigma^2)$ se známou hodnotou σ.*

Řešení: Ke konstrukci použijeme statistiku $\frac{\bar{X}-\mu}{\sigma}\sqrt{n}$, která má rozdělení $N(0,1)$ (viz V.4.2 a II.5.1). Platí tedy rovnost

$$P\left(\frac{\bar{X}-\mu}{\sigma}\sqrt{n} \leq u(1-\frac{\alpha}{2})\right) = 1-\alpha,$$

kde $u(1-\frac{\alpha}{2})$ je $(1-\frac{\alpha}{2})$ - kvantil rozdělení $N(0,1)$. Odtud dostáváme úpravou nerovnosti v závorce vztah

$$P(\bar{X} - \frac{\sigma}{\sqrt{n}}u\ (1-\frac{\alpha}{2})) < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}}u\ (1-\frac{\alpha}{2}) = 1-\alpha,$$

tedy $100(1-\alpha)\%$ intervalem spolehlivosti je interval

$$\bar{X} - \frac{\sigma}{\sqrt{n}}u\ (1-\frac{\alpha}{2}) < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}}u\ (1-\frac{\alpha}{2})$$

Příklad 2.2.7 *Najděte interval spolehlivosti pro střední hodnotu μ, máme-li k dispozici náhodný výběr z normálního rozdělení $N(\mu, \sigma^2)$ s neznámou hodnotou σ.*

Řešení: Neznáme-li hodnotu žádného z parametrů μ a σ v předchozím příkladě, můžeme postupovat obdobně, pouze s tím rozdílem, že použijeme statistiku $\frac{\bar{X}-\mu}{s}\sqrt{n}$. Tato statistika má Studentovo t-rozdělení o $(n-1)$ stupních volnosti (viz V.4.8), a proto

$$P\left(\frac{|\bar{X}-\mu|}{s}\sqrt{n} > t_{n-1}(1-\frac{\alpha}{2})\right) = 1-\alpha,$$

kde $t_{n-1}(1-\frac{\alpha}{2})$ je $(1-\frac{\alpha}{2})$-kvantil rozdělení t o $(n-1)$ stupních volnosti. Tudíž interval

$$\bar{X} - \frac{s}{\sqrt{n}}t_(n-1)(1-\frac{\alpha}{2}) < \mu < \bar{X} + \frac{s}{\sqrt{n}}t_(n-1)(1-\frac{\alpha}{2})$$

je $100(1-\alpha)\%$ intervalem spolehlivosti μ.

2.3 Testování statistických hypotéz

(Upozornění: tento text ještě nebyl upraven do konečné podoby. Proto zde chybějí některé obrázky a objevují se zde odkazy na odstavce, které nejsou označeny odpovídajícím způsobem. Tyto nedostatky budou postupně odstraňovány. Do té doby prosím o trpělivost a omlouvám se za ztížené čtení. GDo)

2.3.1 Test statistické hypotézy

Pod pojmem **statistické hypotézy** si budeme představovat jakékoli tvrzení o jevu statistické povahy. Například tvrzení o délce životnosti výrobku, o nezávislosti výsledku na použité metodě, tvrzení o pravděpodobnostním rozdělení sledované veličiny a podobně. Ověřování, zda hypotéza platí či nikoli, je předmětem statistického testování. Toto provádíme na základě pozorování (měření) nějakého výběru (experimentu).

Test statistické hypotézy H **proti alternativní hypotéze** A je rozhodovací pravidlo, podle něhož na základě realizace náhodného výběru rozhodujeme mezi dvěma tvrzeními - sledovanou hypotézou a doplňkovou, tzv. **alternativní hypotézou** A. Výsledkem našeho rozhodování je buď zamítnutí hypotézy H ve prospěch alternativy A či její nezamítnutí. Skutečnost, že hypotézu nezamítáme, neznamená, že naměřená data tto hypotézu potvrzují, ale pouze to, že ji nevyvracejí. Toto rozhodovací pravidlo je určeno **testovou statistikou** $T(X)$ a množinou ν, které říkáme **kritický obor**. Vlastní rozhodování potom probíhá pomocí indikátorové funkce

$$I_\nu(T(X)) = \begin{cases} 1 & \text{pokud } T(X) \in \nu, \\ 0 & \text{pokud } T(X) \notin \nu. \end{cases}$$

Pokud je hodnota indikátorové funkce rovna 1, tedy $T(X) \in \nu$ potom hypotézu H zamítáme. V opačném případě říkáme, že hypotézu nelze zamítnout. Naše úsilí přitom zaměříme na konstrukci testu, to znamená na určení testové statistiky T a kritického oboru ν, pomocí nichž budeme moci co nejlépe rozhodnout o zamítnutí hypotézy.

Příklad 2.3.1 *Při dodávce rezistorů je pro nás z hlediska použitelnosti rozhodující velikost odporu součástky. Výrobce udává nominální hodnotu, od níž se však většina naměřených hodnot liší. Jak rozhodnout, zda je pro nás dodávka přijatelná.*

Řešení: Naše měření však může podléhat náhodným vlivům. Kontrola dodávky spočívá ve stanovení rozhodovacího pravidla, kterým chceme otestovat hypotézu, že skutečný odpor je roven nominální hodnotě.

Při výše popsaném rozhodovacím pravidle se můžeme dopustit chyby dvěma způsoby. Buď budeme příliš přísní a zamítneme hypotézu, která platí - to je **chyba prvního druhu** - nebo naopak tuto hypotézu nezamítneme, i když je nesprávná - v tomto případě se jedná o **chybu druhého druhu**. Obě mohou mít nepříjemné důsledky, a proto budeme zřejmě za „lepší" test považovat ten test, při kterém bude pravděpodobnost obou chyb co nejmenší. Přitom zpravidla čím menší bude pravděpodobnost chyby 1. druhu, tím větší bude pravděpodobnost chyby 2. druhu a naopak. V takovém případě nelze nalézt test minimalizující obě chyby současně. Proto postupujeme následujícím způsobem.

Při kontrukci testu požadujeme, aby pravděpodobnost chyby 1. druhu byl menší nebo rovna danému číslu α, kterému říkáme **hladina významnosti testu**. Přitom obvykle volíme $\alpha = 0,05; 0,01$ apod. Potom hledáme testovou statistiku $T(X)$ a kritický obor ν_α, tak aby

$$P(T(X)) \in \nu_\alpha | Hplati) \leq \alpha$$

$P(T(X)) \notin \nu_\alpha | H$ neplatí) byla minimální.

Kritický obor je zpravidla interval, ohraničený tzv. **kritickými hodnotami**. Test potom probíhá tak, že spočteme hodnotu testové statistiky, porovnáme ji s kritickými hodnotami, odpovídajícími hladině významnosti α, a rozhodneme o zamítnutí či nezamítnutí hypotézy.

V některých případech - především při testování pomocí počítače - se používá jiný postup. Spočte se hodnota testové statistiky a k ní nejmenší kritický obor, při kterém bychom ještě mohli na základě této hodnoty zamítnout hypotézu proti dané alternativě. Hladina významnosti, odpovídající tomuto kritickému oboru, se nazývá ***p*-hodnota**. Kdybychom volili hladinu významnosti větší, než je tato hodnota, mohli bychom ještě hypotézu zamítnout. Je-li tato p-hodnota příliš malá, hypotézu zamítáme. Například, spočteme-li pro daná data p-hodnotu rovnou $0,005$ znamená to, že pro jakékoliv α větší než $0,005$ bychom měli hypotézu zamítnout. Konkrétně pro $\alpha = 0,01$ už hypotézu zamítáme.

Širokou třídu testů tvoří testy hypotéz o parametrech pravděpodobnostního rozdělení. V tomto případě předpokládáme, že pravděpodobnostní rozdělení je určitého typu a závisí na neznámých parametrech. Předpokládejme, že neznámý parametr θ může nabývat hodnot z nějaké množiny Θ. Uvažujme hypotézu $H : \theta = \theta_0$. Alternativní hypotéza může být buď $A : \theta \neq \theta_0$,

takzvaná **oboustranná alternativa** při této alternativě mluvíme o **oboustranném testu**), nebo $A_1 : \theta \leq \theta_0$ či $A_2 : \theta \geq \theta_0$ tzv. **jednostranné alternativy** (jimž odpovídají tzv. **jednostranné testy**).

Funkci $P_\nu(\theta)$, která každé hodnotě parametru $\theta \in \Theta$ přiřadí pravděpodobnost $P(T(X) \in \nu|\theta)$ nazýváme **silofunkcí testu**. Je to vlastně pravděpodobnost zamítnutí hypotézy H, má-li parametr hodnotu θ. Hodnotu silofunkce v bodě $\theta = \theta_1$ nazýváme **silou testu vzhledem k alternativě** $\theta = \theta_1$ a používáme ji k pro hodnocení kvality testu.

2.3.2 Testy o parametrech normálního rozdělení

V první části tohoto odstavce uvedeme několik parametrických testů, které se používají při výběru $X = X_1, \ldots, X_n$ z rozdělení $N(\mu, \sigma^2)$. Tyto testy se zabývají hypotézami o parametrech μ a σ. Přitom rozlišujeme několik případů:

Předpokládejme, že hodnotu σ známe. Budeme testovat nulovou hypotézu $H : \mu = \mu_0$ proti alternativě $A : \mu \neq \mu_0$ na hladině významnosti α. Test založíme na statistice $\frac{\bar{X}-\mu}{\sigma}\sqrt{n}$, která má, za předpokladu platnosti hypotézy, rozdělení $N(0, 1)$. Kritický obor je potom určen nerovností

$$\frac{|\bar{X} - \mu|}{\sigma}\sqrt{n} > u(1 - \frac{\alpha}{2})$$

kde $u(1 - \frac{\alpha}{2})$ je $(1 - \frac{\alpha}{2})$-kvantil rozdělení $N(0, 1)$. Pokud bude hodnota výběrového průměru $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ získaná z pozorování $x_1, \ldots, x_n$ ležet v intervalu

$$\mu_0 - \frac{\sigma}{\sqrt{n}}u(1 - \frac{\alpha}{2}) < \bar{x} < \mu_0 + \frac{\sigma}{\sqrt{n}}u(1 - \frac{\alpha}{2})$$

hypotézu nezamítneme na hladině významnosti α. V opačném případě hypotézu zamítáme. *Poznámka:* Srovnejte tento výsledek s intervalem spolehlivosti v VI.3.2. Hypotézu H nezamítneme tehdy, když hypotetická hodnota μ_0 bude ležet v $100(1 - \alpha)\%$ intervalu spolehlivosti, zkonstruovaném na základě pozorování $x_1, \ldots, x_n$.

Při jednostranné alternativě $A_1 : \mu < \mu_0$ resp. $A_2 : \mu > \mu_0$ bude kritický obor určen nerovností

$$\frac{\bar{x} - \mu_0}{\sigma}\sqrt{n} \leq u(\alpha), resp. \frac{\bar{x} - \mu_0}{\sigma}\sqrt{n} \geq u(1 - \alpha)$$

Jednovýběrový t-test. Nejčastějším případem je test hypotézy $H : \mu = \mu_0$ proti alternativě $A : \mu \neq \mu_0$ na hladině významnosti α při neznámé hodnotě σ. Neznámou hodnotou σ nahrazujeme jejím odhadem S a test založíme

na statistice $\frac{\bar{x}-\mu_0}{S}\sqrt{n}$. Platí-li hypotéza H, má tato statistika Studentovo t-rozdělení o $(n-1)$ stupních volnosti (viz V.4.6). Kritický obor je potom určen nerovností

$$\frac{|\bar{x}-\mu_0|}{\S}\sqrt{n} > t_{n-1}(1-\frac{\alpha}{2})$$

kde $t_{n-1}(1-\frac{\alpha}{2})$ je $(1-\frac{\alpha}{2})$-kvantil rozdělení t o $n-1$ stupních volnosti.

Pro jednostranné testy proti alternativám $\mu < \mu_0$ resp. $\mu > \mu_0$ bude kritický obor určen nerovnostmi

$$\frac{\bar{x}-\mu_0}{\S}\sqrt{n} \leq -t_{n-1}(1-\alpha) \frac{\bar{x}-\mu_0}{\S}\sqrt{n} \geq t_{n-1}(1-\alpha)$$

Pro výběry o velkém rozsahu n lze t-test použít (přibližně) i bez předpokladu normality výběru. Podle centrální limitní věty IV.2.4 má totiž výběrový průměr $\bar{X}$ v limitě pro $n \to \infty$ normální rozdělení $N(\mu, \sigma^2)$ a tedy statistika $\frac{\bar{x}-\mu_0}{\S}\sqrt{n}$ má přibližně Studentovo t-rozdělení o $(n-1)$ stupních volnosti.

Chceme-li testovat hypotézu o rozptylu $H : \sigma^2 = \sigma_0^2$ proti alternativě $A : \sigma^2 \neq \sigma_0^2$ na hladině významnosti α, můžeme použít statistiku $\frac{(n-1)S^2}{\sigma^2}$, která má za platnosti hypotézy (viz V.4.6) rozdělení χ o $n-1$ stupních volnosti. Označme s^2 hodnotu výběrového rozptylu $s^2 = \frac{1}{n-1}\Sigma_(i=1)^n(x_i - \bar{x})$, získanou z pozorování $x_1, \ldots, x_n$. V tomto případě je kritický obor pro oboustrannou alternativu $A : \sigma^2 \neq sigma_0^2$ určen nerovnostmi

$$s^2 < \frac{\chi_(n-1)^2(\frac{\alpha}{2})\sigma_0^2}{n-1} \text{ a } \frac{\chi_(n-1)^2(\frac{1-\alpha}{2})\sigma_0^2}{n-1} < s^2$$

zatímco pro jednostranné alternativy $A_1 : \sigma^2 < \sigma_0^2$ resp. $A_2 : \sigma^2 > \sigma_0^2$ dostaneme kritické obory

$$s^2 < \frac{\chi_(n-1)^2(\alpha)\sigma_0^2}{n-1} \text{ resp. } \frac{\chi_(n-1)^2(1-\alpha)\sigma_0^2}{n-1} < s^2,$$

symbol $\chi_(n-1)^2(\alpha)$ zde označuje α-kvantil chí-kvadrát rozdělení o $(n-1)$ stupních volnosti.

Párový t-test. Sledujeme-li na jednom objektu dva podobné znaky zároveň, používáme náhodný výběr dvojic náhodných veličin $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$. O veličinách X_i a Y_i předpokládáme, že jsou **párově závislé**. Například měření vlastnosti materiálu před tepelným zpracováním a po něm na vybraných n vzorcích.

Předpokládejme, že $\{X_1, \ldots, X_n\}$ je náhodný výběr z normálního rozdělení $N(\mu_X, \sigma^2)$ a $Y_1, \ldots, Y_n$ je náhodný výběr z rozdělení $N(\mu_Y, \sigma_Y^2)$. Veličiny

X_i a Y_i mohou být párově závislé. Budeme testovat hypotézu o rovnosti středních hodnot $H : \mu_X = \mu_Y$ proti alternativě $A : \mu_X \neq \mu_Y$ na hladině významnosti α. V tomto případě budeme místo původně sledovaných veličin $(X_1, Y_1), \ldots, (X_n, Y_n)$ pracovat s veličinami $Z_1, \ldots, Z_n$, kde $Z_i = Y_i - X_i, i = 1, \ldots, n$. Protože X_i a Y_i mají normální rozdělení, bude se i veličina Z řídit normálním rozdělením se střední hodnotou $\mu_Z = \mu_Y - \mu_X$ a rozptylem σ_Z^2, o jehož vztahu k rozptylům σ_X a σ_Y nelze vzhledem k možné závislosti nic předpokládat. Rovnost středních hodnot X a Y je ekvivalentní nulovosti střední hodnoty rozdílu Z. Pro aritmetický průměr platí $\bar{z} = \bar{x} - \bar{y}$ a hodnotu výběrového rozptylu s_Z^2 spočteme podle vztahu

$$s_Z^2 = \frac{1}{n-1} \Sigma_{n=1}^{n} (x_i - y_i - \bar{x} + \bar{y})^2.$$

K testu hypotézy $H : \mu_Z = 0$ na hladině významnosti α použijeme jednovýběrový t-test (viz VII.2.2), tedy při oboustranné alternativě $A : \mu \neq 0$ hypotézu H zamítneme, pokud

$$\bar{z} = -\frac{s_Z}{\sqrt{n}} t_{n-1}(1 - \frac{\alpha}{2}) \text{ nebo } \frac{s_Z}{\sqrt{n}} t_{n-1}(1 - \frac{\alpha}{2}) < \bar{z}.$$

Jsou-li veličiny X a Y nezávislé, používáme pro srovnání středních hodnot dvou výběrů **dvouvýběrový t-test**. Nechť $X_1, \ldots, X_n$ je výběr z rozdělení $N(\mu_X, \sigma_X^2)$ a $Y_1, \ldots, Y_m$ je výběr z rozdělení $N(\mu_Y, \sigma_Y^2)$ a tyto výběry jsou na sobě nezávislé. Rozlišujeme dva případy:

(**1**) Oba výběry mají **stejné rozptyly** $\sigma_X^2 = \sigma_Y^2$. Potom statistika

$$T = \frac{\bar{X} - \bar{Y}}{S} \sqrt{\frac{mn}{m+n}},$$

kde

$$S^2 = \frac{1}{m+n-2} \left[\Sigma_{i=1}^{n} (X_i - \bar{X})^2 + \Sigma_{j=1}^{m} (Y_j - \bar{Y})^2 \right]$$

má za platnosti nulové hypotézy $H : \mu_X = \mu_Y$ Studentovo t-rozdělení o $m+n-2$ stupních volnosti. Test uvedené hypotézy proti oboustranné alternativě $A : \mu_X \neq \mu_Y$ na hladině významnosti α lze tedy založit na nerovnosti

$$\frac{\bar{y} - \bar{x}}{S} \sqrt{\frac{mn}{m+n}} \geq t_{m+n-2}(1 - \frac{1}{2}\alpha).$$

Test hypotézy H proti jednostranným alternativám $A_1 : \mu_X \leq \mu_Y$, resp. $A_2 : \mu_X \geq \mu_Y$ na hladině významnosti α je založen na nerovnostech

$$\begin{aligned} \frac{\bar{y} - \bar{x}}{S} \sqrt{\frac{mn}{m+n}} &\geq t_{m+n-2}(1 - \alpha), \\ \frac{\bar{y} - \bar{x}}{S} \sqrt{\frac{mn}{m+n}} &\geq t_{m+n-2}(\alpha) = -t_{m+n-2}(\alpha)(1 - \alpha) \end{aligned}$$

(**2**) Oba výběry mají **různé rozptyly** $\sigma_X^2 \neq \sigma_Y^2$. Potom použijeme přibližný test, založený na statistice

$$\widetilde{T} = \frac{\bar{X} - \bar{Y}}{\widetilde{S}}, \text{ kde } \widetilde{S}^2 = \frac{1}{n}s_X^2 + \frac{1}{m}s_Y^2.$$

Test hypotézy $H : \mu_X = \mu_Y$ proti oboustranné alternativě na hladině významnosti α lze založit na nerovnosti

$$\frac{|\bar{y} - \bar{x}|}{\widetilde{S}} \geq \frac{1}{\widetilde{S}^2}\left[\frac{1}{n}s_X^2 t_{n-1}(1 - \frac{\alpha}{2}) + \frac{1}{m}s_Y^2 t_{m-1}(1 - \frac{\alpha}{2})\right]$$

na nerovnosti

$$\frac{\bar{y} - \bar{x}}{\widetilde{S}} \geq \frac{1}{\widetilde{S}^2}\left[\frac{1}{n}s_X^2 t_{n-1}(1 - \alpha) + \frac{1}{m}s_Y^2 t_{m-1}(1 - \alpha)\right]$$

při jednostranné alternativě $A_1 : \mu_X \leq \mu_Y$ a na

$$\frac{\bar{y} - \bar{x}}{\widetilde{S}} \leq -\frac{1}{\widetilde{S}^2}\left[\frac{1}{n}s_X^2 t_{n-1}(1 - \alpha) + \frac{1}{m}s_Y^2 t_{m-1}(1 - \alpha)\right]$$

při alternativě $A_2 : \mu_X \geq \mu_Y$.

Příklad 2.3.2 *Při zpracování je třeba materiál zahřát na vysokou teplotu. Před zpracováním bylo vybráno náhodně 10 vzorků a změřena jejich tvrdost. Po zpracování bylo opět vybráno náhodně jiných 10 vzorků, na nichž byla změřena tvrdost. Naměřené hodnoty jsou v následující tabulce:*

před	*3,15*	*2,98*	*3,00*	*2,75*	*3,21*	*3,33*	*2,95*	*2,81*	*3,26*	*2,88*
po	*3,21*	*2,99*	*3,11*	*2,91*	*3,22*	*3,28*	*3,09*	*3,00*	*3,28*	*2,99*

Testujte hypotézu, že se tvrdost materiálu vlivem zpracování nemění.

Řešení: Je $m = n = 10$. Spočteme $\bar{x} = 3,032, \bar{y} = 3,108, s_x^2 = 0,03875, s_y^2 = 0,018$. Za předpokladu, že rozptyl před i po zpracování zůstává stejný (naměřený rozdíl je nevýznamný), použijeme postup, popsaný v VII.2.6.a). Dostaneme hodnotu $s^2 = 0,02838$ a testové statistiky $T = 1,009$. Při hladině významnosti $\alpha = 0,05$ je $t_18(0,975) = 2,101$ a tedy hypotézu nelze zamítnout.

Příklad 2.3.3 *Uvažujme stejnou úlohu jako v předchozím příkladu, pouze s tím rozdílem, že sledovaná veličina je měřena před i po zpracování na 10 vzorcích, které byly náhodně vybrány před začátkem experimentu. Naměřená data zůstávají stejná.*

Řešení: V tomto případě je třeba vzít do úvahy závislost, která zde může být způsobena dalšími vlastnostmi vzorků. Proto použijeme párový t-test. Dostáváme $\bar{z} = \bar{x} - \bar{y} = -0,076, s_z = 0,07777$. Testová statistika zde bude mít hodnotu $T = 3,09$, kterou budeme srovnávat s číslem $t_9 = (0,975) = 2,262$. V tomto případě hypotézu zamítneme. Uvedené příklady ukazují, jaký vliv na výsledek může mít tzv. *návrh experimentu*. Druhý případ lépe vystihuje skutečnost, že naměřená data nejsou nezávislá a bere do úvahy další možné vlivy, plynoucí z individuality vzorků.

Uvažujme dva nezávislé výběry: $X_1, \ldots, X_n$ z rozdělení $N(\mu_X, \sigma_X^2)$ a $Y_1, \ldots, Y_n$ z rozdělení $N(\mu_Y, \sigma_Y^2)$ můžeme provést tzv. **test shody rozptylů** neboli **F-test**. K testu hypotézy $H : \sigma_X^2 = \sigma_Y^2$ lze použít například statistiku

$$F = \frac{S_X^2}{S_Y^2}$$

Rozdělení statistiky F je podle V.4.6 a V.4.9 za předpokladu H rozdělením F o $(n-1)$ a $(m-1)$ stupních volnosti. Kritický obor pro test hypotézy H proti oboustranné alternativě $A : \sigma_X^2 \neq \sigma_Y^2$ na hladině významnosti α je určen nerovnostmi

$$\frac{s_X^2}{s_Y^2} \geq F_{n-1,m-1}(1 - \frac{\alpha}{2}) a \frac{s_X^2}{s_Y^2} \leq F_{n-1,m-1}(\frac{\alpha}{2}) = \frac{1}{F_{n-1,m-1}(1 - \frac{\alpha}{2})}$$

kde $F_{n,m}(\alpha)$ je α-kvantil rozdělení $F(n,m)$. Tedy hypotézu zamítáme pro malá F blízká nule a pro velká F (při platnosti hypotézy by mělo být F blízké 1). Kritické obory pro test hypotézy H proti alternativám $A_1 : \sigma_X^2 \geq \sigma_Y^2$ $A_2 : \sigma_X^2 \leq \sigma_Y^2$ na hladině významnosti α jsou určeny nerovnostmi

$$\frac{s_X^2}{s_Y^2} \geq F_{n-1,m-1}(1-\alpha) resp. \frac{s_X^2}{s_Y^2} \leq F_{n-1,m-1}(\alpha) = \frac{1}{F_{n-1,m-1}}(1-\alpha)$$

2.3.3 Testy dobré shody

Zatímco při parametrických testech předpokládáme, že typ rozdělení výběru je znám, **testy dobré shody** prověřují hypotézy právě o typu rozdělení, z něhož byl výběr pořízen. Jedná se tedy o hypotézy o *shodě* teoretického a empirického rozdělení.

Test hypotézy, že náhodný výběr pochází z rozdělení se spojitou známou distribuční funkcí $F_0(x)$, můžeme provést pomocí takzvaného **Kolmogorov-Smirnovova testu**. Tento test je založen na statistice

$$D = sup_{x \in R} \Big\{ |F_n(x) - F_0(x)| \Big\},$$

kde empirická distribuční funkce $F_n(x)$ je definována v paragrafu V.3.6. Pomocí této definice lze statistiku D zapsat také ve tvaru

$$D = max_{1\leq i\leq n}\left\{ max\left[\left|F_0(X_{(i)}) - \frac{i}{n}\right|, \left|F_0(X_{(i)}) - \frac{i-1}{n}\right|\right]\right\}$$

Hypotéza $H : F(x) = F_0(x)$ bude zamítnuta na hladině významnosti α ve prospěch alternativy $A : F(x) \neq F_0(x)$ alespoň pro jedno x, jestliže $D \leq D_n(1-\alpha)$, přičemž hodnoty $D_n(1-\alpha)$ jsou malá n tabelovány viz [Ja]) a pro velká $n(n > 100)$ lze použít aproximaci (viz [Zv])

$$D_n(1-\alpha) \cong \sqrt{-\frac{1}{2n} ln \frac{\alpha}{2}}.$$

Tento test lze správně použít pouze pro takové hypotézy, které určují funkci F_0, jednoznačně, včetně jejích parametrů.

χ^2-test dobré shody vychází z třídního rozdělení náhodného výběru. Nejprve tedy provedeme rozklad naměřených hodnot do disjunktních třídních intervalů pomocí zvoleného dělení $a_0 < a_1 < a_2 < \cdots < a_k$ a spočteme četnosti n_i (viz V.2.1). Dále spočteme hypotetické pravděpodobnosti $p_i = F_0(a_i) - F_0(a_{i-1})$. Při volbě třídních intervalů se doporučuje dodržet zásadu aby teoretické četnosti np_i pro všechna i byly větší nebo alespoň rovny číslu 5. Hypotézu, že výběr je z rozdělení s distribuční funkcí $F_0(x)$ potom testujeme pomocí statistiky

$$\chi^2 = \Sigma_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

Při tomto postupu je třeba rozlišit dva případy:
a) Hypotézu určuje distribuční funkci jednoznačně, včetně parametrů. Potom má statistika χ^2 asyptoticky, to znamená přibližně pro velká n, rozdělení $\chi^2(k-1)$. Bude-li tedy

$$\chi^2 \geq \chi^2_{k-1}(1-\alpha)$$

kde $\chi^2_{k-1}(1-\alpha)$ je $(1-\alpha)$-kvantil rozdělení $\chi^2(k-1)$ zamítneme nulovou hypotézu $H : X_1, \ldots, X_n$ *pochází z rozdělení s distribuční funkcí* $F_0(x)$ proti alternativě A :*toto rozdělení je jiné*, na hladině významnosti α
b) Teoretické četnosti p_i závisí na l neznámých parametrech. V takovém případě použijeme při výpočtu p_i odhady těchto parametrů. Přitom se sníží počet stupňů volnosti rozdělení statistiky χ^2 právě o počet odhadnutých parametrů na $(k-1-l)$. Kritický obor při hladině významnosti α potom bude dán nerovností

$$\chi^2 \geq \chi^2_{k-1-l}(1-\alpha)$$

V praxi se často používá tzv. *předpokladu normality*, t.j., že náhodný výběr pochází z normálního rozdělení s určitou střední hodnotou a nějakým, blíže neurčeným rozptylem. K ověření tohoto předpokladu lze použít **testy normality** založené na výběrových koeficientech **šikmosti** A_3 a **špičatosti** A_4. Pro tyto statistiky (viz. V.3.4) platí následující vztahy:

$E(A_3) = 0,\ Var(A_3) = \frac{6(n-2)}{(n+1)(n+3)}$

$E(A_4) = -\frac{6}{n+1},\ Var(A_4) = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}$

přičemž $\sqrt{n}A_3$ a $\sqrt{n}A_4$ mají při $n \to \infty$ přibližně normálního rozdělení. Test založený na šikmosti zamítně hypotézu o normalitě, pokud

$$\frac{|A_3|}{\sqrt{Var(A_3)}} \geq u(1-\frac{\alpha}{2})$$

test založený na špičatosti zamítne hypotézu o normalitě, pokud

$$\frac{|A_4 - E(A_4)|}{\sqrt{Var(A_4)}} \geq u(1-\frac{\alpha}{2})$$

kde $u(1-\frac{\alpha}{2})$ je $(1-\frac{\alpha}{2})$-kvantil rozdělení $N(0,1)$. Hladina významnosti obou testů je asymptoticky rovna α. Při ověřování normality je vhodné provést oba testy. Hypotézu nezamítneme teprve tehdy, pokud ji nelze zamítnout oběma testy zároveň. Tyto testy jsou v některých případech citlivější na porušení normality než χ^2 test, podrobnější informace nalezne čtenář v [An].

Příklad 2.3.4 *Pomocí testů šikmosti a špičatosti testujte normalitu dat z příkladu V.2.3.*

Řešení: Pro výběrové koeficienty šikmosti a špičatosti dostáváme hodnoty $A_3 = -0,005364, E(A_3) = 0, D(A_3) = 0,129325, A_4 = -0,393762, E(A_4) = -0,146341, D(A_4) = 0,414962$. Testová statistika pro test založený na šikmosti nám dává hodnotu 0,015 a statistika pro test založený na špičatosti je rovna $0,384$. Porovnáním s $0,975$-kvantilem standardn9ho normálního rozdělení $u(0,975) = 1,96$ tedy nelze zamítnout hypotézu o normalitě na hladině významnosti $\alpha = 0,05$ ani jedním z testů.

2.3.4 Některé neparametrické testy

Znaménkový test je test o hodnotě mediánu (viz V.3.7). Předpokládáme, že výběr $X_1, \ldots, X_n$ je z rozdělení se spojitou distribuční funkcí $F(x)$. Budeme testovat hypotézu $H : \widetilde{x_{50}} = x_0$ proti alternativě $\widetilde{x_{50}} \neq x_0$ na hladině významnosti α. Vytvoříme posloupnost rozdílů $X_1 - x_0, \ldots, X_n - x_0$.

Označme Z počet členů této posloupnosti s kladným znaménkem a m počet nenulových rozdílů. Z je náhodná veličina s binomickým rozdělením (viz II.3.2) s parametry m a $\frac{1}{2}$. Pro malá m lze tedy stanovit kritický obor pro dané α určením celého čísla c tak, aby byly splněny nerovnosti

$$\sum_{i=1}^{c} \binom{m}{i} \left(\frac{1}{2}\right)^m \leq \frac{\alpha}{2} < \sum_{i=1}^{c+1} \binom{m}{i} \left(\frac{1}{2}\right)^m$$

Hypotézu H potom zamítneme, pokud $Z < c$ nebo $m - c < Z$. Pro větší m lze využít aproximace binomického rozdělení rozdělením normálním (viz věta v IV.2.1) a kritický obor vyjádřit pomocí $(1 - \frac{\alpha}{2})$-kvantilu rozdělení $N(0,1)$ nerovností $\frac{|2Z-m|}{\sqrt{m}} \geq u(1-\frac{\alpha}{2})$. Pro jednostranné alternativy vytvoříme kritický obor analogicky jako v VII.2.1.

Jednovýběrový Wilcoxonův test. Předpokládáme, že $\{X_1, \ldots, X_n\}$ je výběr z rozdělení se spojitou distribuční funkcí $F(x)$, která je symetrická kolem mediánu $\widetilde{x_{50}}$ (neboli $F(\widetilde{x_{50}} - x) = 1 - F(\widetilde{x_{50}} + x)$. Budeme opět testovat hypotézu $H : (\widetilde{x_{50}} = x_0)$ proti alternativě $(\widetilde{x_{50}} \neq x_0)$ na hladině významnosti α. Podobně jako v VII.4.l i v tomto případě vytvoříme posloupnost rozdílů $X_1 - x_0, \ldots, X_n - x_0$ a dále budeme počítat pouze s nenulovými rozdíly, jejichž počet označíme m. Tuto posloupnost uspořádáme vzestupně podle absolutních hodnot a označíme R_i^+ pořadí náhodné veličiny $|X_i - x_0|$. Sečteme-li pořadí R_i^+ pro všechny členy, pro které je $X_i - x_0 > 0$ a tento součet označíme S^+, dostaneme statistiku, pro kterou za platnosti hypotézy platí

$$E(S^+) = \frac{m(m+1)}{4},\ Var(S^+) = \frac{m(m+1)(2m+1)}{24}$$

a pro velká m je její rozdělení přibližně normální. Proto budeme pracovat raději s normovanou veličinou $V = \frac{S^+ - E(S^+)}{\sqrt{Var(S^+)}}$. Hypotézu tedy zamítneme, pokud $|V| \geq v(1 - \frac{\alpha}{2})$. Pro malé hodnoty m jsou kritické hodnoty $v(1 - \frac{\alpha}{2})$ tabelovány, pro velká m lze použít kvantily rozdělení $N(0,1)$.

Dvouvýběrový Wilcoxonův test slouží k testování hypotézy o shodě distribučních funkcí dvou výběrů. Nechť $\{X_1, \ldots, X_n\}$ a $\{Y_1, \ldots, Y_m\}$ jsou dva nezávislé výběry ze dvou spojitých rozdělení. Za platnosti hypotézy jsou tato rozdělení totožná a spojený výběr $\{X_1, \ldots, X_n, Y_1, \ldots, Y_m\}$ lze považovat za výběr z jednoho rozdělení. Označme R_i^X pořadí veličin X_i ve spojeném výběru, uspořádaném podle velikosti a nechť $R^X = \sum_{i=1}^n R_i^X$. Potom je

$$E(R^X) = \frac{m(m+n+1)}{2}, Var(R^X) = \frac{mn(m+n+1)}{12}$$

Test lze založit přímo na statistice R^X a kritický obor je potom určen nerovností $R^X \geq w_{m,n}(1-\frac{\alpha}{2})$, kde kritické hodnoty $w_{m,n}(1-\frac{\alpha}{2})$ jsou tabelovány (viz např. [An], [Sk], [Zv]). Pro přibližný test použijeme normovanou veličinu $W = \frac{R^X - E(R^X)}{\sqrt{D(R^X)}}$, která má za platnosti hypotézy pro velké rozsahy m a n přibližně rozdělení $N(0,1)$.

Oba uvedené testy - znaménkový i Wilcoxonův - jsou častou používány jako testy **párové** namísto párového t-testu (viz VII.2.4).

2.4 Regresní analýza

(Upozornění: tento text ještě nebyl upraven do konečné podoby. Proto zde chybějí některé obrázky a objevují se zde odkazy na odstavce, které nejsou označeny odpovídajícím způsobem. Tyto nedostatky budou postupně odstraňovány. Do té doby prosím o trpělivost a omlouvám se za ztížené čtení. GDo)

2.4.1 Regresní závislost

V matematice vyjadřujeme závislost hodnot jedné proměnné na hodnotách druhé proměnné funkčním vztahem. V praktických úlohách je však situace složitější. Při měření hodnot sledované veličiny, při jejíž realizaci působí řada dalších (náhodných) vlivů, dostáváme soubor naměřených hodnot, které vykazují často jisté odchylky proti hodnotám, které bychom očekávali z teoretického rozboru sledovaného jevu nebo z jakési očekávané pravidelnosti.

Příklad 2.4.1 *Při soustružení vzniká v místě obrábění na nástroji teplota, závislá na rychlosti posuvu nástroje. Mezi teplotou θ měřenou ve stupních Celsia a rychlostí posuvu v v metrech za minutu byl odvozen teoretický vztah $\theta = \alpha v^{\beta}$, kde α a β jsou konstanty, závisející na dalších podmínkách experimentu. Hodnoty, které byly naměřeny při laboratorním měření, však tomuto vztahu odpovídají jen velmi přibližně, jak lze vidět z grafu.*

Předpokládejme, že sledovanou náhodnou veličinu Y lze vyjádřit jako funkci (zpravidla nenáhodných) veličin $X_1, \ldots, X_r$ a náhodné odchylky ϵ jako

$$Y = f(X_1, \ldots, X_r; \theta_1, \ldots, \theta_s) + \epsilon.$$

Funkce f se nazývá **regresní funkce** a $\theta_1, \ldots, \theta_s$ nazýváme **parametry regrese**. O náhodné veličině ϵ, která se často nazývá neprávem „chybou“, předpokládáme, že má symetrické rozdělení se střední hodnotou 0 a rozptylem σ^2. Obvyklý je předpoklad normálního rozdělení $N(0, \sigma^2)$. Uvedený vztah se nazývá **regresní model**. Podle druhu závislosti regresní funkce na neznámých parametrech $\theta_1, \ldots, \theta_s$ potom hovoříme buď o *lineárním regresním modelu* nebo o *nelineárním regresním modelu*. Nadále se budeme zabývat pouze lineárním modelelm.

Střední hodnota $E(Y)$ je potom funkcí hodnot veličin $X_1, \ldots, X_r$ a neznámých parametrů $\theta_1, \ldots, \theta_s$. Tuto vlastnost vyjádříme vztahem

$$E(Y) = f(x_1, \ldots, x_r; \theta_1, \ldots, \theta_s),$$

kde $x_1, \ldots, x_r$ jsou naměřené hodnoty veličin $X_1, \ldots, X_r$ a $\theta_1, \ldots, \theta_s$ jsou parametry.

Náhodné veličině Y se říká **vysvětlovaná proměnná**, veličinám $X_1, \ldots, X_r$ budeme říkat **vysvětlující proměnné**. Podle tvaru regresní funkce budeme mluvit o **přímkové, exponenciální, kvadratické, polynomické** a jiných regresích. V případě přímkové regrese rozlišujeme podle počtu vysvětlujících proměnných tzv. **jednoduchou regresi** s jednou vysvětlující proměnnou a **vícenásobnou regresi** s více vysvětlujícími proměnnými.

V zásadě zde máme dva problémy: určit tvar (typ) regresní funkce a Při vyšetřování regresní závislosti je regresní funkce zpravidla známa (vyplývá z teoretických vztahů) nebo se její tvar odhaduje (opticky, například podle X-Y grafu rozptýlenosti). Proto se v dalším textu omezíme na úlohu odhadu regresních parametrů předpokládané regresní funkce. K tomu nejčastěji používáme tzv. **metodu nejmenších čtverců**. Tato metoda spočívá v provedení n nezávislých měření hodnot veličin $Y, X_1, \ldots, X_r$ a v nalezení hodnot $\hat{\theta_1}, \ldots, \hat{\theta_s}$, při nichž funkce

$$S(\theta_1, \ldots, \theta_s) = \sum_{i=1}^{n} \left[y_i - f(x_{1_i}, \ldots, x_{r_i}; \theta_1, \ldots, \theta_s) \right]^2$$

nabývá svého minima. Vektory $y_i, x_{1_i}, \ldots, x_{r_i}$ označují i-té pozorování vektoru $Y, X_1, \ldots, X_r, i = 1, \ldots, n$.

Nejsme-li si jisti a rozhodujeme-li se mezi několika modely, potom zpravidla volíme ten, v němž je hodnota funkce $S(\theta_1, \ldots, \theta_s)$ – takzvaný **reziduální součet čtverců** – nejmenší.

V případě lineární regresní funkce $f(x, \alpha, \beta) = \alpha + \beta x$ budeme minimalizovat funkci $S(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$. Nutnou podmínkou pro extrém funkce dvou proměnných je nulovost obou parciálních derivací

$$\begin{aligned} \frac{\partial S}{\partial \alpha} &= -2 \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) = 0 \\ \frac{\partial S}{\partial \beta} &= -2 \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) x_i = 0, \end{aligned}$$

což vede k takzvané **soustavě normálních rovnic**

$$\begin{aligned} n\alpha + \beta \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i \\ \alpha \sum_{i=1}^{n} x_i + \beta \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} y_i x_i \end{aligned}$$

jejímž řešením dostaneme bodové odhady a a b parametrů α a β

$$\begin{aligned} b &= \frac{\sum_{i=1}^{n}(x_i-\overline{x})y_i}{\sum_{i=1}^{n}(x_i-\overline{x})^2} \\ a &= \overline{y} - \frac{\sum_{i=1}^{n}(x_i-\overline{x})y_i}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\overline{x} = \overline{y} - b\overline{x} \end{aligned}$$

kde $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Podmínku postačující není třeba vyšetřovat, neboť funkce $S(\alpha, \beta)$ je ryze konvexní.

2.4.2 Jednoduchá přímková regrese

Velmi častým případem regresní závislosti je **přímková regrese**. Předpokládejme regresní vztah $Y = \alpha + \beta X + \epsilon$, kde X je náhodná veličina a ϵ je náhodná veličina s normálním rozdělením $N(0, \sigma^2)$.

Bodové odhady a a b parametrů α a β získáme metodou nejmenších čtverců ve tvaru uvedeném v příkladě VIII.1.6. Naměřené hodnoty $y_1, \ldots, y_n$ lze považovat za hodnoty realizací nezávislých náhodných veličin $Y_1, \ldots, Y_n$ s normálním rozdělením $N(a + bx_i, \sigma^2)$. Z tohoto hlediska jsou bodové odhady a a b odhadovými statistikami, a tedy náhodnými veličinami. Hodnoty $e_i = y_i - a - bx_i, i = 1, \ldots, n$ se nazývají **rezidua** a lze je považovat za odhady hodnot chybového členu ϵ. Číslo $\hat{y}_i = a + bx_i$ je odhadem hodnoty náhodné veličiny Y_i.

Označme $S_R = S(a, b)$ takzvaný **reziduální součet čtverců**

$$S_R = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} y_i - a - bx_i{}^2 = \sum_{i=1}^{n} y_i^2 - a\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i y_i.$$

Bodový odhad s^2 rozptylu σ^2 chybového členu ϵ je potom dán vztahem $s^2 = \frac{S_R}{(n-2)}$ a nazývá se **reziduální rozptyl**.

Pomocí s^2 lze vyjádřit odhady rozptylu obou regresních parametrů

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

Statistiky $T_\alpha = \frac{(a-\alpha)}{S_a}$ a $T_\beta = \frac{(b-\beta)}{S_b}$ mají Studentovo t-rozdělení o $(n-2)$ stupních volnosti. Intervalové odhady pro parametry α a β jsou potom dány nerovnostmi

$$a - S_a t_{n-2}(1-\frac{\gamma}{2}) \leq \alpha \leq a + S_a t_{n-2}(1-\frac{\gamma}{2})$$

$$b - S_b t_{n-2}(1 - \frac{\gamma}{2}) \leq \beta \leq b + S_b t_{n-2}(1 - \frac{\gamma}{2})$$

kde $(1-\gamma)$ je koeficient spolehlivosti a $t_{n-2}(1-\frac{\gamma}{2})$ je $(1-\frac{\gamma}{2})$-kvantil t-rozdělení o $(n-2)$ stupních volnosti.

V některých případech nás zajímá, zda hodnota některého z parametrů se liší významně od nulové hodnoty nebo ne a zda jej lze tudíž v regresní funkci vynechat. Oboustranné testy nulovosti regresních koeficientů lze založit na odhadových statistikách $T_a = \frac{a}{S_a}$ resp. $T_b = \frac{b}{S_b}$ a jim odpovídajícím kritickým oborům tak, že při splnění nerovnosti

$$|T_a| \geq t_{n-2}(1 - \frac{\gamma}{2}), \text{ resp. } |T_b| \geq t_{n-2}(1 - \frac{\gamma}{2}),$$

zamítneme hypotézu o nulovosti parametru α, resp. β, na hladině významnosti γ.

Regresním modelem se snažíme vysvětlit změny - variabilitu - vysvětlované veličiny Y pomocí změn vysvětlující veličiny X. Podíl části variability Y vysvětlené modelem ku celkové variabilitě Y, zpravidla vyjádřený v procentech, se nazývá **koeficient determinace** R^2 a je dán vztahy

$$R^2 = \frac{\sum_{i=1}^{n}(a + bx_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = \left(1 - \frac{S_R}{\sum_{i=1}^{n}(y_i - \overline{y}^2}\right),$$

kde S_R je reziduální součet čtverců.

K úplné regresní analýze patří i **analýza reziduí**. Především by měly vyhovovat předpokladu normality, za kterého byly všechny předchozí výsledky odvozeny. Pokud tomu tak není, nelze výsledky považovat za důvěryhodné. K ověření shody hodnot reziduí s normálním rozdělením lze použít některý z testů, uvedených v odstavci VII.3, nebo pravděpodobnostní papír, který je popsán v kapitole X. Z analýzy reziduí lze detekovat i takzvaná **odlehlá pozorování**. To znamená ty hodnoty, které byly chybně naměřeny nebo indikují nesrovnalosti v modelu, a jimž je třeba věnovat zvláštní pozornost. Ke zjišťování těchto hodnot lze použít například krabicové gragy, popsané v kapitole X.

Model lineární regrese lze použít i v některých případech, kdy závislost mezi veličinami X a Y není lineární. Jsou to případy, kdy lze provést takzvanou **linearizaci modelu**. Vhodnou transformací převedeme nelineární závislost na lineární a použijeme lineární regresní model. Přitom však musíme být velmi opatrní, neboť vše, co bylo odvozeno pro lineární regresní model za předpokladu normality chybového členu ϵ platí pouze pro "linearizovaný model", nikoli pro model původní, a to opět za předpokladu, že náhodná veličina, odpovídající transformovanému chybovému členu v linearizovaném modelu, má normální rozdělení.

Příklad 2.4.2 *Závislost mezi teplotou θ a rychlostí posuvu v v příkladu 2.4.1. lze považovat za regresní závislost ve tvaru $\theta = \alpha.v^{\beta}.\epsilon$, kde α a β jsou regresní koeficienty a ϵ je náhodná veličina se střední hodnotou 1. Provedeme-li transformaci $Y = ln\theta, X = lnv, a = ln\alpha, e = ln\epsilon ab = \beta$, dostaneme $Y = a + bX + e$, tedy lineární vztah.*

Podobně jako v předchozím příkladu lze linearizovat i jiné modely, např. logaritmický, tj. $Y = ln(\alpha + \beta.X)$, reciproký $Y = \frac{1}{\alpha+\beta.X}$ a další.

Příklad 2.4.3 *U 155 automobilů byla sledována spotřeba pohonných hmot v litrech na 100 km a zaznamenáván zdvihový objem válců v cm^3. Analyzujte závislost mezi těmito veličinami.*

Řešení: Analýza přímkové regresní závislosti mezi uvedenými veličinami byla zpracována statistickým programovým systémem STATGRAPHICS. Výsledky jsou uvedeny v následující tabulce:

Absolutní člen = a, směrnice = b, směrodatné odchylky parametrů jsou S_a a S_b, statistika T představuje hodnoty T_a a T_b a p-hodnota se vztahuje k testu nulovosti regresních koeficientů a a b. Graf regresní přímky, proložené naměřenými hodnotami je zobrazen spolu s tzv. „pásy spolehlivosti“ na levém obrázku, na pravém jsou zakreslena rezidua vzhledem k nezávisle proměnné *objem válců*:

2.4.3 Pásy spolehlivosti a predikce v modelu lineární regrese

Kromě obecných úvah o závislosti Y na X se často vyskytuje potřeba odhadu hodnoty $\widetilde{Y}$ závisle proměnné Y pro předem známou hodnotu nezávisle proměnné X, takzvané **předpovědi**, neboli **predikce**. Zpravidla se používá hodnota regresní funkce v daném bodě. To je ovšem pouze bodový odhad a ten jak víme, silně závisí na působení náhodných vlivů v okamžiku měření. Navíc tím nedostaneme odpověď na otázku, zda nějakou teoreticky uvažovanou hodnotu Y lze - za předpokladu platnosti našeho modelu - očekávat při určité hodnotě X se zvolenou pravděpodobností $1 - \alpha$.

Při vyšetřování regresní závislosti často konstruujeme takzvané **pásy spolehlivosti**. V literatuře jich byla navržena celá řad. Uvažujme například sjednocení všech intervalů spolehlivosti kolem hodnoty regresní přímky v jednotlivých hodnotách vysvětlující proměnné, které v našem experimentu přicházejí v úvahu.

Bodový odhad hodnoty regresní funkce $\hat{Y}(x) = a + bx$ dostneme pomocí odhadu regresních koeficientů. Podobně jako v případě intervalových odhadů

pro regresní parametry, i zde lze odvodit statistiku pro nestranný bodový odhad rozptylu $\hat{Y}(x)$ (viz [LM]):

$$S^2_{\hat{Y}} = s^2(\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_{i=1}^n x_i^2 - n\overline{x}^2}),$$

kde s^2 je reziduální rozptyl. Pravděpodobnostní rozdělení statistiky $T_{\hat{Y}} = \frac{(\hat{Y}-a-bx)}{s_{\hat{Y}}}$ je Studentovo t-rozdělení o $(n-2)$ stupních volnosti. Odtud dostaneme hledaný interval spolehlivosti kolem hodnoty regresní přímky pro koeficient spolehlivosti $(1-\alpha)$ ve tvaru:

$$\hat{Y} - S_{\hat{Y}} t_{n-2}(1-\frac{1}{2}\alpha) \leq Y \leq \hat{Y} + S_{\hat{Y}} t_{n-2}(1-\frac{1}{2}\alpha)$$

Horní a dolní meze těchto intervalů budou tvořit dvě křivky, mezi nimiž leží tzv. **pás spolehlivosti kolem regresní přímky**. Zakreslíme-li nyní do grafu nějakou uvažovanou hodnotu X_0, Y_0, můžeme rozhodnout, zda je tato hodnota v souladu s našimi výsledky (je-li tento bod uvnitř pásu), nebo zda se významně liší (je-li mimo tento pás).

Jiný pás spolehlivosti dostaneme použitím **intervalů spolehlivosti pro predikci**. Pro predikci $\widetilde{Y}$ hodnoty Y při dané hodnotě x je ve [Zr] odvozen výraz pro intervalový odhad

$$\hat{Y} - S_{\hat{Y}} t_{n-2}(1-\frac{1}{2}\alpha) \leq Y \leq \hat{Y} + S_{\hat{Y}} t_{n-2}(1-\frac{1}{2}\alpha)$$

kde

$$S^2_{\hat{Y}} = s^2(1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_{i=1}^n x_i^2 - n\overline{x}^2}),$$

Výsledkem sjednocení horních a dolních mezí intervalových odhadů pro různá Y je **pás spolehlivosti pro predikci**. Tento pás je širší, než pás kolem regresní přímky. Na obrázku v příkladu **??** jsou zobrazeny oba uvedené pásy dvojicí přerušovaných čar tak, jak je zobrazuj statistický program STATGRAPHICS. Užší pás (dvojice čar blíže k regresní přímce) je pás spolehlivosti kolem regresní přímky a širší (vzdálenější dvojice přerušovaných čar) je pás spolehlivosti pro predikci.

Často nás zajímá **toleranční pás**, který se týká rovněž predikce. Pokud budeme sledovaný experiment provádět dále za přibližně stejných podmínek, dostaneme řadu takzvaných „budoucích“ pozorování (měření). Toleranční pás pro $100\delta\%$ pozorování na hladině $1-\alpha$ je množina, ve které s pravděpodobností $1-\alpha$ bude ležet alespoň $100\delta\%$ budoucích pozorování. Kostrukce tolerančního pásu je poměrně komplikovaná a k jeho výpočtu je zpravidla třeba použít specializovaný software.

2.5 Závislost a korelace

(Upozornění: tento text ještě nebyl upraven do konečné podoby. Proto zde chybějí některé obrázky a objevují se zde odkazy na odstavce, které nejsou označeny odpovídajícím způsobem. Tyto nedostatky budou postupně odstraňovány. Do té doby prosím o trpělivost a omlouvám se za ztížené čtení. GDo)

Nejčastějším předpokladem při pravděpodobnostních úvahách nebo při statistickém vyšetřování je předpoklad nezávislosti náhodných veličin. Doposud jsme však neuvedli žádný statistický nástroj, jak tuto závislost či nezávislost zjišťovat nebo ověřovat. V této kapitole uvedeme dvě statistické metody, pomocí nichž lze dělat jisté závěry o (lineární) závislosti dvou náhodných veličin na základě jejich pozorování.

2.5.1 Korelační koeficient

Jak jsme uvedli v III.2.3, mírou závislosti dvou náhodných veličin X, Y je jejich korelační koeficient ρ_{XY}. Je-li tento koeficient různý od nuly, říkáme, že veličiny X a Y jsou stochasticky závislé. Nezávislou mohou být pouze při nulovém ρ_{XY}, ale také nemusí, jak ukazuje příklad III.3.5. Korelačnímu koeficientu, který byl definován v III.2.3, odpovídá ve statistickém vyšetřování **výběrový korelační koeficient** r_{XY}, který lze spočítat na základě měření $(x_1, y_1), dots, (x_n, y_n)$ náhodných veličin X, Y podle vztahu

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{(}x))(y_i - \bar{(}y))}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\sum_{i=1}^{n}(x_i y_i - n\bar{x}\bar{y})}{\sqrt{(\sum i = 1^n x_i^2 - n\bar{x}^2)(\sum i = 1^n y_i^2 - n\bar{y}^2)}}$$

Má-li náhodný vektor (X, Y) sdružené normální rozdělení, potom lze snadno dokázat, že nekorelovanost je totéž co nezávilost. Za předpokladu normality lze tedy na základě nulovosti korelačního koeficientu usuzovat na nezávislost. Zpravidla však máme k dispozici pouze statistiku r_{XY}, která je náhodnou veličinou a jejíž hodnota je dána momentálně naměřenými hodnotami výběru. Je-li nenulová, nemusí to ještě znamenat, že i $\rho_{XY} = 0$ a tedy nekorelovanost veličin X a Y. Nulovost korelačního koeficientu se ověřuje pomocí statistického testu. Ten je založen na tom, že za předpokladu, že náhodný výběr $(X_1, Y_1), \ldots, (X_n, Y_n)$ je ze sdruženého normálního rozdělení s $\rho_{XY} = 0$, potom má veličina

$$T = \frac{r_{XY}\sqrt{n-2}}{\sqrt{1 - r_{XY}^2}}$$

Studentovo t-rozdělení o $(n-2)$ stupních volnosti.

Test nulovosti korelačního koeficientu na hladině významnosti α potom zamítá hypotézu $H : \rho_{XY} = 0$ proti alternativě $A : \rho_{XY} \neq 0$, pokud $|T| \geq t_{n-2}(1-\frac{\alpha}{2})$, kde $t_{n-2}(1-\frac{\alpha}{2})$ je $(1-\frac{\alpha}{2})$-kvantil t-rozdělení o $n-2$ stupních volnosti.

Nejsou-li splněny předpoklady pro použití výše uvedeného testu (normální sdružené rozdělení (X,Y), můžeme pro test nezávislosti X a Y použít tzv. **Spearmanův korelační koeficient**. V tomto případě se pouze předpokládá, že náhodný výběr $(X_1,Y_1),\ldots,(X_n,Y_n)$ je ze spojitého dvojrozměrného rozdělení.

Uspořádejmě naměřené hodnoty $x_1,\ldots,x_n$ podle velikosti a jejich pořadí označme $R_1,\ldots,R_n$. Obdobně budeme postupovat s $y_1,\ldots,y_n$, jejichž pořadí označíme $Q_1,\ldots,Q_n$. Nyní spočteme výběrový korelační koeficient pro dvojice $(R_i,Q_i), i=1,\ldots,n$ a označíme jej r^S_{XY}:

$$r^S_{XY} = \frac{\sum_{i=1}^n R_iQ_i - n\bar{R}\bar{Q}}{\sqrt{[\sum_{i=1}^n R_i^2 - n\bar{R}^2][\sum_{i=1}^n Q_i^2 - n\bar{Q}^2]}}$$

kde

$\bar{R} = \frac{1}{n}\sum_{i=1}^n R_i = \bar{Q} = \frac{1}{n}\sum_{i=1}^n Q_i = \frac{n+1}{2}$

a

$\sum_{i=1}^n R_i^2 = \sum_{i=1}^n Q_i^2 = \frac{n(n+1)(2n+1)}{6}$.

Dále je

$$\sum_{i=1}^n R_iQ_i = -\frac{1}{2}\sum_{i=1}^n (R_i-Q_i)^2 + \sum_{i=1}^n R_i^2 + \sum_{i=1}^n Q_i^2 = \\ -\frac{1}{2}\sum_{i=1}^n (R_i-Q_i)^2 + \frac{n(n+1)(2n+1)}{6}. \quad (2.2)$$

Dosazením do vzorce pro r^2_{XY} dostáváme tzv. **Spearmanův korelační koeficient**

$$r^S_{XY} = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^n (R_i-Q_i)^2.$$

Test nezávislosti X a Y na hladině významnosti α je potom založen na nerovnosti

$$|r^S_{XY}| \geq k(\alpha),$$

kde $k(\alpha)$ je kritická hodnota, kterou lze pro malé rozsahy výběru ($n \leq 30$) nalézt v tabulkách (viz např. [An]) nebo při větších rozsazích, tj. pro $n > 30$, ji lze aproximovat hodnotou $k^*(\alpha) \cong \frac{u(1-\frac{\alpha}{2})}{\sqrt{n-1}}$, kde $u(1-\frac{\alpha}{2})$ je $(1-\frac{\alpha}{2})$-kvantil rozdělení $N(0,1)$.

2.5.2 Test nezávislosti v kontingenční tabulce

V odstavci III.3 byla charakterizována nezávislost veličin X a Y pomocí srovnání sdruženého rozdělení s marginálními. Při statistickém šetření je třeba postupovat sice opatrněji, nicméně lze tak učinit podobným způsobem. Pouze teoretická rozložení nahradíme jejich empirickými odhady, tj. poměrnými četnostmi.

Mějme k dispozici nezávislá pozorování $(x_1, y_1), \ldots, (x_n, y_n)$ náhodného vektoru (X, Y). Předpokládejme, že veličina X může nabývat pouze hodnoty z množiny $A_1, \ldots, A_k$ a veličina Y nabývá hodnoty z množiny $B_1, \ldots, B_l$). Spočítáme četnosti n_{rs}, reprezentují počet naměřených dvojic x_i, y_j takových, že $x_i = A_r, , y_j = B_s, r = 1, \ldots, k, s = 1, \ldots, l$. Označme $n_{.j} = \sum_{i=1}^{k} n_{ij} a n_{i.} = \sum_{j=1}^{l} n_{ij}$. Napočítané hodnoty uspořádáme to tzv. **kontingenční tabulky:**
TABULKA

Poměrné četnosti $\frac{n_{ij}}{n}$ lze považovat za odhad sdruženého rozdělení p_{ij} náhodného vektoru (X, Y), $\frac{n_{i.}}{n} a \frac{n_{.j}}{n}$ za odhad marginálních rozdělení q_i, r_j veličin X a Y. Za předpokladu nezávislosti by mělo být $p_{ij} = q_i r_j$ (viz III.3.1) a tedz budeme očekávat, že bude přibližně $\frac{n_{ij}}{n} = \frac{n_{i.} n_{.j}}{n^2}$. To odpovídá požadavku,aby součet $\sum_{i=1}^{k} \sum_{j=1}^{l} (\frac{n_{ij}}{n} - \frac{n_{i.}}{n} \frac{n_{.j}}{n})^2$ byl malý.

Test nezávislosti náhodných veličin X a Y je založen na statistice

která má za předpokladu nezávislosti veličin X a Y při velkých n přibližně rozdělení χ^2 o $m = (k-1)(l-1)$ stupních volnosti. Hypotézu nezávislosti tedy zamítneme, pokud bude $\chi^2 \leq \chi^2_m(1-\alpha)$.

Příklad 2.5.1 *U 155 automobilů byla měřena spotřeba pohonných hmot, výkon a obsah válců. Podle velikosti spotřeby na jednotku výkonu byly automobily rozděleny do tří kategorií: nízká (do 0.15), průměrná (0.15-0.2) a vysoká (více než 2.0). Podle objemu byly rozděleny do sedmi kategorií, první do 1500 ccm a dále po 500 ccm, takže sedmá kategorie byly vozy s objemem větším než 4000 ccm. V tabulce jsou uvedenz četnosti naměřených hodnot:*

spotřeba/	*objem válců*							
jedn.výkonu	*1*	*2*	*3*	*4*	*5*	*6*	*7*	$\sum$
nízká	*5*	*15*	*9*	*8*	*1*	*2*	*10*	*50*
průměrná	*22*	*28*	*16*	*3*	*5*	*11*	*8*	*93*
vysoká	*3*	*2*	*5*	*1*	*0*	*0*	*1*	*12*
$\sum$	*30*	*45*	*30*	*12*	*6*	*13*	*19*	*155*

Existuje statisticky významná závislost spotřeby na výkonu a obsahu válců?

Řešení: Úloha byla zpracována programem MYSTAT, který poskytl tyto výsledky: statistika χ^2 nabývá hodnoty 22,786, počet stupňů volnosti je 12

a p-hodnota je rovna 0,30. Tedy na hladině významnosti $\alpha > 0,03$ bychom mohli ještě hypotézu o nezávislosti zamítnout. Konkrétně pro $\alpha = 0,05$ je $(1-\alpha)$-kvantil rozdělení $\chi^2(12)$ roven (podle [Jn]$\chi^2_{12}(0,95) = 21,0$. Na této hladině významnosti hypotézu o nezávislosti zamítneme. Pokud však položíme $\alpha = 0,01$, pak odpovídající kvantil bude $\chi^2_{12}(0,99) = 26,2$ a tedy na této hladině hypotézu o nezávislosti nelze zamítnout.

Statistika χ^2 slouží pouze pro test nezávislosti a nelze ji používat jako míru závislosti veličin X a Y. Hodnota této statistiky závisí na rozsahu výběru následujícím způsobem. Kdyby byla její hodnota počítána dvakrát na základě dvou výběrů, obsahujících stejné hodnoty, přičemž ve druhém výběru by se každá dvojice (x_i, y_j) opakovala dvakrát, tj. tento výběr by měl dvojnásobný rozsah, pak by druhá vypočítaná hodnota byla dvojnásobná. Přitom míra závislosti měřená například korelačním koeficientem zůstává stejná.

2.6 Grafické metody analýzy dat

Grafické metody jsou ve statistice stále více používány především pro jejich názornost a dostupnost díky rychlé výpočetní technice. Takzvaná *vizualizace dat* je mocným nástrojem při analýze dat, regresní analýze, analýze časových řad a dalších. Grafické metody často nahrazují – především v první fázi statistického vyšetřování – metody analytické.

2.6.1 Frekvenční (četnostní) grafy

Jedním ze základních typů grafického zobrazení ve statistické analýze jsou **frekvenční grafy**. Tyto grafy nám poskytují základní představu o tvaru pravděpodobnostního rozdělení hodnot ve výběru. Frekvenční grafy zobrazují informace obsažené v frekvenční tabulce (viz V.2.3).

Sloupkový diagramje konstruován následujícím způsobem: každé třídě frekvenční tabulky odpovídá jeden obdélník, jehož výška je úměrná četnosti a šířka třídnímu rozpětí. Někdy se do grafu k jednotlivým obdélníkům zakreslují informace o průměru a směrodatné odchylce uvnitř každé třídy. V případě zobrazení relativních četností mluvíme též o **histogramu**. Často se pro srovnání dvou a více výběrů zakresluje do jednoho grafu více histogramů zároveň.

Polygon četností nebo jen **polygon**, je graf, v němž jsou úsečkami spojeny body, odpovídající svými souřadnicemi dvojicím (*střed i-té třídy, četnost i-té třídy*. Také tento graf umožňuje srovnání typu rozložení více výběrů v jednom grafu.

Kruhové diagramy nebo též **koláčové grafy**. Zatímco první dva typy se používají k zobrazení všech druhů četností, kruhové diagramy zobrazují pouze prosté četnosti, nejčastěji v relativním tvaru, vyjádřené v procentech.

Sloupkový diagram a polygon jsou často užívány při usuzování o tvaru hustoty či distribuční funkce pravděpodobnostního rozdělení. Do grafu bývá zakreslena křivka hustoty (v případě poměrných četností) nebo distribuční funkce (zobrazujeme-li poměrné kumulativní četnosti) některého teoretického rozdělení a opticky usuzujeme o její vhodnosti. Tento postup zpravidla předchází dalším odhadům a testům.

2.6.2 Grafické metody průzkumové analýzy dat

Stonek s listy (Steam and leaf diagram) je grafická podoba *frekvenčního histogramu*. Na rozdíl od něho však tento graf uchovává alespoň částečně informaci o pozorovaných hodnotách. Základem je rozdělení dat do tříd (třídních intervalů). Jejich označení se zakresluje nalevo od svislé čáry představující stonek. Napravo v jednotlivých řádcích představujících listy jsou vypsány buď jednotlivé hodnoty, nebo skupiny znaků, které je zastupují. Někdy se ještě zaznamenávají horní a dolní kvartily, medián.

Příklad. Na diagramu jsou metodou „stonek s listy“ zobrazeny naměřené hodnoty spotřeby pohonných hmot automobilů z příkladu VIII.3.3 jak jej zobrazuuje program MYSTAT:

Krabicový graf (Box and whiskers plot, neboli „krabice s vousy“. Data jsou zde zobrazena ve formě obdélníku (krabice), z něhož nahoru a dolů vybíhají úsečky (vousy). Výška obdélníka je rovna (v daném měřítku) velikosti mezikvartilového rozpětí RQ, přičemž dolní, resp. horní, strana odpovídá dolnímu resp. hornímu kvartilu. Uvnitř je obdélník předělen příčkou v místě mediánu Me. „Vousy“ spojují hodnoty, splňující nerovnost $\frac{1}{2}RQ \leq |x-Me| \leq K.RQ$, kde K je konstanta větší než 1. Obvykle $K = 1,5 nebo 1,75$. Hodnoty vně tohoto intervalu se zobrazují jako izolované body (hvězdičkou) za předpokladu normálního rozdělení je lze považovat za tzv. **odlehlá pozorování**.

Vrubové grafy (Notched box plots) jsou konstruovány stejně jako krabicové grafy, navíc však obsahují informaci o intervalu spolehlivosti pro odhad mediánu. Po obou stranách obdélníku jsou v místě mediánu zářezy, tzv. **vruby**, jejichž šířka je úměrná velikosti intervalu spolehlivosti. Označíme-li $\widetilde{x}$ odhad mediánu, $S_{\widetilde{x}} = 0,926RQ.N^{-\frac{1}{2}}$ odhad jeho směrodatné odchylky, pak interval spolehlivosti je dán vztahem $I_{\widetilde{x}} = \widetilde{x} - C.S_{\widetilde{x}}, \widetilde{x} + C.S_{\widetilde{x}}$, kde C je konstanta, která závisí na požadovaném koeficientu spolehlivosti a typu rozdělení. Pro $\alpha = 0,05$ (95% interval spolehlivosti) je doporučena hodnota přibližně $C = 1,7$.

Srovnáním několika vrubových grafů pro různé výběr v jednom měřítku vedle sebe při stejném α, můžeme opticky porovnat výběry z hlediska rovnosti střední hodnoty jejich mediánů na hladině významnosti α. Pokud se vruba zřetelně překrývají, lze soudit o shodě, nepřekrývají-li se vůbec, je to znamení významného rozdílu. Vždy však je třeba tuto hypotézu testovat pomocí statistického testu.

Pro srovnání a hledání závislostí ve vícerozměrných výběrech (kde sledujeme více znaků najednou) existuje řada grafických metod, z nichž nejjednodušší jsou **hvězdicové grafy** a **paprskové grafy**. Příkladem těchto grafů může být zobrazení souboru AUTA, v němž byly u každého automobilu měřeno pět veličin U, V, X, Y, Z. V obou případech je základem grafu hvězdice, jejíž jednotlivé paprsky odpovídají měřeným veličinám, ale jejich rozměry mají v obou případech různou interpretaci. Délka paprsků v **hvězdicovém grafu (Star Plot)** je proměnná podle velikosti naměřených hodnot. Jejich konce jsou spojeny a tvoří "hvězdici":

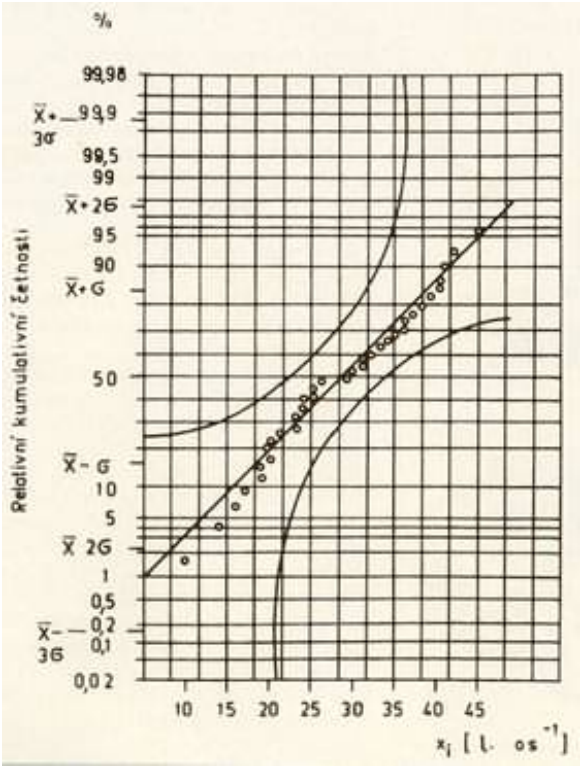
V případě **paprskového grafu (Sun-ray Plot)** délky paprsků odpovídají r/násobku směrodatné odchylky příslušné veličiny, na každém je vyznačena naměřená hodnota a tyto jsou spojeny úsečkami. Hodnoty jsou zobrazeny relativně tak, že délky paprsků reprezentují rozsah hodnot (aritmetický průměr je tedy ve středu paprsku):

2.6.3 Pravděpodobnostní papír

K ověření shody naměřených dat s teoretickým rozdělením pravděpodobnosti a k odhadu jeho parametrů lze použít tzv. **pravděpodobnostní papír**. Je to vlastně graf transformovaných hodnot empirické distribuční funkce, v němž se hodnoty odpovídající teoretické distribuční funkce zobrazují na přímku. Pro tento typ grafu se používá buď papír s předtištěnou mříží, což je obdoba „logaritmického papíru“ s logaritmickým měřítkem na jedné nebo na obou osách nebo jej generujeme přímo na obrazovce počítačového monitoru.

Mějme výběr $X_1, \ldots, X_n$ a označmě $F_n(x)$ jeho empirickou distribuční funkci (viz V.3.6). Nechť $G(x)$ a $G_{-1}(x)$ označují teoretickou ryze monotónní distribuční funkci a funkci k ní inverzní. Pro ověření shody naměřených hodnot s rozdělením s distribuční funkcí $y = G(x)$ použijeme bodový graf hodnot (x_i, y_i), kde $y_i = G^{-1}(F_n(x_i)), i = 1, \ldots, n$. Pokud by funkce F_n a G byly totožné, potom by muselo být $y_i = x_i$ a body (y_i, x_i) by ležely na přímce $y = x$. Je-li skutečné rozdělení výběru blízké teoretickému s distribuční funkcí $y = G(x)$, potom budou body (y_i, x_i) ležet v těsné blízkosti přímky $(y = x)$.

Příklad 2.6.1 *V příkladu* **??** *jsou uvedeny výsledky analýzy lineární regrese*



Obrázek 2.1: Distribuční funkce diskrétní náhodné veličiny z příkladu 1.2.1.

mezi spotřebou a objemem automobilu. Ověřte, zda rezidua v tomto modelu vyhovují normálnímu rozdělení.

Řešení: Tuto shodu lze zobrazit pomocí normálního pravděpodobnostního papíru, jak je ukázáno na obrázku vpravo (graf byl pořízen programem STATGRAPHICS). Z obrázku je patrné, že odklon hodnot od přímky je dosti velký. Proto je třeba provést test normality na zadané hladině významnosti.

Popsanou metodu lze použít i k rychlému a hrubému odhadu parametrů v případě, kdy máme představu o typu teoretického rozdělení. Předpokládejme, že máme výběr $X_1, \ldots, X_n$ z normálního rozdělení $N(\mu, \sigma^2)$ s distribuční funkcí $F(x) = \Phi\frac{x-\mu}{\sigma}$, kde Φ je distribuční funkce rozdělení $N(0,1)$. Potom $\Phi^{-1}(F(x)) = \frac{x-\mu}{\sigma}$. Zobrazíme-li nyní v grafu body o souřadnicích $X_i, \Phi^{-1}(F(X_i))$, za výše uvedeného předpokladu budou v blízkosti přímky $y = \alpha x - \beta$, kde $\alpha = \frac{1}{\sigma}, \beta = \frac{\mu}{\sigma}$. Proložíme jimi tedy přímku a z grafu odečteme její směrnici α a posunutí β, z nich pak spočteme odhady parametrů μ a σ.

2.6.4 Grafy rozptýlenosti

Grafy rozptýlenosti nebo též **rozptylové** či **korelační** grafy jsou vedle frekvenčních grafů nejčastěji používaným zobrazením naměřených dat. Používají se především v regresní analýze, k analýze trendů (závislosti) v časových řadách, ve shlukové analýze a podobně. Používají se nejčastěji jako **bodové, spojnicové** nebo sloupcové.

Indexový graf. Při zobrazení jednorozměrného výběru, například časové řady, zakreslujeme data do pravoúhlých souřadnic, kde x-ové souřadnici na vodorovné ose odpovídá číslo měření (index) i a y-ové souřadnici na svislé ose naměřená hodnota x_i. Při zobrazení časové řady má index význam času měření. Jednotlivé hodnoty se zobrazují buď jako body (bodový graf), spojují se účečkami (spojnicový graf) nebo jako sloupce (sloupcový graf). Použijeme-li namísto pravoúhlých souřadnic polární, dostaneme tzv. **radarový graf**.

Při sledování regresní závislosti mezi veličinami X a Y používáme **dvourozměrný rozptylový graf**, zobrazující naměřené hodnoty jako body (x_i, y_i) v pravoúhlých souřadnicích. Tento typ grafu je často jediným vodítkem při určování typu regresní závislosti. Příklady rozptylových grafů jsou uvedeny v příkladu VIII.2.11.