Poglavje 9

Osnove matematične statistike

9.0 Uvod

V dosedanjih poglavjih smo izdelali način opisa naključnih dogodkov, spremenljivk in funkcij naključnih spremenljivk s pomočjo verjetnosti.

Verjetnost za nastop naključnega dogodka smo opisali s porazdelitveno funkcijo verjetnosti.

Porazdelitveno funkcijo lahko določimo:

1) analitično na osnovi predpostavk o lastnostih dogodkov
2) empirično na podlagi poskusov

Problem na katerega naletimo pri opisu naključnega pojava na osnovi poskusov je, omejenost števila poskusov.

Osnovna naloga statistike je sklepanje o statističnih lastnosti (o porazdelitveni funkciji, momentih) naključne spremenljivke X na osnovi omejenega števila poskusov.

9.1 Osnovni pojmi

Množico vseh možnih izidov poskusa, ki ustreza celotnemu vzorčnemu prostoru S imenujemo populacija.

$\Rightarrow$

Populacijo predstavimo z naključno spremenljivko X.

V večini praktičnih primerov je nemogoče izvesti meritev na celotni populaciji X.

Zadovoljimo se z izbrano podmnožico n vrednosti meritev x_i iz populacije X oziroma vzorcem $\boldsymbol{v}$.

$$\boldsymbol{v} = (x_1, x_2, \ldots, x_n)$$

Pri tem n imenujemo velikost

ali razsežnost vzorca $\boldsymbol{v}$.

Primer: kvaliteta izdelka v serijski proizvodnji.

1- dober, 0-slab:

$$\boldsymbol{v} = (x_1, x_2, \ldots, x_n) = (1,1,0.....1)$$

Ob ponovitvi izbire vzorca $\boldsymbol{v}$ v splošnem dobimo različne realizacije:

$$\begin{aligned} \boldsymbol{v}_1 &= (x_{11}, x_{12}, \ldots, x_{1n}) \\ \boldsymbol{v}_k &= (x_{k1}, x_{k2}, \ldots, x_{kn}) \\ &\ldots\ldots \\ \boldsymbol{v}_m &= (x_{m1}, x_{m2}, \ldots, x_{mn}) \end{aligned}$$

Zato vzorec v splošnem predstavlja naključno spremenljivko:

$$\boldsymbol{V} = (X_1, X_2, \ldots, X_n)$$

V primeru, ko so posamezne naključne spremenljivke X_i , ki nastopajo v vzorcu $\boldsymbol{V}$ medsebojno statistično neodvisne in imajo isto porazdelitev $f_X(x)$, predstavlja vzorec

$$\boldsymbol{V} = (X_1, X_2, \ldots, X_n)$$

naključni vzorec.

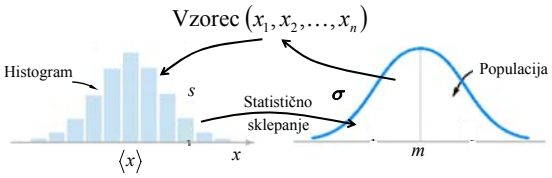
$\Rightarrow$

V praktičnih primerih moramo zagotoviti naključnost vzorca!!

Osnovni naloga statistike je na osnovi izbranega naključnega vzorca

$$\boldsymbol{V} = (X_1, X_2, \ldots, X_n)$$

sklepati na statistične lastnosti obravnavane populacije X.



V ta namen vpeljemo vzorčne karakteristike, ki jih imenujemo statistike.

Primeri statistik:

Vzorčno povprečje:

$$\langle X \rangle_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Vzorčna varianca:

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \langle X \rangle_n\right)^2$$

Popravljena vzorčna varianca:

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \langle X \rangle_n\right)^2$$

Razpon vzorca:

$$\Delta X_n = X_{\max,n} - X_{\min.n}$$

Vzorčna relativna frekvenca dogodka A :

$$p_n(A) = \frac{n(X_i \in A)}{n}$$

Vzorčna porazdelitvena funkcija:

$$F_n(x) = \frac{n(X_i \le x)}{n}$$

V splošnem predstavlja statistiko $\boldsymbol{Z}$ poljubna merljiva funkcija $\boldsymbol{Z}$ opredeljena na naključnem vzorcu $\boldsymbol{V}$.

$$Z = Z(\boldsymbol{V}) = Z(X_1, X_2, \ldots, X_n)$$

$\Rightarrow$

Statistika $\boldsymbol{Z}$ je naključna spremenljivka.

Kar pomeni, da ima $\boldsymbol{Z}$ zalogo vrednosti in neko porazdelitev verjetnosti $F_Z(z)$.

Statistike, ki se uporabljajo za oceno parametrov θ populacije X na osnovi naključnega vzorca $\boldsymbol{V}$ imenujemo tudi cenilke.

Primeri parametrov, ki jih najpogosteje ocenjujemo so $\mathrm{E}(X)$, $\mathrm{Var}(X)$, ….

V splošnem cenilko parametra θ populacije X označimo z $\hat{\Theta}$ vrednost cenilke parametra populacije X pa z $\hat{\boldsymbol{\theta}}$.

Da neko statistiko $Z = \hat{\Theta}$ uporabimo kot cenilko za parametra θ morajo vrednosti statistike $\hat{\boldsymbol{\theta}} = \hat{\Theta}(V)$ imeti lastnost:

$$\lim_{n\to\infty} P\left[\left|\hat{\theta} - \theta\right| < C\right] = 1$$

kjer je C izbrana pozitivna konstanta.

Pri tem sta vrednost cenilke $\hat{\boldsymbol{\theta}}$ naključna in parameter θ deterministična količina.

Splošne lastnosti cenilk

1) Doslednost cenilke

Statistika $\hat{\Theta}$ je dosledna cenilka parametra θ, če za poljubno majhen ε velja:

$$\lim_{n\to\infty} P\left[\left|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right| < \boldsymbol{\varepsilon}\right] = 1$$

Kar pomeni, da vrednost $\hat{\boldsymbol{\theta}}$ cenilke $\hat{\Theta}$ v smislu verjetnosti konvergira k parametru θ.

2) Nepristranost cenilke

Statistika $\hat{\Theta}$ je nepristranska ali centrirana cenilka parametra θ, če je njeno statistično povprečje:

$$\mathrm{E}\left[\hat{\theta}\right] = \boldsymbol{\theta}$$

3) Asimptotska nepristranost cenilke

Statistika $\hat{\Theta}$ je asimptotsko nepristranska cenilka parametra θ če velja:

$$\lim_{n\to\infty} \mathrm{E}\left[\hat{\theta}\right] = \lim_{n\to\infty} \left(\hat{\boldsymbol{\theta}} + O(n)\right) = \boldsymbol{\theta}$$

9.2 Lastnosti pomembnejših statistik

9.2.1 Vzorčna relativna frekvenca

Z A označimo dogodek, da smo pri merjenju naključen spremenljivke X izmerili vrednost , ki leži v intervalu S_A.

$$A = (x_i \in S_A)$$

Izide n ponovitev poskusa zabeležimo z vzorcem

$$\boldsymbol{V} = (X_1, X_2, \ldots, X_n)$$

z n_A označimo število izidov dogodka A.

Vzorčno relativno frekvenca dogodka A opredelimo s statistiko:

$$p_n(A) = \frac{n(X_i \in A)}{n}$$

Nastop dogodka A v vzorcu V obravnavamo kot izid Bernoullijevega poskusa pri katerem se dogodek A zgodi z verjetnostjo p ali pa se z verjetnostjo 1-p ne zgodi.

$\Rightarrow$

Relativno frekvenco $p_n(A)$ lahko obravnavamo kot skupino n Bernoulijevih poskusov oziroma kot binomsko spremenljivko.

Število nastopov n_A dogodka A ima Bernoullijevo ali binomsko porazdelitev verjetnosti:

$$P(n_A;n)=\binom{n}{n_A}p^{n_A}(1-p)^{n-n_A}$$

z $\mathrm{E}[n_A]=np$ in $\mathrm{Var}(n_A)=np(1-p)$.

$\Rightarrow$

$$\mathrm{E}[p_n(A)]=\mathrm{E}\left[\frac{n_A}{n}\right]=\frac{1}{n}\mathrm{E}[n_A]=p$$

$\Rightarrow$

Relativna frekvenca dogodka A je nepristranska cenilka verjetnosti dogodka A.

Statistični raztros relativne frekvence je podan z (glej primer $Y=aX+b$) :

$$\mathrm{Var}(p_n(A))=\mathrm{Var}\left(\frac{n_A}{n}\right)=\frac{1}{n^2}\mathrm{Var}(n_A)=\frac{p(1-p)}{n}$$

in z $n\to\infty$ pada proti 0 .

$\Rightarrow$

Kakovost ocene verjetnosti p_A nastopa dogodka A s pomočjo relativne frekvence $p_n(A)$ narašča z $n\to\infty$.

Verjetnost, da se ocena verjetnosti nastopa dogodka A (s pomočjo statistike relativne frekvence) loči od verjetnosti nastopa dogodka A za več kot ε podamo z neenačbo Čebiševa:

$$P(|X-\mathrm{E}[X]|\geq\varepsilon)\leq\frac{\mathrm{Var}(X)}{\varepsilon^2}$$

$$P(|p_n(A)-p|\geq\boldsymbol{\varepsilon})\leq\frac{\mathrm{Var}(p_n(A))}{\boldsymbol{\varepsilon}^2}=\frac{p(1-p)}{n\boldsymbol{\varepsilon}^2}$$

$\Rightarrow$

$$\lim_{n\to\infty}P[|p_n(A)-p|\geq\boldsymbol{\varepsilon}]=0$$

Relativna frekvenca je dosledna cenilka verjetnosti.

9.2.2 Vzorčna porazdelitvena funkcija

Na razpolago imamo vzorec meritev:

$$\boldsymbol{V}=(X_1,X_2,\ldots,X_n)$$

Z A označimo meritev pri kateri dobimo rezultat $X_i\leq x$.

Z $n_A(x)=n(X_i<x)$ označimo število izmerjenih vrednosti, ki zadoščajo pogoju za nastop dogodka A na izbranem vzorcu $\boldsymbol{V}$:

Vzorčno porazdelitveno funkcijo definiramo z:

$$F_n(X)=\frac{n(X_i\leq x)}{n}$$

Vzorčna porazdelitveno funkcijo uporabimo jo kot cenilko zbirne porazdelitve verjetnosti:

$$F(X)=P(X\leq x)$$

Iz definicije $F_n(X)$ in ugotovljenih lastnosti relativne frekvence lahko pokažemo, da je:

$$F_n(X)=\frac{n(X_i\leq x)}{n}$$

nepristranska in dosledna cenilka $F(X)$.

V praktičnih primerih nas pogosto zanima vrednost naključne spremenljivke x_p pri kateri velja:

$$F_n(X)=\frac{n(X_i\leq x_p)}{n}=p$$

Vrednost x_p imenujemo *fraktil*.

Najpogosteje uporabljamo naslednje fraktile:

mediana: $x_{1/2}$
kvartili : $x_{1/4}$, $x_{1/2}$, $x_{3/4}$
decili: $x_{1/10},\ldots\ldots,x_{9/10}$
centili: $x_{1/100},\ldots\ldots,x_{99/100}$

Vzorčna porazdelitev:

zalogo vrednosti spremenljivke X razdelimo na k intervalov.

Posamezni interval opišemo s *centralno vrednostjo* x_j in širino intervala Δx_j:

$$x_j - \Delta x_j / 2 \le X \le x_j + \Delta x_j / 2 \quad j = 1,....,k$$

Z n_j označimo število vzorčnih vrednosti, ki ležijo v j-tem intervalu.

Z $p_{n,j} = n_j / n$ označimo *intervalne relativne frekvence* .

Z množico intervalnih relativnih frekvenc:

$$\{ p_{n,j} ; \ j=1,....k \}$$

je podana vzorčna porazdelitev.

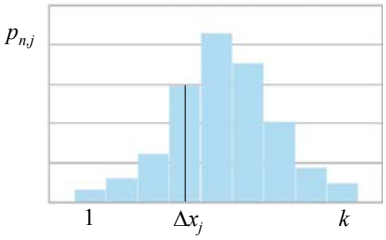
Intervalne relativne frekvence zadoščajo pogoju:

$$\sum_{j=1}^{k} p_{n,j} = 1$$

$\Rightarrow$ Intervalne relativne frekvence so nepristranske in dosledne cenilke verjetnosti, ki pripada izbranemu intervalu:

$$p_j = \int_{x_j - \Delta x_j}^{x_j + \Delta x_j} \mathrm{d}\, P(x)$$

Intervalne frekvence $\{ p_{n,j} ; \ j=1,....k \}$ grafično predstavimo s histogramom:



Razdelitev na intervale je poljubna vendar po priporočilih je število intervalov k med 10 in 100.

9.2.3 Vzorčno povprečje

Iz populacije X izberemo vzorec razsežnosti n:

$$\boldsymbol{V} = (X_1, X_2, ..., X_n)$$

in opredelimo vzorčno povprečje z izrazom:

$$\langle X \rangle_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

$\langle X \rangle_n$ se v splošnem spreminja od vzorca do vzorca

$\Rightarrow \langle X \rangle_n$ je ***naključna spremenljivka***

Opredelimo lahko statistično povprečje vzorčnega povprečja:

$$\mathrm{E}\left[\langle X \rangle_n\right] = \mathrm{E}\left[\frac{1}{n} \sum_{i=1}^{n} X_i\right] = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}[X_i]$$

Komponente vzorca so med seboj neodvisne in imajo enako porazdelitev verjetnosti $\mathrm{E}[X_i] = m$.

$\Rightarrow$

$$\mathrm{E}\left[\langle X \rangle_n\right] = \frac{1}{n} \sum_{i=1}^{n} m = m$$

$\Rightarrow$ Vzorčno povprečje je nepristrana cenilka statističnega povprečja oz. povprečne vrednosti populacije X

Raztros vzorčnega povprečja je podan z:

$$\mathrm{Var}\left(\langle X \rangle_n\right) = \mathrm{E}\left[\left(\langle X \rangle_n - \mathrm{E}\left[\langle X \rangle_n\right]\right)^2\right] = \mathrm{E}\left[\left(\langle X \rangle_n - m\right)^2\right]$$

Komponente vzorca so med seboj neodvisne in imajo enako porazdelitev verjetnosti : (glej primer $Y = aX + b$)

$\Rightarrow$

$$\mathrm{Var}\left[\langle X \rangle_n\right] = \mathrm{Var}\left[\frac{1}{n} \sum_{i=1}^{n} X_i\right]$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}(X) = \frac{\mathrm{Var}(X)}{n} = \frac{\sigma_X^2}{n}$$

Iz neenačbe Čebiševa

$$P(|X-\mathrm{E}[X]|\geq\varepsilon)\leq\frac{\mathrm{Var}(X)}{\varepsilon^2}$$

$$P\left(\left|\langle X\rangle_n-m\right|\geq\boldsymbol{\varepsilon}\right)\leq\frac{\boldsymbol{\sigma}_X^2}{n\boldsymbol{\varepsilon}^2}$$

sledi:

$$\lim_{n\to\infty}P\left(\left|\langle X\rangle_n-m\right|\geq\boldsymbol{\varepsilon}\right)\leq\frac{\boldsymbol{\sigma}_X^2}{n\boldsymbol{\varepsilon}^2}=0$$

$$\lim_{\mathrm{n}\to\infty}P\left(\langle X\rangle_n=m\right)=1$$

$\Rightarrow$

Vzorčno povprečje je dosledna cenilka statističnega povprečja m populacije X.

9.2.4 Vzorčni momenti

Iz populacije X izberemo vzorec razsežnosti n:

$$\boldsymbol{V}=(X_1,X_2,\ldots,X_n)$$

in opredelimo vzorčno varianco:

$$s^2=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\langle X\rangle_n\right)^2$$

in popravljeno vzorčno varianco:

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\langle X\rangle_n\right)^2$$

Statistični povprečji opredeljenih varianc sta podani z:

$$\mathrm{E}[s^2]=\frac{n-1}{n}\boldsymbol{\sigma}^2=\boldsymbol{\sigma}^2+O\left(\frac{1}{n}\right)$$

$$\mathrm{E}[S^2]=\boldsymbol{\sigma}^2$$

$\Rightarrow$

s^2 je asimptotsko nepristranska med tem ko je S^2 nepristranska cenilka variance populacije X.

Varianca nepristranske cenilke S^2 je:

$$\mathrm{Var}(S^2)=\frac{1}{n}\left(\boldsymbol{\mu}^4-\frac{n-3}{n-1}\boldsymbol{\sigma}^2\right)$$

$$\lim_{n\to\infty}\mathrm{Var}(S^2)=0$$

$\Rightarrow$

S^2 je dosledna cenilka variance populacije X.

V splošnem definiramo cenilke vzorčne momentov z izrazom:

$$m_{k,n}=\langle X^k\rangle_n=\frac{1}{n}\sum_{i=1}^{n}X_i^k,\quad k=3,4,\ldots.$$

Cenilke centralnih momentov pa z:

$$\boldsymbol{\mu}_{k,n}=\left\langle\left(X-\langle X\rangle_n\right)^k\right\rangle_n=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\langle X\rangle_n\right)^k$$

Z upoštevanjem neodvisnosti in enakih porazdelitev spremenljivk x_i dobimo:

$$\mathrm{E}[m_{k,n}]=m_k$$

$$\mathrm{Var}(m_{k,n})=\frac{m_{2r}-m_r^2}{n}$$

$$\mathrm{E}[\boldsymbol{\mu}_{k,n}]=\boldsymbol{\mu}_k+O(1/n)$$

$\Rightarrow$

$m_{k,n}$ so nepristranske in dosledne cenilke momentov m_k

$\mu_{k,n}$ so asimptotično nepristranske in dosledne cenike centralnih momentov $\mu_{k,}$

9.3 Verjetnostne porazdelitve statistik

Statistike so funkcije naključnega vzorca $\boldsymbol{V}$:

$$Z = Z(\boldsymbol{V}) = Z(X_1, X_2, ..., X_n)$$

$\Rightarrow$

Statistika je naključna spremenljivka, ki ima ustrezno zalogo vrednosti in porazdelitev verjetnosti oziroma gostoto verjetnosti:

$$f_Z(z)$$

Določitev $f_Z(z)$ ni splošno rešljiv problem ker je rešitev odvisna od funkcije Z.

V primeru, ko je poznana $f_X(x)$ in je statistika opredeljena z funkcijo vsote (glej primer $Z=X+Y$):

$$Z = Z(X) = \sum_{i=1}^{n} Z(X_i)$$

statistično neodvisnih komponent vzorca:

$$\boldsymbol{V} = (X_1, X_2, ..., X_n)$$

je porazdelitev $f_Z(z)$ podana z $n-1$ kratno konvolucijo gostot porazdelitev verjetnosti posameznih komponent.

Pri tem je potrebno predhodno določiti gostoto verjetnosti za funkcijo $Z(X)$, ki je v primeru monotone funkcije podana z:

$$f_Z(z) = f_X(x(z)) \left| \frac{\mathrm{d}\, x(z)}{\mathrm{d}\, z} \right|$$

9.3.1 Porazdelitev statistike vzorčnega povprečja

Vzorčno povprečje naključnega vzorca smo opredelili z:

$$Z_n = \langle X \rangle_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Pri tem so X_i statistično neodvisne in imajo enako gostoto porazdelitve:

$$f_{X_i}(x) = f_{X_j}(x)$$

V primeru, ko je X normalno porazdeljena z:

$$f_{X_i}(x) = f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad \text{in } m_X = 0,\ \sigma_X{}^2 = 1$$

Je za $n = 2$, gostota verjetnosti Z podana s konvolucijo:

$$f_{Z_2}(z) = \int_{-\infty}^{\infty} f_{X_1}(x_1) f_{X_2}(nz - x_1)\, \mathrm{d}\, x_1$$

$$= \frac{1}{\sqrt{2\pi}\sqrt{2}} e^{-\frac{z^2}{4}}$$

$f_{Z_2}(z)$ je normalna z $m_{Z_2} = m_X = 0$ in $\sigma_{Z_2} = \sigma_X / \sqrt{2}$

V splošnem lahko pokažemo, če ima populacija X normalno porazdelitev z:

$$E[X] = m \quad \text{in} \quad \mathrm{Var}(X) = \sigma_X^2$$

Ima statistika vzorčnega povprečja:

$$Z_n = \langle X \rangle_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

normalno porazdelitev z :

$$E\left[\langle X \rangle_n\right] = \frac{1}{n} \sum_{i=1}^{n} m = m \quad \text{in} \quad \mathrm{Var}\left(\langle X \rangle_n\right) = \frac{1}{n^2} \sum_{i=1}^{n} \sigma_X^2 = \frac{\sigma_X^2}{n}$$

Nadalje, za vzorčno povprečje poljubne populacije velja **centralni limitni teorem**, ki pravi:

Če vzorčimo iz populacije X z neznano porazdelitvijo verjetnosti bo porazdelitev vzorčnega povprečja $\langle X \rangle_n$ približno normalna z srednjo vrednostjo in varianco:

$$E[X] = m \quad \text{in} \quad \mathrm{Var}(X) = \sigma_X^2 / n$$

Oziroma če je naključni vzorec

$$\boldsymbol{V} = (X_1, X_2, ..., X_n)$$

velikosti n izbran iz poljubne populacije X z:

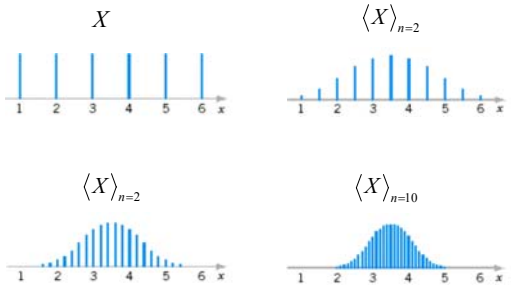
$$E[X] = m \quad \text{in} \quad \mathrm{Var}(X) = \boldsymbol{\sigma}_X^2$$

in če je $\langle X \rangle_n$ vzorčno povprečje potem porazdelitev statistike:

$$Z = \frac{\langle X \rangle_n - m}{\boldsymbol{\sigma} / \sqrt{n}}$$

z $n \to \infty$ limitira k standardni normalni porazdelitvi.

Primer: Povprečje meta n kock.



Primer: Populacija X ima enakomerno zvezno porazdelitev:

$$f_X(x) = \begin{cases} 1/2, & x \in [4,6] \\ 0, & \text{izven} \end{cases}$$

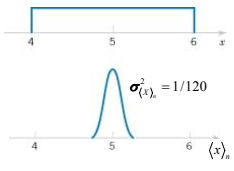
Sprašujemo po porazdelitvi vzorčnega povprečja velikosti vzorca n=40:

$$\mathrm{E}[X] = m = 5 \quad \text{in} \quad \mathrm{Var}(X) = \frac{(b-a)^2}{12} = 1/3$$

Po centralnem limitnem teoremu je porazdelitve $\langle X \rangle_n$ aproksimirana z normalno porazdelitvijo z:

$$\mathrm{E}[\langle X \rangle_n] = \mathrm{E}[X] = m = 5$$

$$\mathrm{Var}(\langle X \rangle_n) = \boldsymbol{\sigma}^2_{\langle X \rangle_n} = \frac{\boldsymbol{\sigma}_X^2}{n} = \frac{1}{3 \cdot 40} = \frac{1}{120}$$



9.3.2 χ^2 Hi-kvadrat porazdelitev

Imamo populacijo X z normalno porazdelitvijo, ki ima:

$$E[X] = m = 0 \quad \text{in} \quad \mathrm{Var}(X) = \boldsymbol{\sigma}_X^2 = 1$$

Zanima nas gostota porazdelitve naključne spremenljivke oziroma statistike Z_n opredeljena z:

$$Z_n = \sum_{i=1}^{n} X_i^2$$

Ker so X_i in s tem X_i^2 statistično neodvisne je porazdelitev vsote $\sum_{i=1}^{n} X_i^2$ podana s konvolucijo posameznih gostot porazdelitve $f_{X_i^2}(x)$:

Porazdelitve so enake:

$$f_{X_1^2}(x) = f_{X_i^2}(x) = f_{X^2}(x) = \quad i = 3,\ldots,n$$

Ker poznamo gostoto porazdelitve X je porazdelitev funkcije naključne spremenljivke $Z=X^2$ podana z:

$$P(Z<z) = P\left(-\sqrt{z} \le X \le \sqrt{z}\right) = \int_{-\sqrt{z}}^{\sqrt{z}} f_X(x)\,\mathrm{d}x$$

Porazdelitvena funkcije naključne spremenljivke $Z=X^2$ podana z:

$$P(Z<z) = P\left(-\sqrt{z} \le X \le \sqrt{z}\right) = \frac{1}{\sqrt{2\pi}} \int_{-\sqrt{z}}^{\sqrt{z}} \mathrm{e}^{-x^2/2}\,\mathrm{d}x$$

Gostoto porazdelitve $f_Z(z)$ dobimo z odvajanjem porazdelitvene funkcije $P(Z<z)$ po spremenljivki z :

$$f_Z(z) = \begin{cases} \frac{1}{\sqrt{2\pi}} \frac{\mathrm{e}^{-z/2}}{\sqrt{z}}, & \text{za } z > 0 \\ 0, & \text{za } z \le 0 \end{cases}$$

Z poznavanjem gostote verjetnosti $f_Z(z)$ in uporabe ***konvolucije*** lahko določimo gostoto verjetnosti vsote dveh kvadratov Z_2

$$f_{Z_2}(z) = \int f_Z(z) f_Z(z-x)\,\mathrm{d}x$$

$$= \frac{1}{(2\pi)^{2/2}} \int_0^z \frac{1}{\sqrt{x}} \frac{1}{\sqrt{z-x}} \mathrm{e}^{-x/2} \mathrm{e}^{-(z-x)/2}\,\mathrm{d}x$$

$$= \begin{cases} \frac{1}{2\pi} \mathrm{e}^{-z/2} \int_0^z \frac{\mathrm{d}x}{\sqrt{x(z-x)}}, & \text{za } z > 0 \\ 0, & \text{za } z \le 0 \end{cases}$$

S substitucijo spremenljivke $x=zv$ in $\mathrm{d}x=z\mathrm{d}v$ v zadnjem integralu dobimo:

$$f_{Z_2}(z) = \begin{cases} \frac{1}{2\pi} \mathrm{e}^{-z/2} z^0 \int_0^1 v^{-1/2}(1-v)^{-1/2}\,\mathrm{d}v, & \text{za } z > 0 \\ 0, & \text{za } z \le 0 \end{cases}$$

Gornji integral lahko izrazimo z beta funkcijo :

$$\mathrm{B}(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \int_0^1 z^{a-1}(1-z)^{b-1}\,\mathrm{d}z$$

Kjer je Γ gama funkcija.

Z uporabo *beta* in *gama* funkcije lahko gostoto verjetnosti $f_{Z_2}(z)$ za $z>0$ zapišemo v obliki:

$$f_{Z_2}(z) = \frac{1}{(2\pi)^{2/2}} z^0 e^{-z/2} \mathrm{B}(1/2, 1/2)$$

$$= \frac{1}{(2\pi)^{2/2}} z^0 e^{-z/2} \frac{\Gamma(1/2)\Gamma(1/2)}{\Gamma(1/2+1/2)}$$

Z upoštevanjem, da je $\Gamma(1/2) = \sqrt{\pi}$ dobimo:

$$f_{Z_2}(z) = \begin{cases} \frac{1}{2^{2/2}\Gamma(2/2)} z^0 \mathrm{e}^{-z/2}, & z > 0 \\ 0, & z \le 0 \end{cases}$$

S pomočjo popolne matematične indukcije lahko pokažemo da velja:

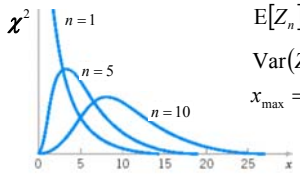
$$f_{Z_n}(z) = \chi^2 = \begin{cases} \frac{1}{2^{n/2}\Gamma(n/2)} z^{\left(\frac{n}{2}-1\right)} \mathrm{e}^{-z/2}, & z > 0 \\ 0, & z \le 0 \end{cases}$$

Dobljena gostota verjetnosti za se imenuje "*hi-kvadrat*" z n prostostnimi stopnjami.

$\Rightarrow$

Naključna spremenljivka opredeljena z vsoto kvadratov standardiziranih normalnih spremenljivk ima *hi-kvadrat* porazdelitev gostote verjetnosti.

Primeri grafov *hi kvadrat* porazdelitve v odvisnosti od n :



$$\mathrm{E}[Z_n] = n$$
$$\mathrm{Var}(Z_n) = 2n$$
$$x_{\max} = n-2, \quad n \geq 2$$

Vrednosti ostalih značilnih parametrov porazdelitve χ^2 :

$$\mathrm{E}[Z_n^r] = m_r = n(n+2)\cdots(n+2r-2)$$
$$g_1 = 2\sqrt{2/n}$$
$$g_2 = 12/n$$

Aditivna lastnost porazdelitve hi-kvadrat

Naj bodo Y_1, Y_2, ..., Y_n, hi-kvadrat neodvisne naključne spremenljivke s prostostnimi stopnjami k_1 , k_2, ..., k_n.

$$Y = Y_1 + Y_2 + \cdots + Y_n$$

je hi-kvadrat naključna spremenljivka z prostostno stopnjo:

$$k = \sum_{i=1}^{n} k_i$$

Primeri spremenljivk z χ^2 porazdelitvijo

1. Poljubni normalni porazdeljeni spremenljivki X z:

$$E[X] = m_X \quad \text{in} \quad \mathrm{Var}(X) = \sigma_X^2$$

lahko priredimo standardni odmik od srednje vrednosti Z :

$$Z = \frac{X - m_X}{\sigma_X}$$

z:

$$E[Z] = m_Z = 0 \quad \text{in} \quad \mathrm{Var}(Z) = \sigma_Z^2 = 1$$

Vzorčni drugi moment spremenljivke Z opredeljen z:

$$m_{2,n} = \frac{1}{n}\sum_{i=1}^{n} Z_i^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - m_X}{\sigma_X}\right)^2 \propto \sum_{i=1}^{n} X_i^2$$

ima χ^2 porazdelitev z n prostostnimi stopnjami.

2. Za poljubno normalno spremenljivki X z:

$$E[X] = m_X \quad \text{in} \quad \mathrm{Var}(X) = \sigma_X^2$$

je z $X_i - \langle X \rangle_n$ podan odmik od vzor. povprečja $\langle X \rangle_n$

Na osnovi odmika vpeljemo spremenljivko:

$$\chi^2_{n-1} = \frac{1}{\sigma_X^2}\sum_{i=1}^{n}\left(X_i - \langle X \rangle_n\right)^2$$

ki ima χ^2 porazdelitve z *n-1* prostostnimi stopnjami.

3. Z uporabo spremenljivke χ^2 lahko izrazimo vzorčni varianci:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \langle X \rangle_n\right)^2 = \frac{\sigma_X^2}{n}\chi^2_{n-1}$$

in

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \langle X \rangle_n\right)^2 = \frac{\sigma_X^2}{n-1}\chi^2_{n-1}$$

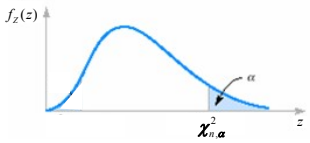
ki imata χ^2 porazdelitve z *n-1* prostostnimi stopnjami.

Vrednosti porazdelitvena funkcija χ^2 so podane tabelarično. Tabela podaja verjetnosti:

$$P\left(X \geq \chi^2_{n,\alpha}\right) = \int_{\chi^2_{n,\alpha}}^{\infty} f_Z(z)\,\mathrm{d}z = \alpha$$

Kjer je $\chi^2_{n,\alpha}$ označuje vrednost *hi-kvadrat* spremenljivke X z n prostostnimi stopnjami, pri kateri je:

$$P\left(X \geq \chi^2_{n,\alpha}\right) = \alpha$$



9.3.3 Studentova porazdelitev *t*

Vzorčimo iz normalne populacije X z:

$$\mathrm{E}[X] = m_X \quad \text{in} \quad \mathrm{Var}(X) = \sigma_X^2$$

Vzorčno povprečje $\langle X \rangle_n$ populacije X ima normalno porazdelitev z:

$$\mathrm{E}\left[\langle X \rangle_n\right] = m_X \quad \text{in} \quad \mathrm{Var}\left(\langle X \rangle_n\right) = \frac{\sigma_X^2}{n}$$

Na osnovi centralnega limitnega teorema ima statistika oziroma naključna spremenljivka:

$$Z = \frac{\left(\langle X \rangle_n - m_X\right)}{\sigma_X / \sqrt{n}}$$

standardno normalno porazdelitev.

Predpostavimo, da variance σ_X populacije X ne poznamo. Kaj se zgodi z porazdelitvijo spremenljivke Z če v njej σ_X nadomestimo z vzorčno varianco S :

$$T = \frac{\left(\langle X \rangle_n - m_X\right)}{S / \sqrt{n}}$$

V splošnem lahko pokažemo, da:

Če je Z normalna spremenljivka in V hi-kvadrat spremenljivka z n prostostnimi stopnjami in če sta Z in V statistično neodvisni, potem ima spremenljivka:

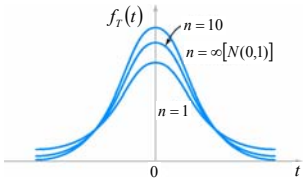
$$T = \frac{Z}{\sqrt{V/n}}$$

gostoto verjetnosti:

$$f_T(t) = \frac{\Gamma\left[(n+1)/2\right]}{\sqrt{\pi n}\,\Gamma(n/2)} \cdot \frac{1}{\left[\left(t^2/n\right)+1\right]^{(n+1)/2}} \qquad -\infty < t < \infty,$$

ki se imenuje *Studentova* ali t_n porazdelitev z n prostostnim stopnjami.

Primeri grafov Studentove gostote verjetnosti:



$$\mathrm{E}[T_n] = 0$$
$$\mathrm{Var}(T_n) = n/(n-2)$$
$$t_{\max} = 0$$

Vrednosti ostalih značilnih parametrov:

$$\mathrm{E}\left[T_n^{2r}\right] = m_{2r} = \frac{1 \cdot 3 \cdots (2r-1)}{(n-2)(n-4)\cdots(n-2r)} n^r$$

$$g_1 = 0$$

$$g_2 = 3\frac{n-2}{n-4} - 3, \quad n > 4$$

V našem primeru imamo spremenljivko:

$$T = \frac{\left(\langle X \rangle_n - m_X\right)}{S/\sqrt{n}} = \frac{\left(\langle X \rangle_n - m_X\right)}{\sqrt{S^2/n}}$$

Kjer je $\langle X \rangle_n - m_X$ normalna in S^2 hi-kvadrat spremenljivka z n-1 prostostnimi stopnjami, ki sta statistično neodvisni.

Zato ima gornja statistika T Studentovo oziroma t porazdelitev z n-1 prostostnimi stopnjami:

$$f_T(t) = \frac{\Gamma\left[(n)/2\right]}{\sqrt{\pi n - 1}\,\Gamma\left((n-1)/2\right)} \cdot \frac{1}{\left[\left(t^2/(n-1)\right)+1\right]^{(n)/2}}$$

Tudi za spremenljivko:

$$T_i = \frac{X_i - m_X}{S}$$

lahko pokažemo, da ima Studentovo porazdelitev z n-1 stopnjami.

Studentova porazdelitev je podana v tabeli, ki podaja:

$$P(T_n \geq t_{n,\alpha}) = \int_{t_{n,\alpha}}^{\infty} f_T(t)\,\mathrm{d}z = \alpha$$