A412

2021. december 2., csütörtök 12:13

A412

A4 Valószínűségszámítás — XII. EA

dr. Keszthelyi Gabriella
Sztochasztika és Dinamikai rendszerek tanszék

2021. december 2.

Momentumgeneráló függvény (ismétlés)

$$M_X(s) = E(e^{sX}) = \mathcal{L}\{f\}(-s) = \int_{-\infty}^{\infty} f(x)e^{sx}\, dx$$

A momentumgeneráló függvény

- egy várható érték s függvényében,
- Laplace transzformáltja a sűrűségfüggvény előjelcserélt változatának,
- és a sűrűségfüggvény/súlyfüggvény kölcsönösen meghatározzák egymást.

Gamma/Erlang momentumgeneráló függvénye

Gamma sűrűségfüggvénye:

$$f(x) = \frac{\lambda(\lambda x)^{p-1}}{\Gamma(p)} e^{-\lambda x}$$

Erlang sűrűségfüggvénye ($p = n$):

$$f(x) = \frac{\lambda(\lambda x)^{n-1}}{(n-1)!} e^{-\lambda x}$$

Momentumgeneráló:

$$M_X(s) = \frac{\lambda^n}{(n-1)!} \int_0^\infty e^{sx} e^{-\lambda x} x^{n-1}\, dx =$$

$$= \left(\frac{\lambda}{\lambda - s}\right)^n \left[\frac{1}{(n-1)!} \int_0^\infty e^{-(\lambda - s)x} x^{n-1} (\lambda - s)^n\, dx\right] = 1$$

$$= \left(\frac{\lambda}{\lambda - s}\right)^n$$

Gamma/Erlang momentumgeneráló függvénye

$$M'_X(s) = \frac{n\lambda^n}{(\lambda - s)^{n+1}}$$

$$M''_X(s) = \frac{n(n+1)\lambda^n}{(\lambda - s)^{n+2}}$$

Tehát

$$E(X) = M'_X(0) = \frac{n}{\lambda}$$

$$D^2(X) = E(X^2) - (E(X))^2 = M''_X(0) - \left(\frac{n}{\lambda}\right)^2 =$$

$$= \frac{n(n+1)}{(\lambda)^2} - \frac{n^2}{(\lambda)^2} = \frac{n}{(\lambda)^2}$$

χ^2-eloszlás

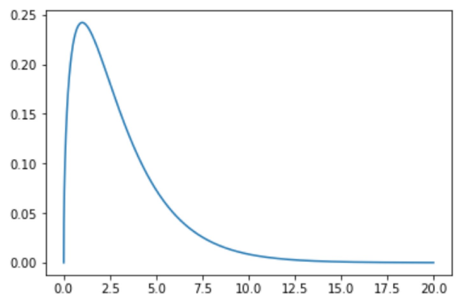
Legyenek $Z_1, Z_2, \ldots, Z_n$ független, standard normál eloszlású változók. Ekkor
$X = Z_1^2 + Z_2^2 + \cdots + Z_n^2$-t χ^2-eloszlásnak hívjuk n szabadsági fokkal.
Jelölés: $X \sim \chi_n^2$
Kiejtés: 'Khí-négyzet'
A χ^2 eloszlás tartja a konvolúciót, azaz $X_1 \sim \chi_{n_1}^2$ és $X_2 \sim \chi_{n_2}^2$, akkor $W = X_1 + X_2 \sim \chi_{n_1+n_2}^2$.

Példa
Tegyük fel, hogy egy tárgyat akarunk lokalizálni a 3D térben, és a 3 koordináta mérési hibái független normális változók 0 várható értékkel és 2 szórással. Mi a vszsége, hogy az objektum és az általunk becsült hely távolsága több, mint 3 méter?

Példa

Jelöljük W-vel a távolságot. Ekkor
$W^2 = X_1^2 + X_2^2 + X_3^2$ ahol X_i az i. koordináta hibája. Legyen Z_i a standardizált változata, azaz $Z_i = X_i/2$.

$$P(W^2 > 9) = P(Z_1^2+Z_2^2+Z_3^2 > 9/4) = P(\chi_3^2 > 9/4) = \underline{0.52216}$$



χ_3^2 sűrűsége

χ^2 momentumgeneráló függvénye

$n = 1$ esetén

$$E(e^{sX}) = E(e^{sZ^2}) = \int_{-\infty}^{\infty} e^{sX^2} f_Z(x)\, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \underbrace{e^{sX^2}}_{e^{sx^2}} e^{-x^2/2}\, dx =$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2(1-2s)/2}\, dx$$

STANDARD NORMÁLIS sűrűsége $= 1$

$$= (1-2s)^{-1/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2\sigma^2}\, dx =$$

$$= \boxed{(1-2s)^{-1/2}}$$

általános n-re

$$E(e^{sX}) = E(e^{s\sum_{i=1}^{n} Z_i^2}) = E(\prod_{i=1}^{n} e^{sZ_i^2}) = (1-2s)^{-n/2}$$

$$(1-2s)^{-n/2}$$

vs.

$$\left(\frac{\lambda}{\lambda - t}\right)^{n}$$

Ez a kettő megegyezik ha $\lambda = 1/2$ és $n = n/2$. Azaz a χ^2 n szabadságfokkal és a Gamma/Erlang $n/2$ és $1/2$ paraméterrel egybeesik.

Mire használjuk a χ^2-eloszlást?
A χ^2-et ritkán használjuk közvetlenül modellezésre, inkább hipotézis-vizsgálat kapcsán merül fel. A χ^2-próba arra jó, hogy egy adatsorról eldöntsük, hogy egy általunk gyanított eloszlásra mennyire jól illeszkedik.

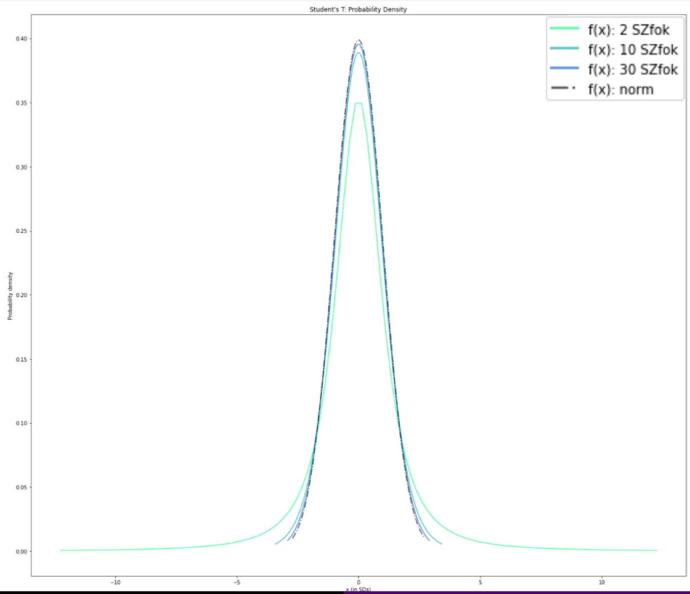
Student-féle T-eloszlás

Legyen Z standard normális és χ^2_n független valváltozók. Ekkor az n szabadságfokú T-eloszlás a kövekezőképp van definiálva:

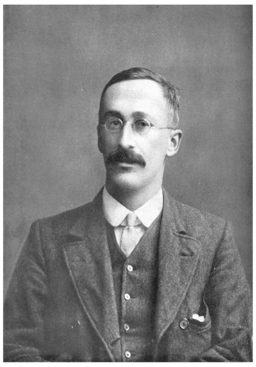
$$T_n = \frac{Z}{\sqrt{\chi^2_n / n}}$$

T-eloszlást akkor használunk, amikor gyanús, hogy normális eloszlású a minta (harang alakú a sűrűség), de kis mintával dolgozunk és a szórás ismeretlen.
A T-eloszlás szabadságfoka (n) nem más, mint a minta elemszáma mínusz 1. A szabadságfok növelésével a Student-eloszlás tart a Standard Normálhoz (30 fölött már egybeesnek).

Student-féle T-eloszlás



Student-féle T-eloszlás



William Sealy Gosset egy angol kémikus volt, aki a Guinness fő sörözfőjeként szolgált, és a modern statisztika úttörője volt.
Ő publikálta **Student** álnéven a T-eloszlást.

$E(T_n) = 0$
$D^2(T_n) = \frac{n}{n-2}$

Paraméterbecslés

Legyen X valváltozó, θ pedig a valváltozó paramétere. Tegyük fel, hogy van egy n elemű mintánk (statisztikai minta) $X_n = (X_1, X_2, \ldots, X_n)$. Ez n darab független kísérletet/megfigyelést jelent az X valváltozóra néve. A mintaelemek eloszlása megegyezik a valváltozó eredeti eloszlásával.

Az X_n pontbecslése:

$$\hat{\Theta}(X_n) = g(X_1, X_2, \ldots, X_n)$$

Ahol $X_1 = x_1, X_2 = x_2, \ldots$ megfigyelések alapján meg tudjuk becsülni a θ paramétert.

Legegyszerűbb becslések

Legyen X várható értéke μ és szórása σ. $\hat{\Theta}(X_n) = \overline{X_n}$ a mintaátlag.

$$E(\overline{X_n}) = \frac{1}{n} E\left(\sum_{j=1}^{n} X_j\right) = \mu$$

$$\overline{X_n} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

$$\frac{1}{n}\left(E(X_1) + E(X_2) + \ldots + E(X_n)\right) = \frac{1}{n} \cdot n \cdot \mu$$

A $\overline{X_n}$ szórása:

$$D^2(\overline{X_n}) = \frac{1}{n^2} D^2\left(\sum_{j=1}^{n} X_j\right) = \frac{\sigma^2}{n}$$

Egy becslést **torzítatlan**nak nevezünk, ha

$$E(\hat{\Theta}) = \theta$$

Ekkor tehát $\overline{X_n}$ egy torzítatlan becslés μ-re.

A minta szórása

A tapasztalati szórásnégyzet

$$\hat{S}^2 = \frac{1}{n} \sum_{j=1}^{n} (X_j - \overline{X_n})^2$$

ahol $\overline{X_n}$ még mindig a mintaátlag.

$$\begin{aligned}
\hat{S}^2 &= \frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2 = \frac{1}{n}\sum_{j=1}^{n}(X_j - \mu + \mu - \overline{X}_n)^2 \\
&= \frac{1}{n}\sum_{j=1}^{n}\left\{(X_j - \mu)^2 + 2(X_j - \mu)(\mu - \overline{X}_n) + (\mu - \overline{X}_n)^2\right\} \\
&= \frac{1}{n}\sum_{j=1}^{n}(X_j - \mu)^2 + \frac{2}{n}(\mu - \overline{X}_n)\sum_{j=1}^{n}(X_j - \mu) + \frac{1}{n}\sum_{j=1}^{n}(\mu - \overline{X}_n)^2 \\
&= \frac{1}{n}\sum_{j=1}^{n}(X_j - \mu)^2 + \frac{2}{n}(\mu - \overline{X}_n)(n\overline{X}_n - n\mu) + \frac{n(\mu - \overline{X}_n)^2}{n}
\end{aligned}$$

$$= \frac{1}{n}\sum_{j=1}^{n}(X_j - \mu)^2 - 2(\overline{X}_n - \mu)^2 + (\overline{X}_n - \mu)^2$$

$$= \frac{1}{n}\sum_{j=1}^{n}(X_j - \mu)^2 - (\overline{X}_n - \mu)^2.$$

$$E[\hat{S}^2] = E\left[\frac{1}{n}\sum_{j=1}^{n}(X_j - \mu)^2 - (\overline{X}_n - \mu)^2\right]$$

$$= \frac{1}{n}\sum_{j=1}^{n}[E[(X_j - \mu)^2] - E[(\overline{X}_n - \mu)^2]]$$

$$= \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2 \quad \neq \frac{\sigma^2}{n}$$

Tehát $\hat{S}^2$ torzított becslése a σ-nak.

Torzítatlan becslés szórásnégyzetre

A becslés torzítása

$$B(\hat{\Theta}) = E(\hat{\Theta}) - \theta \qquad \frac{\sigma^2}{n} - \frac{n-1}{n}\sigma^2$$

Hogyan nyerhetünk ki egy torzítatlan becslést a szórásnégyzetre?

$$\hat{\sigma^2} = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2$$

torzítatlan becslés a szórásn

Maximum Likelihood becslés (MAP)

Legyen $f(x_1, x_2, \ldots, x_n|\theta)$ az együttes sűrűségfüggvénye ez egy likelihoodfüggvény
$X_1, X_2, \ldots, X_n$-nek.
$f(x_1, x_2, \ldots, x_n|\theta)$ egyfajta valószínűséget fejez ki, hogy az $x_1, x_2, \ldots, x_n$ értékek lesznek megfigyelve, amikor θ az igazi értéke a paraméternek.

$$\begin{cases} P(X_1, X_2, \ldots, X_n|\theta), & \text{ha X diszkrét,} \\ f(x_1, x_2, \ldots, x_n|\theta), & \text{ha X folytonos} \end{cases}$$

Mivel X_i-k függetlenek:

$$P(X_1, X_2, \ldots, X_n|\theta) = P(X_1|\theta) \cdot P(X_2|\theta) \cdot \ldots P(X_n|\theta) = \prod_{i=1}^{n} P(X_i|\theta)$$

$$f(x_1, x_2, \ldots, x_n|\theta) = f(x_1|\theta) \cdot f(x_2|\theta) \cdot \ldots f(x_n|\theta) \cdot = \prod_{i=1}^{n} f(x_i|\theta)$$

Log Likelihood

A Maximum Likelihood becslése $\hat{\theta}$-nak, az a θ amely maximalizálja $f(x_1, x_2, \ldots, x_n|\theta)$-t.

$$l(x_1, x_2, \ldots, x_n; \hat{\theta}) = \max_{\theta} l(x_1, x_2, \ldots, x_n; \theta)$$

Néha érdemesebb a **log likelihood**dal dolgozni:

$$L(x_1, x_2, \ldots, x_n; \theta) = \log l(x_1, x_2, \ldots, x_n; \theta) =$$

$$= \begin{cases} \sum_{i=1}^{n} \log P(X_1, X_2, \ldots, X_n|\theta), & \textit{ha X diszkrét,} \\ \sum_{i=1}^{n} \log f(x_1, x_2, \ldots, x_n|\theta), & \textit{ha X folytonos} \end{cases}$$

$$= \sum_{i=1}^{n} L(x_j; \theta)$$

Ahhoz, hogy maximalizáljuk $l(x_1, x_2, \ldots, x_n; \theta)$-t elég maximalizálni $L(x_1, x_2, \ldots, x_n; \theta)$-t:

$$\frac{\partial}{\partial \theta} L(x_1, x_2, \ldots, x_n; \theta) = \frac{\partial}{\partial \theta} \log l(x_1, x_2, \ldots, x_n; \theta) = 0$$

Példa

Legyenek X_i-k független $Exp(\lambda)$ valváltozók, melyeknek nem ismerjük a várható értékét és így λ-t sem: ez lesz a paraméter amit becsülni szeretnénk.

$$f(x_1, x_2, ..., x_n|\lambda) = \lambda e^{-x_1\lambda} \lambda e^{-x_2\lambda} \ldots \lambda e^{-x_n\lambda} = \lambda^n e^{-\sum_{i=1}^n x_i\lambda}$$

$\log(\lambda^n e^{-\sum_{i=1}^n x_i\lambda})$

$$L(x_1, x_2, \ldots, x_n; \lambda) = \sum_{i=1}^{n} (\log(\lambda) - x_i\lambda)$$

Loglikelihood

$$\frac{\partial}{\partial \lambda} L(x_1, x_2, \ldots, x_n; \lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i$$

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

Azaz λ-ra kijött, hogy a mintaátlag reciproka.

Poisson becslése

Példa
A Géza által benyújtott dolgozatban a hibák száma Poisson eloszlású 1 hiba/oldal várható értékkel, míg Béla által írt dolgozatban átlagosan 5 hiba található oldalanként. A beküldött dolgozat egy oldalán 2 hiba található, mi a valószínűbb, melyikük írta?

$\theta = \lambda$

$$P(X = 2|\lambda = 1) = \frac{1^2}{2!}e^{-1} = 0,1839$$

$$P(X = 2|\lambda = 5) = \frac{5^2}{2!}e^{-5} = 0,0842$$

Poisson becslése 2.

$$P(X = k_j|\theta) = \frac{\theta^{k_j}}{k_j!}e^{-\theta}$$

$\log\left(\frac{\theta^{k_j}}{k_j!}e^{-\theta}\right)$

A Loglikelihood függvény:

$$L(k_1, k_2, \dots, k_n; \theta) = \sum_{j=1}^{n} \ln P(X = k_j|\theta) = \sum_{j=1}^{n}(k_j \ln\theta - \theta - \ln k_j!) =$$

$$= \ln\theta \sum_{j=1}^{n} k_j - n\theta - \sum_{j=1}^{n} \ln k_j!$$

$\rightarrow \frac{1}{\theta}\sum k_j - n = 0$

Megkeressük a maximumot:

$$0 = \frac{\partial}{\partial\theta}L(k_1, k_2, \dots, k_n; \theta) = \frac{1}{\theta}\sum_{j=1}^{n} k_j - n$$

$$E(X) = \theta = \frac{1}{n}\sum_{j=1}^{n} k_j = \bar{X}$$

Bernoulli becslése Likelihooddal

Tegyük fel, hogy van n Bernoulli változóm p paraméterrel. Legyen $(k_1, k_2, \dots k_n)$ a kimenetelek, ekkor

$$P(X = k_j|\theta = p) = \underline{p^{k_j}(1-p)^{1-k_j}}$$

k_j: $\nearrow 0$, $\searrow 1$

$P(X=1)=p$

$P(X=0)=1-p$

$$L(k_1, \dots, k_n; \theta) = \ln(\prod p^{k_j}(1-p)^{1-k_j}) = \sum_{j=1}^{n}(k_j \ln p + (1-k_j)\ln(1-p))$$

Megkeressük a maximumot:

$$0 = \frac{\partial}{\partial p} L(k_1, k_2, \dots, k_n; p) = \frac{1}{p}\sum_{j=1}^{n} k_j - \frac{1}{1-p}\sum_{j=1}^{n}(1-k_j) =$$

$$= -\frac{n}{1-p} + \left(\frac{1}{p} + \frac{1}{1-p}\right)\sum_{j=1}^{n} k_j = -\frac{n}{1-p} + \frac{1}{p(1-p)}\sum_{j=1}^{n} k_j$$

$$\theta = p = \frac{1}{n}\sum_{j=1}^{n} k_j = \overline{X}$$

$$0 = -\frac{n}{1-p} + \frac{1}{p(1-p)}\sum_{j=1}^{n} k_j$$

$$\frac{n}{\cancel{1-p}} = \frac{1}{p\cancel{(1-p)}}\sum_{j=1}^{n} k_j$$

$$p = \frac{1}{n}\sum_{j=1}^{n} k_j$$

Konfidencia intervallum

Ahelyett, hogy pontosan meg akarnánk becsülni a keresett paramétert, gyakran elég, ha tudjuk, hogy milyen intervallumban van (bizonyos valószínűséggel). Például, ha a várható értéket akarjuk megbecsülni:

$$P(l(X) \leq \mu \leq u(X)) = 1 - \alpha$$

Azaz a μ a $[l(X), u(X)]$ intervallumban van $1-\alpha$ valószínűséggel, akkor $[l(X), u(X)]$ a **konfidencia intervallum**, $1-\alpha$ pedig a **konfidenciaszint**.

Amikor csak a várható érték ismeretlen

Példa

Legyenek X_j-k független, azonos eloszlású normális változók, ismeretlen μ-vel és ismert σ-val. Ekkor $\overline{X_n}$ is normális μ várható értékkel és $\sigma/\sqrt{n}$ szórással. Mi lesz az $1-\alpha$ szinthez tartozó konfidenciaintervallum?

$$2\Phi(z)-1 = P\left(-z \leq \frac{\overline{X_n}-\mu}{\sigma/\sqrt{n}} \leq z\right) =$$

$$= P\left(\overline{X_n} - \frac{z\sigma}{\sqrt{n}} \leq \mu \leq \overline{X_n} + \frac{z\sigma}{\sqrt{n}}\right)$$

Legyen $z_{\alpha/2}$ amire $\alpha = 2-2\Phi(z_{\alpha/2})$, ekkor

$$\left(\overline{X_n} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}}, \overline{X_n} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}\right)$$

lesz a megfelelő konfidenciaintervallum.

Példa

Az X feszültség egy konstansból és zajból tevődik össze:

$$X = v + N$$

ahol $N \sim N(0,1)$. Találj 95%-os konfidenciaintervallumot v-re ha 100-szor volt megmérve a feszütség és a mintaátlag $5.25\mu V$!

$\alpha = 0,05$, $z_{\alpha/2} = 1.96$ Ekkor a keresett intervallum:

$$\left(5,25 - \frac{1,96 \cdot 1}{10}, 5,25 + \frac{1,96 \cdot 1}{10}\right)$$

$(5,06 ; 5,44)$

$1-\alpha = 0,95$

$\underline{\alpha = 0,05} = 2-2\Phi(z_{\frac{\alpha}{2}})$

$\Phi^{-1}(0,975) = 1,96$

Várható érték és szórás is ismeretlen

Legyenek X_j-k független, azonos eloszlású normális változók, ismeretlen μ-vel és ismeretlen σ-val. Mi lesz μ-nek az $1-\alpha$ szinthez tartozó konfidenciaintervalluma?

A szórást lecseréljük a tapasztalati szórásra:

$$\left(\overline{X_n} - \frac{t\hat{\sigma}}{\sqrt{n}}, \overline{X_n} + \frac{t\hat{\sigma}}{\sqrt{n}}\right)$$

általában nem normális változó

$$P\left(\overline{X_n} - \frac{t\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \overline{X_n} + \frac{t\hat{\sigma}}{\sqrt{n}}\right) = P\left(-t \leq \frac{\overline{X_n} - \mu}{\hat{\sigma}/\sqrt{n}} \leq t\right) =$$

$$= \underbrace{2T_{n-1}(t) - 1}_{\alpha}$$

T-eloszlású változó

Irodalomjegyzék

- Vetier András—Valószínűségszámítás
- Alberto Leon-Garcia—Probability, Statistics, and Random Processes for Electrical Engineering
- Sheldon M. Ross —Introduction To Probability and Statistics for Engineers and Scientists
- Dimitri P. Bertsekas and John N. Tsitsiklis —Introduction to Probability
- E. T. Jaynes —Probability Theory - The Logic of Science

Köszönöm a figyelmet!