# VERJETNOST

## Lastnosti in računanje verjetnosti

$A$...dogodek
$\bar{A}$ ... naspritni dogodek
$A \cap B$ ... presek dogodkov
$A \cup B$...unija dogodkov
$N$...nemogoč dogodek
$G$...gotov dogodek
$P(A)$ ... verjetnost dogodka (realno število med 0 in 1)
$P(N) = 0 ; P(G) = 1$

**IZREK**

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
$P(\bar{A}) = 1 - P(A)$

**DEFINICIJA**

Dogodka A in B sta **nezdružljiva**, če $A \cap B = N ; P(A \cap B) = 0$

**TRDITEV**

$\overline{(A \cap B)} = \bar{A} \cup \bar{B}$
$\overline{(A \cup B)} = \bar{A} \cap \bar{B}$

## Pogojna verjetnost

$P(A|B)$ ... verjetnost, da se zgodi dogodek A, pri pogoju, da se zgodi dogodek B.
Pri uporabi tega zapisa zahtevamo, da je $P(B) > 0$.

**IZREK**

$P(A|B) = \frac{P(A \cap B)}{P(B)}$ , to lahko vzamemo za def. če je $P(B) > 0$
$P(A \cap B) = P(B) * P(A|B)$, če je $P(B) > 0$
$P(A \cap B) = P(A) * P(B|A)$, če je $P(A) > 0$

## Neodvisni dogodki

A in B sta **neodvisna** ntk.

1. $P(A) = P(A|B)$
2. $P(B) = P(B|A)$

**IZREK**

Dogodka A in B sta neodvisna ntk.

$$P(A \cap B) = P(A) * P(B)$$

- To lahko vzamemo tudi za **definicijo** neodvinosti
- Ta formula dela tudi če $P(A) = 0$ ali $P(B) = 0$

**OPAZKA**

- Če je $B = N$ (ali $A = N$), potem zgornji izrek velja.
  To pomeni, da je N neodvisen od vseh dogodkov.
- Če je $B = G$ (ali $A = G$), potem zgornji izrek velja.
  To pomeni, da je G neodvisen od vseh dogodkov.

## Dvofazni poskus in Bayesov obrazec

**Popoln sistem dogodkov** je množica dogodkov: $B_1$, $B_2$, $B_3$, … , $B_k$

- $B_1 \cup B_2 \cup \ldots \cup B_k = G$
- $B_i \cap B_j = N \ (i \neq j)$
- $P(B_i) > 0 \ (za \ vse \ i)$

1. faza : Hipoteze $H_1$, $H_2$, … , $H_k$ , ki sestavljajo popoln sistem dogodkov
  Poznamo $P(H_i)$ za vse i

2. faza: Opazujemo dogodek A
  Poznamo $P(A|H_i)$ za vse i
  Iščemo P(A)

**FORMULA O POLNI VERJETNOSTI**

$$P(A) = P(A|H_1) * P(H_1) + P(A|H_2) * P(H_2) + \cdots + P(A|H_k) * P(H_k)$$

**BAYESOV OBRAZEC**

$$P(H_i|A) = \frac{P(A|H_i) * P(H_i)}{P(A)}$$

Zgled: Kolikšna je verjetnost, da smo v 1. Metu vrgli enico, če smo v drugem metu vrgli več kot v prvem?

## Zaporedja neodvisnih poskusov

- Veliko krat ponovimo isti poskus, ki lahko uspe ali ne uspe (zgodi se dogodek ali pa ne)
- V vsakem poskusu se dogodek A zgodi z isto verjetnostjo p
- V vsakem poskusu se dogodek A zgodi neodvisno od izidov/rezultatov prejšnjih/vseh poskusov

**BERNULLIJEV OBRAZEC**

$$P(n,p,k) = \binom{n}{k} * p^k * (1-p)^{n-k}$$

n … n-krat ponovimo isti poskus
p … dogodek A se v poskusu zgodi z verjetnostjo p
k … verjetnost, da se v n poskusih dogodek A zgodi k-krat

Zgled: $P(10000, \frac{1}{2}, 5000)$ – verjetnost, da v 10000 metih kovanca, grb pade 5000 krat

## Slučajne spremenljivke

So funkcije, katerih rezultat ni odvisen od podatkov, temveč od slučaja. Pravimo, da je rezultat slučajne spremenljivke realno število.

Opis slučajne spremenljivke X:

a) Zalogo vrednosti
b) Za vsak $X_k$ iz zaloge vrednosti, naj bo $p_a$ verjetnost, da slučajna spremenljivka X zavzame vrednost $x_k$

**VERJETNOSTNA SHEMA**

$$X \sim \begin{pmatrix} x_1 & x_2 & \dots & x_n \\ p_1 & p_2 & \dots & p_n \end{pmatrix}$$

$$p_i \geq 0 \, ; \sum_{j=1}^{n} p_i = 1$$

## Matematično upanje

Matematično upanje oz. pričakovana vrednost označujemo z $E(X)$. Pove nam kakšna je vrednost spremenljivke X v povprečju.

$$E(X) = \sum_{i=1}^{n} x_i * p_i$$

## Disperzija

Disperzija slučajne spremenljivke X, $D(X)$, meri, kako »niha« vrednost X okoli »povprečja« $E(X)$.

$$D(X) = E(X^2) - E(X)^2$$

## Standardni odklon

Standardni odklon, označen z $\sigma(X)$, meri povprečno odstopanje od povprečne vrednosti in ima iste enote kot slučajna spremenljivka.

$$\sigma(X) = \sqrt{D(X)}$$

**IZREK**

Če je $X \sim b(n,p)$, potem je:

$$E(X) = n * p$$
$$q = 1 - p$$
$$D(X) = n * p * (1-p) = n * p * q$$
$$\sigma(X) = \sqrt{n * p * (1-p)} = \sqrt{n * p * q}$$

## Neenakost Markova

**IZREK**

Naj bo slučajna spremenljivka $X \geq 0$. Potem je:

$$P[X \geq a] \leq \frac{E(x)}{a}$$

Zgled: Ocenite verjetnost, da pri metu dveh kock pade vsaj 10!

## Neenakost Čebiševa

Neenačba Čebiševa ocenjuje kakšna je verjetnost, da se slučajna spremenljivka veliko razlikuje od matematičnega upanja.

**IZREK**

X slučajna spremenljivka

$$P[|X - E(X)| \geq t] \leq \frac{D(X)}{t^2} \; ; t > 0$$

Zgled: Kovanec vržemo 1000 krat. Oceni verjetnost da, da bo število grbov med 400 in 600.

## Lastnosti matematičnega upanja in disperzije

**TRDITEV**

Če je neka konstanta in $P[X = a] = 1$, potem je $E(X) = a, D(X) = 0$

**IZREK**

Matematično upanje je linearno, če sta X in Y slučajni spremenljivki in $a, b \in \mathbb{R}$, je

$$E(aX + bY) = a * E(X) + b * E(Y)$$

**PROBLEM**: Za disperzijo takšna trditev **ne** velja.

## Kovarianca slučajnih spremenljivk

Naj bosta X in Y slučajni spremenljivki. **Kovarianca** slučajnih spremenljivk X in Y je količina, definirana z:

$$K(X,Y) \coloneqq E\left(\left(X - E(X)\right) * \left(Y - E(Y)\right)\right) = E(X * Y) - E(Y) * E(X)$$

**Komentar:** $K(X,X) = D(X) = \sigma(X)^2$

**Velja:**

$$|K(X,Y)| \leq \sigma(X) * \sigma(Y)$$
$$r(X,Y) \coloneqq \frac{K(X,Y)}{\sigma(X) * \sigma(Y)}$$

r … korelacijski koeficient slučajnih spremenljivk X in Y

**IZREK:**

Za slučajni spremenljivki X in Y velja

$$D(X+Y) = D(X) + D(Y) + 2 * K(X,Y)$$

**Komentar:** Če sta X in Y neodvisni slučajni spremenljivki, potem sta tudi nekorelirani! $K(X,Y) = 0\,, r(X,Y) = 0$
Obratno ni nujno res. Slučajni spremenljivki X in Y sta lahko nekolerirani in nista neodvisni.

**IZREK:**

Za neodvisni (dovolj nekorelirani) slučajni spremenljivki X in Y velja:

$$D(X+Y) = D(X) + D(Y)$$
$$D(a * X) = a^2 * D(X)\,; a \in \mathbb{R}$$
$$\sigma(a * X) = a * \sigma(X)\,; a > 0$$

## Slučajne spremenljivke z neskončno mnogo vrednostmi

$$E(X) = \sum_{i=1}^{\infty} x_i * p_i$$

$$D(X) = E(X^2) - E(X)^2 = \sum_{i=1}^{\infty} {x_i}^2 * p_i - E(X)^2$$

Če vrsta ne konvergira matematično upanje te spremenljivke ne obstaja . Lahko se zgodi da disperzija ne obstaja, upanje pa. Obratno ni mogoče.

**POTENČNA VRSTA:**

$$E(X) = \sum_{i=1}^{\infty} i * \frac{1}{2^i}$$

$$D(X) = \sum_{i=1}^{\infty} i^2 p_i - E(X)$$

Zgledi: Mečemo kovanec ; X … zaporedno številko meta v katerem prvič pade grb
Igralnica: stavimo 1€ na rdeče ; dobimo ali izgubimo ; če izgubimo stavimo 2€ na rdeče

## Zvezno porazdeljene slučajne spremenljivke

X je opisana z gostoto verjetnosti $g_x(x)$

- $g_x(X)$ je funkcija : $\mathbb{R} \to \mathbb{R}$
- $g_x(X) \geq 0$
- $\int_{-\infty}^{\infty} g_x(X)\,dx = 1$
- Odsekana zvezna (zvezna povsod, razen morda v nekaj točkah)

Računanje verjetnosti:

$$P[a < X < b] = \int_a^b g_x(X)\,dx \;; a, b \in \mathbb{R}\,; a < b$$
$$P[X = a] = P[X = b] = 0$$

Matematično upanje:

$$E(X) = \int_{-\infty}^{\infty} x * g_x(X) * dx$$

Disperzija:

$$D(X) = \int_{-\infty}^{\infty} x^2 * g_x(X) * dx - E(X)^2$$

### Enakomerna zvezna slučajna spremenljivka na [a,b]

$$E(X) = \frac{a+b}{2}$$
$$D(X) = \frac{1}{12}(b-a)^2$$
$$\sigma(X) = \frac{1}{\sqrt{12}}(b-a)$$

## Normalna porazdelitev

Normalna porazdelitev je »limita« binomske porazdelitve b(n,p).
$E(X) = n * p =: a$
$D(X) = n * p * (1-p)$
$\sigma(X) = \sqrt{n * p * (1-p)} =: \sigma$

Je t.i. normalna porazdelitev $N(a, \sigma)$ porazdeljena s porazdelitveno gostoto:

$$g(X) = \frac{1}{\sigma\sqrt{2\pi}} * e^{-\frac{(X-a)^2}{2\sigma^2}}$$

### Standardna normalna slučajna spremenljivka

$Z \sim N(0,1)$ je standardna (standardizirana) normalna slučajna spremenljivka. Njena porazdelitvena gostota je enaka:

$$\boldsymbol{\varphi(X) = \frac{1}{\sqrt{2\pi}} * e^{-\frac{x^2}{2}}}$$

**IZREK:**

Naj bo $X \sim N(a, \sigma)$. Potem je slučajna spremenljivka $Z := \frac{X-a}{\sigma}$ porazdeljena standardno normalno $Z \sim N(0,1)$.

**Nauk:**

Če vemo, da je slučajna spremenljivka porazdeljena normalno $[X \sim N(a, \sigma)]$, lahko problem verjetnosti dogodkov v zvezi z X prevedemo na računanje verjetnosti v zvezi **s standardno normalno porazdeljeno slučajno spremenljivko.**

## Centralno limitni izrek

**Centralni limitni izrek** pravi, da vsako zaporedje enako porazdeljenih neodvisnih slučajnih spremenljivk z istim matematičnim upanjem in disperzijo teži k normalni porazdelitvi.

Če vzamemo naključni vzorec n opazovanj iz katerekoli populacije dobimo (n > 30) distribucijo povprečij katera je porazdeljena približno normalno s povprečjem enakim povprečju populacije in standardnim odklonom enakim $\frac{1}{\sqrt{n}}$ standardnega odklona celotne populacije.

Statistikom omogoča aproksimacijo zaporedja podatkov z neznanimi porazdelitvami kot porazdeljene.

$X_1, X_2, X_3, \ldots$ zaporedje **neodvisnih** slučajnih spremenljivk, ki imajo vse isto:

- Matematično upanje $E(X_i) = a$
- Disperzijo $D(X_i) = \sigma^2$

Opazujemo $S_1, S_2, S_3, \ldots$
$S_n = X_1+X_2+X_3+\ldots+X_n$ ($N \geq 30$ če želimo uporabiti CLI)

<u>Vemo:</u>

- $E(S_n) = n * a$
- $D(S_n) = n * \sigma^2$

Za zaporedje $S_n$ velja:

$$\lim_{n\to\infty} P\left(\frac{S_n - E(S_n)}{\sigma(S_n)} < X\right) = \int_{-\infty}^{X} \frac{1}{\sqrt{2\pi}} * e^{-\frac{t^2}{2}} dt \; ; rešitev: \leq 1$$

**BERI:**

Za velike n je

$$\frac{S_n - E(S_n)}{\sigma(S_n)} \sim N(0,1)$$
$$S_n \sim N(n * a, \sqrt{n} * \sigma)$$

CLI velja tudi če:

X1,X2,X3, … nimajo istega matematičnega upanja in iste disperzije morajo pa biti še vedno neodvisne in »primerljive« po velikost

$$S_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$
$$S_n = N(a, \frac{\sigma}{\sqrt{n}})$$

Zgledi:

- Kovanec vržemo 1000 krat. Oceni verjetnost, da grb pade med 400 in 600 krat.

# STATISTIKA

- Deskriptivna statistika (opisna)
- Inferenčna statistika

## Vzorčenje

**Populacija** – velika količina podatkov – N

**Vzorec** - manjša podmnožica – n

**Slučajni vzorec** – slučajno izbrani vzorec, vsak vzorec enake velikosti, ima enako verjetnost, da bo izbran:

- Brez ponavljanja
- S ponavljanjem

Zgledi: Volitve – vzporedne volitve ; kvaliteta proizvodov

Cilj statistične analize: iz lastnosti vzorca sklepati na lastnosti celotne populacije
**stopnja zaupanja –** verjetnost, da populacija res ima takšno lastnost kot napoveduje vzorec

Velja: pri velikih populacijah (velikih N) je skoraj vseeno ali vzorčimo s ponavljanjem ali brez.

Za nas: Vse populacije so dovolj velike, da je ta razlika zanemarljiva.

Če na vzorcu izračunamo neko količino, potem dobljenemu podatku pravimo **statistika**. Če je vzorec slučajen je vsaka statistika slučajna spremenljivka.

**OSNOVNA STATISTIKA**

Vzorčno povprečje spremenljivke Y na populacija
Y – ocena .. podatki
$\bar{Y}$ – povprečje ocen na vzorcu ... slučajna spremenljivka

Na populaciji imamo N vrednosti spremenljivke Y: $Y_1, Y_2, \dots, Y_N$
Izberemo slučajni vzorec velikost n, izbrali smo vrednosti: $Y_{i1}, Y_{i2}, \dots, Y_{in}$

$$\bar{Y} = \frac{1}{n}\sum_{k=1}^{n} Y_{ik}$$

**IZREK**

Matematično upanje vzorčnega povprečja $\bar{Y}$ je enako povprečni vrednosti spremenljivke Y na celi populaciji.

Dosledna statistika: statistika, ki ima matematično upanje enako pravi vrednosti

Vzorčno povprečje je dosledna statistika za povprečje na populaciji.

**IZREK**

Naj bo Y spremenljivka na neki populaciji velikosti N.

$\mu = E(X)$ ... povprečna vrednost spremenljivke
$D = \sigma^2$ ... disperzija spremenljivke Y in standardni odklon

Potem velja za vzorčenje brez ponavljanja:

a) $E(\bar{Y}) = E(Y) = \mu$
b) $D(\bar{Y}) = D(Y) * \frac{N-n}{n*(N-1)} \xrightarrow{N\to\infty} \frac{D(Y)}{n}$
   če je n = 1 potem $D(\bar{Y}) = D(Y)$
   če je $n = N$ potem ima $\bar{X}$ eno samo vrednost $=> D(\bar{X}) = 0$
   $\sigma(\bar{Y}) = \frac{\sigma(Y)}{\sqrt{n}}$
c) Če je Y na populaciji normalno porazdeljen z $N(\mu, \sigma)$, potem je $\bar{Y}$ porazdeljena kar normalno z: $N(\mu, \frac{\sigma}{\sqrt{n}})$
d) Če je Y poljubno porazdeljena je $\bar{Y}$ »približno normalno porazdeljena z: $N(\mu, \frac{\sigma}{\sqrt{n}})$

## Disperzija vzorca

X ... spremenljivka

$\bar{X}$ ... vzorčno povprečje

$S^2$ ... disperzija vzorca (velikost n)

$$S^2 = \frac{1}{n}\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right]$$

Popravljena vrednost

$$\hat{S}^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right]$$

Če ne poznamo disperzije D(X) (torej parametra σ) potem tudi ne poznamo porazdelitve za $\bar{X}$, torej moramo parameter σ oceniti.

Za oceno uporabimo $\hat{S}^2$.

$\hat{S}^2$ ... je cenilka za parameter D(X)
$\hat{S}^2$... je cenilka za parameter σ

# Ocenjevanje parametrov

Ocenjujemo parameter $\mu = E(X)$

N – velikost populacije
n – velikost slučajno izbranega vzorca
X – slučajna spremenljivka

$\mu = E(X)$
$\sigma = \sigma(X)$ $(\sigma^2 = D(X) = D)$

$\bar{X} - \ cenilka\ za\ \mu \sim N(\mu, \frac{\sigma}{\sqrt{n}}) \sim N(\mu, \frac{\hat{S}}{\sqrt{n}})$

**DEFINICIJA**

Stopnja zaupanja je izbrano število γ (npr. $\gamma = 0{,}95$ ali $\gamma = 0{,}99$)

Interval zaupanja je interval $[\bar{X} - k, \bar{X} + k]$ na katerem leži prava vrednsot $\mu$ z verjetnostjo γ.

$$P(\mu \in [\bar{X} - k, \bar{X} + k]) = \gamma$$
$$P(|\mu - \bar{X}| \leq k) = \gamma$$
$$P(\bar{X} \in [\mu - k, \mu + k]) = \gamma$$

$\bar{X}\ in\ \mu$ nastopata simetrično v tej verjetnosti.

# Algoritem za računanje intervala zaupanja

$$X \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$
$$\Phi: N(0,1)$$
$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}\ ; N(0,1)$$

## $\sigma$ znan

1. Izberemo γ
2. Določimo c:

$$\Phi(c) = \frac{\gamma}{2}$$
$$\Phi(c) = P(0 \leq Z \leq c) = \frac{1}{2}(-c \leq Z \leq c)$$

3. Izračunamo $\bar{X}$
4. $k = \frac{c * \sigma}{\sqrt{n}}$
   $[\bar{X} - k, \bar{X} + k]$ interval zaupanja za $\mu$

## $\sigma$ ni znan, velik vzorec $n \geq 30$

$\sigma$ ocenimo s parametrom $\hat{S}$

3. Izračunamo $\bar{X}\ in\ \hat{S}$
4. $k = \frac{c * \hat{S}}{\sqrt{n}}$

### $\sigma$ ni znan, velik vzorec $n \leq 30$

Studentova porazdelitev

t-porazdelitev

## Porazdelitev in interval zaupanja za standardni odklon $(\hat{S})$

Kako statistično ocenimo vrednost σ?

**IZREK**

Porazdelitev spremenljivke

$$\frac{n * \hat{S}^2}{\sigma^2} = \frac{(n-1) * \hat{S}^2}{\sigma^2} \approx \chi^2 \; z \; n-1 \; prostostnimi \; stopnjami$$

### Kaj je $\chi^2$?

$X_1, X_2, \dots, X_n$ … neodvisne porazdeljene normalno

$X_i \sim N(0,1)$

$\chi^2 = \chi_1^2 + \chi_2^2 + \cdots + \chi_n^2$ ; n – število prostostnih stopenj

Porazdelitvena funkcija slučajne spremenljivke $\chi$

$$P(\chi^2 \leq a) = \frac{1}{2^{\frac{n}{2}} * \gamma * \frac{n}{2}} \int_0^a t^{\frac{n}{2}-1} * e^{-\frac{t}{2}} dt$$

Gostota verjetnosti:

$$\frac{1}{2^{\frac{n}{2}} * \gamma * \frac{n}{2}} * \chi^{\frac{n}{2}-1} * e^{-\frac{x}{2}}$$

## Kako izračunati interval zaupanja za $\hat{S}$ (oziroma za $\sigma$)?

Določimo $\gamma$ (gledamo tabelo $\chi^2$ z n prostostnimi stopnjami)

$C_1 : P(\chi^2 < c_1) = \frac{1-\gamma}{2}$

$C_2 : P(\chi^2 < c_2) = \frac{1+\gamma}{2}$

Interval zaupanja:

$$\left[\frac{\sqrt{n} * \hat{S}}{\sqrt{c_2}}, \frac{\sqrt{n} * \hat{S}}{\sqrt{c_1}}\right]$$

# Preverjanje hipotez

**STATISTIČNI TESTI**

Statistična hipoteza:

- O tipu porazdelitve slučajne spremenljivke
- O parametru porazdelitve

$H_0$ ... ničelna hipoteza
$H_{alt}$ ... alternativna hipoteza

Izvedemo test, na podlagi testa $H_0$ sprejmemo ali zavrnemo.

- Neparametrični testi – tip spremenljivke
- Parametrični testi – parametru porazdelitve

Stopnja značilnosti, tveganje:
$\alpha = 1 - \gamma$ ... verjetnost, da bomo pravilno hipotezo z našim testom zavrnili

**Vrste napak:**

- 1. Vrste: pravilno hipotezo $H_0$ zavrnemo
- 2. Vrste: napačno hipotezo $H_0$ sprejmemo.

**Stopnja značilnosti α:** verjetnost, da smo naredili napako 1. Vrste. Lahko si izberemo poljubno majhno.

**Ničelno hipotezo lahko**:

- Zavrnemo
- Ne zavrnemo - $H_0$ velja pri dani stopnji zaupanja γ

## *Parametrični test*

**DVOSTRANSKI**

$H_0 = \frac{1}{2}$
$H_{alt} \neq \frac{1}{2}$

Zgled: število fantkov med N novorojenčki

**ENOSTRANSKI**

## *Neparametrični test $\chi^2$*

Testiramo ali je porazdelitev na populaciji enaka neki dani porazdelitvi.

Zgled: met kocke ; n = 600 ; štejemo frekvence posameznih izidov ; $H_0$ : kocka je poštena

$$\begin{array}{cccc} x = x_1 & x = x_2 & \dots & x = x_k \quad -\, zaloga\ vrenodsti \\ X_1 & X_2 & \dots & X_k \quad -\, frekvence \\ n * p_1 & n * p_2 & \dots & n * p_k \quad -\, pričakovane\ frekvence \end{array}$$

$$\chi^2 = \sum_{i=1}^{k} \frac{(X_i - n * p_i)^2}{n * p_i}$$

**IZREK**

Če je n dovolj velik in če je $n * p_i \geq 5$ za vsak i, potem se $\chi^2$ porazdeljuje približno tako, kot $\chi^2$ s $k - 1$ prostostnimi stopnjami

$\chi^2$ - test: $H_0$ porazdelitev na populaciji je enaka dani porazdelitvi

1. Izberemo α (1% ali 5%)
2. Izračunamo $\chi^2$
3. Primerjamo $\chi^2$ z vrednostjo $\chi^2_{1-\alpha}$ s $k - 1$ prostostnimi stopnjami
4. Če je $\chi^2 > \chi^2_{1-\alpha}$ => zavrnjemo
   Če ke $\chi^2 \leq \chi^2_{1-\alpha}$=> ne zavrnemo

### *Peardonov $\chi^2$-test za ugotavljanje neodvisnosti*

$\chi^2$ test statistične signifikantnosti je serija matematičnih formul, ki primerjajo opazovane frekvence nekega dogodka v vzorcu s frekvencami katere bi pričakovali, če ne bi bilo nobene povezave med spremenljivkama v večji (vzorčeni) populaciji. $\chi^2$ testira naše trenutne rezultate proti ničelni hipotezi in pove ali so rezultati dovolj različni da preidejo določeno verjetnost da so nastali zaradi napačnega vzorčenja.
Test lahko uporabimo za testiranje, če sta dve spremenljivki odvisni ali ne.

Zgled: izbira vrste bureka je neodvisna od starostne skupine kupca

Testiramo: ali je porazdelitev po vrsticah enaka porazdelitvi v spodnji vrstici

$$Y_1 = \sum_{i=1}^{m} X_{i1}$$

Testna statistika:

$$\chi^2 = n * \sum_{j=1}^{n} \sum_{i=1}^{m} \frac{(X_{ij} - \frac{X_i * Y_i}{n})}{X_i * Y_i} \underset{H_0}{\sim} \chi^2\ (\,(n-1)(m-1))$$

**Nato isto kot pri $\chi^2$ tets**

# Regresija

$y^{'} = f(x)$
kaže, kako X vpliva na Y

Regresijskih funkcij je lahko več.

Najbolj pogosta **linearna regresija!**

$y^{'} = a + bx$ ... iščemo parametra a in b, tako da bo ta premica najlepše opisala podatke

**Metoda najmanjših kvadratov!**

$$F(a,b) = \sum_{i=1}^{n}(Y_i - Y_i^{'})^2 = \sum_{i=1}^{n}(Y_i - (a + bX_i))^2$$

iščemo minimum funkcije
$F_a = 0$ in $F_b = 0$ (parcialni odvod)

**Rešitev:**

Regresijska premica:

$$Y^{'} = E(Y) + \frac{K(X,Y)}{D(X)} * (X - E(X))$$

Druga regresijska premica:

$$X^{'} = \alpha Y + \beta = E(X) + \frac{K(X,Y)}{D(X)} * (Y - E(Y))$$

Na vzorcu:

$$Y^{'} = \bar{Y} + \frac{\hat{K}(X,Y)}{(\hat{\sigma}(X))^2} * (X - \bar{X})$$
$$X^{'} = \bar{X} + \frac{\hat{K}(X,Y)}{(\hat{\sigma}(Y))^2} * (Y - \bar{Y})$$

Premici se sekata v točki: $\big(E(X), E(Y)\big)$ ; $(\bar{X}, \bar{Y})$

<u>Ali regresijska premica (funkcija) odraža odvisnost med X in Y ali ne?</u>

**Korelacijski koeficient:**

$$\rho(X,Y) = \frac{K(X,Y)}{\sigma(X), \sigma(Y)}$$
$$\rho(X,Y) \in [-1,1]$$

$\rho = 0$ => X in Y nista korelirani
$\rho = 1$ => X in Y sta močno korelirani

Na vzorcu: $\gamma(X,Y) = \frac{\hat{K}(X,Y)}{\hat{\sigma}(X), \hat{\sigma}(Y)}$