S t a t i s t i k a

Ivan Nagy, Pavla Pecherková

FD ČVUT, Praha

Obsah

1 Počet pravděpodobnosti

O pravděpodobnosti, jako vědecké disciplíně, lze v Evropě hovořit až od 16. století našeho letopočtu. Do té doby se nikdo vážněji studiem počtu pravděpodobnosti nezabýval, což můžeme vysvětlit nepotřebností matematického vysvětlení náhodných jevů a jejich chování. 16. století je bráno jako století rozvoje v oblasti přírodních věd a právě v těchto vědách se již objevila potřebnost počtu pravděpodobnosti. O rozvoj pravděpodobnosti se zasloužili takoví myslitelé té doby jako byl Blaise Pascal, Pierre de Fermat či Christian Huygens.

Bohužel jako většina vědeckých teorií, i počet pravděpodobnosti nemá základ v žádné „vznešené“ myšlence, ale v potřebě rize praktické, a to v sázení. Typickým příkladem, který je oblíbeným problémem až do dnešních dob, je sázení na určitý součet ok na několika kostkách. Čím větší je počet kombinací hodů, tím menší bude výhra. V té době se však nejednalo pouze o peníze, ale i o životy. Z historie je známo několik případů, kdy u soudu nebyly řádné důkazy a tak soudci nechali rozhodnout vyšší síly. V případě, že obviněný hodí na všech třech kostkách jedničku, bude osvobozen jako nevinný v ostatních případech bude rozhodnuto o jeho vině a on bude popraven. Naštěstí pro obžalované, většina z nich neměla ani tušení o tom, jak malou šanci na osvobození dostali a viděli v tomto záchranu.

Cílem této kapitoly je popsat základní pojmy počtu pravděpodobnosti jako je náhodný pokus, náhodné jevy a pravděpodobnost, včetně pravidel, jak se s nimi počítá. Setkáme se tedy s pojmy a postupy, které objevili a dokázali zakladatelé počtu pravděpodobnosti a rozšířili jejich následovníci, ovšem již s ohledem na potřeby a problémy dnešní doby. Ačkoliv jak již bylo řečeno výše, některá problematika či použití je i přes několik staletí stejná či velmi podobná.

1.1 Náhodný pokus

V přírodních i společenských vědách je možno pospat nesčetně činností, které při dodržení stejných podmínek dávají stejné výsledky (na severní polokouli vždy vodní vír rotuje na levou stranu) . Na druhou stranu je kolem nás mnoho dějů a činností, u kterých není možné jednoznačně říci výsledek (brzdná dráha vozidla z rychlosti 40 km/h na 0 km/h). Ději, který můžeme opakovat neomezeně a přitom dává různé výsledky, říkáme náhodný pokus. Lze tedy říci, že náhodným pokusem rozumíme opakovatelnou činnost prováděnou za stejných podmínek, jejíž výsledek je nejistý a závisí na náhodě.

1.1.1 Náhodný pokus

Náhodný pokus je určitý experiment, který i za relativně stálých podmínek dává různé výsledky.

Poznámka

Při definici experimentu jsou důležité podmínky, které experiment provázejí. Většinou jsou tyto podmínky stanoveny implicitně formulací experimentu. Např. pro experiment “hod mincí”

se automaticky uvažuje férová mince (tj. stejná pravděpodobnost pro obě strany) a nemožnost výsledku "hrana" a podobných. Další dodané podmínky většinou změní celý pokus: např. experiment "hod kostkou" s podmínkou "nepadlo liché číslo" definuje experiment hodu kostkou ale s výsledky 2, 4, 6.

Náhodný pokus je základním pojmem pravděpodobnosti i statistiky. V pravděpodobnosti na sebe náhodný pokus bere nejrůznější podoby (házení mincí, kostkou, tažení korálků apod.) Ve statistice má podobu ankety nebo sběru (měření) dat.

Příklady

1. Hod mincí, hod jednou nebo dvěma kostkami, odbočení nebo neodbočení auta v křižovatce.
2. Doba čekání na tramvaj, měření rozměru vyráběné součástky, doba životnosti přístroje.

1.2 Náhodný jev

Naším cílem je nalézt popis náhodného pokusu. Vzhledem k náhodné povaze pokusu, tento popis nemůže spočívat v přesné předpovědi výsledku, který nastane. Jediné, co lze udělat, je vymezit všechny možné výsledky a určit s jakou četností se objevují, a tedy, které výsledky můžeme očekávat více a které méně. K takovému popisu náhodného pokusu směřuje naše další snažení.

Množinu všech možných výsledků náhodného pokusu nazveme **základní prostor** a označíme ji Ω.

Příklady

1. Základní prostor pokusu hod mincí je $\Omega = \{$"rub", "líc"$\}$.
2. Výsledkem pokusu hod dvěma kostkami je uspořádaná dvojice $[i, j]$, kde $i, j \in \{1, 2, \cdots .6\}$. Tedy $\Omega = \{[1,1], [1,2], \cdots, [1,6], [2,1], [2,2], \cdots, [6,6]\}$.
3. Uvažujme křižovatku ve tvaru T, do které zespoda přijíždějí automobily a odbočují buď vpravo (P) nebo vlevo (L). Potom $\Omega = \{P, L\}$.
4. Vezmeme-li jako výsledek náhodného pokusu dobu čekání na tramvaj s pětiminutovým intervalem po náhodném příchodu na zastávku, bude množinou všech výsledků interval $(0, 5)$ na reálné ose. Tedy $\Omega = \{x \in R, x > 0 \wedge x < 5\}$.

Jak jsme již řekli, popis náhodného pokusu není dán jen výčtem možných výsledků, ale také zjištěnou četností jejich výskytu. Pojem četnost výskytu zahrneme pod pojem pravděpodobnost výskytu, který budeme definovat později. V této chvíli nás zajímá, pro které objekty budeme pravděpodobnost definovat. Určitě nás bude zajímat pravděpodobnost všech výsledků náhodného pokusu. To ale není všechno. Představme si, že házíme kostkou. Někdo

přijde, a zeptá se: jaká je pravděpodobnost, že padne šestka, $\cdots$ padne sudé číslo, $\cdots$ padne záporné číslo atd. Chceme tedy znát pravděpodobnostní odpověď nejen na dotaz o výsledku náhodného pokusu, ale na jakýkoli dotaz, týkající se pokusu. Proto, než přistoupíme k definici pravděpodobnosti, uvedeme následující definici náhodného jevu.

1.2.1 Náhodný jev, jevové pole

Náhodný jev je libovolný výrok o výsledku náhodného pokusu. Množina všech jevů se nazývá jevové pole a značí se A.

Komentář k definici

- Náhodný jev jsme definovali jako výrok. To ale není jednoznačné. Např. výrok „padne 5 nebo 6“, „padne větší než 4“ nebo „nepadne 1 až 4“ mají stejný význam a bude jim zřejmě přiřazena stejná pravděpodobnost. Proto budeme dále mluvit o **množinové reprezentaci** jevu.
- Množinovou reprezentací náhodného jevu nazveme množinu všech výsledků pokusu, které příslušný výrok splňují. Tedy např. množinovou reprezentací náhodného jevu „padne sudé číslo“ bude množina $\{2,\, 4,\, 6\}$.
- Řekneme, že **jev nastoupil** (nastal), jestliže výsledek pokusu tento jev splňuje (patří do jeho množinové reprezentace). Např. padne-li na kostce 6, nastal jev „padne sudé číslo“ protože $6 \in \{2, 4, 6\}$.
- Jevové pole A v množinové reprezentaci budeme chápat jako **množinu všech podmnožin** základního prostoru Ω.

Příklad

Zavedené pojmy budeme ilustrovat na jednoduchém příkladě hod mincí s výsledky R a L.

Základní prostor (množina všech výsledků)

$$\Omega = \{R,\, L\}$$

Příklady jevů a jejich množinová reprezentace

„padne rub“	$\{R\}$
„nepadne rub“ nebo „padne líc“	$\{L\}$
„padne cokoli“ nebo „nepadne hrana“	$\{R,\, L\}$
„padne hrana“, „padne 5“ nebo „padne sudé číslo“	$\emptyset$

Jevové pole (množina všech jevů - v množinové reprezentaci)

$$A = \{\emptyset,\, \{R\},\, \{L\},\, \{R,\, L\}\}$$

Další příklady

Rozmyslet, co bude základní prostor a jevové pole pro následující náhodné pokusy:
Hod kostkou (s výsledky 1,2,$\cdots$, 6), odbočení auta (s výsledky odbočí doprava, doleva), hod dvěma kostkami (s výsledky padlo na první, padlo na druhé), čekání na tramvaj (s výsledky z $(0, 5) \subset R$).

Poznámka

Jevové pole nemusí být vždy celá množina podmnožin. Stačí, když tvoří σ-algebru, což je struktura, uzavřená na doplňky a sjednocení.

1.3 Počítání s jevy

Skutečnost, že každý jev má svou množinovou reprezentaci, nám umožňuje zacházet s jevy jako s množinami a využít pro ně množinovou symboliku a množinové operace. Uvažujme tedy jevy J, J_1, J_2,... , J_n a budeme definovat následující operace.

Jev opačný:

$$\overline{J} \;=\; \Omega - J$$

V případě, že nenastane jev J lze tvrdit, že nastal jev opačný $\overline{J}$ k jevu J. Jev opačný $\overline{J}$ je doplňkem jevu J a ve sjednocení tvoří celý základní prostor Ω.

Příklad

Na kostce jev opačný k jevu „padne sudé číslo“ - {2, 4, 6} je jev „padne liché číslo“ - {1, 3, 5}.

Průnik a sjednocení:

$$\begin{aligned} J_p &= J_1 \cap J_2 \;\; ... \;\; \text{průnik} \\ J_s &= J_1 \cup J_2 \;\; ... \;\; \text{sjednocení} \end{aligned}$$

Jev J_p nastane, nastanou-li současně jevy J_1 a J_2. Jev J_p tedy nazveme průnikem jevu J_1 a J_2.

Jev J_s nastane, nastane-li alespoň jeden z jevů J_1 a J_2. Jev J_s tedy nazveme sjednocením jevu J_1 a J_2.

Příklad

Pro hod kostkou definujme jevy J_1: „padne menší než 4“ - $\{1, 2, 3, 4\}$ a J_2: „padne sudé číslo“ -$\{2, 4, 6\}$. Potom jejich průnik je $J_p = \{2, 4\}$ a sjednocení $J_s = \{1, 2, 3, 4, 6\}$.

Neslučitelnost:

$$J_1 \cap J_2 = \emptyset$$

Jevy J_1 a J_2 nazveme neslučitelnými, je-li jejich průnikem prázdná množina. Tedy jevy J_1 a J_2 nemohou nastoupit současně (jsou tvořeny disjunktními množinami).

Příklad:

Jevy J_1: „padne sudé“ a J_2: „padne liché“ jsou neslučitelné, protože $J_1 \cap J_2 = \{2, 4, 6\} \cap \{1, 3, 5\} = \emptyset$.

Podmíněnost:

$$J_c = J_1|J_2$$

J_c je jev J_1 při znalosti o nastoupení jevu J_2. Podmíněný jev je určen sledovaným jevem J_1 při znalosti toho, že nastoupil jev (okolnost) J_2, která sledovaný jev doprovází.

Příklad

Při pokusu hod kostkou definujeme jevy J_1: „padne menší než 5“ a J_2: „padne sudé“. Jaká bude množinová reprezentace podmíněného jevu $J_c = J_1|J_2$: „padne menší než 5, když víme, že padlo sudé“?
Výsledky, na které se ptáme, jsou $\{1, 2, 3, 4\}$. Z podmínky ale víme, že padlo sudé číslo. Podmíněný jev tedy bude dán výsledky $\{2, 4\}$, což je průnik jevů J_1 a J_2.

Poznámka

Není-li dáno jinak, jsou podmínky určeny přímo definicí náhodného pokusu (např. hod kostkou - předpokládáme férovou kostku).

1.4 Pravděpodobnost

Nyní se dostáváme k zavedení ústředního pojmu pravděpodobnost. Nejprve tento pojem zavedeme axiomaticky. Axiomatická definic neříká jak pravděpodobnost počítat, pouze vymezuje, co pravděpodobnost je. Později ukážeme další definice (klasickou a statistickou), které dávají návod pro výpočet pravděpodobnosti.

1.4.1 Axiomatická pravděpodobnost

Pravděpodobnost je nezáporná, normovaná aditivní funkce definovaná na množině jevů (jevovém poli A)

Komentář k definici

Pravděpodobnost je funkce z jevového pole A do množiny reálných čísel R

$$P: A \to R$$

která je:

1. Nezáporná, tj. $P(J) \geq 0 \ \forall J \in A$.
2. Normovaná, tj. $P(\Omega) = 1$.
3. σ-aditivní, tj. pro všechny neslučitelné jevy $J_1, J_2, J_3, \cdots$ platí $P(\cup J_i) = \sum P(J_i)$

Poznámky

Na pravděpodobnost lze pohlížet jako na plochu příslušného jevu. → množinová interpretace.

Z axiomů pravděpodobnosti plyne, že pravděpodobnost je vždy menší než jedna. Důkaz je následující:

$$\Omega = J \cup \bar{J} \quad a \quad J \cap \bar{J} = \emptyset.$$

Je tedy $1 = P(\Omega) = P\left(J \cup \bar{J}\right) = P(J) + P\left(\bar{J}\right)$ *odkud* $P(J) = 1 - P\left(\bar{J}\right)$. *Protože podle prvního axiomu je* $P\left(\bar{J}\right) \geq 0$ *je* $P(J) \leq 1$.

Dále lze na základě axiomů pravděpodobnosti dokázat, že pravděpodobnost prázdné množiny je vždy nula.

Důkaz: Platí $J \cup \emptyset = J$ *a* $J \cap \emptyset = 0$ *pro libovolný jev* J. *Jev* J *a* $\emptyset$ *jsou tedy neslučitelné jevy. Proto platí* $P(J \cup \emptyset) = P(J) + P(J) = P(J)$ *podle prvního vztahu. Odtud plyne dokazovaná vlastnost.*

Trojice {Základní prostor, Jevové pole, Pravděpodobnost} se nazývá **Pravděpodobnostní prostor**. Ten je úplným pravděpodobnostním popisem náhodného pokusu.

Příklad

Ukážeme pravděpodobnostní prostor pro náhodný pokus narození dítěte s výsledky „děvče“ D a „chlapec“ CH. Přitom dlouhodobé statistiky ukazují, že $P(D) = 0.48$ a $P(CH) = 0.52$.

Základní prostor	$\{D, CH\}$
Jevové pole	$\{\emptyset, \{D\}, \{CH\}, \{D, CH\}\}$
Pravděpodobnost	$P(\emptyset) = 0,\ P(\{D\}) = 0.48,\ P(\{CH\}) = 0.52,\ P(\{D, CH\}) = 1$

1.5 Počítání s pravděpodobností

Uvedená axiomatická definice, umožňuje formulovat následující pravidla pro počítání s pravděpodobnostmi.

Pravděpodobnost doplňku

$$P\left(\bar{J}\right) = 1 - P\left(J\right)$$

Důkaz: Platí $J \cup \bar{J} = \Omega$ a $J \cap \bar{J} = \emptyset$. Odtud $1 = P\left(\Omega\right) = P\left(J \cup \bar{J}\right) = P\left(J\right) + P\left(\bar{J}\right)$. A tedy platí dokazovaný vzorec.

Příklad

Pro pokus hod kostkou je doplňkem jevu J_1: „padne sudé číslo“ jev $J_2 = \bar{J_1}$: „padne liché číslo“. Platí $P\left(J_1\right) = \frac{1}{2}$, $P\left(J_2\right) = \frac{1}{2}$. A tedy $P\left(\bar{J_1}\right) = 1 - P\left(J_1\right)$.

Pravděpodobnost průniku

$$P\left(J_1 \cap J_2\right) \tag{1}$$

je pravděpodobnost výsledků, které jsou společné oběma jevům.

Příklad

Při hodu kostkou definujeme jevy J_1: „padne sudé číslo“ a J_2: „padne menší než 5“. Průnik těchto jevů má množinovou reprezentaci $\{2, 4\}$ a jeho pravděpodobnost je $P\left(\{2, 4\}\right) = \frac{1}{3}$.

Pravděpodobnost sjednocení

$$P\left(J_1 \cup J_2\right) = P\left(J_1\right) + P\left(J_2\right) - P\left(J_1 \cap J_2\right).$$

Důkaz: Platí $J_1 \cup J_2 = \left(J_1 - J_2\right) \cup \left(J_1 \cap J_2\right) \cup \left(J_2 - J_1\right)$ přičemž jevy na pravé straně rovnosti jsou neslučitelné a proto $P\left(J_1 \cup J_2\right) = P\left(J_1 - J_2\right) + P\left(J_1 \cap J_2\right) + P\left(J_2 - J_1\right)$. Přitom $J_1 = \left(J_1 - J_2\right) \cup \left(J_1 \cap J_2\right)$ a $J_2 = \left(J_1 \cap J_2\right) \cup \left(J_2 - J_1\right)$. Je tedy $P\left(J_1\right) + P\left(J_2\right) = P\left(J_1 - J_2\right) + P\left(J_1 \cap J_2\right) + P\left(J_1 \cap J_2\right) + P\left(J_2 - J_1\right) = P\left(J_1 \cup J_2\right) + P\left(J_1 \cap J_2.\right)$ A to je dokazovaný vzorec.

Příklad

Při hodu kostkou definujeme jevy J_1: „padne liché číslo“ a J_2: „nepadne 3 a 6“. Potom $J_1 \cap J_2 = \{1, 5\}$ a $J_1 \cup J_2 = \{1, 2, 3, 4, 5\}$. Přitom $P\left(J_1\right) = \frac{1}{2}$, $P\left(J_2\right) = \frac{2}{3}$ a $P\left(J_1 \cap J_2\right) = \frac{1}{3}$. Pravděpodobnost sjednocení je

$$P\left(J_1 \cup J_2\right) = P\left(J_1\right) + P\left(J_2\right) - P\left(J_1 \cap J_2\right) = \frac{1}{2} + \frac{2}{3} - \frac{1}{3} = \frac{5}{6} = P\left(\{1, 2, 3, 4, 5\}\right).$$

Pravděpodobnosti pro neslučitelné jevy

$$P(J_1 \cup J_2) = P(J_1) + P(J_2)$$

Důkaz: Tento vzorec plyne přímo z předchozího vzorce, protože pro neslučitelné jevy J_1 a J_2 platí $P(J_1 \cap J_2) = 0$.

Příklad

Typicky neslučitelné jevy při hodu kostkou jsou jevy J_1: „padne sudé číslo“ a J_2: „padne liché číslo“, oba s pravděpodobností $\frac{1}{2}$. Sjednocením je celý základní prostor Ω, který má pravděpodobnost $P(\Omega) = 1$. Platí tedy $P(J_1 \cup J_2) = P(J_1) + P(J_2) = \frac{1}{2} + \frac{1}{2} = 1$.

Podmíněná pravděpodobnost

Velmi důležitým pojmem v teorii pravděpodobnosti je podmíněná pravděpodobnost. Ta se vztahuje k podmíněnému jevu (viz strana 9). Jedná se o pravděpodobnost určitého jevu, jestliže víme, že nastal jiný jev. Například, při hodu kostkou sledujeme jev J_1: „padne sudé číslo“ - $\{2, 4, 6\}$. Po provedení pokusu, jehož výsledek nevidíme, se dozvíme, že nastal jev J_2: „padne menší než 4“ - $\{1, 2, 3\}$. Na výsledky 4, 5 a 6 můžeme v tuto chvíli zapomenout a tak množina možných výsledků - nový základní prostor - je nyní dána množinou $J_2 = \{1, 2, 3\}$. Množina příznivých výsledků je dána množinou J_1 ovšem jen v rámci nového základního prostoru, a tedy je $J_1 \cap J_2 = \{2\}$. Touto úvahou je motivována následující definice

1.5.1 Podmíněná pravděpodobnost

Nechť J_1 je sledovaný jev a J_2 je pozorovaný jev pro který je $P(J_2) > 0$. Potom pravděpodobnost podmíněného jevu $P(J_1|J_2)$ je dána vzorcem

$$P(J_1|J_2) = \frac{P(J_1 \cap J_2)}{P(J_2)}$$

Příklad

Vezmeme příklad z úvodu k podmíněné pravděpodobnosti. Házíme kostkou a sledujeme jev J_1: „padne sudé číslo“. Po provedení pokusu se dozvíme, že nastal jev J_2: „padne menší než 4“. Jaká je pravděpodobnost $P(J_1|J_2)$?

Průnik jevů $J_1 \cap J_2$ je $\{2\}$. Pravděpodobnost každého jednotlivého výsledku je $\frac{1}{6}$ (protože výsledků je 6 a jsou stejně pravděpodobné). Pravděpodobnost každého jevu je dána součtem pravděpodobností výsledků, které obsahuje (výsledky jsou neslučitelné a platí 3. axiom pravděpodobnosti). Dostáváme tedy $P(J_1 \cap J_2) = \frac{1}{6}$ a $P(J_2) = 3 \cdot \frac{1}{6} = \frac{1}{2}$. Pravděpodobnost podmíněného jevu podle definice tedy bude $P(J_1|J_2) = \frac{1}{6}/\frac{1}{2} = \frac{1}{3}$.

Nezávislost

Velice důležitou vlastností jevů je jejich nezávislost. Volně řečeno, dva jevy jsou nezávislé, jestliže spolu nijak nesouvisí. A to například v tom smyslu, že když víme něco o jednom z nich, nijak to nezmění naši vědomost o druhém - a naopak.

Význam nezávislosti spočívá jednak v tom, že výpočty pro nezávislé jevy se značně zjednoduší (uvidíme dále), ale také v tom, že nezávislost znamená reprezentativnost. Např. jestliže chceme poznávat náhodný pokus podle jeho výsledků, musíme je vybírat nezávisle. Jinak by se mohlo stát, že budeme vybírat stále stejné výsledky a k některým se vůbec nedostaneme.

Definice nezávislosti je následující

1.5.2 Nezávislost

Dva jevy J_1 a J_2 jsou nezávislé, jestliže platí

$P(J_1|J_2) = P(J_1)$ nebo, což je totéž $P(J_2|J_1) = P(J_2)$

Podmínka nezávislosti, ekvivalentní s definicí, je
$P(J_1 \cap J_2) = P(J_1)\,P(J_2)$

Důkaz podmínky: Podle definice je $P(J_1|J_2) = P(J_1 \cap J_2)/P(J_2)$. Pro nezávislé jevy ale platí $P(J_1|J_2) = P(J_1)$. Porovnáním podmíněných pravděpodobností dostáváme dokazovanou podmínku nezávislosti.

Příklad

Uvedeme opět příklad s kostkou, na které sledujeme dva jevy $J_1 = \{2,4,6\}$ (sudé) a $J_2 = \{1,2,3\}$ (menší než 4). Určíme, zda jsou tyto jevy nezávislé.

Pro řešení použijeme podmínku nezávislosti. Protože $J_1 \cap J_2 = \{2\}$ je $P(J_1 \cap J_2) = \frac{1}{6}$ (6 stejně pravděpodobných výsledků, každý má pravděpodobnost $\frac{1}{6}$) a dále $P(J_1) = P(\{2,4,6\}) = \frac{1}{2}$ a $P(J_2) = P(\{1,2,3\}) = \frac{1}{2}$. Protože $P(J_1 \cap J_2) \neq P(J_1)\,P(J_2)$ lze konstatovat, že jevy „padne sudé“ a „padne menší než 4“ jsou závislé.

Uvažujme dále stejný příklad, jev J_1 je opět „padne sudé číslo“, ale jev J_2 bude nyní „padne menší než 5“ - $\{1,2,3,4\}$. Určete opět závislost či nezávislost obou jevů.

Budeme postupovat obdobně jako v předchozí části příkladu. $P(J_1 \cap J_2) = P(\{2,4\}) = \frac{1}{3}$, $P(J_1) = \frac{1}{2}$ a $P(J_2) = \frac{2}{3}$. Tentokrát je $P(J_1 \cap J_2) = P(J_1)\,P(J_2)$ a oba jevy „padne sudé číslo“ a „padne menší než 5“ jsou nezávislé.

Vysvětlení:

Všimněme si rozdílu mezi oběma případy. Základní prostor pro hod kostkou je $\Omega = \{1,2,3,4,5,6\}$. Sudá a lichá čísla jsou v něm zastoupena v poměru 1:1. Omezíme-li základní prostor na jev $\{1,2,3\}$ (jako v prvém případě), poměr 1:1 se rozváží ve prospěch sudých čísel. Znalost jevu „padne menší než 4“ tedy přinese informaci o sledovaném výsledku, kterým je jev „padne

sudé číslo“. V druhém případě se základní prostor omezí na množinu $\{1, 2, 3, 4\}$. Zde je ale stejně tak, jako v samotném základním prostoru poměr mezi sudými a lichými 1:1. Žádná nová informace o možném hodu sudého čísla nepřichází. Oba jevy jsou nezávislé.

Úplná pravděpodobnost

Uvažujme jev J a jevy $K_1, K_2, \cdots, K_n$. Sledujeme jev J, jevy K_i, $i = 1, 2, \cdots, n$ jev J doprovází. Jevy K_i můžeme chápat jako okolnosti za kterých se jev J realizuje. Například, jev J znamená „půjdu na procházku“ a jevy K_i představují počasí: K_1 - „slunce“, K_2 - „zataženo“, K_3 - „déšť“. Je středa a já plánuji sobotu odpoledne. Vím (znám pravděpodobnosti), jak se zachovám za určitého počasí, ale teď ještě nevím, jak přesně bude. Vím jen, že touto dobou je v 50% slunečno, ve 30% zataženo a ve 20% prší. Vzorec pro úplnou pravděpodobnost dává odpověď na základní otázku - jaká je pravděpodobnost, že se bude realizovat určitá hodnota jevu J, a to pro libovolnou okolnost K_i, tedy, že půjdu na procházku aniž bych věděl, jaké bude počasí.

Vzorec (úplná pravděpodobnost)

Uvažujme jev J a a navzájem neslučitelné jevy K_1, $K_2, \cdots$, K_n pro které platí $P(K_i) > 0$ a dále $\sum_{i=1}^{n} P(K_i) = 1$ (jevy K_i tvoří úplný rozklad svého základního prostoru). Potom platí

$$P(J) = \sum_{i=1}^{n} P(J|K_i)\, P(K_i)$$

Důkaz: Z vlastností jevů K_i plyne, že $J = \cup_i (J \cap K_i)$ (kdo nevěří, ať si nakreslí množinový diagram). Přitom jevy $J \cap K_i$, $i = 1, 2, \cdots, n$ jsou neslučitelné. Aplikací pravděpodobnosti na původní rovnost dostáváme $P(J) = P(\cup_i (J \cap K_i)) = \sum_i P(J \cap K_i) = \sum_i P(J|K_i)\, P(K_i)$, kde v poslední rovnosti jsme využili vzorec pro definici podmíněné pravděpodobnosti. První a poslední člen v poslední posloupnosti rovností tvoří úplnou pravděpodobnost.

Bayesův vzorec

Bayesův vzorec se objevuje v podobných souvislostech jako úplná pravděpodobnost. Opět uvažujeme jev J a okolnosti K_i, $i = 1, 2, \cdots, n$. Základní otázkou v úloze řešené pomocí Bayesova vzorce je: zjistili jsme, že jev J nastal. Jaká je pravděpodobnost, že jej doprovázela okolnost K_i? Počítáme tedy obrácenou podmíněnou pravděpodobnost $P(K_i|J)$ zatímco za známé považujeme pravděpodobnosti $P(J|K_i)$.

Oba případy rozeznáme spolehlivě takto: zatímco úplná pravděpodobnost se počítá před provedením experimentu, Bayesův vzorec je možno počítat až po provedení experimentu (až když zjistíme, že jev J nastal).

Vzorec (Bayesův vzorec)

Uvažujme jev J a a navzájem neslučitelné jevy K_1, $K_2, \cdots$, K_n pro které platí $P(K_i) > 0$ a dále $\sum_{i=1}^{n} P(K_i) = 1$ (jevy K_i tvoří úplný rozklad svého základního prostoru). Potom platí

$$P(K_i|J) = \frac{P(J|K_i)\,P(K_i)}{\sum_{j\neq i} P(J|K_j)\,P(K_j)},$$

kde jmenovatel je roven $P(J)$ a je vyjádřen pomocí úplné pravděpodobnosti.
Důkaz: Dosadíme-li do jmenovatele za vzorec pro úplnou pravděpodobnost, dostáváme $P(K_i|J) = P(J|K_i)\,P(K_i)\,/P(J)$. Násobíme $P(J)$ a dostaneme $P(K_i|J)\,P(J) = P(J|K_i)\,P(K_i)$ a odtud $P(K_i, J) = P(J, K_i)$. Tím je Bayesův vzorec dokázán.

1.6 Výpočetní definice pravděpodobnosti

Axiomatická definice pravděpodobnosti 1.4.1 většinou neslouží přímo k výpočtu hodnot pravděpodobnosti. Pouze vymezuje, co pravděpodobnost je. K vyčíslení hodnot pravděpodobnosti slouží následující dvě definice. První z nich - klasická definice - se vztahuje přímo k procesu, který sledujeme a který se snažíme pravděpodobnostním způsobem popsat. Jde o určení kolika způsoby lze sledovaného jevu dosáhnout v poměru ke všem možným způsobům, jak experiment provést. Například, šestka na kostce padne jediným způsobem. Všech možností je šest, odtud pravděpodobnost šestky při hodu kostkou je $\frac{1}{6}$. Druhá z nich - statistická definice - se opírá o experimenty, nebo spíše o data získaná při experimentech. Jedná se o poměr počtu experimentů příznivých danému jevu a počtu provedených experimentů. Například, hodíme $100\times$ kostkou a napočítáme 15 experimentů, kdy padla šestka. Statistická pravděpodobnost tedy bude $\frac{15}{100} = 0.15$.

1.6.1 Klasická pravděpodobnost

Uvažujme náhodný pokus s konečným počtem počtem stejně pravděpodobných výsledků. Klasickou pravděpodobnost jevu definujeme jako podíl počtu možných výsledků, při kterých sledovaný jev nastoupí, a počtu všech možných výsledků.

Komentář k definici

- Protože se klasická definice vztahuje přímo k procesu, o kterém předpokládáme, že má pevnou strukturu, je výsledek klasické definice pro určitý jev vždy stejný. Vzhledem ke sledovanému procesu dívá klasická definice „přesný výsledek“. Její nevýhodou je, že pro složitější procesy je většinou její výpočet nad naše síly.

Příklady

Jaká je pravděpodobnost, že náhodně sestavené čtyřciferné číslo z číslic 0,1,2,3,4,5 bude sudé, jestliže se číslice v sestaveném čísle A) nesmí opakovat, B) mohou opakovat.

A) $V_3(5) + 2\,(V_3(5) - V_2(4)) = 60 + 2\,(60 - 12) = 156$
$V_4(6) - V_3(5) = 300 \quad \rightarrow \quad P = 0.52$

B) $3\left(V_3'(6) - V_2'(6)\right) = 3\,(216 - 36) = 540$
$V_4'(6) - V_3'(6) = 1080 \quad \rightarrow \quad P = 0.5$

Na skladu je 100 žárovek.. Z nich 3 jsou vadné. Náhodně vybereme 5 žárovek. Jaká je pr., že aspoň jedna z vybraných bude vadná?

$$\left(3C_4\left(97\right)+C_2\left(3\right)C_3\left(97\right)+97\right)/C_5\left(100\right)=0.1439.$$

1.6.2 Statistická pravděpodobnost

Statistická pravděpodobnost jevu je dána podílem počtu experimentů, při kterých sledovaný jev nastoupil a počtu všech provedených experimentů.

Příklady

Dotazovali jsme se lidí na jejich platovou skupinu a bydliště. Rozlišujeme platové skupiny P: I, II a III a bydliště B: S a J. Výsledky jsou v tabulce

B\P	I	II	III	Σ
S	235	181	143	**559**
J	251	113	77	**441**
Σ	**486**	**294**	**220**	**1000**

Jaká je (i) pr. těch, co P=I a B=J; (ii) P=II; (iii) P=I když víme, že B=J.

Řešení:

(i) $P\left(P=I,B=J\right)=P_{P,B}\left(I,J\right)=251/1000.$

(ii) $P\left(P=II\right)=P_P\left(II\right)=\left(181+113\right)/1000=294/1000.$

(iii) $P(P=I|B=J)=251/\left(251+113+77\right)=251/441.$

Pozn.: Podmíněná (iii) má jiný základní prostor.

1.7 Příklady

1. Hod 2 kostkami:

(a) Jaká je pravděpodobnost, že při hodu dvěma kostkami padne sudý součet?

(b) Jaká je pravděpodobnost, že při hodu dvěma kostkami padne sudý součet, jestliže na první kostce nepadla 6.

(c) Jaká je pravděpodobnost, že při hodu dvěma kostkami padne sudý součet, jestliže ani na jedné kostce nepadla 6.

Maticové schéma:

	1	**2**	**3**	**4**	**5**	**6**
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

(a) $P = \frac{18}{36} = \frac{1}{2} \cdots$ všechny sudé lomeno všechny

(b) $P = \frac{15}{30} = \frac{1}{2} \cdots$ bez posledního řádku (jsou nezávislé)

(c) $P = \frac{13}{25} \cdots$ bez posledního řádku a sloupce (jsou závislé).

Závislost je dána tím, že v odstraněných prvcích (a tedy u v těch, co zůstaly) je rozvážen původní poměr sudých a lichých!

2. Nezávislé a závislé pokusy - korálky (3m a 5b)

V krabičce máme 3 modré a 5 bílých, stejně velkých, korálků.

(a) Jaká je pr., že při třech tazích vytáhneme 2 modré, jestliže korálky vracíme?

Binomická pravděpodobnost

Tahy jsou nezávislé (vracíme), $P(b) = \frac{5}{8}$, $P(m) = \frac{3}{8}$, jsou tři možnosti tahů

1.	m	m	b
2.	m	b	m
3.	b	m	m

Tedy

$$P_3(2) = 3\left(\frac{3}{8}\right)^2 \frac{5}{8} = 0.264$$

Obecně: n pokusů, p pr. v jednom pokuse, k počet úspěchů

$$P_n(k) = \binom{n}{k} p^k (1-p)^{n-k}, \ k = 0, 1, 2, \cdots$$

(b) Jaká je pr., že při tahu dvou korálků vytáhneme jako druhý modrý, jestliže jsme první nevrátili?

$$P(\textit{druhý}\, m) = P(b, m) + P(m, m) = P(b)\, P(m|b) + P(m)\, P(m|m) = \frac{5}{8}\frac{3}{7} + \frac{3}{8}\frac{2}{7} = \frac{3}{8}$$

Strom

			b; $\frac{4}{7}$	b,b	$\frac{5}{8}\frac{4}{7}$	
		4b3m				
	b; $\frac{5}{8}$		m; $\frac{3}{7}$	b,m	$\frac{5}{8}\frac{3}{7}$	x
5b3m						
	m; $\frac{3}{8}$		b; $\frac{5}{7}$	m,b	$\frac{3}{8}\frac{5}{7}$	
		5b2m				
			m; $\frac{2}{7}$	m,m	$\frac{3}{8}\frac{2}{7}$	x

Výsledek dostaneme součtem pravděpodobností v posledním sloupci, u kterých je křížek (tj. druhý korálek byl vytažen modrý). Je tedy

$$P\left(\text{druhý modrý}\right) = \frac{5}{8}\cdot\frac{3}{7} + \frac{3}{8}\cdot\frac{2}{7} = \frac{3}{8}.$$

3. Úplná pravděpodobnost

Na dálnici v Praze sledujeme provoz. V oblasti ulice Legerova evidujeme stav (S) provozu (i - provoz bez poruchy, ii - blokovaný jeden pruh, iii - blokováno více pruhů) a v oblasti Hlavní nádraží měříme intenzitu (I) proudu (nulová, malá, střední, velká). Z naměřených dat jsme určili pr. $P\left(S\right)$ a $P\left(I|S\right)$ - jak[1]? Naším úkolem je určit pravděpodobnosti intenzit při neznalosti stavu, tj. $P\left(I\right)$.

1. pokus S s výsledky i, ii, iii
2. pokus I s výsledky n, m, s, v.

Známe $P\left(S\right) = \left[P\left(S=i\right), P\left(S=ii\right), P\left(S=iii\right)\right]$ a $P\left(I|S\right)$ tj. matici

$$\begin{bmatrix} P\left(I=n|S=i\right) & P\left(I=n|S=ii\right) & P\left(I=n|S=iii\right) \\ P\left(I=m|S=i\right) & P\left(I=m|S=ii\right) & P\left(I=m|S=iii\right) \\ P\left(I=s|S=i\right) & P\left(I=s|S=ii\right) & P\left(I=s|S=iii\right) \\ P\left(I=v|S=i\right) & P\left(I=v|S=ii\right) & P\left(I=v|S=iii\right) \end{bmatrix}$$

Například:

$$P\left(S\right) = \left[0.2,\ 0.5,\ 0.3\right], \qquad P\left(I,S\right) = \begin{bmatrix} 0.2 & 0.3 & 0.5 \\ 0.1 & 0.2 & 0.2 \\ 0.5 & 0.4 & 0.2 \\ 0.2 & 0.1 & 0.1 \end{bmatrix}$$

Řešení:

$$\begin{aligned} P\left(I\right) &= P\left(I,S=i\right) + P\left(I,S=ii\right) + P\left(I,S=3\right) = \\ &= P\left(I|i\right)P\left(i\right) + P\left(I|ii\right)P\left(ii\right) + P\left(I|iii\right)P\left(iii\right) \end{aligned}$$

např. pro $I=m$ je $P\left(I=m\right) = 0.1\times 0.2 + 0.2\times 0.5 + 0.2\times 0.3 = 0.18$

[1] Např. z relativních četností.

4. Bayesův vzorec

Opět na dálnici. Měříme pouze intenzity u Hlavního nádraží. Po změření $I = m$ máme určit pravděpodobnost nepozorovaného stavu v Legerově ulici, tj. $P(S|I = m)$.

Jako předchozí.

Řešení:

$$P(S|I = m) = \frac{P(I = m|S)\,P(S)}{P(I = m)}$$

Např. pro $S = i$ je $P(S = i|I = m) = \frac{O.2 \times 0.2}{0.18} = 0.22$

1.8 Procvičování

Klasická pravděpodobnost

V dodávce 100 kusů křišťálových váz je 5 vadných. Při kontrole vybereme náhodně 4 kusy. S jakou pravděpodobností
a) jedna vybraná váza je vadná?
b) alespoň jedna z vybraných váz je vadná?}

[a) 0.176, b) 0.188 (kombinace)]

Geometrická pravděpodobnost

Do kruhu o poloměru R je vepsán čtverec. Poté je do kruhu náhodně vhozen bod. S jakou pravděpodobností padne do vepsaného čtverce?

[$\frac{2}{\pi}$ (poměr ploch)]

Hodiny, které nebyly správně nataženy, se zastaví. Jaká je pravděpodobnost, že se velká ručička zastaví mezi šestkou a devítkou?

[$\frac{1}{4}$ (poměr úhlů)]

Nezávislé pokusy

Pravděpodobnost, že zákazník vejde do obchodu v průběhu jedné minuty je 0,01. S jakou pravděpodobností v průběhu 100 minut vejdou do obchodu tři zákazníci?

[0.061 (binomická věta - 100 x opakovaný pokus)]

Přístroj je sestaven z 300 nezávisle pracujících součástek. Pravděpodobnost poruchy každé ze součástek za jednu směnu je 0,01. S jakou pravděpodobností v náhodně vybrané směně bude mít alespoň jedna součástka přístroje poruchu? S jakou bude bez poruchy?

[0.951, $1 - 0.951$ (součiny)]

Pravděpodobnostní strom

V urně je jeden bílý a čtyři černé míčky. Dvě osoby vytahují střídavě a bez vracení vždy po jednom míčku. Vyhrává ten, kdo první vytáhne bílý míček. S jakou pravděpodobností to bude ten, kdo začíná?

[$\frac{3}{5}$ (strom)]

V urně jsou tři bílé, pět černých a dva červené míčky. Dvě osoby vytahují střídavě a bez vracení vždy po jednom míčku. Vyhrává ten, kdo první vytáhne bílý míček a při tahu červeného míčku končí hra nerozhodně. S jakou pravděpodobností vyhraje ten, kdo začíná?

[0.395 (strom)]

Dva hráči házejí postupně mincí. Vyhrává ten, komu padne jako první líc. S jakou pravděpodobností vyhraje první z hráčů?

[$\frac{2}{3}$ (strom - geometrická řada)]

Úplná pravděpodobnost

V dílně pracuje 20 dělníků, kteří vyrábějí stejné součástky. Každý z nich vyrobí za směnu stejné množství. Deset z nich vyrobí 94% výrobků 1.třídy, šest 90% a čtyři 85%. S jakou pravděpodobností náhodně vybraný výrobek bude 1.třídy?

[0.91]

Při sportovní střelbě volí střelec náhodně jednu ze čtyř pušek. Pravděpodobnosti zásahu jednotlivých pušek jsou 0,6; 0,7; 0,8; 0,9. Jaká je pravděpodobnost zásahu při jednom výstřelu?

[0.75]

Bayesův vzorec

Při vyšetřování pacienta je podezření na tři navzájem se vylučující onemocnění. Pravděpodobnost výskytu první choroby je 0,3, druhé 0,5 a třetí 0,2. Laboratorní zkouška je pozitivní u 15% nemocných s první nemocí, 30% nemocných s druhou a 30% nemocných s třetí nemocí. Jaká je pravděpodobnost druhé nemoci, je-li po laboratorním vyšetření výsledek pozitivní?

[0.588]

V dílně pracuje 10 dělníků, kteří za směnu vyrobí stejný počet výrobků. Pět z nich vyrobí 96% standardních výrobků, tři 90% a dva 85%. Náhodně vybereme jeden výrobek a ten je standardní. S jakou pravděpodobností jej vyrobila první skupina dělníků?

[0.52]

2 Náhodná veličina a její popis

2.1 Náhodná veličina

V předchozí kapitole jsme pojednali o náhodném pokusu a zkonstruovali jsme jeho úplný popis - pravděpodobnostní prostor. Důvody, proč zavádíme nový pojem „náhodná veličina“, jsou následující:

1. Konstrukce úplného popis náhodného pokusu není jednoduchá záležitost, kterou jsme schopni provést jen u nejjednodušších pokusů jako je například hod mincí nebo kostkou, tažení korálků různých barev nebo losování karet. Představme si ale pokus, při němž sledujeme na daném místě komunikace rychlosti projíždějících vozidel. Úplná analýza tohoto pokusu by znamenala zjistit vše nejen o řidičích (schopnosti, povahu, zodpovědnost, šikovnost, plány), ale také o automobilech, silnicích, atd. Zde je konstrukce úplného pokusu nemožná. Přitom ale nám často postačí nějaký neúplný popis - charakteristika, jako například průměrná rychlost.

2. Konstrukce charakteristik vyžaduje numerické výpočty. Např. pro výpočet průměru je třeba sčítat a nakonec dělit počtem členů. Výsledky náhodného pokusu však nemusí mít numerickou povahu. např. strana mince - rub, líc, barva na semaforu: zelená, oranžová, červená, a podobně. Tento problém lze však jednoduše řešit. Stačí výsledkům přiřadit čísla, např. rub$\leftarrow$ 0, líc$\leftarrow$ 1. Potom průměrný hod bude 0.5 a to je výsledek, který jsme očekávali. Říká, že průměr je uprostřed a tedy, že obě strany rub i líc jsou stejně pravděpodobné.

Tyto úvahy vedou k definici náhodné veličiny.

2.1.1 Náhodná veličina

Náhodná veličina X je zobrazení z množiny výsledků náhodného pokusu Ω do množiny reálných čísel,

$$X: \ \Omega \to R \tag{2}$$

pro které platí

$$\{e \in \Omega | X(e) \leq x\} \in A, \ \forall x \in R \tag{3}$$

Komentář k definici

Uvedená definice neříká nic jiného než to, že výsledkům náhodného pokusu jsou přiřazena čísla. Jsou to hodnoty náhodné veličiny, kterým se také říká realizace náhodné veličiny.

Navíc je uvedena podmínka, která požaduje, aby intervaly typu $(-\infty, x\rangle$ „byly prvky“ jevového pole, tj. aby jim vždy bylo možno přiřadit pravděpodobnost. Tato vlastnost bude důležitá později, pro definici úplného popisu náhodné veličiny pomocí distribuční funkce.

V tuto chvíli je třeba ještě vysvětlit, co znamená vyjádření požadující aby intervaly $(-\infty, x\rangle$ „byly prvky“ jevového pole. Na tento zápis je třeba pohlížet ve smyslu podmínky (3). Tedy požadujeme, aby množina výsledků e, s hodnotami přiřazenými náhodnou veličinou X, které jsou menší než x vždy (tj. $\forall x \in R$) tvořila jev, patřící do jevového pole. A protože všem prvkům jevového pole jsou přiřazeny pravděpodobnosti, bude přiřazena pravděpodobnost i naší množině.

Nakonec se ještě zmíníme o tom, jakou pozici zaujímá náhodná veličina vzhledem k výstavbě jevové pravděpodobnosti. Základem jevové pravděpodobnosti je náhodný pokus, z něho získáváme výsledky. Výsledkům jsou přiřazena čísla, hodnoty náhodné veličiny. Lze proto říci, že náhodná veličina odpovídá náhodnému pokusu. Hodnoty náhodné veličiny (body na reálné ose) odpovídají výsledkům náhodného pokusu. Souvislost náhodné veličiny s předchozími pojmy lze vyznačit v následujícím schematu

$$\begin{array}{ccc} \text{náhodný pokus} & & \text{náhodná veličina} \\ \downarrow & & \downarrow \\ \text{výsledky} & \rightarrow & \text{realizace} \end{array}$$

Náhodná veličina tedy zastupuje náhodný pokus.

Jevovému poli, tj. množině všech jevů, kterým přiřazujeme pravděpodobnosti, odpovídá (v pojmech náhodné veličiny) borelovské pole na reálné ose (tj. systém všech intervalů, které tvoří borelovské množiny - viz odstavec 13.1).

Příklad

Pro jednoduchost budeme uvažovat pokus hod mincí s výsledky rub (R) a líc (L). Pravděpodobnostní pole $\mathcal{P} = (\Omega, A, P)$ pro tento příklad je

$$\Omega = \{R, L\}$$

$$A = \{\emptyset, \{R\}, \{L\}, \Omega\}$$

$$P(\emptyset) = 0,\ P(\{R\}) = 0.5,\ P(\{L\}) = 0.5,\ P(\Omega) = 1.$$

Náhodnou veličinu budeme definovat např. takto

$$X(R) = 0,\ X(L) = 1$$

což skutečně je zobrazení z Ω do R. Nyní ještě ověříme podmínku.

Pro $x \in (-\infty, 0)$ platí: množina výsledků e pro něž je $X(e) \leq x$ je $\emptyset$ - není žádný výsledek, který by měl přiřazenou hodnotu menší než $x < 0$.

Pro $x \in \langle 0,1)$ je množina výsledků, pro které je $X(e) \leq x$ rovna $\{R\}$, protože $X(R) = 0 \leq x \in \langle 0,1)$.

Pro $x \geq 1$ je tato množina $\{R, L\} = \Omega$, protože $X(R) = 0$, $X(L) = 1 \leq x \in \langle 1, \infty)$.

Množiny $X(e) \leq x$ jsou prvky A, a to $\forall x$. Podmínka (3) na náhodnou veličinu je tedy splněna.

Poznámka

Jak je vidět z uvedeného příkladu, podmínka (3) na náhodnou veličinu není kritická. V případech, které budeme dále uvažovat je vždy splněna a nebudeme ji tedy už dále vyšetřovat.

Typy náhodné veličiny

Hodnoty náhodné veličiny se nazývají **realizace**.

Podle toho, do jaké množiny patří realizace, dělíme náhodnou veličinu na

- **diskrétní náhodná veličina** - realizace jsou z konečné nebo spočetné množiny (např. hod mincí, hod kostkou, výběr z deseti barevných korálků, apod.)
- **spojitá náhodná veličina** - realizace jsou z množiny reálných čísel, tedy má nespočetně mnoho realizací (např. bezporuchová doba funkce přístroje, doba čekání na dopravní prostředek, měření rozměru u vyráběných součástek, apod.)

2.2 Rozdělení náhodné veličiny

Rozdělením náhodné veličiny budeme mít na mysli její úplný pravděpodobnostní popis. Jedná se vymezení přípustných hodnot náhodné veličiny a přiřazení pravděpodobností buď přímo jejím hodnotám, nebo množinám jejích hodnot. Rozdělení nejčastěji definuje **distribuční funkce** nebo **hustota pravděpodobnosti**.

2.3 Distribuční funkce

Nejsnáze definovatelným úplným popisem náhodné veličiny je distribuční funkce. Lze ji zavést přímo použitím pravděpodobnosti a je společná jak pro diskrétní tak i spojitou náhodnou veličinu. Nyní uvedeme její definici.

2.3.1 Distribuční funkce

Distribuční funkce F_X reálného argumentu x pro náhodnou veličinu X je definována vztahem

$$F_X(x) = P(X \leq x) \tag{4}$$

Komentář k definici

Podobně jako pravděpodobnostní prostor určoval pravděpodobnost pro každý prvek jevového pole, tedy pro každý jev, požadujeme po úplném popisu náhodné veličiny, aby určoval pravděpodobnost pro každý prvek borelovského pole, tedy každý interval na reálné ose který představuje borelovslou množinu. Jak jsme se již zmínili v odstavci 13.1, lze každou borelovslou množinu na reálné ose generovat z intervalů typu $(-\infty, x\rangle$, kde $x \in R$. Proto nám stačí přiřazovat pravděpodobnosti právě těmto intervalům a ostatní je již možno z tohoto přiřazení dopočítat.

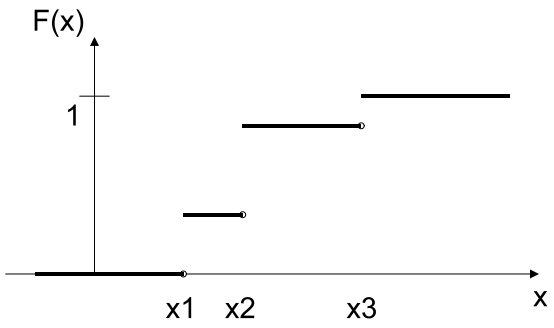
Vlastnosti distribuční funkce

Distribuční funkce je

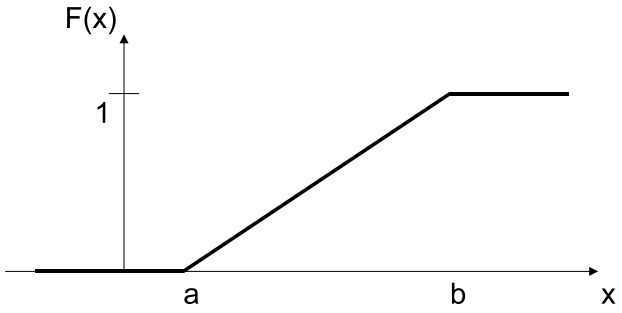
- neklesající v celém oboru,
- limita vlevo je nula, $\lim_{x\to-\infty} F(x) = 0$,
- limita vpravo jedna, $\lim_{x\to\infty} F(x) = 1$.

Příklady distribučních funkcí

Příklad pro diskrétní náhodnou veličinu



Příklad pro spojitou náhodnou veličinu



Příklady

Hod korunou, doba čekání na autobus.

Poznámka

Distribuční funkci jsme označili $F_X(x)$, kde X je označení příslušné náhodné veličiny a x je realizace, tj. reálné číslo - hodnota náhodné veličiny X. Pokud se označení náhodné veličiny shoduje s označením její realizace nebo pokud nemůže dojít k omylu, bývá zvykem index X vynechávat. Potom se rozumí, že distribuční funkce $F(x)$ se vztahuje k náhodné veličině X.

2.4 Pravděpodobnostní funkce

Druhým z používaných úplných popisů diskrétní náhodné veličiny je pravděpodobnostní funkce. Její definice je následující.

2.4.1 Pravděpodobnostní funkce

Pro diskrétní náhodnou veličinu X s realizacemi x_i, $i = 1, 2, \cdots, n$ definujeme **pravděpodobnostní funkci** f_X jako reálnou funkci diskrétního argumentu x_i (tj. posloupnost) vztahem

$$f_X(x_i) = P(X = x_i).$$

Komentář k definici

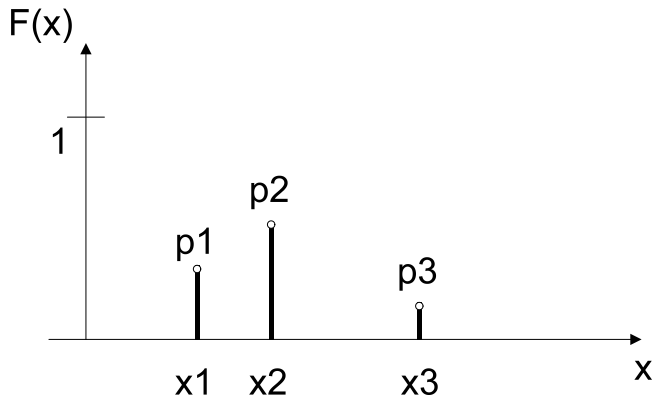
Definice říká, že hodnota pravděpodobnostní funkce v bodě x_i, který je realizací diskrétní náhodné veličiny, se rovná pravděpodobnosti této realizace.

Vlastnosti pravděpodobnostní funkce

Pravděpodobnostní funkce má následující vlastnosti

- všechny její členy jsou nezáporné, $f_X(x_i) \geq 0,\ i = 1, 2, \cdots, n,$
- součet všech jejích členů je roven jedné, $\sum_{i=1}^{n} f_X(x_i) = 1.$

Příklad pravděpodobnostní funkce je na obrázku



Příklad

Hod korunou.

Poznámka

Pro značení opět platí poznámka o značení distribuční funkce. Pokud nemůže dojít k omylu, symbolem $f(x)$ označujeme pravděpodobnostní funkci vztahující se k náhodné veličině X.

2.5 Hustota pravděpodobnosti

Pro spojitou náhodnou veličinu je je obdobou pravděpodobnostní funkce **hustota pravděpodobnosti.** Její definice ale nemůže být stejná jako v případě pravděpodobnostní funkce, která byla dána přímo pravděpodobnostmi jednotlivých realizací. Spojitá náhodná veličina má totiž nekonečně mnoho realizací a když se mezi ně rozdělí jednotka pravděpodobnosti, pak na každou vyjde nula. Touto cestou bychom tedy dostali identicky nulovou funkci. Proto se hustota pravděpodobnosti zavádí takto.

2.5.1 Hustota pravděpodobnosti

Pro **spojitou náhodnou veličinu** X definujeme hustotu pravděpodobnosti f_X jako reálnou funkci reálného argumentu vztahem

$$f_X(x) = \frac{dF_X(x)}{dx} \quad \text{nebo} \quad F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt.$$

Komentář k definici

Oba uvedené vzorce v definici jsou prakticky ekvivalentní. Derivace distribuční funkce vyžaduje existenci derivace na celém definičním oboru. Tento vzorec tedy nelze použít např. pro distribuční funkce po částech lineární, což druhému vzorci nevadí. Pro konečný počet bodů, kde se derivace zleva nerovná derivaci zprava však lze derivaci snadno dodefinovat a oba vzorce jsou pak prakticky stejné.

První vzorec má tu výhodu, že je explicitní. Druhý vzorec je velice důležitý. Říká, že pravděpodobnost intervalu se spočte jako plocha pod hustotou pravděpodobnosti nad tímto intervalem. Snadno se ukáže, že toto tvrzení neplatí jen pro polonekonečné intervaly ale i pro obecné

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

$$= \int_{\infty}^{b} f_X(t)\,dt - \int_{\infty}^{a} f_X(t)\,dt = \int_{a}^{b} f_X(t)\,dt.$$

Vlastnosti hustoty pravděpodobnosti

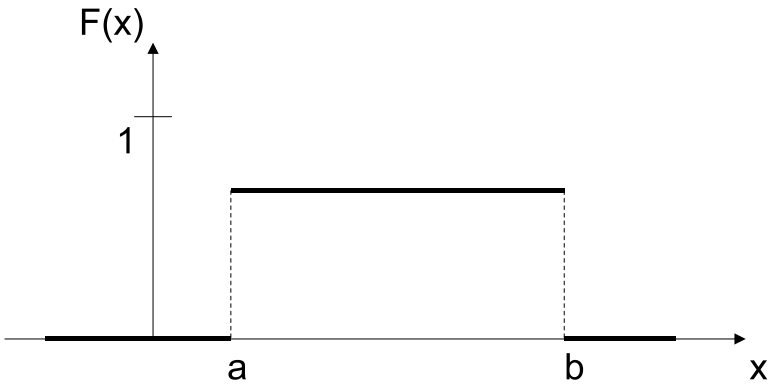
Hustota pravděpodobnosti je

- nezáporná na celém definičním oboru,
- její integrál přes celý definiční obor je jedna, $\int_{-\infty}^{\infty} fX(x)\,dx = 1$

Další důležitá vlastnost hustoty pravděpodobnosti je, že její integrál přes interval (a,b) je roven pravděpodobnosti, že náhodná veličina bude mít realizaci z tohoto intervalu, tj.

$$P(X \in (a,b)) = \int_{a}^{b} f_X(x)\,dx.$$

Příklad hustoty pravděpodobnosti je na obrázku



Poznámka

Pro značení hustoty pravděpodobnosti platí totéž, co pro pravděpodobnostní funkci nebo distribuční funkci. $f(x)$ je označení pro hustotu pravděpodobnosti náhodné veličiny X.

Příklad

Doba čekání na autobus.

2.6 Střední hodnota, rozptyl a směrodatná odchylka

Střední hodnota a rozptyl patří mezi nevýznamnější neúplné popisy náhodné veličiny. Neposkytují veškerou dostupnou informaci o náhodné veličině, zato jsou názorné a jednoduché. Střední hodnota udává určitou hladinu, na které se realizace náhodné veličiny pohybují. Rozptyl vypovídá o variabilitě náhodné veličiny, tj. o tom, jak moc se jednotlivé realizace od sebe liší. Čím je rozptyl větší, tím je větší rozptýlenost realizací náhodné veličiny. Směrodatná odchylka je odmocninou z rozptylu a je to veličina, která je srovnatelná s realizacemi. Lze tedy psát: realizace náhodné veličiny X jsou $E(X) \pm \sigma$.

Pro diskrétní náhodnou veličinu X s realizacemi x_i, $i = 1, 2, \cdots, n$ a pravděpodobnostní funkci f_X definujeme

Střední hodnotu $E(X)$

$$E(X) = \sum_{i=1}^{n} x_i f_X(x_i),$$

Rozptyl $D(X)$

$$D(X) = \sum_{i=1}^{n} (x_i - E(X))^2 f_X(x_i) = \sum_{i=1}^{n} x_i^2 f_X(x_i) - [E(X)]^2,$$

kde druhé vyjádření rozptylu se nazývá výpočetní tvar.

Směrodatnou odchylku σ

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

Příklad

Je dána diskrétní náhodná veličina X s pravděpodobnostní funkcí $f(x_i)$

x_i	3	5	10
$f(x_i)$	0.3	0.5	0.2

Určete střední hodnotu a rozptyl.

Střední hodnota:

$$E[X] = 3 \times 0.3 + 5 \times 0.5 + 10 \times 0.2 = 5.4$$

Rozptyl:

$$D[X] = (3 - 5.4)^2 \times 0.3 + (5 - 5.4)^2 \times 0.5 + (10 - 5.4)^2 \times 0.2 = 6.04$$

Pro spojitou náhodnou veličinu X s realizacemi $x \in R$ a hustotou pravděpodobnosti f_X definujeme

Střední hodnotu $E(X)$

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx$$

Rozptyl $D(X)$

$$D(X) = \int_{-\infty}^{\infty} (x - E(X))^2 f_X(x)\, dx = \int_{-\infty}^{\infty} x^2 f_X(x)\, dx - [E(X)]^2,$$

kde opět, druhé vyjádření se nazývá výpočetní tvar vzorce pro rozptyl.

Směrodatnou odchylku σ

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

což je stejný vzorec jako pro diskrétní náhodnou veličinu.

Příklad

Je dána hustota pravděpodobnosti $f(x)$náhodné veličiny X

$$f(x) = \begin{cases} 0 & \text{pro } x < 1 \\ \frac{1}{4} & \text{pro } x \in (1,5) \\ 0 & \text{pro } x > 5 \end{cases}$$

Určete střední hodnotu a rozptyl.

Střední hodnota:

$$E[X] = \int_1^5 x\frac{1}{4}dx = \left[\frac{x^2}{8}\right]_1^5 = \frac{25-1}{8} = 3$$

Rozptyl:

$$D[X] = \int_1^5 (x-3)^2 \frac{1}{4}dx = \left[\frac{(x-3)^3}{12}\right]_1^5 = \frac{4}{3}$$

2.7 Kvantil a kritická hodnota

Pro náhodnou veličinu x s hustotou pravděpodobnosti $f(x)$a pravděpodobnost α definujeme

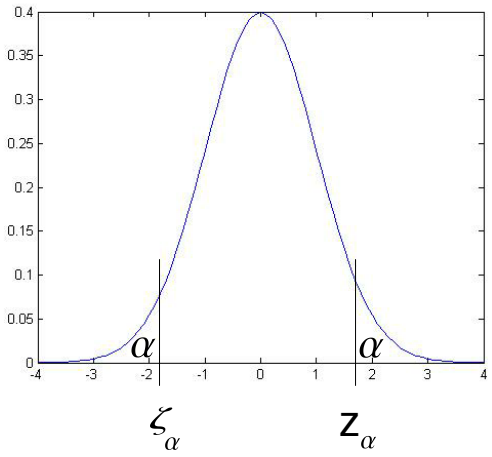
Kvantil ζ_α

$$\int_{-\infty}^{\zeta_\alpha} f(x)\,dx = \alpha.$$

Kritická hodnota z_α

$$\int_{z_\alpha}^{\infty} f(x)\,dx = \alpha.$$

Kvantil a kritická hodnota jsou tedy pravá a levá hranice, které oddělují $\alpha \times 100\%$ nejmenších nebo největších realizací náhodné veličiny. Obě charakteristiky jsou vyznačeny na následujícím obrázku.



Z obrázku je vidět: kvantil ζ_α ohraničuje plochu pod hustotou pravděpodobnosti a velikosti α, a to vlevo, kritická hodnota z_α ohraničuje podobnou plochu, ale vpravo.

Příklad

Určete 5% kvantil $\zeta_{0.05}$ normálního rozdělení $X \sim N(\mu, \sigma^2)$ s $\mu = 5$ a $\sigma^2 = 18$, jestliže statistické tabulky udávají $Pr(Z > 1.645) = 0.05$ a $Z \sim N(0, 1)$.

Protože tabulky udávají hodnoty jen pro normované normální rozdělení, musíme se nejdříve zabývat normováním.

$$X = \sigma Z + \mu$$

Příklad 1

Je dána hp $f(x) = 1 - kx^2$ pro $x \in \left(-\frac{3}{4}; \frac{3}{4}\right)$.

Určete (i) konstantu k, (ii) distribuční funkci, (iii) střední hodnotu a (iv) rozptyl (v) pravděpodobnost $P\left(|X| < \frac{1}{2}\right)$ a kritickou hodnotu $z_{0.05}$ a $z_{0.01}$.

(i) Konstanta

$$\int_{-\frac{3}{4}}^{\frac{3}{4}} 1 - kx^2 dx = \left[x - \frac{k}{3}x^3\right]_{-\frac{3}{4}}^{\frac{3}{4}} = 2\left(\frac{3}{4} - \frac{k}{3}\left(\frac{3}{4}\right)^3\right) = 1, \;\rightarrow\; k = \frac{16}{9}$$

(ii) Distribuční funkce

$$F(x) = \int_{-\frac{3}{4}}^{x} 1 - \frac{16}{9}s^2 ds = \left[s - \frac{16}{27}s^3\right]_{-\frac{3}{4}}^{x} = -\frac{16}{27}x^3 + x + \frac{1}{2},$$

pro $x \in \left(-\frac{3}{4}, \frac{3}{4}\right)$, vlevo nula, vpravo jedna.

(iii) Střední hodnota

$$E[X] = \int_{-\frac{3}{4}}^{\frac{3}{4}} x\left(1 - \frac{16}{9}x^2\right) dx = \left[\frac{1}{2}x^2 - \frac{4}{9}x^4\right]_{-\frac{3}{4}}^{\frac{3}{4}} = 0 \quad \text{(symetrie)}$$

(iv) Rozptyl

$$D[X] = \int_{-\frac{3}{4}}^{\frac{3}{4}} x^2\left(1 - \frac{16}{9}x^2\right) dx = \left[\frac{1}{3}x^3 - \frac{16}{45}x^5\right]_{-\frac{3}{4}}^{\frac{3}{4}} = \frac{9}{80} = 0.1125$$

(v) Pravděpodobnost

$$P\left(|X| < \frac{1}{2}\right) = P\left(X \in \left(-\frac{1}{2}; \frac{1}{2}\right)\right) = F\left(\frac{1}{2}\right) - F\left(-\frac{1}{2}\right) = 2\left(-\frac{16}{27}\frac{1}{8} + \frac{1}{2} + \frac{1}{2}\right) = \frac{23}{27}$$

(vi) Kritická hodnota

$$P(X > z_{0.05}) = 0.05 \;\rightarrow\; 1 - F(z_{0.05}) = 0.05 \;\rightarrow\; F(z_{0.05}) = 0.95$$

$$-\frac{16}{27}z^3 + z + \frac{1}{2} = 0.95, \;\rightarrow\; -\frac{16}{27}z^3 + z - \frac{9}{20} = 0.$$

To nelze analyticky řešit. Buď v Excelu je "Nástroje ▷ Hledat řešení", nebo v Matlabu pomocí programu

```
a=.05; x1=-3/4; x2=3/4;
d=1; i=0;
if fce(x1,a)*fce(x2,a)>0
   disp('Chyba: f(x1) a f(x2) musi mit ruzna znamenka');
   return
end
while d>1e-5
   i=i+1; if i>1000, break, end
   y1=fce(x1,a);
   y2=fce(x2,a);
   xp=(abs(y1)*x2+abs(y2)*x1)/(abs(y1)+abs(y2));
   yp=fce(xp,a);
   if yp>0, x2=xp; else x1=xp; end
   d=abs(yp)
end
xp,i


function y=fce(x,a)
y=-16/27*x^3+x+1/2-(1-a);
```

Odvození: viz Dodatek 13.4 (vzadu).

Příklad 2

Je dána hp $f(x) = \frac{3-|x^2-4x+3|}{8}$ pro $x \in (0;4)$.

Distribuční funkce:

Určete distribuční funkci a $P(X \in (0.5;1.5))$

Do $x = 0$ je $F(x) = 0$.

Pro $x \in (0,1)$

$$\begin{aligned} F(x) &= 0 + \int_0^x (-\frac{1}{8}t^2 + \frac{1}{2}t)dt = -\frac{1}{24}x^3 + \frac{1}{4}x^2 \\ & F(1) = \frac{5}{24} \end{aligned}$$

Pro $x \in (1,3)$

$$\begin{aligned} F(x) &= \frac{5}{24} + \int_1^x (\frac{1}{8}t^2 - \frac{1}{2}t + \frac{3}{4})dt = \frac{5}{24} + \frac{1}{24}x^3 - \frac{1}{4}x^2 + \frac{3}{4}x - (\frac{1}{24} - \frac{1}{4} + \frac{3}{4}) = \\ &= \frac{1}{24}x^3 - \frac{1}{4}x^2 + \frac{3}{4}x - \frac{1}{3} \\ & F(3) = \frac{19}{24} \end{aligned}$$

Pro $x \in (3,4)$

$$\begin{aligned} F(x) &= \frac{19}{24} + \int_3^x (-\frac{1}{8}t^2 + \frac{1}{2}t)dt = \frac{19}{24} - \frac{1}{24}x^3 + \frac{1}{4}x^2 - (-\frac{9}{8} + \frac{9}{4}) = \\ &= -\frac{1}{24}x^3 + \frac{1}{4}x^2 - \frac{1}{3} \end{aligned}$$

Nad $x = 4$ je $F(x) = 1$.

Pravděpodobnost

$$P(X \in (0.5;1.5)) = F(1.5) - F(0.5) =$$

$$= \frac{1}{24}\left(\frac{3}{2}\right)^3 - \frac{1}{4}\left(\frac{3}{2}\right)^2 + \frac{3}{4}\frac{3}{2} - \frac{1}{3} - \left(-\frac{1}{24}\left(\frac{1}{2}\right)^3 + \frac{1}{4}\left(\frac{1}{2}\right)^2\right) = 0.3125$$

Doma: spočítat střední hodnotu a rozptyl (stř.h.=2, rozpt.=1)

1. V závodním automobilu je důležitá součástka, jejíž porucha znamená konec závodu. Proto je zálohována tak, že 2 nové součástky jsou připraveny v záloze a v případě poruchy je pokažená součástka automaticky nahrazena jednou ze zásobních. Pr. poruchy součástky během 1 minuty závodu je 0.002, a to nezávisle na době, se kterou je již používána. Pr. poruchy dvou součástek za 1 minutu považujeme za velmi nepravděpodobnou a zanedbáváme ji. Určete rozdělení pr. pro počet odešlých součástek během 5ti hodinového závodu.

   Výsledky jsou 0, 1, 2 a mají binomické rozdělení pr. s $p = 0.002$ a $n = 5 * 60 = 300$. (Každá minuta je pokus, porucha je úspěch - teoreticky může být až 300 poruch, my ale rozlišujeme jen 0, 1, 2, 3-300). Je tedy

$$P(0) = (1 - 0.002)^{300}; \quad P(1) = 300 \times 0.002 \times (1 - 0.002)^{299};$$

$$P(2) = \binom{300}{2} 0.002^2 (1 - 0.002)^{398}, \quad P(3 - 300) = 1 - P(0) - P(1) - P(2)$$

   což je hezké, ale nespočteme to. Proto se musíme uchýlit k aproximaci Poissonovým rozdělením. Pokračování bude příště.

2. Přístroj se skládá ze tří, nezávisle pracujících částí. Pravděpodobnost poruchy každé části je 0.1. Najděte distribuční funkci náhodné veličiny, popisující počet porouchaných částí.

   Vzhledem k nezávislosti se počet porouchaných součástek řídí binomickou pravděpodobností $P_3(k)$ při $p = 0.1$.

$$P(0) = 0.9^3 = 0.729, \qquad P(1) = 3{\cdot}0.1{\cdot}0.9^2 = 0.243, \qquad P(2) = 0.027 \qquad P(3) = 0.1^3 = 0.001$$

   Distribuční funkci dostaneme jako kumulativní součet pravděpodobností, tj.

$$F(x) = \begin{cases} 0 & \text{pro } x < 0 \\ 0.729 & \text{pro } 0 \le x < 1 \\ 0.972 & \text{pro } 1 \le x < 2 \\ 0.999 & \text{pro } 2 \le x < 3 \\ 1 & \text{pro } x \ge 3 \end{cases}$$

   Nakreslit!

3. Zkoušení probíhá tak, že studentovi jsou kladeny otázky, dokud odpovídá správně. První špatná odpověď zkoušení ukončí a známka je odvozena z počtu dobrých odpovědí. Určete pravděpodobnostní funkci a distribuční funkci náhodné veličiny "počet dobrých odpovědí", jestliže otázky jsou nezávislé a pravděpodobnost dobré odpovědi při jednom dotazu je stálá a rovna 0.9.

   Protože jsou otázky nezávislé, je pravděpodobnost dobré série odpovědí rovna součinu pravděpodobností dobrých odpovědí. Pokud má tato série být na konci ukončena, musí následovat špatná odpověď – tedy na konci bude krát $P(\text{špatně}) = 1 - P(\text{dobře})$.

$$P(0) = 0.1, \quad P(1) = 0.9 \cdot 0.1, \quad P(2) = 0.9^2 \cdot 0.1, \quad \ldots \quad , P(x) = 0.9^x \cdot 0.1, \quad \ldots$$

Jedná se tedy o nekonečnou geometrickou posloupnost s prvním členem rovným 0.1 a kvocientem $q = 0.9$. Protože řada začíná indexem $x = 0$, bude člen $P(x)$ v pořadí x+prvý. Částečný součet takové řady je

$$s_x = \sum_{k=0}^{x} = P(0)\frac{1-q^{(x+1)}}{1-q}.$$

Distribuční funkce (jako kumulativní součet hodnot pravděpodobnostní funkce) je

$$F(x) = \sum_{k=0}^{x} 0.1 \cdot 0.9^k = 0.1\frac{1-0.9^{x+1}}{1-x} = 1 - 0.9^{x+1}.$$

4. Je dána hustota pravděpodobnosti

$$f(x) = \frac{1}{2}\sin(x), \qquad \text{pro } x \in (0,\pi) \text{ a 0 jinde.}$$

Určete distribuční funkci.

Distribuční funkce:

$$F(x) = \begin{cases} 0 & \text{pro } x \leq 0 \\ \frac{1}{2}(1-\cos(x)) & \text{pro } x \in (0,\pi) \\ 1 & \text{pro } x \geq \pi \end{cases}$$

2.8 Normování náhodné veličiny

Dvě nejvýznamnější charakteristiky náhodné veličiny jsou střední hodnota μ a rozptyl σ^2. Pro normální rozdělení jsou tyto dvě charakteristiky zároveň parametry rozdělení.

Připomeňme, že změna střední hodnoty znamená změnu polohy hustoty pravděpodobnosti, změna rozptylu představuje změnu jejího tvaru, nárůst rozptylu znamená rozšíření zatímco pokles vyvolá zúžení.

Pro motivaci k uvedeným vzorcům si všimneme rovnoměrného rozdělení na intervalu $(-1,\ 1)$. Toto rozdělení má nulovou střední hodnotu. Budeme-li z něho generovat, budeme dostávat hodnoty mezi -1 a 1. Když ke každé generované hodnotě přičteme 5, budeme dostávat hodnoty mezi -4 a 6, tedy se střední hodnotou 5. Dále, budeme-li původní (neposunuté) hodnoty násobit dvěma, budeme dostávat hodnoty mezi -2 a 2 - tedy s větším rozptylem. Naopak, násobení číslem 0.1 bude podobně znamenat snížení rozptylu čísel, která obdržíme.

V souladu s předchozími úvahami jsou následující vzorce.

Uvažujeme náhodnou veličinu X, s charakteristikami $E[X] = \mu$ a $D[X] = \sigma^2$ a Z, s charakteristikami $E[Z] = 0$, $D[Z] = 1$. Potom platí:

normování

$$Z = \frac{X-\mu}{\sigma} \tag{5}$$

a **odnormování**

$$X = \sigma Z + \mu \tag{6}$$

První vzorec převádí nenormovanou veličinu X na normovanou Z. Druhý naopak.

Příklad

Určeme interval, ve kterém leží 5% největších hodnot normálního rozdělení se střední hodnotou 20 a rozptylem 81, jestliže kvantil $\zeta_{0.05} = 1.645$.

Hledáme hodnotu x, pro niž platí $P(X > x) = 0.05$. Známe kvantil, který se týká normované veličiny. Proto normujeme, a dosadíme hranici $Z = \zeta_{0.05}$

$$Z = \zeta_{0.05} = 1.645 = \frac{x - \mu}{\sigma} = \frac{x - 20}{9}$$

Odtud je $x = 1.645 \times 9 + 20 = 34.8$.

Interval pěti procent největších čísel je tedy $(34.8, \infty)$.

3 Důležitá rozdělení náhodné veličiny

Co to znamená, když prohlásíme, že jsou nějaká důležitá rozdělení? Rozdělení náhodné veličiny je její popis. A náhodná veličina představuje určitý náhodný pokus (kde výsledky prezentujeme jako čísla). Jedná se tedy o typické náhodné pokusy u nichž je znám jejich pravděpodobnostní popis. Několik takových známých rozdělení jak pro diskrétní, tak i pro spojitou náhodnou veličinu si dále uvedeme.

3.1 Diskrétní rozdělení

Diskrétní rozdělení má konečný nebo spočetný počet realizací.

Příklady Hod mincí, hod kostkou, tažení králku ze 3m a 5b, počet nehod na křižovatce za 1 rok, počet aut zaznamenaných detektorem za 1 hodinu.

3.1.1 Alternativní rozdělení

Jedná se o jeden pokus, který může mít dva různé výsledky - úspěch nebo neúspěch. Nejčastější interpretace je následující: máme sklad s výrobky mezi nimiž je určité procento zmetků. Vybereme náhodně jeden výrobek. Úspěch je, když je výrobek dobrý, neúspěch, je-li výrobek zmetek.

Pravděpodobnostní funkce

$$f(x) = \pi^x (1 - \pi)^{1-x}, \; x = 0, 1$$

kde π je parametr - pravděpodobnost úspěšného pokusu.

Uvedenou pravděpodobnostní funkci lze také vyjádřit tabulkou

x	0	1
$f(x)$	$1-\pi$	π

Z tabulky je přímo vidět, že pravděpodobnost úspěchu ($x = 1$) je π a pravděpodobnost neúspěchu ($x = 0$) je $1 - \pi$. Totéž dostaneme i z analytického vyjádření, když dosadíme za x příslušnou hodnotu.

Příklad

Sledujeme auta v křižovatce. Výsledky: 1 - jede rovně, 0 - odbočuje. Pravděpodobnost jízdy rovně je π a odbočení $1 - \pi$.

3.1.2 Binomické rozdělení

Binomické rozdělení dostaneme když provedeme n nezávislých pokusů s alternativním rozdělením a výsledky sečteme (tj. za výsledek vezmeme počet dosažených úspěchů). Například: V křižovatce projíždějí auta, která s pravděpodobností π jedou rovně, nebo s pravděpodobností $1 - \pi$ odbočí. Binomický pokus spočívá ve sledování n aut a hodnoty odpovídající náhodné veličiny jsou jsou dány počtem aut, která projedou rovně.

Pravděpodobnostní funkce

$$f(x) = \binom{n}{x} \pi^x (1-\pi)^{n-x}, \ x = 0, 1, 2, \cdots, n$$

kde π - pravděpodobnost úspěchu v odpovídajícím alternativní pokusu a n - počet provedených alternativních pokusů, jsou parametry rozdělení.

Vzorec, kterým jsme vyjádřili pravděpodobnostní funkci binomického rozdělení se nazývá vzorec pro binomickou pravděpodobnost. Jeho odvození je poměrně snadné. Člen π^x představuje pravděpodobnost požadovaných x úspěchů (při nezávislých pokusech), člen $1 - \pi^{n-x}$ je pravděpodobnost zbylých pokusů, kdy výsledkem byl neúspěch a kombinační číslo $\binom{n}{x}$ představuje počet způsobů, kterými lze z provedených n pokusů vybrat x úspěchů.

Příklad

Za jednu hodinu projede křižovatkou v průměru 50 aut. Jaká je pravděpodobnost, že 5 z nich odbočí, je-li pravděpodobnost přímé jízdy 0.9?

V tomto příkladě považujeme za úspěch odbočení a jeho pravděpodobnost je $\pi = 1-0.9 = 0.1$. Pravděpodobnost odbočení pěti aut z 350 je

$$f(5) = \binom{50}{5} 0.1^5 0.9^{45} = 0.185$$

Příklad

Ve velikém skladu jsou uloženy výrobky. Mezi nimi jich je 85% první jakosti. Jak veliký je třeba udělat výběr, aby v něm bylo s pravděpodobností 99% alespoň 5 dobrých výrobků?

Jedná se opět o binomické rozdělení s velikostí výběru n a pravděpodobností úspěchu $\pi = 0.85$. Hledáme první n takové, že platí

$$P = \sum_{i=5}^{n} \binom{n}{5} \pi^5 (1-\pi)^{n-5} \geq 0.99$$

To je ale implicitní nerovnice, která by se obecně musela řešit nějakou numerickou metodou. V našem případě, kdy n je diskrétní, stačí počítat hodnoty P postupně pro $n = 5, 6, \cdots$ a za výsledek vezmeme první hodnotu n pro kterou sledovaná nerovnice platí. Výpočty uspořádáme do tabulky

n	5	6	7	8	9
P	0.4437	0.7765	0.9262	0.9786	0.9944

Hledaný výběr tedy bude o rozsahu $n = 9$.

Poznámka

Uvažovaný sklad musí být hodně velký. Základním předpokladem pro binomické rozdělení je, že uvažované pokusy jsou nezávislé. Jestliže ale vybíráme výrobky a nevracíme je zpět, jedná se závislé pokusy, protože počet a tím i pravděpodobnost dobrých a Špatných výrobků se mění. Jestliže je ale sklad hodně veliký a v něm velké množství výrobků, bude relativní vliv vybraných výrobků na celkový počet zanedbatelný a pravděpodobnosti můžeme pokládat za konstantní.

3.1.3 Poissonovo rozdělení

Poissonovo rozdělení je limitní případem binomického, a sice pro $n \to \infty$ a $\pi \to 0$, a to tak, že limita součinu $n \cdot \pi = \lambda$ je konstantní. λ se nazývá intenzita Poissonova rozdělení.

Pravděpodobnostní funkce

$$f(x_i) = \mathrm{e}^{-\lambda} \frac{\lambda^{x_i}}{x_i!}, \; x = 0, 1, 2, \cdots$$

Toto rozdělení má sice nekonečný počet realizací, ale nekonečně spočetný (realizacím lze jednoznačně přiřadit přirozená čísla). Proto patří mezi diskrétní rozdělení.

Podle základních vlastností musí být součet všech členů roven jedné, tj.

$$\sum_{i=0}^{\infty} \mathrm{e}^{-\lambda} \frac{\lambda^{x_i}}{x_i!} = \mathrm{e}^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^{x_i}}{x_i!} = 1$$

což skutečně je, protože suma v prostředním výrazu není nic jiného než Taylorův rozvoj funkce $\mathrm{e}^{-\lambda}$.

Pravděpodobnostní funkci Poissonova rozdělení lze odvodit jako limitu binomického rozdělení. Odvození je uvedeno v Dodatku 13.5

Příklad 1

Uvažujme podobný příklad jako byl u binomického rozdělení ale s extrémějšími hodnotami: Za jeden den projede křižovatkou v průměru 3500 aut. Jaká je pravděpodobnost, že 5 z nich odbočí, je-li pravděpodobnost přímé jízdy 0.999?

Pokud bychom chtěli úlohu řešit pomocí binomického rozdělení (protože se skutečně o binomické rozdělení jedná), dospějeme k výrazu

$$f(5) = \binom{3500}{5} 0.001^5 0.999^{3495}$$

což není moc hezké. Proto aproximujeme Poissonovým rozdělením pro $\lambda = 3500 \cdot 0.001 = 3.5$. Dostaneme

$$f_P(5) = \mathrm{e}^{-3.5} \frac{3.5^5}{5!} = 0.132$$

Příklad 2

Sledujeme silnici s malým provozem. Po dobu jedné minuty zaznamenáváme každou vteřinu přítomnost auta. Po hodině sledování jsme zjistili, že v průměru za jednu minutu projedou 3 auta. Za předpokladu Poissonova rozdělení počtu projetých aut během jedné minuty určete pravděpodobnost, že během jedné minuty projede více než 5 automobilů.

Pro určení λ známe $n = 60,\ \pi = 3/60$. Odtud $\lambda = n\pi = 3$.

$$P(X > 5) = 1 - P(X \leq 5) = 1 - \mathrm{e}^{-3} \left(\frac{3^0}{0!} + \frac{3^1}{1!} + \frac{3^2}{2!} + \frac{3^3}{3!} + \frac{3^4}{4!} + \frac{3^5}{5!} \right) = 0.084$$

Pravděpodobnost projetí více než pěti aut za jednu minutu je 0.084.

3.1.4 Geometrické

Základem pro geometrické rozdělení je opět alternativní pokus, tedy pokus se dvěma výsledky (úspěch a neúspěch) s pevnou pravděpodobností úspěchu, označenou π. Toto rozdělení počítá pravděpodobnost toho, že při opakovaných nezávislých pokusech bude první úspěch předcházet x neúspěchů. Situaci můžeme demonstrovat na následujícím příkladu: Sledujeme silniční síť s řízenými křižovatkami. Sítí projíždí automobil a pravděpodobnost, že ho každý jednotlivý semafor zastaví je stejná a stálá, rovna 0.62. Jaká je pravděpodobnost, že automobil bude dvakrát zastaven, než se mu podaří projet bez zastavení?

Pravděpodobnostní funkce

$$f(x) = \pi(1-\pi)^x, \ x = 0, 1, 2, \cdots$$

Příklad

Dopočítáme příklad z úvodu ke geometrickému rozdělení o projíždění automobilu přes semafory. V tomto příkladě je úspěch projetí automobilu bez zastavení. Pravděpodobnost úspěchu je $\pi = 1 - 0.62 = 0.38$. Sledovaná realizace je $x = 2$ (dvě zastavení před volným průjezdem). Dosazením do pravděpodobnostní funkce dostáváme

$$f(2) = 0.38(1-0.38)^2 = 0.146$$

Pravděpodobnost, že automobil bude dvakrát zastaven než se mu podaří projet bez zastavení je tedy 0.146.

3.1.5 Negativně binomické rozdělení

Negativně binomické rozdělení je rozšířením geometrického. Vychází také z alternativního pokusu a podobně jako geometrické rozdělení určuje pravděpodobnost toho, že x neúspěchů předchází n-tý úspěch. Přitom předchozích $n-1$ úspěchů mohlo nastat kdykoli od začátku provádění pokusů. Ilustrační příklad může být obdobný jako pro geometrické rozdělení. Sledujeme silniční síť s řízenými křižovatkami. Sítí projíždí automobil a pravděpodobnost, že ho každý jednotlivý semafor zastaví je stejná a stálá, rovna 0.62. Jaká je pravděpodobnost, že automobil bude dvakrát zastaven, než se mu podaří potřetí projet bez zastavení?

Pravděpodobnostní funkce

$$f(x) = \binom{x+n-1}{x} \pi^n (1-\pi)^x, \ x = 0, 1, 2, \cdots$$

Příklad

Opět budeme pokračovat v načatém příkladu o autu, které projíždí semafory. Zadání je $x = 2$, $\pi = 0.38$ a $n = 3$. Dosazením dostaneme

$$f(2) = \binom{2+3-1}{2} 0.38^3 (1-0.38)^2 = 0.127$$

3.1.6 Rovnoměrné rozdělení

Diskrétní rovnoměrné rozdělení patří k základním a nejjednodušším. Na něm jsou založeny nejznámější „školní“ příklady o házení mincí nebo kostkou. Jedná se rozdělení k konečným počtem n realizací které nasávají se stejnou pravděpodobností $\frac{1}{n}$. Rovnoměrnost rozdělení vyjadřuje skutečnost, že v realizacích nemáme žádné preference nebo, že o nich nic nevíme.

Pravděpodobnostní funkce

$$f(x) = \frac{1}{n}, \;\; x = 1, 2, \cdots, n$$

Příklad

(i) Hod hrací kostkou - náhodná veličina je číslo, které padne. Pokus má šest realizací, každá má pravděpodobnosti $\frac{1}{6}$. (ii) Záznam poslední cifry na SPZ projíždějících automobilů. Máme deset cifer se stejnou pravděpodobností.

3.2 Spojité rozdělení

> Spojité rozdělení má nekonečný nespočetný počet realizací, tj. jeho realizace jsou z reálné osy.

Příklady: chyby v měření rozměru součástky, doba čekání na tramvaj, bezporuchová doba fungování přístroje, intenzita dopravního proudu (je vlastně diskrétní, ale lze ji aproximovat spojitým - půlky aut)

3.2.1 Rovnoměrné rozdělení

O rovnoměrném rozdělení jsme mluvili už v diskrétních rozděleních. Nyní se dostáváme k jeho spojité variantě.

Rovnoměrné rozdělení má dvě důležité charakteristiky
(i) všechny realizace jsou rovnocenné (žádný výsledek není preferován) a
(ii) ostré hranice pro definiční obor (hranice nelze překročit).

Hustota pravděpodobnosti

$$f(x) = \frac{1}{b-a}, \;\; x \in (a, b), \quad a, b \in R$$

Příklady

(i) Sledujeme příčnou polohu závodního automobilu na dráze. Všechny polohy jsou stejně dobré a přípustné, překročení hranice znamená katastrofu.

(ii) Sledujeme úbytek vzorku na pneumatikách autobusů MHD. Výška vzorku je dobrá a stejně pravděpodobná v daných mezích. Horní mez je dána konstrukcí pneumatiky, dolní mez je určena předpisy a je zakázána (alespoň ve smyslu zaplacení pokuty).

(iii) Sledujeme stav paliva v nádrži automobilu. Horní mez je dána velikostí nádrže, dolní mez je nula (pokud má člověk v autě kanystr s benzínem).

(iv) Sledujeme dobu čekání na tramvaj, která má pevný a stálý interval, pro náhodně přicházející osoby. Dolní hranice je 0 (osoba tramvaj zrovna doběhla) a horní mez je rovna intervalu (dané osobě tramvaj právě ujela a ona musela čekat na další).

3.2.2 Normální rozdělení

Normální rozdělení je nejčastěji používaným rozdělením. Důvody k tomu jsou dva: (i) normální rozdělení vzniká při spolupůsobení velkého množství nezávislých malých náhod, takže řada ne přesně definovaných neurčitostí se mu značně podobají - např. když se sype suchý písek, zrníčka na sebe navzájem mnohokráte narazí a výsledkem je hromada, která dosti připomíná „dvourozměrnou gaussovku“, (ii) normální rozdělení je jedním z mála, pro které se dají analyticky dopočítat i složitější statistické úlohy - jinak je často třeba dělat určité aproximace.

Hustota pravděpodobnosti

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

kde parametry rozdělení jsou μ - střední hodnota a σ^2- rozptyl.

Budeme-li normálně rozdělenou náhodnou veličinu X normovat tak, aby měla střední hodnotu nula a rozptyl jedna, dostaneme náhodnou veličinu Z se standardním normálním rozdělením. Normovací vztah je $z = (x - \mu)/\sigma$ a odpovídající hustota pravděpodobnosti

$$f(z) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}z^2}$$

Komentář

Ačkoli je normální rozdělení tak časté a pro svoje dobré výpočetní vlastnosti oblíbené, je s ním trochu potíž. K hustotě pravděpodobnosti totiž neexistuje primitivní funkce. Pravděpodobnosti intervalů, což jsou integrály z hustoty pravděpodobnosti přes tyto intervaly, a stejně tak hodnoty distribuční funkce nelze analyticky počítat. Hodnoty distribuční funkce pro standardní normální rozdělení byly vypočteny numericky a jsou uvedeny ve statistických tabulkách (viz Tabulky na stránce 147)

Příklad 1

Z vojenského tábora má do sousedního stanoviště dojet nákladní automobil. Po cestě se nachází most o výšce 4 metry. Vojenské nákladní automobily mají výšku s normálním rozdělením se střední hodnotou 3.8 metru a směrodatnou odchylkou 0.3 metru. Jaká je pravděpodobnost, že vyslané nákladní auto pod mostem projede?

Ze zadání tedy plyne, že výška auta je náhodná veličina $X \sim N\left(3.8,\, 0.3^2\right)$, tj. s normálním rozdělením se střední hodnotou $\mu = 3.8$ a rozptylem $\sigma^2 = 0.3^2 = 0.09$. Ptáme se, na pravděpodobnost $P\left(X < 4\right)$. Pro určení této pravděpodobnosti hodláme použít tabulky na straně 147. Ty však platí pouze pro normované normální rozdělení. Musíme tedy úlohu převést do normované oblasti, tj. náhodnou veličinu X a s ní související údaje musíme přepočítat na náhodnou veličinu $Z \sim N\left(0,\, 1\right)$. Vzorec pro přepočet je

$$Z = \frac{X - \mu}{\sigma} \tag{7}$$

Pro hledanou pravděpodobnost nyní platí

$$P\left(X < 4\right) = P\left(\frac{X - \mu}{\sigma} < \frac{4 - 3.8}{0.3}\right) = P\left(Z < 0.67\right) = 0.749$$

Tedy pravděpodobnost, že vojenský automobil projede pod mostem je 0.749.

Příklad 2

Při kontrole výrobků je sledován podélný rozměr, který má střední hodnotu 335 mm a rozptyl 42 mm^2. Jaká je pravděpodobnost, že bude vyroben zmetek, tj. výrobek s podélným rozměrem větším než 350 mm?

Postup řešení je stejný jako v předchozím příkladě. Sledovaná náhodná veličina je $X \sim N\left(335,\, 42\right)$. Ptáme se na pravděpodobnost $P\left(X > 350\right)$. Normováním dostaneme

$$P\left(X > 350\right) = P\left(Z > \frac{350 - 335}{\sqrt{42}}\right) = P\left(Z > 2.31\right)$$

Najdeme tabulky normálního rozdělení a ihned vidíme, že je malér. Tabulky ukazují pravděpodobnosti jen pro $X < x$. Musíme tedy naši pravděpodobnost převést do požadovaného tvaru, což není složité

$$P\left(X > 350\right) = 1 - P\left(X < 350\right) = 1 - 0.99 = 0.001$$

Z tisíce vyrobených výrobků bude tedy v průměru jeden zmetek.

Příklad 3

Na městské komunikaci nedaleko od řízené křižovatky měříme rychlosti projíždějících automobilů a uvažujeme o tom, zda tyto rychlosti můžeme popsat pomocí normálního rozdělení.

Ověření typu rozdělení je předmětem testování, ke kterému se dostaneme až později ve statistice. Nyní se budeme držet jen základních rysů, které by normální rozdělení mělo mít.

Budeme-li situaci sledovat, ihned zjistíme, že rychlosti automobilů se významně liší v situaci, kdy do měřeného místa zasahuje z křižovatky kolona, a jestliže žádná kolona není nebo je tak malá, že na měřené místo nemá vliv. Rozhodneme-li se popisovat rychlosti bez ohledu na kolonu, dostaneme velmi hrubý popis, který nepopisuje dobře ani jednu ze zmíněných situací. Bude proto lépe situaci rozdělit na dva případy: kolona má velký vliv a kolona má malý vliv.

Všimneme si nejprve prvního stavu, kdy kolona prakticky zasahuje na měřené místo a způsobuje, že často naměříme rychlost nula. V důsledku toho, bude střední hodnota blízká nule. Vzhledem k tomu, že normální rozdělení je symetrické, bude jeho hustota pravděpodobnosti svou levou částí zasahovat do záporné poloosy realizací a tím tvrdíme, že mohou nastat i záporné rychlosti (mají nenulovou pravděpodobnost). V tomto případě je tedy normální rozdělení nevhodné.

V druhém případě jsou rychlosti aut větší (zřejmě někde kolem maximální povolení rychlosti). Průměrná rychlost bude taky velká a odchylky od ní budou na obě strany, jak kladné (auta překračující povolenou rychlost), tak i záporné (ti, co telefonují, hledají neznámou adresu atd.). V tomto případě lze dobře předpokládat, že nulové rychlosti se nebudou vyskytovat. Gaussova křivka, která popisuje hustotu pravděpodobnosti normálního rozdělení bude svým vrcholem nad střední hodnotou a levou částí bude zasahovat do záporných hodnot až v místě, kdy je prakticky nulová, a to tak, že i hodnota integrálu přes záporné hodnoty bude přibližně nula. V tomto případě je pro popis rychlostí projíždějících aut možno použít normální rozdělení.

3.2.3 Log-normální rozdělení

Normální rozdělaní, kterým jsme se právě zabývali, má jednu nevýhodu. Je symetrické, definované na celé reálné ose a tedy, připouští i záporné realizace náhodné veličiny. Přitom nezápornost je přirozenou vlastností celé řady reálných veličin, např. délka, počet (tj. intenzita i hustota dopravního proudu), rychlost a mnoho dalších. S touto nevýhodou se částečně vypořádává log-normální rozdělení. Jeho realizace dostaneme transformací normálního rozdělení $U \sim N(\mu, \sigma^2)$ podle rovnice $X = \mathrm{e}^U$. Toto rozdělení se pro velká μ chová přibližně jako normální, pro malá μ je nesymetrické. Obor hodnot, na kterém je jeho hustota pravděpodobnosti nenulová je množina nezáporných reálných čísel. Za toto vylepšení se ale platí složitějším popisem a jednoduchost při úpravách se většinou ztrácí.

Hustota pravděpodobnosti

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

Toto rozdělení se hodí např. pro modelování dopravního proudu. Pro velké intenzity mohou být změny kladné i záporné (blízké normálnímu rozdělení), při nulové intenzitě mohou být změny jen kladné.

3.2.4 Exponenciální rozdělení

Exponenciální rozdělení je spojitou obdobou geometrického, které popisuje počet neúspěchů, které je třeba absolvovat, než dojde ke změně stavu a obdržíme úspěch. Podobně exponenciální rozdělení popisuje např. dobu fungování určitého přístroje, než dojde ke změně stavu - přístroj se porouchá. Tomuto jevu se běžně říká bezporuchová doba fungování přístroje. Zde se ale jedná o velmi speciální případ, kdy porucha je způsobena nikoli stárnutím, ale nějakou vnější příčinou, která přístroj v čase ohrožuje se stále stejnou intenzitou. Z této skutečnosti plyne vlastnost exponenciálního rozdělení, která říká, že toto rozdělení nemá paměť. To lze vyjádřit vztahem $P\left(X > t+s | X > t\right) = P\left(X > s\right)$, který říká: jestliže přístroj fungoval až do času t, pak pravděpodobnost, že bude ještě dále fungovat po dobu s je stejná, jako pravděpodobnost, že bude fungovat po dobu s, aniž byl před tím v provozu. Tato vlastnost je dokázána v Dodatku 13.2.

Hustota pravděpodobnosti

$$f\left(x, \delta\right) = \frac{1}{\delta} \mathrm{e}^{-\frac{x}{\delta}}, \ x \geq 0$$

Příklad: Životnost závodního automobilu: Jaká je pr., že závodní automobil dokončí tříhodinový závod, jestliže jeho střední životnost na plný výkon je 2 hodiny a 15 minut?

Hp je $f\left(x\right) = \frac{1}{2.25} \mathrm{e}^{-x/2.25}$. Dále $P\left(X < 3\right) = \int_0^3 f\left(x\right) dx =$

$$= \int_0^3 \mathrm{e}^{-x/2.25} dx = \frac{1}{2.25} \left(-2.25\right) \left[\mathrm{e}^{-x/2.25}\right]_0^3 = 1 - \mathrm{e}^{-3/2.25} = 0.736$$

3.2.5 χ^2rozdělení

$$\chi^2\left(n\right) = \sum_{i=1}^{n} \left(N_i\left(0,1\right)\right)^2$$

Pozn.: Toto rozdělení doprovází veličiny, které jsou tvořeny kvadráty - rozptyl, kriterium ve tvaru součtu čtverců $\cdots$

3.2.6 Studentovo rozdělení

$$St\left(n\right) = \frac{N\left(0,1\right)}{\chi^2\left(n\right)/n}$$

Pozn.: Generátor má jmenovatel. Bude-li v něm malé číslo (což může), bude hodnota realizace velká - rozdělení s těžkými konci.

3.2.7 F rozdělení (jen generátor i v Matlabu)

$$F(m,n) = \frac{\chi_1^2(m)/m}{\chi_2^2(n)/n}$$

Pozn.: Popisuje veličinu ve tvaru podílu dvou rozptylů. Typicky se jedná o rozptyl vysvětlený a nevysvětlený při zkoumání závislosti dvou nebo více veličin.

Příklady

1. Kvantil ζ_α, kritická hodnota z_α, pr. vlevo $p^-(x)$, pr. vpravo $p^+(x)$

$$\zeta_\alpha : \int_{-\infty}^{\zeta_\alpha} f(x)\,dx = \alpha; \qquad z_\alpha : \int_{z_\alpha}^{\infty} f(x)\,dx$$
$$p^-(x) = P(X \le x) \qquad p^+(x) = P(X > x)$$

   Pro $N(0,1)$ platí:

$$z_\alpha = -\zeta_\alpha = \zeta_{1-\alpha} \quad p^+(x) = 1 - p^-(x) = p^-(-x)$$

   Z tabulek jsme zjistili, že $P(X \le 1.645) = 0.95$. Určete:

   (a) $P(X > 1.645)$ [0.05]; $P(X \le -1.645)$ [0.05]; $P(X > -1.645)$ [0.95];

   (b) $\zeta_{0.05}$ [−1.645]; $z_{0.05}$ [1.645]; $z_{0.95}$ [−1.645] $\cdots$ Nakreslit !

2. V dodávce 1000 výrobků je 50 vadných.

   (a) Jaká je pravděpodobnost, že ve výběru 80 výrobků bude 5 vadných?

   (b) Jaký bude průměrný počet vadných výrobků v opakovaném výběru 80 ks?

   (c) Provedeme 100 výběrů po 80 kusech. Jaký je očekávaný počet výběrů s pěti vadnými výrobky?

   Řešení:

   (a) Úplně **přesně**:

$$P = C_5(50)C_{75}(950)/C_{80}(1000)$$

   (nejde spočítat.)
   Aproximace **binomickým** rozdělením s $n = 80$ a $\pi = 50/1000 = 0.05$. Hledaná pravděpodobnost je

$$P(5) = \binom{80}{5} 0.05^5 0.95^{75} = 0.1603.$$

   Další aproximace **Poissonovým** rozdělením (n velké, π malé) s $\lambda = n\pi = 4$.

$$f(5) = \mathrm{e}^{-4}\frac{4^5}{5!} = 0.1563.$$

(b) Bude to střední (očekávaná) hodnota $E[K] = n\pi = \lambda = 4$.

(c) Podle statistické definice je pravděpodobnost rovna počtu úspěšných pokusů dělenému celkovým počtem pokusů, tedy

$$P(5) = \frac{N^+}{N} \Rightarrow 0.16 = \frac{N^+}{100} \Rightarrow N^+ \doteq 16$$

3. Pravděpodobnost výroby vadného výrobku je 0.01%. Nový pracovník vyrobil 5000 výrobků a z nich byly 3 vadné. Je tento počet normální, nebo lze předpokládat, že je způsoben nezkušeností pracovníka?

Řešení:

Náhodná veličina "počet vadných výrobků v sérii 5000" má Poissonovo rozdělení (n velké, π malé) s $\lambda = 5000 \times 0.0001 = 0.5$. Potom pravděpodobnost, že za běžných okolností (tj. s běžnou pravděpodobností vyrobení vadného výrobku $\pi = 0.01$) je pr. tří a více vadných výrobků rovna

$$P(K > 3) = 1 - P(0) - P(1) - P(2) = 1 - \mathrm{e}^{-0.5}(1 + 0.5 + \frac{0.5^2}{2}) = 0.014$$

Tj. pravděpodobnost, že za normálních okolností budou v sérii 5000 výrobků 3 nebo více vadných je 0.014; jinak řečeno, jen 1.4% pracovníků by mělo tolik vadných výrobků. To ukazuje na nezkušenost nového pracovníka.

4. Měřící přístroj je zatížen jednak systematickou chybou 0.5, jednak náhodnou chybou se směrodatnou odchylkou 0.3.
a) S jakou pravděpodobností bude změřena hodnota s odchylkou v intervalu (0.4; 0.6)?
b) V jakých mezích lze očekávat odchylky od správné hodnoty s pravděpodobností 0.95?

Řešení:

a) Normální rozdělení. Předpokládáme, že máme k dispozici pouze tabulky pravděpodobností, kde jsou uvedeny jen pravděpodobnosti normovaného rozdělení. Budeme proto normovat. (Pozn.: např v Excelu lze řešit bez normování.)

$$Z = \frac{X - \mu}{\sigma}, \quad \Rightarrow \quad z_d = \frac{0.4 - 0.5}{0.3} = -0.33, \quad z_h = \frac{0.6 - 0.5}{0.3} = 0.33.$$

Požadovaná pr. je
$P(X \in (0.4;\ 0.6)) = P(Z \in (-0.33;\ 0.33)) = 1 - 2P(Z < -0.33) = 1 - 2 \times 0.37 \doteq 0.26.$

b) Je-li pravděpodobnost intervalu 0.95, je pravděpodobnost vpravo i vlevo rovna 0.025. Hledáme kritickou hodnotu $z_{0.025}$ (kvantil $\zeta_{0.025}$), tj hodnotu, pro kterou platí

$$P(Z > z_{0.025}) = 0.025, \qquad \text{nebo} \quad P(Z < \zeta_{0.025}) = 0.025$$

Z tabulek: $z_{0.025} = 1.96$ nebo $\zeta_{0.025} = -1.96$. Po odnormování (přechodu k původní náhodné veličině X) dostaneme

$$X = \mu + \sigma \times Z \quad \Rightarrow \quad x_d = 0.5 - 1.96 \times 0.3 = -0.088, \quad x_d = 0.5 + 1.96 \times 0.3 = 1.088.$$

Chyba bude s pr. 0.95 ležet v intervalu (-0.088; 1.088).

POZNÁMKA: T*ento druh výpočtů je velmi důležitý. Bude základem pro řadu úloh ve statistice (intervaly spolehlivosti, testy hypotéz).*

5. Průměrná doba čekání na zákazníka je 50 s. Doba čekání se řídí exponenciálním rozdělením. Jaká je pravděpodobnost, že zákazník bude obsloužen v době kratší než 30 s?

   Řešení:

   Průměrná doba čekání je střední hodnota, tedy $\delta = 50$.

$$P(X < 30) = \int_0^{30} f(\xi)d\xi = F(30) = 1 - \mathrm{e}^{-\frac{1}{50}30} = 0.45.$$

4 Vektorová náhodná veličina

4.1 Náhodný vektor

V komentáři k definici náhodné veličiny jsme uvedli, že náhodnou veličinu lze chápat jako ekvivalent náhodného pokusu. Z pokusu dostáváme výsledky, z náhodné veličiny generujeme čísla, která jsou výsledkům přiřazena. Tato představa však odpovídá jen v případě, kdy experiment se týká jen jedné veličiny. Například: hod kostkou, kde sledovaná veličina je počet ok, která padla, nebo, měření intenzity dopravního proudu, kde sledovanou veličinou je intenzita (počet aut, která projedou za jednotku času). To ale není jediný typ náhodného pokusu. Ten se může týkat více než jedné veličiny. Jedná se například o pokus hod dvěma kostkami, kde první veličina je počet ok na první kostce a druhá veličina jsou oka na druhé kostce. Stejně tak, lze uvažovat experiment, při kterém měříme intenzitu provozu na více místech sledované oblasti. Každé měřené místo je potom spojeno s jednou veličinou.

Obecně lze tuto situaci formulovat takto. Náhodný pokus spočívá ve sledování určitého procesu. Náhodná veličina je pak určena měřícím místem na tomto procesu. Procesem může být např. dopravní síť a náhodné veličiny - intenzity dopravního proudu - jsou spojeny s jednotlivými měřicími místy - měřicími detektory.

Příklad

Uvažujeme čtyř-ramennou křižovatku, kde v každém rameni ve směru vjezdu je umístěn měřící detektor. Na každém z detektorů měříme intenzitu dopravního proudu. Protože všechna měření jsou pod vlivem neurčitosti (drobné poruchy a nepravidelnosti), lze intenzity spojené s jednotlivými detektory považovat za náhodné veličiny. Konkrétní měření jsou realizace těchto náhodných veličin. Jak náhodné veličiny, tak i jejich realizace lze uspořádat do vektorů.

4.1.1 Náhodný vektor

Náhodný vektor X je vektor náhodných veličin $X_1, X_2, \cdots, X_n$,

$$X = [X_1, X_2, \cdots, X_n]'$$

Komentář k definici

Na náhodnou veličinu je možno pohlížet jako na „měřící místo“ na procesu, který sledujeme, který ve svém chování vykazuje neurčitost a který se snažíme poznat. V praktickém případě si můžeme představit např. dopravní oblast, kterou projíždí automobily a my sledujeme jejich rychlosti. Je celkem logické, že pokud je oblast větší, budeme se snažit měřit rychlost nejen na jednom místě, ale na několika místech najednou. Pracujeme tedy nejen s jednou náhodnou veličinou, ale s celou množinou náhodných veličin. Náhodný vektor představuje takovou množinu náhodných veličin, kterou jsme navíc seřadili do uspořádaného útvaru - do vektoru.

Poznámka

Uvažování více náhodných veličin ve tvaru vektoru umožní stručný zápis s využitím maticové symboliky. Vektor uvažujeme jako sloupcový kvůli přehlednosti dalších vzorců (apostrof značí transpozici).

4.2 Rozdělení náhodného vektoru

Rozdělení náhodného vektoru (podobně jako tomu bylo u náhodné veličiny - viz odstavec 2.2) dává úplný popis náhodného vektoru v pravděpodobnostním smyslu. Tento popis se tedy sestává ze specifikace množiny, ze které pocházejí hodnoty náhodného vektoru a z přidělení pravděpodobnosti prvkům příslušného borelovského pole, tedy systému různých podmnožin oboru hodnot náhodného vektoru. Tyto podmnožiny jsou (podobně jako pro skalární náhodnou veličinu) definovány vztahem $X_1 \leq x_1 \wedge X_2 \leq x_2 \wedge \cdots \wedge X_n \leq x_n$.

4.3 Sdružené, marginální a podmíněné rozdělení

Sdružené rozdělení

Stejně jako pro náhodnou veličinu (2), tak i pro náhodný vektor $X = [X_1, X_2, \cdots, X_n]'$ je třeba definovat úplný popis. Při jeho definici vyjdeme z pravděpodobnosti průniku jevů (1) a pro X budeme definovat sdruženou distribuční funkci $F_X(x_1, x_2, \cdots, x_n)$ jako pravděpodobnost oblasti, na které je $X_1 \leq x_1 \wedge X_2 \leq x_2 \wedge \cdots \wedge X_n \leq x_n$. Srovnejte s (4).

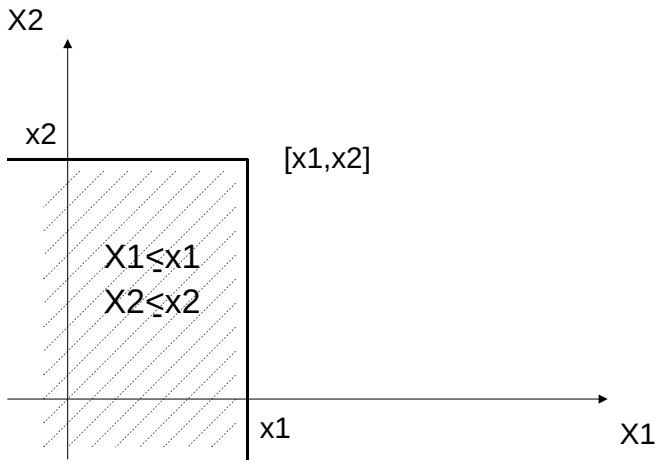
4.3.1 Sdružená distribuční funkce

Pro náhodný vektor $X = [X_1, X_2, \cdots, X_n]$' definujeme sdruženou distribuční funkci jako reálnou funkci n reálných argumentů předpisem

$$F_X(x_1, x_2, \cdots, x_n) = P(X_1 \leq x_1 \wedge X_2 \leq x_2 \wedge \cdots \wedge X_n \leq x_n)$$

Komentář k definici

Sdružená distribuční funkce je definována na borelovském poli v R^n (viz odstavec 13.1). Toto borelovské pole je generováno útvary popsanými nerovnostmi $X_1 \leq x_1 \wedge X_2 \leq x_2 \wedge \cdots \wedge X_n \leq x_n$ ze kterých je možno pomocí operací doplněk, sjednocení a průnik vytvořit všechny potřebné podmnožiny. V případě jedné náhodné veličiny je situace popsána v odstavci 13.1, pro dvě náhodné veličiny je obecný generující prvek naznačeny na následujícím obrázku.



Pomocí takových „rohů“, jejich doplňků, sjednocení a průniků lze konstruovat úsečky, obdélníky, pravoúhlé mnohoúhelníky a v limitním přechodu prakticky libovolné plošné oblasti a jejich sjednocení. Bohatství těchto útvarů, které představují obrazy jevů definovaných na sledovaných náhodných pokusech, naznačuje, že zobrazení z množiny jevů do množiny těchto útvarů (podobně jako tomu bylo pro jednu náhodnou veličinu) prakticky vždy existuje.

Dvourozměrné rozdělení

Obecné n-rozměrné rozdělení nelze jednoduše zobrazit, protože existuje v $(n+1)$-rozměrném prostoru - každá náhodná veličina představuje jednu dimenzi prostoru a v dimenzi $n+1$ bychom zobrazovali vlastní distribuční funkci. Proto se budeme zabývat především dvourozměrným náhodným vektorem $X = [X_1, X_2]$. Jeho **distribuční funkce** je

$$F_X(x_1, x_2) = P(X_1 \leq x_1,\ X_2 \leq x_2),$$

kde čárka , představuje konjunkci $\wedge$.

Hustotu pravděpodobnosti pro diskrétní resp. spojitý náhodný vektor $X = [X_1, X_2]$ dostaneme analogicky podle jedné náhodné veličiny

$$f_X = P(X_1 = x_1, X_2 = x_2), \quad \text{resp.} \quad f_X(x_1, x_2) = \frac{\partial^2 F_X(x_1, x_2)}{\partial x_1 \partial x_2}$$

nebo v integrálním tvaru

$$F_X = \sum_{t_1 \le x_1} \sum_{t_2 \le x_2} f_X(t_1, t_2), \quad \text{resp.},$$

$$F_X(x_1, x_2) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f_X(t_1, t_2)\, dt_1 dt_2$$

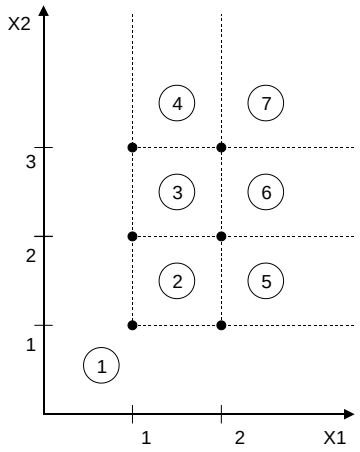
Pro **nezávislé náhodné veličiny** platí

$$f_X(x_1, x_2) = f_{X_1}(x_1)\, f_{X_2}(x_2).$$

Příklad

Uvažujme náhodný pokus, při kterém házíme levou rukou minci a pravou rukou kostku. S hodem mincí je spojena náhodná veličina X_1 s hodnotami 1 (rub) a 2 (líc) a s hodem kostkou náhodná veličina X_2 s hodnotami 1, 2, 3, kde 1 odpovídá výsledku „padne 1“, 2 zahrnuje výsledky „padne 2 nebo 3“ a 3 reprezentuje výsledky „padne 4 nebo 5 nebo 6“. Výsledkem pokusu je uspořádaná dvojice $[X_1, X_2]$. Máme určit distribuční funkci náhodného vektoru $X = [X_1, X_2]$.

Hodnoty distribuční funkce $F_X(x_1, x_2)$ budeme počítat pro jednotlivé oblasti, tvořené body - hodnotami náhodného vektoru $X = [X_1, X_2]$ vymezené nerovnostmi $X_1 \le x_1$ a $X_2 \le x_2$. Při pohledu „shora“ bude zkoumaná oblast $[X_1, X_2]$ rozdělena tak, jak je ukázáno na následujícím obrázku



Černé tečky představují uspořádané dvojice, které tvoří realizace náhodného vektoru X. Jsou to body $[1,1]$, $[1,2]$, $[1,3]$, $[2,1]$, $[2,2]$, $[2,3]$. Pravděpodobnosti jednotlivých bodů jsou

dány součinem pravděpodobnosti jednotlivých složek náhodného vektoru (složky jsou nezávislé):

$$P([1,1]) = P([X_1 = 1 \wedge X_2 = 1]) = P(X_1 = 1)\,P(X_2 = 1) = \frac{1}{2}\cdot\frac{1}{6} = \frac{1}{12},$$
$$P([1,2]) = \frac{1}{6},\ P([1,3]) = \frac{1}{4},\ P([2,1]) = \frac{1}{12},\ P([2,2]) = \frac{1}{6},\ P([2,3]) = \frac{1}{4}.$$

Pravděpodobnosti jednotlivých oblastí, na obrázku naznačených odlišným odstínem než je pozadí a rozlišeny číslem v kroužku, jsou vždy dány součtem pravděpodobností realizací, které jsou směrem vlevo a dolů od dané oblasti.

Tak pravděpodobnost přiřazená oblasti (1) je 0 - protože „vlevo a dolů nic není“.

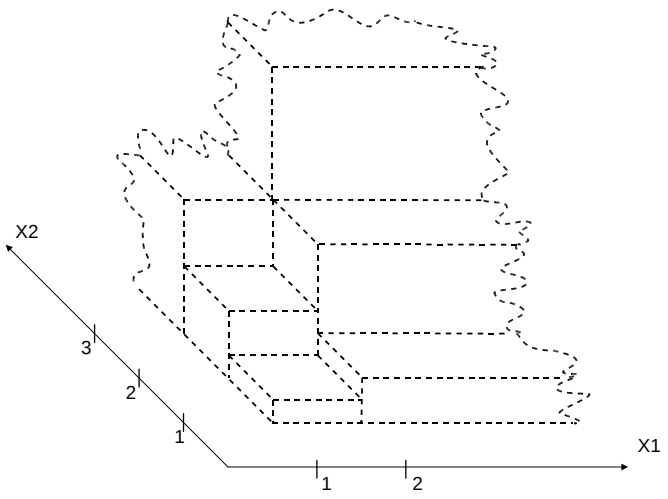
Oblast (2) má pravděpodobnost $\frac{1}{12}$, protože „vlevo a dolů“ leží realizace $[1,1]$, která má pravděpodobnost $\frac{1}{12}$.

Oblast (3) má pravděpodobnost $\frac{1}{12} + \frac{1}{6} = \frac{1}{4}$, protože „vlevo dole“ leží realizace $[1,1]$ a $[1,2]$ a pravděpodobnostmi $\frac{1}{12}$ a $\frac{1}{6}$.

Dále podle stejného principu, budou pravděpodobnosti oblastí následující

$$P((4)) \to \frac{1}{2},\ P((5)) \to \frac{1}{6},\ P((6)) \to \frac{1}{2},\ P((7)) \to 1.$$

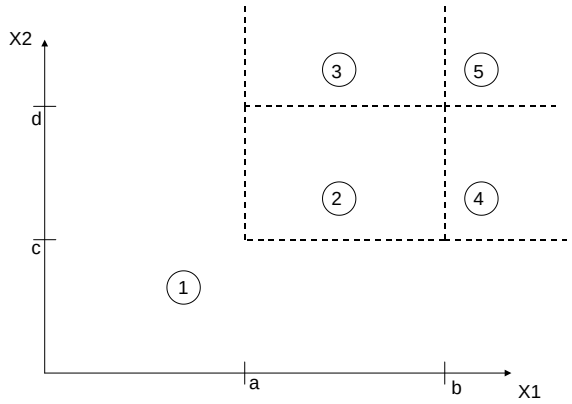
Celkový pohled na distribuční funkci je na obrázku



Příklad

Uvažujme dvě náhodné veličiny X_1 a X_2 tvořící náhodný vektor $X = [X_1, X_2]$. První má hodnoty rovnoměrně rozděleny na intervalu (a, b), druhá také rovnoměrně, ale na intervalu (c, d). Obě náhodné veličiny jsou nezávislé. Máme určit distribuční funkci $F_X(x_1, x_2)$.

Zadání je graficky znázorněno na následujícím obrázku



Opět dostáváme oblasti, na kterých se předpis pro distribuční funkci bude lišit. Tak na oblasti (1) bude distribuční funkce nulová, protože všechny její body $[x_1, x_2]$ mají „vlevo a dole“ (tj. pro dvojice $[X_1, X_2]$, pro které platí $X_1 \leq x_1 \wedge X_2 \leq x_2$) oblasti s nulovou pravděpodobností - $X_1 \leq a$ a $X_2 \leq c$ mají nulovou pravděpodobnost.

Dále se podíváme na oblast (2). Tady má nenulovou pravděpodobnost jak náhodná veličina X_1, tak i X_2. Díka rovnoměrnému rozdělení pravděpodobnosti obou náhodných veličin můžeme říci, že čím postoupíme více doprava a nahoru, tím bude pravděpodobnost oblasti „vlevo a dole“ větší. Konkrétně, označíme-li aktuální bod $x = [x_1, x_2]$, bude platit

$$F_X(x_1, x_2) = P(X_1 \leq x_1 \wedge \ X \leq x_2) = \frac{x_1 - a}{b - a} \cdot \frac{x_2 - c}{d - c}.$$

Budeme-li postupovat nahoru, narazíme na oblast (3). Tady se již dostáváme za pravou mez náhodné veličiny X_2 a pravděpodobnost oblasti dole bude jedna. Pro distribuční funkci tedy platí

$$F_X(x_1, x_2) = \frac{x_1 - a}{b - a}.$$

Obdobně to bude na oblasti (4)

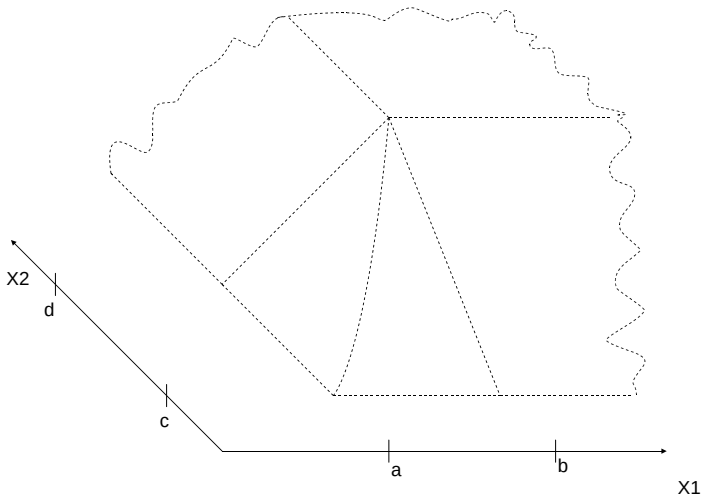
$$F_X(x_1, x_2) = \frac{x_2 - c}{d - c}.$$

Oblast (5) je již za horními hranicemi obou náhodných veličin, a tedy celková pravděpodobnost „vlevo a dole“ musí být $F(x_1, x_2) = 1$.

Dostáváme tedy distribuční funkci ve tvaru

$$F_X(x_1, x_2) = \begin{cases} 0 & \text{pro } x_1 \leq a \wedge x_2 \leq c \\ \frac{x_1-a}{b-a} \cdot \frac{x_2-c}{d-c} & \text{pro } x_1 \in (a,b) \wedge x_2 \in (c,d) \\ \frac{x_1-a}{b-a} & \text{pro } x_1 \in (a,b) \wedge x_2 \geq d \\ \frac{x_2-c}{d-c} & \text{pro } x_1 \geq b \wedge x_2 \in (c,d) \\ 1 & \text{pro } x_1 \geq b \wedge x_2 \geq d \end{cases}$$

Výsledek je schematicky zachycen na dalším obrázku



Příklady

1. Hod dvěma mincemi, kde $R \to 0$ a $L \to 1$ a mince rozlišujeme. Potom

$$\Omega = \{[0;0], [0;1], [1;0], [1,1]\}$$

$$A = \{\emptyset, \{[0;0]\}, \{[0;1]\}, \{[1;0]\}, \{[1,1]\},$$
$$\{[0;0], [0;1]\}, \{[0;0], [1;0]\}, \{[0;0], [1;1]\}, \{[0;1], [1;0]\}, \{[0;1], [1;1]\}, \{[1;0], [1;1]\},$$
$$\{[0;0], [0;1], [1;0]\}, \{[0;0], [0;1], [1;1]\}, \{[0;0], [1;0], [1;1]\}, \{[0;1], [1;0], [1;1]\}, \Omega\}$$

$$P: [0;0] \to \frac{1}{4}, [0;1] \to \frac{1}{4}, [1;0] \to \frac{1}{4}, [1;1] \to \frac{1}{4}.$$

2. Na třech vybraných místech v Praze sledujeme stupeň dopravy $s = 1$ - malá zátěž, 2 - velká zátěž. Sledovaná datová položka (data item) je $x = [s_1, s_2, s_3]'$. Změřený vektor dat považujeme za realizaci náhodného vektoru $X = [S_1, S_2, S_3]'$. Určete pr., že stupně budou tak a tak a tak..

3. Sledujeme provoz na vybrané řízené křižovatce. V každém rameni křižovatky je umístěna kamera, která poskytuje aktuální obraz provozu v křižovatce a tak umožňuje měřit délky kolon. Navíc máme k dispozici řídící plány ze SSZ. Data snímáme v diskrétních okamžicích t jako 5ti rozměrný vektor - $x = [q_1, q_2, q_3, q_4, z, t]'$, kde ... Protože naměřená data nejsou vždy stejná, považujeme je za realizace vektorové náhodné veličiny $X = [Q_1, Q_2, Q_3, Q_4, Z, T]'$, která představuje sledovaná data na křižovatce.
   Pr., že kolony budou tak a tak a zelená bude tak a čas bude tak, ale nevíme, kdy to bylo změřeno.

Marginální rozdělení

Uvažujme následující pokus: Na n místech dopravní sítě měříme rychlosti projíždějících automobilů. Takový pokus je spojen s n-rozměrným náhodným vektorem jehož složky jsou náhodné veličiny odpovídající jednotlivým měřeným místům. Tento náhodný vektor je popsán sdruženou hustotou pravděpodobnosti a tato hustota obsahuje veškerou pravděpodobnostní informaci jak o jednotlivých náhodných veličinách nebo jejich skupinách, tak i o vztazích mezi nimi.

Nejdříve se budeme zabývat otázkou, jak se lze s pomocí sdružené hustoty pravděpodobnosti dostat k popisu jednotlivých náhodných veličin a k popisu jejich různých skupin. Odpověď na tuto otázku dává definice marginální hustoty pravděpodobnosti.

4.3.2 Marginální hustota pravděpodobnosti

Uvažujme náhodný vektor $X = [X_1, X_2, \cdots, X_n]'$ a celé číslo $0 < k < n$. Z indexů 1, 2, $\cdots$, n vybereme k-prvkovou podmnožinu i_1, $i_2, \cdots$, i_k. Zbylé indexy budou i_{k+1}, $i_{k+2}, \cdots$, i_n. Marginální hustotu pravděpodobnosti definujeme vztahem

$$f_{X_{i_1}, X_{i_2}, \cdots, X_{i_k}}(x_{i_1}, x_{i_2}, \cdots, x_{i_k}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_X(x_1, x_2, \cdots, x_n)\, dx_{i_{k+1}} dx_{i_{k+2}} \cdots dx_{i_n}$$

Pro $n = 2$, tedy dvě náhodné veličiny, je sdružená hustota pravděpodobnosti $f_X(x_1, x_2)$ a ta má dvě dvě marginální hustoty pravděpodobnosti

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} f_X(x_1, x_2)\, dx_2 \quad \text{a} \quad f_{X_2}(x_2) = \int_{-\infty}^{\infty} f_X(x_1, x_2)\, dx_1$$

Příklady

1. Jaká je pr, že na první minci padne ...
2. Jaká je pr., že na prvním sledovaném místě bude stupeň 1
3. Jaké je rozdělení pr. délek kolon (bez ohledu na zelené a čas) - ! zejména nezávislost na času je ilustrativní pro výklad marginální hp !

Podmíněné rozdělení

Uvažujme náhodný vektor $X = [x_1, x_2, \cdots, x_n]$ a jeho sdružené rozdělení $f_X(x_1, x_2, \cdots, x_n)$. Již jsme hovořili o situaci, kdy nás zajímá jen určitá podmnožina náhodných veličin, jejíž indexy jsme označili $i_1, i_2, \cdots, i_k$. Nyní se zabýváme situací, kdy pro vybrané veličiny $[x_{i_1}, x_{i_2}, \cdots, x_{i_k}]$ jsme nějakým způsobem zjistili jejich skutečné realizace a s touto znalostí hledáme pravděpodobnosti ostatních, tj. rozdělení vektoru $\left[x_{i_{k+1}}, x_{i_{k+2}}, \cdots, x_{i_n}\right]$. O tomto rozdělení mluvíme jako o podmíněném rozdělení a značíme jej takto

$$f_{X_{i_{k+1}}, X_{i_{k+2}}, \cdots, X_{i_n} | X_{i_1}, X_{i_2}, \cdots, X_{i_k}} \left(x_{i_{k+1}}, x_{i_{k+2}}, \cdots, x_{i_n} | x_{i_1}, x_{i_2}, \cdots, x_{i_k}\right).$$

Jeho výpočet určuje následující definice

4.3.3 Podmíněná hustota pravděpodobnosti

Uvažujme náhodný vektor $X = [X_1, X_2, \cdots, X_n]'$ a celé číslo $0 < k < n$. Z indexů 1, 2, $\cdots$, n vybereme k-prvkovou podmnožinu $i_1, i_2, \cdots, i_k$. Zbylé indexy budou $i_{k+1}, i_{k+2}, \cdots, i_n$. Podmíněnou hustotu pravděpodobnosti, při znalosti realizací vektoru $[x_{i_1}, x_{i_2}, \cdots, x_{i_k}]$, definujeme vztahem

$$f_{X_{i_{k+1}}, X_{i_{k+2}}, \cdots, X_{i_n} | X_{i_{k+1} i_{k+1} i_1}, X_{i_2}, \cdots, X_{i_k}} \left(x_{i_{k+1}}, x_{i_{k+2}}, \cdots, x_{i_n}\right) =$$
$$= \frac{f_X(x_1, x_2, \cdots, x_n)}{f_{X_{i_{k+1}}, X_{i_{k+2}}, \cdots, X_{i_n} | X_{i_1}, X_{i_2}, \cdots, X_{i_k}} \left(x_{i_{k+1}}, x_{i_{k+2}}, \cdots, x_{i_n} | x_{i_1}, x_{i_2}, \cdots, x_{i_k}\right)}.$$

Pro $n = 2$, tedy dvě náhodné veličiny, je podmíněná hustota pravděpodobnosti $f_{X_1|X_2}(x_1|x_2)$ definována takto

$$f_{X_1|X_2}(x_1|x_2) = \frac{f_X(x_1, x_2)}{f_{X_2}(x_2)}. \tag{8}$$

Opačná podmíněná hustota pravděpodobnosti, tedy při znalosti realizace náhodné veličiny X_1 je

$$f_{X_2|X_1}(x_2|x_1) = \frac{f_X(x_1, x_2)}{f_{X_1}(x_1)}. \tag{9}$$

Poznámka

Podmíněná hustota pravděpodobnosti náhodné veličiny X_1 za podmínky znalosti realizace náhodné veličiny X_2 (viz první ze dvou předchozích vztahů) je funkcí x_1 - x_2 je známo a vystupuje tedy jako konstanta. Z uvedené definice plyne, že tvar podmíněné hustoty pravděpodobnosti je dán tvarem hustoty sdružené. Podmíněná pravděpodobnost je ale definována jen na části

celého základního prostoru, a to na části, vymezené podmínkou. Kdybychom tedy ponechali podmíněnou hustotu přesně stejnou jako sdruženou (na podmnožině vymezené podmínkou) nebyla by splněna podmínka jednotkového integrálu. Proto je třeba sdruženou hustotu normovat. To je dáno jmenovatelem definičního vztahu, konkrétně hodnotou $f_{X_2}(x_2)$.

To, co jsme právě popsali budeme ilustrovat na příkladě: Uvažujme pokus „hod kostkou" a s ním svázanou náhodnou veličinu X. Na tomto pokusu definujeme dva jevy. První X_1 „padne menší než šest", tj. $\{1,2,3,4,5\}$ a druhý X_2 „padne sudé číslo", tj. $\{2,4,6\}$. V nepodmíněném případě je $P(X_1) = \frac{1}{6}$, $P(X_2) = \frac{1}{2}$ a pravděpodobnost kteréhokoli jednotlivého čísla $1 \cdots 6$ je $\frac{1}{6}$. Za podmínky X_2 - padlo sudé, je omezený základní prostor roven $\{2,4,6\}$ a tedy kdybychom ponechali každému číslu pravděpodobnost $\frac{1}{6}$, byla by pravděpodobnost celého základního prostoru rovna $\frac{1}{2}$. Tvar sdružené hustoty (tj. rovnoměrnost) ponecháme a hodnoty přeformujeme tak, že každá hodnota bude mít pravděpodobnost $\frac{1}{3} = 2 \times \frac{1}{6}$. Normovací faktor je tedy 2, což je právě $\frac{1}{f(X_2)} = \frac{1}{\frac{1}{2}} = 2$, kde symbol $f(X_2)$ značí $P(\{2,4,6\})$ v původním pravděpodobnostním prostoru.

Příklady

1. Jaka je pr. sudého součtu, když na první minci nepadl R?
2. Jaká je pr., že na prvním sledovaném místě bude stupeň 1, jestliže druhé dva stupně byly 1 a 2.
3. Jaká je pr. délek kolon, když víme kdy bude měřeno a za jakého signálu SSZ.

Řetězové pravidlo

Z definice podmíněné pravděpodobnosti plyne

$$f_X(x_1, x_2) = f_{X_2|X_1}(x_2|x_1) f_{X_1}(x_1)$$

Tento vztah má velice hezkou interpretaci. Říká, že pravděpodobnost, že jevy x_1 a x_2 nastanou současně (levá strana), lze rozložit na dvě části - první je pravděpodobnost jevu x_2 za předpokladu, že nastal jev x_1 a druhá vyjadřuje pravděpodobnost toho, že zkutečně jev x_1 nastal.

Obecný tvar řetězového pravidla je

$$f_X(x_1, \cdots, x_n) = f\left(x_{i_n}|x_{i_1}, \cdots x_{i_{n-1}}\right) f\left(x_{i_{n-1}}|x_{i_1}, \cdots x_{i_{n-2}}\right) \cdots f(x_{i_2}|x_{i_1}) f(x_{i_1})$$

kde jsme pro přehlednost vynechali indexy u f a řídíme se značením v argumentu.

4.4 Ilustrace pro diskrétní náhodné veličiny

Uvažujeme měřenou intenzitu I s hodnotami $i \in \{0,1\}$ (0 znamená do 50 aut/5min, 1 označuje nad 50 aut/5min) a pozorovaný stav dopravy S s hodnotami $s \in \{1,2,3\}$. Z dlouhodobých měřená jsme získali tabulku sdružených pravděpodobností $f_{I,S}(i,s)$. Sčítáním určíme marginály $f_S(s)$ a $f_I(i)$

Sdružená hp $f_{I,S}$ a marginály f_I a f_S

i\s	1	2	3	$f_I(i)$
0	0.18	0.15	0.14	0.47
1	0.13	0.18	0.22	0.53
$f_S(s)$	0.31	0.33	0.36	

Podmíněné hp dostaneme podle vzorců (8) nebo (9). Konkrétně např.

$$f_{I|S}(0|1) = \frac{f_{I,S}(0,1)}{f_S(1)} = \frac{0.18}{0.31} = 0.58$$

nebo

$$f_{S|I}(1|0) = \frac{f_{I,S}(0,1)}{f_I(0)} = \frac{0.18}{0.47} = 0.38$$

Výsledné podmíněné hustoty jsou v následujících tabulkách, vlevo $f_{I|S}$ a vpravo $f_{S|I}$

$f_{I\|S}(I\|S=1)$	$f_{I\|S}(I\|S=2)$	$f_{I\|S}(I\|S=3)$				
1-3 0.58	0.45	0.39	0.38	0.32	0.30	$f_{S\|I}(S\|I=0)$
0.42	0.55	0.61	0.25	0.34	0.41	$f_{S\|I}(S\|I=1)$

4.5 Ilustrace pro spojité náhodné veličiny

Uvažujme hustotu pravděpodobnosti danou předpisem

$$f(x,y) = \sin(x+2y), \quad \text{pro } x \in \left(0, \frac{\pi}{2}\right),\ y \in \left(0, \frac{\pi}{2}\right)$$

Marginály jsou

$$f(x) = \int_0^{\pi/2} \sin(x+2y)\,dy = \left| \begin{array}{c} x+2y=z \\ 2dy=dz \\ z=x\cdots x+\pi \end{array} \right| = \frac{1}{2}\int_x^{x+\pi} \sin(z)\,dz =$$

$$= -\frac{1}{2}\left[\cos(z)\right]_x^{x+\pi} = -\frac{1}{2}\left(\cos(x+\pi) - \cos(x)\right) = \cos(x)$$

protože $\cos(x+\pi) = -\cos(x)$.

$$f(y) = \int_0^{\pi/2} \sin(x+2y)\,dx = \begin{array}{|c|} x+2y=z \\ dx=dz \\ z=2y\cdots 2y+\frac{\pi}{2} \end{array} = \int_{2y}^{2y+\frac{\pi}{2}} \sin(z)\,dz =$$

$$= -\left[\cos(z)\right]_{2y}^{2y+\frac{\pi}{2}} = -\left(\cos\left(2y+\frac{\pi}{2}\right) - \cos(2y)\right) = \cos(2y) + \sin(2y)$$

protože $\cos\left(2y+\frac{\pi}{2}\right) = -\sin(2y)$.

Podmíněné hustoty pravděpodobnosti jsou

$$f(x|y) = \frac{\sin(x+2y)}{\cos(2y)+\sin(2y)}$$

$$f(y|x) = \frac{\sin(x+2y)}{\cos(x)}$$

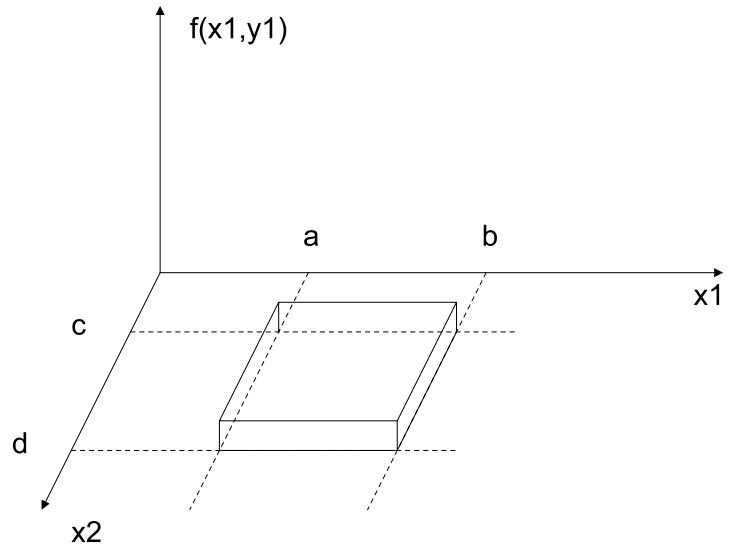
Náhodné veličiny X a Y nejsou nezávislé, protože $f(x,y) \neq f(x)\,f(y)$!

4.6 Dvourozměrné rovnoměrné rozdělení

Uvažujeme náhodnou veličinu $X = [X_1, X_2]'$ a její hustotu pravděpodobnosti

$$f(x_1, x) = k$$

pro $x_1 \in (a,b)$ a $x_2 \in (c,d)$ (viz obrázek)



Chceme určit konstantu k, distribuční funkci F_{X_1,X_2}, střední hodnotu $E[X]$ a kovarianční matici $C[X]$.

Konstanta k

Určíme z podmínky, že integrál přes $(a,b)\,x\,(c,d)$ je roven jedné.

$$\int_a^b \int_c^d k\,dx_1 dx_2 = (b-a)\,(d-c) \quad \rightarrow \quad k = \frac{1}{(b-a)\,(d-c)}$$

Distribuční funkce F_{X_1,X_2}

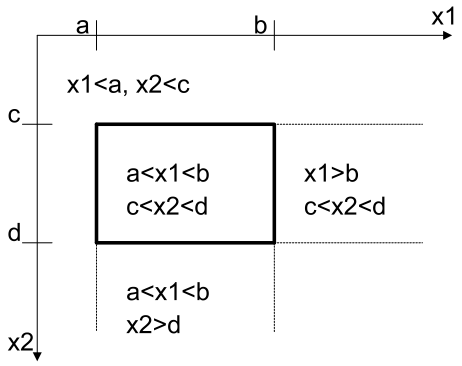
Ze vztahu mezi hp a DF

$$F_{X_1,X_2}\,(x_1,x_2) = \int_a^{x_1} \int_c^{x_2} f(u,v)\,dudv$$

Počítáme postupně:

1. Pro $x < a$ nebo $x_2 < c$ je $F_{X_1,X_2}\,(x,y) = 0$.
2. Pro $x \in (a,b)$ a $y \in (c,d)$ je
$$\begin{aligned} F_{X_1,X_2}\,(x,y) &= \int_a^x \int_c^y \frac{1}{(b-a)\,(d-c)} dvdu = \frac{1}{(b-a)\,(d-c)} \int_a^x [v]_c^y\,du = \\ &= \frac{y-c}{(b-a)\,(d-c)}\,[u]_a^x = \frac{(x-a)\,(y-c)}{(b-a)\,(d-c)} \end{aligned}$$
3. Pro $x \in (a,b)$ a $y > d$ je $F_{X_1,X_2}\,(x,y)$ stejná, jako pro $x \in (a,b)$ a $y = d$
$$F_{X_1,X_2}\,(x,y) = F_{X_1,X_2}\,(x,d) = \frac{x-a}{b-a}$$
4. Pro $x > b$ a $y \in (c,d)$ je podobně
$$F_{X_1,X_2}\,(x,y) = F_{X_1,X_2}\,(b,y) = \frac{y-c}{d-c}$$
5. Pro $x > b$ a $y > d$ je $\quad F_{X_1,X_2}\,(x,y) = 1$.

Nakreslit !!



Střední hodnota *E[X]* a *E[Y]*

Podle definice je

$$E[X] = [E[X_1], E[X_2], \cdots, E[X_n]]'$$

Odvození:

$$E[X] = \int\int \cdots \int_{X^*} [x_1, x_2, \cdots, x_n]' f(x_1, x_2, \cdots, x_n)\, dx_1 dx,_2, \cdots, dx_n =$$

$$= \int\int \cdots \int_{X^*} [x_1 f(\cdots), x_2 f(\cdots), \cdots, x_n f(\cdots)]'_1, dx_2, \cdots, dx_n =$$

$$\left[\int_{X_1^*} x_1 f(x_1)\, dx_1, \cdots \int_{X_n^*} x_n f(x_n)\, dx_n\right]' = [E[X_1], E[X_2], \cdots, E[X_n]]'$$

Marginála

$$f_{X_1}(x_1) = \int_c^d \frac{1}{(b-a)(d-c)} dy = \frac{1}{b-a}, \quad f(y) = \frac{1}{d-c}.$$

Střední hodnota

$$E[X_1] = \int_a^b x_1 \frac{1}{b-a} dx_1 = \frac{1}{b-a}\frac{1}{2}\left[x_1^2\right]_a^b = \frac{a+b}{2}$$

Podobně pro x_2. $\cdots$

Rozptyl

$$D[X_1] = E\left[X_1^2\right] - E[X_1]^2 = \int_a^b x_1^2 \frac{1}{(b-a)} dx_1 - \left(\frac{a+b}{2}\right)^2 = \frac{(a-b)^2}{12}$$

4.7 Dvourozměrné normální rozdělení.

z webu (otáčení)

Uvažujeme náhodný vektor $X = [X_1, X_2]'$, který je normálně rozložen se střední hodnotou $\mu = [\mu_1, \mu_2]'$ a kovarianční maticí

$$R = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix},$$

kde ρ je korelační koeficient $\rho = \frac{C(X_1,X_2)}{\sqrt{D(X_1)D(X_2)}}$. Určete podmíněné a marginální rozdělení.

Sdružená hp

Pro $x = [x_1, x_2]'$ je sdružená hustota pravděpodobnosti

$$f(x) = \frac{1}{2\pi\sqrt{|\,R\,|}} \exp\left\{-\frac{1}{2}(x-\mu)' R^{-1}(x-\mu)\right\}.$$

Determinant kovarianční matice R je

$$|R| = \sigma_1^2\sigma_2^2\left(1-\rho^2\right)$$

a inverze kovarianční matice je

$$R^{-1} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}^{-1} = \frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)} \begin{bmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{bmatrix}$$

Upravíme kvadratickou formu v exponentu hustoty pravděpodobnosti

$$\begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix}' \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}^{-1} \begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix} =$$

$$= -\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right] = *$$

$$\text{substituce: } \xi_1 = x_1-\mu_1,\ \xi_2 = x_2-\mu_2$$

$$* = -\frac{1}{2(1-\rho^2)\sigma_1^2\sigma_2^2}\left[\sigma_2^2\xi_1^2 - 2\rho\sigma_1\sigma_2\xi_1\xi_2 + \sigma_1^2\xi_2^2\right] =$$

$$= -\frac{1}{2(1-\rho^2)\sigma_1^2\sigma_2^2}\left[\sigma_2^2\left\{\xi_1^2 - 2\rho\frac{\sigma_1}{\sigma_2}\xi_1\xi_2 + \frac{\sigma_1^2}{\sigma_2^2}\xi_2^2\right\}\right] =$$

$$= -\frac{1}{2(1-\rho^2)\sigma_1^2\sigma_2^2}\left[\sigma_2^2\left\{\xi_1^2 - 2\rho\frac{\sigma_1}{\sigma_2}\xi_1\xi_2 + \left(\rho\frac{\sigma_1}{\sigma_2}\xi_2\right)^2 - \left(\rho\frac{\sigma_1}{\sigma_2}\xi_2\right)^2 + \frac{\sigma_1^2}{\sigma_2^2}\xi_2^2\right\}\right] =$$

$$-\frac{1}{2(1-\rho^2)\sigma_1^2\sigma_2^2}\left[\sigma_2^2\left(\xi_1 - \rho\frac{\sigma_1}{\sigma_2}\xi_2\right)^2 + \left(1-\rho^2\right)\sigma_1^2\xi_2^2\right] =$$

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

Podmíněná hp

$$f_{X_1|X_2}(x_1|x_2) = \frac{1}{\sqrt{2\pi(1-\rho^2)\sigma_1^2}} \exp\left\{-\frac{1}{2}\frac{1}{(1-\rho^2)\sigma_1^2}\left[x_1 - \left\{\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2-\mu_2)\right\}\right]\right\}$$

a tedy střední hodnota je $\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2-\mu_2)$ a rozptyl $(1-\rho^2)\sigma_1^2$.

Marginální hp

$$f_{X_2}(x_2) = \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left\{-\frac{1}{2}\frac{1}{\sigma_2^2}(x_2-\mu_2)^2\right\}$$

odkud střední hodnota je μ_2 a rozptyl je σ_2^2.

Poznámka

Úpravy (doplnění na čtverec), které jsme prováděli jsou velice důležitou výpočetní technikou při práci s normálním rozdělením. $\cdots$ N hp nemá primitivní funkci, proto integrál při výpočtu marginály se dělá rozkladem na podm. a marg. Tím se vlastně získá integrál.

4.8 Charakteristiky náhodného vektoru

4.8.1 Kovariance

$$C[X_1, X_2] = \int_{X_1^*}\int_{X_2^*}(x_1-\mu_1)(x_2-\mu_2) f(x_1,x_2)\,dx_1dx_2$$

kde $\mu_i = E[X_i]$, $i = 1, 2$.

Popsat význam: 0 nekorelované, +velká pozitivně a -velká negativně korelované.

Srovnat: nekorelovanost a nezávislost.

4.8.2 Střední hodnota a kovarianční matice.

$$E[X] = \begin{bmatrix} E[X_1] \\ E[X_2] \\ \cdots \\ E[X_n] \end{bmatrix}, \quad C[X] = \begin{bmatrix} D[X_1] & C[X_1,X_2] & \cdots & C[X_1,X_n] \\ C[X_2,X_1] & D[X_2] & \cdots & C[X_2,X_n] \\ \cdots & \cdots & \cdots & \cdots \\ C[X_n,X_1] & C[X_n,X_2] & \cdots & D[X_n] \end{bmatrix}$$

Průměrná délka první kolony; prům. délka 1. kolony za poledne, prům. délka 1 kolony při krátké zelené, prům. délka 1 kolony při při dlouhých sousedních kolonách !!! – pokaždé jiný průměr (podmíněný)

4.8.3 Momenty

$k-$tý centrální moment	$k-$tý obecný moment
$\mu_k = \int_{X^*} (x-\mu)^k f(x)\,dx$	$\mu_k' = \int_{X^*} x^k f(x)\,dx$

a analogicky vzájemné momenty a momenty pro náhodné vektory.

4.8.4 Kvantily a kritické hodnoty

kvantil ζ_α	kritická hodnota z_α
$\int_{-\infty}^{\zeta_\alpha} f(x)\,dx = \alpha$	$\int_{z_\alpha}^{\infty} f(x)\,dx = \alpha$

4.9 Příklady

Příklad 1

$$f(x_1,x_2) = \frac{1}{15}(x_1+x_2+1) \quad \text{pro} \quad x_1 = 0,1,2; \quad x_2 = 0,1$$

Řešení:

Marginální hustoty:

$f(x_1,x_2)$	$x_2=0$	$x_2=1$	$f(x_1)$
$x_1=0$	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{1}{5}$
$x_2=1$	$\frac{2}{15}$	$\frac{3}{15}$	$\frac{1}{3}$
$x_3=2$	$\frac{3}{15}$	$\frac{4}{15}$	$\frac{7}{15}$
$f(x_2)$	$\frac{2}{5}$	$\frac{3}{5}$	1

Střední hodnota:

$$\begin{aligned} E(x_1) &= 0\cdot\frac{1}{5} + 1\cdot\frac{1}{3} + 2\cdot\frac{7}{15} = \frac{19}{15} = 1\frac{4}{5} \\ E(x_2) &= 0\cdot\frac{2}{5} + 1\cdot\frac{3}{5} = \frac{3}{5} \end{aligned}$$

Rozptyl:

$$\begin{aligned} D(x_1) &= (0-\frac{19}{15})^2\frac{1}{5} + (1-\frac{19}{15})^2\frac{1}{3} + (2-\frac{19}{15})^2\frac{7}{15} = \frac{134}{225} \\ D(x_2) &= (0-\frac{3}{5})^2\frac{2}{5} + (1-\frac{3}{5})^2\frac{3}{5} = \frac{24}{125} \end{aligned}$$

Kovariance:

$$\begin{aligned} cov(x_1,x_2) &= (0-\frac{19}{15})(0-\frac{3}{5})\frac{1}{15}+(0-\frac{19}{15})(1-\frac{3}{5})\frac{2}{15}+ \\ &+ (1-\frac{19}{15})(0-\frac{3}{5})\frac{2}{15}+(1-\frac{19}{15})(1-\frac{3}{5})\frac{3}{15}+ \\ &+ (2-\frac{19}{15})(0-\frac{3}{5})\frac{3}{15}+(2-\frac{19}{15})(1-\frac{3}{5})\frac{4}{15}=-\frac{2}{75} \end{aligned}$$

Kovarianční matice:

$$C = \begin{bmatrix} \frac{134}{225} & -\frac{2}{75} \\ -\frac{2}{75} & \frac{24}{125} \end{bmatrix}$$

Příklad 2

$$f(x_1,x_2) = 4x_1x_2 \quad \textbf{pro} \quad x_1,x_2 \in (0,1)$$

Řešení:

Marginální hustoty:

$$\begin{aligned} f(x_1) &= \int_0^1 4x_1x_2dx_2 = \left[2x_1x_2^2\right]_0^1 = 2x_1 \\ f(x_2) &= 2x_2 \end{aligned}$$

Střední hodnota:

$$\begin{aligned} E(x_1) &= \int_0^1 x_1 2x_1 dx_1 = \left[\frac{2}{3}x_1^3\right]_0^1 = \frac{2}{3} \\ E(x_2) &= \frac{2}{3} \end{aligned}$$

Rozptyl:

$$\begin{aligned} D(x_1) &= \int_0^1 (x_1-\frac{2}{3})^2 2x_1dx_1 = 2\left[\frac{1}{4}x_1^4-\frac{4}{9}x_1^3+\frac{2}{9}x_1^2\right]_0^1 = \frac{1}{18} \\ D(x_2) &= \frac{1}{18} \end{aligned}$$

Kovariance:

$$cov(x_1,x_2) = \int_0^1\int_0^1 (x_1-\frac{2}{3})(x_2-\frac{2}{3})4x_1x_2dx_1x_2 = 0$$

Kovarianční matice:

$$C = \begin{bmatrix} \frac{1}{18} & 0 \\ 0 & \frac{1}{18} \end{bmatrix}$$

Příklad 3

$$f(x_1, x_2) = x_1 + x_2 \quad \text{pro} \quad x_1, x_2 \in (0, 1)$$

Řešení:

Marginální hustoty:

$$\begin{aligned} f(x_1) &= \int_0^1 (x_1 + x_2) dx_2 = \left[x_1 x_2 + \frac{1}{2} x_2^2 \right]_0^1 = x_1 + \frac{1}{2} \\ f(x_2) &= x_2 + \frac{1}{2} \end{aligned}$$

Střední hodnota:

$$\begin{aligned} E(x_1) &= \int_0^1 x_1 (x_1 + \frac{1}{2}) dx_1 = \left[\frac{1}{3} x_1^3 + \frac{1}{4} x_1^2 \right]_0^1 = \frac{7}{12} \\ E(x_2) &= \frac{7}{12} \end{aligned}$$

Rozptyl:

$$\begin{aligned} D(x_1) &= \int_0^1 (x_1 - \frac{7}{12})^2 (x_1 + \frac{1}{2}) dx_1 = \left[\frac{1}{4} x_1^4 - \frac{2}{9} x_1^3 - \frac{35}{288} x_1^2 + \frac{49}{288} x_1 \right]_0^1 = \frac{3}{16} \\ D(x_2) &= \frac{3}{16} \end{aligned}$$

Kovariance:

$$\begin{aligned} cov(x_1, x_2) &= \int_0^1 \int_0^1 (x_1 - \frac{7}{12})(x_2 - \frac{7}{12})(x_1 + x_2) dx_1 dx_2 = \\ &= \int_0^1 (x_1 - \frac{7}{12}) \left[\frac{1}{3} x_2^3 + \frac{1}{2} x_2^2 \left(x_1 - \frac{7}{12} \right) - \frac{7}{12} x_1 x_2 \right]_0^1 dx_1 = \\ &= \left[-\frac{1}{36} x_1^3 + \frac{13}{288} x_1^2 - \frac{7}{288} x_1 \right]_0^1 = -\frac{1}{144} \end{aligned}$$

Kovarianční matice:

$$C = \begin{bmatrix} \frac{3}{16} & -\frac{1}{144} \\ -\frac{1}{144} & \frac{3}{16} \end{bmatrix}$$

Příklad 4

$$f(x_1, x_2) = e^{-x_1 - x_2} \quad \textbf{pro} \quad x_1, x_2 \in (0, \infty)$$

Řešení:

Marginální hustoty:

$$\begin{aligned} f(x_1) &= \int_0^\infty e^{-x_1-x_2}dx_2 = e^{-x_1}\left[-e^{-x_2}\right]_0^\infty = e^{-x_1} \\ f(x_2) &= e^{-x_2} \end{aligned}$$

Střední hodnota:

$$\begin{aligned} E(x_1) &= \int_0^\infty x_1 e^{-x_1}dx_1 = \left|\begin{array}{ll} u = x_1 & u' = 1 \\ v' = e^{-x_1} & v = -e^{-x_1} \end{array}\right| = \left[-x_1e^{-x_1}\right]_0^\infty + \int_0^\infty e^{-x_1}dx_1 = \\ &= \left[-e^{-x_1}\right]_0^\infty = 1 \\ E(x_2) &= 1 \end{aligned}$$

Rozptyl:

$$\begin{aligned} D(x_1) &= \int_0^\infty (x_1-1)^2e^{-x_1}dx_1 = \int_0^\infty x_1^2e^{-x_1}dx_1 - 2\int_0^\infty x_1e^{-x_1}dx_1 + \int_0^\infty e^{-x_1}dx_1 = \\ &= \int_0^\infty x_1^2e^{-x_1}dx_1 - 1 = \left|\begin{array}{ll} u = x_1^2 & u' = 2x_1 \\ v' = e^{-x_1} & v = -e^{-x_1} \end{array}\right| \\ &= \left[-x_1^2e^{-x_1}\right]_0^\infty + 2\int_0^\infty x_1e^{-x_1}dx_1 - 1 = 1 \\ D(x_2) &= 1 \end{aligned}$$

Kovariance:

$$\begin{aligned} cov(x_1,x_2) &= \int_0^\infty\int_0^\infty (x_1-1)(x_2-1)e^{-x_1-x_2}dx_1dx_2 = \\ &= \int_0^\infty (x_1-1)e^{-x_1}dx_1\int_0^\infty (x_2-1)e^{-x_2}dx_2 = \\ &= \left(\int_0^\infty xe^{-x}dx - \int_0^\infty e^{-x}dx\right)^2 = 0 \end{aligned}$$

Kovarianční matice:

$$C = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Příklad 5

$$f(x_1,x_2) = (a-1)(a-2)(1+x_1+x_2)^{-a} \quad \text{pro} \quad x_1,x_2 \in (0,\infty),\ a > 4$$

Řešení:

Marginální hustoty:

$$\begin{aligned} f(x_1) &= \int_0^\infty (a-1)(a-2)(1+x_1+x_2)^{-a}dx_2 = (a-1)(a-2)\left[\frac{1}{1-a}(1+x_1+x_2)^{1-a}\right]_0^\infty = \\ &= (a-1)(a-2)\frac{1}{a-1}(1+x_1)^{1-a} = (a-2)(1+x_1)^{1-a} \\ f(x_2) &= (a-2)(1+x_2)^{1-a} \end{aligned}$$

Střední hodnota:

$$E[X_1] = \int_0^\infty x_1(a-2)(1+x_1)^{1-a}dx_1 =$$

$$= \left[\begin{array}{ll} u = x_1, & v' = (1+x_1)^{1-a} \\ u' = 1, & v = \frac{1}{2-a}(1+x_1)^{2-a} \end{array}\right] = \frac{1}{(a-3)}$$

Rozptyl:

$$D[X_1] = \int_0^\infty x_1^2(a-2)(1+x_1)^{1-a}dx_1 - \frac{1}{(a-3)^2} =$$

$$= \left[\begin{array}{ll} u = x^2, & v' = (1+x)^{1-a} \\ u' = 2x, & \frac{1}{2-a}(1+x)^{2-a} \end{array}\right] = \cdots = \left[\begin{array}{ll} u = x, & v' = (1+x)^{2-a} \\ u' = 1, & v = \frac{1}{3-a}(1+x)^{3-a} \end{array}\right] = \frac{2}{(a-3)(a-4)}$$

Kovariance:

$$C[X_1, X_2] = \frac{1}{(a-3)(a-4)}$$

Kovarianční matice:

$$C = \frac{1}{(a-3)(a-4)}\left[\begin{array}{cc} 2 & 1 \\ 1 & 2 \end{array}\right].$$

4.10 Transformace hustoty pravděpodobnosti

Hovoříme-li o transformaci náhodné veličiny, pak máme na mysli situaci, kdy zdrojem neurčitosti je určitá náhodná veličina, my se ale zajímáme o jinou náhodnou veličinu, která je deterministickou funkcí původní veličiny. Například, na váze vážíme náhodně různě těžké předměty. Váha je špatně postavena a proto váží o 5 gramů méně. Označíme-li x naváženou hodnotu (v kg), pak správná hodnota y (v kg) bude

$$y = x + 0.005 \tag{10}$$

Pokud budeme navážené hodnoty považovat za realizace náhodné veličiny X, pak Y bude rovněž náhodná veličina, definovaná vztahem (10) jako funkce návodné veličiny X.

Nejčastěji používaným popisem náhodné veličiny je hustota pravděpodobnosti. Budeme proto hledat hustotu transformované náhodné veličiny vyjádřenou pomocí hustota původní náhodné veličiny.

Transformace pomocí rostoucí funkce

Je dána rostoucí funkce $y = h(x)$, která definuje transformaci $Y = h(X)$. Předpokládáme, že známe distribuční funkci $F_X(x)$ původní náhodné veličiny X a hledáme distribuční funkci $F_Y(y)$ transformované náhodné veličiny Y.

Platí

$$F_Y(y) = P(Y \leq y) = P(h(X) \leq y) = P\left(X \leq h^{-1}(y)\right) = F_X\left(h^{-1}(y)\right). \tag{11}$$

Poznámka

Pro klesající transformační funkci platí $F_Y(y) = 1 - F_X(h^{-1}(y))$. *Dokažte!*
Návod: pro klesající transformační funkci se obrací nerovnost, tj. je-li $Y < y$, *pak* $X \geq h^{-1}(y)$. *Potom stačí už jenom si uvědomit, jaká je pravděpodobnost doplňkového jevu.*

Transformace pomocí monotónní funkce

Opět sledujeme postup použitý v (11). V případě, že transformační funkce je klesající, bude $\frac{dF_Y}{dy} = -f_X(h^{-1}(y))\frac{dh^{-1}(y)}{dy}$, ale vzhledem k tomu, že h je klesající, bude také její inverze klesající a tedy její derivace záporná. Formálně můžeme proto odstranit znaménko minus a také minus z derivace (pomocí absolutní hodnoty). Dva minusy v součinu dají plus a výraz se tak nezmění. Odvození je následující

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{d}{dx}F_X\left(h^{-1}(y)\right)\left|\frac{dh^{-1}(y)}{dy}\right| = f_X\left(h^{-1}(y)\right)\left|\frac{dh^{-1}(y)}{dy}\right|$$

Regulární transformace náhodného vektoru

Postup, uvedený v předchozím odstavci, můžeme aplikovat i na n-rozměrný náhodný vektor v případě, že transformujeme na nový vektor stejné dimenze. Transformace tedy probíhá podle soustavy n funkcí n proměnných

$$\begin{aligned} y_1 &= h_1(x_1, x_2 \cdots x_n) \\ y_2 &= h_2(x_1, x_2 \cdots x_n) \\ &\cdots \\ y_n &= h_n(x_1, x_2 \cdots x_n) \end{aligned}$$

Inverzní funkce označíme $x_1 = h_1^{-1}(y), \cdots h_n^{-1}(y)$, tj. $x = h^{-1}(y)$ a determinant jejích parciálních derivací

$$J = \begin{bmatrix} \frac{\partial h_1^{-1}}{\partial y_1} & \frac{\partial h_1^{-1}}{\partial y_2} & \cdots & \frac{\partial h_1^{-1}}{\partial y_n} \\ \frac{\partial h_2^{-1}}{\partial y_1} & \frac{\partial h_2^{-1}}{\partial y_2} & \cdots & \frac{\partial h_2^{-1}}{\partial y_n} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial h_n^{-1}}{\partial y_1} & \frac{\partial h_n^{-1}}{\partial y_2} & \cdots & \frac{\partial h_n^{-1}}{\partial y_n} \end{bmatrix}$$

Determinant z matice J, který označíme $|J|$ se nazývá Jakobián.

Transformační vztah má potom tvar

$$f_Y(y) = f_X\left(h^{-1}(y)\right)|J|$$

Příklady

1. Transformace normování $N(0,1) \to N(\mu, \sigma^2)$

$$f(x) = N(0,1) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}x^2}, \;\; Y = \sigma X + \mu, \;\; X = \frac{Y-\mu}{\sigma} \;\; \frac{dx}{dy} = \frac{1}{\sigma} > 0$$

$$f(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{Y-\mu}{\sigma}\right)^2} \frac{1}{\sigma} = N\left(\mu, \sigma^2\right)$$

2. Transformace $Y = X_1 + X_2$

$$F_Y(y) = P(Y \leq y) = \int\int_{x_1+x_2 \leq y} f_X(x_1, x_2)\, dx_1 dx_2 = \int_{-\infty}^{\infty} \int_{-\infty}^{y-x_2} f_X(x_1, x_2)\, dx_1 dx_2$$

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \int_{-\infty}^{\infty} f(y - x_2, x_2)\, dx_2.$$

   Dopočítat pro f_X rovnoměrné na čtverci $\mu_1 \pm h \times \mu_2 \pm h$ nebo exponenciální $f(x_1, x_2) \propto \mathrm{e}^{-a_1 x_1 - a_2 x_2}$ pro $x_1, x_2 > 0$.

3. Příklad 3. Transformace 2 x 2 - viz LaTex

$$\begin{aligned} Y_1 &= X_1 + X_2 \quad X_1 = Y_1 - Y_2 \\ Y_2 &= \quad X_2 \quad X_2 = {}_2 \end{aligned}$$

$$J = \det \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} = 1, \quad f_Y(y) = f_X(x_1, x_2) = f_X(y_1 - y_2, y_2)$$

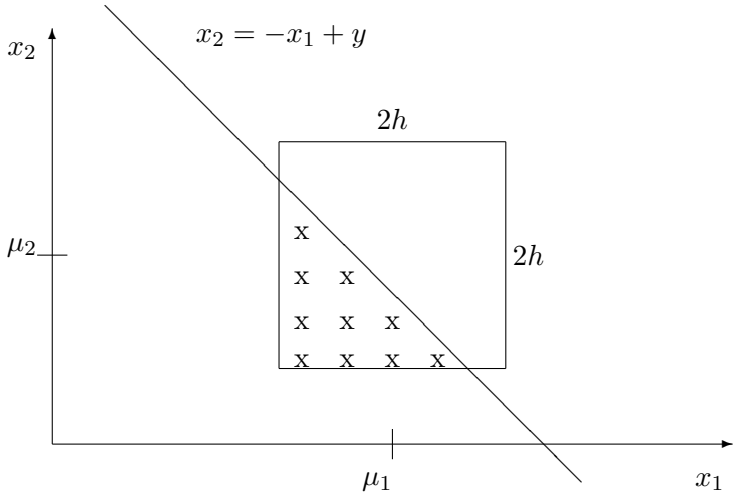
$$f(y_1) = \int_{-\infty}^{\infty} f_Y(y)\, dy_2 = \int_{-\infty}^{\infty} f_X(y_1 - y_2, y_2)\, dy_2$$

   což odpovídá předchozímu výsledku.

Další příklady

Určete distribuční funkci transformace $Y = X_1 + X_2$ pro nezávislé, rovnoměrně rozdělené náhodné veličiny X a X_2 s hustotou pravděpodobnosti $f(x) = \frac{1}{2h}$ pro $x \in \mu \pm h$.

Počítáme na základě definice $F_Y(y) = \int_{X^*} f_X(x)\underset{\sim}{x}$, kde X^* je oblast integrace daná podmínkou $h(x) \leq y$, tj. $x_1 + x_2 \leq y$. Přitom obory integrace pro x jsou $x_1 \in (\mu_1 - h, \mu_1 + h)$ a $x_2 \in (\mu_2 - h, \mu_2 + h)$. Oblast integrace tedy bude podle obrázku



Hustota pravděpodobnosti X je $f_X(x) = \frac{1}{4h^2}$, protože X_1 X_2 jsou nezávislé.

Nejprve budeme integrovat přes dolní trojúhelník, tedy s mezemi

$$x_1 = (\mu_1 - h) : (y - \mu_2 + h) \qquad \text{a} \qquad x_2 = (\mu_2 - h) : (y - x_1)$$

a tedy

$$F_Y(y) = \int_{\mu_1 - h}^{y - \mu_2 + h} \int_{\mu_2 - h}^{y - x_1} \frac{1}{4h^2} dx_2 dx_1 = \frac{1}{4h^2} \int_{\mu_1 - h}^{y - \mu_2 + h} (y - x_1 - \mu_2 + h)\underset{\sim}{x}_1 =$$

$$| \text{ subst. } y - x_1 - \mu_2 + h = z |$$

$$= \frac{1}{h^2} \int_0^{y - \mu_1 - \mu_2 + 2h} z dz = \frac{1}{8h^2}(y - \mu_2 - \mu_2 + 2h)^2$$

pro $y \in (\mu_1 + \mu_2 - 2h, \mu_1 + \mu_2)$.

Pokračujeme v integraci přes horní trojúhelník. Meze integrace jsou

$$x_1 = (y - \mu_2 - h) : (\mu_1 + h) \qquad \text{a} \qquad x_2 = (y - x_1) : (\mu_2 + h)$$

Budeme počítat $P(Y > y)$ a tedy bude $F(y) = 1 - P(Y > y)$.

$$F_Y(y) = 1 - \int_{y-\mu_2-h}^{\mu_1+h} \int_{y-x_1}^{\mu_2+h} \frac{1}{4h^2} dx_2 dx_1 = \frac{1}{4h^2} \int_{y-\mu_2-h}^{\mu_1+h} (\mu_2 + h - y + x_1) dx_1$$

$$| \text{ subst. } y - x_1 - \mu_2 + h = z |$$

$$= 1 - \frac{1}{4h^2} \int_0^{\mu_1+\mu_2+2h-y} z dz = 1 - \frac{1}{8h^2} (y - \mu_1 - \mu_2 - 2h)^2$$

pro $y \in (\mu_1 + \mu_2, \mu_1 + \mu_2 + 2h)$.

Další příklad:

Určete hustotu pravděpodobnosti $f_Y(y)$ pro funkci $Y_1 = \frac{X_1}{X_1+X_2}, Y_2 = X_1 + X_2$ jestliže $f_X(x_1, x_2) = \mathrm{e}^{-x_1-x_2}$ a $x_1, x_2 > 0$.

Meze pro Y jsou $y_1 \in (0, 1)$ a $y_2 > 0$.

Inverzní funkce: $x_1 = y_1 y_2$ a $x_2 = y_2 - y_1 y_2$.

Jakobián

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix} = \begin{vmatrix} y_2 & y_1 \\ -y_2 & 1 - y_1 \end{vmatrix} = y_2$$

Hustota F_Y tedy bude

$$f_Y(y) = f_X(y_1 y_2, y_2 - y_1 y_2)|J| = y_2 \mathrm{e}^{-y_2}.$$

pro $y_1 \in (0, 1)$ a $y_2 > 0$.

4.11 Operátorový počet se střední hodnotou

V předchozích kapitolách jsme definovali charakteristiky náhodné veličiny a náhodného vektoru. Nejdůležitější z nich, a také nejpoužívanější, jsou střední hodnota a rozptyl. Připomeňme jejich definice

Střední hodnota:

$$E[X] = \int_X x f(x) \, dx, \tag{12}$$

Rozptyl:

$$D[X] = E\left[(X - E[X])^2\right], \tag{13}$$

kde $\int_X \cdot dx$ znamená $\int_{-\infty}^{\infty} \cdot dx$ pro spojitou náhodnou veličinu a $\sum_{x \in X}$ pro diskrétní náhodnou veličinu.

Protože se s těmito charakteristikami poměrně často pracuje, vyplatí se (pro stručnost zápisu) odvodit operátorový počet, kde pracujeme s operátory $E[\cdot]$ a $D[\cdot]$ a používáme pro ně pravidla, odvozená na základě definic (12) a (13). Pravidla uvádíme dále, jejich odvození je v Doplňku 13.3.

Uvažujeme náhodné veličiny X, Y a čísla a, b. Potom platí

1. $E[a] = a$
2. $E[a + X] = a + E[X]$
3. $E[aX] = aE[X]$
4. $E[X + Y] = E[X] + E[Y]$
5. $E[aX + bY] = aE[X] + bE[Y]$ linearita
   ale $E[XY] \neq E[X]\,E[Y]$, $E[X^2] \neq E[X]^2$

6. $D[a] = 0$
7. $D[a + X] = D[X]$
8. $D[aX] = a^2 D[X]$
9. $D[X + Y] = D[X] + D[Y]$ pro X, Y nezávislé !
10. $D[aX + bY] = a^2 D[X] + b^2 D[Y]$ pro X, Y nezávislé !

Příklady

Uvedeme dva příklady využití operátorového počtu k odvození statistických vztahů

Výpočetní vzorec pro rozptyl

$$D[X] = E\left[(X - E[X])^2\right] = E\left[X^2\right] - E[X]^2$$

Pro X_i kde $E[X_i] = \mu$ a $D[X_i] = \sigma^2$, $\forall i = 1 \cdots n$, X_i nezávislé platí

$$E\left[\bar{X}\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \mu$$

$$D\left[\bar{X}\right] = D\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{\sigma^2}{n}$$

5 Náhodný výběr

Hlavním úkolem statistiky je rozbor dat, která vykazují náhodné kolísání. Jedná se většinou o data měřená na určitém procesu za účelem (lepšího) poznání tohoto procesu. Např. měříme hustoty a intenzity dopravního proudu na několika místech určité dopravní oblasti, abychom byli schopni (lépe) řídit dopravu v této oblasti.

5.1 Pojem náhodného výběru

Zkoumaný proces chápeme jako náhodnou veličinu s určitým, nám neznámým (nebo ne úplně známým), rozdělením a měřená data jako realizace této náhodné veličiny. Pro jednoduchost a lepší představu se při výkladu omezíme na nejběžnější druh procesu, který nazveme **základním statistickým pokusem** a který spočívá v dotazu vybrané jednotky z dané množiny jednotek na určitou věc. Množinu všech odpovědí nazveme **soubor** a podmnožinu získaných odpovědí nazveme **výběr**.

Například sledování spotřeby automobilů určitého typu, vyrobené v daném roce a při najetí 5000 km. Souborem jsou spotřeby automobilů vyrobených v daném roce. Výběrem jsou spotřeby sledovaných automobilů. Dotaz je symbolický název pro změření spotřeby automobilu. Jednotka je automobil, vyrobený v daném roce (prvek souboru). Množina spotřeb všech automobilů z daného roku je soubor a podmnožina spotřeb sledovaných automobilů) je výběr.

Abychom získali objektivní informace o zkoumaném procesu, je třeba, aby výběr dotazovaných jednotek byl nezávislý.

Například zjišťujeme-li průměrné stáří automobilů, je nesmyslné omezit se na autobazary.

Jestliže výběr, tj. sérii dotazů, opakujeme, dostaneme jiné odpovědi. To je dáno tím, že dotazy jsou adresovány náhodně a při dalším výběru zahrneme jiné jednotky. Abstraktně lze definovat **náhodný výběr** jako uspořádanou množinu (vektor) náhodných veličin, jejíž realizací dostaneme jednu konkrétní realizaci výběru – číselný vektor.

5.1.1 Náhodný výběr

Náhodný výběr $X = [X_1, X_2, \ldots, X_n]$ je vektor **nezávislých** a **stejně rozdělených** náhodných veličin. číslo n se nazývá **rozsah výběru**. Vektor realizací náhodných veličin $x = [x_1, x_2, \ldots, x_n]$ nazveme **realizací náhodného výběru**.

Jedna z typických úloh statistiky je odhad střední hodnoty souboru, přičemž předpokládáme, že typ rozdělení souboru je známý. Odhadujeme jeden z jeho parametrů – střední hodnotu. O té vypovídají všechna data náhodného výběru. Je proto užitečné znát rozdělení celého náhodného výběru.

5.1.2 Distribuční funkce náhodného výběru

$X = [X_1, X_2, \ldots, X_n]$ je náhodný výběr a $F(x_i),\ i = 1, 2, \ldots, n$ je distribuční funkce, určující rozdělení jeho složek. Pak distribuční funkce náhodného výběru $H(x)$ je rovna součinu distribučních funkcí jeho složek

$$H(x) = H(x_1, x_2, \ldots, x_n) = F(x_1)F(x_2)\ldots F(x_n).$$

Tvrzení plyne přímo ze skutečnosti, že složky náhodného výběru jsou nezávislé a distribuční funkce nezávislých náhodných veličin je rovna součinu jejich distribučních funkcí.

5.2 Charakteristiky výběru

Realizace náhodného výběru je datový soubor. Ten lze popsat pomocí charakteristik popisné statistiky. Těmito charakteristikami lze popsat i náhodný výběr, s jednou důležitou odlišností. Charakteristiky datového souboru jsou konstanty (např. pro střední hodnotu platí: sečtu-li daná čísla odpředu nebo odzadu, dostanu vždy totéž). Charakteristiky náhodného výběru jsou náhodné veličiny. Náhodnost zde vzniká vzhledem k opakovaným výběrům. Provedeme-li první výběr a spočteme charakteristiku (např. průměr) této realizace, dostaneme číslo. Další výběr poskytne novou realizaci a protože je náhodný, budou v ní jiná čísla. Proto i charakteristika (průměr) bude mít jinou hodnotu. Tedy: každý výběr dá jinou realizaci a jinou hodnotu charakteristiky – realizaci charakteristiky náhodného výběru, která je náhodnou veličinou.

Různé charakteristiky náhodného výběru budou dále velmi důležité. Každá zkoumaná vlastnost bude mít svou charakteristiku, které budeme říkat **statistika**. Protože statistika je náhodná, má své rozdělení (hustotu pravděpodobnosti statistiky). Ta je základním nástrojem pro veškerá odvození klasické statistiky.

Nejprve se podrobněji podíváme na nejznámější charakteristiku **výběrový průměr**, v příští kapitole si pak uvedeme další základní charakteristiky a jejich vlastnosti.

5.2.1 Výběrový průměr

Pro výběr $X = [X_1, X_2, \ldots, X_n]$ ze spojité náhodné veličiny X se střední hodnotou μ a rozptylem σ^2 je **výběrový průměr** definován vztahem

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{14}$$

Komentář k definici

1. Všimněme si, že ve vzorci pro výběrový průměr figurují velká písmena. Není to tedy průměr z čísel, ale formálně zapsaný průměr z náhodných veličin.

5.2.2 Charakteristiky výběrového průměru

Pro výběrový průměr, jako náhodnou veličinu, uvedeme jeho střední hodnotu a rozptyl. Jsou to vlastně charakteristiky definované na charakteristice výběrový průměr. Tyto charakteristiky se počítají pro všechny možné hodnoty výběrového průměru. Závisí proto na všech realizacích souboru. Jsou to tedy již konstanty (důkladně rozmyslet).

Příklad

Uvažujme soubor s hodnotami $\{1, 2, 3\}$ a výběr z tohoto souboru o rozsahu $n = 2$. Všechny možné výběry jsou uvedeny jako sloupce následující tabulky

možné	1	1	1	2	2	2	3	3	3
výběry	1	2	3	1	2	3	1	2	3
výb. průměry	1	1.5	2	1.5	2	2.5	2	2.5	3

Jestliže zprůměrujeme všechny výběrové průměry, dostaneme střední hodnotu výběrového průměru. Zde je to 2.

Střední hodnota výběrového průměru $\overline{X}$ je rovna střední hodnotě souboru $E[X] = \mu$

$$E[\overline{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} n\mu = \mu. \tag{15}$$

Rozptyl výběrového průměru $\overline{X}$ je roven rozptylu souboru $D[X] = \sigma^2$, dělenému rozsahem výběru n

$$s^2_{\overline{X}} = D[\overline{X}] = D\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2} D\left[\sum_{i=1}^{n} X_i\right] \underbrace{=}_{nezávislost} \frac{1}{n^2}\sum_{i=1}^{n} D[X_i] = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}. \tag{16}$$

5.2.3 Rozdělení výběrového průměru

Pro výběr z náhodné veličiny s normálním rozdělením $N(\mu, \sigma^2)$, nebo pro dostatečně velký výběr (viz Centrální limitní věta) platí

$$\overline{X} \sim N(E[\overline{X}], D[\overline{X}]) = N(\mu, \sigma^2/n)$$

tj, výběrový průměr má normální rozdělení se střední hodnotou μ a rozptylem σ^2/n.

5.2.4 Výběrový rozptyl

Pro výběr $\mathbf{X} = [X_1, X_2, \ldots, X_n]$ ze spojité náhodné veličiny X se střední hodnotou μ a rozptylem σ^2 je **výběrový rozptyl** definován vztahem

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2. \tag{17}$$

5.2.5 Výběrový podíl

Pro výběr $\mathbf{X} = [X_1, X_2, \ldots, X_n]$ z alternativní náhodné veličiny X se souborovým podílem π je **výběrový podíl** definován vztahem

$$p = \frac{n^+}{n}, \tag{18}$$

kde n^+ je počet příznivých pokusů ve výběru a n je rozsah výběru.

5.3 Typy úloh s náhodným výběrem

Výběrový průměr je nejznámější, ale zdaleka ne jedinou charakteristikou výběru. Nyní uvedeme ty charakteristiky, které budeme dále nejčastěji používat. Mohou se týkat jednoho nebo dvou výběrů.

Jeden náhodný výběr

Uvažujeme výběr z jednoho rozdělení, např. zjišťujeme stáří náhodně vybraných automobilů. Budeme sledovat tři základní charakteristiky náhodného výběru: výběrový **průměr**, **rozptyl** a **podíl**. Protože je výběr náhodný, jsou i tyto charakteristiky náhodné a každá z nich má své rozdělení.

5.3.1 *Výběrový průměr při známém rozptylu souboru*

Normovaný výběrový průměr z normálního rozdělení $N(\mu; \sigma^2)$, nebo dostatečně velký výběr z libovolného rozdělení se střední hodnotou μ a rozptylem σ^2 je

$$\boxed{Z = \frac{\overline{X} - \mu}{\sigma} \sqrt{n} \quad \sim N(0,1)} \tag{19}$$

a má normované normální rozdělení $N(0; 1)$.

5.3.2 *Výběrový průměr při neznámém rozptylu souboru*

V případě neznámého rozptylu souboru, ze kterého je výběr pořízen, se neznámý rozptyl odhadne pomocí výběrového rozptylu.

Normovaný výběrový průměr je definován

$$\boxed{t = \frac{\overline{X}-\mu}{S}\sqrt{n} \quad \sim St(n-1)} \tag{20}$$

a má Studentovo rozdělení s $n-1$ stupni volnosti.

5.3.3 *Výběrový rozptyl*

Normovaný výběrový rozptyl je definován vztahem

$$\boxed{\chi^2 = \frac{(n-1)S^2}{\sigma^2} \quad \sim Chi2(n-1)} \tag{21}$$

a má rozdělení rozdělení χ^2 s $n-1$ stupni volnosti.

5.3.4 *Výběrový podíl*

Normovaný výběrový podíl je

$$\boxed{Z = \frac{p-\pi}{\sqrt{p(1-p)}}\sqrt{n} \quad \sim N(0;1)} \tag{22}$$

a má přibližně normální normované rozdělení $N(0;1)$. To platí pro $np > 5$ a zároveň $n(1-p) > 5$.

Dva náhodné výběry

Uvažujeme dva výběry, tj. dvě náhodné veličiny, které chceme zkoumat. Například měříme nahuštění předních pneumatik u náhodně vybraných automobilů. Tlak v levé pneumatice je realizací první náhodné veličiny, tlak v pravé je realizací druhé náhodné veličiny. Charakteristikami budou opět **průměr, rozptyl** a **podíl** a vztahují se na rozdíl obou náhodných veličin. Předpokládáme, že rozptyl rozdělení není znám a je nahrazen výběrovým rozptylem.

Značíme: x_1, x_2 výběry, n_1, n_2 rozsahy výběrů, μ_1, μ_2 střední hodnoty výběrů a σ_1^2, σ_2^2 rozptyly výběrů.

Rozdíl dvou výběrových průměrů při shodných rozptylech

$$\boxed{t = \frac{\overline{X}_1-\overline{X}_2-(\mu_1-\mu_2)}{S_P\sqrt{1/n_1+1/n_2}} \quad \sim St(n-1)}, \qquad S_P^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}. \tag{23}$$

a má Studentovo rozdělení s $n-1$ stupni volnosti.

Rozdíl dvou výběrových průměrů při různých rozptylech

$$\boxed{t = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \quad \sim St(n-1)} \tag{24}$$

a má Studentovo rozdělení s δ stupni volnosti, kde

$$\delta = \frac{(k_1 + k_2)^2}{\frac{k_1^2}{n_1 - 1} + \frac{k_2^2}{n_2 - 1}}, \qquad k_1 = \frac{s_1^2}{n_1}, \quad k_2 = \frac{s_2^2}{n_2}$$

Rozdíl dvou výběrových průměrů při párových výběrech

$$\boxed{t = \frac{\overline{D} - (\mu_1 - \mu_2)}{S_D}\sqrt{n} \quad \sim St(n-1)}\,, \tag{25}$$

kde $D = X_1 - X_2$, $\overline{D}$ je výběrový průměr a S_D výběrový rozptyl náhodného vektoru D.

$$D_i = X_{1,i} - X_{2,i}; \qquad \overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i; \qquad S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2$$

Tato charakteristika má Studentovo rozdělení s $n-1$ stupni volnosti.

Příklad

Sledujeme nahuštění předních pneumatik u osobních automobilů. U každého auta změříme tlak v levé pneumatice (prvek výběru 1) a v pravé pneumatice (prvek výběru 2). Tedy, pro každý objekt (automobil) měříme dvě vlastnosti (pravou a levou pneumatiku). Tak dostaneme párové výběry.

Poznámka

Párový rozdíl výběrových průměrů dostaneme tak, že odečteme položky jednotlivých výběrů, tak dostaneme výběr D a z něho sestavíme obyčejný normovaný výběrový průměr (20).

Podíl dvou výběrových rozptylů

$$\boxed{F = \frac{S_1^2}{S_2^2} \quad \sim F(n_1 - 1, n_2 - 1).} \tag{26}$$

Tato charakteristika má F rozdělení s $n_1 - 1$ stupni volnosti pro čitatele a $n_2 - 1$ stupni volnosti pro jmenovatele.

Rozdíl dvou výběrových podílů p_1 a p_2 pro dvě alternativní rozdělení s podíly π_1 a π_2 je

$$\boxed{Z = \frac{p_1 - p_2 - (\pi_1 - \pi_1)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}} \quad \sim N(0;1)} \tag{27}$$

a má přibližně normální rozdělení $N(0;1)$.

6 Bodové odhady a jejich vlastnosti

6.1 Statistika

Výběr pořizujeme proto, abychom se (více) dověděli o souboru, ze kterého jsme výběr pořídili. Zde se soustředíme na situaci, kdy známe rozdělení souboru až na jeden nebo více parametrů. Např. víme, že rozdělení souboru je normální, ale neznáme jeho střední hodnotu, případně i rozptyl. Z výběru se snažíme hodnoty těchto neznámých parametrů odhadnout. Předpis, pomocí kterého z výběru vypočteme hodnotu neznámého parametru se nazývá **statistika**. V souladu s tím je i následující definice.

6.1.1 Statistika

Statistika $T = T(X)$ je funkce výběru X.

Komentář k definici

1. Statistika určená pro odhadování se nazývá **odhadová statistika**, pro testování **testová statistika**.
2. Definice neříká nic o tom, jak statistiku volit vzhledem k jejímu cílovému využití (odhad, test). Její vhodnost či nevhodnost budeme zkoumat později.

6.2 Bodový odhad parametru rozdělení

6.2.1 Bodový odhad

Sledujeme rozdělení s hustotou pravděpodobnosti $f(x;\theta)$ s neznámým parametrem θ. Provedli jsme realizaci náhodného výběru $x = (x_1, x_2, \ldots, x_n)$ z tohoto rozdělení a definovali statistiku $T(X)$. **Bodový odhad** $\hat{\theta}$ parametru θ pro realizaci náhodného výběru x je hodnota statistiky T s dosazenou realizací náhodného výběru x

$$\hat{\theta} = T(x)$$

Komentář k definici

1. Pro každou novou realizaci výběru obdržíme jiný bodový odhad. Odtud je zřejmé, že bodový odhad nemůže dát úplně přesnou hodnotu parametru.
2. Vlastní volbu statistiky jsme zatím nechali stranou. Lze pro ni použít metodu momentů nebo maximální věrohodnosti, o které se zmíníme. Statistiku je také možno volit heuristicky, potom je však třeba ověřit její vlastnosti.

6.2.2 Příklad

Odhadujeme parametr střední hodnotu μ. Provedli jsme výběr

$$x = \{x_1, x_2, \ldots, x_n\} = \{3, 5, 2, 4, 5\}.$$

Protože víme, že střední hodnotu si lze přibližně představit jako průměr všech možných realizací, usoudíme, že jejím vhodným odhadem bude aritmetický průměr $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Po dosazení realizace výběru dostaneme $\overline{x} = 19/5$. Vlastnosti zvolené statistiky je však třeba ověřit (viz dále).

6.3 Vlastnosti bodových odhadů

Vlastnosti bodového odhadu $\hat{\theta}$ vypovídají o vhodnosti použité statistiky T k odhadu hodnoty parametru θ. Uvedeme tři vlastnosti: **nestrannost**, **konzistenci** a **vydatnost**.

6.3.1 Nestrannost

Statistika T poskytuje **nestranný bodový odhad** parametru θ, jestliže její střední hodnota se rovná tomuto parametru

$$E[T] = \theta \tag{28}$$

Komentář k definici

1. Střední hodnotu $E[T]$ je třeba chápat jako „průměrování“ přes všechny možné výběry. Jestliže bychom chtěli naznačit výpočet této střední hodnoty, bylo by třeba postupovat takto: provedeme první výběr, spočteme bodový odhad, provedeme druhý výběr a opět spočteme bodový odhad, atd. Po provedení všech možných výběrů uděláme průměr ze všech jednotlivých bodových odhadů. To je hledaná střední hodnota.
2. Nestrannost říká, že odhad je „v průměru“ (rozumíme přes všechny možné výběry) přesný.

6.3.2 Příklad

Ověříme nestrannost výběrového průměru vzhledem ke střední hodnotě. Máme dokázat, že platí $E[\overline{X}] = \mu$. Dosadíme definici výběrového průměru a manipulujeme s operátorem střední hodnoty

$$E[\overline{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

Vychýlení bodového odhadu $B(\theta)$ je definováno jako rozdíl mezi střední hodnotou statistiky a odhadovaným parametrem

$$B(\theta) = E(T) - \theta \tag{29}$$

Poznámka

Z definice nevychýlenosti přímo plyne, že je-li statistika T pro odhad parametru θ nevychýlená, pak vychýlení $B(\theta)$ je nulové.

6.3.3 Konzistence

Statistika T dává **konzistentní bodový odhad** parametru θ, jestliže pro rostoucí rozsah výběru se hodnota statistiky (v pravděpodobnosti) neomezeně blíží skutečnému parametru

$$\lim_{n\to\infty} P(|T-\theta|<\epsilon)=1, \quad \forall \epsilon>0 \tag{30}$$

Komentář k definici

Tato vlastnost je limitní. Lze ji sledovat jen při zvětšujícím se rozsahu výběru – např. změříme 10 hodnot, pak 100, atd.

Kriterium konzistence

Odhad je konzistentní, jestliže je

– asymptoticky nestranný,
– jeho rozptyl jde k nule s rozsahem výběru jdoucím k nekonečnu.

Příklad

Ověříme konzistenci výběrového průměru vzhledem k odhadu střední hodnoty. Pro důkaz využijeme Čebyševovu nerovnost, kterou zapíšeme pro výběrová průměr a střední hodnotu

$$P(|\overline{X}-\mu|<\epsilon) > 1-\frac{\sigma^2}{n\epsilon^2}$$

Protože $\lim_{n\to\infty}\frac{\sigma^2}{n\epsilon^2}=0$, je odhad konzistentní.

6.3.4 Vydatnost

Pro dvě nestranné statistiky T a U definujeme jako vydatnější tu z nich, která má menší rozptyl

$$D[T]<D[U] \quad \Longrightarrow \quad T \text{ je vydatnější než } U \tag{31}$$

6.4 Konstrukce bodových odhadů

Řekli jsme co je bodový odhad a jaké u něho sledujeme vlastnosti. Nyní se budeme věnovat otázce, jak lze takový odhad zkonstruovat. Ukážeme dvě základní metody pro konstrukci statistiky, vhodné pro odhad daného parametru.

6.4.1 Metoda momentů

Tato metoda je velmi jednoduchá, obecně však nedává příliš kvalitní výsledky. Spočívá v porovnání obecných (nebo centrálních) momentů souboru a výběru. Podle toho, kolik parametrů odhadujeme, tolik momentů musíme porovnat. Momenty souboru počítáme s pomocí hustoty pravděpodobnosti souboru $f(x, \theta)$. Budou tedy obsahovat neznámý parametr θ. Výběr je množina změřených hodnot. Moment výběru bude tedy číslo. Porovnáním momentů získáme rovnice kde neznámé budou odhadované parametry. Z nich odhad vypočteme.

Budeme odhadovat neznámý parametr δ exponenciálního rozdělení s hustotou pravděpodobnosti $f(x, \delta) = \delta \exp\{-\delta\, x\},\ \delta > 0,\ x \in (0, \infty)$ z výběru $x = [x_1, x_2, \ldots, x_n]$.

Protože odhadujeme jediný parametr, stačí porovnat první momenty, tj. střední hodnotu souboru (exponenciálního rozdělení) a výběrový průměr změřeného výběru.
Střední hodnota souboru je

$$E[X] = \int_0^\infty x\, f(x, \delta)\, dx = \int_0^\infty x\, \delta \exp\{-\delta x\} dx = 1/\delta.$$

Výběrový průměr je

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \text{ číslo.}$$

Porovnáním dostaneme

$$\frac{1}{\delta} = \overline{x} \quad \Rightarrow \quad \hat{\delta} = \frac{1}{\overline{x}},$$

kde symbolem $\hat{\delta}$ jsme označili bodový odhad parametru δ.

6.4.2 Metoda maximální věrohodnosti

Odhad pro obecné rozdělení

Tato metoda dává velmi kvalitní výsledky a je často používána. Pro normální rozdělení souboru je ekvivalentní s metodou nejmenších čtverců, o které budeme mluvit v regresní analýze. Metoda je založena na minimalizaci tzv. **věrohodnostní funkce** nebo jejím logaritmu (což je totéž– proč?).

Věrohodnostní funkce

Pro rozdělení s hustotou pravděpodobnosti $f(x, \theta)$ a realizaci náhodného výběru $x = [x_1, x_2, \ldots, x_n]$ definujeme věrohodnostní funkci $\mathcal{L}_n(\theta)$ vztahem

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(x_i, \theta). \tag{32}$$

Maximálně věrohodný odhad

Maximálně věrohodným odhadem parametru θ rozdělení $f(x,\theta)$ nazveme odhad $\hat{\theta} \in \theta^*$, který maximalizuje (logaritmus) věrohodnostní funkce, tj. platí

$$\ln \mathcal{L}_n(\hat{\theta}) \geq \ln \mathcal{L}_n(\theta), \quad \forall \theta \in \theta^* \tag{33}$$

kde θ^* označuje množinu všech přípustných hodnot parametru θ.

Obecný postup při určení maximálně věrohodného odhadu je následující:

1. Sestavíme věrohodnostní funkci tak, že násobíme hustoty pravděpodobnosti souboru a do každé dosadíme za x jeden prvek výběru x_i.
2. Je-li to výhodné, věrohodnostní funkci logaritmujeme. Protože řada rozdělení má hustotu pravděpodobnosti ve tvaru exponenciály, bývá logaritmus užitečný pro zjednodušení součinu. Není však nutný.
3. Hledáme maximum logaritmu věrohodnostní funkce. V jednoduchém případě např. pomocí derivace, ve složitějším případě numericky.

   Bod, kde leží maximum je maximálně věrohodný odhad.

Odhad pro exponenciální třídu rozdělení

Exponenciální třída rozdělení - řekneme, že hustota pravděpodobnosti patří do exponenciální třídy, jestliže je možno ji zapsat ve tvaru

$$f(x,\theta) = \exp\{Q(\theta)\,U(x) + R(\theta) + V(x)\}, \tag{34}$$

kde Q, R jsou funkcemi jen θ (ne x) a U, V jsou funkcemi jen x (ne θ).

Alternativní rozdělení je z exponenciální třídy, neboť platí

$$f(x,\pi) = \pi^x(1-\pi)^{1-x} = \exp\{\log(\pi^x(1-\pi)^{1-x})\} =$$

$$= \exp\{x\log(\pi) + (1-x)\log(1-\pi)\} = \exp\{[\log(\pi) - \log(1-\pi)]\,x + \log(1-\pi)\}.$$

Zde platí: $Q = \log(\pi) - \log(1-\pi), \quad U = x, \quad R = \log(1-\pi), \quad V = 0.$

Je-li rozdělení z exponenciální třídy, dostáváme následující jednoduché řešení úlohy maximálně věrohodného odhadu.

Max. věr. odhad pro exponenciální třídu

Pro rozdělení s hustotou pravděpodobnosti $f(x,\theta) = \exp\{Q(\theta)\,U(x) + R(\theta) + V(x)\}$ a realizaci výběru $x = [x_1, x_2, \ldots, x_n]$ je extrém logaritmické věrohodnostní funkce dán řešením rovnice

$$Q'(\theta)\,S(x) + n\,R'(\theta) = 0,$$

kde Q', R' jsou derivace podle θ a $S(x) = \sum_{i=1}^{n} U(x_i)$.

Aby nalezený extrém byl maximum, musí být ještě splněna nerovnost

$$Q''(\theta)S(x) + n\,R'' < 0.$$

Důkaz:

Věrohodnostní funkce je

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} \exp\{Q(\theta)\,U(x_i) + R(\theta) + V(x_i)\} = \exp\left\{\sum_{i=1}^{n}(Q(\theta)\,U(x_i) + R(\theta) + V(x_i))\right\} =$$

$$= \exp\left\{Q(\theta)\sum_{i=1}^{n} U(x_i) + nR(\theta) + \sum_{i=1}^{n} V(x_i)\right\}.$$

Tento výraz logaritmujeme a se zavedeným značením dostaneme

$$\log \mathcal{L}_n(\theta) = Q(\theta)S(x) + nR(\theta) + \sum_{i=1}^{n} V(x_i).$$

Po derivaci podle θ poslední výraz zmizí. Anulováním derivace dostáváme podmínku pro extrém. Podmínka pro maximum je záporná druhá derivace.

Příklad

Metodou maximální věrohodnosti určete odhadovou statistiku pro parametr π alternativního rozdělení.

Uvažujeme alternativní rozdělení, jehož exponenciální tvar jsme již odvodili (viz příklad k (34)) a zjistili jsme, že platí $Q = \log(\pi/(1-\pi)),\quad U = x,\ R = \log(1-\pi),\quad V = 0$. Abychom mohli použít příslušné vzorce, potřebujeme ještě spočítat funkci S a její derivace.

$S = \sum_{i=1}^{n} x_i = n\overline{x},\quad Q' = \frac{1}{\pi(1-\pi)},\quad Q'' = -\frac{2\pi-1}{\pi^2(1-\pi)^2},\quad R' = -\frac{1}{1-\pi},\quad R'' = -\frac{1}{(1-\pi)^2}$

Podmínka extrému

$$Q'S + nR' = \frac{1}{\pi(1-\pi)}n\overline{x} - n\frac{1}{1-\pi} = 0 \quad \Rightarrow \quad \hat{\pi} = \overline{x}$$

Podmínka maxima (s dosazeným odhadem $\overline{x} = \hat{\pi}$)

$$-\frac{2\hat{\pi}-1}{\hat{\pi}^2(1-\hat{\pi})^2}n\hat{\pi} - n\frac{1}{(1-\hat{\pi})^2} < 0 \quad \Rightarrow \quad \hat{\pi} < 1$$

což je vždy splněno, neboť $\pi = 1$ je patologický případ.

7 Intervalové odhady

7.1 Pojem intervalu spolehlivosti

V minulé kapitole jsme se zabývali bodovým odhadem neznámého parametru souboru. Tento odhad zkonstruujeme tak, že (i) zvolíme odhadovou statistiku, (ii) ověříme její vhodnost

vzhledem k odhadovanému parametru, (iii) dosadíme realizaci výběru. Hodnota statistiky s dosazeným výběrem je bodový odhad.

Bodové odhady poskytují hodnotu odhadu parametru, ale neříkají nic o jeho přesnosti. Ta je zahrnuta v intervalu spolehlivosti - tj. intervalu, ve kterém leží neznámý parametr s danou pravděpodobností. Přesnost (nebo neurčitost) odhadu je dána šířkou intervalu.

7.1.1 Interval spolehlivosti

Interval $I_\alpha = (\theta_D, \theta_H)$ nazveme α-interval spolehlivosti pro parametru θ, jestliže platí

$$P(\theta \in I_\alpha) = 1 - \alpha. \tag{35}$$

Komentář k definici

1. Interval spolehlivosti budeme zkracovat IS.
2. I_α se také nazývá $100(1-\alpha)$-procentní interval spolehlivosti, tj. 0.05-IS je 95% IS.
3. Pokud je $-\infty < \theta_D$ a $\theta_H < \infty$, tj. $I_\alpha$je zdola i shora omezený, hovoříme o **oboustranném IS**. Je-li $\theta_D = -\infty$, tj. $I_\alpha = (-\infty, \theta_H)$, jedná se o **pravostranný IS**, pro $\theta_H = \infty$, tj. $I_\alpha = (\theta_D, \infty)$ jde o **levostranný IS**.

Uvedená definice říká pouze co interval spolehlivosti je, nikoliv jak ho určíme. Pro konstrukci intervalu spolehlivosti se budeme opírat o bodové odhady, což jsou hodnoty statistiky po dosazení konkrétní realizace výběru. Statistika představuje určité "kukátko", přes které neznámý parametr sledujeme. Jaké kukátko vezmeme, tak neznámý paramertr vidíme. Má-li statistika velký rozptyl, vidíme paramtr s velkými chybami. Je-li statistika vychýlená, vidíme jeho polohu polohu posunutou.

Poznámka

V jednoduchých případech, kterými se zde zabýváme, tj. odhad střední hodnoty, rozptylu a podílu, vychází odhadová statistika jednoznačně. Ve složitějších případech může být situace komplikovanější a závislost bodového odhadu na volbě statistiky je třeba si uvědomit.

Odvození IS přímo z definice není možné. Parametr θ je neznámé číslo, které "pozorujeme" pouze přes hodnoty statistiky. Musíme proto vzít na pomoc vhodnou statistiku. Tou je právě statistika pro konstrukci bodového odhadu.

Podle definice (35) máme určit interval I_α, ve kterém bude ležet neznámý parametr θ s pravděpodobností $1-\alpha$. Tento požadavek vyjádříme přibližně jako interval, do kterého teoreticky padne $(1-\alpha)\times 100\%$ všech bodových odhadů. Tento nový požadavek můžeme vyjádřit vztahem

$$P\left(T\left(X\right) \in I_\alpha\right) = 1 - \alpha,$$

kde T je zvolená odhadová statistika a X je náhodný výběr, který jsme zapsali ve tvaru náhodného vektoru abychom zdůraznili jeho potenciální opakování.

Podle tohoto nového vyjádření lze také konstrukci intervalu spolehlivosti provést. Pro nejjednodušší případ ji budeme ilustrovat v následujícím příkladě.

Příklad

Určete 95% IS pro střední hodnotu μ normálního rozdělení s rozptylem $\sigma^2 = 1$ na základě realizace výběru o rozsahu $n = 100$ s realizací výběrového průměru jako statistiky $\overline{x} = 3.7$.

Podle požadavků na IS musí platit

$$P(\mu_D < \overline{X} < \mu_H) = 1 - \alpha.$$

Normujeme

$$P(\zeta_{\frac{\alpha}{2}} < \frac{\overline{X} - \mu}{\sigma}\sqrt{n} < z_{\frac{\alpha}{2}}) = 1 - \alpha$$

a v argumentu vyjádříme μ a použijeme vztah $\zeta_{\frac{\alpha}{2}} = -z_{\frac{\alpha}{2}}$, zároveň za výběrový průměr dosadíme jeho realizaci

$$-z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \overline{X} - \mu < z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \qquad \Rightarrow \qquad \overline{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}.$$

Upravenou podmínku dosadíme zpět

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

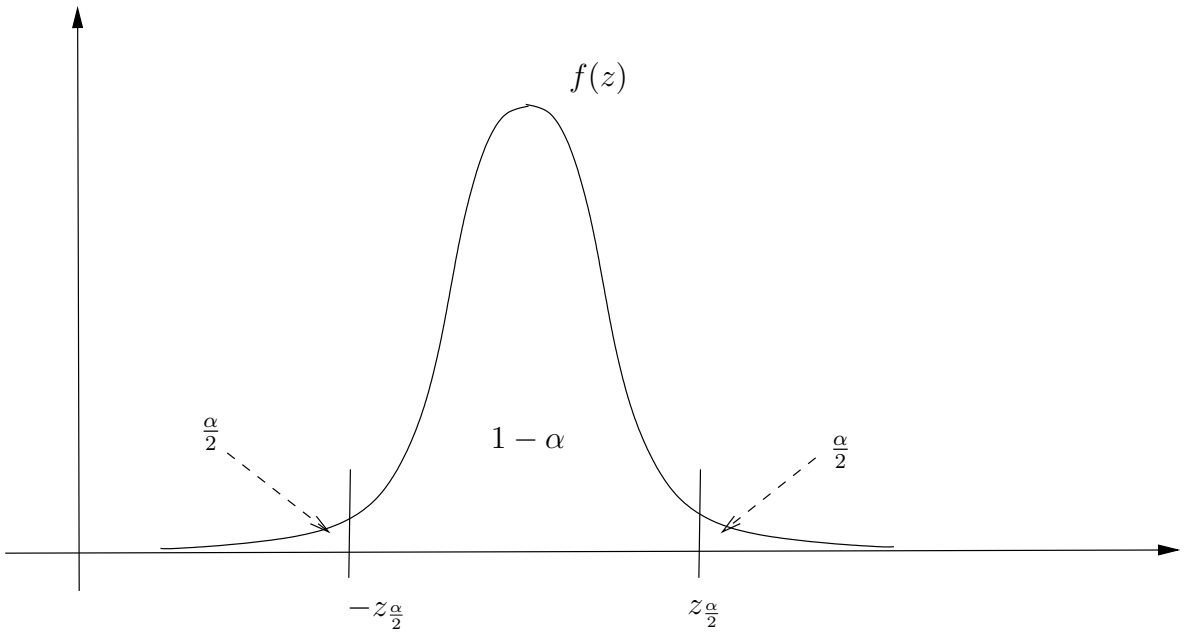
Podle definice a s dosazenou realizací výběrového průměru dostáváme IS ve tvaru

$$I_\alpha = (\overline{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}})$$

nebo v jiném zápisu

$$\mu = \overline{x} \pm z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}.$$

Na obrázku je hustota pravděpodobnosti bodového odhadu, tj. výběrového průměru, a na ní jsou vyznačeny příslušné pravděpodobnosti, vymezující pravděpodobnost $1 - \alpha$. Hranicemi intervalu (v normované oblasti) jsou kvantil $-z_{\frac{\alpha}{2}}$ a kritická hodnota $z_{\frac{\alpha}{2}}$. Jestliže tyto hraniční body "odnormujeme", dostáváme přímo interval spolehlivosti.



Konkrétně, po dosazení zadaných hodnot,

$$\alpha = 0.05/2 = 0.025; \qquad \overline{x} = 3.7; \qquad \sigma = 1; \qquad n = 100; \qquad z_{\frac{\alpha}{2}} = 1.96 \ \text{(z tabulek)}$$

dostáváme interval spolehlivosti ve tvaru

$$I_{0.05} = (3.7 - \frac{1}{10}1.96;\ 3.7 + \frac{1}{10}1.96) = (3.504;\ 3.896).$$

Poznámka

V předchozích úvahách jsme používali názvy normovaná a nenormovaná oblast. O normování a "odnormování" náhodné veličiny jsme již mluvili v pravděpodobnosti. Připomeňme. Pro náhodnou veličinu X se střední hodnotou μ a rozptylem σ^2 určíme normovanou veličinu Z

$$Z = \frac{X - \mu}{\sigma}$$

pro niž platí $E[Z] = 0$ a $D[Z] = 1$. Vzorec pro odnormování dostaneme vyjádřením X

$$X = \sigma Z + \mu.$$

Pro výběrový průměr $\bar{X}$ se střední hodnotou μ a rozptylem $\frac{\sigma^2}{n}$ mají předchozí vzorce tvar

$$\bar{Z} = \frac{\bar{X} - \mu}{\sigma}\sqrt{n} \;\; a \;\; \bar{X} = \frac{\sigma}{\sqrt{n}}\bar{Z} + \mu$$

Význam těchto úprav je následující: nenormovaná oblast je z reálného světa. Například když sleduji rychlost projíždějících automobilů a nazvu ji X, pak $X \in (40, 120)$ znamená, že skutečně vybírám rychlosti mezi 40 km/h a 120 km/h. Nevýhodou ale je, že kdybychom chtěli určit např., jakou pravděpodobnost mají rychlosti větší než 90 km/h, budeme bez počítače v potížích. I když budeme předpokládat nějaké známé rozdělení náhodné veličiny X, v tabulkách potřebné hodnoty nenajdeme. Kdyby totiž měla být tabelována rozdělení pro každou hodnotu svého parametru zvlášť, třeba normální rozdělení pro každou hodnotu μ a σ^2, dostali bychom veletabulky. Proto je tabelováno jen normované rozdělení, tam určíme normované hodnoty a ty se pak "odnormují" do nenormované oblasti která odpovídá realitě.

7.2 Druhy intervalů spolehlivosti

Jak jsme se již také zmínili, souborových parametrů, pro které je možno počítat intervaly spolehlivosti, je celá řada. Od těch nejjednodušších, kterými se zde budeme zabývat, až po velmi komplikované. Snad již bylo patrné z předchozího výkladu, že základním bodem při odvození intervalu spolehlivosti je nalezení vhodné statistiky a jejího rozdělení. A právě v tom, že potřebujeme znát rozdělení statistiky, je hlavní potíž. Hledání takového rozdělení je úloha transformace náhodné veličiny s předpokládaným (většinou normálním rozdělením) podle funkce, kterou je statistika definována. To je ve složitějších případech úloha velmi komplikovaná.

V našem případě se omezíme na dva typy souborů, a to

- spojitou náhodnou veličinu s **normálním rozdělením** se střední hodnotou μ a rozptylem σ^2,
- diskrétní náhodnou veličinu s **alternativním rozdělením** a se souborovým podílem π.

Spojité rozdělení je jasné. Připomeňme situaci spojenou s alternativním rozdělením. Typickým příkladem je sklad s výrobky. Náhodná veličina je definována jako kvalita výrobků s hodnotami 0 = kvalitní výrobek, 1 = nekvalitní výrobek. Potom parametr π označuje podíl nekvalitních výrobků v celkovém počtu výrobků ve skladu.

Parametry o které se budeme zajímat a které budeme odhadovat jsou právě parametry těchto souborů, tedy **střední hodnota** μ, **rozptyl** σ^2 pro spojitou náhodnou veličinu a **podíl** π pro diskrétní náhodnou veličinu.

Při odhadu střední hodnoty spojité náhodné veličiny se objevuje ještě malá komplikace. Je třeba odlišit dvě situace.

1. **Rozptyl souboru známe:** V tomto případě musíme znát skutečně rozptyl souboru, a to z jiných, hlubších, zdrojů, než je pouhý výběr.

2. **Rozptyl souboru neznáme:** Tady rozptyl souboru nená znám a místo něho musíme použít odhad, který se pořídí z výběru.

Poznámka

Tady je potřeba dát pozor. Jde opravdu o znalost souborového rozptylu, tedy o znalost hlubší, než jen tu, založenou na datech z výběru. Jestliže je v zadání příkladu: Změřili jsme výběr a spočítali rozptyl $\cdots$ Odhadněte střední hodnotu $\cdots$, pak, i když je rozptyl číselně zadán, je třeba jej považovat za neznámý, protože je spočten pouze na základě výběru. Jiná situace je v příkladě: Z dlouhodobého pozorování víme, že rozptyl $\cdots$ Provedli jsme výběr a máme odhadnout střední hodnotu. Zde je možno rozptyl považovat za známý, protože jeho znalost je založena na hlubších podkladech než jen na výběru.

Kromě odhadu parametrů jediného souboru uvedeme ještě také tzv. dvouvýběrové odhady, při kterých sledujeme dva soubory (tj. dvě náhodné veličiny). Jestliže náhodné veličiny označíme X_i, $i = 1, 2$ s parametry μ_i, σ_i^2 nebo π_i, $i = 1, 2$, potom budeme odhadovat rozdíl dvou středních hodnot nebo podílů $\mu_1 - \mu_2$ nebo $\pi_1 - \pi_2$ případně podíl dvou rozptylů σ_1^2/σ_2^2.

Poznámka

Proč odhadujeme zrovna tyto výrazy? Protože rozdíl rovný nule nebo podíl rovný jedné vypovídá o shodě obou parametrů. Odhady rozdílů blízké nule nebo odhad podílu blízký jedné svědčí tedy o tom, ze zkoumané parametry jsou si rovny.

Při dvou-výběrovém odhadu rozdílu středních hodnot se opět objevuje malá komplikace. Je třeba rozlišit tři druhy výběrů, ze kterých odhad vychází.

1. **Sdružený výběr:** Výběry se provádí nezávisle (mohou mít i rozdílné délky). Podmínka je, že rozptyly obou souborů jsou přibližně stejné (dejme tomu, že se neliší řádově).

2. **Nesdružený výběr:** Výběry jsou opět nezávislé, avšak o rozptylech souborů platí, že jsou rozdílné.

3. **Párový výběr:** V tomto případě vybíráme prvky výběrů v párech. Většinou se tyto výběry pořizují tak, že bereme jeden objekt za druhým a na každém změříme vždy dvě hodnoty. Hodnoty z různých objektů se pak při dalším zpracování nemíchají.
   Specifickou charakteristikou tohoto výběru je, že i když se hodnoty veličin z různých objektů třeba i značně liší, ve výsledcích se to neprojeví. Rozhodující je vždy jen vztah hodnot z jednotlivých párů.

Jednotlivé typy výběru budeme ilustrovat na příkladě.

Příklad

Sdružený nebo nesdružený výběr (nezávislé výběry):
Testujeme hlučnost vozů Škoda a Peugeot. Náhodně jsme vybrali 30 vozů od každé značky (může být i různý počet) a podrobili je testu $\cdots$

Párový výběr:
Srovnáváme stupeň opotřebení brzdového obložení předních kol vozů Opel. Náhodně jsme vybrali 35 vozů a u každého jsme zaznamenali stupeň ojetí levé a pravé brzdy. Evidentně, počet záznamů pro levou i pravou brzdu musí být stejný.

V následujících odstavcích se budeme věnovat jednotlivým případům odhadu. Uvedeme vždy vzorec pro interval a funkci, kterou je možno ve Scilabu příslušný interval spočítat.

7.2.1 Střední hodnota (známé σ^2)

Vzorec

$$I_\alpha : \quad \mu \in \overline{x} \pm \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \qquad z \sim N(0;1)$$

Funkce:

[dolni_mez, horni_mez]=**z_int** (stredni_hodnota,rozptyl,n,strana,alpha)

7.2.2 Střední hodnota (neznámé σ^2)

Vzorec

$$I_\alpha : \quad \mu \in \bar{x} \pm \frac{s}{\sqrt{n}} t_{\alpha/2}, \qquad t \sim St(n-1)$$

Funkce

[dolni_mez, horni_mez]=**t_int** (stredni_hodnota,rozptyl,n,strana,alpha)

7.2.3 Rozptyl

Vzorec

$$I_\alpha : \quad \left(\frac{(n-1)s^2}{\chi^2_{\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}} \right), \qquad \chi^2 \sim Chi(n-1)$$

Funkce

[dolni_mez, horni_mez]=**var_int**(rozptyl,n,strana,alpha)

7.2.4 Podíl

Vzorec

$$I_\alpha : \quad \pi \in p \pm \sqrt{\frac{p(1-p)}{n}} z_{\alpha/2}, \qquad z \sim N(0;1)$$

Funkce

[dolni_mez, horni_mez]=**prop_int**(p,n,strana,alpha)

Příklad

V jakém intervalu lze očekávat životnost zakoupené pneumatiky s pravděpodobností 0.95, jestliže pro 10 náhodně `vybraných` pneumatik byly zjištěny následující životnosti (v rocích)

$$2 \quad 4 \quad 5 \quad 6 \quad 10 \quad 8 \quad 6 \quad 5 \quad 6 \quad 7.$$

Přípravné výpočty:

$$\bar{x} = 5.9, \qquad s = 2.183, \qquad t_{0.025}(9) = 2.262$$

Interval:

$$I_{0.05} = (4.34;\ 7.46)$$

8 Parametrické testy hypotéz

8.1 Pojem parametrického testu

V předchozí kapitole jsme hovořili o intervalech spolehlivosti. Parametr sledovaného souboru jsme odhadovali intervalem, ve kterém jeho hodnota leží s velkou pravděpodobností $(1 - \alpha$, kde α je malé). To znamená, že odhady, ležící mimo interval spolehlivosti jsou velmi nepravděpodobné nebo, což považujeme za stejné, skutečné hodnoty parametru mimo interval jsou velmi nepravděpodobné.

V této kapitole budeme vlastně přímo navazovat na intervalové odhady. Budeme uvažovat dva soubory. První bude s danou hodnotou sledovaného parametru, o druhém budeme předpokládat, že hodnota jeho parametru je jiná. Otázka je, který z těchto dvou souborů skutečně generoval data, která jsme pořídili výběrem. Řešení je jednoduché. Sestrojíme interval spolehlivosti pro první soubor. Pokud v něm uvažovaná hodnota jeho parametru leží, jsme spokojeni a s touto hodnotou můžeme i nadále počítat. Pokud ale hodnota jeho parametru leží mimo sestrojený interval spolehlivosti, máme buď velikou "smůlu" při pořizování výběru, nebo není pravda, že hodnota parametru je ta, co jsme dosud uvažovali. Na velkou náhodu nevěříme, proto se přikloníme k druhému tvrzení a hodnotu parametru prvního souboru zamítneme.

Ve skutečnosti jsme dva soubory uvedli jen pro lepší představu. Máme jen jeden soubor a na něm sledujeme jeden nebo více parametrů. O hodnotě parametru tohoto souboru pochybujeme a proto o nich vyslovujme tvrzení (hypotézy). První hypotéza (nulová) obhajuje tu hodnotu, která doposud platila. Druhá hypotéza (alternativní) tu první hypotézu popírá. Říká, že parametr je větší, nebo menší, nebo prostě jen že je jiný, než se původně myslilo.

Podobně jako u intervalů spolehlivosti, budeme se zabývat testy parametrů jednoho souboru nebo budeme testovat rozdíl (nebo podíl) parametrů dvou souborů. Budeme mluvit o jednovýběrových nebo dvou-výběrových testech.

8.2 Formulace úlohy testování

Uvažujeme **soubor**, tvořený náhodnou veličinou X s rozdělením $f(x, \theta)$, kde θ je neznámý parametr rozdělení, jehož hodnotu testujeme.

Nulová hypotéza H_0 je tvrzení, které obhajuje stávající stav, tedy říká, že $\theta = \theta_0$.

Alternativní hypotéza H_A popírá tvrzení nulové hypotézy. Říká buď $\theta > \theta_0$ nebo $\theta < \theta_0$ nebo $\theta \neq \theta_0$.

Testová statistika T je stejná jako statistika pro odhad testovaného parametru.

Realizovaná testová statistika T_r je hodnota statistiky vypočítaná pro konkrétní dosazenou realizaci výběru.

Obor přijetí O je normovaný interval spolehlivosti pro odhad parametru podle nulové hypotézy (je to množina hodnot statistiky, pro které nulovou hypotéuzu nezamítáme).

Kritický obor W je doplněk oboru přijetí $W = R - O$ (je to množina hodnot statistiky, pro které nulovou hypotézu zamítáme).

Hladina významnosti je pravděpodobnost α se kterou počítáme interval spolehlivosti - obor přijetí.

Směrování testu určuje, tvar a polohu kritického oboru.

alternativní hypotéza		**kritický obor**
• θ podle H_A **je větší** než podle H_0,	$\theta > \theta_0$	(T_α, ∞)
• θ podle H_A **je menší** než podle H_0,	$\theta < \theta_0$	$(-\infty, T_{1-\alpha})$
• θ podle H_A **je jiný než** podle H_0,	$\theta \neq \theta_0$	$R - \left(T_{1-\frac{\alpha}{2}}, T_{\frac{\alpha}{2}}\right)$

kde θ_0 je hodnoty parametru θ podle nulové hypotézy, T je testová statistika a T_α je její α-kritická hodnota (pro symetrické rozdělení, jako např. normální, bude $T_{1-\alpha} = z_{1-\alpha} = -z_\alpha$).

Příklad

Na volné silnici sledujeme rychlosti projíždějících automobilů značky Škoda. Předpokládáme, že rychlosti mají normální rozdělení u kterého známe rozptyl (třeba daný poměrem nových a starších vozů) a střední hodnotu (z průzkumu, který probíhal před pěti lety). Testovaným parametrem je $\theta = \mu$ - střední hodnota rychlostí.

Nulová hypotéza obhajuje stávající hodnotu, tedy změřenou v průzkumu před pěti lety. Alternativní hypotéza hodnotu z průzkumu popírá. Tvrdí, že za pět let se podstatně změnil automobilový park směrem k silnějším automobilům a že tedy střední hodnota rychlostí je větší.

Poznámka

Z předchozího výkladu je patrné, že testy hypotéz jsou konstruovány na základě intervalového odhadu (parametru podle nulové hypotézy) Jediný formální rozdíl je v tom, že pro konstrukci intervalu spolehlivosti používáme nenormovaný tvar statistiky, např. pro střední hodnotu se známým rozptylem je to výběrový průměr $\bar{X}$, zatímco pro test použijeme normovaný výběrový průměr $z = \frac{\bar{X}-\mu_0}{\sigma}\sqrt{n}$. Připomeňme, že normování náhodné veličiny X na normovanou veličinu Z s nulovou střední hodnotou a jednotkovým rozptylem se obecně provádí podle vzorce

$$Z = \frac{X - \mu}{\sigma},$$

kde $\mu = E[X]$ a $\sigma = \sqrt{D[X]}$.

Poznámka

Zastavme se ještě u významu hladiny významnosti α. Definovali jsme jej jako pravděpodobnost intervalu spolehlivosti, který tvoří obor přijetí testu. Další význam tohoto koeficientu je, že představuje pravděpodobnost tzv. ***chyby prvního druhu****. To je chyba, která spočívá v tom, že nulovou hypotézu zamítneme, zatímco ona ve skutečnosti platí. Pravděpodobnost se kterou se dopustíme tohoto omylu můžeme předem zvolit - volbou koeficientu α. Nejčastěji používaná hodnota je 5% což je pravděpodobnost 0.05. Kromě chyby prvního druhu lze ještě definovat* ***chybu druhého druhu,*** *která spočívá v tom, že nezamítneme nulovou hypotézu a ona ve skutečnosti neplatí. Volbu velikosti chyby druhého druhu ale v rukou nemáme.*

8.3 Postup testování

V předchozím odstavci jsme nastínili podstatu testování hypotéz a uvedli základní názvosloví. Nyní ukážeme obecný postup při provádění testu a budeme jej ilustrovat na konkrétním příkladě, který jsme uvedli v minulém odstavci.

Příklad

Testujeme střední hodnotu rychlostí vozů Škoda. Chceme prokázat, že v současnosti jsou rychlosti větší než 95.7 km/h, jak bylo zjištěno při posledním průzkumu. Změřili jsme rychlosti u 100 náhodně vybraných automobilů a určili výběrový průměr $\bar{X} = 96.4\ km/h$. Rozptyl považujeme za známý $\sigma^2 = 13.8\ (km/h)^2$. Volba hladiny významnosti je na nás. Zvolíme běžnou hodnotu $\alpha = 0.05$.

1. Typ příkladu
   (a) testovaný parametr: jedna střední hodnota, známý rozptyl
   (b) nulová hypotéza: $\mu = \mu_0$
   alternativní hypotéza: $\mu > \mu_0$
   (c) typ testu: pravostranný (podle H_A)
2. Statistika: normální rozdělení
3. Hodnota statistiky
$$T_r = \frac{\bar{X} - \mu_0}{\sigma}\sqrt{n} = \frac{96.4 - 95.7}{\sqrt{13.8}}\sqrt{100} = 1.884$$
4. Kritická hodnota: normální rozdělení
$$z_\alpha = z_{0.05} = 1.645$$
5. Kritický obor
$$W = (z_\alpha, \infty) = (1.645,\ \infty)$$
6. Závěr: $T_r \in W$ - nulovou hypotézu zamítáme.
   Odpověď: Není pravda, že současná střední rychlost automobilů Škoda je 95.7 km/h (jak bylo dříve zjištěno průzkumem).

8.4 P-hodnota

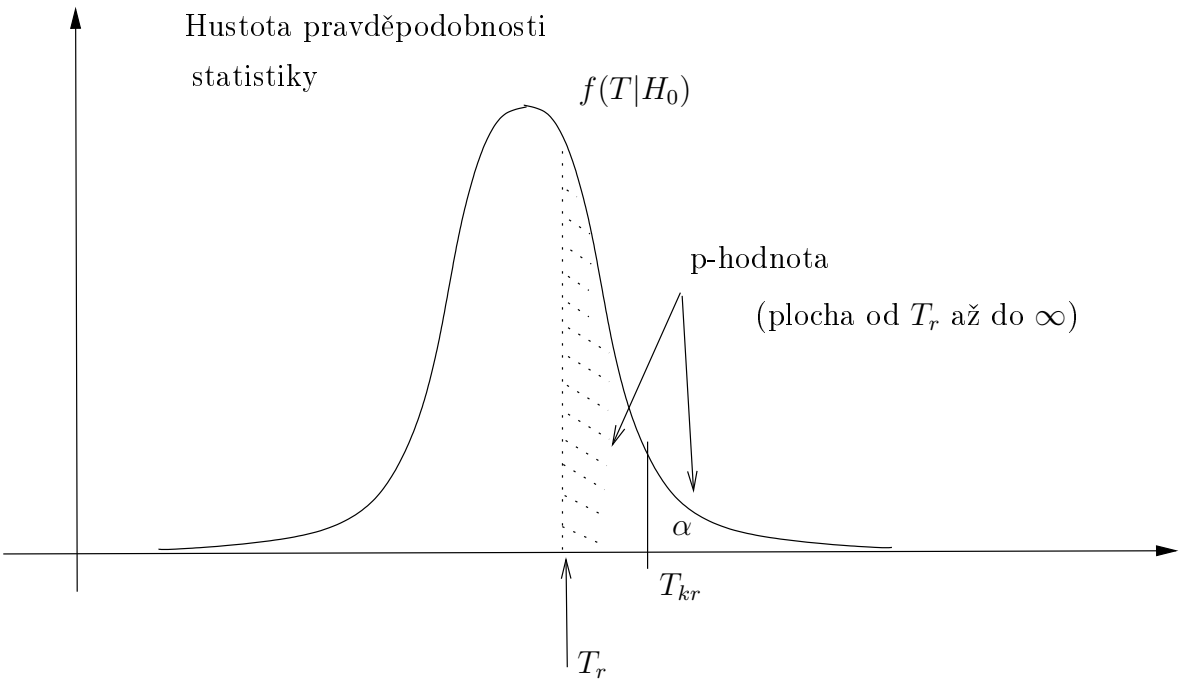
Z předchozího příkladu je patrné, že z výsledku klasického testu nepoznáme, „jak moc“ H_0 zamítáme nebo „jak daleko“ od zamítnutí se H_0 nachází. Abychom tento nedostatek odstranili, a také abychom výsledek testu vyjádřili jediným číslem, zavádíme **p-hodnotu** p_v. Ta je definována

- pro pravostranný test
$$p_v = P(T > T_r|H_0), \tag{36}$$
- pro levostranný test
$$p_v = P(T < T_r|H_0), \tag{37}$$

- pro oboustranný test

$$p_v = 2 \min \{P(T > T_r|H_0),\ P(T < T_r|H_0)\}\,. \tag{38}$$

P-hodnota pravostranného testu je plocha pod hustotou pravděpodobnosti normované statistiky, vpravo od realizované statistiky. Situace je zachycena na následujícím obrázku.



Z obrázku je patrné, že:

- bude-li $p_v = \alpha$, budeme na hranici zamítnutí,
- pro $p_v < \alpha$ leží realizovaná statistika v kritickém oboru W, a tedy H_0 zamítáme,
- pro $p_v > \alpha$ leží realizovaná statistika mimo kritický obor W, a tedy H_0 nezamítáme.

Příklad

Dokončíme příklad o testování rychlostí automobilů Škoda.

Spočítali jsme hodnotu statistiky $T_t = 1.884$. Protože test je pravostranný, bude mít p-hodnota tvar

$$p_v = P\left(T > T_r|H_0\right) = P\left(T > 1.884\right) = 0.03$$

Výsledek buď uzavřeme se spočtenou p-hodnotou (p-hodnota je malá, proto H_0 zamítáme), nebo testujeme na dané hladině významnosti $\alpha = 0.05$ a porom řekneme: $p_v < \alpha$ proto H_0 zamítáme.

Poznámka

Srovnáme-li závěr založený na p-hodnotě se závěrem, který jsme učinili na základě realizované statistiky a kritického oboru vidíme, že výsledek je stejný. Pro p-hodnotu však máme určitou míru, ze kterou je H_0 zamítnuta. H_0 má pravděpodobnost 3%, zatímco my požadujeme minimálně 5% abychom ji mohli přijmout.

8.5 Další příklady

Příklad (test pro dva podíly)

Na dvou pracovištích A a B byly sledovány pracovní prostoje. Pracoviště A bylo sledováno v $n_A = 800$ časových okamžicích a bylo zaznamenáno $n_A^+ = 116$ prostojů, zatímco pracoviště B bylo sledováno $n_B = 1200$ krát a zjištěno $n_B^+ = 138$ prostojů. Na hladině významnosti $\alpha = 0.05$ testujte hypotézu o rovnosti středních podílů prostojů na obou pracovištích.

Řešení provedeme podle uvedeného schématu:

Známe:

Hypotézy: $H_0 : p_A = p_B, \quad H_A : p_A \neq p_B$

Směrování: oboustranný test.

Podíly:

$$p_A = \frac{116}{800} = 0.145, \quad p_B = \frac{138}{1200} = 0.115, \quad p_P = \frac{p_1(1-p_1)}{n_1 - 1} + \frac{p_2(1-p_2)}{n_2 - 1} = 2.4\,10^{-4}$$

Hladina významnosti: $\alpha = 0.05$

Testová statistika:

$$z = \frac{p_1 - p_2}{\sqrt{p_P}} \; N(0;1)$$

Hodnota statistiky: $z_r = 1.94$

Kritický obor $W = (-1.96; 1.96)$

P-hodnota: $p_v = 2 \times 0.0265 = 0.053.$

Závěr: Na hladině 0.05 hypotézu H_0 nezamítáme, podíly prostojů na pracovištích A a B nejsou stejné. P-hodnota navíc ukazuje, že zamítnutí je poměrně těsné.

Budeme sledovat parametrické testy pro střední hodnotu, rozptyl a podíl, jako pro IS.

Střední hodnota (známé σ^2)

[p_hodnota,T,z_alpha]=**z_test**(Mu0,stredni_hodnota,rozptyl,n,strana,alpha)

[p_hodnota,T,z_alpha]=**z_test_2**(alpha,n1,str_hod_1,n2,str_hod_2,rozptyl_1, rozptyl_2, strana)

Střední hodnota (neznámé σ^2)

[p_hodnota,T,t_alpha]=**t_test_2s**(stredni_hodnota_1,rozptyl_1,n1,stredni_hodnota_2,rozptyl_2, n2,strana,alpha),

[p_hodnota,T,t_alpha]=**t_test_2n**(stredni_hodnota_1,rozptyl_1,n1,stredni_hodnota_2,rozptyl_2, n2,strana,alpha),

[p_hodnota,T,t_alpha]=**t_test_2p**(vyber1,vyber2,strana,alpha)

Rozptyl

[p_hodnota,T]=**var_test**(sigma0,rozptyl,strana,n,alpha),

[p_hodnota,T,f_alpha]=**var_test_2**(Rozptyl_S1,n1,Rozptyl_S2,n2,strana,alpha)

Podíl

[p_hodnota,T,z_alpha]=**prop_test**(p0,p,n,strana,alpha),

[p_hodnota,T,z_alpha]=**prop_test_2**(p1,n1,p2,n2,strana,alpha)

9 Neparametrické testy

Dosud jsme se zabývali tzv. parametrickými testy. Jsou to testy hypotéz, které vyslovují tvrzení o parametru sledovaného souboru, tedy o parametru rozdělení náhodné veličiny, která tento soubor představuje. Přitom tyto testy byly úzce svázány s intervaly spolehlivosti, které příslušný parametr odhadovaly. Nyní přikročíme k další skupině testů, jejichž hypotézy se netýkají parametrů, ale vyslovují se o dalších vlastnostech souboru, jako například typu rozdělení souboru nebo nezávislosti dvou souborů.

9.1 Chi-kvadrát testy

První podskupinou těchto tzv. neparametrických testů jsou testy chi-kvadrát (Chi2). Jejich název je odvozen od typu rozdělení statistiky, která je pro ně pro všechny společná.

Chi2 testy pracují s četnostmi. To znamená, že naměřená data je nejprve třeba rozdělit do skupin, podle testované vlastnosti. Například, budeme-li testovat shodu počtu nehod v pracovních dnech, budeme počítat nehody v pondělí, úterý, $\cdots$, až pátek. Dostaneme tak četnosti nehod pro jednotlivé pracovní dny.

Pro test se používají pozorované četnosti O (observed) a očekávané četnosti E (expected).

Pozorované četnosti O získáme z datového výběru tím, že naměřená data roztřídíme do předem daných skupin.

Očekávaní četnosti E konstruujeme pro stejné skupiny jako pozorované četnosti, ale tak, aby tyto četnosti přesně odpovídaly požadavkům nulové hypotézy. Říká-li např. nulová hypotéza, že data jsou rovnoměrně rozdělená, konstruujeme očekávané četnosti tak, aby celkový počet dat v nich obsažený byl stejný jako v datovém vzorku a každá skupina obsahovala počet dat, úměrný její velikosti. Přitom tyto teoretické četnosti dat nemusí být celočíselné.

Příklad

Srovnáváme počet nehod za všední dny a víkendy. Naměřili jsme 238 nehod během týdne a 128 nehod během soboty a neděle. To jsou pozorované četnosti. Očekávané četnosti určíme takto: celkový počet nehod je $238 + 128 = 366$. Tento počet musíme rozdělit v poměru $5 : 2$ (pět všedních dní a dva dny víkendu). Teoretický (očekávaný) počet nehod pro všední dny bude $366\frac{5}{7} = 261.43$ a pro víkendy $366\frac{2}{7} = 104.57$.

Statistika

Pro naměřené četnosti $O_i,\ i = 1, 2, \ldots, n$ a teoretické četnosti $E_i,\ i = 1, 2, \ldots, n$ má statistika tvar

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \sim Chi2(n-1) \tag{39}$$

a její rozdělení je chí-kvadrát s $n - 1$ stupni volnosti.

Poznámka

Uvedená statistika měří vzdálenost mezi pozorovanými a teoretickými četnostmi (je nezáporná). Jsou-li pozorované i očekávané četnosti stejné, rovná se hodnota statistiky nule. Čím více se četnosti liší, tím je hodnota statistiky větší.

Jak jsme již řekli, očekávané četnosti vyjadřují požadavky nulové hypotézy. Budou-li tedy změřené (pozorované) četnosti rovny očekávaným (hodnota statistiky je nula) platí přesně nulová hypotéza. Čím více se budou pozorované četnosti lišit od očekávaných, tím bude hodnota statistiky větší a tím více neplatí nulová hypotéza. V okamžiku, kdy se hodnota statistiky dostane do kritického oboru prohlásíme nulovou hypotézu za neplatnou. Jak je patrné z toho, co jsme právě popsali, leží kritický obor vždy vpravo - test je vždy pravostranný

s **kritickým oborem**

$$W = (\chi^2_\alpha, \infty)$$

a **p-hodnotou**

$$p_v = P(\chi^2 > \chi^2_r).$$

Nejvýznamnější aplikace tohoto testu jsou pro testování typu rozdělení (test dobré shody) a testování nezávislosti dvou rozdělení (test nezávislosti). Oba testy vyložíme na příkladech.

Test dobré shody

Jednou z nejznámějších aplikací chí-kvadrát testu je test dobré shody rozdělení testovaného souboru s rozdělením předpokládaným. Toto předpokládané rozdělení opět „vtělíme" do očekávaných četností. Na předem daných, nebo zvolených, intervalech potřebujeme určit absolutní četnosti výskytu realizací náhodné veličiny s předpokládaným rozdělením. Pro tento úkol lze obecně využít vlastnosti distribuční funkce, jejíž hodnota v bodě x je rovna pravděpodobnosti intervalu $(-\infty, x)$. Pravděpodobnost intervalu (x_1, x_2) pak bude $P((x_1, x_2)) = P(-\infty, x_2) - P(-\infty, x_1) = F(x_2) - F(x_1)$, tedy bude se rovnat rozdílu hodnot distribuční funkce v koncovém a počátečním bodě intervalu. Tak získáme pravděpodobnosti intervalů. Absolutní četnosti dostaneme, násobíme-li tyto pravděpodobnosti celkovým počtem dat - rozsahem výběru

V některých jednoduchých případech lze ale očekávané četnosti vypočítat přímo, aniž bychom potřebovali distribuční funkci. Tento případ budeme ilustrovat ihned v následujícím příkladu.

Test rovnoměrnosti

Sledujeme nehodovost na pražských silnicích během různých dní týdne. Po určité době jsme shromáždili následující údaje

den	počet nehod
Po - Pá	1879
So	421
Ne	406

Na hladině významnosti 0.05 testujte tvrzení (nulové hypotézy), že nehody se během týdne vyskytují rovnoměrně.

Pozorované četnosti jsou zadané.

$$O = [1879,\ 421,\ 406]$$

Teoretické četnosti určíme takto: máme 3 intervaly s délkami [5, 1, 1]. Celkový počet pozorování je $1879 + 421 + 406 = 2706$. Tento počet pozorování máme rozdělit na dané intervaly rovnoměrně. První bude mít 5/7 druhý 1/7 a třetí také 1/7. Bude tedy

$$E = 2706 \times [5/7,\ 1/7,\ 1/7] = [1932.86,\ 386.57,\ 386.57].$$

Hodnota statistiky

$$\chi^2_r = \frac{(1879 - 1932.86)^2}{1932.86} + \frac{(421 - 386.57)^2}{386.57} + \frac{(406 - 386.57)^2}{386.57} = 5.54$$

Kritický obor: $W = (\chi^2_\alpha(3-1); \infty) = (5.99; \infty)$

P-hodnota: $p_v = P(\chi^2 > 5.54) = 0.063$

Závěr: Nulovou hypotézu nelze vyvrátit. Zůstává v platnosti tvrzení: "nehody se vyskytují rovnoměrně".

Test normality

Testujeme, zda rychlosti osobních automobilů jedoucích po nuselském mostě mají normální rozdělení s $\mu = 60$ a $\sigma = 7.6$. Výsledky naměřených rychlostí jsou v tabulce

Interval rychlosti [km/h]	(20-50)	(50-60)	(60-70)	(70-120)
Zjištěná četnost	35	268	315	21

Pozorované četnosti: $O = [35,\ 268,\ 315,\ 21]$

Teoretické četnosti určíme pomocí distribuční funkce normálního rozdělení. Abychom mohli použít distribuční funkci standardního normálního rozdělení, normujeme hranice zadaných intervalů $h_i = [20,\ 50,\ 60,\ 70,\ 120]$

$$x_i = \frac{h_i - \mu}{\sigma}, \quad \Rightarrow \quad x = [-5.263,\ -1.316,\ 0,\ 1.316,\ 7.895]$$

Odpovídající hodnoty distribuční funkce standardního normálního rozdělení jsou

$$F(x) = [0,\ 0.094,\ 0.5,\ 0.906,\ 1].$$

Pravděpodobnosti p_i intervalů dostaneme jako rozdíly $F(x_{i+1}) - F(x_i)$, $i = 2, 3, 4, 5$

$$p_i = [0.094,\ 0.406,\ 0.406,\ 0.094].$$

Z celkového počtu pozorování $35+268+315+21 = 639$ tedy jednotlivým intervalům přísluší teoretické četnosti

$$E = 639 \times p_i = 639 \times [0.094,\ 0.406,\ 0.406,\ 0.094] = [60.06,\ 259.43,\ 259.43,\ 60.06]$$

Hodnota statistiky: $\chi^2_r = 48.05$

Kritický obor: $W = (7.82; \infty)$

P-hodnota: $p_v = P(\chi^2 > 48.05) = 2.10^{-10}$

Závěr: Zjištěné četnosti nepochází z normálního rozdělení se střední hodnotou 60 a směrodatnou odchylkou 7.6.

Funkce ve Scilabu:

[p_hodnota,T,alpha]=**chisquare_test**(O,E,alpha)

Test nezávislosti

Tento test používá kontingenční tabulku[2] absolutních četností dvou náhodných veličin, jejichž nezávislost testujeme. Jednotlivá políčka tabulky tvoří pozorované četnosti O.

Očekávané četnosti E určíme s pomocí definice nezávislosti $f(x, y) = f(x)f(y)$. Tato definice říká, že náhodné veličiny X a Y jsou nezávislé, jestliže se jejich sdružená hustota pravděpodobnosti rovná součinu jednotlivých marginálních hustot. Sdruženou hustotu pravděpodobnosti zkonstruovanou z pořízeného výběru (výběrovou hp) dostaneme normováním tabulky pozorovaných četností tak, aby součet jejích prvku po normování byl roven jedné. Z této tabulky určíme marginální hustoty a z nich zkonstruujeme tabulku nezávislých náhodných veličin - kazdý její prvek je dán součinem odpovídajících prvků marginálních hustot. Tabulku absolutních četností dostaneme renormalizací, tedy násobením celé tabulky původním součtem prvků.

Uvedeme algoritmus konstrukce tabulky očekávaných četností:

- Vezmeme tabulku pozorovaných četností a normalizujeme ji - celou tabulku dělíme součtem všech prvků. Tak dostaneme sdruženou výběrovou hustotu pravděpodobnosti testovaných náhodných veličin.

[2] Kontingenční tabulka má řádky odpovídající skupinám první náhodné veličiny a sloupce skupinám druhé náhodné veličiny. Libovolné políčko tabulky obsahuje četnost výskytu dvojic $[x_1, x_2]$ kde x_1 je ze skupiny určená rádkem políčka a x_2 je ze skupiny určené sloupcem políčka.

- Určíme marginální hustoty pravděpodobnosti tak, že sečteme řádky nebo sloupce normované tabulky. Součet sloupců dá marginálu první náhodné veličiny (jejíž hodnoty se mění po řádcích), součet řádků definuje marginálu druhé náhodné veličiny (s hodnotami po sloupcích).
- Vypočteme tabulku nezávislých pravděpodobností. Prvek nezávislé tabulky s indexem (i, j) je součinem i-tého prvku sloupcové a j-tého prvku řádkové marginály.
- Tabulku re-normalizuje na absolutní četnosti. Celou tabulku násobíme původním součtem všech prvků tabulky pozorovaných četností.

Test je pravostranný a má $(n_x - 1)(n_y - 1)$ stupňů volnosti, kde n_x je počet skupin pro první náhodnou veličinu X a n_y je počet skupin pro druhou náhodnou veličinu Y.

Pomocí statistiky (39) se porovnává původní tabulka s tabulkou absolutních četností nezávislých veličin. Statistiku počítáme pro všechny prvky tabulek (srovnáme obě tabulky do vektorů). Nulová hypotéza H_0 je "jsou nezávislé".

Příklad

Testujeme tvrzení, že u řidičů osobních automobilů nesouvisí věk V [roky] a reakční doba R (měřená časem, za který řidič přehlédne křižovatku). Zjištěné údaje jsou sestaveny do následující tabulky

$R \setminus V$	1 = (18-30)	2 = (30-50)	3 = (50-70)
1 = menší než 2 sec	56	42	23
2 = větší než 2 sec	32	49	37

Nezávislost testujte na hladině významnosti $a = 0.05$.

Řešení:

Pozorované četnosti jsou: $O = [56,\ 32,\ 42,\ 49,\ 23,\ 37]$.

Teoretické četnosti dostaneme podle výše uvedeného postupu:

– součet prvků tabulky

$$56 + 32 + 42 + 49 + 23 + 37 = 239$$

– normalizovaná tabulka $f(r, v)$

$$\begin{matrix} 0.234 & 0.176 & 0.096 \\ 0.134 & 0.205 & 0.155 \end{matrix}$$

– marginální hustoty $f(v)$ a $f(r)$ jsou

$$\begin{array}{ccc} 0.368 & 0.381 & 0.251 \end{array} \qquad \begin{array}{c} 0.50628 \\ 0.49372 \end{array}$$

– tabulka nezávislých pravděpodobností $f_N(r, v)$

$$\begin{array}{ccc} 0.186 & 0.193 & 0.12710 \\ 0.182 & 0.188 & 0.124 \end{array}$$

– tabulka absolutních četností E

$$\begin{array}{ccc} 44.552 & 46.071 & 30.377 \\ 43.448 & 44.929 & 29.623 \end{array}$$

Očekávané četnosti: $E = [44.552,\ 43.448,\ 46.071,\ 44.929,\ 30.377,\ 29.623]$

Hodnota statistiky: $\chi^2_r = 10.315.$

Kritický obor: $W = (\chi^2_\alpha((3-1)(2-1)); \infty) = (5.99; \infty).$

P-hodnota: $p_v = 0.006.$

Závěr: Nulová hypotéza je vyvrácena, reakční doby řidičů jsou závislé na věku.

Funkce ve Scilabu:

[p_hodnota,T,chi_alpha]=**chisquare_test_i**(KT,alpha).

9.2 Znaménkový test

Tento test ověřuje tvrzení o hodnotě mediánu sledovaného souboru.

Uvažujme realizaci výběru $x = [x_1, x_2, \ldots, x_n]$ ze spojitého rozdělení s neznámým mediánem $x_{0.5}$. Testujeme nulovou hypotézu $H_0 :\ x_{0.5} = x_0$, kde x_0 je dané číslo.

Vypočteme rozdíly prvků výběru x_i a testované hodnoty mediánu x_0

$$D_i = x_i - x_0,\ i = 1, 2, \ldots, n$$

a písmenem b označíme počet kladných rozdílů D_i. b je statistika s binomickým rozdělením, kterou lze pro $n \to \infty$ aproximovat normálním rozdělením $N(\frac{n}{2}; \frac{n}{4})$.

Normovaná statistika testu má tvar

$$z = \frac{2b - n}{\sqrt{n}} \quad \sim N(0; 1)$$

a lze ji testovat pomocí oboustranného z**-testu**.

Příklad

(Jen demonstrace postupu! Není $n \to \infty$)

Testujeme, zda výběr

$$x = [5.3,\ 4.2,\ 6.8,\ 5.7,\ 5.1,\ 3.1]$$

pochází z rozdělení s mediánem $x_0 = 5$.

Počet dat výběru: $n = 6$.

Rozdíly $D_i = x_i - x_0$ jsou

$$D_i = [0.3,\ -0.8,\ 1.8,\ 0.7,\ 0.1,\ -1.9]$$

a z nich $b = 4$ jsou kladné.

Normovaná statistika je

$$z = \frac{2 \times 4 - 6}{\sqrt{6}} = 0.817$$

a **p-hodnota** pro oboustranný test je $p_v = 0.207$.

Závěr: Nulová hypotéza "data pochází z rozdělení s mediánem 5" se nepopírá.

[p_hodnota,T,z_alpha]=**sign_test**(X,Y,strana,alpha)

9.3 Test nezávislosti prvků výběru

Tento test, na rozdíl od jiných testů nezávislosti, při kterých se testuje nezávislost dvou souborů (tj. nezávislost dvou náhodných veličin), se zabývá jen jedním souborem. Uvažuje náhodný výběr $X = [X_1, X_2, \cdots, X_n]$ jako posloupnost náhodných veličin[3]. Úkolem tohoto testu je ověřit, zda jednotlivé prvky výběru jsou nezávislé náhodné veličiny. Při provádění testu samozřejmě výběry neopakujeme, ale opíráme se o jedinou realizaci výběru. Postup testu je následující.

Uvažujeme realizaci výběru $x = [x_1, x_2, \cdots, x_n]$ o rozsahu n a vypočteme jeho výběrový medián $\hat{x}_{0.5}$. Test vychází z rozdílů mezi prvky výběru x_i, $i = 1, 2, \ldots, n$ a výběrového mediánu $\hat{x}_{0.5}$ (srovnej znaménkový test)

$$D_i = x_i - \hat{x}_{0.5} \quad i = 1, 2, \ldots, n.$$

Na těchto rozdílech se definuje **série**, tj. souvislé posloupnosti prvků se stejným znaménkem, odděle změnami znaménka. Jako statistiku b definujeme počet sérií v posloupnosti rozdílů D. Tato statistika má přibližně normální rozdělení $N(\frac{n}{2} + 1;\ \frac{\sqrt{n-1}}{2})$.

[3] Připomeňme, že náhodnost jednotlivých prvků náhodného výběru spočívá v potenciální opakovatelnosti výběru. Každá nová realizace výběru bude obsahovat jiná čísla - vybíráme náhodně. A tak, každá složka výběru může nabývat různé realizace - chová se jako náhodná veličina.

Normovaná statistika je

$$z = \frac{2b - (n-2)}{\sqrt{n-1}} \quad \sim N(0;1).$$

Nulovou hypotézu H_0 : "prvky výběru jsou nezávislé" lze testovat pomocí levostranného z-testu.

Příklad

Testujeme nezávislost prvků výběru s realizací

$$x = \{2.4,\ 2.2,\ 1.6,\ 1.8,\ 1.5,\ 1.8,\ 2.2,\ 2.3,\ 2.3,\ 2.5\}$$

o rozsahu $n = 10$. Medián výběru je $\hat{x}_{0.5} = 2.2$ a diference

$$D_i = x_i - \hat{x}_{0.5} = [0.2,\ 0.0,\ -0.6,\ -0.4,\ -0.7,\ -0.4,\ 0.0,\ 0.1,\ 0.1,\ 0.3]\,.$$

V těchto diferencích je možno nalézt $b = 3$ série se stejnými znaménky. Statistika tedy je

$$z = \frac{2 \times 3 - (10-2)}{\sqrt{10-1}} = -0.667$$

a p-hodnota $p_v = 0.25$.

Závěr: Prvky výběru lze považovat za nezávislé.

Poznámka

Tento test je velmi významný nejen pro testování, zda náš výběr je skutečně nezávislý (jak požaduje definice výběru), ale také například pro test reziduí po výpočtu regresní analýzy, pro ověření kvality regrese, kterou se budeme zabývat v následující kapitole.

[p_hodnota,T,z_alpha]=**ordinal_test**(X,alpha)

9.4 Test nezávislosti výběrů

Pearsonův test

Pro dvě náhodné realizace výběrů x a † oba o rozsahu n vypočteme výběrový korelační koeficient $r = s_{xy}/\sqrt{s_x^2 s_y^2}$, kde s_{xy} je výběrová kovariance a s_x^2, s_y^2 jsou příslušné výběrové rozptyly.

Statistika je dána vztahem

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \sim St(n-2)$$

a má Studentovo rozdělení s $n-2$ stupni volnosti.

Pro test nulové hypotézy H_0 : "výběry jsou nezávislé" lze použít oboustranný t-test.

Poznámka

Tento test budeme používat i pro ověření výsledku regresní analýzy.

[p_hodnota,T,z_alpha]=**pearson_test**(X,Y,alpha)

Spearmanův test

Uvažujme dva náhodné výběry X a $\mathcal{Y}$, oba o rozsahu n. Pro oba výběry definujeme **pořadí** P, resp., Q, tj. např. pro $x = [6.2,\ 2.8,\ 4.1]$ je $p = [3,\ 1,\ 2]$, protože 6.2 je na třetím místě uspořádaného výběru x, 2.8 na prvním a 4.1 na druhém. Tato pořadí dosadíme do vzorce pro výběrový korelační koeficient a dostaneme statistiku

$$r_S = \frac{s_{pq}}{\sqrt{s_p^2 s_q^2}} = 1 - \frac{6}{n(n^2-1)} S,$$

kde s_{pq}, s_p^2, s_q^2 jsou druhé výběrové momenty realizací náhodných výběrů P a Q; $S = \sum_{i=1}^{n} (p_i - q_i)^2$.

Tato statistika se testuje podle speciálních tabelovaných hodnot Spearmanova testu.

Nulová hypotéza H_0 : "výběry jsou nezávislé".

[p_hodnota,T,z_alpha]=**spearman_test**(X,Y,alpha)

Kendalův test

Uvažujme dva náhodné výběry X a $\mathcal{Y}$ o rozsahu n a jejich pořadí[4] P a Q. Z pořadí sestavíme dvouřádkovou matici a její sloupce uspořádáme tak, aby v prvním řádku bylo $1, 2, \ldots, n$. Druhý řádek uspořádané matice označíme R a jeho prvky $r_1,\ r_2,\ \ldots,\ r_n$. Písmenem k_i označíme počet všech prvků $r_{i+1},\ r_{i+2},\ \ldots, r_n$, které jsou větší než r_i. Dále označíme $K = \sum k_i$. Statistika pak je

$$r_K = \frac{4K}{n(n-1) - 1}$$

a testuje se opět podle speciálních hodnot Kendalova testu.

[4] Je-li např. $x = [5,\ 7,\ 3]$ bude $p = [2,\ 3,\ 1]$ protože v uspořádaném vektoru je 5 druhá, 7 třetí a 3 první.

9.5 Test typu rozdělení

Kolmogorov-Smirnovův test

Tento test slouží k ověření, zda zkoumané rozdělení má dané rozdělení. Je založen na porovnání distribuční funkce $F(x)$ daného rozdělení X a výběrové distribuční funkce $F_n(x)$[5], určené z výběru o rozsahu n.

Statistika testu je definována vztahem

$$k_s = \sup_{x_i \in X} |F_n(x_i) - F(x_i)|$$

a má různá rozdělení, podle typu testované distribuční funkce. Pro důležitá testovaná rozdělení jsou hodnoty rozdělení k_s tabelovány.

Nulová hypotéza H_0 : "rozdělení má předpokládaný typ".

bude dodán co nevidět :-)

Příklad

Naměřili jsme rychlosti automobilů na dálnici D1 při výjezdu z Prahy. Chceme vědět, jestli je oprávněný předpoklad o normalitě těchto rychlostí.

```
[pval,ks]=KS_test(x,"normal",0,1),
```

kde `alt` je « > ” jako předvolba.

Poznámka

Uvedený test je velmi důležitý, neboť řada jiných statistických procedur vyžaduje určitý typ rozdělení (většinou normální). Není radno dělat závěry ze statistických procedur, pokud nemáme ověřenu platnost předpokladů!

10 Regresní analýza

Regresní analýza poskytuje nástroj k hledání stochastické závislosti mezi dvojicí náhodných veličin X – **nezávisle proměnná** a Y – **závisle proměnná**. V nejběžnější (lineární) podobě zkoumá, zda mezi oběma veličinami existuje **lineární vztah**. Velice jednoduše lze také zkoumat např. polynomický nebo exponenciální vztah. Dále se budeme věnovat především lineární regresi, o ostatních se stručně zmíníme později.

[5] Výběrovou distribuční funkci F_n dostaneme tak, že na reálné ose uspořádáme prvky realizace výběru podle velikosti. Počet prvků označíme n. Začneme kreslit pro $-\infty$ v nule a kde narazíme na hodnotu výběru, uděláme skok o velikosti $1/n$. Skončíme v $+\infty$ v hodnotě jedna. Je to tedy schodovitá funkce s přírůstky $1/n$ v místech změřených dat.

10.1 Lineární regrese

Příklad

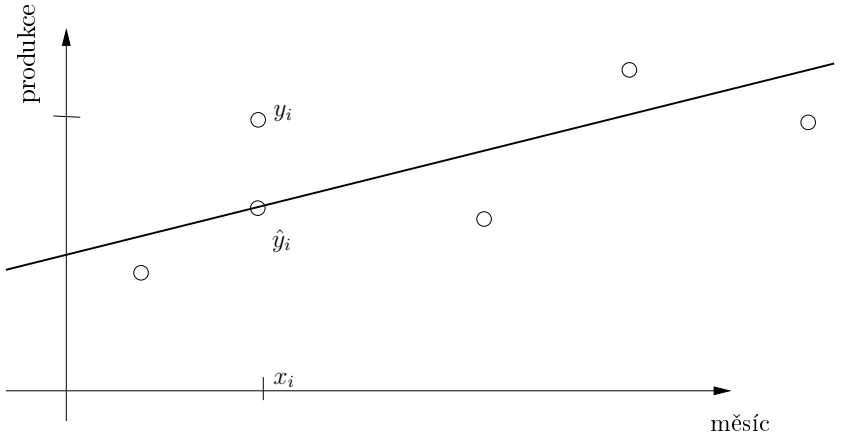
Sledujeme produkci automobilového závodu během půl roku. Produkce v jednotlivých měsících byla

měsíc	1	2	3	4	5	6
produkce (ks×100)	4.3	3.9	4.2	4.5	4.4	5.1

Odhadněte lineární trend těchto dat a určete, zda mají tendenci k růstu nebo poklesu.

Zadaná data jsou vykreslena na obrázku a je jimi "od oka" proložena přímka. Ihned nás napadnou otázky

- jsou tato data vhodná k aproximaci přímkou?
- je nakreslená přímka tou nejlepší, která data aproximuje?



Abychom na otázky z příkladu dokázali odpovědět, budeme datové dvojice $[x_i,\ y_i]$, $i = 1, 2, \ldots, n$ reprezentovat geometricky, jako body v rovině. Přímku, kterou chceme body proložit, budeme uvažovat ve směrnicovém tvaru $y = b_1 x + b_0$, kde $[x,\ y]$ je libovolný bod přímky, b_1 je směrnice a b_0 absolutní člen (úsek na ose y). Optimální přímku nazveme **regresní přímka** a budeme požadovat, aby poloha této regresní přímky vůči datovým bodům minimalizovala určité kriterium vzdálenosti. Abychom mohli být konkrétní, zavedeme následující pojmy a uvedeme vzorce pro výpočet regresních koeficientů:

Predikce $\hat{y}_i$ je bod, jehož x-ová souřadnice je x_ia y-ová $b_1 x_i + b_0$, tj. leží na proložené přímce.

Chyba predikce e_i je rozdíl mezi datovým bodem a predikcí, tj. svislá vzdálenost bodu od přímky.

Model datových bodů lze pomocí proložené přímky vyjádřit takto: datový bod je predikce plus chyba predikce, tj.

$$y_i = b_1 x_i + b_0 + e_i \tag{40}$$

Kritérium optimality pro "ideální regresní přímku" definujeme jako součet kvadrátů všech chyb predikce

$$J = \sum_{i=1}^{n} e_i^2, \tag{41}$$

a požadujeme, aby bylo minimální.

Chyby predikce odpovídající regresní přímce, tj. přímce, pro kterou je kriterium J minimální, nazýváme **rezidua.**

Koeficienty regresní přímky b_1 a b_0 jsou[6]

$$b_1 = \frac{S_{xy}}{S_{xx}}, \qquad b_0 = \overline{y} - b_1 \overline{x}, \tag{42}$$

s označením

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x, \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2, \quad S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}).$$

Korelační koeficient r je

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}, \tag{43}$$

kde $S_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2$.

Poznámka

Uvedený vzorec není v rozporu s tím, který jsme uvedli v souvislosti s Pearsonovým testem. Po vydělení čitatele i jmenovatele výrazem $n-1$ dostaneme totéž.

Význam:

- Koeficient b_1 vypovídá o trendu regresní přímky. Je-li $b_1 > 0$ přímka roste, pro $b_1 < 0$ klesá a v případě $b_1 = 0$ je vodorovná.
- Korelační koeficient r nese informaci o tom, jak silná je vazba mezi daty x a y. Jeho rozsah je $r \in (0,\ 1)$ a je li nulový, veličiny x a y spolu nesouvisí – regrese nemá význam. Je-li kladný přímka roste, je-li záporný, klesá. Čím více se blíží ± 1 tím je vazba silnější. Pro $r = \pm 1$ leží datové body na regresní přímce.

[6] Odvození vzorců je uvedeno v Dodatku **??**.

Příklad

(**pokračování**) Podle zadání našeho příkladu o produkci automobilů bude

$$\overline{x} = 3.5, \quad \overline{y} = 4.4, \quad S_{xx} = 17.5, \quad S_{yy} = 0.8, \quad S_{xy} = 2.9, \quad b_1 = 0.166, \quad b_0 = 3.82, \quad r = 0.775$$

Predikce v bodech měření (nebo i kdekoliv jinde) určíme tak, že do odhadnuté regresní přímky dosadíme příslušná x

$$\hat{y}_i = b_1 x_i + b_0 = 0.166 x_i + 3.82, \quad i = 1, 2, \ldots, n$$

$$\text{tj.} \qquad 3.99,\ 4.15,\ 4.32,\ 4.48,\ 4.65,\ 4.82$$

Hodnota regresního koeficientu $r = 0.775$, stejně jako směrnice regresní přímky $b_1 = 0.166$ potvrzují, že data mají tendenci nárůstu. Regresní koeficient navíc ukazuje, že data lze dosti dobře aproximovat přímkou ($0.775 \to 1$).

Funkce ve Scilabu:

- jednoduchá regrese a predikce

```
p = lin_reg(x,y),

yp=lin_pred(x,p),
```

- vícenásobná regrese a predikce

```
p=lin_reg_n(x,y),

yp=lin_pred_n(x,p)
```

10.2 Nelineární regrese

Přímka je nejznámější, nikoliv však jedinou funkcí, kterou lze měřenými body prokládat. Jako další z možností uvedeme polynomickou a exponenciální regresi. Důležité při tom je, abychom dokázali vyjádřit závisle proměnnou y (nebo její funkci) jako skalární součin vektoru parametrů a regresního vektoru, tj. vektoru, jehož složky jsou funkce měřených dat, pomocí kterých chceme y vyjadřovat. Tak dostaneme obecnou regresní rovnici

$$f_0(y_i) = b_m f_m(x_i) + b_{m-1} f_{m-1}(x_i) + \cdots + b_1 f_1(x_i) + b_0 + e_i = \tag{44}$$

$$= [f_m(x_i),\ f_{m-1}(x_i), \ldots,\ f_1(x_i),\ 1] \times [b_m, b_{m-1}, \ldots, b_1, b_0]' + e_i,$$

kde

$f_k(x_i),\ k = 1, 2, \ldots, m,\ i = 1, 2, \ldots, n$ jsou známé funkce nezávisle proměnných x_i,

$f_0(y_i),\ i = 1, 2, \ldots, n$ je známá funkce závisle proměnných y_i a

symbol_′ značí transpozici.

Regresní koeficienty obdržíme takto:

- Sestavíme **vektor výstupů** a **datovou matici**

$$Y = \begin{bmatrix} f_0(y_1) \\ f_0(y_2) \\ \ldots \\ f_0(y_n) \end{bmatrix}, \quad X = \begin{bmatrix} f_m(x_1), & f_{m-1}(x_1), & \ldots, & f_1(x_1), & 1 \\ f_m(x_2), & f_{m-1}(x_2), & \ldots, & f_1(x_2), & 1 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ f_m(x_n), & f_{m-1}(x_n), & \ldots, & f_1(x_n), & 1 \end{bmatrix}.$$

- Regresní koeficienty jsou
-

$$b = [b_m,\ b_{m-1}, \ldots, b_0]' = (X'X)^{-1} X'Y. \tag{45}$$

Poznámka

Všimněte si, že pro skalární případ tento vzorec souhlasí s dříve uvedeným.

Příklad

(polynomiální regrese) Regresní přímka má tvar

$$y = b_m x^m + b_{m-1} x^{m-1} + \cdots + b_1 x + b_0, \tag{46}$$

který je vzhledem k parametrům lineární.

Datová matice bude

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{bmatrix} X = \begin{bmatrix} x_1^m, & x_1^{m-1}, & \ldots, & x_1, & 1 \\ x_2^m, & x_2^{m-1}, & \ldots, & x_2, & 1 \\ \ldots & \ldots & \ldots & \ldots & \\ x_n^m, & x_n^{m-1}, & \ldots, & x_n, & 1 \end{bmatrix}.$$

Další postup podle obecného návodu výše.

p=**pol_reg**(x,y,k)

yp=**pol_pred**(x,p)

10.3 Exponenciální regrese

Regresní křivka má tvar

$$y = a \exp\{b_1 x\}, \tag{47}$$

který je nelineární. Linearizujeme je logaritmováním

$$\ln y = \ln a + b_1 x \quad \Rightarrow \tilde{y} = b_1 x + b_0, \tag{48}$$

kde $\tilde{y} = \ln y$ a $b_0 = \ln a$.

Dále již postupujeme podle obecného návodu.

Poznámka

Pozor! Z linearizované regresní přímky dostaneme logaritmy predikcí $\ln \hat{y}$. *Skutečné predikce jsou* $\hat{y} = \exp\{\ln \hat{y}\}$.

p=**exp_reg**(x,y),

yp=**exp_pred**(x,p).

10.4 Logistická regrese

Obyčejná regresní analýza, jak jsme ji dosud prezentovali, neumí jako závisle proměnnou y uvažovat diskrétní veličinu, tj veličinu, která může nabývat jen vybraných hodnot (většinou celočíselných). Je to pochopitelné. Když budeme násobit data z regresního vektoru odhadnutými parametry, těžko zajistíme, aby součet těchto součinů patřil do množiny předem vybraných (zpravidla celých) čísel. Přitom případy, kdy potřebujeme modelovat diskrétní veličinu, jsou dosti časté. Tak například nehoda (je, není), (je smrtelná, je se zraněním, jen hmotná škoda, není) nebo kolona (není, je malá, je velká) nemluvě ani o stupni dopravy (1,2,3,4,5), který je přímo jako diskrétní definován.

Jak je patrné z předchozího výčtu, některé veličiny jsou diskrétní již svou povahou, jiné lze chápat jako spojité a vhodně je diskretizovat. Každopádně je možno říci, že statistické postupy pro diskrétní modely jsou jednodušší a lze s nimi řešit i takové úlohy, které by ve spojité formě byly neřešitelné. Nicméně ani diskrétní modely nejsou bez vady. Zvláště při počáteční diskretizaci se snadno dostáváme do prostorů s tak vysokou dimenzí, že výpočetní postupy také zhavarují. Proto je počet hodnot při diskretizaci třeba volit velmi obezřetně a množstvím hodnot při diskretizaci neplýtvat.

Úloha, která modeluje diskrétní veličiny s hodnotami $y = 0, 1$, v závislosti na jiných, ať už diskrétních nebo spojitých, se nazývá **logistická regrese.** Své jméno dostala po logistické křivce

$$y = \log \frac{p}{(1-p)} = \text{logit}\,(p)$$

pro $p \in (0,1)$. Tato funkce provádí transformaci intervalu $(0,1)$ na celou reálnou osu R[7]. Celá logistická regrese se velmi podobá obyčejné regresi až na to, že místo y modelujeme funkci $\text{logit}(p)$, kde $p = P(y = 1|x, b)$, což je podmíněná pravděpodobnost, že $y = 1$ při znalosti regresního vektoru x a regresních parametrů b.

Rovnice logistické regrese má tvar

$$\text{logit}(p) = x'b$$

kde $x'b = b_0 + b_1x_1 + \cdots + b_nx_n$ a x i b jsou sloupcové vektory (vektor b začíná nulovým indexem, vektor x jedničkou).

Jestliže dosadíme za funkci $\text{logit}(\cdot)$, dostaneme uvedenou rovnici v jiném tvaru

$$p = \frac{\exp(x'b)}{1 + \exp(x'b)} \tag{49}$$

Odhadování

Úkolem odhadu je určit hodnoty vektoru parametrů b na základě změřeného datového vzorku y_i, x_i pro $i = 1, 2, \cdots, N$ kde $x_i = [1, x_{1;i}, x_{2;i}; \cdots, x_{n;i}]$. Protože se jedná o nelineární model, nelze použít metodu nejmenších čtverců jako u obyčejné lineární regrese. Pro identifikaci tedy použijeme metodu maximální věrohodnosti. Sestavíme věrohodnostní funkci (součin modelů s dosazeným výběrem) a budeme ji numericky maximalizovat.

Model pro y je podmíněná hustota pravděpodobnosti y s alternativním rozdělením (dva možné výsledky) a s parametrem $p = p(x, b)$ závislým na parametrech b a změřeném regresním vektoru x. Pro změřené y_i a x_i dostáváme

$$f(y_i|x_i, b) = p(x_i, b)^{y_i} [1 - p(x_i, b)]^{1-y_i}$$

a pro logaritmus věrohodnostní funkce $l_N(b) = \log L_N(b)$

$$l_N(b) = \log L_N(b) = \log \prod_{i=1}^{N} f(y_i|x_i, b) = \log \prod_{i=1}^{N} p(x_i, b)^{y_i} [1 - p(x_i, b)]^{1-y_i} =$$

$$= \sum_{i=1}^{N} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] = \cdots$$

$$\cdots = \sum_{i=1}^{N} [y_i x_i b - \log(1 + \exp\{x_i b\})] \tag{50}$$

[7] Skutečně, pro $p = 0$ je $y = -\infty$, pro $p = 0.5$ je $y = 0$ a pro $p = 1$ dostaneme $y = \infty$.

První derivace logaritmické věrohodnostní funkce je

$$\frac{\partial l_N(b)}{\partial b_j} = \sum_{i=1}^{N} \left(y_i x_{j;i} - \frac{1}{1+\exp\{x_i b\}} \exp\{x_i b\} x_{j;i} \right) = \sum_{i=1}^{N} (y_i - p_i) x_{j;i} \tag{51}$$

kde $p_i = p(x_i, b)$ a dále jsme využili vztah (49).

Druhá derivace je

$$\frac{\partial^2 l_N(b)}{\partial b_k \partial b_j} = \frac{\partial l_N(b)}{\partial b_k} \frac{\partial l_N(b)}{\partial b_j} = p_i (1 - p_i) x_{k;i} x_{j;i}$$

Pro maximalizaci lze použít **Newtonovu metodu**. Tu dostaneme pomocí rozvoje maximalizované funkce Taylorovým rozvojem druhého řádu.

$$l(b+\delta) = f(b) + f'(b)\delta + \frac{1}{2} f''(b)\delta^2$$

Derivaci podle δ položíme rovnu nule a dostaneme

$$f'(b) + f''(b)\delta = 0$$

Odtud plyne vzorec pro numerické hledání extrému. Položíme obecně $b_{t+1} = b + \delta$ a $b = b_t$ a dostaneme

$$b_{t+1} = b_t - \frac{f'(b_t)}{f''(b_t)} \quad \text{nebo vektorově} \quad b_{t+1} = b_t - H(b_t)^{-1} G(b_t)$$

kde Hessova matice H je matrice druhých parciálních derivací a gradient G je vektor prvních parciálních derivací

$$H_{j,k} = \frac{\partial^2 l_N(b)}{\partial b_k \partial b_j}, \qquad G_j = \frac{\partial l_N(b)}{\partial b_j}$$

Program realizující zmíněnou optimalizaci je následující

```
% Program pro logistickou regresi zavislosti poctu nehod
%   na vybranych charakteristikach dopravnich komunikaci
% ====================================================
%
matini
load nehody                     % natazení dat

[n m]=size(x);
x1=[ones(n,1) x];

th=rand(m+1,1);                 % startovaci hodnoty parametru
for i=1:1000
  [g h]=gradHessL(y,x1,th);     % generovani gradientu a Hessovy matice
  gg=sqrt(g'*g);
  hh=h+eye(m+1)*1e-5;           % regularizace pro inverzi
  th=th-inv(hh)*g/(1+gg);       % promenná delka kroku
  if gg<1e-3, break, end        % test na ukoncení iteraci
end

yp=logit_inv(x1*th);            % predikce v pravdepodobnostech
yi=round(yp);                   % bodova predikce
SE=sum(abs(y-yp))/n;            % kriterium pro vyhodnoceni kvality
fig,plot(1:n,y,'o',1:n,yp,'x',1:n,yi,'*')
i,th,SE
```

kde procedura, která počítá gradient a Hessovu matici z log-likelihoodu pro logistickou regresi je

```
function [g h]=grHess(y,x,th)
% [g h]=grHess(y,x,th)   gradient and Hessova matice pro
%                        logaritmus logisticke verohodnostni
%                        funkce
%
% g   gradient
% h   Hessova matice
% y   zavisle promenna
% x   nezavisle promenna
% th  parametry

[n m]=size(x);

% gradient
g=zeros(m,1);
for t=1:n
  xt=x(t,:);
  pe=exp(xt*th);
  p=pe/(1+pe);
  g=g+(y(t)-p)*xt';
end

% Hessova matice
h=zeros(m,m);
for t=1:n
  xt=x(t,:);
  pe=exp(xt*th);
  p=pe/(1+pe);
  for i=1:m
    for j=1:m
      h(i,j)=h(i,j)-x(t,i)*x(t,j)*p*(1-p);
    end
  end
end
```

Program pracuje s daty pro analýzu nehodovosti.

Modelovaná veličina y je

nehoda (hmotná škoda, zranění nebo smrt),

v regresním vektoru x jsou veličiny:

doba (den, noc, soumrak)

viditelnost (jasno, mlha, déšť, sníh)

rychlost (normální, překročená)

typ srážky (stejný směr, protisměr, vyjetí z vozovky)

výsledek (ohrožení, střet, pevná překážka, zvěř)

11 Korelační analýza

V předchozí kapitole jsme ukázali, jak k daným realizacím výběru x a y sestrojit regresní přímku. Zároveň jsme konstatovali, že výběrový korelační koeficient r vypovídá o kvalitě regrese. Přesnější vyjádření, založené na intervalech spolehlivosti a testech hypotéz, poskytuje korelační analýza, které se budeme nyní věnovat.

Ideální regresní přímka

Pro další úvahy budeme předpokládat, že skutečně existuje **ideální regresní přímka** s rovnicí $y = \beta_0 + \beta_1 x$, která generuje body $[x_i, y_i]$. My je však nejsme schopni měřit čisté a měříme je se šumem. O tomto šumu předpokládáme, že má nulovou střední hodnotu a konstantní rozptyl σ^2. Naší regresní přímku $y = b_1 x + b_0$ chápeme jako odhad této ideální přímky.

Koeficienty b_1 a b_0 jsou potom bodovými odhady parametrů β_1 a β_0 a predikce $\hat{y}_i = b_0 + b_1 x_i$ jsou bodovými odhady hodnot ideální regresní přímky $\beta_0 + \beta_1 x_i$ v bodech x_i. V přeneseném slova smyslu říkáme, že celá regresní přímka $y = b_0 + b_1 x$ je bodovým odhadem ideální přímky $y = \beta_0 + \beta_1 x$ pro $x \in R$.

11.1 Intervaly spolehlivosti v regresi

Podobně jako v kapitole o intervalech spolehlivosti uvažujeme soubor a jeho parametry. Těmi jsou β_0, β_1 (koeficienty skutečné přímky), $\beta_0 + \beta_1 x$ (hodnota skutečné přímky v libovolném bodě x), ρ (souborový korelační koeficient), případně další. My však tyto parametry neznáme a odhadujeme na základě výběru. Intervalový odhad pak určuje množinu hodnot které parametr může nabýt s danou pravděpodobností $1 - \alpha$.

Interval pro β_1

Jak jsme se zmínili dříve, koeficient regresní přímky u nezávisle proměnné x je směrnicí přímky a jeho znaménko určuje, zda je přímka klesající nebo rostoucí. Tento fakt může být v některých aplikacích rozhodující a je středem zájmu regresní analýzy. Představme si

například, že hodnoty y_i jsou zisky naší firmy zaznamenané v jednotlivých měsících. Zisky se mohou měsíc od měsíce velmi lišit, ale co nás především zajímá je to, zda mají celkovou tendenci růst nebo klesat.

Přitom je třeba si uvědomit, že to, co nás zajímá, je koeficient β_1, což je parametr skutečného procesu (souboru), který sledujeme a ze kterého pořizujeme výběr v podobě bodů - dvojic $[x_i, y_i]$. Statistikou pro odhad tohoto parametru je koeficient b_1 - směrnice regresní přímky, kterou jsme spočetli z pořízeného výběru.

Zkonstruovaný interval potom ukazuje „velmi pravděpodobné" hodnoty parametru β_1. Pokud např. celý interval leží v kladné poloose, je „velmi pravděpodobné" že skutečná přímka roste. Pokud obsahuje jak kladný, tak i záporný pod-interval, máme z poměru délek pod-intervalů alespoň představu o pravděpodobnosti růstu nebo klesání skutečné přímky.

Interval spolehlivosti pro β_1 je oboustranný interval se středem v b_1, ve které bude ležet β_1 s pravděpodobností $1 - \alpha$.

Interval spolehlivosti je dán vzorcem:

$$\beta_1 \in b_1 \pm \frac{s_e}{\sqrt{S_{xx}}} t_{\alpha/2},$$

kde

$$s_e = \sqrt{\frac{S_{yy} - S_{xy}^2/S_{xx}}{n-2}}, \quad \text{a} \quad t_{\alpha/2} \text{ je krit. hod. } St(n-1) \tag{52}$$

Využití: obsahuje-li interval nulu, data nejsou příliš vhodná pro lineární regresi.

Interval pro hodnotu regresní přímky

Regresní přímku konstruujeme proto, abychom odhadli „skutečná y" pro různá x. Pokud bychom se ale zajímaly o hodnoty y, které budeme opakovaně dostávat v jednom bodě $x = x_p$, zjistíme samozřejmě, že se budou navzájem také lišit (měříme se šumem a ten se neustále mění). Neurčitost hodnot y bude dána jednak neznalostí přesné polohy ideální přímky (parametry β_0 a β_1) a dále neznalostí šumu (rozptyl σ^2). Bude-li nás zajímat, v jakém rozsahu můžeme očekávat hodnoty y v daném bodě x_p můžeme zkonstruovat interval spolehlivosti pro hodnoty y v bodě x_p.

Tento interval pro hodnoty y v pevném bodě x_p je oboustranným intervalem ve kterém bude ležet $(1 - \alpha) \times 100\%$ všech hodnot y generovaných v bodě x_p.

Interval spolehlivosti je:

$$y(x_p) \in \hat{y}_p \pm s_e \sqrt{1 + \frac{1}{n} + \frac{(x_p - \overline{x})^2}{S_{xx}}}$$

kde směrodatná odchylka s_e je dána v (52).

Využití: Čím širší je tento interval, tím méně vhodná je lineární regrese.

Poznámka

Často se uvádí celý pás spolehlivosti, tvořený intervaly na husté síti bodů x_p.

[is_e,is_p,pval_a,pval_r]=reg_infe(x,y,xp,alpha,alt),

`is_e,is_p` jsou IS pro střední hodnotu a predikci,

`pval_a,pval_r` jsou p-hodnoty pro směrnici a korelační koeficient,

`x,y` výběry, `xp` pevný bod x, `alpha` hladina významnosti pro test, `alt` směrování testu.

11.2 Testy hypotéz v regresi

Intervaly spolehlivosti, o kterých jsme mluvili v minulém odstavci, vypovídaly o jistých vlastnostech sledovaného souboru, který je v případě regresní analýzy reprezentován dvěma náhodnými veličinami X a Y, mezi nimiž hledáme závislost. Parametry, které jsme odhadovali souvisely právě s vlastnostmi této závislosti.

Vlastnosti sledovaného souboru jsou vyjádřeny lineární závislostí s aditivním šumem ve tvaru

$$y = \beta_0 + \beta_1 x + e$$

kde pro šum e předpokládáme nulovou střední hodnotu a konstantní rozptyl σ^2. Otázkou ale zůstává do jaké míry, nebo zda vůbec, změřená realizace výběru $[x_i,\, y_i]\,,\ i = 1, 2, \cdots, n$ může pocházet z předpokládaného souboru, který generuje lineárně závislá data. Situaci budeme ilustrovat na třech následujících případech:

Ideální regrese

V tomto případě je šum nulový a změřené body leží přesně na regresní přímce.

Normální regrese

Regresní přímka body aproximuje ve smyslu nejmenších čtverců reziduí. Pokud bychom ale polohu přímky jakkoli změnili, dostaneme aproximaci, která je horší. Čím více jsou body kolem přímky „rozházeny", tím je kvalita aproximace přímkou pochybnější.

Nesmyslná regrese

V tomto extrémním případě (kdy změřené body tvoří vrcholy čtverce) je aproximace přímkou nesmyslná. Všechny přímky, procházející středem čtverce, jsou stejně dobre (nebo lépe, stejně špatné).

Dále uvedeme dva testy, které ověřují, zda je regresní analýza v daném případě, tj. pro sebraný datový výběr $[x_i,\, y_i]\,,\ i = 1, 2, \cdots, n$, vhodná.

t-test korelačního koeficientu (Pearsonův test)

První z uvedených testů, Pearsonův test, který jsme mimochodem již uváděli mezi neparametrickými testy, je založen na testu nulovosti souborového korelačního koeficientu ρ. V případě, že je korelační koeficient ρ nulový, jsou náhodné veličiny X a Y nekorelované, není mezi nimi žádná závislost, a tedy ani lineární. Opět, skutečný korelační koeficient ρ neznáme a dozvídáme se o nem jen z výběru a to prostřednictvím výběrového korelačního koeficientu r. Ten je také statistikou pro odhad ρ.

Normovaná statistika:

$$t = \frac{r}{\sqrt{\frac{1-r}{n-2}}} \sim St(n-2)$$

Test je oboustranný, s nulovou hypotézou H_0 : "data nejsou vhodná pro regresi".

[p_hodnota,T,z_alpha]=**pearson_test**(X,Y,alpha)

Poznámka

Většinou testujeme data na regresní analýzu proto, abychom ji mohli provést. Přitom nulová hypotéza říká, že data nejsou vhodná pro regresi. Dostáváme tak trochu zamotaný výsledek: nulovou hypotézu zamítáme, data jsou vhodná pro regresi. Jako by něco negativního, přinášelo pozitivní výsledek.

F-test poměru vysvětleného a nevysvětleného rozptylu

Tento test je velice obecný a má další použití i mimo regresní analýzu - například se užívá i v analýze rozptylu, kterou se budeme zabývat dále. Navíc je založen pouze na změřených hodnotách y_i a na predikcích $\hat{y}_i$, tj. hodnotách změřených na regresní přímce. Pracuje s rozkladem celkového součtu čtverců na regresní součet čtverců (který je vysvětlený) a reziduální součet čtverců (který je nevysvětlený). Jednotlivé pojmy dále vysvětlíme s pomocí následujícího obrázku.

Na obrázku je naznačena situace regresní analýzy se třemi body. Plnou čarou je vyznačena regresní přímka a čárkovanou čarou průměrná hodnota ze všech tří naměřených y. Vpravo nahoře je podrobněji rozkreslena situace pro jeden /třetí) naměřený bod. Dole je hodnota průměru, nad ním změřená hodnota y_3 a ještě dále nahoru na přímce leží predikce $\hat{y}_3$. Přitom platí:

$$(y_3 - \bar{y}) = (\hat{y}_3 - \bar{y}) + (y_3 - \hat{y}_3)$$

Tato rovnice ukazuje rozklad odchylky y_3 od průměru $\bar{y}$ do dvou složek. První, $(\hat{y}_3 - \bar{y})$ je odchylka predikce ležící na regresní přímce od průměrné hodnoty $\bar{y}$. Tato složka je vysvětlená existenci regresní přímky. Druhá složka, $(y_3 - \hat{y}_3)$ je odchylka změřeného bodu od regresní přímky. Tato složka je nevysvětlená. Předpokládáme, že body budou vysvětleny přímkou a přitom tomu tak úplně není, aniž bychom to uměli vysvětlit.

Rovnost, kterou jsme právě popsali pro odchylky, nalezneme i u součtu čtverců těchto odchylek. Platí

$$S_y = S_{\hat{y}} + S_{\hat{e}}$$

kde

- **celkový součet čtverců** odchylek dat od průměru je

$$S_y = \sum_{i=1}^{n} (y_i - \overline{y})^2 = S_{yy},$$

- **regresní součet čtverců** odchylek predikcí od průměru je

$$S_{\hat{y}} = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 = S_{xy}^2 / S_{xx},$$

- **reziduální součet čtverců** odchylek dat od predikcí (tj. od přímky) je

$$S_{\hat{e}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = S_{yy} - S_{xy}^2 / S_{xx} = S_y - S_{\hat{y}}.$$

Regresní součet čtverců ukazuje, kolik původního rozptylu je vysvětleno na základě regrese. Reziduální součet vypovídá o zbylém, tedy nevysvětleném, rozptylu.

Statistika pro popisovaný test je

$$F = \frac{(n-2)S_{\hat{y}}}{S_{\hat{e}}} \quad \sim F(1, n-2),$$

což je podíl rozptylu vysvětleného regresí a nevysvětleného. Statistika má rozdělení F se stupni volnosti 1 a $n-2$. Nulovou hypotézu H_0 : "data nejsou vhodná pro regresi" testujeme pomocí pravostranného F-testu.

[p_hodnota, T, df_num, df_den]=**f_test_reg**(x,y)

z-test nezávislosti reziduí

Dalším ověřením správného přístupu k regresní analýze, zejména k dostatečnosti lineární aproximace naměřených bodů, může být test nezávislosti reziduí. Důkazem toho, že aproximující křivka (většinou přímka) je postačující, je nezávislost „zbytků po aproximaci", tedy reziduí. Jestliže rezidua nejsou nezávislá, znamená to, že mezi nimi existují vazby, které by bylo možno ještě dále modelovat. Například, bude-li data generovat parabola, a my je budeme

aproximovat přímkou, budou rezidua závislá, a jejich závislost bude svědčit o tom, že lineární aproximace není dostačující. Bude třeba hledat lepší křivku - polynom druhého řádu.

Pro test nezávislosti reziduí použijeme Pořadový test nezávislosti podle odstavce

[p_hodnota,T,z_alpha]=**ordinal_test**(X,alpha)

11.3 Test logistické regrese

Vhodnost jednotlivých prvků regresního vektoru

Jedná se o test, který má prokázat, že vybraný prvek regresního vektoru je platný při vysvětlování závisle proměnné y. Za tímto účelem provedeme dvě regresní anylýzy. První s celým regresním vektorem a určíme hodnotu logaritmické věrohodnostní funkce $l(b)$ (50) a druhou s regresním vektorem s vynechaným testovaným prvkem a s logaritmickou věrohodnostní funkcí $l^{-}(b)$.

Statistika testu je

$$G = -2\left[l^{-}(b) - l(b)\right]$$

Rozdělení statistiky je Ch2, s jedním stupněm volnosti. Test je pravostranný, nulová hypotéza: testovaná veličina do regrese nepatří.

Vhodnost celého modelu logistické regrese

Celý logistický model je možno testovat podobně. Statistika G je pak dána ve tvaru

$$G = -2\left[l^{0}(b) - l(b)\right]$$

kde $l^{0}(b)$ je logaritmická věrohodnostní funkce pro model, tvořený pouze konstantou b_0. Statistika má opět Chi2 rozdělení, ale s počtem stupňů volnosti rovným počtu nezávislých proměnných. Test je pravostranný, nulová hypotéza: model není vhodný pro logistickou regresi.

12 Analýza rozptylu (ANOVA)

Podobně jako korelační analýza, která z rozboru celkového rozptylu dat dělá závěry o funkci regrese, pracuje i analýza rozptylu (ANalysis Of VAriance) z celkovým rozptylem dat. Rozkládá jej do jednotlivých tříd a dělá závěry o těchto třídách.

Předpokladem pro použití testu ANOVA je shoda rozptylů dat v jednotlivých třídách. K ověření tohoto předpokladu slouží Bartlettův test.

bude dodán pod jménem bartlett_test()

12.1 ANOVA při jednoduchém třídění

Uvažujeme několik podobných zdrojů dat a chceme se přesvědčit, že všechny tyto zdroje fungují stejně, tj. průměry dat změřených na jednotlivých zdrojích jsou stejné. Postup testu uvedeme podle příkladu.

Příklad

Testujeme tři automobily stejné značky pro závod do vrchu. Vlivem nestejných podmínek (různí řidiči, povrch vozovky, atd.) se naměřené časy liší. Naším úkolem je zjistit, zda rozdíly v různých průměrných časech při opakovaných jízdách jsou způsobeny rozdílnou kvalitou automobilů nebo je lze přičíst na vrub náhodným vlivům. Data, která jsme naměřili, jsou v tabulce

Časy při opakovaných jízdách automobilů (min)

jízda	automobil 1	automobil 2	automobil 3
1	5.32	5.88	5.32
2	5.24	5.31	4.21
3	5.47	4.86	5.44
4	4.98	5.45	5.33
5	5.16	5.12	5.24

Označíme:

a počet tříd (automobilů - počet sloupců tabulky),

n_j počet měření v j-té třídě (počet řádků tabulky - zde stejný počet měření, ale nemusí být),

$N = \sum_{j=1}^{a}$ počet všech měření (počet všech prvků tabulky),

$x_{i,j}$ je měření j-tého automobilu při i-té jízdě (prvek tabulky na pozici (i, j)),

$x_j = [x_{1,j}, x_{2,j}, \ldots, x_{n_j,j}]$ měření od j-tého automobilu (j-tý sloupec tabulky,

$\overline{x}_{\cdot,j} = \frac{1}{n}\sum_{i=1}^{n} x_{i,j}$ průměry dat od j-tého automobilu (průměry ze sloupců),

$\overline{\overline{x}} = \frac{1}{a}\sum_{i=1}^{a} \overline{x}_j$ průměr z průměrů pro jednotlivé automobily (celkový průměr tabulky),

pro $j = 1, 2, \ldots, a$.

Celkový součet čtverců změřených dat bude

$$S_C^2 = \sum_{j=1}^{a}\sum_{i=1}^{n_j} \left(x_{i,j} - \overline{\overline{x}}\right)^2 == \sum_{j=1}^{a}\sum_{i=1}^{n_j} \left(x_{i,j} - \overline{x}_{\cdot,j} + \overline{x}_{\cdot,j} - \overline{\overline{x}}\right)^2 =$$

$$= \sum_{j=1}^{a} \sum_{i=1}^{n_j} (x_{i,j} - \overline{x}_{\cdot,j})^2 + 2 \sum_{j=1}^{a} \sum_{i=1}^{n_j} (x_{i,j} - \overline{x}_{\cdot,j}) \left(\overline{x}_{\cdot,j} - \overline{\overline{x}}\right) + \sum_{j=1}^{a} \sum_{i=1}^{n_j} \left(\overline{x}_{\cdot,j} - \overline{\overline{x}}\right)^2 =$$

$$= \sum_{j=1}^{a} \sum_{i=1}^{n_j} (x_{i,j} - \overline{x}_{\cdot,j})^2 + \sum_{j=1}^{a} n_j \left(\overline{x}_{\cdot,j} - \overline{\overline{x}}\right)^2,$$

protože

- druhá závorka v prostředním členu je nezávislá na i, můžeme ji tedy vytknout před sumu a součet prvního členu dá $(\bar{x}_{\cdot,j} - \bar{x}_{\cdot,j}) = 0$.
- celý třetí člen je nezávislý na i a $\sum_{i=1}^{n_j} 1 = n_j$.

Definujeme:

S_U^2 **- součet čtverců uvnitř tříd**, který po normování dává rozptyl uvnitř tříd - tj. nevysvětlený rozptyl

$$S_U^2 = \sum_{j=1}^{a} \sum_{i=1}^{n_j} (x_{i,j} - \overline{x}_{\cdot,j})^2,$$

a normalizační faktor, stupně volnosti, který je

$$df_U = N - a,$$

protože volných proměnných je $n_1 - 1 + n_2 - 1 + \ldots, n_a - 1 = \sum_{j=1}^{a} n_j - a = N - a$.

S_M^2 **- součet čtverců mezi třídami** je to druhá část rozkladu S_C^2, která po normování dává rozptyl mezi třídami - tj. rozptyl vysvětlený růzností tříd

$$S_M^2 = \sum_{j=1}^{a} n_j (\overline{x}_{\cdot,j} - \overline{\overline{x}})^2$$

a normalizační faktor, stupně volnosti, je

$$df_M = a - 1,$$

protože průměrů je a a mezi nimi existuje jedna vazba $\overline{\overline{x}}$.

Statistika pro test je dána podílem

$$F = \frac{S_M^2/(a-1)}{S_U^2/(N-1)} \quad \sim F(a-1, N-1),$$

což je podíl rozptylu vysvětleného třídami a rozptylu nevysvětleného. Statistika má rozdělení F se stupni volnosti $a - 1$ a $N - 1$.

Nulová hypotéza H_0: "střední hodnoty tříd jsou stejné". Test je pravostranný.

Příklad (dokončení)

V našem příkladě o závodních automobilech je
$a = 3;\ \ n_1 = n_2 = n_3 = 5;\ \ N = 3 \times 5 = 15 \cdots$ průměry $\overline{x}_i$: 5.23, 5.32, 5.11.

součet čtverců mezi třídami $S_M^2 = 0.12;$ součet čtverců uvnitř tříd $s_U^2 = 1.74.$

stupně volnosti mezi třídami $df_M = 2;$ stupně volnosti uvnitř tříd $df_U = 12.$

Statistika: $F_r = \frac{0.1177/2}{1.7435/12} = 0.405$

p-hodnota: $p_v = P(F > F_r) = 0.676$
při použití rozdělení $F(a-1, N-1)) = F(2, 12)$.

Závěr testu: automobily jsou stejné.

[p_hodnota,T]=**anova_1**(y,n)

12.2 ANOVA při dvojném třídění

Uvažujeme podobně jako v předchozím několik zdrojů dat. Rozdíl je v tom, že na tyto zdroje mohou mít vliv ještě další faktory. Cílem je určit, zda rozdíly v datech je možno vysvětlit vlivem těchto dalších faktorů, nebo zda zdroje jsou skutečně různé.

Příklad

Testujeme tři automobily stejné značky pro závod do vrchu a máme k dispozici pět řidičů. Každého z řidičů necháme vyjet závodní dráhu se všemi automobily a zaznamenáme jejich časy. Ty jsou uvedeny v tabulce

Časy při jízdách automobilů s různými řidiči (min)

	automobil 1	automobil 2	automobil 3
řidič 1	5.32	5.88	5.32
řidič 2	5.24	5.31	4.21
řidič 3	5.47	4.86	5.44
řidič 4	4.98	5.45	5.33
řidič 5	5.16	5.12	5.24

Naším úkolem je zjistit, zda rozdíly v různých průměrných časech při jízdách automobilů jsou způsobeny rozdílnou kvalitou automobilů nebo je lze přičíst na vrub rozdílům mezi řidiči.

Poznámka

V příkladě pro jednoduché třídění jsme se snažili rozdílnosti v časech vysvětlit pouze rozdílností automobilů. Nyní rozdíly v časech vysvětlujeme jednak rozdílností automobilů, ale také rozdílností řidičů.

Označíme:

a je počet automobilů (počet sloupců tabulky),

b počet řidičů (počet řádků tabulky),

$x_{i,j}$ je čas j-tého automobilu s i-tým řidičem

$\overline{x}_{\cdot,j} = \frac{1}{b}\sum_{i=1}^{b} x_{i,j}$ je průměrný čas j-tého automobilu (přes všechny řidiče),

$\overline{x}_{i,\cdot} = \frac{1}{a}\sum_{j=1}^{b} x_{i,j}$ je průměrný čas i-tého řidiče (se všemi automobily),

$\overline{\overline{x}}$ je celkový průměrný čas z celé tabulky.

Celkový rozptyl dat a jeho ortogonální rozklad je

$$S_C^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\left(x_{i,j} - \overline{\overline{x}}\right)^2 =$$

$$= \sum_{i=1}^{a}\sum_{j=1}^{b}\left(x_{i,j} - \overline{\overline{x}} + \overline{x}_{\cdot,j} - \overline{x}_{\cdot,j} + \overline{x}_{i,\cdot} - \overline{x}_{i,\cdot} + \overline{\overline{x}} - \overline{\overline{x}}\right)^2 =$$

$$= \sum_{i=1}^{a}\sum_{j=1}^{b}\left[\left(\overline{x}_{\cdot,j} - \overline{\overline{x}}\right) + \left(\overline{x}_{i,\cdot} - \overline{\overline{x}}\right) + \left(x_{i,j} - \hat{x}_{ij}\right)\right]^2 =$$

$$= \sum_{i=1}^{a}\sum_{j=1}^{b}\left(\overline{x}_{\cdot,j} - \overline{\overline{x}}\right)^2 + b\sum_{i=1}^{a}\left(\overline{x}_{i,\cdot} - \overline{\overline{x}}\right)^2 + a\sum_{j=1}^{b}\left(x_{i,j} - \hat{x}_{ij}\right)^2,$$

kde predikce $\hat{x}_{i,j}$ je

$$\hat{x}_{i,j} = \overline{x}_{\cdot,j} + \overline{x}_{i,\cdot} - \overline{\overline{x}} = \left(\overline{x}_{\cdot,j} - \overline{\overline{x}}\right) + \left(\overline{x}_{i,\cdot} - \overline{\overline{x}}\right) + \overline{\overline{x}} \tag{53}$$

a poslední rovnost platí, protože

$$\sum_{j=1}^{b} \left(\overline{x}_{\cdot,j} - \overline{\overline{x}}\right) = 0 \quad \text{a} \quad \sum_{i=1}^{a} \left(\overline{x}_{i,\cdot} - \overline{\overline{x}}\right) = 0$$

Definujeme:

S_A^2 **součet čtverců mezi průměry automobilů** (prostřední člen v rozkladu S_C^2)

$$S_A^2 = b \sum_{i=1}^{a} (\overline{x}_{i,\cdot} - \overline{\overline{x}})^2$$

a odpovídající stupně volnosti

$$df_A = a - 1$$

S_B^2 **součet čtverců mezi průměry řidičů** (poslední člen v rozkladu S_C^2)

$$S_B^2 = a \sum_{j=1}^{b} (\overline{x}_{\cdot,j} - \overline{\overline{x}})^2$$

se stupni volnosti

$$df_B = b - 1$$

S_U^2 **reziduální součet čtverců (uvnitř tříd)** (první člen v rozkladu S_C^2)

$$S_U^2 = \sum_{i=1}^{a} \sum_{j=1}^{b} (x_{i,j} - \hat{x}_{i,j})^2,$$

který je vypočten z rozdílů mezi daty $x_{i,j}$ a jejich předpověďmi $\hat{x}_{i,j}$ viz (53).

Stupně volnosti jsou

$$df_U = (a-1)(b-1).$$

Statistika pro test automobilů je dána podílem

$$F_A = \frac{(b-1)S_A^2}{S_U^2} \quad \sim F(a-1, (a-1)(b-1)),$$

což je podíl rozptylu vysvětleného rozdíly v automobilech a rozptylu nevysvětleného. Statistika má rozdělení F se stupni volnosti $a-1$ a $(a-1)(b-1)$.

Nulová hypotéza H_0: "střední hodnoty pro automobily jsou stejné". Test je pravostranný.

Statistika pro test řidičů je dána podílem

$$F_B = \frac{(a-1)S_B^2}{S_U^2} \quad \sim F(b-1, (a-1)(b-1)),$$

což je podíl rozptylu vysvětleného rozdíly v řidičích a rozptylu nevysvětleného. Statistika má rozdělení F se stupni volnosti $b-1$ a $(a-1)(b-1)$.

Nulová hypotéza H_0: "střední hodnoty pro řidiče jsou stejné". Test je pravostranný.

Funkce ve Scilabu:

[pv_col, pv_row]=**anova_2**(s)

Poznámka

Na první pohled by se mohlo zdát, že ANOVA s dvojným tříděním provádí pouze dva paralelní testy pro dvě veličiny, které mají vliv na sledovaná data. Není to však pravda. V každém z obou testů se berou v úvahu vlivy obou veličin. To co jsme dříve museli prohlásit za nevysvětlený rozptyl (který byl třeba proti vysvětlenému příliš veliký), lze nyní vysvětlit pomocí druhé veličiny. Tím se nevysvětlený rozptyl zmenší a test může dopadnout zcela jinak.

Příklad (dokončení)

V našem příkladě o závodních automobilech je
$a = 3, \quad b = 5,$

průměry pro automobily $\overline{x}_{i,\cdot}$: 5.23 5.32 5.11,

průměry pro řidiče $\overline{x}_{\cdot,j}$: 5.51 4.92 5.26 5.25 5.17.

součet čtverců mezi automobily $S_A^2 = 0.118,$ součet čtverců mezi řidiči $S_B^2 = 0.530,$

reziduální součet čtverců $S_U^2 = 1.213.$

stupně volnosti mezi automobily $df_A = 2,$ součet čtverců mezi řidiči $df_B = 4,$

stupně volnosti pro reziduální součet čtverců $df_U = 8.$

Statistika: $F_A = 0.388,$ p-hodnota$_A = 0.69$

Statistika: $F_B = 0.874,$ p-hodnota$_B = 0.52$

Závěr testu: ani automobily ani řidiči nejsou odlišní.

[pv_col, pv_row] = anova_2(x),

`pv_col, pv_row` jsou p-hodnoty,

`x` data (matice s prvním faktorem ve sloupcích a druhým faktorem v řádcích).

12.3 Souvislost metody ANOVA s regresní analýzou

Úlohu ANOVA lze řešit jako úlohu regresní analýzy. Nejprve se budeme zabývat úlohou ANOVA s jednoduchým tříděním, kdy změřený datový vzorek dělíme na třídy podle jediného faktoru, potom úlohou ANOVA s dvojným tříděním, kdy data dělíme do tříd podle dvou faktorů. Pro jednodušší zápis budeme uvažovat, že v každé třídě daného faktoru je vždy stejný počet měření. To mám umožní změřená data zapsat v kompaktním tvaru matice. Stejný počet měření ve třídách však není podmínkou úlohy.

ANOVA s jednoduchým tříděním

Tato úloha analyzuje datovou matici X s n řádky (měřeními) a a sloupci (třídami), data jednotlivých tříd X_j jsou uložena ve sloupcích, např.

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \\ x_{41} & x_{42} & x_{43} \end{bmatrix}$$

kde $X_j = [x_{1j}, x_{2j}, x_{3j}, x_{4j}]'$ jsou pro $j = 1, 2, 3$ data tříd 1, 2, 3.

Základní **model** pro jeden datový bod je

$$x_{ij} = \mu + a_j + \epsilon_{ij} \tag{54}$$

který říká, že bod, změřený ve třídě j má hodnotu μ, společnou pro všechny třídy *plus* hodmotu a_j, která je společná bodům jedné třídy, ale mezi třídami se může lišit *plus* šum. Čísla a_j, $j = 1, 2, \cdots, a$ tedy vyjadřují, o co se jednotlivé třídy liší. Budou-li všechna a_j nulová, budou třídy stejné.

Z uvedeného modelu dostaneme pro jeden sloupec matice X

$$X_j = \mu O_n + a_j O_n + \epsilon_j \tag{55}$$

kde O_n je sloupcový vektor jedniček o rozměru n a ϵ_j je j-tý sloupec matice šumu ϵ.

Seřadíme-li nyní sloupce X_j matice X pod sebe do vektoru, dostaneme s pomoci vztahu (55) následující model

$$\begin{bmatrix} x_{11} \\ x_{21} \\ x_{31} \\ x_{41} \\ x_{12} \\ x_{22} \\ x_{32} \\ x_{42} \\ x_{13} \\ x_{23} \\ x_{33} \\ x_{43} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{21} \\ \epsilon_{31} \\ \epsilon_{41} \\ \epsilon_{12} \\ \epsilon_{22} \\ \epsilon_{32} \\ \epsilon_{42} \\ \epsilon_{13} \\ \epsilon_{23} \\ \epsilon_{33} \\ \epsilon_{43} \end{bmatrix} \tag{56}$$

který je regresním modelem $x = d\theta + \epsilon$ s obecně parametry $\theta = [\mu, a_1, a_2, \cdots, a_j]'$. Vektor y je složen s měřených dat z jednotlivých tříd postupně pod sebou. Datová matice d je tvořena jedničkami v prvním sloupci a dále má úseky jedniček na místech odpovídajících jednotlivým třídám.

ANOVA s dvojným tříděním

Při úloze ANOVA s dvojným tříděním uvažujeme dva faktory. Zdrojem dat je matice X s n řádky (třídami 2. faktoru) a s a sloupci (třídami 1. faktoru).

Základní **model** pro jeden datový bod je

$$x_{ij} = \mu + a_j + b_i + \epsilon_{ij}$$

kde a_j se mění po sloupcích (v každé sloupci je stejné) a b_i se mění po řádcích (v každém řádku je stejné). a_j tedy modeluje různost tříd 1. faktoru (sloupců) a b_i různost tříd 2. faktoru (řádků).

Seřadíme-li opět matici X do sloupce jako v (56) a použijeme-li stejného příkladu, dostaneme

$$\begin{bmatrix} x_{11} \\ x_{21} \\ x_{31} \\ x_{41} \\ x_{12} \\ x_{22} \\ x_{32} \\ x_{42} \\ x_{13} \\ x_{23} \\ x_{33} \\ x_{43} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ a_1 \\ a_2 \\ a_3 \\ b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{21} \\ \epsilon_{31} \\ \epsilon_{41} \\ \epsilon_{12} \\ \epsilon_{22} \\ \epsilon_{32} \\ \epsilon_{42} \\ \epsilon_{13} \\ \epsilon_{23} \\ \epsilon_{33} \\ \epsilon_{43} \end{bmatrix} \tag{57}$$

to je opět regresní model $x = d\theta$ s parametry $\theta = [\mu, a, \cdots, b, \cdots]$

Regresi testujeme F-testem. Je-li $p_v \to 0$ je regrese vhodná, koef. nejsou blízko 0 a tedy faktory mají vliv. V opačném případě nemají.

13 Dodatky k pravděpodobnosti

13.1 σ-algebra a borelovské pole

σ-algebra

Jevové pole jsme definovali jako množinu všech jevů, tj. množinu všech podmnožin základního prostoru. Ve skutečnosti ale nemusíme pracovat s množinou všech jevů. Stačí uvažovat jen určitou podmnožinu jevů, která ale musí splňovat požadavky na σ-algebru.

Příklad

Uvažujme pokus hod kostkou. Základní prostor tohoto pokusu je $\Omega_k = \{1, 2, 3, 4, 5, 6\}$. Jestliže se rozhodneme použít kostku pro generování výsledků pokusu hod mincí, můžeme například definovat výsledky jako sudé a liché (označíme 1 a 2). Základní prostor nového pokusu bude $\Omega = \{1, 2\}$ a prostor všech jevů $A = \{\emptyset, \{1\}, \{2\}, \Omega\}$. Prostor jevů A je podmnožinou prostoru jevů A_k spojeného s původní úlohou o kostce.

$\sigma-$algebra jevů

Neprázdnou množinu jevů $S \subset \Omega$ nazveme $\sigma-$**algebrou**, jestliže platí:

1. jestliže $J \in S$ pak také $J' \in S$,
2. jestliže $J_i \in S$ pro $i = 1, 2, \cdots$, pak také $\cup_{i=1}^{\infty} J_i \in S$

Výklad definice

Uvedená definice říká, že aby množina jevů byla σ-algebrou, a tedy mohla být považována za jevové pole nějakého náhodného pokusu, musí být uzavřená na doplňky a (spočetná) sjednocení.

Příklad

Ověříme vlastnost uzavřenosti σ-algebry pro jevové pole koruny, kterou jsme obdrželi z kostky. Nejdříve ověříme doplňky: $\emptyset' = \Omega$, $1' = 2$, $2' = 1$, $\Omega' = \emptyset$. Dále ověříme sjednocení dvojice jevů: $\emptyset \cup 1 = 1$, $\emptyset \cup 2 = 2$, $\emptyset \cup \Omega = \Omega$, $1 \cup 2 = \Omega$, $1 \cup \Omega = \Omega$, $2 \cup \Omega = \Omega$. Pokračujeme trojicemi: $\emptyset \cup 1 \cup 2 = \Omega$, $\emptyset \cup 1 \cup \Omega = \Omega$, $\emptyset \cup 2 \cup \Omega = \Omega$, $1 \cup 2 \cup \Omega = \Omega$. A nakonec $\emptyset \cup 1 \cup 2 \cup \Omega = \Omega$. Všechny výsledky patří do A, a tedy A je σ-algebra.

Důsledek definice

Ze skutečnosti, že σ-algebra je uzavřená na doplňky a sjednocení plyne, že je také uzavřená na průniky.

Toto tvrzení dokážeme pro dvojice jevů: Jestliže $J_1, J_2 \in S$, pak také $J_1', J_2' \in S$. Potom $\left(J_1' \cup J_2'\right)' = J_1 \cap J_2 \in S$ - jak jsme chtěli dokázat.

σ-algebra generovaná jevem

Jestliže je dán základní prostor Ω a jeho podmnožina (jev) J, pak množina jevů

$$\{\emptyset,\, J,\, J',\, \Omega\}$$

se nazývá σ-algebra generovaná jevem J.

Tato množina má skutečně vlastnosti σ-algebry - ověřte.[8]

Poznámka

Pojem σ-algebry generované jevem lze rozšířit na pojem σ-algebry generované více jevy. Situaci budeme ilustrovat na příkladě pro dva jevy.

Příklad

Uvažujme základní prostor $\Omega = \{a, b, c, d\}$. Napište σ-algebru generovanou jevy $\{a\}$ a a, b.

Tato σ-algebra bude

$$S = \{\emptyset, \{a\}, \{a, b\}, \{b, c, d\}, \{c, d\}, \{a, c, d\}, \{b\}, \Omega\}$$

Ověřte že podmínky pro σ-algebru jsou splněny a srovnejte množinu S s množinou všech podmnožin základního prostoru Ω. Ověřte také, že množina S je uzavřená rovněž na průniky.

Borelovské pole na reálné ose

Borelovské pole na reálné ose je nejmenší σ-algebra intervalů na reálné ose, generovaná všemi intervaly $(-\infty, a\rangle$, kde $a \in R$. Prvky borelovského pole nazýváme borelovské množiny.

Ukážeme, příklady intervalů, které jsou borelovskými množinami.

1. Borelovské pole obsahuje všechny polo-nekonečné intervaly $(-\infty, a\rangle$, tedy například $(-\infty, 5\rangle$, $(-\infty, \frac{2}{3}\rangle$ nebo $(-\infty, \sqrt{7}\rangle$.
2. Obsahuje také doplňky uvedených intervalů (a, ∞), například $(5, \infty)$, $\left(\frac{2}{3}, \infty\right)$ nebo $\left(\sqrt{7}, \infty\right)$.
3. Dále musí obsahovat i průniky, tedy např. intervaly $(-\infty, b\rangle \cap (a, \infty) = (a, b\rangle$, což pro $a < b$ jsou omezené, zprava uzavřené intervaly.

[8] Nejdříve ověříme doplňky: $\emptyset' = \Omega$, $J' = (J')$, $(J')' = J$, $\Omega' = \emptyset$. Sjednocení dvojic: $\emptyset \cup J = J$, $\emptyset \cup J' = J'$, $\emptyset \cup \Omega = \Omega$, $J \cup J' = \Omega$, $J \cup \Omega = \Omega$, $J' \cup \Omega = \Omega$. Sjednocení trojic: $\emptyset \cup J \cup J' = \Omega$, $\emptyset \cup J \cup \Omega = \Omega$, $\emptyset \cup J' \cup \Omega = \Omega$, $J \cup J' \cup \Omega = \Omega$. Čtveřice: $\emptyset \cup J \cup J' \cup \Omega = \Omega$. Sledovaná množina je uzavřená na doplňky a sjednocení. Je tedy σ-algebrou.

4. Pomocí limitního přechodu lze z předchozího případu odvodit, že i všechny jednobodové intervaly jsou borelovské množiny. Například $\{5\}$, $\frac{2}{3}$ nebo $\sqrt{7}$ patří do borelovského reálného pole.

5. Z předchozího bodu plyne, že i všechny oboustranně uzavřené nebo otevřené omezené intervaly typu $\langle a, b\rangle$ nebo (a, b) jsou borelovské množiny.

Borelovské pole, jak je vidět, je velmi bohatá struktura intervalů. Lze však ukázat, že je menší než množina všech podmnožin reálné osy.

Poznámka

Pojem borelovského pole na reálné ose lze přímočaře rozšířit z reálné osy na prostor R^n. Například pro $n = 2$ (tj. v rovině) budeme místo o intervalech $(-\infty, a\rangle$ hovořit o útvarech $(-\infty, x\rangle \cap (-\infty, y\rangle$ a tak dále.

13.2 Exponenciální rozdělení

Pro exponenciální rozdělení chceme dokázat $P(X > t + s|X > t) = P(X > s)$, kde t a s jsou dvě kladná reálná čísla, která interpretujeme jako délky časových úseků.

Pravděpodobnost hodnot X menších než t je

$$P(X \leq t) = F(t) = \int_0^t \frac{1}{\delta} \mathrm{e}^{-\frac{u}{\delta}} du = 1 - \mathrm{e}^{-\frac{t}{\delta}}$$

Pravděpodobnost hodnot X větších než t je

$$P(X > t) = 1 - \left(1 - \mathrm{e}^{-\frac{t}{\delta}}\right) = \mathrm{e}^{-\frac{t}{\delta}}$$

Podmíněná pravděpodobnosti je podle definice podíl sdružené a marginální. Sdružená popisuje průnik obou jevů, tedy $(t + s, \infty) \cap (t, \infty) = (t + s, \infty)$. Vezmeme-li v úvahu konkrétní tvar pravděpodobností, dostaneme

$$P(X > t + s|X > t) = \frac{P(X > t + s \wedge X > t)}{P(X > t)} = \frac{P(X > t + s)}{P(X > t)} = \frac{\mathrm{e}^{-\frac{t+s}{\delta}}}{\mathrm{e}^{-\frac{t}{\delta}}} = \mathrm{e}^{-\frac{s}{\delta}} = P(X > s)$$

Tím je uvedená vlastnost exponenciálního rozdělení dokázána.

13.3 Odvození vztahů pro operátorový počet

Odvození provedeme pro spojitou náhodnou veličinu. Pro diskrétní veličinu je odvození stejné, jen integrál se nahradí sumací.

1. $E[a] = \int_{-\infty}^{\infty} af(x)\,dx = a\int_{-\infty}^{\infty} f(x)\,dx = a$, protože integrál přes celý obor hustoty pravděpodobnosti je vždy 0.

2. $E[a+X] = \int_{-\infty}^{\infty} (a+x) f(x)\,dx = \int_{-\infty}^{\infty} af(x)\,dx + \int_{-\infty}^{\infty} xf(x)\,dx = a + E[X]$, kde první úprava v posledním kroku je podle vztahu 1. a druhá podle definice.

3. $E[aX] = \int_{-\infty}^{\infty} axf(x)\,dx = a\int_{-\infty}^{\infty} xf(x)\,dx = aE[X]$,

4. $$\begin{aligned} E[X+Y] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x+y) f(x,y)\,dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [xf(x,y) + yf(x,y)]\,dxdy = \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xf(x,y)\,dxdy + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yf(x,y)\,dxdy = \\ &= \int_{-\infty}^{\infty} xf(x)\,dx + \int_{-\infty}^{\infty} yf(y)\,dy = E[X] + E[Y] \end{aligned}$$

5. plyne z 1. a 2.

6. $D[a] = E\left[(a - E[a])^2\right] = E\left[(a-a)^2\right] = 0$, při úpravě jsme využili vztah 1.

7. $$D[a+X] = E\left[(a+X-E[a+X])^2\right] = E\left[(a+X-a-E[X])^2\right] =$$
$$= E\left[(X-E[X])^2\right] = D[X],$$

8. $$D[aX] = E\left[(aX-E[aX])^2\right] = E\left[(aX-aE[X])^2\right] = E\left[a^2(X-E[X])^2\right] =$$
$$= a^2 E\left[(X-E[X])^2\right] = a^2 D[X]$$

9. $$D[X+Y] = E\left[(X+Y-E[X+Y])^2\right] = E\left[(X-E[X]+Y-E[Y])^2\right] =$$
$$E\left[(X-E[X])^2\right] + 2E\left[(X-E[X])(Y-E[Y]) + E\left[(Y-E[Y])^2\right]\right] =$$
$$D[X] + 2C[X,Y] + D[Y]$$
V případě, kdy X a Y jsou nezávislé, jsou také nekorelované a platí $C[X,Y] = 0$. Potom platí
$$D[X+Y] = D[X] + D[Y].$$

10. Plyne z 6. a 9.

13.4 Numerické hledání kořene nelineární rovnice

Hledáme řešení rovnice $f(x) = 0$. Najdeme (někde blízko kolem předpokládaného řešení) dva body x_1 a x_2 tak, aby $f(x_1)$ a $f(x_2)$ měly různá znaménka. Označíme $f_1 = \text{abs}(f(x_1))$ a $f_2 = \text{abs}(f(x_2))$. Body $[x_1, f(x_1)]$a $[x_2, f(x_2)]$ spojíme přímkou. Pro x-ovou souřadnici x průsečíku s touto přímkou platí

$$x = \frac{f_1 x_2 + f_2 x_1}{f_1 + f_2}.$$

Spočteme $f(x)$ a jestliže je kladné, pak položíme $x_2 = x$ je-li $f(x_2) > 0$ nebo $x_1 = x$ je-li $f(x_2) < 0$. To opakujeme tak dlouho, dokud $f(x)$ není dostatečně malé (třeba 10^{-5}). Řešením je poslední bod x.

13.5 Poisson jako limita binomického

Binomické: $\binom{n}{k} p^k (1-p)^{n-k}$. Poisson je limita pro $n \to \infty$ při $\lambda = np = const. \to p = \frac{\lambda}{n}$.

$$f_P = \lim_{n\to\infty} f_{B;} = \lim_{n\to\infty} \frac{n(n-1)\cdots(n-k+1)}{k!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{k}\right)^{n-k} =$$

$$= \lim_{n\to\infty} \frac{n(n-1)\cdots(n-k+1)}{n^k} \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{k}\right)^n \left(1 - \frac{\lambda}{k}\right)^{-k} = 1 \cdot \frac{\lambda^k}{k!} \mathrm{e}^{-\lambda} \cdot 1 = \mathrm{e}^{-\lambda} \frac{\lambda^k}{k!}$$

13.6 Gama a beta funkce

Gama funkce

$$\Gamma(x) = \int_0^\infty t^{x-1} \mathrm{e}^{-t} dt$$

Platí: $\Gamma(x+1) = x\Gamma(x)$.

Pro $x = n$ celé je: $\Gamma(n+1) = n!$.

V Matlabu: GAMMA(X)

Beta funkce

$$B(x,y) = \int_0^1 p^{x-1} (1-p)^{y-1} \, dt = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$$

V Matlabu: BETA(X,Y)

13.7 Doplnění na čtverec

Pro skalární veličinu

$$ax^2+2bxy+cy^2 = a\left(x^2+2\frac{b}{a}xy+\left(\frac{b}{a}\right)^2y^2-\left(\frac{b}{a}\right)^2y^2\right)+cy^2 = a\left(x+\frac{b}{a}y\right)^2+\left(c-\frac{b^2}{a}\right)y^2 \tag{58}$$

Pro vektorovou veličinu

$$\begin{gathered} X'AX+2X'BY+Y'CY = \\ = X'AX+2X'AA^{-1}BY+Y'B'\left(A'\right)^{-1}AA^{-1}BY-Y'B'A^{-1}BY+Y'CY = \\ \left(X+A^{-1}BY\right)'A\left(X+A^{-1}BY\right)+Y'\left(C-B'A^{-1}B\right)Y \end{gathered}$$

Pozn.: A je symetrická, proto $\left(A'\right)^{-1}=A^{-1}$.

13.8 Dvourozměrné normální rozdělení

Gauss 1 (s nulovou střední hodnotou)

$$f\left(x\right)=k.\exp\left\{-\frac{1}{2}ax^2\right\},$$

kde $k=\frac{1}{\sqrt{2\pi\sigma^2}}$ a $a=\frac{1}{\sigma^2}$ (ale to je teď jedno)

Gauss 2 (s nulovou střední hodnotou)

Kvadratická forma

$$\begin{bmatrix} x \\ y \end{bmatrix}'\begin{bmatrix} a & b \\ b & c \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix}=ax^2+2bxy+cy^2$$

matice uprostřed je inverze kovarianční matice.

Hustota

$$f\left(x,y\right)=K.\exp\left\{-\frac{1}{2}\left(ax^2+2bxy+cy^2\right)\right\}$$

Závorku v exponentu doplníme na čtverec v x (viz 58) a dostaneme

$$\begin{aligned} f\left(x,y\right) &= K.\exp\left\{-\frac{1}{2}\left[a\left(x+\frac{b}{a}y\right)^2+\left(c-\frac{b^2}{a}\right)y^2\right]\right\}= \\ &= K_1\exp\left\{-\frac{1}{2}\left[a\left(x+\frac{b}{a}y\right)^2\right]\right\}.K_2\exp\left\{-\frac{1}{2}\left(c-\frac{b^2}{a}\right)y^2\right\} \end{aligned}$$

První hp závisí jek na x, tak na y. Je to podmíněná hp pro x s podmínkou y. Druhá závisí jen na y a je to marginální hp.

14 Dodatky ke statistice

14.1 Odvození vzorců pro koeficienty regresní přímky

Pro naměřená data $x = [x_1, x_2, \cdots, x_n]$ a $y = [y_1, y_2, \cdots, y_n]$ chceme minimalizovat kriterium

$$\tilde{J} = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \overline{e^2},$$

kde $e_i = y_i - b_0 - b_1 x_i$ - vyjádřené z modelu regresní analýzy.

Poznámka

Je pravda, že jako kriterium pro konstrukci regresní přímky jsme zavedli kriterium J jako součet kvadrátů reziduí. Tady použijeme modifikované kriterium $\tilde{J}$ ve tvaru průměru z reziduí. Je to z důvodu přehlednějšího značení a na poloze minima to nic nemění.

Odvození začneme tím, že dosadíme do kriteria za e_i, umocníme a upravíme (budeme přitom používat operátorový počet pro střední hodnotu ve výběrovém tvaru - tedy pro $\bar{x}$).

$$\tilde{J} = \overline{(y - b_0 - b_1 x)^2} = \overline{y^2 + b_0^2 + b_1^2 x^2 - 2yb_0 - 2yb_1 x + 2b_0 b_1 x} =$$

$$= \overline{y^2} + b_0^2 + b_1^2\overline{x^2} - 2b_0\bar{y} - 2b_1\overline{xy} + 2b_0 b_1 \bar{x}$$

Poznámka

Při této úpravě je třeba si uvědomit, že posloupnosti, přes které středujeme jsou pouze x a y. Koeficienty b_0a b_1 jsou konstanty. Platí tedy např. $\overline{b_0} = b_0$ nebo $\overline{b_1 y} = b_1\bar{y}$. Porovnejte s vzorci pro operátorový počet se střední hodnotou. Roli náhodných veličin zde hrají navzorkovaná data - výběry, konstanty jsou stejné.

Pokračujeme v úpravě kriteria. Chceme kriterium minimalizovat. Jedna z možností, jak postupovat u kvadratické funkce (kterou naše kriterium je), je doplnit ji na úplný čtverec. Protože naše kriterium je funkcí dvou proměnných (b_0 a b_1), začneme postupně. Jako první vezmeme proměnnou b_0.

$$\tilde{J} = b_0^2 - 2b_0\left(\bar{y} - b_1\bar{x}\right) + \left(\bar{y} - b_1\bar{x}\right)^2 - \left(\bar{y} - b_1\bar{x}\right)^2 + \overline{y^2} + b_1^2\overline{x^2} - 2b_1\overline{xy} =$$

$$= \left(b_0 - [\bar{y} - b_1\bar{x}]\right)^2 - \bar{y}^2 + 2b_1\bar{x}\bar{y} - b_1^2\bar{x}^2 + \overline{y^2} + b_1^2\overline{x^2} - 2b_1\overline{xy} =$$

$$= \left(b_0 - [\bar{y} - b_1\bar{x}]\right)^2 + b_1^2\overline{x^2} - b_1^2\bar{x}^2 - 2b_1\overline{xy} + 2b_1\bar{x}\bar{y} + \overline{y^2} - \bar{y}^2 =$$

$$= (b_0 - [\bar{y} - b_1\bar{x}])^2 + b_1^2 s_x^2 - 2b_1 s_{xy} + s_y^2$$

V předchozích řádcích jsme doplnili na čtverec členy, které obsahovali b_0 a urovnali to, co zbylo. Přitom jsme použili vzorec pro výpočetní tvar rozptylu a kovariance ($D(X) = E(X^2) - E(X)^2 \rightarrow s_x^2 = \overline{x^2} - \bar{x}^2$ atd.)

Kriterium $\tilde{J}$ jsme zatím upravili do tvaru, ve kterém se b_0 vyskytuje jen v kvadratickém členu. Zbytek kriteria jeho volbou ovlivnit nemůžeme a nejmenší hodnota kvadrátu je nula. Tu dostaneme pro volbu

$$b_0 = \bar{y} - b_1\bar{x}. \tag{59}$$

Tím jsme provedli první část minimalizace podle b_0. Kvadrát „zmizí" a můžeme pokračovat v minimalizaci podle b_1.

$$\tilde{J} = b_1^2 s_x^2 - 2b_1 s_{xy} + s_y^2 = s_x^2 \left(b_1^2 - 2b_1 \frac{s_{xy}}{s_x^2} + \left(\frac{s_{xy}}{s_x^2}\right)^2 - \left(\frac{s_{xy}}{s_x^2}\right)^2 \right) + s_y^2 =$$

$$= s_x^2 \left(b_1 - \frac{s_{xy}}{s_x^2} \right)^2 + s_y^2 - \frac{(s_{xy})^2}{s_x^2}$$

A jsme opět ve stejné situaci. b_1 se objevuje jen pod kvadrátem, který jeho volbou

$$b_1 = \frac{s_{xy}}{s_x^2} \tag{60}$$

můžeme minimalizovat. Hodnota kriteria $\tilde{J}$ po minimalizaci je

$$\tilde{J} = s_y^2 - \frac{(s_{xy})^2}{s_x^2}.$$

Vztahy (59) a (60) odpovídají těm, které jsme uvedli v kapitole o regresní analýze.

14.2 Obecný vzorec pro regresní koeficienty

Tento vzorec má tvar

$$b = (X'X)^{-1} X'Y \tag{61}$$

kde pro případ základní statistické lineární regrese s jednou vysvětlující proměnnou (tj. pro model $y = b_0 + b_1 x$) a N měření je

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_N \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{bmatrix}$$

Dosadíme do jednotlivých částí vzorce (61). Nejdříve do výrazu v mimo inverzi

$$X'Y = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_N \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_N \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

a pak v inverzi

$$X'X = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_N \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdots & \cdots \\ 1 & x_N \end{bmatrix} = \begin{bmatrix} N & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}$$

Inverze je

$$(X'X)^{-1} = \frac{1}{N\sum x_i^2 - (\sum x_i)} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & N \end{bmatrix}$$

Celý výraz

$$b = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \frac{1}{N\sum x_i^2 - (\sum x_i)} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & N \end{bmatrix} \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix} =$$

$$= \frac{1}{N\sum x_i^2 - (\sum x_i)^2} \begin{bmatrix} \sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i \\ N\sum x_i y_i - \sum x_i \sum y_i \end{bmatrix}$$

Druhý člen

$$b_1 = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{N\sum x_i^2 - (\sum x_i)^2} = \frac{\frac{1}{N}\sum x_i y_i - \frac{1}{N}\sum x_i \frac{1}{N}\sum y_i}{\frac{1}{N}\sum x_i^2 - \left(\frac{1}{N}\sum x_i\right)^2} = \frac{s_{xy}}{s_x^2}$$

První člen (odečteme a zase přičteme člen $\frac{1}{N}\sum x_i \sum x_i \sum y_i$)

$$b_0 = \frac{\sum x_i^2 \sum y_i - \frac{1}{N}\sum x_i \sum x_i \sum y_i + \frac{1}{N}\sum x_i \sum x_i \sum y_i - \sum x_i \sum x_i y_i}{N\sum x_i^2 - (\sum x_i)^2} =$$

$\cdots$ vhodně rozdělíme $\cdots$

$$= \frac{\sum x_i^2 \sum y_i - \frac{1}{N}\sum x_i \sum x_i \sum y_i}{N\sum x_i^2 - (\sum x_i)^2} + \frac{\frac{1}{N}\sum x_i \sum x_i \sum y_i - \sum x_i \sum x_i y_i}{N\sum x_i^2 - (\sum x_i)^2} =$$

$\cdots$ vytkneme $\cdots$

$$= \frac{\frac{1}{N}\sum y_i \left(N\sum x_i^2 - \sum x_i \sum x_i\right)}{N\sum x_i^2 - \left(\sum x_i\right)^2} + \frac{\frac{1}{N}\sum x_i \left(\sum x_i \sum y_i - N\sum x_i y_i\right)}{N\sum x_i^2 - \left(\sum x_i\right)^2} =$$

$\cdots$ první člen zkrátíme, v ruhém se objeví b_1 $\cdots$

$$= \bar{y} + \bar{x}\frac{\left(\frac{1}{N}\sum x_i \frac{1}{N}\sum y_i - \frac{1}{N}\sum x_i y_i\right)}{\frac{1}{N}\sum x_i^2 - \left(\frac{1}{N}\sum x_i\right)^2} = \bar{y} - b_1\bar{x}$$

15 Vzorce

Statistiky pro intervalové odhady a testy hypotéz

Normální rozdělení

Odhad/test μ, **(σ^2 známe),** rozdělení $N(0;1)$

Interval spolehlivosti

$$\mu = \bar{x} \pm \frac{\sigma}{\sqrt{n}} . z_{\alpha/2}$$

Test hypotézy

$$h = \frac{\bar{x} - \mu}{\sigma} \sqrt{n}$$

Vzorce

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

Odhad/test μ, **(σ^2 neznáme)**, rozdělení $St(n-1)$

Interval spolehlivosti

$$\mu = \bar{x} \pm \frac{S}{\sqrt{n}} . t_{\alpha/2}$$

Test hypotézy

$$h = \frac{\bar{x} - \mu}{S} \sqrt{n}$$

Vzorce

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

Odhad/test σ^2, **(μ známe)**, rozdělení $Chí^2(n)$

Interval spolehlivosti

$$\frac{n.s_0^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{n.s_0^2}{\chi^2_{1-\alpha/2}}$$

Test hypotézy

$$h = \frac{n.s_0^2}{\sigma^2}$$

Vzorce

$$S_0^2 = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \mu \right)^2$$

Odhad/test σ^2, **(μ neznáme),** rozdělení $Chí^2(n-1)$

Interval spolehlivosti

$$\frac{(n-1).s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1).s^2}{\chi^2_{1-\alpha/2}}$$

Test hypotézy

$$h = \frac{(n-1).s^2}{\sigma^2}$$

Odhad/test μ_1-μ_2, **($\sigma_1^2 = \sigma_2^2$ neznáme)**, rozdělení $St(n_1+n_2-2)$

Interval spolehlivosti

$$\mu_1-\mu_2=\bar{x}_1-\bar{x}_2\pm s_P\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}.t_{\alpha/2}$$

Test hypotézy

$$h=\frac{\bar{x}_1-\bar{x}_2-(\mu_1-\mu_2)}{s_P\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

Vzorce

$$S_P=\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}$$

Odhad/test μ_1-μ_2, **($\sigma_1^2 \neq \sigma_2^2$ neznáme)**, rozdělení $St(\nu)$

Interval spolehlivosti

$$\mu_1-\mu_2=\bar{x}_1-\bar{x}_2\pm\sqrt{K_1+K_2}.t_{\alpha/2}$$

Test hypotézy

$$h=\frac{\bar{x}_1-\bar{x}_2-(\mu_1-\mu_2)}{\sqrt{K_1+K_2}}$$

Vzorce

$$\nu=\frac{(K_1+K_2)^2}{\frac{K_1^2}{n_1-1}+\frac{K_2^2}{n_2-1}},\qquad K_1=\frac{S_1^2}{n_1},\quad K_2=\frac{S_2^2}{n_2}$$

Odhad/test $\Delta = \mu_1$-μ_2, **(σ_1^2, σ_2^2 neznáme), párové výběry**, rozdělení $St(n-1)$

Interval spolehlivosti

$$\Delta=\overline{D}\pm\frac{S_D}{\sqrt{n}}.t_{\alpha/2}$$

Test hypotézy

$$h=\frac{\overline{D}-\Delta}{S_D}\sqrt{n}$$

Vzorce

$$D_i=X_{1,i}-X_{2,i}$$

$$S_D^2=\frac{1}{n-1}\sum_{i=1}^{n}(D_i-\overline{D})^2$$

ALTERNATIVNÍ ROZDĚLENÍ

Odhad/test π, rozdělení $N(0,1)$

Interval spolehlivosti

$$\pi=p\pm\sqrt{\frac{p(1-p)}{n}}.z_{\alpha/2}$$

Test hypotézy

$$h=\frac{p-\pi}{\sqrt{\pi(1-\pi)}}\sqrt{n}$$

Vzorce

$$p=\frac{n_1}{n}$$ (n_1 je poč. klad. výsl.)

Odhad/test π_1-π_2, rozdělení $N(0,1)$

Interval spolehlivosti

$$\pi_1-\pi_2=p_1-p_2\pm\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}.z_{\alpha/2}$$

Test hypotézy

$$h=\frac{p_1-p_2}{P_P\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

Vzorce

$$P_P=\sqrt{\bar{p}(1-\bar{p})};\quad \bar{p}=\frac{n_1p_1+n_2p_2}{n_1+n_2}$$

Regresní analýza

Odhad β_1 / test $\beta_1 = 0$, rozdělení $St(n-2)$

Interval spolehlivosti

$$\beta_1 = b_1 \pm \frac{s_e}{\sqrt{S_{xx}}} . t_{\alpha/2}$$

Test hypotézy

$$h = \frac{b_1}{s_e}\sqrt{S_{xx}}$$

Odhad E[Y], rozdělení $St(n-2)$

Interval spolehlivosti

$$E[Y] = \hat{y} \pm s_e \cdot \sqrt{\frac{1}{n} + \frac{\left(x_p - \bar{x}\right)^2}{S_{xx}}} . t_{\alpha/2}$$

Vzorce

$$S_{xx} = \sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 , \quad S_{xy} = \sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) ,$$

$$\hat{y} = b_0 + b_1 . x_p , \qquad s_e^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 = \frac{1}{n-2} S_R$$

Odhad Y, rozdělení $St(n-2)$

Interval spolehlivosti

$$Y = \hat{y} \pm s_e \cdot \sqrt{1 + \frac{1}{n} + \frac{\left(x_p - \bar{x}\right)^2}{S_{xx}}} . t_{\alpha/2}$$

Vzorce: (směrnice) (abs. člen) (výb. korelační koef.)

$$b_1 = \frac{S_{xy}}{S_{xx}}, \qquad b_0 = \bar{Y} - b_1 \bar{X} , \qquad r = \frac{S_{xy}}{\sqrt{S_{xx} . S_{yy}}}$$

Korelační analýza

Test ρ=0, rozdělení $St(n-2)$

Test hypotézy

$$h = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$$

Vzorce: (korelační koef.) (kovariance)

$$\rho_{XY} = \frac{\mathrm{cov}(X,Y)}{\sqrt{D[X].D[Y]}} , \quad \mathrm{cov}(X,Y) = E\left[\left(X - E[X]\right).\left(Y - E[Y]\right)\right]$$

16 Tabulky

Pravděpodobnost vlevo pro N(0;1) tj. P(Z<z)

z	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0,500	0,504	0,508	0,512	0,516	0,520	0,524	0,528	0,532	0,536
0,1	0,540	0,544	0,548	0,552	0,556	0,560	0,564	0,567	0,571	0,575
0,2	0,579	0,583	0,587	0,591	0,595	0,599	0,603	0,606	0,610	0,614
0,3	0,618	0,622	0,626	0,629	0,633	0,637	0,641	0,644	0,648	0,652
0,4	0,655	0,659	0,663	0,666	0,670	0,674	0,677	0,681	0,684	0,688
0,5	0,691	0,695	0,698	0,702	0,705	0,709	0,712	0,716	0,719	0,722
0,6	0,726	0,729	0,732	0,736	0,739	0,742	0,745	0,749	0,752	0,755
0,7	0,758	0,761	0,764	0,767	0,770	0,773	0,776	0,779	0,782	0,785
0,8	0,788	0,791	0,794	0,797	0,800	0,802	0,805	0,808	0,811	0,813
0,9	0,816	0,819	0,821	0,824	0,826	0,829	0,831	0,834	0,836	0,839
1	0,841	0,844	0,846	0,848	0,851	0,853	0,855	0,858	0,860	0,862
1,1	0,864	0,867	0,869	0,871	0,873	0,875	0,877	0,879	0,881	0,883
1,2	0,885	0,887	0,889	0,891	0,893	0,894	0,896	0,898	0,900	0,901
1,3	0,903	0,905	0,907	0,908	0,910	0,911	0,913	0,915	0,916	0,918
1,4	0,919	0,921	0,922	0,924	0,925	0,926	0,928	0,929	0,931	0,932
1,5	0,933	0,934	0,936	0,937	0,938	0,939	0,941	0,942	0,943	0,944
1,6	0,945	0,946	0,947	0,948	0,949	0,951	0,952	0,953	0,954	0,954
1,7	0,955	0,956	0,957	0,958	0,959	0,960	0,961	0,962	0,962	0,963
1,8	0,964	0,965	0,966	0,966	0,967	0,968	0,969	0,969	0,970	0,971
1,9	0,971	0,972	0,973	0,973	0,974	0,974	0,975	0,976	0,976	0,977
2	0,977	0,978	0,978	0,979	0,979	0,980	0,980	0,981	0,981	0,982
2,1	0,982	0,983	0,983	0,983	0,984	0,984	0,985	0,985	0,985	0,986
2,2	0,986	0,986	0,987	0,987	0,987	0,988	0,988	0,988	0,989	0,989
2,3	0,989	0,990	0,990	0,990	0,990	0,991	0,991	0,991	0,991	0,992
2,4	0,992	0,992	0,992	0,992	0,993	0,993	0,993	0,993	0,993	0,994
2,5	0,994	0,994	0,994	0,994	0,994	0,995	0,995	0,995	0,995	0,995
2,6	0,995	0,995	0,996	0,996	0,996	0,996	0,996	0,996	0,996	0,996
2,7	0,997	0,997	0,997	0,997	0,997	0,997	0,997	0,997	0,997	0,997
2,8	0,997	0,998	0,998	0,998	0,998	0,998	0,998	0,998	0,998	0,998
2,9	0,998	0,998	0,998	0,998	0,998	0,998	0,998	0,999	0,999	0,999

Krit. hodnoty St(n) tj. takové t, že P(T>t)=alfa

n / alfa	0,2	0,1	0,05	0,025	0,01	0,005	0,0025	0,001
1	1,376	3,078	6,314	12,706	31,821	63,657	127,321	318,309
2	1,061	1,886	2,920	4,303	6,965	9,925	14,089	22,327
3	0,978	1,638	2,353	3,182	4,541	5,841	7,453	10,215
4	0,941	1,533	2,132	2,776	3,747	4,604	5,598	7,173
5	0,920	1,476	2,015	2,571	3,365	4,032	4,773	5,893
6	0,906	1,440	1,943	2,447	3,143	3,707	4,317	5,208
7	0,896	1,415	1,895	2,365	2,998	3,499	4,029	4,785
8	0,889	1,397	1,860	2,306	2,896	3,355	3,833	4,501
9	0,883	1,383	1,833	2,262	2,821	3,250	3,690	4,297
10	0,879	1,372	1,812	2,228	2,764	3,169	3,581	4,144
11	0,876	1,363	1,796	2,201	2,718	3,106	3,497	4,025
12	0,873	1,356	1,782	2,179	2,681	3,055	3,428	3,930
13	0,870	1,350	1,771	2,160	2,650	3,012	3,372	3,852
14	0,868	1,345	1,761	2,145	2,624	2,977	3,326	3,787
15	0,866	1,341	1,753	2,131	2,602	2,947	3,286	3,733
16	0,865	1,337	1,746	2,120	2,583	2,921	3,252	3,686
17	0,863	1,333	1,740	2,110	2,567	2,898	3,222	3,646
18	0,862	1,330	1,734	2,101	2,552	2,878	3,197	3,610
19	0,861	1,328	1,729	2,093	2,539	2,861	3,174	3,579
20	0,860	1,325	1,725	2,086	2,528	2,845	3,153	3,552
21	0,859	1,323	1,721	2,080	2,518	2,831	3,135	3,527
22	0,858	1,321	1,717	2,074	2,508	2,819	3,119	3,505
23	0,858	1,319	1,714	2,069	2,500	2,807	3,104	3,485
24	0,857	1,318	1,711	2,064	2,492	2,797	3,091	3,467
25	0,856	1,316	1,708	2,060	2,485	2,787	3,078	3,450
26	0,856	1,315	1,706	2,056	2,479	2,779	3,067	3,435
27	0,855	1,314	1,703	2,052	2,473	2,771	3,057	3,421
28	0,855	1,313	1,701	2,048	2,467	2,763	3,047	3,408
29	0,854	1,311	1,699	2,045	2,462	2,756	3,038	3,396
30	0,854	1,310	1,697	2,042	2,457	2,750	3,030	3,385
40	0,851	1,303	1,684	2,021	2,423	2,704	2,971	3,307
50	0,849	1,299	1,676	2,009	2,403	2,678	2,937	3,261
60	0,848	1,296	1,671	2,000	2,390	2,660	2,915	3,232
80	0,846	1,292	1,664	1,990	2,374	2,639	2,887	3,195

Krit. hodnoty Chí2(n) tj. takové t, že P(T>t)=alfa

n / alfa	0,995	0,99	0,975	0,95	0,05	0,025	0,01	0,005
1	0,000	0,000	0,001	0,004	3,841	5,024	6,635	7,879
2	0,010	0,020	0,051	0,103	5,991	7,378	9,210	10,597
3	0,072	0,115	0,216	0,352	7,815	9,348	11,345	12,838
4	0,207	0,297	0,484	0,711	9,488	11,143	13,277	14,860
5	0,412	0,554	0,831	1,145	11,070	12,833	15,086	16,750
6	0,676	0,872	1,237	1,635	12,592	14,449	16,812	18,548
7	0,989	1,239	1,690	2,167	14,067	16,013	18,475	20,278
8	1,344	1,646	2,180	2,733	15,507	17,535	20,090	21,955
9	1,735	2,088	2,700	3,325	16,919	19,023	21,666	23,589
10	2,156	2,558	3,247	3,940	18,307	20,483	23,209	25,188
11	2,603	3,053	3,816	4,575	19,675	21,920	24,725	26,757
12	3,074	3,571	4,404	5,226	21,026	23,337	26,217	28,300
13	3,565	4,107	5,009	5,892	22,362	24,736	27,688	29,819
14	4,075	4,660	5,629	6,571	23,685	26,119	29,141	31,319
15	4,601	5,229	6,262	7,261	24,996	27,488	30,578	32,801
16	5,142	5,812	6,908	7,962	26,296	28,845	32,000	34,267
17	5,697	6,408	7,564	8,672	27,587	30,191	33,409	35,718
18	6,265	7,015	8,231	9,390	28,869	31,526	34,805	37,156
19	6,844	7,633	8,907	10,117	30,144	32,852	36,191	38,582
20	7,434	8,260	9,591	10,851	31,410	34,170	37,566	39,997
21	8,034	8,897	10,283	11,591	32,671	35,479	38,932	41,401
22	8,643	9,542	10,982	12,338	33,924	36,781	40,289	42,796
23	9,260	10,196	11,689	13,091	35,172	38,076	41,638	44,181
24	9,886	10,856	12,401	13,848	36,415	39,364	42,980	45,559
25	10,520	11,524	13,120	14,611	37,652	40,646	44,314	46,928
26	11,160	12,198	13,844	15,379	38,885	41,923	45,642	48,290
27	11,808	12,879	14,573	16,151	40,113	43,195	46,963	49,645
28	12,461	13,565	15,308	16,928	41,337	44,461	48,278	50,993
29	13,121	14,256	16,047	17,708	42,557	45,722	49,588	52,336

pokračováni v další tabulce

Krit. hodnoty Chí2(n) tj. takové t, že P(T>t)=alfa

pokračování

n / alfa	0,995	0,99	0,975	0,95	0,05	0,025	0,01	0,005
30	13,787	14,953	16,791	18,493	43,773	46,979	50,892	53,672
31	14,458	15,655	17,539	19,281	44,985	48,232	52,191	55,003
32	15,134	16,362	18,291	20,072	46,194	49,480	53,486	56,328
33	15,815	17,074	19,047	20,867	47,400	50,725	54,776	57,648
34	16,501	17,789	19,806	21,664	48,602	51,966	56,061	58,964
35	17,192	18,509	20,569	22,465	49,802	53,203	57,342	60,275
36	17,887	19,233	21,336	23,269	50,998	54,437	58,619	61,581
37	18,586	19,960	22,106	24,075	52,192	55,668	59,893	62,883
38	19,289	20,691	22,878	24,884	53,384	56,896	61,162	64,181
39	19,996	21,426	23,654	25,695	54,572	58,120	62,428	65,476
40	20,707	22,164	24,433	26,509	55,758	59,342	63,691	66,766
41	21,421	22,906	25,215	27,326	56,942	60,561	64,950	68,053
42	22,138	23,650	25,999	28,144	58,124	61,777	66,206	69,336
43	22,859	24,398	26,785	28,965	59,304	62,990	67,459	70,616
44	23,584	25,148	27,575	29,787	60,481	64,201	68,710	71,893
45	24,311	25,901	28,366	30,612	61,656	65,410	69,957	73,166
50	27,991	29,707	32,357	34,764	67,505	71,420	76,154	79,490
55	31,735	33,570	36,398	38,958	73,311	77,380	82,292	85,749
60	35,534	37,485	40,482	43,188	79,082	83,298	88,379	91,952
65	39,383	41,444	44,603	47,450	84,821	89,177	94,422	98,105
70	43,275	45,442	48,758	51,739	90,531	95,023	100,425	104,215
75	47,206	49,475	52,942	56,054	96,217	100,839	106,393	110,286
80	51,172	53,540	57,153	60,391	101,879	106,629	112,329	116,321
85	55,170	57,634	61,389	64,749	107,522	112,393	118,236	122,325
90	59,196	61,754	65,647	69,126	113,145	118,136	124,116	128,299
95	63,250	65,898	69,925	73,520	118,752	123,858	129,973	134,247
100	67,328	70,065	74,222	77,929	124,342	129,561	135,807	140,169

Krit. hodnoty F(m;n) pro alfa=0.05 tj. takové z, že P(Z>z)=0.05

n / m	1	2	3	4	5	6	7	8	9	10
1	161,448	199,500	215,707	224,583	230,162	233,986	236,768	238,883	240,543	241,882
2	18,513	19,000	19,164	19,247	19,296	19,330	19,353	19,371	19,385	19,396
3	10,128	9,552	9,277	9,117	9,013	8,941	8,887	8,845	8,812	8,786
4	7,709	6,944	6,591	6,388	6,256	6,163	6,094	6,041	5,999	5,964
5	6,608	5,786	5,409	5,192	5,050	4,950	4,876	4,818	4,772	4,735
6	5,987	5,143	4,757	4,534	4,387	4,284	4,207	4,147	4,099	4,060
7	5,591	4,737	4,347	4,120	3,972	3,866	3,787	3,726	3,677	3,637
8	5,318	4,459	4,066	3,838	3,687	3,581	3,500	3,438	3,388	3,347
9	5,117	4,256	3,863	3,633	3,482	3,374	3,293	3,230	3,179	3,137
10	4,965	4,103	3,708	3,478	3,326	3,217	3,135	3,072	3,020	2,978

Krit. hodnoty F(m;n) pro alfa=0.025 tj. takové z, že P(Z>z)=0.025

n / m	1	2	3	4	5	6	7	8	9	10
1	647,789	799,500	864,163	899,583	921,848	937,111	948,217	956,656	963,285	968,627
2	38,506	39,000	39,165	39,248	39,298	39,331	39,355	39,373	39,387	39,398
3	17,443	16,044	15,439	15,101	14,885	14,735	14,624	14,540	14,473	14,419
4	12,218	10,649	9,979	9,605	9,364	9,197	9,074	8,980	8,905	8,844
5	10,007	8,434	7,764	7,388	7,146	6,978	6,853	6,757	6,681	6,619
6	8,813	7,260	6,599	6,227	5,988	5,820	5,695	5,600	5,523	5,461
7	8,073	6,542	5,890	5,523	5,285	5,119	4,995	4,899	4,823	4,761
8	7,571	6,059	5,416	5,053	4,817	4,652	4,529	4,433	4,357	4,295
9	7,209	5,715	5,078	4,718	4,484	4,320	4,197	4,102	4,026	3,964
10	6,937	5,456	4,826	4,468	4,236	4,072	3,950	3,855	3,779	3,717