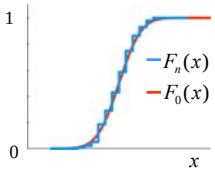
Testiranje hipotez

11.2.1 Prilagoditveni neparametrični test χ^2

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



$$H_0(F_n = F_0)$$

Uporabimo 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.

11.2.3 Preverjanje neodvisnih vplivov

Hipotezi H_0 in H_1 zapišemo v obliki:

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

Uporabimo statistiko:

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

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

Kontingenčna tabele za preverjanje neodvisnosti:

		Y				
		y_1	y_2	...	y_c	n_{i*}
X	x_1	n_{11}	n_{12}	...	n_{1c}	n_{1*}
	x_2	n_{21}	n_{22}	...	n_{1c}	n_{2*}
	⋱					⋱
	⋱					⋱
	⋱					⋱
	x_r	n_{r1}	n_{r2}	...	n_{rc}	n_{r*}
	n_{*j}	n_{*1}	n_{*2}	...	n_{*c}	n

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

11.2.4 Preverjanje homogenosti skupin

Hipotezi H_0 in H_1 zapišemo v obliki:

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

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

Uporabimo statistiko:

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

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

Kontingenčna tabele za preverjanje homogenosti:

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

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

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

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

11.2.5 Test enakosti dveh normalnih pojavov

Hipotezi H_0 in H_1 zapišemo v obliki:

$$H_0(m_1 = m_2) \quad H_1(m_1 \neq m_2) \qquad \text{in} \qquad H_0(\sigma_1 = \sigma_1) \quad H_1(\sigma_1 \neq \sigma_1)$$

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

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

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

Poglavje 12

Analiza variance

12.1 Osnovni pojmi

Pogosto imamo opraviti z analizo sistemov na katerih stanje in dinamiko hipotetično vplivajo različni parametri.

Obraba orodja: različna oplaščenje rezalne ploščice
Storilnost: različni delavci
Trdnost jekla: različni odstotki primesi mangana

Namen je ugotoviti značilnost vpliva izbranega parametra na povprečno vrednost stanja sistema.

Parametre, katerih vpliv analiziramo imenujemo *faktorji vpliva*.

Vrednosti posameznega parametra, ki jih želimo preveriti imenujemo *nivoji faktorja*.

Analiza vpliva enega faktorja

V primeru, dvonivojskega faktorja za analizo uporabimo že poznane metode.

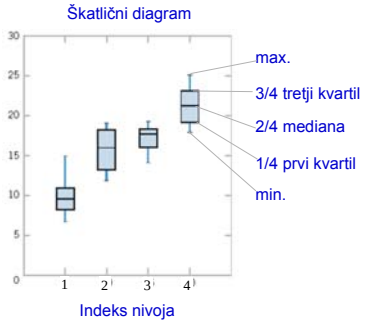
V primeru, več nivojskega faktorja pa uporabimo analizo variance.

Za izvedbo analize variance je potrebno predhodno opraviti ***naključni poskus***, ki se v splošnem sestoji iz naključne ***priprave vzorcev*** in ***izvedbe meritev***.

Izide meritev predstavimo v tabeli:

		Meritev
		1 . . . . j . . . . n_i
Nivoji faktorja	v_1	$x_{11} \ldots x_{1j} \ \ldots \ x_{1nl}$
	.	. . . . .
	v_i	$x_{i1} \ldots x_{ij} \ \ldots \ x_{1ni}$
	.	. . . . .
	v_r	$x_{r1} \ldots x_{rj} \ \ldots \ x_{nr}$

Grafični prikaz meritev naključnega poskusa:



12.1 Analiza variance

Analiza variance ali zasnova poskusa z enim faktorjem vpliva.

Predpostavimo da imamo r nivojski faktor vpliva. Izide x_{ij} naključnega poskusa predstavimo v tabeli:

		Meritev
		1 . . . . j . . . . n_i
Nivoji faktorja	v_1	$x_{11} \ldots x_{1j} \ \ldots \ x_{1nl}$
	.	. . . . .
	v_i	$x_{i1} \ldots x_{ij} \ \ldots \ x_{1ni}$
	.	. . . . .
	v_r	$x_{r1} \ldots x_{rj} \ \ldots \ x_{nr}$

Izidi poskusa x_{ij} nam predstavljajo realizacijo naključne spremenljivke X.

Vpliv faktorja opišemo z r vzorčnimi povprečji po posameznih nivojih faktorja:

$$\langle X_i \rangle = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij} = m_i, \quad i = 1, \ldots r$$

V splošnem je $m_i \neq m_l$

Zanima nas ali so te razlike posledica naključja ali vpliva faktorja.

Postavimo domnevo oziroma hipotezo, da faktor nima vpliva:

$$H_0 (m_1 = m_2 = \cdots m_r)$$

Poleg tega predpostavimo:

1) Naključna spremenljivka X je normalno porazdeljena s std σ.

2) Vzorčne variance glede na nivo faktorja so enake oziroma $\sigma_i = \sigma_l$.

Testiranje hipoteze H_0 ob predpostavkah 1 in 2 imenujemo *analiza variance*.

Ime analiza variance izvira iz dejstva, da je postopek analize zasnovan na razcepitvi odstopanj vzorčnih vrednosti od vzorčnega povprečja celotne skupine.

Vzorčno povprečje spremenljivke X po vsej skupini opredelimo z izrazom:

$$\langle X \rangle = \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{n_i} x_{ij} = \frac{1}{n} \sum_{i=1}^{r} n_i m_i = m$$

kjer je:

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

Odstopanje x_{ij} od m lahko zapišemo:

$$x_{ij} - m = x_{ij} - m_i + m_i - m = (x_{ij} - m_i) + (m_i - m)$$

S tem odstopanje $x_{ij} - m$ sestavimo iz:

$x_{ij} - m_i$ odstopnaja x_{ij} od vzorčnega povprečja m_i po i-tem nivoju faktorja in

$m_i - m$ odstopanja vzorčnega povprečja m_i po i-tem nivoju faktorja od celotnega povprečja m

Kot statistiko odstopanja celotne skupine izmerjenih vrednosti x_{ij} od srednje vrednosti m uporabimo vsoto kvadratov posameznih odstopnaj:

$$q = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (x_{ij} - m)^2 = \sum_{i=1}^{r} \sum_{j=1}^{n_i} \left[(x_{ij} - m_i) + (m_i - m) \right]^2$$

$$= \sum_{i=1}^{r} \sum_{j=1}^{n_i} \left[(x_{ij} - m_i)^2 + 2(x_{ij} - m_i)(m_i - m) + (m_i - m)^2 \right]$$

Drugi člen v tej vsoti je:

$$\sum_{i=1}^{r} \sum_{j=1}^{n_i} (x_{ij} - m_i)(m_i - m) = \sum_{i=1}^{r} (m_i - m) \sum_{j=1}^{n_j} (x_{ij} - m_i) = 0$$

Tretji člen pa lahko zapišemo v obliki:

$$\sum_{i=1}^{r}\sum_{j=1}^{n_i}(m_i-m)^2=\sum_{i=1}^{r}n_i(m_i-m)^2$$

Z upoštevanjem obeh izrazov statistiko q zapišemo v obliki:

$$Q=\sum_{i=1}^{r}n_i(m_i-m)^2+\sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij}-m_i)^2=Q_1+Q_2$$

Kjer je:

$$Q_1=\sum_{i=1}^{r}n_i(m_i-m)^2$$

Posledica odstopanj vzorčnega povprečja i-tega nivoja faktorja m_i od celotnega povprečja m.

in :

$$Q_2=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij}-m_i)^2$$

Posledica odstopanj izmerjenih vrednosti x_{ij} od vzorčnega povprečja i-tega nivoja faktorja m_i.

Na osnovi opredeljenih naključnih spremenljivk Q, Q_1 in Q_2 , ki so določene z vrednostmi x_{ij} naključne spremen- ljivke X lahko vplejemo naslednjo statistiko:

$$F=\frac{Q_1(n-k)}{Q_2(r-1)}=\frac{S_1^2}{S_2^2}$$

kjer je $S_1^2=Q_1/(r-1)$ in $S_2^2=Q_2/(n-k)$.

Pri vrednosti statistike:

$$f=\frac{Q_1(n-r)}{Q_2(r-1)}>1$$

je celotno odstopanje Q v večji meri posledica odstopanj Q_1 oziroma vzorčnih povprečij m_i i-tih nivojev faktorja od celotnega povprečja m .

Z drugimi besedami celotno odstopanje Q je v večji meri posledica $m_i \neq m_l \neq m$ oziroma vpliva faktorja.

Ker so Q, Q_1 in Q_2 , funkcije naključne spremenljivke X za statistično ovrednotenje hipoteze:

$$H_0(m_1=m_2=\cdots m_r)$$

potrebujemo porazdelitev statistike:

$$F=\frac{Q_1(n-r)}{Q_2(r-1)}=\frac{S_1^2}{S_2^2}$$

Za katero lahko ob upoštevanju predpostavk 1) in 2) pokažemo, da ima Snedekorjevo ali F porazdelitvijo s prostostnima stopnjama u=r-1in v=n-r.

S poznano porazdelitvijo statistike F in podanega tveganja α lahko na osnovi enačbe:

$$P(F>f_{u,v;\alpha})=\alpha$$

Določimo spodnjo mejo intervala zavračaja:

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

Za izračunano vrednost statistike:

$$f\in S_c=(f_{u,v;\alpha},\infty)$$

Hipotezo $H_0(m_1=m_2=\cdots m_r)$ zavrnemo.

Rezultate računa pri analizi variance prikažemo v *Tabeli analize variance*:

Odstopanje	Vsota kvadratov odstopanj	Število prostostnih stopnen	Srednji kvadrat odstopnaj	Statistika
Med nivoji	Q_1	r-1	$S_1^2=Q_1/(r\text{-}1)$	$F=S_1^2/S_2^2$
Znotraj nivojev	Q_2	n-r	$S_2^2=Q_1/(n\text{-}r)$	
Celotno	Q	n-1	$S^2=Q/(n\text{-}1)$	

Pri tem za računanje uporabimo izpeljanke:

$$Q=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij}-m)^2=\left(\sum_{i=1}^{r}\sum_{j=1}^{n_i}x_{ij}^2\right)-nm^2$$

$$Q_1 = \sum_{i=1}^{r}\sum_{j=1}^{n_i} n_i (m_i - m)^2 = \left(\sum_{i=1}^{r} n_i m_i^2\right) - nm^2$$

$$Q_2 = Q - Q_1$$

Primer : Enotedenski zapis storilnosti treh delavcev je podan v tabeli:

		Dan				
		1	2	3	4	5
Delavec	D_1	20	19	21	22	22
	D_2	19	18	21	22	23
	D_3	21	20	22	19	20

Zanima nas ali lahko s stopnjo tveganja 0.05 sklepamo, da so delavci enako učinkoviti:

$$H_0(m_1 = m_2 = \cdots m_r)$$

Iz podane tabele, določimo srednje vrednosti količine izdelkov in vsote kvadratov za vsakega delavca:

i	$m_i = \frac{1}{n}\sum_{j=1}^{n} x_{ij}$	$\sum_{j=1}^{n} x_{ij}^2$
1	20.8	2170
2	20.2	2055
3	20.4	2086

Izračunamo srednjo vrednost in vsoto kvadratov za celotno skupino:

$$m = \frac{1}{r}\sum_{i=1}^{r} m_i = 20.467 \qquad \sum_{i=1}^{r}\sum_{j=1}^{n_i} x_{ij}^2 = 6311$$

Določimo vrednosti Q, Q_1 in Q_2 :

$$Q = \sum_{i=1}^{r}\sum_{j=1}^{n_i} x_{ij}^2 - nm^2 = 6311 - 15\cdot(20.467)^2 = 27.529$$

$$Q_1 = \sum_{i=1}^{r} n_i m_i^2 - nm^2 = 6284.2 - 15\cdot 20.467^2 = 0.729$$

$$Q_2 = Q - Q_1 = 26.8$$

Izračunamo srednjo vrednost kvadratičnih odstopanj:

$$S_1^2 = \frac{Q_1}{r-1} = \frac{0.729}{2} = 0.365$$

$$S_2^2 = \frac{Q_2}{n-r} = \frac{26.8}{12} = 2.233$$

Vrednosti predstavimo v tabeli analize variance

Odstopanje	Vsota kvadratov odstopanj	Število prostostnih stopnen	Srednji kvadrat odstopnaj	Statistika
Med nivoji	Q_1=0.729	r-1=2	S_1^2=0.365	f=0.163
Znotraj nivojev	Q_2=26.80	n-r=12	S_2^2=2.233	
Celotno	Q=27.529	n-1=14	S^2=1.966	

Vrednost statistike:

$$f = \frac{S_1^2}{S_2^2} = 0.163$$

Iz tabele z porazdelitev F določimo kritično vrednost pri stopnji tveganja 0.05:

$$f_{u,v;\alpha} = f_{2,12;0.05} = 3.885$$

Pripadajoči interval zavračanja:

$$S_c = (f_{u,v;\alpha}, \infty) = (3.885, \infty)$$

Ker:

$$f = 0.163 \notin S_c = (f_{u,v;\alpha}, \infty) = (3.885, \infty)$$

Ničelne hipoteze o enaki učinkovitosti vseh treh delavcev:

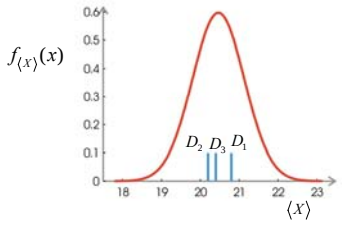
$$H_0(m_1 = m_2 = m_3)$$

ne zavrnemo.

Čeprav je povprečno število izdelkov prvega delavca večje od ostalih dveh je razlika ni dovolj značilna, da bi lahko sklepali o njegovi večji storilnosti.

To nakazuje tudi podatek, da se srednje vrednosti nahajajo znotraj področja, ki ga pokriva gostota porazdelitve povprečnega števila izdelanih izdelkov:

$$f_{\langle X\rangle}(x) = N(m, S_2/\sqrt{n_i})$$



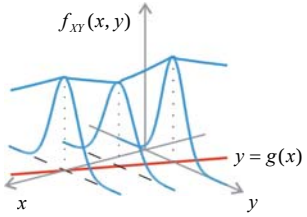
Poglavje 13

Cenilke Funkcij

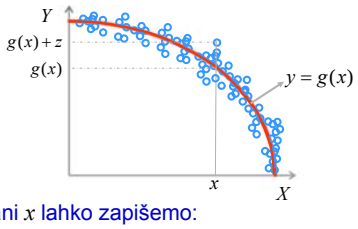
13.1 Osnovni pojmi

Opazujemo naključni pojav, ki ga opišemo z dvema naključnima spremenljivkama X in Y.

V nekaterih primerih lahko opazimo:



Opaženo vzročno povezanost med naključnima spremenljivkama X in Y je v takem primeru smiselno opisati analitično.



Za izbrani x lahko zapišemo:

$$y = g(x) + z$$

Za izbrani x imamo lahko različne y_i in različne razlike z_i-$g(x)$.

X, Y in Z so naključne spremenljivke in v splošnem lahko povezavo med X in Y zapišemo:

$$Y = g(X) + Z = Y_C + Z$$

Kjer Y_c=$g(X)$ imenujemo *cenilka funkcije* ali *prediktor* medsebojne odvisnosti spremenljivk X in Y.

V splošnem lahko danemu y-*nu* ustrezata različna $g(x)$ in z.

Za enolično določitev $g(x)$ in z je potrebno podati dodatni pogoj, ki ga morata izpolnjevati $g(x)$ in z.

Pogoj minimalne povprečne vrednosti kvadrata razlike:

$$\mathrm{E}\left[Z^2\right] = \mathrm{E}\left[(Y - Y_c)^2\right] = \Delta^2 = \min$$

Z^2 imenjemo tudi *kvadratična napaka ocene*.

Povprečno kvadratično napako Δ^2 lahko zapišemo:

$$\Delta^2 = \mathrm{E}\left[Z^2\right] = \mathrm{E}\left[(Y - Y_c)^2\right]$$
$$= \iint (y - g(x))^2 f_{XY}(x, y)\,\mathrm{d}x\,\mathrm{d}y = I(g(x))$$

Naloga je poiskati funkcijo $g(x)$ pri kateri je $I(g(x))$ in s tem Δ^2 minimalen:

$$g_o(x) \rightarrow I(g_o(x)) = \min$$

$g_o(x)$ določimo z *variacijskim postokom*.

Določitev optimalne cenilke funkcije $g_o(x)$:

Predpostavimo, da je $g_o(x)$ funkcija pri kateri je integral:

$$I(g_o(x)) = I_o = \min$$

Poljubno neoptimalno funkcijo lahko zapišemo z:

$$g(x) = g_o(x) + \delta g(x)$$

Kjer $\delta g(x)$ predstavlja dodatek ali variacijo funkcije $g(x)$.

Z uvedbo variacije velja neglede na obliko variacije:

$$I(g(x)) = I(g_o(x) + \delta g(x)) \geq I(g_o(x))$$

Variacijo $\delta g(x)$ lahko zapišemo v obliki:

$$\delta g(x) = \varepsilon \cdot h(x)$$

kjer z ε opišemo amplitudi z $h(x)$ pa obliko variacije $g(x)$.

Z uporabljenim opisom variacije postane $I(g(x))$ funkcija amplitude ε :

$$I(g(x)) = I(g_o(x) + \varepsilon h(x))$$

Pogoj za minimum je zato podan z:

$$\left.\frac{\partial I}{\partial \varepsilon}\right|_{\varepsilon=0} = 0$$

V ta namen odvajamo integral:

$$I(g_o(x) + \varepsilon h(x)) = \iint (y - g_o(x) + \varepsilon h(x))^2 f_{XY}(x, y)\,\mathrm{d}x\,\mathrm{d}y$$

Z odvajanjem in upoštevanjem pogoja dobimo:

$$\left.\frac{\partial I}{\partial \varepsilon}\right|_{\varepsilon=0} = -2\iint (y - g_o(x)) h(x) f_{XY}(x, y)\,\mathrm{d}x\,\mathrm{d}y = 0$$

iz česar nadalje velja:

$$\iint y h(x) f_{XY}(x, y)\,\mathrm{d}x\,\mathrm{d}y = \iint g_o(x) h(x) f_{XY}(x, y)\,\mathrm{d}x\,\mathrm{d}y$$
$$= \int g_o(x) h(x) f_X(x)\,\mathrm{d}x$$

Če upoštevamo da velja:

$$f_{XY}(x, y) = f_{Y|X}(y \mid x) f_X(x)$$

Lahko zapišemo:

$$\iint y h(x) f_{Y|X}(y \mid x) f_X(x)\,\mathrm{d}x\,\mathrm{d}y = \int g_o(x) h(x) f_X(x)\,\mathrm{d}x$$

Gornja enačba je ispolnjena za poljuben $h(x)$ če je:

$$\boxed{g_o(x) = \int y f_{Y|X}(y \mid x)\,\mathrm{d}x = \mathrm{E}[Y \mid X = x]}$$

Pogojno povprečje predstavlja optimalno cenilko, ki minimizira kvadratično napako.

Do enekaga pogoja pridemo tudi, če ocenimo točkovno optimalno cenilko pri danem x :

Pri danem x je statistično povprečje kvadratična napake podana z:

$$\Delta^2 = \mathrm{E}\left[(Y - Y_c)^2 \mid x\right]$$

Minimum je podan z:

$$2\,\mathrm{E}\left[(Y \mid x - Y_c)\right] = \mathrm{E}\left[Y \mid x\right] - Y_C = 0$$

Iz česar sledi:

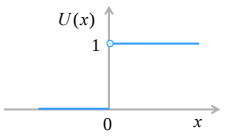
$$Y_C = \mathrm{E}\left[Y \mid X = x\right]$$

Ocena pogojnega povprečja v primeru diskretne spremenljivke

Zapis gostote verjetnost diskretne spremenljivke

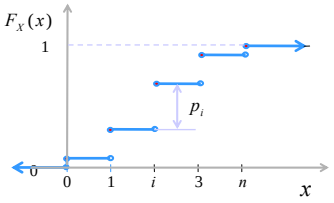
Opredelimo enotsko stopničasto funkcijo $U(x)$:

$$U(x) = \begin{cases} 0, & za\ x < 0 \\ 1, & za\ x \geq 0 \end{cases}$$



S pomočjo $U(x)$ lahko izrazimo ***kumulativno porazdelitev verjetnosti*** $F_X(x_i)$ diskretne spremenljivke X_i :

$$F_X(x) = \sum_{i=1}^{n} p_i U(x - x_i)$$



Gostota verjetnosti je opredeljena z:

$$f_X(x) = \frac{F_X}{\mathrm{d}\,x} = \frac{\mathrm{d}}{\mathrm{d}\,x}\sum_{i=1}^{n} p_i U(x - x_i) = \sum_{i=1}^{n} p_i \frac{\mathrm{d}}{\mathrm{d}\,x} U(x - x_i)$$

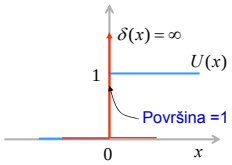
Posredno preko integrala vpeljemo z:

$$U(x) = \int_{-\infty}^{x} \delta(t)\,\mathrm{d}\,t = \begin{cases} 0, & za\ x < 0 \\ 1, & za\ x \geq 0 \end{cases}$$

simbolično podamo odvod stopničaste funkcije $U(x)$:

$$\delta(x) = \frac{\mathrm{d}\,U}{\mathrm{d}\,x} = \begin{cases} 0, & za\ x < 0 \\ \infty, & za\ x \geq 0 \end{cases}$$

Kjer z $\delta(x)$ označimo Dirakovo funkcijo, ki jo grafično ponazorimo :



Z upoštevanjem lastnosti Dirakove funkcije lahko gostoto verjetnost simbolično zapišemo:

$$f_X(x) = \frac{F_X}{\mathrm{d}\,x} = \sum_{i=1}^{n} p_i \frac{\mathrm{d}}{\mathrm{d}\,x} U(x - x_i) = \sum_{i=1}^{n} p_i \delta(x - x_i)$$

Pogojnega povprečja v primeru diskretne spremenljivke

Podana je množica izmerjenih točk $\{X_i, Y_i\ ;\ i=1,\ldots,n\}$

Splošni izraz za empirično povprečje je podano z:

$$\langle h(X,Y)\rangle = \frac{1}{n}\sum_{i=1}^{n} h(x_i, y_j) = \iint h(x,y) f_{XY}(x,y)\,\mathrm{d}\,x\,\mathrm{d}\,y$$

Iz česar sledi:

$$f_{XY}(x,y) = \frac{1}{n}\sum_{i=1}^{n} \delta(x - x_i)\delta(y - y_i)$$

S pomočjo izraza za $f_{XY}(x,y)$ lahko formalno izrazimo gostoto pogojne verjetnosti:

$$f_{Y|X}(y\,|\,x)=\frac{f_{XY}(x,y)}{f_X(x)}=\frac{\sum_{i=1}^{n}\delta(x-x_i)\delta(y-y_i)}{\sum_{i=1}^{n}\delta(x-x_i)}$$

In pogojno povprečje za $Y|X=x$:

$$Y_C(x)=\mathrm{E}[Y\,|\,X=x]=\int y f_{Y|X}(y\,|\,x)\,\mathrm{d}y=\frac{\sum_{i=1}^{n}y_i\delta(x-x_i)}{\sum_{i=1}^{n}\delta(x-x_i)}$$

Pri oceni empirične gostote povezane $f_{XY}(x,y)$ in pogojne $f_{X/Y}(x/y)$ verjetnosti nastopita dva problema :

1) pri vrednostih vzorčnih točkah sta izraza $f_{XY}(x,y)$ in $f_{X/Y}(x/y)$ zaradi nastopanja Dirakovih funkcij singularna
2) Izven vzorčnih točk pa neopredeljena

Problemom se skušamo izogniti tako da:

1) verzeli med diskretnimi točkami poskušamo zapolniti na osnovi predpostavke o povezanosti med merskimi točkami. To nas vodi do *neparametričnega modeliranja* empiričnih povezav na osnovi izmerjenih podatkov.
2) na osnovi poznavanja povezav predpostavimo splošni model zakona.

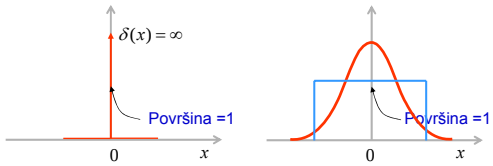
$$Y=g(X,p)$$

Parametre p prilagodimo glede na podatke. V tem primeru govorimo o *parametričnem modeliranju*

13.2 Neparametrična regresija

Težavam z singularnostjo in nepovezanostjo merskih podatkov se lahko izognemo če Dirakovo $\delta(x)$ funkcijo nadomestimo:

$$\delta(x)\to w(x,\sigma)$$



Kjer je w zvezna odvedljiva funkcije, ki izpolnjuje pogoje funkcije gostote verjetnosti in σ st. deviacija.

S tem privzamemo, da so naša merjenja nenatančna, z raztrosom σ.

Izraz za gostoto pogojne verjetnosti zapišemo:

$$f_{Y|X}(y\,|\,x)=\frac{\sum_{i=1}^{n}w(x-x_i,\sigma)w(y-y_i,\sigma)}{\sum_{i=1}^{n}w(x-x_i,\sigma)}$$

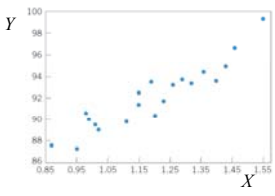
Cenilka pa dobi obliko:

$$Y_C(x)=\int y f_{Y|X}(y\,|\,x)\,\mathrm{d}y=\frac{\sum_{i=1}^{n}y_i w(x-x_i,\sigma)}{\sum_{i=1}^{n}w(x-x_i,\sigma)}$$

Dobljeni izraz imenujemo neparametrična regresija.

13.3 Linearna regresija

Kadar opazimo, da so meritve naključnih spremenljivk X in Y linearno povezane:

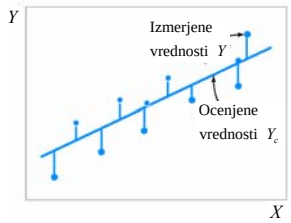


cenilko funkcije za $Y_c(X)$ podamo v obliki:

$$Y_c=g_c(X,p)=aX+b$$

Parametra a in b določimo tako, da bo kvadratična napaka:

$$\Delta^2 = \mathrm{E}\left[(Y - Y_c)^2\right] = \mathrm{E}\left[(Y - aX - b)^2\right] = \min$$



Minimalno napako dobimo ob pogojih:

$$\frac{\partial \Delta^2}{\partial a} = -2\,\mathrm{E}\left[(Y - aX - b)X\right] = 0$$
$$\frac{\partial \Delta^2}{\partial b} = -2\,\mathrm{E}\left[(Y - aX - b)\right] = 0$$

Kar nam da sistem enačb:

$$b = \mathrm{E}[Y] - a\,\mathrm{E}[X]$$
$$0 = \mathrm{E}\left[(Y - \mathrm{E}[Y] - a(X - \mathrm{E}[X]))(X - \mathrm{E}[X])\right]$$

Rešitev sisteam nam da vrednosti parametrov a in b :

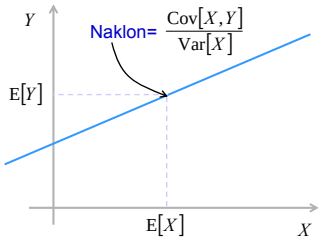
$$a = \frac{\mathrm{Cov}[X,Y]}{\mathrm{Var}[X]}$$
$$b = \mathrm{E}[Y] - \frac{\mathrm{Cov}[X,Y]}{\mathrm{Var}[X]}\mathrm{E}[X]$$

S čimer zapišemo izraz za linearno cenilko:

$$Y_C = \mathrm{E}[Y] + \frac{\mathrm{Cov}[X,Y]}{\mathrm{Var}[X]}(X - \mathrm{E}[X])$$

Graf cenilke:

$$Y_C = \mathrm{E}[Y] + \frac{\mathrm{Cov}[X,Y]}{\mathrm{Var}[X]}(X - \mathrm{E}[X])$$



Z upoštevanjem izrazov za parametra a in b zapišemo izraz za kvadratično napako:

$$\Delta^2 = \mathrm{E}\left[(Y - aX - b)^2\right]$$
$$= \mathrm{E}\left[\left(Y - \mathrm{E}[Y] - \frac{\mathrm{Cov}[X,Y]}{\mathrm{Var}[X]}(X - \mathrm{E}[X])\right)^2\right]$$
$$= (1 - r)^2\,\mathrm{Var}[Y]$$

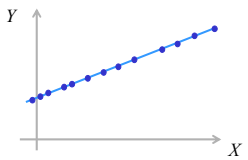
Kjer smo z r označili *korelacijski koeficient*:

$$r = \frac{\mathrm{Cov}[X,Y]}{\sqrt{\mathrm{Var}[X]\mathrm{Var}[Y]}}$$

Iz izraza za kvadratično napako:

$$\Delta^2 = (1 - r)^2\,\mathrm{Var}[Y]$$

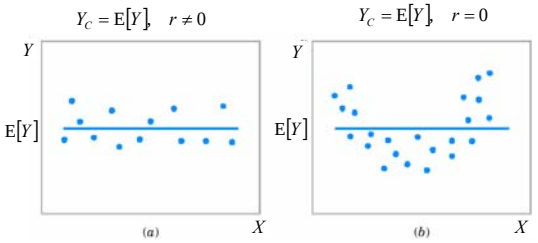
Vidimo da je pri r=1 kvadratična napaka Δ^2=0.
⇒ Povezava med X in Y je deterministična oziroma točke ležijo na premici.



Napaka je največja pri r=0 kadar sta ceniki nekorelirani.

Vrednost cenilke je takrat:

$$Y_C = \mathrm{E}[Y]$$



Regresijska premica je primerna za ocenjevanje medsebojne odvisnosti, če je $r \approx 0.75$.