

Verjetnost

Nataša Kejžar

Inštitut za biostatistiko in medicinsko informatiko
Medicinska fakulteta, Univerza v Ljubljani

Natasa.Kejzar at mf.uni-lj.si

Kaj je verjetnost nekega dogodka?

- relativna frekvenca opazovanega dogodka
- delež pri vseh ponovitvah poskusa
 - če lahko ponovimo poskus velikokrat

$$\text{relativna frekvenca} = \frac{\text{frekvenca dogodka}}{\text{število vseh poskusov}}$$

- delež ugodnih izidov (izmed vseh izidov)
 - če so vsi izidi enako verjetni

$$P(A) = \frac{m}{n} = \frac{\text{število ugodnih izidov}}{\text{število vseh izidov}}$$

- kako močno verjamemo, da se bo dogodek zgodil ('degree of belief')

Zakaj potrebujemo verjetnost?



športniki



ostali

Povprečje populacije:

μ_s

???

$\mu_s = \mu_{ns}$

???

Povprečje populacije:

μ_{ns}

Namen: Primerjati povprečen pulz v mirovanju pri športnikih in ostalih v populaciji

Osnove verjetnosti

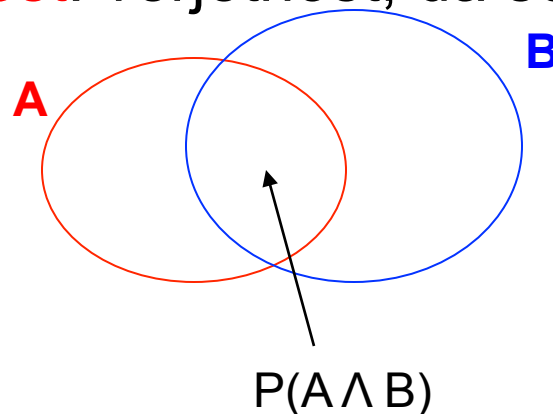
- E: **dogodek** (event): pojav, ki se lahko zgodi
- A={Izberemo študenta, ki se aktivno ukvarja s športom}
- B={Rodi se deklica}
- C={Pri metu kocke pade enka}

- $0 \leq P(E) \leq 1$: velja za vsak dogodek

- Če se E ne more zgoditi: $P(E)=0$ (**nemogoč dogodek**)
- Če se E vedno zgodi : $P(E)=1$ (**gotov dogodek**)

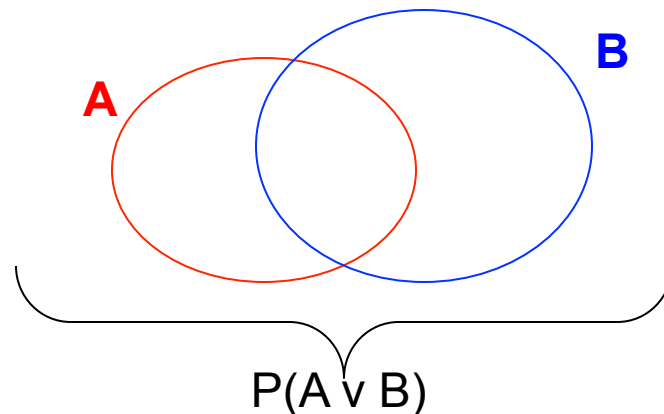
Osnove verjetnosti

- \wedge : “in”, “presek” – Produkt dogodkov
- $E=A \wedge B$, dogodek, ki se zgodi, ko se zgodita A in B
 - $P(A \wedge B) = P(A) P(B)$, če A in B sta neodvisna
 - $P(A \wedge B) = 0$, če A in B se ne moreta zgodita hkrati
 - $P(A \wedge B) = P(A) P(B|A)$, če A in B nista neodvisna
- $P(B|A)$: pogojna verjetnost: verjetnost, da se zgodi B, če se zgodi A



Osnove verjetnosti

- \vee : "ali", "unija" – Unija dogodkov
- $E=A \vee B$, dogodek, ki se zgodi, ko se zgodi A ali B
 - $P(A \vee B) = P(A)+P(B)$, če A in B se **medsebojno izključujeta, se ne moreta zgoditi hkrati (nezdružljivi dogodki)**
 - $P(A \vee B) = P(A)+P(B)-P(A \wedge B)$ če se A in B **medsebojno ne izključujeta**

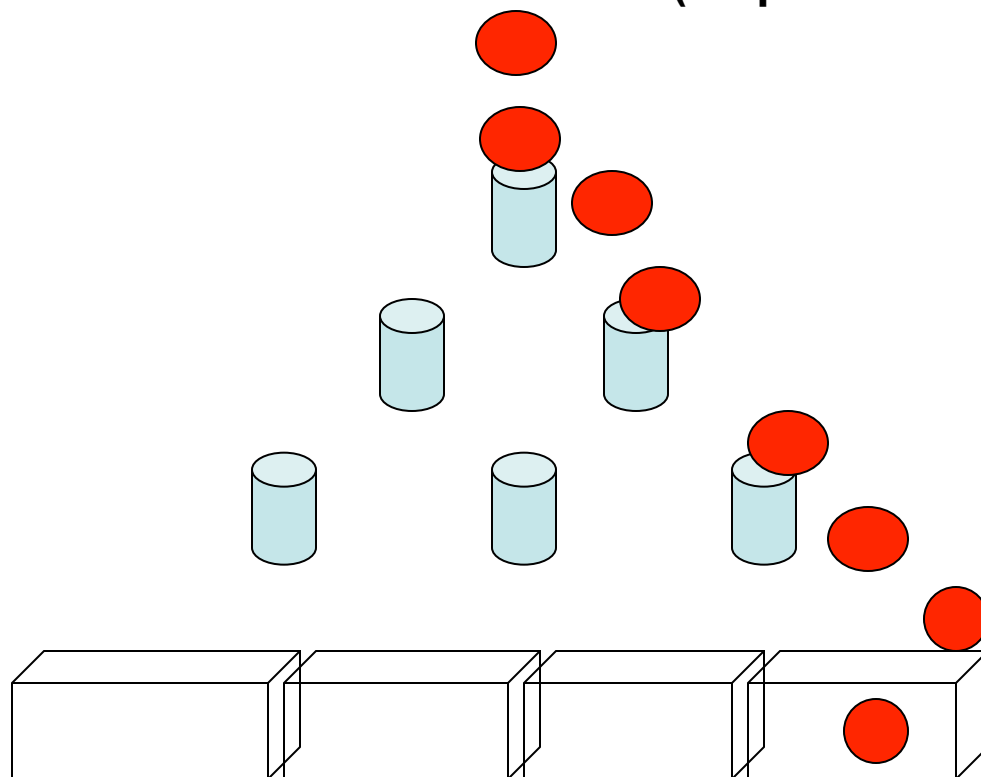


Slučajne spremenljivke (Random Variables)

- Slučajne spremenljivke so **numerični izid nekega poskusa**
 - Merjena količina, katere vrednosti naključno variirajo
 - Primeri:
 - Število grbov, ko mečemo kovanec 10 krat ($G, g=0,1,\dots,10$)
 - Število metov kocke, preden dobimo prvo enko
 - Prisotnost bolezni
 - Višina
 - Število srčnih utripov v eni minuti
 - Genska izraženost
 - ...
- Če poznamo vse možne izide in verjetnost vsakega izida, poznamo **verjetnostno porazdelitev** ene slučajne spremenljivke
- STATISTIKA: predpostavljamo porazdelitev, preverjamo, ali je predpostavka dovolj verjetna

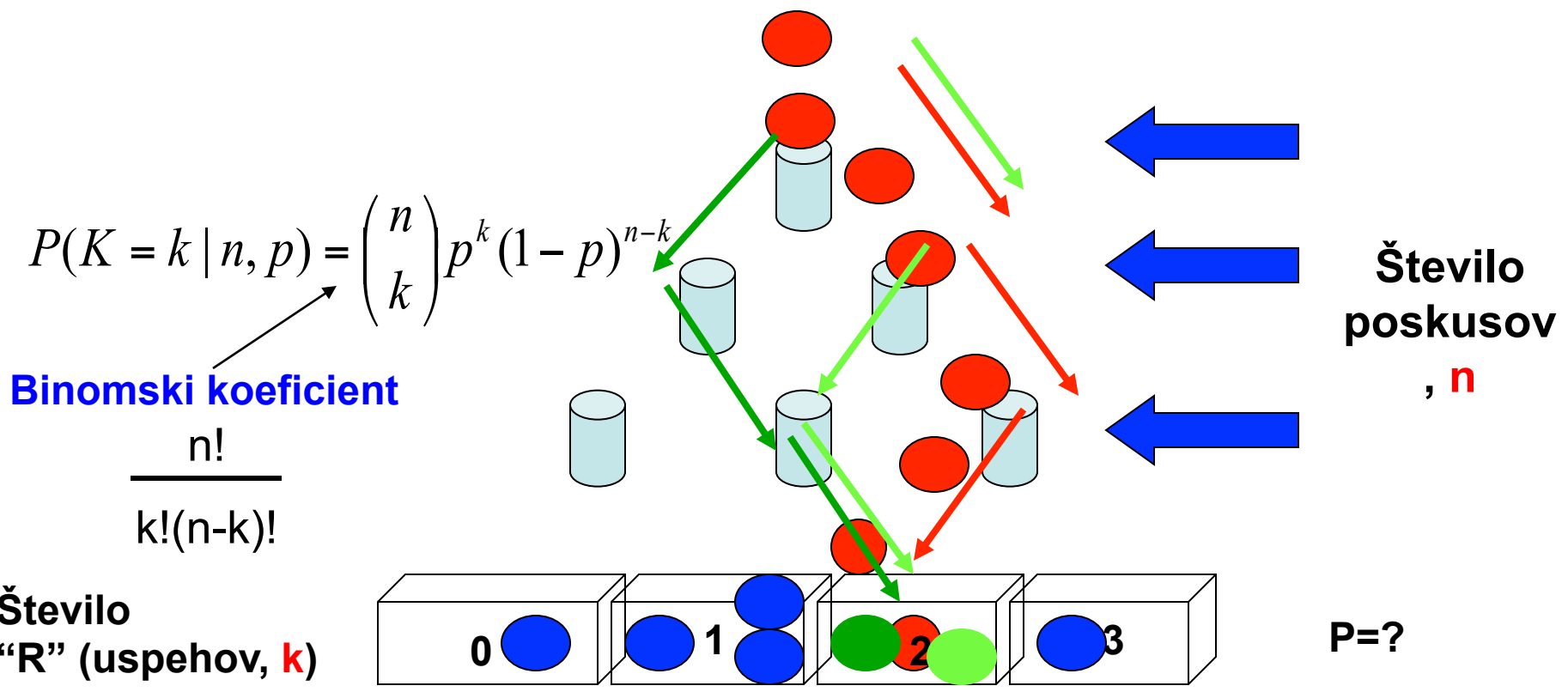
Koliko bi stavili (p), da se bo žogica ustavila v skrajno desni škatli?

... da lahko dobite 1 enoto
(na primer 1 €)



(... ampak ne izberete!) – kakšna je verjetnost? (p)

Koliko bi stavili?



$$P(K = k | n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

Binomski koeficient

$$\frac{n!}{k!(n-k)!}$$

Število "R" (uspehov, **k**)

Število poskusov, **n**

P=?

Verjetnost, da gre desno ("right", R) = ? **0.50 = p**

Verjetnost **rdeče poti** = ?

neodvisna dogodka

$$P(R \wedge R \wedge L) = P(R) P(R) P(L) = p \cdot p \cdot (1-p) = p^2 (1-p)^1 = 0.50 \cdot 0.50 \cdot 0.50 = .125$$

nezdružljivi

$$P(3 \cdot \text{škatla}) = P(\text{pot1} \vee \text{pot2} \vee \text{pot3}) = P(\text{pot1}) + P(\text{pot2}) + P(\text{pot3}) = 0.125 + ? + ? = 3Pr(\text{pot1})$$

Primer

- Kako lahko izračunamo verjetnost, da ima ženska s tremi otroki 0, 1, 2 ali 3 punčke?
 - Kaj merimo: spol vsakega otroka
 - X_1 =spol prvega otroka
 - X_2 =spol drugega otroka
 - X_3 =spol tretjega otroka
 - Izid: Punčka/Fantek; Da/Ne; 1/0;

Primer (nadaljevanje)

- X =punčka
- $P(X=1)=P(\text{punčka})=p$ (=0.48)
- $P(X=0)=P(\text{fantek})=1-p$ (=0.52)
- p je enak za vsak porod in spol vsakega otroka je neodvisen od spola drugih otrok
- $K=X_1+X_2+X_3$: število punčk
- $P(K=k)=?$
 - Možni izidi: $k=0,1,2,3$

	X_1	X_2	X_3	K	P(Out)
Out1	0	0	0	0	$P(X_1=0 \wedge X_2=0 \wedge X_3=0) = P(X_1=0)P(X_2=0)P(X_3=0) = (1-p)(1-p)(1-p) = (1-p)^3$ $P(K=0) = (1-p)^3$
Out2	1	0	0	1	$p(1-p)^2$
Out3	0	1	0	1	$p(1-p)^2$
Out4	0	0	1	1	$p(1-p)^2$
Out5	1	1	0	2	$p^2(1-p)$
Out6	1	0	1	2	$p^2(1-p)$
Out7	0	1	1	2	$p^2(1-p)$
Out8	1	1	1	3	p^3

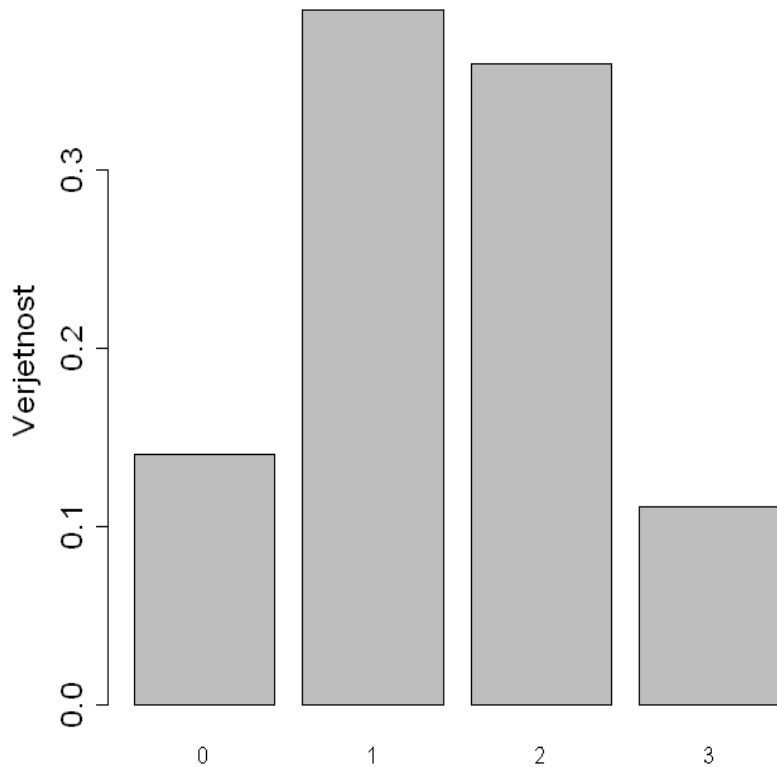
$$P(K=1) = 3p(1-p)^2$$

$$P(K=2) = 3p^2(1-p)$$

$$P(K=3) = p^3$$

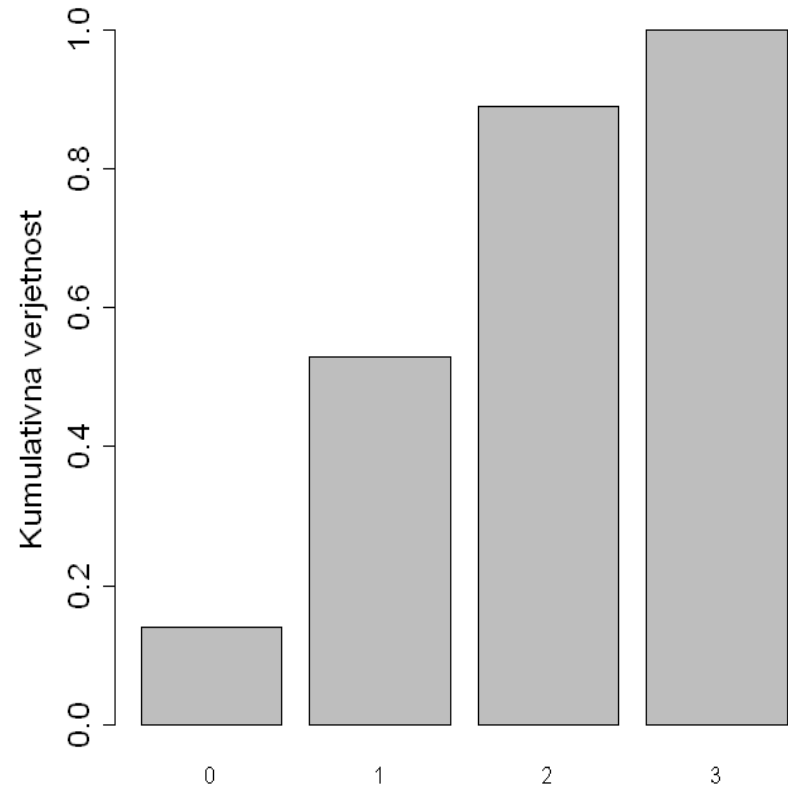
Porazdelitev slučajne spremenljivke K “število punčk za žensko s tremi otroki”

$$P(K=k) = p(k)$$



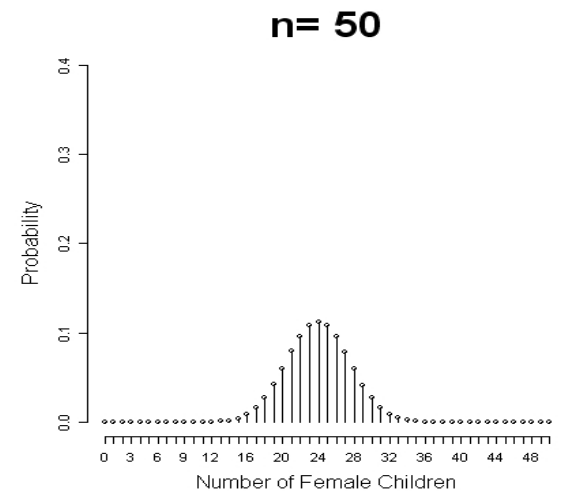
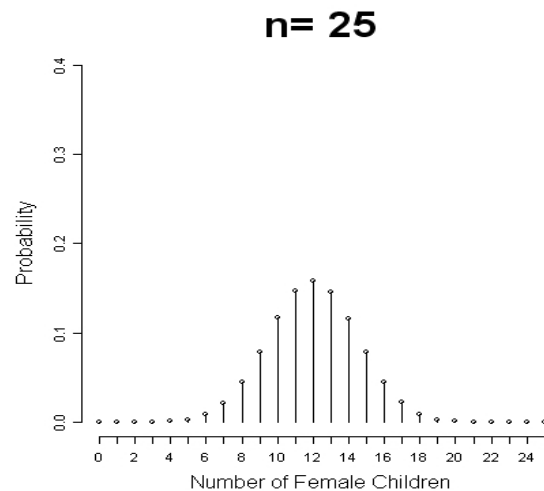
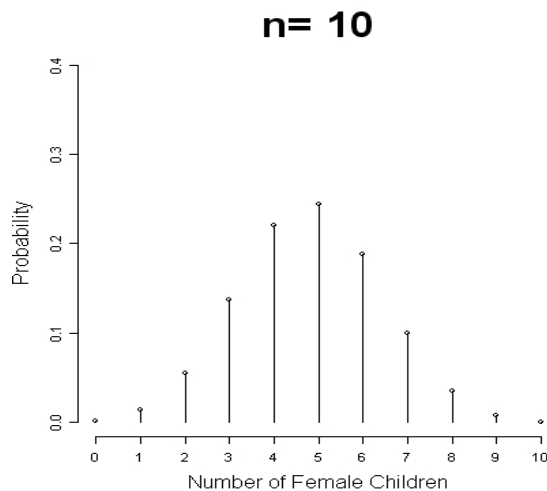
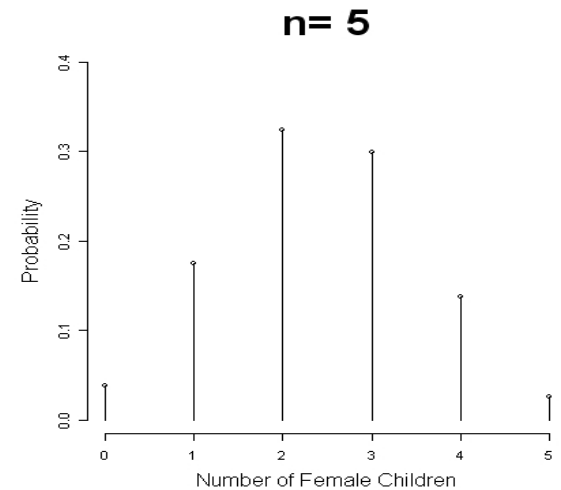
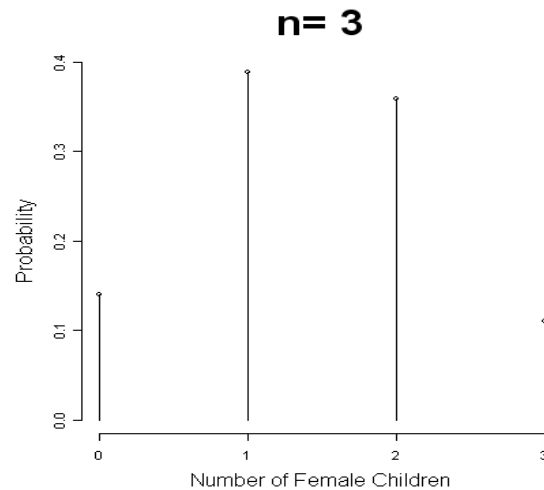
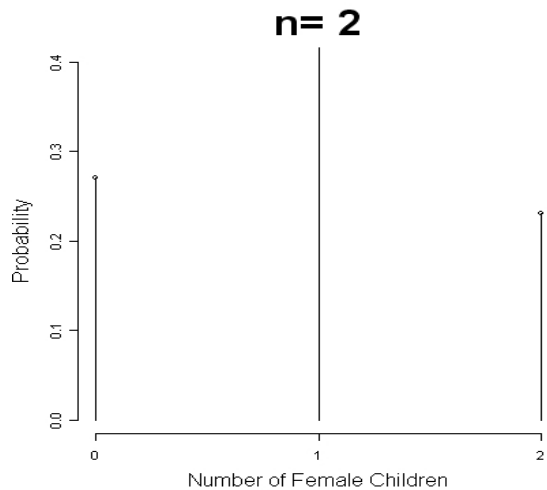
Verjetnostna porazdelitev (stolpični diagram za verjetnosti vsakega izida)

$$P(K \leq k) = F(k)$$



Porazdelitvena funkcija
(stolpični diagram za kumulativne verjetnosti)

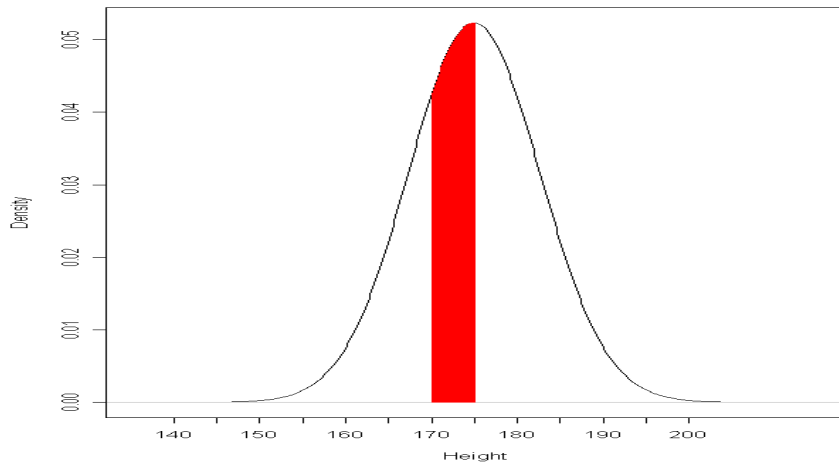
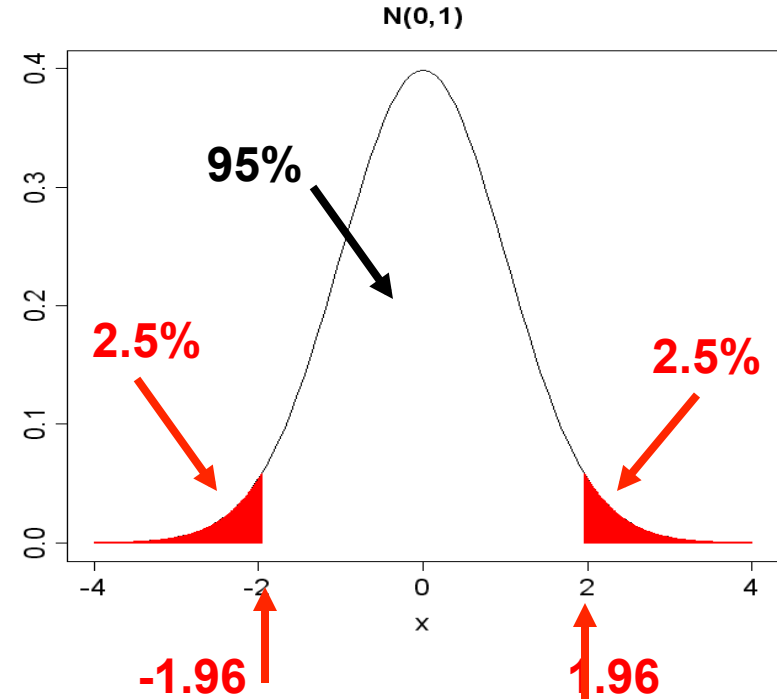
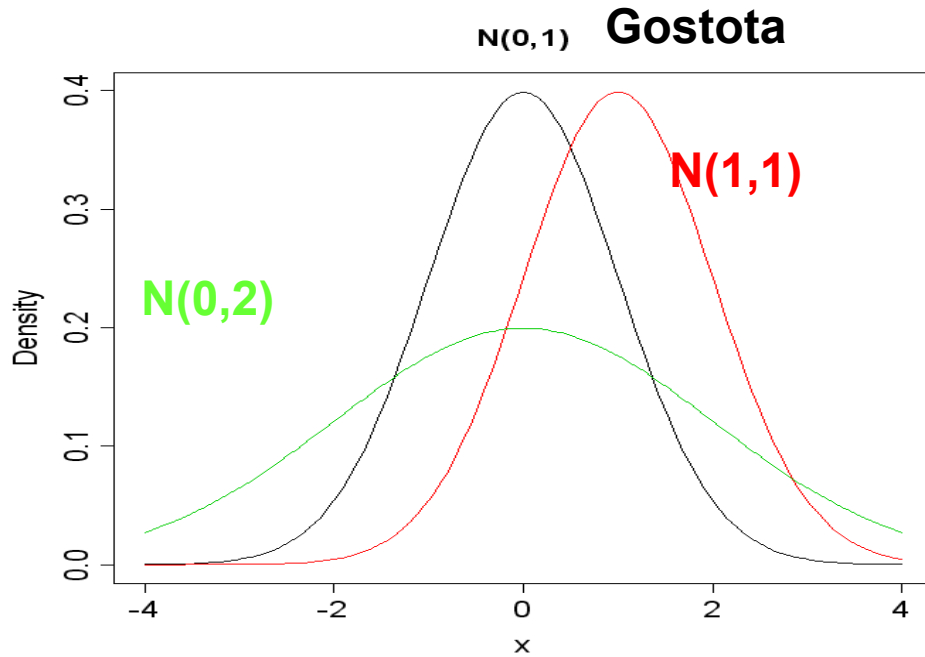
Če povečamo število seštetiH spremenljivk (otroci), kaj se zgodi s porazdelitvijo?



p=0.48

np in n(1-p)>5

Normalna porazdelitev – zvezna



$P(170 < X < 175 | \mu, \sigma) = ?$

2,5 percentil

$$z_{0,025} = -z_{0,975}$$

97,5 percentil

$$z_{0,975} = z_{1-0,05/2}$$

$$z_{0,025} : P(Z \leq z_{0,025}) = 0,025$$

- $X \sim N(\mu, \sigma)$, $Z = (X - \mu) / \sigma$
- $Z \sim N(0, 1)$ standardna normalna porazdelitev

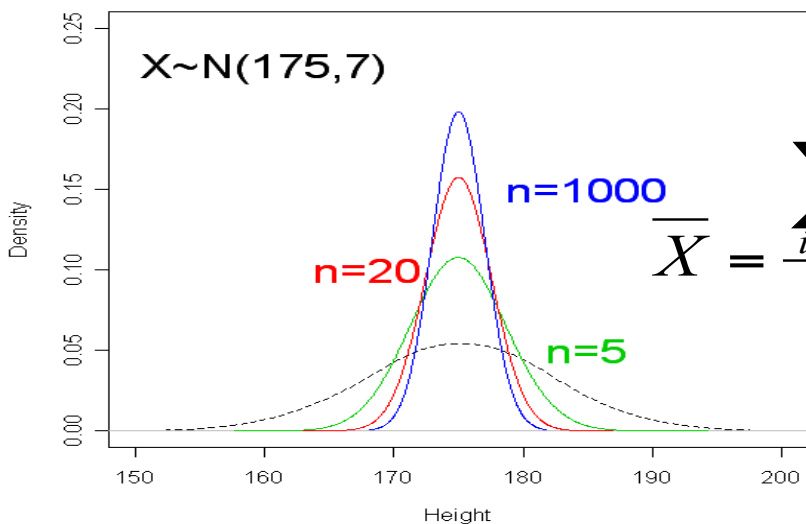
Zakaj je normalna porazdelitev tako pomembna?

Centralni limitni izrek

- Povprečje neodvisnih spremenljivk**, ki imajo enako porazdelitev, se porazdeli z normalno porazdelitvijo (če imamo veliko spremenljivk – velik n)

Neodvisni X_i , $E(X_i)=\mu$, $SD(X_i)=\sigma$

Standardna napaka (standard error of the mean, SE)



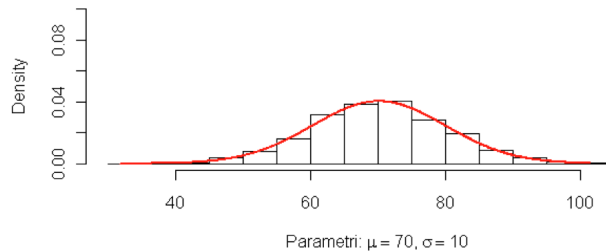
$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} \sim N(\mu, \sigma/\sqrt{n})$$

$$K = \sum_{i=1}^n X_i \sim N(n\mu, \sqrt{n} \sigma)$$

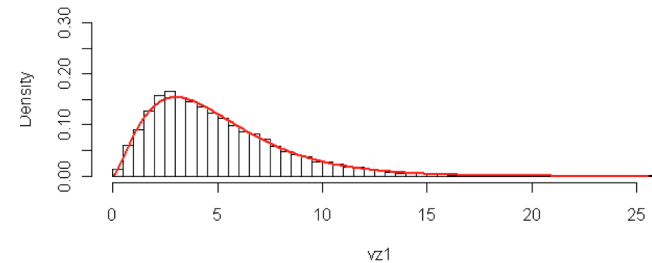
CLI - nadaljevanje

- CLI pojasnjuje, zakaj v praksi tako pogosto srečamo (skoraj) normalno porazdeljene slučajne spremenljivke - **spremenljivka**, na katero vpliva **večje število neodvisnih slučajnih dejavnikov** bo skoraj normalno porazdeljena.

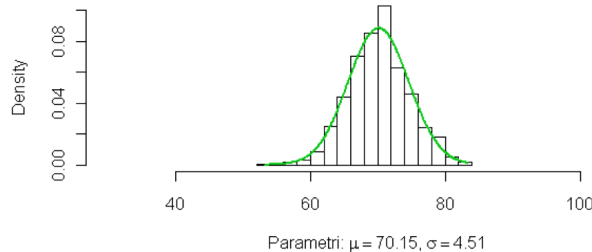
Porazdelitev vseh vzorčnih vrednosti



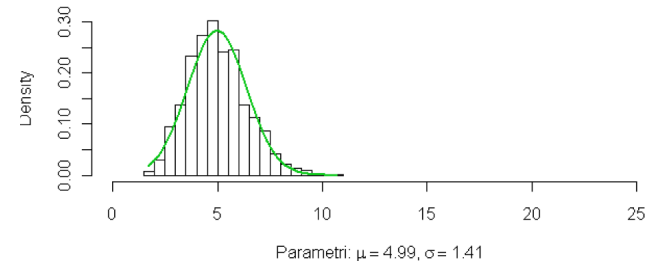
Porazdelitev vseh vzorčnih vrednosti



Porazdelitev vzorčnih povprečij



Porazdelitev vzorčnih povprečij



Raziskovalno vprašanje



športniki



ostali

Povprečje populacije:

μ_s

???

μ_s

=

μ_{ns}

???

Povprečje populacije:

μ_{ns}

Je natančnost raziskave odvisna od velikosti vzorca?

Intervali zaupanja

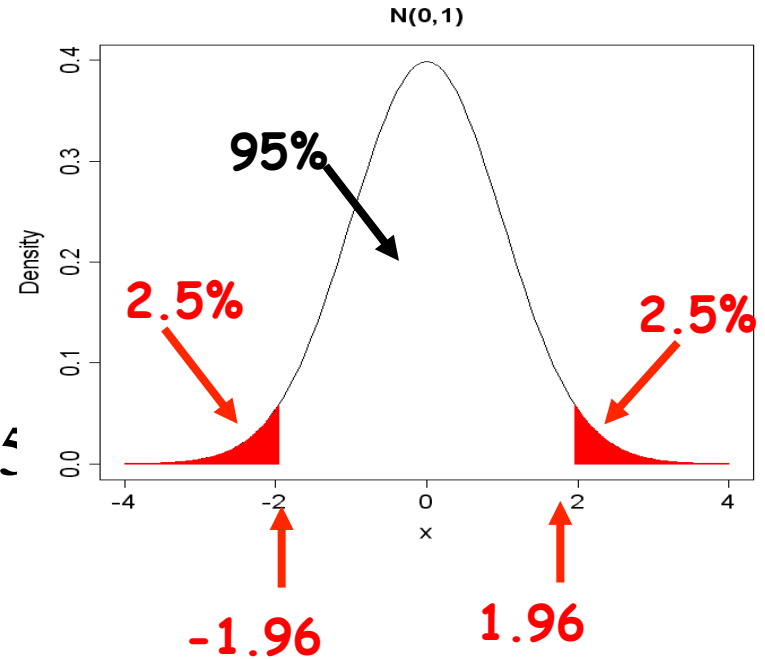
$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

$$P(-1,96 < Z < 1,96) = 0,95$$

$$P\left(-1,96 < \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < 1,96\right) = 0,95$$

$$P\left(\bar{X} - 1,96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1,96 \frac{\sigma}{\sqrt{n}}\right) = 0,95$$

$$P(\bar{X} - 1,96SE < \mu < \bar{X} + 1,96SE) = 0,95$$



95% interval zaupanja za populacijsko povprečje

Primer: interval zaupanja za populacijsko povprečje

- 1, 2, 3, 4, 5, $\sigma=2$, spremenljivka se porazdeli normalno
- Vzorčno povprečje=3
- $n=5$
- $SE = \sigma/\sqrt{n} = 2/\sqrt{5} = 0,89$
- 95% interval zaupanja za μ
 - Spodnji limit $3 - 1,96 * 0,89 = 1,25$
 - Zgornji limit $3 + 1,96 * 0,89 = 4,75$
- Od 1,25 do 4,75

Populacija vs. vzorec

Populacija	Vzorec
N = velikost populacije	n = velikost vzorca
$\mu = \frac{X_1 + X_2 + \dots + X_N}{N}$ <i>populacijsko povprečje</i>	$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$ <i>vzorčno povprečje</i>
$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$ <i>populacijska varianca</i>	$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ <i>vzorčna varianca</i>
$\sigma = \sqrt{\sigma^2}$ <i>populacijski standardni odklon</i>	$s = \sqrt{s^2}$ <i>vzorčni standardni odklon</i>