Porazdelitve vzorčnih statistik

Denimo, da je v populaciji N enot in
da iz te populacije slučajno izbiramo n enot
v enostavni slučajni vzorec ali na kratko slučajni vzorec
(vsaka enota ima enako verjetnost, da bo izbrana v vzorec, tj. $1/N$).

Če hočemo dobiti slučajni vzorec, moramo izbrane
enote pred ponovnim izbiranjem vrniti v populacijo
(vzorec s ponavljanjem).

Če je velikost vzorca v primerjavi s populacijo majhna,
se ne pregrešimo preveč, če imamo za slučajni vzorec tudi vzorec,
ki nastane s slučajnim izbiranjem brez vračanja.

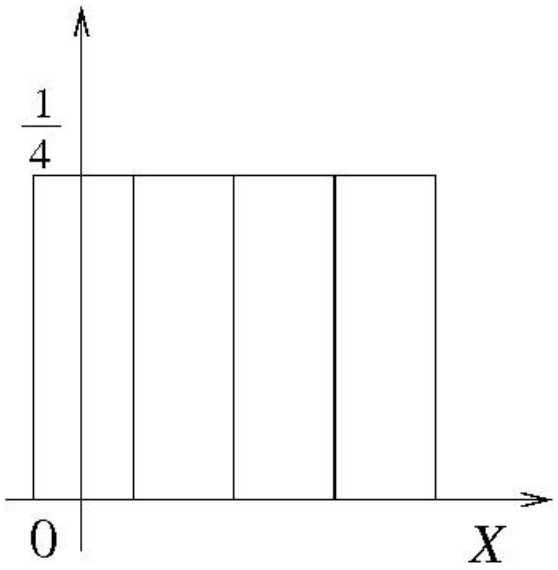
Predstavljajmo si, da smo iz populacije izbrali vse možne vzorce.
Dobili smo populacijo vseh možnih vzorcev.
Teh je v primeru enostavnih slučajnih vzorcev (s ponavljanjem) N^n;
kjer je N število enot v populaciji in n število enot v vzorcu.

Število slučajnih vzorcev brez ponavljanja pa je $\binom{N}{n}$,
če ne upoštevamo vrstnega reda izbranih enot v vzorcu,
oziroma $\binom{N+n-1}{n}$, če upoštevamo vrstni red.

Primer: Vzemimo populacijo z $N = 4$ enotami, ki imajo naslednje vrednosti spremenljivke X:

$$0, 1, 2, 3$$

Grafično si lahko porazdelitev spremenljivke X predstavimo s histogramom:



in izračunamo populacijsko aritmetično sredino in varianco:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{3}{2}$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \frac{5}{4}$$

Sedaj pa tvorimo vse možne vzorce velikosti $n = 2$ s ponavljanjem in na vsakem izračunajmo vzorčno aritmetično sredino $\overline{X}$

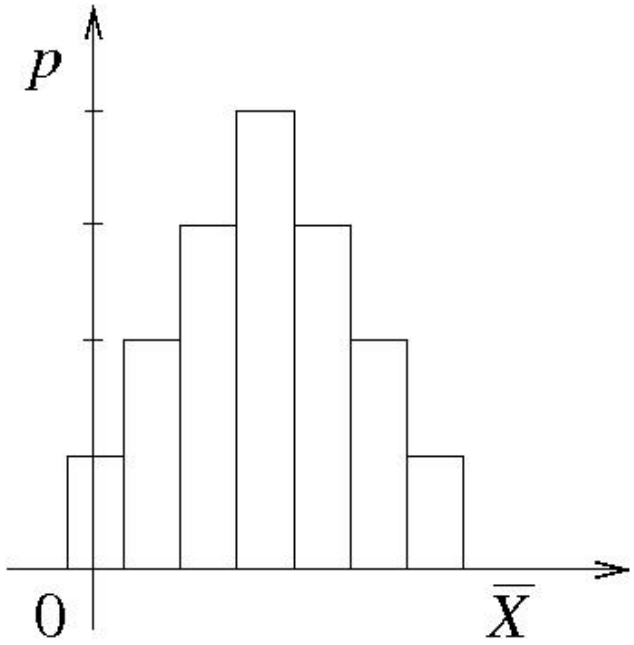
$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

vzorci	$\overline{X}$	vzorci	X
$0,0$	0	$2,0$	1
$0,1$	$0,5$	$2,1$	$1,5$
$0,2$	1	$2,2$	2
$0,3$	$1,5$	$2,3$	$2,5$
$1,0$	$0,5$	$3,0$	$1,5$
$1,1$	1	$3,1$	2
$1,2$	$1,5$	$3,2$	$2,5$
$1,3$	2	$3,3$	3

Zapišimo verjetnostno shemo slučajne spremenljivke vzorčno povprečje $\overline{X}$:

$$\overline{X} : \begin{pmatrix} 0 & 0,5 & 1 & 1,5 & 2 & 2,5 & 3 \\ 1/16 & 2/16 & 3/16 & 4/16 & 3/16 & 2/16 & 1/16 \end{pmatrix}$$

Grafično jo predstavimo s histogramom:



... in izračunajmo matematično upanje ter disperzijo vzorčnega povprečja:

$$E(X) = \sum_{i=1}^{m} \overline{X}_i p_i = \frac{0+1+3+6+6+5+3}{16} = \frac{3}{2}$$

$$D(X) = \sum_{i=1}^{m} \left(\overline{X}_i - E(\overline{X}) \right)^2 p_i = \frac{5}{8}$$

S tem primerom smo pokazali, da je statistika ‘vzorčna aritmetična sredina’ slučajna spremenljivka s svojo porazdelitvijo. Poglejmo, kaj lahko rečemo v splošnem o porazdelitvi vzorčnih aritmetičnih sredin.

Vzorčna porazdelitev povprečja

Centralni limitni izrek

Če je naključni vzorec velikosti n izbran iz populacije s končnim povprečjem μ in varianco σ^2, potem je lahko, če je n dovolj velik, vzorčna porazdelitev povprečja $\overline{y}$ aproksimirana z gostoto normalne porazdelitve.

Naj bo $y_1, y_2, \ldots, y_n$ naključni vzorec, ki je sestavljen iz n meritev populacije s končnim povprečjem μ in končnim standardnim odklonom σ. Potem sta povprečje in standardni odklon vzorčne porazdelitve $\overline{y}$ enaka

$$\mu_{\overline{Y}} = \mu, \qquad \text{and} \qquad \sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}}.$$

Porazdelitve vzorčnih statistik

Vzorčna statistika je poljubna simetrična funkcija (vrednost neodvisna od permutacije argumentov) vzorca

$$Y = g(X_1, X_2, X_3, \ldots, X_n)$$

Tudi vzorčna statistika je slučajna spremenljivka, za katero lahko določimo porazdelitev iz porazdelitve vzorca. Najzanimivejši sta značilni vrednosti njeno matematično upanje $\mathsf{E}Y$ in standardni odklon σY, ki mu pravimo tudi *standardna napaka* statistike Y (angl. standard error – zato oznaka $\mathrm{SE}(Y)$).

Vzorčno povprečje

Vzorčno povprečje je določeno z zvezo

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Recimo, da ima spremenljivka X parametra $\mathsf{E}X = \mu$ in $\mathsf{D}X = \sigma^2$. Tedaj je

$$\mathsf{E}\overline{X} = \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}X_i = \mu$$

$$\mathsf{D}\overline{X} = \frac{1}{n^2}\sum_{i=1}^{n} \mathsf{D}X_i = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}$$

Iz druge zveze vidimo, da standardna napaka $\sigma\overline{X} = \dfrac{\sigma}{\sqrt{n}}$ statistike $\overline{X}$ pada z naraščanjem velikosti vzorca – $\overline{X} \to \mu$;
(enako nam zagotavlja tudi krepki zakon velikih števil).

Denimo, da se spremenljivka X na populaciji porazdeljuje normalno $N(\mu, \sigma)$. Na vsakem vzorcu (s ponavljanjem) izračunamo vzorčno aritmetično sredino $\overline{X}$. Dokazati se da, da je **porazdelitev vzorčnih aritmetičnih sredin** normalna, kjer je

- matematično upanje vzorčnih aritmetičnih sredin enako aritmetični sredini spremenljivke na populaciji

$$E(\overline{X}) = \mu,$$

- standardni odklon vzorčnih aritmetičnih sredin

$$\mathrm{SE}(\overline{X}) = \frac{\sigma}{\sqrt{n}}.$$

Če tvorimo vzorce iz končne populacije brez vračanja, je standardni odklon vzorčnih aritmetičnih sredin

$$\mathrm{SE}(\overline{X}) = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}.$$

Za dovolj velike vzorce ($n > 30$) je porazdelitev vzorčnih aritmetičnih sredin približno normalna, tudi če spremenljivka X ni normalno porazdeljena. Če se statistika X porazdeljuje vsaj približno normalno s standardno napako $\mathrm{SE}(X)$, potem se

$$Z = \frac{X - E(X)}{\mathrm{SE}(X)}$$

porazdeljuje standardizirano normalno.

Vzorčno povprečje in normalna porazdelitev

Naj bo $X : N(\mu, \sigma)$. Tedaj je $\sum_{i=1}^{n} X_i : N(n\mu, \sigma\sqrt{n})$ in dalje $\overline{X} : N(\mu, \sigma/\sqrt{n})$. Tedaj je vzorčna statistika

$$Z = \frac{\overline{X} - \mu}{\sigma}\sqrt{n} : N(0, 1)$$

Kaj pa če porazdelitev X ni normalna? Izračun porazdelitve se lahko zelo zaplete. Toda pri večjih vzorcih ($n > 30$), lahko uporabimo centralni limitni izrek, ki zagotavlja, da je spremenljivka Z porazdeljena skoraj standardizirano normalno. Vzorčno povprečje

$$\overline{X} = \frac{\sigma}{\sqrt{n}}Z + \mu$$

ima tedaj porazdelitev približno $N(\mu, \sigma/\sqrt{n})$.

Zgled

Odgovorimo na vprašanje: Kolikšna je verjetnost, da bo pri 36 metih igralne kocke povprečno število pik večje ali enako 4 ?

X je slučajna spremenljivka z vrednostmi 1,2,3,4,5,6 in verjetnostmi $1/6$. Zanjo je $\mu = 3,5$ in standardni odklon $\sigma = 1,7$. Vseh 36 ponovitev meta lahko obravnavamo kot slučajni vzorec velikost 36. Tedaj je

$$P(\overline{X} \geq 4) = P(Z \geq (4-\mu)\sqrt{n}/\sigma) = P(Z \geq 1,75) \approx 0,04.$$

```
> x <- 1:6
> m <- mean(x)
> s <- sd(x)*sqrt(5/6)
> z <- (4-m)*6/s
> p <- 1-pnorm(z)
> cbind(m,s,z,p)
        m        s       z          p
[1,] 3.5 1.707825 1.75662 0.03949129
```

Vzorčna disperzija

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

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

ki jo lahko takole preoblikujemo

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

. . . Vzorčna disperzija

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

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

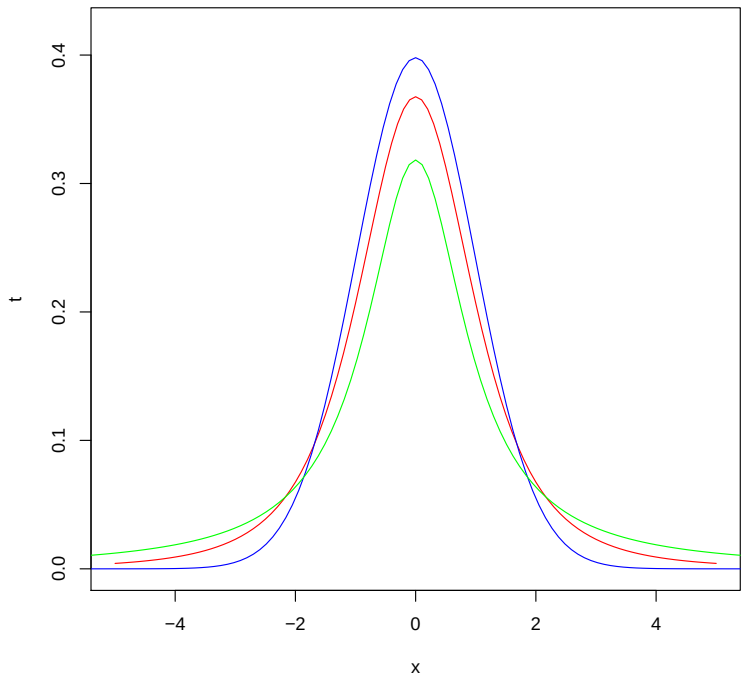
Pri normalno porazdeljeni slučajni spremenljivki X je tudi porazdelitev $\overline{X}$ normalna, in sicer $N(\mu, \frac{\sigma}{\sqrt{n}})$. Statistika $Z = \frac{\overline{X}-\mu}{\sigma}\sqrt{n}$ je potem porazdeljena standardizirano normalno.

Pri ocenjevanju parametra μ z vzorčnim povprečjem $\overline{X}$ 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 $\frac{\overline{X}-\mu}{S}\sqrt{n}$ ni več $N(0,1)$ (razen, če je n zelo velik in S skoraj enak σ). Kakšna je porazdelitev nove vzorčne statistike

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

. . . 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

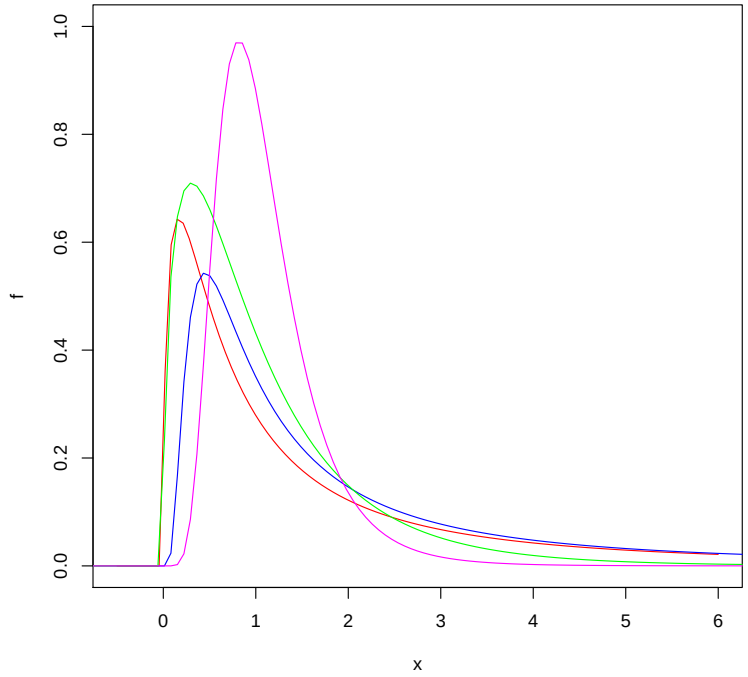
Poskusimo najti še porazdelitev kvocienta $Z = \frac{U}{V}$, kjer sta $U : \chi^2(m)$ in $V : \chi^2(n)$ ter sta U in V neodvisni.

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 $\overline{X}$ 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 $\overline{X}$ 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 $\overline{X}$ 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.

Disperzija nepristranskih cenilk

Izmed nepristranskih cenilk istega parametra ζ je boljša tista, ki ima manjšo disperzijo – v povprečju daje bolj točne ocene.

Če je razred cenilk parametra ζ *konveksen* (vsebuje tudi njihove konveksne kombinacije), obstaja v bistvu ena sama cenilka z najmanjšo disperzijo:

Naj bo razred nepristranskih cenilk parametra ζ konveksen. Če sta C in C' nepristranski cenilki, obe z najmanjšo disperzijo σ^2, je $C = C'$ z verjetnostjo 1.

Za to poglejmo

$$\mathsf{D}(\frac{1}{2}(C+C')) = \frac{1}{4}(\mathsf{D}C+\mathsf{D}C'+2\mathsf{Cov}(C,C')) \leq (\frac{1}{2}(\sqrt{\mathsf{D}C}+\sqrt{\mathsf{D}C'}))^2 = \sigma^2$$

Ker sta cenilki minimalni, mora biti tudi $\mathsf{D}(\frac{1}{2}(C+C')) = \sigma^2$ in dalje $\mathsf{Cov}(C,C') = \sigma^2$ oziroma $r(C,C') = 1$. Torej je $C' = aC + b$, $a > 0$ z verjetnostjo 1. Iz $\mathsf{D}C = \mathsf{D}C'$ izhaja $a = 1$, iz $\mathsf{E}C = \mathsf{E}C'$ pa še $b = 0$.

Srednja kvadratična napaka

Včasih je celo bolje vzeti pristransko cenilko z manjšo disperzijo, kot jo ima druga, sicer nepristranska, cenilka z veliko disperzijo.

Mera *učinkovitosti* cenilk parametra ζ je *srednja kvadratična napaka*

$$q(C) = \mathsf{E}(C - \zeta)^2$$

Ker velja

$$q(C) = \mathsf{E}(C - \mathsf{E}C + \mathsf{E}C - \zeta)^2 = \mathsf{E}(C - \mathsf{E}C)^2 + (\mathsf{E}C - \zeta)^2$$

jo lahko zapišemo tudi v obliki

$$q(C) = \mathsf{D}C + B(C)^2$$

Za nepristranske cenilke je $B(C) = 0$ in zato $q(C) = \mathsf{D}C$.

Če pa je disperzija cenilke skoraj 0, je $q(C) \approx B(C)^2$.

Rao-Cramérjeva ocena

Naj bo f gostotna ali verjetnostna funkcija slučajne spremenljivke X in naj bo odvisna še od parametra ζ, tako da je $f(x;\zeta)$ njena vrednost v točki x. Združeno gostotno ali verjetnostno funkcijo slučajnega vzorca $(X_1, X_2, X_3, \ldots, X_n)$ označimo z L in ji pravimo *funkcija verjetja* (tudi *zanesljivosti*, angl. *likelihood*)

$$L(x_1, x_2, x_3, \ldots, x_n; \zeta) = f(x_1;\zeta) f(x_2;\zeta) f(x_3;\zeta) \cdots f(x_n;\zeta)$$

Velja (*): $\int \int \ldots \int L(x_1, x_2, \ldots, x_n; \zeta) dx_1 dx_2 \ldots dx_n = 1.$

$L(X_1, X_2, X_3, \ldots, X_n)$ je funkcija vzorca – torej slučajna spremenljivka.

Privzemimo, da je funkcija L vsaj dvakrat zvezno odvedljiva po ζ na nekem intervalu I in naj na tem intervalu tudi integral odvoda L po ζ enakomerno konvergira.

. . . Rao-Cramérjeva ocena

Odvajajmo enakost (*) po ζ in upoštevajmo $\frac{\partial \ln L}{\partial \zeta} = \frac{1}{L}\frac{\partial L}{\partial \zeta}$ pa dobimo

$$\int \int \ldots \int \frac{\partial \ln L}{\partial \zeta} L dx_1 dx_2 \ldots dx_n = 0$$

kar lahko tolmačimo kot $\mathsf{E}\frac{\partial \ln L}{\partial \zeta} = 0$.

Naj bo sedaj C nepristranska cenilka parametra ζ, torej $\mathsf{E}C = \zeta$, oziroma zapisano z integrali $\int \int \ldots \int CL dx_1 dx_2 \ldots dx_n = \zeta$.

Ker C ni odvisna od ζ, dobimo z odvajanjem po ζ:

$$\int \int \ldots \int C\frac{\partial \ln L}{\partial \zeta} L dx_1 dx_2 \ldots dx_n = 1$$

kar pomeni $\mathsf{E}(C\frac{\partial \ln L}{\partial \zeta}) = 1$.

. . . Rao-Cramérjeva ocena

Če to enakost združimo s prejšnjo (pomnoženo s ζ), dobimo:

$$\mathsf{E}((C-\zeta)\frac{\partial \ln L}{\partial \zeta}) = 1$$

Od tu po $(\mathsf{E}XY)^2 \leq \mathsf{E}X^2\mathsf{E}Y^2$ izhaja naprej

$$1 = (\mathsf{E}((C-\zeta)\frac{\partial \ln L}{\partial \zeta}))^2 \leq \mathsf{E}(C-\zeta)^2\mathsf{E}(\frac{\partial \ln L}{\partial \zeta})^2 = \mathsf{D}C\ \mathsf{E}(\frac{\partial \ln L}{\partial \zeta})^2$$

kar da *Rao-Cramérjevo oceno*

$$\mathsf{D}C \geq (\mathsf{E}(\frac{\partial \ln L}{\partial \zeta})^2)^{-1} = (-\mathsf{E}\frac{\partial^2 \ln L}{\partial \zeta^2})^{-1} = (n\mathsf{E}(\frac{\partial \ln f}{\partial \zeta})^2)^{-1}$$

Učinkovitost cenilk

Rao-Cramérjeva ocena da absolutno spodnjo mejo disperzije za vse nepristranske cenilke parametra ζ (v dovolj gladkih porazdelitvah). Ta meja ni nujno dosežena. Cenilka, ki jo doseže, se imenuje *najučinkivitejša cenilka* parametra ζ in je ena sama (z verjetnostjo 1).

Kdaj pa je ta spodnja meja dosežena?

V neenakosti $(\mathsf{E}XY)^2 \leq \mathsf{E}X^2\mathsf{E}Y^2$, ki je uporabljena v izpeljavi Rao-Cramérjeve ocene, velja enakost natanko takrat, ko je $Y = cX$ z verjetnostjo 1.

. . . Učinkovitost cenilk

Torej velja v Rao-Cramérjevi oceni enakost natanko takrat, ko je

$$\frac{\partial \ln L}{\partial \zeta} = A(\zeta)(C - \zeta)$$

kjer je $A(\zeta)$ konstanta, odvisna od ζ in neodvisna od vzorca.

Zato je tudi

$$(\mathsf{D}C)^{-1} = \mathsf{E}(\frac{\partial \ln L}{\partial \zeta})^2 = A(\zeta)^2 \mathsf{E}(C - \zeta)^2 = A(\zeta)^2 \mathsf{D}C$$

oziroma končno

$$\mathsf{D}C = |A(\zeta)|^{-1}$$

Najučinkovitejše cenilke za parametre normalne porazdelitve

Naj bo $X : N(\mu, \sigma)$. Tedaj je

$$L = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-((\frac{X_1-\mu}{\sigma})^2+\cdots+(\frac{X_n-\mu}{\sigma})^2)/2}$$

in

$$\ln L = \ln \frac{1}{(2\pi)^{n/2}\sigma^n} - ((\frac{X_1-\mu}{\sigma})^2 + \cdots + (\frac{X_n-\mu}{\sigma})^2)/2$$

ter dalje

$$\frac{\partial \ln L}{\partial \mu} = \frac{X_1-\mu}{\sigma^2} + \cdots + \frac{X_n-\mu}{\sigma^2} = \frac{n}{\sigma^2}(\overline{X}-\mu)$$

Torej je vzorčno povprečje $\overline{X}$ najučinkovitejša cenilka za μ z disperzijo $\mathsf{D}\overline{X} = \frac{\sigma^2}{n}$.

. . . normalna porazdelitev

Prvi člen v izrazu za $\ln L$ lahko zapišemo tudi $-\frac{n}{2}(\ln 2\pi + \ln \sigma^2)$. Tedaj je, če privzamemo, da je μ znano število

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}((X_1-\mu)^2 + \cdots + (X_n-\mu)^2) = \frac{n}{2\sigma^4}(S_\mu^2 - \sigma^2)$$

To pomeni, da je $S_\mu^2 = \frac{1}{n}\sum_{i=1}^n (X_i-\mu)^2$ najučinkovitejša cenilka za parameter σ^2 z disperzijo $\mathsf{D}S_\mu^2 = \frac{2\sigma^4}{n}$.

Poissonova porazdelitev

Za Poissonovo porazdelitev $P(\lambda)$ s parametrom λ, $p_k = \lambda^k \frac{e^{-\lambda}}{k!}$ je

$$L = e^{-n\lambda} \frac{\lambda^{x_1+\cdots+x_n}}{x_1! \cdots x_n!}$$

in dalje

$$\ln L = -n\lambda + (x_1 + \cdots + x_n) \ln \lambda - \ln(x_1! \cdots x_n!)$$

ter končno

$$\frac{\partial \ln L}{\partial \lambda} = -n + \frac{x_1 + \cdots + x_n}{\lambda} = \frac{n}{\lambda}(\overline{X} - \lambda)$$

Najučinkovitejša cenilka za parameter λ je $\overline{X}$ z disperzijo $\mathsf{D}\overline{X} = \frac{\lambda}{n}$.

Učinkovitost cenilke

Naj bo C_0 najučinkovitejša cenilka parametra ζ in C kaka druga nepristranska cenilka. Tedaj je *učinkovitost* cenilke C določena s predpisom

$$e(C) = \frac{\mathsf{D}C_0}{\mathsf{D}C}$$

Učinkovitost najučinkovitejše cenilke je $e(C_0) = 1$.

Če najučinkovitejša cenilka ne obstaja, vzamemo za vrednost $\mathsf{D}C_0$ desno stran v Rao-Cramérjevi oceni.

Primer: Naj bo $X : N(\mu, \sigma)$. Pri velikih n-jih je vzorčna mediana $\tilde{X}$ – ocena za μ, porazdeljena približno po $N(\mu, \sigma\sqrt{\frac{\pi}{2n}})$. Torej je

$$e(\tilde{X}) = \frac{\mathsf{D}\overline{X}}{\mathsf{D}\tilde{X}} = \frac{\frac{\sigma^2}{n}}{\frac{\pi\sigma^2}{2n}} = \frac{2}{\pi} \approx 0.64$$

. . . Učinkovitost cenilke

Primer: Naj bo $X : N(\mu, \sigma)$. Če poznamo μ, je najučinkovitejša cenilka za σ^2 statistika S_μ^2 z disperzijo $\mathsf{D}S_\mu^2 = \frac{2\sigma^4}{n}$. Popravljena vzorčna disperzija S^2 pa je nepristranska cenilka istega parametra z disperzijo $\mathsf{D}S^2 = \frac{2\sigma^4}{n-1}$. Torej je učinkovitost S^2

$$e(S^2) = \frac{\mathsf{D}S_\mu^2}{\mathsf{D}S^2} = \frac{\frac{2\sigma^4}{n}}{\frac{2\sigma^4}{n-1}} = \frac{n-1}{n}$$

Iz tega vidimo, da $e(S^2) \to 1$, ko $n \to \infty$. Pravimo, da je cenilka S^2 *asimptotično najučinkovitejša cenilka* za σ^2.

Metoda momentov

Recimo, da je za zvezno slučajno spremenljivko X njena gostota f odvisna od m parametrov $f(x;\zeta_1,\zeta_2,\zeta_3,\ldots,\zeta_m)$ in naj obstajajo momenti

$$z_k = \int_{-\infty}^{\infty} x^k f(x;\zeta_1,\zeta_2,\zeta_3,\ldots,\zeta_m)dx$$

za $k = 1, 2, 3, \ldots, m$. Če se dajo iz teh enačb enolično izračunati parametri ζ_1, ζ_2, ζ_3, …,ζ_m kot funkcije momentov z_1, z_2, z_3, …, z_m

$$\zeta_k = \varphi_k(z_1, z_2, z_3, \ldots, z_m)$$

potem so

$$C_k = \varphi_k(Z_1, Z_2, Z_3, \ldots, Z_m)$$

cenilke parametrov ζ_k po *metodi momentov*. k-ti vzorčni začetni moment $Z_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$ je cenilka za ustrezni populacijski moment z_k.

Cenilke, ki jih dobimo po metodi momentov so dosledne.

. . . Metoda momentov

Naj bo $X : N(\mu, \sigma)$. Tedaj je $z_1 = \mu$ in $z_2 = \sigma^2 + \mu^2$. Od tu dobimo $\mu = z_1$ in $\sigma^2 = z_2 - z_1^2$. Ustrezni cenilki sta $Z_1 = \overline{X}$ za μ in

$$Z_2 - Z_1^2 = \overline{X^2} - \overline{X}^2 = S_0^2$$

za σ^2 – torej vzorčno povprečje in disperzija.

Metoda največjega verjetja

Funkcija verjetja

$$L(x_1, x_2, x_3, \ldots, x_n; \zeta) = f(x_1; \zeta) f(x_2; \zeta) f(x_3; \zeta) \cdots f(x_n; \zeta)$$

je pri danih $x_1, x_2, x_3, \ldots, x_n$ odvisna še od parametra ζ. Izberemo tak ζ, da bo funkcija L dosegla največjo vrednost. Če je L vsaj dvakrat zvezno odvedljiva, mora veljati $\frac{\partial L}{\partial \zeta} = 0$ in $\frac{\partial^2 L}{\partial \zeta^2} < 0$. Največja vrednost parametra je še odvisna od $x_1, x_2, x_3, \ldots, x_n$:
$\zeta_{max} = \varphi(x_1, x_2, x_3, \ldots, x_n)$. Tedaj je cenilka za parameter ζ enaka

$$C = \varphi(X_1, X_2, X_3, \ldots, X_n)$$

Metodo lahko posplošimo na večje število parametrov.

Pogosto raje iščemo maksimum funkcije $\ln L$.

Če najučinkovitejša cenilka obstaja, jo dobimo s to metodo.

. . . Metoda največjega verjetja - binomska

Naj bo $X : B(1, p)$. tedaj je $f(x; p) = p^x(1-p)^{1-x}$, kjer je $x = 0$ ali $x = 1$. Ocenjujemo parameter p. Funkcija verjetja ima obliko $L = p^x(1-p)^{n-x}$, kjer je sedaj $x \in \{0, 1, 2, \ldots, n\}$. Ker je $\ln L = x \ln p + (n-x)\ln(1-p)$, dobimo

$$\frac{\partial \ln L}{\partial p} = \frac{x}{p} - \frac{n-x}{1-p},$$

ki je enak 0 pri $p = \frac{x}{n}$. Ker je v tem primeru $\frac{\partial^2 \ln L}{\partial p^2} = -\frac{x}{p^2} - \frac{n-x}{(1-p)^2} < 0$, je v tej točki maksimum. Cenilka po metodi največjega verjetja je torej $P = \frac{X}{n}$, kjer je X binomsko porazdeljena spremenljivka – frekvenca v n ponovitvah. Cenilka P je nepristranska, saj je $\mathsf{E}P = \frac{\mathsf{E}X}{n} = p$. Ker za $n \to \infty$ gre $\mathsf{D}P = \frac{\mathsf{D}X}{n^2} = \frac{p(1-p)}{n} \to 0$, je P dosledna cenilka. P je tudi najučinkovitejša $\frac{\partial \ln L}{\partial p} = \frac{X}{p} - \frac{n-X}{1-p} = \frac{n}{p(1-p)}(\frac{X}{n} - p) = \frac{n}{p(1-p)}(P - p)$.

. . . Metoda največjega verjetja - Poissonova

Za Poissonovo porazdelitev $P(\lambda)$ s parametrom λ, $p_x = \lambda^x \frac{e^{-\lambda}}{x!}$ je

$$L = e^{-n\lambda} \frac{\lambda^{x_1+\cdots+x_n}}{x_1! \cdots x_n!}$$

in dalje

$$\ln L = -n\lambda + (x_1 + \cdots + x_n) \ln \lambda - \ln(x_1! \cdots x_n!)$$

ter končno

$$\frac{\partial \ln L}{\partial \lambda} = -n + \frac{x_1 + \cdots + x_n}{\lambda} = \frac{n}{\lambda}(\overline{X} - \lambda)$$

Odvod je enak 0 za $\lambda = \overline{X}$. Drugi odvod v tej točki je $\frac{\partial^2 \ln L}{\partial \lambda^2} = -\frac{x_1+\cdots+x_n}{\lambda^2} < 0$. V točki je maksimum.

Cenilka za λ po metodi največjega verjetja je vzorčno povprečje $\overline{X}$.

Je tudi najučinkovitejša cenilka za λ z disperzijo $\mathsf{D}\overline{X} = \frac{\lambda}{n}$.

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 $Z = \frac{\overline{X}-\mu}{\sigma}\sqrt{n}$. Tedaj je

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

Označimo $A = \overline{X} - \frac{b\sigma}{\sqrt{n}}$ in $B = \overline{X} - \frac{a\sigma}{\sqrt{n}}$. 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

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

z verjetnostjo $1 - \alpha$ je $|\overline{X} - \mu| < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. 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 $T = \frac{\overline{X}-\mu}{S}\sqrt{n}$. Ustrezni interval je tedaj

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

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 $\int_a^b p(t)dt = 1 - \alpha$.