Meritve konstantnih količin/Statistika

Nekaj statistike smo delali že prej, zato je to bolj grob refresher. Posvetili bomo pozornost šumu. Privzemimo, da je ta porazdeljen po Gaussovi porazdelitvi:

$$r = (z - Hx) = \mathcal{N}(a, \sigma)$$

Ocenjujemo pravzaprav parametra a in σ. Zanima nas ali je porazdelitev omejena če ji "fittamo" Gaussovko, ko gre $n \to \infty$.

Imamo vzorec z_{in}. Vzorčni statistiki sta **povprečje**:

$$\bar{z} = \frac{1}{n} \sum_i z_i$$

in **varianca**:

$$s^2 = \frac{1}{n-1} \sum_i (z_i - \bar{z})^2$$

T-statistika

Definirajmo si novo statistiko takole:

$$T = \frac{\bar{z} - a}{\sqrt{s^2}} \sqrt{n}$$

Imenujemo jo **T-statistika**. Opazimo, da je **odvisna od izbire parametra** a. Poglejmo njeno verjetnostno gostoto:

$$\frac{dP}{dT} = \frac{1}{\sqrt{n-1}\mathrm{B}\left(\frac{n-1}{2}, \frac{1}{2}\right)} \left(1 + \frac{T^2}{n-1}\right)^{-n/2} = \mathrm{S}\left(n-1\right)$$

kjer je:

$$\mathrm{B}\left(x, y\right) = \int_0^1 t^{x-1}(1-t)^{y-1}\, dt$$

Vidimo iz zadnjega člena, da nam da zvonasto obliko in da je normiran. Ker je ta porazdelitev tako pogosto uporabljena jo poimenujejo tudi **Studentova porazdelitev**. Studentov zakon je potem, da je verjetnostna gostota po T statistiki takole:

$$\frac{dP}{dT} = \mathrm{S}\left(n-1\right)$$

Izgleda pa nekako takole:

Postopek uporabe

1. Acquire vzorec meritev z_{in}
2. Izberemo a glede na $\mathcal{N}(a, \sigma)$
3. Izračunamo $\bar{z}$ in s^2

4. Izračunamo T-statistiko T
5. Iz tabeliranih vrednosti za vrednost n izberemo $t_>$ in postavimo stopnjo tveganja α

$$P(|t| > t_>) = \alpha$$

1. Če je $T > t_>$ potem rečemo, da **"na stopnji α zavržemo parameter a"**. V nasprotnem primeru pa rečemo, da **"na stopnji tveganja α parameter a ne moremo zavreči"**.

Interval zaupanje

Interval zaupanje oz. angl. confidence interval definiramo takole:

$$\int_{t_<}^{t_>} \frac{dP}{dT}\, dT = (1-\alpha)$$

Torej $T \in [t_<, t_>]$ na stopnji $(1-\alpha)$. Bolj pravilno rečeno mogoče tako, da lahko **izven** tega intervala pričakujemo T z gotovostjo α

$a \in [a_<, a_>]$ na stopnji $(1-\alpha)$ oz. a lahko izven tega intervala na stopnji tveganja α zavržemo. Pretvorba je slednja:

$$t_< = \frac{\bar{z} - a_>}{s}\sqrt{n} \qquad t_> = \frac{\bar{z} - a_<}{s}\sqrt{n}$$

Porazdelitev χ^2

Imejmo tako kot prej vzorec z_{in} za katerega predpostavimo, da je porazdeljen po Gaussovi porazdelitvi $\mathcal{N}(a, \sigma)$. Sestavimo si statistiko χ^2:

$$\chi^2 = (n-1)\frac{s^2}{\sigma^2}$$

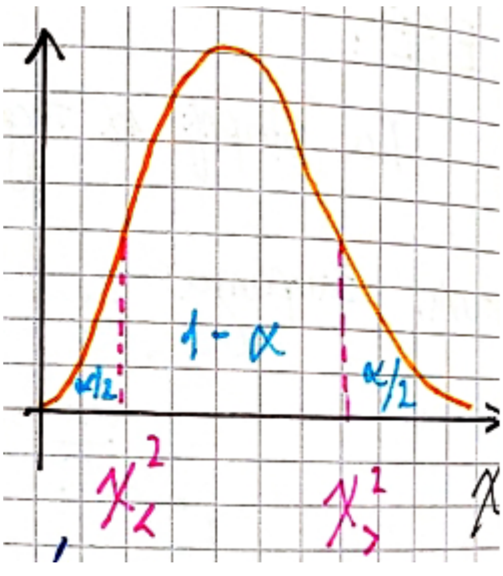
Opazimo, da je odvisna od σ. K sreči ko pogledamo njeno gostoto verjetnosti ni več odvisnosti od σ in je $\frac{dP}{d\chi^2}$ univerzalna funkcija in je tabelirana. Zakon pravi pa takole:

$$\chi^2(n-1) = \frac{dP}{d\chi^2} = \frac{1}{2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)}(\chi^2)^{\frac{n-3}{2}} e^{-\frac{\chi^2}{2}}$$

kjer se spomnimo, da je Gamma funkcija:

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx$$

Narišimo jo na hitro:



Tudi ta porazdelitev je normirana, kar pomeni, da tako kot prej lahko postavimo meje in stopnjo tveganja. Torej z verjetnostjo $(1-\alpha)$ pričakujemo, da bo $\chi^2 \in [\chi^2_<, \chi^2_>]$. V nasprotnem če $\chi^2 \notin [\chi^2_<, \chi^2_>]$ lahko parameter σ na stopnji tveganja α zanemarimo.

Izrek za vsoto kvadratov..

Izrek neznanega imena ali avtorja pravi takole:

Vsota kvadratov n neodvisnih standardizirano normalno naključno porazdeljenih spremenljivk je porazdeljena po zakonu χ^2 z n prostostnimi stopnjami.

Dokaz:

Iz dokaza bomo videli, da je to ekvivalentno naši prejšnji definiciji. Torej imamo x porazdeljen po $\mathcal{N}(0,1)$. Veljata vzorčni statistiki:

$$\sum x^2 = \chi^2 \qquad s^2 = \frac{1}{n-1}\sum(z_i - \bar{z})^2$$

Če imamo nek vzorec z_{in} porazdeljen po $\mathcal{N}(\bar{z},\sigma)$ potem moramo prvo vzorec normalizirati. To storimo takole:

$$\frac{z_i - \bar{z}}{\sigma}$$

To bo porazdeljeno po $\mathcal{N}(0,1)$. Torej če združimo znanje vemo, da bo:

$$\sum_i \frac{(z_i - \bar{z})^2}{\sigma^2}$$

porazdeljen po $\chi^2(n-1)$. Iz tega pa po hitrem obračanju sledi

$$\chi^2 = (n-1)\frac{s^2}{\sigma^2}$$

Izrek za mešanje porazdelitev..

Parafraziram ampak izrek pravi nekako takole:

Naj bo X porazdeljena po $\mathcal{N}(0,1)$ in Y porazdeljena po $\chi^2(n)$. Iz nju lahko tvorimo T-statistiko, ki bo porazdeljena po $\mathrm{S}(n)$ po predpisu:

$$T = \frac{X}{\sqrt{\frac{Y}{n}}}$$

Dokaz:

Dokažimo kar direktno:

$$T = \frac{\bar{z}-a}{s}\sqrt{n} = \frac{(\bar{z}-a)\sqrt{n}}{\sigma\sqrt{\frac{s^2}{\sigma^2}}} = \frac{\frac{\bar{z}-a}{\sigma}}{\frac{\sigma}{\sqrt{n}}}\frac{\sqrt{n-1}}{\sqrt{\frac{s^2}{\sigma^2}(n-1)}}$$

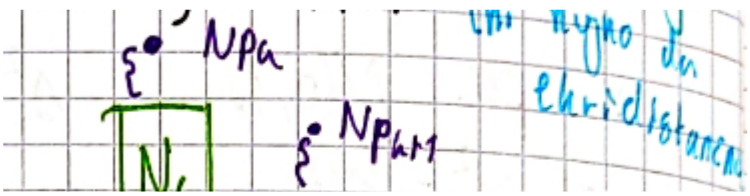
V spodnjem korenu prepoznamo definicijo statistike χ^2. To je (nekako, nisem čisto 100%) porazdeljeno po $\mathrm{S}(n-1)$.

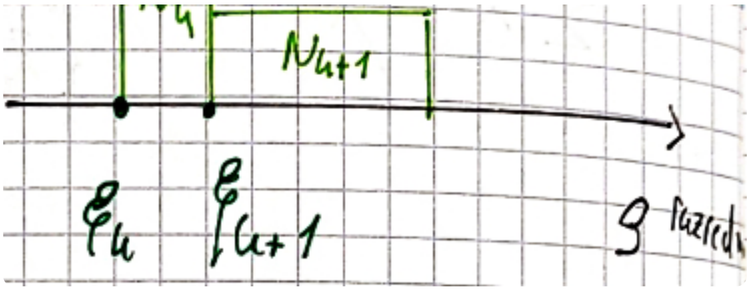
Oblikovni testi

Pearsonov χ^2 test

Namen Pearsonovega χ^2 testa je, da ugotovimo kako dobro se neka določena verjetnostna gostota $\frac{dP}{d\xi}$ ujema z našimi podatki.

Imejmo spet vzorec z_{in}, ki je porazdeljen po **poljubnem porazdelitvenem zakonu** $\frac{dP}{d\xi}$. Sedaj pa naredimo **razrede**. Grafično si to predstavljamo kot histogram samo da ni treba, da so bin-i ekvidistančni. Takole:





Za njih mora veljati slednje:

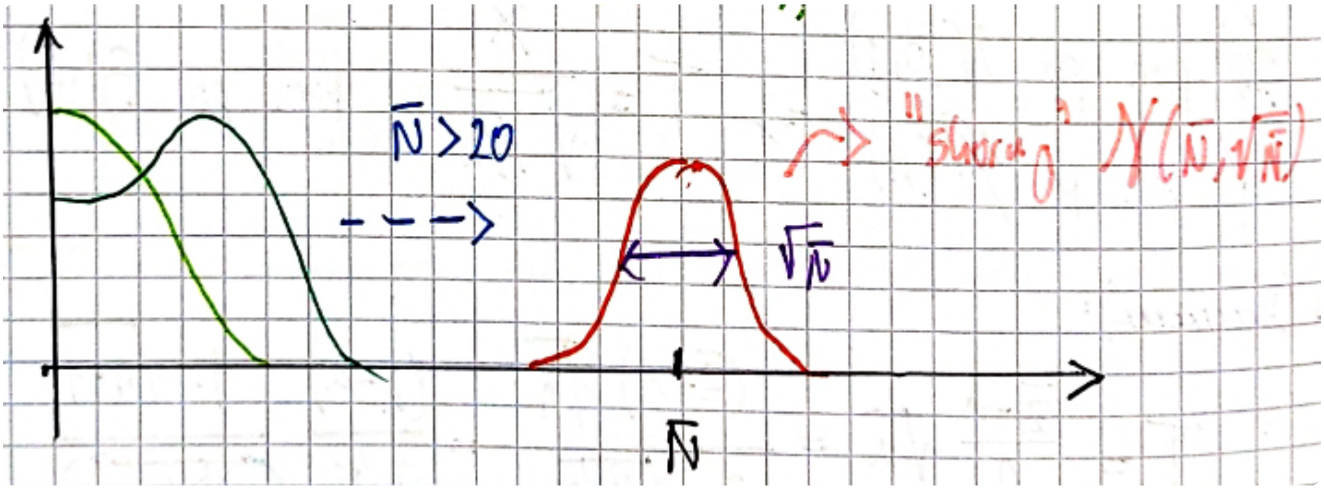
1. Morajo biti brez prekrivanja in brez “lukenj”. Sestavimo jih ρ
2. N_k izmerkov pade v k-ti razred
3. Verjetnost, da pade v en interval:

$$N_{P_k} = N \int_{\xi_k}^{\xi k+1} \frac{dP}{d\xi}\, d\xi$$

1. Če smo si dobro izbrali $(\frac{dP}{d\xi})$ bo $(N_k - N_{P_k})$ le statističen šum. Tako lahko štejemo dogodke $\bar{N}$ in dobimo **Poissonovo** porazdelitev $\frac{dP}{dN}$:

$$w_v = \frac{dP}{dN} = \frac{\bar{N}^v e^{-\bar{N}}}{v!}$$

Pri dovolj velikem $\bar{N}$ se ta porazdelitev bliža $\mathcal{N}(\bar{N}, \sqrt{\bar{N}})$. Visual aid:



in tako pridemo do zaključka, da je:

$$\chi^2 = \sum_{k=1}^{\rho} \frac{N_k - N_{P_k}}{N_{P_k}}$$

porazdeljena po $\chi^2(\rho - 1)$

[Zgled: Radioaktiven razpad]

Opazujemo pojav radioaktivnega razpada. Beležimo čas med sosednjima sunkoma t_k. Recimo da imamo vzorec meritev t_{k100}, torej $N = 100$ in da nikoli ni čas med sunkoma daljši od 1.6s. Ali lahko na stopnjo tveganja $\alpha = 5\%$ ovržemo hipotezo $\tau = 1$s?

Tu preizkušamo če je eksponentna porazdelitev dobra. Torej, ali velja:

$$dP = \frac{1}{\tau} e^{-t/\tau}\, dt$$

Ustvarimo zdaj dva razreda. Prvi razred sega do $t = 1.6$s, kar pomeni, da bomo v njega zajeli vseh $N_0 = 100$ meritev, ostalo do neskončnosti pa je drugi razred, kamor pade $N_1 = 0$ meritev. Kakšne so verjetnosti za razreda?

$$N_{P_0} = N \int_0^{1.6} \frac{1}{\tau} e^{-t/\tau}\, dt = 100 \cdot 1 - e^{-1.6} = 80$$

$$N_{P_1} = N \int_{1.6}^{\infty} \frac{1}{\tau} e^{-t/\tau}\, dt = -100 \cdot e^{-1.6} = 20$$

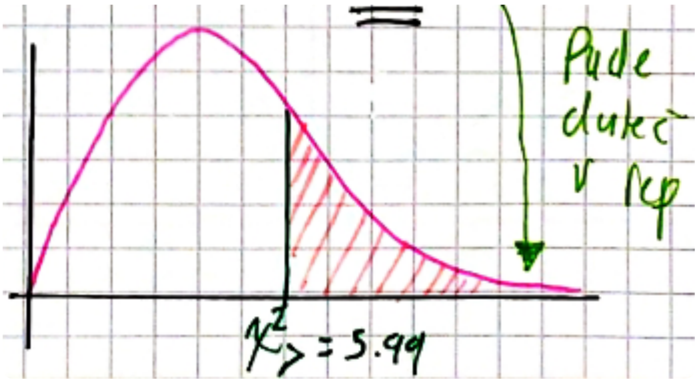
Tvorimo zdaj χ^2:

$$\chi^2 = \frac{(100-80)^2}{80} + \frac{(0-20)^2}{20} = 5 + 20 = 25$$

Pogledamo tabelirano vrednost za $\rho = 2$ in $\alpha = 5\%$, ki znaša:

$$\chi^2_>(2-1)^{5\%} = 5.99$$

To pa je manjše od dobljenega χ^2. Grafično prikazano:



Torej: **lahko zavržemo hipotezo $\tau = 1\text{s}$ na stopnjo tveganja** 5%.

Fischerjev test in funkcija zanesljivosti

Želimo testirati porazdelitveni zakon pri čemer želimo parametre določiti optimalno

$$\frac{dP}{dz}(q_1, q_2, \dots, q_m)$$

Torej imamo m izmerkov, ki jih razdelimo v ρ razredov $[z_{k-1}, z_k)$. Sestavimo **funkcijo zanesljivosti**:

$$L^\star = \prod_k^\rho [P_k(q_i)]^{N_k}$$

kjer je verjetnost za k-ti razred:

$$P_k = \int_{z_{k-1}}^{z_k} \frac{dP}{dz}\, dz = P_k(q_i;\ i = 1, \dots, m)$$

Funkcija zanesljivosti nam podaja verjetnost, da dobimo točno tak histogram, kot smo ga izmerili oz. da velja privzeta porazdelitev. Iz

$$\frac{\partial(mL^\star)}{\partial q_i} = 0$$

dobimo ocene za parametre $q_1^\star, \dots, q_m^\star$ in potem sestavimo χ^2:

$$\chi^2 = \sum_{k=1}^{\rho} \frac{(N_k - N_{P_k}^\star)^2}{N_{P_k}^\star}$$

ki pa je porazdeljena po zakonu $\chi^2(\rho - 1 - m)$.

Lahko pa tudi ne delimo izmerkov na razrede ko računamo funkcijo zanesljivosti in dobimo tako **poenostavljeno funkcijo zanesljivosti**:

$$L = \prod_{k=1}^{N} \frac{dP}{dz}(z_k, q_1, \ldots, q_m)$$

To izračunamo v vrednostih meritev z_k. Potem isto zahtevamo:

$$\frac{\partial(mL)}{\partial q_i} = 0 \quad \Rightarrow \quad \hat{q}_1, \ldots, \hat{q}_m$$

in z njimi tvorimo statistiko:

$$\chi^2 = \sum_{k=1}^{\rho} \frac{(N_k - N_{\hat{P}_k})}{N_{\hat{P}_k}}$$

kjer je verjetnost za k-ti razred definirana kot prej. Pri velikem N je dobljeni χ^2 porazdeljen nekje med $\chi^2(\rho - 1 - m)$ in $\chi^2(\rho - 1)$.

[Zgled: Radioaktivni razpad]

Opazujemo spet radioaktivni razpad za katerega smo določili porazdelitev:

$$\frac{dP}{dt} = \frac{1}{\tau} e^{-t/\tau}$$

Radi bi zdaj našli optimalen τ. Tvorimo poenostavljeno funkcijo zanesljivosti:

$$L = \prod_{k=1}^{N} \frac{1}{\tau} e^{-t_k/\tau} = \frac{1}{\tau^N} \exp\left[-\sum_k \frac{t_k}{\tau}\right]$$

To logaritmirajmo in odvajajmo po parametru, da zahtevamo ekstremalen problem:

$$\ln L = -N \ln\left(\frac{1}{\tau}\right) - \sum_k \frac{t_k}{\tau}$$

$$\frac{\partial \ln L}{\partial \tau} = -\frac{N}{\tau} + \frac{1}{\tau^2} \sum_k t_k = 0 \quad \Rightarrow \quad \frac{1}{\tau} \sum_k t_k = N$$

$$\hat{t} = \frac{1}{N} \sum_{k}^{N} t_k$$

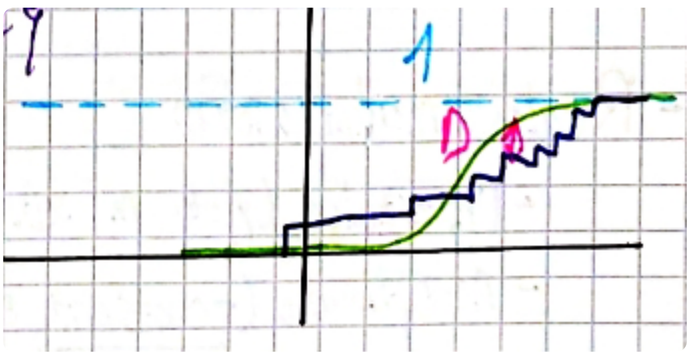
Vidimo, da smo v temu primeru dobili na koncu kar **povprečevanje**, da bo parameter $\hat{\tau}$ optimalen.

Test Kolmogorova oz. Kumulativni test

Imejmo spet vzorec $z_{i\,N}$ in privzet porazdelitveni zakon $\frac{dP}{d\xi}$. Sestavimo **kumulativno porazdelitveno funkcijo**:

$$F(z) = \int_{-\infty}^{z} \frac{dP}{d\xi}\, d\xi$$

Zelo grob grafičen prikaz:



Sestavimo **eksperimentalno kumulativno funkcijo**:

$$f(z) = \frac{k(z)}{N}$$

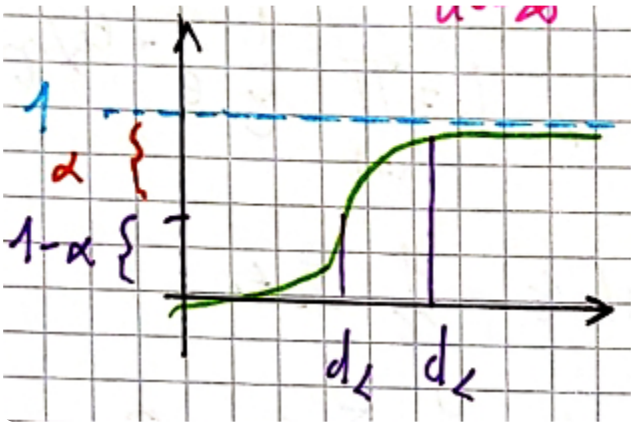
kjer je $k(z)$ funkcija, ki nam pove število vseh izmerkov, manjših od z. Testiramo maksimalni odmik:

$$D = \sup_{-\infty<z<\infty} |F(z) - f(z)|$$

Če $\frac{dP}{d\xi}$ dobro opiše meritve, je D le izjemoma zelo velik. Kolmagorov je napisal takole:

$$\lim_{N\to\infty} P(D\sqrt{N} < d) = \sum_{k=-\infty}^{\infty} (-1)^k e^{2k^2d^2}$$

Zadeva zgleda nekako tako:



Izberemo tak mejni $d_<$, da je $P(D\sqrt{N} < d) = 1 - \alpha$.
Če je $D\sqrt{N} > d_<$, lahko s tveganjem α zavrnemo hipotezo, da $F(z)$ ustreza izmerkom.