Statističko zaključivanje

15.12.2023.

Rosa Karlić

Naputci za dobre grafove

- Ne zatrpavajte graf nepotrebnim dodacima (uzorci, 3D efekti, nepotrebne legende) – što jednostavnije to bolje
- Uključite sve potrebne informacije (označite obje osi). Simboli, boje i uzorci trebaju biti definirani u legendi (na slici ili ispod slike). Rezultati na grafu moraju se moći razumjeti bez čitanja glavnog teksta.
- Budite oprezni s korištenjem boja u grafovima (trebaju biti podobne za crno-bijelo printanje i za daltoniste) .
- Nemojte pokušavati prikriti ili krivo interpretirati rezultate.

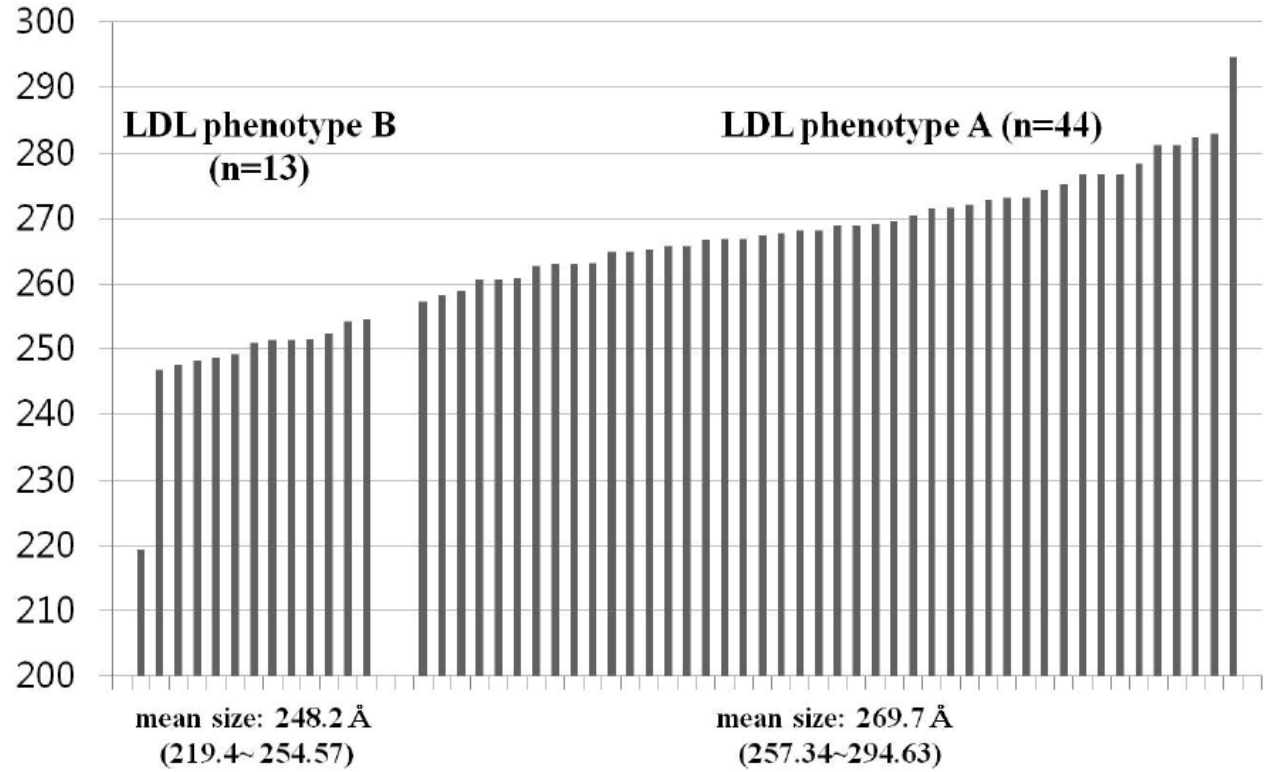


Fig. 1. Distribution of low-density lipoprotein (LDL) particle size in all study subjects (LDL phenotypes A and B). *LDL phenotype A group* (mean size: 269.7 Å, n = 44), subjects with buoyant-mode profiles [peak LDL particle diameter $\geq$ 264 Å] including intermediate LDL subclass pattern [256 Å $\leq$ peak LDL particle diameter $\leq$ 263 Å]; *LDL phenotype B group* (mean size: 248.2 Å, n = 13), subjects with dense-mode profiles [peak LDL particle diameter $\leq$ 255 Å]

https://www.biostat.wisc.edu/~kbroman/topten_worstgraphs/

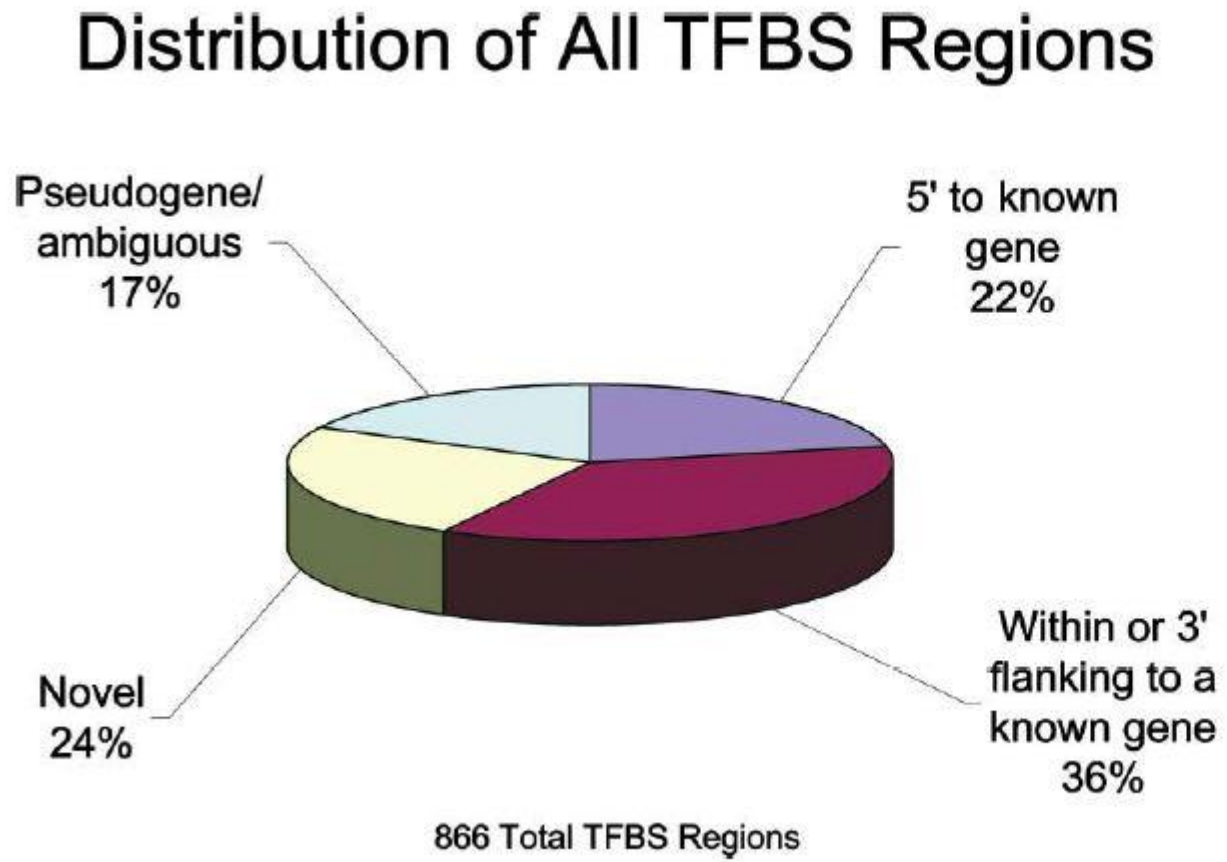


Figure 1. Classification of TFBS Regions

TFBS regions for Sp1, cMyc, and p53 were classified based upon proximity to annotations (RefSeq, Sanger hand-curated annotations, GenBank full-length mRNAs, and Ensembl predicted genes). The proximity was calculated from the center of each TFBS region. TFBS regions were classified as follows: within 5 kb of the 5′ most exon of a gene, within 5 kb of the 3′ terminal exon, or within a gene, novel or outside of any annotation, and pseudogene/ambiguous (TFBS overlapping or flanking pseudogene annotations, limited to chromosome 22, or TFBS regions falling into more than one of the above categories).

https://www.biostat.wisc.edu/~kbroman/topten_worstgraphs/

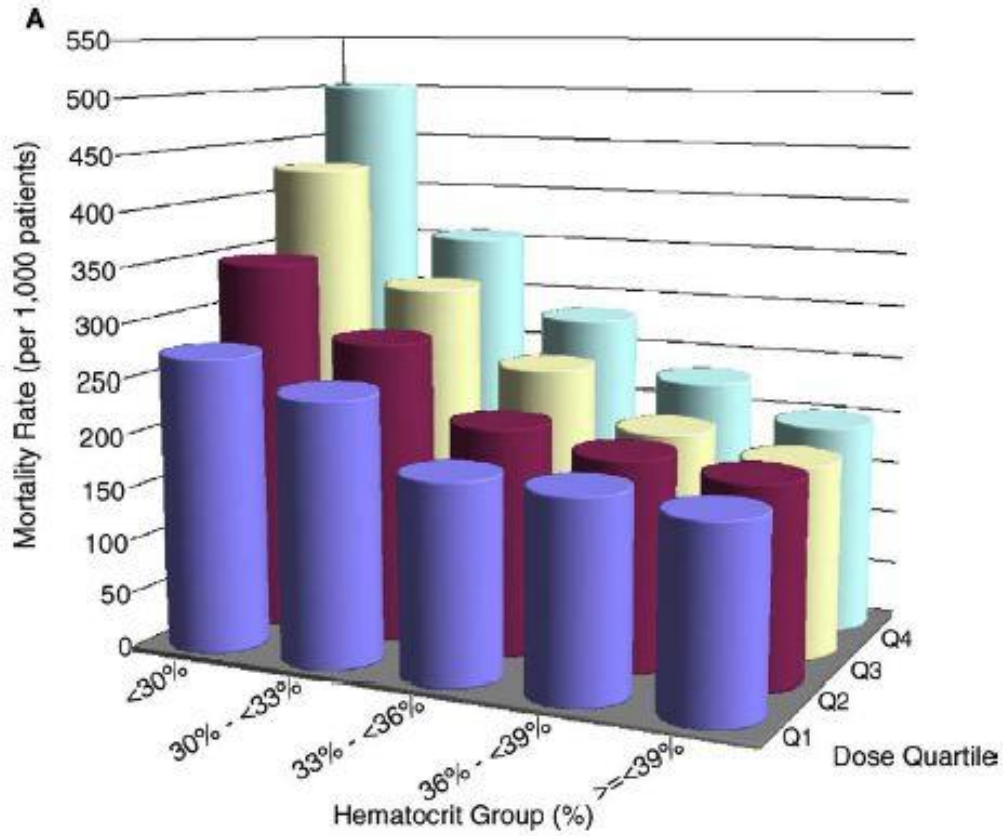
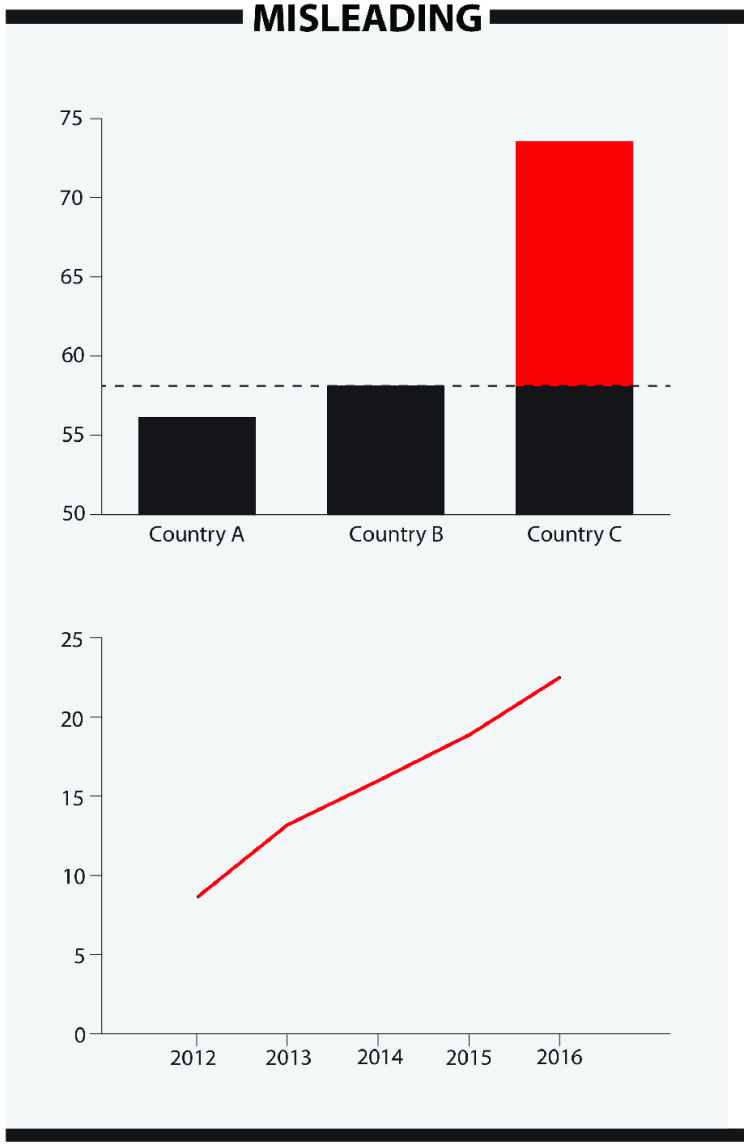
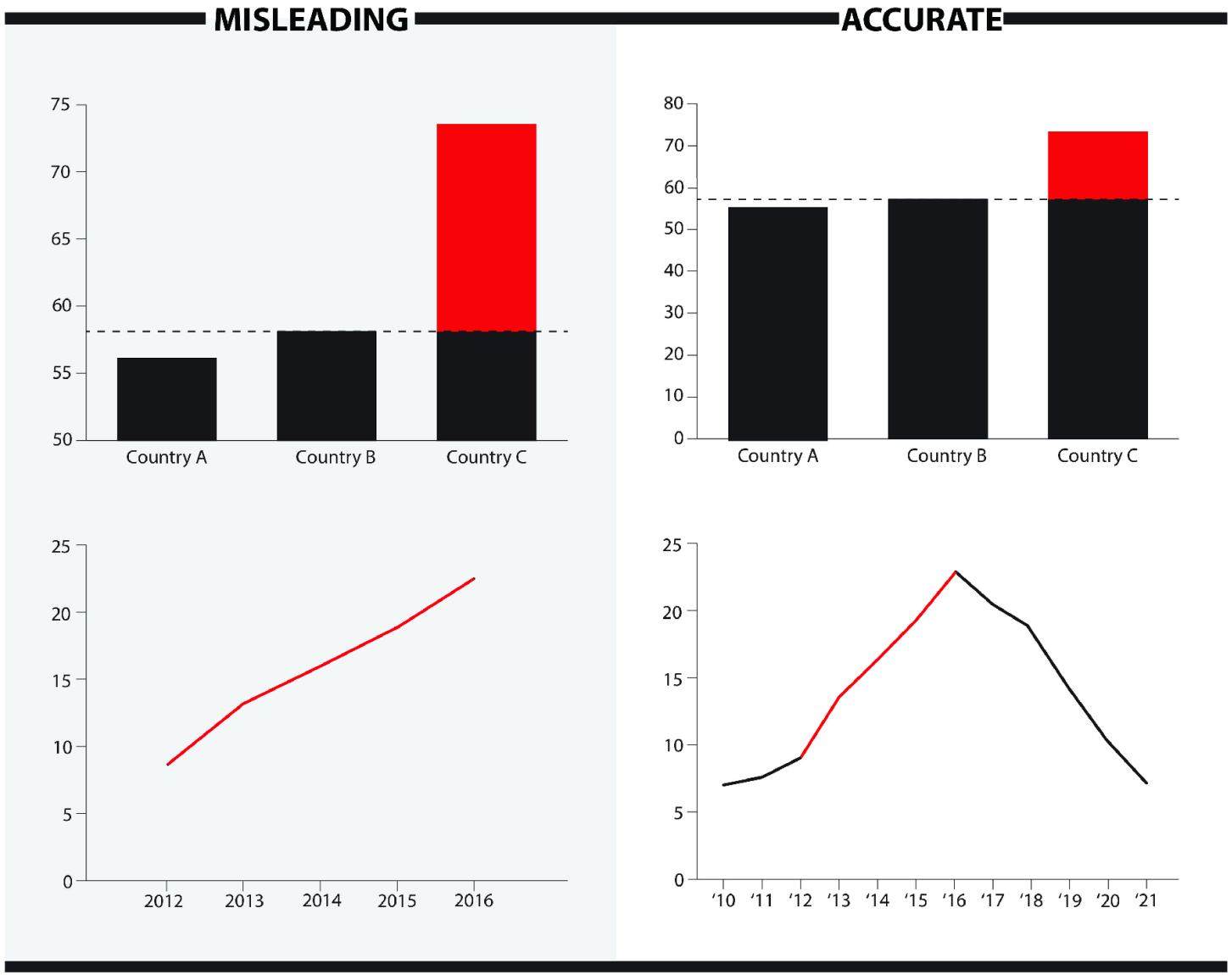


Fig. 2. (A) Unadjusted 1-year mortality rates by hematocrit group disaggregated by epoetin dose quartile. Within each epoetin dose quartile, there is a trend toward increasing mortality as the observed study hematocrit decreases, most notably in the fourth quartile (>21,692 units/wk). Similarly, there is a trend toward increasing mortality as the epoetin dose increases within each observed study hematocrit range, most notably in the lowest (<30%) hematocrit range. (B) Relative risk of death by hematocrit group

https://www.biostat.wisc.edu/~kbroman/topten_worstgraphs/


MISLEADING
75
70
65
60
55
50
Country A
Country B
Country C
25
20
15
10
5
0
2012
2013
2014
2015
2016


MISLEADING
ACCURATE
75
70
65
60
55
50
Country A
Country B
Country C
80
70
60
50
40
30
20
10
0
Country A
Country B
Country C
25
20
15
10
5
0
2012
2013
2014
2015
2016
25
20
15
10
5
0
'10
'11
'12
'13
'14
'15
'16
'17
'18
'19
'20
'21

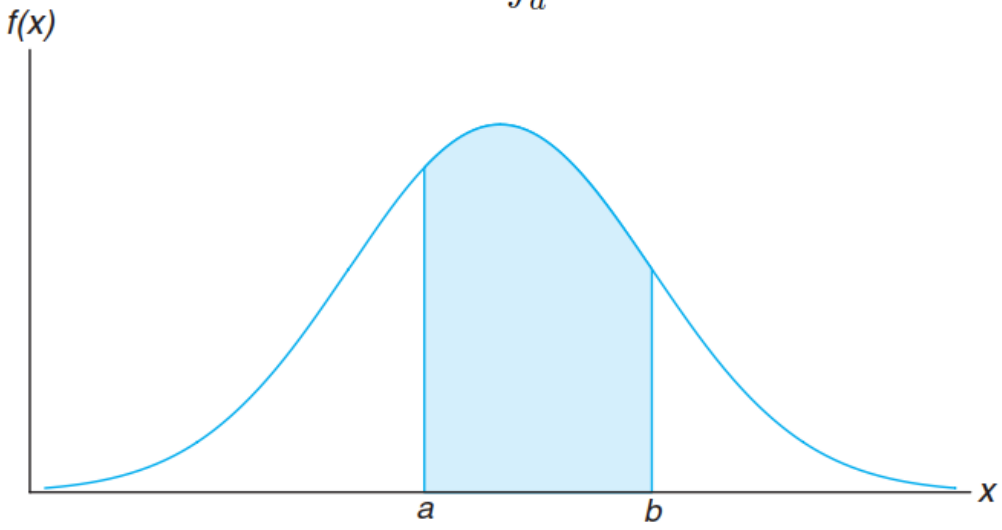
Slučajna varijabla

- Slučajne varijable mogu biti:
    - Diskretne
        - Starost pacijenata u godinama X = {0, 1, 2, 3, …, N}
        - Broj mutacija u DNA lancu X = {0, 1, 2, 3, …, N}
    - Kontinuirane
        - Visina pacijenta X = (0, M)
        - Tjelesna temperatura pacijenta X = (M, N)
- Opisujemo ih distribucijama vjerojatnosti – vjerojatnost da će slučajna varijabla poprimiti određenu vrijednost ili se nalaziti u određenom intervalu

Kontinuirane slučajne varijable

- Mogu poprimiti beskonačno mnogo vrijednosti
- Područje vrijednosti – interval na brojevnom pravcu ili cijeli brojevni pravac
- Određujemo vjerojatnost da će se vrijednost kontinuirane slučajne varijable nalaziti unutar nekog intervala (vjerojatnost da će poprimiti točno određenu vrijednost je 0)
- Opisuju se funkcijama gustoće vjerojatnosti

$$P(a < X < b) = \int_a^b f(x)\,dx.$$

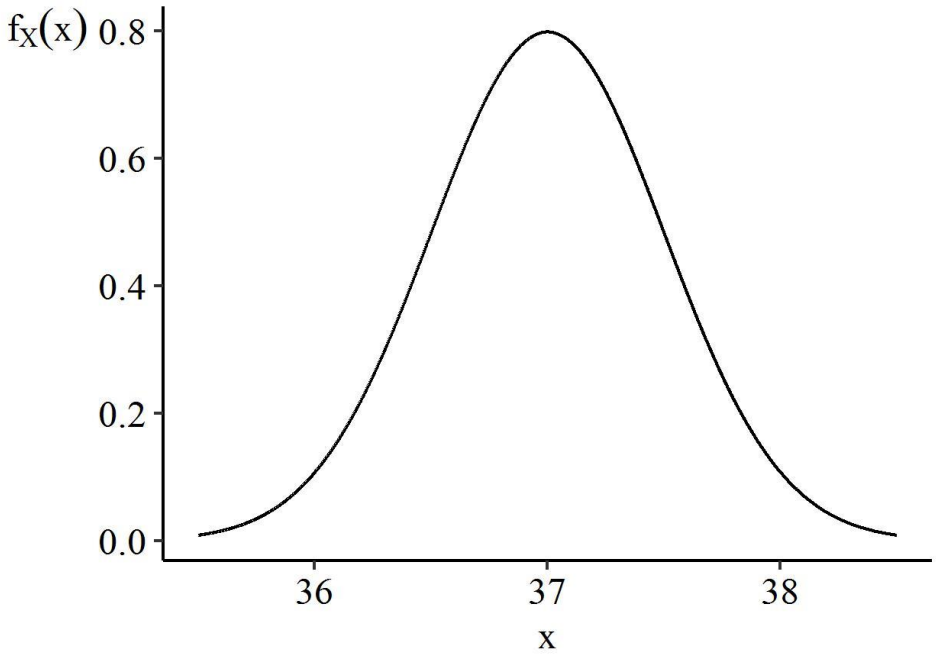


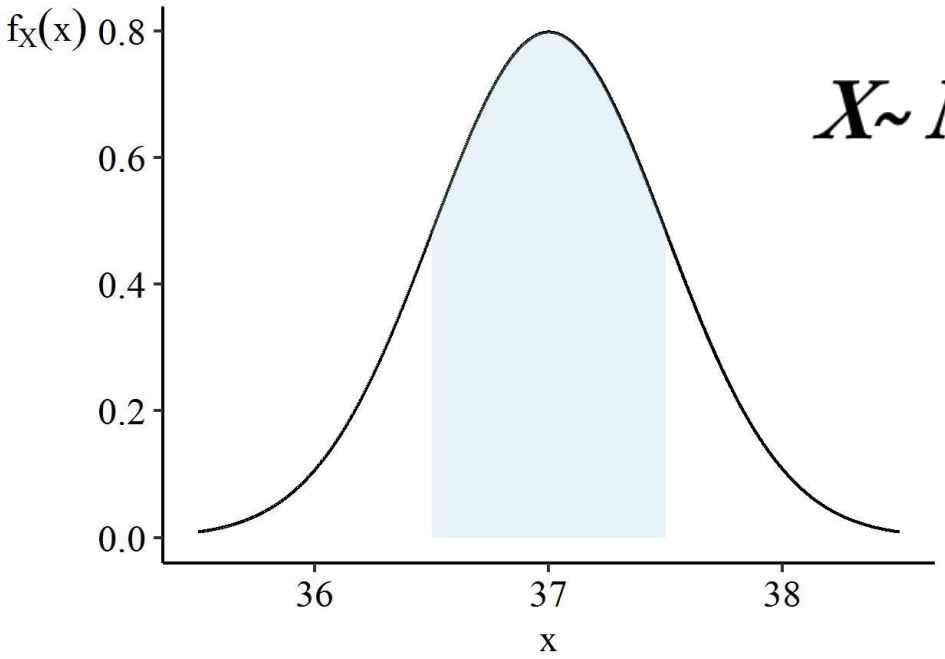
Walpole, R. E., Myers, R. H., Myers, S. L. and Ye, K. (2012)
Probability and Statistics for Engineers and Scientists

Kontinuirane slučajne varijable - primjer

- Tjelesna temperatura ispitanika
- μ (srednja vrijednost) and σ (standardna devijacija) određuju lokaciju i oblik distribucije

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$





$$X \sim N(\mu, \sigma^2)$$

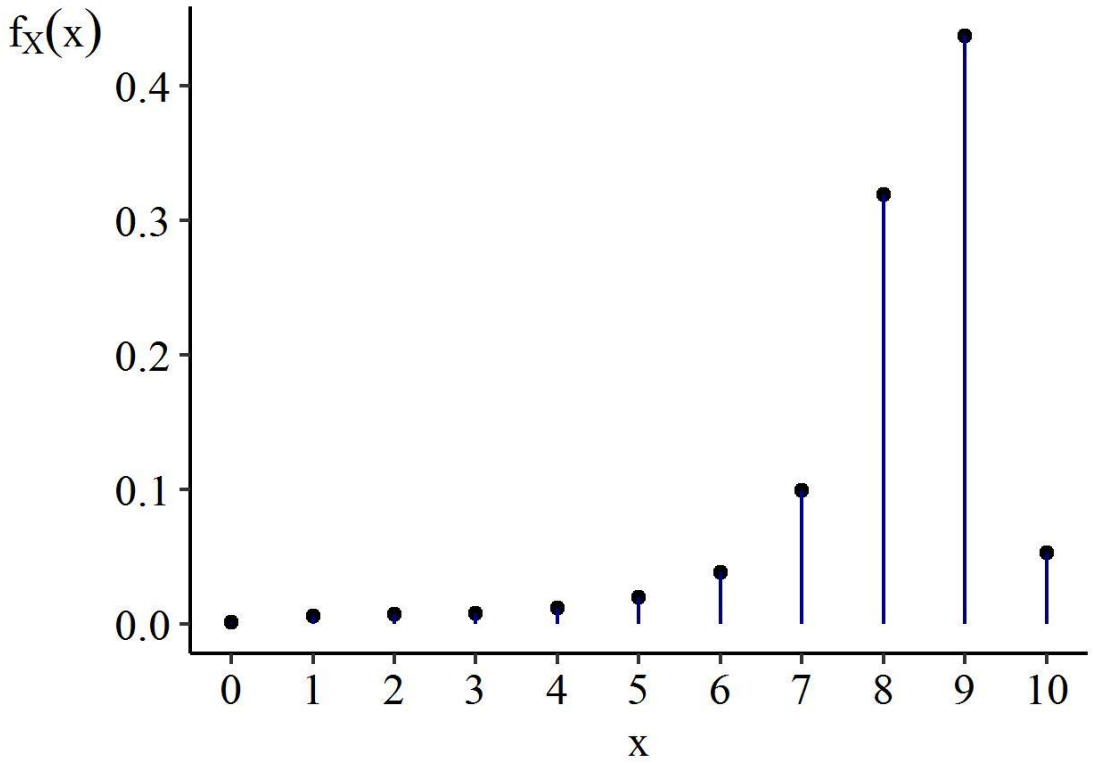
Diskretne slučajne varijable

- Mogu poprimiti prebrojivo mnogo diskretnih vrijednosti
- Svaka vrijednost ima konačnu vjerojatnost
- Opisuju se funkcijama mase vjerojatnosti

$$f_X(x) = P(X = x)$$

Diskretne slučajne varijable - primjer

- Apgar ocjena



$F_x(x)$	x
0.001	0
0.006	1
0.007	2
0.008	3
0.012	4
0.020	5
0.038	6
0.099	7
0.319	8
0.437	9
0.053	10

Kumulativna funkcija distribucije

- Cumulative distribution function (CDF)
- Vjerojatnost da je vrijednost slučajne varijable *X* manja ili jednaka od *x*

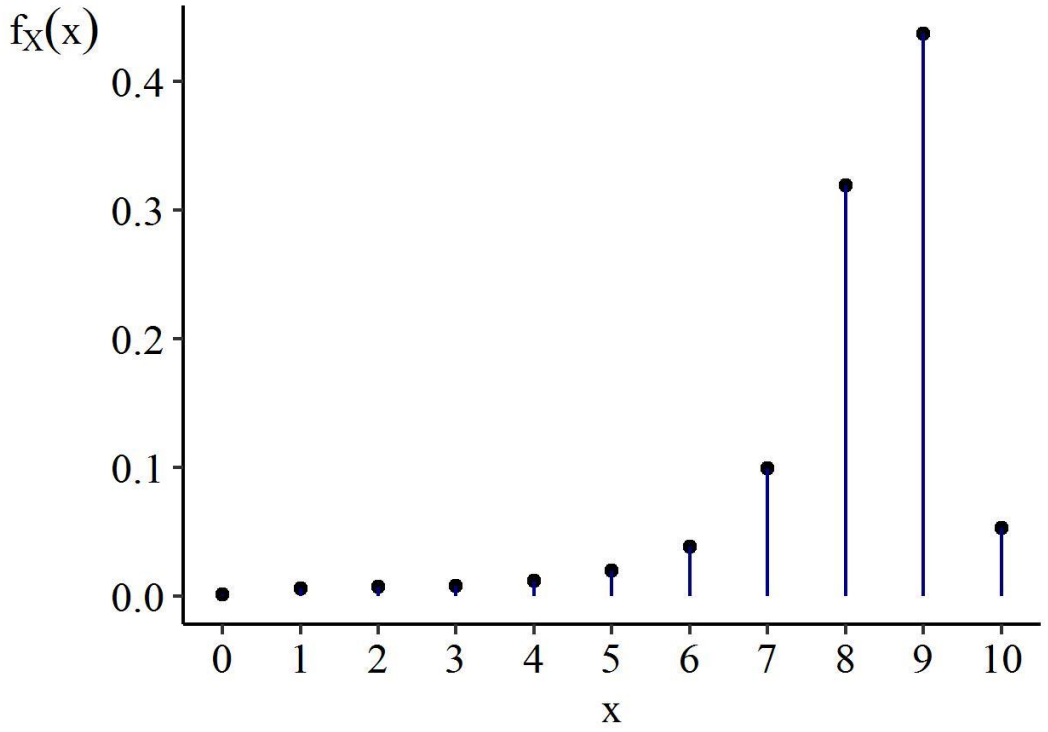
$$F(x) = P(X \leq x)$$

Kumulativna funkcija distribucije

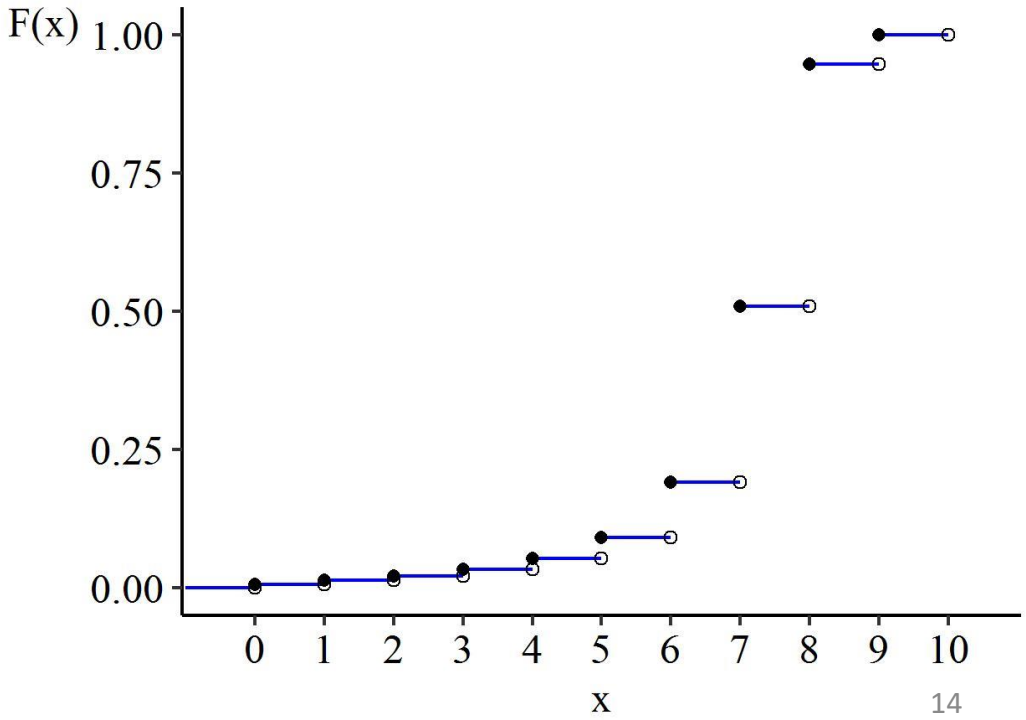
- Diskretne slučajne varijable
- Monotono se povećava od 0 do 1

$$F(x) = P(X \leq x)$$

PMF



CDF

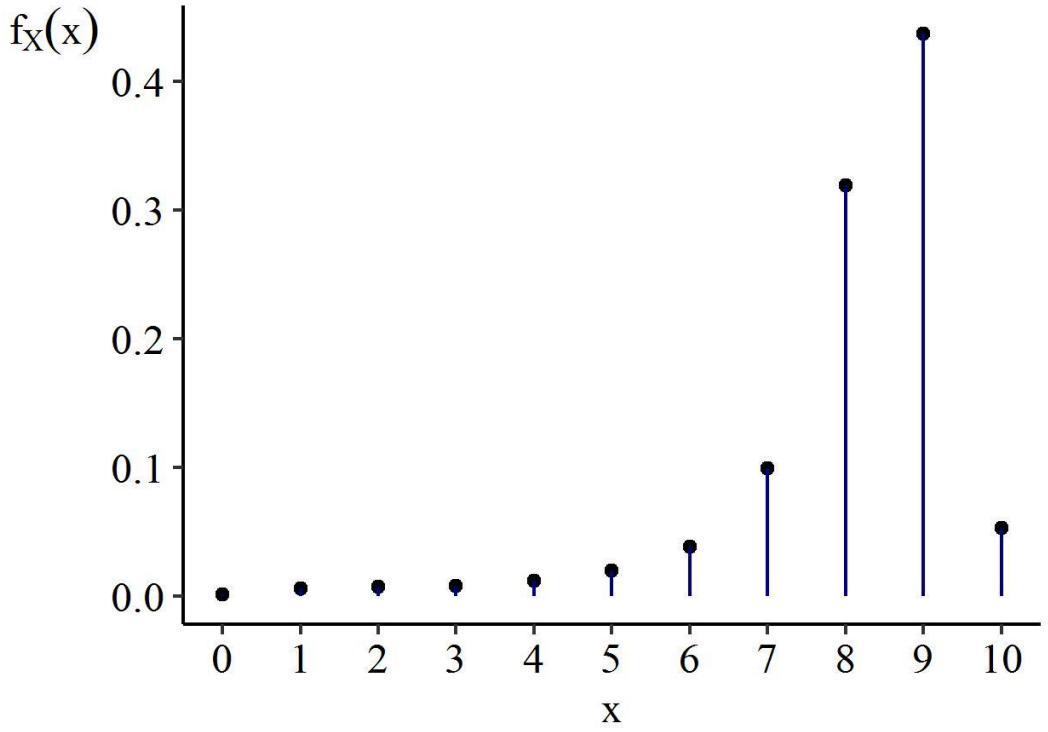


Kumulativna funkcija distribucije

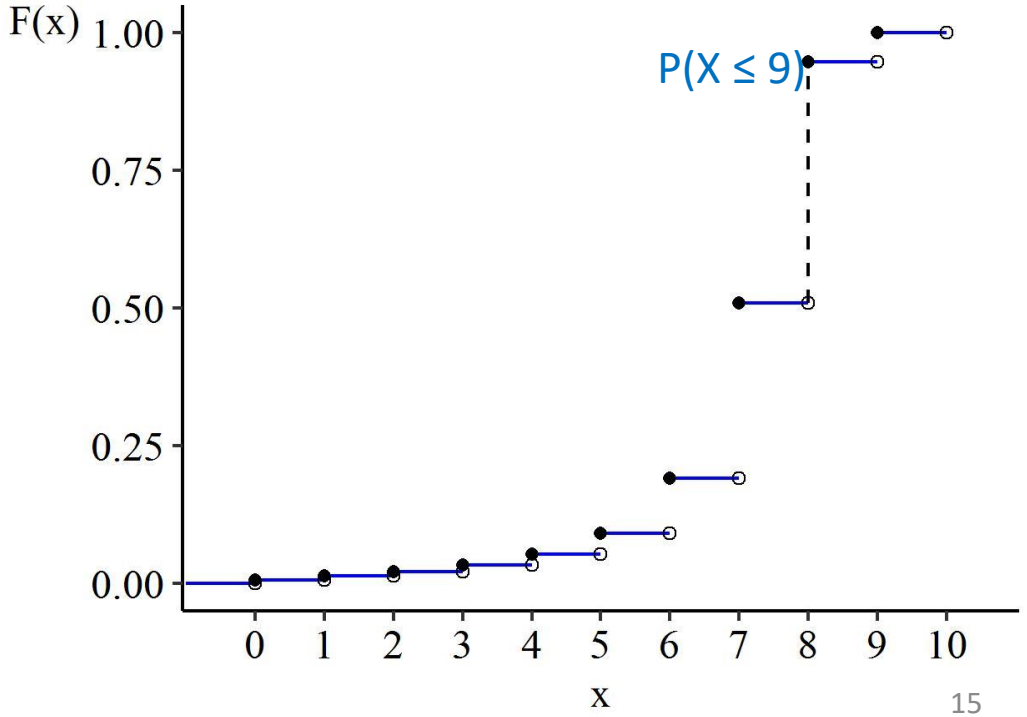
- Diskretne slučajne varijable
- Monotono se povećava od 0 do 1

$$F(x) = P(X \leq x)$$

PMF



CDF

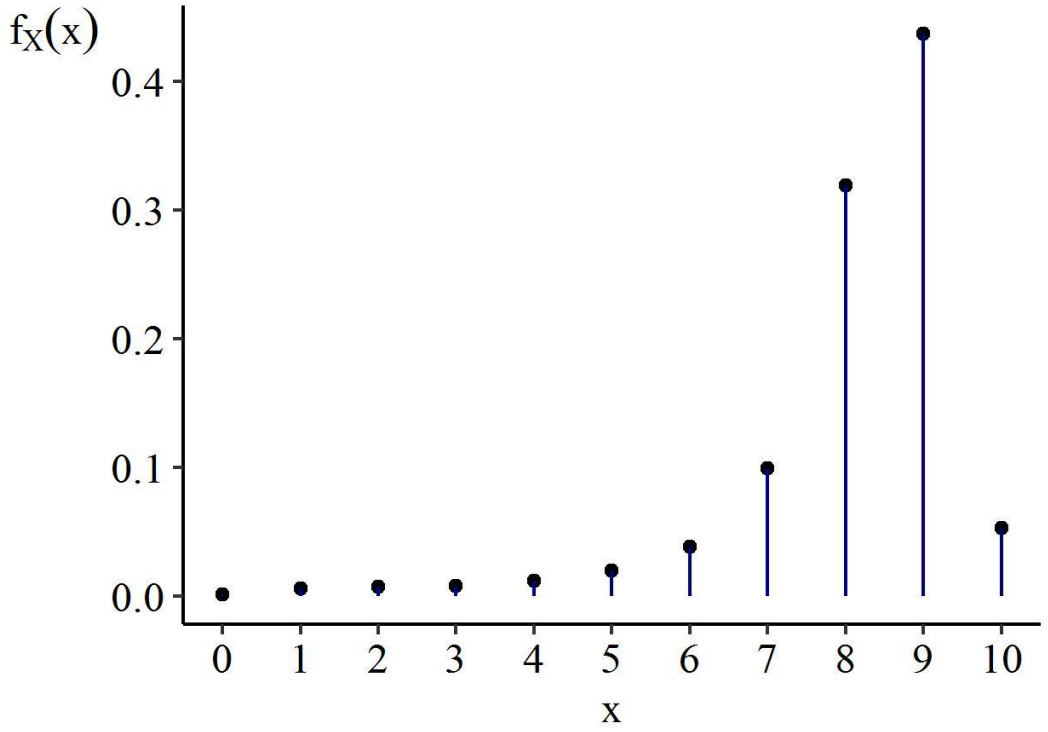


Kumulativna funkcija distribucije

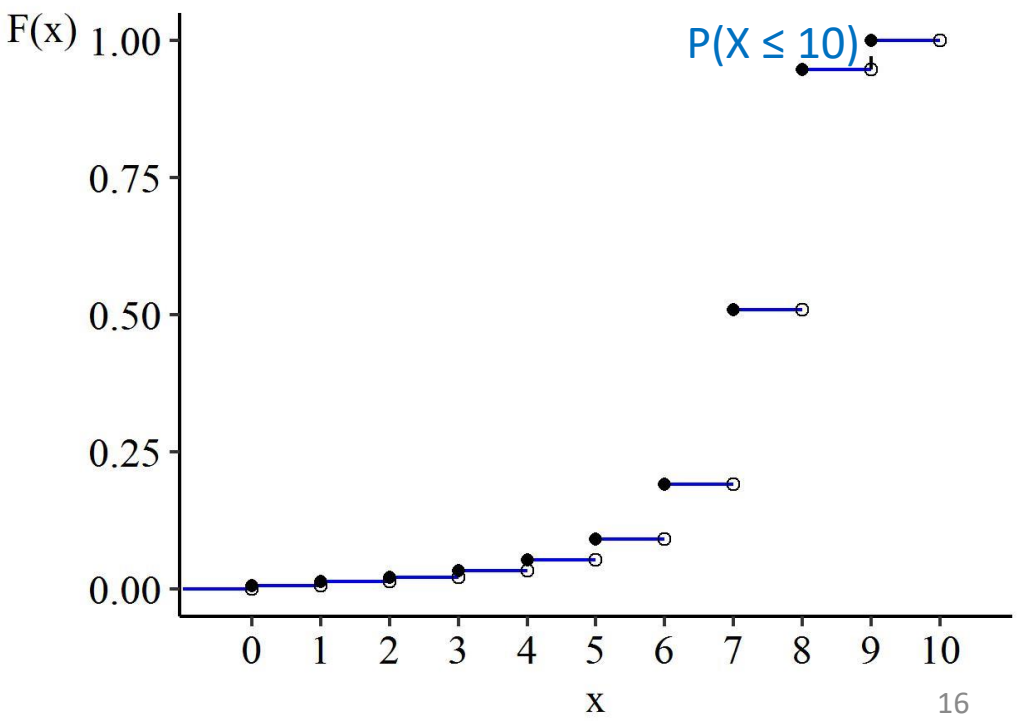
- Diskretne slučajne varijable
- Monotono se povećava od 0 do 1

$$F(x) = P(X \leq x)$$

PMF



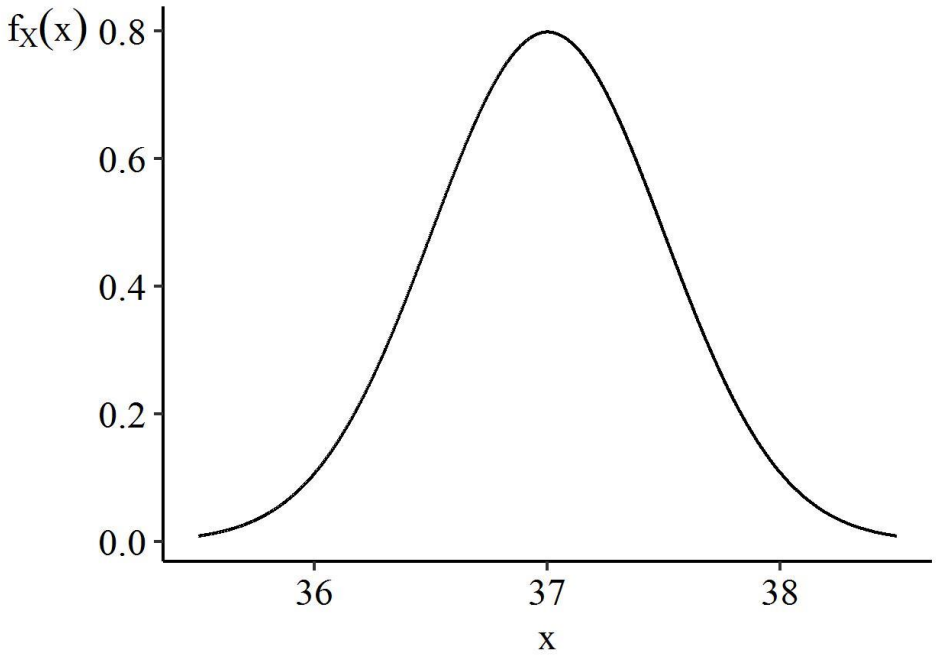
CDF

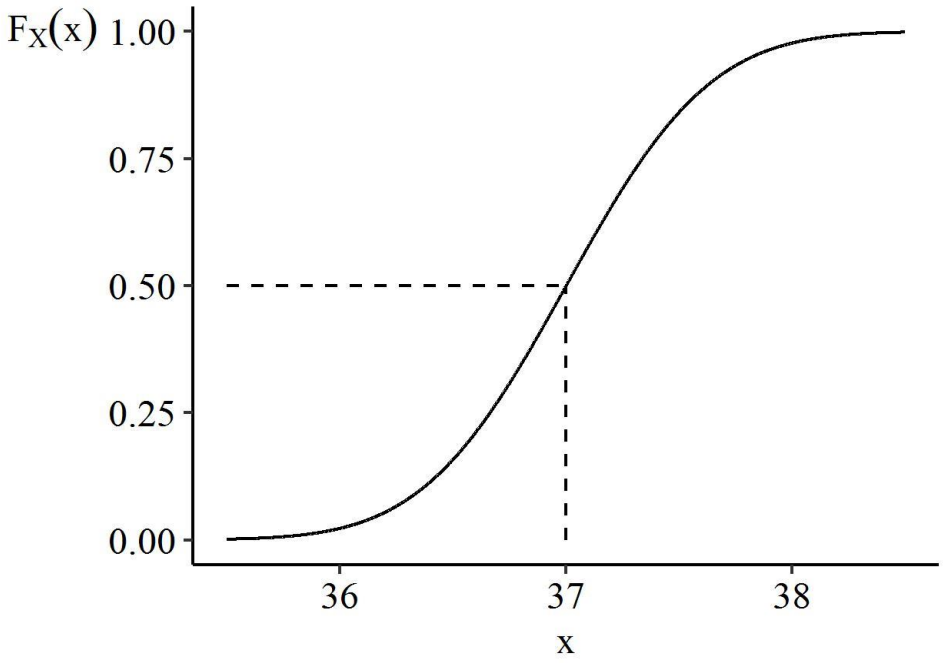


Kumulativna funkcija distribucije

- Kontinuirane slučajne varijable
- Monotono se povećava od 0 do 1

$$F(x) = P(X \leq x)$$





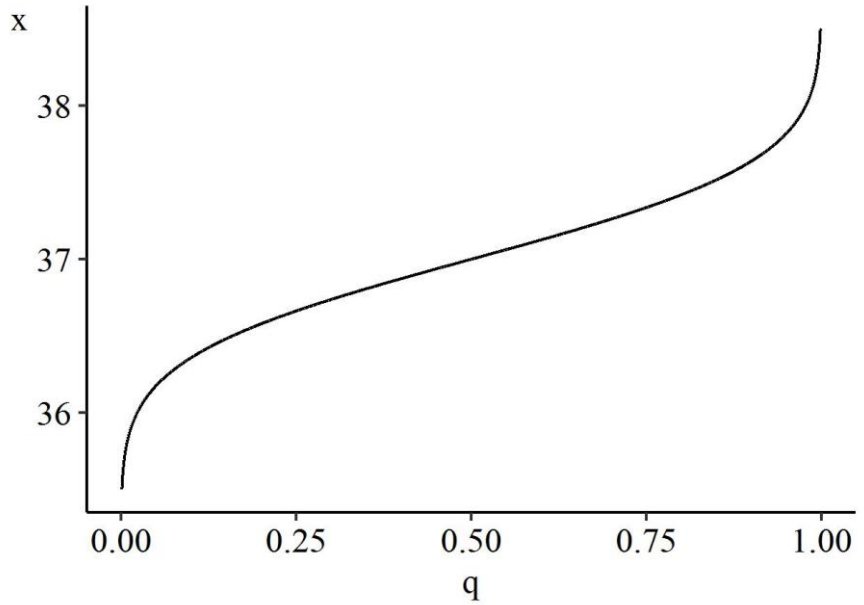
Kvantilna funkcija (Inverse CDF)

- Ako je X slučajna varijabla sa CDF F. Kvantilna funkcija (inverse CDF) je definirana kao:

za $$F^{-1}(q) = \inf\left\{x : F(x) \geq q\right\}$$

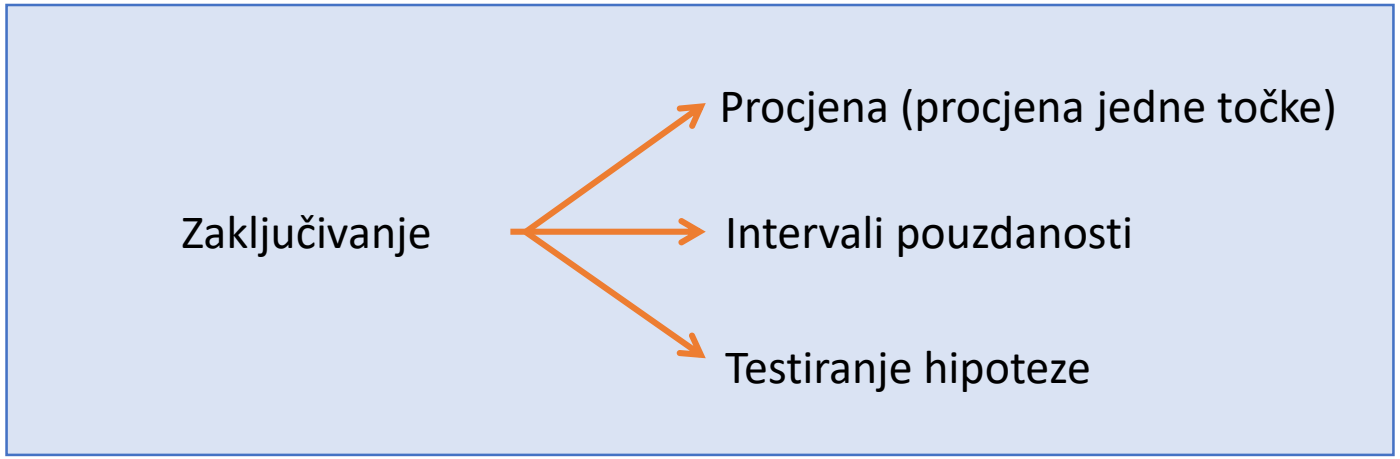
$$q \in [0,1]$$

- $F^{-1}(1/4)$ – prvi kvartil
- $F^{-1}(1/2)$ - medijan
- $F^{-1}(3/4)$ – treći kvartil



Statističko zaključivanje

- Proces korištenja podataka za donošenje zaključaka o distribuciji koja je generirala podatke ili nekoj značajki te distribucije, kao što je srednja vrijednost
- Dominantni pristupi: frekventističko zaključivanje i Bayesovo zaključivanje



Procjena parametara

Zaključujemo nešto o populaciji na temelju informacija dobivenih iz uzorka

- Statistike se koriste kao procjene parametara

- Procjena u jednoj točki
- Procjena u obliku intervala

Intervali pouzdanosti

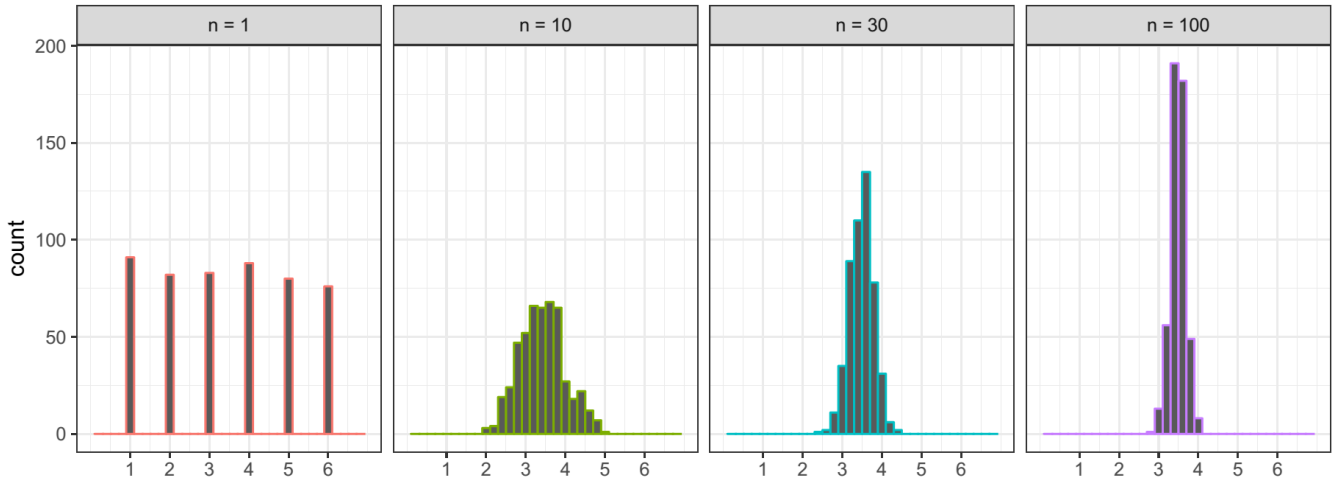
- Koristi se za izražavanje preciznosti i nesigurnosti povezanih s određenom metodom uzorkovanja.
- Sastoji se od tri dijela:
  - Razina pouzdanosti - opisuje nesigurnost metode uzorkovanja
  - Statistika
  - Margina greške

Definiraju procjenu intervala koji opisuje preciznost metode

- Razina pouzdanosti - koliko čvrsto vjerujemo da će neka metoda uzorkovanja proizvesti interval pouzdanosti koji uključuje stvarni parametar populacije.

Centralni granični teorem

- Distribucija procjena statistika zbroja ili prosjeka i.i.d. slučajne varijable bit će normalne ili gotovo normalne ako je veličina uzorka dovoljno velika



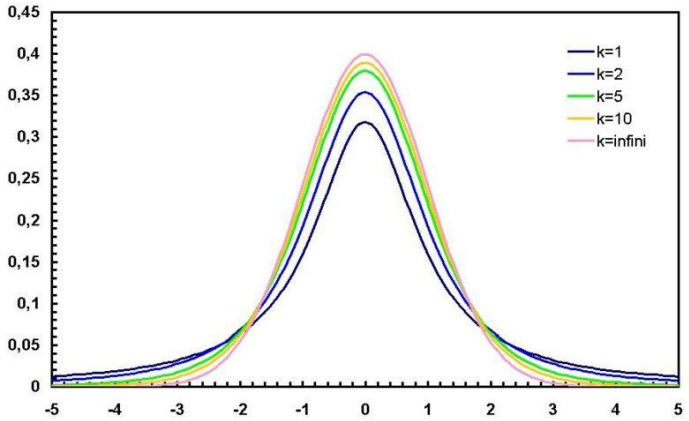
- Statistiku možemo izraziti kao t-vrijednost ili kao z-vrijednost (koristite t-vrijednostkada je veličina uzorka mala ili je standardna devijacija populacije nepoznata)
  - Standardna devijacija statistike
  - Standardna greška statistike

$$\sigma_{\bar{x}} = \sigma / \sqrt{n}$$

$$SE_{\bar{x}} = s / \sqrt{n}$$

Studentova t-distribucija

- Obitelj sličnih distribucija vjerojatnosti.
- Specifična t-distribucija ovisi o stupnjevima slobode, v.
- Stupnjevi slobode – broj neovisnih opažanja u uzorku podataka koji su dostupni za procjenu parametra populacije iz koje je taj uzorak izvučen
- Srednja vrijednost t-distribucije je nula.
- Kako v raste, t-distribucija postaje normalnija.



$$f_\nu(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}(1+\frac{t^2}{\nu})^{-(\nu+1)/2}$$

$$-\infty \le t \le +\infty$$

T-test na jednom uzorku

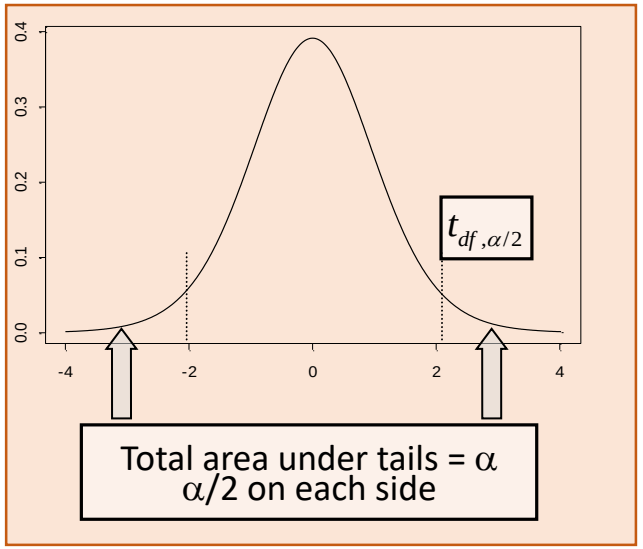
- Pretvorimo statistiku u t-vrijednost:

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}} \qquad \Pr\left[-t_{df,\alpha/2} \leq \frac{\bar{x} - \mu}{s / \sqrt{n}} \leq t_{df,\alpha/2} \right] = 1 - \alpha$$

$$\Pr\left[-t_{df,\alpha/2}(s/\sqrt{n}) \leq \bar{x} - \mu \leq t_{df,\alpha/2}(s/\sqrt{n}) \right] = 1 - \alpha$$

$$\Pr\left[t_{df,\alpha/2}(s/\sqrt{n}) \geq -\bar{x} + \mu \geq -t_{df,\alpha/2}(s/\sqrt{n}) \right] = 1 - \alpha$$

$$\Pr\left[\bar{x} - t_{df,\alpha/2}(s/\sqrt{n}) \leq \mu \leq \bar{x} + t_{df,\alpha/2}(s/\sqrt{n}) \right] = 1 - \alpha$$



$$(1-\alpha)\%\ C.I.: \quad \bar{x} \pm t_{df,\alpha/2} SE$$

$$SE_{\bar{x}} = s / \sqrt{n}$$

Kako konstruirati interval pouzdanosti

- Identificirajte statistiku uzorka.
- Odaberite razinu pouzdanosti.
- Pronađite marginu greške.
  - Margina greške = kritična vrijednost * standardna devijacija statistike
  - Margina greške = kritična vrijednost * standardna greška statistike
- Odredite interval pouzdanosti. Nesigurnost je označena razinom pouzdanosti.

Interval pouzdanosti = statistika uzorka ± margina greške

Margina greške

= kritična vrijednost * standardna devijacija (ili standardna greška) statistike

Primjer

- Želimo procijeniti prosječnu težinu odraslih muškaraca u Zagrebu. Odaberemo nasumičan uzorak 1,000 muškaraca iz populacije od 1,000,000 muškaraca i izvažemo ih. Izmjerili smo da je prosječna težina našeg uzorka 92 kg, a standardna devijacija uzorka je 14kg. Izračunajte 95% interval pouzdanosti.

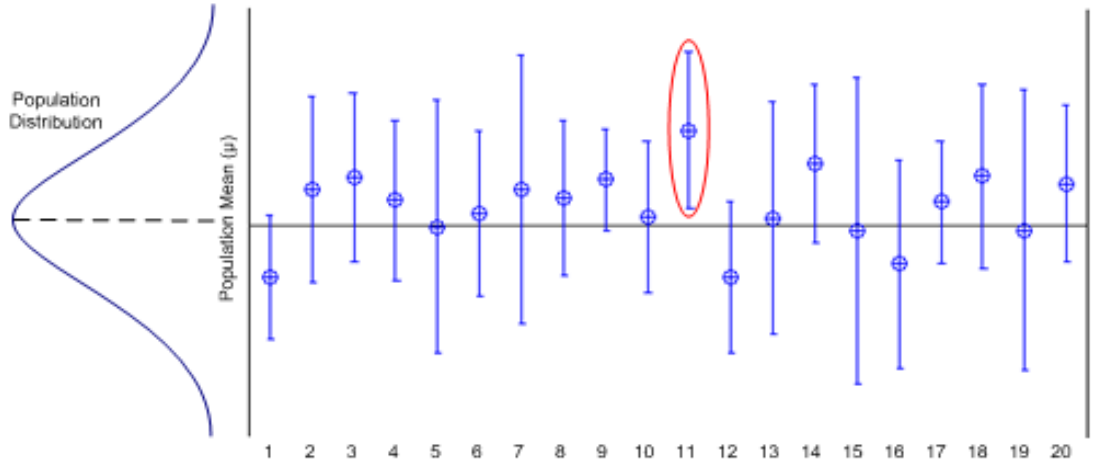
$\alpha = 0.05;\ \bar{x} = 92;\ s = 14;\ n = 1000;\ df=999$

df – degrees of freedom (stupnjevi slobode)

95% interval pouzdanosti: $92 \pm t_{999,0.025}(14/\sqrt{1000}) = 92 \pm 0.87$

Interpretacija intervala pouzdanosti

- Kad bi se uzeli ponovljeni uzorci i izračunao interval pouzdanosti od 95% za svaki uzorak, 95% intervala sadržavalo bi srednju vrijednost populacije.



- Standardna devijacija – koliko su raspršena mjerenja
- Standardna greška – preciznost procjene određene mjere
- Interval pouzdanosti – preciznost metode uzorkovanja