Empiri

I verkligheten har vi n oberoende observationer $x_1, \dots, x_n$ av någon variabel x, som är det vi observerar när vi mäter en parameter θ.

Teori

I teorin har vi ett stickprov ("random sample") $X_1, \dots, X_n$ från X:s fördelning och denna har åtminstone en okänd parameter θ. Det kan finnas fler.

Fördelningen för X ska spegla de variationer vi ser i x då vi mäter θ.

Om observationerna är tider, är kanske $\mathrm{Exp}(\lambda)$ en bra modell.

Om observationerna är någon slags mätvärden är ofta $\mathrm{N}(\mu, \sigma)$ en bra modell. Man tänker sig då att

$$X = \mu + \sigma\epsilon \quad \text{där} \quad \epsilon \sim \mathrm{N}(0, 1)$$

Om man räknar ovanliga händelser i ett tidsintervall av längd t är ofta $\mathrm{Poi}(\lambda t)$ en bra modell. Ofta har man då bara en observation x, som alltså är antalet händelser i ett tidsintervall av längd t.

Om man vill skatta en okänd sannolikhet p ska man använda $\mathrm{Bin}(n, p)$-modellen. Här observerar man

$$x = \sum_{i=1}^{n} x_i$$

där $x_1, \dots, x_n$ är oberoende $\mathrm{Ber}(p)$-observationer.

Tre exempel

I syfte att estimera tillförlitligheten av 16-kbit dynamiska RAM av ett visst fabrikat, valde man slumpmässigt ut 100 st som testades under 1000 timmars kontinuerlig drift. Man fann att $x = 91$ av de 100 testobjekten fungerade felfritt under hela testtiden. Teoretisk modell för detta försök är $X \sim \mathrm{Bin}(n, p)$, där $n = 100$ och p är sannolikheten att ett RAM av den testade typen fungerar felfritt de första 1000 driftstimmarna.

Man har i 4 lika stora vattenprov nedströms en avfallsanläggning räknat antalet kolibakterier och fått

$$x_1 = 12, \; x_2 = 15, \; x_3 = 16, \; x_4 = 17$$

Teoretisk modell för detta försök skulle kunna vara $X \sim \mathrm{Poi}(\lambda)$.

Man har i 5 oberoende jordprov i en gammal industritomt mätt kromhalten i mg/kg och fått

$$x_1 = 237.5, \; x_2 = 201.9, \; x_3 = 253.6, \; x_4 = 258.3, \; x_5 = 216.9$$

Teoretisk modell för detta försök skulle kunna vara $X \sim \mathrm{N}(\mu, \sigma)$. Även $\ln X \sim \mathrm{N}(\mu, \sigma)$ skulle kunna vara en bra modell.

Ytterligare ett exempel

Man har under 1350 timmar (drygt 56 dygn) studerat belastningen y_t av ett visst tekniskt system. Under denna tidsrymd gick belastningen över en kritisk nivå $u = 5$ sammanlagt $n = 8$ gånger. Maximala nivån x_i under dessa 8 överskridanden var

$$6.85 \;\; 5.11 \;\; 5.61 \;\; 5.28 \;\; 6.87 \;\; 6.40 \;\; 5.16 \;\; 5.12$$

Teoretisk modell för detta försök skulle kunna vara att överskridandena sker enl en Poissonprocess N med intensitet λ [st/dygn] och att deras resp maximala nivåer X är Paretofördelade, vilket innebär att tätheten är

$$f(x) = \frac{\alpha}{u}\left(\frac{x}{u}\right)^{-(\alpha+1)} \quad \text{för} \quad x > u$$

för något $\alpha > 0$.

Exempel: Estimator och estimat

Vi skall använda det observerade (empiriska) medelvärdet $\bar{x}$ till att *estimera* parametern $\mu = E[X]$. I vardagligt tal säger man ofta att man *skattar* μ. Vi kallar $\bar{x}$ för μ:s *estimat*.

Att $\bar{x}$ är μ:s estimat skrives $\widehat{\mu} = \bar{x}$.

Motsvarande stokastiska variabel

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

kallas μ:s *estimator*.

Att $\bar{X}$ är vår estimator av μ skrives $\widehat{\mu} = \bar{X}$.

Estimatet är alltså det uträknade (observerade) värdet, medan estimatorn är en stokastisk variabel.

I svenskan använder man ofta ordet *skattning* och sammanhanget avgör om man avser den stokastiska variabeln eller det observerade värdet.

Definition: Estimator and estimate

A statistic used to approximate or estimate a population parameter θ is called a *(point) estimator* for θ and is denoted by $\widehat{\theta}$ or $\widetilde{\theta}$ or θ^* or $\ldots$; the numerical value assumed by this statistic when evaluated for a given sample is called a *(point) estimate* for θ.

Att statistikan $\widehat{\theta}$, som vi använder som estimator av θ, beror av stickprovet $X_1, \ldots, X_n$ brukar man inte skriva ut explicit.

Exempel

Antag att vi vid n oberoende mätningar av en viss anläggnings driftstid (tiden från start till nästa produktionsstopp) erhöll medelvärdet $\bar{x} = 732.9$ timmar. Låt μ beteckna den förväntade (teoretiska) driftstiden. Då är 732.9 timmar vårt estimat av μ och vi skriver

$$\widehat{\mu} = 732.9$$

Definition 7.1.1: Väntevärdesriktighet

En estimator $\widehat{\theta}$ sägs vara väntevärdesriktig för θ, om

$$E[\widehat{\theta}] = \theta$$

Det är önskvärt att en estimator $\widehat{\theta}$ har liten varians. Ju mindre $\mathrm{Var}[\widehat{\theta}]$ är, desto bättre är estimatorn.

Theorem 7.1.1, 7.1.2

Stickprovsmedelvärdet $\bar{X}$ är en väntevärdesriktig estimator av μ. Dess varians är

$$\mathrm{Var}[\bar{X}] = \frac{\sigma^2}{n}$$

Definition 7.1.2: Standard error of the mean

The standard deviation of $\bar{X}$ is given by $\sigma/\sqrt{n}$ and is called the standard error of μ.

Theorem 7.1.3

Stickprovsvariansen S^2 är en väntevärdesriktig estimator av σ^2.

I exemplet med överskridandedata (se OH no 3) förekommer en parameter α. Vi ska nu med två olika metoder härleda en skattning av α.

A: Momentmetoden

Vi börjar med att m.h.a. X.s täthet

$$f(x) = \frac{\alpha}{u}\left(\frac{x}{u}\right)^{-(\alpha+1)} \quad \text{för} \quad x > u$$

beräkna

$$\mu = E[X] = \int_u^\infty x \frac{\alpha}{u}\left(\frac{x}{u}\right)^{-(\alpha+1)} dx = \cdots = \frac{\alpha}{\alpha-1}u$$

Denna beräkning förutsätter att $\alpha > 1$, så det som följer är bara giltligt om så är fallet. Inverteras likheten

$$\mu = \frac{\alpha}{\alpha-1}u$$

får vi

$$\alpha = \frac{\mu}{\mu-u}$$

Således är momentestimatet (momentskattningen) av α lika med

$$\widehat{\alpha} = \frac{\bar{x}}{\bar{x}-u} = \frac{5.8}{5.8-5.0} = 7.25$$

Momentestimatorn av α är

$$\widehat{\alpha} = \frac{\widehat{\mu}}{\widehat{\mu}-u} = \frac{\bar{X}}{\bar{X}-u}$$

B: Trolighetsmetoden

Vi börjar med att konstatera att täthetens värde i punkten x är ett mått på hurpass troligt det är att vi observerar just värdet x. Att vi observerade värdena $x_1, \ldots, x_n$ i n oberoende försök blir då så här troligt:

$$\prod_{i=1}^n f(x_i) = \frac{\alpha}{u}\left(\frac{x_1}{u}\right)^{-(\alpha+1)} \cdot \ldots \cdot \frac{\alpha}{u}\left(\frac{x_n}{u}\right)^{-(\alpha+1)}$$

För givna observationer $x_1, \ldots, x_n$ är detta en funktion av parametern α som vi kallar trolighetsfunktionen eller bara troligheten och betecknar $L(\alpha)$. Här ser vi att

$$L(\alpha) = \left(\frac{\alpha}{u}\right)^n \left(\prod_{i=1}^n \frac{x_n}{u}\right)^{-(\alpha+1)}$$

Trolighetsmetodens skattning av α är

$$\widehat{\alpha} = \underset{\alpha}{\operatorname{argmax}}\, L(\alpha)$$

Lösningen av detta maximeringsproblem är

$$\widehat{\alpha} = \frac{n}{\sum_{i=1}^n \ln(x_i/u)}$$

Trolighetsestimatorn är alltså

$$\widehat{\alpha} = \frac{n}{\sum_{i=1}^n \ln(X_i/u)} = \frac{1}{\frac{1}{n}\sum_{i=1}^n \ln X_i - \ln u}$$

och trolighetsskattningen (trolighetsestimatet) är

$$\widehat{\alpha} = \frac{1}{0.1407} \approx 7.105$$

Således:

Antag att vi har k parametrar $\theta_1, \ldots, \theta_k$ i den aktuella fördelningen för X (det som vi observerar eller mäter).
Ex 1: $k = 1$ och $\theta_1 = \lambda$ om $X \sim \mathrm{Exp}(\lambda)$
Ex 2: $k = 2$ och $\theta_1 = \mu, \theta_2 = \sigma$ om $X \sim \mathrm{N}(\mu, \sigma)$
Vår uppgift är att hitta en skattning av $\theta = g(\theta_1, \ldots, \theta_k)$.

Momentmetoden bygger på identiteten

$$\theta = h(m_1, m_2, \ldots, m_k)$$

där $m_i = E[X^i]$ och momentskattningen av θ är

$$\widehat{\theta} = h(\widehat{m}_1, \widehat{m}_2, \ldots, \widehat{m}_k)$$

där

$$\widehat{m}_j = \frac{1}{n}\sum_{i=1}^{n} x_i^j$$

Trolighetsmetoden maximerar trolighetsfunktionen

$$L(\theta_1, \ldots, \theta_k) = \prod_{i=1}^{n} f(x_i)$$

och trolighetsskattningen av $\theta_1, \ldots, \theta_k$ är

$$\widehat{\theta}_1, \ldots, \widehat{\theta}_k = \underset{\theta_1, \ldots, \theta_k}{\mathrm{argmax}}\, L(\theta_1, \ldots, \theta_k)$$

Trolighetsskattningen av $\theta = g(\theta_1, \ldots, \theta_k)$ är $\widehat{\theta} = g(\widehat{\theta}_1, \ldots, \widehat{\theta}_k)$.

PS Använd i första hand trolighetsmetoden. Om detta ej går, härled din skattning på något annat sätt, t.ex med momentmetoden.

Exempel

Ur $n = 7$ mätningar $x_1, \ldots, x_n$ har vi erhållit $\bar{x} = 7.4$, $s^2 = 4.50$ ($s \approx 2.1213$). Vi ska skatta 95%-kvantilen $x_{0.95}$ och har kommit fram till att $X \sim \mathrm{N}(\mu, \sigma)$ är en vettig modell. Ur definitionen $F(x_{0.95}) = 0.95$ härleder vi $x_{0.95} = \mu + 1.645\sigma$.

<u>Momentmetoden:</u> Ur $m_1 = \mu$, $m_2 = \mu^2 + \sigma^2$ får vi $\mu = m_1$, $\sigma = \sqrt{m_2 - m_1^2}$. Således gäller $x_{0.95} = m_1 + 1.645\sqrt{m_2 - m_1^2}$.
Momentskattningen av 95%-kvantilen är alltså

$$\widehat{x}_{0.95} = \widehat{m}_1 + 1.645\sqrt{\widehat{m}_2 - \widehat{m}_1^2}$$
$$= 7.4 + 1.645\sqrt{58.617 - 7.4^2} = 10.63$$

ty $\widehat{m}_1 = (1/n)\sum_i x_i = \bar{x} = 7.4$, $\widehat{m}_2 = (1/n)\sum_i x_i^2$ $= (n-1)s^2/n - \bar{x}^2 = 58.617$.

<u>Trolighetsmetoden:</u> Maximering av troligheten

$$L(\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x_i-\mu)^2/2\sigma^2} = (2\pi)^{-n/2}\sigma^{-n} e^{-(1/2)\sum_i (x_i-\mu)^2/\sigma^2}$$

ger $\widehat{\mu} = \bar{x} = 7.4$, $\widehat{\sigma} = \sqrt{\frac{1}{n}\sum_i (x_i - \bar{x})^2} = \sqrt{\frac{n-1}{n}}\, s = 1.964$.
Trolighetsskattningen av 95%-kvantilen är alltså

$$\widehat{x}_{0.95} = \widehat{\mu} + 1.645\widehat{\sigma} = 7.4 + 1.645 \cdot 1.964 = 10.63$$

Fast i praktiken använder man gärna den naturliga skattningen

$$\widehat{x}_{0.95} = \bar{x} + 1.645 s = 7.4 + 1.645 \cdot 2.1213 = 10.89$$

Theorem 7.3.1

Antag att X och Y är stokastiska variabler och med m.g.f $m_X(t)$ och $m_Y(t)$. Om $m_X(t) = m_Y(t)$ för alla t i en öppen omgivning till $t = 0$, så har X och Y samma fördelning.

Theorem 7.3.2

Låt X_1 och X_2 vara två oberoende stokastiska variabler med m.g.f $m_{X_1}(t)$ och $m_{X_2}(t)$. Låt $Y = X_1 + X_2$. Då ges Y:s m.g.f av

$$m_Y(t) = m_{X_1}(t) m_{X_2}(t)$$

Theorem 7.3.3

Låt X vara en stokastisk variabel med m.g.f $m_X(t)$. Låt $Y = \alpha + \beta X$. Då ges Y:s m.g.f av

$$m_Y(t) = e^{\alpha t} m_X(\beta t)$$

Theorem 7.3.4: Distribution of $\bar{X}$—normal population

Let $X_1, \ldots, X_n$ be a random sample of size n from a normal distribution with mean μ and standard deviation σ. Then $\bar{X}$ is normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$.

I exemplet med jordproverna (se OH no 2) är $\bar{x} = 233.64$ vårt estimat av väntevärdet μ. Antag att vi bestämt oss för modellen $X \sim \mathrm{N}(\mu, \sigma)$. Estimatorn $\bar{X}$ är enl. sats 7.3.4 $\mathrm{N}(\mu, \sigma/\sqrt{n})$, där $n = 5$ är antalet observationer. Observera att för att teorin ska fungera är det väsentligt att jordproverna är oberoende.

Notera nu att

$$P\left[-1.96 \le \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \le 1.96\right] = 0.95$$

Men olikheten

$$-1.96 \le \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \le 1.96$$

är ekvivalent med

$$\bar{X} - 1.96\sigma/\sqrt{n} \le \mu \le \bar{X} + 1.96\sigma/\sqrt{n}$$

Således gäller

$$P\left[\bar{X} - 1.96\sigma/\sqrt{n} \le \mu \le \bar{X} + 1.96\sigma/\sqrt{n}\right] = 0.95$$

Anta att det är känt att $\sigma = 30$ för den här typen av analyser. Då är $1.96\sigma/\sqrt{n} = 26.30$ och vi får intervallestimatet

$$(233.64 - 26.30 =) \quad 207.34 \le \mu \le 259.94 \quad (= 233.64 + 26.30)$$

Intervallet $[207.34, 259.94]$ är ett konfidensintervall för μ med konfidensgraden 95%.

Ofta skriver man

$$\mu = 233.64 \pm 26.30$$

Definition 7.4.1: Confidence interval

A $100(1-\alpha)\%$ confidence interval for a parameter θ is a random interval $[L_1, L_2]$ such that

$$P[L_1 \le \theta \le L_2] = 1-\alpha$$

regardless of the value of θ.

To construct a $100(1-\alpha)\%$ confidence interval for a parameter θ, we shall find a random variable whose expression involves θ and whose probability distribution is known (at least approximately).

Theorem 7.4.1: $100(1-\alpha)\%$ confidence interval on μ (σ known)

Let $X_1, \ldots, X_n$ be a random sample of size n from a normal distribution with mean μ and standard deviation σ. A $100(1-\alpha)\%$ confidence interval on μ when σ is known, is given by

$$\mu = \bar{X} \pm z_{\alpha/2}\sigma/\sqrt{n}$$

where the quantile $z_{\alpha/2}$ is determined by the requirement

$$P[Z > z_{\alpha/2}] = \alpha/2$$

for $Z \sim \mathrm{N}(0, 1)$.

Ibland vill man ha ett ensidigt konfidensintervall av typen $\theta < L$.

I normalfördelningsfallet utgår man då ifrån

$$P\left[-z_\alpha < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right] = 1-\alpha$$

och härleder

$$P\left[\mu < \bar{X} + z_\alpha\sigma/\sqrt{n}\right] = 1-\alpha$$

I jordprovsexemplet (se OH no 2 och 12) får vi (med $\sigma = 30$) om $\alpha = 0.05$ att $z_\alpha = 1.645$ vilket ger $z_\alpha\sigma/\sqrt{n} = 22.07$ och intervallestimatet med konfidens 95% blir

$$\mu \le 255.71 \quad (= 233.64 + 22.07)$$

J.f.r med

$$\mu = 233.64 \pm 26.30 \quad \Leftrightarrow \quad 207.34 \le \mu \le 259.94$$

Theorem 7.4.2: Central Limit Theorem

Let $X_1, \ldots, X_n$ be a random sample of size n from a distribution with mean μ and standard deviation σ. Then for large n, $\bar{X}$ is approximately normal with mean μ and standard deviation $\sigma/\sqrt{n}$. Hence for large n, the standardized random variable

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

is approximately standard normal.

Men, då n är stort, gäller även att $s \approx \sigma$ och följaktligen har vi då också att

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \overset{\text{ap}}{\approx} \mathrm{N}(0, 1)$$

Denna approximation är naturligtvis sämre än den i sats 7.4.2.

Normalapproximation av $\mathrm{Bin}(n, p)$

Här är $X_1, \ldots, X_n$ ett stickprov på $\mathrm{Ber}(p)$.

Då är $\mu = p$ och $\sigma^2 = p(1-p)$.

Enl c.g.s är $\bar{X}$ approximativt $\mathrm{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$ då n är stort.

Men då är också $S = \sum_{i=1}^{n} X_i$ approximativt $\mathrm{N}(np, \sqrt{np(1-p)})$.

Notera också att $S \sim \mathrm{Bin}(n, p)$.

Således gäller att $\mathrm{Bin}(n, p) \approx \mathrm{N}(np, \sqrt{np(1-p)})$.

Hur stort behöver n vara för att approximationen ska fungera?

Milton & Arnold, sida 121 (avs 4.6): ”For most practical purposes the approximation is acceptable for values of n and p such that either $p \le 0.5$ and $np > 5$ or $p > 0.5$ and $n(1-p) > 5$.”

M.a.o är $n \min(p, 1-p) > 5$ en lämplig regel.

Se exempel 4.6.1.