U N I V E R Z A N A P R I M O R S K E M
Fakulteta za matematiko, naravoslovje in informacijske tehnologije

Janez Žibert

SKRIPTA ZA PREDMET

VERJETNOST IN STATISTIKA V TEHNIKI IN NARAVOSLOVJU

NA ŠTUDIJSKEM PROGRAMU
RAČUNALNIŠTVA IN INFORMATIKE NA 2. STOPNJI
UP FAMNIT

PRVA IZDAJA

Koper, 2012

Predgovor

Ta skripta so nastala iz prosojnic, ki sem jih uporabljal na predavanjih pri predmetu *Verjetnost in statistika v tehniki in naravoslovju* na 2. stopnji študijske smeri Računalništva in informatike na Fakulteti za matematiko, naravoslovje in informacijske tehnologije na Univerzi na Primorskem v letih od 2009 do 2012.

Skripta predstavljajo uvod v statistiko, ki naj bi jo poznali vsi, ki se ukvarjajo z obdelavo in analizo podatkov v naravoslovju in tehniki. Statistične metode so predstavljene v knjigi predvsem z vidika uporabe na konkretnih primerih, ne ukvarjamo pa se s strogim matematičnim izpeljevanjem in dokazovanjem posameznih statistični metod in njihovih lastnosti. V skriptih so predstavljena osnovna orodja in metode za statistično obdelavo podatkov, kjer se seznanimo z različnimi testi za preverjanje hipotez, regresijsko analizo in analizo variance.

Skripta so nastala s pomočjo naslednje literature:

- Douglas C. Montgomery, George C. Runger: *Applied Statistics and Probability for Engineers, 4th ed.*,
- Michael J. Crawley: *Statistics: An Introduction using R*,
- dokumentacija programskega paketa *Statistics Toolbox*^{TM} programa MATLAB, The MathWorks, Inc.

Podatki, ki jih uporabljamo za primere statistične analize, so v glavnem povzeti iz:

- podatkovnih zbirk programskega paketa *Statistics Toolbox*^{TM} programa MATLAB, The MathWorks, Inc.,
- knjige: Michael J. Crawley: *Statistics: An Introduction using R*, ki jih lahko najdemo na spletni strani: http://www3.imperial.ac.uk/naturalsciences/research/statisticsusingr

Ljubljana, september 2012 Janez Žibert

Kazalo

1 Osnovne motivacije za uporabo statistike 3

2 Osnovne statistične lastnosti podatkov 19

3 Naključne spremenljivke in porazdelitve 39

4 Statistična analiza vzorcev in testiranje hipotez 59

5 Regresijska analiza 100

6 Analiza variance in analiza kovariance 142

A Osnove programskega jezika MATLAB 173

1 Osnovne motivacije za uporabo statistike

V tem poglavju podamo nekaj osnovnih motivacij za uporabo statistične analize na podatkih. Seznanimo se s splošnimi pojmi statistične analize, kot so variabilnost, statistična značilnost in testiranje domnev, statistično modeliranje, pridobivanje podatkov in interpretacija rezultatov, načrtovanje in analiza eksperimentov ter ugotavljanje odvisnosti ali neodvisnosti med podatki.

Primer namesto uvoda

- **Merimo višino 10-letnih deklic in dečkov. Imamo 15 dečkov in 15 deklic.**
  - Kaj merimo?
    - Višino. To je količina, ki se spreminja = spremenljivka Y .
  - Rezultati meritev?
    - Izmerjene višine otrok = vrednosti spremenljivke Y.
  - Kakšen tip izmerjenih vrednosti?
    - Realne vrednosti – Y zavzame realne vrednosti = zvezna spremenljivka.
  - Katere kategorije obravnavamo?
    - Dečke in deklice – kategorija = spremenljivka X diskretna (številne vrednosti)
    - Lahko tudi: vsak otrok je kategorija zase:
      spr. X zavzame vrednosti od 1 do 30, še vedno diskretna spr.

Primer namesto uvoda

- **Merimo višino 10-letnih deklic in dečkov. Imamo 15 dečkov in 15 deklic.**
  - Kaj nas zanima?
    - Y v odvisnosti od X.
      Y je odzivna spremenljivka (ang. *response variable*)
    - Ali, kako se vrednosti Y spreminjajo glede na vrednosti X.
      X je pojasnjevalna spremenljivka (ang. *explanatory variable*)
  - Denimo, da ugotovimo, da so dečki v povprečju višji od deklic v tej skupini. Vprašanja:
    - Ali gre naši ugotovitvi **verjeti**?
    - Ali lahko sklepamo, da je tako pri vsej populaciji 10-letnikov?
    - Torej: imamo **vzorec** 30 primerkov in bi radi na podlagi teh meritev sklepali na celotno **populacijo**. Ali je vzorec dovolj velik, reprezentativen?
    - Ali lahko razvrstimo dečke in deklice na podlagi njihove višine v dva razreda?

Variabilnost

- **Statistika se ukvarja z variabilnostjo!**
  - Če merimo iste količine dvakrat zapored, bomo dobili različne rezultate. ► **časovna različnost**
  - Če merimo iste količine na različnih mestih, bomo dobili različne rezultate. ► **prostorska različnost**

- Potrebujemo znanje (orodja) za preučevanje variabilnosti:
  - da ugotavljamo, kaj vpliva na rezultate,
  - da ugotavljamo, kako nastajajo rezultati,
  - da ugotavljamo, kako so rezultati medsebojno odvisni.

Statistična značilnost ugotovitev/rezultatov

- Kaj pomeni, ko rečemo, ta rezultat je signifikanten?
  - Slovar tujk: pomemben, značilen.
  - V statistiki: rezultat se zelo verjetno ni zgodil po naključju.

- Pojmi:
  - 'se zelo verjetno ni zgodil' - kaj to pomeni?
    Npr., da se dogodek zgodi v manj kot 5% primerov. **Verjetnost**.

  - 'po naključju' – torej ni v povezavi s tistim, kar smo predvidevali?
    Predpostavljamo neko odvisnost naših meritev od nekih pojavov, lastnosti in preverjamo ali ta odvisnost obstaja. Imamo neko **hipotezo**, ki je preverjamo ali drži ali ne.

Dobre in slabe hipoteze

- Karl Popper (avstrijski filozof 1902-1994):
  Dobra hipoteza je tista, ki jo je moč zavrniti.

- Primer dveh hipotez:
  - Obe hipotezi obravnavajo veverice v parku.
    Vendar eno hipotezo je možno izpodbijati.

V parku so veverice.	**V parku ni veveric.**
Gremo v park in opazujemo veverice. Pa jih ni. Ali to pomeni, da veveric v parku ni? Mogoče pa se skrivajo in jih ne boš nikoli opazil. Vse kar lahko rečemo je, bil sem v parku, pa nisem opazil nobene veverice. **To, da nimaš dokazov, ni dokaz, da dokazov ni.**	Gremo v park in opazujemo veverice. Brž, ko opazimo eno veverico, lahko zavrnemo hipotezo.

Hipoteze

- **Ničta hipoteza**: *nič se ne dogaja*
- **Alternativna hipoteza**: *nekaj se dogaja*

- Primeri ničtih hipotez:
  - primerjamo povprečja dveh vzorcev podatkov:
    ničta hipoteza: *povprečja obeh vzorcev sta enaka.*
  - imamo podatke dveh količin *Y* in *X:*
    ničta hipoteza: količini *Y* in *X* sta neodvisni.

- Ničto hipotezo zavrnemo, ko iz podatkov (vzorca) ugotovimo, da se hipoteza zelo verjetno ne more zgoditi.
  - primer: V parku ni veveric.

Napake pri napovedovanju pravilnosti hipoteze

- Pri napovedovanju pravilnosti hipoteze lahko naredimo dva tipa napak:
  - hipotezo lahko zavrnemo, čeprav je pravilna,
  - hipotezo lahko sprejmemo, čeprav je nepravilna.

- Napake Tipa I (false negative) in Tipa II (false positive)

	dejansko stanje	
napovedano stanje	***pravilno***	***nepravilno***
pravilno	pravilna odločitev	tip II
nepravilno	tip I	pravilna odločitev

Moč testa

- Moč testa je verjetnost zavrnitve ničte hipoteze, ko je ta napačna. Povezan je z napako tipa II.

- Označimo: β je verjetnost sprejetja ničte hipoteze, ko je ta napačna.

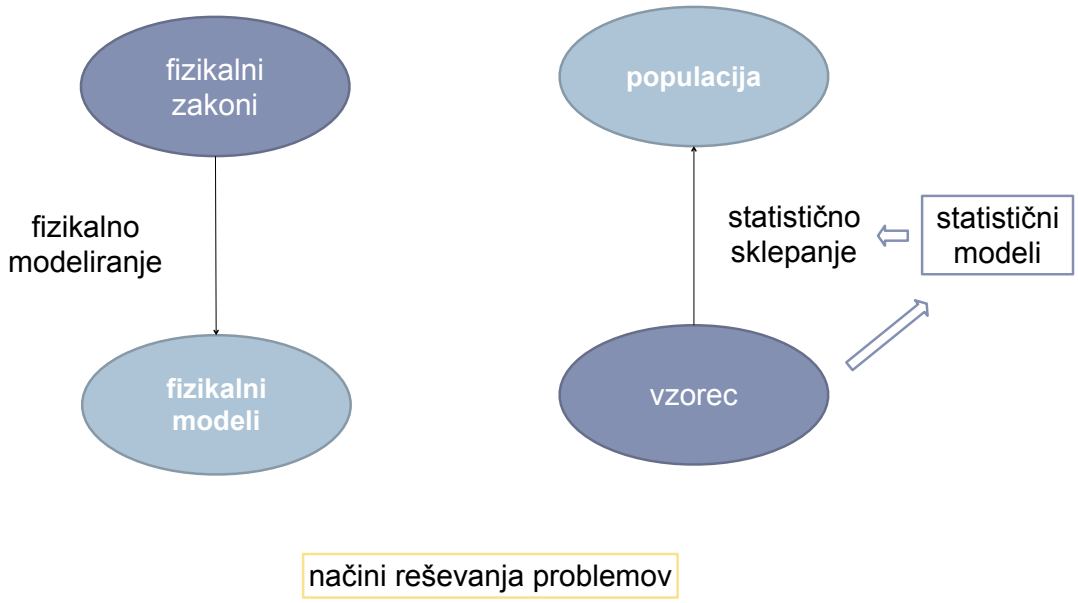
- V idealnem primeru bi morala biti β čim manjša, toda pri tem bi lahko povzročali napake tipa I (zavrnitve pravilne hipoteze).

- Potreben je kompromis:
  - običajno $\alpha = 0.05$ (tip I), $\beta = 0.2$ (tip II).

- Moč testa: $1- \beta$

- Iz tega lahko izračunamo velikost vzorcev za potrditev/zavrnitev hipoteze.

Modeliranje problemov



Statistično modeliranje

- Pri statističnem modeliranju gre za ocenjevanje parametrov modelov (matematičnih funkcij) glede na podatke, ki so na voljo.
- Zakaj to delamo?
  - da iz podatkov dobimo neko znanje, ki ga uporabimo na novih – neznanih podatkih,
  - statistično sklepanje: napovedovanje in razvrščanje
- Pri statističnem modeliranju:
  - izbiramo med različnimi modeli
  - in med njimi izberemo tistega, ki z najmanj parametri, najboljše opiše dane podatke.

► minimalnost modela
► najboljše ujemanje modela s podatki - optimalnost

Stat. modeliranje: optimalnost modela

- Kaj pomeni najboljše ujemanje modela s podatki?

- Modeliranje:
  - imamo podatke,
  - izbrali smo model,

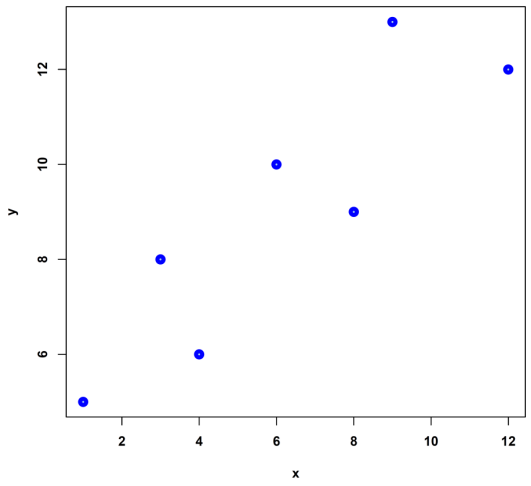
  - Ena možnost: **metoda najmanjše kvadratne napake**
    - parametre modela določimo tako, da bo skupna razdalja med podatki in modelom najmanjša.

  - Druga možnost: **metoda maksimalnega verjetja**
    - parametre modela določimo tako, da bodo dani podatki najbolj verjetno generirani iz našega modela.

Primer: Metoda maksimalnega verjetja

- Imamo podatke:

```
x<-c(1,3,4,6,8,9,12)
y<-c(5,8,6,10,9,13,12)
plot(x,y)
```

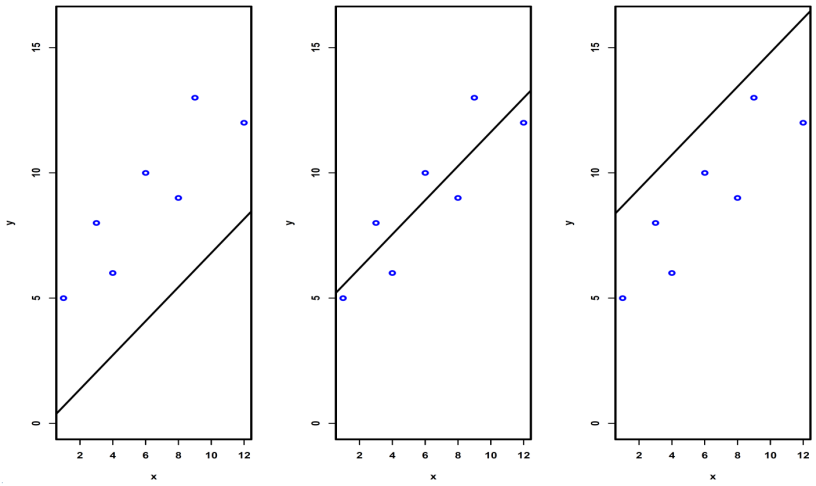


Primer: Metoda maksimalnega verjetja

- Izberemo model: npr. regresijsko premico

$$y = a + bx$$

  - Predpostavimo, da poznamo koeficient ***b = 0.68 (naklon)***

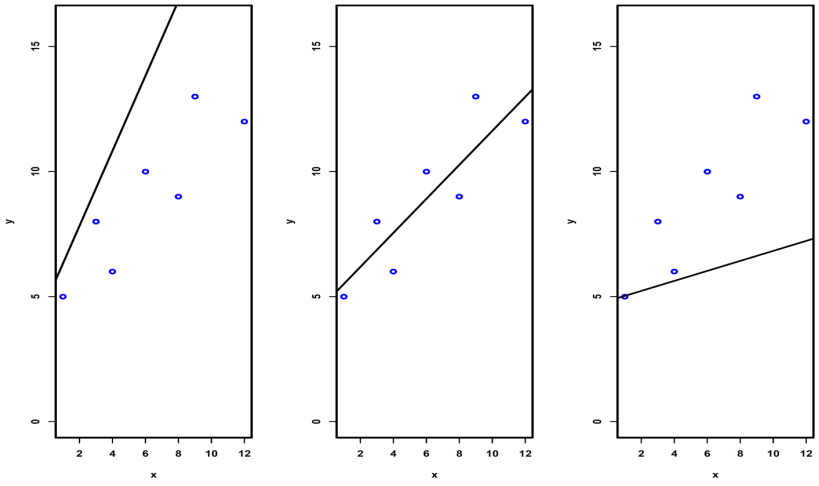


Primer: Metoda maksimalnega verjetja

- Izberemo model: npr. regresijsko premico

$$y = a + bx$$

  - Predpostavimo, da poznamo koeficient ***a = 4.827 (odsek na osi y)***



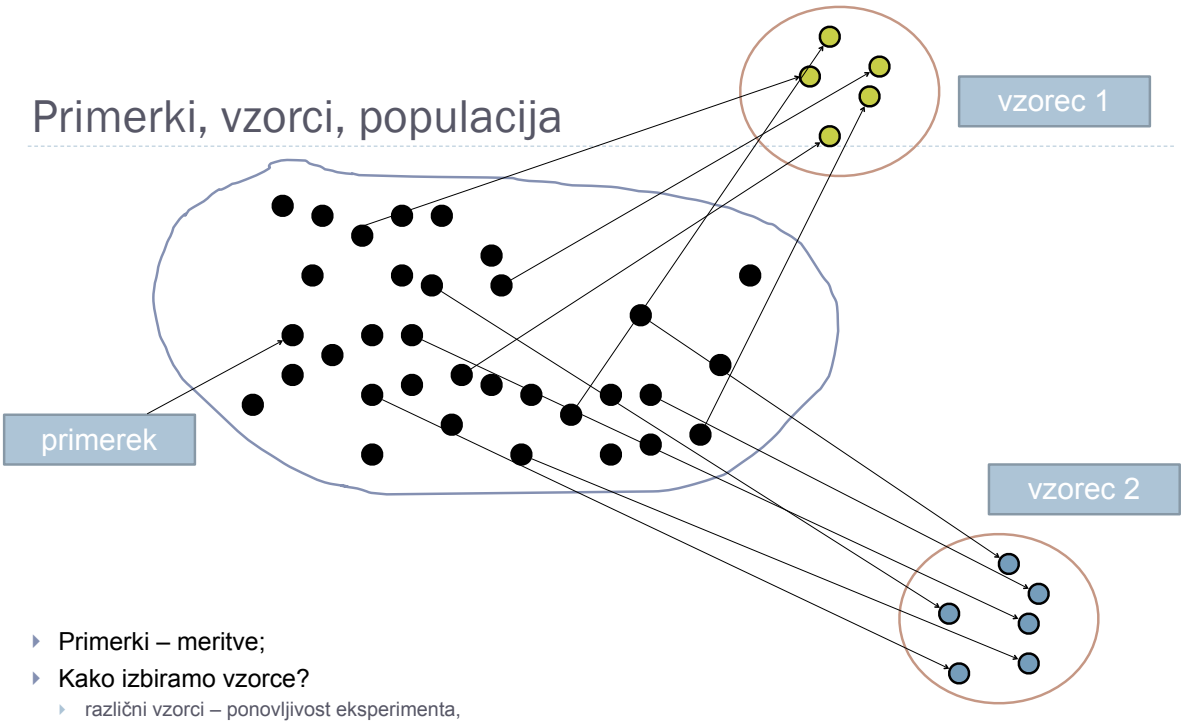
Stat. modeliranje: izbira modela

- Princip ekonomičnosti (Occam's razor):
  - *Prava razlaga (problema) je najbolj enostavna razlaga (problema).*
- Princip ekonomičnosti v statistiki:
  - modeli naj imajo čim manj parametrov,
  - bolje uporabljati linearne (preproste) modele kot nelinearne (bolj kompleksne),
  - modeli, ki so odvisni od manj predpostavk, so boljši od tistih, kjer je predpostavk več,
  - preprosta razlaga je boljša od kompleksne,
  - modele izbiramo tako, da manjšamo število parametrov, vendar kljub temu ohranjamo visoko ujemanje z danimi podatki,
- Einstein:
  - *Model mora biti preprost kar se dá, vendar ne več kot to.*

Načrtovanje eksperimentov in statistika

- Vzorec – populacija
- Ponovitve in naključnost pridobivanja vzorcev
  - ponovitve eksperimentov (več vzorcev) povečajo zanesljivost rezultatov,
  - naključnost izbire primerkov za vzorec zmanjša pristranskost pri ocenjevanju rezultatov.
- Kontrola
  - brez nadzorovanja eksperimenta, ne moremo sprejemati ustreznih zaključkov.
- Preverjanje hipotez in sklepanje
  - znanstveni problemi:
    - postavimo hipotezo, testiramo, na podlagi rezultatov sklepamo o pravilnosti hipoteze.
- Izogibanje psevdo ponovitvam
  - imamo več meritev istega pojava, primerka
- Izbira opazovanih/merjenih količin:
  - začetni pogoji
  - neodvisnosti in odvisnosti med podatki

Primerki, vzorci, populacija



- Primerki – meritve;
- Kako izbiramo vzorce?
  - različni vzorci – ponovljivost eksperimenta,
  - izbira primerkov za vzorce – naključnost
  - koliko primerkov za zanesljive ocene – moč vzorca
  - relacije med primerki v vzorcu – odvisnost, neodvisnost med različnimi meritvami

Ponovitve eksperimentov – različnost vzorcev

- Različni vzorci lahko ‘povejo’ različne stvari o populaciji.
- Potrebno je pazljivo izbirati primerke iz populacije, da dobimo dejanske ugotovitve o populaciji.
- Ena možnost povečanja zanesljivosti ugotovitev je ponovljivost poskusov, kjer v vsakem poskusu pridobimo svoj vzorec. Na ta način lahko ovrednotimo variabilnost rezultatov med vzorci in s tem sklepamo na zanesljivost naših ugotovitev.
- Kaj štejemo za različne vzorce – ‘prave’ ponovitve?
  - vzorci morajo biti pridobljeni iz poskusov, ki so med seboj neodvisni;
  - različni vzorci ne smejo biti sestavljeni iz podatkov, ki so pridobljeni iz meritev na istih primerkih, vendar v drugem času (primer: signali);
  - različni vzorci ne smejo biti sestavljeni iz podatkov, ki so pridobljeni iz meritev na primerkih, ki so prostorsko blizu skupaj (primer: populacija gozd, vzorci: drevesa ob potoku);
  - podatki različnih vzorcev morajo biti med seboj primerljivi (enake količine, enake skale meritev).

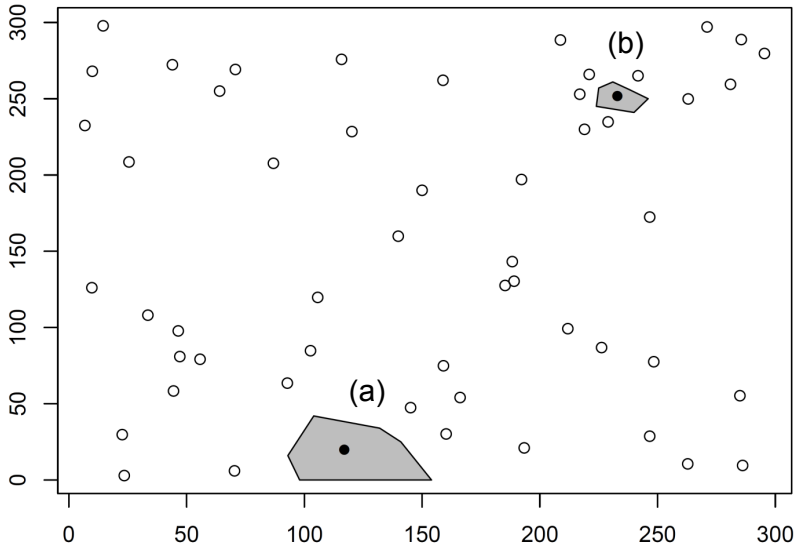
Koliko podatkov v vzorcu?

- Koliko podatkov (primerkov) naj vsebuje vzorec za zanesljive ocene parametrov?
  - Čim več.
- Kaj pa je spodnja meja števila podatkov za ‘še’ dovolj dobre zanesljive ocene?
  - Različni testi za moč vzorca za statistično signifikantnost ocene.
- Možne ovire pri pridobivanju podatkov (vzorcev):
  - kakšne eksperimente je zelo težko ponavljati,
  - nekje lahko dobimo različne vzorce, drugje pa ne,
  - mogoče imamo nadzor nad eksperimentom, se pravi, da ga opazujemo in ustrezno popravljamo razmere v eksperimentu, drugje pa imamo podatke, ki so bili že pridobljeni in eksperimenta ne moremo ponoviti – torej imamo podatke takšne kot so.
  - …

Naključnost izbire

- Kako izbiramo primerke iz populacije, da tvorimo vzorec?
- **Vsak primerek mora biti enako verjetno izbran iz populacije – naključnost izbire. Sicer je izbira pristranska.**
- *Primer:*
  - izbiramo drevo v gozdu, da bi merili učinke fotosinteze,
  - Kako izbirati drevesa, da bomo zagotovili naključnost izbire in bomo tako ocenjevali učinke fotosinteze čim manj pristransko?
    - izbiramo drevesa, ki so najbližje našemu laboratoriju,
    - izbiramo drevesa, ki imajo veje čim nižje, da bomo lahko dosegle liste, kjer bomo postavili merilne senzorje,
    - izbiramo drevesa, ki izgledajo na prvi pogled zdrava,
  - Izberemo drevo na naslednji način:
    - naključno izberemo par zemljepisnih koordinat,
    - gremo do teh koordinat v gozdu in
    - izberemo drevo, ki je najbližje tem koordinatam.
  - A je to naključna izbira?

Primer: naključnost izbire



Katero drevo je bolj verjetno, da ga izberemo?
A je to res naključna izbira dreves?

Primer: naključnost izbire

- Torej, kako dosežemo popolno naključnost izbire.
  - Označimo vsa drevesa v gozdu z zaporednimi številkami (!!!).
  - Med števili izberemo eno naključno, nato med preostalimi naslednje naključno itn. Nato gremo k drevesom, ki so označeni s temi številkami.
  - **Samo to je popolna naključnost izbire**. Alternative ni.

Psevdo ponovitve – psevdo različnost vzorcev

- Dva tipa psevdo ponovitev:
  - časovne psevdo ponovitve:
    - vključujejo ponovitve meritev istega primerka ob različnih časih
  - krajevne psevdo ponovitve
    - vključujejo ponovitve meritev v bližnji okolici istega primerka

- Zakaj je to problem?
  - osnovna predpostavka večine statističnih analiz je **neodvisnost napak posameznih meritev.**
  - Pri časovnih ponovitvah ne moremo zagotoviti neodvisnosti posameznih meritev, saj merimo isti primerek ob različnih časih in zato meritve vsebujejo posebnosti tega primerka (so časovno korelirane).
  - Pri krajevnih ponovitvah ne moremo zagotoviti neodvisnosti posameznih meritev, saj merimo (različne) primerke v isti okolici in zato meritve vsebujejo posebnosti tiste okolice (so krajevno korelirane).

Psevdo ponovitve: primer

- Primer:
  - Preučujemo vpliv delovanja insekticida na rast rastlin.
  - Imamo 20 gredic, v vsaki gredici 50 rastlin (iste sorte).
  - 10 gredic poškropimo z insekticidom, 10 pa ne.
  - Meritve opravimo 5-krat v obdobju rasti rastlin.

  - Koliko imamo meritev?
    - 20 x 50 x 5 = 5000
  - A so to med seboj neodvisne meritve?
  - Torej imamo 5000 različnih primerkov. NE.
    - Meritve istega primerka ob različnih časih niso novi primerki.
    - Meritve različnih primerkov znotraj ene gredice imajo podobne pogoje rasti, torej imajo skupne lastnosti zaradi skupne lege. Niso neodvisni primerki.
    - Neodvisne meriteve (primerki) so: 10 poškropljenih gredic in 10 nepoškropljenih.

Kako se poskušamo znebiti psevdo ponovitvam?

- Prejšnji primer:
  - če bi imeli 5000 podatkov in bi merili učinkovitost škropljenja z insekticidom, imamo 1 prostostno stopnjo za škropljenje in 4999 za napake.
  - če pa gledamo samo gredice:
    - imamo 10 škropljenih in 10 neškropljenih,
    - iz vsake gredice dobimo samo eno meritev (npr. odstotek pojedenih listov) za vsako obdobje meritev, torej imamo 20 meritev za vsako obdobje.
    - Opravimo analizo za vsako obdobje.

- Kako se poskušamo znebiti psevdo ponovitvam?
  - povprečimo meritve na psevdo primerkih in izvajamo stat. analizo samo na teh povprečenih vrednostih,
  - izvajamo stat. analizo na primerkih, ki smo jih izmerili v istem časovnem obdobju (znotraj enega časovnega obdobja)
  - uporabimo metode primerne za obdelavo časovno pogojenih meritev (obdelava signalov, DNA, ipd.)

Načrtovanje poskusov za sprejemanje ugotovitev

- Sprejemanje ugotovitev o nekem pojavu/sistemu/procesu lahko dosežemo s pravilnim načrtovanjem poskusov:
  - pripraviti je potrebno kontrolirano okolje za izvajanje poskusa,
  - identificirati je potrebno dejavnike vpliva,
  - potrebno je upoštevati in kontrolirati začetne pogoje poskusa,
  - potrebno je upoštevati zveze/relacije med različnimi meritvami,
  - izvajati je potrebno določeno število meritev, da lahko dovolj zanesljivo sklepamo o neki predpostavki/hipotezi,
  - postavit je potrebno jasno hipotezo, ki jo s poskusom preverjamo.

Začetni pogoji

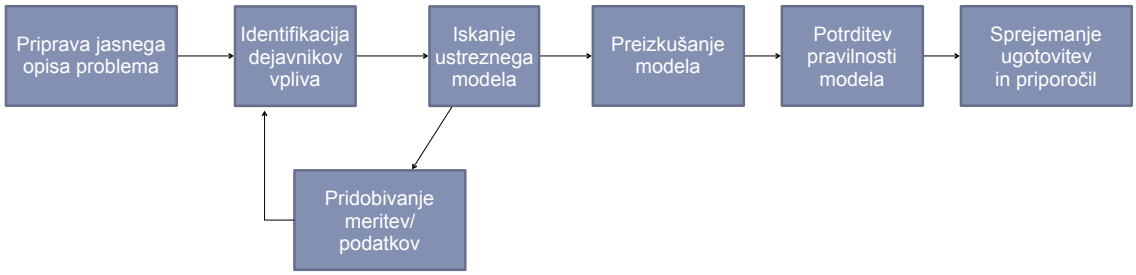
- Začetni pogoji so lahko pomemben dejavnik, ki vplivajo na končne rezultate poskusa.
- Kako lahko povemo, da se je kaj spremenilo, če ne vemo, v kakšnem stanju je bilo na začetku?
  - Primer opazovanja rasti nekih organizmov,
    - Ali lahko sklepamo, kako veliki postanejo organizmi, če ne vemo kakšnih velikosti so bili na začetku?

Soodvisnost med različnimi meritvami

- Potrebno je detektirati soodvisnosti med posameznimi faktorji vpliva.
- Posamezne količine so lahko med seboj odvisne ali neodvisne.
  - neodvisnost pomeni:
    - da sprememba ene količine ne vpliva na spremembo druge količine,
    - da če poznamo lastnosti ene količine, ne moremo nič povedati o lastnostih druge.
- Statistična analiza mora upoštevati soodvisnosti med količinami.

Reševanje problemov v tehniki in naravoslovju

- **Inženir** – nekdo, ki rešuje probleme v interesu javnosti (družbe, okolja) z “učinkovito” uporabo znanstvenih spoznanj in odkritij.
- **Inženirske ali znanstvene metode** so postopki formulacije in reševanja teh problemov.



inženirski pristop reševanja problema

2 Osnovne statistične lastnosti podatkov

V tem poglavju obravnavamo osnovne statistične lastnosti podatkov, ki jih običajno podajamo pri osnovni statistični analizi podatkov. Obravnavamo jih tudi pod pojmom opisna statistika.

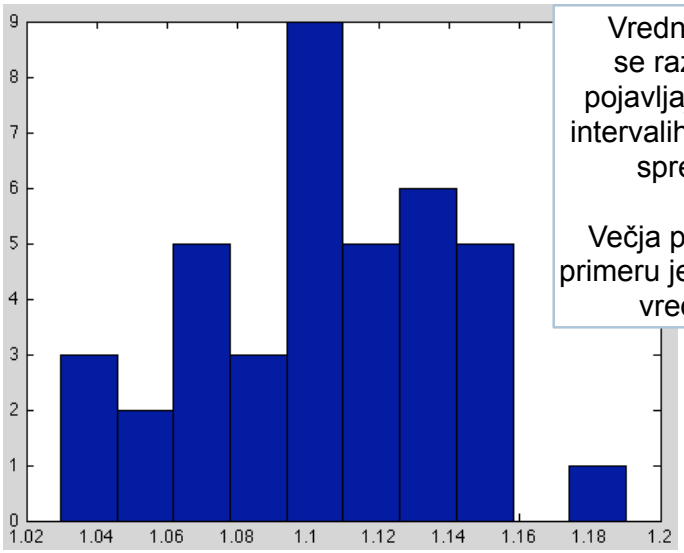
Poglavje obravnava:

- osnovne mere centralne tendence,
- osnovne mere razpršitve vrednosti ter
- nekaj osnovnih grafov, ki jih v statistiki uporabljamo za prikazovanje podatkov.

Pogostnost določenih vrednosti v podatkih

```
data = importdata('datasets/yvalues.txt');
y = data.data;
hist(y);
```

Porazdelitev vrednosti spr. y



Vrednosti v podatkih se različno pogosto pojavljajo v posameznih intervalih zaloge vrednosti spremenljivke y.

Večja pogostnost v tem primeru je v okolici središča vrednosti spr. y.

Aritmetična sredina

- Aritmetična sredina – povprečje:

$$\overline{y} = \frac{\sum y}{n}$$

- V Matlabu:

```
function m = m_mean(x)
    m = sum(x)/length(x);
end

>> mean(y)
ans =
    1.1035

>> m_mean(y)
ans =
    1.1035
```

Mediana

- Slabost aritmetične sredine je, da je zelo občutljiva na točke, ki zelo odstopajo (ang. *outliers*). Več kot je takih točk, slabša je ocena.
- Središče, ki je manj občutljivo na takšne točke, je **mediana = to je vrednost podatka, ki leži točno na sredini vseh podatkov, ki so urejeni po svojih vrednostih.**
- Izračun:

liho število podatkov

```
>> sort_y = sort(y);
>> length(y)/2
ans =
   19.5000
>> ceil(length(y)/2)
ans =
    20
>> sort_y(ceil(length(y)/2))
ans =
    1.1088
```

sodo število podatkov

```
>> y_even = y(1:38);
>> sort_y = sort(y_even);
>> ly = sort_y(19)
ly =
    1.1040
>> ry = sort_y(20)
ry =
    1.1088

>> (ly + ry)/2
ans =
    1.1064
```

Mediana

- Funkcija v Matlabu

```
function m = m_median(x)

x_sort = sort(x);
len_x = length(x);

odd_even = mod(len_x,2);

if (odd_even == 0)
    m = (x_sort(len_x/2) + x_sort(len_x/2+1)) / 2;
else
    m = x_sort(ceil(len_x/2));
end


>> median(y)
ans =
    1.1088

>> m_median(y)
ans =
    1.1088
```

Geometrična sredina

- Primer:
  - Štejemo insekte na rastlinah v 5-ih gredicah.
  - V prvi smo prešteli 10, v drugi 1, v tretji 1000, v četrti 1 in v peti 10 insektov.
  - Izračunajmo povprečje: (10+1+1000+1+10)/5 = 204.4
  - A je to zadovoljiva ocena ‘povprečnega’ dogajanja v gredicah? Težava je 1000 (insekti se razmnožujejo zelo hitro, mogoče jih je bilo včeraj samo 10, danes jih je pa že 1000. Proces se spreminja multiplikativno.)

- Geometrična sredina

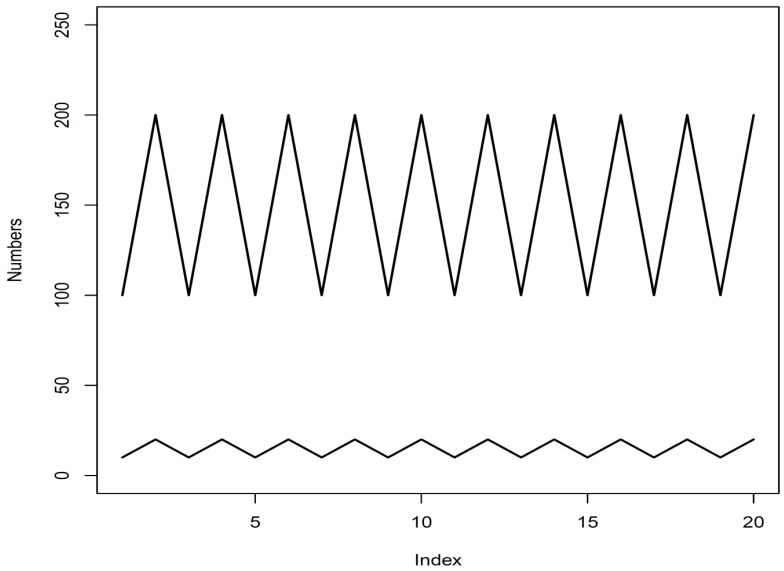
$$\hat{y} = \sqrt[n]{\prod y}$$

Geometrična sredina

- V Matlabu:

```
>> insekti = [1,10,1000,10,1];
>> mean(insekti)
ans =
  204.4000

>> exp(mean(log(insekti)))
ans =
   10.0000

>> geomean(insekti)
ans =
   10.0000
```

Logaritemska skala



- Imamo dva procesa. Oba alternirajoče spreminjata vrednosti.
- Katera bolj varira? A gre za enak proces?

Harmonična sredina

- Primer:
  - Slon se giblje v območju, ki je v obliki kvadrata s stranico 2km.
  - Vsak dan slon enkrat obhodi svoje območje:
    - Zjutraj, da se pretegne, gre po eni stranici s hitrostjo 1 km/h.
    - Potem nadaljuje po drugi stranici s hitrostjo 2 km/h (je že malce zbujen).
    - Na kar poveča hitrost na 4 km/h (je že ogret) na tretji stranici.
    - Potem pa je že utrujen in njegova hitrost pade na 1 km/h na zadnji stranici.
  - Kakšna je njegova povprečna hitrost?
    - (1+2+4+1)/4 = 8/4 = 2 km/h. NAPAKA.
  - Hitrost se izračuna kot pot/čas.
    - Pot: 4x2 km = 8 km.
    - Čas:
      - prva stranica 2km in hitrost 1km/h, torej 2h
      - druga stranica 2km in hitrost 2km/h, torej 1h
      - tretja stranica 2km in hitrost 4km/h, torej 0.5h
      - četrta stranica 2km in hitrost 1km/h, torej 2h
      - SKUPAJ: 5.5 h
    - Povprečna hitrost je torej: 8km / 5.5h = **1.4545 km/h**.

Harmonična sredina

- Tak primer rešimo z izračunom harmonične sredine.
- **Harmonična sredina je recipročna vrednost od povprečja recipročnih vrednosti:**

$$\tilde{y} = \frac{n}{\sum \frac{1}{y}}$$

```
>> v = [1,2,4,1];
>> length(v)/sum(1./v)
ans =
    1.4545

>> 1/mean(1./v)
ans =
    1.4545

>> harmmean(v)
ans =
    1.4545
```

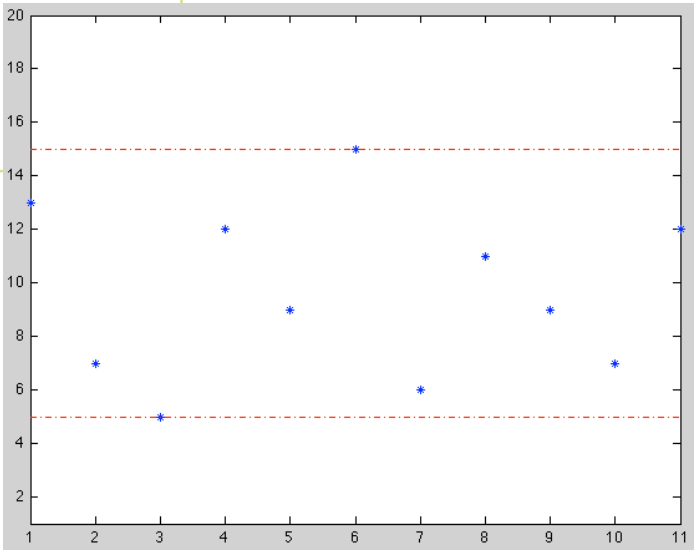
Variabilnost v podatkih

- Količina za merjenje variabilnosti ali razpršenosti v podatkih je **varianca**.

- Večja kot je variabilnost v podatkih:
  - večja bo negotovost v ocenjene statistične parametre,
  - manjša bo zanesljivost potrjevanja (zavrnitve) hipoteze.

Mera razpršenosti: rang

```
>> y = [13,7,5,12,9,15,6,11,9,7,12];
plot(y, '*')
hold on;
plot(min(y).*ones(1,11), 'r-.')
plot(max(y).*ones(1,11), 'r-.')
hold off
>> range(y), min(y), max(y)
ans =
    10
ans =
     5
ans =
    15
```

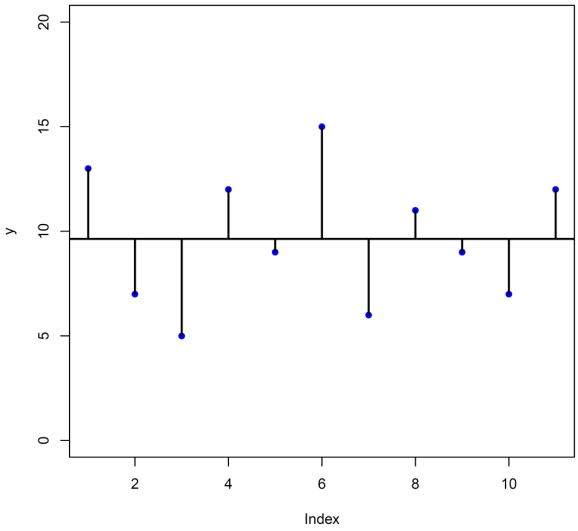


Mera razpršenosti: razlika od povprečne vrednosti

- Razlika od povprečne vrednosti:

$$\sum (y - \overline{y})$$

- A je to dobra mera?
- NE.
  Zgornja vsota je vedno 0.



Mera razpršenosti: razlika od povprečne vrednosti

- Vsota absolutnih vrednosti razlik od povprečja:

$$\sum |y - \overline{y}|$$

- Vsota kvadratov razlik od povprečja:

$$\sum (y - \overline{y})^2$$

  - Enote: dolžina [mm] - ploščina [mm^2]
  - Problem, vsakič ko dodamo novo razliko na kvadrat se vsota poveča. Torej mera je odvisna od števila podatkov. To ni v redu.
  - Naredimo povprečje kvadratov razlik. Boljše. Vendar ocena pristranska.

Prostostne stopnje

- Denimo, da poznamo vsoto 5-ih števil, ki znaša 20. Koliko števil lahko prosto določimo, da bo vsota 20?

2				
2	7			
2	7	4		
2	7	4	0	

- Zadnja možnost je samo ena, ker mora biti vsota 20.

2	7	4	0	7

- Število prostostnih stopenj, če poznamo vsoto 5-ih števil, je 4.

Prostostne stopnje

- Če ocenimo povprečje iz *n* podatkov, nam ostane samo še *n-1* prostostnih stopenj.

- Splošno:
  Število prostostnih stopenj je enako številu podatkov n (primerkov v vzorcu) minus številu ocenjenih parametrov p iz teh podatkov.

Varianca

$$\text{varianca} = s^2 = \frac{\sum (y - \bar{y})^2}{n - 1}$$

- Matlab:

```
function s = m_var(x)

s = sum((x-mean(x)).^2) / (length(x)-1);


>> m_var(y)
ans =
   10.2545

>> var(y)
ans =
   10.2545
```

Kako izračunamo vsoto razlike kvadratov?

- Kvadrat razlike:

$$(y-\overline{y})^2 = (y-\overline{y})(y-\overline{y}) = y^2 - 2y\overline{y} + \overline{y}^2$$

- Vsota kvadratov razlike:

$$\sum y^2 - 2\overline{y}\sum y + n\overline{y}^2 = \sum y^2 - 2\frac{\sum y}{n}\sum y + n[\frac{\sum y}{n}]^2$$

- Izpeljava:

$$\sum y^2 - 2\frac{[\sum y]^2}{n} + n[\frac{\sum y}{n}]^2 = \sum y^2 - \frac{1}{n}[\sum y]^2$$

- Torej za izračun potrebujemo samo:

$$\sum y^2 \quad \text{in} \quad \sum y$$

Primer: uporabe variance

```
>> garden = importdata('datasets/gardens.txt')
garden =
          data: [10x3 double]
      textdata: {1x3 cell}
    colheaders: {1x3 cell}

>> header = garden.textdata
header =
    'gardenA'    'gardenB'    'gardenC'

>> data = garden.data
data =
     3     5     3
     4     5     3
     4     6     2
     3     7     1
     2     4    10
     3     4     4
     1     3     3
     3     5    11
     5     6     3
     2     5    10
```

```
>> mean(data(:,1))
ans =
     3
>> var(data(:,1))
ans =
    1.3333

>> mean(data(:,2))
ans =
     5
>> var(data(:,2))
ans =
    1.3333

>> mean(data(:,3))
ans =
     5
>> var(data(:,3))
ans =
   14.2222
```

Primer

```
>> mean(data(:,1))
ans =
     3
>> var(data(:,1))
ans =
    1.3333

>> mean(data(:,2))
ans =
     5
>> var(data(:,2))
ans =
    1.3333

>> mean(data(:,3))
ans =
     5
>> var(data(:,3))
ans =
   14.2222
```

- Kaj lahko rečemo na podlagi teh rezultatov?
- Varianci vrta A in B sta enaki.
  Povprečja različna.
  Ali sta vzorca enaka?
  - Studentov t-test: dva vzorca.
  - ANOVA: tri ali več vzorcev.
- Povprečja vrta B in C sta enaka.
  Variance različne.
  A lahko rečemo, da so vzorci z enakimi povprečji enaki? NE!
  - Denimo, da je prag za poškodovanje rastlin zaradi ozona 8 pphm.
  - Ker so izmerjene vrednosti v obeh vrtovih B in C v povprečju pod pragom, je vse v redu. PA NI.
  - Poglejmo vrt C:

```
>> data(:,3)'
     3     3     2     1    10     4     3    11     3    10
```

  - V 30% meritev so rastline podvržene višjim koncentracijam ozona od mejnih.

Uporaba variance

- Varianco uporabljamo v dveh primerih:
  - **za merjenje zaupanja v ocene parametrov – intervali zaupanja v ocene,**
  - **za testiranje hipotez.**

Momenti višjega reda

- Povprečje (moment prvega reda)

$$\overline{y} = \frac{\sum y}{n}$$

- Varianca (moment drugega reda)

$$s^2 = \frac{\sum (y - \overline{y})^2}{n - 1}$$

- Momenti višjega reda vključujejo razlike višjih potenc (>2)

$$\sum (y - \overline{y})^3 \qquad \sum (y - \overline{y})^4$$

Koeficient asimetričnosti (skewness)

- Tretji moment:

$$m_3 = \frac{\sum (y - \overline{y})^3}{n}$$

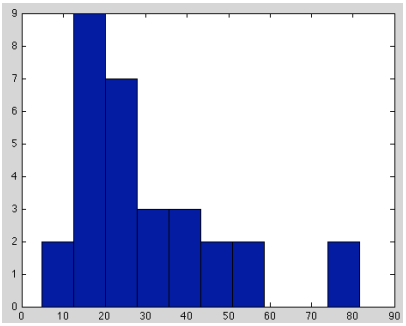
- Koeficient asimetričnosti:

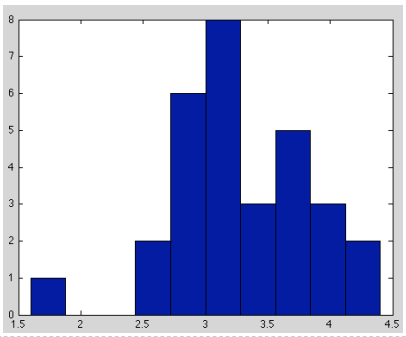
$$\gamma_1 = \frac{m_3}{s^3}$$

  - Mera asimetričnosti porazdelitve: ali ima porazdelitev daljši rep na levi (koeficient je negativen, negativna asimetrija) ali na desni (koeficient je pozitiven, pozitivna asimetrija).
  - Normalna porazdelitev je simetrična zato je koeficient = 0.

Koeficient asimetričnosti

```
function s = m_skew(x)

m3 = sum((x-mean(x)).^3) / length(x);
s3 = std(x)^3;

s = m3/s3;



>> data = importdata('datasets/skewdata.txt');
>> data = data.data;
>> hist(data)
>> m_skew(data)
ans =
    1.3877
>> skewness(data)
ans =
    1.3877


>> hist(log(data))
>> m_skew(log(data))
ans =
   -0.3106
>> skewness(log(data))
ans =
   -0.3106
```





Koeficient sploščenosti (kurtosis)

- Četrti moment:

$$m_4 = \frac{\sum (y - \bar{y})^4}{n}$$

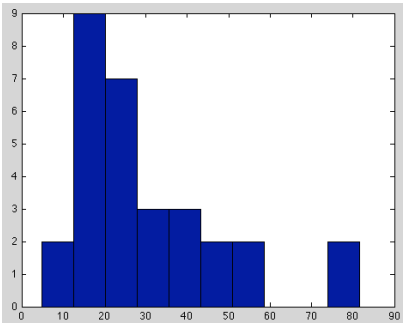
- Koeficient sploščenosti:

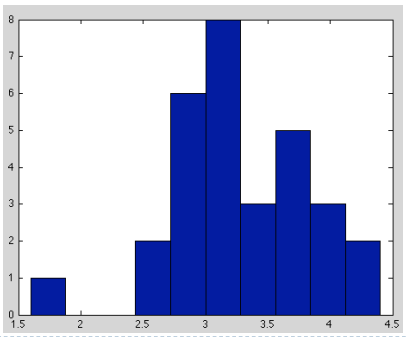
$$\gamma_2 = \frac{m_4}{s^4} - 3$$

  - Normalna porazdelitev ima koeficient $m_4/s^4 = 3$ zato odštejemo 3.
  - Mera sploščenosti porazdelitve: Normalna porazdelitev ima po popravku koeficient = 0.
  - Bolj sploščena porazdelitev od normalne ima negativen koeficient, bolj točkasta ima pozitiven koeficient.

Koeficient sploščenosti (kurtosis)

```
function s = m_kurtosis(x)

m4 = sum((x-mean(x)).^4) / length(x);
s4 = std(x,1)^4;

s = m4/s4-3;



>> data = importdata('datasets/skewdata.txt');
>> data = data.data;
>> hist(data)
>> m_kurtosis(data)
ans =
    1.5993
>> kurtosis(data)-3
ans =
    1.5993

>> hist(log(data))
>> m_kurtosis(log(data))
ans =
    0.9785
>> kurtosis(log(data))-3
ans =
    0.9785
```





Mere oblike podatkov

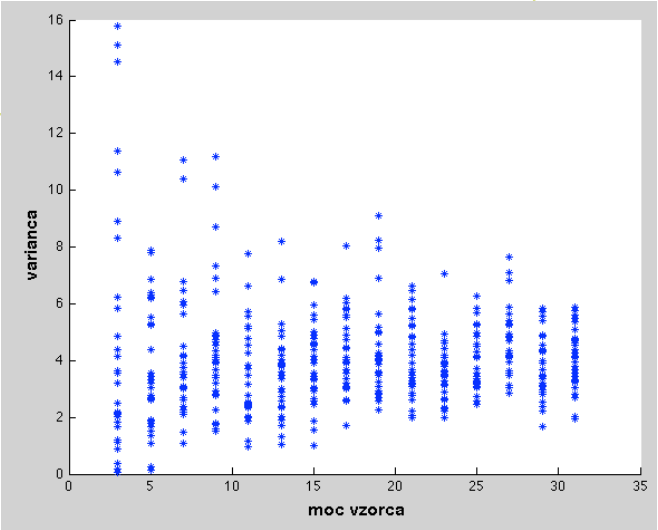
- Kvantili
  - kvantil za $0 \leq p \leq 1$ je vrednost v podatkih, pri katerih je p-ti delež razvrščenih vrednosti na levi strani in (1-p)-ti delež razvrščenih vrednosti na desni strani.
- Percentili
  - percentil je kvantil, kjer so vrednosti p zapisane v odstotkih, torej $0 \leq p \leq 100$.

```
>> data = importdata('datasets/das.txt');
>> data = data.data;

>> quantile(data,[.025 .25 .50 .75 .975])
ans =
    1.9371    2.2409    2.4141    2.5696    2.9223

>> prctile(data, [2.5, 25, 50, 75, 97.5])
ans =
    1.9371    2.2409    2.4141    2.5696    2.9223
```

Ocenjevanje parametrov: ocena variance in moč vzorca

- Primer:

```
hold on;
for df = 3 : 2 : 31
    for i = 1:30
        data =  random('Normal',10,2, 1, df);
        plot(df, var(data), '*')
    end
end
hold off;
```

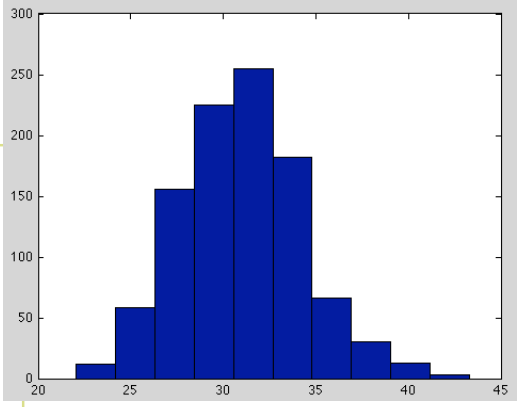


- Varianca je 4.
- Ocene variance so prikazane na sliki:
- **Manjša kot je moč vzorca, slabša je ocena variance.**

Zaupanje v oceno parametra: bootstraping

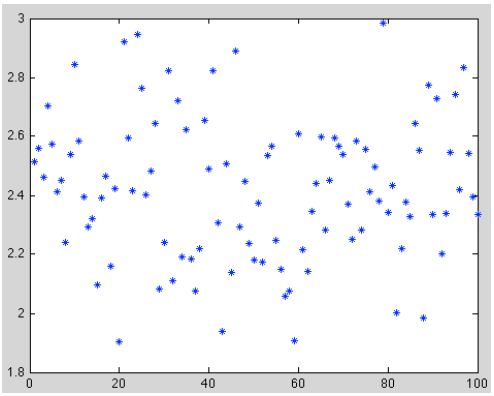
- Ocenjevanje intervalov zaupanja se lahko izvaja na več načinov.
  - lahko se predpostavi neko porazdelitev podatkov in se na podlagi tega ocenjuje zaupanje v oceno parametrov te porazdelitve
  - drugi način: **metoda ponovnega vzorčenja (ang. bootstraping)**
    - Iz danih podatkov večkrat naključno izberemo podatke, ki se lahko tudi ponavljajo in vsakič ocenimo vrednost parametra.
    - Tako dobimo različne ocene parametra.
    - Lahko ocenimo parameter in interval zaupanja v oceno.
    - Primer:

```
>> data = importdata('datasets/skewdata.txt');
>> data = data.data;
>> mean(data)
ans =
   30.9687

>> [m, bootsamp] = bootstrp(1000,@mean,data);
>> hist(m)
>> quantile(m, [0.025, 0.5, 0.975])
ans =
   24.8522   30.9459   38.2191
```

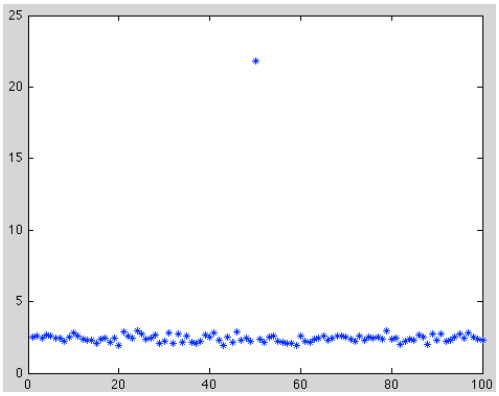


Prikazovanje podatkov s točkami v ravnini

```
>> data = importdata('datasets/das.txt');
>> data = data.data;
>> plot(data, '*')
```

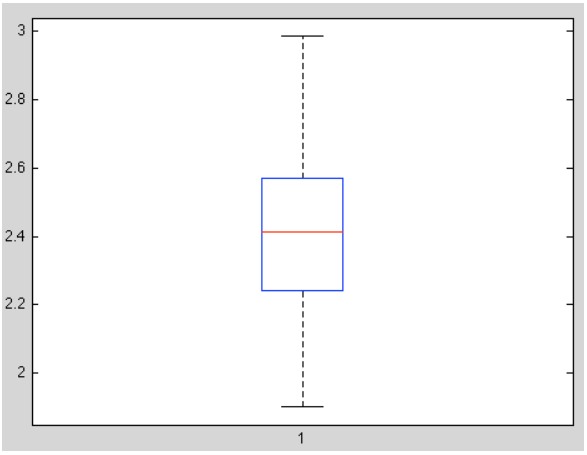



```
>> data(50) = 21.79386;
>> plot(data, '*')
>> find(data > 10)
ans =
     50
>> data(50) = 2.179386;
```

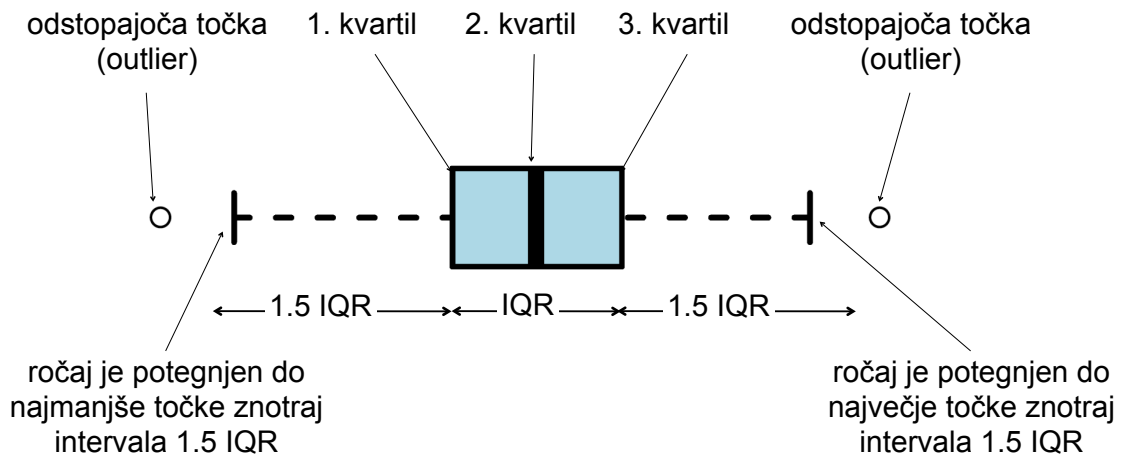


Okvir z ročaji

```
>> boxplot(data)
```

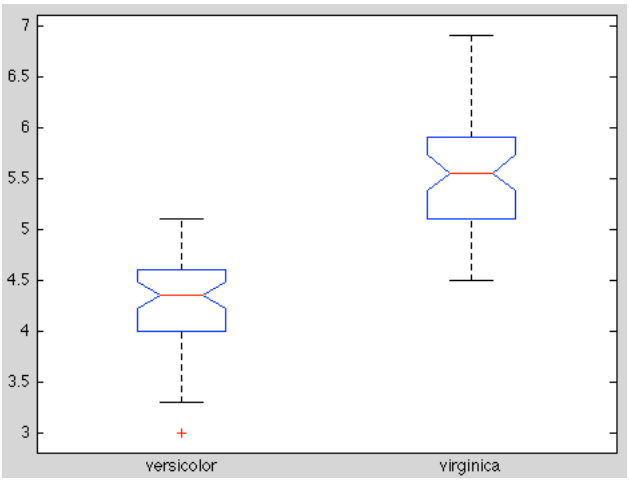


Okvir z ročaji



Okvir z ročaji

```
>> load fisheriris
>> s1 = meas(51:100,3);
>> s2 = meas(101:150,3);
>> boxplot([s1 s2],'notch','on',...
        'labels',{'versicolor','virginica'})
```



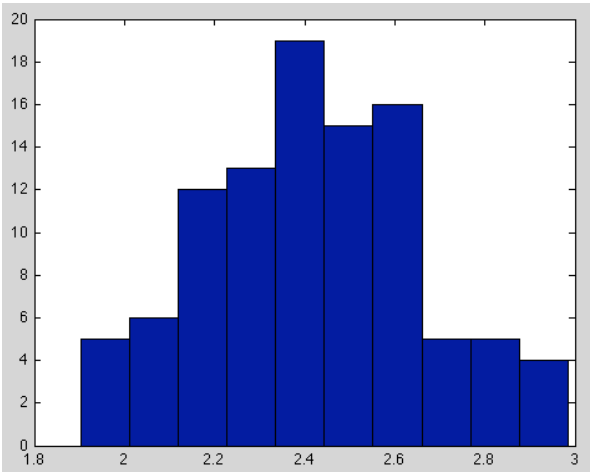
Zareza je namenjena primerjanju ocen median med dvema vzorcema.

Širina zareze je izračunana tako, da v primeru, ko se zarezi ne prekrivata, pomeni, da se mediani razlikujeta s 5% statistično značilnostjo (ob predpostavki normalne porazdelitve).

To lahko potrdimo s t-testom (v nadaljevanju).

Histogram

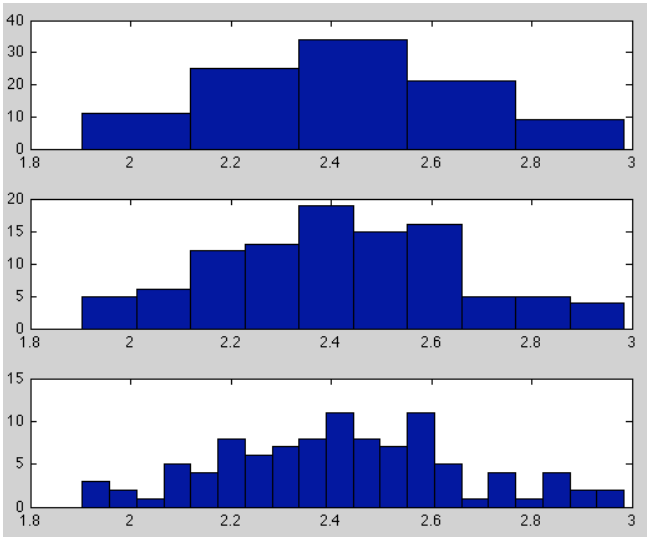
```
>> data = importdata('datasets/das.txt');
>> data = data.data;
>> hist(data)
```



- Histogram:
  - dobimo, če razdelimo rang vrednosti podatkov na manjše podintervale in štejemo število podatkov, ki padejo v posamezne podintervale.
  - na ta način dobimo porazdelitev podatkov po vrednostih.

Histogram

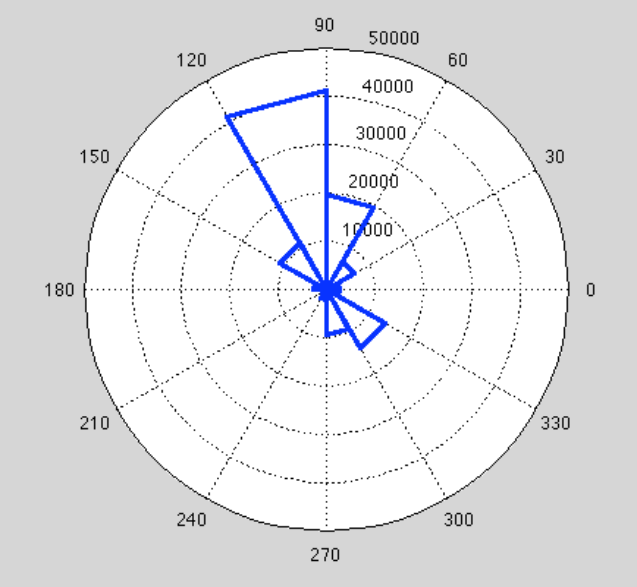
- Občutljivost histograma na izbiro števila podintervalov:




```
>> subplot(3,1,1)
>> hist(data,5)

>> subplot(3,1,2)
>> hist(data,10)

>> subplot(3,1,3)
>> hist(data,20)
```

Histogram podatkov v krogu: roža

```
load datasets/wind_dir_Koper.mat %8 let vetra na 15 min: Koper
rose(wind_dir_Koper, 12) %smeri v rad, 12 odsekov na kroznici
```



Grafi medsebojne razpršitve podatkov

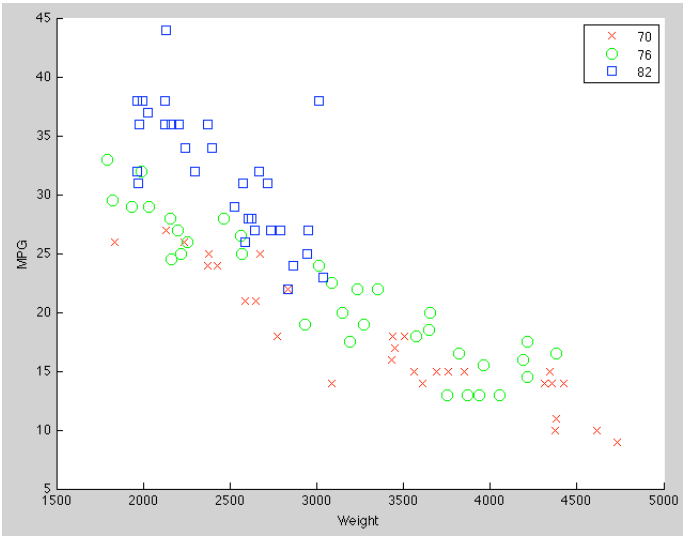
- Z grafi medsebojne razpršitve podatkov lahko opazujemo medsebojno odvisnost med različnimi tipi podatkov, ki so zajeti v vzorcu.

```
>> load carsmall
>> gscatter(Weight,MPG,Model_Year,'','xos')
```

Name	Value	Min
Acceleration	<100x1 double>	8
Cylinders	<100x1 double>	4
Displacement	<100x1 double>	85
Horsepower	<100x1 double>	NaN
MPG	<100x1 double>	NaN
Mfg	<100x13 char>	
Model	<100x33 char>	
Model_Year	<100x1 double>	70
Origin	<100x7 char>	
Weight	<100x1 double>	1795

Graf prikazuje medsebojno odvisnost med težo avtomobila in porabo goriva (število prevoženih milj na galono).

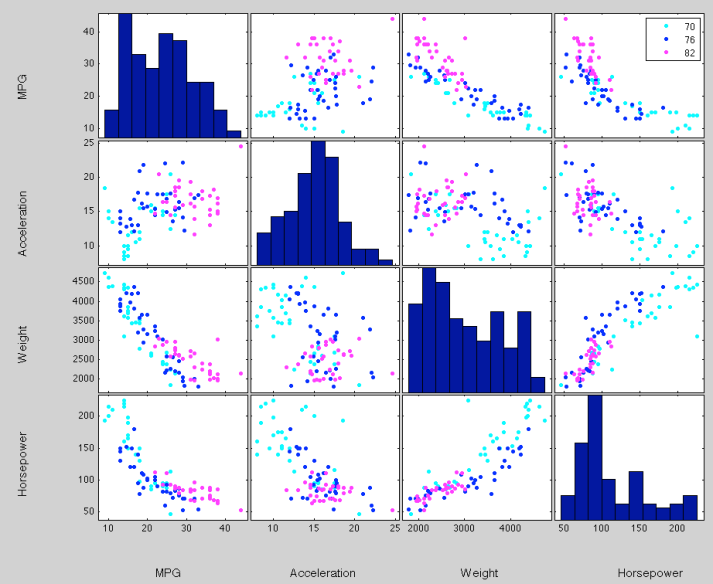
Poleg tega imamo različno obarvane avtomobile po letnikih.



Grafi medsebojne razpršitve podatkov

- Prikazovanje matrike grafov razpršitve v primeru več tipov podatkov

```
X = [MPG,Acceleration,Weight,Horsepower];
varNames = {'MPG'; 'Acceleration'; 'Weight'; 'Horsepower'};
gplotmatrix(X,[],Model_Year,['c' 'b' 'm' 'r'],[],[],'on');
text([.10 .32 .60 .86 ], repmat(-.1,1,4), varNames, 'FontSize',12);
text(repmat(-.12,1,4), [.83 .52 .31 .05 ], varNames, 'FontSize',12, 'Rotation',90);
```



Delo s podatki v tabelah

- Primer tabele: meritve cvetnih listov pri različnih tipih ene vrste rože

	species	SL	SW	PL	PW
Obs28	setosa	5.2	3.5	1.5	0.2
Obs29	setosa	5.2	3.4	1.4	0.2
Obs30	setosa	4.7	3.2	1.6	0.2
Obs89	versicolor	5.6	3	4.1	1.3
Obs90	versicolor	5.5	2.5	4	1.3
Obs91	versicolor	5.5	2.6	4.4	1.2
Obs131	virginica	7.4	2.8	6.1	1.9
Obs132	virginica	7.9	3.8	6.4	2

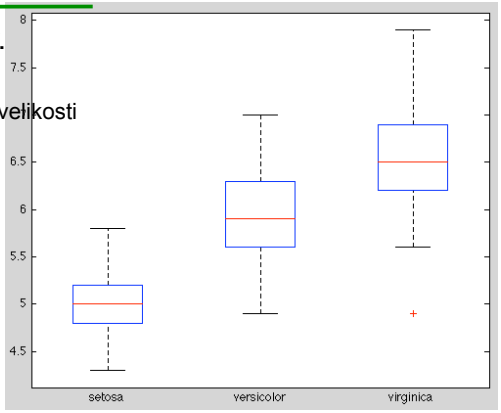
kategorijska spr.

nominalne spr.
(nimamo definirane urejenosti)

zvezne spr.

ordinalne spr.
podatki so urejeni po velikosti

```
>> load fisheriris
>> boxplot(meas(:,1), species)
```



3 Naključne spremenljivke in porazdelitve

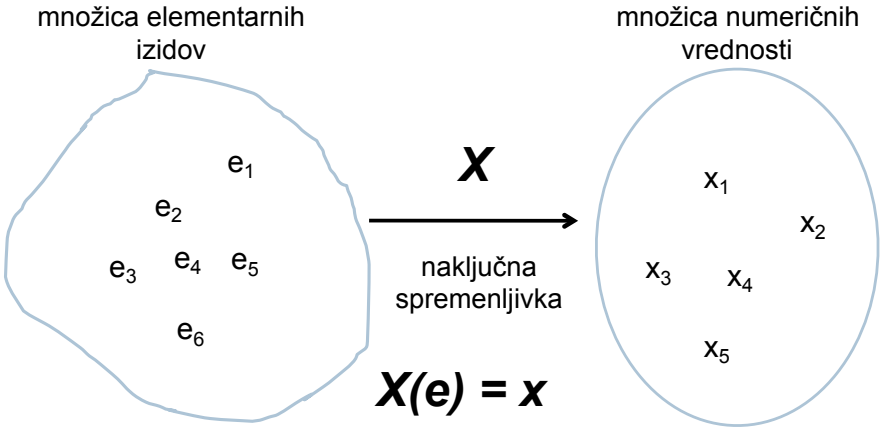
V tem poglavju pregledamo nekaj osnovnih pojmov verjetnostnega računa, definiramo naključne spremenljivke in porazdelitve naključnih spremenljivk. Tu se ne poglabljamo v detaljne izpeljave in dokazovanje posameznih lastnosti naključnih spremenljivk, ampak podajamo le najnujnejše, osnovne, lastnosti, ki jih potrebujemo pri statistični analizi.

Poglavje obravnava:

- definicija naključne spremenljivke in porazdelitve naključnih spremenljivk
- različne porazdelitve naključnih spremenljivk,
- primeri porazdelitev in njihovo uporabo,
- iskanje ujemanja porazdelitev podatkov z znanimi porazdelitvami,
- ocenjevanje parametrov porazdelitev po metodi maksimalnega verjetja in
- neparametrične porazdelitve.

Naključna spremenljivka

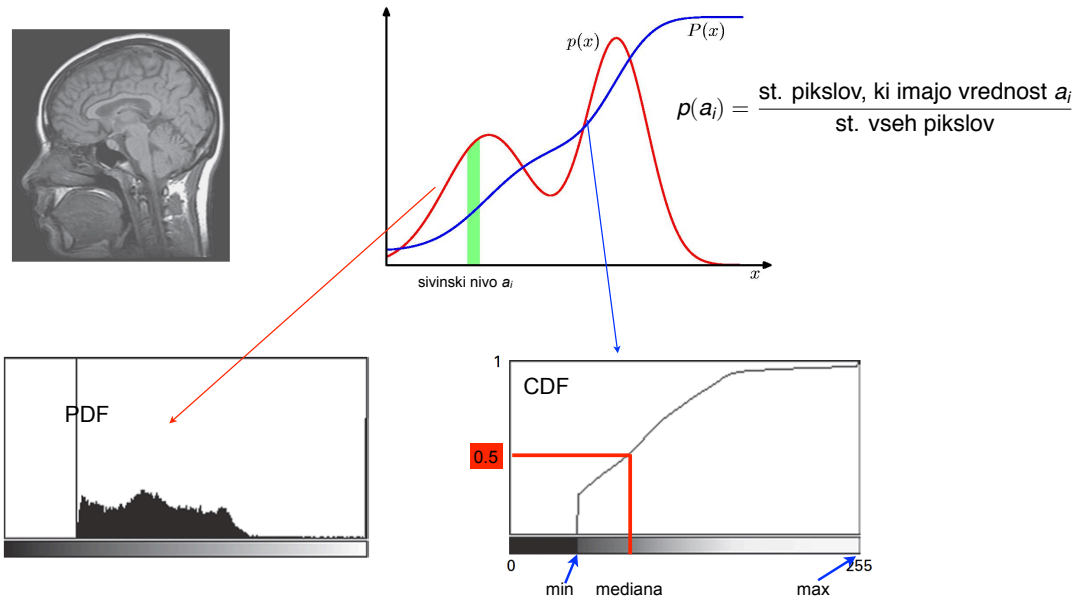
- Elementarnim izidom nekega poskusa določimo numerično vrednost.



- Če je zaloga vrednosti končna ali števno neskončna, govorimo o ***diskretni naključni spremenljivki***.
- Če je zaloga vrednosti npr. množica realnih števil, ***zvezna naključna spremenljivka***.

Primer porazdelitve

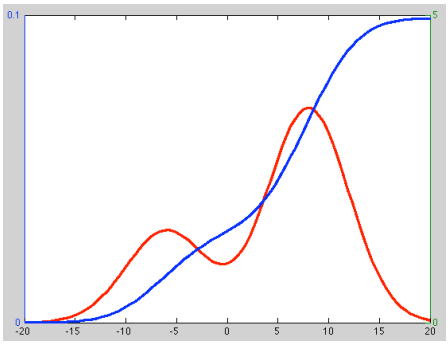
- Porazdelitev sivinskih nivojev na sliki

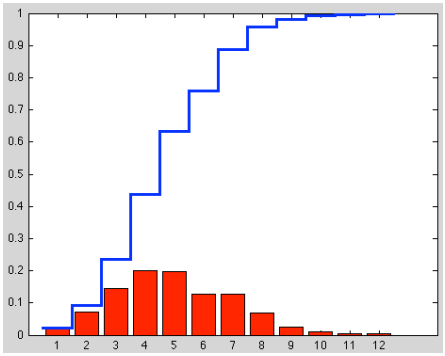


histogram sivinskih nivojev

kumulativna funkcija porazdelitve siv.n.

Porazdelitvene funkcije naključnih spremenljivk





$$p(x \in (a,b)) = \int_a^b p(x)dx$$

$$p(x \in (k,l)) = \sum_{i=k}^{l} f(x_i)$$

$$P(z) = \int_{-\infty}^{z} p(x)dx$$

kumulativna funkcija porazdelitve verjetnosti

$$P(k) = \sum_{i=1}^{k} f(x_i)$$

$$p(x) \geq 0 \quad \int_{-\infty}^{\infty} p(x)dx = 1$$

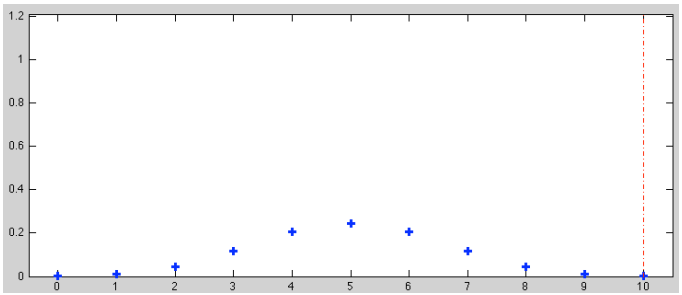
$$f(x_i) \geq 0 \quad \sum_{i=1}^{n} f(x_i) = 1$$

Različne porazdelitve

- Porazdelitve diskretnih naključnih spremenljivk:
  - Binomska porazdelitev:
    - verjetnost k uspehov v n poskusih, če je verjetnost enega poskusa p

$$f(k;n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

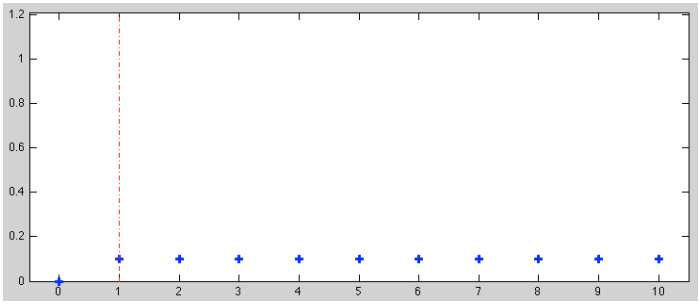
Primer porazdelitve:
n = 10
p = 0.5



Različne porazdelitve

- Porazdelitve diskretnih naključnih spremenljivk:
  - Diskretna enakomerna porazdelitev (uniformna):
    - vsi dogodki so enako verjetni

Primer porazdelitve:
n = 10

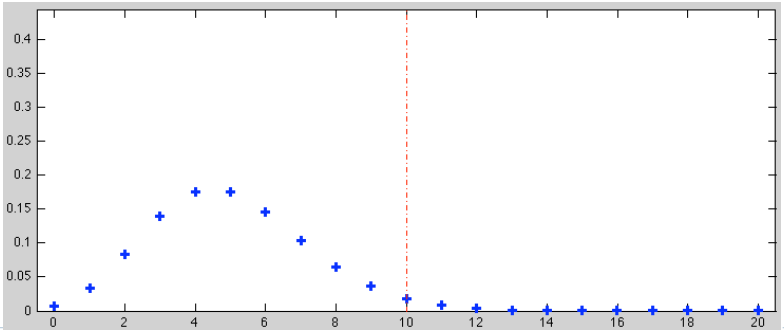


Različne porazdelitve

- Porazdelitve diskretnih naključnih spremenljivk:
  - Poissonova porazdelitev:
    - če je pričakovano število dogodkov v nekem času enako λ, potem je verjetnost, da se bo zgodilo k dogodkov enaka:

$$f(k;\lambda) = \frac{\lambda^k}{k!}e^{-\lambda}$$

Primer porazdelitve:
lam = 5

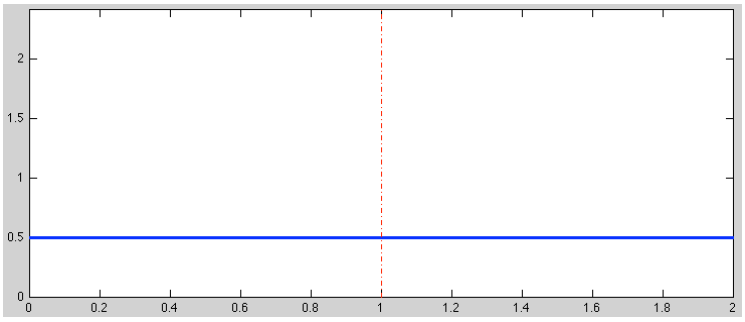


Različne porazdelitve

- Porazdelitve zveznih naključnih spremenljivk:
  - Zvezna enakomerna porazdelitev (uniformna):

$$p(x; a, b) = \frac{1}{b-a} \quad \text{za } a \leq x \leq b$$

Primer porazdelitve:
a = 0
b = 2

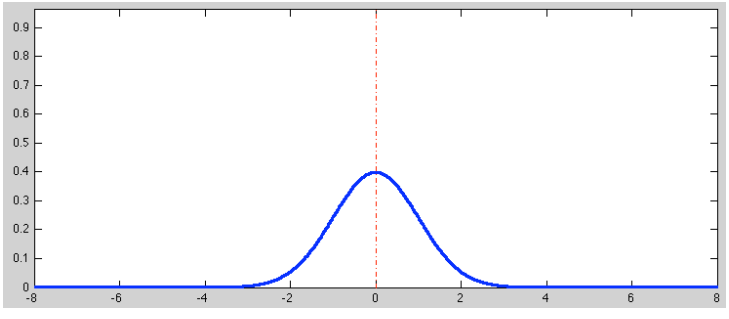


Različne porazdelitve

- Porazdelitve zveznih naključnih spremenljivk:
  - (Standardna) normalna porazdelitev:

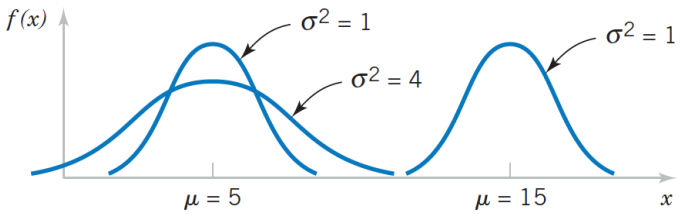
$$p(x; a, b) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Primer porazdelitve:
mu = 0
sigma = 1

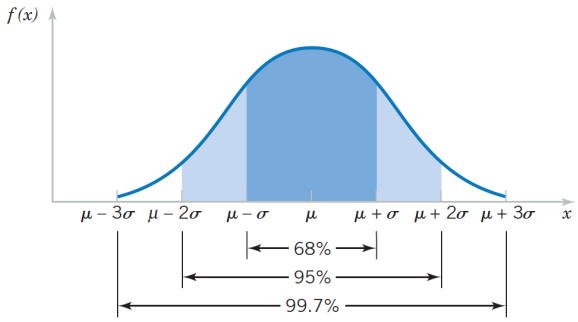


Lastnosti normalne porazdelitve

- Simetričnost.
- Maksimalna vrednost v povprečju.

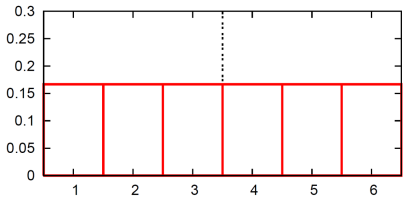


- Večina vrednosti leži znotraj 3 standardnih odklonov od povprečja.



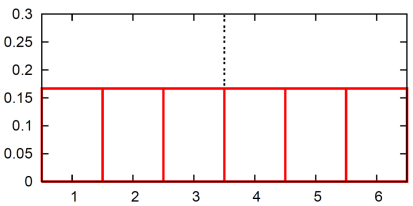
Centralni limitni izrek

Naj bo $X_1, X_2, \ldots$ zaporedje naključnih spr., ki ustrezajo metom kocke. Pri tem naj velja $p_{X_i}(x) = 1/6$ za vsak met X_i, $i \in \mathbb{N}$, pri čemer je $x \in \{1, 2, 3, 4, 5, 6\}$.



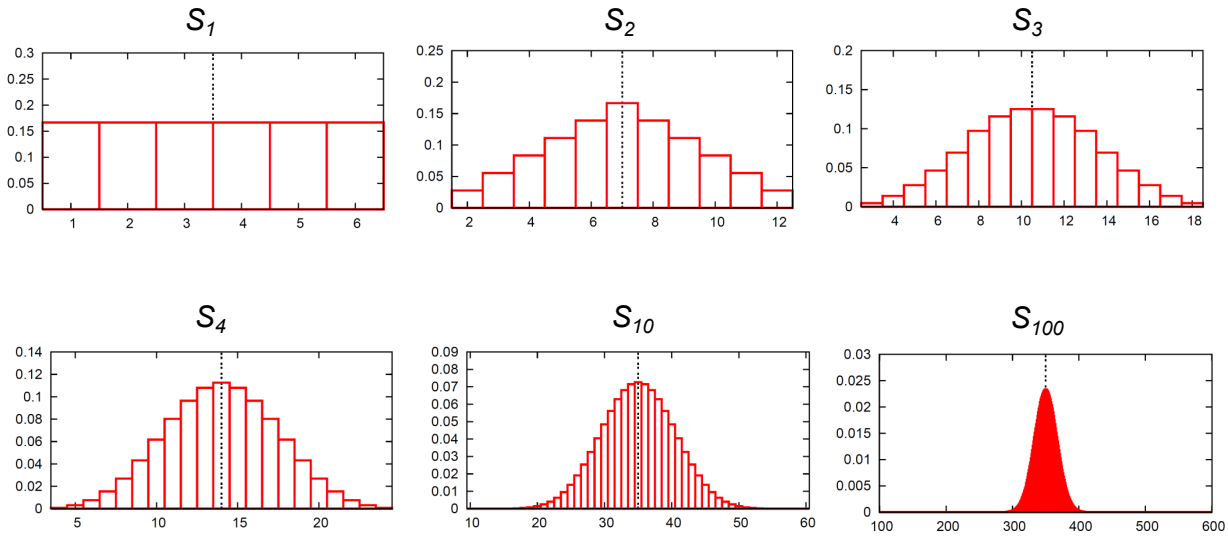
Naj bo $S_n = \sum_{i=1}^{n} X_i$ vsota prvih n metov kocke. Porazdelitev naključne spr. S_n lahko zapišemo kot

$$P_{S_n}(x) = \frac{\text{število načinov, da dobimo vsoto } x \text{ z } n \text{ kockami}}{6^n}$$



porazdelitev S_1

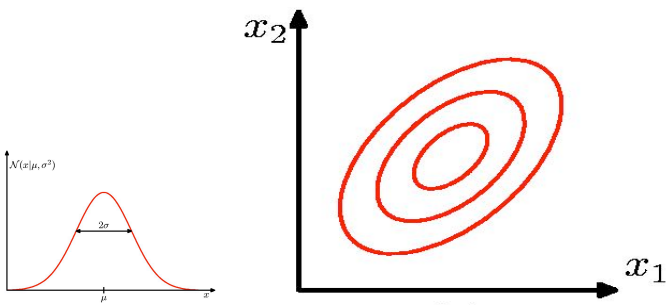
Centralni limitni izrek

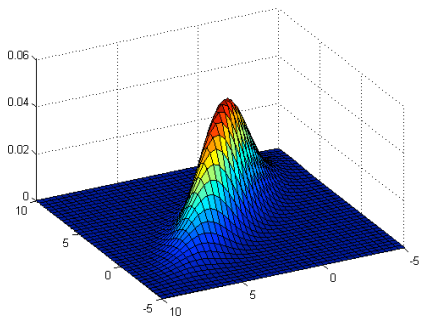


Različne porazdelitve

- Porazdelitev več zveznih naključnih spremenljivk:
  - Normalna porazdelitev (več dimenzionalna):

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$





$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

Različne porazdelitve

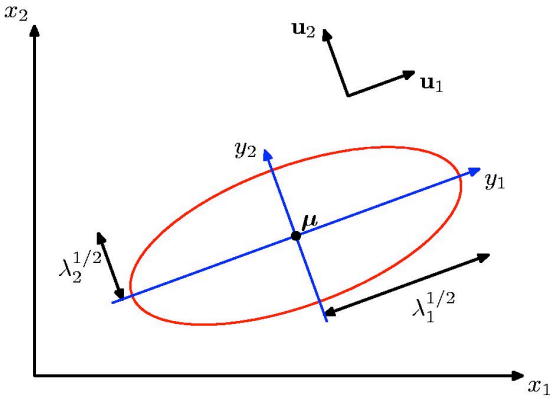
- Geometrija večdimenzionalne normalne porazdelitve:

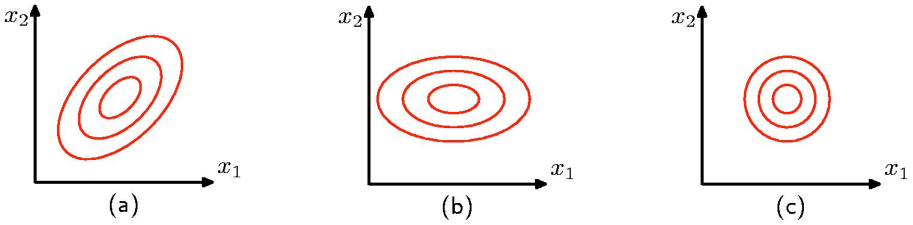
$$\Delta^2 = (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

$$\boldsymbol{\Sigma}^{-1} = \sum_{i=1}^{D}\frac{1}{\lambda_i}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}$$

$$\Delta^2 = \sum_{i=1}^{D}\frac{y_i^2}{\lambda_i}$$

$$y_i = \mathbf{u}_i^{\mathrm{T}}(\mathbf{x}-\boldsymbol{\mu})$$





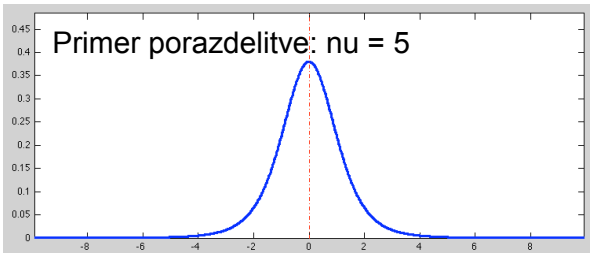
Različne porazdelitve

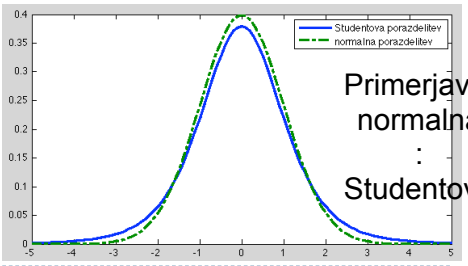
- Porazdelitve naključnih spr., ki jih uporabljamo v statistiki:
  - Studentova t-porazdelitev:
    - Če je x naključno izbran vzorec z močjo n iz populacije, ki je porazdeljena po normalni porazdelitvi s povprečjem μ, potem je spr.

$$t = \frac{\overline{x}-\mu}{s/\sqrt{n}} \quad \overline{x} \text{ in } s \text{ ocenjeno povprečje in standardni odklon iz vzorca } x$$

porazdeljena po naslednji porazdelitvi

$$p(x;\nu) = \frac{\left(\frac{\nu}{x^2+\nu}\right)^{\frac{\nu+1}{2}}}{\sqrt{\nu}B\left(\frac{\nu}{2},\frac{1}{2}\right)} \quad \text{kjer je } \nu = n-1 \text{ št. prostostnih stopenj}$$





Primerjava: normalna : Studentova

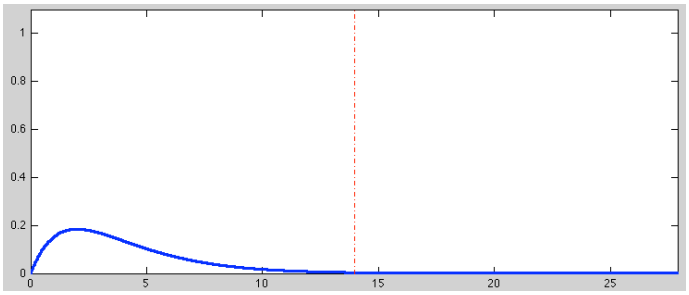
Različne porazdelitve

- Porazdelitve naključnih spr., ki jih uporabljamo v statistiki:
  - Hi-kvadrat porazdelitev:

    Če iz n vzorcev, ki so porazdeljeni normalno z varianco σ^2, ocenjujemo varianco s^2, potem je spremenljivka $x = \frac{(n-1)s^2}{\sigma^2}$ porazdeljena s $p(x, n-1)$, kjer je

$$p(x; n) = \frac{e^{-x/2} 2^{-n/2} x^{\frac{n}{2}-1}}{\Gamma\left(\frac{n}{2}\right)}$$

Primer porazdelitve:
n = 4



Različne porazdelitve

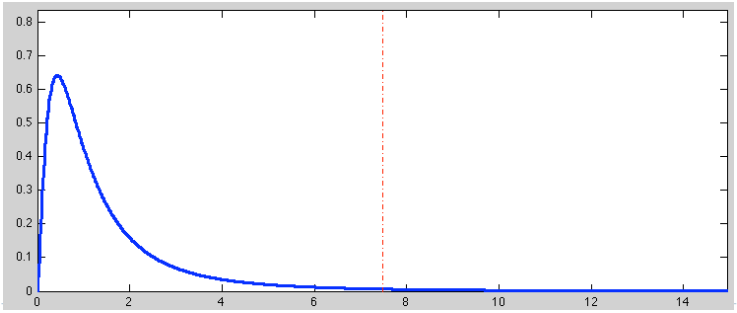
- Porazdelitve naključnih spr., ki jih uporabljamo v statistiki:
  - F-porazdelitev:

    Če imamo dve χ^2 porazdelitvi s prostostnima stopnjama n_1 in n_2, potem je spremenljivka

$$x = \frac{\chi_1^2(x; n_1)/n_1}{\chi_2^2(x; n_2)/n_2}$$

    porazdeljena po F-porazdelitvi.

Primer porazdelitve:
n_1 = 5, n_2=5



Matematično upanje

$$\mathbb{E}[f] = \sum_x p(x)f(x)$$

matematično upanje diskretne sl. spr.

$$\mathbb{E}[f] = \int p(x)f(x)dx$$

matematično upanje zvezne sl. spr.

- Primer izračuna matematičnega upanja:
  - Kolikšno je matematično upanje pri metu kocke?

$$p(x_i) = \frac{1}{6} \quad i = 1, \ldots, 6$$

$$\mathbb{E}[x] = \sum_{i=1}^{6} p(x_i)x_i = \sum_{i=1}^{6} \frac{1}{6} \cdot i = \frac{1}{6}(1+2+3+4+5+6) = \frac{21}{6} = 3.5$$

  - Matematično upanje normalno porazdeljene spr. x?

$$p(x) = \mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad \mathbb{E}[x] = \int_{-\infty}^{\infty} p(x)x dx = \mu$$

Varianca in kovarianca

$$var[f] = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$$

$$\begin{aligned} cov[x, y] &= \mathbb{E}_{x,y}\left[(x - \mathbb{E}[x])(y - \mathbb{E}[y])\right] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y] \end{aligned}$$

$$\begin{aligned} cov[\mathbf{x}, \mathbf{y}] &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{y}^T - \mathbb{E}[\mathbf{y}^T])\right] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^T] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^T] \end{aligned}$$

Primeri izračunov variance

- Varianca pri metu kocke $p(x_i) = \frac{1}{6} \quad i = 1, \ldots, 6$

$$var[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sum_{i=1}^{6} p(x_i) x_i^2 - \left(\sum_{i=1}^{6} p(x_i) x_i \right)^2$$

$$= \frac{1}{6}(1 + 4 + 9 + 16 + 25 + 36) - 3.5^2 = \frac{91}{6} - 12.25 = 2.9167$$

- Normalna porazdelitev

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) x^2 dx = \mu^2 + \sigma^2$$

$$var[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2$$

Povprečje in varianca večdimenzionalne Gaussove porazdelitve

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}$$

- 1. moment – povprečje

$$\begin{aligned} \mathbb{E}[\mathbf{x}] &= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \int \exp\left[-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \right] \mathbf{x} d\mathbf{x} \\ &= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \int \exp\left[-\frac{1}{2}\mathbf{z}^T \Sigma^{-1} \mathbf{z} \right] (\mathbf{z} + \mu) d\mathbf{z} = \mu \end{aligned}$$

- 2. moment – varianca

$$\begin{aligned} \mathbb{E}[\mathbf{x}\mathbf{x}^{\mathbf{T}}] &= \mu\mu^T + \boldsymbol{\Sigma} \\ cov[\mathbf{x}] &= \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^T] = \boldsymbol{\Sigma} \end{aligned}$$

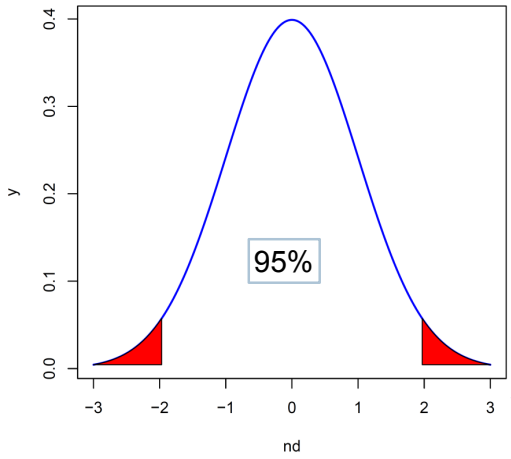
Delo s porazdelitvami

- V katerem intervalu leži 95% vseh podatkov, če so standardno normalno porazdeljeni?

- To pomeni, da bo 5% vrednosti ležalo izven intervala.
- Ker je norm. porazdelitev simetrična, pomeni, da bo ležalo 2.5% vrednosti na levi in 2.5% na desni izven našega intervala.

- Kvantili normalne porazdelitve:

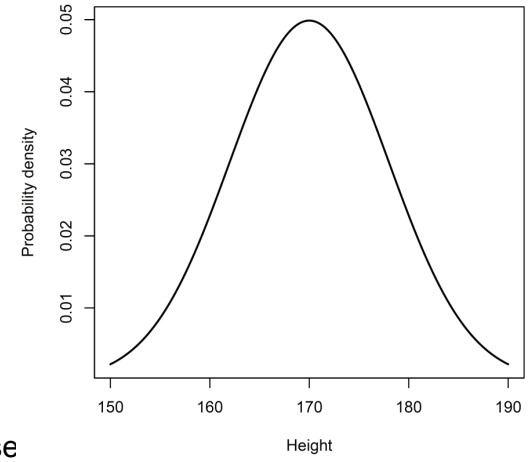
```
>>  norminv([0.025 0.975],0,1)
ans =
   -1.9600    1.9600
```



Statistična analiza z normalno porazdelitvijo

- Primer:
  - Merimo višino 100 osebam. Ugotovili smo, da je povprečna višina 170 cm in standardni odklon 8 cm.

  - Vprašanja:
    - Kakšna je verjetnost, da bo naključno izbrana oseba nižja od določene višine?
    - Kakšna je verjetnost, da bo oseba višja od določene višine?
    - Kakšna je verjetnost, da bo višina ose na intervalu med eno in drugo višino?



Statistična analiza z normalno porazdelitvijo

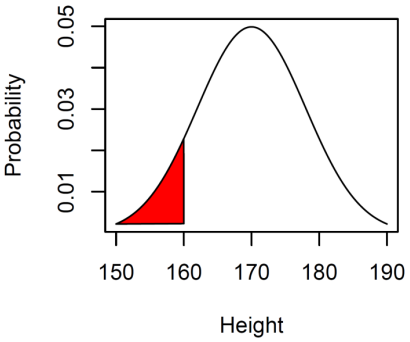
- 1. vprašanje: Kakšna je verjetnost, da bo naključno izbrana oseba nižja od 160 cm?

- Najprej **normaliziramo podatke**: $z = \frac{y - \overline{y}}{s}$ ***z ~ N(0,1)***

  Primer: $z = \frac{160 - 170}{8} = -1.25$

- Izračunamo verjetnost:

```
>> z = (160 - 170)/8
z =
   -1.2500
>> normcdf(z, 0, 1)
ans =
    0.1056
```



- Brez normalizacije v Matlabu

```
>> normcdf(160, 170, 8)
ans =
    0.1056
```

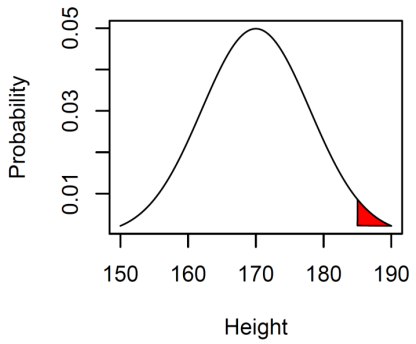
Statistična analiza z normalno porazdelitvijo

- 2. vprašanje: Kakšna je verjetnost, da bo naključno izbrana oseba večja od 185 cm?

- Najprej **normaliziramo podatke**: $z = \frac{y - \overline{y}}{s}$ ***z ~ N(0,1)***

- Izračunamo verjetnost:

```
>> z = (185 - 170)/8
z =
    1.8750
>> normcdf(z, 0, 1)
ans =
    0.9696

>> 1 - normcdf(z, 0, 1)
ans =
    0.0304
```



Brez normalizacije v Matlabu

```
>> 1 - normcdf(185, 170, 8)
```

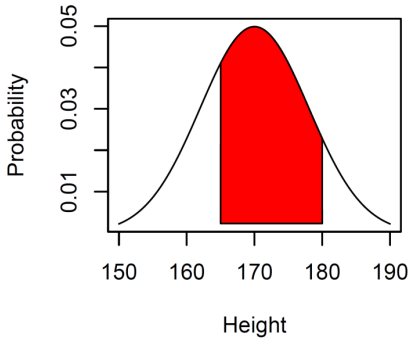
Statistična analiza z normalno porazdelitvijo

- 3. vprašanje: Kakšna je verjetnost, da bo naključno izbrana oseba večja od 165 cm in manjša od 180 cm?

- **Normaliziramo podatke**: $z = \frac{y - \overline{y}}{s}$ $z \sim N(0,1)$

- Izračunamo verjetnost:

```
>> z1 = (165 - 170)/8
z1 =
   -0.6250
>> z2 = (180 - 170)/8
z2 =
    1.2500

>> normcdf(z2, 0, 1) - normcdf(z1, 0, 1)
ans =
    0.6284
```



Brez normalizacije v Matlabu

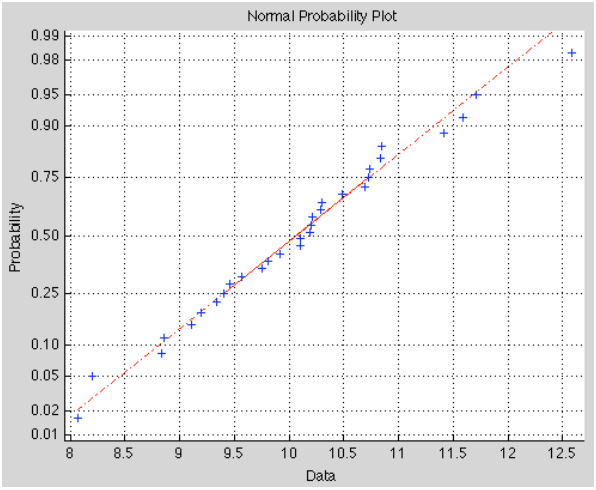
```
>> normcdf(180, 170, 8) - normcdf(165, 170, 8)
```

Delo s porazdelitvami: grafi ujemanja podatkov s porazdelitvami

- Grafi ujemanja podatkov z normalno porazdelitvijo nam povedo ali so podatki normalno porazdeljeni.
- veliko statističnih ocen in testov predpostavlja normalnost vzorcev. S takšnimi grafi lahko takoj preverimo, če je to res.

```
>> x = normrnd(10,1,30,1);
>> normplot(x)
```

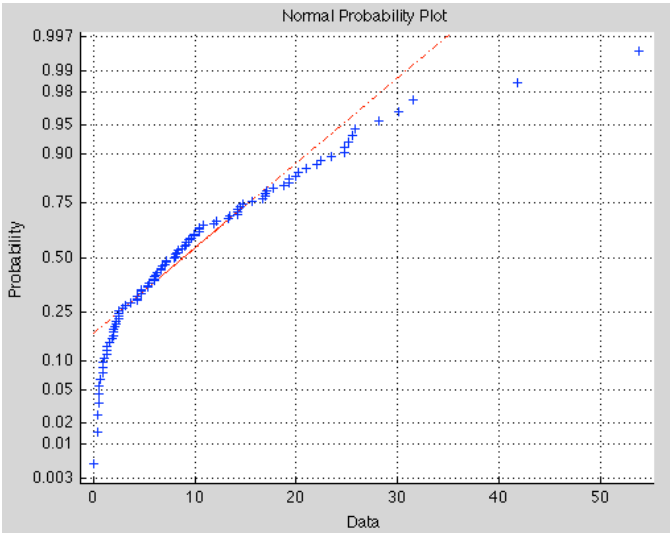
- Generirali smo 30 vrednosti, ki so normalno porazdeljene.
- S funkcijo normplot preverjamo normalnost vzorca.
- Če vrednosti ležijo na premici, to kaže na normalnost vzorca.



Delo s porazdelitvami: grafi ujemanja podatkov s porazdelitvami

- Preverimo ali so vzorci porazdeljeni po normalni porazdelitvi.

```
>> x = exprnd(10,100,1);
>> normplot(x)
```

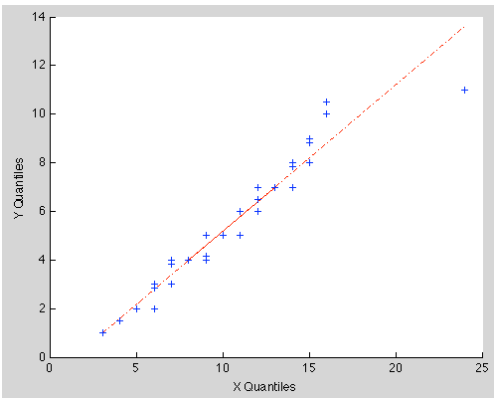


- Če je vzorec normalno porazdeljen, se bo ujemal s premico, sicer **imamo odstopanja v oblik črke S ali banane**.

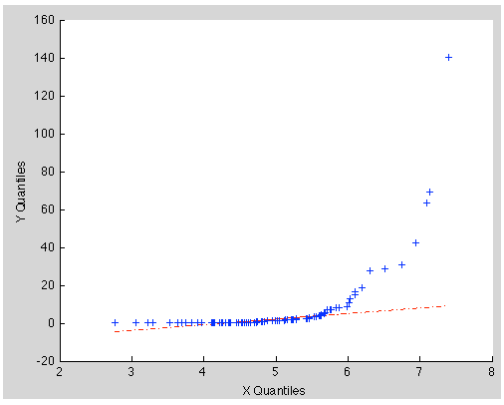
Delo s porazdelitvami: grafi ujemanja podatkov s porazdelitvami

- Graf kvantil-kvantil**:**

```
>> x = poissrnd(10,75,1);
>> y = poissrnd(5,100,1);
>> qqplot(x,y);
```




```
>> x = normrnd(5,1,100,1);
>> y = wblrnd(2,0.5,100,1);
>> qqplot(x,y);
```



- Z grafom kvantil-kvantil preverimo ujemanje dveh vzorcev, ali izhajajo iz enake porazdelitve.
- Povezana rdeča črta povezuje točki prvega in tretjega kvartila ocenjeni iz obeh vzorcev (koordinate x iz prvega vzorca, koordinati y pa iz drugega).
- Če se podatki porazdeljujejo okoli premice, potem prihajajo iz enake družine porazdelitev (tudi če imajo različne parametre), če pa ne, pa niso iz enake družine porazdelitev.

Ocenjevanje parametrov porazdelitve

- Kako oceniti parametre porazdelitev iz podatkov?
  - Imamo torej zaporedje podatkov $\mathbf{x} = (x_1, x_2, \ldots, x_N)$.
  - Oceniti moramo porazdelitev p s parametri $\theta_1, \theta_2, \ldots, \theta_M$.
  - Če predpostavljamo enako porazdeljenost in neodvisnost lahko zapišemo:

$$p(\mathbf{x}; \theta_1, \theta_2, \ldots, \theta_M) = \prod_{i=1}^{N} p(x_i; \theta_1, \theta_2, \ldots, \theta_M)$$

  - **Princip maksimalnega verjetja (ang. maximum likelihood)**: Iščemo take parametre porazdelitve, da bo verjetje maksimalno:

$$\theta_1^{ML}, \theta_2^{ML}, \ldots, \theta_M^{ML} = \arg\max p(\mathbf{x}; \theta_1, \theta_2, \ldots, \theta_M)$$

  - Ker je $\ln(.)$ monotono naraščajoča funkcija, lahko izvedemo logaritem verjetja in **iščemo maksimalni logaritem verjetja (ang. maximum log-likelihood)**

$$\theta_1^{ML}, \theta_2^{ML}, \ldots, \theta_M^{ML} = \arg\max \ln p(\mathbf{x}; \theta_1, \theta_2, \ldots, \theta_M)$$

Ocenjevanje parametrov porazdelitve

- Imamo podatke $\mathbf{x} = (x_1, \ldots, x_N)^T$, ki so med seboj neodvisni in enako porazdeljeni:

$$p(\mathbf{x}|\mu, \sigma) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2) = \prod_{n=1}^{N} \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}}$$

- Logaritem verjetja glede na te podatke je:

$$\ln p(\mathbf{x}|\mu, \sigma) = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2$$

- Iščemo takšna μ_{ML} in σ_{ML}, da bo $\ln p(\mathbf{x}|\mu_{ML}, \sigma_{ML})$ maksimalen.

Ocenjevanje parametrov porazdelitve

Odvajamo logaritem verjetja po parametru μ:

$$\frac{\partial}{\partial \mu} \ln p(\mathbf{x}|\mu, \sigma) = -\frac{1}{2\sigma^2} \sum_{n=1}^{N} 2(x_n - \mu)(-1)$$

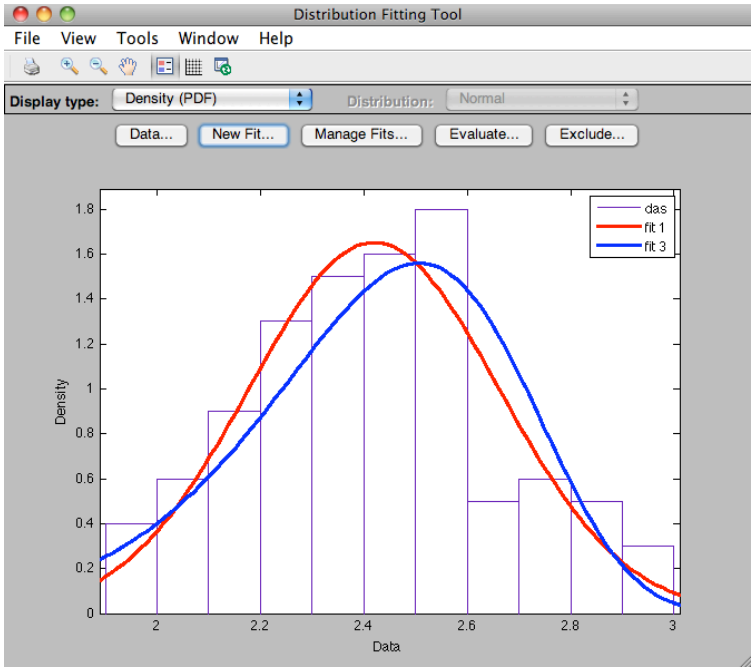
in iščemo rešitev, ko je odvod enak nič (maksimalne vrednosti):

$$\frac{\partial}{\partial \mu} \ln p(\mathbf{x}|\mu, \sigma) = 0 \qquad \mu_{ML} = \frac{1}{N} \sum_{n=1}^{N} x_n$$

Podobno odvajamo tudi po σ in dobimo:

$$\sigma^2_{ML} = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu_{ML})^2.$$

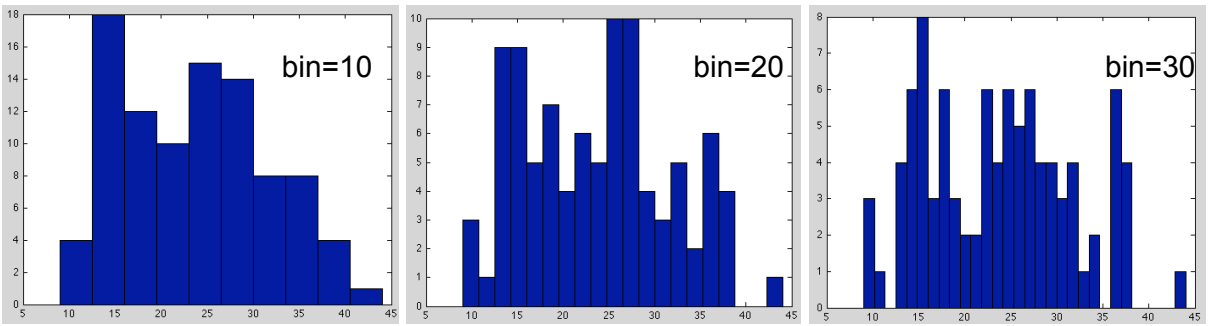
Numerično ocenjevanje parametrov porazdelitve



Neparametrične porazdelitve

- Porazdelitev lahko ponazorimo grafično tudi s histogramom:

```
>> cars = load('carsmall','MPG','Origin');
>> MPG = cars.MPG;
>> hist(MPG,10)
>> hist(MPG,20)
>> hist(MPG,30)
```

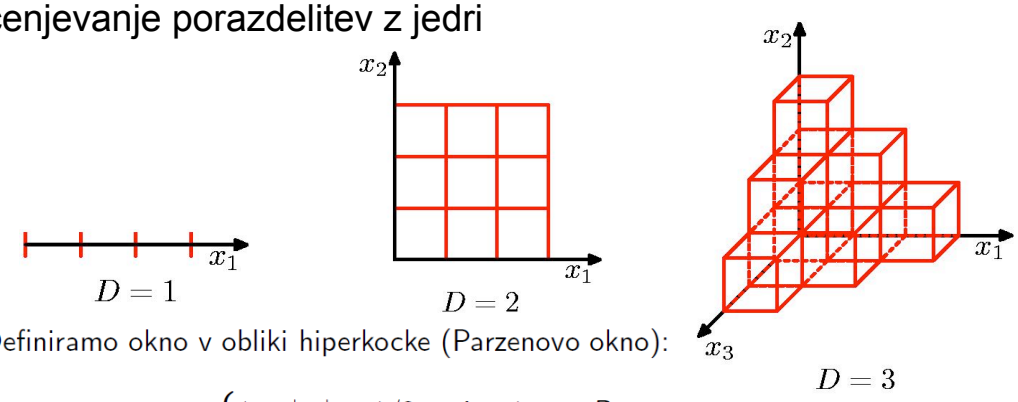


Pri histogramu določimo število podintervalov, ki so enako veliki, in štejemo, koliko primerkov je v danem podintervalu. To izrišemo v obliki stolpcev.

Verjetnost v posameznem podintervalu ocenimo kot $p_i = \frac{n_i}{N\Delta_i}$, kjer je n_i število primerkov v podintervalu Δ_i in N število vseh primerkov.

Neparametrične porazdelitve

- Ocenjevanje porazdelitev z jedri



- Definiramo okno v obliki hiperkocke (Parzenovo okno):

$$k(\mathbf{u}) = \begin{cases} 1, & |u_i| \leq 1/2, \quad i = 1, \ldots, D, \\ 0, & \text{sicer} \end{cases}$$

- In jo premikamo po prostoru vrednosti slučajne spr. $\mathbf{x}$ ter štejemo, koliko vrednosti leži znotraj take hiperkocke:

$$K = \sum_{n=1}^{N} k\left(\frac{\mathbf{x} - \mathbf{x}_n}{h}\right)$$

Neparametrične porazdelitve

- Ocenjevanje porazdelitev z jedri
  - Ocena verjetnosti je podobno kot pri histogramu normalizirana frekvenca števila točk, ki ležijo znotraj hiperkocke

$$p(\mathbf{x}) = \frac{K}{NV}.$$

  - V našem primeru:

$$p(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{h^D}k\left(\frac{\mathbf{x}-\mathbf{x}_n}{h}\right)$$

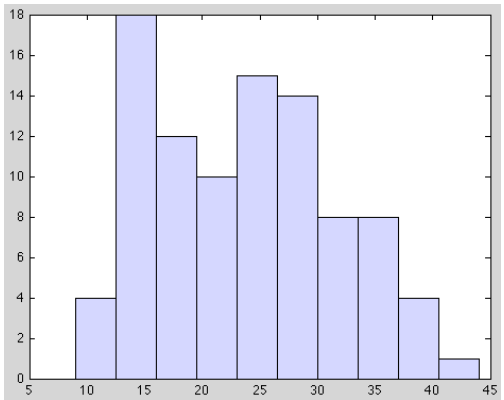
  - Lahko vzamemo, kakšno drugo okno, tako ki je bolj zvezno (bolj gladko, manj preskokov), npr. Gaussovo jedro:

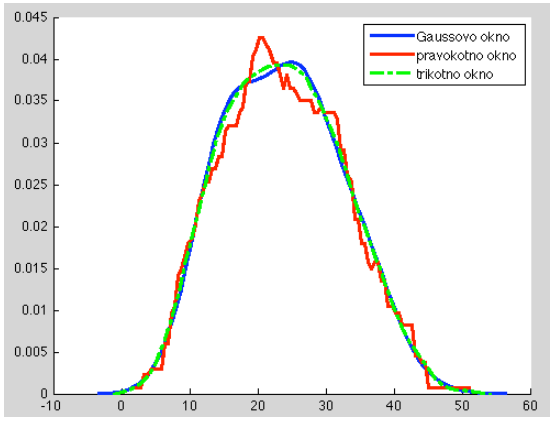
$$p(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{(2\pi h^2)^{1/2}}\exp\left(-\frac{||\mathbf{x}-\mathbf{x}_n||^2}{2h^2}\right)$$

Neparametrične porazdelitve

- Primer ocenjevanja porazdelitve z različnimi jedri:

```
hold on;
[f,x] = ksdensity(MPG, 'kernel', 'normal');
plot(x,f, 'b', 'LineWidth', 2);
[f,x] = ksdensity(MPG, 'kernel', 'box');
plot(x,f, 'r', 'LineWidth', 2);
[f,x] = ksdensity(MPG, 'kernel', 'triangle');
plot(x,f, 'g-.', 'LineWidth', 2);
hold off;
```

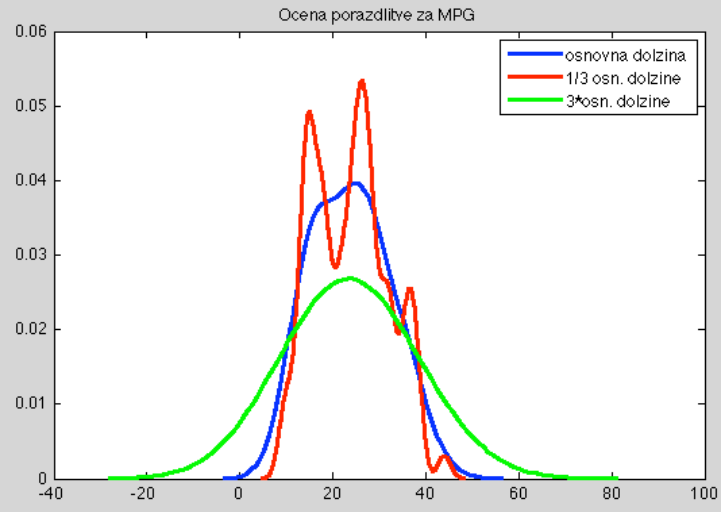




Neparametrične porazdelitve

- Vpliv širine okna na oceno porazdelitve:

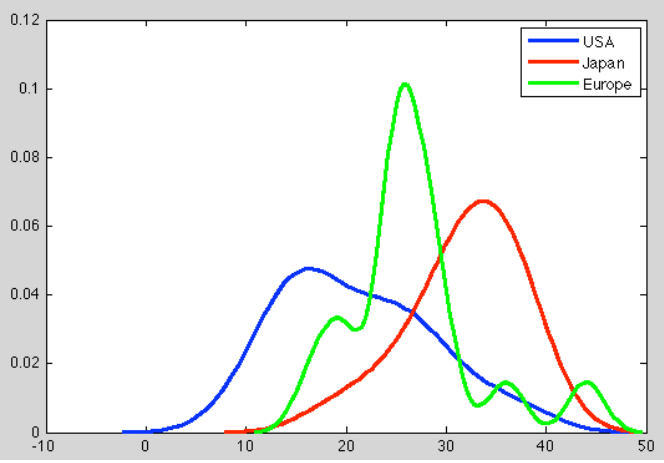
```
>> [f,x,u] = ksdensity(MPG);
>> u
u = 4.1143
>> plot(x,f)
>> title('Ocena porazdlitve za MPG')
>> hold on
>>[f,x] = ksdensity(MPG,'width',u/3);
>> plot(x,f,'r');
>>[f,x] = ksdensity(MPG,'width',u*3);
>> plot(x,f,'g');
>> legend('osnovna dolzina','1/3 osn.
dolzine','3*osn. dolzine')
>> hold off
```



Neparametrične porazdelitve

- Primer:
  - ocena porazdelitev MPG glede na državo proizvajalko avtomobilov

```
Origin = cellstr(cars.Origin);
I = strcmp('USA',Origin);
J = strcmp('Japan',Origin);
K = ~(I|J);
MPG_USA = MPG(I);
MPG_Japan = MPG(J);
MPG_Europe = MPG(K);

[fI,xI] = ksdensity(MPG_USA);
plot(xI,fI,'b')
hold on

[fJ,xJ] = ksdensity(MPG_Japan);
plot(xJ,fJ,'r')

[fK,xK] = ksdensity(MPG_Europe);
plot(xK,fK,'g')

legend('USA','Japan','Europe')
hold off
```



4 Statistična analiza vzorcev in testiranje hipotez

Poglavje obravnava:

- **Osnovne motivacije za testiranje hipotez**
- **Definicija pojmov pri testiranju hipotez:**
  - ničta in alternativna hipoteza,
  - eno- in dvo-stranski testi,
  - p-vrednost,
  - statistična značilnost,
  - intervali zaupanja v oceno parametrov statistike.
- **Statistike enega vzorca:**
  - t-test in z-test (ocena povprečij),
  - χ^2 -test (ocena variance),
  - testiranje porazdelitve,
  - test predznaka (ocena mediane) - neparametrična statistika.
- **Statistike dveh vzorcev:**
  - t-test2: primerjava povprečij dveh vzorcev z normalno porazdeljeno populacijo,
  - F-test: primerjava varianc dveh vzorcev z normalno porazdeljeno populacijo,
  - kontingenčne tabele: test enakosti porazdelitev,
  - Wilcoxonov test: primerjava povprečij dveh vzorcev z ne-normalno porazdeljenimi napakami,
  - test Kolmogorova in Smirnova: porazdelitev dveh naključnih spremenljivk.

Primer

- Primer, da ugotavljamo povprečno ceno 1kg belega kruha v Sloveniji.
- Denimo, da je nekdo ugotovil, da je povprečna cena enaka 1.15 EUR.
- Kako lahko ugotovimo, ali je ta ocena pravilna?
  - Lahko gremo po vseh trgovinah in ugotavljamo ceno kruha.
    - To je najboljša varianta. Vendar je časovno neizvedljiva.
  - Lažja varianta: Naključno izberemo nekaj trgovin in na podlagi njihovih cen kruha izračunamo povprečno vrednost.
    - Ocenjena vrednost iz našega vzorca znaša 1.18 EUR.
    - Ali je razlika 0.03 EUR posledica naše izbire vzorca?
    - Ali je razlika takšna, da lahko trdimo, da je povprečna ocena kruha višja od ocenjene povprečne vrednosti 1.15 EUR?
- Pri testiranju hipotez se ukvarjamo s takšnimi vprašanji.

Terminologija pri testiranju hipotez

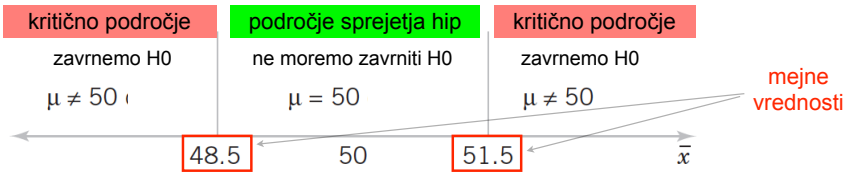
- Hipoteza:
  - Hipoteza je lastnost neke populacije, ki jo želimo testirati.
  - **Ničta hipoteza:**
    - Populacija ne izpolnjuje lastnosti.
    - "Nič se ne dogaja".
    - V našem primeru:
      - H0: povprečna vrednost kruha je 1.15 EUR.
  - **Alternativna hipoteza:**
    - Nasprotje ničti hipotezi, ki ga je mogoče statistično ovrednotiti.
    - "Nekaj se dogaja."
    - V našem primeru več možnih alternativnih hipotez:
      - Povprečje ni enako 1.15 EUR. (dvo-stranski test)
      - Povprečje je večje od 1.15 EUR. (eno-stranski test: desni)
      - Povprečje je manjše od 1.15 EUR. (eno-stranski test: levi)
  - Izvedba testiranja hipoteze:
    - Iz populacije izberemo **naključno določen vzorec**.
    - Izračunamo ustrezno **testno statistiko**, ki ustreza hipotezi.
      Testne statistike so različne glede na tip testa, ki ga uporabljamo.
    - Predpostavka pri vseh testih je: **poznamo porazdelitev testne statistike ob ničti hipotezi**.

Terminologija pri testiranju hipotez

- Statistična značilnost
  - Primer:
    - Denimo, da preverjamo, ali je povprečna vrednost neke populacije enaka 50.
    - Hipotezi: $H_0: \mu = 50$

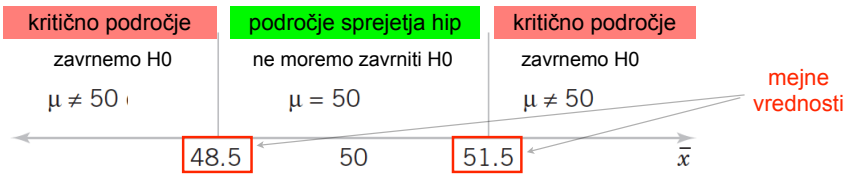
      $H_1: \mu \neq 50$
    - Naključno smo izbrali vzorec iz 10-ih primerkov in ocenili povprečno vrednost $\bar{x}$.
    - Verjetno lahko zaključimo:
      - da v primeru, če je ocenjena vrednost povprečja nekje okoli 50, potem je hipoteza H0 bolj verjetna.
      - in nasprotno: če je ocenjena vrednost povprečja občutno različna od 50, potem je hipoteza H1 bolj verjetna.
    - Denimo, da predpostavimo, če
      - $48.5 \leq \bar{x} \leq 51.5$, potem velja hipoteza H0.
      - $\bar{x} < 48.5$ ali $\bar{x} > 51.5$, potem velja hipoteza H1.

Torej:

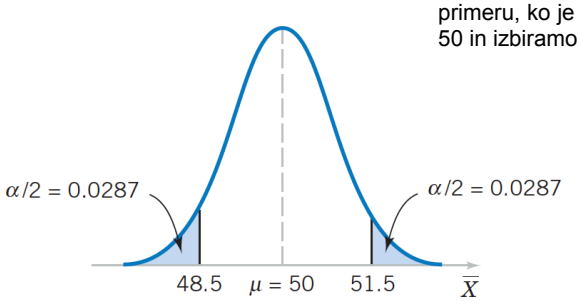


Terminologija pri testiranju hipotez

- Statistična značilnost:



Porazdelitev ocen povprečne vrednosti v primeru, ko je dejanska povprečna vrednost 50 in izbiramo vzorce z 10 primerki.

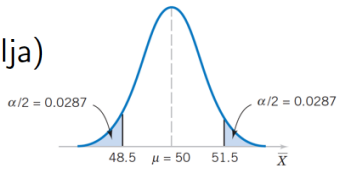


V primeru, da se odločimo, da delamo napako zavrnitve H0 (tip I) 5%, dobimo v tem primeru mejne vrednosti 48.5 in 51.5.

Terminologija pri testiranju hipotez

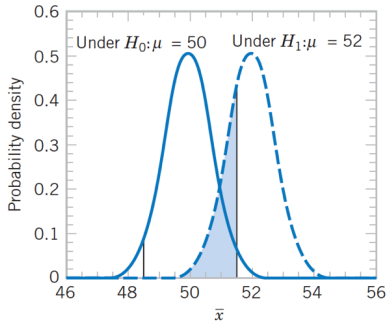
- Napake pri testiranju hipotez:
  - sprejetje hipoteze z možnostjo napake zavrnitve pravilne hipoteze 5%. (**napaka tipa I**, α = 0.05)

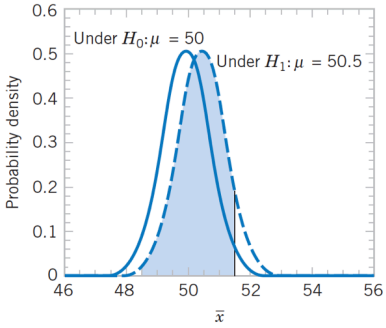
$$\alpha = P(\text{napaka tipa I}) = P(\text{zavrnemo } H_0, \text{ ko } H_0 \text{ velja})$$



  - **napaka tipa II**

$$\beta = P(\text{napaka tipa II}) = P(\text{sprejmemo } H_0, \text{ ko } H_0 \text{ ne velja})$$



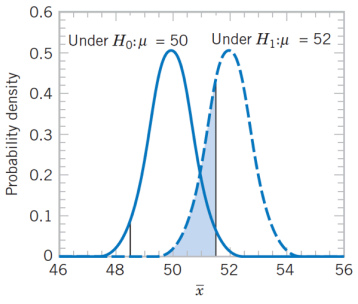


Kolikšno napako delamo, če smo ocenili zgornjo mejo na 51.5, vendar je dejanska ocena povprečja v prvem primeru 52, v drugem pa 50.5, če je moč vzorca n=10.

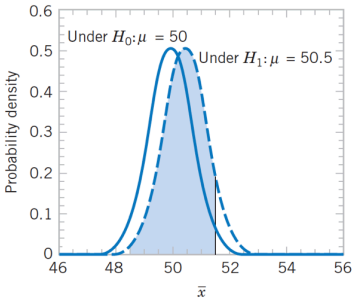
Terminologija pri testiranju hipotez

- Napake pri testiranju hipotez: moč vzorca in dejansko povprečje

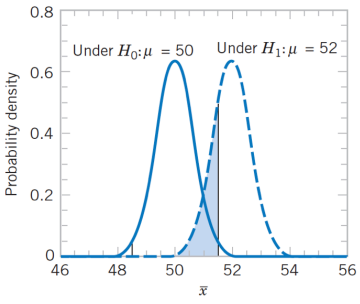
n = 10



n = 10



n = 16



$$\text{moč testa} = 1 - \beta$$

Acceptance Region	Sample Size	α	β at μ = 52	β at μ = 50.5
$48.5 < \bar{x} < 51.5$	10	0.0576	0.2643	0.8923
$48 < \bar{x} < 52$	10	0.0114	0.5000	0.9705
$48.5 < \bar{x} < 51.5$	16	0.0164	0.2119	0.9445
$48 < \bar{x} < 52$	16	0.0014	0.5000	0.9918

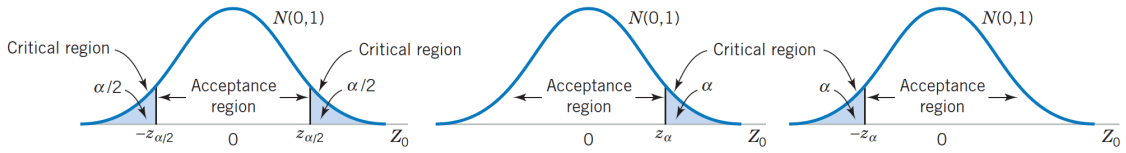
Terminologija pri testiranju hipotez

- p-vrednost:
  - P-vrednost je verjetnost, da testna statistika ob predpostavki, da ničta hipoteza velja, zavzame vrednost, ki je večja ali enaka izračunani testni statistiki iz vzorca.
  - P-vrednost pove več kot statistična značilnost. Pri statistični značilnosti povemo, da smo npr. v primeru napake zavrnitve 5% sprejeli hipotezo. Tu pa še izvemo, kolikšna je verjetnost, da bi ob izračunani testni statistiki naredili napako zavrnitve hipoteze.
  - Z drugimi besedami: napaka zavrnitve 5% je prag, do katerega sprejmemo hipotezo, p-vrednost pa nam pove še kolikšna je verjetnost, da smo naredili napako zavrnitve hipoteze ob izračunani statistiki iz danega vzorca.
    Če je p-vrednost nad 5%, potem ne moremo zavrniti ničte hipoteze.

Sample Size n	P-value When $\bar{x} = 50.5$	Power (at $\alpha = 0.05$) When True $\mu = 50.5$
10	0.4295	0.1241
25	0.2113	0.2396
50	0.0767	0.4239
100	0.0124	0.7054
400	5.73×10^{-7}	0.9988
1000	2.57×10^{-15}	1.0000

Terminologija pri testiranju hipotez

- Enostranski in dvostranski testi



$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$

$H_0: \mu = \mu_0$ $H_1: \mu > \mu_0$

$H_0: \mu = \mu_0$ $H_1: \mu < \mu_0$

  - dvostranski test:
    - možnost napake zavrnitve pravilne hipoteze razdelimo na oba konca porazdelitve testne statistike.
  - enostranski test:
    - možnost napake zavrnitve pravilne hipoteze obravnavamo na eni ali drugi strani porazdelitve testne statistike:
      - levi in desni enostranski test

Testiranje hipotez

- Splošen postopek pri testiranju hipotez:
  - 1. Pri danem problemu določimo lastnost (količino), ki jo želimo testirati.
  - 2. Določimo ničto hipotezo H0.
  - 3. Določimo ustrezno alternativno hipotezo H1.
  - 4. Odločimo se, na podlagi katerega kriterija bomo preverjali ničto hipotezo (statistična značilnost, p-vrednost).
  - 5. Izberemo pravilno testno statistiko za preverjanje hipoteze.
  - 6. Določimo področje, kjer na podlagi te statistike lahko zavrnemo hipotezo.
  - 7. Izračunamo testno statistiko iz vzorca(ev), ki ga (jih) imamo.
  - 8. Odločimo, ali ničto hipotezo lahko zavrnemo

Intervali zaupanja v oceno parametrov

- Interval zaupanja v oceno parametra je interval, v katerem lahko z veliko verjetnostjo trdimo, da je dejanska vrednost parametra populacije.

- Spodnjo in zgornjo mejo intervala izračunamo iz ocenjene vrednosti parametra iz danega vzorca populacije ob predpostavki znane porazdelitve ocenjenih vrednosti iz vzorca.
  - Običajno predpostavimo, da se ocene parametrov iz vzorcev porazdeljujejo normalno (po centralnem limitnem izreku).

- Široki intervali zaupanja pomenijo slabe (manj zanesljive) ocene parametrov.
  Ožji intervali zaupanja pomenijo dobro (bolj zanesljive) ocene parametrov.

- Intervali zaupanja v oceno so odvisni predvsem od moči vzorca.

Intervali zaupanja v oceno parametrov

- Napaka pri oceni povprečja:
  - večja kot je varianca, večja je lahko napaka pri oceni povprečja,
  - več podatkov imamo (večja moč vzorca), manjšo napako pri oceni delamo.
- **Torej:** $$\text{napaka pri oceni } \bar{y} \propto \frac{s^2}{n}$$
- **Standardna napaka pri oceni povprečja:**

$$SE_{\bar{y}} = \sqrt{\frac{s^2}{n}}$$

Intervali zaupanja v oceno parametrov

- V primeru treh vrtov:

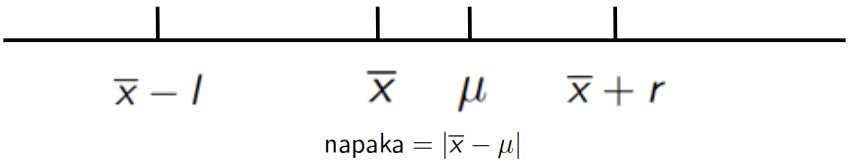
```
>> garden = importdata('datasets/gardens.txt')
>> data = garden.data
data =
     3     5     3
     4     5     3
     4     6     2
...
>> mean(data(:,1))
ans = 3
>> sqrt(var(data(:,1))/10)
ans = 0.3651

>> mean(data(:,2))
ans = 5
>> sqrt(var(data(:,2))/10)
ans = 0.3651

>> mean(data(:,3))
ans = 5
>> sqrt(var(data(:,3))/10)
ans =1.1926
```

  - Povprečje koncentracije ozona v vrtu A je 3.0 ± 0.365 (1 s.e. n = 10).
  - Povprečje koncentracije ozona v vrtu B je 5.0 ± 0.365 (1 s.e. n = 10).
  - Povprečje koncentracije ozona v vrtu C je 5.0 ± 1.193 (1 s.e. n = 10).

Intervali zaupanja v oceno parametrov

- Interval zaupanja v oceno povprečja ob neznani varianci pri normalno porazdeljeni populaciji:
  - Predpostavimo normalno porazdeljeno populacijo s povprečjem μ (ki pa ga ne poznamo).
  - Iz dane populacije smo naključno izbrali vzorec z močjo n in ocenili povprečje $\overline{x}$ in varianco s^2.
  - Radi bi ocenili zgornjo in spodnjo mejo intervala zaupanja v pravo vrednost povprečja iz naše ocene povprečja, torej kakšna l in r potrebujem, da

$$\overline{x} - l \leq \mu \leq \overline{x} + r$$



napaka $= |\overline{x} - \mu|$

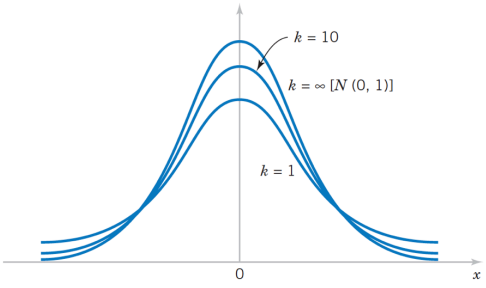
Intervali zaupanja v oceno parametrov

- Interval zaupanja v oceno povprečja ob neznani varianci pri normalno porazdeljeni populaciji:
  - Definirajmo statistiko

$$T = \frac{\overline{x} - \mu}{s/\sqrt{n}}$$

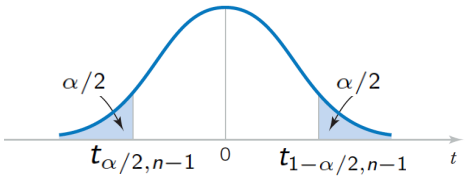
.

  - Ta spremenljivka je porazdeljena po Studentovi t-porazdelitvi s prostostno stopnjo $n - 1$.



Intervali zaupanja v oceno parametrov

- Interval zaupanja v oceno povprečja ob neznani varianci pri normalno porazdeljeni populaciji:
  - Če hočemo imeti interval zaupanja v oceno, ki bo pravilen v 95% primerov (α = 0.05, ali pravilnost 1-α). Potem:



$$P(t_{\alpha/2,n-1} \leq T \leq t_{1-\alpha/2,n-1}) = 1-\alpha$$
$$P(\overline{x} - t_{\alpha/2,n-1} \cdot s/\sqrt{n} \leq \mu \leq \overline{x} + t_{\alpha/2,n-1} \cdot s/\sqrt{n}) = 1-\alpha$$

Interval zaupanja v oceno povprečja pri $100(1-\alpha)$ odstotni pravilnosti s številom prostostnih stopenj $n-1$ je

$$\overline{x} - t_{\alpha/2,n-1} s/\sqrt{n} \leq \mu \leq \overline{x} + t_{\alpha/2,n-1} s/\sqrt{n}$$

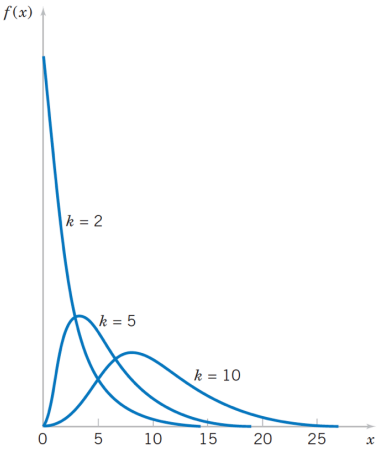
Intervali zaupanja v oceno parametrov

- Interval zaupanja v oceno variance pri normalno porazdeljeni populaciji:

  - Za oceno intervala zaupanja v oceno variance definiramo statistiko

$$X^2 = \frac{(n-1)s^2}{\sigma^2},$$

    kjer je s^2 ocenjena vrednost variance iz vzorca z močjo n in σ^2 dejanska varianca populacije, ki je porazdeljena normalno.
  - Statistika X^2 je porazdeljena po χ^2 porazdelitvi.

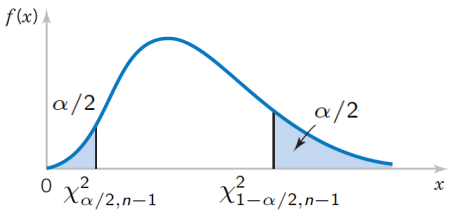


Intervali zaupanja v oceno parametrov

- Interval zaupanja v oceno variance pri normalno porazdeljeni populaciji:

$$P(\chi^2_{\alpha/2,n-1} \leq X^2 \leq \chi^2_{1-\alpha/2,n-1}) = 1-\alpha$$

$$P(\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}) = 1-\alpha$$



Interval zaupanja v oceno variance pri $100(1-\alpha)$ odstotne pravilnosti s številom prostostnih stopenj $n-1$ je

$$\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}$$

Statistike enega vzorca

- z-test:
  - Testiramo, ali je **ocenjeno povprečje vzorca** iz populacije, ki je **normalno porazdeljena**, **enako predpostavljenemu povprečju populacije** ob **znani varianci**.
  - Hipotezi:

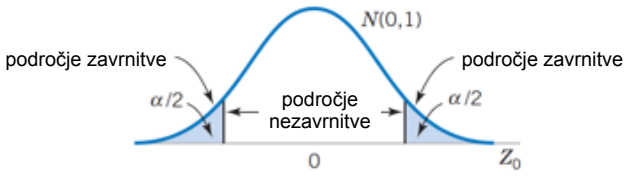
$$H_0: \mu = \mu_0$$

$$H_1: \mu \neq \mu_0$$

  - Testna statistika:

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

  - Porazdelitev testne statistike : normalna



Statistike enega vzorca

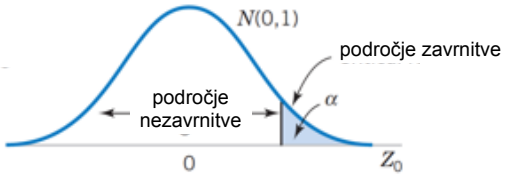
- z-test:
  - Testiramo, ali je **ocenjeno povprečje vzorca** iz populacije, ki je **normalno porazdeljena**, **enako predpostavljenmu povprečju populacije** ob **znani varianci**.
  - Hipotezi:

$$H_0: \mu = \mu_0$$
$$H_1: \mu > \mu_0$$

  - Testna statistika:

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

  - Porazdelitev testne statistike : normalna



Statistike enega vzorca

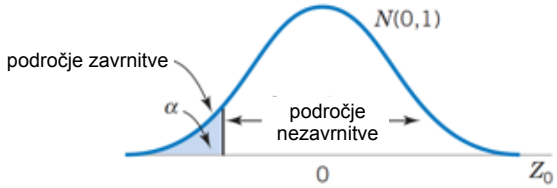
- z-test:
  - Testiramo, ali je **ocenjeno povprečje vzorca** iz populacije, ki je **normalno porazdeljena**, **enako predpostavljenemu povprečju populacije** ob **znani varianci**.
  - Hipotezi:

$$H_0: \mu = \mu_0$$
$$H_1: \mu < \mu_0$$

  - Testna statistika:

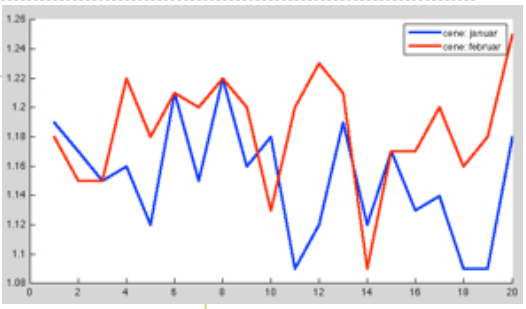
$$z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

  - Porazdelitev testne statistike : normalna



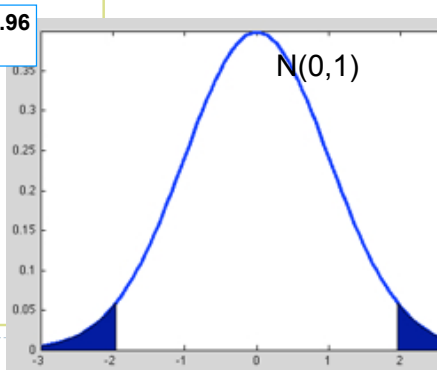
Statistike enega vzorca

- z-test: primer



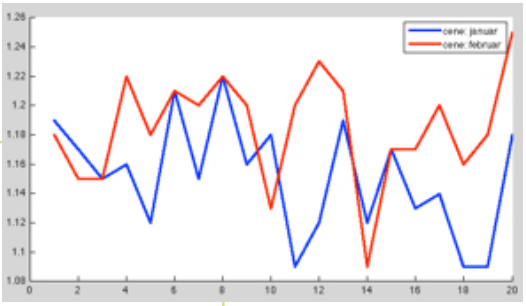

```
load('datasets/cene.mat')
hold on; plot(cene_jan, 'b'); plot(cene_feb, 'r'); hold off

m0 = 1.16;      % predpostavljeno povprecje
s0 = 0.04;      % znani std. odklon
n = 20;         % stevilo primerkov

m1 = mean(cene_jan)    % izracunano povprecje: januar
m2 = mean(cene_feb)    % izracunano povprecje: februar


z1 = (m1 - m0) / (s0/sqrt(n))  % testna stat. za m1
z2 = (m2 - m0) / (s0/sqrt(n))  % testna stat. za m2

meja = norminv([0.025 0.975],0,1) %leva in desna meja porazdelitve za alf=0.05

m1 = 1.1515
m2 = 1.1850

z1 = -0.9503
z2 = 2.7951

meja =
   -1.9600    1.9600

>> p1 = normcdf(z1)              % p-vrednost za obojestranski test
p1 = 0.1710
p1 = 2*min(p1,1-p1)
p1 = 0.3419
% ali v tem primeru
p1 = 2*(1-normcdf(abs(z1)))      % p-vrednost za obojestranski test
p2 = 2*(1-normcdf(abs(z2)))      % p-vrednost za obojestranski test
p1 =
    0.3419
p2 =
    0.0052
```

Primer 1: H0 ne moremo zavrniti: -1.96 < z1 < 1.96
Primer 2: H0 lahko zavrnemo, saj: z2 > 1.96



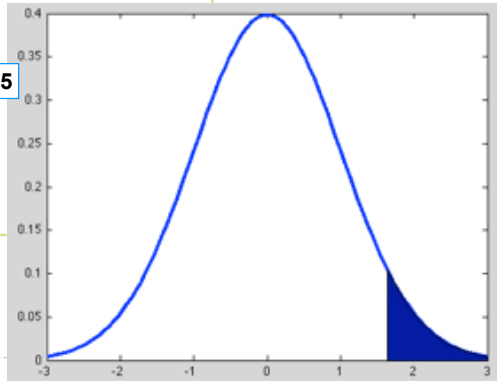
Statistike enega vzorca

- z-test:
  - primer : H0: povprečje 2 = 1.16
    H1: povprečje2 > 1.16




```
m0 = 1.16;      % predpostavljeno povprecje
s0 = 0.04;      % znani std. odklon
n = 20;         % stevilo primerkov

m2 = mean(cene_feb)    % izracunano povprecje: februar

z2 = (m2 - m0) / (s0/sqrt(n))  % testna stat. za m2

meja = norminv(0.95,0,1)    % meja za hipotezo H1 povp > m0 pri alf = 0.05

m2 = 1.1850

z2 = 2.7951

meja = 1.6449


>> p = (1-normcdf(z2))       % p-vrednost za H0 m>m0
p = 0.0026
```

Primer 2: H0 lahko zavrnemo, saj: z2 > 1.65

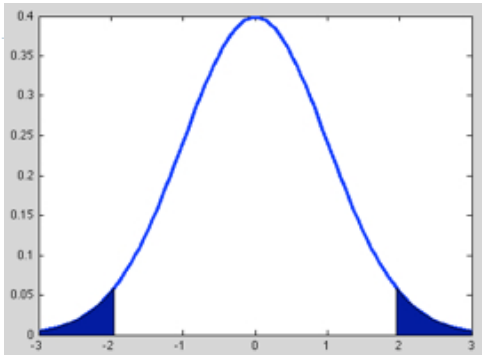


Statistike enega vzorca

- z-test:
  - primer : H0: povprečje1 = 1.16
    H1: povprečje1 ~= 1.16

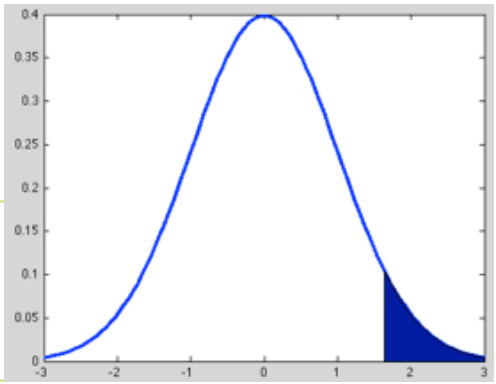
    H0: povprečje2 = 1.16
    H1: povprečje2 ~= 1.16

```
m0 = 1.16;      % predpostavljeno povprecje
s0 = 0.04;      % znani std. odklon

>> [h,pvalue,ci] = ztest(cene_jan, m0, s0)
h = 0
pvalue = 0.3419
ci = 1.1340
     1.1690

>> [h,pvalue,ci] = ztest(cene_feb, m0, s0)
h = 1
pvalue = 0.0052
ci = 1.1675
     1.2025
```



  - primer : H0: povprečje2 = 1.16
    H1: povprečje2 > 1.16

```
m0 = 1.16;      % predpostavljeno povprecje
s0 = 0.04;      % znani std. odklon

>> [h,pvalue,ci] = ztest(cene_feb, m0, s0, 0.05, 'right')
h = 1
pvalue = 0.0026
ci = 1.1703
       Inf
```



Statistike enega vzorca

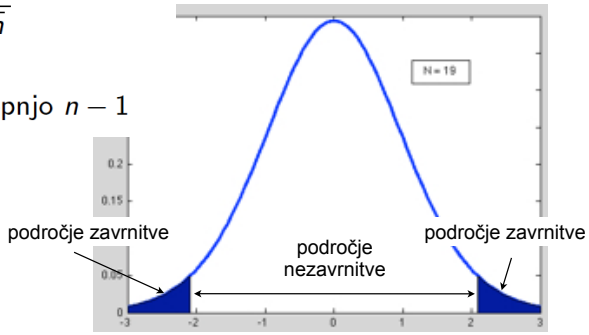
- t-test:
  - Testiramo, ali je **ocenjeno povprečje vzorca** iz populacije, ki je **normalno porazdeljena**, **enako predpostavljenemu povprečju populacije** ob **neznani varianci**.
    - Hipotezi:

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

    - Definirajmo statistiko:

$$t_0 = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

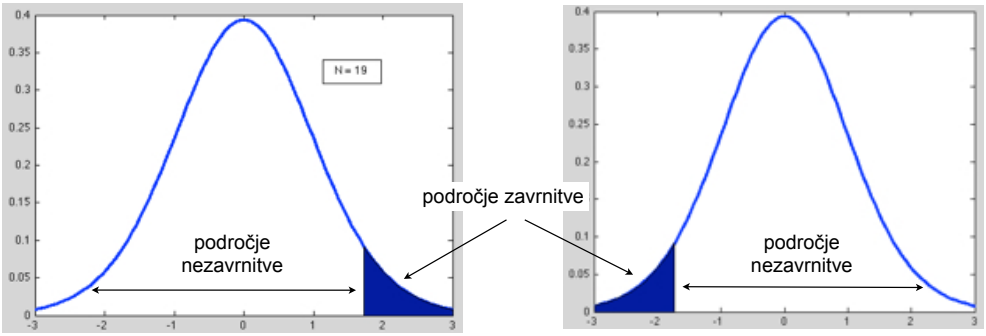
    - Porazdelitev testne statistike:
      Studentova t-porazdelitev s prostostno stopnjo $n - 1$



Statistike enega vzorca

- t-test:
  - Testiramo, ali je **ocenjeno povprečje vzorca** iz populacije, ki je **normalno porazdeljena**, **enako predpostavljenemu povprečju populacije** ob **neznani varianci**.
  - Hipotezi:

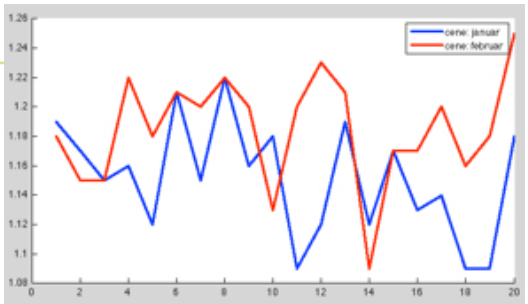
$$H_0 : \mu = \mu_0 \qquad H_0 : \mu = \mu_0$$
$$H_1 : \mu > \mu_0 \qquad H_1 : \mu < \mu_0$$



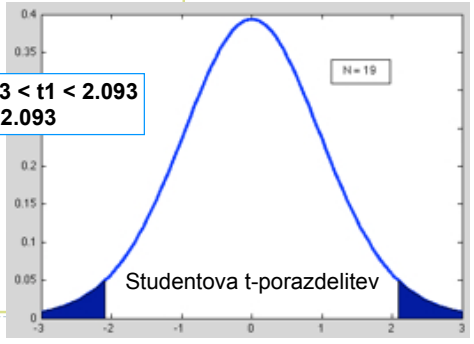
Statistike enega vzorca

- t-test: **neznana varianca**

```
load('datasets/cene.mat')
hold on; plot(cene_jan, 'b'); plot(cene_feb, 'r'); hold off

m0 = 1.16;      % predpostavljeno povprecje
n = 20;         % stevilo primerkov

m1 = mean(cene_jan)    % izracunano povprecje: januar
m2 = mean(cene_feb)    % izracunano povprecje: februar
s1 = std(cene_jan)     % izracunajmo std. odklon: januar
s2 = std(cene_feb)     % izracunajmo std. odklon: februar


t1 = (m1 - m0) / (s1/sqrt(n))  % testna stat. za m1
t2 = (m2 - m0) / (s2/sqrt(n))  % testna stat. za m2

meja = tinv([0.025 0.975],n-1) %leva in desna meja porazdelitve za alf=0.05

m1 =    1.1515
m2 =    1.1850

s1 =    0.0387
s2 =    0.0373

t1 =   -0.9823
t2 =    2.9937

meja =
   -2.0930    2.0930

>> p1 = tcdf(t1,n-1);    % p-vrednost za obojestranski test
>> p1 = 2*min(p1,1-p1)

p2 = tcdf(t2, n-1);      % p-vrednost za obojestranski test
p2 = 2*min(p2,1-p2)

p1 = 0.3383
p2 = 0.0075
```

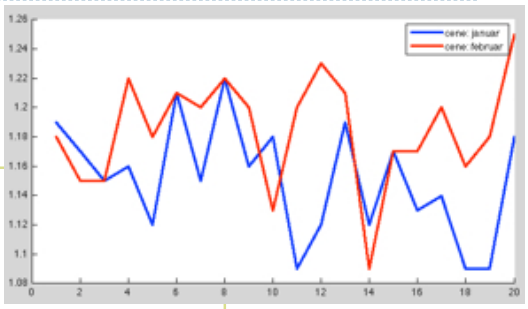


Primer 1: H0 ne moremo zavrniti: -2.093 < t1 < 2.093
Primer 2: H0 lahko zavrnemo, saj: t2 > 2.093



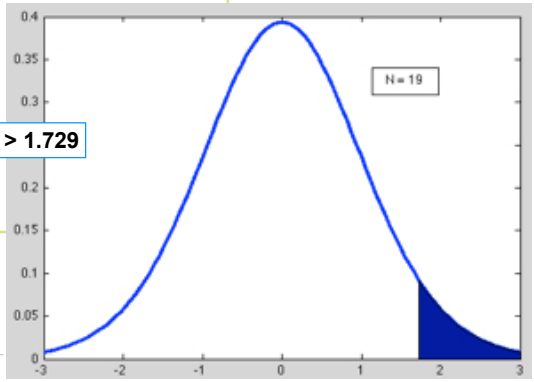
Statistike enega vzorca

- t-test: **neznana varianca**
  - primer : H0: povprečje 2 = 1.16
    H1: povprečje2 > 1.16




```
m0 = 1.16;      % predpostavljeno povprecje
n = 20;         % stevilo primerkov

m2 = mean(cene_feb)    % izracunano povprecje: februar
s2 = std(cene_feb)     % izracunajmo std. odklon: februar

t2 = (m2 - m0) / (s2/sqrt(n))  % testna stat. za m2

meja = tinv(0.95,n-1)   % meja za hipotezo H1 povp > m0 pri alf = 0.05

m2 =    1.1850

s2 =    0.0373

t2 =    2.9937

meja =  1.7291

>> p = 1-tcdf(t2,n-1)       % p-vrednost za H0 m>m0
p = 0.0037
```

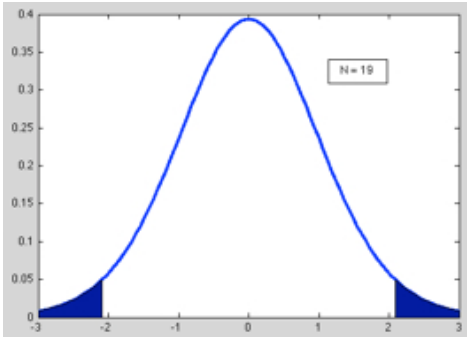
Primer 2: H0 lahko zavrnemo, saj: t2 > 1.729



Statistike enega vzorca

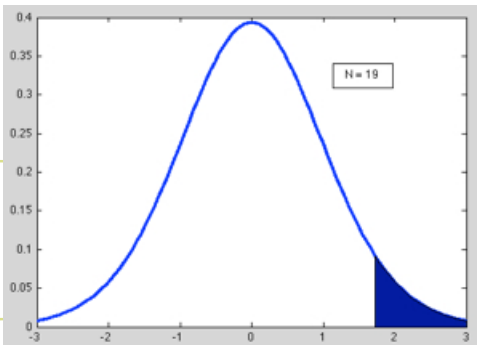
- t-test:
  - primer : H0: povprečje1 = 1.16
    H1: povprečje1 ~= 1.16

    H0: povprečje2 = 1.16
    H1: povprečje2 ~= 1.16




```
m0 = 1.16;      % predpostavljeno povprecje

>> [h,pvalue,ci] = ttest(cene_jan, m0)
h =      0
pvalue =    0.3383
ci = 1.1334
     1.1696

>> [h,pvalue,ci] = ttest(cene_feb, m0)
h =      1
pvalue = 0.0075
ci = 1.1675
     1.2025
```

  - primer : H0: povprečje2 = 1.16
    H1: povprečje2 > 1.16




```
m0 = 1.16;      % predpostavljeno povprecje

>> [h,pvalue,ci] = ttest(cene_feb, m0, 0.05, 'right')
h =      1
pvalue = 0.0037
ci = 1.1703
       Inf
```

Statistike enega vzorca

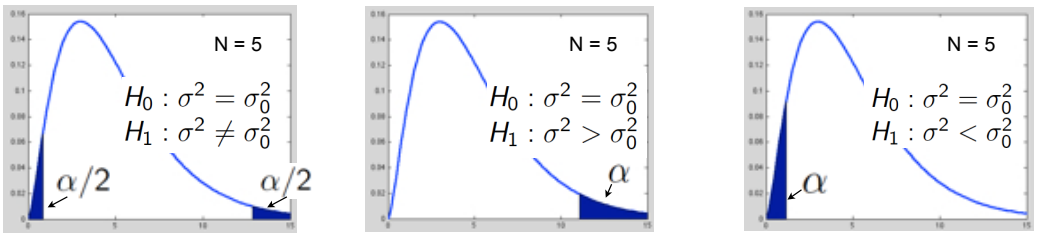
- hi-kvadrat test:
  - Testiramo, ali je **ocenjena varianca vzorca** iz populacije, ki je **normalno porazdeljena**, **enaka predpostavljeni varianci populacije**.
    - Hipotezi:

$$H_0 : \sigma^2 = \sigma_0^2$$
$$H_1 : \sigma^2 \neq \sigma_0^2$$

    - Testna statistika

$$X_0^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

    - Porazdelitev testne statistike χ^2-porazdelitev s prostostno stopnjo $n-1$.



Statistike enega vzorca

- hi-kvadrat test: primer

```
load('datasets/cene.mat')
hold on; plot(cene_jan, 'b'); plot(cene_feb, 'r'); hold off

s0 = 0.04;      % predpostavljen std. odklon
n = 20;         % stevilo primerkov

s1 = std(cene_jan)    % izracunajmo std. odklon: januar
s2 = std(cene_feb)    % izracunajmo std. odklon: februar

hi1 = (n-1)*s1^2 / s0^2  % testna stat. za s1
hi2 = (n-1)*s2^2 / s0^2  % testna stat. za s2

meja = chi2inv([0.025 0.975],n-1) %leva in desna meja porazdelitve za alf=0.05

s1 =    0.0387
s2 =    0.0373

hi1 =   17.7844
hi2 =   16.5625

meja =
    8.9065   32.8523

p1 = chi2cdf(hi1,n-1);          % p-vrednost za obojestranski test
p1 = 2*min(p1,1-p1)

p2 = chi2cdf(hi2, n-1);          % p-vrednost za obojestranski test
p2 = 2*min(p2,1-p2)

p1 = 0.9262
p2 = 0.7610
```

Primer 1: H0 ne moremo zavrniti: 8.9 < hi1 < 32.8
Primer 2: H0 ne moremo zavrniti: 8.9 < hi2 < 32.8

Statistike enega vzorca

- hi-kvadrat test: primer

```
>> [h, p, ci, stat] = vartest(cene_jan, 0.04^2)
h =     0
p =    0.9262
ci =  0.0009    0.0032
stat =
    chisqstat: 17.7844
           df: 19
```

Matlab test.

Statistike enega vzorca

- Hi-kvadrat test: za testiranje porazdelitve populacije (ang. goodness of fit)
  - Hipotezi: H0 : Ocenjena porazdelitev ustreza predpostavljeni porazdelitvi.
    H1 : Ocenjena porazdelitev ne ustreza predpostavljeni porazdelitvi.

  - Testna statistika:
    Iz podatkov iz vzorca z močjo n naredimo histogram s k podintervali in ocenimo frekvence za posamezne podintervale O_i.
    Iz predpostavljene porazdelitve ocenimo vrednosti za histogram (E_i).
    Primerjamo med seboj:

$$X_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

  - Porazdelitev testne statistike χ^2-porazdelitev s prostostno stopnjo $k - p - 1$ (p je število parametrov porazdelitve).

Statistike enega vzorca

- Hi-kvadrat test: za testiranje porazdelitve populacije - primer
  - Pri preverjanju kakovosti nekega izdelka smo ugotovili naslednje:
    - Vzeli smo 60 proizvodov in ugotovili, da

število napak	število proizvodov
0	32
1	15
2	9
3	4

  - Predpostavimo Poissonovo porazdelitev.
    - V tem primeru ocenimo parametre Poiss. porazdelitve. Potrebno je oceniti povprečno vrednost napake:

$$(32 \cdot 0 + 15 \cdot 1 + 9 \cdot 2 + 4 \cdot 3)/60 = 0.75$$

Statistike enega vzorca

- Hi-kvadrat test: za testiranje porazdelitve populacije - primer
  - Izračunajmo verjetnosti in frekvence Poissonove porazdelitve ob povprečni vrednosti 0.75.

$$p_1 = P(X = 0) = \frac{e^{-0.75}(0.75)^0}{0!} = 0.472$$

$$p_2 = P(X = 1) = \frac{e^{-0.75}(0.75)^1}{1!} = 0.354$$

$$p_3 = P(X = 2) = \frac{e^{-0.75}(0.75)^2}{2!} = 0.133$$

$$p_4 = P(X \geq 3) = 1 - (p_1 + p_2 + p_3) = 0.041$$

  - Naredimo tabelo glede na Poisssonovo porazdelitev.
  - ß

$$E_i = n \cdot p_i$$

število napak	verjetnost	pričak. frekvenca
0	0.472	28.32
1	0.354	21.24
2	0.133	7.98
3 ali več	0.041	2.46*

* Če je frekvenca pod 3, potem združimo s prejšnjim podintervalom, zaradi napak pri hi-kvadrat statistiki.

Statistike enega vzorca

- Hi-kvadrat test: za testiranje porazdelitve populacije - primer
  - Združimo obe tabeli in primerjamo vrednosti:

število napak	število proizvodov	pričak. frekvenca
0	32	28.32
1	15	21.24
2 ali več	13	10.44

  - Izračunamo testno statistiko:

$$X_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

  - Mejna vrednost pri alfa = 0.05: $\chi_0^2 = 2.94$
  - Hipotezo H0 ne moremo zavrniti. $\chi_0^2 > \chi_{0.05,1}^2 = 3.84$

Statistike enega vzorca

- Hi-kvadrat test: za testiranje porazdelitve populacije - primer
  - Matlab:

```
>> O = [32 15 13];
>> E = [28.32 21.24 10.44];
>> sum((O - E).^2 ./ E)
ans =     2.9392

>> chi2inv(0.95, 1)
ans =    3.8415
```

  - Ali:

```
bins = 0:3; obsCounts = [32 15 9 4];
n = sum(obsCounts);
lambdaHat = sum(bins.*obsCounts)/n;
expCounts = n*poisspdf(bins,lambdaHat);

[h,p,st] = chi2gof(bins,'ctrs',bins,'frequency',obsCounts,'expected',expCounts, 'nparams',1)

h = 0
p = 0.0719
st =     chi2stat: 3.2387
            df: 1
         edges: [-0.5000 0.5000 1.5000 3.5000]
             O: [32 15 13]
             E: [28.3420 21.2565 9.9640]
```

Statistike enega vzorca: neparametrične

- Test predznaka (neparametrična statistika):
  - Testiramo, ali je **ocenjena mediana vzorca** iz populacije **enaka predpostavljeni mediani populacije**.
    - Hipotezi:

$$
\begin{aligned}
H_0 &: M = M_0 \quad \text{ali } \theta = P(X > M_0) = P(X < M_0) = 0.5 \\
H_1 &: M \neq M_0 \quad \text{ali } \theta = P(X > M_0) \neq P(X < M_0) \\
H_1 &: M > M_0 \quad \text{ali } \theta = P(X > M_0) > P(X < M_0) \\
H_1 &: M < M_0 \quad \text{ali } \theta = P(X > M_0) < P(X < M_0)
\end{aligned}
$$

    - Testna statistika: K je število $(X_i - M_0) > 0$
    - Porazdelitev testne statistike: Binomska porazdelitev s parametri moč vzorca n in $\theta = 0.5$.

Statistike enega vzorca

- Test predznaka: primer

```
data = importdata('datasets/yvalues.txt');
x = data.data;

>> M0 = 1.12;     % predpostavljena mediana
>> median(x)      % izracunana mediana

ans = 1.1088

>> K = sum(x-M0 > 0)    % testna statistika
K = 14

>> binoinv([0.025 0.975], 39, 0.5)  %mejne vrednosti pri alf=0.05 obojestransko
ans =
    13    26                          H0 ne moremo zavrniti: 13 < K < 26

>> p = binocdf(K,39,0.5);          % p-vrednost za obojestranski test
>> p = 2*min(p,1-p)

p = 0.1081

>> [p, h, stat] = signtest(x, 1.12)  % sig test v matlabu
p =
    0.1081
h =     0

stat =
    zval: NaN
    sign: 14
```

Statistike dveh vzorcev

- Prej: statistike enega vzorca:
  - Obravnavamo lastnosti enega vzorca.
  - Predpostavimo neko lastnost, ki jo testiramo na danem vzorcu.

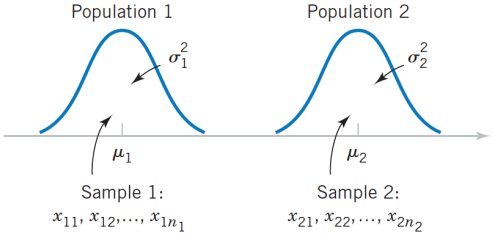
- Statistike dveh vzorcev
  - Primerjamo lastnosti dveh vzorcev.
  - Hipoteza 0: lastnosti obeh vzorcev se ne razlikujeta.
    - Prej: primerjali smo lastnost z neko predpostavljeno lastnostjo.
    - Sedaj: primerjamo isto lastnost na dveh različnih vzorcih.

Statistike dveh vzorcev

- z-test in t-test:
  - prej: testirali smo enakost povprečja ocenjenega iz enega vzorca s predpostavljenim povprečjem ob normalni porazdelitvi z znano (z-test) in neznano varianco (t-test)
  - sedaj: testiramo enakost ocenjenih povprečij iz dveh vzorcev ob predpostavki normalne porazdelitve populacije z znano (z-test) in neznano varianco (t-test).

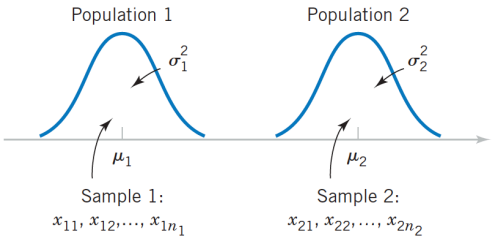
  - Predpostavke:

    - Vzorec 1: naključno izbrani primerki iz populacije 1.
    - Vzorec 2: naključno izbrani primerki iz populacije 2.
    - Obe populaciji, predstavljeni z vzorcev X_1 in X_2, sta neodvisni.
    - Obe populaciji sta normalno porazdeljeni.



Statistike dveh vzorcev

- z-test: znana varianca
  - Prej: hipotezi: $H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$ testna statistika: $z_0 = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
  - Sedaj:



    - Hipotezi:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2 \quad \text{ali } \mu_1 > \mu_2 \quad \text{ali } \mu_1 < \mu_2$$

    - Testna statistika:

$$z_0 = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

    - Porazdelitev testne statistike: Normalna porazdelitev $N(0, 1)$.

Statistike dveh vzorcev

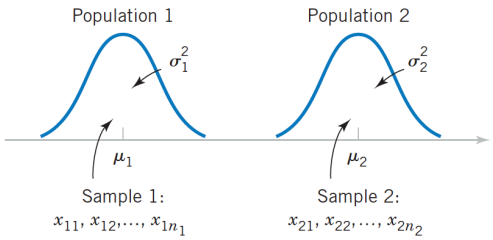
- z-test: primer
  - Pri razvoju barve za kovino testiramo uporabo sredstva za hitrejše sušenje barve.
  - Izvedemo eksperiment, kjer v prvem poskusu testiramo običajno barvo brez dodanega sredstva za sušenje, v drugem poskusu pa dodamo barvi sredstvo za sušenje.
  - V obeh poskusih smo pobarvali po 10 primerkov in merili čas sušenja barve. Že iz izkušenj poznamo std. odklon sušenja, ki znaša 8 minut.
  - V prvem primeru smo izmerili povprečje sušenja 121 minut, v drugem pa 112 minute.
  - Kaj lahko ugotovimo iz naših poskusov?

Statistike dveh vzorcev

- z-test: primer
  - Hipoteza $H_0 : \mu_1 = \mu_2$
  - Hipoteza $H_1 : \mu_1 > \mu_2$. Hipotezo H_0 lahko zavrnemo, če je prvo povprečje višje od drugega.
  - $\alpha = 0.05$
  - Testna statistika:
$$z_0 = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}},$$
kjer je $\sigma_1^2 = \sigma_2^2 = 8^2 = 64$ in $n_1 = n_2 = 10$.
  - Izračunamo testno statistiko:
$$z_0 = \frac{121 - 112}{\sqrt{\frac{8^2}{10} + \frac{8^2}{10}}} = 2.52$$
  - Izračunamo mejno vrednost za $\alpha = 0.05$ iz normalne porazdelitve $z_{0.95} = 1.645$ (enostranski test).
  - Zaključek: Ker velja $z_0 = 2.52 > 1.645$, lahko zavrnemo hipotezo H_0 z možnostjo napačne zavrnitve $\alpha = 0.05$ in lahko zaključimo, da dodajanje sredstva za sušenje statistično značilno izboljša hitrost sušenja.

Statistike dveh vzorcev

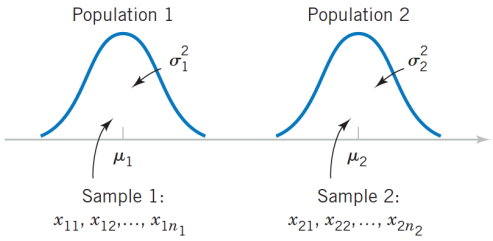
- t-test: neznana varianca - enake variance
  - Prej: hipotezi: $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$ testna statistika: $t_0 = \dfrac{\overline{x} - \mu}{s/\sqrt{n}}$
  - Sedaj:



    - Hipotezi:
$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2 \quad \text{ali } \mu_1 > \mu_2 \quad \text{ali } \mu_1 < \mu_2$$
    - Testna statistika v primeru enake variance:
$$t_0 = \frac{\overline{X}_1 - \overline{X}_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$
    - Porazdelitev testne statistike v primeru enake variance: Studentova t-porazdelitev s prostostno stopnjo $n_1 + n_2 - 2$.

Statistike dveh vzorcev

- t-test: neznana varianca - različne variance



  - Hipotezi:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2 \quad \text{ali } \mu_1 > \mu_2 \quad \text{ali } \mu_1 < \mu_2$$

  - Testna statistika v primeru različnih varianc:

$$t_0 = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

  - Porazdelitev testne statistike v primeru različnih varianc: Studentova t-porazdelitev s prostostno stopnjo

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}}$$

Statistike dveh vzorcev

- t-test: primer
  - Primer ugotavljanja koncentracij arzena v pitni vodi v zvezni državi Phoenix v ZDA.
  - Izvedene so bile meritve v mestnih predelih in na podeželju. Meritve so naslednje:

Metro Phoenix ($\bar{x}_1 = 12.5, s_1 = 7.63$)	Rural Arizona ($\bar{x}_2 = 27.5, s_2 = 15.3$)
Phoenix, 3	Rimrock, 48
Chandler, 7	Goodyear, 44
Gilbert, 25	New River, 40
Glendale, 10	Apachie Junction, 38
Mesa, 15	Buckeye, 33
Paradise Valley, 6	Nogales, 21
Peoria, 12	Black Canyon City, 20
Scottsdale, 25	Sedona, 12
Tempe, 15	Payson, 1
Sun City, 7	Casa Grande, 18

  - Preučimo možnost ali se koncentracije na podeželju razlikujejo od koncentracij v urbanih središčih.

Statistike dveh vzorcev

- t-test: primer
  - Hipoteza $H_0 : \mu_1 = \mu_2$
  - Hipoteza $H_1 : \mu_1 \neq \mu_2$.
  - $\alpha = 0.05$
  - Testna statistika:

$$t_0 = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}},$$

    kjer so $\overline{X}_1, S_1^2$ in $\overline{X}_2, S_2^2$ ocenjena povprečja in variance iz prvega in drugega vzorca. Izračunamo testno statistiko.
  - Izračunamo mejni vrednosti za zavrnitev hipoteze H_0 pri $\alpha = 0.05$. Če sta različni varianci, moramo oceniti število prostostnih stopenj:

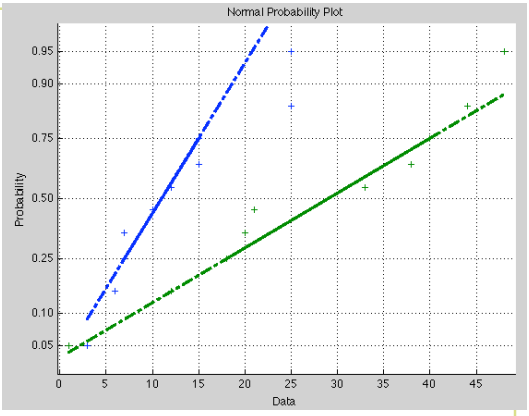
$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}}.$$

  - Zaključek: Če velja $t_0 < t_{0.025,\nu}$ ali $t_0 > t_{0.975,\nu}$, lahko zavrnemo hipotezo H_0 z možnostjo napačne zavrnitve $\alpha = 0.05$.

Statistike dveh vzorcev

- t-test: primer

```
metro_phx = [3 7 25 10 15 6 12 25 15 7];
rural_phx = [48 44 40 38 33 21 20 12 1 18];

normplot([metro_phx(:), rural_phx(:)])

n1 = length(metro_phx);    % stevilo meritev metro
n2 = length(rural_phx);    % stevilo meritev rural

m1 = mean(metro_phx)   % izracunano povprecje: metro
m2 = mean(rural_phx)   % izracunano povprecje: rural

s1 = std(metro_phx)    % izracunan std: metro
s2 = std(rural_phx)    % izracunan std: rural

>>
m1 = 12.5000
m2 = 27.5000

s1 = 7.6340
s2 = 15.3496
>>
t0 = (m1 - m2) / sqrt(s1^2/n1 + s2^2/n2)  % testna statistika
nu = (s1^2/n1 + s2^2/n2)^2 / ((s1^2/n1)^2/(n1-1) + (s2^2/n2)^2/(n2-1)) % stevilo prostostnih stopenj
>>
t0 = -2.7669
nu = 13.1956
>>
nu = round(nu);
tinv([0.025, 0.975], nu)     % meje za zavrnitev H0 pri alf = 0.05
ans =   -2.1604    2.1604
```



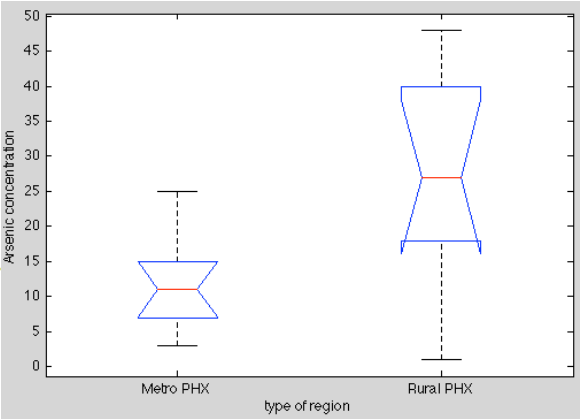
H0 lahko zavrnemo, saj: t0 < -2.16

Statistike dveh vzorcev

- t-test: primer

```
metro_phx = [3 7 25 10 15 6 12 25 15 7];
rural_phx = [48 44 40 38 33 21 20 12 1 18];

[h, p, ci, stat] = ttest2(metro_phx, rural_phx, 0.05, 'both', 'unequal') % ttest v matlabu
>>
h = 1
p = 0.0158

ci =
  -26.6941   -3.3059

stat =
    tstat: -2.7669
       df: 13.1956
       sd: [7.6340 15.3496]


boxplot([metro_phx(:), rural_phx(:)],1)
set(gca,'XTick',[1 2])
set(gca,'XtickLabel',{'Metro PHX','Rural PHX'})
xlabel('type of region')
ylabel('Arsenic concentration')
```

H0 lahko zavrnemo.



Statistike dveh vzorcev

- Test na vzorcih, sestavljenih iz vrednosti v parih: **parni t-test**
  - Imamo dva vzorca, ki sta sestavljena iz podatkov, ki nastopajo v parih, tako da so med seboj odvisni.
  - Primer: opazujemo skupino ljudi z visokim krvnim pritiskom
    - Prvi vzorec: merimo pritisk pred terapijo
      Drugi vzorec: merimo pritisk po terapiji.
    - Ugotavljamo, če je terapija uspešna.

- Hipotezi:

$$H_0 : \mu_D = \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_D \neq 0 \quad \text{ali } \mu_D > 0 \quad \text{ali } \mu_D < 0$$

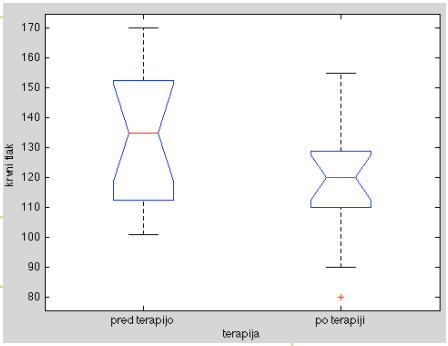
- Testna statistika:

$$t_0 = \frac{\overline{D}}{S_D/\sqrt{n}}$$

- Porazdelitev testne statistike:
  Studentova t-porazdelitev s prostostno stopnjo $n - 1$.

Statistike dveh vzorcev:
Test na vzorcih, sestavljenih iz vrednosti v parih

```
data = importdata('datasets/preassure.txt');
X = data.data;

boxplot(X,1)
set(gca,'XTick',[1 2])
set(gca,'XtickLabel',{'pred terapijo','po terapiji'})
xlabel('terapija')
ylabel('krvni tlak')
```




```
>> [h,p,ci, stat] = ttest(X(:,1), X(:,2)) % parni t-test
h =    1
p =    5.1261e-04
ci = 8.2114   23.2553
stat =
    tstat: 4.4862
       df: 14
       sd: 13.5829

>> [h,p,ci, stat] = ttest2(X(:,1), X(:,2)) % t-test: dva vzorca
h = 0
p = 0.0672
ci = -1.1907   32.6574
stat =
    tstat: 1.9043
       df: 28
       sd: 22.6266
```

- V prvem primeru je hipoteza H0 zavrnjena, torej obstaja stat. značilna razlika krvnega tlaka pred in po terapiji.
- V drugem primeru je p=7%, torej ne moremo zavrniti H0.
- To je zaradi upoštevane pozitivne korelacije med meritvami (osebki so isti, odvisnost).

Statistike dveh vzorcev: primerjava varianc

- Primerjamo, ali sta ocenjeni varianci iz dveh vzorcev enaki ob predpostavljeni normalni porazdelitvi populacije.

- **Izpeljava:**
  - Vzorec 1 iz populacije 1, ki je normalno porazdeljena s povprečjem μ_1 in varianco σ_1^2.
  - Vzorec 2 iz populacije 2, ki je normalno porazdeljena s povprečjem μ_2 in varianco σ_2^2.
  - Predpostavimo statistično neodvisnost obeh populacij.
  - Ocenimo varianco S_1^2 iz vzorca 1 in varianco S_2^2 iz vzorca.
  - Primerjajmo kvocienta S_1^2/σ_1^2 in S_2^2/σ_2^2:

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

  - Kvocient F je porazdeljen po F-porazdelitvi s prostostnima stopnjama $n_1 - 1$ in $n_2 - 1$.

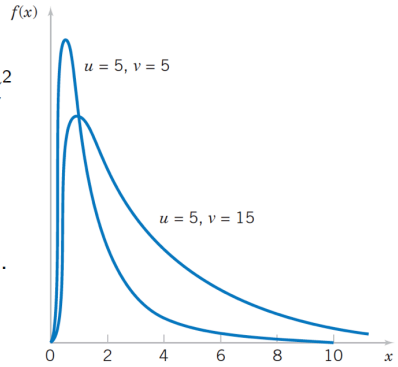
Statistike dveh vzorcev: primerjava varianc

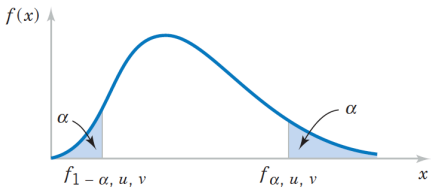
- F-porazdelitev:

- Če sta W in Y naključni spremenljivki, ki sta porazdeljeni po χ^2 porazdelitvi s prostostnima stopnjama u in v. Potem je

$$F = \frac{W/u}{Y/v}$$

porazdeljena po F-porazdelitvi s prostostnima stopnjama u in v.





- Velja:

$$f_{1-\alpha,u,v} = \frac{1}{f_{\alpha,u,v}}$$

- Primer:

$$f_{0.95,10,5} = \frac{1}{f_{0.05,10,5}} = \frac{1}{0.211} = 4.47$$

Statistike dveh vzorcev: primerjava varianc

- Hipotezi:

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2 \quad \text{ali } \sigma_1^2 > \sigma_2^2 \quad \text{ali } \sigma_1^2 < \sigma_2^2$$

- Testna statistika:

$$F_0 = \frac{S_1^2}{S_2^2}$$

- Porazdelitev testne statistike:
F-porazdelitev s prostostnima stopnjama $n_1 - 1$ in $n_2 - 1$.

Statistike dveh vzorcev: primerjava varianc

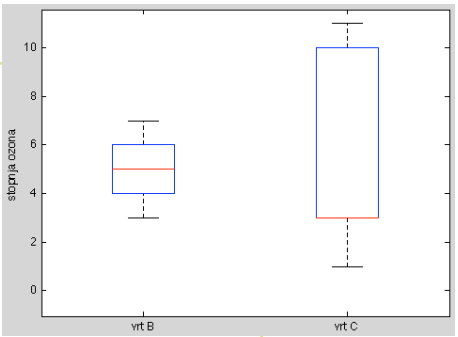
- Izvedba testiranja primerjave varianc:
  - Imamo dva vzorca in ocenimo varianci S_1^2 in S_2^2.
  - Predpostavimo normalno porazdeljenost populacije in neodvisnost obeh populacij.
  - Testiramo hipotezo $H_0 : \sigma_1^2 = \sigma_2^2$ ob alternativni hipotezi $H_1 : \sigma_1^2 \neq \sigma_2^2$ ali pa v ostalih dveh primerih.
  - Predpostavimo $\alpha = 0.05$ za zavrnitev hipoteze H_0.
  - Izračunamo testno statistiko
$$f_0 = \frac{S_1^2}{S_2^2}.$$
  - Izračunamo mejni vrednosti za zavrnitev hipoteze H_0 pri α, torej v našem primeru $f_{1-\alpha,u,v}$ in $f_{\alpha,u,v}$ s prostostnima stopnjama $u = n_1 - 1$ in $v = n_2 - 1$.
  - Zaključek: Če velja $f_0 < f_{\alpha,u,v}$ ali $f_0 > f_{1-\alpha,u,v}$, lahko zavrnemo hipotezo H_0 z možnostjo napačne zavrnitve α.

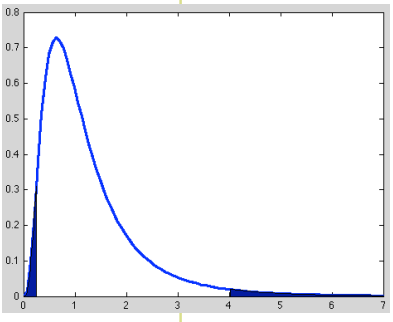
Statistike dveh vzorcev: primerjava varianc

- Testiranje varianc: primer

```
>> data = importdata('datasets/f.test.data.txt')
data =
          data: [10x2 double]
      textdata: {'gardenB'  'gardenC'}
    colheaders: {'gardenB'  'gardenC'}

>> X = data.data;
>> boxplot(X)
>> set(gca,'XTick',[1 2])
>> set(gca,'XtickLabel',{'vrt B','vrt C'})
>> ylabel('stopnja ozona')

V1 = var(X(:,1));      % ocenjena varianca vrta B
V2 = var(X(:,2));      % ocenjena varianca vrta C

n1 = length(X(:,1));   % stevilo meritev vrt B
n2 = length(X(:,2));   % stevilo meritev vrt C

F0 = V1 / V2           % testna statistika

alf = 0.05;            % stopnja zavrnitve
finv([alf/2, 1-alf/2], n1-1, n2-1)    %mejne vrednosti zavrnitve H0

>> F0 = 0.0938
ans =
    0.2484    4.0260

>> p = fcdf(F0,n1-1,n2-1);          % p-vrednost za obojestranski test
p = 2*min(p,1-p)

p = 0.0016
```

H0 lahko zavrnemo, ker F0 < 0.248.





Statistike dveh vzorcev: primerjava varianc

- Testiranje varianc: primer

```
>> [h,p,ci, stat] = vartest2(X(:,1), X(:,2), 0.05, 'both')
h =
     1
p =
    0.0016

ci =
    0.0233
    0.3774

stat =
    fstat: 0.0938
      df1: 9
      df2: 9
```

H0 je zavrnjena, ker p=0.0016.

Statistike dveh vzorcev

- Pravilo:
  - najprej preverjamo variance dveh vzorcev, potem pa šele povprečja.

Statistike dveh vzorcev: primerjava podatkov iz deležev

- Primer:
  - Denimo, da so v neki delovni organizaciji napredovale samo 4 ženske in 196 moških. Ali lahko rečemo, da gre za neenakopravnost med spoloma?
  - Preden to lahko rečemo, moramo povedati, koliko je zaposlenih moških in žensk. Denimo, da je zaposlenih 3270 moških in 40 žensk. To pomeni, da je napredovalo 10% žensk in le 6% moških.
  - Preverimo, ali je favoriziranje žensk statistično značilno, ali je bolj plod naključja.
  - To naredimo s testom, prirejenim za testiranje podatkov iz deležev.

Statistike dveh vzorcev: primerjava podatkov iz deležev

- Izberemo dva vzorca iz populacije z močjo n_1 in n_2. Naj predstavlja X_1 število primerkov, ki ustrezajo neki lastnosti iz vzorca 1, ki jo preučujemo. Podobno za vzorec 2 predstavlja število X_2.
- Izračunamo deleže $\hat{P}_1 = X_1/n_1$ in $\hat{P}_2 = X_2/n_2$.
- Hipotezi:

$$H_0 : p_1 = p_2$$
$$H_1 : p_1 \neq p_2 \quad \text{ali } p_1 > p_2 \quad \text{ali } p_1 < p_2$$

- Testna statistika:

$$Z_0 = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{P}(1-\hat{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

  kjer je $\hat{P} = \frac{X_1+X_2}{n_1+n_2}$.
- Porazdelitev testne statistike:
  Normalna porazdelitev $N(0,1)$.

Statistike dveh vzorcev: primerjava podatkov iz deležev

- Primer:
  - Napredovalo je 196 moških od 3270 zaposlenih in 4 ženske od 40 zaposlenih. Torej imamo $\hat{p}_1 = 196/3270 = 0.0599$ in $\hat{p}_2 = 4/40 = 0.1$.
  - Hipotezi: $H_0 : p_1 = p_2$ in $H_1 : p_1 \neq p_2$.
  - Izberemo stopnjo stat. značilnosti $\alpha = 0.05$.
  - Izračunamo testno statistiko (pred tem še $\hat{p} = \frac{x_1+x_2}{n_1+n_2} = \frac{196+4}{3270+40} = 0.0604$):

$$z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{0.0599 - 0.1}{\sqrt{0.0604(1-0.0604)\left(\frac{1}{3270}+\frac{1}{40}\right)}} = -1.0569$$

  - Izračunamo mejni vrednosti za zavrnitev hipoteze H_0 pri $\alpha = 0.05$, torej v našem primeru $z_{0.025} = -1.96$ in $z_{0.975} = 1.96$.
  - Ker $z_{0.025} < z_0 < z_{0.975}$, ne moremo zavrniti hipoteze H_0.

Statistike dveh vzorcev: brez predpostavke o normalnosti populacije

- Za testiranje povprečij smo uporabili:
  - z-test: v primeru, ko poznamo varianco populacije
  - t-test: v primeru, ko varianc ne poznamo
  - ob testa imata predpostavko, da so populacije normalno porazdeljene.
- Kaj pa če populacije, ki jih testiramo, niso normalno porazdeljene?
  - **Wilcoxonov test,**
    - testiramo, ali imata dva vzorca, ki izhajata iz populacije z neko porazdelitvijo, enaki mediani, ali da se mediani razlikujeta.

Wilcoxonov test

- To je neparametričen test, ki ga uporabimo v primeru, da populacije niso nujno normalno porazdeljene.

- Wilcoxonovo statistiko W izračunamo na naslednji način:
  - Vrednosti obeh vzorcev združimo v en seznam.
  - Gremo po seznamu in za vsako vrednost določimo, na katerem mestu je (po velikosti). Če imamo dve vrednosti enaki, potem si delijo mesto (ang. ties). To lahko obravnavamo na različne načine, pri Wilcoxonovem testu rečemo, da so na mestu, ki je povprečje obeh mest, ki si ju razdelita.

```
>> tiedrank([10 20 30 40 20])
ans =
    1.0000    2.5000    4.0000    5.0000    2.5000
```

  - Seštejemo vsa mesta prvega vzorca skupaj in vsa mesta drugega vzorca skupaj.
  - Primerjamo vsoti.
  - Če se vsoti stat. značilno razlikujeta, to pomeni, da hipotezo, da sta mediani enaki lahko zavrnemo (testna statistika za za večje vzorce aproksimira z normalno porazdelitvijo).

Wilcoxonov test

```
garden = importdata('datasets/gardens.txt');
data = garden.data;

% primerjamo vrt A in vrt B

>> xrank = tiedrank([data(:,1); data(:,2)]) % izracunamo vrstni red posameznih vrednosti

   6.0000   10.5000   10.5000    6.0000    2.5000    6.0000    1.0000 6.0000   15.0000
2.5000   15.0000   15.0000   18.5000   20.0000 10.5000   10.5000    6.0000   15.0000
18.5000   15.0000


>> sum(xrank(1:10))   % sestejemo vrstna mesta za vrt A
ans = 66

>> sum(xrank(11:20))  % sestejemo vrstna mesta za vrt B
ans = 144


>> [p,h, stat] = ranksum(data(:,1), data(:,2))  % Wilcoxonov test v matlabu
p = 0.0030
h = 1
stat =zval: -2.9689
      ranksum: 66

>> [h, p , ci, stat] = ttest2(data(:,1), data(:,2)) % primerjava s ttestom
h = 1
p = 0.0011
```

- Primerjava s t-testom:
  - Wilcoxonov test je bolj konzervativen. To pomeni, da če bomo s tem testom ugotovili, da sta povprečja signifikantno različna (p= 0.003), se bo to izrazilo še bolj pri Studentovem testu (p= 0.0011)
  - Po drugi strani ima t-test predpostavko o normalnosti populacije, kar ni vedno res.

Kontingenčne tabele

- Primer:
  - hočemo ugotoviti zvezo med barvo oči in barvo las pri ljudeh.
  - imamo npr. dve možni barvi oči (rjave in modre) in dve barvi las (svetla in temna)
  - Ljudi razvrščamo v te štiri kategorije in sicer štejemo število ljudi, ki spada v eno izmed kategorij.
  - Sestavimo tabelo:

	modre oči	rjave oči
svetli lasje	38	11
temni lasje	14	51

  - To imenujemo
    kontingenčna tabela = tabela frekvenc posameznih dogodkov v vzorcu.

Ocene verjetnosti iz kontingenčnih tabel

- Kakšna je verjetnost, da izberemo v takšnem vzorcu osebo s svetlimi lasmi?
  - Vse skupaj imamo 49 (38+11) ljudi s svetlimi lasmi in vseh ljudi je skupaj 114.
    Torej je verjetnost 49/114.
    Verjetnost, da izberemo oseb s temnimi lasmi je 65/114.
- Kakšna je verjetnost, da izberemo v takšnem vzorcu osebo z modrmi očmi?
  - Verjetnost 52/114.
  - Z rjavimi očmi 62/114.

	modre oči	rjave oči	**skupaj**
svetli lasje	38	11	49
temni lasje	14	51	65
skupaj	52	62	114

Kontingenčna tabela: test neodvisnosti

- Kakšna je verjetnost, da ima oseba svetle lase in modre oči?
- **Predpostavimo, da sta barva oči in barva las neodvisni količini.**
- **V tem primeru je verjetnost tega dogodka, produkt verjetnosti, da ima oseba svetle lase, z verjetnostjo, da ima modre oči**.
  - 49/114 x 52/114
- Podobno lahko naredimo še za ostala polja v tabeli.

	modre oči	rjave oči	**skupaj**
svetli lasje	49/114 x 52/114	49/114 x 62/114	
temni lasje	65/114 x 52/114	65/114 x 62/114	
skupaj			

Kontingenčna tabela: test neodvisnosti

- Ocenimo pričakovane frekvence v tabeli:
  - Pričakovano število ljudi s svetlimi lasmi in modrimi očmi je E = 49/114 x 52/114 x 114 = 22.35
  - To je mnogo manj od naših 38 ljudi. Zgleda, da je bila predpostavka o neodvisnosti med obema količinama prenagljena.
  - Podobno ocenimo še ostale frekvence.

	modre oči	rjave oči	**skupaj**
svetli lasje	22.35	26.65	49
temni lasje	29.65	35.35	65
skupaj	52	62	114

Kontingenčna tabela: test neodvisnosti

- Opazili smo, da se pričakovane frekvence (pod predpostavko neodvisnosti) in izmerjene frekvence razlikujejo.
- Ali je razlika statistično značilna?
- Izračunamo statistiko

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

- O je izmerjena frekvenca, E je pričakovana frekvenca, vsota po vseh elementih v tabeli.

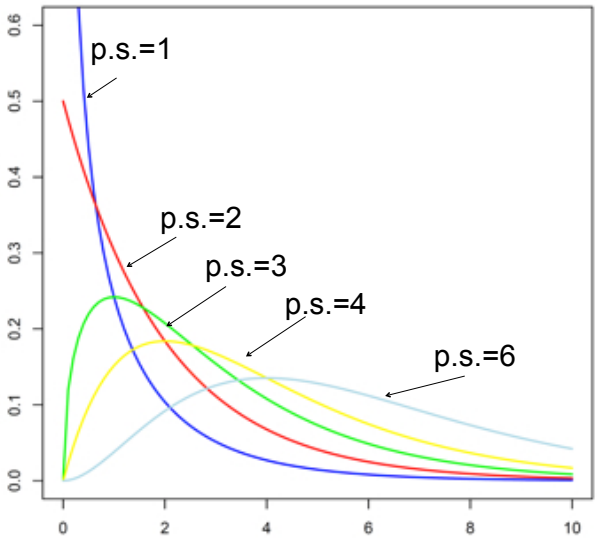
	O	E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
svetli lasje in modre oči	38	22.35	244.96	10.96
svetli lasje in rjave oči	11	26.65	244.92	9.19
temni lasje in modre oči	14	29.65	244.91	8.26
temni lasje in rjave oči	51	35.35	244.98	6.93

Kontingenčna tabela: test neodvisnosti

- Ko seštejemo skupaj, dobimo $\chi^2 = 35.33$
- Ali je testna statistika ustrezna, da bi lahko sprejeli ali zavrnili hipotezo, da sta ti dve količini med seboj neodvisni?
- Preden to ugotovimo, moramo vedeti naslednje:
  - po kakšni porazdelitvi se porazdeljuje testna statistika
  - koliko prostostnih stopenj nastopa v podatkih,
  - kakšna so mejne vrednosti za zavrnitev hipoteze ob znani porazdelitvi in predpostavljeni napaki zavrnitve alfa.

Kontingenčna tabela: test neodvisnosti

- Testna statistika se porazdeljuje po hi-kvadrat porazdelitvi:



Prostostne stopnje v kontingenčni tabeli

- V splošnem ima kontingenčna tabela *r* vrstic in *c* stolpcev, torej je število prostostnih stopenj = *(r-1)x(c-1)*
  - *V našem primeru (2-1)(2-1) = 1*

- Preverimo, če je to res.

	modre oči	rjave oči	**skupaj**
svetli lasje		11	49
temni lasje			65
skupaj	52	62	114

- Ker so robne vsote določene, imamo res v 2x2 tabeli samo eno prosto izbiro.

Kontingenčna tabela in hi-kvadrat test

- Kontingenčna tabela:

		Columns			
		1	2	…	c
Rows	1	O_{11}	O_{12}	…	O_{1c}
	2	O_{21}	O_{22}	…	O_{2c}
	⋮	⋮	⋮	⋮	⋮
	r	O_{r1}	O_{r2}	…	O_{rc}

- **Hipotezi:** **H0:** Količini izmerjeni po vrsticah in po stolpcih sta neodvisni.
  H1: Količini izmerjeni po vrsticah in po stolpcih nista neodvisni.
- Testna statistika: $$\chi_0^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$
- Porazdelitev testne statistike je hi-kvadrat s prostostno stopnjo (r-1)(c-1).
- Pogoj za zavrnitev hipoteze H0, pri določeni alfa: $\chi_0^2 > \chi^2_{1-\alpha,(r-1)(c-1)}$

Kontingenčna tabela: test neodvisnosti

- Naš primer:
  - Hipotezi:
    H_0 : Količini barva las in barva oči sta neodvisni.
    H_1 : Količini barva las in barva oči nista neodvisni.
  - V našem primeru smo izračunali:
    $$\chi_0^2 = 35.33$$
  - Mejna vrednost za zavrnitev hipoteze H_0 pri $\alpha = 0.05$:
    Število prostostnih stopenj je $(2-1)(2-1) = 1$.
    $$\chi^2_{0.95,1} = 3.84$$
  - Ker je $\chi_0^2 > \chi^2_{0.95,1}$, hipotezo H_0 lahko zavrnemo.

Kontingenčna tabela: test neodvisnosti

- V Matlabu:

```
% Pearson's Chi-squared test

% lasje oci
X = [38 11;  14 51];

e=zeros(1,4);
tot=sum(sum(X));
e(1)=(X(1,1)+X(1,2)).*(X(1,1)+X(2,1))/tot;
e(2)=(X(1,1)+X(1,2)).*(X(1,2)+X(2,2))/tot;
e(3)=(X(2,1)+X(2,2)).*(X(1,1)+X(2,1))/tot;
e(4)=(X(2,1)+X(2,2)).*(X(1,2)+X(2,2))/tot;

e
o=reshape(X', 1, 4);

chi2_0 = sum((o-e).^2 ./ e)

meja = chi2inv(0.95, 1)

%p-vrednost
p = 1-chi2cdf(chi2_0,1)

>>
e =
   22.3509   26.6491   29.6491   35.3509
o =
    38    11    14    51

chi2_0 = 35.3338

meja = 3.8415

p = 2.7777e-09
```

Povezava med količinami

- Kako sta povezani količini? V kakšni relaciji sta? Kakšna je njuna korelacija?
  - To lahko pogledamo iz kontingenčne tabele.

	modre oči	rjave oči
svetli lasje	38	11
temni lasje	14	51

dejanske vrednosti

	modre oči	rjave oči
svetli lasje	22.35	26.65
temni lasje	29.65	35.35

pričakovane vrednosti

  - Svetli lasje in modre oči so bile izmerjene 38-krat, pričakovana vrednost pa je bila 22.35. Skoraj 2-krat, torej je relacija pozitivna.
  - Podobno za temne lase in rjave oči.
  - In simetrično.

Test Kolmogorova in Smirnova

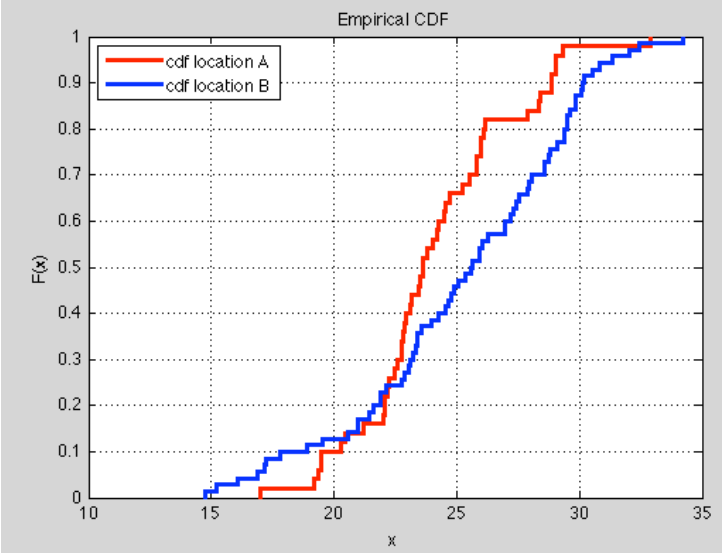
- Test Kolmogorova in Smirnova nam pomaga odgovoriti na naslednji vprašanji:
  - Prej: Ali je ocenjena porazdelitev iz vzorca enaka neki vnaprej napovedani porazdelitvi, s katero želimo modelirati populacijo? - statistika enega vzorca.
  - Sedaj: Ali sta porazdelitvi dveh različnih vzorcev enaki, ali sta statistično različni? - statistika dveh vzorcev (v nadaljevanju)

- Dve porazdelitvi sta lahko različni,
  - če imata različna povprečja,
  - če imata enaka povprečja, pa imata različni varianci ali momente višjega reda, itn.
  - če imata različno kumulativno funkcijo porazdelitve verjetnosti = test KS

Test Kolmogorova in Smirnova

- Test Kolmogorova in Smirnova primerja med seboj **kumulativne funkcije porazdelitve verjetnosti:**
  - Opazujemo velikost kril letečih žuželk v dveh geografsko različnih regijah.
  - Hočemo preveriti ali sta porazdelitvi verjetnosti dolžine kril, ocenjeni iz vzorcev žuželk iz obeh regij, enaki.

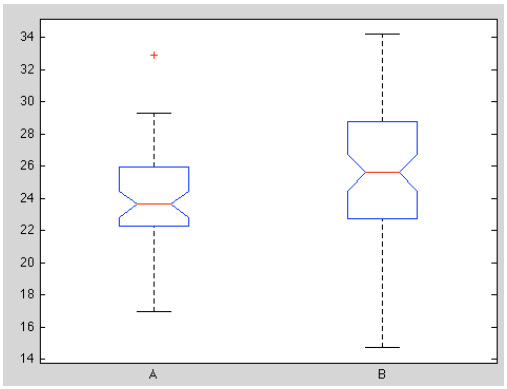
```
load datasets/wings.mat


F1 = cdfplot(wings(location=='A'));
hold on
F2 = cdfplot(wings(location=='B'))
set(F1,'LineWidth',2,'Color','r')
set(F2,'LineWidth',2)
legend([F1 F2],'cdf location A','cdf
location B','Location','NW')
hold off
```



Test Kolmogorova in Smirnova

- Primer:

Studentov t-test:
ne moremo zavrniti hipoteze H0, da je dolžina kril na različnih lokacijah enaka (p=0.13).




```
>> boxplot(wings, location, 'notch', 1)

>> [h, p, ci, stat] = ttest2(wings(location=='A'), wings(location=='B'))

h = 0
p = 0.1312
ci =
   -2.5736
    0.3384
stat =
    tstat: -1.5200
       df: 118
       sd: 3.9708
```

Test Kolmogorova in Smirnova

- Test Kolmogorova in Smirnova

```
>> [h,p,stat] = kstest2(wings(location=='A'), wings(location=='B'))
h =    1
p =    0.0287
stat = 0.2629
```

  - Test KS pokaže, da sta porazdelitvi dolžine kril signifikantno različni (p<0.05).
- Kje je razlika v porazdelitvah? Poglejmo variance.

```
>> var(wings(location=='A'))
>> var(wings(location=='B'))
ans =
    9.9701
ans =
   19.8841


>> [h,p,stat] = vartest2(wings(location=='A'), wings(location=='B'))
h =    1
p =    0.0119
stat =
    0.3007
    0.8560
```

5 Regresijska analiza

V tem poglavju je predstavljena regresijska analiza s poudarkom na linearnih regresijskih modelih, ki se uporabljajo tako za modeliranje podatkov, napovedovanje vrednosti iz podatkov in za ugotavljanje vplivnosti posameznih podatkov na modelirane količine.

Poglavje obravnava:

- **Linearna regresijska analiza:**
  - linearna regresija ene spremenljivke,
  - posplošeni linearni modeli - primer logistične regresije,
  - multipla linearna regresija,
  - koračna metoda določanja linearnega regresijskega modela.
- **Ne-linearni regresijski modeli:**
  - splošen model,
  - regresijsko drevo - primer neparametričnega regresijskega modela.

Statistično modeliranje

- Statistično modeliranje se uporablja za določanje modelov iz vzorcev.
- Statističen model je običajno matematična funkcija (v primeru parametričnega modela), ki ji na podlagi vzorca določimo parametre, tako da se kar najbolje ujema s podatki v vzorcu.
- Statističen model uporabljamo:
  - za bolj zgoščeno opisovanje podatkov iz vzorca,
  - za napovedovanje dogodkov,
  - za razvrščanje novih primerkov v različne populacije (razrede).
- Zahteve pri statističnih modelih so:
  - da se čim bolj natančno ujemajo s podatki, s katerih so ocenjeni,
  - da imajo lastnost posploševanja,
  - da so čim manj kompleksni.

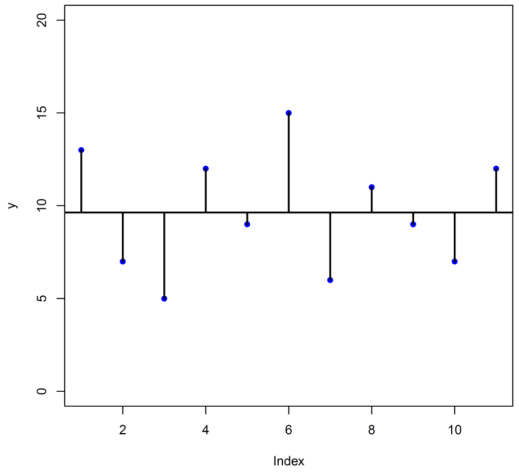
Regresijska analiza

- Regresijska analiza je postopek določitve modelov, da se ujemajo s podatki:
  - postopek modeliranja (regresije) je odvisen od modela
    - če je model parametričen, potem z regresijsko analizo ocenjujemo parametre modela,
    - če je model linearen, potem s postopki linearne algebre lahko določamo parametre modela z minimizacijo napake ujemanja modela s podatki,
    - če je model nelinearen in parametričen, uporabimo drugačne postopke iskanja optimalnih parametrov modela (nelinearna optimizacija).
    - če je model neparametričen (npr. regresijsko drevo), uporabimo drugačne postopke.
  - pri regresiji določamo odvisnost med odzivno spremenljivko Y in opisno spremenljivko X:
    - Y in X sta zvezni spremenljivki (če X diskretna (kategorijska) in Y zvezna = analiza variance)
    - če X vektor, imamo multiplo regresijo
    - če Y vektor, imamo multivariatno regresijo

Regresijska analiza

- Primer preprostega modela:
  - Imamo primerke v nekem vzorcu: $y_i, i = 1, \ldots, N$
  - Model je konstanta a.
  - Kako moramo določiti parameter a, da se bo najbolje "ujemal" s podatki?
  - Metoda najmanjših kvadratov:

$$\min_a \sum_{i=1}^{N} (y_i - a)^2$$



Regresijska analiza: primer

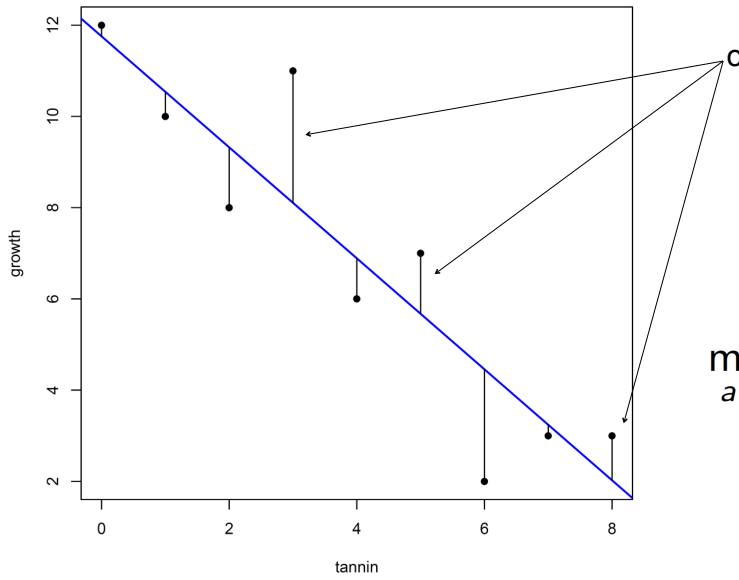
$$\begin{aligned}
\frac{\partial}{\partial a} \sum_{i=1}^{N} (y_i - a)^2 &== 0 \\
\sum_{i=1}^{N} 2(y_i - a)(-1) &== 0 \\
\sum_{i=1}^{N} (y_i - a) &== 0 \\
\sum_{i=1}^{N} y_i - \sum_{i=1}^{N} a &== 0 \\
\sum_{i=1}^{N} y_i &== Na \\
a &= \frac{1}{N} \sum_{i=1}^{N} y_i
\end{aligned}$$

To je povprečje.

Regresijska analiza: primer linearne zveze

- Model

$$y = a + bx$$



$$\min_{a,b} \sum_{i=1}^{N} (y_i - a - bx_i)^2$$

Regresijska analiza: primer linearne zveze

- Določitev parametrov a in b po metodi najmanjših kvadratov:

$$\min_{a,b} \sum_{i=1}^{N} (y_i - a - bx_i)^2$$

$$\frac{\partial}{\partial a} \sum_{i=1}^{N} (y_i - a - bx_i)^2 == 0$$

$$\frac{\partial}{\partial b} \sum_{i=1}^{N} (y_i - a - bx_i)^2 == 0$$

Regresijska analiza: primer linearne zveze

$$\frac{\partial}{\partial a}\sum_{i=1}^{N}(y_i - a - bx_i)^2 == 0$$ Izpeljava parametra a

$$\begin{aligned}
\sum_{i=1}^{N} 2(y_i - a - bx)(-1) &= 0 \\
\sum_{i=1}^{N} y_i - \sum_{i=1}^{N} a - \sum_{i=1}^{N} bx_i &= 0 \\
Na &= \sum_{i=1}^{N} y_i - b\sum_{i=1}^{N} x_i \\
a &= \frac{1}{N}\sum_{i=1}^{N} y_i - b\frac{1}{N}\sum_{i=1}^{N} x_i \\
a &= \overline{y} - b\overline{x}
\end{aligned}$$

Regresijska analiza: primer linearne zveze

$$\frac{\partial}{\partial b}\sum_{i=1}^{N}(y_i - a - bx_i)^2 == 0$$ izpeljava parametra b

$$\sum_{i=1}^{N} 2(y_i - a - bx_i)(-x_i) = 0$$

$$\sum x_i y_i - \sum a x_i - \sum b x_i^2 = 0$$

$$\sum x_i y_i - a\sum x_i - b\sum x_i^2 = 0$$

$$\sum x_i y_i - (\frac{1}{N}\sum y_i - b\frac{1}{N}\sum x_i)\sum x_i - b\sum x_i^2 = 0$$

$$\sum x_i y_i - \frac{1}{N}\sum y_i \sum x_i + b\frac{1}{N}\sum x_i \sum x_i - b\sum x_i^2 = 0$$

$$b = \frac{\sum x_i y_i - \frac{1}{N}\sum y_i \sum x_i}{\sum x_i^2 - \frac{1}{N}\sum x_i \sum x_i}$$

Regresijska analiza: primer linearne zveze

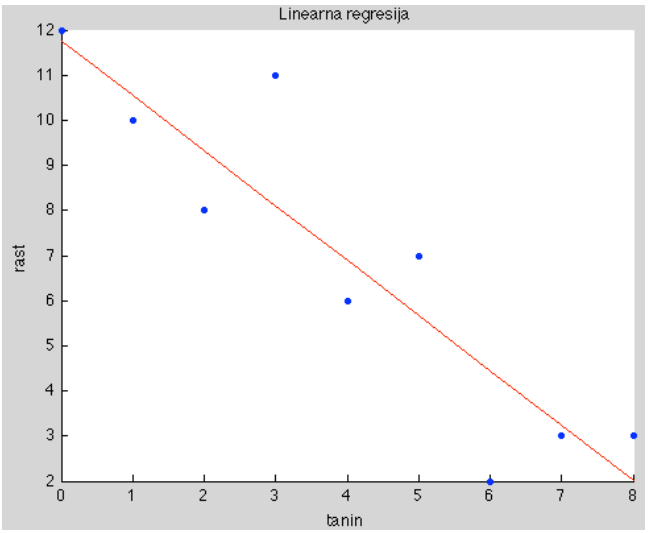
- Matlab:

```
tanin = importdata('datasets/tannin.txt');
data = tanin.data;

x = data(:,2); %tanin
y = data(:,1); %rast

SSX=sum(x.^2)-sum(x)^2/length(x)
SSXY=sum(x.*y)-sum(x)*sum(y)/length(x)

b = SSXY/SSX
a = mean(y) - b*mean(x)

>>
SSX = 60
SSXY = -73

b = -1.2167
a = 11.7556
```

Regresijska analiza: primer linearne zveze

- Matlab:

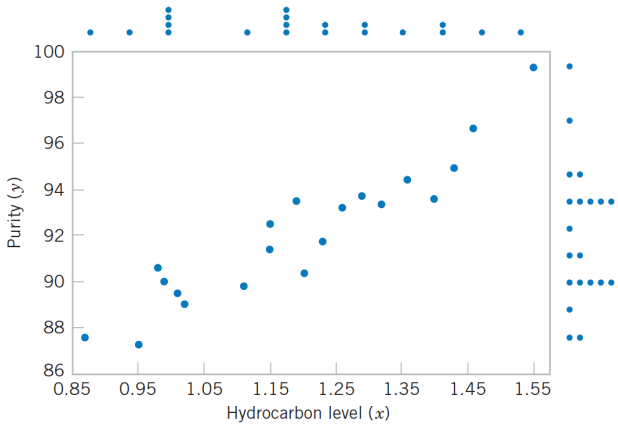
```
X = [ones(length(x),1), x];
b = regress(y,X)
>>
b =
   11.7556
   -1.2167

>>
hold on;
plot(x,y, '.')
plot(x, b(1) + b(2).*x, 'r-')
title('Linearna regresija')
xlabel('tanin'); ylabel('rast')
hold off;
```



Teorija linearne regresijske analize

- Denimo, da imamo primer, ko opazujemo stopnjo čistosti destilacije (y) od vsebnosti hidrokarbnskih snovi (x) v postopku destilacije:

Observation Number	Hydrocarbon Level x(%)	Purity y(%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33



Teorija linearne regresijske analize

- Potem lahko predpostavimo naslednjo linearno zvezo med Y in x:

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

kjer je ϵ napaka, ki jo lahko obravnavamo kot naključno spr. porazdeljeno s povprečjem 0 in varianco σ^2.

- Če izračunamo matematično upanje $E(Y|x)$, potem dobimo

$$E(Y|x) = E(\beta_0 + \beta_1 x + \epsilon) = \beta_0 + \beta_1 x + E(\epsilon) = \beta_0 + \beta_1 x.$$

Torej je povprečna vrednost Y pri danem x ob takšnem modelu enaka $\beta_0 + \beta_1 x$.

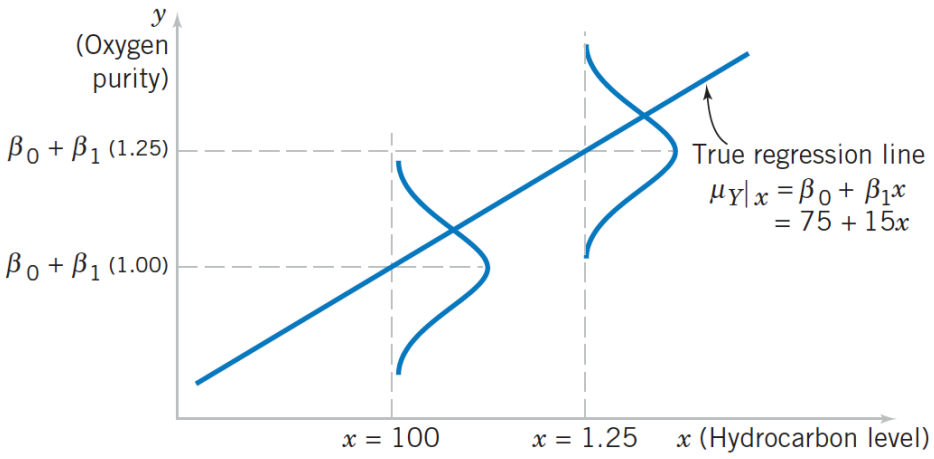
- Varianca Y pri danem x, pa

$$V(Y|x) = V(\beta_0+\beta_1 x+\epsilon) = V(\beta_0+\beta_1 x)+V(\epsilon) = 0+\sigma^2 = \sigma^2$$

Teorija linearne regresijske analize

- V našem primeru po metodi najmanjših kvadratov dobimo:

$$Y = 75 + 15x$$

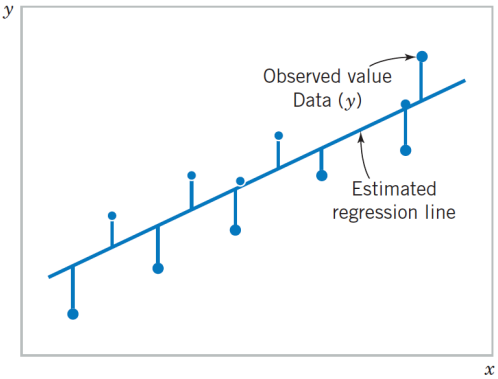


Teorija linearne regresijske analize

- Ocena parametrov linearne regresijske zveze po metodi najmanjših kvadratov:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, 2, \ldots, n$$

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$



$$\left.\frac{\partial L}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\left.\frac{\partial L}{\partial \beta_1}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$$

Teorija linearne regresijske analize

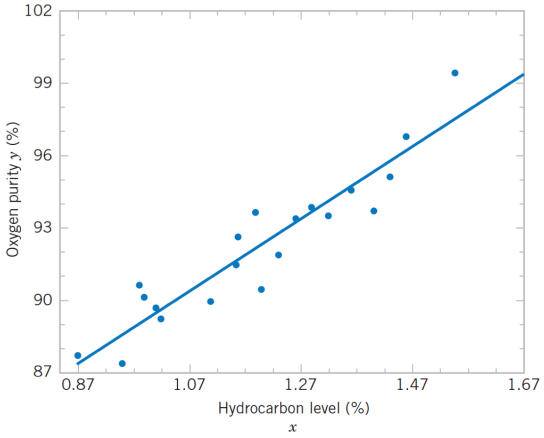
- Ocena parametrov linearne regresijske zveze po metodi najmanjših kvadratov:

$$SSXY = \sum x_i y_i - \frac{1}{N}\sum y_i \sum x_i$$
$$SSX = \sum x_i^2 - \frac{1}{N}\sum x_i \sum x_i$$
$$SSY = \sum y_i^2 - \frac{1}{N}\sum y_i \sum y_i$$

$$\hat{\beta}_1 = \frac{SSXY}{SSX} \qquad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i,\ \overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$



Ocenjevanje linearnega regresijskega modela

- Dobili smo ocene y-ov pri danih x-ih. Torej:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- Vendar imamo napake (residuali):

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \epsilon_i \qquad \epsilon_i = y_i - \widehat{y}_i$$

- Lastnosti napake:
  - povprečje: $\sum \epsilon_i = 0 \Rightarrow \overline{\epsilon} = 0$
  - varianca: $SSE = \sum (y_i - \widehat{y}_i)^2$

$$s^2 = var(\epsilon) = \frac{SSE}{n-2}$$

Ocenjevanje linearne regresijske zveze

- Preveriti želimo ali je ocenjena zveza med y in x statistično značilna.

- **Ničta hipoteza:**
  - Kvocient regresijske premice (beta1) je 0.
    (ni linearne regresijske zveze med y in x).
- **Alternativna hipoteza:**
  - Kvocient regresijske premice (beta1) ni nič.

- Lahko preverjamo z različnimi testi:
  - t-testom
    - **ali je kvocient enak 0?**
  - analiza variance
    - **ali variabilnost dejanskih meritev ustreza variabilnosti napovedanih meritev?**

Ocenjevanje linearne regresijske zveze

- Nekaj definicij:

$$SSR = \sum (\hat{y}_i - \overline{y})^2$$

vsota kvadratov razlike napovedanih vrednosti in povprečja

$$SSE = \sum (y_i - \hat{y}_i)^2$$

vsota kvadratov razlike dejanskih in napovedanih vrednosti

$$SSY = \sum (y_i - \overline{y})^2$$

vsota kvadratov razlike dejanskih vrednosti in povprečja

- Velja zveza:

$$SSY = SSR + SSE$$

Variabilnost dejanskih meritev je enaka variabilnosti napovedanih meritev + variabilnosti napake.

Ocenjevanje linearne regresijske zveze

	vsota kvadratov razlik	prostostne stopnje	povprečje kvadratov razlik	kvocient F
regresija	SSR	1	MSR = SSR/1	MSR/MSE
napake	SSE	n-2	MSE = SSE/(n-2)	
skupaj	SSY	n-1		

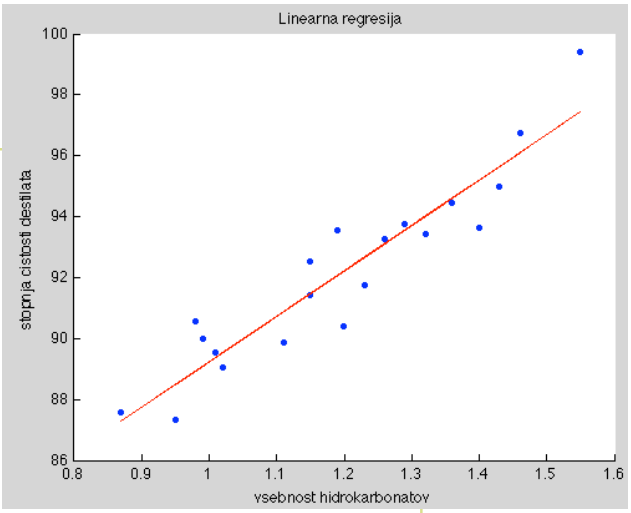
Primerjamo MSR/MSE z mejno vrednostjo F-porazdelitve ob predpostavki napačne zavrnitve hipoteze alfa = 0.05.

Če je izračunana vrednost večja od mejne vrednosti, lahko zavrnemo ničto hipotezo.

Ocenjevanje linearne regresijske zveze

- Primer:

```
data = importdata('datasets/destilacija.txt');

x = data(:,1); %hidrokarbonati
y = data(:,2); %stopnja cistosti destilata

X = [ones(length(x),1), x];
b = regress(y,X)

>>
b =

   74.2833
   14.9475

hold on;
plot(x,y, '.')
plot(x, b(1) + b(2).*x, 'r-')
title('Linearna regresija')
xlabel('vsebnost hidrokarbonatov'); ylabel('stopnja cistosti destilata')
hold off;
```



Ocenjevanje linearne regresijske zveze

- Primer:

	vsota kvadratov razlik	prostostne stopnje	povprečje kvadratov razlik	kvocient F
regresija	152.13	1	152.13	128.86
napake	21.25	18	1.18	
skupaj	173.38	19		

```
SSR = sum((y_n - mean(y)).^2)
SSE = sum((y - y_n).^2)
SSY = sum((y - mean(y)).^2)
>>
SSR =  152.1271
SSE =   21.2498
SSY =  173.3769

>>
n = length(y)
MSR = SSR/1
MSE = SSE/(n-2)
>>
n =    20
MSR =  152.1271
MSE =    1.1805

>> MSR/MSE
ans =  128.8617
```

Določimo mejno vrednost z možnostjo napake 5%.

```
>> finv(0.95, 1, 18)
% mejna vrednost za zavrnitev hipoteze
ans = 4.4139
```

Naša vrednost (128.86) je mnogo večja od mejne. Hipotezo zavrnemo.

p-vrednost:

```
>> p = fcdf(128.86, 1, 18)
>> min(p, 1-p)
ans =
   1.2274e-09
```

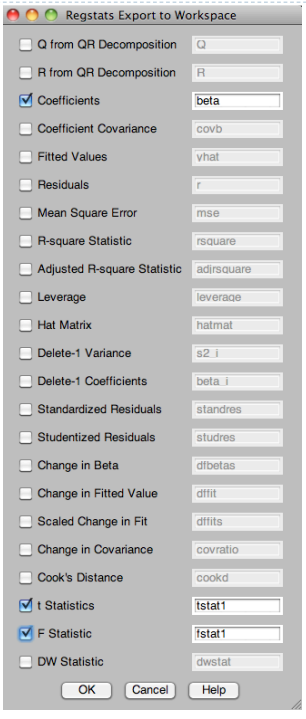
Ocenjevanje linearne regresijske zveze

- V Matlabu:

```
regstats(y,x, 'linear')
```

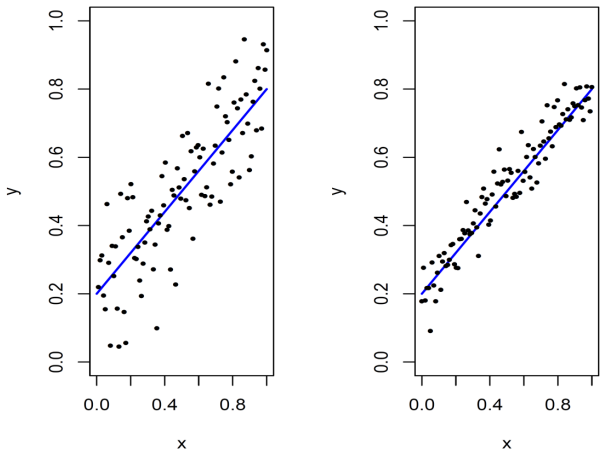
```
>> fstat

fstat =

     sse: 21.2498
     dfe: 18
     dfr: 1
     ssr: 152.1271
       f: 128.8617
    pval: 1.2273e-09
```

```
>> tstat
tstat =

    beta: [2x1 double]
      se: [2x1 double]
       t: [2x1 double]
    pval: [2x1 double]
     dfe: 18

>> tstat.beta, tstat.t, tstat.pval
ans =
   74.2833
   14.9475
ans =
   46.6172
   11.3517
ans =
   1.0e-08 *
    0.0000
    0.1227
```



Kvaliteta regresijskega modela

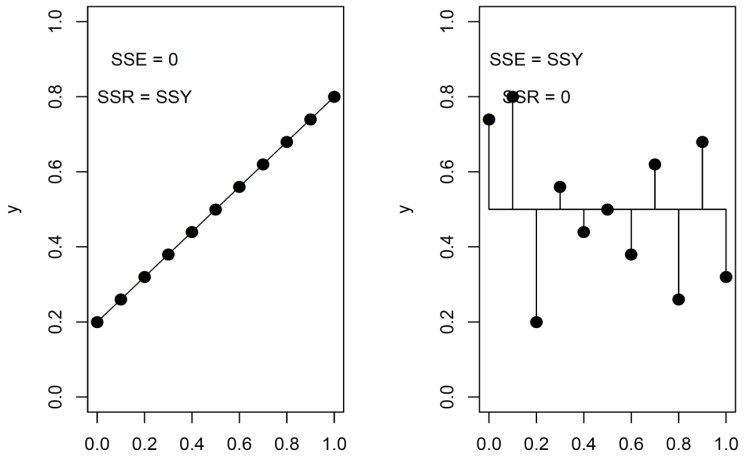
- Poglejmo dva primera:



  - Regresijski premici sta v obeh primerih enaki, toda ocenjeni sta bili iz različnih vzorcev.
  - Kateri model bolje opisuje podatke?
  - Potrebno je oceniti kvaliteto modela glede na ujemanje s podatki.

Kvaliteta regresijskega modela

- V ekstremnih primerih imamo dve situaciji:
  - vse točke ležijo natanko na ocenjeni premici – ujemanje je popolno, razpršitve glede na premico NI.
  - z x-i ne moremo opisati variabilnosti y-i: regresijski koeficient je 0, razpršitev glede na premico je popolna.



Kvaliteta regresijskega modela

- Poglejmo vrednosti SSR (varianca regresije), SSE (varianca napake) in SSY (varianca y, skupna varianca):
  - v prvem primeru je SSE = 0 in zato, ker je SSY = SSR + SSE, sledi SSR = SSY.
  - v drugem primeru je SSR = 0 in zato, ker je SSY = SSR + SSE, sledi SSE = SSY.

- Predlagana mera: r kvadrat – determinacijski koeficient
  - razmerje med varianco regresije in skupno varianco

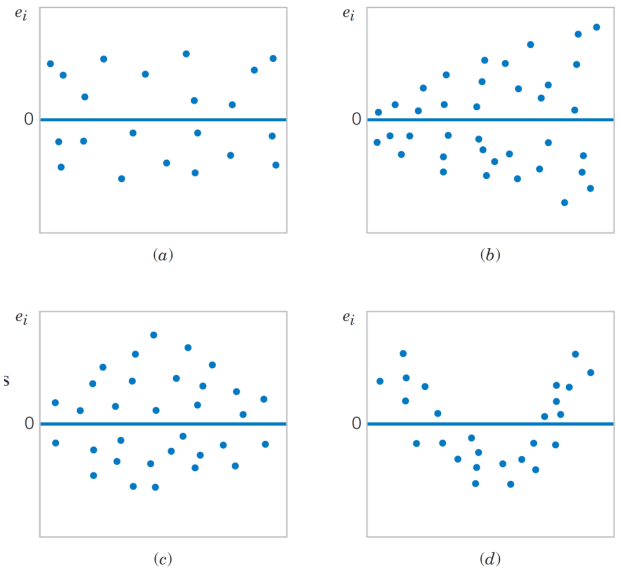
$$r^2 = \frac{SSR}{SSY}$$

  - Bolj kot je diskriminacijski koeficient blizu 1, boljši je model.

Kvaliteta regresijskega modela

- Druga mera:
  - preverjamo porazdelitev napake (residualov):
    - residuali morajo imeti povprečje 0 in konstantno varianco

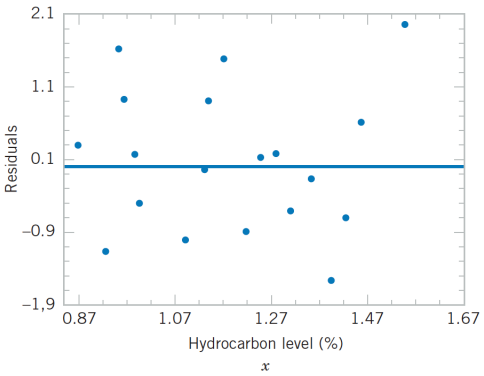
$$\epsilon_i = y_i - \hat{y}_i$$



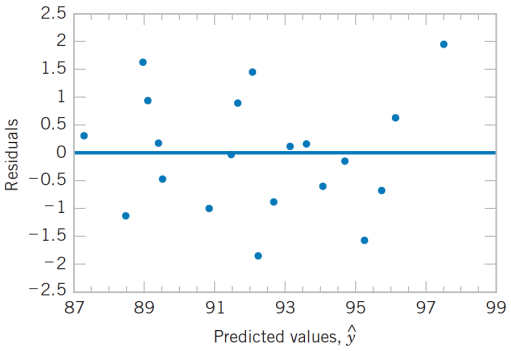
- Izris residualov v odvisnosti od opisne spr. x.
- Graf (a) je v redu:
  - residuali so enakomerno razpršeni po celotnem grafu.
    Varianca residualov je konstantna.
- Grafi (b),(c),(d) pa ne:
  - varianca residualov ni konstantna: **heteroskedastičnost**
  - v primeru (b) narašča, z večanjem indeksa (po času),
    - to lahko odpravimo s transformacijo odzivne spremenljivke (log(y), sqrt(y), 1/ y, ...)
  - v primeru (c) varianca residualov najprej narašča potem pada,
  - v primeru (d) podobno:
    - nakazuje neprimernost modela, potrebno je dodati člene višje stopnje v model.

Kvaliteta regresijskega modela

- Druga mera:
  - Včasih rišemo residuale tudi od napovedanih y.



residuali v odvisnosti od x

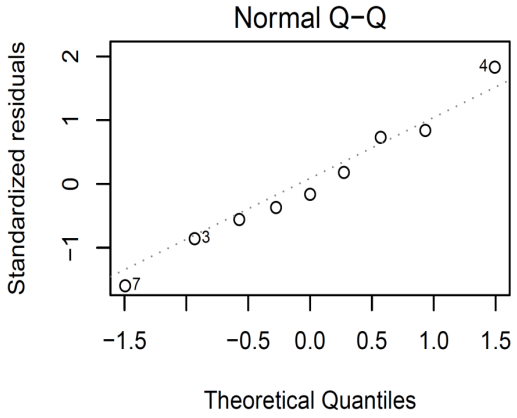


residuali v odvisnosti od yhat

Kvaliteta regresijskega modela

- Preverjanje normalnosti porazdelitve residualov:

$$d_i = e_i / \sqrt{\hat{\sigma}^2}$$



označene so točke, ki najbolj izstopajo od premice

Kvaliteta regresijskega modela

- Detekcija točk z največjim vplivom na ocenjene parametre modela:
  - največji vpliv na oceno parametrov imajo točke odstopanja (outliers):
    - ker zelo odstopajo od povprečja ostalih točk, prinesejo največ k oceni parametrov regresije po metodi najmanjših kvadratov.

- Cookova razdalja (v primeru linearne regresije):

točke z razdaljo več kot 1 so problematične

ocena y z modelom brez i-te meritve

$$D_i = \frac{\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_{j(i)})^2}{p \cdot \frac{\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}{n}}$$

število parametrov modela

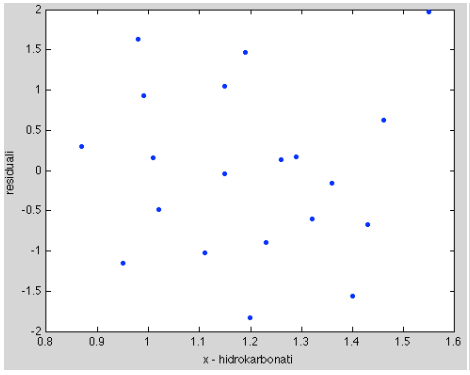
SSE/n

Kvaliteta regresijskega modela

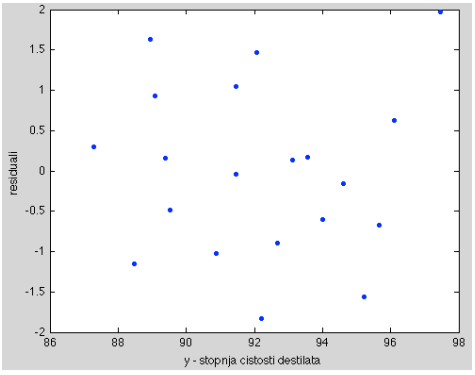
```
regstats(y,x, 'linear')
```

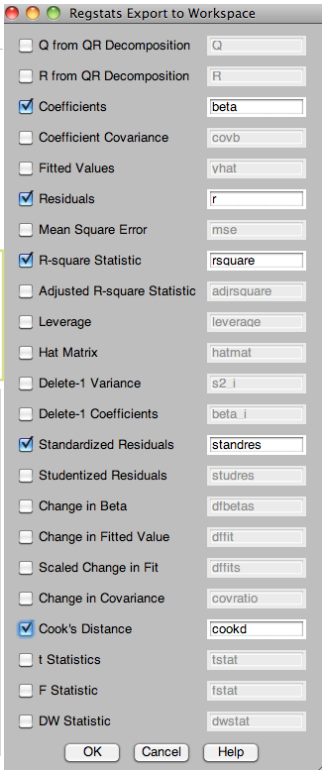
```
>> rsquare
rsquare = 0.8774
```

```
% primerjamo porazdelitev residualov
plot(x, r, '.')
xlabel('x - hidrokarbonati')
ylabel('residuali')
```




```
% primerjamo porazdelitev residualov
plot(yhat, r, '.')
xlabel('y - stopnja cistosti destilata')
ylabel('residuali')
```

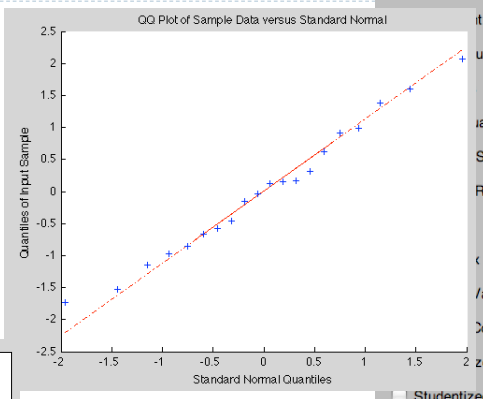


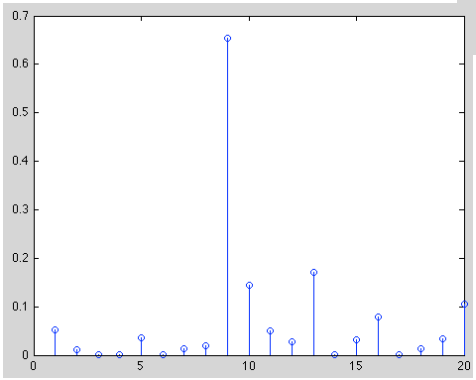


Kvaliteta regresijskega modela

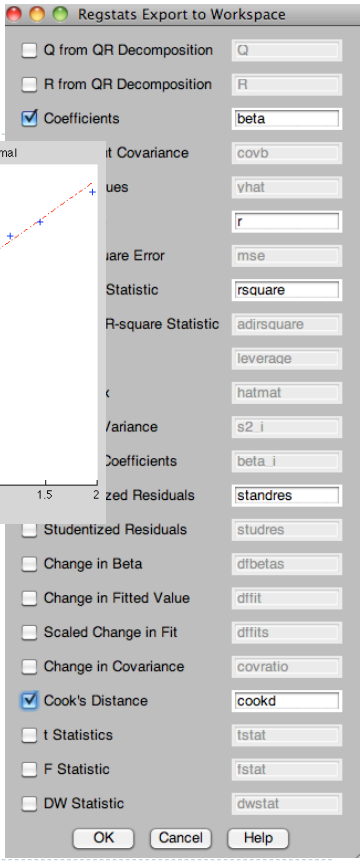
```
regstats(y,x, 'linear')
```

```
% preverjamo normalnost porazdelitve
std. residualov
qqplot(standres)
```





```
% Cookova razdalja in tocke,
% ki najbolj vplivajo na oceno
parametrov
stem(cookd)
```



Druge oblike modelov linearne regresije

- Včasih ni dovolj, da imamo samo linearno regresijsko zvezo. To opazimo bodisi iz grafa x in y meritev ali pa že poznamo naravo procesa, ki povezuje obe spremenljivki.
- Vendar se kljub temu nekatere zveze dajo pretvoriti v linearne zveze. Takšna je npr. eksponentna odvisnost med spremenljivkami:

$$Y = \beta_0 e^{\beta_1 x} \epsilon.$$

Takšno zvezo lahko spremenimo v linearno zvezo z logaritmiranjem:

$$\ln Y = \ln \beta_0 + \beta_1 x + \ln \epsilon.$$

Seveda je potrebno v tem primeru tudi zahtevati, da je porazdelitev napake porazdeljena s povprečjem 0 in konstantno varianco σ^2.

- Druga možna oblika je:

$$Y = \beta_0 + \beta_1 \frac{1}{x} + \epsilon$$

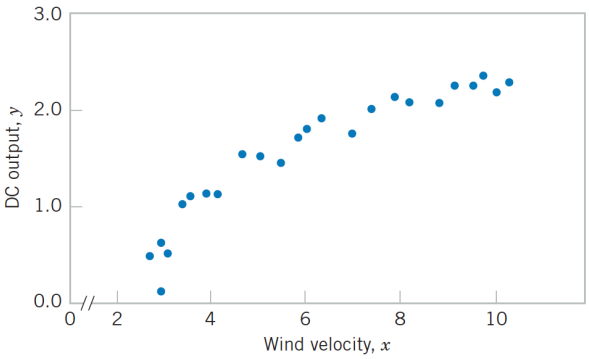
Prevedba v linearno obliko je možna z vpeljavo nove spremenljivke $z = 1/x$. Tako dobimo linearno regresijsko zvezo

$$Y = \beta_0 + \beta_1 z + \epsilon$$

ob enakih predpostavkah kot prej.

Primer regresijske zveze

- Primer nelinearne regresije v Matlabu:
  - Preučujemo povezavo med hitrostjo vetra in proizvedeno električno energijo pri vetrnih elektrarnah.
  - Kot lahko vidimo iz slike zveza ni ravno linearna, ampak se ob večjih hitrostih vetra asimptotično približuje vrednosti okoli 2.5.



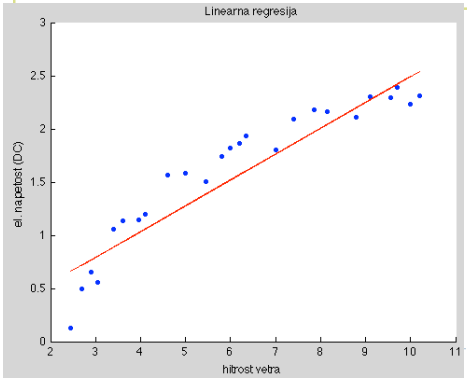
Observation Number, i	Wind Velocity (mph), x_i	DC Output, y_i
1	5.00	1.582
2	6.00	1.822
3	3.40	1.057
4	2.70	0.500
5	10.00	2.236
6	9.70	2.386
7	9.55	2.294
8	3.05	0.558
9	8.15	2.166
10	6.20	1.866
11	2.90	0.653
12	6.35	1.930
13	4.60	1.562
14	5.80	1.737
15	7.40	2.088
16	3.60	1.137
17	7.85	2.179
18	8.80	2.112
19	7.00	1.800
20	5.45	1.501
21	9.10	2.303
22	10.20	2.310
23	4.10	1.194
24	3.95	1.144
25	2.45	0.123

Primer regresijske zveze

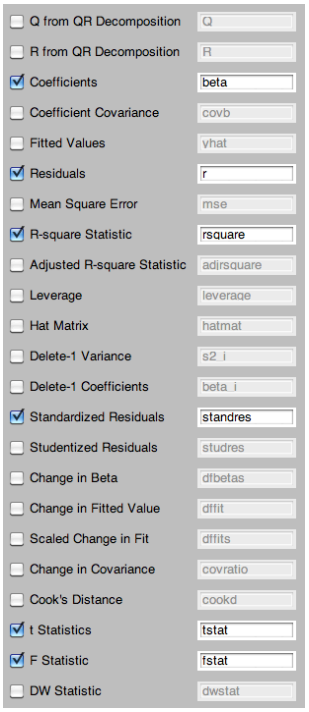
```
data = importdata('datasets/windmill.txt');
x = data(:,1); %hitros vetra
y = data(:,2); %el. napetost
```

```
%linearna regresija
regstats(y,x)
```

```
hold on;
plot(x,y, '.')
plot(x, yhat, 'r-')
title('Linearna regresija')
xlabel('hitrost vetra'); ylabel('el. napetost (DC)')
hold off;
```

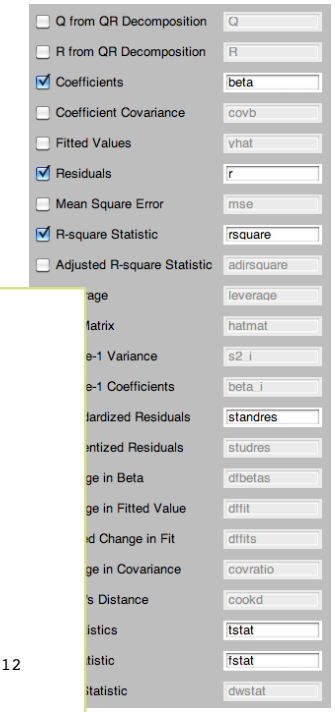



```
beta =
    0.1309
    0.2411
```

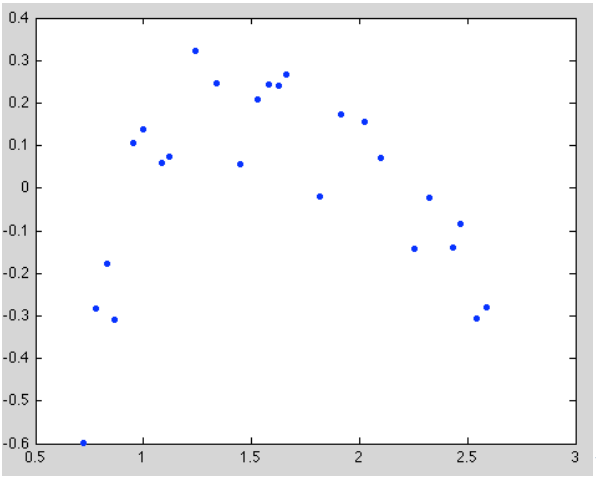


Primer regresijske zveze

```
%linearna regresija
regstats(y,x)

%kvaliteta modela
plot(yhat, r, '.')
```



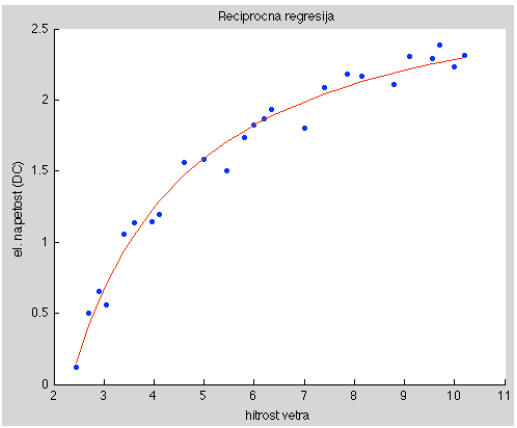
Porazdelitev residualov ni enakomerna.
Ta model ni najbolj primeren.



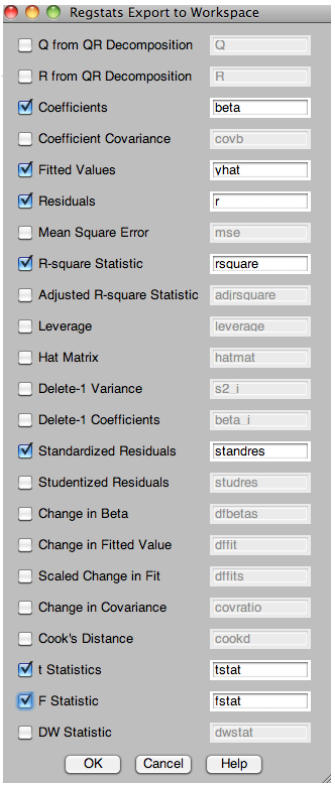

```
beta =
    0.1309
    0.2411
>> rsquare
rsquare =
    0.8745

>> mse
mse =
    0.0557
>> fstat
fstat =
     sse: 1.2816
     dfe: 23
     dfr: 1
     ssr: 8.9296
       f: 160.2571
    pval: 7.5455e-12
```

Primer regresijske zveze

```
%reciprocna regresija
regstats(y,1./x)

hold on;
plot(x,y, '.')
[sx, xi] = sort(x);
plot(sx, yhat(xi), 'r-')
title('Reciprocna regresija')
xlabel('hitrost vetra'); ylabel('el. napetost (DC)')
hold off;
```

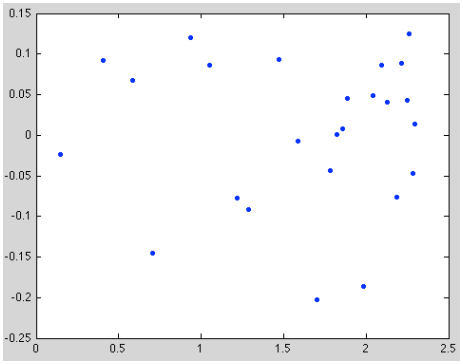



```
beta =
    2.9789
   -6.9345
```



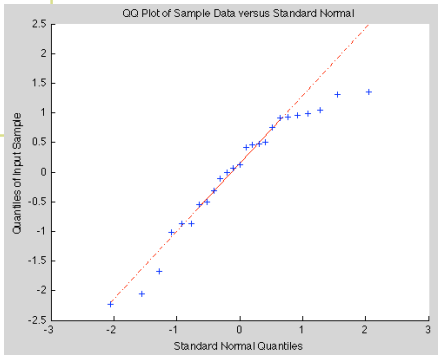
Primer regresijske zveze

```
%reciprocna regresija
regstats(y,1./x)

%kvaliteta modela
plot(yhat, r, '.')
```




```
rsquare =
    0.9800

mse =
    0.0089

fstat =
      sse: 0.2040
      dfe: 23
      dfr: 1
      ssr: 10.0072
        f: 1.1284e+03
     pval: 4.7425e-21
```

```
qqplot(standres)
```



Posplošeni linearni modeli

- Pri modelu linearne regresije smo imeli linearno zvezo med opisno in odzivno spremenljivko.
- Obstajajo pa tudi nelinearne regresijske zveze, ki se jih da linearizirati:
  - v prejšnjem primeru smo zvezo prediktorja preoblikovali tako, da smo z vpeljavo nove spr. dobili linearno zvezo:

$$y = \beta_0 + \beta_1 f(x)$$

  - lahko pa preoblikujemo tudi odzivno spremenljivko, da dobimo linearno zvezo:

$$f(y) = \beta_0 + \beta_1 x$$

- Pri linearni regresijski zvezi imamo naslednji model:
  - linearno regresijsko zvezo lahko zapišemo tudi kot $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, kjer je matrika $\mathbf{X}$ v našem primeru sestavljena iz $\mathbf{X} = [\mathbf{1}, \mathbf{x}]$ in $\beta = [\beta_0, \beta_1]$.
  - napovedane vrednosti $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$ so normalno porazdeljene in velja $\hat{y} = E(y|x)$ ob predpostavki normalne porazdelitve napake ϵ s povprečjem 0.

Posplošeni linearni modeli

- Pri posplošenemu linearnemu modelu predpostavimo naslednje:
  - linearno regresijsko zvezo lahko zapišemo kot $f(\mathbf{Y}) = \mathbf{X}\beta + \epsilon$, kjer je matrika $\mathbf{X}$ v našem primeru sestavljena iz $\mathbf{X} = [\mathbf{1}, \mathbf{x}]$ in $\beta = [\beta_0, \beta_1]$.
  - napovedane vrednosti $\hat{\mathbf{Y}}$ so lahko različno porazdeljene: normalno, binomsko, Poissonovo, ..., vse pa vključujejo povprečje $\hat{y} = E(y|x)$ v parametrih svojih porazdelitev.
  - regresijska zveza je $f(\hat{\mathbf{Y}}) = \mathbf{X}\hat{\beta}$.

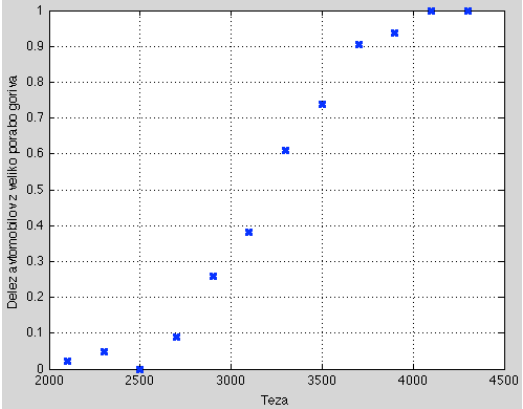
- Primer: logistična regresija.

Logistična regresija

- Imejmo primer avtomobilov različne teže:
  - Količina w predstavlja teže avtomobilov, v spr. total so prešteti avtomobili določene teže, v spr. poor pa so avtomobili, ki imajo veliko porabo goriva:

```
w = [2100 2300 2500 2700 2900 3100 ...
     3300 3500 3700 3900 4100 4300]';
total = [48 42 31 34 31 21 23 23 21 16 17 21]';
poor = [1 2 0 3 8 8 14 17 19 15 17 21]';
```

  - Narišimo deleže avtomobilov z veliko porabo goriva po teži:

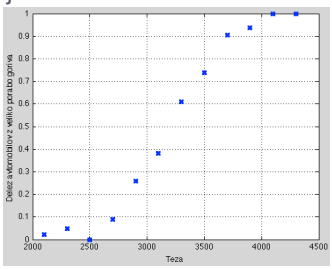


Logistična regresija

- Nadaljevanje:
  - Porazdelitev je v obliki sigmoidne funkcije, kar lahko zapišemo v naslednji obliki:

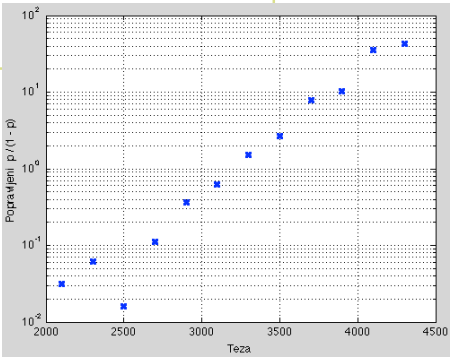
$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_2 \cdot w$$

kjer je p delež ”slabih” avtomobilov in w teža



  - Pri takšnem zapisu imamo nekaj težav, saj je funkcija v primeru p=0 in p = 1 nedefinirana. Zato deleže malo popravimo:

```
p_adjusted = (poor+.5)./(total+1);
semilogy(w,p_adjusted./(1-p_adjusted),'x','LineWidth',2)
grid on
xlabel('Teza')
ylabel('Popravljeni  p / (1 - p)')
```



Logistična regresija

- Nadalje lahko predpostavimo:
  - Binomsko porazdelitev spr. poor:
    - spr. poor predstavlja delež “slabih” avtomobilov od vseh avtomobilov določene teže = število cifer (poor) pri n (total) metih kovanca.
- V takem primeru zveza:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_2 \cdot w$$

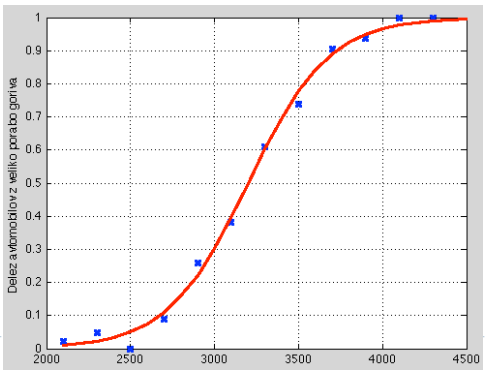
$$p = \frac{1}{1 + \exp(-\beta_0 - \beta_1 w)}$$

predstavlja model **logistično regresije** med w in p, funkcija log(p/(1-p)) se imenuje **logit funkcija**

- Izračun koeficientov v Matlabu:

```
b = glmfit(w,[poor total],'binomial','link','logit')
>>
b =-13.3801
    0.0042
```

```
x = 2100:100:4500;
y = glmval(b,x,'logit');

plot(w,poor./total,'x','LineWidth',2)
hold on
plot(x,y,'r-','LineWidth',2)
grid on
xlabel('Teza')
ylabel('Delez avtomobilov z veliko porabo goriva')
```



Posplošeni linearni modeli

- Možne funkcije posplošene linearne regresije v glmfit

`'link'`	`'identity'`, default for the distribution `'normal'`	$\mu = Xb$
	`'log'`, default for the distribution `'poisson'`	$\log(\mu) = Xb$
	`'logit'`, default for the distribution `'binomial'`	$\log(\mu/(1-\mu)) = Xb$
	`'probit'`	$\text{norminv}(\mu) = Xb$
	`'comploglog'`	$\log(-\log(1-\mu)) = Xb$
	`'reciprocal'`	$1/\mu = Xb$
	`'loglog'`, default for the distribution `'gamma'`	$\log(-\log(\mu)) = Xb$
	p (a number), default for the distribution `'inverse gaussian'` (with $p = -2$)	$\mu^p = Xb$
	cell array of the form `{FL FD FI}`, containing three function handles, created using `@`, that define the link (`FL`), the derivative of the link (`FD`), and the inverse link (`FI`).	User-specified link function

Multipla linearna regresija

- Multipla regresija podaja zvezo med eno odzivno zvezno spremenljivko (Y) in več zveznimi opisnimi spremenljivkami (X_i).

- Model multiple linearne regresije
  - k regresorjev – k opisnih spremenljivk x:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

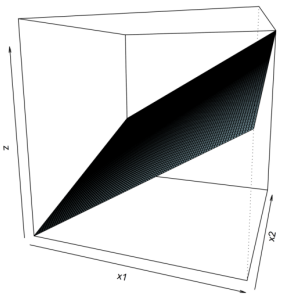
  - vključeni so samo linearni členi, zato linearna regresija.

  - ε je napaka pri oceni odzivne spremenljivke Y.

Multipla linearna regresija

- Primer multiple linearne regresije dveh spr.:

$$E(Y) = 50 + 10x_1 + 7x_2$$



- Primer s členi višjega reda:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$$

  - preuredimo jih tako, da dobimo samo linearne člene

$$x_1 = x, x_2 = x^2, x_3 = x^3$$

  in zapišemo:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

Multipla linearna regresija

- Primer s členi višjega reda:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon$$

  - preuredimo jih tako, da dobimo samo linearne člene

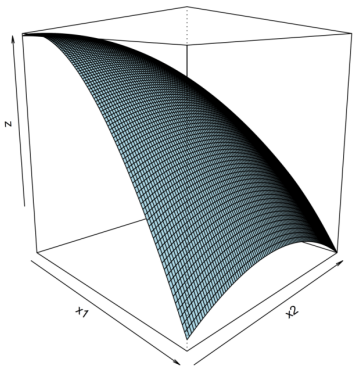
$$x_3 = x_1^2, x_4 = x_2^2, x_5 = x_1 x_2, \beta_3 = \beta_{11}, \beta_4 = \beta_{22}, \beta_5 = \beta_{12}$$

  in zapišemo:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \epsilon$$

  - Primer:

$$E(Y) = 800 + 10x_1 + 7x_2 - 8.5x_1^2 - 5x_2^2 + 4x_1 x_2$$



MLR: Ocenjevanje parametrov modela

- Metoda najmanjših kvadratov:
  - Denimo, da imamo n meritev in k (n>k) opisnih vrednosti x_i za vsako meritev y_i:

$$(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i), \qquad i = 1, 2, \ldots, n$$

  - Vsaka meritev ustreza zvezi:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$$
$$= \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \epsilon_i \qquad i = 1, 2, \ldots, n$$

  - Optimizacijska funkcija po metodi najmanjših kvadratov je:

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2$$

    - iščemo parametre $\beta_0, \beta_1, \ldots, \beta_k$ funkcije ***L,*** pri katerih bo funkcija dosegla minimum.

MLR: Ocenjevanje parametrov modela

- Iskanje minimuma funkcije L glede na ‘bete’:

$$\left.\frac{\partial L}{\partial \beta_0}\right|_{\hat{\beta}_0,\hat{\beta}_1,\ldots,\hat{\beta}_k} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k}\hat{\beta}_j x_{ij}\right) = 0$$

$$\left.\frac{\partial L}{\partial \beta_j}\right|_{\hat{\beta}_0,\hat{\beta}_1,\ldots,\hat{\beta}_k} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k}\hat{\beta}_j x_{ij}\right)x_{ij} = 0 \quad j = 1, 2, \ldots, k$$

- Dobimo linearen sistem enačb:

$$\begin{aligned}
n\hat{\beta}_0 + \hat{\beta}_1\sum_{i=1}^{n}x_{i1} + \hat{\beta}_2\sum_{i=1}^{n}x_{i2} + \cdots + \hat{\beta}_k\sum_{i=1}^{n}x_{ik} &= \sum_{i=1}^{n}y_i \\
\hat{\beta}_0\sum_{i=1}^{n}x_{i1} + \hat{\beta}_1\sum_{i=1}^{n}x_{i1}^2 + \hat{\beta}_2\sum_{i=1}^{n}x_{i1}x_{i2} + \cdots + \hat{\beta}_k\sum_{i=1}^{n}x_{i1}x_{ik} &= \sum_{i=1}^{n}x_{i1}y_i \\
\vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad & \vdots \qquad \vdots \\
\hat{\beta}_0\sum_{i=1}^{n}x_{ik} + \hat{\beta}_1\sum_{i=1}^{n}x_{ik}x_{i1} + \hat{\beta}_2\sum_{i=1}^{n}x_{ik}x_{i2} + \cdots + \hat{\beta}_k\sum_{i=1}^{n}x_{ik}^2 &= \sum_{i=1}^{n}x_{ik}y_i
\end{aligned}$$

Matrični zapis multiple linearne regresije

- Imamo zvezo:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i \qquad i = 1, 2, \ldots, n$$

- Lahko jo zapišemo v matrični obliki:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

  - pri čemer:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad \text{and} \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Matrični zapis multiple linearne regresije

- Optimizacijska funkcija po metodi najmanjših kvadratov v matrični obliki:

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \boldsymbol{\epsilon}'\boldsymbol{\epsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

- Iskanje minimuma:

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = \mathbf{0}$$

- Rešitev:

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$$

ali:

$$\hat{\boldsymbol{\beta}} = \underbrace{(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'}\mathbf{y}$$

psevdo inverz

Model multiple linearne regresije

- Ko izračunamo 'bete' – parametre modela, je naš model za opisovanje odzivne spremenljivke Y na podlagi opisnih spremenljivk x:

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \qquad i = 1, 2, \ldots, n$$

ali v matrični obliki:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

- Napaka modela: residuali

$$e_i = y_i - \hat{y}_i$$

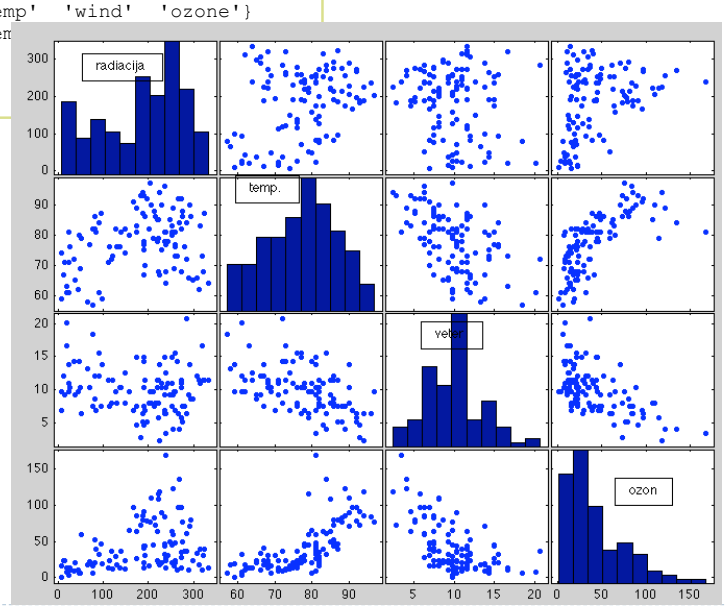
ali kot vektor:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$$

MLR: primer

- Preučevanje koncentracije ozona v odvisnosti od temperature, vetra in sončnega sevanja:

```
>> data = importdata('datasets/ozone.data.txt')
data =
          data: [111x4 double]
      textdata: {'rad'  'temp'  'wind'  'ozone'}
    colheaders: {'rad'  'tem

>> X = data.data;
>> gplotmatrix(X)
```



MLR: primer

- Preučevanje koncentracije ozona v odvisnosti od temperature, vetra in sonč. sevanja:

```
X = [ones(111,1), X];
>> X
X =
    1.0000  190.0000   67.0000    7.4000   41.0000
    1.0000  118.0000   72.0000    8.0000   36.0000
    1.0000  149.0000   74.0000   12.6000   12.0000
    1.0000  313.0000   62.0000   11.5000   18.0000
...
```

- Izračunajmo regresijske koeficiente:

```
y = X(:,4);
>> X = [ones(111,1), X(:,1:3)];
>> X'*X
   1.0e+06 *
    0.0001    0.0205    0.0086    0.0011
    0.0205    4.7048    1.6239    0.1993
    0.0086    1.6239    0.6817    0.0840
    0.0011    0.1993    0.0840    0.0124
>> X'*y
   1.0e+05 *

    0.0467
    9.7980
    3.8789
    0.3846
```

$$\mathbf{X'X\hat{\beta} = X'y}$$

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$$

```
>> X'*X \ X'*y
ans =
  -64.2321
    0.0598
    1.6512
   -3.3376
```

Funkcija v Matlabu

```
[beta] = regress(y,X)
beta =
  -64.2321
    0.0598
    1.6512
   -3.3376
```

MLR: Lastnosti ocenjenih parametrov

- Varianca napake:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-p} = \frac{SS_E}{n-p}$$

n … število podatkov
p … število parametrov regresije

- Vsota kvadratov napake:

$$SS_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \mathbf{e}'\mathbf{e}$$

- Vsota kvadratov razlike napovedanih vrednosti in povprečja:

$$SS_T = \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2 / n$$

spomnimo se pri regresiji

$$SSY = \sum (y_i - \bar{y})^2$$

- Vsota kvadratov razlike napovedanih vrednosti in povprečja:

Ali:

$$SS_E = SS_T - SS_R$$

$$SS_T = SS_R + SS_E$$

$$SSR = \sum (\hat{y}_i - \bar{y})^2$$

MLR: Statistična ustreznost regresijskega modela

- Test za statistično značilnost modela regresije je test, da obstaja (linearna) zveza med opisnimi spremenljivkami x in odzivno spremenljivko y.

- Ničta hipoteza:
- Alternativna hipoteza:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$$
$$H_1: \beta_j \neq 0 \quad \text{za vsaj en } j$$

	vsota kvadratov razlik	prostostne stopnje	povprečje kvadratov razlik	kvocient F
regresija	SS_R	k	MSR = SS_R/k	MSR/MSE
napake	SS_E	n-p	MSE = SS_E/(n-p)	
skupaj	SS_T	n-1		

Primerjamo MSR/MSE z mejno vrednostjo iz kvantila F-porazdelitve.
Če je izračunana vrednost večja od mejne vrednosti pri določeni napaki (alfa) napačne zavrnitve, lahko zavrnemo ničto hipotezo.

MLR: Statistična ustreznost regresijskega modela

- Primer:

	vsota kvadratov razlik	prostostne stopnje	povprečje kvadratov razlik	kvocient F
regresija	73837.79	3	24612.60	54.9066
napake	47964.12	111-4=107	448.2628	
skupaj	121801.9	110		

```
>> SST = y'*y - sum(y)^2/length(y)
SST =
   1.2180e+05

>> SSE = sum( (y - (beta' * X')').^2)
SSE =
   4.7964e+04

>> SSR = SST - SSE
SSR =

   7.3838e+04
```

```
>> MSR=SSR/3
MSR =
   2.4613e+04

>> MSE = SSE/107
MSE =
  448.2628

>> F0 = MSR/MSE
F0 =
   54.9066
```

MLR: Statistična ustreznost regresijskega modela

- Primer:

Določimo mejno vrednost za napačno zavrnitev hipoteze po Fisherjevi porazdelitvi z možnostjo napake 5%.

```
>> finv(0.95, 3, 107)
ans =
    2.6895
```

Naša vrednost (54.9) je mnogo večja od mejne. Hipotezo zavrnemo.

To lahko preverimo tudi s p-vrednostjo

```
>> 1-fcdf(54.9, 3, 107)
ans =     0
```

Hipotezo zavrnemo:
Koncentracija ozona je linearno odvisna ali od sevanja, ali od vetra, ali od temperature, ali od kakršnekoli kombinacije vsote teh treh količin.

Ne vemo pa od katere.

MLR: Koeficient multiple determinacije

- Druga mera za preverjanje ustreznosti regresijskega modela je koeficient multiple determinacije:

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

  - Primer:

```
>> R2 = SSR/SST
R2 = 0.6062
```

To pomeni, da smo z našim modelom pokrili 60% vse variance v podatkih.

  - Pomanjkljivost: z večanjem števila parametrov v model, se povečuje pokritost variance (več parametrov, boljše ujemanje s podatki).

- Zato: **popravljeni koeficient multiple determinacije**:

$$R^2_{\text{adj}} = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)}$$

  - Primer:

```
>> R2a = 1 - (SSE/107) / (SST/110)
R2a = 0.5952
```

MLR: Vrednotenje posameznih parametrov regresijskega modela

- Standardna napaka ocene parametrov:

$$SE(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 \cdot C_{jj}}$$

kjer je C_{jj} j-ti diagonalni element matrike $(\mathbf{X}'\mathbf{X})^{-1}$

- Hipoteza za vrednotenje posameznih parametrov multiple linearne regresije:

$$H_0: \beta_j = 0$$
$$H_1: \beta_j \neq 0$$

- t-statistika za preverjanje hipoteze:

$$t = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

- Hipotezo lahko zavrnemo, če je t statistika pod spodnjo vrednost t-porazdelitve ob alfa/2 oziroma nad mejno vrednost t-porazdelitve (1-alfa/2).

MLR: Vrednotenje posameznih parametrov regresijskega modela

- Primer:

```
>> invXX = inv(X'*X)

invXX =

   1.184429302588710   0.000059493837245  -0.012276710095694  -0.023280146840504
   0.000059493837245   0.000001198344622  -0.000003513403259  -0.000000767967750
  -0.012276710095694  -0.000003513403259   0.000143260023842   0.000179237632158
  -0.023280146840504  -0.000000767967750   0.000179237632158   0.000953709815145

SEb = sqrt(SSE/107 * 1.184429302588710)
T0 = beta(1) / SEb

p = tcdf(T0, 107);
p = 2*min(p, 1-p)

>>
SEb =   23.0420
T0 =   -2.7876

p = 0.0063
```

standardna napaka pri oceni parametra β_0

t-statistika za preverjanje, če je $\beta_0 = 0$

p-vrednost statistike

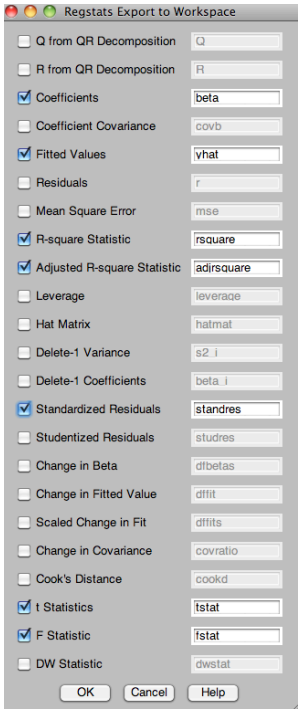
```
SEb = sqrt(SSE/107 * invXX(2,2));
T0 = beta(2) / SEb
p = tcdf(T0, 107);
p = 2*min(p, 1-p)
>>
T0 = 2.5800
p = 0.0112
```

```
SEb = sqrt(SSE/107 * invXX(3,3));
T0 = beta(3) / SEb
p = tcdf(T0, 107);
p = 2*min(p, 1-p)
>>
T0 = 6.5159
p = 2.4292e-09
```

```
SEb = sqrt(SSE/107 * invXX(4,4));
T0 = beta(4) / SEb
p = tcdf(T0, 107);
p = 2*min(p, 1-p)
>>
T0 = -5.1046
p =   1.4497e-06
```

MLR: Vrednotenje posameznih parametrov regresijskega modela

```
>> regstats(y, X)

>> rsquare
rsquare = 0.6062

>> adjrsquare
adjrsquare = 0.5952

>> fstat

fstat =

     sse: 4.7964e+04
     dfe: 107
     dfr: 3
     ssr: 7.3838e+04
       f: 54.9066
    pval: 1.4457e-21

>> tstat.t
ans =
   -2.7876
    2.5800
    6.5159
   -5.1046

>> tstat.pval
ans =
    0.0063
    0.0112
    0.0000
    0.0000
```

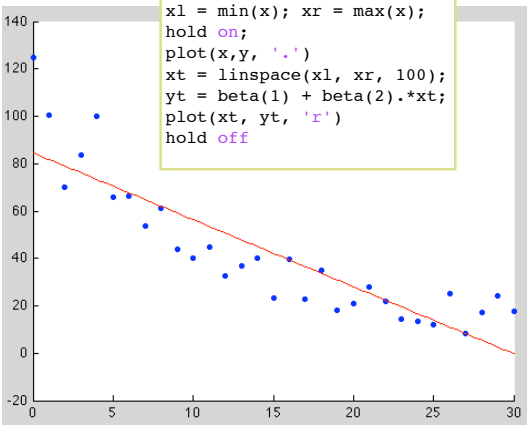


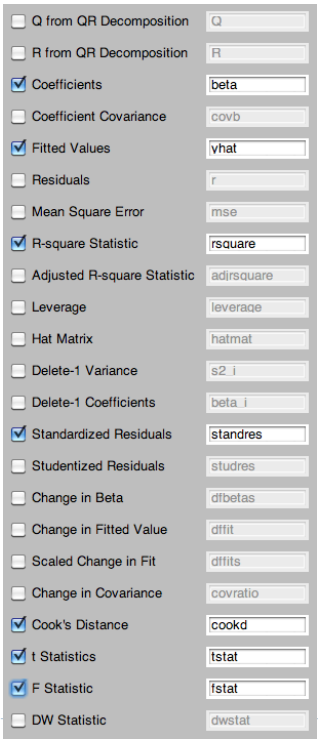
MLR: primer polinomske regresije

- Primer:

```
data = importdata('datasets/decay.txt')

x = data.data(:,1);
y = data.data(:,2);

regstats(y,x) %linearna regresija
```

```
xl = min(x); xr = max(x);
hold on;
plot(x,y, '.')
xt = linspace(xl, xr, 100);
yt = beta(1) + beta(2).*xt;
plot(xt, yt, 'r')
hold off
```





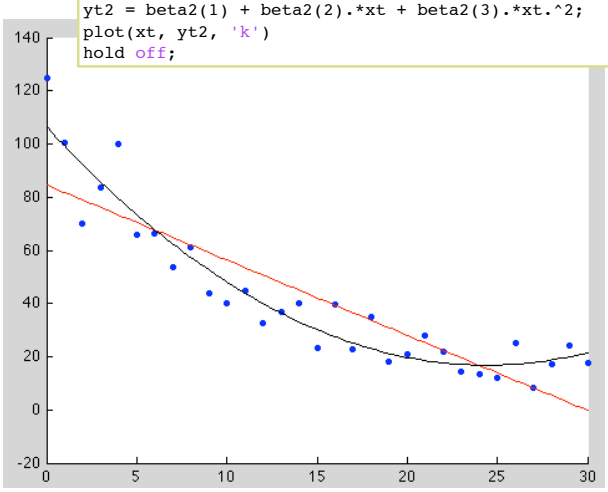
Komentar:
Večina napak na obeh robovih je pozitivna. Na sredini pa negativna. To ni najbolje.

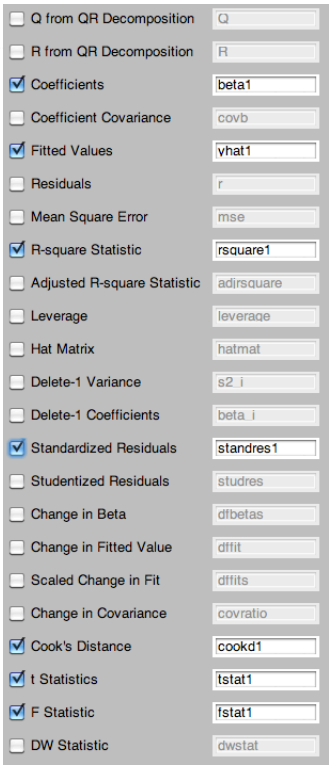
MLR: primer polinomske regresije

- Primer: kvadratna regresijska zveza

```
X = [x, x.^2];
regstats(y,X) %kvadratna regresija
```

```
hold on;
yt2 = beta2(1) + beta2(2).*xt + beta2(3).*xt.^2;
plot(xt, yt2, 'k')
hold off;
```





Komentar:
Kvadratna regresijska zveza deluje bolje kot linearna.

MLR: primer polinomske regresije

- Analiza kvadratne zveze:

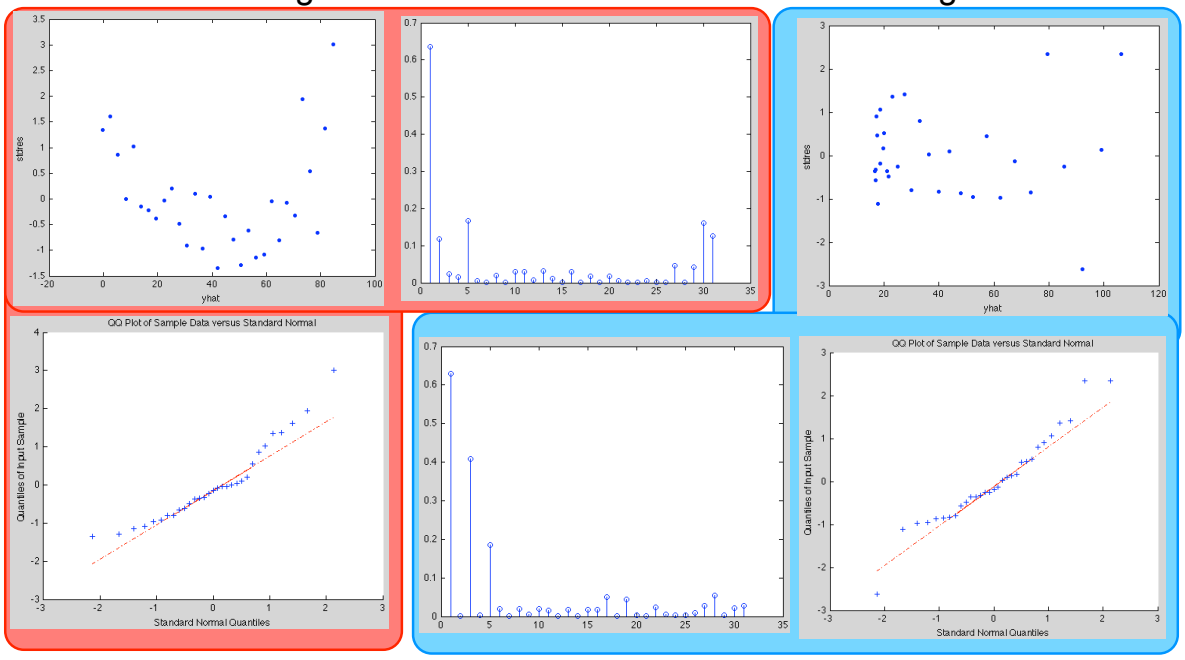
```
fprintf('Coefficients:\n');
fprintf('            \tEstimate\t\t t value\t\t p-val\n');
for i=1:3
    fprintf('beta(%d)\t\t%f\t\t%f\t\t%e\n', i, beta2(i), tstat2.t(i), tstat2.pval(i));
end
fprintf('----\n')
fprintf('Multiple R-squared: %f,      Adjusted R-squared: %f\n', rsquare2,adjrsquare2);
fprintf('F-statistic: %.2f on %d and %d DF,  p-value: %e\n', fstat2.f, fstat2.dfr, fstat2.dfe, fstat2.pval);

>>
Coefficients:
                    Estimate            t value            p-val
beta(1)             106.388804          22.848511          1.201127e-19
beta(2)             -7.344849           -10.223347         5.897173e-11
beta(3)             0.150589            6.507245           4.727037e-07
----
Multiple R-squared: 0.907979,      Adjusted R-squared: 0.901406
F-statistic: 138.14 on 2 and 28 DF,  p-value: 3.121963e-15
```

$$y = 106.388 - 7.34485\,x + 0.15059\,x^2$$

p-vrednost kvadratnega člena je manjša od 0.05, kar štejemo za statistično značilno. Torej je uvedba kvadratnega člena smiselna.

Primerjava obeh modelov:

- Kvaliteta linearnega modela:
- Kvaliteta kvadratnega modela:



Iskanje pravega regresijskega modela

- Kako poiščemo primeren regresijski model?

- Iskanje pravega modela:
  - Katere opisne spremenljivke vključiti v model in kakšne kombinacije?
  - Kakšni so medsebojni odnosi med opisnimi spremenljivkami?
  - Korelacije med opisnimi spr.
  - Ali je mogoče zmanjšati število prediktorjev in s tem določiti jasne povezave med opisni in odzivnimi spr. - iskanje minimalnega modela.

- Kot smo že videli, nam ocene o ustreznosti nekaj povedo o kvaliteti modela:
  - koliko variance uspemo opisati z modelom,
  - kateri členi so statistično značilni,
  - kako se porazdeljuje napaka med dejanskimi in napovedanimi vrednostmi.

Koračna metoda grajenja regresijskega modela

- Postopek koračne izgradnje regresijskega modela:

  - izberemo si začetni model in ocenimo njegove parametre, potem pa primerjamo ustreznost modelov z več ali manj dodanimi členi.

  - v vsakem koraku izračunamo F-statistiko in ocenimo p-vrednost ter primerjamo vrednosti med modelom z ali brez obravnavanega člena.

  - če člen ni vključen v model, potem obravnavamo ničto hipotezo, da bi bil koeficient beta pri tem členu enak 0. Če lahko to hipotezo zavrnemo, potem člen vključimo v model.

  - Podobno, če je člen že vključen v model in je ničta hipoteza, daj je koeficient beta pri tem členu enak 0, in če ugotovimo, da ne moremo zavrniti hipoteze, potem ta člen odstranimo iz modela.

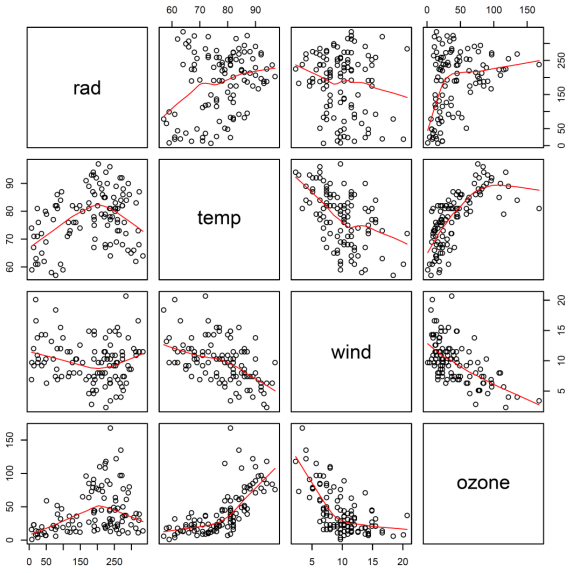
Koračna metoda grajenja regresijskega modela

- Postopek
  1. Naredimo začetni model in ocenimo parametre modela.
  2. Če v modelu nimamo členov, ki imajo p-vrednost, ki je večja od neke začetne tolerančne meje (to pomeni, da bi bil beta ob tem členu različen od 0, če bi bil člen v modelu), dodamo člen z najmanjšo p-vrednostjo v model in ponavljamo ta korak, sicer gremo na korak 3.
  3. Če imamo v modelu člene s p-vrednostjo večjo od predpisane izstopne tolerančne meje (to pomeni, da je zelo verjetno, da je beta enak 0), izločimo člen z največjo p-vrednostjo in gremo na korak 2, sicer končamo.

- Lokalna optimalnost (ne globalna):
  - ustreznost modela je odvisna od začetnega modela, saj glede na začetni model dodajamo ali odvzemamo člene,
  - to pomeni, da je končna ustreznost modela odvisna od začetnega modela in ni nujno, da vedno dobimo globalno optimalno rešitev.

Primer grajenja regresijskega modela

- Preučevanje koncentracije ozona v odvisnosti od temperature, vetra in sončnega sevanja:




```
data = importdata('datasets/ozone.data.txt')
```

- Odzivna spr. je koncentracija ozona.
- Opisne spr. so sončno sevanje, temperatura, veter.
- Na diagramu lahko vidimo porazdelitev meritev za posamezno kombinacijo spr.
- Zanima nas zadnja vrstica:
  - kako se obnaša koncentracija ozona v odvisnosti od ostalih količin.

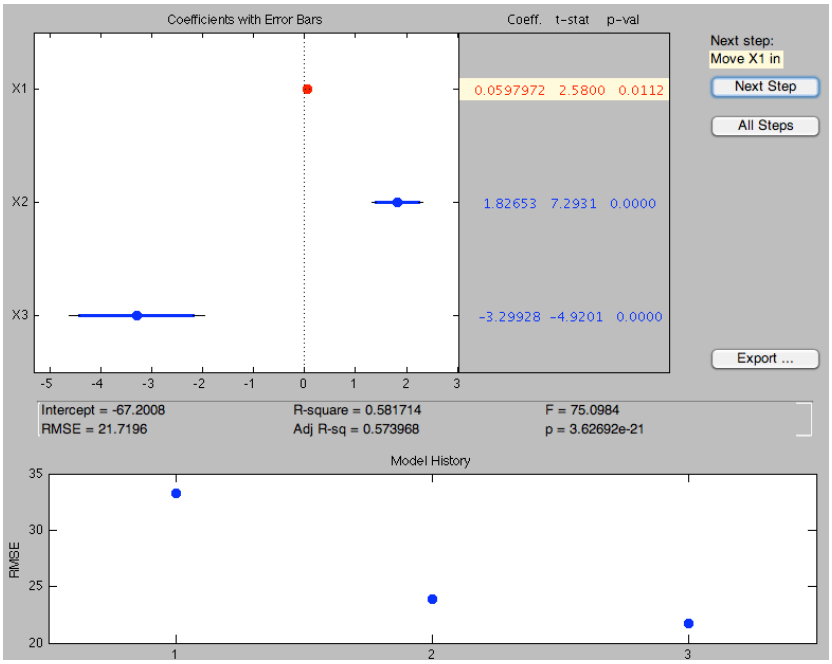
- Vidimo lahko:
  - negativno korelacijo med hitrostjo vetra in koncentracijo ozona (več vetra, manjša koncentracija)
  - pozitivno korelacijo med temperaturo in koncentr. ozona (višja temperatura, večja koncentracija)
  - nejasno zvezo med sončnim sevanjem in koncentracijo ozona.

Primer grajenja regresijskega modela

- Matlab:

```
X = data.data;

ozone = X(:,4);
rad   = X(:,1);
temp  = X(:,2);
wind  = X(:,3);
```

```
% osnovni model
X = [rad, temp, wind];
stepwise(X, ozone);
```

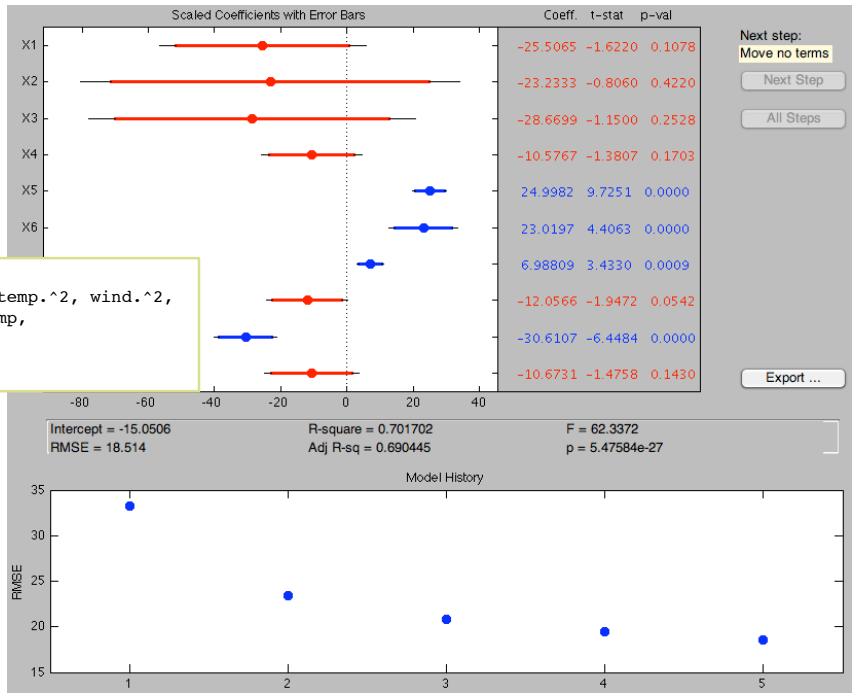


Primer grajenja regresijskega modela

- Matlab:

```
X = data.data;

ozone = X(:,4);
rad   = X(:,1);
temp  = X(:,2);
wind  = X(:,3);
```

```
% bolj kompleksni model
X = [rad, temp, wind, rad.^2, temp.^2, wind.^2,
rad.*temp, rad.*wind, wind.*temp,
rad.*temp.*wind];
stepwise(X, ozone);
```



Primer grajenja regresijskega modela

```
X = [rad, temp, wind];
[betahat1, se1, pval1, outModel1, stats1] = stepwisefit(X, ozone, 'penter', 0.05, 'premove', 0.10);

>> stats1.rmse
ans =   21.1722
```

```
X = [rad, temp, wind, rad.^2, temp.^2, wind.^2, rad.*temp, rad.*wind, wind.*temp, rad.*temp.*wind];
[betahat2, se2, pval2, outModel2, stats2] = stepwisefit(X, ozone, 'penter', 0.05, 'premove', 0.10);

>> stats2.rmse
ans =   18.5140
```

```
X = [rad, temp, wind, rad.^2, temp.^2, wind.^2, rad.*temp, rad.*wind, wind.*temp, rad.*temp.*wind];
initialModel = [false true true false false false false false false false];
[betahat3, se3, pval3, outModel3, stats3] = stepwisefit(X, ozone, 'penter', 0.05, 'premove', 0.10,
'inmodel', initialModel);

>> stats3.rmse
ans =   18.1369
```

Nelinearni regresijski modeli

- Ne-linearna regresija:
  - pri linearni regresiji ponavadi nimamo nekega znanja, da bi vedeli, kako se tvorijo odzivne meritve, zato predvidevamo neko zvezo, ki jo potem ustrezno ovrednotimo in dodamo še ostale člene, če je potrebno.
  - pri nelinearni regresiji pa običajno poznamo zvezo med opisnimi in odzivno spremenljivko - poznamo (fizikalni) model, oceniti moramo parametre.
  - Parametrični modeli:
    - Model:
      $$y = f(X, \beta) + \epsilon$$
    - y je vektor n meritev odzivne spremenljivke,
    - X je matrika velikosti $[n, p]$, kjer je podana zveza med med odzivnimi in opisnimi spremenljivkami, n je število meritev, p je število členov, ki nastopa v modelu,
    - β je vektor dimenzije p, ki predstavlja število parametrov, ki jih je potrebno oceniti,
    - f je model, ki povezuje opisne spr. x in parametre β z odzivno spr. y,
    - ϵ je vektor dimenzije n, ki predstavlja napako med dejanskimi meritvami in modelom. Predpostavimo, da je napaka enako porazdeljena.

Nelinearni regresijski modeli

- Ne-linearna regresija:
  - neparametrični modeli
    - če ne poznamo jasne zveze med opisnimi in odzivno spremenljivko, potem ne moremo predpostaviti modela s parametri,
    - ena rešitev: neparametrični modeli
      - **Primer: regresijsko drevo**

Regresijsko drevo

- Pri parametričnih regresijskih modelih poznamo oz. predpostavljamo neko zvezo opisnimi in odzivnimi spremenljivkami:
  - definiramo model (linearen, polinomski, logistična regresija) in potem ocenjujemo parametre tega modela ter ugotavljamo ustreznost modela
  - pri tem moramo zadostiti nekaterim predpostavkam, ki jih preverjamo ob gradnji modela.
- V veliko primerih pa ne poznamo zveze med opisnimi in odzivnimi spremenljivkami:
  - ena možnost je regresijsko drevo, ki predstavlja neparametrični regresijski model
  - v primeru, ko imamo odzivne spremenljivke kategorijske, govorimo o klasifikacijskem drevesu.

Regresijsko drevo

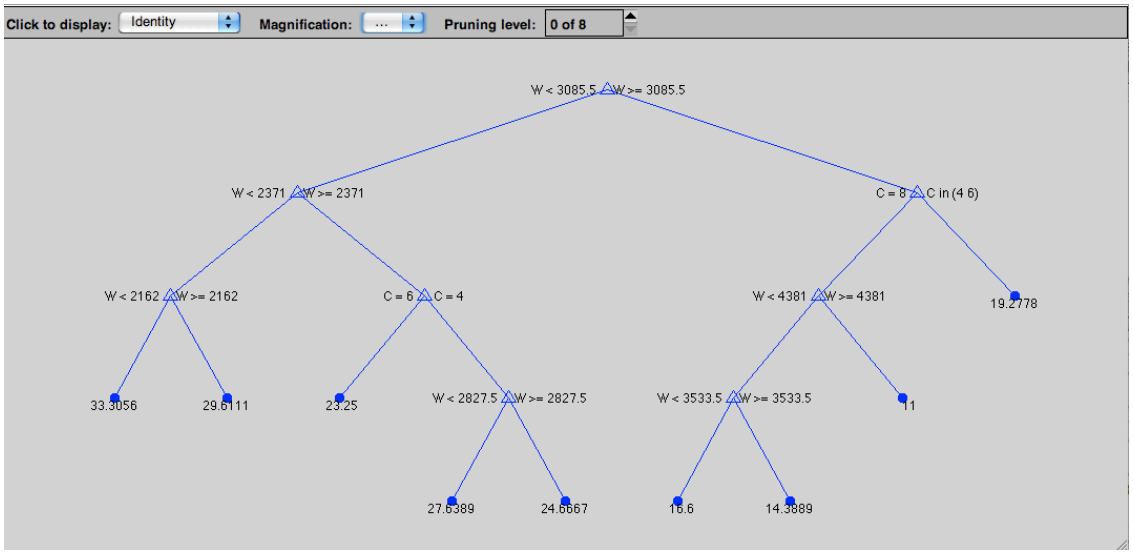
- Primer:
  - naredimo regresijsko drevo, kjer na na podlagi teže avtomobila (Weight, zvezna spr.) in števila cilindrov (Cylinders, kategorijska spr.) napovedujemo porabo goriva (MPG, zvezna spr.).

```
load carsmall

t = classregtree([Weight, Cylinders],MPG,...
                 'cat',2,'splitmin',20,...
                 'names',{'W','C'})
>>
t =
Decision tree for regression
 1  if W<3085.5 then node 2 elseif W>=3085.5 then node 3 else 23.7181
 2  if W<2371 then node 4 elseif W>=2371 then node 5 else 28.7931
 3  if C=8 then node 6 elseif C in {4 6} then node 7 else 15.5417
 4  if W<2162 then node 8 elseif W>=2162 then node 9 else 32.0741
 5  if C=6 then node 10 elseif C=4 then node 11 else 25.9355
 6  if W<4381 then node 12 elseif W>=4381 then node 13 else 14.2963
 7  fit = 19.2778
 8  fit = 33.3056
 9  fit = 29.6111
10  fit = 23.25
11  if W<2827.5 then node 14 elseif W>=2827.5 then node 15 else 27.2143
12  if W<3533.5 then node 16 elseif W>=3533.5 then node 17 else 14.8696
13  fit = 11
14  fit = 27.6389
15  fit = 24.6667
16  fit = 16.6
17  fit = 14.3889
```

Regresijsko drevo

- Primer:

```
view(t)
```

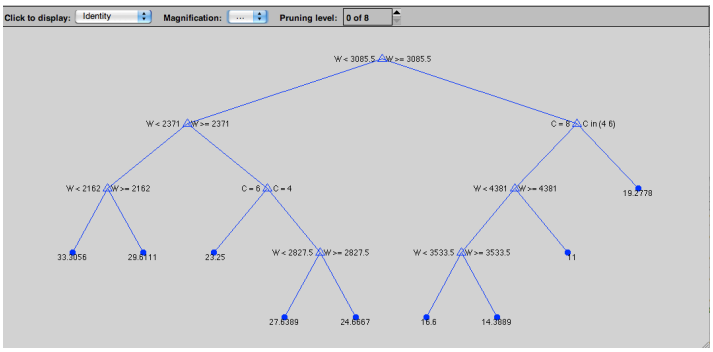


Regresijsko drevo

- Primer:

```
>> mileage2K = t([2000 4; 2000 6; 2000 8])

mileage2K =

   33.3056
   33.3056
   33.3056
```

```
>> mileage4K = t([4000 4; 4000 6; 4000 8])

mileage4K =

   19.2778
   19.2778
   14.3889
```




```
var3 = cutvar(t,3) % Po kateri spr. se delijo podatki v vozliscu st. 3.
>>
var3 = 'C'
```

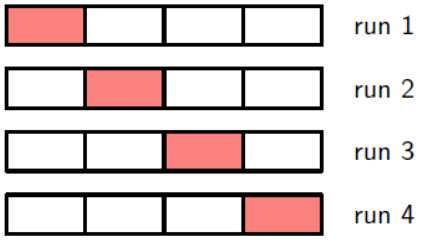
```
type3 = cuttype(t,3) % Kaksen tip delitve je?
>>
type3 = 'categorical'
```

```
c = cutcategories(t,3) % Kateri razredi so v levem otroku in kateri v desnem?
>> c{1}
ans =  8

>> c{2}
ans =
     4     6
```

Regresijsko drevo in navzkrižni test

- Z regresijskim drevesom lahko zelo dobro modeliramo učne podatke, vendar imamo lahko zelo veliko težav s podatki, ki jih moramo napovedati, pa niso vključeni v učenje parametrov:
  - regresijsko drevo je zelo občutljivo na odstopajoče točke (outlier-je).

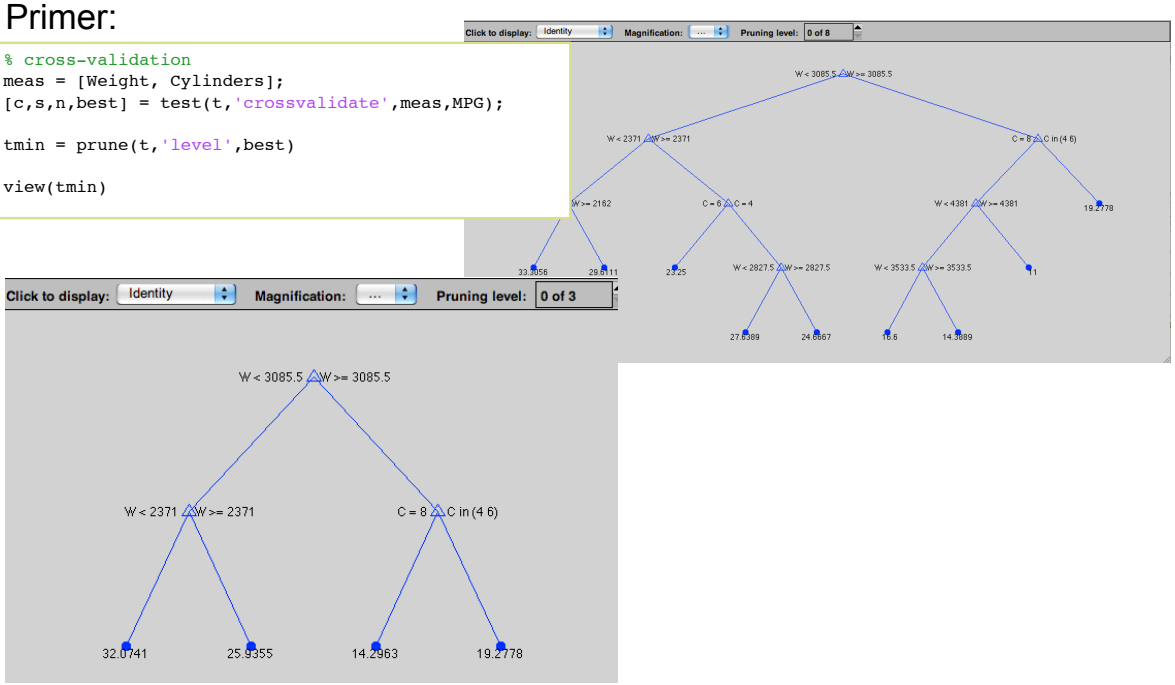
- Zato ponavadi učimo drevo s t.i. postopkom navzkrižnega testiranja (ang. cross validation):
  - podatke razdelimo na n približno enakih delov (npr. 10)
  - potem pa za vsak del naredimo test tako, da naučimo drevo na ostalih delih in testiramo predikcijo na tem delu.
  - cena:
    - povprečna kvadratna napaka med napovedanimi in dejanskimi vrednostmi v posameznem vozlišču.



Regresijsko drevo in navzkrižni test

- Primer:

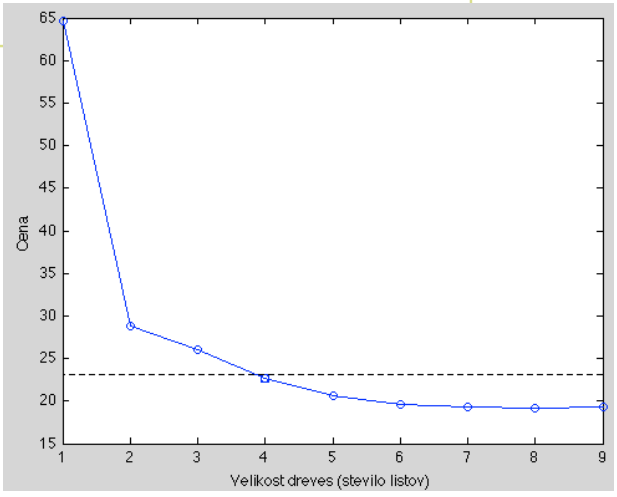
```
% cross-validation
meas = [Weight, Cylinders];
[c,s,n,best] = test(t,'crossvalidate',meas,MPG);

tmin = prune(t,'level',best)

view(tmin)
```



Regresijsko drevo in navzkrižni test

- Primer: Katero drevo izbrati?

```
[mincost,minloc] = min(c);
plot(n,c,'b-o',...
     n(best+1),c(best+1),'bs',...
     n,(mincost+s(minloc))*ones(size(n)),'k--')
xlabel('Velikost dreves (stevilo listov)')
ylabel('Cena')
```



6 Analiza variance in analiza kovariance

Pri analizi variance se ukvarjamo z ugotavljanjem, ali vzorci izhajajo iz ene populacije ali iz več različnih populacij glede na ocenjene vrednosti variance. Opisne spremenljivke v tem primeru so kategorijske spremenljivke, ki jih imenujemo tudi faktorji. Glede na stopnje znotraj faktorjev in glede na število faktorjev poznamo različne postopke analize variance, ki jih podrobno predstavimo v tem poglavju.

V poglavju se ukvarjamo tudi z analizo kovariance, ki vključuje obravnavo odzivnih in opisnih spremenljivk, ki so lahko zvezne in/ali kategorijske spremenljivke. Analiza medsebojne odvisnosti zato vključuje postopke iz analize variance in regresijske analize.

Poglavje obravnava:

- **Analiza variance - ANOVA:**
  - enosmerna ANOVA:
    - analiza vpliva posamičnih stopenj
  - več-smerna ANOVA:
    - dvosmerna ANOVA,
    - splošni model analize variance vpliva več faktorjev.
  - neparametrični testi analize variance.
- **Analiza kovariance - ANCOVA**

Analiza variance

- Analizo variance uporabljamo, ko so opisne spremenljivke kategorijske spremenljivke (npr. barvo, spol, ...).
- V tem primeru rečemo opisnim spremenljivkam **faktorji**.
- V vsakem faktorju imamo na izbiro dve ali več možnosti – stopnje.
- Če imamo samo en faktor z:
  - dvema stopnjama: Studentov t-test (test enakih povprečij) ali F-test (test enake variance)
  - s tremi ali več stopnjami: enosmerna ANOVA
- Če imamo več faktorjev:
  - dvosmerna, večsmerna ANOVA (odvisno od števila opisnih spremenljivk)
- Faktorski poskus:
  - preučujemo medsebojno odvisnost vseh faktorjev vpliva

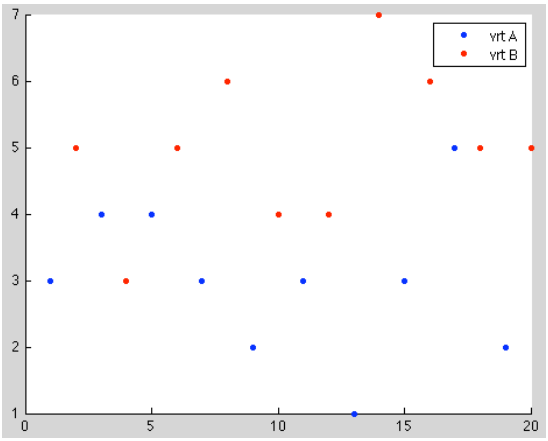
Enosmerna analiza variance

- Analiza variance spr. y (zvezna), ki je odvisna od ene opisne spremenljivke x, ki je kategorija – 1 faktor.
- Vrednosti x lahko izbiramo med dvema ali več stopnjami.
- Kaj hočemo narediti?
  - Ugotoviti želimo, ali se ocenjena povprečja za vsako stopnjo faktorja statistično značilno razlikujejo, ali ne.
  - To storimo s pomočjo analize variance po posameznih stopnjah.

Enosmerna analiza variance

- Primer:

```
fid = fopen('datasets/oneway.gardens.txt', 'r');
data = textscan(fid, '%d %s', 'HeaderLines', 1);
fclose(fid);

vrt = data{2};
ozon = data{1};


hold on;
num_obs = 1:20;
plot(num_obs(vrt=='A'), ozon(vrt=='A'), 'b.')
plot(num_obs(vrt=='B'), ozon(vrt=='B'), 'r.')
legend('vrt A', 'vrt B')
hold off
```



tip vrta meritve

dve stopnji →

```
vrt A   3   4   4   3   2   3   1   3   5   2
vrt B   5   5   6   7   4   4   3   5   6   5
```

Enosmerna analiza variance

```
vrt A   3   4   4   3   2   3   1   3   5   2
vrt B   5   5   6   7   4   4   3   5   6   5
```

Treatment	Observations				Totals	Averages
1	y_{11}	y_{12}	$\cdots$	y_{1n}	$y_{1\cdot}$	$\bar{y}_{1\cdot}$
2	y_{21}	y_{22}	$\cdots$	y_{2n}	$y_{2\cdot}$	$\bar{y}_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots\vdots\vdots$	$\vdots$	$\vdots$	$\vdots$
a	y_{a1}	y_{a2}	$\cdots$	y_{an}	$y_{a\cdot}$	$\bar{y}_{a\cdot}$
					$y_{\cdot\cdot}$	$\bar{y}_{\cdot\cdot}$

- Modeliramo povprečja:

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij} \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, n \end{cases} \qquad \epsilon_{ij} \sim N(0, \sigma^2)$$

Enosmerna analiza variance

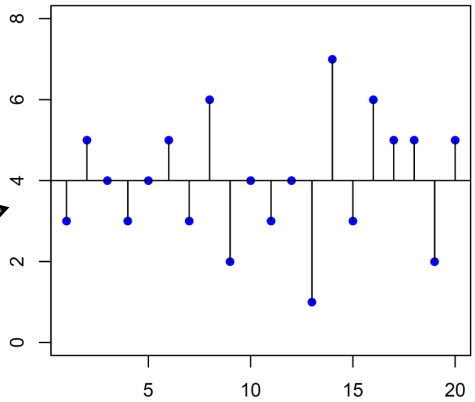
- Ničta hipoteza: $H_0: \tau_1 = \tau_2 = \cdots = \tau_a = 0$
- Alternativna hipoteza: $H_1: \tau_i \neq 0$ za vsaj en i
- Predpostavka: $\sum_{i=1}^{a} \tau_i = 0$

- Kaj to pomeni?
  - Če velja ničta hipoteza, potem lahko vsako meritev sestavimo iz globalnega povprečja in napake ocene epsilon.
  - To pomeni: da se meritve porazdeljuejo po normalni porazdelitvi s povprečjem μ in varianco σ^2.
  - Z drugimi besedami, različne stopnje v faktorju nimajo vpliva na končne rezultate.

Enosmerna analiza variance

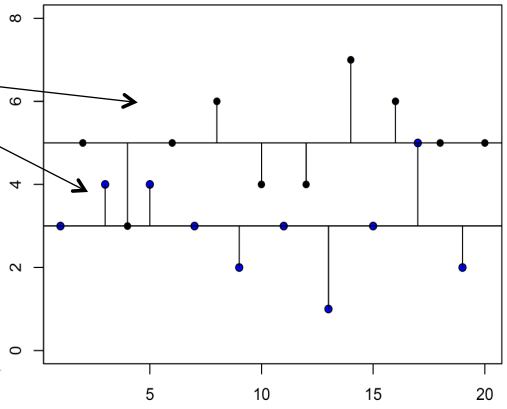
- Pomembne oznake:
  - Vsota kvadratov razlik glede na skupno povprečje:

$$SSY = \sum (y - \overline{y})^2$$



  - Vsota kvadratov razlik glede na posamična povprečja:

$$SSE = \sum_{j=1}^{a} \sum (y - \overline{y}_j)^2$$



  - Razlika obeh vsot:

$$SSA = SSY - SSE$$

Enosmerna ANOVA: statistike

- Matematično upanje SSY:

$$\mathbb{E}(SSY) = (a-1)\sigma^2 + n\sum_{i=1}^{a}\tau_i^2$$

- Matematično upanje SSE:

$$\mathbb{E}(SSE) = a(n-1)\sigma^2$$

- Prostostne stopnje:
  - SSY: an – 1 SSE: a(n – 1)
- Povprečja vsote kvadratov: vsota kvadratov/ prostostne stopnje

$$MSA = \frac{SSA}{a-1} \qquad MSE = \frac{SSE}{a(n-1)}$$

- F-statistika:

$$F0 = \frac{MSA}{MSE}$$

Tabela enosmerne ANOVA

	vsota kvadratov razlik	prostostne stopnje	povprečje kvadratov razlik	F0
faktor	SSA	a-1	MSA = SSA/(a-1)	MSA/MSE
napake	SSE	a(n-1)	MSE = SSE/(a(n-1))	
skupaj	SSY	an-1		

- F0 primerjamo z mejno vrednostjo F-porazdelitve pri tveganju napačne zavrnitve, določenim z alfa.
- Če je F0 večji od mejne vrednosti, potem lahko zavrnemo ničto hipotezo, da so povprečja po stopnjah enaka 0.
- Dodatno lahko izračunamo tudi verjetnost (p-vrednost), da ob predpostavki ničte hipoteze dobimo vrednost F0.

Tabela enosmerne ANOVA

	vsota kvadratov razlik	prostostne stopnje	povprečje kvadratov razlik	kvocient F
faktor	20.0	1	20.0	15.0
napake	24.0	18	1.3333	
skupaj	44.0	19		

```
SSY = sum((ozon - mean(ozon)).^2)

SSE1 = sum( (ozon(vrt=='A')  - mean(ozon(vrt=='A')) ).^2  )
SSE2 = sum( (ozon(vrt=='B')  - mean(ozon(vrt=='B')) ).^2  )

SSE = SSE1 + SSE2
SSA = SSY - SSE

MSA = SSA / 1
MSE = SSE / 18

F0 = MSA/MSE
```

```
>>
SSY = 44
SSE1 = 12
SSE2 = 12
SSE = 24

SSA = 20

MSA = 20
MSE = 1.3333

F0 = 15
```

Tabela enosmerne ANOVA

	vsota kvadratov razlik	prostostne stopnje	povprečje kvadratov razlik	kvocient F
faktor	20.0	1	20.0	15.0
napake	24.0	18	1.3333	
skupaj	44.0	19		

```
finv(0.95, 1, 18)   % mejna vrednost za zavrnitev hipoteze

p = fcdf(15.0, 1, 18);
min(p, 1-p)
>>
ans = 4.4139

ans = 0.0011
```

mejna vrednost

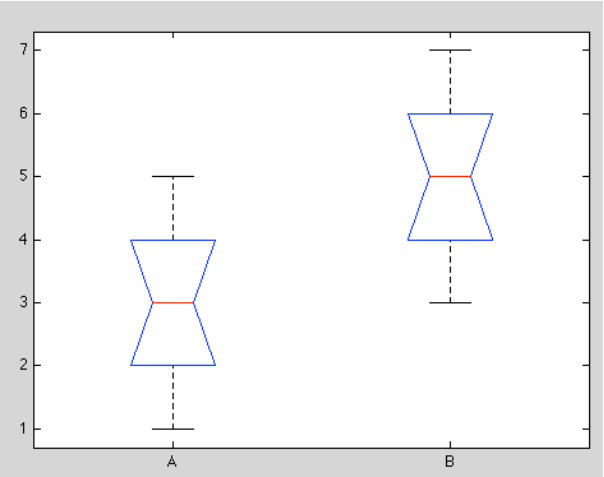
verjetnost, da dobimo 15.0 ob predpostavki ničte hipoteze

ANOVA v MATLABu

```
[p,table,stats] = anova1(ozon, vrt);
```

ANOVA Table

Source	SS	df	MS	F	Prob>F
Groups	20	1	20	15	0.0011
Error	24	18	1.33333		
Total	44	19			



Analiza vpliva posamičnih stopenj

- Primer:
  - meritve količine bakterij v posameznih pošiljkah mleka

```
>> load hogg
hogg
    24    14    11     7    19
    15     7     9     7    24
    21    12     7     4    19
    27    17    13     7    15
    33    14    12    12    10
    23    16    18    18    20
```

    - vrstice predstavljajo količino bakterij v posameznem litru mleka, ki smo ga naključno izbrali iz pošiljke: imamo 6 pošiljk in 5 naključno izbranih meritev v posamezni pošiljki.

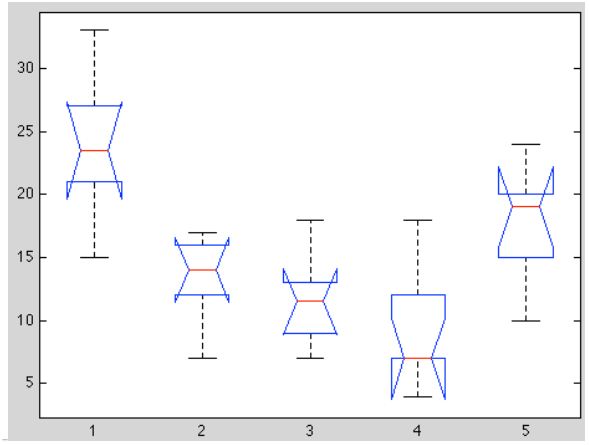
  - Zanima nas,
    - ali kakšna pošiljka statistično značilno odstopa po številu bakterij od ostalih
    - in če odstopa, katera pošiljka je to?

Analiza vpliva posamičnih stopenj

- 1. vprašanje: Ali kakšna pošiljka statistično značilno odstopa po številu bakterij od ostalih?
- Eno-smerna ANOVA: `[p,tbl,stats] = anova1(hogg);`

ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	803	4	200.75	9.01	0.0001
Error	557.17	25	22.287		
Total	1360.17	29			



Analiza vpliva posamičnih stopenj

- 2. vprašanje: Katere pošiljke se med seboj statistično značilno razlikujejo?
- Primerjamo s t-testi povprečja paroma med posameznimi pošiljkami.
  A je to v redu? NE:
  - Če s t-testom primerjamo povprečja dveh vzorcev, je v primeru, da sta povprečja enaka verjetnost, da bo t-statistika presegla mejne vrednosti zelo mala (recimo 5%). V primeru, da sta povprečja različna pa je verjetnost, da bo t-statistika presegla mejno vrednost zelo velika.
  - V našem primeru imamo 5 pošiljk, torej bi naredili 10 primerjav med povprečji. V primeru, da nimamo razlik med povprečji (povprečja so enaka) in da je alfa 5%, potem lahko pri vsakem testu naredimo v 5% napako, da zavrnemo hipotezo, da sta povprečja enaka. To lahko storimo v vsakem izmed 10-ih testov. To pa pomeni, da je možnost napake, da vsaj enkrat napačno zavrnemo hipotezo o enakih povprečjih mnogo večja od zahtevanih 5% (torej ne moremo govoriti o statistični značilnosti).
- Zato obstajajo drugačni načini testiranja:
  - postopki medsebojnih večkratnih primerjav (*ang. multiple comparison methods*)

Ena možnost: postopek LSD (*ang. least significant difference*)

- Ničta hipoteza: $\mu_i = \mu_j$
- Statistika t: $t = \frac{\text{razlika povprečij}}{\text{standardna napaka razlike}}$ $\quad t_0 = \frac{\bar{y}_{i\cdot} - \bar{y}_{j\cdot}}{\sqrt{\frac{2MS_E}{n}}}$
- Alternativna hipoteza: dva povprečja bosta signifikantno različna, če bo veljalo:

$$|\bar{y}_{i\cdot} - \bar{y}_{j\cdot}| > \text{LSD}$$

- Statistika LSD: $LSD = t_{1-\alpha/2,a(n-1)}\sqrt{\frac{2MSE}{n}}$
  - v primeru različno velikih vzorcev $LSD = t_{1-\alpha/2,N-a}\sqrt{MSE(\frac{1}{n_1} + \frac{1}{n_2})}$

Postopek LSD

- Primer: pošiljke mleka

```
[p,tbl,stats] = anova1(hogg);
>> tbl
    'Source'     'SS'            'df'    'MS'          'F'         'Prob>F'
    'Columns'    [  803.0000]    [ 4]    [200.7500]    [9.0076]    [1.1971e-04]
    'Error'      [  557.1667]    [25]    [ 22.2867]          []            []
    'Total'      [1.3602e+03]    [29]            []          []            []

>> stats
stats =
    gnames: [5x1 char]
         n: [6 6 6 6 6]
    source: 'anova1'
     means: [23.8333 13.3333 11.6667 9.1667 17.8333]
        df: 25
         s: 4.7209
```

MSE

število meritev za vsako stopnjo

povprečja za vsako stopnjo

razlike med povprečij

```
abs(stats.means(1) - stats.means(2)) = 10.50
abs(stats.means(1) - stats.means(3)) = 12.17
abs(stats.means(1) - stats.means(4)) = 14.67
abs(stats.means(1) - stats.means(5)) = 6.00
abs(stats.means(2) - stats.means(3)) = 1.67
abs(stats.means(2) - stats.means(4)) = 4.17
abs(stats.means(2) - stats.means(5)) = 4.50
abs(stats.means(3) - stats.means(4)) = 2.50
abs(stats.means(3) - stats.means(5)) = 6.17
abs(stats.means(4) - stats.means(5)) = 8.67
```

Statistika LSD

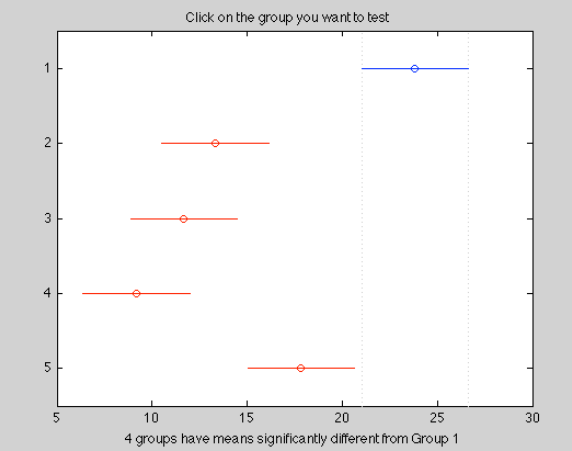
```
LSD = tinv(0.975, 25) * sqrt(2*22.2867/6)
>> LSD = 5.6135
```

Vse razlike niso večje od LSD!!!

Analiza vpliva posamičnih stopenj

- Multicompare (MATLAB)

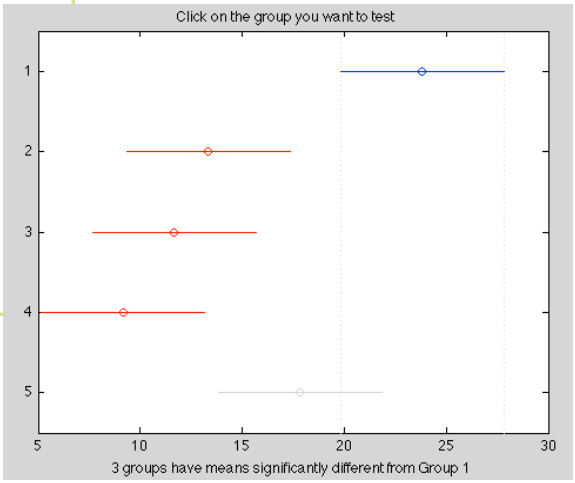
```
[p,tbl,stats] = anova1(hogg);

[c,m] = multcompare(stats, 'ctype', 'lsd')
c =
    1.0000    2.0000    4.8865   10.5000   16.1135
    1.0000    3.0000    6.5532   12.1667   17.7801
    1.0000    4.0000    9.0532   14.6667   20.2801
    1.0000    5.0000    0.3865    6.0000   11.6135
    2.0000    3.0000   -3.9468    1.6667    7.2801
    2.0000    4.0000   -1.4468    4.1667    9.7801
    2.0000    5.0000  -10.1135   -4.5000    1.1135
    3.0000    4.0000   -3.1135    2.5000    8.1135
    3.0000    5.0000  -11.7801   -6.1667   -0.5532
    4.0000    5.0000  -14.2801   -8.6667   -3.0532

m =
   23.8333    1.9273
   13.3333    1.9273
   11.6667    1.9273
    9.1667    1.9273
   17.8333    1.9273
```



Analiza vpliva posamičnih stopenj

- Multicompare: drugi kriterij primerjave (MATLAB)

```
>> [c,m] = multcompare(stats)
c =
    1.0000    2.0000    2.4953   10.5000   18.5047
    1.0000    3.0000    4.1619   12.1667   20.1714
    1.0000    4.0000    6.6619   14.6667   22.6714
    1.0000    5.0000   -2.0047    6.0000   14.0047
    2.0000    3.0000   -6.3381    1.6667    9.6714
    2.0000    4.0000   -3.8381    4.1667   12.1714
    2.0000    5.0000  -12.5047   -4.5000    3.5047
    3.0000    4.0000   -5.5047    2.5000   10.5047
    3.0000    5.0000  -14.1714   -6.1667    1.8381
    4.0000    5.0000  -16.6714   -8.6667   -0.6619

m =
   23.8333    1.9273
   13.3333    1.9273
   11.6667    1.9273
    9.1667    1.9273
   17.8333    1.9273
```



Analiza vpliva posamičnih stopenj

- Multicompare: kriteriji primerjave (MATLAB)

Value	Description
`'hsd'` or `'tukey-kramer'`	Use Tukey's honestly significant difference criterion. This is the default, and it is based on the Studentized range distribution. It is optimal for balanced one-way ANOVA and similar procedures with equal sample sizes. It has been proven to be conservative for one-way ANOVA with different sample sizes. According to the unproven Tukey-Kramer conjecture, it is also accurate for problems where the quantities being compared are correlated, as in analysis of covariance with unbalanced covariate values.
`'lsd'`	Use Tukey's least significant difference procedure. This procedure is a simple t-test. It is reasonable if the preliminary test (say, the one-way ANOVA *F* statistic) shows a significant difference. If it is used unconditionally, it provides no protection against multiple comparisons.
`'bonferroni'`	Use critical values from the t distribution, after a Bonferroni adjustment to compensate for multiple comparisons. This procedure is conservative, but usually less so than the Scheffé procedure.
`'dunn-sidak'`	Use critical values from the t distribution, after an adjustment for multiple comparisons that was proposed by Dunn and proved accurate by Sidák. This procedure is similar to, but less conservative than, the Bonferroni procedure.
`'scheffe'`	Use critical values from Scheffé's S procedure, derived from the F distribution. This procedure provides a simultaneous confidence level for comparisons of all linear combinations of the means, and it is conservative for comparisons of simple differences of pairs.

Več-smerna ANOVA in faktorski poskus

- Večsmerna ANOVA izvaja analizo variance v primeru, ko imamo več opisnih spremenljivk, ki so kategorične narave – imamo več faktorjev.
- Ti faktorji imajo lahko dva ali več opisnih stopenj.
- Pri faktorskem poskusu preučujemo **vpliv posameznih stopenj** faktorjev in **vpliv medsebojne interakcije stopenj** faktorjev na rezultate poskusa.
- Pri enosmerni ANOVI smo preučevali vpliv enega faktorja (z več stopnjami) na končne rezultate poskusa. Interakcij zato ni bilo.

Tabela faktorskega poskusa z dvema faktorjema

- Denimo, da imamo dva faktorja A in B:
  - v faktorju A imamo *a* stopenj, v faktorju B imamo *b* stopenj
  - imamo n ponovitev poskusa, pri vsakem poskusu pridobimo meritve vseh kombinacij, torej *ab* meritev

		Factor *B*					
		1	2	⋯	*b*	Totals	Averages
Factor *A*	1	$y_{111}, y_{112}, \dots, y_{11n}$	$y_{121}, y_{122}, \dots, y_{12n}$		$y_{1b1}, y_{1b2}, \dots, y_{1bn}$	$y_{1\cdot\cdot}$	$\bar{y}_{1\cdot\cdot}$
	2	$y_{211}, y_{212}, \dots, y_{21n}$	$y_{221}, y_{222}, \dots, y_{22n}$		$y_{2b1}, y_{2b2}, \dots, y_{2bn}$	$y_{2\cdot\cdot}$	$\bar{y}_{2\cdot\cdot}$
	⋮						
	a	$y_{a11}, y_{a12}, \dots, y_{a1n}$	$y_{a21}, y_{a22}, \dots, y_{a2n}$		$y_{ab1}, y_{ab2}, \dots, y_{abn}$	$y_{a\cdot\cdot}$	$\bar{y}_{a\cdot\cdot}$
Totals		$y_{\cdot 1\cdot}$	$y_{\cdot 2\cdot}$		$y_{\cdot b\cdot}$	$y_{\cdots}$	
Averages		$\bar{y}_{\cdot 1\cdot}$	$\bar{y}_{\cdot 2\cdot}$		$\bar{y}_{\cdot b\cdot}$		$\bar{y}_{\cdots}$

y_{ijk} pomeni k-to meritev stopnje i faktorja A in stopnje j faktorja B

Statistična analiza faktorskega poskusa z dvema faktorjema

- Pri poskusu imamo tako *abn* meritev.
- Meritve lahko modeliramo z linearnim modelom:

$$Y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \epsilon_{ijk} \begin{cases} i = 1, 2, \dots, a \\ j = 1, 2, \dots, b \\ k = 1, 2, \dots, n \end{cases}$$

  - kjer je μ dejansko globalno povprečje, τ_i je prispevek i-te stopnje faktorja A k pvoprečju, β_j je prispevek j-te stopnje k povprečju $(\tau\beta)_{ij}$ je prispevek interakcije med faktorjem A in B, ε_{ijk} pa je naključna napaka, ki je standardno normalno porazdeljena s povprečjem 0 in varianco σ^2.

- Zanimajo nas vplivi posameznih faktorjev A in B ter njune interakcije na končne rezultate (meritve) Y.
  - **Ničta hipoteza:** faktor A nima vpliva, faktor B nima vpliva, interakcija A in B nima vpliva

- Ker imamo dva faktorja, imenujemo to analizo tudi dvosmerna ANOVA.

Statistična analiza faktorskega poskusa z dvema faktorjema

- Fiksno določena faktorja A in B: stopnje faktorja A in faktorja B so vnaprej določene in jih ne izbiramo naključno iz faktorjev.
- V tem primeru lahko privzamemo:

$$\sum_{i=1}^{a}\tau_i = 0, \sum_{j=1}^{b}\beta_j = 0, \sum_{i=1}^{a}(\tau\beta)_{ij} = 0 \quad \sum_{j=1}^{b}(\tau\beta)_{ij} = 0$$

- Analiza variance: preverjamo vpliv faktorja A in B ter interakcije obeh na končne rezultate.
- Hipoteze, ki jih bomo testirali:

1. $H_0: \tau_1 = \tau_2 = \cdots = \tau_a = 0$ faktor A nima vpliva
   $H_1:$ vsaj en $\tau_i \neq 0$
2. $H_0: \beta_1 = \beta_2 = \cdots = \beta_b = 0$ faktor B nima vpliva
   $H_1:$ vsaj en $\beta_j \neq 0$
3. $H_0: (\tau\beta)_{11} = (\tau\beta)_{12} = \cdots = (\tau\beta)_{ab} = 0$ interakcija nima vpliva
   $H_1:$ vsaj en $(\tau\beta)_{ij} \neq 0$

Dvosmerna ANOVA

- Definiramo vsote:

$$y_{i\cdot\cdot} = \sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk} \qquad \bar{y}_{i\cdot\cdot} = \frac{y_{i\cdot\cdot}}{bn} \qquad i = 1, 2, \ldots, a$$

$$y_{\cdot j\cdot} = \sum_{i=1}^{a}\sum_{k=1}^{n} y_{ijk} \qquad \bar{y}_{\cdot j\cdot} = \frac{y_{\cdot j\cdot}}{an} \qquad j = 1, 2, \ldots, b$$

$$y_{ij\cdot} = \sum_{k=1}^{n} y_{ijk} \qquad \bar{y}_{ij\cdot} = \frac{y_{ij\cdot}}{n} \qquad \begin{matrix} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, b \end{matrix}$$

$$y_{\cdots} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk} \qquad \bar{y}_{\cdots} = \frac{y_{\cdots}}{abn}$$

- Vsote kvadratov:

$$SS_T = SS_A + SS_B + SS_{AB} + SS_E$$

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}^2 - \frac{y_{\cdots}^2}{abn}$$

$$SS_A = \sum_{i=1}^{a}\frac{y_{i\cdot\cdot}^2}{bn} - \frac{y_{\cdots}^2}{abn}$$

$$SS_B = \sum_{j=1}^{b}\frac{y_{\cdot j\cdot}^2}{an} - \frac{y_{\cdots}^2}{abn}$$

$$SS_{AB} = \sum_{i=1}^{a}\sum_{j=1}^{b}\frac{y_{ij\cdot}^2}{n} - \frac{y_{\cdots}^2}{abn} - SS_A - SS_B$$

Prostostne stopnje pri dvosmerni ANOVI

- Prostostne stopnje faktorja SS_A: a – 1
- Prostostne stopnje faktorja SS_B: b – 1
- Prostostne stopnje interakcije SS_{AB}: (a – 1)(b – 1)
- Prostostne stopnje SS_T: abn – 1
- Prostostne stopnje napake SSE: ab(n – 1)
  - znotraj vsake celice imamo n ponovitev meritev, torej je v vsaki celici št. prostostnih stopenj n – 1. Skupaj imamo *ab* celic, torej je št. vseh prostostnih stopenj za napako ab(n – 1)

- Velja zveza: abn-1 = (a-1) +(b-1)+(a-1)(b-1)+ab(n-1)

F – statistike pri 2-smerni ANOVI

- Izračunamo povprečja vsot kvadratov:

$$MS_A = \frac{SS_A}{a-1} \qquad MS_B = \frac{SS_B}{b-1} \qquad MS_{AB} = \frac{SS_{AB}}{(a-1)(b-1)} \qquad MS_E = \frac{SS_E}{ab(n-1)}$$

- Za preverjanje hipotez izračunajmo še matematična upanja:

$$E(MS_A) = E\left(\frac{SS_A}{a-1}\right) = \sigma^2 + \frac{bn\sum_{i=1}^{a}\tau_i^2}{a-1} \qquad E(MS_B) = E\left(\frac{SS_B}{b-1}\right) = \sigma^2 + \frac{an\sum_{j=1}^{b}\beta_j^2}{b-1}$$

$$E(MS_{AB}) = E\left(\frac{SS_{AB}}{(a-1)(b-1)}\right) = \sigma^2 + \frac{n\sum_{i=1}^{a}\sum_{j=1}^{b}(\tau\beta)_{ij}^2}{(a-1)(b-1)}$$

$$E(MS_E) = E\left(\frac{SS_E}{ab(n-1)}\right) = \sigma^2$$

F – statistike pri 2-smerni ANOVI

- Hipoteza 1 (vpliv faktorja A): $H_0: \tau_1 = \tau_2 = \cdots = \tau_a = 0$

$$F_0 = \frac{MS_A}{MS_E}$$

- Hipoteza 2 (vpliv faktorja B): $H_0: \beta_1 = \beta_2 = \cdots = \beta_b = 0$

$$F_0 = \frac{MS_B}{MS_E}$$

- Hipoteza 3 (vpliv interakcije AB): $H_0: (\tau\beta)_{11} = (\tau\beta)_{12} = \cdots = (\tau\beta)_{ab} = 0$

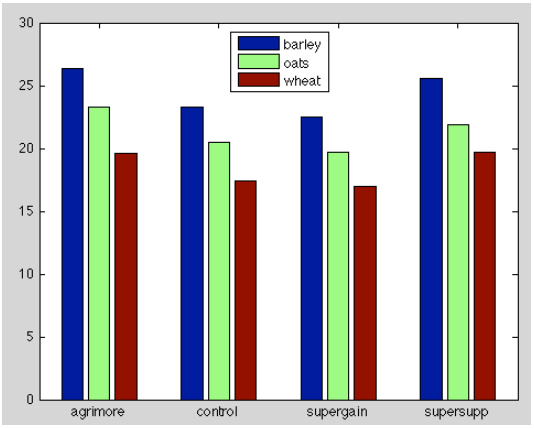
$$F_0 = \frac{MS_{AB}}{MS_E}$$

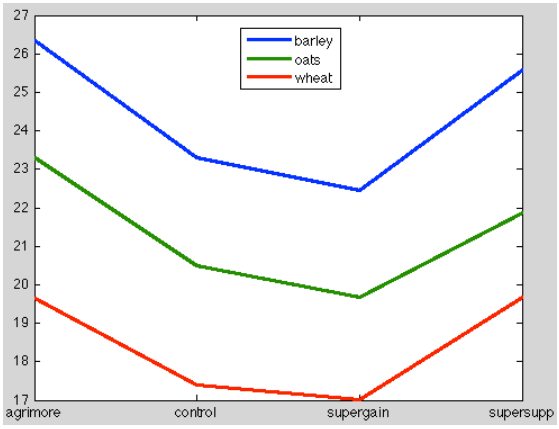
Tabela za dvosmerno ANOVO

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
A treatments	SS_A	$a - 1$	$MS_A = \frac{SS_A}{a-1}$	$\frac{MS_A}{MS_E}$
B treatments	SS_B	$b - 1$	$MS_B = \frac{SS_B}{b-1}$	$\frac{MS_B}{MS_E}$
Interaction	SS_{AB}	$(a-1)(b-1)$	$MS_{AB} = \frac{SS_{AB}}{(a-1)(b-1)}$	$\frac{MS_{AB}}{MS_E}$
Error	SS_E	$ab(n-1)$		
Total	SS_T	$abn - 1$	$MS_E = \frac{SS_E}{ab(n-1)}$	

Primer dvosmerne ANOVE

- Primer: preučujemo vpliv prehrane na rast živali
  - Imamo dva faktorja vpliva (najbolj enostaven primer):
    - **vrsta prehrane**: stopnje: *ječmen (barley), oves (oats), pšenica (wheat)*
    - **dodatki k prehrani**: stopnje (vrste dodatkov): *agrimore, control, supergain, supersupp*
  - **Odzivna spremenljivka**: *teža živali po 6 tednih hranjenja.*





Primer dvosmerne ANOVE

- Preučujemo vpliv prehrane na rast živali
  - Imamo dva faktorja vpliva (najbolj enostaven primer):
    - **faktor A: vrsta prehrane**: stopnje: *ječmen (barley), oves (oats), pšenica (wheat)*
    - **faktor B: dodatki k prehrani**: stopnje (vrste dodatkov): *agrimore, control, supergain, supersupp*
  - **Odzivna spremenljivka**: *teža živali po 6 tednih hranjenja.*

```
load datasets/growth.mat

factor1_level = unique(ordinal(diet))
factor2_level = unique(ordinal(supplement))
>>
factor1_level =
     barley
     oats
     wheat

factor2_level =
     agrimore
     control
     supergain
     supersupp
```

Tabela faktorskega poskusa dveh faktorjev

	vsota kvadratov razlik	pr. stopnje	povprečje kvadratov razlik	F0	p
faktor A					
faktor B					
interakcija AB					
napake					
skupaj					

```
>> SY = zeros(3,4);
for i = 1 : length(factor1_level)
    for j =  1 : length(factor2_level)
        SY(i,j) = sum( gain(ordinal(diet) == factor1_level(i
                    & ordinal(supplement)==factor2_level(j)

    end
end

>> SY
        agrimore  control supergain supersupp
barley 105.39391 93.18660  89.86449 102.30121
oats    93.19354 81.97466  78.65201  87.44094
wheat   78.55629 69.62207  68.04973  78.67336
```

```
%vsote po faktorjih A
YA1 = sum(SY(1,:));
YA2 = sum(SY(2,:));
YA3 = sum(SY(3,:));
%vsote po faktorijh B
YB1 = sum(SY(:,1));
YB2 = sum(SY(:,2));
YB3 = sum(SY(:,3));
YB4 = sum(SY(:,4));
%vsote po interakciji
YAB = SY;
%skupna vsota
YT = YB1+YB2+YB3+YB4;
```

Tabela faktorskega poskusa dveh faktorjev

	vsota kvadratov razlik	pr. stopnje	povprečje kvadratov razlik	F0	p
faktor A	287.1711	2			
faktor B	91.88101	3			
interakcija AB	3.405791	6			
napake	61.89025	36			
skupaj	444.3481	47			

```
a = 3;
b = 4;
n = 4;
>> SST = sum(gain.^2) - YT.^2/(a*b*n)
SST = 444.3481
```

```
SST = sum(gain.^2) - YT.^2/(a*b*n)
SSA = (YA1.^2 + YA2.^2 + YA3.^2)/(b*n) - YT.^2/(a*b*n)
SSB = (YB1.^2 + YB2.^2 + YB3.^2 + YB4.^2)/(a*n) - YT.^2/(a*b*n)
SSAB = sum(sum(YAB.^2/n)) -  YT.^2/(a*b*n) - SSA - SSB
SSE = SST - SSA - SSB -SSAB
>>
SST = 444.3481
SSA = 287.1711
SSB = 91.8810
SSAB = 3.4058
SSE = 61.8902
```

Tabela faktorskega poskusa dveh faktorjev

	vsota kvadratov razlik	pr. stopnje	povprečje kvadratov razlik	F0	p
faktor A	287.1711	2	143.5855	83.5201	
faktor B	91.88101	3	30.62700	17.815	
interakcija AB	3.405791	6	0.5676318	0.33018	
napake	61.89025	36	1.719174		
skupaj	444.3481	47			

```
MSA = SSA/(a-1)
MSB = SSB/(b-1)
MSAB = SSAB/((a-1)*(b-1))
MSE = SSE/(a*b*(n-1))
>>
MSA = 143.5855
MSB =  30.6270
MSAB =  0.5676
MSE =   1.7192
```

```
FA0 = MSA/MSE
FB0 = MSB/MSE
FAB0 = MSAB/MSE
>>
FA0 = 83.5201
FB0 = 17.8150
FAB0 = 0.3302
```

Tabela faktorskega poskusa dveh faktorjev

	vsota kvadratov razlik	pr. stopnje	povprečje kvadratov razlik	F0	p
faktor A	287.1711	2	143.5855	83.5201	3.0e-14
faktor B	91.88101	3	30.62700	17.815	2.95e-07
interakcija AB	3.405791	6	0.5676318	0.33018	0.9166
napake	61.89025	36	1.719174		
skupaj	444.3481	47			

```
1-fcdf(FA0, a-1, a*b*(n-1))
1-fcdf(FB0, b-1, a*b*(n-1))
1-fcdf(FAB0, (a-1)*(b-1), a*b*(n-1))
>>
ans =
   2.9976e-14
ans =
   2.9519e-07
ans =
    0.9166
```

Tabela faktorskega poskusa dveh faktorjev

- V Matlabu:

```
X = zeros(12, 4);  %vse mozne kombinacije stopenj
                   %in stevilo meritev za vsako stopnjo
rn = 0;
for i = 1 : length(factor1_level)
    for j =  1 : length(factor2_level)
        rn = rn + 1;
        X(rn, :) = gain(ordinal(diet) == factor1_level(i)...
                   & ordinal(supplement)==factor2_level(j));
    end
end
% spremeniti moramo vrstni red za anovo v matlabu
Y = [X(1:4,:)'; X(5:8,:)'; X(9:12,:)' ]
[p, tbl, stats] = anova2(Y, 4)  %4 ponovitve za vsako kombinacijo stopenj
```

ANOVA Table

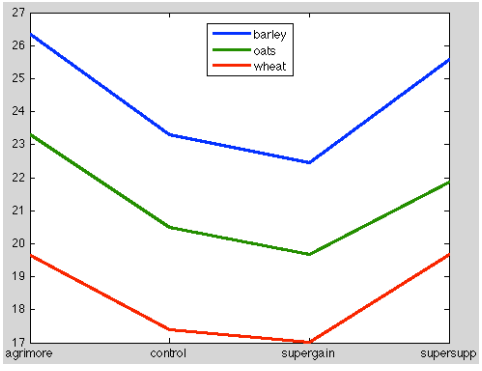
Source	SS	df	MS	F	Prob>F
Columns	91.881	3	30.627	17.81	0
Rows	287.171	2	143.586	83.52	0
Interaction	3.406	6	0.568	0.33	0.9166
Error	61.89	36	1.719		
Total	444.348	47			

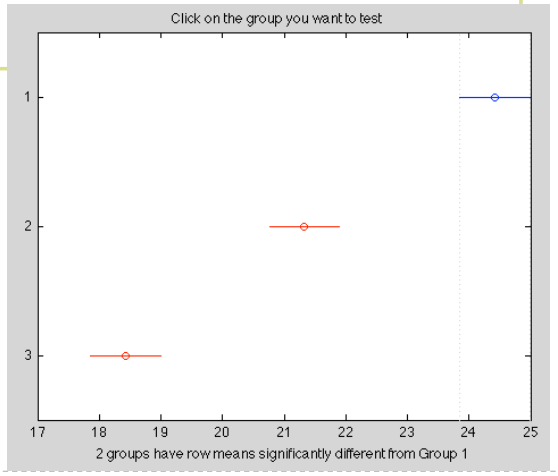
glede na rezultate, bi lahko modelirali
gain ~ diet + supplement brez interakcije

Analiza ustreznosti posameznih stopenj

Pregledamo še vse kombinacije stopenj faktorja A in B
in ugotavljamo statistično značilnost posameznih stopenj.

```
>> [c, m] = multcompare(stats, 'estimate', 'row')    %stopnje faktorja A (vrstice)
c =
    1.0000    2.0000    1.9597    3.0928    4.2259
    1.0000    3.0000    4.8572    5.9903    7.1234
    2.0000    3.0000    1.7644    2.8975    4.0306
m =
   24.4216    0.3278
   21.3288    0.3278
   18.4313    0.3278
```

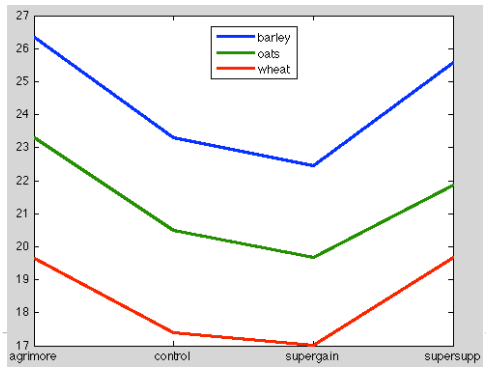


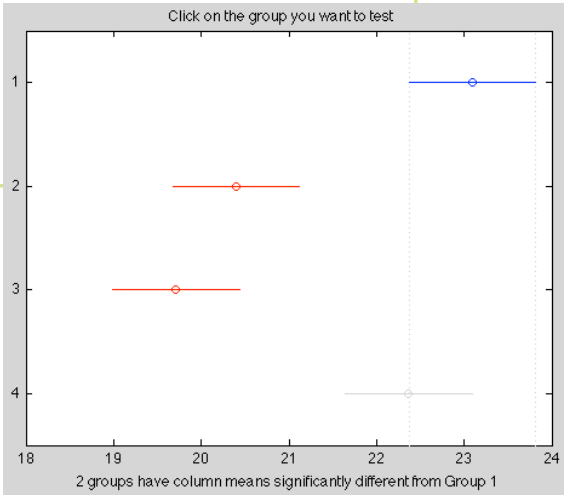


Analiza ustreznosti posameznih stopenj

Pregledamo še vse kombinacije stopenj faktorja A in B in ugotavljamo statistično značilnost posameznih stopenj.

```
>>[c, m] = multcompare(stats, 'estimate', 'column') %stopnje faktorja B (stolpci)

c =
    1.0000    2.0000    1.2551    2.6967    4.1383
    1.0000    3.0000    1.9398    3.3815    4.8231
    1.0000    4.0000   -0.7143    0.7274    2.1690
    2.0000    3.0000   -0.7569    0.6848    2.1264
    2.0000    4.0000   -3.4110   -1.9693   -0.5277
    3.0000    4.0000   -4.0957   -2.6541   -1.2125
m =
   23.0953    0.3785
   20.3986    0.3785
   19.7139    0.3785
   22.3680    0.3785
```





N-smerna ANOVA

- N-smerna ANOVA je posplošitev dvosmerne ANOVE.
- Za tri faktorje lahko zapišemo model:

$$y_{ijkl} = \mu + \alpha_{.j.} + \beta_{i..} + \gamma_{..k} + \alpha\beta_{ij.} + \alpha\gamma_{i.k} + \beta\gamma_{.jk} + \alpha\beta\gamma_{ijk} + \epsilon_{ijkl}$$

  - Parametri $\alpha\beta_{ij.}, \alpha\gamma_{i.k}, \beta\gamma_{.jk}$ predstavljajo interakcijo med dvema faktorjema. Parameter $\alpha\beta\gamma_{ijk}$ pa interakcijo med vsemi tremi faktorji.

- Faktorski poskus:
  - Model za ANOVO ima lahko vse možne parametre (vse faktorje in vse kombinacije faktorjev), lahko pa tudi samo določene. Npr. vključimo samo interakcije dveh faktorjev, interakcije vseh treh faktorjev pa ne upoštevamo.
  - Faktorski poskus:
    - postavitev modela (katere kombinacije faktorjev bomo upoštevali) in
    - pridobivanje meritev za ta model.

Matlab: ANOVAN

- Funkcija v Matlabu za n-smerno ANOVO: anovan
  - drugačna kot anova1 in anova2: nimamo podatkov v obliki tabele, ampak v obliki vektorja meritev odzivne spremenljivke in posebnega vektorja, kjer povemo, h kateremu faktorju pripadajo posamezne meritve.
- Primer: ANOVA2 in ANOVAN

```
>> m = [23 15 20;27 17 63;43 3 55;41 9 90]
>> anova2(m,2)
m = 23    15    20
    27    17    63
    43     3    55
    41     9    90

ans =
    0.0197    0.2234    0.2663
```

ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	4232.67	2	2116.33	8.1	0.0197
Rows	481.33	1	481.33	1.84	0.2234
Interaction	868.67	2	434.33	1.66	0.2663
Error	1567	6	261.17		
Total	7149.67	11			

Matlab: ANOVAN

- Primer: ANOVA2 in ANOVAN
  - definicija faktorjev in stopenj

```
>> cfactor = repmat(1:3,4,1)
cfactor =
     1     2     3
     1     2     3
     1     2     3
     1     2     3
>> rfactor = [ones(2,3); 2*ones(2,3)]
rfactor =
     1     1     1
     1     1     1
     2     2     2
     2     2     2
```

```
>> m = m(:);
cfactor = cfactor(:);
rfactor = rfactor(:);

[m cfactor rfactor]

ans =

    23     1     1
    27     1     1
    43     1     2
    41     1     2
    15     2     1
    17     2     1
     3     2     2
     9     2     2
    20     3     1
    63     3     1
    55     3     2
    90     3     2
```

```
>> anovan(m,{cfactor rfactor},2)
ans =
    0.0197
    0.2234
    0.2663
```

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
X1	4232.67	2	2116.33	8.1	0.0197
X2	481.33	1	481.33	1.84	0.2234
X1*X2	868.67	2	434.33	1.66	0.2663
Error	1567	6	261.17		
Total	7149.67	11			

Matlab: ANOVAN

- Večji primer: Analiza podatkov o 406 avtih iz zbirke carbig:

```
>> load carbig
>> whos
  Name                Size              Bytes  Class     Attributes

  Acceleration      406x1                3248  double
  Cylinders         406x1                3248  double
  Displacement      406x1                3248  double
  Horsepower        406x1                3248  double
  MPG               406x1                3248  double
  Mfg               406x13              10556  char
  Model             406x36              29232  char
  Model_Year        406x1                3248  double
  Origin            406x7                5684  char
  Weight            406x1                3248  double
  cyl4              406x5                4060  char
  org               406x7                5684  char
  when              406x5                4060  char
```

  - Analizirali bomo 4 tipe podatkov:
    - odzivna spr.: porabo goriva (MPG),
    - 1. faktor: cyl4 – kategorija: avto ima 4 cilindre ali ne
    - 2. faktor: org – kategorija: država proizvodnje: Evropa, ZDA, Japonska
    - 3. faktor: when – kategorija: leto proizvodnje: prej, vmes, potem

Matlab: ANOVAN

- Večji primer: Analiza podatkov o 406 avtih iz zbirke carbig:

```
varnames = {'Origin';'4Cyl';'MfgDate'};
anovan(MPG,{org cyl4 when},3,3,varnames)
```

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
# Origin	416.8	1	416.77	29.34	0
# 4Cyl	0	0	0	0	NaN
# MfgDate	1112.3	1	1112.27	78.31	0
# Origin*4Cyl	2.1	1	2.07	0.15	0.7032
# Origin*MfgDate	301.2	3	100.41	7.07	0.0001
# 4Cyl*MfgDate	22.7	1	22.68	1.6	0.2072
# Origin*4Cyl*MfgDate	20.3	3	6.77	0.48	0.699
Error	5411.8	381	14.2		
Total	24252.6	397			

  - Parametri označeni z #: to pomeni, da nimamo pri vseh avtih podatka o porabi goriva.
  - Pri 4cyl pa se tudi ne more izračunati p-vrednosti, kar pomeni, da verjetno nimamo podatkov za vse možne stopnje tega faktorja.

```
>> [table, chi2, p, factorvals] = crosstab(org,when,cyl4)
table(:,:,1) =
    82    75    25
     0     4     3
     3     3     4
table(:,:,2) =
    12    22    38
    23    26    17
    12    25    32
```

```
chi2 =  207.7689
p =   8.0973e-38

factorvals =
    'USA'       'Early'     'Other'
    'Europe'    'Mid'       'Four'
    'Japan'     'Late'          []
```

Matlab: ANOVAN

- Večji primer: Analiza podatkov o 406 avtih iz zbirke carbig:
  - ker smo ugotovili, da ni potrebno imeti parametra trojne interakcije, naredimo ANOVO samo z dvojnimi interakcijami.

```
>> [p,tbl,stats,terms] = anovan(MPG,{org cyl4 when},2,3,varnames);
>> terms
terms =
     1     0     0
     0     1     0
     0     0     1
     1     1     0
     1     0     1
     0     1     1
```

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
Origin	532.6	2	266.29	18.82	0
4Cyl	1769.8	1	1769.85	125.11	0
MfgDate	2887.1	2	1443.55	102.05	0
Origin*4Cyl	12.5	2	6.27	0.44	0.6422
Origin*MfgDate	350.4	4	87.59	6.19	0.0001
4Cyl*MfgDate	31	2	15.52	1.1	0.3348
Error	5432.1	384	14.15		
Total	24252.6	397			

Matlab: ANOVAN

- Večji primer: Analiza podatkov o 406 avtih iz zbirke carbig:
  - odstranimo tiste člene, ki niso stat. značilni:

```
>> terms
terms =
     1     0     0
     0     1     0
     0     0     1
     1     1     0
     1     0     1
     0     1     1
```

```
>> terms([4 6], :) = []
terms =
     1     0     0
     0     1     0
     0     0     1
     1     0     1
```

  - Ponovimo ANOVO samo s členi, ki jih modeliramo:

```
>> anovan(MPG,{org cyl4 when},terms,3,varnames)
```

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
Origin	686.7	2	343.36	24.34	0
4Cyl	4206.2	1	4206.17	298.19	0
MfgDate	3590.7	2	1795.34	127.28	0
Origin*MfgDate	336.8	4	84.19	5.97	0.0001
Error	5473	388	14.11		
Total	24252.6	397			

  - Poraba goriva je odvisna od vseh treh faktorjev, vpliv pa ima tudi, kje in kdaj je bil avto proizveden.

Neparametrični testi ANOVE

- Poleg parametričnih testov ANOVE obstajajo tudi neparametrični testi enosmerne in dvosmerne ANOVE:
  - neparametrični testi nimajo predpostavke o normalni porazdelitvi podatkov, zato so primerni v primerih, ko nimamo takšne predpostavke,
  - so manj zanesljivi od klasičnih metod (ker imajo manj predpostavk)

- Neparametrični test ANOVE:
  - enosmerna ANOVA: Kruskal-Wallisov test
  - dvosmerna ANOVA: Friedmanov test

Kruskall-Wallisov test

- Pri enosmerni ANOVI smo imeli primer meritev količine bakterij v posameznih pošiljkah mleka:

```
>> load hogg
hogg
    24    14    11     7    19
    15     7     9     7    24
    21    12     7     4    19
    27    17    13     7    15
    33    14    12    12    10
    23    16    18    18    20
```

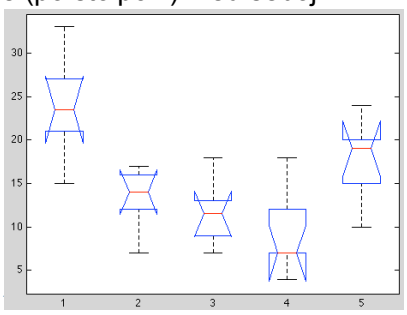
vrstice predstavljajo količino bakterij v posameznem litru mleka, ki smo ga naključno izbrali iz pošiljke:
imamo 6 pošiljk
in 5 naključno izbranih meritev
v posamezni pošiljki.

- preučevali smo, ali število bakterij varira med posameznimi pošiljkami mleka. Pri enosmerni ANOVI smo predpostavili, da so meritve (po stolpcih) med seboj neodvisne in da so porazdeljene normalno z enako varianco in 'fiksnimi' povprečji.

```
>> [p,tbl,stats] = anova1(hogg);
```

ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	803	4	200.75	9.01	0.0001
Error	557.17	25	22.287		
Total	1360.17	29			



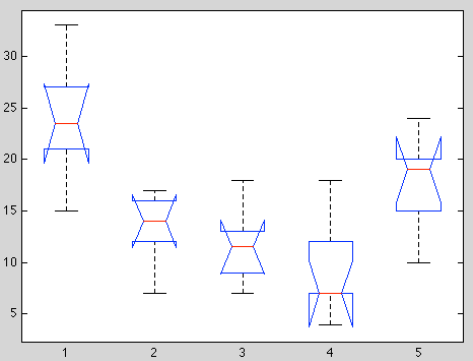
Kruskall-Wallisov test

- Ponovimo enosmerno ANOVO s Kruskall-Wallisovim testom.
  - Kruskall-Wallisov test je neparametrična verzija enosmerne ANOVE.
  - Predpostavka je, da so meritve zvezno porazdeljene, ne pa nujno normalno.
  - Test se ne računa neposredno na meritvah, ampak na vrstnem redu (rangu) meritev (podobno, kot pri neparametričnih testih hipotez).

```
>> p = kruskalwallis(hogg)
```

Kruskal-Wallis ANOVA Table

Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Columns	1302.25	4	325.562	16.91	0.002
Error	930.75	25	37.23		
Total	2233	29			



Friedmanov test

- Friedmanov test je neparametrična oblika dvosmerne ANOVE.
- V našem primeru dvosmerne ANOVE smo preučevali vpliv prehrane in dodatkov na rast živali:
  - Imeli smo dva faktorja: vrsto prehrane (3 stopnje: oves, pšenica, ječmen) in dodatke (4 stopnje).
  - Preučevali smo, pri katerem izmed faktorjev ali interakciji obeh znamo statistično značilne razlike pri rasti živali.
  - Izvedli smo dvosmerno ANOVO in ugotovili, da faktorja statistično značilno vplivata na rast živali, interakcija pa ne.

```
>> load datasets/growth.mat
>> Y = [X(1:4,:)'; X(5:8,:)'; X(9:12,:)' ]
>> [p, tbl, stats] = anova2(Y, 4)  %4 ponovitve za vsako kombinacijo stopenj
```

ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	91.881	3	30.627	17.81	0
Rows	287.171	2	143.586	83.52	0
Interaction	3.406	6	0.568	0.33	0.9166
Error	61.89	36	1.719		
Total	444.348	47			

Friedmanov test

- Friedmanov test je neparametrična oblika dvosmerne ANOVE.
  - za razliko od anove2 Friedmanov test v Matlabu ne testira vrstic in stolpcev posebej, ampak testira različnost samo med stolpci, saj se testna statistika izračuna za različnost med stolpci iz razlik vrstic, ki so urejene po vrstnem redu (zopet ne primerjamo neposredne meritve, ampak vrstni red meritev).
  - S tem preučujemo vpliv faktorja, ki je zapisan po vrsticah (v našem primeru vrste prehrane).
  - Interakcije obeh faktorjev ne moremo analizirati.

```
>> [p, tbl, stats] = friedman(Y, 4)  %4 ponovitve za vsako kombinacijo stopenj
```

Friedman's ANOVA Table

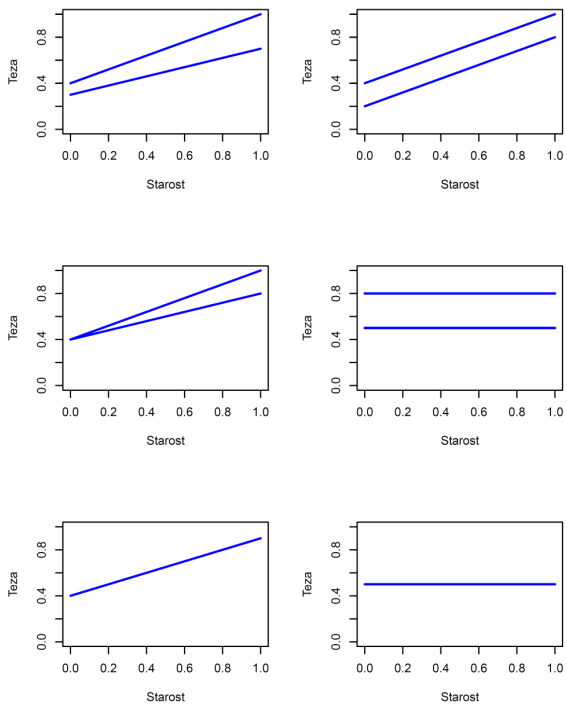
Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Columns	675.167	3	225.056	29.79	1.53019e-06
Interaction	19.333	6	3.222		
Error	325.5	36	9.042		
Total	1020	47			

  - Vrsta prehrane ima statistično značilen različen vpliv na rast živali.
  - Za ugotavljanje vpliva dodatkov je potrebno matriko meritev ustrezno preoblikovati.

Analiza kovariance: ANCOVA

- Analiza kovariance vključuje analizo variance in regresijo.
- Odzivna spr. je zvezna.
- Opisne spr.:
  - vsaj ena zvezna,
  - vsaj ena kategorijska (faktorska).
- Maksimalni model:
  - regresijski model za vsako stopnjo faktorja.
- Minimizacija modela:
  - zmanjševanje števila parametrov modela.

Analiza kovariance: primer



- Denimo, da modeliramo težo ljudi v odvisnosti od spola in starosti:
  - spol je faktor (kategorija) z dvema stopnjama: moški, ženske
  - starost je zvezna opisna spr.
- Maksimalni model je torej lahko:
  - regresijska premica za moške in regresijska premica za ženske:

$$\text{teza}_{\text{moski}} = a_{\text{moski}} + b_{\text{moski}} \times \textit{starost}$$

$$\text{teza}_{\text{zenska}} = a_{\text{zenska}} + b_{\text{zenska}} \times \textit{starost}$$

- Minimizacija modela:
- imamo 6 možnih modelov:
  - različni a in b za oba spola
  - enak naklon, različni a
  - različni nakloni, skupen a
  - brez naklona in različni a
  - skupni a in b,
  - en skupen a (brez b), ki je povprečje

Minimizacija modela

- Odločitve o zmanjševanju parametrov modela sprejemamo na podlagi statistične analize posameznih parametrov modela.

- Minimizacija:
  - Če opisuje poenostavljen model varianco odzivne spr. enako dobro kot ne-poenostavljen model (razlika ni statistično značilna), potem je poenostavljeni model boljši.

- Preverjanje kvalitete modela izvedemo z analizo variance med enim in drugim modelom:
  - če je p-vrednost manjša od 0.05 je razlika statistično značilna (p-vrednost v tem primeru meri verjetnost, da se zgodi razlika pri opisovanju variance odzivne spr. ob hipotezi, da sta modela enaka).

ANCOVA: primer

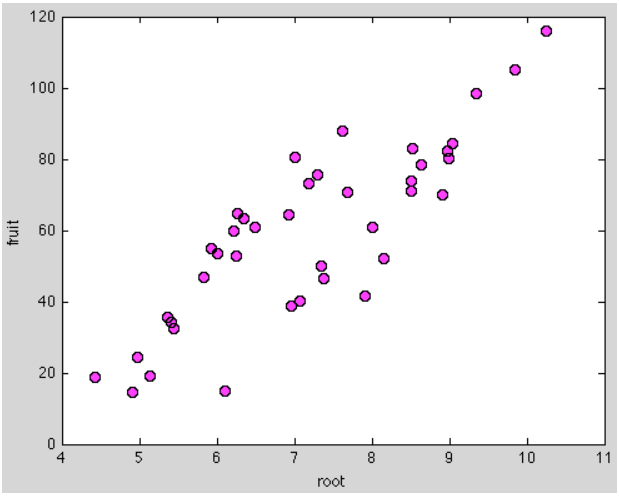
- Preučevanje vpliva paše živine na poraščenost pašnikov. Poskus vključuje opazovanje količine sadežev, ki jih proizvedejo rastline – dvoletnice v odvisnosti od začetne velikosti rastlin pred pašo in glede na izpostavljenost paši ali pa ne.

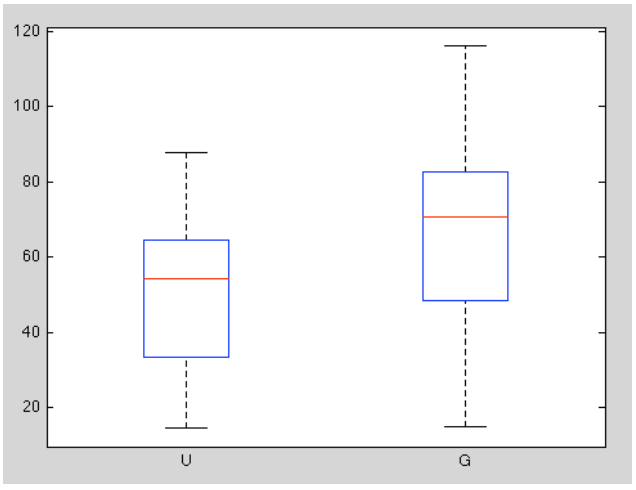
- Odzivna spremenljivka:
  - skupna količina sadežev na rastlino: fruit

- Opisni spremenljivki:
  - začetna velikost rastline pred pašno sezono (zv. spr.): root
  - kategorijska spremenljivka – faktor z dvema stopnjama: paša, brez paše (grazed, ungrazed)

```
>> load datasets/compensation.mat

Root      Fruit     Grazing
6.225     59.77     Ungrazed
6.487     60.98     Ungrazed
…
10.253    116.05    Grazed
6.958     38.94     Grazed
8.001     60.77     Grazed
```

ANCOVA: primer



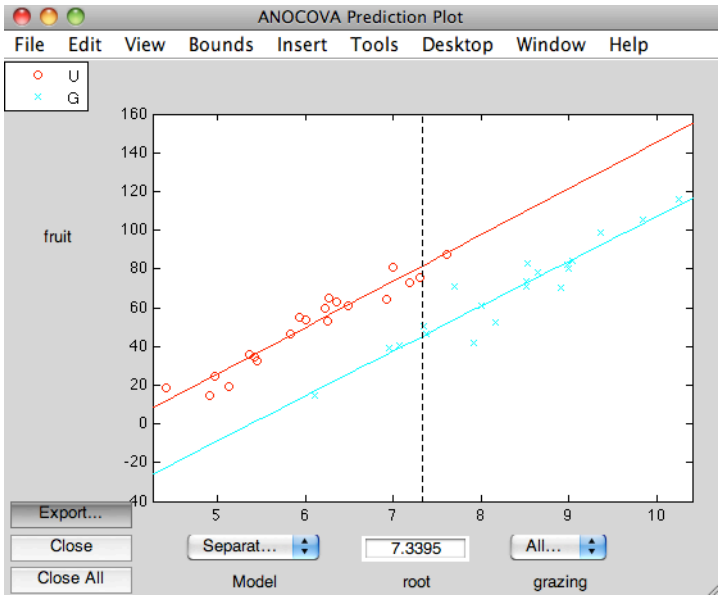



```
>>plot(root, fruit, 'o', 'MarkerEdgeColor','k', 'MarkerFaceColor','m','MarkerSize',7);
>> xlabel('root'); ylabel('fruit')

>> boxplot(fruit, grazing)
```

ANCOVA primer

- Statistična analiza v Matlabu:

```
[h,atab,ctab,stats] = aoctool(root,fruit,grazing);
```



ANCOVA primer

- Statistična analiza v Matlabu:

```
[h,atab,ctab,stats] = aoctool(root,fruit,grazing);
```

Coefficient Estimates

Term	Estimate	Std. Err.	T	Prob>|T|
Intercept	-109.77	8.42091	-13.04	0
U	15.4	8.42091	1.83	0.0757
G	-15.4	8.42091	-1.83	0.0757
Slope	23.62	1.17706	20.07	0
U	0.38	1.17706	0.32	0.75
G	-0.38	1.17706	-0.32	0.75

Model:

linearni model za U: *fruit = (-109.77+15.4) + (23.62+0.38)root*
linearni model za G: *fruit = (-109.77-15.4) + (23.62-0.38)root*

ANCOVA primer

- Statistična analiza v Matlabu:

```
[h,atab,ctab,stats] = aoctool(root, fruit, grazing);
```

ANOVA Table

Source	d.f.	Sum Sq	Mean Sq	F	Prob>F
grazing	1	5264.4	5264.4	112.83	0
root	1	19148.9	19148.9	410.42	0
grazing*root	1	4.8	4.8	0.1	0.75
Error	36	1679.6	46.7		

SSR

SSA

SSRrazlike

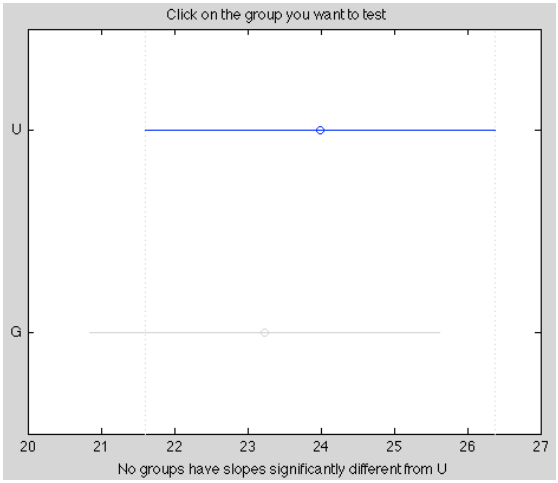
SSE

SSY = SSR + SSA + SSRrazlike + SSE

ANCOVA primer

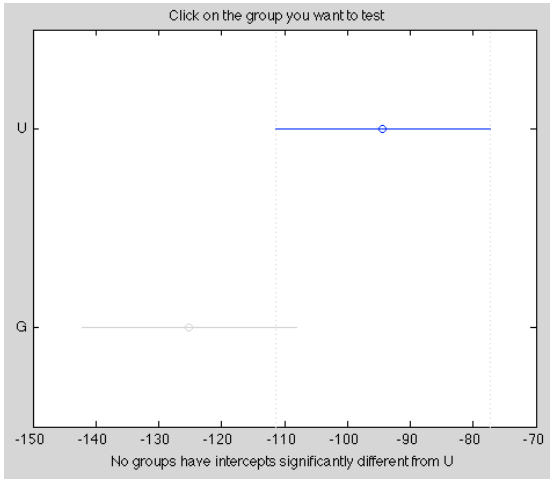
- Statistična ustreznost posameznih parametrov:

```
%analiza signifikantnosti smernega koeficienta
multcompare(stats, 0.05, 'on', '', 'slope')
```




```
1.0000    2.0000   -4.0183    0.7560    5.5304
```

```
% analiza signifikantnosti odseka
multcompare(stats, 0.05, 'on', '', 'intercept')
```

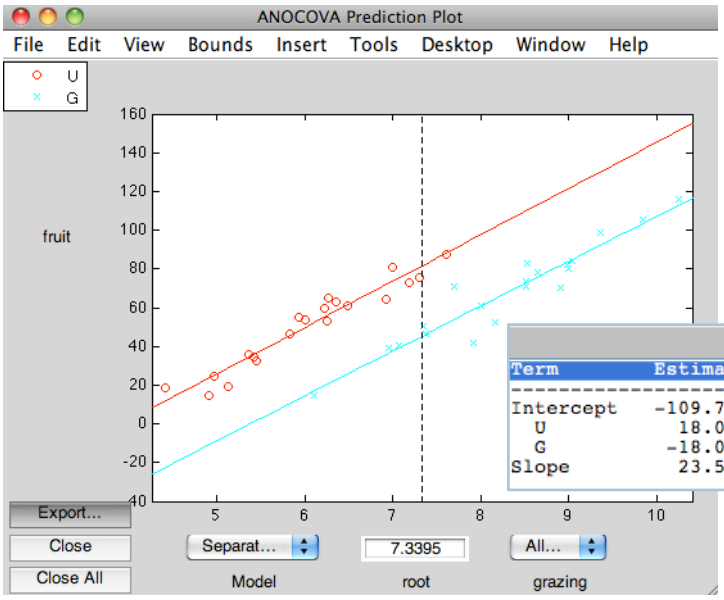



```
1.0000    2.0000   -3.3511   30.8057   64.9625
```

ANCOVA primer

- Statistična analiza v Matlabu: manjši model.

```
[h,atab,ctab,stats] = aoctool(root,fruit,grazing);
```



Coefficient Estimates

Term	Estimate	Std. Err.	T	Prob>|T|
Intercept	-109.78	8.31819	-13.2	1.44703e-15
U	18.05	1.6787	10.75	6.10729e-13
G	-18.05	1.6787	-10.75	6.10729e-13
Slope	23.56	1.14877	20.51	8.40823e-22

ANCOVA: interpretacija rezultatov

- Na začetku smo presenetljivo ugotovili, da je produkcija sadežev na popasenih rastlinah višja kot na ne-popasenih rastlinah. In če bi pozabili na začetno velikost rastlin, bi lahko to ugotovitev tudi potrdili:

```
>> mean(fruit(grazing=='U'))
ans = 50.8805
>> mean(fruit(grazing=='G'))
ans = 67.9405

>> anova1(fruit, grazing)
```

ANOVA Table

Source	SS	df	MS	F	Prob>F
Groups	2910.4	1	2910.44	5.31	0.0268
Error	20833.4	38	548.25		
Total	23743.8	39			

- Ob upoštevanju začetne velikosti pa je ocenjena vrednost ob povprečju začetne velikosti:
  - Ungrazed:
  - Grazed:

```
>> -109.78+18.05 + 23.56*mean(root)
 77.4579

>> -109.78-18.05 + 23.56*mean(root)
 41.3579
```

Term	Estimate
Intercept	-109.78
U	18.05
G	-18.05
Slope	23.56

A Osnove programskega jezika MATLAB

V dodatku predstavimo nekaj osnov programskega jezika, ki se uporablja v programskemu paketu MATLAB. Matlab je izvrstno orodje za vse inženirje, ki potrebujejo zmogljivo matematiko, predvsem numerično. Matlab je sestavljen iz več programskih paketov, t.i. toolbox-ov, ki imajo različne funkcionalnosti, med njimi je tudi paket *Statistics Toolbox*, ki ga uporabljamo pri izvajanju primerov statistične analize in modeliranja v tej knjigi.

Osnove programskega jezika MATLAB

- Imena spremenljivk so občutljiva na male in velike črke.
- Začeti se morajo s črko, lahko sledijo številke, podčrtaji.

- Primer:

```
>> x = 2;
>> abc_123 = 0.005;
>> 1ab = 2;
Error: Unexpected MATLAB expression
```

Osnove programskega jezika MATLAB

- Posebne že definirane spremenljivke:

  - pi ... vrednost števila pi = 3.1416...
  - eps ... najmanjša vrednost, kjer še ločujemo decimalna števila
  - inf ... oznaka za neskončno, npr. 1/0
  - NaN ... to ni število (NaN = not a number), npr. 0/0 ali manjkajoča vrednost
  - i ali j ... koren števila (-1), imaginarna enota
  - realmin ... najmanjše realno število (=2.2251e-308)
  - realmax... največje realno število (=1.7977e+308)

Osnove programskega jezika MATLAB

- Relacije med števili:
  - < manjše kot
  - <= manjše ali enako kot
  - > večje kot
  - >= večje ali enako kot
  - == enako kot
  - ~= različno kot (ne tako kot pri C-ju !=)
- Logične operacije:
  - ~ negacija
  - & logični in
  - | logični ali

Osnove programskega jezika MATLAB

- Matrike in vektorji
  - V Matlabu obravnavamo vse spremenljivke kot matrike ali več dimenzionalna polja:
    - Vektorji so posebna oblika matrike - stolpec ali vrstica.
    - Skalarji so matrike z eno vrstico in enim stolpcem.
  - Primeri:
    - skalar

```
>> x = 23;
```

    - vektor kot vrstica

```
>> y = [12,10,-3]
y =
    12    10    -3
```

    - vektor kot stolpec

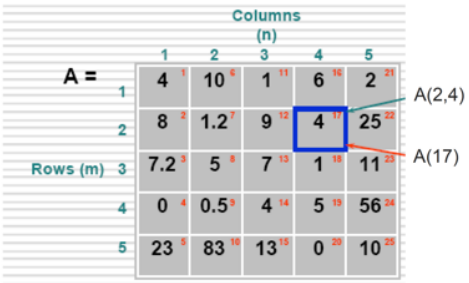
```
>> z = [12;10;-3]
z =
    12
    10
    -3
```

Matrika:

```
>> X = [1,2,3;4,5,6;7,8,9]
X =
     1     2     3
     4     5     6
     7     8     9
```

Osnove programskega jezika MATLAB

- Matrike:



- Dostopanje do elementov matrike:

```
>> X = [1,2,3;4,5,6;7,8,9]
X =
    1    2    3
    4    5    6
    7    8    9

>> X22 = X(1:2 , 2:3)
X22 =
       2    3
       5    6
```

```
>> X13 = X(3,1:3)
X13 =
       7    8    9

>> X21 = X(1:2,1)
X21 =
       1
       4
```

Osnove programskega jezika MATLAB

- Delo z matrikami:

```
>> a = [1,2i,0.56]
a =
    1   0+2i   0.56
>> a(2,4) = 0.1
a =
    1   0+2i   0.56   0
    0    0      0    0.1
```

- repmat - podvoji in razdeli matrike

```
>> b = [1,2;3,4]
b =
    1   2
    3   4
>> b_rep = repmat(b,1,2)
b_rep =
        1   2   1   2
        3   4   3   4
```

- Združevanje matrik:

```
>> a = [1,2;3,4]
a =
    1   2
    3   4
>> a_cat =[a,2*a;3*a,2*a]
a_cat =
        1   2   2   4
        3   4   6   8
        3   6   2   4
        9   12  6   8
```

Dimenzije matrike se morajo ujemati.

Osnove programskega jezika MATLAB

- Matrike: seštevanje
  - povečanje vseh vrednosti za 5:

```
>> x = [1,2;3,4]
x =
    1    2
    3    4
>> y = x + 5
y =
    6    7
    8    9
```

- Seštevanje dveh matrik:

```
>> xsy = x + y
xsy =
       7     9
      11    13
>> z = [1,0.3]
z =
    1    0.3
>> xsz = x + z
??? Error using => plus
Matrix dimensions must
agree
```

Osnove programskega jezika MATLAB

- Matrike: množenje:

```
>> a = [1,2;3,4];     (2x2)
>> b = [1,1];         (1x2)
>> c = b*a
c =
      4    6

>> c = a*b
??? Error using ==> mtimes
Inner matrix dimensions
must agree.
```

- Množenje po elementih matrike:

```
>> a = [1,2;3,4];
>> b = [1,½;1/3,¼];
>> c = a.*b
c =
    1    1
    1    1
```

Osnove programskega jezika MATLAB

- Operacije po elementih matrike:

```
>> a = [1,2;1,3];
>> b = [2,2;2,1];
```

- Deljenje:

```
>> c = a./b
c =
    0.5    1
    0.5    3
```

- Množenje:

```
>> c = a.*b
c =
     2     4
     2     3
```

- Potenciranje:

```
>> c = a.^2
c =
     1     4
     1     9
```

```
>> c = a.^b
c =
     1     4
     1     3
```

Osnove programskega jezika MATLAB

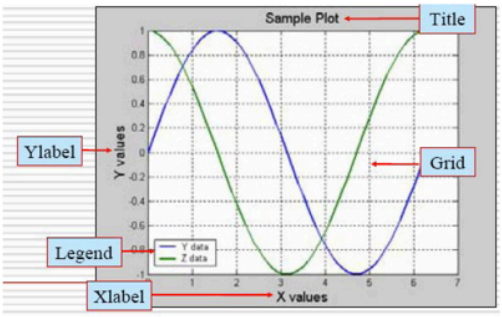
- Funkcije za delo z matrikami:
  - zeros : creates an array of all zeros, Ex: x = zeros(3,2)
  - ones : creates an array of all ones, Ex: x = ones(2)
  - eye : creates an identity matrix, Ex: x = eye(3)
  - rand : generates uniformly distributed random numbers in [0,1]
  - diag : Diagonal matrices and diagonal of a matrix
  - size : returns array dimensions
  - length : returns length of a vector (row or column)
  - det : Matrix determinant
  - inv : matrix inverse
  - eig : evaluates eigenvalues and eigenvectors
  - rank : rank of a matrix
  - find : searches for the given values in an array/matrix.

  - sum,prod - summation and product of elements
  - max,min - maximum and minimum of arrays
  - mean,median – average and median of arrays
  - std,var - Standard deviation and variance

- Osnove grafike:
  - Primer sinusa in kosinusa na intervalu od 0 do 2*pi:

1. možnost:

```
>> x = linspace(0,2*pi,1000);
>> y = sin(x);
>> z = cos(x);
>> hold on;
>> plot(x,y,'b');
>> plot(x,z,'g');
>> xlabel 'X values';
>> ylabel 'Y values';
>> title 'Sample Plot';
>> legend ('Y data','Z data');
>> hold off;
```

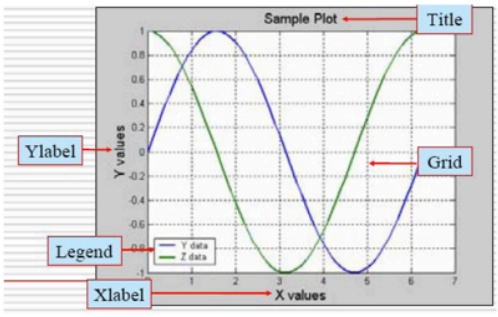


Osnove programskega jezika MATLAB

- Osnove grafike:
  - Primer sinusa in kosinusa na intervalu od 0 do 2*pi:

2. možnost:

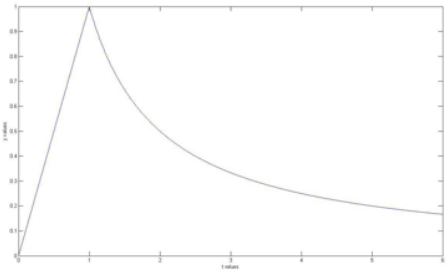
```
>> x = 0:0.01:2*pi;
>> y = sin(x);
>> z = cos(x);
>> figure
>> plot (x,y,x,z);
>> xlabel 'X values';
>> ylabel 'Y values';
>> title 'Sample Plot';
>> legend ('Y data','Z data');
>> grid on;
```



- Osnove grafike:
  - Primer vejnate funkcije: $y = \begin{cases} t & 0 \le t \le 1 \\ 1/t & 1 \le t \le 6 \end{cases}$

1. možnost:

```
>> t1 = linspace(0,1,1000);
>> t2 = linspace(1,6,1000);
>> y1 = t1;
>> y2 = 1./ t2;
>> t = [t1,t2];
>> y = [y1,y2];
>> figure
>> plot(t,y);
>> xlabel 't values', ylabel 'y values';
```

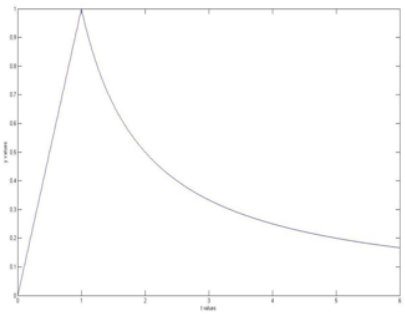


Osnove programskega jezika MATLAB

- Osnove grafike:
  - Primer vejnate funkcije: $y = \begin{cases} t & 0 \le t \le 1 \\ 1/t & 1 \le t \le 6 \end{cases}$

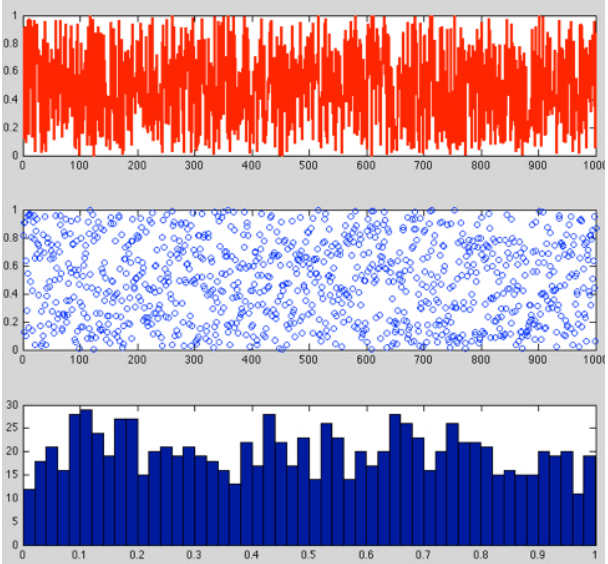
2. možnost:

```
>> t = linspace(0,6,1000);
>> y = zeros(1,1000);
>> y(t()<=1) = t(t()<=1);
>> y(t()>1) = 1./ t(t()>1);
>> figure
>> plot(t,y);
>> xlabel't values';
>> ylabel'y values';
```



Osnove programskega jezika MATLAB

- Osnove grafike:
  - več grafov skupaj na eni sliki : `>> subplot(rows, columns, index)`



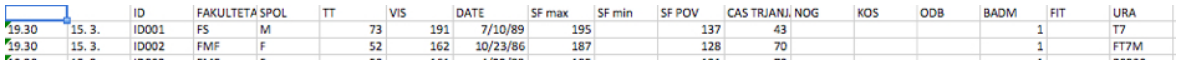
```
>> sig = rand(1000,1);
>> subplot(3,1,1)
>> plot(sig, 'r', 'LineWidth', 1.5)


>> subplot(3,1,2)
>> plot(1:1000, sig, 'o')



>> subplot(3,1,3)
>> hist(sig, 50)
```

Osnove programskega jezika MATLAB

- Branje in shranjevanje podatkov
  - Uporaba funkcij load in save:
    - `load filename`
      - v delovni prostor naložimo vse spr. iz datoteke filename
    - `load filename x`
      - v delovni prostor naložimo spr. x iz datoteke filename
    - `load filename a*`
      - v delovni prostor naložimo spr., ki se začnejo na a
    - za več informacij `help load` ali `doc load`
    - `save filename`
      - v datoteko filename.mat shranimo vse spr. iz delovnega prostora
    - `save filename x,y`
      - v datoteko filename.mat shranimo spr. x in y
    - za več informacij `help load` ali `doc load`

Osnove programskega jezika MATLAB

- Branje podatkov iz excel tabele



		ID	FAKULTETA	SPOL	TT	VIS	DATE	SF max	SF min	SF POV	CAS TRJANJ	NOG	KOS	ODB	BADM	FIT	URA
19.30	15. 3.	ID001	FS	M	73	191	7/10/89	195		137	43				1		T7
19.30	15. 3.	ID002	FMF	F	52	162	10/23/86	187		128	70				1		FT7M

- Naložimo podatke iz excel tabele:
  - `>> [num,txt,raw] = xlsread('testiranje_fs_pop.xls');`

```
>> raw(1,:)
ans =
  Columns 1 through 8
    [NaN]    [NaN]    'ID'    'FAKULTETA'    'SPOL'    'TT'    'VIS'    'DATE'
  Columns 9 through 15
    'SF max'    'SF min'    'SF POV'    'CAS TRJANJA'    'NOG'    'KOS'    'ODB'
  Columns 16 through 18
    'BADM'    'FIT'    'URA'
>> raw(2,:)
ans =
  Columns 1 through 8
    '19.30'    '15. 3.'    'ID001'    'FS'    'M'    [73]    [191]    [32699]
  Columns 9 through 17
    [195]    [NaN]    [137]    [43]    [NaN]    [NaN]    [NaN]    [1]    [NaN]
  Column 18
    'T7'
```

Osnove programskega jezika MATLAB

- Pisanje podatkov v excel tabelo

```
>> A = rand(3,3)

A =

    0.6312    0.2242    0.3872
    0.3551    0.6525    0.1422
    0.9970    0.6050    0.0251

>> xlswrite('a.xls', A, 'A2:C4')
```

Osnove programskega jezika MATLAB

- Branje podatkov v drugih formatih

```
Field.Name      Area    Slope   Vegetation      Soil.pH Damp    Worm.density
Nashs.Field     3.6     11      Grassland       4.1     F       4
Silwood.Bottom  5.1     2       Arable  5.2     F       7
Nursery.Field   2.8     3       Grassland       4.3     F       2
Rush.Meadow     2.4     5       Meadow  4.9     T       5
Gunness.Thicket 3.8     0       Scrub   4.2     F       6
```

```
>> dezevniki = importdata('datasets/dezevniki.txt')
dezevniki =
        data: [5x1 double]
    textdata: {6x7 cell}

>> dezevniki.textdata
ans =

    'Field.Name'          'Area'    'Slope'    'Vegetation'    'Soil.pH'    'Damp'    'Worm.density '
    'Nashs.Field'         '3.6'     '11'       'Grassland'     '4.1'        'F'       ''
    'Silwood.Bottom'      '5.1'     '2'        'Arable'        '5.2'        'F'       ''
    'Nursery.Field'       '2.8'     '3'        'Grassland'     '4.3'        'F'       ''
    'Rush.Meadow'         '2.4'     '5'        'Meadow'        '4.9'        'T'       ''
    'Gunness.Thicket'     '3.8'     '0'        'Scrub'         '4.2'        'F'       ''

>> dezevniki.data
ans =
     4
     7
     2
     5
     6
```

Osnove programskega jezika MATLAB

- Pisanje in branje podatkov: binarne datoteke

```
>> m5w = magic(5)
m5w =
    17    24     1     8    15
    23     5     7    14    16
     4     6    13    20    22
    10    12    19    21     3
    11    18    25     2     9
>> fid = fopen('magic5.bin', 'w');
>> fwrite(fid, m5w);
>> fclose(fid);
```

```
>> fid = fopen('magic5.bin');
>> m5 = fread(fid, [5, 5], '*uint8');
>> fclose(fid);
>> m5
m5 =
   17   24    1    8   15
   23    5    7   14   16
    4    6   13   20   22
   10   12   19   21    3
   11   18   25    2    9
```

- Pisanje in branje podatkov: tekstovne datoteke:

```
>> x = 0:.1:1;
>> y = [x; exp(x)];
>> fid = fopen('exp.txt', 'w');
>> fprintf(fid, '%6.2f %12.8f\n', y);
>> fclose(fid);
```

```
>> fid = fopen('exp.txt');
>> A = fscanf(fid, '%g %g', [2 inf]);
>> fclose(fid);
```

Osnove programskega jezika MATLAB

- Kontrolni stavki v Matlabu
  - if, switch, for, while, break

- If:

```
>> if i == j
>>    a(i,j) = 2;
>> elseif i >= j
>>    a(i,j) = 1;
>> else
>>    a(i,j) = 0;
>> end
```

```
>> if (attn>0.9)&(grade>60)
>>    pass = 1;
>> end
```

- Switch:

```
>> x = 2, y = 3;
 >> switch x
>>  case x==y
>>   disp('x and y are equal');
>>  case x>y
>>   disp('x is greater than y');
>>  otherwise
>>   disp('x is less than y');
>> end
x is less than y
```

Osnove programskega jezika MATLAB

- Kontrolni stavki v Matlabu
  - if, switch, for, while, break

- For:

```
for x = 0:0.05:1
   printf('%d\n',x);
end
```

```
a = zeros(n,m);
for i = 1:n
  for j = 1:m
      a(i,j) = 1/(i+j);
  end
end
```

- While:

```
n = 1;
y = zeros(1,10);
while n <= 10
   y(n) = 2*n/(n+1);
   n = n+1;
end
```

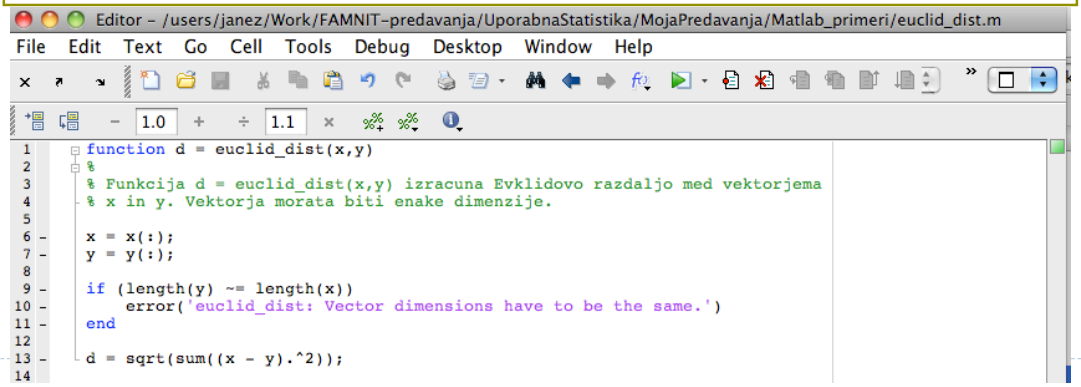
Break:

```
>> y = 3;
>> for x = 1:10
>>    printf('%5d',x);
>>    if (x>y)
>>        break;
>>    end
>> end
1     2     3     4
```

Osnove programskega jezika MATLAB

- Funkcije v Matlabu:

```
>> edit euclid_dist.m
>> euclid_dist(ones(3,1), zeros(3,1))
ans =
    1.7321
>> euclid_dist(ones(3,1), zeros(2,1))
??? Error using ==> euclid_dist at 10
euclid_dist: Vector dimensions have to be the same.

>> help euclid_dist
  Funkcija d = euclid_dist(x,y) izracuna Evklidovo razdaljo med vektorjema
  x in y. Vektorja morata biti enake dimenzije.
```



Osnove programskega jezika MATLAB

- Učinkovito programiranje v MATLABu
  - Izogibati se je potrebno zankam. Čim manj gnezdenih zank.
  - V veliki večini primerov se lahko izognemo uporabi zank z uporabo matričnih operacij.
  - Ko izvajamo prirejanje, npr. y = x; se v bistvu izvaja kopiranje podatkov, zato raje ponovno uporabljamo že alocirane spremenljivke. Zelo slabo je povečevati matrike (in druge podatkovne strukture) postopoma, ampak je bolje vnaprej predpisati velikost matrike (če seveda poznamo velikost), npr. z uporabo A = zeros(N,M).
  - Uporabljamo funkcije, ki so že narejene v MATLABu, saj so v večini primerov učinkovito narejene.

Osnove programskega jezika MATLAB

- Učinkovito programiranje v MATLABu

- Primer: Izračunajmo primer ene vsote (FIR filtra) pri danih x in h za n=1,2,3:

$$y[n] = \sum_{k=0}^{19} h[k]x[n+k]$$

```
% 1. moznost
y = zeros(1,3);
for n = 1:3
    for k = 0:19
        y(n)= y(n)+h(k)*x(n+k);
    end
end
```

```
% 2. moznost
y = zeros(1,3);
for n = 1:3
    y(n) = h.'*x(n:(n+19));
end
```

```
% 3. moznost
X= [x(1:20),x(2:21),x(3:22)];
y = h.'*X;
```

Osnove programskega jezika MATLAB

- Učinkovito programiranje v MATLABu

- Primer: Izračunajmo naslednji izraz za n= 1 do 20:

$$y(n) = 1^3*(1^3+2^3)*(1^3+2^3+3^3)*\ldots*(1^3+2^3+\ldots+n^3)$$

```
%1. moznost
y = zeros(20,1);
y(1) = 1;
for n = 2:20
    for m = 1:n
        temp = temp + m^3;
    end
    y(n) = y(n-1)*temp;
    temp = 0;
end
```

```
% 2. moznost
y = zeros(20,1);
y(1) = 1;
for n = 2:20
    temp = 1:n;
    y(n) = y(n-1)*sum(temp.^3);
end
```

```
% 3. moznost
X = tril(ones(20)*diag(1:20));
x = sum(X.^3,2);
Y = tril(ones(20)*diag(x))+ triu(ones(20)) - eye(20);
y = prod(Y,2);
```

Osnove programskega jezika MATLAB

- Nadaljnja literatura:
  - doc, help v MATLABu
  - **Na internetu:**
  - http://www.mathworks.com/support/
  - http://www.mathworks.com/products/demos/#
  - http://www.math.siu.edu/MATLAB/tutorials.html
  - http://matlab.wikia.com/wiki/FAQ