Statistická rozdělení

Václav Adamec
vadamec@mendelu.cz

Úvod

- Náhodná proměnná: matematická veličina, jejíž hodnoty oscilují. Produkt náhodného procesu – lze charakterizovat funkcí
- Hodnoty proměnné v oboru přípustných hodnot
- Rozdělení definují funkční vztah mezi hodnotami náhodné proměnné a četnostmi jejich výskytu
- Spojitá rozdělení: nekonečný počet možných hodnot, funkce pravděpodobnostní hustoty (p.d.f., f(y))
- Nespojitá rozdělení: konečný počet možných hodnot, pravděpodobnostní funkce (p.m.f., p(y))
- Kumulativní distribuční funkce (c.d.f., F(y))

Typy proměnných

- Náhodné proměnné: numerické (kvantitativní)
  nominální (kvalitativní): barva, pohlaví
- Numerické: kardinální (měřitelné)
  ordinální (pořadové): třídy jakosti, stup. klasifikace
- Kardinální: spojité (kontinuální): hmotnost, masná užitkovost
  nespojité (diskrétní): počty mláďat, defektů

Funkce popisující rozdělení

- Pravděpodobnostní funkce (p.m.f., p(y)) vyjadřuje pravděpost výskytu diskrétní hodnoty y v oboru možných hodnot

$$p(y) = p(Y = y) = F(y_i) - F(y_{i-1}) \qquad \sum_y p(y) = 1.0$$

- Funkce pravděpodobnostní hustoty (p.d.f., f(y)) spojité proměnné

$$f(y) = \frac{dF(y)}{dy} = F'(y)$$

- Kumulativní distribuční funkce (c.d.f., F(y)) vyjadřuje pravděpod. výskytu hodnoty y menší nebo rovné Y.

$$F(y) = P(y \le Y) = \sum_{y \le Y} p(y)$$ nespojitý případ

$$F(y) = P(y \le Y) = \int_{-\infty}^{Y} f(y)dy$$ spojitý případ

Střední hodnota (Expectation)

- Expectation (E) definujeme jako <u>první obecný moment</u>.
- Pro <u>spojitou</u> náhodnou proměnnou:

$$M_1' = \int_{-\infty}^{+\infty} y^1 f(y)dy \qquad E(Y) = \int_{-\infty}^{+\infty} yf(y)dy$$

- Pro <u>nespojitou</u> náhodnou proměnnou:

$$M_1' = \sum_y y^1 P(y) \qquad E(Y) = \sum_{i=1}^{k} y_i P(y_i)$$

Variance

- Varianci (Var) definujeme jako <u>druhý centrální moment</u>
- Pro <u>spojitou</u> proměnnou:

$$Var(y) = M_2 = \int_{-\infty}^{+\infty} [y - M_1']^2 f(y)dy$$

- Pro <u>nespojitou</u> proměnnou:

$$Var(y) = M_2 = \sum_y [y - M_1']^2 P(y)$$

- Obecně platí:

$$Var(Y) = E(y - E(y))^2 = E(y^2) - (E(y))^2$$

Příklad

- Je dáno rozdělení pravděpodobností diskrétní proměnné Y: Jaká je střední hodnota a variance ?

y	p(y)	F(y)
0	0,15	0,15
1	0,25	0,40
2	0,25	0,65
3	0,35	1,00

E(y) = 0 * 0,15 + 1 * 0,25 + 2 * 0,25 +3 * 0,35 = 1,8

Var(y) = $(0-1,8)^2$ * 0,15 + $(1-1,8)^2$ *0,25 + $(2-1,8)^2$ *0,25 + $(3-1,8)^2$ * 0,35 = 0 + 0,16 + 0,01 + 0,504 = 1,16

Pravidla pro expectation

$$E(c) = c$$
$$E(Y_i) = \mu_y$$
$$E(cY) = cE(Y) = c\mu_y$$
$$E(Y \pm c) = E(Y) \pm E(c) = \mu_y \pm c$$
$$E(Y \pm X) = E(Y) \pm E(X) = \mu_y \pm \mu_x$$
$$E(y_1 + y_2) = \mu_y + \mu_y = 2\mu_y$$
$$E(g(Y)) = \sum_{i=1}^{k} g(y_i)P(y_i) \qquad \text{Y diskrétní}$$
$$E(g(Y)) = \int_{-\infty}^{+\infty} g(y)f(y)dy \qquad \text{Y spojitá}$$

Pravidla pro varianci

$$Var(c) = 0$$
$$Cov(Y, c) = 0$$
$$Var(y_i) = \sigma_y^2$$
$$Var(cY) = c^2 Var(Y) = c^2 \sigma_y^2$$
$$Var(Y \pm c) = Var(Y) + Var(c) \pm 2\,\mathrm{cov}(Y, c) = \sigma_y^2 + 0 \pm 0 = \sigma_y^2$$
$$Var(Y \pm X) = Var(Y) + Var(X) \pm 2\,\mathrm{cov}(Y, X) = \sigma_y^2 + \sigma_x^2 \pm 2\sigma_{yx}$$

Najděte $E(\bar{y}) = E\left(\frac{\sum_{i=1}^{n} y_i}{n}\right)$ $\quad Var(\bar{y}) = Var\left(\frac{\sum_{i=1}^{n} y_i}{n}\right)$

Bernoulliho proměnná

- Binární nespojitá proměnná: pacient přežil (y = 1)
  pacient nepřežil (y = 0)

$$Y \sim Bernoulli\ (\pi)$$

kde π je parametr rozdělení (pravděpodobnost přežití, y = 1)
a 1 - π je pravděpodobnost nepřežití y = 0

Pravděpodobnostní f-ce: $P(y) = \pi^y (1-\pi)^{1-y}$

$$E(y) = \pi$$
$$Var(y) = \pi(1-\pi)$$

Jaké jsou hodnoty F(y=0) a F(y=1) ?

Binomické rozdělení

- Bernoulliho experiment opakovaný n - krát

$$Y \sim Bin\ (n, \pi)$$

- Bernoulliho opakování jsou vzájemně nezávislá
- Parametr π je stálý (Bernoulliho pokusy jsou identické, s vracením)
- Pravděpodobnostní funkce:

$$P(y) = \binom{n}{y} \pi^y (1-\pi)^{n-y}; y \geq 0; n > 0$$

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

$$E(y) = n\pi$$

$$Var(y) = n\pi(1-\pi)$$

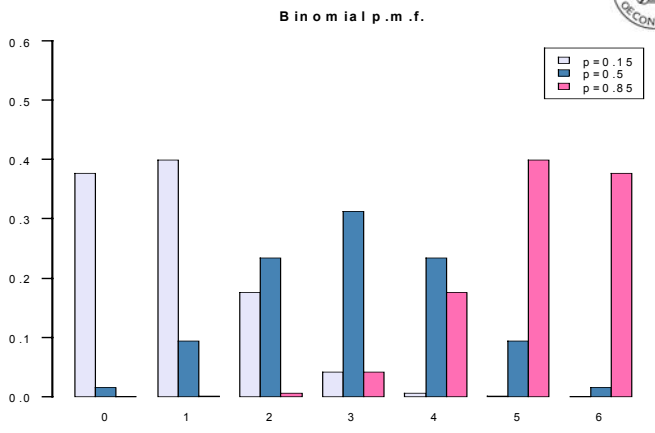
Binomické rozdělení

Bin(n = 6, $\pi = 0{,}5$)

y	c	p(y)	F(y)
0	1	0,015625	0,015625
1	6	0,093750	0,109375
2	15	0,234375	0,343750
3	20	0,312500	0,656250
4	15	0,234375	0,890625
5	6	0,093750	0,984375
6	1	0,015625	1,000000

- Pravděpodobnosti nejvíce 2 synů, nejméně 2 synů ?
- Pravděpodobnost nejvíce 2 dcer ? E(y) ? Var(y) ?

Binomická rozdělení pro různá π



Multinomické rozdělení

- Rozšíření binomického opakování na více (k>2) možných výstupů

$$Y_1, Y_2, ..., Y_k \sim Multinom(n, \pi_1, ..., \pi_k)$$

- Opakování jsou opět nezávislá
- Parametry π_1, ..., π_k jsou stálé
- Pravděpodobnostní funkce:

$$P(y_1, y_2, ..., y_k) = \frac{n!}{\prod_{i=1}^{k} y_i!} \cdot \prod_{i=1}^{k} \pi_i^{y_i}; \forall y_i \geq 0$$

$$E(y) = \sum_i y_i p_i \qquad Var(y) = n \prod_i \pi_i$$

Příklad: multinomické rozdělení

- Pravděpodobnost jedináčka $\pi_1 = 0,6$; dvojčat $\pi_2 = 0,3$; trojčat $\pi_3 = 0,1$
- Jaká je pravděpodobnost, že ve vrzích 13 matek bude 7x jedináček, 4x dvojče a 2x trojče ?

  $P(y_1=7, y_2=4, y_3=2 | \pi) = (13! /(7!4!2!)) * 0,6^7 * 0,3^4 * 0,1^2 = 0,0584$

- Jaká je pravděpodobnost, že ve vrzích 13 matek nebude ani jednou vrh s trojčaty ?
  Výsledek = 0,2542

Poissonovo rozdělení

- Proměnná: Počty bez přirozeného jmenovatele

$$Y \sim Poisson\ (\lambda)$$

- Binomické případy s $n \to \infty$ a s malým π
- Distribuční parametr $\lambda = n\pi$ z Binomického rozdělení
- Parametr λ je stálý
- Pravděpodobnostní funkce:

$$P(y) = \frac{e^{-\lambda} \lambda^y}{y!}; y \geq 0 \qquad E(y) = Var(y) = \lambda$$

- Příklad: Na části chromozomu o dané délce se vyskytují rekombinace v průměru (=λ) 1,05x za meiozi. Jaké jsou pravděpodobnosti výskytu y = 0,1,2,...,9 crossoverů na úseku?

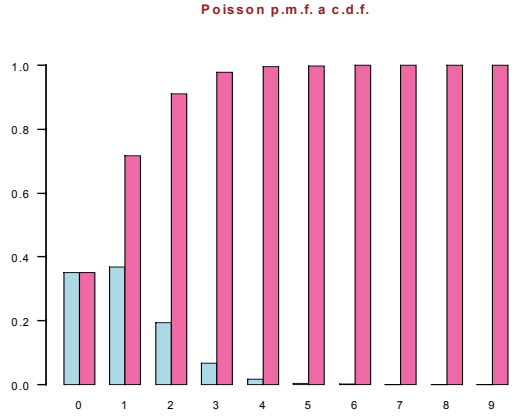
Poissonovo rozdělení

Poisson(λ = 1,05)

y	p(y)	F(y)
0	0,349938	0,34994
1	0,367435	0,71737
2	0,192903	0,91028
3	0,067516	0,97779
4	0,017723	0,99552
5	0,003722	0,99924
6	0,000651	0,99989
7	0,000098	0,99999
8	0,000013	1,00000
9	0,000001	1,00000

- Přesvědčete se, že E(y) = Var(y) = λ

Poissonovo rozdělení



Gaussovo rozdělení

- Spojitá proměnná Y generovaná polyfaktoriální sumací
- Určujících faktorů je mnoho a jsou nezávislé
- Možné hodnoty Y v oboru reálných čísel od -∞ do + ∞

$$Y \sim N(\mu, \sigma^2)$$

- Funkce pravděpodobnostní hustoty (p.d.f.):

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

- Hodnota f-ce pravděpodobnostní hustoty f(y) není pravděpodobnost !
- P(y = Y) = 0 !

$$E(y) = \mu_y \qquad Var(y) = \sigma_y^2$$

Gaussovo rozdělení

- Atributy:
  - Normálních rozdělení je nekonečně mnoho
  - Parametry μ a σ^2 definují každé normální rozdělení
  - Rozdělení je symetrické podle osy procházející průměrem
  - Lokační míry průměr, medián a modus jsou totožné
  - Plocha pod Gaussovou křivkou odpovídá P = 1,0
  - Pravidlo 34 – 14 – 2 se týká pravděpodobnosti výskytu hodnot (%) mezi μ a σ, σ a 2σ, 2σ a + ∞

Standardizované Gaussovo rozdělení

- Hodnoty z každého normálního rozdělení lze standardizovat
- Standardní normální proměnná z:

$$z = \frac{y - \mu_y}{\sigma_y} \sim N(\mu_z = 0, \sigma_z^2 = 1)$$

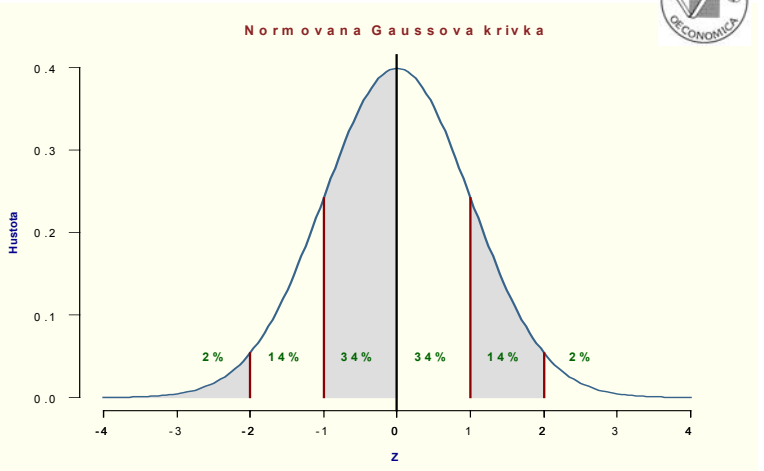
- P.d.f. Std. normálního rozdělení se značí $\phi(z)$

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

- C.d.f. Std. normálního rozdělení se značí $\Phi(z)$

$$\Phi(z) = P(z \le Z) = \int_{-\infty}^{Z} f(z)dz$$

Proměnná Z ~IID, N(0,1)



Kalkulace pravděpodobností

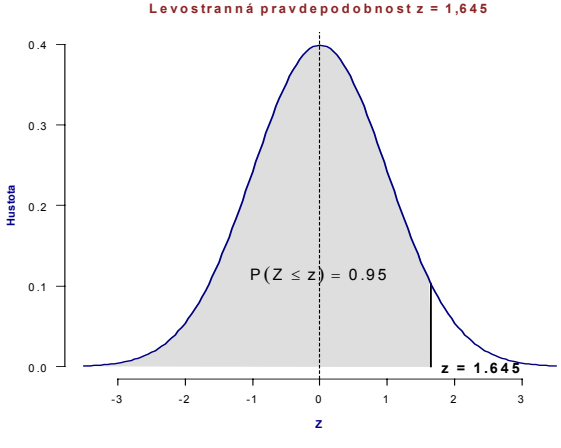
- Řešíme integrálem

$$P(z \le Z) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{Z} e^{\frac{-(z-\mu)^2}{2\sigma^2}} dz$$

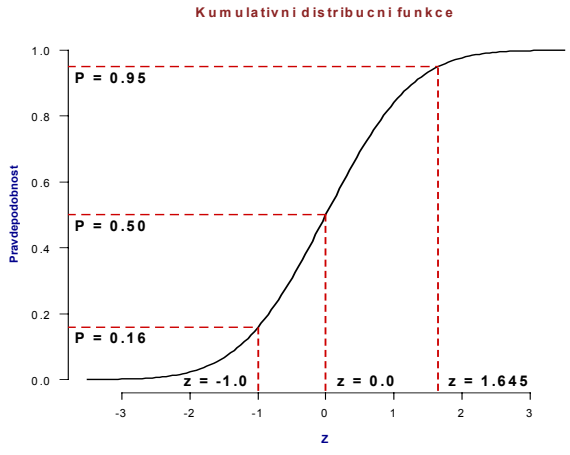
$$P(a \le z \le b) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{a}^{b} e^{\frac{-(z-\mu)^2}{2\sigma^2}} dz$$

- Platí že: $\phi(z) = \phi(-z)$ (důsledek souměrnosti)
  $\Phi(-z) = 1 - \Phi(z)$
  $z_{1-p} = -z_p$ (vyplývá z předchozího výrazu)

Kalkulace pravděpodobností



Kumulativní distribuční funkce F(z)



Pravděpodobnostní výrazy

- Princip:
  Kvantil z lze převést na levostrannou pravděpost P a obráceně při využití souměrnosti rozdělení Z ~ N(0,1)

- Kolik % dojnic se nachází v populaci s průměrem 4500 l a směrodatnou odchylkou 650 l mezi 3800 l až 5000 l ?

  z_1 = (3800 - 4500) / 650 = -1,07692

  z_2 = (5000 - 4500) / 650 = 0,76923

  1 - P(-1,07692) – (1 - P(0,76923))

  1 – 0,140758 – (1 - 0,779122)

  1 – 0,140758 – 0,220878 = 0,638364, tedy 64 %
- Jaká je pravděpodobnost výskytu dojnice s užitkovostí nad 6000 l ?

  z_1 = (6000 - 4500) / 650 = 2,30769

  1 - P(2,30769) = 1 - 0,989492 = 0,0105082, tedy 1,05 %

Pravděpodobnostní výrazy

- Jaká je pravděpodobnost výskytu dojnice s užitkovostí pod 3300 l ?

  z_1 = (3300-4500) / 650 = -1,84615

  P(-1,84615) = 1- P(1,84615) = 0,0324352, tedy 3,2 %
- 5 % nejlepších dojnic budou využity v ET. Stanovte selekční limit užitkovosti.

  z(0,95) = 1,64485

  4500 + 1,64485 * 650 = 5569,15 l , tedy 5570 l
- 15 % nejhorších dojnic nebudou zapojeny do reprodukce stáda. Stanovte limit užitkovosti pro vyřazení.

  z(0,15) = -1,03643

  4500 - 1,03643 * 650 = 3826,32, tedy 3830 l

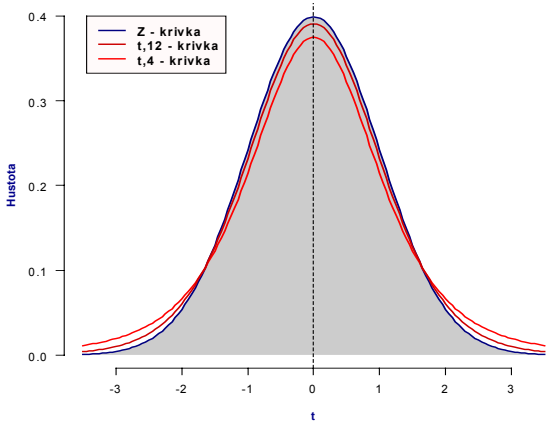
Studentovo t - rozdělení

- Gossettovo t - rozdělení
- Spojité rozdělení derivátů výběrových veličin mající vztah výběrovému rozptylu s limitovanými stupni volnosti ν
- Možné hodnoty t v oboru reálných čísel od -∞ do + ∞
- Rozdělení je unimodální a souměrné kolem nuly
- Platí, že $t_{1-p;\nu} = -t_{p;\nu}$
- Tvar p.d.f. definován parametrem ν (stupně volnosti)
- Vztah k proměnné Z dán výrazem

$$z_p = t_{p;\nu=\infty}$$

- V praktických případech, je-li přibližně $\nu > 120$

Studentovo t - rozdělení



Rozdělení Chí-kvadrát (Pearsonovo)

- Chí - kvadrát χ^2 je spojité rozdělení (p.d.f.) nezáporné veličiny
- Součet čtverců standardních normálních odchylek má Chí-kvadrát rozdělení s $\nu = n - 1$ stupni volnosti

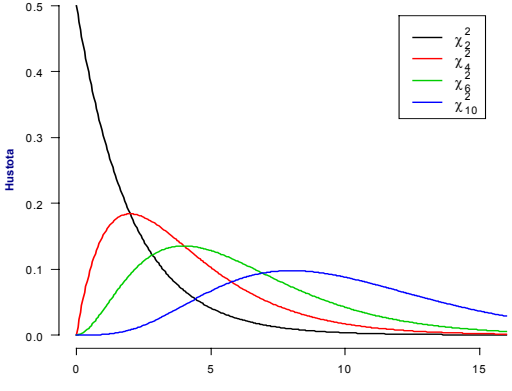
$$\sum_{i=1}^{n} z_i^2 = z_1^2 + z_2^2 + \ldots + z_n^2 \sim \chi^2_{n-1}$$

$$\sum_{i=1}^{n} z_i^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\sigma^2} = \frac{(n-1)s^2_{n-1}}{\sigma^2} \sim \chi^2_{\nu=n-1}$$

- Parametr rozdělení: stupně volnosti ν dány počtem nezávislých odchylek od průměru
- Počet stupňů volnosti ν určuje tvar křivky p.d.f.
- Užitečné při testování rozptylu a jeho derivátů (sumy čtverců)

$$E(\chi^2_\nu) = \nu \qquad Var(\chi^2_\nu) = 2\nu$$

Rozdělení Chí-kvadrát (Pearsonovo)



Fisherovo - Snedecorovo rozdělení

- Je spojité rozdělení pro podíl dvou nezávislých nezáporných veličin (rozptylů, součtu čtverců)
- U každé z veličin se předpokládá Chí-kvadrát rozdělení $\chi^2_{\nu 1}$ a $\chi^2_{\nu 2}$
- Podíl těchto veličin má F rozdělení se stupni volnosti $\nu 1$ (proměnná v čitateli) a $\nu 2$ (proměnná ve jmenovateli)
- F-rozdělení je vždy asymetrické
- Platí, že:

$$F_{p;\nu 1;\nu 2} = \frac{1}{F_{1-p;\nu 2;\nu 1}}$$

$$F_{p;1;\nu 2} = t^2_{p/2;\nu 2}$$

$F_{0.95,3,7} = 1/ F_{0.05,7,3} = 1/ 0.230053 = 4.34683$

$F_{0.95,1,4} = t^2_{0.975,4} = 2.77645^2 = 7.70865$

F-rozdělení