

Биостатистика и анализа података

Марија Цупарић

Седми час



Интервал поверења за параметар средње вредности

Видели смо да је узорачка средина $\bar{X}$ логична оцена средње вредности m .
Која је расподела од $\bar{X}$?

Теорема: Нека је $X_1, \dots, X_n$ узорак обима n из нормалне расподеле са параметрима m и σ^2 . Тада $\bar{X}$ има нормалну расподелу чија је средња вредост m и дисперзија $\frac{\sigma^2}{n}$.

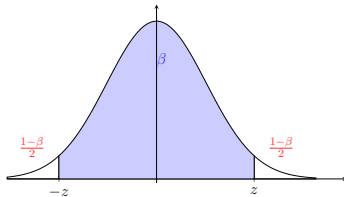
На основу стандардизације, $Z = \frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}}$ има стандардну нормалну расподелу.

Уколико узорак није из нормалне, него из неке друге расподеле чија је средња вредност m , а дисперзија σ^2 , онда Z добијено горњом формулом нема нормалну расподелу, али за велико n ($n > 30$) има приближно нормалну расподелу (што следи из ЦТГ).

Интервал поверења за m када је σ^2 познато

Нека је $X_1, \dots, X_n$ случајни узорак обима n из нормалне расподеле с параметрима m и познатом вредношћу σ^2 .

- ★ случајна променљива: $Z = \frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} \in \mathcal{N}(0, 1)$
- ★ треба одредити константу z тако да $P\{-z < Z < z\} = \beta$



- ★ z је такво да важи $P\{Z > z\} = \frac{1-\beta}{2}$ или $z = F^{-1}(\frac{1+\beta}{2})$
- ★ $-z < Z < z \iff -z < \frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} < z \iff \bar{X}_n - z \frac{\sigma}{\sqrt{n}} < m < \bar{X}_n + z \frac{\sigma}{\sqrt{n}}$

Интервал поверења је $(\bar{X}_n - z \frac{\sigma}{\sqrt{n}}, \bar{X}_n + z \frac{\sigma}{\sqrt{n}})$, а његова дужина је $d = 2z \frac{\sigma}{\sqrt{n}}$.

Интерпретација интервала поверења: Генерисаћемо узорке обима 10 из нормалне $\mathcal{N}(0, 1)$ и формирати 95% интервале поверења (за параметар m када је σ^2 познато). Када се ова процедура понови велики број пута у приближно 95% случајева стварна вредност параметра m (у нашем случају 0) треба да се нађе унутар реализованог интервала.

```
intervalPoverenjaM <- function(n,m,sigma,beta){
  xsr <- mean(rnorm(n,m,sigma))
  L <- xsr - qnorm((1+beta)/2)*sigma/sqrt(n)
  U <- xsr + qnorm((1+beta)/2)*sigma/sqrt(n)
  return(c(L,U))
}
brojac<-0
N<-10000
for( i in 1:N){
  intpov<-intervalPoverenjaM(10,0,1,0.95)
  if((intpov[1]<0)&&(intpov[2]>0))
    brojac<-brojac+1
}
brojac/N
```

Резултат је: 0.947



Пример Претпоставимо да студент мери температуру кључања (y °C) одређене течности и добио је резултате 102.5, 101.7, 103.1, 100.9, 100.5 и 102.2 на 6 различитих узорака течности. Ако температура кључања течности прати нормалну расподелу са стандардним одступањем 1.2 степена, одредити 95% интервал поверења за очекивање.

Интервал поверења је $(\bar{X}_n - z \frac{\sigma}{\sqrt{n}}, \bar{X}_n + z \frac{\sigma}{\sqrt{n}})$.

Одредимо реализоване вредности:

$$\star \bar{x}_6 = 101.8167$$

$$\star z = \Phi^{-1}\left(\frac{1+\beta}{2}\right) = \Phi^{-1}(0.975) = 1.96 \quad (\text{qnorm}(0.975))$$

$$\Rightarrow \left(101.8167 - 1.96 \frac{1.2}{\sqrt{6}}, 101.8167 + 1.96 \frac{1.2}{\sqrt{6}}\right) = (100.8565, 102.7769)$$



```
intervalPoverenjaM <- function(x,sigma,beta){
  xsr <- mean(x)
  n<-length(x)
  L <- xsr - qnorm((1+beta)/2)*sigma/sqrt(n)
  U <- xsr + qnorm((1+beta)/2)*sigma/sqrt(n)
  return(c(L,U))
}
x<-c(102.5, 101.7, 103.1, 100.9, 100.5, 102.2)
intervalPoverenjaM(x,1.2,0.95)
```

Студент се пита колики узорак би требало да узме да би 95% интервал поверења био ужи од 1.5.

Дакле, треба да важи

$$d < 1.5 \iff 2z \frac{\sigma}{\sqrt{n}} < 1.5 \iff n > 4z^2 \frac{\sigma^2}{1.5^2} \iff n > 4 \cdot 1.96 \frac{1.2^2}{1.5^2} = 9.83$$

Треба узети узорак од бар 10 елемената.

Интервал је добијен за конкретан узорак. За други узорак добили бисмо и други интервал.

Да ли очекујемо да је краћи 90% или 95% интервал поверења?



Шта да радимо ако σ^2 није познато? Оцењујемо га узорачком дисперзијом S^2 и онда се расподела мења.

Теорема: Нека је $X_1, \dots, X_n$ узорак обима n из нормалне расподеле са параметрима m и σ^2 . Тада $\frac{\sqrt{n}(\bar{X}_n - m)}{S}$ има Студентову t_{n-1} расподелу.

Особине Студентових расподела

- ★ Свака Студентова расподела има један параметар ν , број степени слободе
- ★ Студентова расподела је непрекидна
- ★ График је симетричан око нуле, средња вредност је нула
- ★ Параметар ν утиче на дисперзију, што је он већи, дисперзија је мања
- ★ Када је ν велико, Студентова расподела је приближна стандардној нормалној

Читање таблица стандардне Студентове расподеле

Студентова расподела - вредности x такве да је $P\{T_\nu < x\} = p$

	p											
ν	0.600	0.667	0.750	0.800	0.875	0.900	0.950	0.975	0.990	0.995	0.999	
1	0.325	0.577	1.000	1.376	2.414	3.078	6.314	12.706	31.821	63.657	318.31	
2	0.289	0.500	0.816	1.061	1.604	1.886	2.920	4.303	6.965	9.925	22.327	
3	0.277	0.476	0.765	0.978	1.423	1.638	2.353	3.182	4.541	5.841	10.215	
4	0.271	0.464	0.741	0.941	1.344	1.533	2.132	2.776	3.747	4.604	7.173	
5	0.267	0.457	0.727	0.920	1.301	1.476	2.015	2.571	3.365	4.032	5.893	
6	0.265	0.453	0.718	0.906	1.273	1.440	1.943	2.447	3.143	3.707	5.208	
...	...	...	...	...	...	...	...	...	...	...	...	
28	0.256	0.435	0.683	0.855	1.175	1.313	1.701	2.048	2.467	2.763	3.408	
29	0.256	0.435	0.683	0.854	1.174	1.311	1.699	2.045	2.462	2.756	3.396	
30	0.256	0.435	0.683	0.854	1.173	1.310	1.697	2.042	2.457	2.750	3.385	
...	...	...	...	...	...	...	...	...	...	...	...	
55	0.255	0.433	0.679	0.848	1.163	1.297	1.673	2.004	2.396	2.668	3.245	
60	0.254	0.433	0.679	0.848	1.162	1.296	1.671	2.000	2.390	2.660	3.232	
∞	0.253	0.431	0.674	0.842	1.150	1.282	1.645	1.960	2.326	2.576	3.090	

- ★ Тражимо x такво да је $P\{T_5 < x\} = 0.95$, из таблице: $x = 2.015$.
- ★ Тражимо x такво да је $P\{T_5 < x\} = 0.05$, површина је мања од 0.5, па је x негативно $P\{T_5 < -x\} = 0.95$, па је $x = -2.015$.
- ★ Тражимо x такво да је $P\{T_2 > x\} = 0.025$, онда је $P\{T_2 < x\} = 0.975$, па је $x = 4.303$

Случајна величине T_ν има Студентову t_ν расподелу

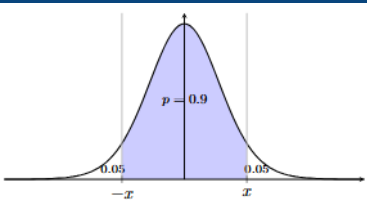
- ★ Вероватноћу догађаја $P\{T_\nu < x\}$ рачунамо функцијом $pt(x, df=\nu)$.

R `> pt(1.25, df=10)`
0.8801197

- ★ Вредност x за коју је $P\{T_\nu < x\} = p$ рачунамо функцијом $qt(p, df=\nu)$.

R `> qt(0.95, df=5)`
2.015048

Пример Ако X има Студентову расподелу са 18 степени слободе, одредити x тако да важи $P\{-x < X < x\} = 0.9$.

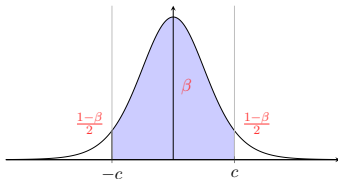


Видимо да је површина графика лево од x једнака 0.95 па x налазимо као $qt(0.95, df = 18)$, а то је 1.734.

Интервал поверења за m када је σ^2 непознато

Нека је $X_1, \dots, X_n$ случајни узорак обима n из нормалне расподеле с параметрима m и непознатом вредношћу σ^2 .

- ★ случајан величина: $T_n = \frac{\bar{X}_n - m}{\frac{S}{\sqrt{n}}} \in t_{n-1}$
- ★ треба одредити константу c тако да $P\{-c < T_n < c\} = \beta$



- ★ c је такво да важи $P\{T_n > c\} = \frac{1-\beta}{2}$
- ★ $-c < T_n < c \iff \bar{X}_n - c \frac{S}{\sqrt{n}} < m < \bar{X}_n + c \frac{S}{\sqrt{n}}$

Интервал поверења је $(\bar{X}_n - c \frac{S}{\sqrt{n}}, \bar{X}_n + c \frac{S}{\sqrt{n}})$.



Пример Како би се проучила ефикасност правосуђа проучава се обележје X које представља време протекло од подизања оптужнице до суђења. Претпоставља се да наведено време прати нормалну расподелу. На основу узорка од $n = 400$ времена, добијено је да $\bar{x} = 10.1$ месеци са узорачком дисперзијом $s^2 = 4.3$. Одредити 90% интервал поверења за просечно време које протекне од подизања оптужнице до суђења.

Интервал поверења је $(\bar{X}_n - c \frac{S}{\sqrt{n}}, \bar{X}_n + c \frac{S}{\sqrt{n}})$, где је $c = F_{t_{n-1}}^{-1}(\frac{1+\beta}{2})$.

С обзиром да је n велико, Студентова расподела се апроксимира нормалном: $c = F_{t_{399}}^{-1}(0.95) \approx \Phi^{-1}(0.95)$

У R -у се може одређивати или квантил стандардне нормалне или квантил Студентове расподеле.

$$c = 1.65$$

$$\Rightarrow \left(10.1 - 1.65 \sqrt{\frac{4.3}{400}}, 10.1 + 1.65 \sqrt{\frac{4.3}{400}} \right) = (9.93, 10.27).$$

Закључак је да верујемо, са поверењем од 90% да је време које протекне између подизања оптужнице и суђења између 9.93 и 10.27 месеци.

- ★ Тестирање статистичких хипотеза је поступак којим се, применом статистичких метода, утврђује да ли се на основу узорка може, и са којом вероватноћом, прихватити претпоставка о конкретној вредности неког параметра или о расподели обележја.
- ★ У сваком тесту се морају дефинисати:
 - ★ нулта H_0 и алтернативна H_1 хипотеза;
 - ★ критична област теста (или област одбацивања нулте хипотезе) W - део узорачког простора који указује на H_1 , ако реализовани узорак $(x_1, \dots, x_n) \in W$, онда се одбацује нулта хипотеза, у супротном се не одбацује;
 - ★ тест статистика T_n - на основу њених реализованих вредности доносимо закључке, јавља се у критичној области као функција узорка.



- ★ Од избора хипотеза зависи целокупна процедура
- ★ Алтернативна (или истраживачка) хипотеза је оно што тврдимо и желимо да статистички проверимо (обично садржи речи као веће, мање, зависи...)
- ★ Нулта хипотеза је супротна алтернативној (обично садржи речи једнако, мање или једнако, не зависи...)
- ★ Разликујемо
 - ★ параметарске хипотезе - тестирање је везано за параметар расподеле,
 - ★ непараметарске хипотезе - тестирање је везано за саму расподелу обележја.
- ★ Разликујемо
 - ★ просте хипотезе - у потпуности одређују статистички модел,
 - ★ сложене хипотезе - статистички модел зависи од неких параметара.
- ★ Оно што заправо желимо да покажемо је најбоље да буде у алтернативној хипотези



Да бисмо нешто одбацили довољно је да једном добијемо негативан резултат тестирања, док је за прихватање потребно да се прихвати са доста тестова!

Пример Размотримо експеримент довођења оптуженог на суђење. Обично се каже: „невин док се не докаже супротно“, односно суђење је једна врста статистичког теста. Поставимо хипотезе:

$$H_0(\text{невин}) \quad H_1(\text{крив}).$$

Сврха суђења је да се покаже да је крив.

Примери хипотеза у општем случају:

- ★ $H_0(X \in \mathcal{N}(m_0, \sigma_1^2)) \quad H_1(X \notin \mathcal{N}(m_0, \sigma_1^2))$
- ★ $H_0(\text{обележја } X \text{ и } Y \text{ су независна}) \quad H_1(\text{обележја } X \text{ и } Y \text{ нису независна})$
- ★ ако X има расподелу F_θ , онда $H_0(\theta = \theta_0) \quad H_0(\theta < \theta_0)$
- ★ $\vdots$

Пример На крају суђења, на основу одбране и оптужбе, доноси се један од закључака:

- ★ Невина особа је проглашена кривом
- ★ Крива особа је проглашена кривом
- ★ Невина особа је проглашена невином
- ★ Крива особа је проглашена невином

Две од четири одлуке су погрешне.

одлука суда	оптуђени је	
	невин	крив
крив	×	✓
невин	✓	×

Слично важи и у општем случају.

закључак тестирања	стварно стање	
	H_0 тачна	H_0 нетачна
одбацимо H_0	грешка прве врсте	✓
не одбацимо H_0	✓	грешка друге врсте



- ★ Закључак доносимо на основу вредности неке статистике, коју називамо тест статистиком. Ако она у нашем узорку узме вредност која је неуобичајена ако важи H_0 , онда одбацујемо H_0 .
- ★ Одбацивање H_0 је статистички значајна одлука, значи да смо скупили довољно доказа у прилог нашој алтернативној хипотези.
- ★ Неодбацивање H_0 није статистички значајан резултат. То може да значи да стварно важи H_0 , или да важи H_1 али да немамо довољно доказа њој у прилог.

Како донети одлуку?

- ▷ Једна могућност је задати унапред вредност α (најчешће 0.05), који називамо мером или прагом значајности теста, који ће нам фиксирати вероватноћу грешке прве врсте. Уколико нам тест статистика узме вредност која под H_0 има вероватноћу мању од α , одбацујемо H_0 .
- ▷ p —вредност теста је вероватноћа да извучемо неки узорак који је бољи доказ у корист наше алтернативне хипотезе од оног који смо већ извукли. Уколико је та вероватноћа мала, то значи да су наши докази одлични па одбацујемо H_0 . Граница је поново обично на 5%.



Тестирање хипотеза у вези са параметром средње вредности

Примери

- (1) Аутоматски постављач чуњева за куглање треба да постави чуњеве 4 секунде након што је прва лопта закотрљана. Ако је машина сувише спора, куглаши постају нестрпљиви, а ако је пребрза удариће цуњеве док падају. Тестирамо да ли машина ради како треба

$$H_0(m = 4) \quad H_1(m \neq 4)$$

- (2) Имамо рачунар на коме је за наш програм потребно 45 секунди да се изврши. Приликом куповине новог рачунара желимо да будемо сигурни да је он бољи. Тестирамо

$$H_0(m \geq 45) \quad H_1(m < 45)$$

- (3) У једној популацији јединки тврди се да је просечан животни век 5 година. Сумња се да је, услед побољшаних услова средине, нови услови довели до тога да је просечан животни век већи од претходно утврђеног. Тестирамо

$$H_0(m \leq 5) \quad H_1(m > 5)$$

ХИПОТЕЗЕ:

- ★ $H_0(m = m_0)$ против $H_1(m \neq m_0)$ (двострани тест)
- ★ $H_0(m \geq m_0)$ против $H_1(m < m_0)$ (једнострани леви)
- ★ $H_0(m \leq m_0)$ против $H_1(m > m_0)$ (једнострани десни)

ТЕСТ СТАТИСТИКА:

- ★ ако је σ^2 познато: $T_n = \frac{\bar{X}_n - m_0}{\frac{\sigma}{\sqrt{n}}}$ при H_0 има $\mathcal{N}(0, 1)$
- ★ ако је σ^2 непознато: $T_n = \frac{\bar{X}_n - m_0}{\frac{s}{\sqrt{n}}}$ при H_0 има t_{n-1}

p-вредност теста:

- ★ код једностраног левог теста је површина лево од вредности t_0 коју статистика T_n узме у узорку;
- ★ код једностраног десног теста је површина десно од вредности t_0 коју статистика T_n узме у узорку;
- ★ код двостраног теста је двострука површина лево од вредности t_0 , ако је $t_0 < 0$ или двострука површина десно од вредности t_0 , ако је $t_0 > 0$.

Ако се оцена $\bar{X}_n$ превише разликује од m_0 , имамо разлога да одбацимо H_0 .

Пример Аутоматски постављач чуњева за куглање треба да постави чуње-
ве 4 секунде након што је прва лопта закотрљана. Ако је машина сувише
спора, куглаши постају нестрпљиви, а ако је пребрза удариће цуњеve док
падају. Обе ситуације су непожељне. Машина је тестирана да би се про-
верио њен рад. Мерено је време постављања чуњева и добијени су резул-
тате 4.1, 2.5, 3.5, 3.8, 3.2, 4.6, 4.1, 3.0, 3.5, 4.1, 4.3, 3.6, 4.0, 3.7, 4.5, 3.9.
Претпоставља се да време рада има нормалну расподелу. Да ли овој ма-
шини треба поправка?

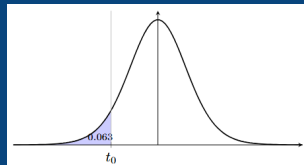
Поставимо хипотезе: $H_0(m = 4)$ $H_1(m \neq 4)$.

Тест статистика: $T_n = \sqrt{n} \frac{\bar{X} - 4}{S}$ при H_0 има Студентову t_{15}

Реализована вредност тест статистике: $t_0 = \sqrt{16} \frac{3.775 - 4}{0.5543} = -1.62$.

p —вредност теста:

$$\begin{aligned} 2P_{H_0}\{T_n \leq -1.62\} &= 2F_{t_{n-1}}(-1.62) \\ &= 2 \cdot 0.063 \\ &= 0.126. \end{aligned}$$



Пошто је p —вредност теста велика (већа од 0.05), закључујемо да нема
доказа да машини треба поправка.

```
x<-c(4.1, 2.5, 3.5, 3.8, 3.2, 4.6, 4.1, 3.0, 3.5, 4.1, 4.3, 3.6, 4.0,
      3.7, 4.5, 3.9)
R (mean(x)-4)/(sd(x)/sqrt(length(x)))
qt(1-0.05/2,15)
2*min(pt(-1.62,15),1-pt(-1.62,15))
t.test(x, alternative = "two.sided", mu=4)
```

One sample t-test

```
data: x
t = -1.6234, df = 15, p-value = 0.1253
alternative hypothesis: true mean is not equal to 4
95 percent confidence interval:
 3.479594 4.070406
sample estimates:
mean of x
 3.775
```

Функција у R —у за тест када је дисперзија позната:

```
R z.test() % paket BSDA
```

Пример На рачунару који поседујемо време извршавања одређениог програма је 45 секунди. Да бисмо купили нови рачунар потребно је да на њему време извршавања буде мање. Ако је на основу узорка од 30 рачунара одређеног типа добијено $\bar{x} = 44.5$, $s = 2$, испитати да ли треба купити рачунар те фирме.

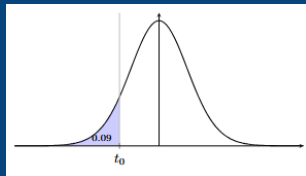
Поставимо хипотезе: $H_0(m \geq 45)$ $H_1(m < 45)$.

Тест статистика: $T_n = \sqrt{n} \frac{\bar{X} - 45}{S}$ има Студентнову t_{29}

Реализована вредност тест статистике: $t_0 = \sqrt{30} \frac{44.5 - 45}{2} = -1.37$.

p —вредност теста:

$$P_{H_0}\{T_n \leq -1.37\} = F_{t_{29}}(-1.37) = 0.09.$$



Одлука је на нама, ако сматрамо да је грешка од 9% велика, закључићемо да не треба одбацити H_0 и не треба купити нови рачунар, а ако мислимо да је мала, онда ћемо купити нови рачунар.

Пример За једну врсту биљака познато је да је просечна висина 50 *cm*. Сумња се да су нове агроеколошке мере довеле до повећања просечне висине. У експерименту је измерена висина 15 биљака (*y cm*): 44, 54, 48, 46, 53, 53, 50, 54, 56, 54, 56, 53, 54, 51, 53. Да ли се може закључити да су нове мере довеле до повећања просечне висине биљака?

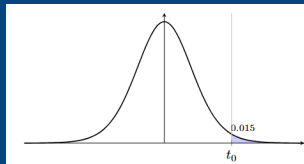
Поставимо хипотезе: $H_0(m \leq 50)$ $H_1(m > 50)$.

Тест статистика: $T_n = \sqrt{n} \frac{\bar{X} - 50}{S}$ има Студентнову t_{14}

Реализована вредност тест статистике: $t_0 = \sqrt{15} \frac{51.93 - 50}{3.51} = 2.13$.

p –вредност теста:

$$1 - P_{H_0}\{T_n \leq 2.13\} = 1 - F_{t_{14}}(2.13) = 0.026.$$



Закључујемо да треба одбацити H_0 , односно модел јесте побољшан.



```
t.test(x, alternative = "greater", mu=50)
```

Доношење одлука на основу прага значајности, односно максималне грешке коју толеришемо при одбацавању нулте хипотезе.

Ако је дат праг α , онда ако је p —вредност теста мања од α одбацујемо нулту хипотезу, а ако је p —вредност теста већа од α , немамо довољно доказа да одбацимо нулту хипотезу.

Примери (наставак) Нека је $\alpha = 0.05$

- (1) Аутоматски постављач чуњева за куглање: добијена је p —вредност теста 0.126, што значи да не одбацујемо нулту хипотезу.
- (2) Рачунар: добијена је p —вредност теста 0.09, што значи да не одбацујемо нулту хипотезу при прагу α .
- (3) Висина биљке: добијена је p —вредност теста 0.026, што значи да одбацујемо нулту хипотезу.