# Primijenjena i inženjerska matematika

## Statistika - predavanje 6

***Statističko zaključivanje - dvije varijable***

Zaključivanje o nezavisnosti

17.11.2023.

# Dvodimenzionalni slučajni vektor

- $x_1, \ldots, x_n$ je uzorak iz slučajne varijable $X$
- $y_1, \ldots, y_n$ je uzorak iz slučajne varijable $Y$
- parove $(x_1, y_1), \ldots, (x_n, y_n)$ nazivamo uzorkom iz **slučajnog vektora** $(X, Y)$
- svaki par $(x_i, y_i)$ je realizacija jednog slučajnog vektora $(X_i, Y_i)$ jednako distribuiranog kao $(X, Y)$
- cilj je utvrditi postoje li neke ovisnosti među varijablama ili su one neovisne jedna o drugoj

# Dvodimenzionalni slučajni vektor

Takvi uzorci obično nastaju kada:

- promatramo isto obilježje na istim jedinkama, ali pod različitim uvjetima (npr. masa osobe prije i poslije treninga)
- promatramo različita obilježja na istim jedinkama (npr. mjerimo istovremeno tjelesnu masu i visinu – dva različita, ali vezana obilježja pa ih promatramo u paru)

# Dvodimenzionalni slučajni vektor - primjer

- Primjer: broj pogrešno zavarenih vrećica na dvije linije za pakiranje bombona

| sat | prva linija - broj grešaka | druga linija - broj grešaka |
|---|---|---|
| 1 | 0 | 0 |
| 2 | 1 | 0 |
| 3 | 2 | 2 |
| ⋮ | ⋮ | ⋮ |
| 400 | 3 | 1 |

# Dvodimenzionalni slučajni vektor - primjer

- **Zajednička tablica frekvencija** za broj pogrešno zavarenih vrećica na prvoj i drugoj liniji

| | | druga linija $(Y)$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 | zbroj |
| prva linija $(X)$ | 0 | 22 | 12 | 13 | 12 | 7 | 66 |
| | 1 | 20 | 24 | 14 | 30 | 10 | 98 |
| | 2 | 15 | 20 | 30 | 10 | 7 | 82 |
| | 3 | 6 | 5 | 10 | 32 | 20 | 73 |
| | 4 | 5 | 7 | 13 | 31 | 25 | 81 |
| zbroj | | 68 | 68 | 80 | 115 | 69 | 400 |

# Dvodimenzionalni slučajni vektor - primjer

- **Empirijska distribucija slučajnog vektora** $(X, Y)$ (zajednička tablica relativnih frekvencija)

| | | $Y$ | | | | |
|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 |
| | 0 | 0.0550 | 0.0300 | 0.0325 | 0.0300 | 0.0175 |
| | 1 | 0.0500 | 0.0600 | 0.0350 | 0.0750 | 0.0250 |
| $X$ | 2 | 0.0375 | 0.0500 | 0.0750 | 0.0250 | 0.0175 |
| | 3 | 0.0150 | 0.0125 | 0.0250 | 0.0800 | 0.0500 |
| | 4 | 0.0125 | 0.0175 | 0.0325 | 0.0775 | 0.0625 |

# Dvodimenzionalni slučajni vektor - primjer

- **Empirijske distribucije slučajnih varijabli $X$ i $Y$** (tablice relativnih frekvencija)

| vrijednost od $X$ | $0$ | $1$ | $2$ | $3$ | $4$ |
|---|---|---|---|---|---|
| relativna frekvencija | $0.165$ | $0.245$ | $0.205$ | $0.1825$ | $0.2025$ |

| vrijednost od $Y$ | $0$ | $1$ | $2$ | $3$ | $4$ |
|---|---|---|---|---|---|
| relativna frekvencija | $0.17$ | $0.17$ | $0.2$ | $0.2875$ | $0.1725$ |

# Dvodimenzionalni slučajni vektor - primjer

- Zajednička tablica relativnih frekvencija s marginalnim distribucijama

| | | $Y$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 | zbroj |
| $X$ | 0 | 0.0550 | 0.0300 | 0.0325 | 0.0300 | 0.0175 | 0.165 |
| | 1 | 0.0500 | 0.0600 | 0.0350 | 0.0750 | 0.0250 | 0.245 |
| | 2 | 0.0375 | 0.0500 | 0.0750 | 0.0250 | 0.0175 | 0.205 |
| | 3 | 0.0150 | 0.0125 | 0.0250 | 0.0800 | 0.0500 | 0.1825 |
| | 4 | 0.0125 | 0.0175 | 0.0325 | 0.0775 | 0.0625 | 0.2025 |
| zbroj | | 0.17 | 0.17 | 0.2 | 0.2875 | 0.1725 | 1 |

# Distribucija diskretnog slučajnog vektora

- jedna realizacija dvodimenzionalnog slučajnog vektora – uređeni par realnih brojeva
- **diskretan slučajni vektor** – realizacija može biti samo konačno ili prebrojivo mnogo
- radi jednostavnosti promatat ćemo samo slučajne vektore s konačnim skupom svih mogućih vrijednosti
- zadati distribuciju slučajnog vektora znači zadati vjerojatnosti na skupu svih njegovih mogućih realizacija:

$$P(X = x_i, Y = y_j), \text{ za sve } x_i \in \mathcal{R}(X), y_j \in \mathcal{R}(Y)$$

- $P(X = x_i, Y = y_j)$ – vjerojatnost da je *istovremeno* $X = x_i$ i $Y = y_j$
- te brojeve organiziramo u tablicu distribucije

# Tablica distribucije diskretnog slučajnog vektora

- neka je $(X, Y)$ slučajni vektor takav da je $\mathcal{R}(X) = \{x_1, \dots, x_m\}$ i $\mathcal{R}(Y) = \{y_1, \dots, y_n\}$
- distribucija slučajnog vektora dana je sljedećom **tablicom distribucije**

| | | $Y$ | | | |
|---|---|---|---|---|---|
| | | $y_1$ | $y_2$ | $\dots$ | $y_n$ |
| | $x_1$ | $P(X = x_1, Y = y_1)$ | $P(X = x_1, Y = y_2)$ | $\dots$ | $P(X = x_1, Y = y_n)$ |
| $X$ | $x_2$ | $P(X = x_2, Y = y_1)$ | $P(X = x_2, Y = y_2)$ | $\dots$ | $P(X = x_2, Y = y_n)$ |
| | $\vdots$ | $\vdots$ | $\vdots$ | | $\vdots$ |
| | $x_m$ | $P(X = x_m, Y = y_1)$ | $P(X = x_m, Y = y_2)$ | $\dots$ | $P(X = x_m, Y = y_n)$ |

# Tablica distribucije diskretnog slučajnog vektora

- **Broj** $P(X = x_i, Y = y_j)$ **je vjerojatnost da slučajna varijabla** $X$ **primi vrijednost** $x_i$ **i slučajna varijabla** $Y$ **vrijednost** $y_j$**, tj. vjerojatnost da se dogode oba događaja** $\{X = x_i\}$ **i** $\{Y = y_j\}$**:**

$$P\left(\{X = x_i\} \cap \{Y = y_j\}\right) = P(X = x_i, Y = y_j).$$

- distribucije slučajnih varijabli koje čine ovaj slučajni vektor (tj. posebno distribucija od $X$ i distribucija od $Y$) mogu se dobiti iz tablice distribucije slučajnog vektora zbrajanjem vjerojatnosti u odgovarajućim redovima, odnosno stupcima

# Tablica distribucije diskretnog slučajnog vektora

| | | $Y$ | | | |
|---|---|---|---|---|---|
| | | $y_1$ | $\dots$ | $y_n$ | zbroj |
| | $x_1$ | $P(X = x_1, Y = y_1)$ | $\dots$ | $P(X = x_1, Y = y_n)$ | $P(X = x_1)$ |
| $X$ | $x_2$ | $P(X = x_2, Y = y_1)$ | $\dots$ | $P(X = x_2, Y = y_n)$ | $P(X = x_2)$ |
| | $\vdots$ | $\vdots$ | | $\vdots$ | |
| | $x_m$ | $P(X = x_m, Y = y_1)$ | $\dots$ | $P(X = x_m, Y = y_n)$ | $P(X = x_m)$ |
| | zbroj | $P(Y = y_1)$ | $\dots$ | $P(Y = y_n)$ | $1$ |

- distribucije u posljednjem retku, odnosno stupcu, zovemo **marginalne distribucije** slučajnog vektora $(X, Y)$
- na osnovu podataka distribuciju možemo samo procijeniti

# Empirijska distribucija diskretnog slučajnog vektora

- **empirijsku distribuciju diskretnog slučajnog vektora** dobijemo tako da elemente zajedničke tablice frekvencija dobivene temeljem nezavisnih mjerenja realizacija slučajnog vektora $(X, Y)$ podijelimo ukupnim brojem mjerenja $N$
- primjer 6.11, djelatnici.sta

# Uvjetne distribucije

- **uvjetne distribucije** – distribucija jedne komponente slučajnog vektora ako je poznata realizacija njegove druge komponente
- uvjetnu vjerojatnost za dva skupa $A$ i $B$ označavamo

$$P(X \in A | Y \in B)$$

  vjerojatnost da se $X$ realizira u skupu $A$ **ako se** $Y$ realizirao u skupu $B$
- neka je $y_j \in \mathcal{R}(Y)$ takav da je $P(Y = y_j) \neq 0$
- **uvjetna distribucija slučajne varijable $X$ uz uvjet da se dogodio događaj** $\{Y = y_j\}$

$$P(X = x_i | Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)}, \ i = 1, \ldots, m$$

# Uvjetne distribucije

- neka je $x_i \in \mathcal{R}(X)$ takav da je $P(X = x_i) \neq 0$
- **uvjetna distribucija slučajne varijable $Y$ uz uvjet da se dogodio događaj $\{X = x_i\}$**

$$P(Y = y_j | X = x_i) = \frac{P(X = x_i, Y = y_j)}{P(X = x_i)},\ j = 1, \ldots, n$$

- uvjetne distribucije možemo promatrati kao distribucije slučajnih varijabli

$$X|_{Y=y_j}, \quad y_j \in \mathcal{R}(Y), \quad j = 1, \ldots, , n$$

$$Y|_{X=x_i}, \quad x_i \in \mathcal{R}(X), \quad i = 1, \ldots, m.$$

# Nezavisnost

- za slučajne varijable $X$ i $Y$ kažemo da su **nezavisne** ako za sve $i = 1, \ldots, m$, $j = 1, \ldots, n$ vrijedi da je

$$P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j)$$

  tj. vjerojatnosti iz distribucije slučajnog vektora mogu se dobiti množenjem odgovarajućih vjerojatnosti iz marginalnih distribucija
- u suprotnom kažemo da su slučajne varijable $X$ i $Y$ **zavisne**

## Nezavisnost

- ako su slučajne varijable $X$ i $Y$ nezavisne, tada za svaki $y_j \in \mathcal{R}(Y)$ vrijedi da je

$$\begin{aligned} P(Y = y_j | X = x_i) &= \frac{P(X = x_i, Y = y_j)}{P(X = x_i)} \\ &= \frac{P(X = x_i)P(Y = y_j)}{P(X = x_i)} \\ &= P(Y = y_j) \end{aligned}$$

- ako su $X$ i $Y$ nezavisne, tada vrijedi:
  - distribucija slučajne varijable $Y$ i uvjetna distribucija $Y|_{X=x_i}$ su jednake za sve $x_i \in \mathcal{R}(X)$ za koje je $P(X = x_i) \neq 0$
  - distribucija slučajne varijable $X$ i uvjetna distribucija $X|_{Y=y_j}$ su jednake za sve $y_j \in \mathcal{R}(Y)$ za koje je $P(Y = y_j) \neq 0$,
- marginalne i odgovarajuće uvjetne distribucije slučajnog vektora $(X, Y)$ su jednake

# Nezavisnost

- na temelju podataka uvjetne distribucije možemo samo procijeniti
- primjeri 6.12, 6.13

# Analiza zavisnosti

- na temelju podataka možemo odrediti empirijsku distribuciju slučajnog vektora $(X, Y)$, marginalne empirijske distribucije, kao i uvjetne empirijske distribucije koje koristimo za procjenu odgovarajućih stvarnih distribucija
- zavisnost slučajnih varijabli definirana je na temelju pravih, a ne empirijskih distribucija
- procjene odstupaju od stvarnih distribucija – kako provjeriti nezavisnost?
- $\chi^2$ test – statistički test kojim možemo testirati hipotezu o nezavisnosti slučajnih varijabli
- hipoteze:

$$H_0 : \text{varijable su nezavisne}$$
$$H_1 : \text{varijable su zavisne}$$

# Analiza zavisnosti

| | | $Y$ | | | | |
|---|---|---|---|---|---|---|
| | | $y_1$ | $y_2$ | $\dots$ | $y_n$ | $zbroj$ |
| $X$ | $x_1$ | $n(x_1, y_1)$ | $n(x_1, y_2)$ | $\dots$ | $n(x_1, y_n)$ | $n_X(x_1)$ |
| | $x_2$ | $n(x_2, y_1)$ | $n(x_2, y_2)$ | $\dots$ | $n(x_2, y_n)$ | $n_X(x_2)$ |
| | $\vdots$ | $\vdots$ | $\vdots$ | | $\vdots$ | $\vdots$ |
| | $x_m$ | $n(x_m, y_1)$ | $n(x_m, y_2)$ | $\dots$ | $n(x_m, y_n)$ | $n_X(x_m)$ |
| | $zbroj$ | $n_Y(y_1)$ | $n_Y(y_2)$ | $\dots$ | $n_Y(y_n)$ | $N$ |

Table 1: Zajednička tablica frekvencija slučajnog vektora $(X, Y)$.

# Analiza zavisnosti

| | | $Y$ | | | | |
|---|---|---|---|---|---|---|
| | | $y_1$ | $y_2$ | $\ldots$ | $y_n$ | $zbroj$ |
| | $x_1$ | $\widehat{p}(x_1, y_1)$ | $\widehat{p}(x_1, y_2)$ | $\ldots$ | $\widehat{p}(x_1, y_n)$ | $\widehat{p}_X(x_1)$ |
| $X$ | $x_2$ | $\widehat{p}(x_2, y_1)$ | $\widehat{p}(x_2, y_2)$ | $\ldots$ | $\widehat{p}(x_2, y_n)$ | $\widehat{p}_X(x_2)$ |
| | $\vdots$ | $\vdots$ | $\vdots$ | | $\vdots$ | $\vdots$ |
| | $x_m$ | $\widehat{p}(x_m, y_1)$ | $\widehat{p}(x_m, y_2)$ | $\ldots$ | $\widehat{p}(x_m, y_n)$ | $\widehat{p}_X(x_m)$ |
| | $zbroj$ | $\widehat{p}_Y(y_1)$ | $\widehat{p}_Y(y_2)$ | $\ldots$ | $\widehat{p}_Y(y_n)$ | $1$ |

Table 2: Empirijska distribucija slučajnog vektora $(X, Y)$.

# Testiranje hipoteze o nezavisnosti

- hipotezu nezavisnosti slučajnih varijabli $X$ i $Y$ nezavisne možemo zapisati kao

$$H_0 : P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j)$$

za svaki $i = 1 \ldots, m$, $j = 1, \ldots, n$,

- za dovoljno velike uzorke hipoteza se može testirati $\chi^2$ testom
- temelji se na usporedbi očekivanih frekvencija po poljima tablice u uvjetima istinitosti nul-hipoteze s frekvencijama koje u tom polju stvarno imamo na osnovi podataka
- očekivana frekvencija $ij$-tog polja tablice u uvjetima istinitosti nul-hipoteze je

$$E_{ij} = N\widehat{p}_X(x_i)\widehat{p}_Y(y_j) = \frac{n_X(x_i)n_Y(y_j)}{N},$$

dok je eksperimentalna (utvrđena) frekvencija $n_{ij} = n(x_i, y_j)$.

# Testiranje hipoteze o nezavisnosti

- Ako su $X$ i $Y$ nezavisne slučajne varijable, test-statistika

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

  ima $\chi^2$ distribuciju s $(n-1)(m-1)$ stupnjeva slobode. Na temelju realizacije test statistike određujemo pripadnu $p$-vrijednost na uobičajeni način i usporedbom dobivene $p$-vrijednosti s razinom značajnosti $\alpha$ donosimo odluku:
  - ako je $p \leq \alpha$, odbacujemo nul-hipotezu i na razini značajnosti $\alpha$ prihvaćamo alternativnu hipotezu, tj. kažemo da podaci potvrđuju postojanje zavisnosti između varijabli $X$ i $Y$ na razini značajnosti $\alpha$
  - ako je $p > \alpha$, nemamo dovoljno argumenata koji bi poduprli odluku o odbacivanju nul-hipoteze, tj. kažemo da podaci ne daju potvrdu o postojanju zavisnosti među varijablama $X$ i $Y$.

# Testiranje hipoteze o nezavisnosti

- veličina uzorka koja je dovoljna za primjenu ovog testa – npr. znamo da je uzorak dovoljno velik ako su očekivane frekvencije u svakom polju tablice frekvencija veće od $5$.
- valja napomenuti da zavisnost slučajnih varijabli još uvijek ne znači i uzročnu vezu
- može se dogoditi da varijable nisu uzročno povezane, ali imaju neku zajedničku varijablu koja je s objema u uzročnoj vezi.
- **napomena**: test se može koristiti samo za diskretne varijable
- primjer 6.14, zadatak djeca.sta

$\leftrightsquigarrow$