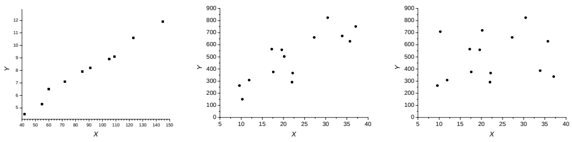
Korelacije

Razni stupnjevi rasipanja:



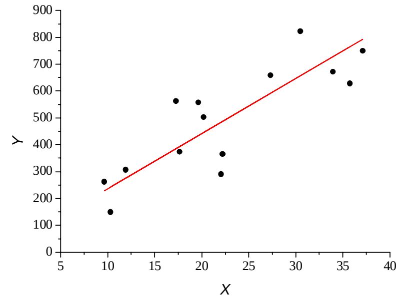
Proučavamo srednju sliku:

ovisnost Y o X:

$$y = ax + b$$

$$a = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - \bar{x}^2}$$

$$b = \frac{\overline{x^2}\,\bar{y} - \bar{x}\,\overline{xy}}{\overline{x^2} - \bar{x}^2}$$



$$a = \frac{m_{11} - m_{10}m_{01}}{m_{20} - m_{10}^2} = \frac{M_{11}}{M_{20}} = \frac{\sigma_{XY}}{\sigma_X^2}$$

kovarijanca :

$$M_{11} = \sigma_{XY} = \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y})$$

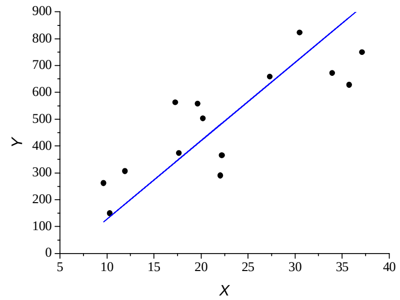
$$M_{11} = \frac{1}{n}\sum_i x_i y_i - \overline{x}\overline{y} = \frac{1}{n}\sum_i x_i y_i - \frac{1}{n^2}\sum_i x_i \sum_i y_i$$

Ovisnost X o Y:

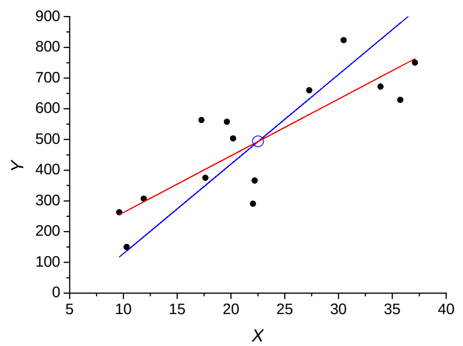
$$x = cy + d$$

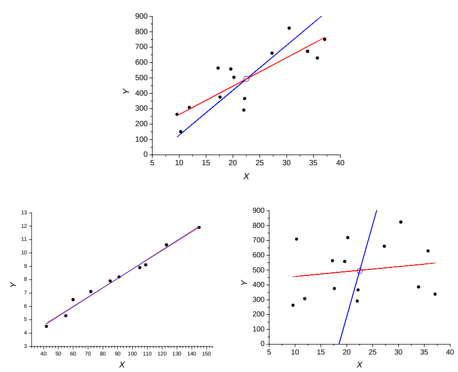
$$c = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{y^2} - \bar{y}^2}$$

$$d = \frac{\overline{y^2}\,\bar{x} - \bar{y}\,\overline{xy}}{\overline{y^2} - \bar{y}^2}$$



$$c = \frac{m_{11} - m_{10}m_{01}}{m_{02} - m_{01}^2} = \frac{M_{11}}{M_{02}} = \frac{\sigma_{XY}}{\sigma_Y^2}$$





Jednadžbe pravaca regresije pomoću kovarijance

Ovisnost Y o X:

jednadžba pravca: $(y-\bar{y})=\frac{\sigma_{XY}}{\sigma_X^2}(x-\bar{x})$

Ovisnost X o Y:

jednadžba pravca: $(x-\bar{x})=\frac{\sigma_{XY}}{\sigma_Y^2}(y-\bar{y})$

standardizirane jednadžbe pravaca:

$$\frac{(y-\bar{y})}{\sigma_Y}=\frac{\sigma_{XY}}{\sigma_X\sigma_Y}\frac{(x-\bar{x})}{\sigma_X}$$

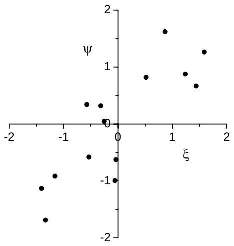
$$\frac{(x-\bar{x})}{\sigma_X}=\frac{\sigma_{XY}}{\sigma_X\sigma_Y}\frac{(y-\bar{y})}{\sigma_Y}$$

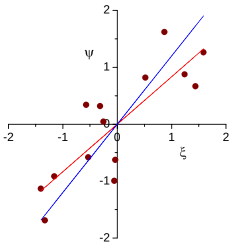
bezdimenzionalne varijable: $\xi=\frac{(x-\bar{x})}{\sigma_X}$ i $\psi=\frac{(y-\bar{y})}{\sigma_Y}$

novi pravci regresije: $\psi=\rho\xi$ $\quad\quad$ $\xi=\rho\psi$

Def.: **koeficijent korelacije:** $\boxed{\rho=\frac{\sigma_{XY}}{\sigma_X\sigma_Y}}$

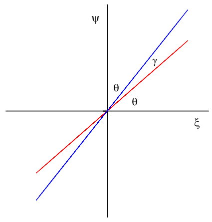
bezdimenzionalno
ne ovisi o jedinicama





Značenje koeficijenta korelacije

poprima vrijednosti: $-1\le\rho\le 1$



$\rho=\mathrm{tg}\,\theta$

$\gamma=\frac{\pi}{2}-2\theta$

$\mathrm{tg}\,\gamma=\frac{1-\rho^2}{2\rho}$

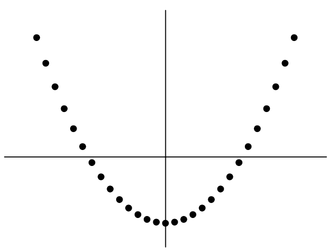
X i Y nezavisni $\Rightarrow$ $\gamma=\frac{\pi}{2}$, $\rho=0$

X i Y linearno deterministički vezani $\Rightarrow \gamma=0$, $\rho=-1$ ili $\rho=1$

Dogovor: dobra korelacija $\rightarrow |\rho|>0{,}5$

Napomena:
Kad su X i Y nezavisne varijable, koeficijent korelacije težit će nuli ($\rho\approx 0$). Međutim, činjenica da je $\rho=0$ ne znači nužno da su varijable nezavisne.

Primjer:



Provjeravanje (testiranje) hipoteza

Provjera hipoteze vrlo je bitan dio statističkog zaključivanja.

Da bi se takva provjera formulirala, potrebno je postaviti neku teoriju koja se želi dokazati.
Npr.:

- Novi lijek bolji je za liječenje određenih simptoma od starog.
- Igraća kocka ima pomaknuto težište
- U svijetu se rađa više muškaraca nego žena

U svakom takvom problemu postavljamo dvije tvrdnje (hipoteze) od kojih je točno jedna istinita:

H_0= nul-hipoteza

H_1= alternativna hipoteza

One se ne tretiraju ravnopravno:

U sudskom procesu: H_0: optuženi je nevin

H_1: optuženi je kriv

Hipoteza H_0 smatra se ispravnom dok se ne dokaže H_1.

Za H_0 obično se uzima stara, postojeća teorija:

Npr., za lijek: H_0= stari lijek jednako je dobar kao i novi

za kocku: H_0= kocka je poštena

Ako ne odbacimo H_0, to ne znači da je ona ispravna, nego samo da nemamo dovoljno dokaza da ju odbacimo.

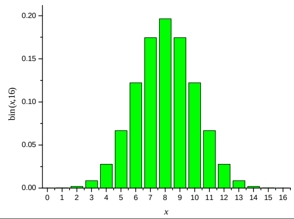
Primjer: ispitujemo simetričnost Galtonove daske sa 16 redova.

1. Postavljanje problema, izricanje hipoteza

H_0 : daska je simetrična, tj. X~Bin(16, ½)

H_1 : daska je nagnuta udesno

Raspodjela dana nul-hipotezom:



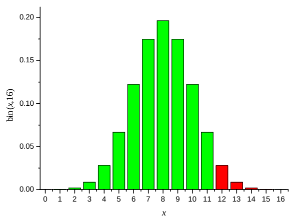
---

2. Postavljanje problema, kritično područje

Prije uzimanja uzorka podijelimo skup mogućih ishoda u dva područja:

A = područje prihvaćanja H_0

B = područje odbacivanja H_0 = **kritično područje**



Odlučimo se: A={0,1,2,…,11} B={12,13,…,16}

Vrste pogrešaka

		odluka	
		odbaci H_0	prihvati H_0
istina	H_0	pogreška I. vrste	ispravan zaključak
	H_1	ispravan zaključak	pogreška II. vrste

Smatra se da je pogreška I. vrste mnogo ozbiljnija od pogreške II. vrste.

Vjerojatnost pogreške I. vrste:

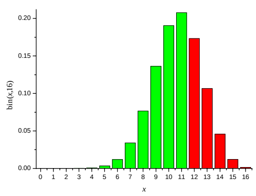
α = vjerojatnost odbacivanja H_0 kada je istinita = $P(B/H_0)$

Vjerojatnost pogreške II. vrste:

β = vjerojatnost prihvaćanja H_0 kada nije istinita = $P(A/H_1)$

U našem primjeru je $\alpha = 0{,}038 \approx 4\%$.

Međutim, čak i da je daska toliko nagnuta da je p = 2/3, raspodjela varijable X bi bila



pa bi vjerojatnost pogreške II. vrste uz naš odabir kritičnog područja bila $\beta = 0{,}6488$.

Veliki β se događa kada je uzorak premalen (n = 16).

Signifikantnost testa (važnost, značajnost)

Def:

Za postupak provjere kažemo da ima **razinu signifikantnosti** (važnosti) α ako je

$$P(\text{pogreška I. vrste}) \le \alpha .$$

Kažemo da je to **test razine α**.

Naš primjer je test razine 0,04.

Tradicionalno se kao razine signifikantnosti uzimaju vrijednosti 0,01; 0,05 ili 0,10.

Za α = 0,05 kažemo da je test signifikantan, a za α = 0,01 kažemo da je test vrlo signifikantan.

Moć testa

Moć statističkog testa mjeri sposobnost testa da odbaci nul-hipotezu kad je uistinu pogrešna, tj. da učini ispravnu odluku.

$$P = 1-\beta$$

U našem primjeru je P = 0,34. Idealno je P = 1. Za snažniji test morali bismo imati veći uzorak.

Opća pravila odabira testa

Prije uzimanja uzorka:

1)Izreci nul-hipotezu i alternativnu hipotezu.
2)Razmotri odgovarajuću raspodjelu danu nul-hipotezom.
3)Odluči o razini signifikantnosti testa.
4)Odredi kritično područje (odluči o kriteriju odbacivanja nul-hipoteze)

Sada uzmi uzorak!

5)Očitaj rezultat testa (izračunaj vrijednost statistike testa)
6)Učini odluku:
Ako je vrijednost statistike testa u kritičnom području, odbaci H_0!
Ako vrijednost statistike testa nije u kritičnom području, nemoj odbaciti H_0!

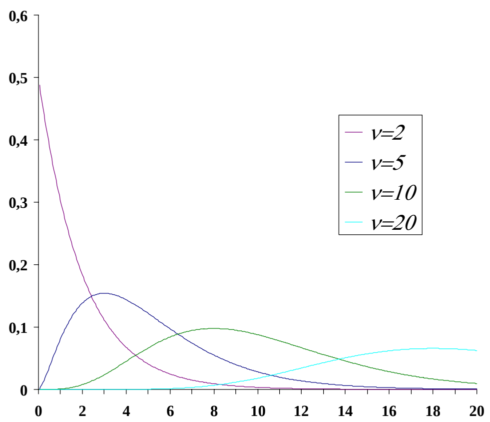
Testiranje dobrote prilagodbe

χ^2 raspodjela

Neka je $\nu \in \mathbf{N}$. Kontinuirana slučajna varijabla X ima χ^2 **raspodjelu** s parametrom ν ako je njezina funkcija gustoće vjerojatnosti dana s

$$f(x;\nu) = \begin{cases} \dfrac{x^{\nu/2-1}e^{-x/2}}{2^{\nu/2}\Gamma(\nu/2)} & , \quad x \geq 0 \\ 0 & , \quad x < 0 \end{cases}$$

Parametar ν zove se “broj stupnjeva slobode” varijable X.

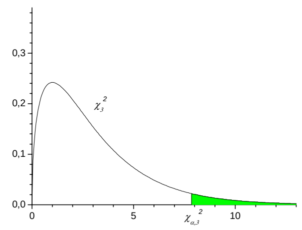


Vjerojatnost da varijabla X poprimi vrijednost veću od neke određene vrijednosti

$$\chi^2_{\alpha,\nu}$$

dana je površinom ispod repa krivulje:

$$P\left(X > \chi^2_{\alpha,\nu}\right) = \int_{\chi^2_{\alpha,\nu}}^{\infty} \chi^2_{\nu}(x)dx$$



Indeks α označava iznos te vjerojatnosti. Ako X poprimi vrijednost veću od $\chi^2_{\alpha,\nu}$, kažemo da je u kritičnom području.

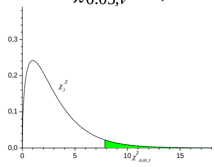
Obično unaprijed odredimo koliku vjerojatnost α želimo pa na osnovu toga određujemo kritičnu vrijednost $\chi^2_{\alpha,\nu}$.

Te vjerojatnosti je teško izračunati pa se služimo tablicama.
U testiranju hipoteza najčešće nas zanimaju vjerojatnosti:

$$P\left(X > \chi^2_{0{,}05,\nu}\right) = 0{,}05 \qquad P\left(X > \chi^2_{0{,}01,\nu}\right) = 0{,}01$$

Npr., za ν =3, vrijednosti su:

$$\chi^2_{0.05,\nu} = 7{,}82 \qquad \chi^2_{0.01,\nu} = 11{,}34$$



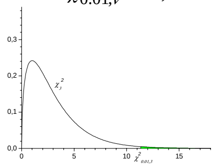
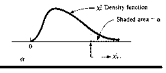


Table A.6 Critical Values $X^2_{\alpha,\nu}$ for the Chi-Squared Distribution



ν \ α	.995	.99	.975	.95	.90	.10	.05	.025	.01	.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.843	5.025	6.637	7.882
2	0.010	0.020	0.051	0.103	0.211	4.605	5.992	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.344	12.837
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.832	15.085	16.748
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.440	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.012	18.474	20.276
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.534	20.090	21.954
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.022	21.665	23.587
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.724	26.755
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.041	19.812	22.362	24.735	27.687	29.817
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.600	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.577	32.799
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.407	7.564	8.682	10.085	24.769	27.587	30.190	33.408	35.716
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.843	7.632	8.906	10.117	11.651	27.203	30.143	32.852	36.190	38.580
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.033	8.897	10.283	11.591	13.240	29.615	32.670	35.478	38.930	41.399
22	8.643	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289	42.796
23	9.260	10.195	11.688	13.090	14.848	32.007	35.172	38.075	41.637	44.179
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.558
25	10.519	11.523	13.120	14.611	16.473	34.381	37.652	40.646	44.313	46.925
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.807	12.878	14.573	16.151	18.114	36.741	40.113	43.194	46.962	49.642
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.120	14.256	16.147	17.708	19.768	39.087	42.557	45.772	49.586	52.333
30	13.787	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
31	14.457	15.655	17.538	19.280	21.433	41.422	44.985	48.231	52.190	55.000
32	15.134	16.362	18.291	20.072	22.271	42.585	46.194	49.480	53.486	56.328
33	15.814	17.073	19.046	20.866	23.110	43.745	47.400	50.724	54.774	57.646
34	16.501	17.789	19.806	21.664	23.952	44.903	48.602	51.966	56.061	58.964
35	17.191	18.508	20.569	22.465	24.796	46.059	49.802	53.203	57.340	60.272
36	17.887	19.233	21.336	23.269	25.643	47.212	50.998	54.437	58.619	61.581
37	18.584	19.960	22.105	24.075	26.492	48.363	52.192	55.667	59.891	62.880
38	19.289	20.691	22.878	24.884	27.343	49.513	53.384	56.896	61.162	64.181
39	19.994	21.425	23.654	25.695	28.196	50.660	54.572	58.119	62.426	65.473
40	20.706	22.164	24.433	26.509	29.050	51.805	55.758	59.342	63.691	66.766

For $\nu > 40$, $\chi^2_{\alpha,\nu} \approx \nu\left(1 - \frac{2}{9\nu} + z_\alpha\sqrt{\frac{2}{9\nu}}\right)^3$

χ^2-test

Razmotrimo pokus koji daje n opaženih frekvencija f_i.
Želimo li postaviti hipotezu da ta opažanja slijede neku teorijsku raspodjelu izračunamo očekivane teorijske vrijednosti f_{ti}.
Opažene frekvencije odstupaju od teorijskih, a mi želimo donijeti odluku mogu li se odstupanja pripisati slučaju ili ne.
Nul-hipoteza je: "*Opažanja slijede teorijsku raspodjelu*".

Definiramo statistiku: $$\chi^2_{\text{op}} = \sum_{i=1}^{n} \frac{(f_i - f_{ti})^2}{f_{ti}}$$

Teorem (K. Pearson):

Statistika χ^2_{op} približno je raspodijeljena prema χ^2 raspodjeli sa stupnjem slobode ν koji ovisi o broju opažanja (ili razreda) i o broju postavljenih ograničenja.

Napomene:
- Uzimanjem χ^2 raspodjele, diskretnu raspodjelu aproksimiramo kontinuiranom. Ta aproksimacija ne vrijedi ako je očekivana frekvencija manja od 5. Taj problem prevladavamo grupiranjem nekoliko razreda male frekvencije u jedan razred dovoljno velike frekvencije.
- Ako je vrijednost χ^2_{op} vrlo mala, mudro je upitati se nisu li opaženi rezultati namješteni.

Stupnjevi slobode

Parametar ν definiran je kao:
ν = broj nezavisnih varijabli uključenih u izračun

Nalazimo ga na sljedeći način:
ν = broj razreda – broj ograničenja

Normalna raspodjela, μ i σ poznati

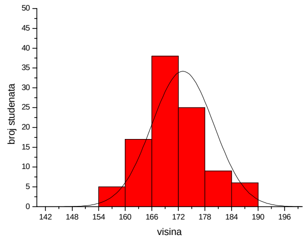
Primjer:
Dugogodišnje statistike pokazuju da je visina studenata na nekom sveučilištu normalno raspodijeljena s očekivanjem μ = 173 cm i standardnom devijacijom σ = 7 cm.
Iz jedne generacije studenata na tom sveučilištu izdvojeno je 100 studenata, mjerene su njihove visine i svrstane u razrede.
Opažene su frekvencije:

x	154-160	160-166	166-172	172-178	178-184	184-190	Total
$f(x)$	5	17	38	25	9	6	100

Provjeri s 5% signifikantnosti je li raspodjela normalna s s očekivanjem μ = 173 cm i standardnom devijacijom σ = 7 cm !

Rješenje:

H_0: *Raspodjela je Gaussova s očekivanjem* μ = 173 cm *i standardnom devijacijom* σ = 7 cm.



Imamo jedno ograničenje: $\Sigma f_i = \Sigma f_{ti} = N = 100$

Uvodimo standardnu normalnu slučajnu varijablu $z_i = \dfrac{x-\mu}{\sigma}$.

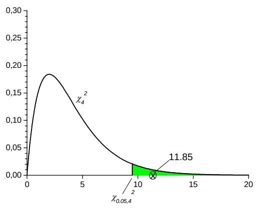
Za te vrijednosti z_i određujemo iz tablica funkciju raspodjele $F(z)$ i vjerojatnosti razreda $f(z)$:

razred *i*	f_i	z_i	$F(z)$	$f(z)$	f_{ti}	$(f_i-f_{ti})^2/f_{ti}$
154-160	5	z<-1,86	0,0314	0,0314	3	
160-166	17	-1,86<z<-1	0,1587	0,1273	13	2,25
166-172	38	-1<z<-0,14	0,4443	0,2856	28	3,57
172-178	25	-0,14<z<0,71	0,7611	0,3168	32	1,53
178-184	9	0,71<z<1,57	0,9418	0,1807	18	4,5
184-190	6	z>1,57	1	0,0582	6	0
zbroj	100			1	100	11,85

Dakle, imamo pet razreda i jedno ograničenje pa je ν =5-1=4

Za 4 stupnja slobode i 5% signifikantnosti, kritična vrijednost (iz tablica) je

$$\chi^2_{0.05,4} = 9.49$$



Izračunata vrijednost $\chi^2_{op} = 11.85$ pada u kritično područje.

Stoga hipotezu H_0 **odbacujemo**.

Normalna raspodjela, μ i σ nepoznati

Isti primjer:

Odbacili smo hipotezu o očekivanju i standardnoj devijaciji, ali i dalje vjerujemo da je raspodjela normalna.

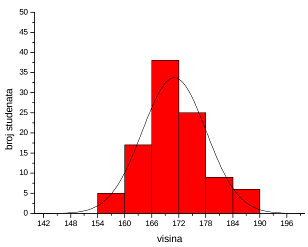
Provjeri s 5% signifikantnosti je li raspodjela normalna!

Rješenje:

H_0: *Raspodjela je Gaussova s očekivanjem* $\mu = \bar{x}_{uzorka}$ *i standardnom devijacijom* $\sigma = \sigma_{uzorka}$.

Izračunamo:

$$\bar{x}_{uzorka} = \frac{1}{N}\sum f_i x_i = 171\,\text{cm} \qquad \sigma_{uzorka} = \sqrt{\frac{1}{N}\sum f_i (x_i - \bar{x}_{uzorka})^2} = 7{,}1\,\text{cm}$$



razred *i*	f_i	z_i	$F(z)$	$f(z)$	f_{ti}	$(f_i-f_{ti})^2/f_{ti}$
154-160	5	z<-1,55	0,0606	0,060	6	0,167
160-166	17	-1,55<z<-0,70	0,2420	0,181	18	0,056
166-172	38	-0,70<z<0,14	0,5557	0,314	32	1,125
172-178	25	0,14<z<0,99	0,8389	0,283	28	0,321
178-184	9	0,99<z<1,83	0,9664	0,128	13	0,063
184-190	6	z>1,83	1	0,033	3	
zbroj	100			1	100	1,732

Dakle, imamo pet razreda i tri ograničenja: $N = 100$

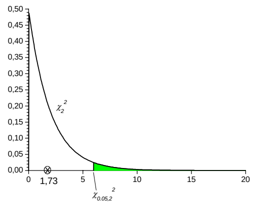
$\mu = \bar{x}_{uzorka}$

$\sigma = \sigma_{uzorka}$

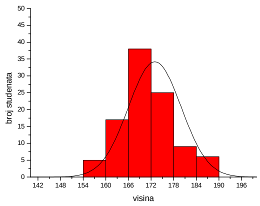
pa je ν = 5-3 = 2

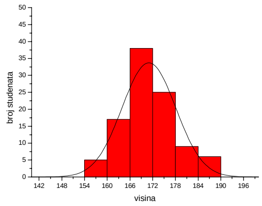
Za 2 stupnja slobode i 5% signifikantnosti, kritična vrijednost je

$$\chi^2_{0.05,2} = 5{,}99$$



Izračunata vrijednost $\chi^2_{op} = 1{,}73$ ne pada u kritično područje.
Stoga hipotezu H_0 **zadržavamo**.





ODBACUJEMO!

PRIHVAĆAMO!