Povzetek prejšnjega predavanja

Trditev oziroma domnevo o neznanem porazdelitvenem zakonu populacije imenujemo statistična hipoteza in jo označimo z $H()$.

Za odločanje o hipoteze H_0 oblikujemo pravilo

Imamo naključno spremenljivko X, naključni vzorec $\mathbf{V}$ in izbrano statistiko Z opredeljeno na vzorcu $\mathbf{V}$.

- Z W označimo zalogo vrednosti te statistike $Z(\mathbf{V}) \in W$.
- Izberemo neko podmnožico $S_c \subset W$, ki jo imenujemo kritično področje.

Glede na vrednost statistike $Z(\mathbf{V})$ opredelimo pravilo:

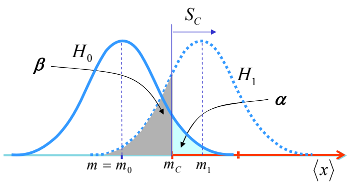
če je:

$Z(\mathbf{V}) \in S_c$ potem H_0 zavrnemo

če je:

$Z(\mathbf{V}) \in W - S_c$ potem H_0 sprejmemo

z upoštevanjem pravila hipotezo H_0 bodisi zavrnemo ali sprejmemo, zato opredeljeno pravilo odločanja imenujemo statistični test hipoteze H_0



Napaka I vrste:

$$\alpha = P\left(\langle X \rangle \in S_C \,|H_0 \text{ je pravilna}\right) = \int_{S_C} f_{\langle X \rangle | H_0(m=m_0)}\left(\langle x \rangle\right) \mathrm{d}\langle x \rangle$$

Napaka II vrste:

$$\beta = P\left(\langle X \rangle \notin S_C \,|H_0 \text{ je napačna}\right)$$
$$= P\left(\langle X \rangle \notin S_C \,|H_1(m = m_1 > m_0)\right) = \int_{\notin S_C} f_{\langle X \rangle | H_1(m=m_1)}\left(\langle x \rangle\right) \mathrm{d}\langle x \rangle$$

Da bi se pri odločanju izognili možnosti napake druge vrste pravilo preoblikujemo v naslednje:

Glede na vrednost statistike $Z(\mathbf{V})$ opredelimo pravilo:

če je:

$Z(\mathbf{V}) \in S_c$ potem H_0 zavrnemo

če je:

$Z(\mathbf{V}) \in W - S_c$ potem se o H_0 ne odločimo

Tako preoblikovano pravilo oziroma test hipoteze H_0 imenujemo *preskus značilnosti*.

Povzetek vpeljanih karakteristik pri testiranju hipotez:

α verjetnost, da pravilno hipotezo H_0 zavrnemo in s tem storimo napako I vrste, imenujemo stopnja značilnosti.

$1-\alpha$ verjetnost, da pravilno hipotezo H_0 ne zavrnemo.

β verjetnost, da napačno hipotezo H_0 sprejmemo in s tem storimo napako II vrste.

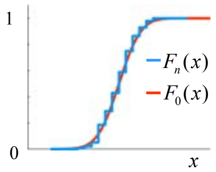
$1-\beta$ verjetnost, da napačno hipotezo H_0 zavrnemo imenujemo jakost testa $1-\beta$

Splošni postopek pri testiranji parametričnih hipotez

1. Iz vsebine problema določimo parameter obravnave.
2. Opredelimo ničelno hipotezo H_0.
3. Opredelimo ustrezno alternativno hipotezo H_1.
4. Izberemo stopnjo značilnosti α .
5. Opredelimo testno statistiko
6. Opredelimo področje zavračanja S_C .
7. Iz vzorca izračunamo potrebne količine, jih vstavimo v statistiko Z, in izračunamo vrednost statistike.
8. Na osnovi vrednosti statistike H_0 zavrnemo ali pa ne in odločitev utemeljimo.

11.2.1 Prilagoditveni neparametrični test χ^2

Domnevamo, da vzorčno porazdelitveno funkcijo $F_n(x)$ lahko opišemo s teoretično porazdelitvijo $F_0(x)$.



Testiramo hipotezo:

$$H_0(F_n = F_0)$$

Za primerjavo porazdelitev pogostosti n_i in hipotetičnih pogostosti np_{i0} opredelimo statistiko:

$$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - n_{i_0})^2}{n_{i_0}} = \sum_{i=1}^{r} \frac{(n_i - np_{i_0})^2}{np_{i_0}}$$

$\chi^2(n \to \infty) \Rightarrow$ hi-kvadrat porazdelitev s :

$$k=r-l-1$$

prostostnimi stopnjami.

S poznavanjem porazdelitve statistike pri dani stopnji značilnosti α določimo χ^2_C :

$$P(\chi^2 \geq \chi^2_C = \chi^2_{k,\alpha}) = \alpha$$

Če je iz vzorca izračunana statistika:

$$\chi^2 > \chi^2_C$$

F_0 izrazito odstopa od empirične F_n in H_0 zavrnemo.

V primeru, ko je $\chi^2 < \chi^2_C$, H_0 ne zavrnemo.

11.2.2 Test Kolmogorova

Za hitro oceno ujemanja med predpostavljeno teoretično F_0 in na osnovi vzorca $\boldsymbol{v}$ razsežnosti n izmerjeno empirično porazdelitvijo F_n verjetnosti populacije X .

Vpeljemo statistiko:

$$\Lambda_n = D_n\sqrt{n}\ ,$$

kjer je:

$$D_n = \max_{-\infty<x<\infty} |F_n(x) - F_0(x)|$$

Statistiki:

$$\Lambda_n = D_n\sqrt{n}\ ,$$

ustreza asimptotična porazdelitev verjetnosti Kolomogorov-Smirnova:

$$\lim_{n\to\infty} = P(\Lambda_n < \lambda) = Q(\lambda) = \sum_{k=-\infty}^{\infty} (-1)^k\, e^{-2k^2\lambda^2}$$

Pri tem n predstavlja velikost vzorca $\boldsymbol{v}$. Porazdelitev K-S je podana v tabeli.

Postopek uporabe testa Kolmogorova:

1) Imamo vzorec $\boldsymbol{v}$ velikosti n iz katerega tvorimo empirično porazdelitev F_n.
2) Na osnovi poteka F_n predpostavimo, da porazdelitev populacije X lahko opišemo z F_0 .
3) Izračunamo maksimalno razliko:

$$D_n = \max_{-\infty<x<\infty} |F_n(x) - F_0(x)|$$

4. Za podano stopnjo značilnosti oziroma zavračanja testa α iz tabele določimo kritično vrednost λ_c .
5. Če je izračunana statistika:

$$\Lambda_n = D_n \sqrt{n} > \lambda_c$$

hipotezo o ujemanju porazdelitev F_n in F_0 zavrnemo.

V primeru ko:

$$\Lambda_n = D_n \sqrt{n} < \lambda_c$$

pa nimamo razloga za zavrnitev hipoteze.

Test K-S lahko uporabimo tudi za oceno ali vzorca $\boldsymbol{v}_1$ in $\boldsymbol{v}_2$ razsežnosti n_1 in n_2 pripadata isti poulaciji X.

In uporabimo modificirano statistiko Λ_n:

$$\Lambda_{n_1,n_2} = D_{n_1,n_2} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}, \quad \text{za } n_1 \neq n_2$$

kjer je

$$D_{n_1,n_2} = \max_{-\infty < x < \infty} \left| F_{n_1}(x) - F_{n_2}(x) \right|$$

vzorčnih porazdelitev F_{n1} in F_{n2} .

11.2 Preverjanje neparametričnih hipotez

Primer: Imamo dva stroja, S_1 in S_2 , ki izdelujeta enake izdelke dolžine X_1 in X_2 Na voljo imamo dva vzorca $\boldsymbol{v}_1$ in $\boldsymbol{v}_2$ velikosti n , po enega za vsak stroj. Iz vzorcev smo določili porazdelitvi pogostosti $F_{1,n}$ in $F_{2,n}$.

x	9	10	11	12	13	14	15	16
$n_{1,i}$	0	2	12	31	52	72	88	100
$n_{2,i}$	0	1	11	29	49	73	87	100

Zanima nas ali lahko pri stopnji zavračanja α=0.05 zavrnemo hipotezi:

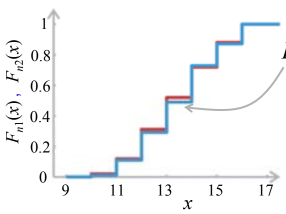
I. Obe porazdelitvi F_{n1} in F_{n2} sta enaki:

$$H_0(F_{n1} = F_{n2})$$

II. Porazdelitev $F_{n1} = F_{n2}$ je normalna:

$$H_0(F_{n1} = F_{n2} = \mathcal{N})$$

I.) Grafa porazdelitev F_{n1} in F_{n2} :



$$D_{n_1,n_2} = \left| F_{n,1}(13) - F_{n,2}(13) \right| = 0.03$$

$$\lambda_n = D_n \sqrt{n/2} = 0.21$$

Iz tabele porazdelitve K-S za α=0.05 preberemo kritično vrednost λ_α statistike Λ_{n_1,n_2} :

$$\lambda_\alpha = \lambda_{0.05} = 1.358$$

Interval zavračanja:

$$S_C = (1.358, \infty)$$

Ker:

$$\lambda_n = 0.21 \notin S_C = (1.358, \infty)$$

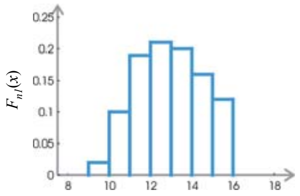
Hipoteze $H_0(F_{n1} = F_{n2})$ o enakosti obeh porazdelitev ***ne zavrnemo***.

II.) Za testiranje druge hipoteze $H_0(F_{n1} = F_{n2} = \mathcal{N})$ je potrebno oceniti srednjo vrednost in varianco populacije X_1 ali X_2 .

V ta namen iz določimo porazdelitev relativnih frekvenc:

$$f_{1i} = \Delta n_{1i} / n$$

x	9.5	10.5	11.5	12.5	13.5	14.5	15.5
$f_{1i}=\Delta n_{1i}/n$	0.02	0.10	0.19	0.21	0.20	0.16	0.12



Iz vzorčne porazdelitve verjetnosti ocenimo srednjo vrednost in varianco:

$$\langle x \rangle = \sum_{i=1}^{8} x_i f_i = 12.94$$

$$S^2 = \frac{n}{n-1} \sum_{i=1}^{n} \left(x_i - \langle x_i \rangle\right)^2 f_i = 2.53 \Rightarrow S = 1.59$$

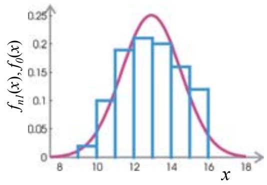
S pomočjo vzorčne srednje vrednosti in variance opredelimo standardni odmik:

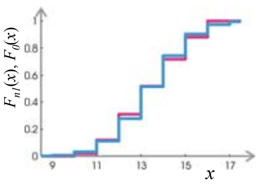
$$z = \frac{x - \langle x \rangle}{S}$$

In hipotetitično normalno porazdelitev:

$$F_0(z) = 0.5 + \Phi(z)$$

Grafa empirične in hipotetične porazdelitve gostote:





Izračunamo standardizirane vrednosti F_{n1}(z) , F_0(z)= $\mathcal{N}$ in razlike |F_{n1}(z) - F_0(z) |:

x_i	9.5	10.5	11.5	12.5	13.5	14.5	15.5
z_i	-1.85	-1.22	-0.59	-0.04	0.67	1.30	1.93
$F_0(z)$	0.03	0.11	0.28	0.52	0.75	0.90	0.97
$F_{n1}(z)$	0.02	0.12	0.31	0.52	0.72	0.88	1.00
$\lvert F_0(z) - F_{n1}(z) \rvert$	0.01	0.01	0.03	0.00	0.03	0.02	0.03

Vrednost K-S statistike je:

$$\lambda_n = D_n \sqrt{n/2} = 0.03\sqrt{100} = 0.3$$

Ker:

$$\lambda_n = 0.3 \notin S_C = (1.358, \infty)$$

Hipoteze $H_0(F_{n1} = F_{n2} = \mathcal{N})$, da je F_{n1} normalna pri izbrani stopnji zavračanja α=0.05 ***ne zavrnemo***.

11.2.3 Preverjanje neodvisnih vplivov

Pogostokrat lahko n elementov vzorca $\mathbf{v}$ razdelimo glede na dva različna kriterija. Glede na kriterij dobimo dve podmnožici $\mathbf{v}_x$ in $\mathbf{v}_y$ vzorca $\mathbf{v}$. Zanima nas ali sta dobljeni podmnožici $\mathbf{v}_x$ in $\mathbf{v}_y$ medsebojno statistično neodvisni.

Primer: Obravnavamo vzorec $\mathbf{v}$, n univerziteno zobraženih ljudi. Vzorec razdelimo glede na vrsto izobrazbe $\mathbf{v}_x$ in glede na velikost začetne plače $\mathbf{v}_y$. Zanima nas ali je začetna plača neodvisna od vrste univerzitetne izobrazbe.

Obravnavajmo dve naključni spremenljivki X in Y

Opredelimo r razredov $(x_1, x_2, \ldots, x_r)$ spremenljivke X, in c razredov $(y_1, y_2, \ldots, y_c)$ spremenljivke Y.

Na vzorcu razsežnosti n določimo pogostosti n_{ij} dogodkov

$$(X, Y) = (x_i, y_j)\ .$$

Vzorčne pogostosti n_{ij} predstavimo s kontingenčno tabelo.

Primer kontingenčne tabele:

		Y				n_{i*}
		y_1	y_2	$\ldots$	y_c	
X	x_1	n_{11}	n_{12}	$\ldots$	n_{1c}	n_{1*}
	x_2	n_{11}	n_{12}	$\ldots$	n_{1c}	n_{2*}
	$\vdots$					$\vdots$
	x_r	n_{r1}	n_{r2}	$\ldots$	n_{rc}	n_{r*}
	n_{*j}	n_{*1}	n_{*2}	$\ldots$	n_{*c}	n

$$n = \sum_{j=1}^{c} n_{i*} = \sum_{i=1}^{k} n_{*j}$$

Kjer so: $n_{i*} = \sum_{j=1}^{c} n_{ij}$, in $n_{*j} = \sum_{i=1}^{r} n_{ij}$ robne pogostosti.

Opredelitev ničelne hipoteze o pojavu

Postavimo domnevo H_0, da sta spremenljivki X in Y statistično neodvisni. In alternativno hipotezo H_1 s katero trdimo, da sta X in Y statistično odvisni.

Za neodvisni spremenljivki velja:

$$p_{ij} = P((X = x_i) \cap (Y = y_i))$$
$$= p_i p_j = P(X = x_i) P(Y = y_i)$$

Hipotezi H_0 in H_1 lahko zapišemo v obliki:

$$H_0(p_{ij} = p_i p_j), \qquad H_1(p_{ij} \neq p_i p_j)$$

Opredelitev testne statistike

Pri prilagoditvenem testu χ^2 smo primerjali vzorčno in predpostavljeno teoretično porazdelitev na osnovi razlike med vzorčno pogostostijo n_i in izračunano teoretično n_{i0} :

$$(n_i - n_{i0})^2 \quad \text{oziroma} \quad (n_i - np_{i0})^2$$

Podobno pri preverjanju neodvisnosti spremnljivk X in Y primerjamo vzorčno pogostost n_{ij} s predpostavljeno n_{ij0}

$$(n_{ij} - n_{ij0})^2 \quad \text{oziroma} \quad (n_{ij} - np_{ij0})^2$$

Za neodvisni spremenljivki X in Y velja:

$$p_{ij_0} = p_{i_0} p_{j_0}$$

Za p_{i_0} in p_{j_0} lahko podamo ocene:

$$\hat{p}_{i_0} = \frac{n_{i*}}{n} \quad \text{in} \quad \hat{p}_{j_0} = \frac{n_{*j}}{n}$$

In sledi:

$$p_{ij_0} = p_{i_0} p_{j_0} = \frac{n_{i*}}{n} \frac{n_{*j}}{n} = \frac{n_{i*} n_{*j}}{n^2}$$

oziroma:

$$n_{ij_0} = np_{ij_0} = \frac{n_{i*} n_{*j}}{n^2} = \frac{n_{i*} n_{*j}}{n}$$

Z upoštevanjem izrazov za n_{ij0} in p_{ij0} opredelimo statistiko:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - n_{ij_0})^2}{n_{ij_0}} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - np_{ij_0})^2}{np_{ij_0}}$$

Statistika χ^2 ima porazdelitev χ^2 z $k=(r-1)(c-1)$ prostostnimi stopnjami.

Za neodvisni spremenljivki X in Y je $\chi^2 \approx 0$.

Iz tabele za χ^2 porazdelitev, ki podaja vrednosti za verjetnost:

$$P(\chi^2 \geq \chi^2_{(r-1)(c-1);\alpha}) = \alpha$$

za vrednosti stopnje značilnost α določimo interval zavračanja:

$$S_C = (\chi^2_{(r-1)(c-1);\alpha}, \infty)$$

Hipotezo $H_0(p_{ij}=p_i p_j)$ zavrnemo s stopnjo značilnosti α če:

$$\chi^2 \in S_C = (\chi^2_{(r-1)(c-1);\alpha}, \infty)$$

Primer: Na dve strojih y_1 in y_2 izdelujemo tri vrste izdelkov x_1 , x_2 in x_3 . Pri nadzoru smo dobili podatke o številu slabih izdelkov, ki so podani v tabeli:

		Stroj-*j*	
		y_1	y_2
Izdelek-*i*	x_1	18	16
	x_2	15	20
	x_3	16	15

Ali lahko s stopnjo značilnosti α=0.05 zavrnemo hipotezo H_0, da je kakovost izdelka neodvisna od stroja.

Iz dopolnjene kontingenčne tabele izračunamo hipotetičen pogostosti n_{ij0} :

$$n_{ij0} = n_{i*} \cdot n_{*j} / n$$

		Stroj-*j*		
		y_1	y_2	n_{i*}
Izdelek-*i*	x_1	18	16	34
	x_2	15	20	35
	x_3	16	15	31
	n_{*j}	49	51	n=100

		Stroj-*j*	
		y_1	y_2
Izdelek-*i*	x_1	16.66	17.43
	x_2	17.15	17.85
	x_3	15.19	15.81

Kjer je:

$$n_{i*} = \sum_{j=1}^{c} n_{ij}, \quad n_{*j} = \sum_{i=1}^{r} n_{ij} \quad \text{in} \quad n = \sum_{j=1}^{c} n_{i*} = \sum_{i=1}^{k} n_{*j}$$

S podatki iz obeh tabel izračunamo vrednost statistike:

$$\chi^2 = \sum_{i=1}^{r=3}\sum_{j=1}^{c=2} \frac{(n_{ij} - n_{ij_0})^2}{n_{ij_0}} = \frac{(18-16.66)^2}{16.66} + \cdots + \frac{(15-15.81)^2}{15.81}$$
$$= 0.82$$

ki ima k=(r-1)(c-1)= (3-1)(2-1)=2 prostostnih stopenj.

Iz tabele za χ^2 porazdelitev preberemo vrednost:

$$\chi^2_{(r-1)(c-1);\alpha} = \chi^2_{2;0.05} = 5.991$$

In določimo interval zavračanja:

$$S_C = (\chi^2_{(r-1)(c-1);\alpha}, \infty)$$

In določimo interval zavračanja:

$$S_C = (\chi^2_{(r-1)(c-1);\alpha}, \infty) = (5.991, \infty)$$

Ker izračunana vrednost statistike χ^2 =0.82 ne leži v intervalu zavračanja,

$$\chi^2 = 0.82 \notin S_C = (5.991, \infty)$$

hipoteze H_0, da je kakovost izdelka neodvisna od stroja pri α=0.05 ***ne zavrnemo***.

To nadalje pomeni, da je glede kakovosti vseeno na katerem stroju izdelujemo posamezni izdelek.

11.2.4 Preverjanje homogenosti skupin

Obravnavajmo vzorce v *skupin* X_i, i=1,...,v . Elemente vzorcev v razdelimo glede na izbrani *kriterij* v r *razredov* y_j, j=1,...,r.

Primer : *skupine* x_i, i=1,...,v predstavljajo generacije študentov v letih 99, 00, ... ,03
Kot *kriterij* izberemo uspešnost pri izbranem predmetu.
r *razredov* y_j, j=1,...,r predstavlja r stopenj uspešnosti: zelo uspešen, manj uspešen ... . neuspešen

Skupine x_i (generacije) so homogene, če imajo v poljubnem razredu (uspešnosti: zelo uspešen, uspešen, manj uspešen) y_j vse enako relativno frekvenco:

$$p_{1,j} = p_{2,j} = \cdots = p_{v,j} \qquad \text{za } j = 1, \ldots, r$$

Opis skupin po posameznih razredih kriterija podamo s kontingenčno tabelo.

Primer tabele:

		Kriterij				
		y_1	y_2	$\ldots$	y_r	n_{i*}
Skupina	x_1	n_{11}	n_{12}	$\ldots$	n_{1r}	$n_{1*}=n_1$
	x_2	n_{11}	n_{12}	$\ldots$	n_{1r}	$n_{2*}= n_2$
	$\ddots$					$\ddots$
	$\ddots$					$\ddots$
	$\ddots$					$\ddots$
	x_v	n_{v1}	n_{v2}	$\ldots$	n_{vr}	$n_{r*}= n_v$
	n_{*j}	n_{*1}	n_{*2}	$\ldots$	n_{*r}	n

Pri tem je:

n_{ij} število elementov vzorca skupine i, ki izpolnjuje pogoje razreda j izbranega kriterija.

x_i je število elementov n_{i*} posmezne skupine, ki je določeno vnaprej z velikostjo n_i vzorca v_i skupine x_i i=1,...,v .

Med tem, ko je skupno število elementov n_{*j} v posameznem razredu y_j naključno odvisno od vzorcev v_i .

Opredelitev ničelne hipoteze

Z ničelno hipotezo H_0, postavimo domnevo, da so skupine x_i i=1,...,v homogene glede na r razredov y_j izbranega kriterija.

Skupine x_i so homogene, če imajo v poljubnem razredu y_j vse enako relativno frekvenco:

$$p_{1,j} = p_{2,j} = \cdots = p_{v,j} \qquad \text{za } j = 1, \ldots, r$$

Hipotezo H_0 zapišemo v obliki:

$$H_0(p_{1,j} = p_{2,j} = \cdots = p_{v,j}) \qquad \text{za vsak } j$$

Kot alternarnativno hipotezo H_1 opredelimo domnevo, da skupine x_i niso homogene.

$$H_1(p_{i,j} \neq p_{k,j}) \quad \text{za vsaj eno trojico } i, j, k$$

Opredelitev testne statistike

Podobno kot pri preverjanju neodvisnosti spremnljivk X in Y primerjamo vzorčno pogostost n_{ij} z predpostavjeno n_{ij0}

$$(n_{ij} - n_{ij0})^2 \quad \text{oziroma} \quad (n_{ij} - np_{ij0})^2$$

Kjer n_{ij0} predstavlja pričakovanega število elementov in p_{ij0} pričakovano verjetnost j-tega razreda i-te skupine.

Pričakovano verjetnost p_{*j0} za razred kriterija y_j ocenimo z:

$$p_{*j_0} = \frac{n_{*j}}{n}$$

Kjer je n število elementov vseh vzorcev oziroma skupin :

$$n = \sum_{i=1}^{v} n_i$$

S pomočjo ocene p_{*j0} in podanega števil elementov n_{j*} celotne skupine je ocena pričakovanega število n_{ij0} elementov j-tega razreda i-te skupine podana z:

$$n_{ij_0} = p_{*j_0} n_{*i}$$

Pričakovana verjetnost

$$p_{ij_0} = \frac{n_{ij_0}}{n} = \frac{n_{i*}p_{*j}}{n} = \frac{n_{i*}n_{*j}}{n \cdot n} = \frac{n_{i*}n_{*j}}{n^2}$$

in pričakovano število n_{ij0} :

$$n_{ij_0} = np_{ij_0} = n\frac{n_{i*}n_{*j}}{n^2} = \frac{n_{i*}n_{*j}}{n}$$

Z upoštevanjem omenjenih povezav za preizkus homogenosti uporabimo statistiko:

$$\chi^2 = \sum_{i=1}^{v}\sum_{j=1}^{k}\frac{\left(n_{ij} - n_{ij_0}\right)^2}{n_{ij_0}} = \sum_{i=1}^{v}\sum_{j=1}^{k}\frac{\left(n_{ij} - np_{ij_0}\right)^2}{np_{ij_0}}$$

Statistika χ^2 ima porazdelitev χ^2 z $(v\text{-}1)(k\text{-}1)$ prostostnimi stopnjami.

Statistika ima v primeru homogenosti skupin x_i $i=1,\ldots,v$ vrednost $\chi^2 \approx 0$.

Za vrednost stopnje značilnost α iz tabele za χ^2, ki podaja vrednosti za verjetnost:

$$P(\chi^2 \geq \chi^2_{(v-1)(k-1);\alpha}) = \alpha$$

določimo interval zavračanja:

$$S_C = (\chi^2_{(r-1)(c-1);\alpha}, \infty)$$

Hipotezo

$$H_0(p_{1,j} = p_{2,j} = \cdots = p_{v,j}) \quad \text{za vsak } j$$

zavrnemo s stopnjo značilnosti α če statistika:

$$\chi^2 \in S_C = (\chi^2_{(r-1)(c-1);\alpha}, \infty)$$

Primer: Podatki o uspešnosti treh generacij študentov x_i pri izpitu izbranega predmeta so podani v tabeli:

		Uspešnost			
		ZU	U	MU	NU
Gener.	x_1	24	38	14	10
	x_2	28	45	20	3
	x_3	16	40	10	15

Za stopnjo značilnosti $\alpha=0.05$ preverimo hipotezo H_0, da so vse tri generacije študentov enako uspešne pri, opravljanju izpita.

Iz dopolnjene kontingenčne tabele v kateri je:

$$n_{i*} = \sum_{j=1}^{c} n_{ij}, \quad n_{*j} = \sum_{i=1}^{r} n_{ij} \quad \text{in} \quad n = \sum_{j=1}^{c} n_{i*} = \sum_{i=1}^{k} n_{*j}$$

		Uspešnost				
		ZU	U	MU	NU	n_{i*}
Gener.	x_1	24	38	14	10	86
	x_2	28	45	20	3	96
	x_3	16	40	10	15	81
	n_{*j}	68	123	44	28	n=263

Izračunamo hipotetične pogostosti n_{ij0} :

$$n_{ij0} = n_{i*} \cdot n_{*j} / n$$

		Uspešnost				
		ZU	U	MU	NU	n_{i*}
Gener.	x_1	22.24	40.22	14.39	9.16	86
	x_2	24.82	44.90	16.06	10.22	96
	x_3	20.94	37.88	13.55	8.62	81
	n_{*j}	68	123	44	28	n=263

S podatki iz obeh tabel izračunamo vrednost statistike:

$$\chi^2 = \sum_{i=1}^{v=3}\sum_{j=1}^{c=4}\frac{\left(n_{ij}-n_{ij_0}\right)^2}{n_{ij_0}} = \frac{(24-22.24)^2}{22.24}+\cdots+\frac{(15-8.62)^2}{8.62}$$
$$=13.76$$

ki ima $k=(v\text{-}1)(c\text{-}1)= (3\text{-}1)(4\text{-}1)=6$ prostostnih stopenj.

Iz tabele za χ^2 porazdelitev preberemo vrednost:

$$\chi^2_{(v-1)(c-1);\alpha} = \chi^2_{6;0.05} = 12.592$$

In določimo interval zavračanja:

$$S_C = (\chi^2_{6;0.05}, \infty) = (12.592, \infty)$$

Ker izračunana vrednost statistike $\chi^2 = 13.76$ leži v intervalu zavračanja,

$$\chi^2 = 13.76 \in S_C = (12.59, \infty)$$

hipoteze H_0, o homogenosti skupin x_i glede na razrede y_j kriterija pri $\alpha = 0.05$ ***zavrnemo***.

Oziroma obravnavane generacije študentov ***niso homogene*** glede na uspešnost pri izbranem izpitu.

11.2.5 Test enakosti dveh normalnih pojavov

V praksi se pogosto srečamo s problemom ocene enakosti dveh naključnih pojavov.

Primer : Enakost lastnosti X izdelkov, ki jih izdelujemo na dveh različnih strojih.

Če poznamo vzorčne porazdelitev lastnosti X_1 izdelkov na prvem in X_2 na drugem stroju lahko v ta namen uporabimo bodisi:

- hi-kvadrat prilagoditveni test ali
- test K-S

Najpogosteje porazdelitvenih funkcij $F_{n1}(X_1)$ in $F_{n2}(X_2)$ ne poznamo.

Zadovoljimo se že, če poznamo vzorčni povprečji $\langle X_1 \rangle$ in $\langle X_2 \rangle$ ter pripadajoči vzorčni varianci S_1^2 in S_2^2 .

V takem primeru hipotezo o enakosti obeh porazdelitev prevedemo na hipotezo o enakosti srednjih vrednosti in enakosti standardnih deviacij:

$$H\left(F_1(X_1) = F_2(X_2)\right) \Rightarrow H\left(m_1 = m_2; \sigma_1 = \sigma_1\right)$$

Ker sta m in σ statistično neodvisna parametra je hipoteza:

$$H\left(m_1 = m_2; \sigma_1 = \sigma_1\right)$$

pravilna če sta hkrati pravilni hipotezi:

$$H\left(m_1 = m_2\right) \quad \text{in} \quad H\left(\sigma_1 = \sigma_1\right)$$

Zato lahko testiranje hipotezi obravnavamo ločeno.

Preverjanje hipoteze H_0 o enakosti srednjih vrednosti dveh normalnih populacij:

$$H_0\left(m_1 = m_2\right)$$
$$H_1\left(m_1 \neq m_2\right)$$

1. Poznamo varianci σ_1 in σ_2 populacij X_1 in X_2 s sredjima vrednostima m_1 in m_2 :

2. Ne poznamo varianci σ_1 in σ_2 populacij X_1 in X_2 s sredjima vrednostima m_1 in m_2 :

1. Poznamo varianci σ_1 in σ_2 populacij X_1 in X_2 s sredjima vrednostima m_1 in m_2 :

Na razpolago imamo neodvisna naključna vzorca $\mathbf{V}_{X1}$ in $\mathbf{V}_{X2}$ velikosti n_1 in n_2 populacij X_1 in X_2 .

$\Rightarrow$

Porazdelitev razlike vzorčnih povprečij $\langle X_1\rangle_{n1} - \langle X_2\rangle_{n2}$ je normalna s srednjo vrednostjo $(m_1 - m_2)$ in varianco $(\sigma_1^2/n_1 + \sigma_2^2/n_2)$:

$$\langle X_1\rangle_{n1} - \langle X_1\rangle_{n1} \approx N\left(m_1 - m_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

Zato lahko za testiranje hipoteze $H_0(m_1 = m_2)$ kot statistiko uporabimo standardni odklon:

$$Z = \frac{\langle X_1\rangle_{n1} - \langle X_2\rangle_{n2}}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

Oziroma v primeru, ko je $\sigma_1^2 = \sigma_2^2 = \sigma^2$:

$$Z = \frac{\langle X_1\rangle_{n1} - \langle X_2\rangle_{n2}}{\sigma\sqrt{1/n_1 + 1/n_2}}$$

Statistika Z ima normalno porazdelitev in je njena vrednost v primeru ko $(m_1 \approx m_2)$ enaka $z \approx 0$.

2. Ne poznamo varianci σ_1^2 in σ_2^2 populacij X_1 in X_2 s sredjima vrednostima m_1 in m_2 :

V tem primeru σ_1^2 in σ_2^2 ocenimo iz vzorcev $\mathbf{V}_{X1}$ in $\mathbf{V}_{X2}$ z vzorčnima variancama S_1^2 in S_2^2 :

$$S_j^2 = \frac{1}{n-1}\sum_{i=1}^{n_j}\left(x_{ji} - \langle X_j\rangle_{n_j}\right)^2 \qquad j = 1,2$$

Kot statistiko uporabimo:

$$T^* = \frac{\langle X_1\rangle_{n1} - \langle X_2\rangle_{n2}}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$$

Porazdelitev statistika T^* lahko aproksimiramo s Studentovo ali t-porazdelitvijo, s številom prostostnih stopnej k ocenimo z:

$$k = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(S_1^2/n_1\right)^2}{n_1+1} + \frac{\left(S_2^2/n_2\right)^2}{n_2+1}} - 2$$

V primeru ko je $\sigma_1^2 = \sigma_2^2$, σ^2 ocenimo s pomočjo utežene ocene S_P^2 na osnovi vzorčnih varianc S_1^2 in S_2^2 z izrazom:

$$S_P^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1-1)+(n_2-1)}$$

Če v enačbo vstavimo izraza za S_1^2 in S_2^2 dobimo:

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1-1)+(n_2-1)}$$
$$= \frac{\sum_{i=1}^{n_1}(x_{1i} - \langle X_1\rangle_{n_1} + \sum_{i=1}^{n_2}(x_{2i} - \langle X_2\rangle_{n_2}}{n_1 + n_2 - 2}$$

Z uporabo sestavljene ocene S_P^2 za varianco σ^2 kot statistiko za testiranje hipoteze $H_0(m_1 = m_2)$ uporabimo:

$$T = \frac{\langle X_1\rangle_{n1} - \langle X_2\rangle_{n2}}{S_p\sqrt{1/n_1 + 1/n_2}}$$

V tem primeru je porazdelitev statistike T enaka Studentovi ali t-porazdelitvi z $n_1 + n_2 - 2$ prostostnimi stopnjami.

Preverjanje hipoteze o enakosti standardnih deviacij oziroma varianc dveh normalnih populacij:

$$H_0(\sigma_1 = \sigma_2) = H_0(\sigma_1^2 = \sigma_2^2)$$
$$H_1(\sigma_1 \neq \sigma_2) = H_1(\sigma_1^2 \neq \sigma_2^2)$$

Na razpolago imamo neodvisna naključna vzorca $\boldsymbol{V}_{X1}$ in $\boldsymbol{V}_{X2}$ velikosti n_1 in n_2 normalnih populacij X_1 in X_2.

Kot cenilko varianc uporabimo vzorčni varianci S_1^2 in S_2^2.

$$S_j^{\,2} = \frac{1}{n-1}\sum_{i=1}^{n_j}\left(x_{ji} - \left\langle X_j \right\rangle_{n_j}\right)^2 \qquad j = 1,2$$

Kot statistiko za testiranja hipoteze

$$H_0(\sigma_1^2 = \sigma_2^2)$$

uporabimo razmerje F vzorčnih varianc:

$$F = \frac{S_1^2}{S_2^2}$$

Vpeljana statistika *F* ima tako imenovano *Snedecorjevo* ali *F porazdelitev* z devema prostostnima stopnjama:

$u = n_1 - 1$ in $v = n_2 - 1$

Kjer sta velikosti n_1 in n_2 velikosti vzorcev $\boldsymbol{V}_{X1}$ in $\boldsymbol{V}_{X2}$

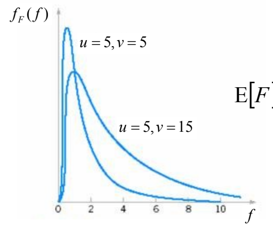
Snedecorjevo ali *F porazdelitev* je opredeljena z gostoto:

$$f_F(f) = \frac{\Gamma\left(\frac{u+v}{2}\right)\left(\frac{u}{v}\right)^{u/2} f^{(u/2)-1}}{\Gamma\left(\frac{u}{2}\right)\Gamma\left(\frac{v}{2}\right)\left[\left(\frac{u}{v}\right)f + 1\right]^{(u+v)/2}}$$

Kjer je Γ(r) gama funkcija:

$$\Gamma(r) = \int_0^\infty x^{r-1}\,\mathrm{e}^{-x}\,\mathrm{d}x, \quad \text{za } r > 0$$

Primeri grafov porazdelitev F za različne n_1 in n_2:

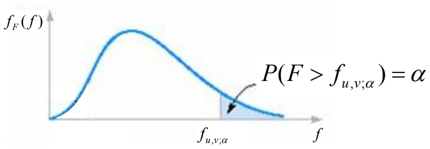


$$\mathrm{E}[F] = \frac{v}{v-2}, \quad \text{za } v > 2$$

Varianca je opredeljena z :

$$\mathrm{Var}(F) = \frac{2v^2(u+v-2)}{u(v-2)^2(v-4)}, \quad \text{za } v > 4$$

Porazdelitev F je tabelirana za različne stopnje značilnosti α in podaja verjetnosti P :



Iz vrednosti f statistike F, ki podaja razmerje vzorčnih varianc ugotovimo ali je :

$$f > f_{u,v;\alpha}$$

Kar je razlog za zavrnitev hipoteze H_0, o enakostih

Za $S_1^2 > S_2^2$ je $f > 1$ in je interval zavračanja opredeljen z:

$$S_C = (f_{u,v;\alpha}, \infty)$$

Ker je F tabelirana samo ta $f > 1$ v primeru, ko je $S_1^2 < S_2^2$ vpeljemo statistiko:

$$F = \frac{S_2^2}{S_1^2}$$

Pri kateri je zopet $f > 1$.

Primer : Imamo podatke o merjenju mas izdelkov iz različnih tovarn.

i	1	2	3	4	5	6	7	8	9	10
$x_{1,i}$	102	99	101	98	103	104	101	98	100	102
$x_{2,i}$	101	100	99	100	104	97	102	100	101	98

Ali lahko s tveganjem α=0.05 zavrnemo hipotezo o enakosti obeh serij izdelkov, če vemo, da je porazdelitev normalna.

Testirati je potrebno ničelno hipotezo:

$$H_0(m_1 = m_2, \sigma_1 = \sigma_2)$$

Vzorčni povprečji in varianci obeh populacij ocenimo iz podanih vzorcev:

Vzorčni povprečji sta:

$$\langle x_1 \rangle = \frac{1}{10}\sum_{i=1}^{10} x_{1i} = 100.8$$

$$\langle x_2 \rangle = \frac{1}{10}\sum_{i=1}^{10} x_{2i} = 100.2$$

Vzorčni varianci :

$$S_1^2 = \frac{1}{10-1}\sum_{i=1}^{10}\left(x_{1i} - \langle x_1 \rangle\right)^2 = 4.18$$

$$S_2^2 = \frac{1}{10-1}\sum_{i=1}^{10}\left(x_{2i} - \langle x_2 \rangle\right)^2 = 3.96$$

Preverimo najprej hipotezo o enakostih varianc.

$$H_0(\sigma_1 = \sigma_2)$$

V ta namen uporabimo statistiko:

$$F = \frac{S_1^2}{S_2^2} = \frac{4.18}{3.96} = 1.06,$$

ki ima *F* ali Snedecorjevo porazdelitev z u=9 in $v = 9$ prostostnimi stopnjami.

Pri stopnji zaupanja α=0.05 določimo enostranski interval zavračanja:

$$S_C = (f_{u,v;\alpha}, \infty) = (f_{9,9;0.05}, \infty) = (3{,}179, \infty)$$

Ker:

$$f = 1.06 \notin S_C = (3{,}179, \infty)$$

ničelne hipoteze o enakosti varianc ne zavrnemo.

Preverimo še hipotezo o enakostih srednjih vrednosti:

$$H_0(m_1 = m_2)$$

Zaradi enakosti varianc uporabimo statistiko:

$$T = \frac{\langle X_1 \rangle_{n1} - \langle X_2 \rangle_{n2}}{S_p\sqrt{1/n_1 + 1/n_2}}$$

Kjer je:

$$S_p^{\,2} = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1+n_2-2)} = \frac{1}{2}\left(S_1^2 + S_2^2\right) = 4.07$$

In izračunamo vrednost statistike:

$$t = \frac{\langle x_1 \rangle - \langle x_2 \rangle}{S_p\sqrt{1/n_1 + 1/n_2}} = 0.67,$$

ki ima Studentovo porazdelitev z n_1+n_2-2 =18 prostostnimi stopnjami.

Pri stopnji zaupanja α=0.05 določimo dvostranski interval zavračanja:

$$S_c = (-\infty, -t_{n;\alpha/2}) \quad \text{in} \quad (t_{n;\alpha/2}, \infty)$$

$$= (-\infty, -t_{18;0.025}) \quad \text{in} \quad (t_{18;0.025}, \infty)$$

$$= (-\infty, -2.1009) \quad \text{in} \quad (2.1009, \infty)$$

Ker vrednost statistike t=0.67 ne leži v intervalu zavračanja S_c

ničelne hipoteze o enakosti srednjih vrednosti ne zavrnemo.

Glede na to, da ne moremo zavrniti nobene od ničelnih hipotez:

$$H_0(m_1 = m_2),\ \ H_0(\sigma_1 = \sigma_2)$$

Tudi ne moremo zavrniti hipoteze o ujemanju obeh porazdelitev populacij

$$H_0(m_1 = m_2, \sigma_1 = \sigma_2)$$

oziroma enakosti izdelkov, ki prihajajo iz različnih tovarn.

11.2.6 Uporaba monograma

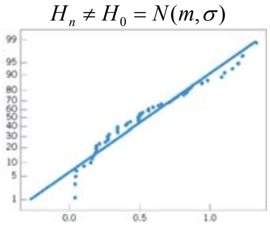
Najpogosteje skušamo iz oblike krivulje vzorčne porazdelitve verjetnosti sklepati na teoretični tip porazdelitve H_0.

Teoretično porazdelitev verjetnosti transformiramo tako, da jo grafično lahko predstavimo z premico.

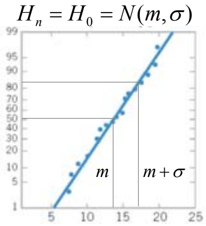
Čim manjša so odstopanja vzorčne porazdelitve H_n od premice, tem boljše je ujemanje le te s teoretično H_0.

Primera normalnega nomograma $H_0 = N(m, \sigma)$

$H_n \neq H_0 = N(m, \sigma)$



$H_n = H_0 = N(m, \sigma)$



$$m = F(50)$$

$$\sigma = F(84) - F(50)$$

Konstrukcija normalnega nomograma

Vrednosti meritev vzorca:

$$\mathbf{v} = (x_1, x_2, \cdots x_n)$$

razvrstimo od najmanjše do največje vrednosti

$$(x_{(1)}, x_{(2)}, \cdots x_{(n)}), \quad x_{i-1} < x_i$$

Vzorčna kumulativna porazdelitev vzorca v je podana z:

$$F_n(X < x_{(i)}) = \frac{i - 0.5}{n}$$

Primer : Podanih je 10 meritev življenjske dobe v minutah baterij prenosnih računalnikov.

$$\mathbf{v} = (176, 191, 214, 220, 205, 192, 201, 190, 183, 185)$$

Domnevamo, da je porazdelitev normalna.

$$F_n(X < x_{(i)}) = \frac{i - 0.5}{n} 100$$

i	1	2	3	4	5	6	7	8	9	10
x_i	176	183	185	190	191	192	201	205	214	220
$F_n(x_i)$	0.05	0.15	0.25	0.35	0.45	0.55	0.65	0.75	0.85	0.95

Graf normalnega nomograma: