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
$\varphi(x) = a.x, a > 0$ podobnost	razmernostna	masa temperatura (K)
$\varphi(x) = a.x + b, a > 0$	razmična	temperatura (C,F) čas (koledar)
$x \geq y \Leftrightarrow \varphi(x) \geq \varphi(y)$ strogo naraščajoča	urejenostna	šolske ocene, kakovost 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.

Vzorci

Iz raznih razlogov (obsežnost, cena, nedostopnost, roki, uničenje enot pri merjenju, ...) ponavadi opazujemo/merimo lastnosti le na razmeroma majhnih vzorcih. 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*

. . . Vzorci

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*.

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 $(x_1, x_2, x_3, \dots, x_n)$ in $x \in \mathbb{R}$ postavimo $K(x) = |\{x_i : x_i < x, i = 1, \dots, n\}|$ 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 $k/n,\ k = 0, \dots, n$, 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

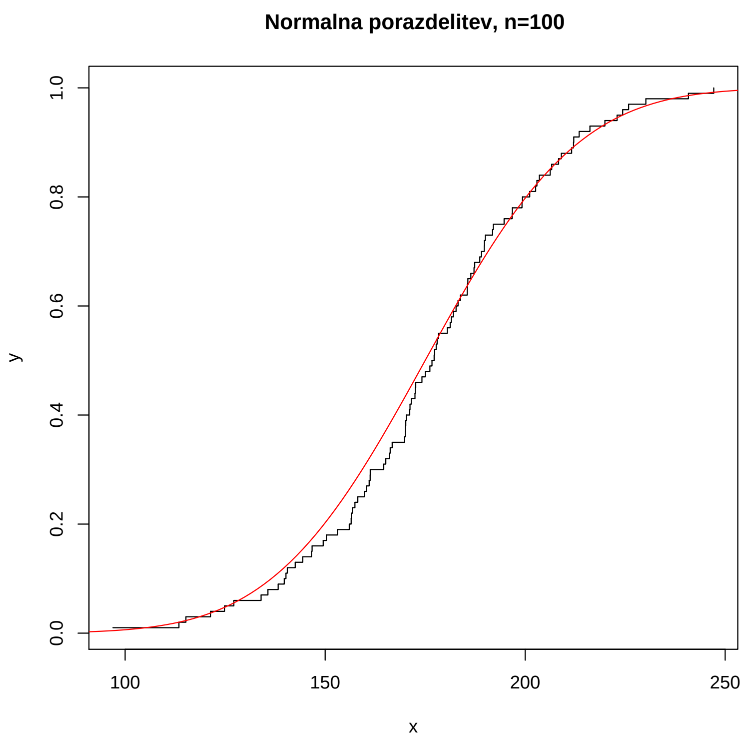
$$V_n(x) = \frac{1}{n} \sum_{i=1}^{n} Y_i(x).$$

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.

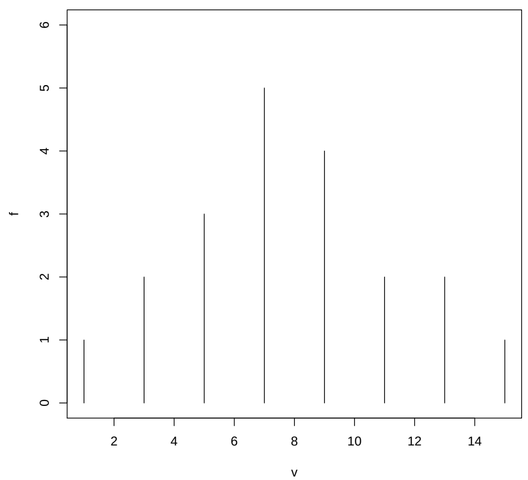
Ogiva: grafična predstavitev kumulativne frekvenčne porazdelitve s poligonom, kjer v koordinatni sistem nanašamo točke $(x_{i,min}, F_i)$.

Nekaj ukazov v R-ju

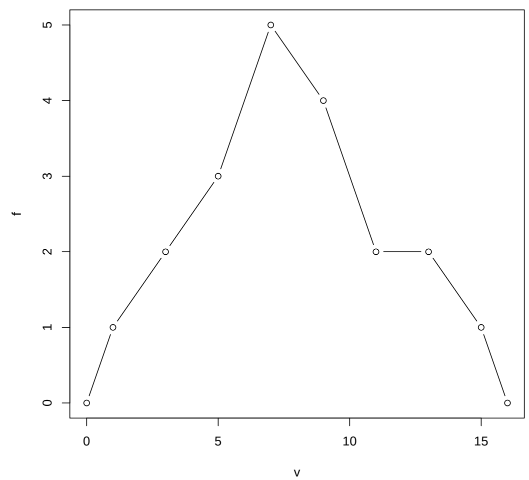
```
> X <- c(5,11,3,7,5,7,15,1,13,11,9,9,3,13,9,7,7,5,9,7)
> n <- length(X)
> t <- tabulate(X)
> t
 [1] 1 0 2 0 3 0 5 0 4 0 2 0 2 0 1
> v <- (1:max(X))[t>0]
> f <- t[t>0]
> rbind(v,f)
  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
v    1    3    5    7    9   11   13   15
f    1    2    3    5    4    2    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

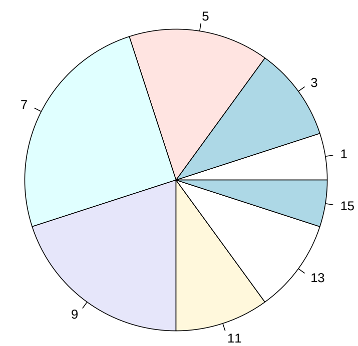
stolpci



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

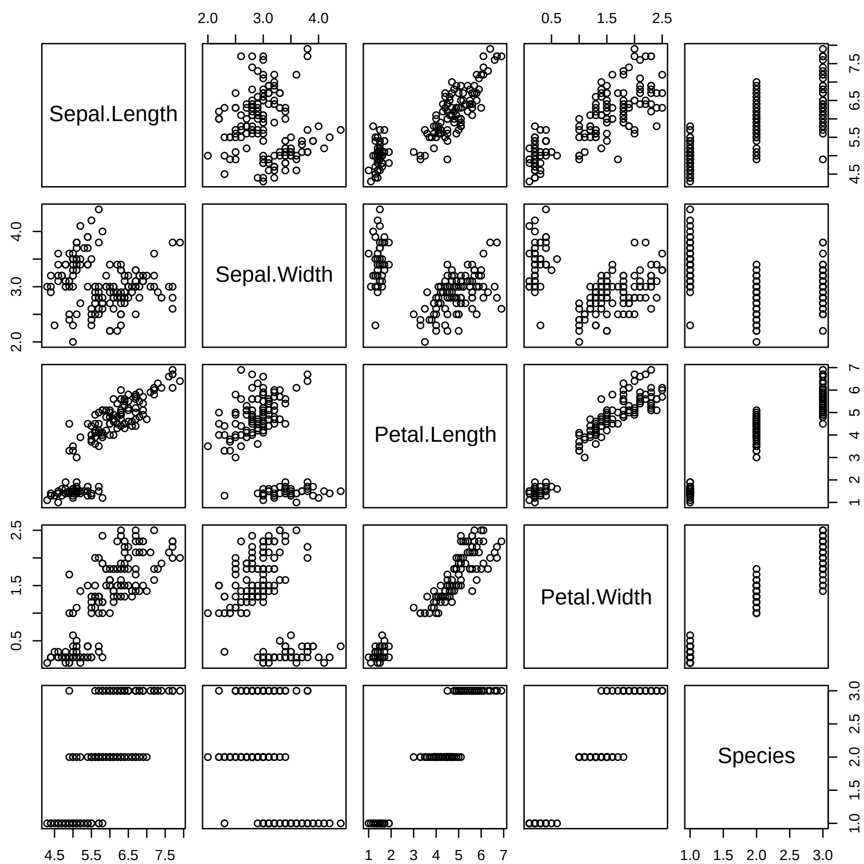
ter ustrezna *vzorčna odklona* s_0 in s.

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.

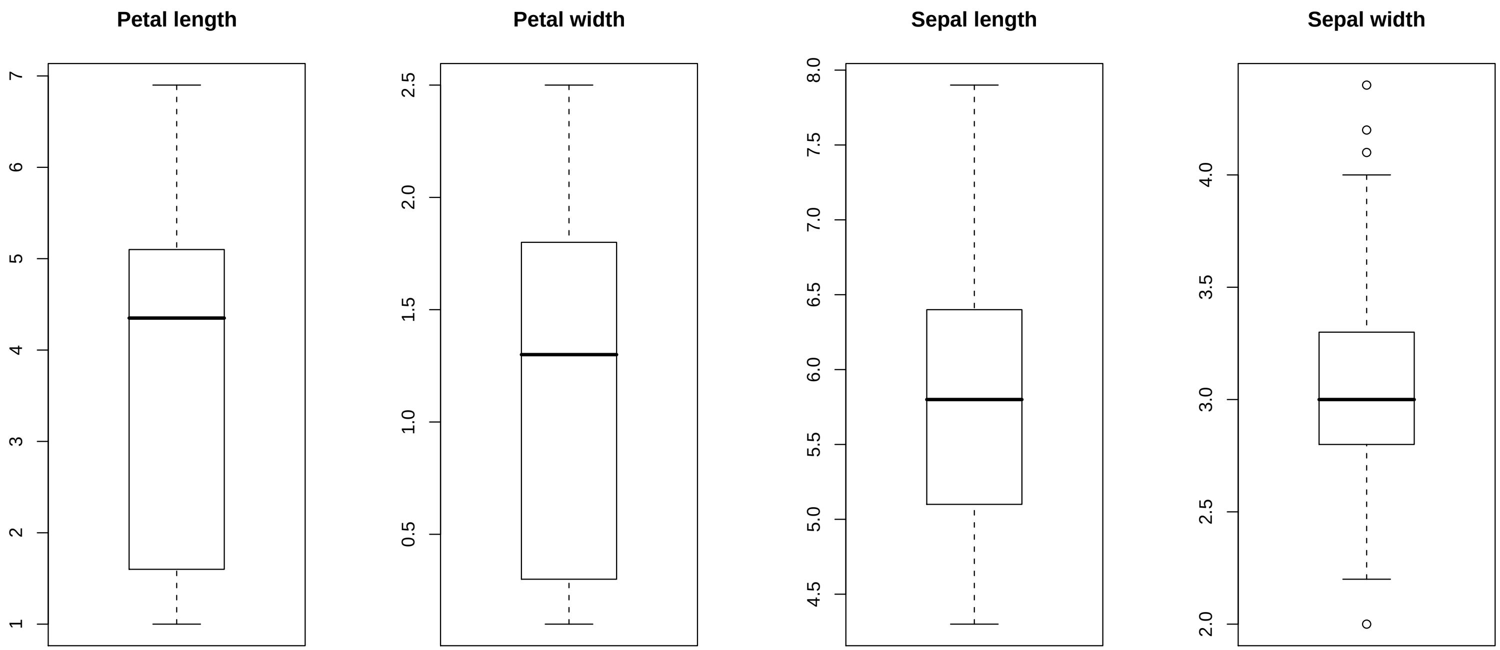
Q-Q-prikaz `qqnorm` je namenjen prikazu normalnosti porazdelitve danih n podatkov. Podatke uredimo in prikažemo pare točk sestavljene iz vrednosti k-tega podatka in pričakovane vrednosti k-tega podatka izmed n normalno porazdeljenih podatkov. Če sta obe porazdelitvi normalni, ležijo točke na premici.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

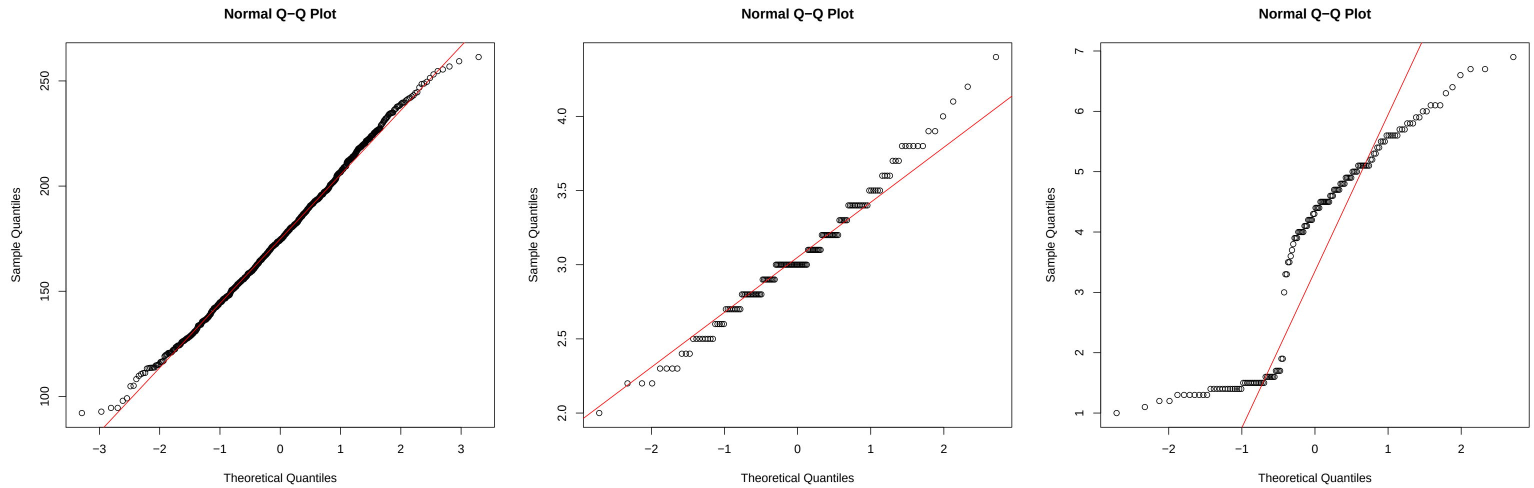
```
> hist(iris$Petal.Length)
> hist(iris$Sepal.Width)
```

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')
```

Porazdelitve vzorčnih statistik

Vzorčna statistika je poljubna simetrična funkcija (vrednost neodvisna od permutacije argumentov) vzorca

$$Y = g(X_1, X_2, X_3, \ldots, X_n)$$

Tudi vzorčna statistika je slučajna spremenljivka, za katero lahko določimo porazdelitev iz porazdelitve vzorca. Tudi za vzorčno statistiko sta najzanimivejši značilni vrednosti njeno matematično upanje $\mathsf{E}Y$ in standardni odklon σY, ki mu pravimo tudi *standardna napaka* statistike Y.

Vzorčno povprečje

Vzorčno povprečje je določeno z zvezo

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

Recimo, da ima spremenljivka X parametra $\mathsf{E}X = \mu$ in $\mathsf{D}X = \sigma^2$. Tedaj je

$$\mathsf{E}\bar{X} = \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}X_i = \mu$$

$$\mathsf{D}\bar{X} = \frac{1}{n^2}\sum_{i=1}^{n} \mathsf{D}X_i = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}$$

Iz druge zveze vidimo, da standardna napaka $\sigma\bar{X} = \frac{\sigma}{\sqrt{n}}$ statistike $\bar{X}$ pada z naraščanjem velikosti vzorca – $\bar{X} \to \mu$; kar nam zagotavlja tudi krepki zakon velikih števil.

Vzorčno povprečje in normalna porazdelitev

Naj bo $X : N(\mu, \sigma)$. Tedaj je $\sum_{i=1}^{n} X_i : N(n\mu, \sigma\sqrt{n})$ in dalje $\bar{X} : N(\mu, \sigma/\sqrt{n})$. Tedaj je vzorčna statistika

$$Z = \frac{\bar{X} - \mu}{\sigma}\sqrt{n} : N(0,1)$$

Kaj pa če porazdelitev X ni normalna? Izračun porazdelitve se lahko zelo zaplete. Toda pri večjih vzorcih ($n > 30$), lahko uporabimo centralni limitni izrek, ki zagotavlja, da je spremenljivka Z porazdeljena skoraj standardizirano normalno. Vzorčno povprečje

$$\bar{X} = \frac{\sigma}{\sqrt{n}} Z + \mu$$

ima tedaj porazdelitev približno $N(\mu, \sigma/\sqrt{n})$.

Zgled

Odgovorimo na vprašanje: Kolikšna je verjetnost, da bo pri 36 metih igralne kocke povprečno število pik večje ali enako 4 ?

X je slučajna spremenljivka z vrednostmi 1,2,3,4,5,6 in verjetnostmi $1/6$. Zanjo je $\mu = 3.5$ in standardni odklon $\sigma = 1.7$. Vseh 36 ponovitev meta lahko obravnavamo kot slučajni vzorec velikost 36. Tedaj je $P(\bar{X} \geq 4) = P(Z \geq (4 - \mu)\sqrt{n}/\sigma) = P(Z \geq 1.75) \approx 0.04$.

```
> x <- 1:6
> m <- mean(x)
> s <- sd(x)*sqrt(5/6)
> z <- (4-m)*6/s
> p <- 1-pnorm(z)
> cbind(m,s,z,p)
       m        s       z          p
[1,] 3.5 1.707825 1.75662 0.03949129
```