

Анализа података у биологији

Марија Цупарић

Девети час



Поређење средњих вредности две нормалне популације

Пример

- ★ У истраживању једне врсте биљака посматрају се две групе: биљке узгајане у стандардним условима (група А) и биљке узгајане уз примену нове методе ђубрења (група Б). Занима нас да ли биљке из ове две групе у просеку имају исту висину, при чему су дисперзије висине у обе групе познате из ранијих истраживања.
- ★ Професор тврди да нема разлике у резултатима на тесту између студената и студенткиња. Међутим, његов колега тврди да студенткиње у просеку имају боље резултате од студената.

У општем случају, имамо обележја X са нормалном $\mathcal{N}(m_1, \sigma_1^2)$ расподелом и Y са нормалном $\mathcal{N}(m_2, \sigma_2^2)$ расподелом. Желимо да оценимо или тестирамо разлику њихових очекивања, $m_1 - m_2$.

- ★ Тачкаста оцена је разлика узорачких средина, $\widehat{m_1 - m_2} = \bar{X}_{n_1} - \bar{Y}_{n_2}$.
- ★ За одређивање интервала поверења и тестирање хипотеза морамо разликовати да ли су узорци независни или спарени.

Нека су дати независни узорци $X_1, \dots, X_{n_1}$ из популације чије обележје X има нормалну $\mathcal{N}(m_1, \sigma_1^2)$ расподелу и $Y_1, \dots, Y_{n_2}$ из популације чије обележје Y има нормалну $\mathcal{N}(m_2, \sigma_2^2)$ расподелу.

Разликујемо случајеве:

- (1) када су σ_1^2 и σ_2^2 познате
- (2) када су σ_1^2 и σ_2^2 непознате и једнаке
- (3) када су σ_1^2 и σ_2^2 непознате и неједнаке

У случајевима (2) и (3), важно је најпре испитати једнакост дисперзија и примењујемо тест о једнакости дисперзија (прошли час).

С обзиром да је последица већа ако одлучимо да не одбацимо H_0 , тестирамо с неуобичајено великим прагом значајности $\alpha = 0.2$, тако да нам треба p -вредност теста од бар 20% да бисмо закључили да су дисперзије једнаке. Уколико закључимо да су дисперзије једнаке, примењујемо процедуру с оцењивањем заједничке дисперзије, а у супротном Сатервајтову процедуру.



Поређење очекивања ако су σ_1^2 и σ_2^2 познате

Теорема: Нека су дати независни узорци $X_1, \dots, X_{n_1}$ из популације чије обележје X има нормалну $\mathcal{N}(m_1, \sigma_1^2)$ расподелу и $Y_1, \dots, Y_{n_2}$ из популације чије обележје Y има нормалну $\mathcal{N}(m_2, \sigma_2^2)$ расподелу. Тада

$$\frac{\bar{X}_{n_1} - \bar{Y}_{n_2} - (m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \in \mathcal{N}(0, 1).$$

$100\beta\%$ интервал поверења за разлику средњих вредности $m_1 - m_2$ је

$$\left(\bar{X}_{n_1} - \bar{Y}_{n_2} - z \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \bar{X}_{n_1} - \bar{Y}_{n_2} + z \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right),$$

где је z тако да је $P\{Z > z\} = \frac{1-\beta}{2}$ за $Z \in \mathcal{N}(0, 1)$.

Тестирање хипотеза о разлици очекивања ако су σ_1^2 и σ_2^2 познате

хипотезе:

- ★ $H_0(m_1 - m_2 = m_0)$ против $H_1(m_1 - m_2 \neq m_0)$ (двострани тест)
- ★ $H_0(m_1 - m_2 = m_0)$ против $H_1(m_1 - m_2 < m_0)$ (једнострани леви)
- ★ $H_0(m_1 - m_2 = m_0)$ против $H_1(m_1 - m_2 > m_0)$ (једнострани десни)

тест статистика: $T_{n_1, n_2} = \frac{\bar{X}_{n_1} - \bar{Y}_{n_2} - m_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ при H_0 има $\mathcal{N}(0, 1)$

p -вредност теста:

- ★ код једностраног левог теста је површина лево од вредности t_0 коју статистика T_{n_1, n_2} узме у узорку;
- ★ код једностраног десног теста је површина десно од вредности t_0 коју статистика T_{n_1, n_2} 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 $t_0 > 0$.



Пример У истраживању једне врсте биљака посматрају се две групе: биљке узгајане у стандардним условима (група А) и биљке узгајане уз примену нове методе ђубрења (група Б). Занима нас да ли биљке из ове две групе у просеку имају исту висину. Узети су узорци од 50 биљака из групе А и 40 биљака из групе Б и добијени су резултати $\bar{x}_A = 52.1$ cm и $\bar{y}_B = 50.8$ cm. Из ранијих истраживања познато је да су дисперзије $\sigma_A^2 = 4$ и $\sigma_B^2 = 9$.

тачкаста оцена: $\widehat{m_A - m_B} = \bar{x}_A - \bar{y}_B = 1.3$

95% интервал поверења: $z = \Phi^{-1}(0.975)$ (qnorm(0.975))

$$\left(\bar{x}_A - \bar{y}_B - z \sqrt{\frac{\sigma_A^2}{n_1} + \frac{\sigma_B^2}{n_2}}, \bar{x}_A - \bar{y}_B + z \sqrt{\frac{\sigma_A^2}{n_1} + \frac{\sigma_B^2}{n_2}} \right) = (0.2176, 2.3824)$$

тестирање хипотезе $H_0(m_A = m_B)$:

алтернатива: $H_1(m_A \neq m_B)$

реализована вредност тест статистике: $\frac{\bar{x}_A - \bar{y}_B - 0}{\sqrt{\frac{\sigma_A^2}{n_1} + \frac{\sigma_B^2}{n_2}}} = 2.3539$

p-вредност теста: $2P\{T > 2.3539\} = 0.0185$

закључак: са нивоом значајности од 5% одбацујемо нулту хипотезу

С обзиром да интервал поверења не садржи 0, закључак на основу интервала се поклапа са закључком добијеним тестирањем.

Поређење очекивања ако су σ_1^2 и σ_2^2 непознате и једнаке

Теорема: Нека су дати независни узорци $X_1, \dots, X_{n_1}$ из популације чије обележје X има нормалну $\mathcal{N}(m_1, \sigma^2)$ расподелу и $Y_1, \dots, Y_{n_2}$ из популације чије обележје Y има нормалну $\mathcal{N}(m_2, \sigma^2)$ расподелу. Тада

$$\frac{\bar{X}_{n_1} - \bar{X}_{n_2} - (m_1 - m_2)}{S_z \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \in t_{n_1+n_2-2},$$

где је $S_z = \frac{(n_1-1)S_{n_1}^2 + (n_2-1)S_{n_2}^2}{n_1+n_2-2}$.

100 $\beta\%$ интервал поверења за разлику средњих вредности $m_1 - m_2$ је

$$\left(\bar{X}_{n_1} - \bar{Y}_{n_2} - tS_z \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \bar{X}_{n_1} - \bar{Y}_{n_2} + tS_z \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right),$$

где је t тако да је $P\{T > t\} = \frac{1-\beta}{2}$ за $T \in t_{n_1+n_2-2}$.

Тестирање хипотеза о разлици очекивања ако је $\sigma_1^2 = \sigma_2^2 = \sigma^2$ непозната

хипотезе:

- ★ $H_0(m_1 - m_2 = m_0)$ против $H_1(m_1 - m_2 \neq m_0)$ (двострани тест)
- ★ $H_0(m_1 - m_2 = m_0)$ против $H_1(m_1 - m_2 < m_0)$ (једнострани леви)
- ★ $H_0(m_1 - m_2 = m_0)$ против $H_1(m_1 - m_2 > m_0)$ (једнострани десни)

тест статистика: $T_{n_1, n_2} = \frac{\bar{X}_{n_1} - \bar{X}_{n_2} - m_0}{S_z \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$, где је $S_z^2 = \frac{(n_1 - 1)S_{n_1}^2 + (n_2 - 1)S_{n_2}^2}{n_1 + n_2 - 2}$, при

H_0 има $t_{n_1 + n_2 - 2}$

p -вредност теста:

- ★ код једностраног левог теста је површина лево од вредности t_0 коју статистика T_{n_1, n_2} узме у узорку;
- ★ код једностраног десног теста је површина десно од вредности t_0 коју статистика T_{n_1, n_2} 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 $t_0 > 0$.



Пример Професор тврди да нема разлике у резултатима на тесту између студената и студенткиња. Његов колега тврди да студенткиње у просеку имају боље резултате од студената. Да би се ово испитало, узети су резултати обе групе

студенти	50	55	70	120	150	80	81	30	65	90
студенткиње	62	65	75	40	100	110	90	87	57	140

Испитати ко је у праву. Претпоставља се да резултати прате нормалну расподелу.

X — резултати теста студената, $X \in \mathcal{N}(m_1, \sigma_1^2)$

Y — резултати теста студенткиња, $Y \in \mathcal{N}(m_2, \sigma_2^2)$

тачкаста оцена: $\widehat{m_1 - m_2} = \bar{x} - \bar{y} = -3.5$

Пошто не знамо да ли су дисперзије једнаке за даљу анализу, проверићемо прво то.



Тестирамо: $H_0 \left(\frac{\sigma_1^2}{\sigma_2^2} = 1 \right)$ $H_1 \left(\frac{\sigma_1^2}{\sigma_2^2} \neq 1 \right)$, за ниво значајности ћемо узети 0.2

Тест статистика: $T_{n_1, n_2} = \frac{S_{n_1}^2}{S_{n_2}^2}$ при H_0 има $\mathcal{F}_{9,9}$

Реализована вредност тест статистике: $\frac{s_{n_1}}{s_{n_2}} = \frac{1215.878}{853.8222} = 1.42$.

p -вредност теста је приближно: $2P\{T_{n_1, n_2} > 1.42\} = 0.61$

Одлука: пошто је p -вредност велика, нема доказа да треба одбацити H_0 , односно можемо претпоставити да су дисперзије једнаке



```
x<-c(50, 55, 70, 120, 150, 80, 81, 30, 65, 90)
y<-c(62, 65, 75, 40, 100, 110, 90, 87, 57, 140)
2(1-pf(1.42,9,9))
var.test(x,y)
```

F test to compare two variances

```
data: x and y
F = 1.424, num df = 9, denom df = 9, p-value = 0.6069
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.3537116 5.7331804
sample estimates:
ratio of variances
 1.424041
```

Вратимо се основном проблему!

95% интервал поверења (једностран): $t = F_{t_{18}}^{-1}(0.95) = qt(0.95, 18)$,

$$s_z^2 = \frac{(n_1-1)s_{n_1}^2 + (n_2-1)s_{n_2}^2}{n_1+n_2-2} = \frac{9 \cdot 1215.878 + 9 \cdot 853.8222}{18} = 1034.85$$

$$\Rightarrow \left(-\infty, \bar{x} - \bar{y} + ts_z \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right) = (-\infty, 22.79647)$$

Интервал поверења садржи нулу, па не можемо рећи да се средње вредности значајно разликују.

тестирање хипотезе $H_0(m_1 = m_2)$:

алтернатива: $H_1(m_1 < m_2)$

тест статистика: $T_{n_1, n_2} = \frac{\bar{X} - \bar{Y} - m_0}{s_z \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ при H_0 има $t_{n_1+n_2-2}$

реализована вредност тест статистике: $\frac{\bar{x} - \bar{y} - 0}{s_z \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{79.1 - 82.6}{32.17 \cdot 0.47} = -0.23$

p-вредност теста: $P\{T_{n_1, n_2} < -0.23\} = 0.41$

закључак: нема доказа да треба одбацити нулту хипотезу, односно не може се рећи да студенткиње имају боље резултате у просеку





```
(mean(x)-mean(y))/(sqrt((9*var(x)+9*var(y))/18)*sqrt(2/9))  
pt(-0.23,18)  
t.test(x,y, var.equal = TRUE, alternative = "less")
```

Two sample t-test

```
data: x and y  
t = -0.24328, df = 18, p-value = 0.4053  
alternative hypothesis: true difference in means is less than 0  
95 percent confidence interval:  
-Inf 21.44702  
sample estimates:  
mean of x mean of y  
79.1 82.6
```



Поређење очекивања ако су σ_1^2 и σ_2^2 непознате и различите

Теорема: Нека су дати независни узорци $X_1, \dots, X_{n_1}$ из популације чије обележје X има нормалну $\mathcal{N}(m_1, \sigma_1^2)$ расподелу и $Y_1, \dots, Y_{n_2}$ из популације чије обележје Y има нормалну $\mathcal{N}(m_2, \sigma_2^2)$ расподелу. Тада

$$\frac{\bar{X}_{n_1} - \bar{Y}_{n_2} - (m_1 - m_2)}{\sqrt{\frac{S_{n_1}^2}{n_1} + \frac{S_{n_2}^2}{n_2}}}$$

има приближно Студентову t_η расподелу, где је $\eta = \frac{\left(\frac{S_{n_1}^2}{n_1} + \frac{S_{n_2}^2}{n_2}\right)^2}{\frac{\left(\frac{S_{n_1}^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_{n_2}^2}{n_2}\right)^2}{n_2 - 1}}$.

- ★ У пракси η неће бити цео број, па се заокружује на најближи цео број
- ★ Наведени поступак назива се Сатервајтовом процедуром

100 β % интервал поверења за разлику средњих вредности $m_1 - m_2$ је

$$\left(\bar{X}_{n_1} - \bar{Y}_{n_2} - t \sqrt{\frac{S_{n_1}^2}{n_1} + \frac{S_{n_2}^2}{n_2}}, \bar{X}_{n_1} - \bar{Y}_{n_2} + t \sqrt{\frac{S_{n_1}^2}{n_1} + \frac{S_{n_2}^2}{n_2}} \right),$$

где је t тако да је $P\{T > t\} = \frac{1-\beta}{2}$ за $T \in t_\eta$.



Тестирање хипотеза о разлици очекивања ако је $\sigma_1^2 \neq \sigma_2^2$ и непознате су

хипотезе:

- ★ $H_0(m_1 - m_2 = m_0)$ против $H_1(m_1 - m_2 \neq m_0)$ (двострани тест)
- ★ $H_0(m_1 - m_2 = m_0)$ против $H_1(m_1 - m_2 < m_0)$ (једнострани леви)
- ★ $H_0(m_1 - m_2 = m_0)$ против $H_1(m_1 - m_2 > m_0)$ (једнострани десни)

тест статистика:

$$T_{n_1, n_2} = \frac{\bar{X}_{n_1} - \bar{Y}_{n_2} - m_0}{\sqrt{\frac{s_{n_1}^2}{n_1} + \frac{s_{n_2}^2}{n_2}}} \text{ при } H_0 \text{ има } t_\eta, \text{ где је } \eta = \frac{\left(\frac{s_{n_1}^2}{n_1} + \frac{s_{n_2}^2}{n_2}\right)^2}{\frac{\left(\frac{s_{n_1}^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_{n_2}^2}{n_2}\right)^2}{n_2 - 1}}$$

p -вредност теста:

- ★ код једностраног левог теста је површина лево од вредности t_0 коју статистика T_{n_1, n_2} узме у узорку;
- ★ код једностраног десног теста је површина десно од вредности t_0 коју статистика T_{n_1, n_2} 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 $t_0 > 0$.

Пример Постоји извесна забринутост да рекламне паузе на телевизији постојају дужине. Забележене су дужине рекламних блокова (у минутима) током две сезоне.

популација I	2.42	2.32	2.40	2.23	2.00	1.84	1.47	1.95
(прва сезона)	1.17	2.16	2.82	1.38	1.18	2.35	2.04	2.42
популација II	2.28	2.64	2.29	2.25	2.36	2.62	2.39	2.31
(друга сезона)	2.05	2.39	2.11	2.44	2.45	2.63	2.04	2.57

Да ли дати подаци поткрепљују тврдњу?

X — резултати на популацији I, $X \in \mathcal{N}(m_1, \sigma_1^2)$

Y — резултати на популацији II, $Y \in \mathcal{N}(m_2, \sigma_2^2)$

тачкаста оцена: $\widehat{m_1 - m_2} = \bar{x} - \bar{y} = -0.354375$

Пошто не знамо да ли су дисперзије једнаке за даљу анализу, проверићемо прво то.



Тестирамо: $H_0 \left(\frac{\sigma_1^2}{\sigma_2^2} = 1 \right)$ $H_1 \left(\frac{\sigma_1^2}{\sigma_2^2} \neq 1 \right)$, за ниво значајности ћемо узети 0.2

Тест статистика: $T_{n_1, n_2} = \frac{S_{n_1}^2}{S_{n_2}^2}$ при H_0 има $\mathcal{F}_{15, 15}$

Реализована вредност тест статистике: $\frac{S_{n_1}^2}{S_{n_2}^2} = 6.26$.

p -вредност теста је приближно: $2P\{T_{n_1, n_2} > 6.26\} = 0.001$

Одлука: пошто је p -вредност мала, треба одбацити H_0 , односно не можемо претпоставити да су дисперзије једнаке



```
x<-c(2.42, 2.32, 2.40, 2.23, 2.00, 1.84, 1.47, 1.95, 1.17, 2.16, 2.82,
      1.38, 1.18, 2.35, 2.04, 2.42)
y<-c(2.28, 2.64, 2.29, 2.25, 2.36, 2.62, 2.39, 2.31, 2.05, 2.39, 2.11,
      2.44, 2.45, 2.63, 2.04, 2.57)
var.test(x,y)
```

F test to compare two variances

```
data: x and y
F = 6.2609, num df = 15, denom df = 15, p-value = 0.001003
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 2.187514 17.919178
sample estimates:
ratio of variances
 6.260866
```

Вратимо се основном проблему!

95% интервал поверења (једнострани): $\eta = \frac{\left(\frac{s_{n_1}^2}{n_1} + \frac{s_{n_2}^2}{n_2}\right)^2}{\left(\frac{s_{n_1}^2}{n_1}\right)^2 + \left(\frac{s_{n_2}^2}{n_2}\right)^2} \approx 20,$

$$t = F_{t_{20}}^{-1}(0.95) = 1.72 \text{ (qt(0.95, 20))},$$

$$\Rightarrow \left(-\infty, \bar{x}_{n_1} - \bar{y}_{n_2} + t\sqrt{\frac{s_{n_1}^2}{n_1} + \frac{s_{n_2}^2}{n_2}}\right) = (-\infty, -0.1287)$$

Интервал поверења не садржи нулу, па можемо рећи да се средње вредности значајно разликују.

тестирање хипотезе $H_0(m_1 = m_2)$:

алтернатива: $H_1(m_1 < m_2)$

тест статистика: $T_{n_1, n_2} = \frac{\bar{x}_{n_1} - \bar{y}_{n_2} - m_0}{\sqrt{\frac{s_{n_1}^2}{n_1} + \frac{s_{n_2}^2}{n_2}}}$ при H_0 има t_η

реализована вредност тест статистике: $\frac{\bar{x}_{n_1} - \bar{y}_{n_2} - 0}{\sqrt{\frac{s_{n_1}^2}{n_1} + \frac{s_{n_2}^2}{n_2}}} = -2.7$

p -вредност теста: $P\{T_{n_1, n_2} < -2.7\} = 0.007$

закључак: треба одбацити нулту хипотезу, односно може се рећи да је у просеку дужина реклама у другој сезони већа



```
t.test(x,y, var.equal = FALSE, alternative = "less")
```

welch Two Sample t-test

data: x and y

t = -2.7082, df = 19.672, p-value = 0.006831

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

-Inf -0.1285062

sample estimates:

mean of x mean of y

2.009375 2.363750



- ★ Некада је природно да сваки елемент једног узорака има свој пар у другом узорку
- ★ Спаривање умањује утицај неке спољне променљиве која нам може сметати да откријемо стварну разлику у средњим вредностима

Пример Да би се проверило да ли одређена крема за сунчање X је ефикаснија од конкурентске Y узето је 25 испитаника којима је на једну руку и једну ногу намазана прва крема, а на другу руку и другу ногу конкурентска крема. Након 3 сата излагања сунцу проверен је степен опекотина мерењем температуре и боје коже на обе стране. На овај начин неутралисан је утицај типа коже на резултате.

У овом случају ми заправо имамо узорак парова $(X_1, Y_1), \dots, (X_n, Y_n)$



Теорема: Нека су $X_1, \dots, X_n$, из популације чије обележје X има нормалну $\mathcal{N}(m_1, \sigma_1^2)$ расподелу, и $Y_1, \dots, Y_n$, из популације чије обележје Y има нормалну $\mathcal{N}(m_2, \sigma_2^2)$ расподелу, спарени узорци. Дефинишемо променљиву $D = X - Y$ која има нормалну $\mathcal{N}(m_D, \sigma_D^2)$, $m_D = m_1 - m_2$, расподелу. Тада

$$T_n = \frac{\bar{D}_n - (m_1 - m_2)}{\frac{S_D}{\sqrt{n}}} \in t_{n-1}.$$

$100\beta\%$ интервал поверења за разлику средњих вредности $m_1 - m_2$ је

$$\left(\bar{D} - t \sqrt{\frac{S_D^2}{n}}, \bar{D} + t \sqrt{\frac{S_D^2}{n}} \right),$$

где је t тако да је $P\{T > t\} = \frac{1-\beta}{2}$ за $T \in t_{n-1}$.



хипотезе:

- ★ $H_0(m_1 - m_2 = m_0)$ против $H_1(m_1 - m_2 \neq m_0)$ (двострани тест)
- ★ $H_0(m_1 - m_2 = m_0)$ против $H_1(m_1 - m_2 < m_0)$ (једнострани леви)
- ★ $H_0(m_1 - m_2 = m_0)$ против $H_1(m_1 - m_2 > m_0)$ (једнострани десни)

Хипотезе се свакако могу дефинисати и као $H_0(m_D = m_0)$ и аналогно за све остале.

тест статистика: $T_n = \frac{\bar{D}_n - m_0}{\frac{s_D}{\sqrt{n}}}$ 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 $t_0 > 0$.



Пример Да би се проверило да ли одређена крема за сунчање X је ефикаснија од конкурентске Y узето је 25 испитаника којима је на једну руку и једну ногу намазана прва крема, а на другу руку и другу ногу конкурентска крема. Након 3 сата излагања сунцу проверен је степен опекотина мерењем температуре и боје коже на обе стране. Добијени су резултати:

нова крема	1.3	6.0	4.3	19.1	7.5	2.0	5.0	7.9	8.9	9.2	6.2	3.0	6.9
	7.6	8.2	15.3	14.9	6.1	7.9	17.5	6.1	5.1	13.7	14.2	18.1	
конкурентска крема	7.1	7.5	2.0	19.3	4.3	7.5	6.0	8.3	8.7	11.3	7.5	2.5	7.1
	8.3	6.9	15.7	13.8	7.3	8.3	17.9	7.3	4.9	13.5	17.1	19.2	

Да ли се може тврдити да је нова крема боља?

Ако пређемо на случајну величину $D = X - Y$ имамо узорак:

-5.8, -1.5, 2.3, -0.2, 3.2, -5.5, -1.0, -0.4, 0.2, -2.1, -1.3, 0.5, -0.2, -0.7, 1.3, -0.4, 1.1, -1.2, -0.4, -0.4, -1.2, 0.2, 0.2, -2.9, -1.1

тачкаста оцена: $\widehat{m}_D = \bar{d} = -0.692$

95% интервал поверења (једностранни): $t = F_{t_{24}}^{-1}(0.95) = 1.71$
 $(qt(0.95, 20)),$

$$\Rightarrow \left(-\infty, \bar{d} + t \sqrt{\frac{s_D^2}{n}} \right) = (-\infty, -0.016)$$

Поставимо хипотезе: $H_0(m_X = m_Y)$ $H_1(m_X < m_Y)$.

тестирање хипотезе $H_0(m_D = 0)$:

алтернатива: $H_1(m_D < 0)$

тест статистика: $T_n = \frac{\bar{D}_n - m_0}{\frac{s_D}{\sqrt{n}}}$ при H_0 има t_{n-1}

реализована вредност тест статистике: $\frac{\bar{d}_n - 0}{\frac{s_d}{\sqrt{n}}} = \frac{-0.692}{\frac{1.98}{\sqrt{25}}} = -1.75$

p -вредност теста: $P\{T_n < -1.75\} = 0.046$

закључак: са нивоом значајности 0.05, закључујемо да наша крема, у просеку, ефикасније штити кожу од конкурентске



```
x<-c(1.3, 6.0, 4.3, 19.1, 7.5, 2.0, 5.0, 7.9, 8.9, 9.2, 6.2, 3.0, 6.9,
      7.6, 8.2, 15.3, 14.9, 6.1, 7.9, 17.5, 6.1, 5.1, 13.7, 14.2, 18.1)
y<-c(7.1, 7.5, 2.0, 19.3, 4.3, 7.5, 6.0, 8.3, 8.7, 11.3, 7.5, 2.5, 7.1,
      8.3, 6.9, 15.7, 13.8, 7.3, 8.3, 17.9, 7.3, 4.9, 13.5, 17.1, 19.2)
t.test(x,y, paired = TRUE, alternative = "less")
```

Paired t-test

```
data: x and y
t = -1.7504, df = 24, p-value = 0.04641
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
 -Inf -0.01561205
sample estimates:
mean of the differences
 -0.692
```

- ★ У популацији сваки члан се може класификовати у односу на то да ли има одређено својство или га нема
- ★ Параметар p представља проценат (тј. удео) популације који има то својство
- ★ Случајна променљива X узима вредност 1 на елементима популације који имају то својство, а 0 на оним који га немају

Пример У истраживању једне популације јединки жели се проверити да ли је проценат јединки које имају одређену карактеристику већи од 2%. Посматрана променљива означава да ли јединка има ту карактеристику или не.



Тачкаста оцена параметра p је

$$\hat{p} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i,$$

где је $\sum_{i=1}^n X_i$, у ствари, број елемената узорка који имају испитивано својство

Пример Узето је 500 јединки и утврђено је да 18 има карактеристику. Ако је p проценат јединки са карактеристиком, оцена процента је $\bar{x} = \frac{18}{500} = 0.036$. Дакле, процена је да је 3.6% јединки са карактеристиком.

- ★ Случајна величина $\sum_{i=1}^n X_i$ има биномну $\mathcal{B}(n, p)$ расподелу
- ★ $\bar{X}$ је непристрасна оцена параметра p , $E\bar{X} = p$
- ★ $D\bar{X} = \frac{p(1-p)}{n}$

Интервал поверења за пропорцију

★ случајна величина: $T_n = \frac{\bar{X}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathcal{N}(0, 1)$ за велико n

★ треба одредити константу c тако да $P\{-c < T_n < c\} = \beta$

★ c је такво да $P\{Z > c\} = \frac{1-\beta}{2}, Z \in \mathcal{N}(0, 1)$

$$\begin{aligned} \star -c < T_n < c &\iff |T_n|^2 < c^2 \iff \left(\frac{\bar{X}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \right)^2 < c^2 \\ &\iff \bar{X}_n^2 - 2p\bar{X}_n + p^2 < c^2 \frac{p}{n} - c^2 \frac{p^2}{n} \\ &\iff p^2 \left(\frac{c^2}{n} + 1 \right) - p \left(\frac{c^2}{n} + 2\bar{X}_n \right) + \bar{X}_n^2 < 0 \end{aligned}$$

Интервал поверења је (p_1, p_2) , где су p_1 и p_2 решења претходне квадратне једначине.



Када је n велико и када p није блиско 0 или 1 (односно када $n\hat{p}$ није много мало, најчешће није мање од 5):

★ случајна величина: $T_n = \frac{\bar{X}_n - p}{\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}} \sim \mathcal{N}(0, 1)$

★ треба одредити константу c тако да $P\{-c < T_n < c\} = \beta$

★ c је такво да $P\{Z > c\} = \frac{1-\beta}{2}$, $Z \in \mathcal{N}(0, 1)$

★ $-c < T_n < c \iff \bar{X}_n - c\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} < p < \bar{X}_n + c\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}$

Интервал поверења је $\left(\bar{X}_n - c\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}, \bar{X}_n + c\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} \right)$ а дужина интервала је $d = 2c\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}$.

Ако је доња граница мања од 0 или горња граница већа од 1, интервал треба редуковати.



Пример У истраживању једне популације јединки жели се проверити да ли је проценат јединки које имају одређену карактеристику већи од 2%. Узето је 500 јединки и утврђено је да је 18 са карактеристиком. Одредити 95% интервал поверења за удео јединки са карактеристиком.

С обзиром да је $\hat{p} = 0.036$, а $n\hat{p} = 18$ и $n(1 - \hat{p}) = 482$, може користити приближни интервал поверења за параметар p .

Интервал поверења је $(\bar{X}_n - c\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}, \bar{X}_n + c\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}})$, где је $c = \Phi^{-1}(\frac{1+\beta}{2}) = 1.96$ (`qnorm(0.975)`).

$$\left(0.036 - 1.96\sqrt{\frac{0.036 \cdot 0.964}{500}}, 0.036 + 1.96\sqrt{\frac{0.036 \cdot 0.964}{500}} \right) = (0.020, 0.052).$$

Одредимо и интервал поверења са тачном дисперзијом: треба решити неједначину

$$p^2 \left(\frac{1.96^2}{500} + 1 \right) - p \left(\frac{1.96^2}{500} + 2 \cdot 0.036 \right) + 0.036^2 < 0$$

$$\iff 1.008p^2 - 0.080p + 0.001 < 0$$

Решења одговарајуће квадратне једначине су: $p_1 = 0.023$, $p_2 = 0.056$.

Тестирање хипотеза у вези са пропорцијом

хипотезе:

- ★ $H_0(p = p_0)$ против $H_1(p \neq p_0)$ (двострани тест)
- ★ $H_0(p \geq p_0)$ против $H_1(p < p_0)$ (једнострани леви)
- ★ $H_0(p \leq p_0)$ против $H_1(p > p_0)$ (једнострани десни)

тест статистика:

$$★ T_n = \frac{\bar{X}_n - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

при H_0 (приближно) има $\mathcal{N}(0, 1)$

$$★ T_n = \frac{\bar{X}_n - p_0}{\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}}$$

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 $t_0 > 0$.

Пример У истраживању једне популације јединки жели се проверити да ли је проценат јединки које имају одређену карактеристику већи од 2%. Узето је 500 јединки и утврђено је да је 18 са карактеристиком. Какав се закључак може донети на основу доступног узорка?

Поставимо хипотезе: $H_0(p = 0.02)$ $H_1(p > 0.02)$.

Реализована вредност тест статистике: $\frac{\bar{x}_n - 0.02}{\sqrt{\frac{0.02(1-0.02)}{500}}} = 2.55$.

p —вредност теста: $P_{H_0}\{T_n > 2.55\} = 0.0054$

Закључујемо да можемо рећи да је више од 2% јединки са карактеристиком.



```
binom.test(18,500,p=0.02,alternative = "greater")
```

Exact binomial test

```
data: 18 and 500
number of successes = 18, number of trials = 500, p-value =
0.01339
alternative hypothesis: true probability of success is greater than 0.02
95 percent confidence interval:
 0.02339513 1.00000000
sample estimates:
probability of success
      0.036
```

Зашто се јавља разлика у p —вредности рачунато „пешке“ и коришћењем R-а? Зато што функција у R-у користи праву биномну расподелу за израчунавање вероватноће, али је закључак исти.