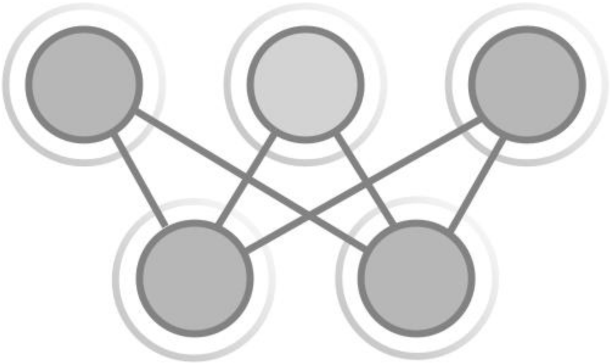
Prof.dr.sc. Bojana Dalbelo Bašić

Fakultet elektrotehnike i računarstva
Zavod za elekroniku, mikroelektroniku, računalne i inteligentne sustave

www.zemris.fer.hr/~bojana
bojana.dalbelo@fer.hr

Uvod u statističko zaključivanje



FER
rujan 2009.

UVOD U STATISTIČKO ZAKLJUČIVANJE

riječ STATISTIKA (lat. *status* = stanje)
riječ VARIJABLA (engl. *variable* = *variate* = *factor*)

Statistika

- deskriptivna
- inferencijalna (*intervalne procjene, testiranje hipoteza*)

- univarijatna
- bivarijatna
- multivarijatna (testiranje, eksporativna statistika)

- parametarska/neparametarska

Varijable:

- zavisne, kriterijske varijable *(engl. dependent, criterion, response variable)*

- nezavisne, prediktorske varijable *(engl. independent, predictior, controlled, regressor variable)*

UVOD U STATISTIČKO ZAKLJUČIVANJE

Ciljevi statističke analize:

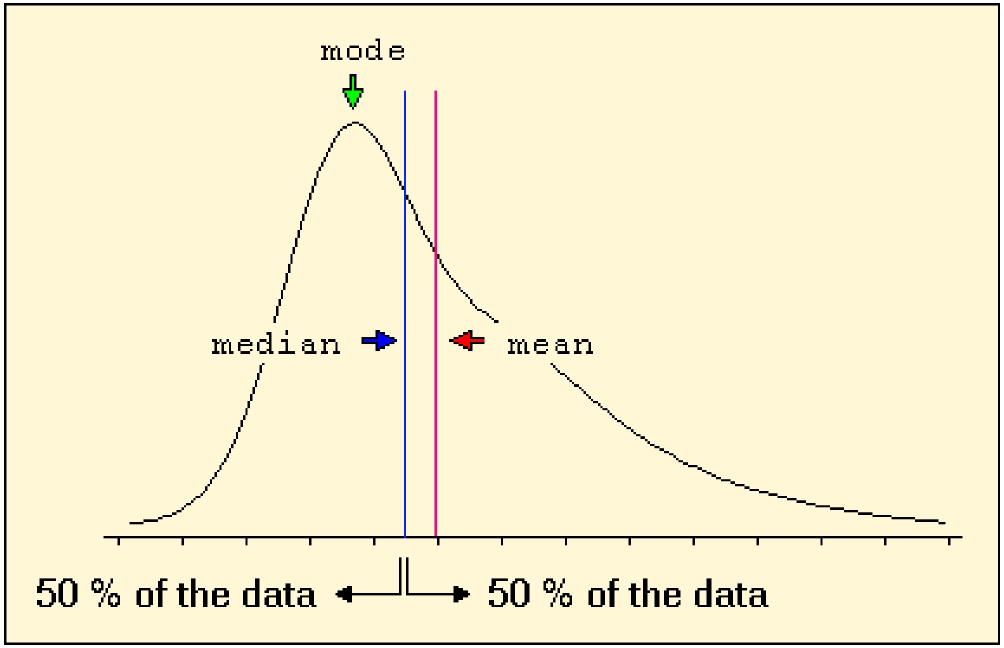
skupljanje podataka → manipulacija podacima → interpretacija

- redukcija podataka
- alat za inferencijalno zaključivanje
- identifikacija relacija ili asocijacija (grupiranja) među podacima

OSNOVNI POJMOVI

- **Objekti**
- **Varijable (obilježja, značajke)**
- **Mjerne skale:**
  - kvalitativne
    - nominalna
    - uređajna (relacija $<$) [primjer: skala tvrdoće]
  - kvantitativne
    - intervalna (operacije + i -) [primjer °C, °F]
    - racionalna (operacije +, -, *, /) [primjer °K]
- *Zadnje dvije su metrička skala*
- Podjela prema vrijednostima koje poprimaju:
  - kvalitativne vs. kvantitativne
  - diskretne vs. kontinuirane

- Mjere centralne tendencije:
  - aritmetička sredina (engl. mean)
  - medijan
  - mod



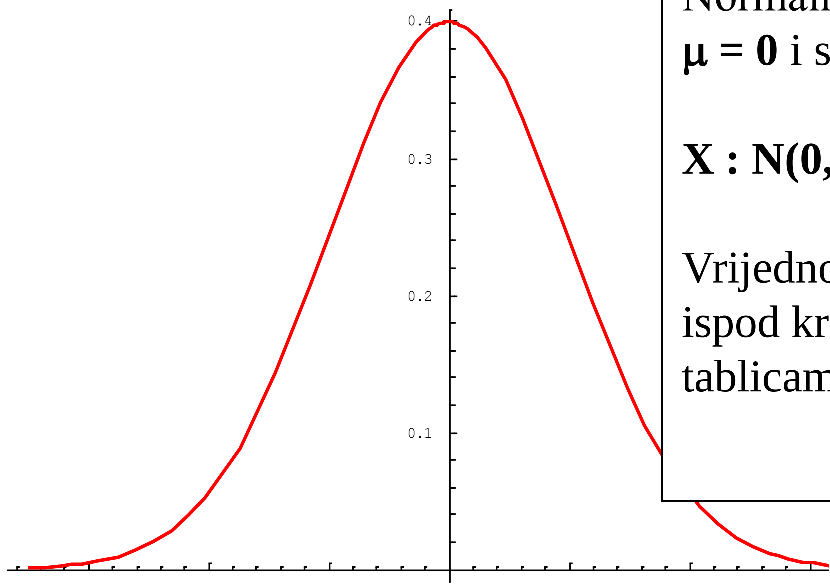
DESKRIPTIVNA STATISTIKA

- Mjere rasipanja
  - varijanca (srednje kvadratno odstupanje)
  - standardna devijacija
  - rang
  - interkvartilni rang
  - srednje apsolutno odstupanje
- Distribucije frekvencija
  - relativne
  - kumulativne
- Grafički prikazi empirijskih distribucija:
  - histogrami
  - poligoni
  - *box and whisker* plot

Teorijske Distribucije

- binomna
- normalna
- t–distribucija
- χ^2
- F distribucija

- **Normalna distribucija**



Normalna distribucija s očekivanjem **μ = 0** i standardnom devijacijom **σ = 1.**

X : N(0,1) - Označavamo je s **U** ili **Z**.

Vrijednosti vjerojatnosti tj. površina ispod krivulje **U** dani su u statističkim tablicama.

- Uobičajena oznaka za normalnu distribuciju s parametrima: očekivanjem μ i varijancom σ^2 je $N(\mu, \sigma^2)$.

DESKRIPTIVNA STATISTIKA

- Velika većina obilježja u prirodi distribuirana je prema normalnoj razdiobi
- Iz tablica za jediničnu normalnu distribuciju U očitavamo:

  $P(-1.65 < U < 1.65) = 90\%$

  $P(-1.96 < U < 1.96) = 95\%$

  $P(-2.58 < U < 2.58) = 99\%$

- Neka je X normalno distribuirana, tj. X: $N(\mu, \sigma^2)$. Vrijedi transformacija:

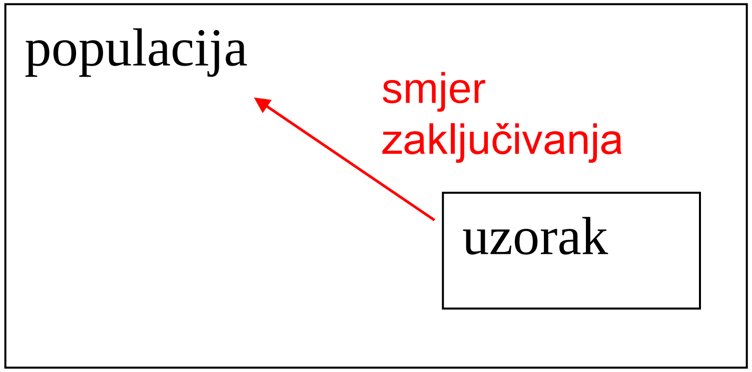
  STANDARDIZACIJA - svođenje X: $N(\mu, \sigma^2)$ na U sa

  transformacijom $$U = \frac{X - \mu}{\sigma}$$

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

- $P(\mu - 1.65\sigma < X < \mu + 1.65\sigma) = 90\%$
- $P(\mu - 1.96\sigma < X < \mu + 1.96\sigma) = 95\%$
- $P(\mu - 2.58\sigma < X < \mu + 2.58\sigma) = 99\%$

INFERENCIJALNA STATISTIKA



- **Populacija** – skup svih mogućih vrijednosti slučajne varijable
- **Uzorak** – podskup populacije
- **Parametar** je bilo koja funkcija populacije, neko svojstvo populacije koje nas zanima, npr. srednja vrijednost, standardna devijacija, proporcija, itd. Parametar se odnosi na populaciju i označava se malim grčkim slovima (μ, σ, π itd).

INFERENCIJALNA STATISTIKA

- Često nam vrijednosti parametra populacija nisu dostupne te ih procjenjujemo na temelju uzorka
- Procjena nekog parametra populacije na temelju uzorka naziva se **statistika ili procjenitelj** i označava se malim slovima (χ, s, p, …). Općenito je vrijednost statistike i nepoznatog parametra populacije dana izrazom:
- Statistika = Parametar_populacije ± *pogreška*

nepoznato

- Taj se izraz može zapisati u obliku:

$$\boxed{\textbf{Parametar\_populacije = Statistika} \pm \textbf{\textit{pogreška}}}$$

INFERENCIJALNA STATISTIKA

- Ono što želimo znati je s kojom točnošću (preciznošću) i s kojom pouzdanošću (vjerojatnosti), neka statistika procjenjuje parametar populacije

- Primjer: $\mu = \bar{x} \pm$ pogreška

- Parametar je svojstvo populacije
- Statistika je funkcija uzorka (podskupa te populacije) te za svaki novi uzorak izvučen iz iste populacije (istog osnovnog skupa) možemo dobiti različitu vrijednost statistike!!!

INFERENCIJALNA STATISTIKA

- Ali, ako znamo kako je statistika uzorka distribuirana – tj. ako znamo kako je distribuirana vrijednost statistike ($\bar{x}$) na temelju beskonačno mnogo uzoraka iste veličine izvučenih iz te populacije (to je distribucija vjerojatnosti statistike uzorka) – tada uz pomoć vjerojatnosti možemo procijeniti s kojom pouzdanošću se parametar populacije nalazi u određenim granicama. Dakle, možemo odrediti granice oko $\bar{x}$ u kojima se nalazi parametar $\boldsymbol{\mu}$ i pridruženu vjerojatnost za takvo odstupanje. ($\mu = \bar{x} \pm$ pogreška, uz određenu vjerojatnost, tj. pouzdanost)

Ta se distribucija vjerojatnosti statistike uzorka naziva se

DISTRIBUCIJA UZORKOVANJA

(*engl. sampling distribution*)

INFERENCIJALNA STATISTIKA

- Poznavanje distribucije uzorkovanja neke statistike temelj je za inferencijalno statističko zaključivanje (intervale procjene parametara populacije, testiranje hipoteza)
- Svaki parametar populacije (srednja vrijednost μ, proporcija π, varijanca σ^2, …) ima svoju distribuciju uzorkovanja
- Važno svojstvo distribucije vjerojatnosti statistike su njezino očekivanje i standardna devijacija. Ta se standardna devijacija distribucije neke statistike naziva STANDARDNA POGREŠKA (SE)
- Od posebnog je značenja distribucija uzorkovanja srednje vrijednosti

Pristranost procjenitelja

- **Pristranost procjenitelja** Y za neki proizvoljan parametar populacije **p** je

$$E[Y] - p.$$

- Ako je razlika E[Y] – p = 0 kažemo da je **procjenitelj nepristran**.
  - Napomena: E[Y] je očekivanje slučajne varijable Y

Primjer:

  - Srednja vrijednost uzorka je nepristran procjenitelj prave srednje vrijednosti populacije μ.
  - Proporcija uzorka **x/n** je nepristran procjenitelj prave proporcije π.

DISTRIBUCIJA UZORKOVANJA SREDNJE VRIJEDNOSTI

- CENTRALNI GRANIČNI TEOREM

Neka je dana populacija sa srednjom vrijednošću μ i standardnom devijacijom σ.

Neka je μ_srednja vrijednost od *n* slučajno odabranih nezavisnih opservacija iz te populacije.

Distribucija uzorkovanja srednje vrijednosti približava se normalnoj sa očekivanjem _μ i standardnom devijacijom $SE = \sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}}$ kada $n \to \infty$

pogledati tekst i simulacije na adresi http://www.ruf.rice.edu/~lane/rvls.html, posebno za ilustraciju CGT pogledati animirani primjer na adresi http://www.ruf.rice.edu/~lane/stat_sim/sampling_dist/index.html

DISTRIBUCIJA UZORKOVANJA SREDNJE VRIJEDNOSTI

- **Standardna pogreška** *(engl. standard error)* nekog parametra je sandardna devijacija distribucije uzorkovanja tog parametra. Ponekad se označava sa **SE**

Primjer

- Standardna pogreška distribucije uzorkovanja srednje vrijednosti je $SE = \sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}}$, gdje je X slučajna varijabla osnovnog skupa (na primjer: visine populacije studenata Zagrebačkog Sveučilišta, težine proizvoda koje proizvede neka tvornica, itd.). Često se umjesto σ_X piše samo σ

DISTRIBUCIJA UZORKOVANJA SREDNJE VRIJEDNOSTI

- Prije smo napomenuli da je

  Parametar = Statistika ± pogreška

- Ako za statistiku koja nas zanima odaberemo srednju vrijednost tada je

  $\mu = \bar{x} \pm$ pogreška

DISTRIBUCIJA UZORKOVANJA SREDNJE VRIJEDNOSTI

- Prije smo napomenuli da je

  Parametar = Statistika ± pogreška

- Ako za statistiku koja nas zanima odaberemo srednju vrijednost tada je

$$\mu = \bar{x} \pm \text{pogreška}$$

- No sada, na temelju centralnog graničnog teorema koji nam kaže da su srednje vrijednosti uzoraka veličine n također distribuirane normalno sa standardnom pogreškom SE = $\sigma / \sqrt{n}$, možemo pisati:

$$\mu = \bar{x} \pm u_p \, \sigma / \sqrt{n}, \text{ sa pouzdanošću } I(u_p),$$

gdje je u_0 vrijednost jedinične normalne razdiobe, a $I(u_p)$ pripadna pouzdanost (vjerojatnost). Te se vrijednosti očitavaju u statističkim tablicama.

DISTRIBUCIJA UZORKOVANJA SREDNJE VRIJEDNOSTI

u_p	1.64	1.96	2.58
pouzdanost $I(u_p)$	90%	95%	99%

- Intervalna procjena očekivanja (za velike uzorke, $n \geq 30$)

  $\mu = \bar{x} \pm u_p \, \sigma / \sqrt{n}$, sa pouzdanošću $I(u_p)$,

  dok je za male uzorke umjesto vrijednosti u_p jedinične normalne distribucije vrijednost studentove t-distribucije koja se očitava iz statističkih tablica za zadani broj stupnjeva slobode k, gdje je $k=n-1$, a n je broj elemenata u uzorku.

  $\mu = \bar{x} \pm t(k) \, \sigma / \sqrt{n}$, s pouzdanošću ovisnom o $t(k)$

DISTRIBUCIJA UZORKOVANJA SREDNJE VRIJEDNOSTI

- Centralni granični teorem (CGT) $U = \dfrac{\overline{X} - \mu}{\sigma / \sqrt{n}}$
- Iz tvrdnje CGT-a slijede formule za **intervalne procjene očekivanja μ (za velike n, n>30):**

$\overline{X} : N(\mu, \sigma^2/n)$

- $P(\bar{x} - 1.65 \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.65 \frac{\sigma}{\sqrt{n}}) = 90\%$
- $P(\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}) = 95\%$
- $P(\bar{x} - 2.58 \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 2.58 \frac{\sigma}{\sqrt{n}}) = 99\%$

DISTRIBUCIJA UZORKOVANJA SREDNJE VRIJEDNOSTI

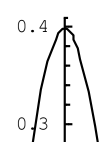
Primjer:

- Naka je X je normalno distribuirana sl. varijabla (kraće ćemo reći normalna distribucija).
- Neka su parametri od X - očekivanje 34 i standardna devijacija 4
- To zapisujemo X: $N(34, 4^2)$
- Kolika je vjerojatnost da slučajno izvučen primjer iz te distribucije poprimi vrijednost veću od 30?

$P(X > 30) = P(U > (30 – 34)/4)) = P(U > -1) =$ (očitavamo iz stat. tablica) $= 0.841$

DISTRIBUCIJA UZORKOVANJA SREDNJE VRIJEDNOSTI

- Ako sada izvlačimo uzorak od 16 elemenata iz zadane distribucije X: N(34, 4^2) i računamo srednju vrijednost, kolika je vjerojatnost da srednja vrijednost izračunata iz tog uzorka bude veća od 30?
- Prema CGT, $\overline{X}$ je distribuirano s očekivanjem 34 i standardnom devijacijom SE = $4/\sqrt{16}$ = 1, dakle
- P($\overline{X}$ > 30) =(standardizacija)=P(U > (30 – 34)/1)) = P(U > -4) = 1



BINOMNA DISTRIBUCIJA

- Slučajni pokus: dva moguća ishoda, A i nonA.
- Vjerojatnost događaja A, P(A) = π i vjerojatnost da se ne dogodi A, P(nonA) = 1 - π

Primjer

- Promatramo jedan proizvod: proizvod je ispravan s vjerojatnošću π. Mogući događaji:
- A = proizvod je ispravan, P(A) = π
- non A = proizvod je neispravan, P(nonA) = 1 - π

- Pretpostavimo da imamo nizove od n takvih nezavisnih pokusa (Bernoullijevi nizovi)

BINOMNA DISTRIBUCIJA

- Kolika je vjerojatnost da će se događaj A pojaviti točno x puta u tom nizu?

Primjer

- Uzorak od n proizvoda, kolika je vjerojatnost da točno x od n proizvoda ($0 \le x \le n$) bude ispravno?

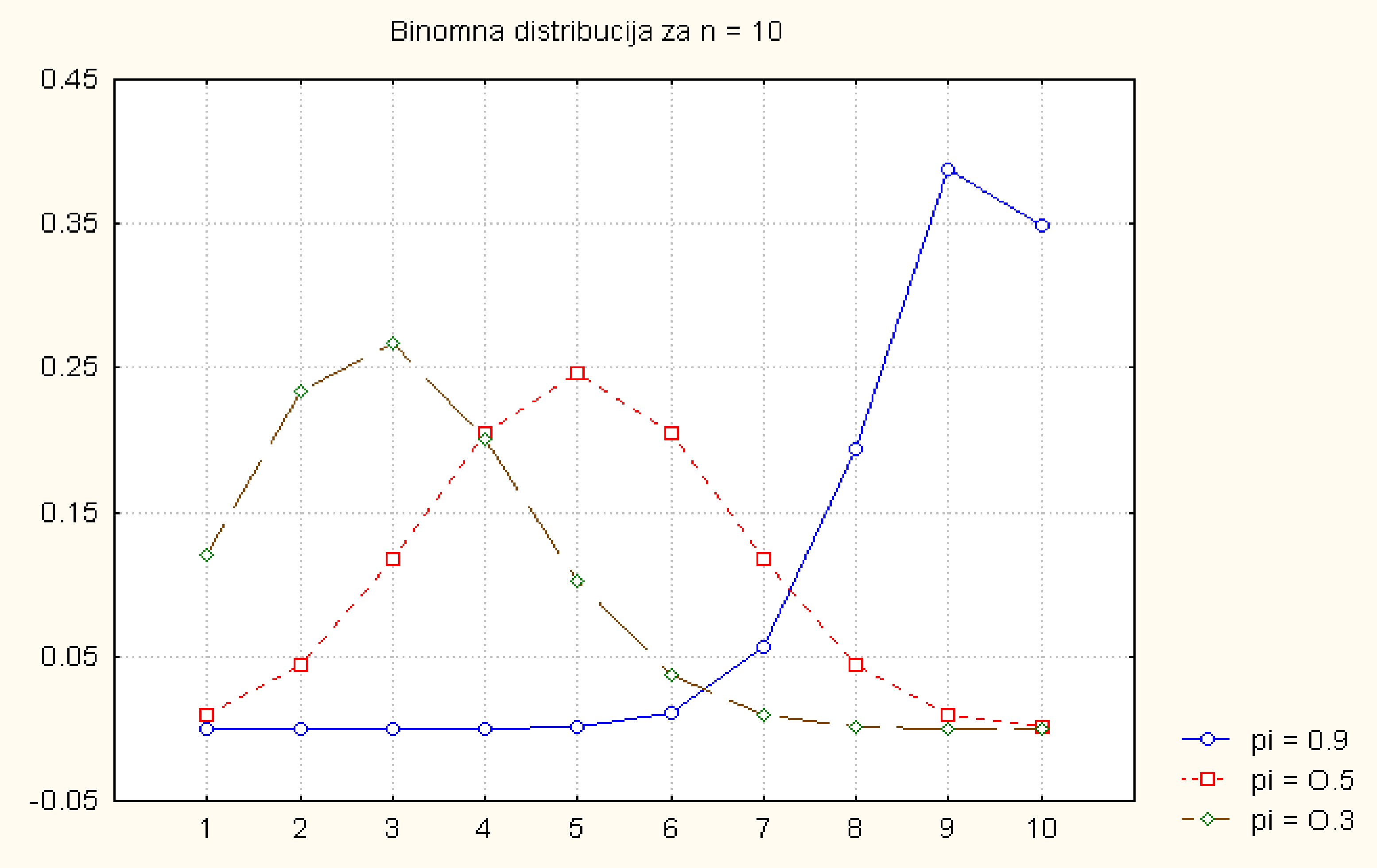
BINOMNA DISTRIBUCIJA

- **Binomna slučajna varijabla s parametrima n i π**

- Kolika je vjerojatnost da će se događaj A pojaviti točno x puta u tom nizu? tj. Kolika je vjerojatnost da slučajna varijabla X poprimi vrijednost x?

$$P(X = x) = \binom{n}{x} \pi^{x} (1-\pi)^{n-x} \qquad (1)$$

BINOMNA DISTRIBUCIJA



BINOMNA DISTRIBUCIJA

- Kolika je vjerojatnost da će se događaj A pojaviti između x_1 i x_2 puta u tom nizu od n pokusa?
- $P(x_1 \leq X \leq x_2) = \sum_{i=x_1}^{x_2} \binom{n}{i} \pi^i (1-\pi)^{n-i}$. (2)
- Očekivanje binomne slučajne varijable X je $E(X) = n\pi$
- Varijanca binomne slučajne varijable $V(X) = \sigma^2 = n\pi(1-\pi)$

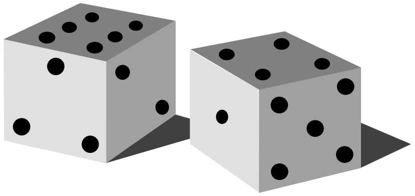
Primjer:

- $n = 15$, $\pi = 0.2$ $x = 4$
- $P(X = 4) = 0.188$

BINOMNA DISTRIBUCIJA

Primjer:

- Da li je povoljno kladiti se da će u 24 uzastopna bacanja dvije igraće kocke barem jednom pasti dvostruka šestica?



Primjer:

- N = 300, $\pi = 0.2$, P(100 > X > 50) = ?

- **Aproksimacija binomne normalnom** (Moivre-Laplaceova formula)

$$P(x_1 \le X \le x_2) \approx P\left(\frac{x_1 - n\pi}{\sqrt{n(1-\pi)\pi}} \le U \le \frac{x_2 - n\pi}{\sqrt{n(1-\pi)\pi}}\right)$$

Uz uvjet $n\pi > 5$ i $n\pi(1-\pi) > 5$.

Proporcija

- X binomna slučajna varijabla s parametrima n i π, tj.
- X: B(π, n)
- Proporcija je omjer P = X/n

DISTRIBUCIJA UZORKOVANJA PROPORCIJE

- Promatramo nizove od **n** elemenata
- Zanima nas broj elemenata u tom nizu od **n** koji imaju neko svojstvo A. Označimo taj broj s **x**

  (**Bernoullijevi nizovi)**
- **Proporcija P** je omjer **P** = **X/n**

DISTRIBUCIJA UZORKOVANJA PROPORCIJE

Primjer

- Neka je dana neka hipoteza h
- Pretpostavimo da imamo uzorak sastavljen 14 elemenata tj. primjera za učenje. Ako 8 od 14 primjera zadovoljava hipotezu h tada je proporcija uspjeha hipoteze h na tom skupu (uzorku) jednaka $p_1 = x/n = 8/14$
- Uzmimo neki drugi uzorak tj. skup primjera za učenje i neka je na tom skupu proporcija valjanosti hipoteza $p_2 = 5/14$
- Neka je dan neki treći skup primjera za učenje iste veličine i neka je na njemu $p_3 = 7/14$
- Ako nastavimo s tim postupkom u ∞ dobivamo **distribuciju uzorkovanja proporcije koju označavamo s P**

DISTRIBUCIJA UZORKOVANJA PROPORCIJE

(Posljedica Moivre -Laplaceovog teorema - CGT)

- **Distribucija uzorkovanja proporcije** za velike n približava se normalnoj distribuciji s

očekivanjem π i

standardnom devijacijom $\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}}$

INTERVALNE PROCJENE PROPORCIJE

- Parametar populacije = statistika_uzorka $\pm$ pogreška
- Primjeri: π = p $\pm$ pogreška, $\mu = \ \pm$ pogreška.
- Na temelju poznate distribucije uzorkovanja proporcije izvode se intervalne procjene proporcije

$$P\left(p - 2.58\sqrt{\frac{\pi(1-\pi)}{n}} < \pi < p + 2.58\sqrt{\frac{\pi(1-\pi)}{n}} \right) = 99\%$$

$$P\left(p - 1.96\sqrt{\frac{\pi(1-\pi)}{n}} < \pi < p + 1.96\sqrt{\frac{\pi(1-\pi)}{n}} \right) = 95\%$$

INTERVALNE PROCJENE PROPORCIJE

Primjer

- Jedan strijelac je pogodio 5 puta u metu od 10 pokušaja
- Drugi strijelac je pogodio 50 puta u metu od 100 pokušaja
- Što možemo reći o pravoj proporciji pogodaka jednog i drugog strijelca?

$$P\left(0.5 - 1.96\sqrt{\frac{0.5(1-0.5)}{10}} < \pi < 0.5 + 1.96\sqrt{\frac{0.5(1-0.5)}{10}}\right) = 95\%$$

$$P\left(0.5 - 1.96\sqrt{\frac{0.5(1-0.5)}{100}} < \pi < 0.5 + 1.96\sqrt{\frac{0.5(1-0.5)}{100}}\right) = 95\%$$

INTERVALNE PROCJENE PROPORCIJE

Prvi strijelac:

- $P(0.5 - 1.96*0.158 < \pi < 0.5 + 1.96*0.158) = 95\%$
- $P(0.5 - 0.31 < \pi < 0.5 + 0.31) = 95\%$
- **$P(0.19 < \pi < 0.81) = 95\%$**

Drugi strijelac:

- $P(0.5 - 1.96*0.05 < \pi < 0.5 + 1.96*0.05) = 95\%$
- $P(0.5 - 0.098 < \pi < 0.5 + 0.098) = 95\%$
- **$P(0.402 < \pi < 0.598) = 95\%$**

TESTIRANJE HIPOTEZA

1. direktno statističko zaključivanje (inferencijalno): *točkovne ili intervalne procjene*
   - uzorak koristimo za procjenu parametra populacije

2. indirektno: *testiranje hipoteza*

- Uzorak podržava ili diskreditira a priori postavljenu tvrdnju ili pretpostavku o stvarnoj vrijednosti parametra populacije
- Hipoteza o populacionom parametru proizlazi iz
  - prethodnih ispitivanja
  - teoretskih pretpostavki

TESTIRANJE HIPOTEZA

- Ako postupkom testiranja nađemo da je H_0 neprihvatljiva s aspekta vjerojatnosti, tada prihvaćamo (vjerujemo) u alternativnu hipotezu

- Isto kao što ne možemo naći 100% interval pouzdanosti tako ni testiranje ne daje 100% sigurnost u ispravnost odluke već su pouzdanosti s kojim radimo 90, 95, 99%. Naime, u postupku testiranja unaprijed zadajemo (i time kontroliramo) pogrešku (tj. rizik s kojim radimo statistički test) a to je vjerojatnost odbacivanja istinite hipoteze. Ta se vjerojatnost naziva **nivo signifikantnosti** (nivo značajnosti) ili pogreška prvog reda i označava se α

TESTIRANJE HIPOTEZA

Postupak:

- Postavljaju se dvije međusobno isključive hipoteze koje zajednički iscrpljuju sve mogućnosti

dvostrani test $\Rightarrow$

$H_0 \quad \theta = a$
$H_1 \quad \theta \neq a$

ili

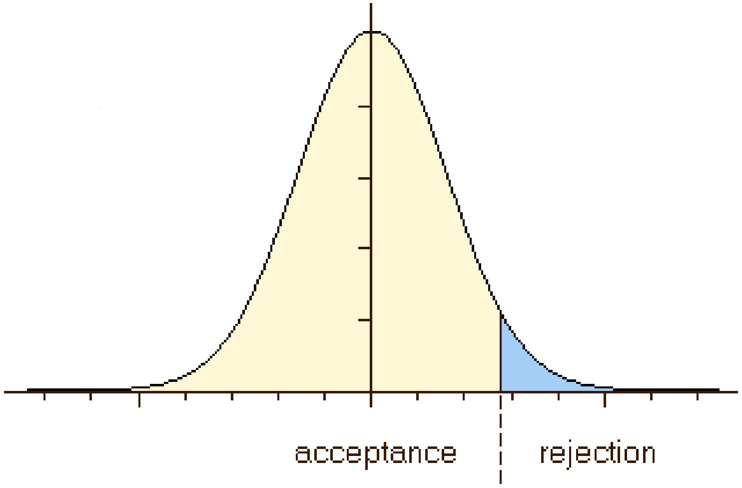
jednostrani testovi $\Rightarrow$

$H_0 \quad \theta = a$
$H_1 \quad \theta < a$

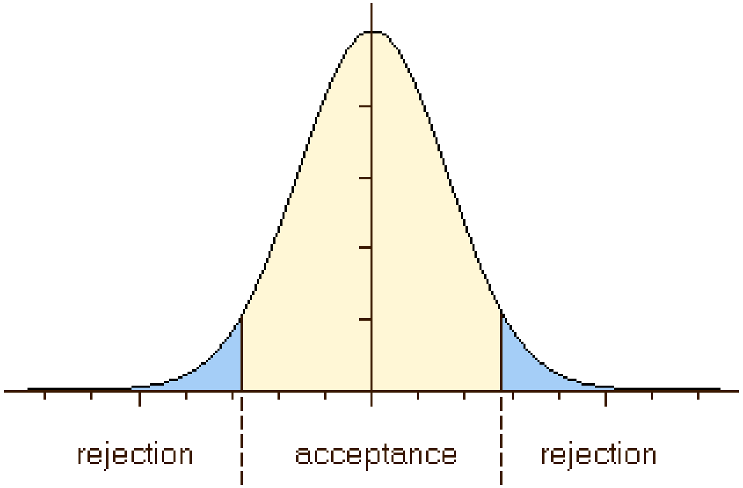
ili

$H_0 \quad \theta = a$
$H_1 \quad \theta > a$

- U zadnja dva slučaja moramo biti sigurni da **θ > a, θ < a**, nije moguće!



jednostrani test



dvostrani test

TESTIRANJE HIPOTEZA

- Postavljanje hipoteza dešava se na logičkoj razini, tj. vezano je za problem poznavanja područja problema
- Prihvaćanje hipoteze tj. vjerovanje u određenu hipotezu je stvar statističke odluke

Primjer

$$H_0 \quad \mu = 20$$
$$H_1 \quad \mu \neq 20$$

Uzorak od 100 elemenata dao je

a) $\bar{x} = 19.1$

b) $\bar{x} = 19.9$

c) $\bar{x} = 16$

- Pretpostavimo da znamo da je st.dev. populacije $\boldsymbol{\sigma = 3}$

TESTIRANJE HIPOTEZA

- Pitanje je da li je moguće, tj. koliko je vjerojatno da dobijemo srednju vrijednost uzorka $\bar{x}$ = 19.1 ako je μ = 20. Ako je ta vjerojatnost mala onda smo skloni ne vjerovati u pretpostavku iz nulte hipoteze.
- Pitanje je *koliko je to "malo vjerojatno"* ?
- Obično je to 1% ili 5% i naziva se **nivo značajnosti (signifikantnosti)** i označava se s α**.**
- α je vjerojatnost odbacivanja istinite hipoteze!
- Rizik testiranja koji se određuje unaprijed!

TESTIRANJE HIPOTEZA

a) $\bar{x}$ = 19.1, odaberemo α = 0.05 tj. 5%. Radimo dvostrani U - test

$$U = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{19.1 - 20}{3 / \sqrt{100}} = -3$$

- Vjerojatnost da je P (U < -3) je praktički jednaka 0 (pa onda i 2*P (U < -3) $\approx$ 0, jer radimo dvostrani test pa gledamo površine u oba repa), tj. ta je vjerojatnost puno manja od 0.05 (koliki je nivo signifikantnosti testa) pa odbacujemo nultu hipotezu

- **Interpretacija**: Vjerojatnost da na temelju uzorka od 100 elemenata dobijemo srednju vrijednost 19.1, ako je prava vrijednost 20, je praktički nula pa smo stoga skloni NE vjerovati u nultu hipotezu tj. odbacujemo je.

b) $\bar{x}$ = 19.9, odaberemo α = 0.05 tj. 5%. Radimo dvostrani U - test

- U = $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \dfrac{19.1-20}{3/\sqrt{100}}$ = - 0.33

- Iz statističkih tablica slijedi da je vjerojatnost 2*P (U < - 0.33) = 2*0.37 = 0.74 što je puno veće od α = 0.05 (koliki je nivo signifikantnosti testa) => prihvaćamo nultu hipotezu.

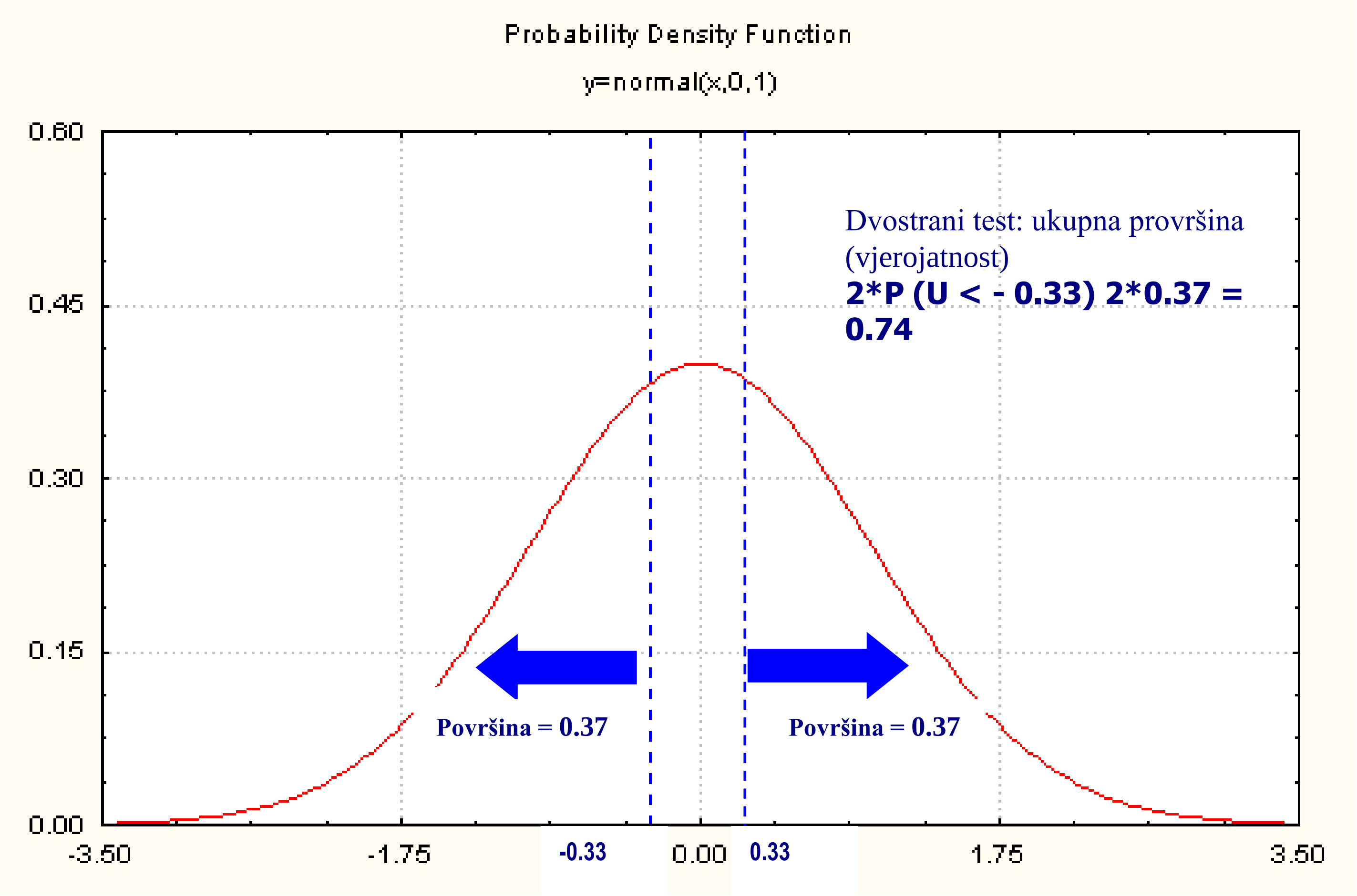
- **Interpretacija**: Nemamo razloga, na temelju predočenog uzorka (uzorak od 100 elemenata čija je srednja vrijednost $\bar{x}$ =19.9), sumnjati u istinitost nulte hipoteze!

TESTIRANJE HIPOTEZA

- Vjerojatnost da dobijemo srednju vrijednost uzorka (po apsolutnoj vrijednosti jednaku ili veću od) $\overline{x}$ = 19.1 je 0.74, ako je stvarna srednja vrijednost populacije 20. To je puno veća vjerojatnost od 0.05 što je granična vjerojatnost s kojom radimo testiranje.

- Mogli bi reći da uzorak podržava tvrdnju iz nulte hipoteze s vjerojatnošću 0.74

TESTIRANJE HIPOTEZA



POGREŠKE PRVOG I DRUGOG REDA

- Prilikom testiranja možemo učiniti dva tipa pogrešaka: greške I i II reda
- Usporedba postupka statističkog testiranja i pravosudnog postupka:
- H_0 Osumnjičeni je nevin
- H_1 Osumnjičeni je kriv

Odluka suda	**Stvarno stanje**		**Zaključak**	**Stvarno stanje**	
	Nevin	Kriv		H_0 je istina	H_0 je laž
Nevin	√	**pogreška**	H_0 prihvaćamo	√	β (greška II reda)
Kriv	**pogreška**	√	H_0 odbacujemo	α (greška I reda)	√

POGREŠKE PRVOG I DRUGOG REDA

- **Pogreška I reda ili** α je pogreška koju uvijek možemo kontrolirati prilikom statističkog zaključivanja
- Zadaje se unaprijed, a hipoteze se formuliraju tako da ona pogreška koja nam je važnija bude pogreška prvog reda α
- Na primjer, u pravosudnom postupku možemo učiniti dvije pogreške, da nevinog čovjeka osudimo ili da krivog oslobodimo. Možemo se odlučiti da je važnije kontrolirati vjerojatnost pogreške da nevinog čovjeka osudimo

POGREŠKE PRVOG I DRUGOG REDA

Formuliramo hipoteze:

- H_0 Osumnjičeni je nevin i
- H_1 Osumnjičeni je kriv

- Pogreška prvog reda ili α je vjerojatnost odbacivanja hipoteze H_0 kada je ona zapravo istinita, tj. u ovom slučaju vjerojatnost da nevinog čovjeka proglasimo krivim

- Kada bi obrnuli hipoteze i stavili H_0 *Osumnjičeni je kriv*, tada bi zadavali unaprijed i time kontrolirali pogrešku da krivog čovjeka oslobodimo

POGREŠKE PRVOG I DRUGOG REDA

Pogreška II reda ili β

- Vjerojatnost prihvaćanja hipoteze H_0 kada je H_1 istina (dakle H_0 je laž)!
- U našem primjeru postavljenih hipoteza:
  - H_0 Osumnjičeni je nevin
  - H_1 Osumnjičeni je kriv
- To je slučaj kada je osumnjičeni zaista kriv no mi ga proglasimo nevinim

POGREŠKE PRVOG I DRUGOG REDA

β ovisi o:

- pravoj vrijednosti parametra o kojem raspravljamo (alternativna hipoteza), *β pada kada je veća razlika između pretpostavljene i prave vrijednosti parametra koji se testira (distribucije su razdijeljene)*
- pogrešci α, tj. *β raste kada α pada i obrnuto, te* jednostranom ili dvostranom testu, β
- standardnoj devijaciji populacije, *β se povećava što je st. dev. populacije veća*
- veličini uzorka, *β se smanjuje kada veličina uzorka raste.*

Zadnja dva parametra određuju standardnu pogrešku SE

TESTIRANJE PROPORCIJA

1. Formuliranje statističke hipoteze

$$H_0 \quad \pi = 0.005$$

$$H_1 \quad \pi < 0.005$$

(jednostrani, lijevi test – područje odbacivanja hipoteze je na lijevo)

2. Odredi statistiku za testiranje: proporcija P

Znamo da vrijedi $U = \dfrac{P - \pi}{\sqrt{\dfrac{p(1-p)}{n}}}$

TESTIRANJE PROPORCIJA

3. Odaberi nivo značajnosti testa tj. pogrešku prvog reda α, neka je α = 5% i pripadnu kritičnu vrijednost očitaj iz tablica.

   Za odabrani nivo značajnosti i jednostrani test u_{krit}=-1.64

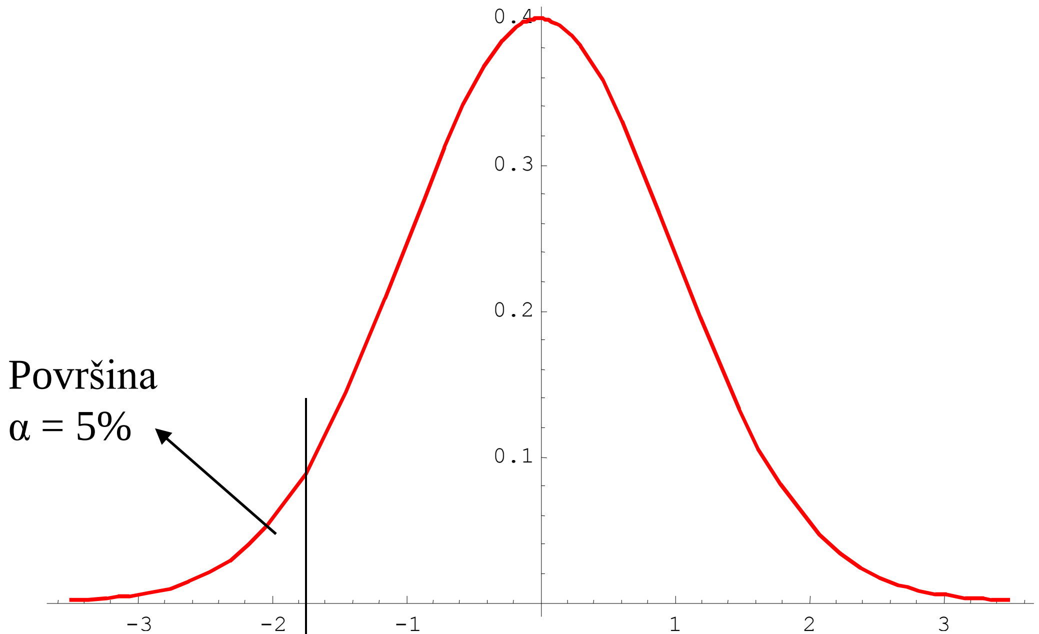
4. Uzmi slučajan uzorak n=2000 i izračunaj vrijednost statistike P na njemu, tj. p=3/2000=0.0015

$$u = \frac{0.0015 - 0.005}{\sqrt{\frac{0.005(1-0.005)}{2000}}} = -2.26$$

5. Donesi odluku:

   Ako je izračunata vrijednost statistike $u < u_{krit}$ odbaci nultu hipotezu. Kako je $-2.26 < -1.64$ H_0 odbacujemo!

TESTIRANJE PROPORCIJA



Područje odbacivanja H_0	Područje prihvaćanja H_0
	-1.64 0.0
-2.26	

χ^2 test

- Neparametarski test

- Koristi se za dvije kategorije testova:
- Testiranje ponašanja po distribuciji (engl. *goodness of fit*)
- Testiranje nezavisnosti klasifikacija: kontigencijske tablice (engl. *contigency tables*)

- H_0 dvije kvalitativne populacijske varijable su nezavisne

- RxS tablice (R-retci, S-stupci)

$$\chi^2 = \Sigma\, (f_{obs} - f_{izracunata})^2 / f_{izracunata}$$

χ^2 test

	PUŠAĆI	NEPUŠAĆI	total
MUŠKARCI	110	90	200
ŽENE	104	96	200
total	214	186	400

χ^2 test

	PUŠAĆI	NEPUŠAĆI	total
MUŠKARCI	107= (214*200/400) (110)	93 (90)	200
ŽENE	107 (104)	93 (96)	200
total	214	186	400

χ^2 = $(110-107)^2/107 + (104-107)^2/107 + (90-93)/93 + (96-93)/93 =$
$0.084 + 0.084 + 0.097 + 0.097 =$ **0.362**

koristiti statističke tablice ili program

Broj stupnjeva slobode = (R-1)(S-1)=1(u ovom slučaju)

Da li postoji statistički značajna razlika između spolova u odnosu na pušenje?