Vzorčna disperzija

Imejmo normalno populacijo $N(\mu, \sigma)$. Kako bi določili porazdelitev za vzorčno disperzijo $S_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ ali ali popravljeno vzorčno disperzijo $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Raje izračunamo porazdelitev za statistiko

$$\chi^2 = \frac{nS_0^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

ki jo lahko takole preoblikujemo

$$\chi^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu + \mu - \bar{X})^2 =$$

. . . Vzorčna disperzija

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 + \frac{2}{\sigma^2}(\mu-\bar{X})\sum_{i=1}^{n}(X_i-\mu) + \frac{n}{\sigma^2}(\mu-\bar{X})^2 =$$

in, ker je $\sum_{i=1}^{n}(X_i-\mu) = -n(\bar{X}-\mu)$, dalje

$$= \sum_{i=1}^{n}(\frac{X_i-\mu}{\sigma})^2 - \frac{1}{n}(\sum_{i=1}^{n}\frac{X_i-\mu}{\sigma})^2$$

oziroma

$$\chi^2 = \sum_{i=1}^{n} Y_i^2 - \frac{1}{n}(\sum_{i=1}^{n} Y_i)^2$$

kjer so $Y_1, Y_2, \ldots, Y_n$ paroma neodvisne standardizirano normalno porazdeljene slučajne spremenljivke, $Y_i = \frac{X_i-\mu}{\sigma}$.

. . . Vzorčna disperzija

Porazdelitvena funkcija za χ^2 je

$$F_{\chi^2} = P(\chi^2 < z) = \int\int \ldots \int_{\sum_{i=1}^n y_i^2 - \frac{1}{n}(\sum_{i=1}^n y_i)^2 < z} e^{-(y_1^2+y_2^2+\ldots+y_n^2)/2} dy_n \ldots dy_1$$

z ustrezno ortogonalno transformacijo v nove spremenljivke $z_1, z_2, \ldots, z_n$ dobimo po nekaj računanja

$$F_{\chi^2} = \frac{1}{(2\pi)^{(n-1)/2}} \int\int \ldots \int_{\sum_{i=1}^{n-1} z_i^2 < z} e^{-(z_1^2+z_2^2+\ldots+z_{n-1}^2)/2} dz_{n-1} \ldots dz_1$$

Pod integralom je gostota vektorja $(Z_1, Z_2, \ldots, Z_{n-1})$ z neodvisnimi standardizirano normalnimi členi. Integral sam pa ustreza porazdelitveni funkciji vsote kvadratov $Z_1^2 + Z_2^2 + \ldots + Z_{n-1}^2$. Tako je porazdeljena tudi statistika χ^2.

. . . Vzorčna disperzija

Kakšna pa je ta porazdelitev? Ker so tudi kvadrati Z_1^2, Z_2^2, . . . ,Z_{n-1}^2 med seboj neodvisni in porazdeljeni po zakonu $\chi^2(1)$, je njihova vsota porazdeljena po zakonu $\chi^2(n-1)$. Tako je torej porazdeljena tudi statistika χ^2.

Ker vemo, da je $\mathsf{E}\chi^2(n) = n$ in $\mathsf{D}\chi^2(n) = 2n$, lahko takoj izračunamo

$$\mathsf{E}S_0^2 = \mathsf{E}\frac{\sigma^2\chi^2}{n} = \frac{(n-1)\sigma^2}{n} \qquad \mathsf{E}S^2 = \mathsf{E}\frac{\sigma^2\chi^2}{n-1} = \sigma^2$$

in

$$\mathsf{D}S_0^2 = \mathsf{D}\frac{\sigma^2\chi^2}{n} = \frac{2(n-1)\sigma^4}{n^2} \qquad \mathsf{D}S^2 = \mathsf{D}\frac{\sigma^2\chi^2}{n-1} = \frac{2\sigma^4}{n-1}$$

. . . Vzorčna disperzija

Če je n zelo velik, je po centralnem limitnem izreku statistika χ^2 porazdeljena približno normalno in sicer po zakonu $N(n-1, \sqrt{2(n-1)})$, vzorčna disperzija S_0^2 približno po $N(\frac{(n-1)\sigma^2}{n}, \frac{\sqrt{2(n-1)}\,\sigma^2}{n})$ in popravljena vzorčna disperzija S^2 približno po $N(\sigma^2, \sqrt{\frac{2}{n-1}}\,\sigma^2)$.

Studentova porazdelitev

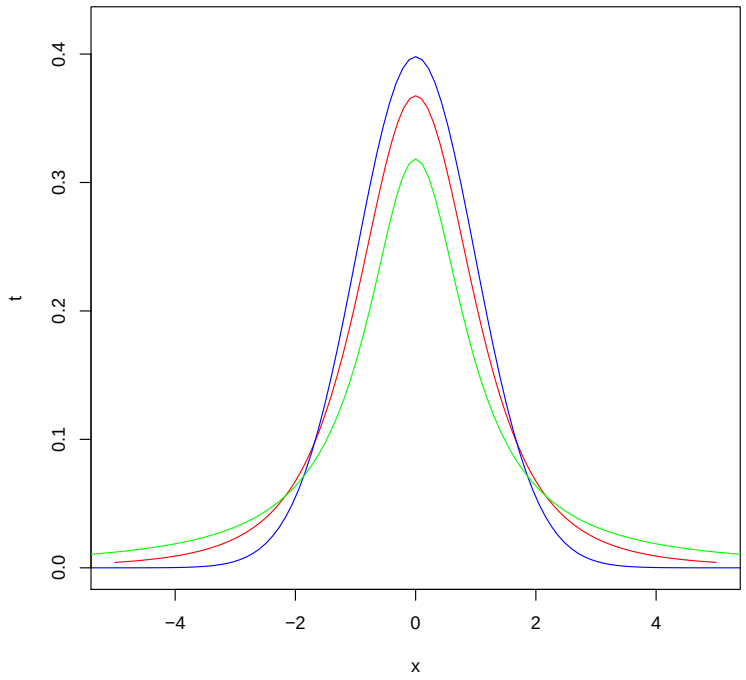
Pri normalno porazdeljeni slučajni spremenljivki X je tudi porazdelitev $\bar{X}$ normalna, in sicer $N(\mu, \frac{\sigma}{\sqrt{n}})$. Statistika $Z = \frac{\bar{X}-\mu}{\sigma}\sqrt{n}$ je potem porazdeljena standardizirano normalno.

Pri ocenjevanju parametra μ z vzorčnim povprečjem $\bar{X}$ to lahko uporabimo le, če poznamo σ; sicer ne moremo oceniti standardne napake – ne vemo, kako dobra je ocena za μ.

Kaj lahko naredimo, če σ ne poznamo? Parameter σ lahko ocenimo s S_0 ali S. Toda S je slučajna spremenljivka in porazdelitev statistike $\frac{\bar{X}-\mu}{S}\sqrt{n}$ ni več $N(0,1)$ (razen, če je n zelo velik in S skoraj enak σ). Kakšna je porazdelitev nove vzorčne statistike

$$T = \frac{\bar{X}-\mu}{S}\sqrt{n} \qquad ?$$

. . . Studentova porazdelitev



Leta 1908 je W.S. Gosset (1876-1937) pod psevdonimom 'Student' objavil članek, v katerem je pokazal, da ima statistika T porazdelitev $S(n-1)$ z gostoto

$$p(t) = \frac{1}{\sqrt{n-1}B(\frac{n-1}{2}, \frac{1}{2})}(1+\frac{t^2}{n-1})^{-\frac{n}{2}}$$

Tej porazdelitvi pravimo *Studentova porazdelitev* z $n-1$ prostostnimi stopnjami.

```
> plot(function(x) dt(x,df=3),-5,5,ylim=c(0,0.42),ylab='t',
  col='red')
> curve(dt(x,df=100),col='blue',add=T)
> curve(dt(x,df=1),col='green',add=T)
```

. . . Studentova porazdelitev

Za $S(1)$ dobimo Cauchyevo porazdelitev z gostoto

$$p(t) = \frac{1}{\pi(1+t^2)}$$

Za $n \to \infty$ pa gre $\frac{1}{\sqrt{n-1}B(\frac{n-1}{2},\frac{1}{2})} \to \sqrt{2\pi}$ in $(1+\frac{t^2}{n-1})^{-\frac{n}{2}} \to e^{-\frac{t^2}{2}}$.
Torej ima limitna porazdelitev gostoto

$$p(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$$

standardizirane normalne porazdelitve.

Če zadnji sliki dodamo

```
> curve(dnorm(x),col='magenta',add=T)
```

ta pokrije modro krivuljo.

Snedecorjeva porazdelitev

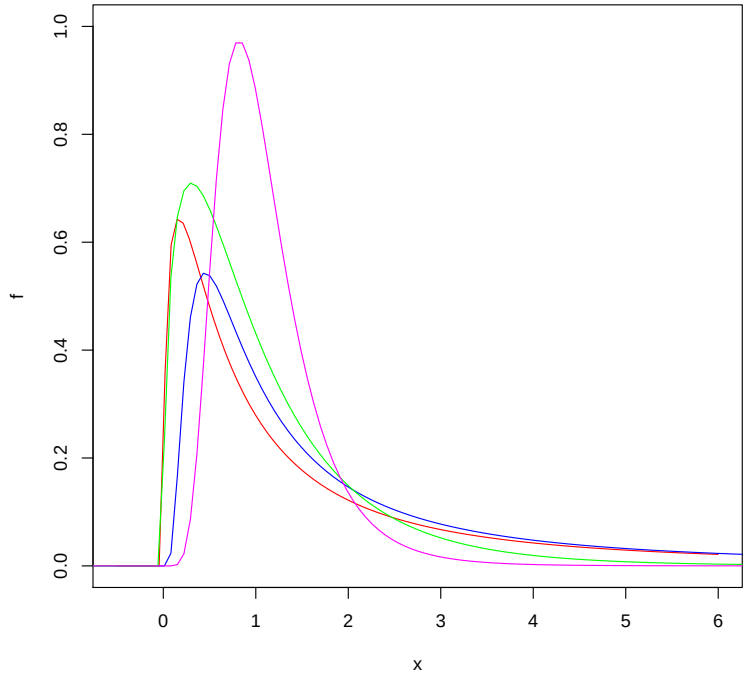
Poskusimo najti še porazdelitev kvocienta $Z = \frac{U}{V}$, kjer sta $U : \chi^2(m)$ in $V : \chi^2(n)$ ter sta U in V neodvisni.

Z nekaj računanja (glej Hladnik) je mogoče pokazati, da je za $x > 0$ gostota ustrezne porazdelitve $F(m, n)$ enaka

$$p(x) = \frac{m^{\frac{m}{2}} n^{\frac{n}{2}}}{B(\frac{m}{2}, \frac{n}{2})} \frac{x^{\frac{m}{2}-1}}{(n + mx)^{\frac{m+n}{2}}}$$

in je enaka 0 drugje.

. . . Snedecorjeva porazdelitev



Porazdelitvi $F(m,n)$ pravimo *Snedecorjeva* (ali tudi Fisherjeva) porazdelitev F z (m,n) prostostnimi stopnjami.

```
> plot(function(x) df(x,df1=3,df2=2),-0.5,6,ylim=c(0,1),ylab='f',
  col='red')
> curve(df(x,df1=20,df2=2),col='blue',add=T)
> curve(df(x,df1=3,df2=20),col='green',add=T)
> curve(df(x,df1=20,df2=20),col='magenta',add=T)
```

. . . Snedecorjeva porazdelitev

Po zakonu $F(m-1, n-1)$ je na primer porazdeljena statistika

$$F = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}$$

saj vemo, da sta spremenljivki

$$U = (m-1)S_X^2/\sigma_X^2 \quad \text{in} \quad V = (n-1)S_Y^2/\sigma_Y^2$$

porazdeljeni po χ^2 z $m-1$ oziroma $n-1$ prostostnimi stopnjami in sta neodvisni.

Velja še:

če je $U : F(m,n)$, je $1/U : F(n,m)$,

če je $U : S(n)$, je $U^2 : F(1,n)$.

Cenilke

Cenilka parametra ζ je vzorčna statistika $C = C(X_1, X_2, X_3, \ldots, X_n)$, katere porazdelitveni zakon je odvisen le od parametra ζ, njene vrednosti pa ležijo v prostoru parametrov.

Od cenilke običajno pričakujemo, da je simetrična – njena vrednost je enaka za vse permutacije argumentov. Seveda je odvisna tudi od velikosti vzorca n .

Primeri: vzorčna mediana $\tilde{X}$ in vzorčno povprečje $\bar{X}$ sta cenilki za populacijsko povprečje μ; popravljena vzorčna disperzija S^2 pa je cenilka za populacijsko disperzijo σ^2 .

Doslednost

Cenilka C parametra ζ je *dosledna*, če z rastočim n zaporedje C_n verjetnostno konvergira k ζ, to je, za vsak $\epsilon > 0$ velja

$$\lim_{n\to\infty} P(|C_n - \zeta| < \epsilon) = 1$$

Primeri: vzorčno povprečje $\bar{X}$ je dosledna cenilka za populacijsko povprečje μ. Tudi vsi *vzorčni začetni momenti*

$$Z_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

so dosledne cenilke ustreznih začetnih populacijskih momentov $z_k = \mathsf{E}X^k$, če le-ti obstajajo.

Vzorčna mediana $\tilde{X}$ je dosledna cenilka za populacijsko mediano.

. . . Doslednost

Če pri pogoju $n \to \infty$ velja $\mathsf{E}C_n \to \zeta$ in $\mathsf{D}C_n \to 0$, je C_n dosledna cenilka parametra ζ.

To sprevidimo takole:

$$1-P(|C_n-\zeta| < \epsilon) = P(|C_n-\zeta| \geq \epsilon) \leq P(|C_n-\mathsf{E}C_n|+|\mathsf{E}C_n-\zeta| \geq \epsilon) \leq$$

upoštevajmo še, da za dovolj velike n velja $|\mathsf{E}C_n - \zeta| < \epsilon/2$, in uporabimo neenakost Čebiševa

$$P(|C_n - \mathsf{E}C_n| \geq \epsilon/2) \leq \frac{4\mathsf{D}C_n}{\epsilon^2} \to 0$$

Primeri: Naj bo $X : N(\mu, \sigma)$. Ker za $n \to \infty$ velja $\mathsf{E}S_0^2 = \frac{(n-1)\sigma^2}{n} \to \sigma^2$ in $\mathsf{D}S_0^2 = \frac{2(n-1)\sigma^4}{n^2} \to 0$, je vzorčna disperzija S_0^2 dosledna cenilka za σ^2.

Nepristranost

Cenilka C_n parametra ζ je *nepristranska*, če je $\mathsf{E}C_n = \zeta$ (za vsak n); in je *asimptotično nepristranska*, če je $\lim_{n\to\infty} \mathsf{E}C_n = \zeta$.

Količino $B(C_n) = \mathsf{E}C_n - \zeta$ imenujemo *pristranost* (angl. *bias*) cenilke C_n.

Primeri: vzorčno povprečje $\bar{X}$ je nepristranska cenilka za populacijsko povprečje μ ; vzorčna disperzija S_0^2 je samo asimptotično nepristranska cenilka za σ^2, popravljena vzorčna disperzija S^2 pa je nepristranska cenilka za σ^2.

Intervalsko ocenjevanje parametrov

Naj bo X slučajna spremenljivka na populaciji G z gostoto verjetnosti odvisno od parametra ζ.

Slučajna množica $M \subset \mathbb{R}$, ki je odvisna le od slučajnega vzorca, ne pa od parametra ζ, se imenuje *množica zaupanja* za parameter ζ, če obstaja tako število α, $0 < \alpha < 1$, da velja $P(\zeta \in M) = 1 - \alpha$. Število $1 - \alpha$ imenujemo tedaj *stopnja zaupanja*; število α pa *stopnja tveganja*.

Stopnja zaupanja je običajno 95% ali 99% – $\alpha = 0.05$ ali $\alpha = 0.01$.
Pove nam, kakšna je verjetnost, da M vsebuje vrednost parametra ζ ne glede na to, kakšna je njegova dejanska vrednost.

Če je množica M interval $M = [A, B]$, ji rečemo *interval zaupanja* (za parameter ζ).
Njegovi krajišči sta funkciji slučajnega vzorca – torej statistiki.

. . . Intervalsko ocenjevanje parametrov

Naj bo $X : N(\mu, \sigma)$ in recimo, da poznamo parameter σ in ocenjujemo parameter μ. Izberimo konstanti a in b, $b > a$, tako da bo $P(a \leq Z \leq b) = 1 - \alpha$, kjer je $Z = \frac{\bar{X}-\mu}{\sigma}\sqrt{n}$. Tedaj je

$$P(\bar{X} - \frac{b\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} - \frac{a\sigma}{\sqrt{n}}) = 1 - \alpha$$

Označimo $A = \bar{X} - \frac{b\sigma}{\sqrt{n}}$ in $B = \bar{X} - \frac{a\sigma}{\sqrt{n}}$. Za katera a in b je interval $[A, B]$ najkrajši? Pokazati je mogoče (Lagrangeova funkcija), da mora biti $a = -b$ in $\Phi(b) = (1 - \alpha)/2$; oziroma, če označimo $b = z_{\alpha/2}$, velja $P(Z > z_{\alpha/2}) = \alpha/2$. Iskani interval je torej

$$A = \bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \qquad B = \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

z verjetnostjo $1 - \alpha$ je $|\bar{X} - \mu| < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. Od tu dobimo, da mora biti za to, da bo z verjetnostjo $1 - \alpha$ napaka manjša od ε, $n > (\frac{z_{\alpha/2}\sigma}{\varepsilon})^2$.

. . . Intervalsko ocenjevanje parametrov

Če pri porazdelitvi $X : N(\mu, \sigma)$ tudi parameter σ ni znan, ga nadomestimo s cenilko S in moramo zato uporabiti Studentovo statistiko $T = \frac{\bar{X}-\mu}{S}\sqrt{n}$. Ustrezni interval je tedaj

$$A = \bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}}, \qquad B = \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}$$

kjer je $P(T > t_{\alpha/2}) = \alpha/2$.

Če pa bi ocenjevali parameter σ^2, uporabimo statistiko $\chi^2 = (n-1)\frac{S^2}{\sigma^2}$, ki je porazdeljena po $\chi^2(n-1)$. Tedaj sta

$$A = \frac{(n-1)S^2}{b}, \qquad B = \frac{(n-1)S^2}{a}$$

Konstanti a in b včasih določimo iz pogojev $P(\chi^2 < a) = \alpha/2$ in $P(\chi^2 > b) = \alpha/2$; najkrajši interval pa dobimo, ko velja zveza $a^2 p(a) = b^2 p(b)$ in seveda $\int_a^b p(t)dt = 1 - \alpha$.