Ogledni primjer pitanja u 1. kolokviju iz vjerojatnosti i statistike 2008/2009

1. Za skup podataka $x_1,x_2,...,x_n$:
   (i) napišite formulu za aritmetičku sredinu i navedite značenje.
   (ii) napišite formule za varijancu i standardnu devijaciju i navedite značenje.
   (iii) napišite formule za korigiranu varijancu i standardnu devijaciju i navedite uporabu.
   (iv) izračunajte sve za skup podataka 1,1,2,3,4,4,4,5.

2. (i) Definirajte i opišite eksponencijalnu razdiobu, nacrtajte grafove funkcija gustoće i distribucije.
(ii) Definirajte i opišite normalnu razdiobu i jediničnu normalnu razdiobu, nacrtajte grafove funkcija gustoće i distribucije.
(iii) Formulirajte, objasnite i grafički predočite pravilo *tri sigme*. Posebno za $X\sim N(5,2^2)$.
(iv Definirajte i opišite binomnu razdiobu.
(v) Definirajte i opišite Poissonovu razdiobu.

3. (i) Kako procjenjujemo očekivanje, a kako varijancu populacije?
   (ii) Napišite formule i predočite interval pozdanosti za očekivanje uz vjerojatnost 1-2p, posebno ako je vjerojatnost 0.95.
   (iii) Objasnite značenje intervala pouzdanosti.
   (iii) Objasnite značenje nivoa signifikantnosti.

4. (i) Opišite i predočite crtežima provjeru hipoteze $\mu = \mu_0$ (uz razne alternativne hipoteze).
(ii) Opišite i predočite crtežima provjeru hipoteze $\mu_1 = \mu_2$ (uz razne alternativne hipoteze).
.(iii) Zapišite formulu za $\chi^2_{\exp}$, za broj stupnjeva slobode, objasnite značenje kritične vrijednosti i predočite kritično područje pri testiranju *hikvadrat* testom.

5. (i) Navedite načelo na kojemu se zasniva metoda najmanjih kvadrata.
   (ii) Zapišite funkciju cilja za linearnu vezu i podatke:

x_i	1	2	4	5	6
y_i	-2	1	8	10	13

(iii) Predočite grafički ove podatke i procijenite parametre.
(iv) Objasnite značenje koeficijenta korelacije.
(v) Izračunajte parametre i koeficijent korelacije za navedene podatke.

Rješenja

1. (i) $\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$. Značenje je prosječna vrijednost podataka, a fizikalno značenje je težište sustava masa na pravca.

(ii) **Varijanca je prosječno kvadratno odstupanje od prosjeka**:

$$(s')^2 := \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + ... + (x_n - \bar{x})^2}{n}$$

Standardna devijacija uzorka s' je drugi korijen iz varijance uzorka:

$$s' := \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + ... + (x_n - \bar{x})^2}{n}}$$

I varijanca i standardna devijacija su mjere rasipanja podataka oko aritmetičke sredine (što je standardna devijacija veća, podatci su raspršeniji). Fizikalno značenje varijance je moment inercije sustava masa na pravcu.

(iii) **Korigirana varijanca**

$$s^2 := \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + ... + (x_n - \bar{x})^2}{n-1}$$

Korigirana standardna devijacija

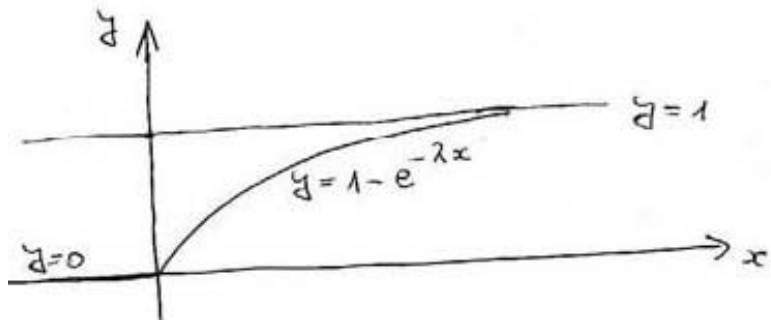
$$s := \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + ... + (x_n - \bar{x})^2}{n-1}}.$$

Korigirane varijanca i standardna devijacija imaju slične uloge kao i obična varijanca i standardna devijacija. Dodatno pomoću njih procjenjujemo varijancu i standardnu devijaciju populacije.

(iv) $\bar{x} = 3$, $(s')^2 = 2$, $s^2 = \frac{16}{7}$.

2. (i) Eksponencijalna razdioba X~E(λ) je kontinuirana razdioba s parametrom λ (vrijedi E(X) = 1/λ); s funkcijom gustoće f(x):= $\lambda\, e^{-\lambda x}$, za x>0:

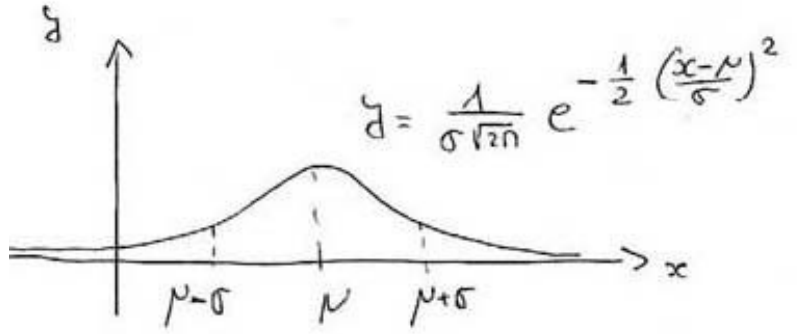
i funkcijom distribucije F(x)= 1- $e^{-\lambda x}$, za x>0:



Pojavljuje se (približno) pri mjerenju vremena između dviju poruka na nekoj adresi, vremena između dvaju uzastopnih kvarova na nekom uređaju i sl

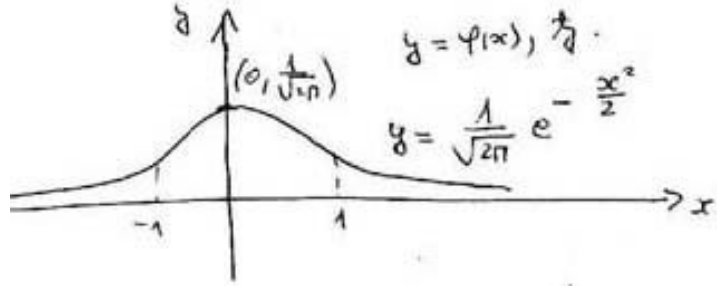
(ii) Normalna razdioba X~N (μ, σ^2) je kontinuirana razdioba s parametrima μ (ima značenje očekivanja) i σ^2 (ima značenje varijance). Pojavljuje se (približno) pri mjerenju mase, visine, inteligencije, prema njoj se ponašaju grješke pri mjerenju i sl.

Funkcija gustoće vjerojatnosti je $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$:



a funkcija distribucije ne može se eksplicitno izraziti.

Jedinična normalna razdioba je X~N($0,1^2$):



(iii) Vjerojatnost da slučajna varijabla X~N(μ, σ^2) poprimi vrijednost u intervalu $< \mu - 3\sigma, \mu + 3\sigma >$ približno je jednaka 1 (oko 0.997). Kraće,

p($\mu - 3\sigma < X < \mu + 3\sigma$)>0.997.

Slično treba napraviti i za *dva sigma* i *jedan sigma*.

Ako je X~N(5,2^2), onda je $< \mu - 3\sigma, \mu + 3\sigma >$ = <-1,11> itd.

(iv) Binomna razdioba X~B(n,p) s parametrima *n* (u pravilu ima značenje broja izvođenja pokusa) i *p* (u pravilu ima značenje vjerojatnosti uočenog događaja) je diskretna razdioba koja prima vrijednosti i=0,1,2,...,n uz vjerojatnosti p(X=i) = $\binom{n}{i} p^i(1-p)^{n-i}$

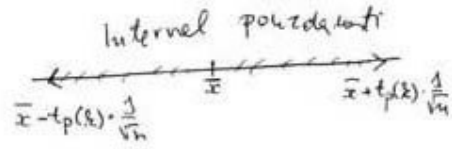
U ovoj interpretaciji slučajna varijabla X registrira broj pojavljivanja uočenog događaja. Vrijedi E(X) = np, V(X)=np(1-p).

(v) Poissonova razdioba X~P(a) s parametrom *a*>0 (koji ima značenje očekivanja od X) je diskretna razdioba koja prima vrijednosti i=0,1,2,... uz vjerojatnosti

p(X=i) = $e^{-a}\frac{a^i}{i!}$. Pojavljuje se pri brojenju poruka na nekoj adresi u fiksiranom vremenskom intervalu i sl.

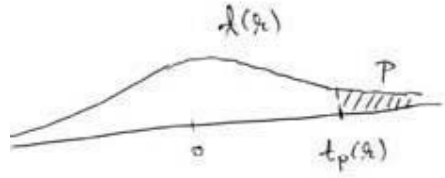
3. (i) Očekivanje populacije procjenjujemo aritmetičkom sredinom uzorka $\bar{x}$ (napisati formulu!!), a varijancu korigiranom varijancom uzorka s^2 (napisati formulu!!).

(ii) To je interval $< \bar{x} - z_p \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_p \cdot \frac{\sigma}{\sqrt{n}} >$ (ako je poznata standardna devijacija σ), odnosno $< \bar{x} - t_p(k)\frac{s}{\sqrt{n}}, \bar{x} + t_p(k)\frac{s}{\sqrt{n}} >$ ako σ nije poznato, već ga procjenjujemo pomoću *s* iz *n* mjerenja:



Broj z_p određen je time što je površina ispod grafa funkcije jedinične normalne razdiobe od tog broje nadalje jednaka p.

Broj $t_p(k)$, ima svojstvo da je površina ispod grafa funkcije Studentove razdiobe t(k), uz k=n-1 stupnjeva slobode, od tog broja nadalje jednaka p:



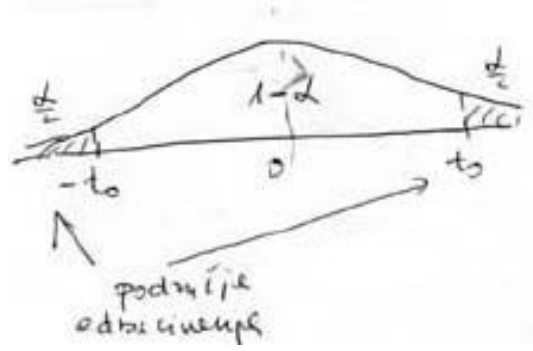
(iii) Interval pouzdanosti za očekivanje uz vjerojatnost 1-2p je simetrični interval oko $\bar{x}$ u kojemu se s vjerojatnošću 1-2p nalazi nepoznato očekivanje μ. Na primjer, uz 1-2p=0.95 u pripadnom bi se intervalu, pri velikom broju nezavisnih ponavljanja n mjerenja, očekivanje μ našlo približno u 95% slučajeva (a u odprilike 5% slučajeva palo bi izvan tako odredjenog intervala).

(iv) Nivo signifikantnosti (razina značajnosti) je broj α koji je vjerojatnost da se nula hipoteza H_0 odbaci, uz predpostavku da je ona istinita. To je površina područja odbacivanja početne hipoteze. Obično se uzima α =0.05.

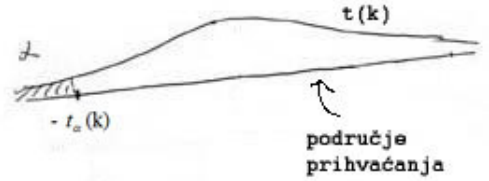
4. (i) Predpostavimo da je X normalno distribuirana slučajna veličina s očekivanjem μ i varijancom σ^2.
Neka smo na osnovi *n* mjerenja dobili procjene:
$\bar{x}$ za njeno očekivanje μ,
s^2 za njenu varijancu σ^2.
Testiramo hipotezu: H_0: $\mu = \mu_0$, gdje je μ_0 neka deklarirana vrijednost.

Prije toga trebali provjeriti hipotezu o bliskosti varijanca. a nakon što testiranje varijanaca pozitivno prođe, možemo pristupiti testiranju očekivanja (iako se i tada može nastaviti, ali s drukčijim postupkom). Ima tri slučaja, ovisno o izboru alternativne hipoteze.
I slučaj (dvostrani t-test)
H_0: $\mu = \mu_0$
H_a: $\mu \neq \mu_0$

1. Računamo $t_{\exp} = \dfrac{\bar{x} - \mu_0}{\dfrac{s}{\sqrt{n}}}$.
2. U tablici t-razdiobe određujemo kritičnu vrijednost $t_0 = t_p(k)$, ovisno o broju stupnjeva slobode k=n-1 i nivou signifikantnosti α =2p, što je obično 0.05.
3. Ako je $-t_p(k) < t_{exp} < t_p(k)$, hipotezu H_0 prihvaćamo, inače je odbacujemo:



II slučaj (jednostrani t-test; provodimo ga u pravilu onda ako su svi rezultati mjerenja ili gotovo svi, manji od deklarirane):
H_0: $\mu = \mu_0$
H_a: $\mu < \mu_0$
Na razini značajnosti α, slutnju odbacujemo ako je $t_{exp} < -t_\alpha(k)$:

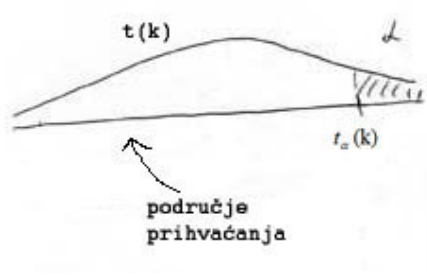


III slučaj (jednostrani t-test; provodimo ga u pravilu onda ako su svi rezultati mjerenja ili gotovo svi, veći od deklarirane):

H_0: $\mu = \mu_0$

H_a: $\mu > \mu_0$

Na razini značajnosti α, slutnju odbacujemo ako je $t_{exp} > t_\alpha(k)$:



(ii) Tom testu u pravilu predhodi F-test. Nakon što taj prođe nastavlja se s t-testom (testiranju očekivanja), tj. s testiranjem hipoteze:

H_0: $\mu_1 = \mu_2$ (nulta hipoteza)

Gdje je μ_1 procijenjenjeno s $\bar{x}_1$, a σ_1 s korigiranom standardnom devijacijom s_1 iz n_1 podataka, dok je μ_2 procijenjenjeno s $\bar{x}_2$, a σ_2 s korigiranom standardnom devijacijom s_2 iz n_2 podataka

Hipoteza se, primjenom t-testa, provodi ovako:

1. Izračuna se:

$$t_{\exp} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1-1)\ s_1^2 + (n_2-1)\ s_2^2}{n_1+n_2-2}}\sqrt{\frac{n_1+n_2}{n_1 n_2}}}$$

gdje obično označavamo: $s_d = \sqrt{\frac{(n_1-1)\ s_1^2 + (n_2-1)\ s_2^2}{n_1+n_2-2}}\sqrt{\frac{n_1+n_2}{n_1 n_2}}$

2. Odredi se broj stupnjeva slobode $k=n_1+n_2-2$.
3. Prihvati se neki nivo signifikantnosti α (obično α=0.05, ali može i α=0.01 ili α=0.1)
4. Odredi se se kritična vrijednost pomoću koje odredjujemo upada li izračunata vrijednost t_{exp} u kritično područje. Kritična vrijednost ovisi o nivou signifikantnosti α, o broju stupnjeva slobode (dakle o broju mjerenja), ali i o našoj kontrahipotezi koja može biti:

a) $\mu_1 \neq \mu_2$ (kad testiramo jesu li te dvije veličine jednake ili različite). Tada kritična vrijednost $t_0 = t_{\frac{\alpha}{2}}(k)$ ima značenje: $P(|t(k)|>t_0) = \alpha$, gdje $t(k)$ označava Studentovu (t-razdiobu) uz k stupnjeva slobode. To je isto kao i da smo rekli da je $P(t(k)>t_0) = \frac{\alpha}{2}$.

Hipotezu prihvaćamo ako je $|t_{exp}|<t_0$ tj. ako je $-t_0<t_{exp}< t_0$ (inače je odbacujemo).

b) (koja ima smisla samo ako je $\bar{x}_1 > \bar{x}_2$).

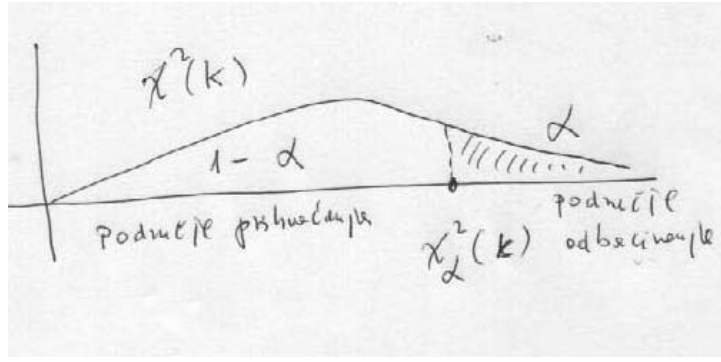
Tada kritična vrijednost t_0 ima značenje: $P(t>t_0) = \alpha$, hipotezu prihvaćamo ako je $t_{exp}<t_0$, inače je odbacujemo.

c) $\mu_1 < \mu_2$ (koja ima smisla samo ako je $\bar{x}_1 < \bar{x}_2$) .
Tada kritična vrijednost t_0 takodjer ima značenje: $P(t>t_0) = \alpha$.
Hipotezu prihvaćamo ako je $t_{exp} > - t_0$, inače je odbacujemo.

(iii) Podatke grupiramo u L razreda. Pripadne frekvecije označimo s f_i. Uz predpostavku (nula hipotezu) da se podatci ravnaju prema nekoj teoretskoj razdiobi, odredimo pripadne teoretske frekvencije f_{ti}. Kao mjera bliskosti eksperimentalnih i teoretskih podataka služi

$$\chi^2_{\exp} := \frac{(f_1 - f_{t1})^2}{f_{t1}} + \frac{(f_2 - f_{t2})^2}{f_{t2}} + ... + \frac{(f_L - f_{ttL})^2}{f_{tL}}$$

Hipotezu prihvaćamo na razini značajnosti α (obično je α=0.05) ako je $\chi^2_{\exp} < \chi^2_\alpha(k)$, gdje je k= $L - 1 - l$ (l je broj parametara u teoretskoj razdiobi), a $\chi^2(k)$ je hi kvadrat razdioba s k stupnjeva slobode, a $\chi^2_\alpha(k)$ je broj iza kojega je ispod grafa funkcije gustoće od $\chi^2(k)$ razdiobe površina jednaka α :



5. (i) Metoda najmanjih kvadrata zasniva se na načelu da
suma kvadrata razlika eksperimentalnih i teoretskih podataka bude minimalna.
Kraće, $\sum D_i^2 \rightarrow$ min, gdje je $D_i := y_i - f(x_i, a,b)$ odstupanje između eksperimentalnih i teoretskih podataka.
(ii) Tu je funkcija cilja $F(a,b) := \sum D_i^2$, dakle
$F(a,b)= [-2-(a\cdot 1+b)]^2+[1-(a\cdot 2+b)]^2+[8-(a\cdot 4+b)]^2+[10-(a\cdot 5+b)]^2+[13-(a\cdot 6+b)]^2$
(iii) pogledajte odgovarajuće mjesto u lekciji.
(iv) Koeficijent korelacije *r* je mjera linearne zavisnosti dviju veličina (dviju serija podataka). Taj je broj između -1 i 1. Ako je *r* blizu 1, to je visoka pozitivna, a ako je blizu -1 to je visoka negativna koreliranost. Ako je, pak, *r* blizu nule koreliranost je vrlo niska.
(v) Pogledajte formule i rješenje primjera u lekciji.