I.9. Karakteristične funkcije in limitni izreki

Karakteristična funkcija

Naj bo Z kompleksna slučajna spremenljivka, tj. $Z = X + iY$ za slučajni spremenljivki X in Y.

Njeno upanje izračunamo z

$$E(Z) = E(X) + iE(Y),$$

disperzijo pa z

$$D(Z) = E(|Z - E(Z)|^2) = D(X) + D(Y),$$

Kompleksna funkcija realne slučajne spremenljivke je kompleksna slučajna spremenljivka, npr. e^{iX}.

... 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 vedno obstajajo in so močno računsko orodje.

Posebej pomembni lastnosti sta:

Če obstaja začetni moment 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.

Reprodukcijska lastnost normalne porazdelitve

Vsaka linearna kombinacija *neodvisnih*
in *normalno* porazdeljenih slučajnih spremenljivk
je tudi sama **normalno** porazdeljena.

Če so slučajne spremenljivke $X_1, \ldots, X_n$ neodvisne in normalno porazdeljene $N(\mu_i, \sigma_i)$, potem je njihova vsota tudi normalno porazdeljena:

$$N\left(\sum \mu_i, \sqrt{\sum \sigma_i^2}\right).$$

Da ne bi vsota povprečij rastla z n, nadomestimo vsoto spremenljivk X_i z njihovim provprečjem $\overline{X}$ in dobimo

$$N\left(\overline{\mu}, \sqrt{\sum \left(\frac{\sigma_i}{n}\right)^2}\right).$$

Če privzamemo $\mu_i = \mu$ in $\sigma_i = \sigma$, dobimo $N\left(\mu, \sigma/\sqrt{n}\right)$.

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 konvergira skoraj gotovo $X_n \to X$,
potem konvergira tudi verjetnostno $X_n \to X$.

Šibki in krepki zakon velikih števil

Naj bo $X_1, \ldots, X_n$ 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 gre verjetnostno $Y_n \to 0$, tj., če $\forall \varepsilon > 0$ 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 gre skoraj gotovo $Y_n \to 0$, tj., če velja
$$P\left(\lim_{n\to\infty} \frac{S_n - \mathsf{E}S_n}{n} = 0\right) = 1.$$

Če za zaporedje $X_1, \ldots, X_n$ velja krepki zakon, velja tudi šibki.

Neenakost Čebiševa

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

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

Dokaz: Pokažimo jo za zvezne spremenljivke

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

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

Neenakost Čebiševa – posledice

(**Čebišev**) Če so slučajne spremenljivke X_i paroma nekorelirane in so vse njihove disperzije omejene z isto konstanto C, tj.

$$\mathsf{D}X_i < C \quad \text{za vsak } i,$$

velja za zaporedje šibki zakon velikih števil.

(**Markov**) Če gre za zaporedje slučajnih spremenljivk X_i izraz

$$\frac{\mathsf{D}S_n}{n^2} \to 0,$$

ko gre $n \to \infty$, velja za zaporedje šibki zakon velikih števil.

Dokaz Bernoullijevega izreka

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,
potem velja zanje šibki zakon velikih števil, tj. 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 izrek (CLI)

Leta 1810 je Pierre Laplace (1749-1827) študiral anomalije orbit Jupitra in Saturna, ko je izpeljal razširitev De Moivrovega limitnega izreka,

"Vsaka vsota ali povprečje, če je število členov dovolj veliko,
je približno normalno porazdeljena."

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.

Skica dokaz centralnega limitnega izreka

Naj bo $Z_i = \dfrac{X_i - \mu}{\sigma}$. Potem je

$$M_Z(t) = 1 - \frac{t^2}{2!} + \frac{t^3}{3!} E(Z_i^3) + \cdots$$

Za $Y_n = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \dfrac{1}{\sqrt{n}} \dfrac{\sum_{i=1}^n X_i - n\mu}{\sigma} = \dfrac{1}{\sqrt{n}} \displaystyle\sum_{i=1}^n Z_i$ velja

$$M_n(t) = \left[M_Z\left(\frac{t}{\sqrt{n}}\right) \right]^n = \left(1 - \frac{t^2}{2\,n} + \frac{t^3}{3!\,n^{3/2}}\, k + \cdots \right)^n,$$

kjer je $k = E(Z_i^3)$.

... Dokaz centralnega limitnega izreka

$$\log M_n(t) = n \log\left(1 - \frac{t^2}{2\,n} + \frac{t^3}{3!\,n^{3/2}}\,k + \cdots\right)$$

Za $x = \left(-\frac{t^2}{2\,n} + \frac{t^3}{3!\,n^{3/2}} + \cdots\right)$ velja

$$\log M_n(t) = n \log(1+x) = n(x - \frac{x^2}{2} + \cdots) =$$

$$n\left[\left(-\frac{t^2}{2n} + \frac{t^3}{3!\,n^{3/2}} + \cdots\right) - \frac{1}{2}\left(-\frac{t^2}{2n} + \frac{t^3}{3!\,n^{3/2}} + \cdots\right)^2 + \cdots\right]$$

in od tod končno še

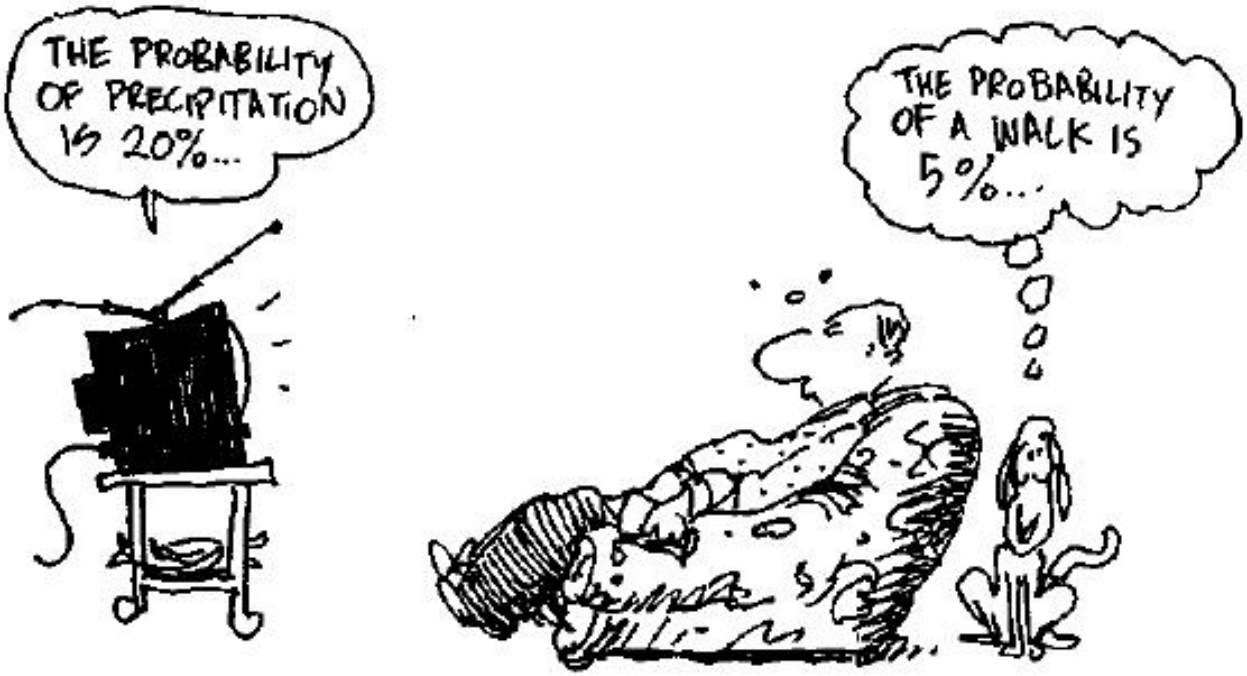
$$\lim_{n\to\infty} \log M_n(t) = -\frac{t^2}{2} \qquad \text{oziroma} \qquad \lim_{n\to\infty} M_n(t) = e^{-t^2/2}.$$

... Dokaz centralnega limitnega izreka

Iz konvergence karakterističnih funkcij φ_{Y_n}
proti karakteristični funkciji standardizirano normalne porazdelitve
lahko sklepamo po obratnem konvergenčnem izreku,
da tudi porazdelitvene funkcije za Y_n konvergirajo
proti porazdelitveni funkciji standardizirano normalne porazdelitve.

Torej velja centralni limitni zakon. ■

I.10. Uporaba

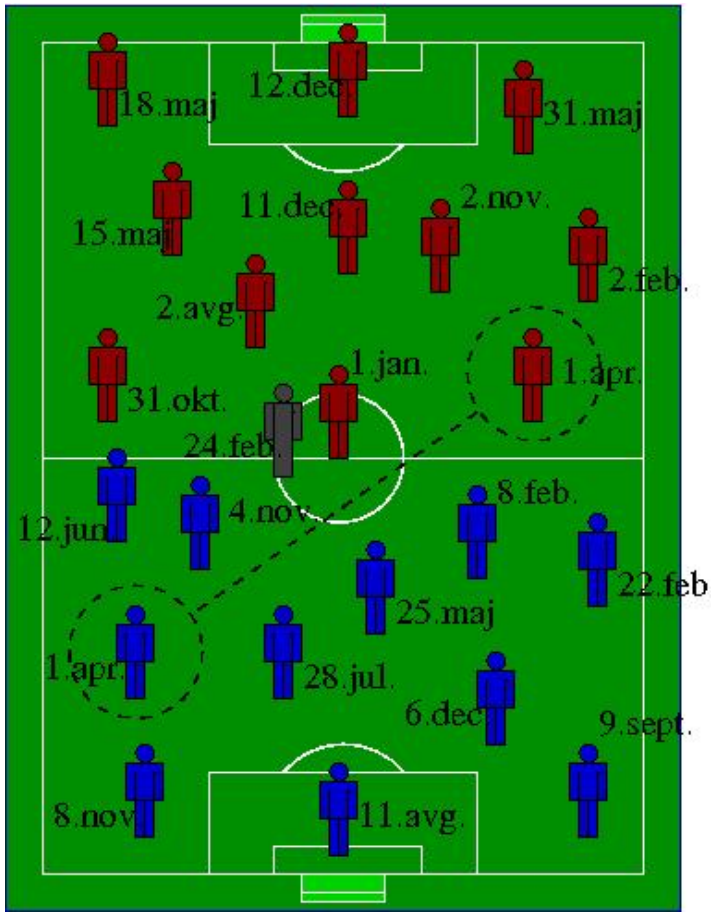

THE PROBABILITY OF PRECIPITATION IS 20%...
THE PROBABILITY OF A WALK IS 5%...

Kakšno naključje!!! Mar res?

Na nogometni tekmi sta
na igrišču dve enajsterici
in sodnik, skupaj
23 oseb.

Kakšna je verjetnost,
da imata **dve osebi**
isti rojstni dan?

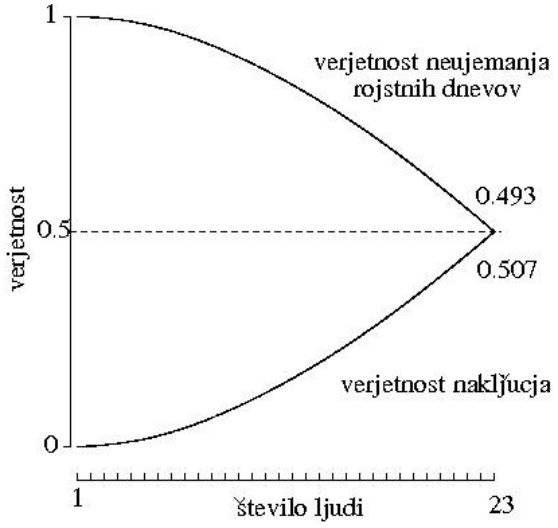
Ali je ta verjetnost lahko
večja od **0,5**?



Ko vstopi v sobo k-ta oseba, je verjetnost, da je vseh k rojstnih dnevov različnih enaka:

$$\frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} \times \cdots \times \frac{365-k+1}{365} =$$

$$= \begin{cases} 0,493; & \text{če je } k=22 \\ 0,507; & \text{če je } k=23 \end{cases}$$



**V poljubni skupini 23-ih ljudi je verjetnost,
da imata vsaj dva skupni rojstni dan $> 1/2$.**

Čeprav je 23 majhno število, je med 23 osebami 253 različnih parov.
To število je veliko bolj povezano z iskano verjetnostjo.

Testirajte to na zabavah z več kot 23 osebami.

Organizirajte stave in dolgoročno boste gotovo na boljšem,
na velikih zabavah pa boste zlahka zmagovali.

Napad s pomočjo paradoksa rojstnih dnevov

(angl. Birthday Attack)

To seveda ni paradoks, a vseeno ponavadi zavede naš občutek.

Ocenimo še splošno verjetnost.

Mečemo k žogic v n posod in gledamo,
ali sta v kakšni posodi vsaj dve žogici.

Poiščimo spodnjo mejo za verjetnost zgoraj opisanega dogodka:

$$\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{k-1}{n}\right)=\prod_{i=1}^{k-1}\left(1-\frac{i}{n}\right)$$

Iz Taylorjeve vrste

$$e^{-x} = 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \cdots$$

ocenimo $1 - x \approx e^{-x}$ in dobimo

$$\prod_{i=1}^{k-1}\left(1 - \frac{i}{n}\right) \approx \prod_{i=1}^{k-1} e^{\frac{-i}{n}} = e^{\frac{-k(k-1)}{2n}}.$$

Torej je verjetnost trčenja

$$1 - e^{\frac{-k(k-1)}{2n}}.$$

Potem velja

$$e^{\frac{-k(k-1)}{2n}} \approx 1-\varepsilon$$

oziroma

$$\frac{-k(k-1)}{2n} \approx \log(1-\varepsilon), \quad \text{tj.} \quad k^2-k \approx 2n\log\frac{1}{1-\varepsilon}$$

in če ignoriramo $-k$, dobimo končno

$$\boxed{k \approx \sqrt{2n\log\frac{1}{1-\varepsilon}}\,.}$$

Za $\varepsilon = 0,5$ je

$$k \approx 1,17\sqrt{n},$$

kar pomeni, da, če zgostimo nekaj več kot $\sqrt{n}$ elementov,
je bolj verjetno, da pride do trčenja kot da ne pride do trčenja.

V splošnem je k proporcionalen s $\sqrt{n}$.

Raba v kriptografiji

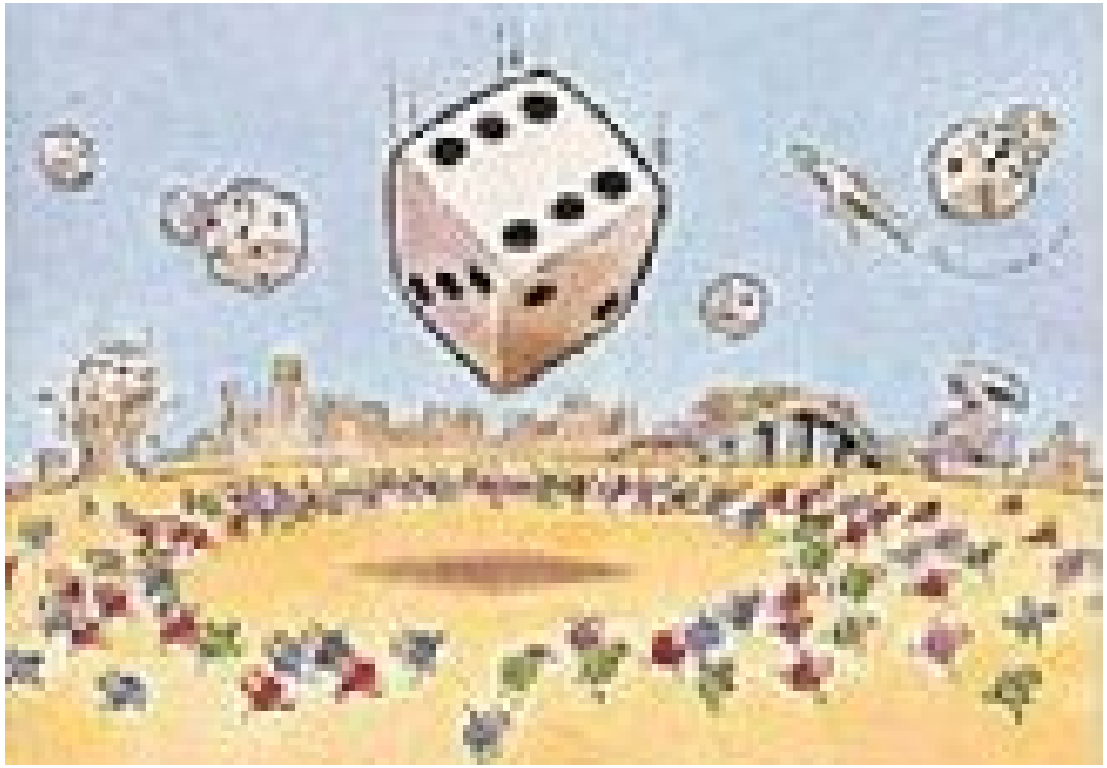
Napad s pomočjo paradoksa rojstnih dnevov s tem določi spodnjo mejo za velikost zaloge vrednosti zgoščevalnih funkcij, ki jih uporabljamo v kriptografiji in računalniški varnosti.

40-bitna zgostitev ne bi bila varna, saj bi prišli do trčenja z nekaj več kot 2^{20} (se pravi milijon) naključnimi zgostitvami z verjetnostjo vsaj $1/2$.

V praksi je priporočena najmanj 128-bitna zgostitev in standard za shema digitalnega podpisa (160 bitov) to vsekakor upošteva.

Podobno si lahko pomagamo tudi pri napadih na DLP in še kje.

II. STATISTIKA



II.1. Osnovni pojmi

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* (računalniško in matematično) statistiko.

... Osnovni pojmi

(Statistična) enota – posamezna proučevana stvar ali pojav.
Primer: redni študent na Univerzi v Ljubljani v študijskem letu 2008/09.

Populacija – množica vseh proučevanih enot; pomembna je natančna opredelitev populacije (npr. časovno in prostorsko).
Primer: vsi redni študentje na UL v študijskem letu 2008/09

Vzorec – podmnožica populacije, na osnovi katere ponavadi sklepamo o lastnostih celotne populacije.
Primer: vzorec 300 slučajno izbranih rednih študentov na UL v l. 2008/09.

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*
$f(x) = x$ (identiteta)	absolutna	štetje
$f(x) = a.x, a > 0$ podobnost	razmernostna	masa temperatura (K)
$f(x) = a.x + b, a > 0$	razmična	temperatura (C,F) čas (koledar)
$x \geq y \Leftrightarrow f(x) \geq f(y)$ strogo naraščajoča	urejenostna	šolske ocene, kakovost zraka, trdost kamnin
f 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.

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.

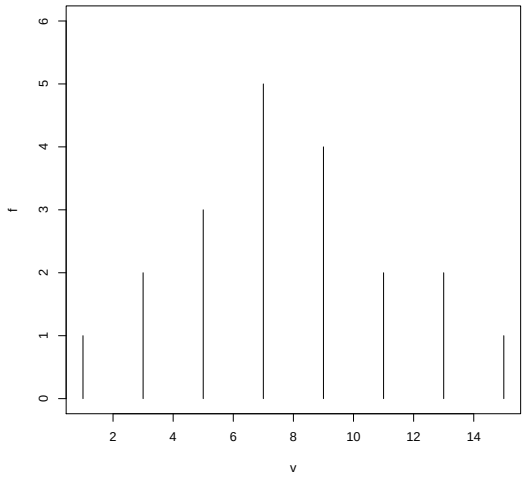
Ogiva: grafična predstavitev kumulativne frekvenčne porazdelitve s poligonom, kjer v koordinatni sistem nanašamo točke $(x_{i,min}, F_i)$.

Nekaj ukazov v R-ju

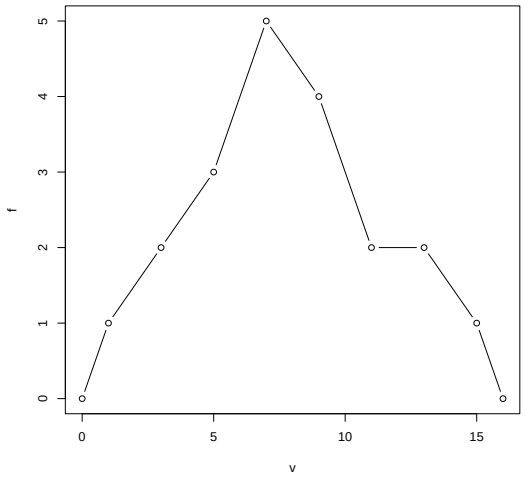
```
> X <- c(5,11,3,7,5,7,15,1,13,11,9,9,3,13,9,7,7,5,9,7)
> n <- length(X)
> t <- tabulate(X)
> t
 [1] 1 0 2 0 3 0 5 0 4 0 2 0 2 0 1
> v <- (1:max(X))[t>0]
> f <- t[t>0]
> rbind(v,f)
  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
v    1    3    5    7    9   11   13   15
f    1    2    3    5    4    2 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

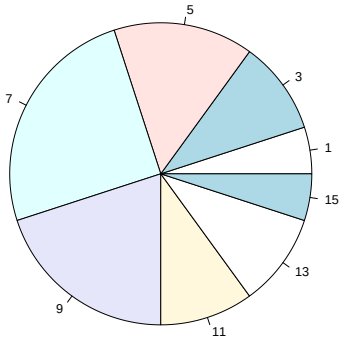
stolpci



poligon



strukturni krog

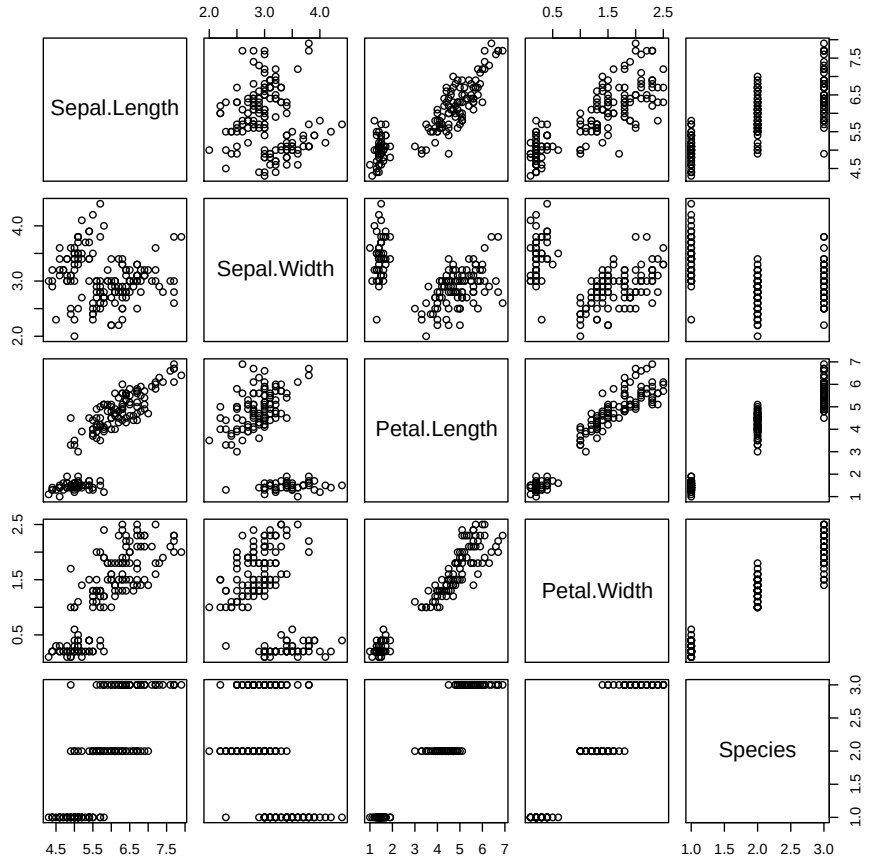


Še nekaj ukazov v R-ju

```
> 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.

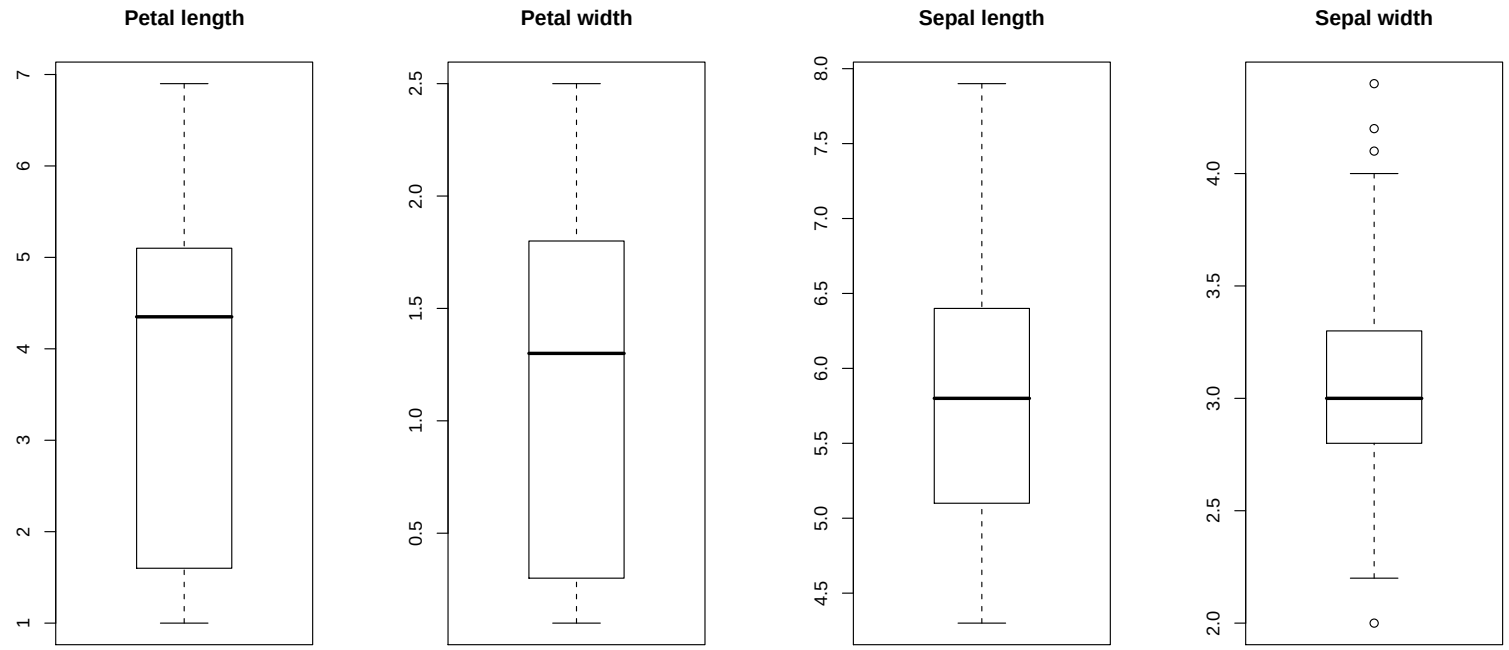
Škatle in Q-Q-prikazi

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.

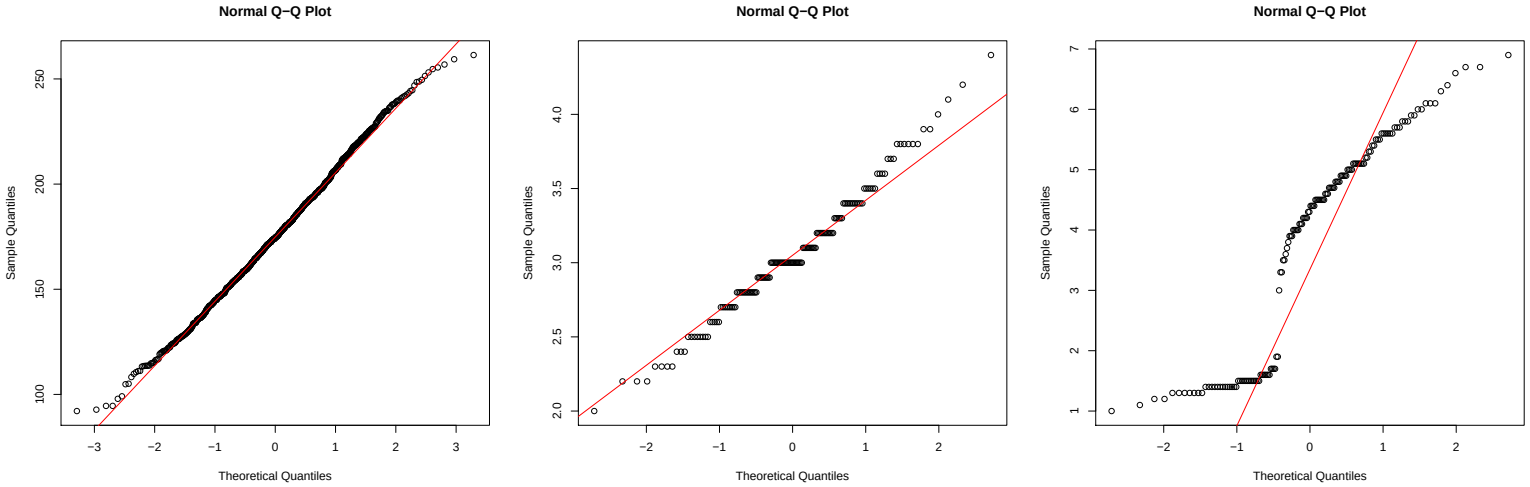
Obstaja tudi splošnejši ukaz `qqplot`, ki omogoča prikaz povezanosti poljubnega para porazdelitev. S parametrom `datax=T` zamenjamo vlogo koordinatnih osi.

Škatle



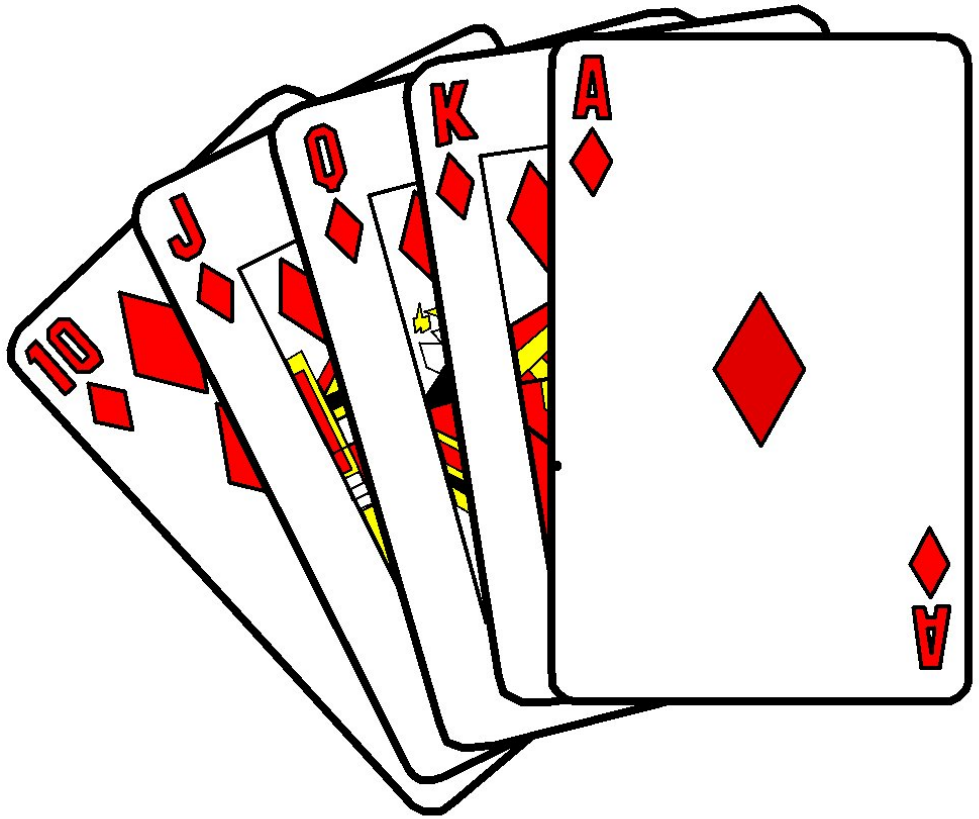

```
> par(mfrow=c(1,2))
> boxplot(iris$Petal.Length,main="Petal length")
> boxplot(iris$Petal.Width,main="Petal width")
> boxplot(iris$Sepal.Length,main="Sepal length")
> boxplot(iris$Sepal.Width,main="Sepal width")
> par(mfrow=c(1,1))
```

Q-Q-prikaz



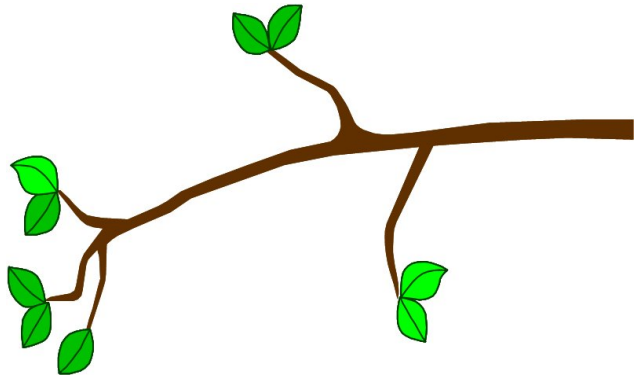

```
> qqnorm(x)
> qqline(x,col="red")
> qqnorm(iris$Sepal.Width)
> qqline(iris$Sepal.Width,col="red")
> qqnorm(iris$Petal.Length)
> qqline(iris$Petal.Length,col="red")
```

II.2. 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.

. . . Vzorčenje

Kdaj vzorec dobro predstavlja celo populacijo?

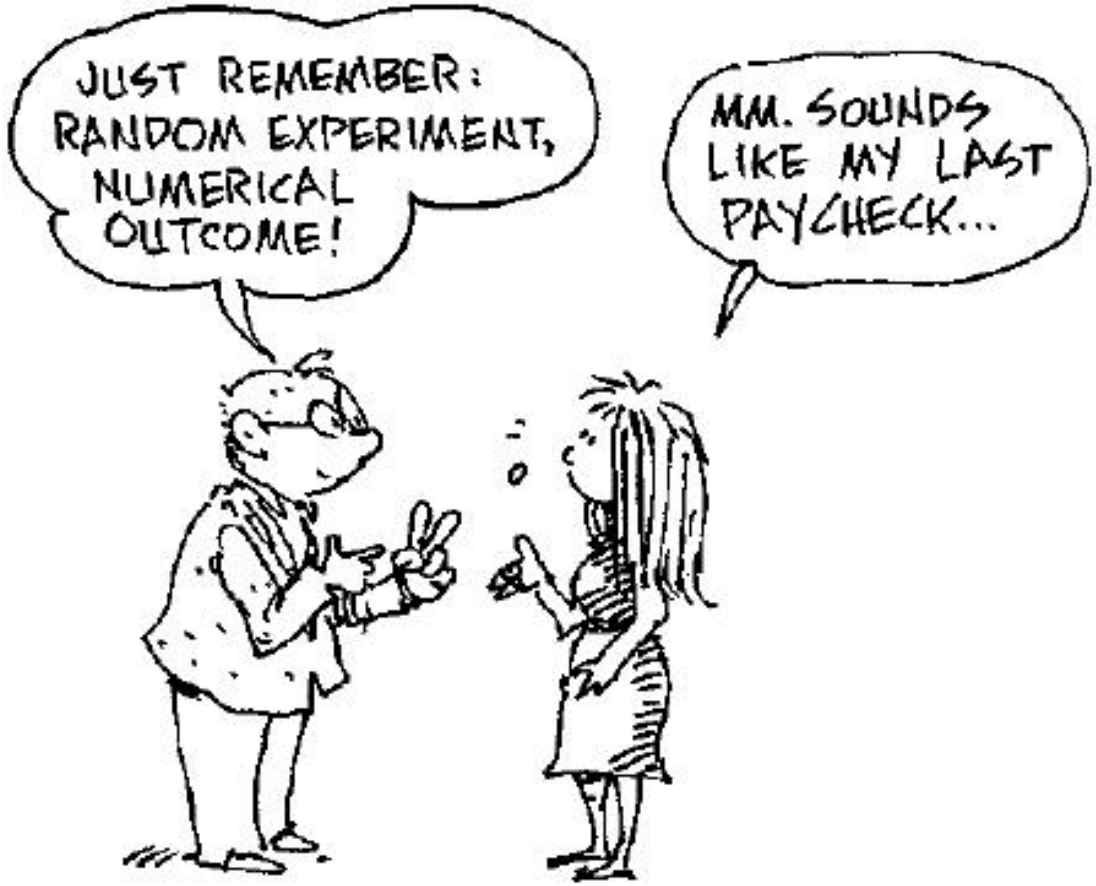
Preprost odgovor je:

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

Recimo, da merimo spremenljivko X, tako da n-krat naključno izberemo neko enoto in na njej izmerimo vrednost spremenljivke X. Postopku ustreza slučajni vektor

$$(X_1, X_2, \ldots, X_n),$$

ki mu rečemo *vzorec*. Število n je *velikost* vzorca.


JUST REMEMBER: RANDOM EXPERIMENT, NUMERICAL OUTCOME!
MM. SOUNDS LIKE MY LAST PAYCHECK...

. . . Vzorčenje

Ker v vzorcu merimo isto spremenljivko in posamezna meritev
ne sme vplivati na ostale, lahko predpostavimo:

1. vsi členi X_i vektorja imajo ***isto*** porazdelitev, kot spremenljivka X,
2. členi X_i so 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

$$K(x) = \left|\{x_i : x_i < x, i = 1, \ldots, n\}\right| \quad \text{in} \quad V_n(x) = K(x)/n.$$

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

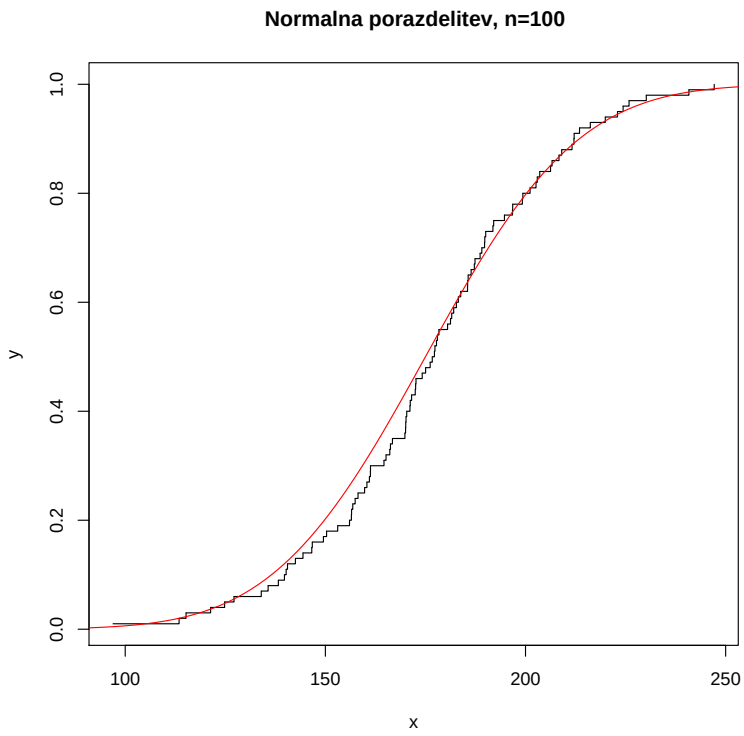
$$V_n(x) = \frac{1}{n} \sum_{i=1}^{n} Y_i(x).$$

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

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

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

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

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

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

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

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 vrednost, glede na urejenost, (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.

Porazdelitve vzorčnih statistik

Denimo, da je v populaciji N enot in
da iz te populacije slučajno izbiramo n enot
v enostavni slučajni vzorec ali na kratko slučajni vzorec
(vsaka enota ima enako verjetnost, da bo izbrana v vzorec, tj. $1/N$).

Če hočemo dobiti slučajni vzorec, moramo izbrane
enote pred ponovnim izbiranjem vrniti v populacijo
(vzorec s ponavljanjem).

Če je velikost vzorca v primerjavi s populacijo majhna,
se ne pregrešimo preveč, če imamo za slučajni vzorec tudi vzorec,
ki nastane s slučajnim izbiranjem brez vračanja.

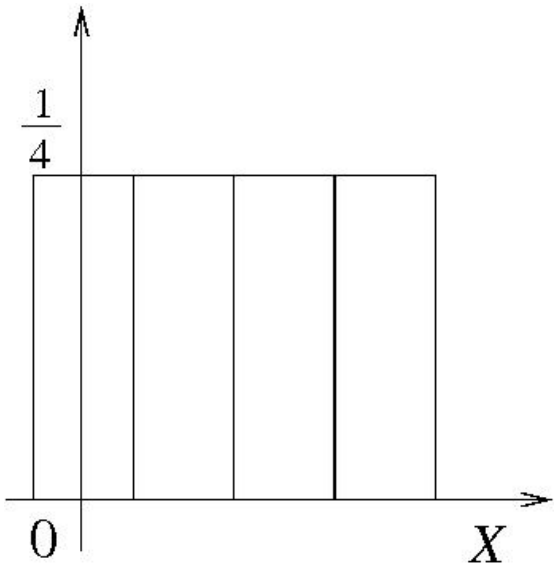
Predstavljajmo si, da smo iz populacije izbrali vse možne vzorce.
Dobili smo populacijo vseh možnih vzorcev.
Teh je v primeru enostavnih slučajnih vzorcev (s ponavljanjem) N^n;
kjer je N število enot v populaciji in n število enot v vzorcu.

Število slučajnih vzorcev brez ponavljanja pa je $\binom{N}{n}$,
če ne upoštevamo vrstnega reda izbranih enot v vzorcu,
oziroma $\binom{N+n-1}{n}$, če upoštevamo vrstni red.

Primer: Vzemimo populacijo z $N = 4$ enotami, ki imajo naslednje vrednosti spremenljivke X:

$$0, 1, 2, 3$$

Grafično si lahko porazdelitev spremenljivke X predstavimo s histogramom:



in izračunamo populacijsko aritmetično sredino in varianco:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{3}{2},$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \frac{5}{4}.$$

Sedaj pa tvorimo vse možne vzorce velikosti $n = 2$ s ponavljanjem, in na vsakem izračunajmo vzorčno aritmetično sredino $\overline{X}$:

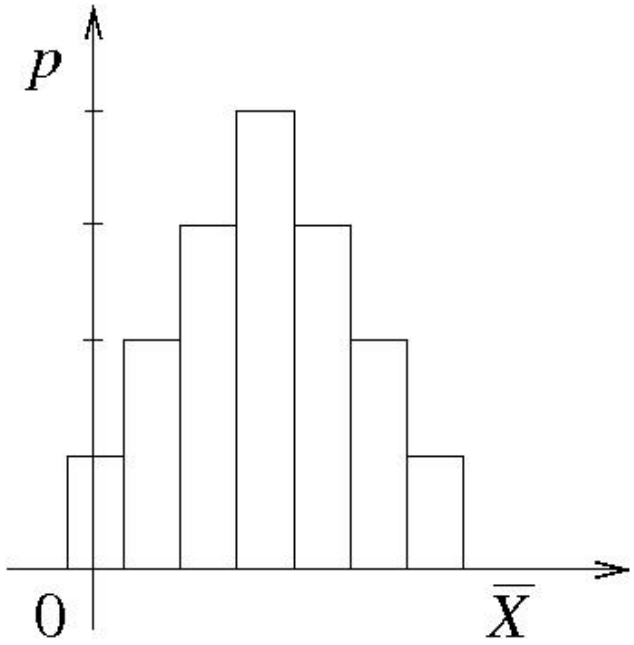
$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{2}(X_1 + X_2).$$

vzorci	$\overline{X}$	vzorci	$\overline{X}$
$0,0$	0	$2,0$	1
$0,1$	$0,5$	$2,1$	$1,5$
$0,2$	1	$2,2$	2
$0,3$	$1,5$	$2,3$	$2,5$
$1,0$	$0,5$	$3,0$	$1,5$
$1,1$	1	$3,1$	2
$1,2$	$1,5$	$3,2$	$2,5$
$1,3$	2	$3,3$	3

Zapišimo verjetnostno shemo slučajne spremenljivke vzorčno povprečje $\overline{X}$:

$$\overline{X} : \begin{pmatrix} 0 & 0,5 & 1 & 1,5 & 2 & 2,5 & 3 \\ 1/16 & 2/16 & 3/16 & 4/16 & 3/16 & 2/16 & 1/16 \end{pmatrix}$$

Grafično jo predstavimo s histogramom:



... in izračunajmo matematično upanje ter disperzijo vzorčnega povprečja:

$$E(\overline{X}) = \sum_{i=1}^{m} \overline{X}_i \, p_i = \frac{0+1+3+6+6+5+3}{16} = \frac{3}{2},$$

$$D(\overline{X}) = \sum_{i=1}^{m} \left(\overline{X}_i - E(\overline{X}) \right)^2 p_i = \frac{5}{8}.$$

S tem primerom smo pokazali, da je statistika ‘vzorčna aritmetična sredina’ slučajna spremenljivka s svojo porazdelitvijo. Poglejmo, kaj lahko rečemo v splošnem o porazdelitvi vzorčnih aritmetičnih sredin.

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

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

Hitrost centralne tendence pri CLI

Dokaz CLI je precej tehničen, kljub temu pa nam ne da občutka kako velik mora biti n, da se porazdelitev slučajne spremenljivke

$$X_1 + \cdots + X_n$$

približa normalni porazdelitvi.

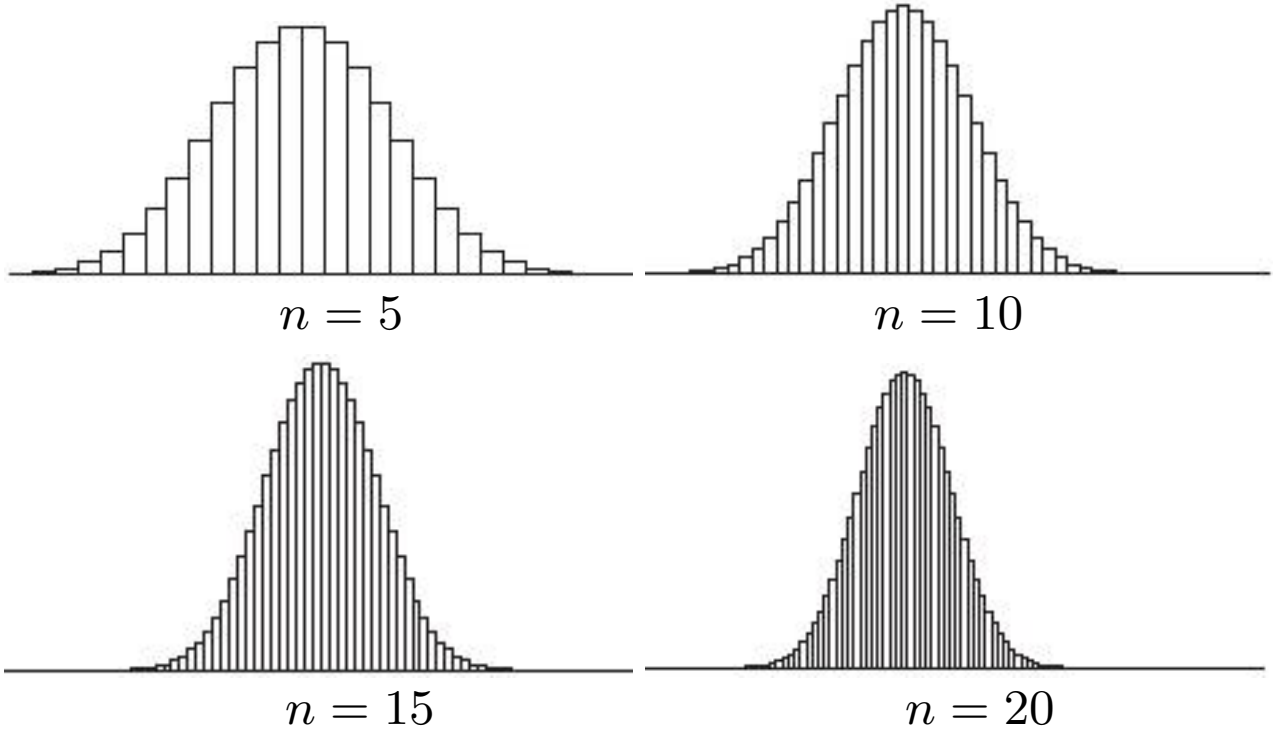
Hitrost približevanja k normalni porazdelitvi je odvisna od tega kako simetrična je porazdelitev.

To lahko potrdimo z eksperimentom: mečemo (ne)pošteno kocko, X_k naj bo vrednost, ki jo kocka pokaže pri k-tem metu.

Centralna tendenca za pošteno kocko

$p_1 = 1/6, \quad p_2 = 1/6, \quad p_3 = 1/6, \quad p_4 = 1/6, \quad p_5 = 1/6, \quad p_6 = 1/6.$

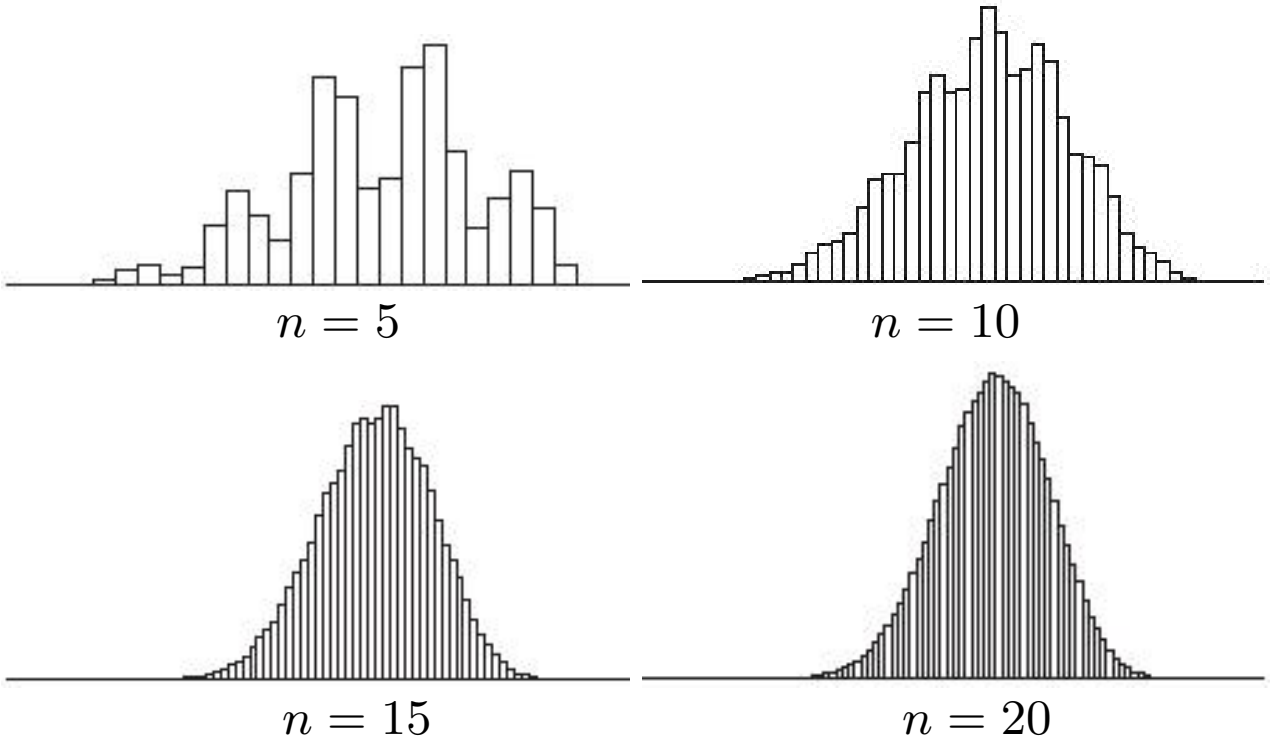
in slučajno spremenljivko $X_1 + X_2 + \cdots + X_n$:



Centralna tendenca za goljufivo kocko

$$p_1 = 0,2, \quad p_2 = 0,1, \quad p_3 = 0, \quad p_4 = 0, \quad p_5 = 0,3, \quad p_6 = 0,4.$$

in slučajno spremenljivko $X_1 + X_2 + \cdots + X_n$:



II.3. Cenilke

Vzorčna statistika

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

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

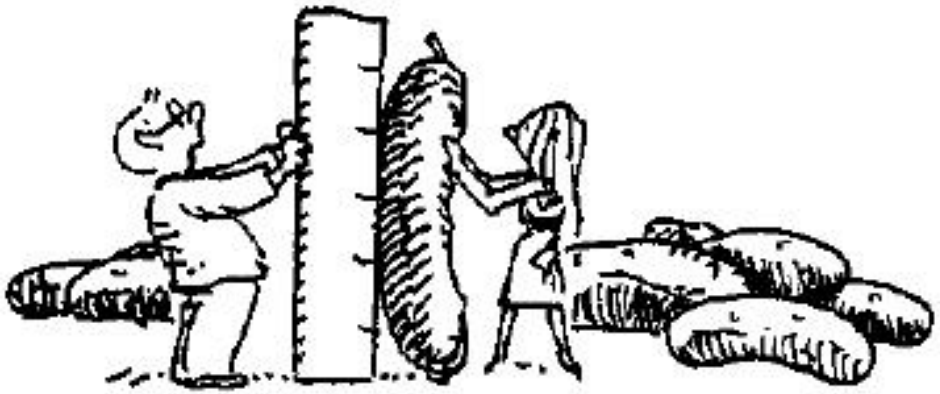
Tudi vzorčna statistika je slučajna spremenljivka, za katero lahko določimo porazdelitev iz porazdelitve vzorca.

Najzanimivejši sta značilni vrednosti

- njeno matematično upanje $\mathsf{E}Y$,
- standardni odklon σY, ki mu pravimo tudi *standardna napaka* statistike Y (angl. standard error – zato oznaka $\mathrm{SE}(Y)$).

(A) Vzorčno povprečje

Proizvajalec embalaže kumare bi rad ugotovil **povprečno dolžino** kumarice (da se odloči za velikost embalaže), ne da bi izmeril dolžino čisto vsake.



Zato naključno izbere n kumar in izmeri njihove dolžine $X_1, \ldots, X_n$. Sedaj nam je že blizu ideja, da je vsaka dolžina X_i **slučajna spremenljivka** (numerični rezultat nalkjučnega eksperimenta).

Če je μ (iskano/neznano) povprečje dolžin, in je σ standardni odklon porazdelitve dolžin kumar, potem velja

$$\mathsf{E}X_i = \mu, \qquad \text{in} \qquad \mathsf{D}X_i = \sigma^2,$$

za vsak i, ker bi X_i bila lahko dolžina katerekoli kumare.

. . . Vzorčno povprečje

Oglejmo si *vzorčno povprečje*, določeno z zvezo

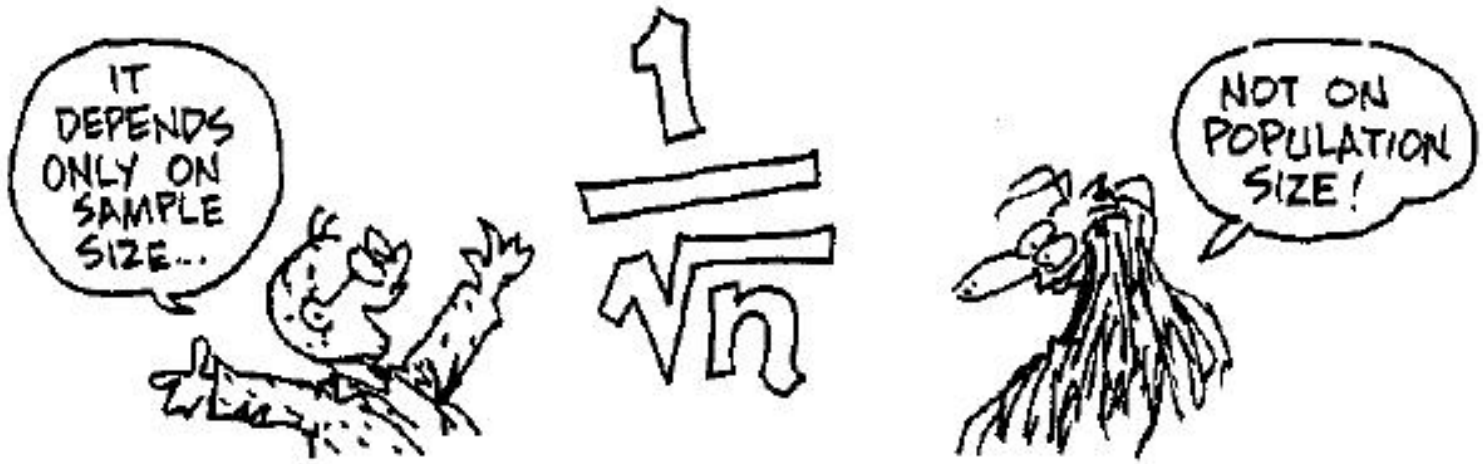
$$\overline{X} = \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

ki je tudi slučajna spremenljivka. Tedaj je

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

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

Iz druge zveze vidimo, da standardna napaka $\sigma\overline{X} = \dfrac{\sigma}{\sqrt{n}}$ statistike $\overline{X}$ pada z naraščanjem velikosti vzorca, tj. $\overline{X} \to \mu$;
(enako nam zagotavlja tudi krepki zakon velikih števil).


IT DEPENDS ONLY ON SAMPLE SIZE...
$\frac{1}{\sqrt{n}}$
NOT ON POPULATION SIZE!

Denimo, da se spremenljivka X na populaciji porazdeljuje normalno $N(\mu, \sigma)$. Na vsakem vzorcu (s ponavljanjem) izračunamo vzorčno aritmetično sredino $\overline{X}$. Dokazati se da, da je **porazdelitev vzorčnih aritmetičnih sredin** normalna, kjer je

- matematično upanje vzorčnih aritmetičnih sredin enako aritmetični sredini spremenljivke na populaciji

$$E(\overline{X}) = \mu,$$

- standardni odklon vzorčnih aritmetičnih sredin

$$\mathrm{SE}(\overline{X}) = \frac{\sigma}{\sqrt{n}}.$$

Če tvorimo vzorce iz končne populacije brez vračanja, je standardni odklon vzorčnih aritmetičnih sredin

$$\mathrm{SE}(\overline{X}) = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}.$$

Za dovolj velike vzorce ($n > 30$) je porazdelitev vzorčnih aritmetičnih sredin približno normalna, tudi če spremenljivka X ni normalno porazdeljena. Če se statistika X porazdeljuje vsaj približno normalno s standardno napako $\mathrm{SE}(X)$, potem se

$$Z = \frac{X - E(X)}{\mathrm{SE}(X)}$$

porazdeljuje standardizirano normalno.

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 $\overline{X} : N(\mu, \sigma/\sqrt{n})$. Tedaj je vzorčna statistika

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

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

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

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

$$P(\overline{X} \geq 4) = P(Z \geq (4-\mu)\sqrt{n}/\sigma) = P(Z \geq 1,75) \approx 0,04.$$

```
> 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
```