

# TESTIRANJE HIPOTEZA O DISTRIBUCIJI

# Pearsonov $\chi^2$ test

Istraživač je slučajno izabrao uzorak od 100 ispitanika.

U uzorak je izabrano 55 muških i 45 ženskih osoba.

Nakon provedenog istraživanja recenzent mu je prigovorio da udio žena i muškaraca u uzorku nije jednak i da je to utjecalo na rezultate istraživanja.

Je li recenzent u pravu?

Je li disproporcija u uzorku posljedica slučajnog izbora ili pristranosti istraživača?

Udio žena u populaciji je 0.5.

Očekivana frekvencija žena i muškaraca u uzorku je 50.

Da li frekvencija spola u uzorku odgovara populaciji?

Promatramo diskretno obilježje s konačno mnogo vrijednosti:

$x_1, x_2, \dots, x_k$ .

Slučajnim izborom iz populacije definirana je slučajna varijabla  $X$  s distribucijom

$$P(X = x_i) = p_i, \quad i = 1, \dots, k$$

gdje  $p_i$  odgovara relativnim frekvencijama obilježja u populaciji.

Za uzorak veličine  $n$ :  $X_1, X_2, \dots, X_n$  distribucija slučajne varijable  $X_i$  je

$$X_i \sim \begin{pmatrix} x_1 & x_2 & \dots & x_k \\ p_1 & p_2 & \dots & p_k \end{pmatrix}$$

Želimo usporediti frekvenciju obilježja na uzorku s distribucijom obilježja u populaciji.

# Jednostruka klasifikacija

Problem jednostruke klasifikacije - unaprijed su zadane teorijske proporcije za pojedine kategorije.

Cilj nam je ustvrditi da li razdioba obilježja u populaciji jednako zadanoj razdiobi.

Hipoteza:

distribucija obilježja u populaciji jednaka je zadanoj distribuciji:

$$\begin{pmatrix} x_1 & x_2 & \dots & x_k \\ p_1 & p_2 & \dots & p_k \end{pmatrix}$$

Za zadanu razdiobu

$$\begin{pmatrix} x_1 & x_2 & \dots & x_k \\ p_1 & p_2 & \dots & p_k \end{pmatrix}$$

i uzorak veličine  $n$  definiramo **teorijske frekvencije**:

$$E_i = n \cdot p_i.$$

S  $F_i$  ćemo označiti frekvencije u uzorku ( **opažene frekvencije** ).

Rezultat možemo prikazati tablično

| Kategorija | Frekvencije |           |
|------------|-------------|-----------|
|            | Opazena     | Teorijska |
| $x_1$      | $F_1$       | $E_1$     |
| $x_2$      | $F_2$       | $E_2$     |
| $\vdots$   | $\vdots$    | $\vdots$  |
| $x_k$      | $F_k$       | $E_k$     |
| Ukupno     | $n$         | $n$       |

Ukoliko je hipoteza točna, tada je statistika

$$\chi^2 = \sum_i \frac{(F_i - E_i)^2}{E_i}$$

distribuirana prema  $\chi^2$  razdiobi s  $k - 1$  stupnjeva slobode:

$$\chi^2 \sim \chi^2(k - 1).$$

**Pretpostavka  $\chi^2$  testa:**

**Barem 80 % očekivanih frekvencije u klijetkama trebaju biti veće ili jednake od 5 a sve moraju biti veće ili jednake od 1.**

**Primjer.** Da li je uzorak iz uvodnog primjera (55 muškaraca i 45 žena) iz populacije s jednakim udjelom muškaraca i žena.

Varijabla spol  $\rightarrow$  Vrijednosti  $x_1 = \text{M}$ ,  $x_2 = \text{Ž}$ .

Udjeli:  $p_1 = 0.5$ ,  $p_2 = 0.5$

Očekivane frekvencije:

$$E_1 = n \cdot p_1 = 100 \cdot 0.5 = 50,$$

$$E_2 = n \cdot p_2 = 100 \cdot 0.5 = 50$$

Opažene frekvencije:  $F_1 = 55$ ,  $F_2 = 45$

Tablica:

| Kategorija | Frekvencije |           |
|------------|-------------|-----------|
|            | Opažena     | Teorijska |
| M          | 55          | 50        |
| Ž          | 45          | 50        |
| Ukupno     | 100         | 100       |

Računamo statistiku:

$$\chi^2 = \sum_i \frac{(F_i - E_i)^2}{E_i} = \frac{(55 - 50)^2}{50} + \frac{(45 - 50)^2}{50} = \frac{25}{50} + \frac{25}{50} = 1$$

Broj vrijednosti:  $k = 2$ .  $\rightarrow$  Broj stupnjeva slobode:  $k - 1 = 1$

Distribucija:  $\chi^2(1)$ .

$$\chi_{0.95}^2(1) = 3.84$$

$$\chi^2 = 1 < 3.84 = \chi_{0.95}^2(1) \rightarrow \text{Hipotezu ne odbacujemo.}$$

Primjedba recenzenta nema osnove.

Razlika u proporcijama je posljedica slučajnog izbora.

**Napomena.** Ovu hipotezu smo mogli testirati i  $t$ -testom za proporcije (hipoteza  $p = 0.5$ ).

Za dihotomne varijable  $t$ -test i  $\chi^2$ -test su ekvivalentni.

Kod  $\chi^2$ -testa s dvije kategorije očekivana frekvencija treba biti veća od 10 u svakoj klijetki.

**Primjer.** Istraživač za slučajni izbor u uzorak koristi igraću kocku. Prije početka izbora uzorka htio je provjeriti je li njegova kocka simetrična (svi brojevi imaju iste vjerojatnosti). Kocku je bacio 120 puta i zabilježio frekvencije brojeva:

| Vrijednost | Opažena frekvencija |
|------------|---------------------|
| 1          | 23                  |
| 2          | 15                  |
| 3          | 16                  |
| 4          | 28                  |
| 5          | 17                  |
| 6          | 21                  |
| Ukupno     | 120                 |

Da li je kocka simetrična?

Ukoliko je kocka simetrična, vjerojatnost pojedinog broja je  $p_i = 1/6$ .

Za  $n = 120$  bacanja, očekivane frekvencije pojedinog broja su jednake  $E_i = n \cdot p_i = 120 \cdot 1/6 = 20$ .

Tablica:

| Vrijednost | Frekvencije |           |
|------------|-------------|-----------|
|            | Opažena     | Teorijska |
| 1          | 23          | 20        |
| 2          | 15          | 20        |
| 3          | 16          | 20        |
| 4          | 28          | 20        |
| 5          | 17          | 20        |
| 6          | 21          | 20        |
| Ukupno     | 120         | 120       |

Računanje statistike:

$$\begin{aligned}\chi^2 &= \sum_i \frac{(F_i - E_i)^2}{E_i} = \\&= \frac{(23 - 20)^2}{20} + \frac{(15 - 20)^2}{20} + \frac{(16 - 20)^2}{20} + \\&\quad + \frac{(28 - 20)^2}{20} + \frac{(17 - 20)^2}{20} + \frac{(21 - 20)^2}{20} = \\&= \frac{9}{20} + \frac{25}{20} + \frac{16}{20} + \frac{64}{20} + \frac{9}{20} + \frac{1}{20} = 6.2\end{aligned}$$

Broj vrijednosti:  $k = 6$ .  $\rightarrow$  Broj stupnjeva slobode:  $k - 1 = 5$

Distribucija:  $\chi^2(5)$ .

$$\chi_{0.95}^2(5) = 11.07$$

$\chi^2 = 6.2 < 11.07 = \chi_{0.95}^2(5) \rightarrow$  Hipotezu ne odbacujemo.

Kocka je simetrična.

# Rješenje primjera u R-u

```
chisq.test(x, p)
```

- x - vektor opaženih frekvencija
- p - vektor teorijskih (očekivanih) proporcija

```
> observed <- c(23, 15, 16, 28, 17, 21)
> expected <- c(1/6, 1/6, 1/6, 1/6, 1/6, 1/6)
> chisq.test(x=observed,p=expected)
```

Chi-squared test for given probabilities

```
data:  observed
```

```
X-squared = 6.2, df = 5, p-value = 0.2872
```

**Primjer.** Istraživač je bilježio broj zgoditaka na uzorku od 200 nogometnih utakmica. Dobio je sljedeći rezultat:

| Broj zgoditaka | 0  | 1  | 2  | 3  | 4  | 5 | 6 | 7 | Ukupno |
|----------------|----|----|----|----|----|---|---|---|--------|
| Frekvencija    | 22 | 53 | 58 | 39 | 20 | 5 | 2 | 1 | 200    |

Može li se na osnovu ovih podataka zaključiti da Poissonova razdioba adekvatno opisuje razdiobu broja zgoditaka?

Poissonova razdioba ( $X \sim \text{Po}(\lambda)$ ):

$$P(X = i) = p_i = \frac{\lambda^i}{i!} e^{-\lambda}$$

$\lambda$  nije zadan pa ga procjenjujemo iz podataka.

Za  $X \sim \text{Po}(\lambda)$  je  $E(X) = \lambda$ .

$\lambda$  ćemo procijeniti s  $\bar{X}$ :  $\lambda = \bar{X} = 2.05$

Provjeravamo da li se radi o  $\text{Po}(2.05)$  razdiobi.

Izračunamo vjerojatnosti i očekivane frekvencije:

| Br. zgod. | 0     | 1     | 2     | 3     | 4     | 5     | 6           | 7+          |
|-----------|-------|-------|-------|-------|-------|-------|-------------|-------------|
| $p_i$     | 0.129 | 0.264 | 0.271 | 0.185 | 0.095 | 0.039 | 0.013       | 0.005       |
| $E_i$     | 25.75 | 52.78 | 54.10 | 36.97 | 18.95 | 7.77  | <b>2.65</b> | <b>1.03</b> |

Preko 20% klijetki ima očekivanu frekvenciju manju od 5.

Spajamo kategorije.

| Br. zgod. | 0     | 1     | 2     | 3     | 4     | 5+    | Ukupno |
|-----------|-------|-------|-------|-------|-------|-------|--------|
| $F_i$     | 22    | 53    | 58    | 39    | 20    | 8     | 200    |
| $E_i$     | 25.75 | 52.78 | 54.10 | 36.97 | 18.95 | 11.46 | 200    |

Statistika: 
$$\chi^2 = \sum_i \frac{(F_i - E_i)^2}{E_i} = 2.04$$

Broj stupnjeva slobode:

- 6 kategorija
- 1 parametar procijenjen iz podataka
- broj stupnjeva slobode = broj kategorija - broj procijenjenih parametara = 6-1=5

Distribucija.  $\chi^2(5)$

$p$ -vrijednost:  $p = 0.843629$

Ne odbacujemo hipotezu.

Broj zgoditaka je dobro opisan Poissonovom distribucijom.

# Testiranje hipoteza o nezavisnosti i homogenosti

## Dvostruka klasifikacija.

### Testiranje homogenosti

Želimo provjeriti hipotezu da je (diskretno) obilježje jednako distribuirano u više populacija.

$c$  - broj populacija u kojem promatramo obilježje  $X$ .

$X_j$  - rezultat mjerenja obilježja  $X$  u  $j$ -toj populaciji

$$X_j \sim \begin{pmatrix} x_1 & x_2 & \dots & x_k \\ p_{1j} & p_{2j} & \dots & p_{kj} \end{pmatrix}, \quad j = 1, \dots, c$$

Želimo provjeriti da li su sve ove slučajne varijable jednako distribuirane, tj. da li vrijedi

Hipoteza o homogenosti:

$$p_{i1} = p_{i2} = p_{i3} = \dots = p_{ic} \quad \text{za} \quad i = 1, \dots, k.$$

Da bismo provjerili hipotezu, u svakoj populaciji izaberemo uzorak.

Veličine uzoaraka:  $n_{y1}, n_{y2}, \dots, n_{yc}$

U svakom uzorku odredimo frekvencije  $n_{ij}$  - broj pojavljivanja obilježja  $x_i$  u  $j$ -toj populaciji.

Rezultat prikazujemo u **kontingencijskoj tablici**:

|               | 1        | 2        | 3        | ... | c        |
|---------------|----------|----------|----------|-----|----------|
| $x_1$         | $n_{11}$ | $n_{12}$ | $n_{13}$ | ... | $n_{1c}$ |
| $x_2$         | $n_{21}$ | $n_{22}$ | $n_{23}$ | ... | $n_{2c}$ |
| $x_3$         | $n_{31}$ | $n_{32}$ | $n_{33}$ | ... | $n_{3c}$ |
| $\vdots$      | $\vdots$ | $\vdots$ | $\vdots$ |     | $\vdots$ |
| $x_k$         | $n_{k1}$ | $n_{k2}$ | $n_{k3}$ | ... | $n_{kc}$ |
| <b>Ukupno</b> | $n_{y1}$ | $n_{y2}$ | $n_{y3}$ | ... | $n_{yc}$ |

Za svako obilježje  $x_i$  prebrojimo broj pojavljivanja (frekvencije)  $n_{xi}$  u svih  $c$  uzoraka:

|               | 1        | 2        | 3        | ... | c        | Ukupno   |
|---------------|----------|----------|----------|-----|----------|----------|
| $x_1$         | $n_{11}$ | $n_{12}$ | $n_{13}$ | ... | $n_{1c}$ | $n_{x1}$ |
| $x_2$         | $n_{21}$ | $n_{22}$ | $n_{23}$ | ... | $n_{2c}$ | $n_{x2}$ |
| $x_3$         | $n_{31}$ | $n_{32}$ | $n_{33}$ | ... | $n_{3c}$ | $n_{x3}$ |
| $\vdots$      | $\vdots$ | $\vdots$ | $\vdots$ |     | $\vdots$ |          |
| $x_k$         | $n_{k1}$ | $n_{k2}$ | $n_{k3}$ | ... | $n_{kc}$ | $n_{xk}$ |
| <b>Ukupno</b> | $n_{y1}$ | $n_{y2}$ | $n_{y3}$ | ... | $n_{yc}$ | $n$      |

Budući da pretpostavljamo jednaku razdiobu u svim populacijama:

$$p_i = p_{i1} = p_{i2} = p_{i3} = \dots = p_{ic} \quad \text{za} \quad i = 1, \dots, k.$$

$p_i$  procjenjujemo na osnovu cijelog uzorka s  $\frac{n_{xi}}{n}$ .

Sada je očekivana frekvencija vrijednosti  $x_i$  u  $j$ -tom uzorku dana s:

$$E_{ij} = \frac{n_{xi}}{n} n_{yj}.$$

Ukoliko je hipoteza točna, statistika

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

distribuirana prema  $\chi^2$  razdiobi.

## Broj stupnjeva slobode

U svakom stupcu imamo  $k - 1$  stupanj slobode (jer je suma  $n_{yi}$  fiksirana)

→ sve skupa  $c \cdot (k - 1)$  stupnjeva slobode

Iz uzorka određujemo  $k$  parametara  $n_{x1}, n_{x2}, \dots, n_{xk}$ .

Jer je  $\sum_i n_{xi} = n \rightarrow k - 1$  stupnjeva slobode u parametrima

Ukupno je

$$\text{d.f.} = c \cdot (k - 1) - (k - 1) = (c - 1) \cdot (k - 1).$$

Statistika  $\chi^2$  je distribuirana:

$$\chi^2 \sim \chi^2((c - 1) \cdot (k - 1))$$

**Primjer.** Provedena je anketa da bi se ustanovila pojavnost alkoholizma u različitim sportovima. Intervjuirani su slučajni uzorci sportaša iz 4 sporta. Dobiveni rezultati su prikazani u tablici

|                | <b>Alkoholičari</b> | <b>Nealkoholičari</b> | <b>Veličina uzorka</b> |
|----------------|---------------------|-----------------------|------------------------|
| <b>Sport 1</b> | 32                  | 268                   | 300                    |
| <b>Sport 2</b> | 51                  | 199                   | 250                    |
| <b>Sport 3</b> | 67                  | 233                   | 300                    |
| <b>Sport 4</b> | 83                  | 267                   | 350                    |

Konstruirajte test da biste provjerili da li je udio alkoholičara jednak u sva 4 sporta

Izračunajmo očekivane frekvencije:

|                | <b>Alkoholičari</b> |         | <b>Nealkoholičari</b> |          | <b>Veličina uzorka</b> |
|----------------|---------------------|---------|-----------------------|----------|------------------------|
| <b>Sport 1</b> | 32                  | (58.25) | 268                   | (241.75) | 300                    |
| <b>Sport 2</b> | 51                  | (48.54) | 199                   | (201.46) | 250                    |
| <b>Sport 3</b> | 67                  | (58.25) | 233                   | (241.75) | 300                    |
| <b>Sport 4</b> | 83                  | (67.96) | 267                   | (282.04) | 350                    |
| <b>Ukupno</b>  | 233                 |         | 967                   |          | 1200                   |

U zagradama su prikazane očekivane frekvencije dobivene množenjem suma stupca i redka te podijeljenih s ukupnom veličinom uzorka (1200).

Ukoliko s  $p_1, p_2, p_3, p_4$  označimo udio alkoholičara u pojedinom sportu, želimo testirati hipotezu

$$H_0 : p_1 = p_2 = p_3 = p_4.$$

Računamo vrijednost statistike

$$\chi^2 = \frac{(32 - 58.25)^2}{58.25} + \frac{(268 - 241.75)^2}{241.75} + \dots + \frac{(267 - 282.04)^2}{282.04} = 20.59$$

Budući da je  $d.f. = (2 - 1) \cdot (4 - 1) = 3$ , za  $\alpha = 0.05$  je

$$\chi_{0.95}^2(3) = 7.815.$$

Jer je  $\chi^2 > \chi_{0.95}^2(3)$  hipotezu o jednakosti proporcija odbacujemo.

Nakon odbacivanja hipoteze, poželjno je provjeriti zbog čega je ona odbačena.

Pogledajmo doprinos svake grupe  $\chi^2$  statistici:

|                | $n_{ij} - E_{ij}$ |        | $(n_{ij} - E_{ij})^2 / E_{ij}$ |        |
|----------------|-------------------|--------|--------------------------------|--------|
|                | Alk.              | Nealk. | Alk.                           | Nealk. |
| <b>Sport 1</b> | -26.25            | 26.25  | <b>11.83</b>                   | 2.85   |
| <b>Sport 2</b> | 2.46              | -2.46  | 0.12                           | 0.03   |
| <b>Sport 3</b> | 8.75              | -8.75  | 1.31                           | 0.32   |
| <b>Sport 4</b> | 15.04             | -15.04 | 3.33                           | 0.80   |

Najveći doprinos daje dio koji odgovara udjelu alkoholičara u Sportu 1.

Odstupanje od homogenosti je u najvećoj mjeri uzrokovano malim udjelom alkoholičara u Sportu 1.

## Rješenje primjera u R-u

Podaci:

```
> podaci <- matrix(c(32, 268, 51, 199, 67, 233, 83,  
267), ncol=2, byrow=TRUE)  
  
> colnames(podaci) <- c("Alk", "Nealk")  
  
> rownames(podaci) <- c("Sport 1", "Sport 2", "Sport  
3", "Sport 4")
```

Treba nam tip *table*:

```
> data <- as.table(podaci)
```

```
> podaci
```

```
      Alk Nealk
```

```
Sport 1  32  268
```

```
Sport 2  51  199
```

```
Sport 3  67  233
```

```
Sport 4  83  267
```

$\chi^2$ -test:

```
> chisq.test(podaci)
```

Pearson's Chi-squared test

```
data: podaci
```

```
X-squared = 20.597, df = 3, p-value = 0.0001277
```

# Testiranje hipoteze o nezavisnosti

**Primjer.** Na uzorku od 95 osoba promatrana je boja očiju i boja kose. Rezultati su prikazani u tablici:

|            |        | Boja kose |       |        |
|------------|--------|-----------|-------|--------|
|            |        | Svjetla   | Tamna | Ukupno |
| Boja očiju | Plave  | 32        | 12    | 44     |
|            | Smeđe  | 14        | 22    | 36     |
|            | Ostalo | 6         | 9     | 15     |
| Ukupno     |        | 52        | 43    | 95     |

Jesu li boja očiju i boja kose povezani?

Promatramo dva obilježja,  $X$  i  $Y$  koja poprimaju vrijednosti  $x_1, \dots, x_k$  i  $y_1, \dots, y_c$ .

U uzorku od  $n$  jedinki promatramo frekvencije pojavljivanja svih kombinacija  $(x_i, y_j)$  i označimo ih s  $n_{ij}$

Promatramo i frekvencije svakog obilježja posebno:

- $n_{xi}$  frekvencija za  $x_i$
- $n_{yj}$  frekvencija za  $y_j$

Vrijedi

$$n_{xi} = \sum_j n_{ij} \quad \text{i} \quad n_{yj} = \sum_i n_{ij}$$

Rezultat prebrojavanja se prikazuje u kontingencijskoj tablici:

|               |          | <b>Y</b> |          |          |         |          |               |
|---------------|----------|----------|----------|----------|---------|----------|---------------|
|               |          | $y_1$    | $y_2$    | $y_3$    | $\dots$ | $y_c$    | <b>Ukupno</b> |
| <b>X</b>      | $x_1$    | $n_{11}$ | $n_{12}$ | $n_{13}$ | $\dots$ | $n_{1c}$ | $n_{x1}$      |
|               | $x_2$    | $n_{21}$ | $n_{22}$ | $n_{23}$ | $\dots$ | $n_{2c}$ | $n_{x2}$      |
|               | $x_3$    | $n_{31}$ | $n_{32}$ | $n_{33}$ | $\dots$ | $n_{3c}$ | $n_{x3}$      |
|               | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |         | $\vdots$ | $\vdots$      |
|               | $x_k$    | $n_{k1}$ | $n_{k2}$ | $n_{k3}$ | $\dots$ | $n_{kc}$ | $n_{xk}$      |
| <b>Ukupno</b> |          | $n_{y1}$ | $n_{y2}$ | $n_{y3}$ | $\dots$ | $n_{yc}$ | $n$           |

Slučajnim izvlačenjem definirane su slučajne varijable  $X$  i  $Y$ .

Njihova razdioba je dana s

$$X \sim \begin{pmatrix} x_1 & x_2 & \dots & x_k \\ p_{x1} & p_{x2} & \dots & p_{xk} \end{pmatrix} \quad \text{i} \quad Y \sim \begin{pmatrix} y_1 & y_2 & \dots & y_c \\ p_{y1} & p_{y2} & \dots & p_{yc} \end{pmatrix}.$$

Označimo

$$p_{ij} = P(X = x_i, Y = y_j).$$

- $\frac{n_{ij}}{n}$  je procjenitelj za  $p_{ij}$ .
- $\frac{n_{xi}}{n}$  je procjenitelj za  $p_{xi}$ .
- $\frac{n_{yj}}{n}$  je procjenitelj za  $p_{yj}$ .

Obilježja  $X$  i  $Y$  ne ovise jedno o drugome

$\longleftrightarrow$  Slučajne varijable  $X$  i  $Y$  su nezavisne:

$$p_{ij} = P(X = x_i, Y = y_j) = P(X = x_i) \cdot P(Y = y_j) = p_{xi} \cdot p_{yj}.$$

Testiramo hipotezu:

Hipoteza o nezavisnosti:

$$H_0 : p_{ij} = p_{xi} \cdot p_{yj} \quad \text{za sve } i, j.$$

Za testiranje hipoteze definiramo očekivane frekvencije

$$E_{ij} = n \cdot \frac{n_{xi}}{n} \frac{n_{yj}}{n} = \frac{n_{xi} \cdot n_{yj}}{n}$$

Ovdje smo iskoristili pretpostavku o nezavisnosti slučajnih varijabli  $X$  i  $Y$ .

Ukoliko je hipoteza točna tada je statistika

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

distribuirana prema  $\chi^2$  razdiobi.

## Broj stupnjeva slobode

Za test smo procijenili parametre  $p_{x1}, p_{x2}, \dots, p_{xk}$  i  $p_{y1}, p_{y2}, \dots, p_{yc}$ .

No kako je

$$\sum_i p_{xi} = 1 \quad \text{i} \quad \sum_j p_{yj} = 1$$

imamo samo  $k - 1$  i  $c - 1$  stupnjeva slobode, tj.  $k + c - 2$  procijenjenih parametara.

Ukupno imamo  $k \cdot c$  frekvencija ( $n_{ij}$ ).

Jer je

$$\sum_i \sum_j n_{ij} = n$$

to daje  $k \cdot c - 1$  stupnjeva slobode.

Ukupan broj stupnjeva slobode je

$$\text{d.f.} = k \cdot c - 1 - (k + c - 2) = k \cdot c - k - c + 1 = (c - 1) \cdot (k - 1).$$

Statistika  $\chi^2$  je distribuirana:

$$\chi^2 \sim \chi^2((c - 1) \cdot (k - 1))$$

Uočimo da su testovi za homogenost i nezavisnost isti.

Razlikuju se po interpretaciji i načinu određivanja uzorka.

- test homogenosti -  $n_{yj}$  je zadan (veličina uzorka)
- test nezavisnosti -  $n_{yj}$  je izračunat iz uzorka

U R-u se za oba testa koristi ista procedura.

Vratimo se primjeru s početka.

Izračunamo očekivane frekvencije  $E_{ij} = \frac{n_{xi} \cdot n_{yj}}{n}$  (prikazane u zagradama).

|            |        | Boja kose |        |       |        | Ukupno |
|------------|--------|-----------|--------|-------|--------|--------|
|            |        | Svjetla   |        | Tamna |        |        |
| Boja očiju | Plave  | 32        | (24.1) | 12    | (19.9) | 44     |
|            | Smeđe  | 14        | (19.7) | 22    | (16.3) | 36     |
|            | Ostalo | 6         | (8.2)  | 9     | (6.8)  | 15     |
| Ukupno     |        | 52        |        | 43    |        | 95     |

$$\chi^2 = \frac{7.9^2}{24.1} + \frac{(-7.9)^2}{19.9} + \frac{(-5.7)^2}{19.7} + \frac{5.7^2}{16.3} + \frac{(-2.2)^2}{8.2} + \frac{2.2^2}{6.8} = 10.67.$$

$$d.f. = (3 - 1) \cdot (2 - 1) = 2$$

$$\chi_{0.95}^2(2) = 5.99$$

Jer je  $\chi^2 > \chi_{0.95}^2(2)$  odbacujemo hipotezu o nezavisnosti.

Boja očiju i boja kose su povezani.

## Narušavanje pretpostavki $\chi^2$ testa

Pretpostavka  $\chi^2$  testa:

Barem 80 % očekivanih frekvencije u klijetkama trebaju biti veće ili jednake od 5 a sve moraju biti veće ili jednake od 1.

Pretpostavka  $2 \times 2$   $\chi^2$  testa:

Sve frekvencije su veće ili jednake od 5.

Ukoliko pretpostavke nisu zadovoljene treba koristiti **Fisherov egzaktni test**.

Ili spojiti kategorije da bismo povećali očekivane frekvencije.

# Fisherov egzaktni test

```
> fisher.test(podaci )
```

Fisher's Exact Test for Count Data

data: podaci

p-value = 5.281e-05

alternative hypothesis: two.sided

$\chi^2$ -test:

```
> chisq.test(podaci)
```

Pearson's Chi-squared test

data: podaci

X-squared = 20.597, df = 3, p-value = 0.0001277

# Testiranje hipoteze o normalnosti

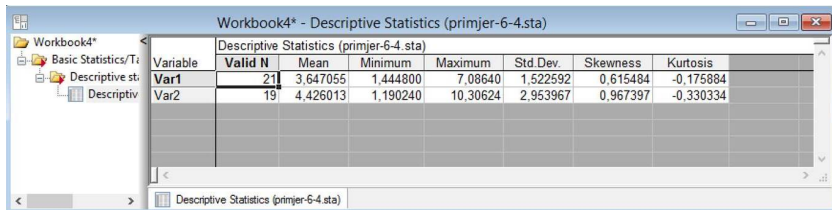
Normalna distribucija obilježja u populaciji je osnovna pretpostavka  $t$ -testa i ANOVA testa.

Kako provjeriti ovu pretpostavku?

**Primjer.** Promatrat ćemo dva primjera mjerenja:

| V1      | V2      | V1      | V2       |
|---------|---------|---------|----------|
| 3.32304 | 6.35024 | 5.93744 | 4.52704  |
| 4.69216 | 7.83632 | 4.42384 | 10.30624 |
| 1.86448 | 1.38976 | 4.00416 | 9.47376  |
| 3.83904 | 1.62368 | 2.79328 | 5.39392  |
| 1.89888 | 1.80944 | 5.0568  | 3.63264  |
| 4.58208 | 2.65568 | 2.73824 | 2.3736   |
| 3.32304 | 3.59824 | 3.27488 | 3.19232  |
| 3.08224 | 9.79024 | 2.18784 | 3.8528   |
| 6.03376 | 1.19024 | 1.4448  | 2.6144   |
| 7.0864  | 2.48368 | 3.13728 |          |
| 1.86448 |         |         |          |

# Deskriptivna statistika



Workbook4\* - Descriptive Statistics (primjer-6-4.sta)

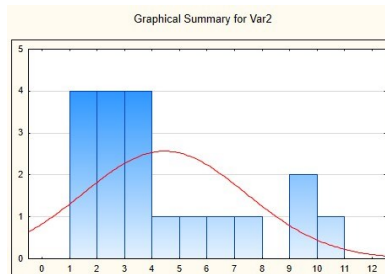
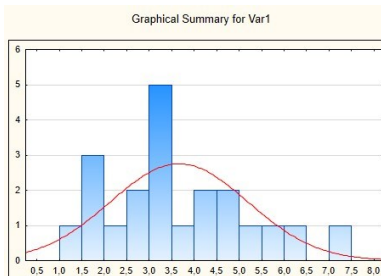
| Variable | Valid N | Mean     | Minimum  | Maximum  | Std.Dev. | Skewness | Kurtosis  |
|----------|---------|----------|----------|----------|----------|----------|-----------|
| Var1     | 21      | 3,647055 | 1,444800 | 7,08640  | 1,522592 | 0,615484 | -0,175884 |
| Var2     | 19      | 4,426013 | 1,190240 | 10,30624 | 2,953967 | 0,967397 | -0,330334 |

Za normalnu razdiobu je

- zakošenost = 0
- spljoštenost = 0 (kurtosis je 3 ali se koristi pomak od -3)

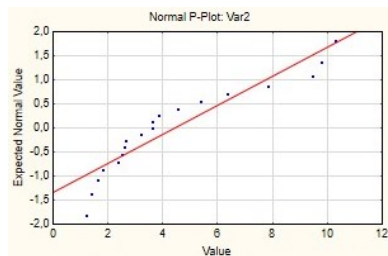
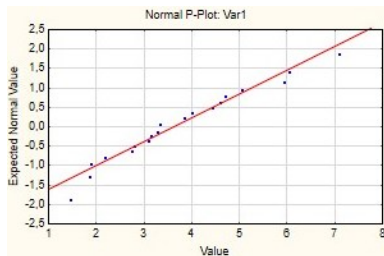
## Grafički prikaz

### Normalna krivulja



## Normalni plot

Normalna razdioba je prikazana kao pravac.

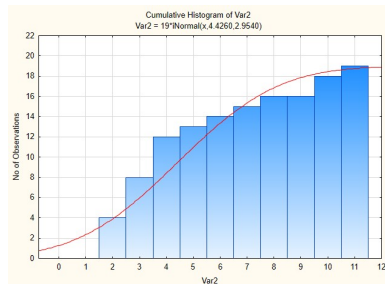
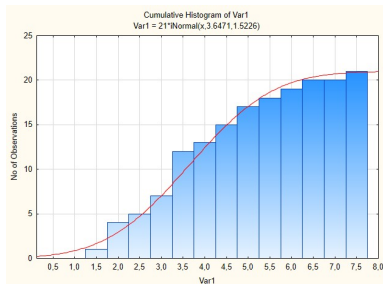


# Kolmogorov-Smirnov test

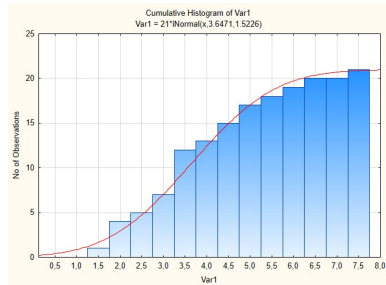
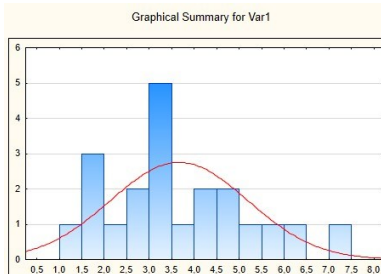
Na osnovu uzorka odredimo eksperimentalnu funkciju distribucije.

Izračunamo  $\bar{X}$  i  $S^2$ .

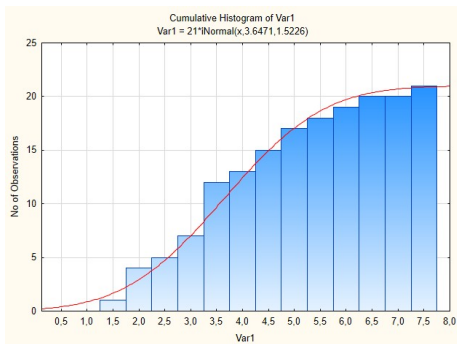
Usporedimo eksperimentalnu funkciju distribucije s funkcijom distribucije normalne razdiobe  $N(\bar{X}, S^2)$ .



## Funkcija gustoće i funkcija distribucije



Kolmogorov-smirnov test koristi statistiku  $D$ =najveće odstupanje eksperimentalne funkcije distribucije i očekivane normalne funkcije distribucije.



Kolmogorov-Smirnov test se može koristiti za proizvoljnu razdiobu.

**Shapiro-Wilkov test** - testira hipotezu o normalnosti populacije.

Snaga je veća nego kod Kolmogorov-Smirnov testa.

Ako populacija nije normalna:

- izabrati test koji nema pretpostavku o normalnosti populacije (neparametarski testovi)
- transformirati podatke (npr. logaritamska transformacija, Box-Coxova transformacija, korijen, inverzna transformacija, ...)