Regresijska funkcija

Preslikavo $x \mapsto \mathsf{E}(Y|x)$ imenujemo *regresija* slučajne spremenljivke Y glede na slučajno spremenljivko X.

Primer: Naj bo $(X, Y) : N(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$.
Tedaj je, kot vemo $p_X(x|y) : N(\mu_x + \rho\frac{\sigma_x}{\sigma_y}(y - \mu_y), \sigma_x\sqrt{1-\rho^2})$.
Torej je pogojno matematično upanje

$$\mathsf{E}(X|y) = \mu_x + \rho\frac{\sigma_x}{\sigma_y}(y - \mu_y)$$

in prirejena spremenljivka

$$\mathsf{E}(X|Y) = \mu_x + \rho\frac{\sigma_x}{\sigma_y}(Y - \mu_y).$$

Na podoben način vpeljemo regresijo slučajne spremenljivke X glede na slučajno spremenljivko Y. Za dvorazsežno normalno porazdelitev dobimo

$$\mathsf{E}(Y|X) = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(X - \mu_x)$$

Obe regresijski funkciji sta **linearni**.

Kovariančna matrika

Matematično upanje slučajnega vektorja $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ je vektor $\mathsf{E}\boldsymbol{X} = (\mathsf{E}X_1, \mathsf{E}X_2, \ldots, \mathsf{E}X_n)$.

Primer: Za $(X, Y) : N(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$ je $\mathsf{E}(X, Y) = (\mu_x, \mu_y)$.

Matematično upanje slučajne spremenljivke Y, ki je linearna kombinacija spremenljivk $X_1, X_2, \ldots, X_n$, je potem

$$\mathsf{E}Y = \mathsf{E}(\textstyle\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} a_i \mathsf{E}X_i$$

Za disperzijo spremenljivke Y pa dobimo $\mathsf{D}Y = \mathsf{E}(Y - \mathsf{E}Y)^2 =$

$$\mathsf{E}\Big(\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j (X_i - \mathsf{E}X_i)(X_j - \mathsf{E}X_j)\Big) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \mathsf{Cov}(X_i, X_j) = \boldsymbol{a}^T \boldsymbol{K} \boldsymbol{a},$$

kjer je $\mathsf{Cov}(X_i, X_j) = \mathsf{E}((X_i - \mathsf{E}X_i)(X_j - \mathsf{E}X_j))$ kovarianca spremenljivk X_i in X_j, $\boldsymbol{K} = [\mathsf{Cov}(X_i, X_j)]$ *kovariančna matrika* vektorja X, ter $\boldsymbol{a} = (a_1, a_2, \ldots, a_n)^T$.

Lastnosti kovariančne matrike

Kovariančna matrika $\boldsymbol{K} = [K_{ij}]$ je *simetrična*: $K_{ij} = K_{ji}$.

Diagonalne vrednosti so disperzije spremenljivk: $K_{ii} = \mathsf{D}X_i$.

Ker je $\boldsymbol{a}^T \boldsymbol{K} \boldsymbol{a} = \mathsf{D}Y \geq 0$, je pozitivno semidefinitna matrika.

Naj bo $\boldsymbol{a}$, $||\boldsymbol{a}|| = 1$ lastni vektor, ki pripada lastni vrednosti λ kovariančne matrike $\boldsymbol{K}$ – velja $\boldsymbol{K}\boldsymbol{a} = \lambda \boldsymbol{a}$. Tedaj je $0 \leq \mathsf{D}Y = \boldsymbol{a}^T \boldsymbol{K} \boldsymbol{a} = \lambda$ – vse lastne vrednosti kovariančne matrike so nenegativne.

Če je kaka lastna vrednost enaka 0, je vsa verjetnost skoncentrirana na neki hiperravnini – porazdelitev je *izrojena*. To se zgodi natanko takrat, ko kovariančna matrika $\boldsymbol{K}$ ni obrnljiva, oziroma ko je $\det \boldsymbol{K} = 0$.

Primer: Za $(X, Y) : N(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$ je $\boldsymbol{K} = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}$.

Ker je $|\rho| < 1$, je $\det \boldsymbol{K} = \sigma_x^2\sigma_y^2(1-\rho^2) > 0$ in je potemtakem porazdelitev vedno neizrojena. Za $N(\boldsymbol{\mu}, \boldsymbol{A})$ je $\boldsymbol{K} = \boldsymbol{A}^{-1}$.

. . . Lastnosti kovariančne matrike

Poglejmo še, kako se spremeni kovariančna matrika pri linearni transformaciji vektorja $X' = \boldsymbol{A}X$, kjer je $\boldsymbol{A}$ poljubna matrika reda $n \times n$.

Vemo, da je $\mathsf{D}(\boldsymbol{a}^T X) = \boldsymbol{a}^T \boldsymbol{K} \boldsymbol{a}$.

Tedaj je, če označimo kovariančno matriko vektorja X' s $\boldsymbol{K}'$,

$$\boldsymbol{a}^T \boldsymbol{K}' \boldsymbol{a} = \mathsf{D}(\boldsymbol{a}^T X') = \mathsf{D}(\boldsymbol{a}^T \boldsymbol{A} X) = \mathsf{D}((\boldsymbol{A}^T \boldsymbol{a})^T X) =$$

$$(\boldsymbol{A}^T \boldsymbol{a})^T \boldsymbol{K} (\boldsymbol{A}^T \boldsymbol{a}) = \boldsymbol{a}^T \boldsymbol{A} \boldsymbol{K} \boldsymbol{A}^T \boldsymbol{a}$$

in potemtakem

$$\boldsymbol{K}' = \boldsymbol{A} \boldsymbol{K} \boldsymbol{A}^T.$$

Višji momenti

Višji momenti so posplošitev pojmov matematičnega upanja in disperzije.

Moment reda $k \in \mathbb{N}$ *glede na točko* $a \in \mathbb{R}$ imenujemo količino

$$m_k(a) = \mathsf{E}((X-a)^k).$$

Moment obstaja, če obstaja matematično upanje $\mathsf{E}(|X-a|^k) < \infty$. Za $a = 0$ dobimo *začetni moment* $z_k = m_k(0)$; za $a = \mathsf{E}X$ pa *centralni moment* $m_k = m_k(\mathsf{E}X)$. Primera: $\mathsf{E}X = z_1$ in $\mathsf{D}X = m_2$.

Če obstaja moment $m_n(a)$, potem obstajajo tudi vsi momenti $m_k(a)$, $k < n$.

Če obstaja moment z_n, obstaja tudi moment $m_n(a)$ za vse $a \in \mathbb{R}$.

$$m_n(a) = \mathsf{E}((X-a)^n) = \sum_{k=0}^{n} \binom{n}{k} (-a)^{n-k} z_k$$

. . . Višji momenti

Posebej za centralni moment velja

$$m_n = m_n(z_1) = \sum_{k=0}^{n} \binom{n}{k} (-z_1)^k z_{n-k}$$

$m_0 = 1, m_1 = 0, m_2 = z_2 - z_1^2, m_3 = z_3 - 3z_2z_1 + 2z_1^3, \ldots$

Asimetrija spremenljivke X imenujemo količino $A(X) = \dfrac{m_3}{\sigma^3}$.

Sploščenost spremenljivke X imenujemo količino $K(X) = \dfrac{m_4}{\sigma^4} - 3$,

kjer je $\sigma = \sqrt{m_2}$.

Za simetrično glede na $z_1 = \mathsf{E}X$ porazdeljene spremenljivke so vsi lihi centralni momenti enaki 0.

. . . Višji momenti

Za $X : N(\mu, \sigma)$ so $m_{2k+1} = 0$ in $m_{2k} = (2k-1)!!\sigma^{2k}$. Zato sta tudi $A(X) = 0$ in $K(X) = 0$.

Če sta spremenljivki X in Y neodvisni, je $m_3(X+Y) = m_3(X) + m_3(Y)$.

Za binomsko porazdeljeno spremenljivko $X : B(n,p)$ je $m_3(X) = npq(q-p)$ in dalje $A(X) = \frac{q-p}{\sqrt{npq}}$.

Kadar spremenljivka nima momentov, uporabljamo kvantile.

Kvantil reda $p \in (0,1)$ je vsaka vrednost $x \in \mathbb{R}$, za katero velja $P(X \le x) \ge p$ in $P(X \ge x) \ge 1-p$ oziroma $F(x) \le p \le F(x+)$. Kvantil reda p označimo z x_p. Za zvezno spremenljivko je $F(x_p) = p$.

Kvantil $x_{\frac{1}{2}}$ imenujemo *mediana*; $x_{\frac{i}{4}}$, $i = 0,1,2,3,4$ so *kvartili*.

Kot nadomestek za standardni odklon uporabljamo *kvartilni razmik* $\frac{1}{2}(x_{\frac{3}{4}} - x_{\frac{1}{4}})$.

Karakteristična funkcija

Karakteristična funkcija realne slučajne spremenljivke X je kompleksna funkcija $\varphi_X(t)$ realne spremenljivke t določena z zvezo $\varphi_X(t) = \mathsf{E}e^{itX}$.

Karakteristične funkcije so močno računsko orodje. Posebej pomembni lastnosti sta:

Če obstaja z_n je karakteristična funkcija n-krat odvedljiva v vsaki točki in velja $\varphi_X^{(k)}(0) = i^k z_k$.

Za neodvisni spremenljivki X in Y je $\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t)$.

Pojem karakteristične funkcije lahko posplošimo tudi na slučajne vektorje.

Limitni izreki

Zaporedje slučajnih spremenljivk X_n *verjetnostno konvergira* k slučajni spremenljivki X, če za vsak $\varepsilon > 0$ velja

$$\lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0$$

ali enakovredno

$$\lim_{n\to\infty} P(|X_n - X| < \varepsilon) = 1$$

Zaporedje slučajnih spremenljivk X_n *skoraj gotovo konvergira* k slučajni spremenljivki X, če velja

$$P(\lim_{n\to\infty} X_n = X) = 1$$

. . . Limitni izreki

Če zaporedje slučajnih spremenljivk X_n skoraj gotovo konvergira k slučajni spremenljivki X, potem za vsak $\varepsilon > 0$ velja

$$\lim_{m\to\infty} P(|X_n - X| < \varepsilon \quad \text{za vsak} \quad n \geq m) = 1$$

Od tu izhaja: če skoraj gotovo $X_n \to X$, potem tudi verjetnostno $X_n \to X$.

Šibki in krepki zakon velikih števil

Naj bo X_k zaporedje spremenljivk, ki imajo matematično upanje. Označimo $S_n = \sum_{k=1}^{n} X_k$ in

$$Y_n = \frac{S_n - \mathsf{E}S_n}{n} = \frac{1}{n}\sum_{k=1}^{n}(X_k - \mathsf{E}X_k) = \frac{1}{n}\sum_{k=1}^{n} X_k - \frac{1}{n}\sum_{k=1}^{n}\mathsf{E}X_k$$

Pravimo, da za zaporedje slučajnih spremenljivk X_k velja:

šibki zakon velikih števil, če verjetnostno $Y_n \to 0$; če za za vsak $\varepsilon > 0$ velja

$$\lim_{n\to\infty} P\left(\left|\frac{S_n - \mathsf{E}S_n}{n}\right| < \varepsilon\right) = 1$$

krepki zakon velikih števil, če skoraj gotovo $Y_n \to 0$; če velja

$$P\left(\lim_{n\to\infty} \frac{S_n - \mathsf{E}S_n}{n} = 0\right) = 1$$

Če za zaporedje X_k velja krepki zakon, velja tudi šibki.

Neenakost Čebiševa

Če ima slučajna spremenljivka X končno disperzijo $\mathsf{D}X < \infty$, velja za vsak $\varepsilon > 0$ *neenakost Čebiševa*

$$P(|X - \mathsf{E}X| \geq \varepsilon) \leq \frac{\mathsf{D}X}{\varepsilon^2}.$$

Pokažimo jo za zvezne spremenljivke

$$P(|X - \mathsf{E}X| \geq \varepsilon) = \int_{|x-\mathsf{E}X|\geq\varepsilon} p(x)dx = \frac{1}{\varepsilon^2} \int_{|x-\mathsf{E}X|\geq\varepsilon} \varepsilon^2 p(x)dx$$

$$\leq \frac{1}{\varepsilon^2} \int_{-\infty}^{\infty} (x - \mathsf{E}X)^2 p(x)dx = \frac{\mathsf{D}X}{\varepsilon^2}.$$

Neenakost Čebiševa – posledice

(**Markov**) Če za zaporedje slučajnih spremenljivk X_i gre izraz $\frac{\mathsf{D}S_n}{n^2}$ proti 0, ko gre $n \to \infty$, velja za zaporedje šibki zakon velikih števil.

(**Čebišev**) Če so slučajne spremenljivke X_i paroma nekorelirane in so vse njihove disperzije omejene z isto konstanto C, $\mathsf{D}X_i < C$, velja za zaporedje šibki zakon velikih števil.

Za Bernoullijevo zaporedje X_i so spremenljivke paroma neodvisne, $\mathsf{D}X_i = pq$, $S_n = k$. Pogoji izreka Čebiševa so izpolnjeni in dobimo:

(**Bernoulli** 1713) Za vsak $\varepsilon > 0$ velja

$$\lim_{n\to\infty} P\left(\left|\frac{k}{n} - p\right| < \varepsilon\right) = 1$$

Še nekaj izrekov

(**Hinčin**) Če so neodvisne slučajne spremenljivke X_i enako porazdeljene in imajo matematično upanje $\mathsf{E}X_i = a$ za vsak i, velja zanje šibki zakon velikih števil. Za vsak $\varepsilon > 0$ je

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - a\right| < \varepsilon\right) = 1$$

(**Kolmogorov**) Če so slučajne spremenljivke X_i neodvisne, imajo končno disperzijo in velja $\sum_{n=1}^{\infty} \frac{\mathsf{D}S_n}{n^2} < \infty$, potem velja krepki zakon velikih števil

$$P\left(\lim_{n\to\infty} \frac{S_n - \mathsf{E}S_n}{n} = 0\right) = 1$$

. . . Še nekaj izrekov

(**Kolmogorov**) Če so slučajne spremenljivke X_i neodvisne, enako porazdeljene in imajo matematično upanje $\mathsf{E}X_i = \mu$, potem velja krepki zakon velikih števil

$$P\left(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mu\right) = 1$$

(**Borel** 1909) Za Bernoullijevo zaporedje velja

$$P\left(\lim_{n\to\infty} \frac{k}{n} = p\right) = 1$$

Centralni limitni zakon

Opazujmo sedaj zaporedje standardiziranih spremenljivk

$$Z_n = \frac{S_n - \mathsf{E}S_n}{\sigma(S_n)}$$

Za zaporedje slučajnih spremenljivk X_i velja *centralni limitni zakon*, če porazdelitvene funkcije za Z_n gredo proti porazdelitveni funkciji standardizirane normalne porazdelitve, to je, če za vsak $x \in \mathbb{R}$ velja

$$\lim_{n\to\infty} P\left(\frac{S_n - \mathsf{E}S_n}{\sigma(S_n)} < x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt$$

(**Osnovni CLI**) Če so slučajne spremenljivke X_i neodvisne, enako porazdeljene s končnim matematičnim upanjem in končno disperzijo, potem zanje velja centralni limitni zakon.

STATISTIKA

Statistika je veda, ki proučuje množične pojave.

Ljudje običajno besedo *statistika* povezujejo z zbiranjem in urejanjem podatkov o nekem pojavu, izračunom raznih značilnosti iz teh podatkov, njih predstavitvijo in razlago. To je najstarejši del statistike in ima svoje začetke že v antiki – z nastankom večjih združb (držav) se je pojavila potreba po poznavanju stanja – 'računovodstvo', astronomija, ... Sama beseda ***statistika*** naj bi izvirala iz latinske besede ***status*** – v pomenu država. Tej veji statistike pravimo *opisna statistika*.

Druga veja, *inferenčna statistika*, poskuša spoznanja iz zbranih podatkov posplošiti (razširiti, podaljšati, napovedati, ...) in oceniti kakovost teh pospošitev.

Statistiko lahko razdelimo tudi na *uporabno* in *teoretično* (matematično in računalniško) statistiko.

Osnovni pojmi

(Statistična) enota – posamezna proučevana stvar ali pojav.
Primer: redni študent na Univerzi v Ljubljani v študijskem letu 1994/95.

Populacija – množica vseh proučevanih enot; pomembna je natančna opredelitev populacije (npr. časovno in prostorsko).
Primer: vsi redni študentje na UL v študijskem letu 1994/95.

Vzorec – podmnožica populacije, na osnovi katere ponavadi sklepamo o lastnostih celotne populacije.
Primer: vzorec 300 slučajno izbranih rednih študentov na UL v l. 1994/95.

Spremenljivka – lastnost enot; označujemo jih npr. z X, Y, X_1.
Vrednost spremenljivke X na i–ti enoti označimo z x_i.
Primer: spol, uspeh iz matematike v zadnjem razredu srednje šole, izobrazba matere in višina mesečnih dohodkov staršev študenta.

. . . Osnovni pojmi

Posamezne spremenljivke in odnose med njimi opisujejo ustrezne porazdelitve.

Parameter – značilnost populacije; običajno jih označujemo z malimi grškimi črkami.

Statistika – značilnost vzorca; običajno jih označujemo z malimi latinskimi črkami. Vrednost statistike je lahko za različne vzorce različna.

Eno izmed osnovnih vprašanj statistike je, kako z uporabo ustreznih statistik oceniti vrednosti izbranih parametrov.

Vrste spremenljivk

Vrste spremenljivk glede na vrsto vrednosti:

1. **opisne** (ali atributivne) spremenljivke – vrednosti lahko opišemo z imeni razredov (npr. poklic, uspeh);
2. **številske** (ali numerične) spremenljivke – vrednosti lahko izrazimo s števili (npr. starost).

. . . Vrste spremenljivk

Vrste spremenljivk glede na vrsto merske lestvice:

1. **imenske** (ali nominalne) spremenljivke – vrednosti lahko le razlikujemo med seboj: dve vrednosti sta enaki ali različni (npr. spol);
2. **urejenostne** (ali ordinalne) spremenljivke – vrednosti lahko uredimo od najmanjše do največje (npr. uspeh);
3. **razmične** (ali intervalne) spremenljivke – lahko primerjamo razlike med vrednostima dvojic enot (npr. temperatura);
4. **razmernostne** spremenljivke – lahko primerjamo razmerja med vrednostima dvojic enot (npr. starost).
5. **absolutne** spremenljivke – štetja (npr. število prebivalcev).

. . . Vrste spremenljivk

dovoljene transformacije	*vrsta lestvice*	*primeri*
$\varphi(x) = x$ (identiteta)	absolutna	štetje
$\varphi(x) = a.x, a > 0$	razmernostna	masa
podobnost		temperatura (K)
$\varphi(x) = a.x + b, a > 0$	razmična	temperatura (C,F)
		čas (koledar)
$x \geq y \Leftrightarrow \varphi(x) \geq \varphi(y)$	urejenostna	šolske ocene, kakovost
strogo naraščajoča		zraka, trdost kamnin
φ je povratno enolična	imenska	barva las, narodnost

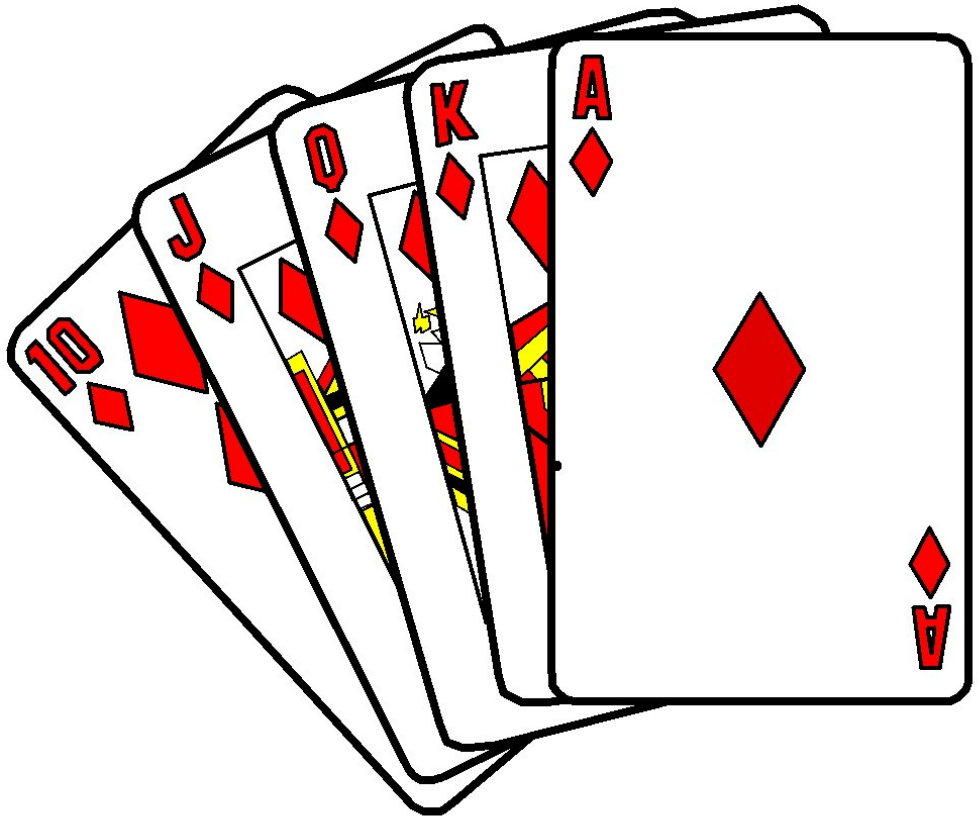
. . . Vrste spremenljivk

Vrste spremenljivk so urejene od tistih z najslabšimi merskimi lastnostmi do tistih z najboljšimi. Urejenostne spremenljivke zadoščajo lastnostim, ki jih imajo imenske spremenljivke; in podobno razmernostne spremenljivke zadoščajo lastnostim, ki jih imajo razmične, urejenostne in imenske spremenljivke.

$$\text{absolutna} \subset \text{razmernostna} \subset \text{razmična} \subset \text{urejenostna} \subset \text{imenska}$$

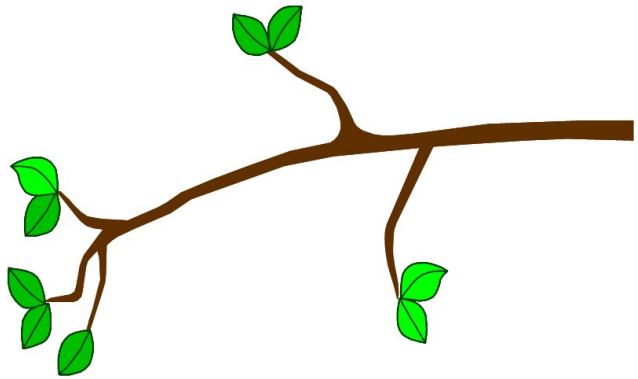
Posamezne statistične metode predpostavljajo določeno vrsto spremenljivk. Največ učinkovitih statističnih metod je razvitih za številske spremenljivke. V teoriji merjenja pravimo, da je nek stavek *smiseln*, če ohranja resničnost/lažnost pri zamenjavi meritev z enakovrednimi (glede na dovoljene transformacije) meritvami.

Vzorčenje



Vzorčenje

Analitična statistika je veja statistike, ki se ukvarja z uporabo vzorčnih podatkov, da bi z njimi naredili zaključek (inferenco) o populaciji.



Zakaj vzorčenje?

- cena
- čas
- destruktivno testiranje

Glavno vprašanje statistike je: kakšen mora biti vzorec, da lahko iz podatkov zbranih na njem veljavno sklepamo o lastnostih celotne populacije.

Kdaj vzorec dobro predstavlja celo populacijo? Preprost odgovor je:

- vzorec mora biti izbran *nepristransko*
- vzorec mora biti *dovolj velik*

. . . Vzorčenje

Recimo, da merimo spremenljivko X, tako da n-krat naključno izberemo neko enoto in na njej izmerimo vrednost spremenljivke X. Postopku ustreza slučajni vektor $(X_1, X_2, X_3, \ldots, X_n)$, ki mu rečemo *vzorec*. Število n je *velikost* vzorca. Ker v vzorcu merimo isto spremenljivko, lahko predpostavimo, da imajo vsi členi X_i vektorja ***isto*** porazdelitev, kot spremenljivka X. Ker posamezna meritev ne sme vplivati na ostale, lahko predpostavimo še, da so členi X_i med seboj ***neodvisni***. Takemu vzorcu rečemo *enostavni slučajni vzorec*. Večina statistične teorije temelji na predpostavki, da imamo opravka enostavnim slučajnim vzorcem.

Če je populacija končna, lahko dobimo enostavni slučajni vzorec, tako da slučajno izbiramo (z vračanjem) enote z enako verjetnostjo.

Z vprašanjem, kako sestaviti dobre vzorce v praksi, se ukvarja posebno področje statistike – *teorija vzorčenja*.

Načini vzorčenja

- ocena
  - priročnost
- naključno
  - enostavno: pri enostavnem naključnem vzorčenju je vsak član populacije izbran/vključen z *enako verjetnostjo.*
  - deljeno: razdeljen naključni vzorec dobimo tako, da razdelimo populacijo na disjunktne množice oziroma dele (razrede) in nato izberemo enostavne naključne vzorce za vsak del posebej.
  - grozdno: takšno vzorčenje je enostavno naključno vzorčenje skupin ali klastrov/grozdov elementov.

Osnovni izrek statistike

Spremenljivka X ima na populaciji G porazdelitev $F(x) = P(X < x)$. Toda tudi vsakemu vzorcu ustreza neka porazdelitev.

Za realizacijo vzorca $(x_1, x_2, x_3, \ldots, x_n)$ in $x \in \mathbb{R}$ postavimo $K(x) = |\{x_i : x_i < x, i = 1, \ldots, n\}|$ in $V_n(x) = K(x)/n$. Slučajni spremenljivki $V_n(x)$ *pravimo vzorčna porazdelitvena funkcija*. Ker ima, tako kot tudi $K(x)$, $n + 1$ možnih vrednosti $k/n,\ k = 0, \ldots, n$, je njena verjetnostna funkcija $B(n, F(x))$

$$P(V_n(x) = k/n) = \binom{n}{k} F(x)^k (1 - F(x))^{n-k}$$

. . . Osnovni izrek statistike

Če vzamemo n neodvisnih Bernoullijevih spremenljivk

$$Y_i(x) : \begin{pmatrix} 1 & 0 \\ F(x) & 1 - F(x) \end{pmatrix}$$

velja

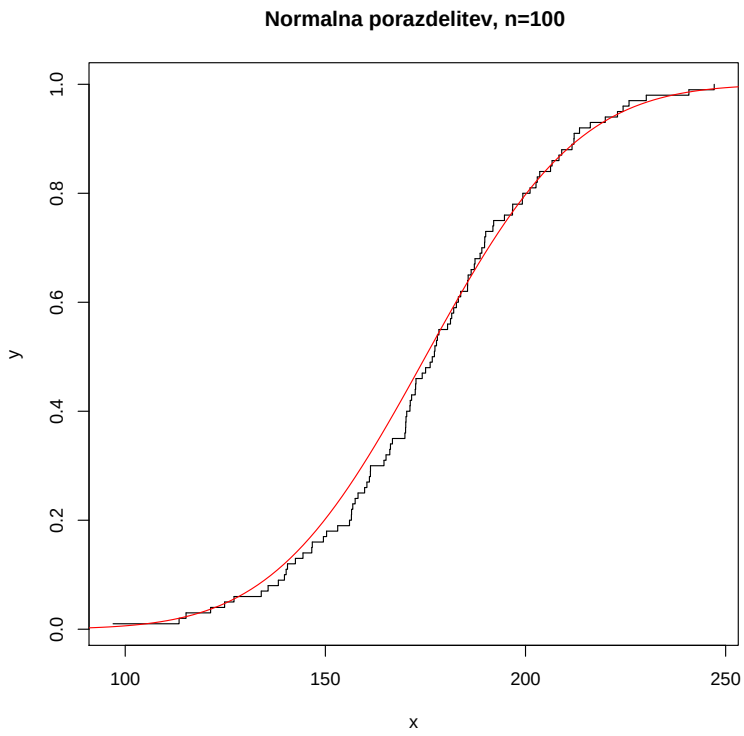
$$V_n(x) = \frac{1}{n} \sum_{i=1}^{n} Y_i(x).$$

Krepki zakon velikih števil tedaj zagotavlja, da za vsa x velja

$$P(\lim_{n \to \infty} V_n(x) = F(x)) = 1$$

To je v bistvu Borelov zakon, da relativna frekvenca dogodka $(X < x)$ skoraj gotovo konvergira proti verjetnosti tega dogodka.

. . . Osnovni izrek statistike



Velja pa še več. $V_n(x)$ je stopničasta funkcija, ki se praviloma dobro prilega funkciji $F(x)$.

Odstopanje med $V_n(x)$ in $F(x)$ lahko izmerimo s slučajno spremenljivko

$$D_n = \sup_{x \in \mathbb{R}} |V_n(x) - F(x)|$$

za $n = 1, 2, 3, \ldots$ Zanjo lahko pokažemo *osnovni izrek statistike*

$$P(\lim_{n \to \infty} D_n = 0) = 1$$

Torej se z rastjo velikosti vzorca $V_n(x)$ enakomerno vse bolje prilega funkciji $F(x)$ – vse bolje povzema razmere na celotni populaciji.

Frekvenčna porazdelitev

Število vseh možnih vrednosti proučevane spremenljivke je lahko preveliko za pregledno prikazovanje podatkov. Zato sorodne vrednosti razvrstimo v skupine. Posamezni skupini priredimo ustrezno reprezentativno vrednost, ki je nova vrednost spremenljivke. Skupine vrednosti morajo biti določene **enolično**: vsaka enota s svojo vrednostjo je lahko uvrščena v natanko eno skupino vrednosti.

Frekvenčna porazdelitev spremenljivke je ***tabela***, ki jo določajo ***vrednosti ali skupine vrednosti*** in njihove ***frekvence***.

Če je spremenljivka vsaj urejenostna, vrednosti (ali skupine vrednosti) uredimo od najmanjše do največje.

Skupine vrednosti številskih spremenljivk imenujemo *razredi*.

. . . Frekvenčna porazdelitev

x_{min} in x_{max} – *najmanjša* in *največja* vrednost spremenljivke X.

$x_{i,min}$ in $x_{i,max}$ – *spodnja* in *zgornja meja* i-tega razreda.

Meje razredov so določene tako, da velja $x_{i,max} = x_{i+1,min}$.

Širina i-tega razreda je $d_i = x_{i,max} - x_{i,min}$. Če je le mogoče, vrednosti razvrstimo v razrede enake širine.

Sredina i-tega razreda je $x_i = \frac{x_{i,min}+x_{i,max}}{2}$ in je značilna vrednost – predstavnik tega razreda.

Kumulativa (ali nakopičena frekvenca) je frekvenca do spodnje meje določenega razreda. Velja $F_{i+1} = F_i + f_i$, kjer je F_i kumulativa in f_i frekvenca v i-tem razredu.

Slikovni prikazi

Stolpčni prikaz: Na eni osi prikažemo (urejene) razrede. Nad vsakim naredimo stolpec/črto višine sorazmerne frekvenci razreda.

Krožni prikaz: Vsakemu razredu priredimo krožni izsek s kotom $\alpha_i = \frac{f_i}{n} 360$ stopinj.

Histogram: drug poleg drugega rišemo stolpce – pravokotnike, katerih ploščina je sorazmerna frekvenci v razredu. Če so razredi enako široki, je višina sorazmerna tudi frekvenci.

Poligon: v koordinatnem sistemu zaznamujemo točke (x_i, f_i), kjer je x_i sredina i–tega razreda in f_i njegova frekvenca. K tem točkam dodamo še točki $(x_0, 0)$ in $(x_{k+1}, 0)$, če je v frekvenčni porazdelitvi k razredov. Točke zvežemo z daljicami.

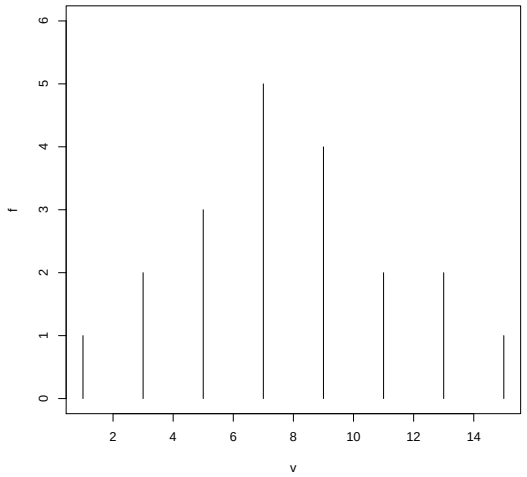
Ogiva: grafična predstavitev kumulativne frekvenčne porazdelitve s poligonom, kjer v koordinatni sistem nanašamo točke $(x_{i,min}, F_i)$.

Nekaj ukazov v R-ju

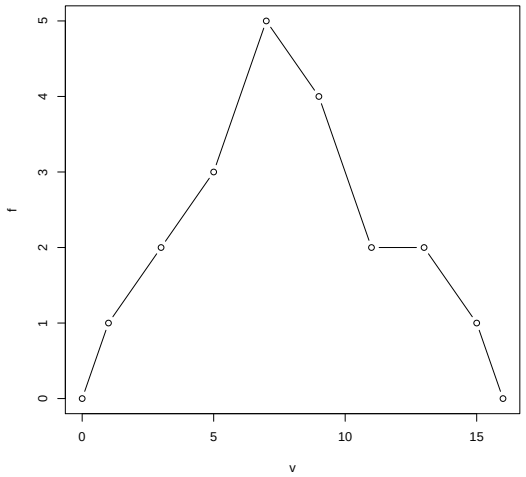
```
> X <- c(5,11,3,7,5,7,15,1,13,11,9,9,3,13,9,7,7,5,9,7)
> n <- length(X)
> t <- tabulate(X)
> t
 [1] 1 0 2 0 3 0 5 0 4 0 2 0 2 0 1
> v <- (1:max(X))[t>0]
> f <- t[t>0]
> rbind(v,f)
  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
v    1    3    5    7    9   11   13   15
f    1    2    3    5    4    2    2    1
> plot(v,f,type='h')
> plot(c(0,v,16),c(0,f,0),type='b',xlab='v',ylab='f')
> pie(f,v)
> plot(c(0,v,16),c(0,cumsum(f)/n,1),col='red',type='s',
  xlab='v',ylab='f')
> x <- sort(rnorm(100,mean=175,sd=30))
> y <- (1:100)/100
> plot(x,y,main='Normalna porazdelitev, n=100',type='s')
> curve(pnorm(x,mean=175,sd=30),add=T,col='red')
```


Slikovni prikazi

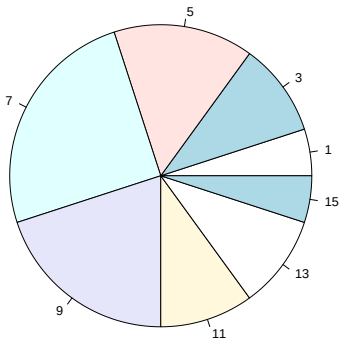
stolpci



poligon



strukturni krog



Vzorčne ocene

Najpogostejša parametra, ki bi ju radi ocenili sta:

sredina populacije μ glede na izbrano lastnost – matematično upanje spremenljivke X na populaciji; in

povprečni odklon od sredine σ – standardni odklon spremenljivke X na populaciji.

Statistike/ocene za te parametre so izračunane iz podatkov z vzorca. Zato jim tudi rečemo *vzorčne ocene*.

Sredinske mere

Kot sredinske mere se pogosto uporabljajo:

Vzorčni modus – najpogostejša vrednost (smiselna tudi za imenske).

Vzorčna mediana – srednja, glede na urejenost, vrednost (smiselna tudi za urejenostne).

Vzorčno povprečje – povprečna vrednost (smiselna za vsaj razmične)

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

Vzorčna geometrijska sredina – (smiselna za vsaj razmernostne)

$$G(x) = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

Mere razpršenosti

Za oceno populacijskega odklona uporabljamo *mere razpršenosti*.

Vzorčni razmah $= \max_i x_i - \min_i x_i$.

Vzorčna disperzija $s_0^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

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

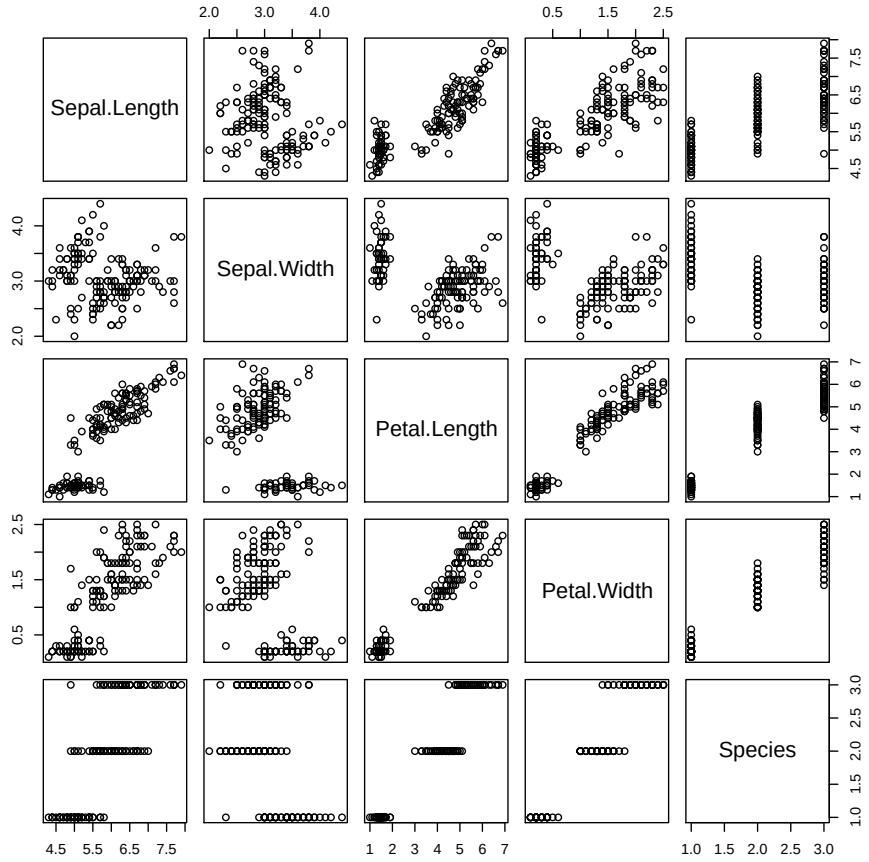
ter ustrezna *vzorčna odklona* s_0 in s.

Še nekaj ukazov v R-ju

```
> x <- rnorm(1000,mean=175,sd=30)
> mean(x)
[1] 175.2683
> sd(x)
[1] 30.78941
> var(x)
[1] 947.9878
> median(x)
[1] 174.4802
> min(x)
[1] 92.09012
> max(x)
[1] 261.3666
> quantile(x,seq(0,1,0.1))
       0%       10%       20%       30%
 92.09012 135.83928 148.33908 158.53864
      40%       50%       60%       70%
166.96955 174.48018 182.08577 191.29261
      80%       90%      100%
200.86309 216.94009 261.36656

> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  92.09  154.20  174.50  175.30  195.50  261.40
> hist(x,freq=F)
> curve(dnorm(x,mean=175,sd=30),add=T,col='red')
```


Fisherjeve oziroma Andersonove perunike (Iris data)




```
> data()
> data(iris)
> help(iris)
> summary(iris)
  Sepal.Length    Sepal.Width
 Min.   :4.300   Min.   :2.000
 1st Qu.:5.100   1st Qu.:2.800
 Median :5.800   Median :3.000
 Mean   :5.843   Mean   :3.057
 3rd Qu.:6.400   3rd Qu.:3.300
 Max.   :7.900   Max.   :4.400
  Petal.Length    Petal.Width
 Min.   :1.000   Min.   :0.100
 1st Qu.:1.600   1st Qu.:0.300
 Median :4.350   Median :1.300
 Mean   :3.758   Mean   :1.199
 3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :6.900   Max.   :2.500
       Species
 setosa    :50
 versicolor:50
 virginica :50
> pairs(iris)
```

Parni prikaz.

Škatle in Q-Q-prikazi

Škatle (box-and-whiskers plot; grafikon kvantilov) `boxplot`: škatla prikazuje notranja kvartila razdeljena z mediansko črto. Daljici – brka vodita do robnih podatkov, ki sta največ za 1.5 dolžine škatle oddaljena od nje. Ostali podatki so prikazani posamično.

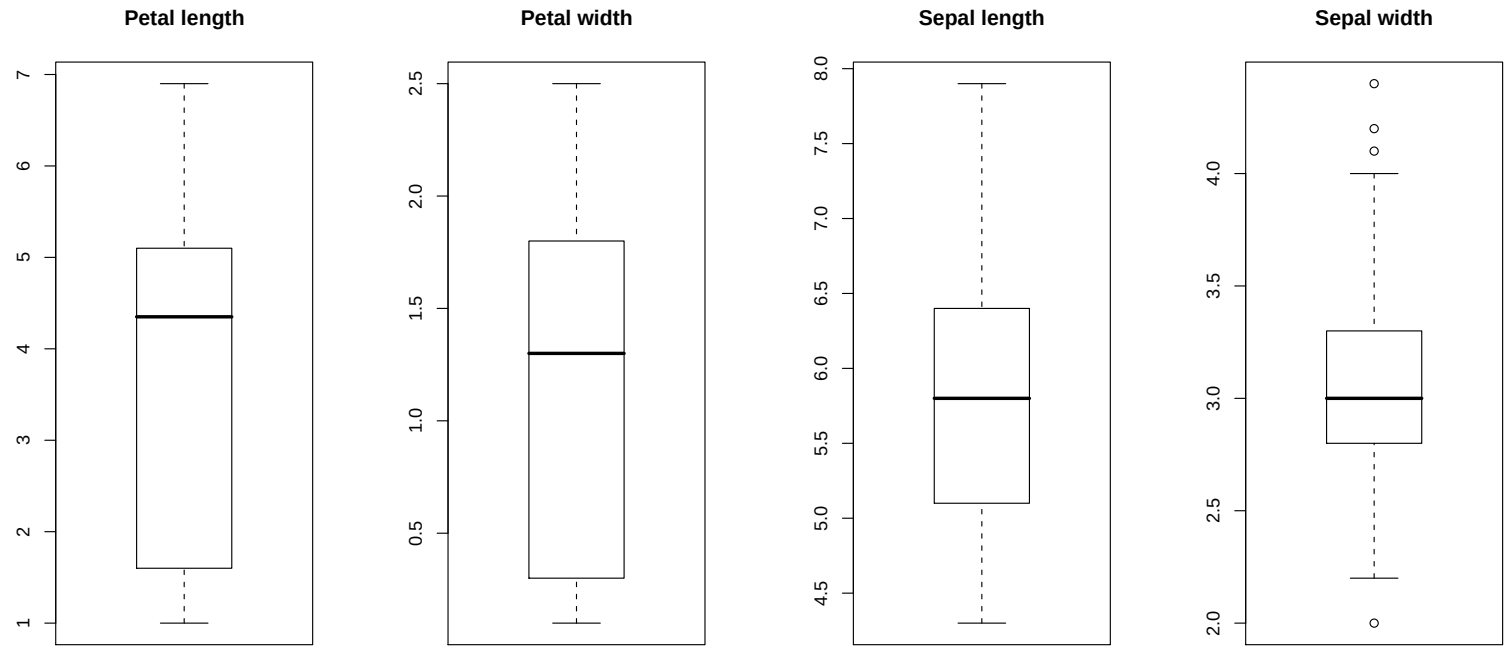
Q-Q-prikaz `qqnorm` je namenjen prikazu normalnosti porazdelitve danih n podatkov. Podatke uredimo in prikažemo pare točk sestavljene iz vrednosti k-tega podatka in pričakovane vrednosti k-tega podatka izmed n normalno porazdeljenih podatkov. Če sta obe porazdelitvi normalni, ležijo točke na premici. Premica `qqline` nariše premico skozi prvi in tretji kvartil.

Obstaja tudi splošnejši ukaz `qqplot`, ki omogoča prikaz povezanosti poljubnega para porazdelitev. S parametrom `datax=T` zamenjamo vlogo koordinatnih osi.

Histogram

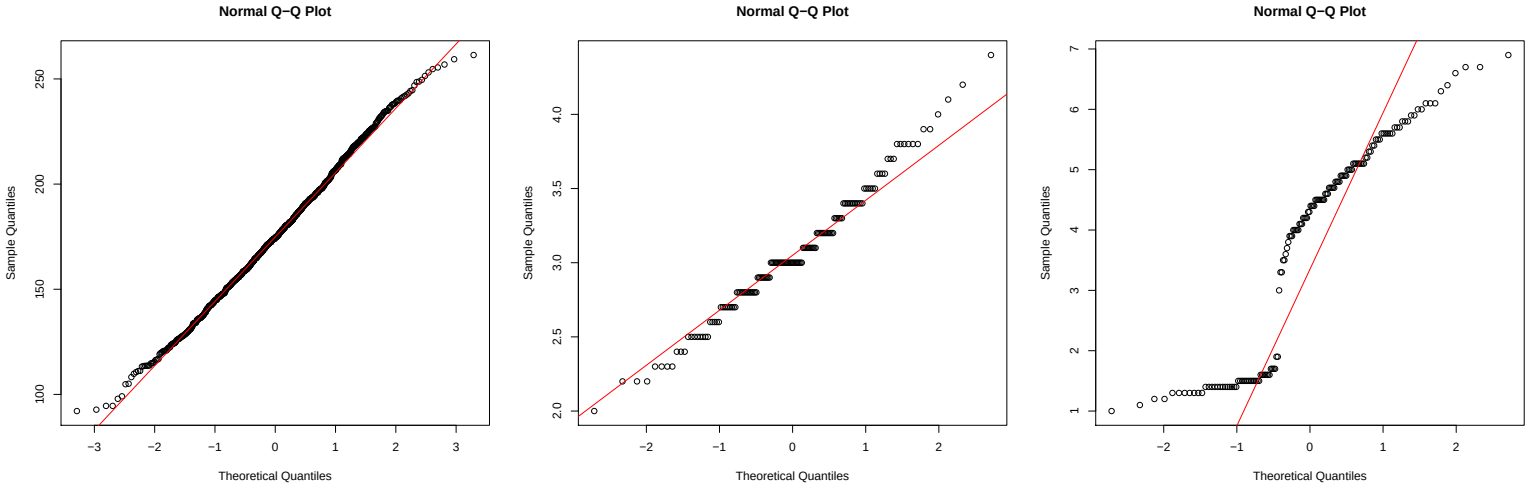
```
> hist(iris$Petal.Length)
> hist(iris$Sepal.Width)
```

Škatle




```
> par(mfrow=c(1,2))
> boxplot(iris$Petal.Length,main='Petal length')
> boxplot(iris$Petal.Width,main='Petal width')
> boxplot(iris$Sepal.Length,main='Sepal length')
> boxplot(iris$Sepal.Width,main='Sepal width')
> par(mfrow=c(1,1))
```

Q-Q-prikaz




```
> qqnorm(x)
> qqline(x,col='red')
> qqnorm(iris$Sepal.Width)
> qqline(iris$Sepal.Width,col='red')
> qqnorm(iris$Petal.Length)
> qqline(iris$Petal.Length,col='red')
```

Populacija: 1,2,3,4 ($n = 4$)

$$\mu = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1+2+3+4}{4} = 2,5$$

$$\begin{aligned}
\sigma &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i-\mu)^2} \\
&= \sqrt{\frac{(1-2,5)^2+(2-2,5)^2+(3-2,5)^2+(4-2,5)^2}{4}} \\
&= \sqrt{\frac{(-1,5)^2+(-0,5)^2+(0,5)^2+(1,5)^2}{4}} \\
&= \sqrt{\frac{5}{4}} = 1,118
\end{aligned}$$

Vsi možni vzorci velikosti 2:

1,1	2,1	3,1	4,1
1,2	2,2	3,2	4,2
1,3	2,3	3,3	4,3
1,4	2,4	3,4	4,4

1,1	1,0
1,2	1,5
1,3	2,0
1,4	2,5
2,1	1,5
2,2	2,0
2,3	2,5
2,4	3,0
3,1	2,0
3,2	2,5
3,3	3,0
3,4	3,5
4,1	2,5
4,2	3,0
4,3	3,5
4,4	4,0

Velikost populacije: $N = 16$

$$1, 01, 51, 52, 02, 02, 02, 52, 52, 52, 53, 03, 03, 03, 53, 54, 0$$

$$\begin{aligned} \mu_{\overline{Y}} &= \frac{1,0 + 2 * 1,5 + 3 * 2,0 + 4 * 2,5 + 3 * 3,0 + 2 * 3,5 + 4,0}{16} \\ &= \frac{40}{16} = 2,5 = \mu. \end{aligned}$$

$$\sigma_{\overline{Y}} = \sqrt{\frac{\sum_{i=1}^{N} (\overline{Y}_i - \mu_{\overline{Y}})^2}{N}} = \sqrt{\frac{10}{16}} = 0,79.$$

Vzorčna porazdelitev povprečja

Centralni limitni izrek

Če je naključni vzorec velikosti n izbran iz populacije s končnim povprečjem μ in varianco σ^2, potem je lahko, če je n dovolj velik, vzorčna porazdelitev povprečja $\overline{y}$ aproksimirana z gostoto normalne porazdelitve.

Naj bo $y_1, y_2, \ldots, y_n$ naključni vzorec, ki je sestavljen iz n meritev populacije s končnim povprečjem μ in končnim standardnim odklonom σ. Potem sta povprečje in standardni odklon vzorčne porazdelitve $\overline{y}$ enaka

$$\mu_{\overline{Y}} = \mu, \qquad \text{and} \qquad \sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}}.$$