Západočeská univerzita v Plzni
Fakulta aplikovaných věd
Katedra informatiky a výpočetní techniky

Vyhodnocení výsledků stochastické metody Monte Carlo

Plzeň, 2004 David Hartman

Abstract

This essay deal with the stochastic simulations and processing of its data. The main aim is about the analysis of this data by the stochastic approach and estimates of its values or trends. The work is consisted of three parts. The first on eis deal with the simulation theory basis. The second one is about probability-statistic theory and the last one concerning in the estimates and trstiny of the hypothesis.

Obsah

Seznam obrázků

Seznam tabulek

1. Úvod

Simulace jsou dnes již klasickou součástí analýzy reálných systémů, jejich návrhu, řízení, verifikace nebo predikce jejich parametrů. Tento obor je hojně využíván v jiných oblastech lidských činností (ekonometrické metody, strojírenské verifikační metody, biologické procesní simulace, dopravní simulace, …). Důležitou součástí simulace je sběr a analyzování výstupních dat, neboť z tohoto důvodu se vlastně samotná simulace provádí. Získaná data jsou ukazateli vlastností reálných simulovaných systémů. Z těchto důvodů je důležité této problematice věnovat velkou pozornost. Tato stať se bude zabývat obecnou analýzou simulačních dat. K tomuto cíli bude potřebovat nadefinovat některé pojmy a uvést poznatky

z oblastí simulací a matematické pravděpodobnosti a statistiky. Vlastní analýza dat je vlastně přímou aplikací matematicko statistických metod analýzy, odhadu a testování, které jsou hlavními tématy tohoto textu

2. Základy teorie simulací

Jak již bylo zmíněno v úvodu je oblast simulací hojně využívanou součástí lidských činností. Co však simulací jako takovou rozumíme. Existuje velké množství definic (viz 4.6), my se však pokusíme o intuitivní nástin dané problematiky.

2.1. Simulace a model

Simulací rozumíme získávání informací o reálném světě prostřednictvím experimentu napodobujícího reálnou situaci. Postup provedení takového experimentu můžeme rozdělit na několik kroků. Nejprve musíme vytvořit prostředí blížící se reálnému systému, neboli definovat soubor entit reálného světa a vztahy mezi nimi a navrhnout jejich interpretaci v experimentu. Nic nebrání tomu, aby jednotlivé entity byly přímo entitami reálného světa, avšak vždy nastane situace, kdy budeme muset některou z reálných entit nahradit jejím substitutem. Spolu s entitami a vztahy mezi nimi musíme také definovat možné operace, které jsou nad entitami proveditelné. Tento postup má velice blízko k metodě objektové analýzy a návrhu používané v programovací praxi. Také v praxi člověka zabývajícího se simulacemi má tento styl nahlížení na svět své nezastupitelné místo. Ukázal se totiž jako velice vhodný pro tuto problematiku. Také velké množství simulačních jazyků má svůj základ v objektové interpretaci světa (např. SIMULA). Celý tento soubor entit, vztahů a operací, který jsme zde popsali, pak nazveme *modelem*. Nad tímto modelem pak můžeme provádět všechny nadefinované operace (operací musíme rozumět i nastavení určitých hodnot).

2.2. Průběh simulace a Monte Carlo

Teď se dostáváme k samotnému popisu provádění simulace. Tímto rozumíme provedení posloupnosti možných operací nad systémem. Jelikož nás často zajímají situace, jejichž důvody vzniku nelze nijak určit, musíme pro realizaci simulace používat náhodných hodnot (náhodné hodnoty můžeme použít i z jiných důvodů např. protestování větší množiny možných situací, atd.). Tato metoda se často označuje jako metoda *Monte Carlo* (viz 4.6). Jednoduše řečeno, náhodně generujeme posloupnost vstupních dat a sbíráme výstupní data po provedení modelu. Tyto data posléze analyzujeme a vyhodnocujeme z nich patřičné závěry. Tato metoda se dá s úspěchem použít v různých oblastech. Jeden ze známých problémů je *Buffonova úloha*, jejíž realizací lze experimentálně určit hodnotu Ludolfova čísla π (viz 4.6). Z předchozího textu plyne, že další důležitou součástí simulací je generování pseudonáhodných čísel. Předpona pseudo u označení náhodnosti čísel značí fakt, že generaci provádíme z velké části na deterministicky založeném číslicovém počítači a posloupnost generovaných čísel tudíž vykazuje náhodnost pouze na určité hladině přesnosti. Existují však i generátory, založené na fyzikální podstatě (například emitace částic), které se mají náhodný charakter i pro velkou přesnost. Vyžadují však speciální hardware. My se tímto problémem do hloubky zabývat nebudeme. Pro zájemce lze nalézt více o tomto problému a jednotlivých algoritmech viz 4.6 nebo 4.6 popř. 4.6.

2.3. Rozdělení simulačních modelů

Zmínili jsme se o tom, že v modelech často používáme náhodných hodnot, ovšem zde musíme říci, že existují i realizace, kdy postupujeme čistě deterministicky. Jedná se vlastně o matematicko-analytické modely, jejichž výsledky jsou většinou ve formě vztahů. Z tohoto pohledu pak můžeme rozdělit modely na:

- *Analytické modely*.
- *Simulační modely*.

U prvního *analytického modelu* je realizace založená více na vztazích, nad kterými je definovaná určitá algebra operací, umožňující pro daný soubor vstupních dat určit data výstupní. Za příkladem takového modelu může být brána newtonovská mechanika, která je jistě pouze modelem reálné situace. Některé její nedostatky řeší např. teorie relativity, kterou je ovšem opět nutné chápat jako model. Takto by se dalo postupovat stále dále, až bychom zjistili, že každá teorie je pouze modelem reality s nějakou, z podstaty věci (nedosažitelnost absolutního poznání reality) neznámou, dávkou přesnosti. Nezapomínejme zde, že i empirie a měření je pouze nástinem reality, který nám metody měření a vnímání dovolují poznat. Ovšem ve své množině definovaných axiómů, operací a vztahů může být realizace analytického modelu čistě deterministická. Determinismem se vyznačuje i průběh simulace nad *simulačním modelem*, ale zde se jedná o determinismus vnitřního průběhu (nezapomínejme, že je simulace realizována na počítači a i generace náhodných hodnot je se vyznačuje předurčeností) a nikoliv o vnější podobu běhu či výsledků. Ty je nutné chápat jako náhodné.

V obou typech modelu je důležitý další velice častý fenomén určující způsob náhledu na procesy v daném systému při jeho převodu na model. Nazvěme tento pohled *průhledností systému*. Tímto pojmem rozumíme detailnost převodu vnitřních mechanizmů zkoumaného reálného systému do simulačního modelu. Často se pro pohled na převáděný systém používá označení „černá schránka“ popř. „bílá schránka“, kde barva bílá označuje úplnou realizaci vnitřních mechanismů (opět nutné chápat omezenou přesnost popisu mechanismů) a barva černá pouze přístup funkční, neboli pohled při kterém je systém pouze funkcí poskytující výstupní data pro danou množinu dat vstupních. Mezi těmito dvěma stupni lze samozřejmě definovat libovolné množství „šedých odstínů“.

2.4. Simulační čas

Velké množství modelů realizujících nějaký reálný problém je závislých na čase. Čas pro ně tvoří jakousi proměnnou, na které je závislé velké množství stejných atributů jednotlivých entit. Někdy se systémy podle závislosti na čase rozdělují na:

- Statické … Nezávislé na čase.
- Dynamické … Závislé na čase.

Často je čas i řídící veličinou celého modelu, tj. určuje jednotlivé kroky simulace. Z toho přímo vyplývá, že existují dva přístupy k chápání času:

- Spojitý čas.
- Diskrétní čas.

Spojitý čas se nejčastěji používá u regulačních metod, kdy modelem je soustava definovaných diferenciálních rovnic. Tyto se dnes samozřejmě řeší na digitálním počítači a je tudíž u nich nutná nějaká metoda diskretizace. To ovšem nesouvisí s časem definovaným pro dané rovnice, neboť to je čas simulační, který má význam pouze pro model a ne pro výpočet některých jeho částí. Z tohoto důvodu musíme zadefinovat:

- *Simulační čas*.
- *Reálný čas*.

Reálný čas je čas výpočtu procesoru (nebo jiné realizační jednotky, která provádí simulaci), kdežto čas simulační je časem, který je právě platný pro model i přes více kroků času reálného.

Pro popis času v diskrétním případě se dobře hodí nahlížení na simulaci z pohledu *teorie systémů* (viz 4.6), kde model i realita jsou systémy (s jednotlivými prvky a vztahy). Stavem systému pak nazveme specifické nastavení daných prvků. Pak převod mezi nimi je chápán jako zobrazení z prostoru definovaného všemi prvky, relacemi a stavy reálného systému do prostoru definovaného prvky, relacemi a stavy modelu. Přechody mezi těmito stavy jsou pak realizovány v diskrétních okamžicích, které lze považovat za čas.

3. Pravděpodobnostní a statistické metody

Data se kterými operují simulace a které jsou jimi obráběny nebo poskytovány jako výstupy mají velice často stochastický charakter. Přesto je často nutné znát o dané náhodnosti určité upřesňující informace nebo naopak tyto informace z dané množiny náhodných dat umět získat. Prvním problémem se zabývá z větší části matematická teorie pravděpodobnosti a používá se nejvíce v již zmiňovaném *generování pseudonáhodných čísel* (viz str. 4).Druhým problémem, který bude naším hlavním tématem, se zase zabývá matematická statistika a je určen pro analyzování výstupních dat simulace. V této části probereme základní pojmy a vztahy z teorie matematické pravděpodobnosti a statistiky bez důrazu na jejich exaktní popis a důkaz z matematického pohledu. Jde nám o přehled nejdůležitějších věcí z této oblasti.

3.1. Pravděpodobnost

Základ teorie pravděpodobnosti tvoří experiment nebo také *náhodný pokus* (např. hod kostkou). *Realizací náhodného pokusu* pak nazveme každý výsledek jeho provedení. Výsledkem náhodného pokusu je *elementární náhodný jev* ω. Množina všech možných výsledků náhodného procesu se označuje Ω. Při realizace experimentů nás zajímají hlavně podmnožiny množiny Ω nazývané *náhodné jev*. Na množině Ω se dají provádět klasické množinové operace a jsou zde i podobná označení některých podmnožin:

- *Nemožný jev* $\emptyset$ … Nemůže nastat při žádné realizaci náhodného pokusu.
- *Jistý jev* Ω … Jev který nastane při každé realizaci náhodného procesu.
- *Rozdíl jevů* $A \setminus B$ … Jev, který nastane pokud nastane jev A a nikoliv jev B.
- *Sjednocení jevů* $A \cup B$ … Jev, který nastane pokud nastane alespoň jeden z jevů.
  $A \cup \Omega = A$
- *Průnik jevů* $A \cap B$ … Jev, který nastane pokud nastanou oba jevy současně.
  $A \cap \Omega = \Omega$
- *Jev opačný* $\overline{A}$ … Jev, který nastane v případě, že nenastane jev A.
  $\overline{\emptyset} = \Omega, \quad \overline{\Omega} = \emptyset, \quad \overline{\overline{A}} = A, \quad A \cup \overline{A} = \Omega, \quad A \cap \overline{A} = \emptyset$
- *De Morganovy vzorce* $\overline{A \cup B} = \overline{A} \cap \overline{B}$
  $\overline{A \cap B} = \overline{A} \cup \overline{B}$
- *Distributivní zákony* $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$
  $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$

Cílem teorie pravděpodobnosti je zavést funkci, která by nějak kvantifikovala podmnožiny množiny Ω a měla pro všechny stejné vlastnosti. Ukázalo se nejvýhodnějším definovat takovou funkci, která daným množinám přiřadí reálná čísla z intervalu <0, 1>, kde vzrůstající hodnota bude indikovat větší možnost uskutečnění jevu. Nula by pak byla přiřazena jevu nemožnému a jednička jevu jistému.

3.1.1. Statistická definice pravděpodobnosti

Pro definici pravděpodobnosti musíme nejprve definovat pojem *četnosti*. Představme si, že realizujeme nějaký experiment (náhodný pokus) a sledujeme kolikrát se vyskytne námi sledovaná náhodná událost, neboli kolikrát vyjde sledovaný výsledek (padne šestka na kostce). Tato hodnota se pak nazývá četností. Její význam je ovšem značně ovlivněn rozměrem souboru náhodných pokusů a proto zavádíme ještě pojem *relativní četnosti*, která již není závislá na této hodnotě, neboť je definována jako:

$$h(A) = \frac{N_A}{N},$$

kde N je počet realizací náhodného procesu a N_A ty realizace při nichž nastal sledovaný jev A. Intuitivně vidíme, že se vzrůstajícím počtem realizací a za podmínek kladených na pravděpodobnostní funkci se pravděpodobnost rovná:

$$P(A) = \lim_{n \to \infty} h(A) = \lim_{n \to \infty} \frac{n_A}{n}.$$

Takto vidíme první poznatek výhodný i pro oblast simulačních modelů. Z této definice totiž vyplývá, že při vzrůstajícím počtu pokusů se relativní četnost blíží pravděpodobnosti. Tuto skutečnost budeme probírat podrobněji dále.

3.1.2. Klasická definice pravděpodobnosti

Jelikož se zabýváme simulačními modely byla statistická definice uvedena na prvním místě. Z pohledu matematického výkladu by možná někdo chtěl spíše na prvním místě uvést definici klasickou, neboť statistická platí v obecnějším případě. Klasická definice totiž předpokládá, že je množina možných výsledků konečná a vyjádřitelná. Pak můžeme definovat pravděpodobnost s požadovanými vlastnostmi jako

$$P(A) = \frac{N_A}{N},$$

kde tentokráte N značí množinu všech možných výsledků náhodného pokusu a N_A její podmnožinu, ve které nastal jev A. Například pro vrh kostkou je počet všech možných výsledků roven 6. Jev A pak může být jevem: padlo sudé číslo, čemuž odpovídá hodnota 3. Pravděpodobnost takového jevu tedy bude 3/6 = 1/3.

3.1.3. Axiomatická definice pravděpodobnosti

Posledním nejobecnějším typem pravděpodobnosti je axiomatická definice, která definuje pravděpodobnost na axiomatickém základě. Uvedeme zde jen zkrácenou definici. Se všemi detaily lze definici nalézt například v 4.6.

Nejprve definujeme *elementární jev E*. Je to takový jev, pro který platí, že

$$\emptyset \neq A \subset E \Rightarrow A = E,$$

neboli takový jev, který již nelze rozložit na jiné podjevy. Pro elementární jevy platí, že jsou *párově neslučitelné*, tj. odlišné jevy mají prázdný průnik.

Dále musíme definovat *σ-algebru jevů* jako systém Σ podmnožin množiny Ω, pro který platí

1. $\Omega \in \Sigma$
2. je-li $A \in \Sigma$, pak i $\overline{A} = \Omega \setminus A \in \Sigma$
3. jsou-li $A_i \in \Sigma \quad i = 1,2,\ldots$ pak i $\bigcup_{i=1}^{\infty} A_i \in \Sigma$

Nyní můžeme nadefinovat pravděpodobnost jako funkci $P : \Sigma \to R$ s vlastnostmi:

1. $0 \le P(A) \le 1$ pro $\forall A \in \Sigma$
2. $P(\Omega) = 1$

3. Pro každou posloupnost {Ai} takovou, že $A_i \in \Sigma, i = 1,2,...$ a $A_i \cap A_j = \emptyset$ pro $i \neq j$, platí

$$P(A_1 \cup A_2 \cup ...) = P(A_1) + P(A_2) + ...$$

Potom trojici (Ω, Σ, P) nazýváme *pravděpodobnostní prostor*.

3.1.4. Podmíněná pravděpodobnost

Dalším důležitým pojmem v teorii pravděpodobnosti je *podmíněná pravděpodobnost*. Jedná se vlastně o pravděpodobnost nějakého jevu podmíněnou platností jevu jiného. Jeho výpočet lze vyjádřit jako:

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

Pochopení tohoto vztahu je jednoduché. Uvědomíme-li si, že předpoklad je platnost jevu *B*, pak musíme platné jevy vybírat jen z množiny, kdy platí jev *B*. Z této množiny pak vybereme jen ty jevy, kdy zároveň nastal jev *A*, tj. $A \cap B$.

Na základě této veličiny můžeme definovat pojem *nezávislých jevů A a B*, pro které platí

$$P(A|B) = P(A),$$

z čehož vyplývá, že jevy jsou nezávislé, právě tehdy když

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

Toto tvrzení se dá zobecnit pro obecně *n* jevů:

$$P\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} A_i$$

a v kombinaci s předešlou definicí nezávislých jevů platí:

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

3.2. Náhodná proměnná

Důležitým pojmem v pravděpodobnosti i statistice je *náhodná proměnná*. Zde opět zdůrazníme důležitost v oboru simulací. Je nám jasné, že většina veličin, které budeme potřebovat získávat nebo generovat budou nabývat obecně reálných hodnot. Proto je výhodné definovat náhodnou proměnnou jako funkci realizující tento požadavek. Pokud (Ω, Σ, P) bude pravděpodobnostním prostorem, pak náhodná proměnná bude funkcí $X : \Omega \to R$, kde pro každé $x \in R$ je

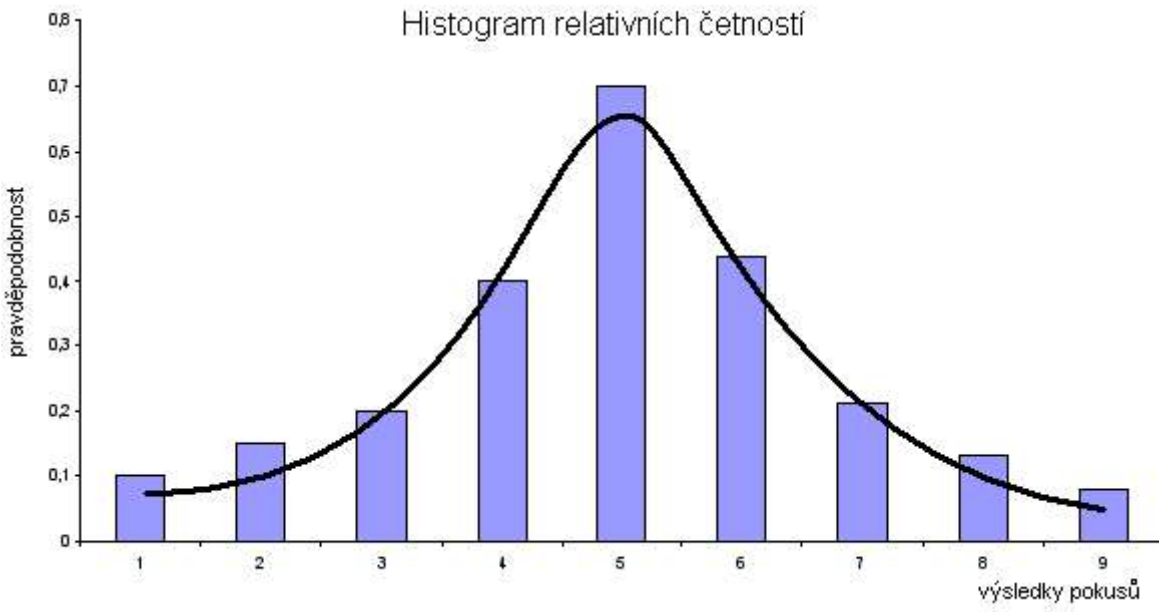
$$X^{-1}((-\infty, x)) = \{\omega \in \Omega | X(\omega) \leq x\} \in \Sigma.$$

Máme-li náhodnou proměnnou je vhodné taky definovat *rozdělení náhodné proměnné*, které bude definovat charakter nahodilosti dané stochastické funkce. Tato funkce bude vlastně každému možnému výsledku přiřazovat pravděpodobnost s jakou nastane a bude definovaná jako:

$$P^X((-\infty, x)) = X^{-1}((-\infty, x)).$$

Princip této funkce lze dobře pochopit na definici *histogramu relativních četností*. Tento bude mít význam i pro simulační techniky, což uvidíme již vzápětí. Řekněme, že provádíme nějaký experiment a chceme zjistit charakteristiky nahodilosti jeho výsledků (tj. pravděpodobnosti jeho jednotlivých hodnot). Provádíme tedy pokus stále dokola a získáváme hodnoty. Celý interval možných výsledků rozdělíme na několik podintervalů, přičemž poslední z nich bude (v případě, že jsou hodnoty z celého R) vést až do nekonečna. Posléze provádíme experiment a zaznamenáváme hodnoty do tabulky výsledků. Po provedení nějakého počtu pokusů jsme

schopni nakreslit sloupcový graf. Když půjdeme v úvaze ještě dál a budeme uvažovat, že velikost intervalů se bude limitně blížit nule a počet pokusů nekonečnu, pak získáme spojitou funkci, která se dá prohlásit za rozdělení náhodné veličiny. Celý tento postup je vidět na následujícím obrázku:



Obrázek 1. Histogram relativních četností

Z definice histogramu relativních četností vyplývá jedno důležité rozdělení náhodných proměnných. Jedná se o rozdělení na *diskrétní* a *spojité náhodné proměnné*. Rozdíl mezi nimi je v množině možných funkčních hodnot. U spojité proměnné je to spojitý interval hodnot, kdežto u diskrétní je to množina diskrétních funkčních hodnot a její rozdělení vypadá v podstatě jako histogram relativních četností. U jednotlivých proměnných se používá následující

Kromě rozdělení náhodné veličiny lze definovat ještě *distribuční funkci*, která vyjadřuje, jaká je pravděpodobnost, že vygenerovaná hodnota bude menší než zadaná, neboli:

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

Z tohoto se dá také odvodit důležité tvrzení, že:

$$P(a < X \le b) = F(b) - F(a)$$

Distribuční funkce většinou plně udává stochastický charakter náhodné proměnné a je jedním z hlavních hledaných vztahů z případných experimentů.

3.2.1. Diskrétní náhodná proměnná

U diskrétní náhodné proměnné lze v návaznosti na rozdělení definovat *pravděpodobnostní funkci P(X)* jako:

$$P(X) = P(X = x_i) \qquad \text{pro } i = 1,2,...$$

$$P(X) = 0 \qquad \text{jinak,}$$

pro níž platí

$$P(a \le X \le b) = \sum_{\{\, i,\, a \le x_i \le b \,\}} P(x_i)$$

Další důležitou funkcí je *distribuční funkce*, která v případě diskrétní proměnné se dá také vyjádřit jako:

$$F(x) = \sum_{\{\, i,\, x_i \le x \,\}} P(x_i)\,,$$

pro níž platí

$$\sum_i P(x_i) = 1\,.$$

3.2.2. Spojitá náhodná proměnná

Pro spojitou náhodnou proměnnou rozšíříme definiční obor na celý přístupný interval a sumy se promění na integrály a tudíž z rozdělení vznikne funkce, kterou nazýváme *hustotou pravděpodobnosti f(x)*, pro kterou platí:

1. $P(a < X \le b) = P(a \le X < b) = P(a < X < b) = P(a \le X \le b) = \int_a^b f(x)\,dx$
2. $\int_{-\infty}^{\infty} f(x)\,dx = 1$

Z tohoto vidíme, že pravděpodobnost, že daná proměnná leží v zadaném intervalu se rovná obsahu plochy určené hustotou pravděpodobnosti. Je také jasné, že platí pro všechny body *x*, že

$$P(X = x_0) = 0\,.$$

Distribuční funkce je pak definovaná jako

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

3.3. Více náhodných veličin a náhodný vektor

3.3.1. Obecné definice s náhodným vektorem

Náhodné proměnné mohou existovat i ve větších souborech, ve kterých se mohou vzájemně ovlivňovat. *Náhodným vektorem* budeme rozumět vektor $X = (X_1, \ldots, X_n)^T$ náhodných proměnných X_i.

K náhodnému vektoru lze snadno dodefinovat příslušnou *sdruženou distribuční funkci náhodného vektoru X* předpisem:

$$F(x_1, \ldots, x_n) = P(X_1 \le x_1, \ldots, X_n \le x_n)\,,$$

který současně znamená splnění podmínek, kdy náhodná proměnná X_i je menší nebo rovna x_i pro všechna i.

O takto nadefinované funkci *F* náhodného vektoru *X* lze tvrdit, že pro interval $I \in \langle a_1, b_1 \rangle \times \langle a_2, b_2 \rangle \subset R^2$ platí

$$P(X \in I) = F(b_1, b_2) - F(a_1, b_2) - F(b_1, a_2) + F(a_1, a_2)$$

Podobně můžeme nadefinovat pro diskrétní náhodný vektor *sdruženou pravděpodobnostní funkci náhodného vektoru X* $P(x) = P(x_1, \ldots, x_n)$ s množinou $M = \{x_1, x_2, \ldots\} \subset R^n$ udávající obor *hodnot vektoru X* jako

$$P(x_i) = P(X = x_i) \quad \text{pro } i = 1,2,\ldots$$
$$P(x) = 0 \qquad\qquad \text{pro } x \notin M\,,$$

pro kterou platí klasické upravené věty:

1. $F(X) = F(x_1,...,x_n) = \sum_{\substack{t_1,...,t_n \in M \\ t \le x_1,...,t_n \le x_{n1}}} P(t_1,....,t_n)$
2. $\sum_{x \in M} P(x) = 1$
3. $P(X \in A) = \sum_{x \in A} P(x)$

Stejně tak pro spojitou náhodnou proměnnou platí vztah pro *sdruženou hustotu pravděpodobnosti náhodného vektoru X f(x)* mající tvar:

$$F(x) = F(x_1,...,x_n) = \int_{-\infty}^{x_1} ... \int_{-\infty}^{x_n} f(t_1,...,t_n)\, dt_n ... dt_1 ,$$

pro níž opět platí očekávané vztahy:

1. $f(x_1,...,x_n) = \frac{\partial^n F(x_1,...,x_n)}{\partial x_1 ... \partial x_n}$
2. $\int_{R^N} ... \int f(x) dx_1 ... dx_n = 1$
3. $P(X \in A) = \int_A ... \int f(x) dx_1 ... dx_n$

3.3.2. Sdružené a marginální rozdělení dvou proměnných

Nejjednodušším případem takového uspořádání je existence dvou náhodných proměnných. Na tomto příkladu se dají popsat pravidla platící v jisté míře i ve vyšším počtu proměnných. V případě nespojitých veličin můžeme jejich rozdělení popsat *sdruženou (simultánní) pravděpodobnostní funkcí*.

$$P(X = x \cap Y = y) = P(x, y)$$

udávající pravděpodobnost současného splnění obou požadavků. Sdružené pravděpodobnosti se uspořádávají do tzv. *korelační tabulky*.

tabulka 1. Korelační tabulka.

X	**y**				**součet**
	y_1	**y_2**	**…**	**y_s**	
x_1	$P(x_1, y_1)$	$P(x_1, y_2)$	…	$P(x_1, y_S)$	$P_1(x_1)$
x_2	$P(x_2, y_1)$	$P(x_1, y_2)$	…	$P(x_2, y_S)$	$P_1(x_2)$
...	…	….	…	…	…
X_R	$P(x_R, y_1)$	$P(x_R, y_2)$	…	$P(x_R, y_S)$	$P_1(x_R)$
součet	$P_2(y_1)$	$P_2(y_2)$	...	$P_2(y_S)$	1

Řádkové a sloupcové součty jsou hodnoty takzvaných marginálních funkcí udávajících pravděpodobnosti nezávislé na druhé proměnné.a platí pro tudíž, že

$$P_1(x)=\sum_y P(x,y) \qquad \text{a} \qquad P_2(y)=\sum_x P(x,y)$$

a také tudíž vyplývající věta

$$\sum_x\sum_y P(x,y)=\sum_x P_2(y)=\sum_y P_1(x)=1$$

Při platnosti těchto vztahů se potom pravděpodobnost, že proměnná *X* bude v intervalu <*x1,x2*> a proměnná *Y* v intervalu <*y1,y2*> bude rovnat

$$P(x_1\le X\le x_2\cap y_1\le Y\le y_2)=\sum_{x=x_1}^{x_2}\sum_{y=y_1}^{y_2}P(x,y)\,.$$

Nyní definujeme *sdruženou distribuční funkci* těchto dvou veličin výrazem

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

s klasickými vlastnostmi distribuční funkce:

$$F(-\infty,y)=F(x,-\infty)=F(-\infty,-\infty)=0 \qquad \text{a} \quad F(\infty,\infty)=1\,.$$

Pak zmiňovaná pravděpodobnost příslušnosti do intervalů bude definována jako

$$P(x_1\le X\le x_2\cap y_1\le Y\le y_2)=F(x_2,y_1)-F(x_1,y_2)-F(x_2,y_1)+F(x_1,y_1)\,.$$

Také můžeme v návaznosti definovat i *marginální distribuční funkce*

$$F_1(x)=F(x,\infty) \qquad \text{a} \qquad F_2(y)=F(\infty,y)\,.$$

Pro spojitou veličinu můžeme definovat *sdruženou hustotu pravděpodobnosti* f(x,y) splňující podmínku

$$\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty}f(x,y)\,dx\right]dy=\int_{\Omega_y}\left[\int_{\Omega_x}f(x,y)\,dx\right]dy=1$$

Její vztah k distribuční funkci je stejný jako v jednorozměrném případě a tudíž

$$F(x,y)=\int_{-\infty}^{y}\left[\int_{-\infty}^{x}f(t,u)\,dt\right]du \quad \text{a} \qquad f(x,y)=\frac{\partial^2F(x,y)}{\partial x\partial y}$$

Nakonec definujeme také *marginální hustoty pravděpodobnosti*

$$f_1(x)=\int_{\Omega_x}f(x,y) \qquad \text{a} \qquad f_2(y)=\int_{\Omega_y}f(x,y)$$

3.3.3. Podmíněná rozdělení pravděpodobnosti pro dvě proměnné

Podmíněné rozdělení náhodné veličiny je vlastně rozdělením první veličiny za předpokladu, že druhá nabývá nějaké hodnoty a je definováno jako podíl sdruženého a marginálního rozdělení.

Pro diskrétní náhodnou veličinu definujeme *podmíněné pravděpodobnostní funkce* jako

$$P(x|y)=\frac{P(x,y)}{P_2(y)},\quad P_2(y)\neq 0 \qquad \text{a} \qquad P(y|x)=\frac{P(x,y)}{P_1(x)},\quad P_1(x)\neq 0\,.$$

K tomuto přidáme odpovídající definice *podmíněných diskrétních distribučních funkcí*

$$F(x|y)=\frac{\sum_{t<x}P(x,y)}{P_2(y)},\quad P_2(y)\neq 0 \qquad \text{a} \quad F(y|x)=\frac{\sum_{t<y}P(x,y)}{P_1(x)},\quad P_1(x)\neq 0$$

Pro tyto rozdělení platí vztahy

1. $P_1(x)>0 \;\Rightarrow\; P(x,y)=P(x|y)P_1(x)$

2. $P_2(y) > 0 \quad \Rightarrow \quad P(x,y) = P(y|x)P_2(y)$
3. *Bayerův vzorec* $P(y|x) = \dfrac{P(x|y)P_2(y)}{P_1(x)}$ (odvození viz 4.1. Základní experiment a odhad pravděpodobnosti)

Pro spojité náhodné veličiny lze zase definovat obdobně *podmíněné hustoty pravděpodobnosti* jako

$$f(x|y) = \frac{f(x,y)}{f_2(y)}, \quad f_2(y) \neq 0 \qquad \text{a} \quad f(y|x) = \frac{f(x,y)}{f_1(x)}, \quad f_1(x) \neq 0$$

a k tomu *podmíněné spojité distribuční funkce*

$$F(x|y) = \frac{\int_{-\infty}^{x} f(t,y)dt}{f_2(y)}, \quad f_2(y) \neq 0 \qquad \text{a} \quad F(y|x) = \frac{\int_{-\infty}^{x} f(y,u)du}{f_1(x)}, \quad f_1(x) \neq 0 .$$

Pro tyto funkce zase platí následující vztahy

4. $f_1(x) > 0 \quad \Rightarrow \quad f(x,y) = f(x|y)f_1(x)$
5. $f_2(y) > 0 \quad \Rightarrow \quad f(x,y) = f(y|x)f_2(y)$
6. *Bayerův vzorec* $f(y|x) = \dfrac{f(x|y)f_2(y)}{f_1(x)}$

Z podmíněných rozdělení přímo vyplývá pojem *nezávislosti proměnných*. Řekneme, že dvě náhodné proměnné *X* a *Y* jsou nezávislé, když pro každé *y*, pro něž platí $P_2(y) > 0$ (resp. $f_2(y) > 0$) platí

$$P(x|y) = P_1(x) \qquad \text{resp.} \qquad f(x|y) = f_1(x) .$$

Potom ovšem pro nezávislé jevy platí i následující tvrzení

1. $f(x,y) = f_1(x)f_2(x)$
2. $P(x,y) = P_1(x)P_2(x)$
3. $F(x,y) = F_1(x)F_2(x)$

3.4. Charakteristiky náhodných veličin

Jak již jsme zmínili, je distribuční funkce hlavní hledanou charakteristikou stochastického průběhu náhodné veličiny. V praxi však mnohdy stačí získat některé číselné charakteristiky rozdělení k jeho postačujícímu popisu. Charakteristiky lze rozdělit na několik skupin.

V prvním dělení rozlišujeme podle tvaru charakteristiky

- Charakteristiky polohy … Měří úroveň náhodné veličiny.
- Charakteristiky variability … Měří rozptýlení funkce náhodné veličiny.
- Charakteristiky dynamiky … Jedná se hlavně o charakteristiky šikmosti a strmosti.

V dalším rozdělení jde o způsob, jakým se charakteristiky získává

- Charakteristiky momentové … Jsou funkcí všech hodnot, které může náhodná veličina nabýt.

- Charakteristiky kvantilové … Jsou to určité hodnoty náhodné veličiny, které vyhovují zadané podmínce.

3.4.1. Charakteristiky polohy

Charakteristiky polohy definují střed, kolem kterého kolísají náhodné veličiny. Obecně lze nadefinovat *k-obecný moment náhodné* diskrétní (resp. spojité) proměnné X jako:

$$\mu_k'(x) = \sum_x x^k \cdot P(x) \qquad \text{resp.} \quad \mu_k'(x) = \int_{-\infty}^{\infty} x^k \cdot P(x)\,.$$

Nicméně nejdůležitějším obecným momentem je moment první, který se nazývá *střední hodnota*, jež definujeme jako:

$$E(x) = \sum_x x \cdot P(x) \qquad \text{resp.} \quad E(x) = \int_{-\infty}^{\infty} x \cdot P(x)\,.$$

Každá hodnota náhodné proměnné je započítávána s takovou silou jaká je její pravděpodobnost určená jejím rozdělením. Z toho vyplívá, že nejvíce se ve střední hodnotě projeví hodnoty s největší pravděpodobností. Dále pro střední hodnotu platí:

- $E(aX+b) = aE(X)+b$
- $E(X_1 + X_2 + ... + X_n) = E(X_1) + E(X_2) + ... + E(X_n)$
- X_i jsou vzájemně nezávislé $\Rightarrow E(X_1 X_2 ...) = E(X_1)E(X_1)\,....\,E(X_1)$

3.4.2. Charakteristiky variability

Charakteristiky polohy nadefinovaly hodnotu, kolem které kolísají výsledky náhodné veličiny. K tomu, abychom mohli mít lepší představu o tom, jak výsledná funkce vypadá, je dobré mít navíc informaci o tom jak moc jsou jednotlivé hodnoty náhodné veličiny rozptýleny kolem tohoto středu. K tomuto účelu slouží charakteristiky variability.

Definujme nejprve obecně *k-tý centrální moment náhodné proměnné*.

$$\mu_k(x) = \sum_x [x - E(x)]^k P(x) \qquad \text{a} \qquad \mu_k'(x) = \int_{-\infty}^{\infty} [x - E(x)]^k f(x)\,dx\,.$$

Dá se ukázat, že první centrální moment je roven nule, nicméně již druhý centrální moment se hojně využívá. Jedná se o tzv. *rozptyl* a je definován jako (pro diskrétní a spojitou veličinu):

$$D(X) = \mu_2(X) = \sum_x [x - E(x)]^2 P(x) \qquad \text{a} \qquad D(X) = \mu_2(X) = \int_{-\infty}^{\infty} [x - E(x)]^2 f(x)\,dx\,.$$

Rozptyl tedy udává variabilitu náhodné veličiny a to tím způsobem, že jeho hodnota rovna nule znamená, že daná hodnota bude vycházet stále. Zvětšující se hodnota rozptylu značí větší rozptýlenost hodnot.

Někdy potřebujeme odchylku funkce ve stejných jednotkách jako byla původní funkce (rozptyl udává odchylku ve čtvercích). Proto je zavedena další veličina s názvem *směrodatná odchylka*.

$$\sigma(X) = \sqrt{D(X)}$$

Podle této veličiny se někdy rozptyl značí $\sigma^2(X)$. Pro rozptyl opět uvedeme věty, které o něm platí při počítání s ním.

1. $D(aX+b)=a^2D(X)$
2. $D(a)=0$
3. $D(X\pm Y)=D(X)+D(Y)$
4. $D(X)=E(X^2)-[E(X)]^2$ … tzv. *výpočtový tvar rozptylu*

3.4.3. Charakteristiky dynamiky

Charakteristiky dynamiky uvádějí informace o tvaru křivky a můžeme z nich zjistit jak hodnoty funkce rychle stoupají nebo klesají a jiné informace.

Pro definici charakteristik dynamiky musíme nejprve definovat pojmy *normované náhodné veličiny* a *normovaného rozptylu*.

Normovaná náhodná veličina je náhodnou veličinou, která je upravená charakteristikami na takový tvar, aby měla následující hodnoty střední hodnoty a rozptylu

$$E(U)=0 \qquad \text{a} \qquad D(U)=1.$$

Toho docílíme úpravou původní veličiny následujícím způsobem

$$U=\frac{X-E(X)}{\sigma(X)}$$

Pro tuto veličinu pak můžeme nadefinovat obecné momenty, které se budou zde označovat jako *k-té normované momenty*.

$$\mu_k'(x)=\sum_x\left(\frac{X-E(X)}{\sigma(X)}\right)^k\cdot P(x) \qquad \text{a} \qquad \mu_k'(x)=\int_{-\infty}^{\infty}\left(\frac{X-E(X)}{\sigma(X)}\right)^k\cdot P(x).$$

Důležitým faktem daných normovaných charakteristik i momentů je to, že jsou bezrozměrné. Další důležitou vlastností je fakt, že každý normovaný moment lze vyjádřit pomocí centrálních momentů a ty zase pomocí momentů obecných. Tvar převodu z normovaného na centrální je pro sudé n:

$$\mu_n(U)=\frac{\mu_n(X)}{[\mu_2(X)]^{n/2}}$$

a pro liché n (dělení n / 2 ve jmenovateli je celočíselné)

$$\mu_n(U)=\frac{\mu_n(X)}{[\mu_2(X)]^{n/2}\cdot\sqrt{\mu_2(X)}}$$

Důležitý normovaný moment je třetí normovaný moment $\mu_3(U)$ *koeficient šikmosti*. U symetrického rozdělení je tato veličina rovna nule. Jeho veličiny tak informují o zešikmení rozdělení na příslušnou stranu podle znaménka koeficientu.

1. $\mu_3(U)=0$ … Symetrické rozdělení.
2. $\mu_3(U)>0$ … Zešikmení do leva.
3. $\mu_3(U)<0$ … Zešikmení do prava.

Další důležitou charakteristikou je *špičatost*, která se měří čtvrtým normovaným koeficientem. Častěji ovšem uvádíme jako míru špičatosti koeficient špičatosti, který se spočte jako

$$\mu_4(X)-3.$$

Při této definici má tento koeficient následující vlastnosti

1. $\mu_4(X)-3=0$ … Špičatost shodná s normálním rozdělením.

2. $\mu_4(X) - 3 > 0$ … Špičatost větší než u normálního rozdělení.
3. $\mu_4(X) - 3 < 0$ … Špičatost menší než u normálního rozdělení.

3.4.4. Kvantilové charakteristiky – kvantily

Kvantily se používají hlavně pro spojité veličiny. Jedná se hlavně o *100P% kvantit* x_p je taková hodnota náhodné veličiny, která je určena podmínkou

$$F(x_p) = P\,.$$

Řečeno slovy, hledáme takovou hodnotu, pro kterou všechny menší hodnoty z oboru hodnot náhodné proměnné (tj. $x_i \le x_p$) tvoří 100p% celého souboru. Také se dá říci, že hodnota kvantitu znamená, že náhodná veličina nabude s pravděpodobností p hodnoty z intervalu $\langle -\infty, x_p \rangle$, neboli pro spojitou proměnnou obsah plochy nad tímto intervalem omezený hustotou se bude rovnat p. Zde jsou některé nejdůležitější kvantilové hodnoty:

- $X_{0,5}$ … *medián*
- $X_{0,25}$ … *dolní kvartil*
- $X_{0,75}$ … *horní kvartil*
- $X_{k/10}$ … *k-tý decil*
- $X_{k/100}$ … *k-tý percentil*

Kromě těchto nejznámějších veličin se používá například hodnota veličiny kvadrilové odchylky mající hodnotu

$$Q(X) = \frac{1}{2}\left(x_{0,75} - x_{0,25}\right).$$

3.4.5. Charakteristiky vícerozměrné náhodné veličiny

Tyto charakteristiky lze rozdělit do několika skupin.

První skupinou jsou charakteristiky marginálních rozdělení, které mají vlastně stejné vzorce pro výpočty charakteristik jako jednorozměrné veličiny.

$$E(x) = \sum_x x \cdot P_1(x) \qquad \text{resp.} \quad E(x) = \int_{-\infty}^{\infty} x \cdot P_1(x)$$

a stejně tak pro rozptyl

$$D(X) = \sum_x \left[x - E(x)\right]^2 P_1(x) \qquad \text{a} \qquad D(X) = \int_{-\infty}^{\infty} \left[x - E(x)\right]^2 f_1(x)\,dx\,.$$

Obdobné vztahy platí i pro veličinu Y.

Další třídou jsou charakteristiky podmíněných rozdělení. Zde pro rozptyl platí následující vztahy:

$$E(X|y) = \sum_x x \cdot P(x|y) \qquad \text{resp.} \quad E(X|y) = \int_{-\infty}^{\infty} x \cdot P(x|y)$$

a pro rozptyl platí

$$D(X|y) = \sum_x \left[x - E(X|y)\right]^2 P(x|y) \qquad \text{a} \qquad D(X|y) = \int_{-\infty}^{\infty} \left[x - E(X|y)\right]^2 f(x|y)\,dx\,.$$

Obdobně i pro obrácenou závislost.

Poslední třídou jsou charakteristiky uvádějící vzájemné vazby jednotlivých veličin. Prvním z nich je *kovariance*. Je definována jako střední hodnota součinu odchylek veličin od jejich středních hodnot, neboli

$$C(X,Y) = E\{[X - E(X)][Y - E(Y)]\}.$$

Pro výpočty se však často používá vzorec

$$C(X,Y) = E(XY) - E(X)E(Y).$$

Mnohem více využívanou mírou těsnosti je ovšem *koeficient korelace*, který je definován jako

$$\rho(X,Y) = \frac{C(X,Y)}{\sigma(X)\sigma(Y)}.$$

Výhodou tohoto koeficientu je, že nabývá pouze hodnot od -1 do 1 (kovariance má obor hodnot všechna reálná čísla). Potom mají jednotlivé extrémy následující významy:

1. ± 1 … Přímá popř. nepřímá lineární závislost.
2. 0 … Veličiny jsou lineárně nezávislé, tzv. *nekorelované*. Nemusí však být úplně nezávislé.

Pokud máme dán náhodný vektor, popisuje se toto rozdělení většinou pomocí

1. *Vektoru středních hodnot* $\mu = [\mu_1 \dots \mu_n]$ a
2. *kovarianční maticí* $C = \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1S} \\ c_{21} & c_{22} & \dots & c_{2S} \\ \dots & \dots & \dots & \dots \\ c_{R1} & c_{R2} & \dots & c_{RS} \end{bmatrix}$

Kovarianční matice je symetrická matice, která má na diagonále rozptyly jednotlivých veličin a ostatní prvky jsou kovariance.

3.4.6. Momentová vytvořující a charakteristická funkce

Další pomocnou konstrukcí při práci s rozděleními náhodné proměnné tvoří *momentová vytvořující funkce*. Ta je definována jako

$$m_x(z) = E\left(e^{zX}\right)$$

neboli pro diskrétní resp. spojitou veličinu

$$m_x(z) = \sum_x e^{zx} P(x) \qquad \text{resp.} \quad m_x(z) = \int_{-\infty}^{\infty} e^{zx} f(x)dx.$$

Takto definovaná funkce je jednoznačná pro každé rozdělení a tak ji lze pokládat za formu jeho popisu. Momentová vytvořující funkce se používá na odvozování rozdělení náhodných veličin nebo na výpočet momentů. Na výpočet rozdělení lze použít vztahů vyjadřující momentovou vytvořující funkci pro veličinu *Y*, která je funkcí veličiny *X*.

$$m_y(z) = \sum_x e^{zy(x)} P(x) \qquad \text{resp.} \quad m_y(z) = \int_{-\infty}^{\infty} e^{zy(x)} f(x)dx.$$

Pro výpočet obecných momentů je dobré zase použít vztahu pro k-tou derivaci momentové vytvořující funkce.

$$m_x^{(k)}(z)\big|_{z=0} = E\left(X^k e^{zX}\right)\big|_{z=0} = \mu_k'(X).$$

Kromě momentové vytvořující funkce existuje ještě *funkce charakteristická*, která je momentové velice podobná (liší se pouze zavedením komplexní jednotky) a proto uvedeme pouze příslušné vztahy

$$\varphi_x(z) = E\left(e^{izX}\right)$$

neboli pro diskrétní resp. spojitou veličinu

$$m_x(z) = \sum_x e^{izx} P(x) \qquad \text{resp.} \quad m_x(z) = \int_{-\infty}^{\infty} e^{izx} f(x)dx .$$

Pro výpočty momentů

$$m_x^{(k)}(z)\big|_{z=0} = i^k \mu_k'(X) .$$

Navíc lze z charakteristické funkce určit hustotu pravděpodobnosti jako

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{izx} \varphi_x(z)dz .$$

Poslední tvrzení se dá použít pro zjišťování rozdělení náhodné veličiny. Platí, že náhodné veličiny mají stejné rozdělení, právě když mají stejnou charakteristickou a momentovou vytvořující funkci.

3.5. Některá důležitá rozdělení náhodných veličin

V praxi se často setkáváme s náhodnými jevy velice podobného charakteru. Proto pro určité charakteristické situace byly vytvořeny předpisy pro rozdělení náhodných veličin. Pro přesnější určení rozdělení pro danou situaci jsou všechna tato standardní rozdělení závislá na parametrech, které je upřesňují.

3.5.1. Diskrétní veličina

Při vyjmenování rozdělení budeme postupně uvádět: vysvětlení rozdělení, definici pravděpodobnostní funkce, hodnoty charakteristik, aproximace jinými rozděleními a případně důležité vlastnosti. Při definování pravděpodobnostní funkce budeme uvádět pouze nenulové hodnoty. Rozumíme tím, že v ostatních bodech jsou funkce nulové.

3.5.1.1. Binomické rozdělení Bi(n, p)

Binomické rozdělení vyjadřuje kolik bude úspěšných pokusů, s pravděpodobností úspěšnosti *p*, při *n* nezávislých opakování.

Dalším případem použití tohoto rozdělení je situace, kdy provádíme náhodný výběr s vracením. Mějme například N prvků s tím, že pravděpodobnost vybrání prvku s nějakou vlastností je p. Při vybírání prvky vracíme do množiny. Počet prvků se sledovanou vlastností se řídí tímto rozdělením.

$$P(X) = \binom{n}{x} p^x (1-p)^{n-x} \qquad \text{pro } x = 0,1,\ldots .$$

Charakteristiky:

$$E(X) = np$$

$$D(X) = np(1-p)$$

$$\mu_3(U) = \frac{1-2p}{\sqrt{np(1-p)}}$$

$$\mu_4(U) - 3 = \frac{1-6p(1-p)}{np(1-p)}$$

Aproximace:

Binomické rozdělení jde aproximovat těmito rozděleními při těchto podmínkách:

- Pro velké *n* a malé *p* ($n \geq 30,\ p \leq 0{,}1$) lze použít aproximaci Poissonovým rozdělením. Pak platí $Bi(n,p) \approx Po(\lambda)$, kde $\lambda = np$
- Pro případy $D(X) \geq 9$ lze aproximovat normálním rozdělením. Pak pro provedenou operaci platí $Bi(n,p) \approx N(\mu,\sigma^2)$, kde $\mu = np, \sigma^2 = np(1-p)$ (viz 3.5.2.3. Normální (Gaussovo) rozdělení N(μ, σ2)).

Alternativní rozdělení A(p)

Je pouze speciálním případem binomického rozdělení s parametrem *n = 1* (tj. *A(p) = Bi(1,p)*). Náhodná proměnná s tímto rozdělením může nabývat pouze hodnot 0 nebo 1. Jde tedy o to, zda-li zkoumaná událost nastane či nikoliv během jednoho pokusu.

Důležité vlastnosti:

Suma n rozdělení A(p) dává rozdělení Bi(n, p).

3.5.1.2. Poissonovo rozdělení Po(λ)

Poissonovo rozdělení se používá pro určení kolikrát nastane daná událost, víme-li, že jsou události zcela náhodné s průměrným výskytem λ za časovou jednotku.

Dále se toto rozdělení používá pro určení počtu poruch přístroje.

$$P(X) = \frac{\lambda^x}{x!} e^{-\lambda} \qquad \text{pro } x = 0,1,\ldots .$$

Charakteristiky:

$$E(X) = \lambda$$

$$D(X) = \lambda$$

$$\mu_3(U) = \frac{1}{\sqrt{\lambda}}$$

$$\mu_4(U) - 3 = \frac{1}{\lambda}$$

Aproximace

- Pro případy $\lambda \geq 9$ lze aproximovat normálním rozdělením. Pak pro provedenou operaci platí $Po(\lambda) \approx N(\mu,\sigma^2)$, kde $\mu = \lambda, \sigma^2 = \lambda$ (viz 3.5.2.3. Normální (Gaussovo) rozdělení N(μ, σ2)).

Důležité vlastnosti:

Suma náhodných proměnných s poissonovým rozdělením je rovna náhodné proměnné s poissonovým rozdělením a parametrem daným součtem původních parametrů.

3.5.1.3. Hypergeometrické rozdělení HG(N, M, n)

Hypergeometrické rozdělení vyjadřuje náhodné výběry bez vracení. Pokud budeme mít N prvků z nichž M bude mít sledovanou vlastnost, pak při výběru n prvků bez vracení se bude počet prvků se sledovanou vlastností řídit hypergeometrickým rozdělením.

$$P(X) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} e^{-\lambda} \quad \text{pro } 0 \le x \le n, \quad x \le M, \quad n-x \le N-M \ .$$

Charakteristiky:

$$E(X) = n\frac{M}{N}$$

$$D(X) = \frac{M}{N}\left(1-\frac{M}{N}\right)\frac{N-n}{N-1}$$

Aproximace:

- Je-li n malé vůči M a také vůči N-M neovlivní nevracení prvků pravděpodobnosti $p = \frac{M}{N}$ a $1-p = \frac{N-M}{N}$. Lze tudíž provést aproximaci rozdělením výběru s vracením, tj. binomickým rozdělením. Získané rozdělení pak bude v následujícím vztahu s původním: $HG(N,M,n) \approx Bi(n,p)$, kde $p = \frac{M}{N}$.

3.5.1.4. Geometrické rozdělení G(p)

Geometrické rozdělení popisuje kolik neúspěšných pokusů předchází úspěšnému (úspěšný pokus má pravděpodobnost p).

$$P(X) = p(1-p)^x \qquad \text{pro } x = 0,1,\ldots .$$

Charakteristiky:

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

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

3.5.1.5. Negativní binomické rozdělení NG(n, p)

Negativní binomické rozdělení popisuje kolik úspěšných pokusů předchází n-tému úspěšnému, tj. je obecnějším případem geometrického rozdělení.

$$P(X) = \binom{x+n-1}{n-1} p^n (1-p)^x \qquad \text{pro } x = 0,1,\ldots .$$

Charakteristiky:

$$E(X) = \frac{n(1-p)}{p}$$

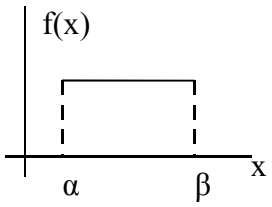
$$D(X) = \frac{n(1-p)}{p^2}$$

3.5.2. Spojitá náhodná veličina

Při definování spojitých rozdělení se budeme držet stejného schématu jako u rozdělení diskrétních, tj. viz 3.5.1. Diskrétní veličina.

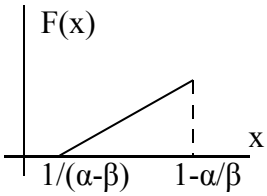
3.5.2.1. Rovnoměrné rozdělení R(α, β)

Rovnoměrné rozdělení na určitém intervalu vyjadřuje pro náhodnou proměnnou vlastnost, že při výběru z daného intervalu budou mít všechny možnosti stejnou pravděpodobnost.



$$f(X) = \frac{1}{\alpha - \beta} \qquad \text{pro } \alpha < x < \beta$$

$$F(X) = \begin{cases} \dfrac{x-\alpha}{\beta-\alpha} & \textit{pro } \alpha < x < \beta \\ 1 & \textit{pro } x \geq \beta \end{cases}$$

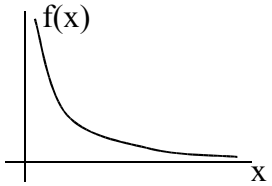


$$E(X) = \frac{\alpha + \beta}{2}$$

$$D(X) = \frac{(\beta - \alpha)^2}{12}$$

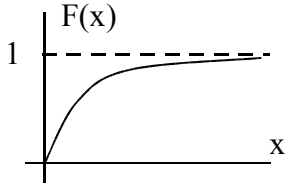
3.5.2.2. Exponenciální rozdělení E(A, λ)

Exponenciální rozdělení se používá v případech, kdy chceme modelovat dobu čekání na určitou událost (např. porucha součástky, příjezd vozidla do ulice apod.).



$$f(X) = \lambda e^{-\lambda(x-A)} \quad \text{pro } x > A$$

$$F(X) = 1 - \lambda e^{-\lambda(x-A)} \quad \text{pro } x > A$$



$$E(X) = A + \frac{1}{\lambda}$$

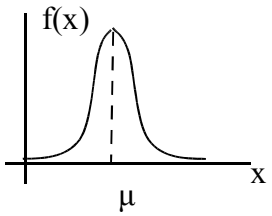
$$D(X) = \frac{1}{\lambda^2}$$

Důležité vlastnosti:

- Často se používá exponenciální rozdělení s parametrem A = 0 a označuje se jako Exp (λ). Všechny charakteristiky jsou stejné s tím, že za A dosadíme nulu.
- Součet náhodných proměnných s exponenciálním rozdělením Exp(λ) se rovná náhodné proměnné s rozdělením gama (viz 4.6).

3.5.2.3. Normální (Gaussovo) rozdělení N(μ, σ²)

Gaussovo normální rozdělení je jedno z nejdůležitějších spojitých rozdělení. Normálním rozdělením se řídí náhodné proměnné, které vzniknou součtem velkého počtu proměnných s různými rozděleními, z nichž žádné nemá dominantní vliv.



$$f(X) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad \text{pro } \sigma^2 > 0, \quad -\infty < \mu < \infty$$

$$E(X) = \mu$$

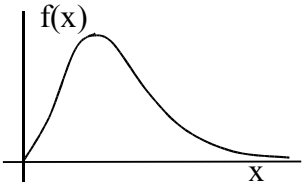
$$D(X) = \sigma^2$$

Důležité vlastnosti:

- Často se používá normální rozdělení s parametry 0 a 1. Vzniklé rozdělení pak pojmenujeme jako *normované normální rozdělení N(0, 1)*. Jeho distribuční funkce je tabelována a značíme ji jako Φ.
- Pokud chceme převést normální rozdělení na normované normální rozdělení, musíme z proměnné vytvořit normovanou proměnnou transformací: $U = \frac{X - \mu}{\sigma}$. Ta pak bude mít normované normální rozdělení.
- Pro distribuční funkci normovaného rozdělení platí, že její hodnota se dá určit z tabelovaných hodnot distribuční funkce normovaného normálního rozdělení jako $F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$.
- Pro distribuční funkci normované náhodné proměnné platí $\Phi(-k) = 1 - \Phi(k)$
- Je-li X náhodná proměnná s normovaným rozdělením, pak pravděpodobnost, že náhodná veličina X nabude hodnoty z intervalu $\mu \pm k\sigma$ je rovna
  $P(\mu - k\sigma < x < \mu + k\sigma) = F(\mu + k\sigma) - F(\mu - k\sigma) = \Phi(k) - \Phi(-k) = 2\Phi(k) - 1$
- Uvedeme tři klasické případy zjišťování příslušnosti výsledku náhodné veličiny do intervalu.
  - $P(X \le a) = F(a) = \Phi\left(\frac{a-\mu}{\sigma}\right)$
  - $P(X > a) = 1 - F(a) = 1 - \Phi\left(\frac{a-\mu}{\sigma}\right)$
  - $P(a < X < b) = F(b) - F(a) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$

3.5.2.4. Logaritmicko-normální rozdělení LN(μ, σ)

Logaritmicko normální rozdělení je upraveným normálním rozdělení. Oproti normálnímu rozdělení jsme zavedli zešikmení rozdělení. Použití je pro situace, kdy je náhodnost veličiny výsledkem velkého množství drobných vlivů, které se navzájem násobí.



$$f(X) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(\ln x - \mu)^2} \qquad \text{pro } x > 0$$

$$F(x) = \Phi\left(\frac{\ln x - \mu}{\sigma}\right) \qquad \text{pro } x \in (0, \infty)$$

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

$$D(X) = e^{2\mu + 2\sigma^2} - e^{2\mu + \sigma^2}$$

Důležité vlastnosti:

- Mají-li X a Y logaritmicko-normální rozdělení, pak součin a podíl těchto veličin mají rozdělení rovna: $X \cdot Y \sim LN(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$ a $X / Y \sim LN(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$.

3.6. Centrální limitní věta

Nyní již jsme si popsali různá rozdělení náhodné veličiny a jsme připraveni vyslovit *centrální limitní větu*.Jednoduše řečeno je-li náhodná proměnná výsledkem většího počtu nezávislých vlivů, pak konverguje k normálnímu rozdělení.

Přesněji řečeno, pokud veličina X je daná součtem veličin X_i mající konečné střední hodnoty a rozptyly, pak normovaná veličina U

$$U = \frac{X - \sum_{i=1}^{n} E(X_i)}{\sqrt{\sum_{i=1}^{n} D(X_i)}}$$

při platnosti *Ljapunovy podmínky*

$$\lim_{n\to\infty} \frac{\sqrt[3]{\sum_{i=1}^{n} \mu_3(X_i)}}{\sqrt{\sum_{i=1}^{n} D(X_i)}} = 0$$

konverguje k normovanému normálnímu rozdělení

$$\lim_{n\to\infty} P(U < u) = \Phi(u) \qquad \text{pro } x \in (-\infty, \infty).$$

Z této věty vyplývá, že pro velké n lze rozdělení proměnné $\sum_{i=1}^{n} X_i$ aproximovat jako $N(n\eta, n\sigma^2)$ a rozdělení proměnné $\sum_{i=1}^{n} \frac{X_i}{n}$ jako $N\left(\eta, \frac{\sigma^2}{n}\right)$.

4. Odhady v náhodných procesech

Simulační úlohy jsou z nějaké části stochastického charakteru a při jejich implementaci se často uplatňují metody, které nám umožňují z pravděpodobnostního hlediska blíže popsat charakter jejich výsledků. Simulace, podobně jako většina užitečných procesů, má svá vstupní a výstupní data. Tyto data mohou být stochastického charakteru jak na vstupu tak na výstupu. My se zde budeme zabývat analýzou výstupních dat simulace. Na tyto data se dají aplikovat různé metody matematické statistiky. My se pokusíme probrat některé z nich s tím, že u některých se zastavíme a popíšeme je podrobněji a jiné pouze hrubě nastíníme. Zpracování výsledků se liší pro různé případy provádění simulačních experimentů, a proto budeme probírat metody podle rozměru experimentů od nejjednodušších až po ty složité.

4.1. Základní experiment a odhad pravděpodobnosti

Předpokládejme, že provádíme experiment, jehož výsledek B je nějak závislý na vzájemně nezávislých alternativách A_i. Jsme tudíž schopni určit pravděpodobnosti $P(B|A_i)$ pro $\forall i$. My však nevíme, která z alternativ je platná a víme naopak jen to, že je splněn výsledek B a chceme tudíž určit pravděpodobnost $P(A_i|B)$ pro $\forall i$. Z nich pak vybereme největší hodnotu pravděpodobnosti a tu pak prohlásíme za odhad skutečné alternativy. Vyjdeme tedy z definice podmíněné pravděpodobnosti

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

Z tohoto vzorce je ještě nutné vyjádřit nějak pravděpodobnost výsledku B. Vyjdeme z předpokladu, že jsou alternativy nezávislé. Lze tedy prohlásit, že

$$\bigcup_{i=1}^{n} A_i = \Omega$$

pomocí čehož lze vyjádřit výsledek pokusu B jako

$$B = B \cap \Omega = B \cap \bigcup_{i=1}^{n} A_i = \bigcup_{i=1}^{n} (A_i \cap B).$$

Dosazením tohoto vztahu do pravděpodobnosti dostáváme

$$P(B) = P\left[\bigcup_{i=1}^{n} (A_i \cap B)\right] = \sum_{i=1}^{n} P(A_i \cap B) = \sum_{i=1}^{n} P(A_i)P(B|A_i)$$

Tento výsledek dosadíme do původního vztahu a získáme tzv. *Bayesův vzorec* (viz také 3.3.3. Podmíněná rozdělení pravděpodobnosti pro dvě proměnné):

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

Poznámky:

1. Pravděpodobnosti jednotlivých alternativ volíme s ohledem na aplikaci metodiky. Nicméně často se používá tzv. *Bayesův postulát*, který je prohlašuje za stejné a tudíž rovny hodnotě:

$$P(A_i) = \frac{1}{n} \qquad \text{pro } \forall i = 1,..,n$$

2. Pravděpodobnosti $P(A_i)$ známe a nebo odhadujeme ještě před výsledkem pokusu, proto se nazývají *apriorní pravděpodobnosti*. Naproti tomu pravděpodobnosti $P(A_i|B)$

zjišťujeme až na základě provedeného pokusu a tudíž je označujeme jako *aposteriorní pravděpodobnosti*.

4.2. Obecnější experiment a odhad pravděpodobnosti

Předpokládejme, že provádíme experiment s počtem realizací výsledku v pokusech charakterizovaným náhodnou proměnnou μ (pro ní se používá označení četnost). Pravděpodobnost, že veličina μ nabude při n nezávislých opakování nějakého zcela konkrétního výsledku k se řídí binomickým rozdělením a bude tedy rovna:

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

Při tomto postupu ovšem předpokládáme, že známe pravděpodobnost. Nezbývá než navrhnout pokus tak, aby hledaná veličina byla vyjádřitelná jako pravděpodobnost. Ta pak bude předmětem zkoumání daného experimentu, a proto musíme opět použít Bayesův vztah. Pravděpodobnost je však spojitou veličinou a musíme tudíž nahradit apriorní i aposteriorní pravděpodobnosti hustotou. Výsledný vztah se rovná:

$$f(p|\mu = k) = \frac{f(p) \cdot P(\mu = k|p)}{\int_0^1 f(p) \cdot P(\mu = k|p) dp}.$$

Zde opět musíme rozhodnout o hodnotách apriorních pravděpodobností. Použijeme Bayesův postulát a tedy každá hodnota p bude stejně možná, což výustí ve vztah:

$$f(p) = \begin{cases} 1 & \dots \quad pro \;\; p \in (0,1) \\ 0 & \dots \quad jinde \end{cases},$$

který po dosazení do původního vztahu jej změní na následující vzorec

$$f(p|\mu = k) = \frac{P(\mu = k|p)}{\int_0^1 P(\mu = k|p) dp}$$

Do tohoto vztahu dosadíme původní pravděpodobnosti $P(\mu = k|p)$ a po úpravách dostaneme

$$f(p|\mu = k) = \frac{(n+1)!}{k!(n-k)!} p^k (1-p)^{n-k} \qquad \text{pro } k = 0,1,2,...,n$$

Nyní jsme dostali hustotu pravděpodobnosti hledané veličiny. Z ní můžeme určit nejpravděpodobnější hodnotu této veličiny jako maximum funkce a tudíž položíme první derivaci rovnou nule. Z toho nám vyjde následující vztah pro odhad veličiny p označovaný jako $\dot{p}$:

$$f'(p|k) = 0 \qquad \Rightarrow \quad \dot{p} = \frac{k}{n}$$

Z toho vidíme, že nejlepším odhadem hledané pravděpodobnosti je relativní četnost. Částečně šlo tento výsledek očekávat už ze statistické teorie pravděpodobnosti (viz 3.1.1. Statistická definice pravděpodobnosti), kde počet pokusů bychom prohlásili za dostatečně veliký. Tímto způsobem jsme vlastně hledali *modus* rozdělení, což je nejčetnější hodnota v prováděném výběru. Odhad by však šel provést ještě dalším způsobem, kdy najdeme střední hodnotu daného rozdělení a tudíž:

$$\dot{p} = E(p|k) = \frac{(n+1)!}{k!(n-k)!} \int_0^1 p^{k+1} (1-p)^{n-k} dp = \frac{k+1}{n+2}.$$

4.3. Bodový odhad

V předcházejících kapitolách jsme si řekli, jak určit, která z alternativ je nejpravděpodobněji platná a jak odhadnout pravděpodobnost výsledku. Naznačili jsme také, že pokusy se musí koncipovat tak, aby hledané výsledky byli ve formě pravděpodobnosti. Nyní se pokusíme všechny tyto poznatky shrnout. K popisu tohoto budeme potřebovat některé pojmy z matematické statistiky.

4.3.1. Matematicko-statistický základ

Předpokládejme, že prvky vybíráme ze statistického souboru o velikosti *N*. Jej budeme nazývat *základní soubor*. Soubor prvků vybraných, tedy $n < N$ se nazývá *výběrový soubor*. Víme, že pro velkou diferenci *N-n* lze výběr s vracením a bez vracení ztotožnit. Lépe však předpokládat výběr s vracením, aby základní soubor byl vždy stejný.

Dále označíme *náhodným výběrem rozsahu n* posloupnost *n* náhodných nezávislých veličin X_i se stejným rozdělením pravděpodobnosti. Jejich číselná realizace x_i je *statistickým souborem*.

Poslední důležitý pojem je pojem *statistika*. Statistikou rozumíme každou veličinu, kterou lze vyjádřit hodnotami náhodného výběru.

4.3.2. Princip odhadu a jeho typy

Jak již jsme se zmínili, experiment by měl být stavěn tak, aby zkoumané veličiny byli závislé na pravděpodobnosti. V obecnějším případě lze říci, že hledáme závislost na nějaké charakteristice rozdělení náhodné proměnné. Jinak řečeno náhodný výběr $X_1,...,X_2$ závisí na nějakém parametru Θ. Tento parametr pak bude předmětem odhadu. Odhadem budeme rozumět funkci $g(X_1,...,X_2)$, která je funkcí proměnných náhodného výběru. Pro odhad potom používáme značení $\hat{\Theta}$.

Odhady mohou být různého typu, podle toho, jak moc se blíží k výsledku:

1. Nestranný odhad … $E(\hat{\Theta}) = \Theta \qquad \forall\Theta \in \Omega$
2. Asymptoticky nestranný odhad … $E(\hat{\Theta}) \xrightarrow{n\to\infty} \Theta \qquad \forall\Theta \in \Omega$
3. Vychýlený odhad … $\hat{\Theta}$ není nestranný.

Posledním důležitým pojmem je konzistence odhadu. Odhad je *konzistentní*, jestliže

$$P\left(\left|\hat{\Theta}_n - \Theta\right| < \varepsilon\right) \to 1 \qquad \text{pro } n \to \infty$$

Z této definice lze odvodit, že

1. Je-li odhad nestranný nebo asymptoticky nestranný a rozptyl konverguje k nule pro $n \to \infty$, pak je odhad konzistentní.
2. Má-li veličina X první 4 momenty konečné, pak následující odhady jsou konzistentní

$$E(X) \cong \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \textit{výběrový průměr}$$

$$D(X) \cong S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \qquad \textit{výběrový rozptyl}$$

4.3.3. Používané metody

Pro odhady se používá několik metod. My se zde zmíníme o odhadu metodou momentů a o odhadu metodou maximální věrohodnosti. Oboje mají společné to, že hledáme odhad hodnoty parametru Θ na němž je závislé rozdělení náhodné proměnné.

4.3.3.1. Metoda odhadu momentů

Princip této metody byl naznačen již dříve. Jedná se o odhad, kdy předpokládáme, že na Θ závisí nějaký moment. Hodnotu tohoto momentu odhadneme z výběru a pak již jen dopočítáme hodnotu příslušného odhadu $\hat{\Theta}$. Nejčastěji se bere podmínka $\hat{\Theta} = E(X)$.

4.3.3.2. Metoda maximální věrohodnosti

Tato metoda je založena na hledání *maximálně věrohodného odhadu*. Tento odhad maximalizuje tzv. *věrohodnostní funkci* $L(\Theta)$. Jedná se o funkci, která je definována jako součin hustot v bodech náhodného výběru $x_1,...,x_n$. To znamená, že hodnota odhadu, pro kterou je tato funkce maximální je nejvěrohodnější odhad.

4.3.4. Přesnost odhadu

Přesnost odhadu můžeme vyjádřit pomocí *střední kvadratické chyby*, která je definována jako

$$MSE(\hat{\Theta}) = E\left(\left[\hat{\Theta} - \Theta\right]^2\right) = D(\hat{\Theta}) + \left[E(\hat{\Theta}) - \Theta\right]^2$$

Je samozřejmé, že se snažíme chybu co nejvíce zmenšit.

Vydatnost odhadu můžeme měřit také tím, že sledujeme její rozptyl. To je vidět i z definice střední kvadratické chyby. Snažíme se tedy rozptyl minimalizovat. Za určitých, dosti obecných podmínek (viz 4.6), můžeme stanovit dolní mez rozptylu z tzv. *Rao-Cramerovy nerovnosti*:

$$D(T) \geq \frac{1}{nI(\Theta)},$$

kde $I(\Theta)$ je *Fischerova míra informace* rovna pro diskrétní resp. spojitou náhodnou veličinu:

$$I(\Theta) = E\left[\frac{\partial \ln f(X,\Theta)}{\partial \Theta}\right]^2 \qquad \text{resp.} \qquad I(\Theta) = \int_{-\infty}^{\infty}\left(\frac{\partial \ln f(X,\Theta)}{\partial \Theta}\right) f(X,\Theta)\,dx$$

4.3.5. Odlehlá pozorování

Poslední zmínkou v této kapitole bude krátké pozastavení nad *odlehlými pozorováními*. Tato pozorování jsou charakteristická tím, že jsou zatížena nějakou hrubou chybou a nespadají do rozdělení hledané náhodné veličiny. Někdy je však nutné s nimi počítat. Jak lze tuto nepříjemnost vyřešit?

Pokud jsou ve výběrovém souboru odlehlá pozorování (pro některé výběry je nutné to předpokládat), pak musíme provést nějako redukci, kdy se do statistik budou započítávat jen pozorování korektnější. To se obvykle dociluje *metodou useknutých průměrů*. Jedná se o průměry, ze kterých vynecháme nějaké procento nejmenších nebo největších pozorování. To má souvislost s kvantily náhodných veličin. Extrémním případem je medián, u kterého byly vynechány všechny postraní hodnoty, ten má ovšem značné problémy s nesymetrickými rozděleními.

4.4. Intervalový odhad

V předchozí kapitole jsme odhadovali některé veličiny bodovým odhadem a získávali jsme tak informaci o zkoumaném rozdělení. Touto metodou jsme získávali několik odhadů, ze kterých jsme vybírali ten nejlepší, nebo jeden odhad u kterého jsme byli schopni říct, že je ten nejvěrohodnější. Často však bývá výhodnější odhadovat tyto charakteristiky tím způsobem, že předem řekneme jakou chybu může mít náš odhad, čímž můžeme i korigovat obtížnost výpočtu a tak i rychlost experimentu.

Označme tentokrát odhad charakteristiky Θ jako Θ_n. Opět máme k dispozici realizace náhodné proměnné $X: x_1,...,x_n$. Také odhad se dá tentokrát pokládat za funkci těchto realizací, tj. $\Theta_n(x_1,...,x_n)$. Chceme tudíž, aby platilo $\Theta_n(x_1,...,x_n)=\Theta$, což z podstaty věci platit nikdy nebude, ale my se budeme pokoušet o minimalizaci rozdílu těchto hodnot. K tomu si zvolíme hodnotu *přípustné odchylky* ε. Potom budeme požadovat, aby platilo:

$$|\Theta_n-\Theta|<\varepsilon\ .$$

Jelikož platnost tohoto vztahu je sama o sobě náhodnou veličinou, můžeme prohlásit, že má svou pravděpodobnost. Tu potom položíme rovnu tzv. *spolehlivosti* α. Tato spolehlivost bude jistě závislá na počtu opakování a velikosti přípustné odchylky. Celkový vztah těchto veličin je následující:

$$P(|\Theta_n-\Theta|<\varepsilon)=\alpha\ .$$

Pak interval $(\Theta_n-\varepsilon,\Theta_n+\varepsilon)$ nazýváme *intervalovým odhadem na hladině spolehlivosti* α.

Jak vidíme v intervalových odhadech se používají metody z odhadů bodových. Ukážeme si to na nástinu odvození. Sledujeme-li platnost nějakého jevu při n nezávislých pozorování, je jeho četnost v náhodné proměnné μ daná binomickým rozdělením.

$$P(\mu=k|p)=\binom{n}{k}p^k(1-p)^{n-k},$$

jehož charakteristiky mají následující tvar:

$$E(\mu)=np \quad \text{a} \qquad D(\mu)=np(1-p)\ .$$

Pro relativní četnost pak získáme

$$E\left(\frac{\mu}{n}\right)=p \quad \text{a} \qquad D\left(\frac{\mu}{n}\right)=\frac{p(1-p)}{n}\ .$$

Pro další postup je nutné nejprve nadefinovat *Čebyševovu nerovnost*. Tato nerovnost platí pro náhodnou veličinu *X* s libovolným rozdělením s konečnou střední hodnotou a rozptylem a má tvar

$$P(|X-E(X)|<\varepsilon)\geq 1-\frac{D(X)}{\varepsilon^2}\ .$$

Když do této nerovnosti dosadíme, podmínky splňující, relativní četnost dostaneme vztah, který je nazýván *Bornoulliho nerovnost*:

$$P\left(\left|\frac{\mu}{n}-p\right|<\varepsilon\right)\geq 1-\frac{p(1-p)}{n\varepsilon^2}$$

Jelikož také $p(1-p)\leq\frac{1}{4}$, pak se dá psát upravený vztah

$$P\left(\left|\frac{\mu}{n}-p\right|<\varepsilon\right)\geq 1-\frac{1}{4n\varepsilon^2}$$

a tvrdit také, že

$$\lim_{n\to\infty}P\left(\left|\frac{\mu}{n}-p\right|<\varepsilon\right)=1\ .$$

Z těchto odvození nakonec vyplývá Bernoulliho limitní věta ve tvaru

$$1-\frac{1}{4n\varepsilon^2}\geq\alpha$$

Dostali jsme se teď k cílovému vztahu, který nám umožňuje určit počet pokusů nutných k dosažení zadané spolehlivosti a přípustné odchylky:

$$n \geq \frac{1}{4n\varepsilon^2(1-\alpha)}.$$

Je-li interval symetrický, můžeme využít normálního rozdělení a centrální limitní věty (viz 3.6. Centrální limitní věta). Pak dostaneme vztah

$$\lim_{n\to\infty} P\left(\left|\Theta_n - \Theta\right| < \frac{\sigma_n}{\sqrt{n}} t_\alpha\right) = 2\Phi(t_\alpha) - 1,$$

u kterého můžeme využít pro výpočet α tabelovaných hodnot normálního rozdělení.

4.5. Redukce rozptylu

Metody odhadu hodnot, které jsme popsali jsou ve svých principech podobné. Provádíme experiment a z něj vyšlé hodnoty umisťujeme do nějakého vzorce, z nějž usuzujeme na hodnotě odhadované charakteristiky.

Tento popsaný způsob by se dal nazvat *naivní simulací*. Nesnaží se totiž během experimentu zlepšovat svůj odhad jinými dostupnými informacemi. Princip je jednoduchý. Nahradíme-li v jakékoliv fázi odhadovanou hodnotu přesnou hodnotou, pak snížíme rozptyl výsledku. Tuto přesnou hodnotu můžeme získat například z paralelního analytického výpočtu. Nebo z jiných dostupných zdrojů (záleží na charakteru úlohy).

4.5.1. Metoda antitetických veličin

Tato metoda využívá pro získání charakteristiky toho, že k sobě připojíme dva odhady, které jsou záporně korelovány. Když budeme mít odhad $\dot{\Theta}'$ a odhad $\dot{\Theta}''$, pak výsledný odhad bude mít hodnotu

$$\dot{\Theta} = \frac{\dot{\Theta}' + \dot{\Theta}''}{2}.$$

Pro jejich rozptyl bude platit

$$D(\dot{\Theta}) = D\left(\frac{\dot{\Theta}' + \dot{\Theta}''}{2}\right) = \frac{1}{4} D(\dot{\Theta}' + \dot{\Theta}'') = \frac{1}{4} D(\dot{\Theta}') + \frac{1}{4} D(\dot{\Theta}'') + \frac{1}{2}\operatorname{cov}(\dot{\Theta}', \dot{\Theta}'').$$

Potom při $\operatorname{cov}(\dot{\Theta}', \dot{\Theta}'') < 0$ bude platit, že $D(\dot{\Theta}) < D(\dot{\Theta}')$.

4.5.2. Metoda stratifikovaných výběrů

V tomto přístupu se snažíme opravit chybu vzniklou tím, že odhadujeme střední hodnotu pro celou funkci a tudíž zákonitě budeme mít větší rozptyl nežli bychom měli, kdybychom tuto hodnotu spočetli pro jednotlivé části rozdělení.

Princip této metody je tedy založen na rozdělení množiny Ω na jednotlivé intervaly, ve kterých budeme odhadovat. Volba intervalu silně závisí na charakteru úlohy. Můžeme použít například stejnou velikost. Také je ovšem výhodné velikost přizpůsobit tomu, jak se od sebe střední hodnoty v jednotlivých intervalech liší. Celkový odhad pak získáme sečtením jednotlivých odhadů vynásobených vahami udávajícími velikost intervalu. Tedy, pokud velikost intervalu pro i-tý člen je k_i, pak

$$\bar{X}_i = \sum_i k_i \bar{X}_i \qquad \text{s tím, že } \sum_i k_i = 1.$$

4.5.3. Metoda řídících veličin

Poslední námi popisovaná metoda je založena na znalosti veličiny *C* (jiné krom zkoumané *X*) u které známe odhadovanou charakteristiku, např. střední hodnotu. Střední hodnotu zkoumané veličiny můžeme vyjádřit statistikou

$$X(b) = X - b(C - E(C)),$$

přičemž platí $E(X(b)) = E(X)$, takže jde o nestranný odhad a pro rozptyl platí

$$D[X(b)] = D(X) - 2b\,\mathrm{cov}(X, C) + b^2 D(C).$$

Minima rozptylu dosáhneme při volbě parametru b tímto způsobem:

$$b = \frac{\mathrm{cov}(Y, C)}{D(C)}.$$

Pro reálné aplikace budeme počítat tuto hodnotu z generovaných dvojic (X_i, C_i) jako

$$\overline{b} = \frac{\sum_i (X_i - \overline{X})(C_i - \overline{C})}{\sum_i (C_i - \overline{C})^2}.$$

4.6. Testování hypotéz

Další možností jak můžeme zpracovávat data z experimentu je statistické *testování hypotéz*. Při této aktivitě budeme testovat předpoklady o rozdělení náhodných veličin. Můžeme tedy nejprve odhadnout z jednoho pokusu nějaký předpoklad a ten pak na dalším, popř. sérii dalších ověřit.

Při tomto testování stojí proti sobě vždy *testovaná hypotéza* , značíme H_0 ,a *alternativní hypotéza*, značíme H_1. Hypotézami nejčastěji testujeme nějakou charakteristiku náhodné veličiny. U ní se pak ptáme, zda-li je rovna nějaké hodnotě, popř. zda-li spadá do nějakého intervalu. Předpokládejme, že testovací kritérium je $H_0 : \Theta = \Theta_0$, pak můžeme provádět tyto testy:

1. *Oboustranný test* … Alternativní hypotéza je $H_1 : \Theta \neq \Theta_0$, neboli testujeme, jestli $\Theta > \Theta_0 \quad \vee \quad \Theta > \Theta_0$.
2. *Jednostranný test* … Alternativní hypotéza je $H_1 : \Theta > \Theta_0$ nebo $H_1 : \Theta < \Theta_0$ a to v případech, kdy můžeme některou z variant z nějakého důvodu vyloučit.

K testování použijeme charakteristiku T, nazývanou *testovací kritérium*. Její obor hodnot pak rozdělíme na *obor přijetí testovaného kritéria* Ψ a *kritický obor* Ξ. Potom experimentem ověříme testovací kritérium a pokud padne do oboru přijetí, pak přijmeme testovací hypotézu. V opačném případě přijímáme hypotézu alternativní.

Při takovémto testu se můžeme dopustit chyb dvojího druhu. Označujeme je jako

- *Chyba I. druhu* … Chybné zamítnutí testované hypotézy.
- *Chyba II. Druhu* … Chybné přijetí testované hypotézy.

Pro celý proces testování jsou důležité zejména pravděpodobnosti vzniku těchto chyb, neboť nám umožňují testovat se zadanou přesností, což jsme uvítali už v případě intervalových testů.

Pravděpodobnost chyby I. druhu se nazývá *hladinou významnosti* a vyjadřuje:

$$\alpha = P(T \in \Xi | H_0).$$

Pravděpodobnost chyby druhého druhu má odpovídající tvar

$$\beta = P(T \in \Psi | H_1),$$

avšak častěji se vyjadřuje ve svém doplňku k jedné, který se nazývá *síla testu*:

$$1-\beta = P\left(T \in \Xi \middle| H_0\right).$$

Při testování se snažíme minimalizovat pravděpodobnosti chyb. Tyto jsou však na sobě závislé a snížení jedné vede ke zvýšení druhé., proto se musí nejprve volit jedna z dalších výsledků posléze vybrat ten s minimální hodnotou té druhé. Dále je třeba říct, že síla testu může záviset na zkoumaném parametru Θ. Tuto funkční závislost pak označujeme jako *silofunkci*.

Důležitou součástí testování hypotéz je výběr kritického oboru. Při tom nám může pomoct *Neymanova-Pearsova věta*, která říká, že kritický obor je množina splňující vztah:

$$\frac{L\left(x, H_0\right)}{L\left(x, H_1\right)} \leq k_\alpha .$$

Kde konstantou k_α zabezpečujeme splnění pravděpodobnosti chyby I. druhu, tj.

$$P\left(\frac{L\left(x, H_0\right)}{L\left(x, H_1\right)} \leq k_\alpha\right) = \alpha .$$

Zdroje a literatura

[L1] *R. Hušek, J. Lauber*: **Simulační modely**. STNL/ALFA, 1987

[L2] *S. Racek, M.Roubín*: **Pravděpodobnostní modely počítačů**. ZČU 1996

[L3] *F. Štulajter*: **Odhady v náhodných procesech**. STNL/ALFA 1990

[L4] *F. Fabian, Z. Kluiber*: **Metoda Monte Carlo a možnosti jejího uplatnění** PROSPECTRUM 1998

[L5] *J. Reif*: **Metody matematické statistiky**. ZČU 2002

[L6] *P.Hebák, J.Kahounová*: **Počet pravděpodobnosti v příkladech**. STNL 1988

[L7] *J. Reif, Z. Kobeda*: **Úvod do pravděpodobnosti a spolehlivosti**. ZČU 2000

[L8] *L. Cyhelský*: **Úvod do teorie statistiky** STNL/ALFA 1981

[L9] *J. Brousek, Z. Ryjáček*: **Sbírka řešených příkladů z počtu pravděpodobnosti** ZČU 1999

[L10] *F. Tůma*: **Kybernetika** ZČU 1996

[L11] *J.Hátle-J.Likeš*: **Základy počtu pravděpodobnosti a matematické statistiky** STNL 1972

[L12] *M. Friesl*: **Pravděpodobnost a statistika hypertextově: http://home.zcu.cz/~friesl/hpsb/**

[L13] www.zcu.cz 2003

Rejstřík