Tõenäosusteooria ja statistika II

Loeng 3

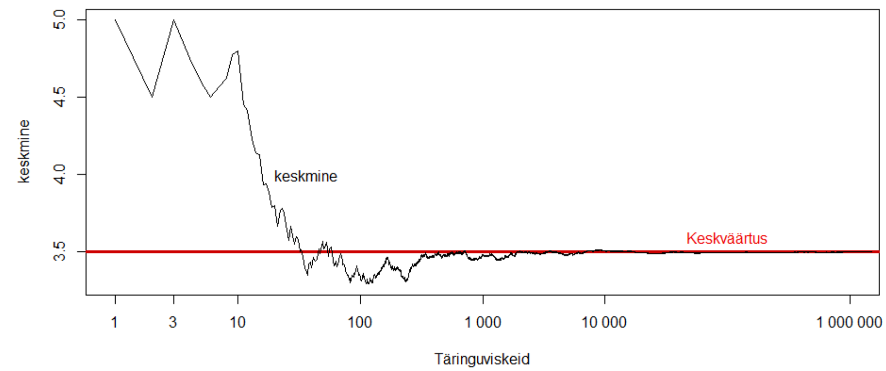
Keskväärtusest

Märt Möls

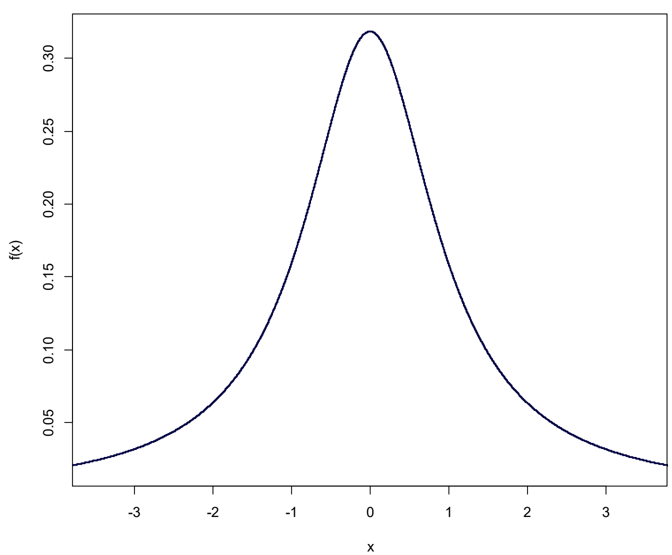
mart.mols@ut.ee

Keskväärtus

Keskmine koondub keskväärtuseks -
keskväärtuse definitsioon (statistiline ehk von Misese tõenäosus) või järeldus suurte arvude seadusest (Kolmogorovi tõenäosus)



Cauchy jaotus

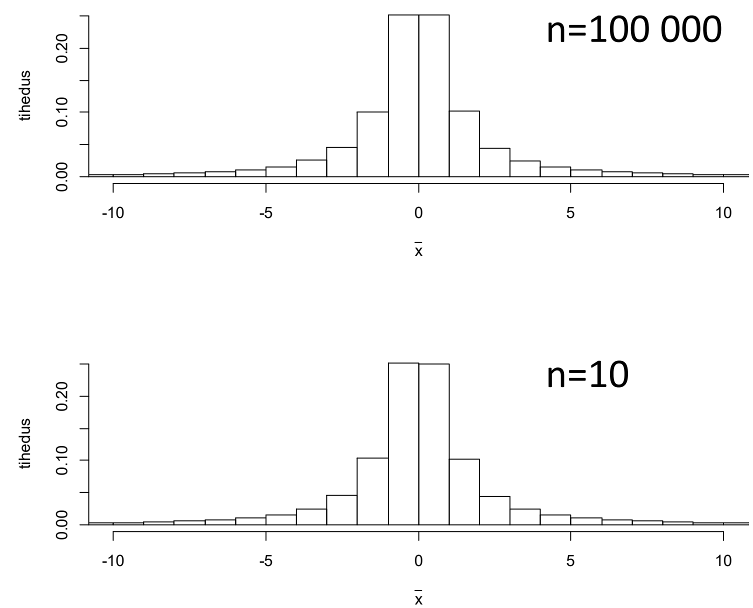


mediaan=0

sümmeetriline:
$f(x)=f(-x)$

Keskväärtus ei ole 0!

valimikeskmise jaotus



Diskreetne juhuslik suurus

$$\mathrm{E}(X) = \sum_i x_i P(X = x_i)$$

Näide 1: Bernoulli jaotus, *X*~B(1; *p*) või *X*~Be(*p*)

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

$$\mathrm{E}(X) = 1\,p + 0\,(1-p) = p$$

Bernoulli jaotusega on 'jah'/'ei' tunnused, kus 'jah' võib tähendada mingi arvamuse, haiguse jm. olemasolu.

Diskreetne juhuslik suurus (binoomjaotus)

$$\mathrm{E}(X) = \sum_i x_i P(X = x_i)$$

Näide 2: binoomjaotus, *X*~B(*n*; *p*)

(Tüüpiline) interpretatsioon:

n - katseseeria pikkus

p – meid huvitava sündmuse toimumise tõenäosus

X – mitu korda toimus meid huvitav sündmus *n* katse jooksul

$$P(X=x) = C_n^x p^x (1-p)^{n-x} \qquad \text{kui } x \in \{0; \ldots; n\}$$

$$C_n^x = \frac{n!}{x!(n-x)!}$$

Diskreetne juhuslik suurus (binoomjaotus)

$$\mathrm{E}(X) = \sum_i x_i P(X = x_i) = \sum_{i=0}^{n} i C_n^x p^i (1-p)^{n-i}$$

$$= \sum_{i=1}^{n} \cancel{i} \frac{n!}{\underset{(i-1)!}{\cancel{i!}}\,(n-i)!} p^i (1-p)^{n-i}$$

$$= n\,p \sum_{i=1}^{n} \frac{(n-1)!}{(i-1)!\,(n-1-(i-1))!} p^{i-1} (1-p)^{n-1-(i-1)}$$

$$= n\,p \sum_{i=0}^{n-1} \underbrace{\frac{(n-1)!}{i!\,(n-1-i)!} p^i (1-p)^{n-1-i}}_{P(Y=i),\ \text{kui } Y \sim B(n-1;p)}$$

$$= n\,p$$

Lisaküsimus (vabatahtlik): kas oskad binoomjaotust päriselus kasutada?

Teame ühe ohtliku bakteri genoomist 60 tähte pikka DNA lõiku (DNA koosneb tähtedest A,C,T,G), mis esineb teadaolevalt ainult ohtlikus bakteris ja mitte üheski teises elusolendis.

Loeme (sekveneerime) proovis olevat bakterit ja vaatame, kas leiame loetud tähejadas meie poolt otsitavat 60 tähte pikka ohtliku bakteri DNA lõiku. Kui leiame, siis loeme ohtliku bakterit leituks (test osutub positiivseks).

Paraku teeb tänapäevane tehnoloogia DNA lugemisel tõenäosusega 0,01 tähevea (lugemisvead teineteisest sõltumatult).

a) Mitu korda peaksime proovis oleva bakteri genoomi lugema, kui sooviksime, et avastaksime ohtliku bakteri (kui ta esineb proovis) vähemalt tõenäosusega 0,99?

Oletame, et proovis võib esineda ka sarnane ohutu bakter, millel on peaaegu samasugune 60 tähte pikk DNA lõik – vaid 1 täht kogu lõigu kohta on erinev.

b) Milline on tõenäosus anda ekslikult häiret, kui proovis on ohtliku bakteri asemel tema ohutu sugulane (ja me üritame tagada avastamistõenäosust 0,99)?

c) Kuidas saavutada ohtliku bakteri avastamistõenäosus 0,99 (kui ohtlik bakter tegelikult on proovis) ja samal ajal tagada, et ohutu sugulasliigi korral oleks valepositiivse testitulemuse tõenäosus väiksem kui 0,01?

Diskreetne juhuslik suurus

$$\mathrm{E}(g(X)) = \sum_i g(x_i) P(X = x_i)$$

$$\mathrm{E}(X) = 2{,}8$$

x	1	2	4
P(X=x)	0,2	0,3	0,5

$$\mathrm{E}(X^2) = 9{,}4$$

x	1	4	16
P(X^2=x)	0,2	0,3	0,5

Teoreem

Olgu X ja Y juhuslikud suurused ühisjaotusega $\{(x_i, y_j, p_{ij}) : i \in I,\ j \in J\}$ ning olgu $g : \mathbb{R}^2 \to \mathbb{R}$ selline funktsioon, et juhuslik suurus $g(X, Y)$ omab lõplikku keskväärtust. Sel juhul kehtib võrdus

$$E[g(X, Y)] = \sum_{i \in I} \sum_{j \in J} g(x_i, y_j) p_{ij}.$$

Märkus. Sarnane tulemus kehtib ka siis, kui vaatleme funktsiooni g rohkem kui kahest diskreetsest juhuslikust suurusest.

Näide. Ühisjaotus.

y \ x	1	2	4
1	0,1	0,1	0,2
2	0,3	0,2	0,1

$$\mathrm{E}(\mathrm{abs}(X - Y)) = 1{,}2$$

Pidev juhuslik suurus

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

Näide: ühtlane jaotus, *X*~U(2; 3): $f_X(x) = \begin{cases} 1, & x \in [2 \ldots 3] \\ 0, & \text{mujal} \end{cases}$

$$\mathrm{E}(X) = \int_{-\infty}^{2} x f_X(x) dx + \int_{2}^{3} x f_X(x) dx + \int_{3}^{\infty} x f_X(x) dx$$

$$= \int_2^3 x dx = x^2/2 \Big|_{x=2}^{x=3} = 9/2 - 4/2 = 5/2 = 2{,}5$$

Kui $g : \mathbb{R}^2 \to \mathbb{R}$ on "piisavalt heade omadustega" funktsioon (nt pidev või selline, mille valemit me oskame kirja panna) ning

$$\iint_{\mathbb{R}^2} |g(x, y)| f_{X,Y}(x, y)\, dx\, dy < \infty,$$

siis

$$E(g(X, Y)) = \iint_{\mathbb{R}^2} g(x, y) f_{X,Y}(x, y)\, dx\, dy.$$

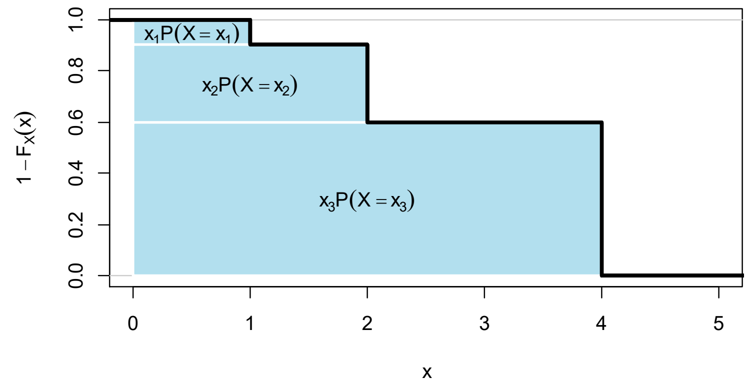
Keskväärtus, universaalsem

$$\mathrm{E}(X) = -\int_{-\infty}^{0} F_X(x)dx + \int_{0}^{\infty} (1 - F_X(x))dx$$

Keskväärtus, universaalsem

$$\mathrm{E}(X) = \underbrace{-\int_{-\infty}^{0} F_X(x)dx}_{0} + \int_{0}^{\infty} (1 - F_X(x))dx$$

x	1	2	4
P(X=x)	0,1	0,3	0,6



Keskväärtus, X>0 pidev

$$\mathrm{E}(X) = \underbrace{-\int_{-\infty}^{0} F_X(x)dx}_{0} + \int_{0}^{\infty} (1 - F_X(x))dx$$

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f_X(x)dx = \int_{0}^{\infty} x f_X(x)dx$$

sest X>0

$$\int uv'dx = uv - \int vu'dx$$

$$u = x \quad u' = 1$$

$$v' = f(x) \quad v = -(1 - F(x))$$

$$= \lim_{b\to\infty} -x(1 - F(x))\Big|_0^b - \int_0^\infty -(1 - F(x))dx$$

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f_X(x)dx = \int_{0}^{\infty} x f_X(x)dx$$

sest X>0

$$\int uv'dx = uv - \int vu'dx$$

$$u = x \quad u' = 1$$

$$v' = f(x) \quad v = -(1 - F(x))$$

$$= \lim_{b\to\infty} -x(1 - F(x))\Big|_0^b - \int_0^\infty -(1 - F(x))dx$$

$$= -\lim_{b\to\infty} b(1 - F(b)) + \int_0^\infty (1 - F(x))dx$$

$$b\int_b^\infty f_X(x)dx = \int_b^\infty b f_X(x)dx$$

$$\leq \int_b^\infty x f_X(x)dx$$

Keskväärtuse (triviaalsed) omadused

Olgu *a*, *b* konstandid ja *X* ja *Y* juhuslikud suurused. Siis

1. E(*a* + *b X*) = *a* + *b* E(*X*)
2. E(X + Y) = E(X) + E(Y)

$$\mathrm{E}(X+Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)f_{X,Y}(x,y)dxdy$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xf_{X,Y}(x,y)dxdy + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}yf_{X,Y}(x,y)dxdy$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xf_{X,Y}(x,y)dydx + \int_{-\infty}^{\infty}y\int_{-\infty}^{\infty}f_{X,Y}(x,y)dxdy$$

$$\mathrm{E}(X+Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)f_{X,Y}(x,y)dxdy$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xf_{X,Y}(x,y)dxdy + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}yf_{X,Y}(x,y)dxdy$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xf_{X,Y}(x,y)dydx + \int_{-\infty}^{\infty}y\int_{-\infty}^{\infty}f_{X,Y}(x,y)dxdy$$

$$= \int_{-\infty}^{\infty}x\int_{-\infty}^{\infty}f_{X,Y}(x,y)dydx + \int_{-\infty}^{\infty}yf_Y(y)dy$$

$$= \int_{-\infty}^{\infty}xf_X(x)dx + \mathrm{E}(Y) \quad = \mathrm{E}(X) + \mathrm{E}(Y)$$

Keskväärtuse (triviaalsed) omadused

Olgu *a*, *b* konstandid ja *X* ja *Y* juhuslikud suurused. Siis

1. E(*a* + *b X*) = *a* + *b* E(*X*)
2. E(X + Y) = E(X) + E(Y)
3. E(*X Y*) = E(*X*) E(*Y*), kui *X* ja *Y* on sõltumatud

$$\mathrm{E}(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xyf_X(x)f_Y(y)dxdy$$

$$= \int_{-\infty}^{\infty}yf_Y(y)\int_{-\infty}^{\infty}xf_X(x)dxdy \ = \int_{-\infty}^{\infty}yf_Y(y)\mathrm{E}(X)dy$$

$$= \mathrm{E}(Y)\mathrm{E}(X)$$

Lemma

Kui $X_i, \ i = 1, 2, \ldots, n$ on sõltumatud juhuslikud suurused ja $f_i, \ i = 1, \ldots, n$ on sellised ühe muutuja funktsioonid, et $f_i(X_i)$ on lõplikku keskväärtust omavad juhuslikud suurused, siis

$$E[f_1(X_1) \cdot f_2(X_2) \cdots f_n(X_n)] = E[f_1(X_1)]E[f_2(X_2)] \cdots E[f_n(X_n)].$$

Tõestus. Olgu X diskreetne juhuslik suurus võimalike väärtustega $\{x_i, \ i \in I\}$ ning Y diskreetne juhuslik suurus võimalike väärtustega $\{y_j, \ j \in J\}$, olgu nende ühisjaotus $(x_i, y_j, p_{ij}), \ i \in I, \ j \in J$. Siis

Definitsioon

Keskväärtust omava diskreetse juhusliku suuruse X tinglikuks keskväärtuseks tingimusel, et sündmus $\{Y = y_j\}$ (mille korral $P(\{Y = y_j\}) > 0$) toimus, nimetatakse arvu

$$E(X \mid \{Y = y_j\}) = \sum_{i \in I} x_i \Pr(\{X = x_i\} \mid \{Y = y_j\}),$$

kus $(x_i, \Pr(\{X = x_i\} \mid \{Y = y_j\}))$, $i \in I$ on juhusliku suuruse X tinglik jaotus tingimusel, et $\{Y = y_j\}$ toimus.

Tuletame meelde, et kehtib seos:

$$\Pr(\{X = x_i\} \mid \{Y = y_j\})) = \frac{\Pr(\{X = x_i\}, \{Y = y_j\}))}{\Pr(\{Y = y_j\}))}.$$

1. Millega võrdub summa Y keskväärtus (teiste sõnadega kahel täringul oodatav silmade arvude summa) juhul, kui on teada, et esimesel tuli 2 silma?

$$E(Y|\{X_1 = 2\}) = \sum_j y_j P(\{Y = y_j\}|\{X_1 = 2\}) =$$

$$= 2 \cdot 0 + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + \quad 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} + 7 \cdot \frac{1}{6} + 8 \cdot \frac{1}{6} + 9 \cdot 0 + \ldots\ 12 \ \cdot 0 = 11/2.$$

2. Millega võrdub esimese täringu keskmine saadud silmade arv juhul kui on teada, et summa tuli 5?

$$E(X_1|\{Y = 5\}) = \sum_i x_i P(\{X_1 = x_i\}|\{Y = 5\})$$

$$= \sum_i x_i \cdot \frac{P(\{X_1 = x_i\} \cap \{Y = 5\})}{P\{Y = 5\}} = \sum_{x=1}^{4} x \frac{1/36}{4/36} = \frac{5}{2}.$$

Teoreem

Olgu B_j, $j \in J$ (kus J on kas lõplik või loenduv hulk) sündmuste täissüsteem. Siis kehtib võrdus

$$E(X) = \sum_{j \in J} E(X \mid B_j) P(B_j).$$

Tavaliselt on sobivateks sündmusteks B_j mingi teise juhusliku suuruse Y väärtustele vastavad sündmused $\{Y = y_j\}$.

Teoreem

Olgu B_j, $j \in J$ (kus J on kas lõplik või loenduv hulk) sündmuste täissüsteem. Siis kehtib võrdus

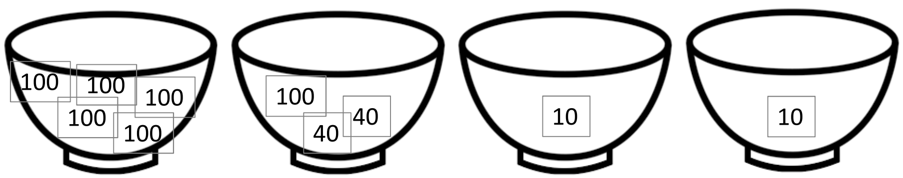
$$E(X) = \sum_{j \in J} E(X \mid B_j) P(B_j).$$

Tavaliselt on sobivateks sündmusteks B_j mingi teise juhusliku suuruse Y väärtustele vastavad sündmused $\{Y = y_j\}$.

$$\sum_{j \in J} \underbrace{\mathrm{E}(X|B_j)}_{\sum_i x_i P(X = x_i|B_j)} P(B_j) = \sum_{j \in J} \sum_i x_i P(X = x_i|B_j) P(B_j)$$

$$= \sum_i x_i \sum_{j \in J} P(X = x_i|B_j) P(B_j) = \sum_i x_i P(X = x_i)$$

Nelja anuma vahel on ära jagatud 10 võitu (keskmine võidu suurus 70 EUR). Seotud silmadega mängijal lastakse juhuslikult välja valida üks anum ja anumas olevatest võitudest üks konkreetne võit. Juhuslik suurus *X* tähistab mänguri poolt saadud võidu suurust. Leia E(*X*)!



Definitsioon

Olgu $(X,\ Y)$ *kahemõõtmeline pidev juhuslik vektor. Siis juhusliku suuruse* X *tinglik keskväärtus tingimusel* $\{Y = y\}$ *on*

$$E\left(X \mid \{Y = y\}\right) = \int_{-\infty}^{+\infty} x \cdot f_X\left(x \mid \{Y = y\}\right) dx,$$

kus

$$f_X\left(x \mid \{Y = y\}\right) = \frac{f(x,y)}{f_Y(y)}.$$

Näide: Olgu juhusliku suuruse $(X,\ Y)$ ühistihedus on antud järgmise valemiga:

$$f(x,y) = \begin{cases} 8xy, & \text{kui } 0 < x < y < 1, \\ 0, & \text{vastasel juhul.} \end{cases}$$

Leida $E(X \mid \{Y = y\})$.

- Paneme tähele, et $E(X|\{Y = y\})$ on muutuja y funktsioon. Tähistame seda funktsiooni korraks $h(y)$. Eelmises näites $h(y) = \frac{2y}{3}$.
- Defineerime keskväärtust $E(X \mid Y)$ kui juhusliku suurust $h(Y)$. Eelmises näiteks oleks $h(Y) = \frac{2Y}{3}$.
- Teisiti öeldes, $E(X|\{Y = y\})$ on mingi arv ühe konkreetse y väärtuse korral. Näiteks, kui eelmises näites võtta $y = 1/2$, siis $E(X|\{Y = 1/2\}) = \frac{2 \cdot 0.5}{3} = 1/3$.
- Samal ajal avaldis $E(X \mid Y)$ on juhuslik suurus, mis sõltub juhuslikust suurusest Y.
- Kuna $E(X \mid Y)$ on juhuslik suurus, siis saame rääkida selle keskväärtusest $E\left(E(X \mid Y)\right)$.
- Sisemine keskväärtus on võetud arvestades juhusliku suuruse X jaotust tingimusel $\{Y = y\}$. Välimine keskväärtus on aga võetud arvestades Y jaotust.

Lemma

Tingliku keskväärtuse kohta kehtib järgmine tulemus:

$$\mathrm{E}(\mathrm{E}(X|Y)) = \mathrm{E}(X)$$

Tõestus pidevate j.s. korral:

$$\mathrm{E}(X|Y=y) = \int_{-\infty}^{\infty} x f_{X|Y}(x|Y=y)\, dx = \int_{-\infty}^{\infty} x \frac{f_{X,Y}(x,y)}{f_Y(y)}\, dx$$

$$\mathrm{E}\left(\mathrm{E}(X|Y)\right) = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} x \frac{f_{X,Y}(x,y)}{\cancel{f_Y(y)}}\, dx \right) \cancel{f_Y(y)}\, dy$$

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

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