6. lekcija: procjenjivanje i testiranje

1. Uvod

Ponovimo neke pojmove iz deskriptivne statistike.
Populacija je skup svih entiteta koje razmatramo, na primjer svi studenti nekog sveučilišta čine populaciju.
Razmatramo neko statistističko obilježje populacije, na primjer visinu. Visina je slučajna veličina.
Uzorak je neki podskup populacije slučajno odabran, na primjer slučajno odabranih 300 studenata.
Neka je **n** duljina uzorka, na primjer n=300.
Mjerenjem slučajne veličine X na uzorku dobijemo *n* podataka:
$x_1, x_2, \ldots, x_n$.
Interpretiramo ih kao n slučajnih vrijednosti slučajne varijable X.

Primjer 1. Da bismo procijenili količinu kemikalije u posudama koje se automatski pune, izaberemo slučajno 10 posuda i provjeravamo količinu kemikalije u njima. Dobivamo podatke koji (nakon sređivanja, od manjeg prema većem) možemo zapisati ovako:
0.98, 0.98, 0.98, 0.99, 0.99, 1.00, 1.01, 1.01, 1.01, 1.02.
Tu slučajna veličina X mjeri količinu kemikalije u posudi,
uzorak čine odabrane posude,
n=10,
x_1, do x_{10} jesu podatci 0.98,…, 1.02; to su vrijednosti slučajne veličine X na uzorku.

Neka slučajna veličina X (u primjeru ili općenito) ima očekivanje μ i varijancu σ^2:
$E(X) = \mu$
$V(X) = \sigma^2$
(takve ćemo oznake imati i onda ako X nema normalnu razdiobu – približno normalnu razdiobu, već neku drugu, iako u pravilu razmatramo samo slučajne veličine normalno distribuirane).
Ta su nam dva parametra od X nepoznata pa ih **procjenjujemo** na osnovi mjerenja.

Očekivanje E(X) procjenjujemo aritmetičkom sredinom podataka

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

V(X) procjenjujemo izrazom

$$s^2 = \frac{(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1},$$ (u nazivniku je *n*-1, a ne *n*)

U gornjem je primjeru:

$$\bar{x} = \frac{3 \cdot 0.98 + 2 \cdot 0.99 + 1.00 + 3 \cdot 1.01 + 1.02}{10}$$

$= 0.997$

$$s^2 = \frac{3(0.98-0.997)^2 + 2(0.99-0.997)^2 + (1.00-0.997)^2 + 3(1.01-0.997)^2 + (1.02-0.997)^2}{10-1}$$

$= 0.000233$

$s = 0.014944$

Slučajne vrijednosti slučajne veličine (varijable) X.

Sad ćemo sve izreći malo drukčijim jezikom.

U predhodnom primjeru slučajna veličina X registrira količinu kemikalije u staklenki iz populacije. Kako je populacija konačna i ta je slučajna veličina konačna. Budući da je populacija vrlo velika, tu slučajnu veličinuu u pravilu zamjenjujemo kontinuiranom. Tu kontinuiranu slučajnu varijablu takodjer ćemo označavati oznakom X. Intuitivno je jasno da je ta X normalno distribuirana (poslije ćemo vidjeti pobliže što to znači).

Slučajna veličina X može postići bilo koju svoju vrijednost.

Neka je $x_1, x_2, ..., x_n$ nekih n slučajnih vrijednosti slučajne veličine X. Postavlja se pitanje približne rekonstrukcije slučajne varijable X iz ovih *n* slučajnih vrijednosti. Općenito, pitanje slučajnih vrijednosti slučajne varijable je vrlo važno i teško praktično i teoretsko pitanje. Postoje algoritmi za približno određivanje slučajnih vrijednosti, koje se obično nazivaju **pseudoslučajne vrijednosti**. Na primjer, u programskom paketu Mathematica, pseudoslučajne vrijednosti dobivaju se naredbom `RandomArray`.

Primijenom te naredbe na normalnu distribuciju s očekivanjem $\mu = 175$ i standardnom devijacijom $\sigma = 5$, za 100 slučajnih vrijednosti dobilo se:

```
{172.245,171.528,175.126,181.414,170.207,178.076,172.062,172.4
66,163.106,172.987,178.936,170.424,188.639,174.808,172.607,170
.222,176.149,171.733,179.166,172.677,169.084,179.869,179.148,1
63.325,174.914,170.227,170.328,173.236,169.499,183.918,177.506
,174.083,179.498,163.901,181.032,178.373,180.085,162.944,172.3
93,176.77,183.359,175.51,165.857,175.806,173.678,173.769,170.8
66,165.969,180.366,169.439,178.993,178.954,166.12,173.062,176.
924,179.091,173.304,165.135,181.489,179.646,183.993,169.244,17
2.846,169.152,177.249,173.359,177.106,182.76,174.611,177.011,1
65.135,173.365,170.879,177.681,170.9,177.904,179.597,170.347,1
75.311,176.744,179.578,181.396,178.267,178.185,175.475,184.13,
166.898,178.865,170.939,181.221,175.353,176.94,181.164,177.516
,173.84,171.767,173.072,172.221,172.539,183.831}
```

Da bismo bolje uočavali ove podatke primijenimo naredbu **Sort**, kojom dobijemo uređenu listu:

```
{160.234,163.649,164.078,165.826,165.905,166.895,166.908,166.9
87,167.01,167.226,167.259,167.604,168.073,168.17,168.684,168.6
88,169.749,170.298,170.31,170.398,170.43,170.622,170.778,170.8
34,171.,171.227,171.446,171.549,171.694,171.72,171.758,171.832
,172.038,172.323,172.38,172.81,172.889,173.16,173.194,173.255,
173.357,173.514,173.662,173.858,173.93,173.95,174.034,174.073,
174.115,174.171,174.336,174.545,174.621,174.627,174.713,174.78
1,175.104,175.376,175.571,175.579,175.631,175.714,175.771,176.
```

047,176.069,176.296,176.332,176.477,176.485,176.56,176.572,176.632,176.995,177.663,178.582,178.625,178.67,178.718,179.48,179.74,179.778,179.884,179.887,180.1,180.169,180.585,180.62,180.749,180.927,180.994,181.07,181.334,181.863,181.93,182.203,182.534,183.427,184.706,185.716,187.524}

Sad se možemo uvjeriti u funkcioniranje pravila *tri sigme*. Naime, prebrojavanjem dobijemo:
u intervalu <170,180> je 66 podataka (idealno bi trebalo biti 68)
u intervalu <165,185> je 95 podataka (idealno bi trebalo biti također 95)
u intervalu <160,190> je svih 100 podataka (kako bi trebalo biti i idealno).

U ovom primjeru pošli smo od poznate slučajne varijable (tj. poznate distribucije) i njenih 100 slučajnih vrijednosti. U praksi se pojavljuje situacija da znademo samo konačno mnogo podataka (slučajnih vrijednosti), a da ne znamo izvornu distribuciju.

Tada očekivanje μ procjenjujemo aritmetičkom sredinom podataka $\bar{x} = \dfrac{x_1 + x_2 + ... + x_n}{n}$

U ovom slučaju, primjenom naredbe **Mean**, dobijemo (na jednu decimalu)
$\bar{x}$**=176.0**
Zaključimo: μ **= 175**, a procjenom iz 100 slučajno odabranih vrijednosti dobili smo $\bar{x}$**=176.0**, pa procjenjujemo $\mu \approx 176.0$.
Dakle, iako je broj podataka bio relativno velik, pri procjenjivanju je došlo do grješke.

Slično postupamo pri procjeni varijance, odnosno standardne devijacije.

Pokazuje se (vidite prošireni tekst) da je $s^2 = \dfrac{(x_1 - \bar{x})^2 + ... + (x_n - \bar{x})^2}{n-1}$

nepristrana procjena varijance σ^2. Time se objašnjava činjenica što je u nazivniku *n-1* (korigirana varijanca uzorka), a ne *n* (varijanca uzorka).
Nastavljajući s početnim primjerom, koristeći se naredbom **Variance[data]**, dobivamo, približno na dvije decimale,
s^2 = 23,80
odnosno,
s = 4.88,
što je vrlo dobra procjena stvarne standardne devijacije σ=5.

2. Interval pouzdanosti za očekivanje – prava vrijednost mjerene veličine.

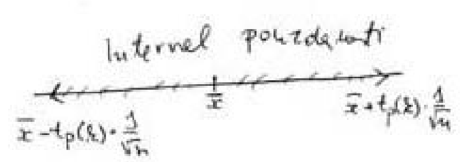
Očekivanje procjenjujemo aritmetičkom sredinom podataka, ali aritmetička sredina ne mora biti (i u pravilu nije) jednaka (nepoznatom) očekivanju. Zato nas zanima **interval** oko $\bar{x}$ unutar kojega će, uz određenu sigurnost, biti očekivanje μ. To je **interval pouzdanosti**.
Intuitivno je jasno:

1. širina intervala pouzdanosti ovisi o razini sigurnosti da se očekivanje nađe u njemu (što je ta razina veća, interval pouzdanosti je širi).
2. interval pouzdanosti je uži ako je broj mjerenja n veći (naravno, uz istu razinu sigurnosti).

3. Interval pouzdanosti je širi ako znamo neki dodatni podatak (na primjer, ako već poznajemo standardnu devijaciju σ, a ne samo njenu procjenu s).

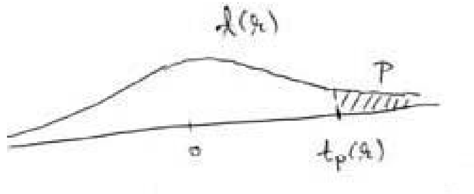
Uz pretpostavku da slučajna veličina X ima normalnu razdiobu, interval pouzdanosti, uz vjerojatnost 1-2p, je

$$< \bar{x} - t_p(k)\frac{s}{\sqrt{n}}, \bar{x} + t_p(k)\frac{s}{\sqrt{n}} > .$$



Tu je:
n broj podataka (duljina uzorka),
$\bar{x}$ je aritmetička sredina podataka,
s je korigirana standardna devijacija podataka,
k=n-1
t_p (k) je broj za koji vrijedi P(t(k)> t_p (k)) = p, gdje je *t(k)* Studentova razdioba s k stupnjeva slobode. Taj podatak dobivamo uporabom prikladnog statističkog kompjutorskog paketa.

Izraz $\frac{s}{\sqrt{n}}$ je procjena **standardne grješke** $\frac{\sigma}{\sqrt{n}}$.



Primjer 1. Iz n=16 mjerenja dobiveno je $\bar{x} = 12.44$, $s = 1.54$.
Odredimo interval pouzdanosti za vjerojatnost:
a) 0.95
b) 0.90

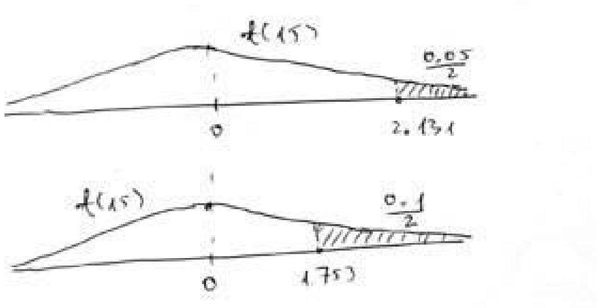
k=16-1 = 15

a) Tu je, prema prihvaćenim oznakama, 2p=0.05 , $t_{0.025}(15)$=2.131, $\frac{s}{\sqrt{n}} = \frac{1.54}{4}$

Interval pouzdanosti je:

$$< 12.44 - 2.131\frac{1.54}{4}, 12.44 + 2.131\frac{1.54}{4} >$$
= <11.62; 13.26>.

b) 2p=0.1, $t_{0.05}(15) = 1.753$.
Interval pouzdanosti je
$$< 12.44 - 1.753\frac{1.54}{4}, 12.44 + 1.753\frac{1.54}{4} >$$
= <11.77; 13.11>.
Taj je interval uži nego prethodni
(što je jasno jer je sad vjerojatnost manja).



Da je bilo n=4, a ostali podatci isti kao i prije, intervali pouzdanosti, uz istu vjerojatnost bili bi dva puta širi (jer bismo u standardnoj grješki dijelili s 2 umjesto s 4). To je prirodno (jer interval pouzdanosti treba biti to uži što je broj mjerenja veći).
Smisao intervala pouzdanosti (na primjer za razinu 95%) nije da se očekivanje μ u njemu nalazi s vjerojatnošću 0.95 (naime μ nije slučajna veličina i nalazi se ili ne nalazi u tom intervalu). Taj se smisao može interpretirati na primjer tako da bi se odprilike u 95 od 100 ponavljanja ovih n mjerenja, aritmetička sredina $\bar{x}$ našla u intervalu pouzdanosti.

Prava vrijednost mjerene veličine.
Gornji postupak mogli smo interpretirati i kao n mjerenja neke vrijednosti (na primjer, n mjerenja postotka šećera u krvi). Taj postotak je nepoznata veličina μ, a $x_1, x_2, ..., x_n$ su postotci dobiveni iz n nezavisnih ispitivanja, nekom metodom (i oni su približno jednaki stvarnom postotku). Prirodno je da kao najvjerodostojniji rezultat uzmemo aritmetičku sredinu $\bar{x}$ tih podataka, a interval pouzdanosti je interval unutar kojega se, uz odgovarajuću vjerojatnost, nalazi prava vriijednost μ (stvarni postotak šećera u krvi).

Pojašnjenje nastanka formule za interval pouzdanosti očekivanja.
Strogo matematički može se pokazati da se broj $\dfrac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$ može interpretirati kao slučajna vrijednost Studentove razdiobe t(k), gdje je k=n-1. Kako se taj broj ravnopravno pojavljuje i među negativnim i među pozitivnim brojevima, zbog simetričnosti t-razdiobe (Studentove),

vjerojatnost da se taj podatak nađe u intervalu <-$t_p(k)$,$t_p(k)$> jednaka je 1-2p. To se može (istina malo neprecizno) pisati kao $P(-t_p(k) < \frac{\mu - \bar{x}}{\frac{s}{\sqrt{n}}} < t_p(k)) = 1-2p$

To je dalje ekvivalentno s

$P(-t_p(k)\frac{s}{\sqrt{n}} < \mu - \bar{x} < t_p(k)\frac{s}{\sqrt{n}}) = 1-2p$, a to s

$P(\bar{x} - t_p(k)\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_p(k)\frac{s}{\sqrt{n}}) = 1-2p$,

A to je upravo ono što smo i htjeli

Predpostavka poznate standardne devijacije.
Sad ćemo ilustrirati što bi se dogodilo ako znademo standardnu devijaciju σ (iako je ta predpostavka nerealna). U tom slučaju ne radimo s procjenom *s* već sa σ. Efekt će biti suženje intervala pouzdanosti.

Veličina $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ zove se **standardna grješka**. Ona je to manja što je *n* veći (što je prirodno, jer što je broj mjerenja veći sigurnost prosjeka treba biti veća). Sad bi za interval pouzdanosti, umjesto $\frac{s}{\sqrt{n}}$ imali $\frac{\sigma}{\sqrt{n}}$, a umjesto $t_p(k)$ imali bismo z_p , a to je broj analogan onome $t_p(k)$, samo što se gleda na jediničnoj normalnoj razdiobi (i on ne ovisi o *k*).

Interval pouzdanosti uz vjerojatnost 1-2p bi bio: $< \bar{x} - z_p \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_p \cdot \frac{\sigma}{\sqrt{n}} >$.

Na primjer, za 95%-nu vjerojatnost, interval pouzdanosti je

$$< \bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}} >$$

jer je $z_{0.025}$ (bez obzira o kojem je broju mjerenja *n* riječ).

Napomena 1. Ako je *n* velik (obično se uzima ako je n>30), onda, ovako možemo postupiti bez obzira je li X bila normalno distribuirana, tj. interval pouzdanosti za očekivanje u tom slučaju približno je jednak intervalu

$$< \bar{x} - t_p(k)\frac{s}{\sqrt{n}}, \bar{x} + t_p(k)\frac{s}{\sqrt{n}} >.$$

Napomena 2. Ovdje smo razmatrali interval pouzdanosti za očekivanje. Analogno, mogli smo razmatrati i interval pouzdanosti za standardnu devijaciju (to nećemo obrađivati, spomenimo samo da tada interval ne bi bio simetričan oko *s*).