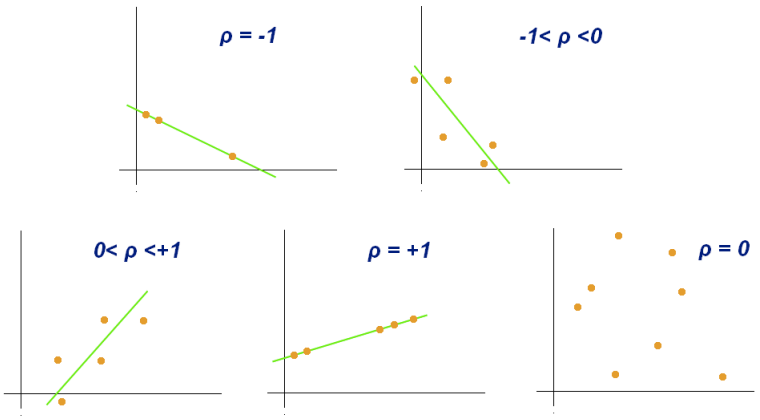
Korelacija med slučajnima spremenljivkama (Pearsonov korelacijski koeficient)

Linearno odvisnost slučajnih spremenljivk X in Y v statistiki pogosto merimo s korelacijskim koeficientom $corr(X, Y) = \rho_{X,Y}$. Velja $-1 \leq \rho_{X,Y} \leq 1$.



Korelacijski koeficient je definiran kot kvocient kovariance obeh spremenljivk in produkta njunih standardnih odklonov. Za celo populacijo je to

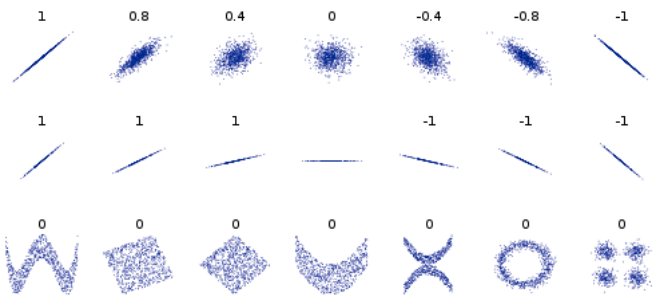
$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y},$$

za vzorec pa

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$

Korelacijski koeficient vzorca je dobra ocena za korelacijski koeficient populacije, če so povprečni vrednosti, varianci in kovarianca vzorca konsistentni.

Interpretacija korelacijskih koeficientov je močno odvisna od konteksta in namena. Korelacija 0.9 je npr. lahko zelo nizka, če preverjamo fizikalni zakon z visoko zanesljivimi instrumenti, lahko pa je zelo visoka v družboslovju, ko imamo veliko različnih dejavnikov.



V programskem paketu R imamo na razpolago ukaze: `cor()`, ...

Intervali zaupanja (angl. confidence interval)

Neznane parametre porazdelitve, kot sta povprečna vrednost in standardni odklon, lahko ocenjujemo na podlagi vzorca. To lahko delamo točkovno (metoda momentov, metoda največjega verjetja) ali intervalno. Pri točkovni oceni ne vemo, kako dobra je dobljena ocena, pri intervalni pa dobimo nek **interval zaupanja** za vrednost parametra t oblike

$$L <= t <= U.$$

Vrednosti l in u slučajnih spremenljivk L in U sta na osnovi danega vzorca določeni tako, da je verjetnost izbire vzorca opazovane porazdelitve z vrednostjo t v intervalu zaupanja $[l, u]$ enaka $1 - \alpha$. Torej

$$P(l \leq t \leq u) = 1 - \alpha,$$

kjer $1 - \alpha$ oziroma $(1 - \alpha) \cdot 100\,\%$ imenujemo **stopnja zaupanja**, l in u pa **spodnja** in **zgornja meja zaupanja**.

Neznane parametre porazdelitev za različne vrste porazdelitev določamo različno.

a.) **Normalno porazdeljena slučajna spremenljivka $X \sim N(\mu, \sigma)$, ocena za povprečno vrednost μ, če poznamo standardni odklon σ.**

Ker vemo, da je spremenljivka $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ standardno normalno porazdeljena, velja

$$P\left[-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right] = 1 - \alpha$$

in zato

$$P\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha,$$

kjer je $z_{\alpha/2}$ zgornji $\alpha/2$ kvantil standardne normalne porazdelitve. Na osnovi vzorca velikosti n dobimo vzorčno povprečje $\bar{x}$ in interval zaupanja s stopnjo zaupanja $(1 - \alpha) \cdot 100\,\%$:

$$l = \bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = u.$$

Enostranska intervala zaupanja s stopnjo zaupanja $(1 - \alpha) \cdot 100\,\%$:

$$l = \bar{x} - z_{\alpha}\frac{\sigma}{\sqrt{n}} \leq \mu \quad \text{in} \quad \mu \leq \bar{x} + z_{\alpha}\frac{\sigma}{\sqrt{n}} = u.$$

Izračun v programskem paketu R:

```
alpha=0.05
sig=...
vzorec=...
n=length(vzorec)
xp=mean(vzorec)
z=qnorm(alpha/2,0,1,lower.tail=FALSE)
l=xp-z*sig/sqrt(n)
u=xp+z*sig/sqrt(n)
print(paste("Interval",(1-alpha)*100,"% zaupanja: [",l,",",u,"]"))
```

b.) **Normalno porazdeljena slučajna spremenljivka $X \sim N(\mu, \sigma)$, ocena za povprečno vrednost μ, če ne poznamo standardnega odklona σ.**

Dan je majhen vzorec ($n < 40$)

Potem je $T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ Studentova t-porazdeljena slučajna spremenljivka z $n-1$ stopnjami prostosti, zato velja:

$$P\left[-t_{\alpha/2,n-1} \leq \frac{\bar{X}-\mu}{S/\sqrt{n}} \leq t_{\alpha/2,n-1}\right] = 1-\alpha$$

in

$$P\left[\bar{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right] = 1-\alpha,$$

kjer je $t_{\alpha/2,n-1}$ zgornji $\alpha/2$ kvantil Studentove t-porazdelitve z $n-1$ stopnjami prostosti. Na osnovi vzorca velikosti n dobimo vzorčno povprečje $\bar{x}$, standardni odklon s in interval zaupanja s stopnjo zaupanja $(1-\alpha)\cdot 100\,\%$:

$$l = \bar{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} = u.$$

Enostranska intervala zaupanja s stopnjo zaupanja $(1-\alpha)\cdot 100\,\%$:

$$l = \bar{x} - t_{\alpha,n-1}\frac{s}{\sqrt{n}} \leq \mu \quad \text{in} \quad \mu \leq \bar{x} + t_{\alpha,n-1}\frac{s}{\sqrt{n}} = u.$$

Izračun v programskem paketu R:

```
alpha=0.05
vzorec=...
n=length(vzorec)
xp=mean(vzorec)
s=sd(vzorec)
t=qt(alpha/2,n-1,lower.tail=FALSE)
l=xp-t*s/sqrt(n)
u=xp+t*s/sqrt(n)
print(paste("Interval",(1-alpha)*100,"% zaupanja: [",l,",",u,"]"))
```

Tukaj si lahko pomagamo z dodatno vgrajeno funkcijo `t.test()`:

```
alpha=0.05
vzorec=...
t.test(vzorec,conf.level=1-alpha)
```

Dan je velik vzorec ($n \geq 40$)

Ko se število prostostnih stopenj veča, se t-porazdelitev približuje standardni normalni porazdelitvi. Spremenljivka $Z = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ je torej približno standardno normalno porazdeljena, zato velja

$$P\left[-z_{\alpha/2} \leq \frac{\bar{X}-\mu}{S/\sqrt{n}} \leq z_{\alpha/2}\right] = 1-\alpha$$

in

$$P\left[\bar{X} - z_{\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}\right] = 1 - \alpha,$$

kjer je $z_{\alpha/2}$ zgornji $\alpha/2$ kvantil standardne normalne porazdelitve. Na osnovi vzorca velikosti n dobimo vzorčno povprečje $\bar{x}$, standardni odklon s in interval zaupanja s stopnjo zaupanja $(1-\alpha)\cdot 100\,\%$:

$$l = \bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}} = u.$$

Enostranska intervala zaupanja s stopnjo zaupanja $(1-\alpha)\cdot 100\,\%$:

$$l = \bar{x} - z_{\alpha}\frac{s}{\sqrt{n}} \leq \mu \quad \text{in} \quad \mu \leq \bar{x} + z_{\alpha}\frac{s}{\sqrt{n}} = u.$$

c.) **Normalno porazdeljena slučajna spremenljivka $X \sim N(\mu, \sigma)$, ocena za standardni odklon σ in varianco σ^2.**

Ker vemo, da je $X^2 = \frac{(n-1)S^2}{\sigma^2}$ χ^2-porazdeljena slučajna spremenljivka z $n-1$ prostostnimi stopnjami, velja:

$$P\left[\chi^2_{1-\alpha/2,n-1} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{\alpha/2,n-1}\right] = 1 - \alpha$$

in

$$P\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right] = 1 - \alpha,$$

kjer sta $\chi^2_{\alpha/2,n-1}$ in $\chi^2_{1-\alpha/2,n-1}$ zgornji in spodnji $\alpha/2$ kvantil χ^2-porazdelitve. Na osnovi vzorca velikosti n dobimo vzorčni standardni odklon s in interval zaupanja s stopnjo zaupanja $(1-\alpha)\cdot 100\,\%$:

$$l = \frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}} = u.$$

Enostranska intervala zaupanja s stopnjo zaupanja $(1-\alpha)\cdot 100\,\%$:

$$l = \frac{(n-1)s^2}{\chi^2_{\alpha,n-1}} \leq \sigma^2 \quad \text{in} \quad \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha,n-1}} = u.$$

d.) **Binomsko porazdeljena slučajna spremenljivka $X \sim Binom(n, p)$, ocena za delež populacije p z določeno lastnostjo.**

Ker vemo, da je spremenljivka $Z = \frac{X-np}{\sqrt{np(1-p)}}$ približno standardno normalno porazdeljena, če je n velik (!), velja

$$P\left[-z_{\alpha/2} \leq \frac{X - np}{\sqrt{np(1-p)}} \leq z_{\alpha/2}\right] = 1 - \alpha$$

in

$$P\left[\hat{P} - z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \leq p \leq \hat{P} + z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}\right] = 1-\alpha,$$

kjer je $\hat{P} = \frac{X}{n}$ in $z_{\alpha/2}$ je zgornji $\alpha/2$ kvantil standardne normalne porazdelitve. Na osnovi vzorca (velikosti n) dobimo vzorčni delež $\hat{p} = \frac{x}{n}$ in interval zaupanja s stopnjo zaupanja $(1-\alpha)\cdot 100\,\%$:

$$l = \hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = u.$$

Enostranska intervala zaupanja s stopnjo zaupanja $(1-\alpha)\cdot 100\,\%$:

$$l = \hat{p} - z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \quad \text{in} \quad p \leq \hat{p} + z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = u.$$

VAJA 8: Korelacije in intervali zaupanja

1. V grafičnem vmesniku programskega paketa R odprite novo skriptno datoteko.

2. V datoteki 'cena-poraba.csv' je podana napoved porabe energije v odvisnosti od cene in to v časovnih intervalih ene ure.

   a) Izračunajte stopnjo linearne odvisnosti (korelacijski koeficient) cene in porabe energije.

   b) Korelacijo cene in porabe energije prikažite grafično. Premico dodajte na graf z ukazom `abline(lsfit(cena,poraba),...)`.

   c) Podatke razdelite na dneve (po 24 vrstic) in izračunajte korelacijske koeficiente za vsak dan posebej.

   d) Dnevne korelacijske koeficiente prikažite grafično v obliki časovne vrste.

   R: a) 0.7660446 b) - c) $0.9864534, 0.8526022, 0.6871524, ..., 0.9274652$ d) -

3. Raziskovalci so ugotovili, da je za zgornji krvni tlak pri ženskah, starih od 25 do 35 let, normalno porazdeljen. V vzorcu velikosti $n = 25$ smo dobili vzorčno povprečno vrednost zgornjega krvnega tlaka $123,7$ mm in vzorčni standardni odklon $s = 15$ mm.

   a) Izračunajte spodnjo in zgornjo mejo 95% in 99% intervala zaupanja za povprečni zgornji krvni tlak.

   b) Kakšna je stopnja zaupanja za interval $[120, 127.4]$?

   c) Izračunajte še 95% in 99% intervala zaupanja za povprečni zgornji krvni tlak ob predpostavki, da je bila velikost vzorca $n = 250$.

   d) Določite še 95% in 99% intervala zaupanja za varianco in standardni odklon povprečnega zgornjega krvnega tlaka pri $n = 25$.

   R: a) $[117.51, 129.89]$ in $[115.31, 132.09]$ b) 77.06 % c) $[121.84, 125.56]$ in $[121.26, 126.14]$ d) za varianco $[137.18, 435.44]$ in $[118.53, 546.21]$, za standardni odklon $[11.71, 20.87]$ in $[10.89, 23.37]$

4. Klicni studio Episcenter je 9. novembra 2012 anketiral 975 opredeljenih volilnih upravičencev (od skupaj 1.711.781 vseh volilnih upravičencev), na podlagi česar je bil napovedan naslednji volilni izid v 1. krogu predsedniških volitev 2012:

kandidat	napoved (%)	izid (%)	interval 95 % zaupanja
Borut Pahor	$36,1$	$40,0$	
dr. Danilo Türk	$40,5$	$35,8$	
dr. Milan Zver	$23,4$	$24,2$	

   Število glasov, ki jih prejme kandidat A, je binomsko porazdeljena slučajna spremenljivka (uspeh = izbira kandidata A, neuspeh = izbira drugega kandidata).

a) Dopolnite tabelo (izračunajte intervale zaupanja s stopnjo zaupanja 95% za volilne napovedi). Ali so dejanski izidi v tem intervalu?

b) Največ kako velik vzorec bi moral dati te napovedi, da bi bil dejanski izid Boruta Pahorja v 95% intervalu zaupanja?

c) Najmanj kakšna stopnja zaupanja vključi dejanski izid Boruta Pahorja v interval zaupanja pri danem vzorcu velikosti $n = 975$?

R: a) $[0.33, 0.39]$, $[0.37, 0.44]$, $[0.21, 0.26]$ b) 582 c) 98.88

5. Vsebino skriptne datoteke shranite pod imenom 'vaja8.r'.