Pravděpodobnost a statistika

Filip Peterek

15. květen 2022

1 Pravděpodobnost

1.1 Zakladni vzorce

Variace bez opakovani:

$$V(n,k) = \frac{n!}{(n-k)!}$$

Kombinace bez opakovani:

$$C(n,k) = \frac{n!}{k!(n-k)!}$$

Permutace:

$$P(n) = n!$$

Variace s opakovanim:

$$V*(n,k) = n^k$$

Kombinace s opakovanim:

$$C*(n,k) = C*(n+k-1,k) = \frac{(n+k-1)!}{(n-1)!*k!}$$

Permutace s opakovanim:

$$P*(n_1,n_2,...,n_k) = \frac{P(n)}{P(n_1)*P(n_2)*...*P(n_k)} = \frac{n!}{n_1!*n_2!*...*n_3!}$$

Prunik jevu:
$$P(A \cap B) = P(A|B) * P(B)$$

Prunik nezavislych jevu:
$$P(A \cap B) = P(A) * P(B)$$

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

1.2 Bayesuv vzorec

Nastal jev A, hledam pravdepodobnost, ktery z jevu B_i jev A zpusobil.

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

1.3 Nahodna velicina

Stredni hodnota:
$$\mu = \sum_{(i)} x_i * P(x_i)$$
$$\mu = \int_{-\infty}^{\infty} x_i * P(x_i)$$

$$E(aX + b) = aE(X) + b$$
$$E(\sum_{i}^{n} X_i) = \sum_{i}^{n} E(X_i)$$

Centralni moment r-teho radu:

$$\mu'_r = \sum_{(i)} (x_i - E(X))^r * P(x_i)$$
$$\mu'_r = \int_{-\infty}^{\infty} (x_i - E(X))^r * P(x_i)$$

Variance:

$$D(X) = \sum_{(i)} (x_i - E(X))^2 * P(x_i)$$

$$D(X) = \int_{-\infty}^{\infty} (x_i - E(X))^2 * P(x_i)$$

$$D(X) = E(X^2) - (E(X))^2$$

$$D(aX + b) = a^2 D(X)$$

Smerodatna odchylka:

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

Sikmost:

$$\alpha_3 = \frac{\mu_3}{\sigma^3}$$

Spicatost:

$$\alpha_4 = \frac{\mu_4}{\sigma^4}$$

Modus: nejpocetnejsi prvek, prvek s nejvyssi pravdepodobnosti

1.4 Nahodny vektor

Vektor, jehoz slozky jsou nahodne veliciny

Vztahy jsou ekvivalentni nahodne velicine, ale upravene pro vektor

Ukazka:

Necht $\boldsymbol{X} = (X, Y)$ je nahodny vektor. Potom plati:

$$E(\boldsymbol{X}) = (E(X), E(Y))$$

1.5 Nezavislost nahodnych velicin

Necht $\boldsymbol{X} = (X, Y)$ je nahodny vektor. X, Y jsou nezavisle, prave kdyz plati:

$$F(x, y) = F_X(x) * F_Y(y)$$

1.6 Kovariance a koeficient korelace

Kovariance $cov(X, Y)$

$$cov(X, Y) = E[(X - E(X)) * (Y - E(Y))]$$

Kladna hodnota kovariance: zvysi se X $\implies$ pravdepodobne se zvysi Y Zaporna hodnota kovariance: zvysi se X $\implies$ pravdepodobne se snizi Y

$$cov(X, Y) = E(XY) - E(X) * E(Y)$$
$$cov(X, X) = D(X)$$
$$cov(a_1X + b_1, a_2X + b_2) = a_1a_2cov(X, Y)$$

Jsou-li X, Y nezavisle $\implies$ $cov(X, Y) = 0$

Korelacni koeficient $\rho(X, Y)$

$$\rho(X, Y) = \begin{cases} \frac{cov(X,Y)}{\sqrt{D(X)*D(Y)}}, & D(X), D(Y) \neq 0 \\ 0, & \text{jinak.} \end{cases} \tag{1}$$

Korelacni koeficient je mirou linearni zavislosti dvou slozek nahodneho vektoru.

$$\rho(X, Y) = \rho(Y, X)$$

$$\rho(X, X) = 1$$

$$X, Y \text{ jsou nezavisle } \implies \rho(X, Y) = 0$$

Implikace, naopak predchozi vztah neplati

$$\rho(X, Y) = 0 \implies X, Y \text{jsou nekorelovane}$$

1.7 Alternativni rozdeleni

Pouze dve moznosti, kazde ma svou pravdepodobnost

$$P(X = 1) = p$$
$$P(X = 0) = 1 - p$$

$$E(X) = p, D(X) = p * (1 - p)$$

1.8 Binomicke rozdeleni

$$X \to Bi(n, p)$$

$n \to$ velikost vyberu

p $\to$ pravdepodobnost uspechu

Provadim nezavisle pokusy (Bernoulliho pokusy), pravdepodobnost uspechu je konstantni

Binomicke rozdeleni - pravdepodobnost, ze v x pokusech se objevi y uspechu

Negativne binomicke rozdeleni – pocet pokusu do k-teho uspechu vcetne

$$X \to NB(k, p)$$

$k \to$ k - pocet uspechu

p $\to$ pravdepodobnost uspechu

```
20 pokusu
Pravdepodobnost jednoho uspechu je 0.3
Pravdepodobnost, ze uspechu bude pet a mene ziskame

pbinom(5, 20, 0.3)

Pravdepodobnost, ze uspechu bude nad pet

1 - pbinom(5, 20, 0.3)
```

```
Pocet uspechu je 6
Pozadovana pravdepodobnost pro 6 uspechu je 0.7
Pravdepodobnost uspechu pri jednom pokusu je 0.3

qnbinom(0.7, 6, 0.3) + 6

Je treba pricist 6, bo R pocita jen neuspechy,
kdezto my chceme vsechny pokusy

nbinom - negativne binomicke rozdeleni
```

1.9 Hypergeometricke rozdeleni

Popisuje pocet uspechu v **zavislych pokusech**.

$$X \to H(N, M, n)$$

N – velikost DS

M – pocet prvku s danou vlastnosti

n – velikost vyberu

`[r|d|p|q]hyper()`

Je-li $\frac{n}{N} < 0.05$, lze hypergeom. rozdeleni nahradit binomickym s param. n a (M/N)

1.10 Poissonovo rozdeleni

Modelujeme vyskyty udalosti na intervalu (plocha, cas, jakykoliv jiny interval)

- **Ordinarita** – pravdepodobnost vyskytu v limitne kratkem intevralu ($t \to 0$) je nulova
- **Stacionarita** – pravdepodobnost vyskytu zavisi pouze na delce intervalu
- **Nezavisle prirustky** – pocty udalosti v disjunktnich intervalech jsou nezavisle
- **Beznaslednost** – pravdepodobnost vyskytu nezavisi na case, ktery uplynul od minule udalosti

Rychlost vyskytu udalosti: λ

$$x \to Po(\lambda t)$$

$$E(X) = D(X) = \lambda t$$

$$(n > 30 \land p < 0.1) \implies Bi(n, p) \sim Po(np)$$

Priklad:

Pocet vyskytu: 30 za hodinu

Sledovane obdobi: 20 minut

$$\lambda = 30$$

$$t = 20min$$

$$\lambda t = \frac{30}{3} = 10$$

```
Pocet udalosti: 5
Casove obdobi: 15

ppois(5, lambda * 15)
```

1.11 Rovnomerne rozdeleni

Pravdepodobnost je konstantni na intervalu $(a; b)$, jinde je nulova

$$X \to R(a; b)$$

$$E(X) = \frac{a+b}{2}, D(X) = \frac{(a-b)^2}{12}$$

1.12 Exponencialni rozdeleni

Mejme Poissonuv proces.

Potom **doba do vyskytu prvni udalosti, pripadne doba mezi udalostmi,** je modelovatelna exponencialnim rozdelenim.

Bezpametove rozdeleni $\to$ doba do vyskytu udalosti nezavisi na predchozich vyskytech.

$$X \to Exp(\lambda)$$

$$f(t) = \begin{cases} \lambda * e^{-\lambda t}, & t > 0 \\ 0, & \text{jinak.} \end{cases} \tag{2}$$

$$f(t) = \begin{cases} 1 - e^{-\lambda t}, & t > 0 \\ 0, & \text{jinak.} \end{cases} \tag{3}$$

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

Intezita poruch:

$$\lambda(t) = \frac{f(t)}{1 - F(t)}$$

```
lambda = 1/30
interval = 10

pexp(10, 1/30)   -- get probability

quantile = 0.05

qexp(0.05, 1/30)  -- get desired interval

Pravdepodobnost, ze k T udalostem dojde drive nez v case X


Prumerna doba: 10
Pozadovany pocet udalosti = 50

mu_ti = 10
sigma_ti^2 = 10
sigma_ti = sqrt(10)

T = suma 100 Ti

T ~ N(50 * 10, sqrt(10) * 100)
T ~ N(udalosti * mu, sigma_ti * udalosti)

pnorm(X, udalosti*mu, sigma_ti * udalosti)

Vyplotovani:

x = seq(0, 1000, by=1)
y = dnorm(x, mean=udalosti*mu, sd=sigma_ti*udalosti)
png(file="aaa.png")
```

```
plot(x, y, type="l")
dev.off()
```

1.13 Weibullovo rozdeleni

Modelovani doby do vyskytu udalosti, umoznuje modelovat obdobi casnych poruch a obdobi starnuti

$$X \to W(\Theta, \beta)$$

$\Theta \to$ parametr meritka

$\beta \to$ parametr tvaru

$$\lambda(t) = konstatnta * t^{\beta-1}$$

Distribucni funkce =

$$F(t) = \begin{cases} 1 - e^{-\frac{t}{\Theta}^{\beta}}, & t > 0 \\ 0, & \text{jinak.} \end{cases} \tag{4}$$

Hustota p.

$$F(t) = \begin{cases} \frac{\beta}{\Theta^{\beta}} t^{\beta-1} e^{-\frac{t}{\Theta}^{\beta}}, & t > 0 \\ 0, & \text{jinak.} \end{cases} \tag{5}$$

Intenzita poruch:

$$\lambda(t) = \frac{\beta}{\Theta^{\beta} t^{\beta-1}}$$

Priklad: $\Theta = 50$

Intenzita poruch je linearni a rostouci, tedy

$\beta = 2$

Hodnotu β ziskame z nasledujiciho vzorce

$$\lambda(t) = konstatnta * t^{\beta-1}$$

Casovy interval - deset

$$X \rightarrow W(\Theta = 50, \beta = 2)$$

Intenzita poruch – dosazenim do vzorce

$$\lambda(10) = \frac{2}{50^2} 10^{2-1} = 0.008$$

Pravdepodobnost, ze system bude 100 hodin bezporuchovy

$$P(X > 100) = 1 - F(100)$$

```
pweibull(100, beta, Theta)

1 - pweibull(100, 2, 50)
```

1.14 Erlangovo rozdeleni

Doba vyskytu do k-te udalosti v Poissonove procesu

k – pocet udalosti

λ – meritko

$$X_k \rightarrow Erlang(k, \lambda)$$

$$E(X_k) = \frac{k}{\lambda}$$

$$D(X_k) = \frac{k}{\lambda^2}$$

1.15 Normalni rozdeleni

μ – stredni hodnota

σ^2 – rozptyl

$$X \to N(\mu; \sigma^2)$$

Pravidlo 3σ

99.8 % prvku spada do intervalu $< \mu - 3\sigma; \mu + 3\sigma >$

Q-Q graf = graficky nastroj pro overeni normality

```
qqline(data, col="blue")
```

2 Statistika

Promenne

- Kvalitativni
  - Nominalni = nelze sortit
  - Ordinalni = lze sortit
- Kvantitativni
  - Diskretni
  - Spojite

2.1 Nominalni hodnota

Cetnost

Relativni cetnost

$$p_i = \frac{n_i}{n}$$

Modus – nejcastejsi prvek

Histogram, vysecovy graf

2.2 Ordinalni promenna

Kumulativni cetnost, kumulativni relativni cetnost

Soucet prvku varianty x nebo nizsi

Lorenzova krivka – vynasim kumulativni cetnosti

Paretova analyza, Paretuv princip - pravidlo $\frac{20}{80}$

2.3 Numericke promenne

Mira polohy a variability

Prumer = $\bar{x}$

Vlastnosti:

- Soucet odchylek od prumeru je roven nule
- Pricteme-li ke vsem hodnotam stejne cislo, o stejne cislo se zvedne prumer
- vynasobime-li vsechny hodnoty stejnym cislem, stejnym pomerem se zvysi prumer

Harmonicky prumer

Cast z celku, typicky uloha o spolecne praci

$$\bar{x}_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Vazeny prumer

$$\bar{x}_H = \frac{\sum_{i-1}^{k} n_i}{\sum_{i=1}^{n} \frac{n_i}{x_i}}$$

Geometricky prumer

Relativni zmena

$$\bar{x}_G = \sqrt[n]{x_1^{n_1} * x_2^{n_2} * ... * x_n^{n_k}}$$

Modus:

Pro diskretni hodnotu to je nejcetnejsi hodnota

Pro spojitou to je hodnota, okolo ktere je nejvyssi koncentrace hodnot – urcujeme pomoci **shorthu -** co nejkratsi interval takovy, ze v nem lezi alespon 50 % hodnot. Modus je potom stred shorthu.

Kvantil – rozdeluje dataset na dve casti, mensi nez kvantil a vetsi nez kvantil

Interkvartilove rozpeti IQR

$$IQR = x_{0.75} - x_{0.25}$$

MAD

Mean Absolute Deviation

median absolutnich odchylek kazde hodnoty od medianu

Vyberovy rozptyl

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

Suma kvadratu odchylek od prumeru podeleno velikosti datasetu bez jedne

Priceteme-li ke vsem hodnotam konstantu, rozptyl se nezmeni

Vynasobime-li vsechny hodnoty konstantou, rozptyl se prenasobi kvadratem konstanty

Vyberova smerodatna odchylka

$$\sqrt{s^2}$$

Variacni koeficient

Vyjadruje miru variability promenne x, lze stanovit pro promenne, ktere nabyvaji pouze kladnych hodnot pomoci vztahu

$$V_x = /fracV\bar{x}$$

2.4 Identifikace outliers

Vnitrni hradby

$$x_i < x_{0.25} - 1.5 * IQR$$

nebo

$$x_i > x_{0.75} + 1.5 * IQR$$

Pak x_i je outlier

z-souradnice

z-skore$_i$ = $\frac{x_i - \bar{x}}{s}$

|z-skore$_i$| $> 3 \implies ||\frac{x_i-\bar{x}}{s}| > 3 \implies |x_i - \bar{x}| > 3s \implies x_i$ je outlier

$x_{0.5}$-souradnice

$$|x_{0.5} - skore_i| = \left|\frac{x_i - x_{0.5}}{1.483MAD}\right| > 3 \implies x_i \text{ je outlier}$$

Odlehla a extremni pozorovani

Odlehla:

$$h_D = x_{0.25} - 1.5IQR$$
$$h_D = x_{0.75} + 1.5IQR$$

Extremni:

$$H_D = x_{0.25} - 3IQR$$
$$H_D = x_{0.75} + 3IQR$$

Vyberova sikmost

$$a = \frac{n}{(n-1)(n-2)} * \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}$$

a ¿ 0 ... prevazuji hodnoty mensi nez prumer a ¡ 0 ... prevazuji hodnoty vetsi nez prumer a = 0 ... symetricke rozlozeni

Vyberova spicatost

b = 0 ... odpovida normalnimu rozdeleni b ¿ 0 ... spicate rozdeleni b ¡ 0 ... ploche rozdeleni

2.5 Graficke znazorneni

Box plot

```
png(file='boxplot.png')
boxplot(machine1, machine2, machine3, machine4,
        main='Prumery lozisek', ylab='mm', names=machine.names)
dev.off()
```

2.6 Vyberove charakteristiky

Vyberovy prumer

Zakon velkych cisel

S rostoucim rozsahem vyberu se vyberovy prumer koncentruje stale vice okolo skutecneho prumeru

Centralni limitni veta

Nezavisle na rozdeleni, z ktereho X_i pochazi, se pro dostatecne velky vyber rozdeleni prumeru blizi normalnimu rozdeleni

Vyber neobsahuje odlehla pozorovani a rozsah je alespon 30

Viz priklad u Exp rozdeleni

2.7 Relativni cetnost

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

Vlastnosti

$$E(p) = \mu_p$$
$$D(p) = \sigma_p^2 = \frac{\pi(1-\pi)}{n}$$

$$p \sim N(\mu_p, \sigma_p^2)$$

Rozdil vyberovych prumeru

Vyber je max. dvacetina populace

Vybery jsou nezavisle

Platí predpoklady CLV – vybery pochazi z norm. rozdeleni, nebo jsou dostatecne velke (30+)

Pak plati:

$$E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2$$

$$D(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{sigma_2^2}{n_2}$$

Rozdil relativnich cetnosti

Rozsah kazde z populaci je dostatecne velky (vyber je max desetina populace)

Pro modelovani rozdilu lze pouzit norm. rozdeleni (dostatecne velke vybery)

Vybery jsou nezavisle

Pak plati:

$$E(p_1 - p_2) = \pi_1 - \pi_2$$

$$D(p_1 - p_2) = \frac{\pi_1(1 - \pi_1)}{n1} + \frac{\pi_2(1 - \pi_2)}{n2}$$

$$(p_1 - p_2) \sim N(E(p_1 - p_2), D(p_1 - p_2))$$

2.8 χ^2 rozdeleni

Soucet ctvercu nahodnych velicin s normovanych normalnim rozdelenim

Pocet nahodnych velicin = ν stupnu volnosti

$$X = \sum_{i=1}^{v} Z_i^2 \rightarrow \chi^2$$

Plati:

$$\frac{(n-1)S^2}{\sigma^2} = \chi^2_{n-1}$$

$$E(X) = \nu, D(X) = 2\nu$$

Pouziti:

Test, zda rozptyl souboru s norm. rozdelenim je roven σ_0^2

Overeni nezavislosti kategorialnich promennych

Test dobre shody - zda nahodne veliciny pochazi z urciteho rozdeleni

Priklad:

$\mu = 5\text{let}$

$\sigma = 6\text{mesicu}$

$P(S > 7) =?$

$$\frac{(n-1)S^2}{\sigma^2} \to \chi^2_{n-1}$$

$X = \frac{(n-1)S^2}{\sigma^2}$

$$X \to \chi^2_{19}$$

$(X > \frac{19.7^2}{36}),\ \text{tedy}\ (X > 25.86)$

$$1 - F_{\chi^2_{19}} = 0.134$$

```
1 - pchisq(q, df)
1 - pchisq(25.86, 19)
```

2.9 Studentovo rozdeleni

Z - nahodna velicina o norm. rozdeleni V - nahodna velicina o χ^2 rozdeleni s ν stupni volnosti