

Citat ur Devore & Farnum, s 271.

The general objective of statistical inference is to use sample information as a basis for drawing various types of conclusions. In an estimation problem, we want to make an educated guess about the value of some population characteristic or parameter, such as the population mean battery lifetime μ , the proportion π of all components of a certain type that need service while under warranty, or the difference $\mu_1 - \mu_2$ between the population mean lifetimes for two different types of batteries. The simplest type of estimate is a point estimate, a single number that represents our best guess for the value of the parameter. Thus we might report a point estimate of 758 hours for the population mean lifetime of all brand X 100-watt lightbulbs; we are not saying that $\mu = 758$, only that sample data suggests 758 as a very plausible value for μ .

Vi skriver $\hat{\mu} = 758$ för att markera att 758 är vår (punkt-)skattning av parametern μ .

Exempel 1 Vi mäter en teknisk storhet (hållfasthet, sträckgräns, e dyl) som vi just nu kan kalla μ , och antar att variabeln vi observerar, x , är normalfördelad med parametrar μ och okänd standardavvikelse σ . Detta innebär att vi tänker oss att

$$x = \mu + \epsilon$$

där ϵ är felet i en enskild mätning.

Exempel 2 Vi vill bestämma en viss mätutrustnings noggrannhet, så vi mäter på en "normal" som vi vet är μ enheter stor. Vi skulle kunna anta att variabeln vi observerar, x , är normalfördelad med parametrar μ och σ .

Exempel 3 Vi vill bestämma olycksfrekvensen för en viss typ av vägavsnitt i Sverige. Den här typen av vägar är relativt vanlig, så det kostar för mycket att analysera alla. Därför väljer man ut n st och bestämmer för var och en av dessa frekvensen olyckor under mars månad ett visst år. Här kan det vara rimligt att tänka sig att variabeln vi mäter, x , är Poissonfördelad med parameter λ (enhet: olyckor/månad). Uppgiften är att bestämma λ med så stor noggrannhet som möjligt.

Exempel 4 Vi vill bestämma värdet av sannolikheten $p = P(A)$ att en viss händelse A inträffar då vi gör ett slumpmässigt försök. Vi utför därför n oberoende upprepningar av försöket och låter $x_i = 1$ om A inträffar i det i te försöket och 0 annars.

Definition 1 (s 272) En punktskattning av en parameter θ är ett reellt tal $\hat{\theta}$, beräknat ur data x_1, \dots, x_n , som vi kan använda som en (mer eller mindre begävad) gissning av θ 's värde.

En punktskattning $\hat{\theta}$ är alltså en funktion av observationerna. Detta betyder att den är en stokastisk variabel. Man brukar kalla sådana stokastiska variabler som beror av ett stickprov för *stickprovsvariabler*.

Exempel på punktskattningar är

- medelvärdet \bar{x}
- medianen m
- en percentil, vilken som helst
- variansen s^2 och standardavvikelsen s

Definition 2 (s 274) En punktskattning $\hat{\theta}$ av θ är väntevärdesriktig ("unbiased") om

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

Definition 3 (s 275) En punktskattning $\hat{\theta}$ av θ är konsistent, om

$$P(|\hat{\theta} - \theta| > \epsilon) \rightarrow 0 \text{ då } n \rightarrow \infty$$

Man kan visa att punktskattningen $\hat{\theta}$ av θ är konsistent, om dess varians går mot noll då antalet observationer går mot oändligheten, d v s om

$$\text{Var}[\hat{\theta}] \rightarrow 0 \text{ då } n \rightarrow \infty$$

Exempel 5 Medelvärden \bar{x} är en väntevärdesriktig och konsistent skattning av väntevärdet μ , ty

$$E[\bar{X}] = \mu$$

och

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

Exempel 6 Stickprovsvariansen S^2 är en skattning av variansen σ^2 och S är en skattning av σ . Senare i kursen ska vi se S^2 är väntevärdesriktig, och vi noterar redan nu att S ej är det.

Hemövning 1 Anta att snötäckets utbredning (j.f.r övn 5, s 68) är lognormalfördelad med parametrar μ och σ . Skatta dessa parametrar m.h.a data i övn 5, s 68.

Sida 278:

Konfidensintervall för μ med approximativ konfidensgrad 95%

$$\mu = \bar{x} \pm 1.96 s / \sqrt{n}$$

Regeln förutsätter att $n \geq 30$.

Sida 281:

Konfidensintervall för μ med approximativ konfidensgrad 90%, 95% eller 99%

$$\mu = \bar{x} \pm z_c s / \sqrt{n}$$

där $z_c = 1.645, 1.96$ eller 2.575 beroende på om konfidensgraden är 90%, 95% eller 99%. Regeln förutsätter att $n \geq 30$.

Sida 286:

Konfidensintervall för p med approximativ konfidensgrad 90%, 95% eller 99%

$$p = \hat{p} \pm z_c \sqrt{\hat{p}(1-\hat{p})/n}$$

där $\hat{p} = f/n$ är relativa frekvensen och $z_c = 1.645, 1.96$ eller 2.575 beroende på om konfidensgraden är 90%, 95% eller 99%. Regeln ska endast användas då man är hyfsat övertygad om att både $np \geq 5$ och att $n(1-p) \geq 5$ gäller.

Exempel 7 I en undersökning av hur gamla V-teknologer är när de avlägger examen erhöles man i ett stickprov medelvärden $\bar{x} = 28.2$ år. Standardavvikelsen uppmättes till $s = 3.37$ och antalet observationer var $n = 35$. Punkt- och intervallskatta den förväntade åldern då examen avlägges.

Exempel 8 I en enkätundersökning i syfte att förutsäga svaret på en fråga i en folkomröstning fann man att $f = 469$ utav de $n = 1037$ som svarade på enkäten, besvarade enkätfrågan med "JA". Totalt tillfrågades 1100 väljare. Punkt- och intervallskatta utgången av folkomröstningen. Önskad konfidensgrad: 99%.

Hemövning 2 Man testade 25 vajrar m.a.p draghållfasthet och beräknade deras resp sträckgräns. Därvid erhöles $\bar{x} = 25.3$ och $s = 1.72$. Punkt- och intervallskatta förväntad sträckgräns. Bortse från att det hade varit önskvärdt med ytterligare ca 5 mätningar. Önskad konfidensgrad: 95%.

Hemövning 3 Laboratorieassistent Emil Vson kontrollerade en viss tillverkningsprocess dagligen kl 09.00. Under 100 på varandra följande dagar godkändes processen i 87 dagar. De övriga dagarna fann man att det var nödvändigt att justera diverse inställningar. Antag att resultaten av Emils test är oberoende av varandra. Skatta sannolikheten p att processen behöver justeras en viss godtycklig dag. Önskad konfidensgrad i skattningen ska vara 90%.

Antag nu att vi har två stickprov och är intresserade utav differensen av deras respektive väntevärden.

Stickprov nr 1 består av n_1 mätningar av en variabel med väntevärde μ_1 och standardavvikelse σ_1 . Stickprovets medelvärde betecknas \bar{x}_1 och dess standardavvikelse s_1 . Analogt beteckningar har vi för stickprov nr 2.

Sida 289:

Fakta om fördelningen för en differens av medelvärden

- $E[\bar{X}_1 - \bar{X}_2] = \mu_1 - \mu_2$
- $\text{Var}[\bar{X}_1 - \bar{X}_2] = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$
- $\bar{X}_1 - \bar{X}_2$ är normalfördelad om de två stickproven består av normalfördelade observationer
- $\bar{X}_1 - \bar{X}_2$ är approximativt normalfördelad om stickprovsstorlekarna n_1, n_2 ej är små

Sida 290:

Konfidensintervall för en differens av väntevärden

$$\mu_1 - \mu_2 = \bar{x}_1 - \bar{x}_2 \pm z_c \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

där $z_c = 1.645, 1.96$ eller 2.575 beroende på om konfidensgraden är approximativt 90%, 95% eller 99%.

Glöm inte av att stickprovsstorlekarna ej får vara små. Då fungerar inte normalapproximationen så bra.

Hemövning 4 Man var inte nöjd med kvaliteten hos de vajrar som testades i en tidigare hemuppgift ($n = 25$, $\bar{x} = 25.3$, $s = 1.72$), så man gjorde en del förändringar i tillverkningen av dem. Innan produktionen startas på nytt är det viktigt att veta huruvida förändringarna förbättrade kvaliteten eller ej. Man testade därför 20 vajrar på samma sätt som tidigare och beräknade deras resp sträckgränser. Därvid erhöles $\bar{x} = 26.1$ och $s = 1.58$. Skatta förändringen. Önskad konfidsgrad: 90%.

Sats 1 (s 294) Om vårt stickprov består av normalfördelade observationer, så är det standardiserade medelvärdet,

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

exakt normalfördelat med parametrar 0 och 1. Om man i det byter den teoretiska standardavvikelsen σ mot den uppmätta s , och får

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

har vi istället en t -fördelad variabel med $n - 1$ frihetsgrader.

Obs. att för fixt $x > 0$ gäller att

$$P(-x \leq Z \leq x) < P(-x \leq T \leq x) \searrow P(-x \leq Z \leq x)$$

då antalet observationer $n \rightarrow \infty$. Härur följer att om z_α och $t_\alpha(n-1)$ uppfyller $P(Z > z_\alpha) = P(T > t_\alpha(n-1)) = \alpha$, så gäller att

$$z_\alpha < t_\alpha(n-1) \searrow z_\alpha$$

då $n \rightarrow \infty$.

Sida 296:

Konfidensintervall för μ med exakt konfidsgrad $1 - \alpha$

$$\mu = \bar{x} \pm t_{\alpha/2} s / \sqrt{n}$$

där $t_{\alpha/2}$ uppfyller $P(T > t_{\alpha/2}) = \alpha/2$ för $T \sim t(n-1)$. Här förutsätts att stickprovet består av $N(\mu, \sigma)$ -fördelade observationer.

Hemövning 5 Skatta parametern μ i lognormalmodellen av snötäckets utbredning i oktober.

Sats 2 (s 298) Antag att \bar{x} , s vara medelvärde och standardavvikelse av n oberoende observationer x_1, \dots, x_n av $X \sim N(\mu, \sigma)$. Låt X_{n+1} vara nästa ännu ej sedda observation. Då

$$\frac{X_{n+1} - \bar{x}}{s\sqrt{1+1/n}} \mid x_1, \dots, x_n \sim t(n-1)$$

Sida 298:

Prediktionsintervall Givet att de n första observationerna har medelvärdet \bar{x} och standardavvikelsen s , gäller för observation nr $n+1$, att

$$P\left(-t_{\alpha/2} \leq \frac{X_{n+1} - \bar{x}}{s\sqrt{1+1/n}} \leq t_{\alpha/2}\right) = 1 - \alpha$$

Hemövning 6 Prediktera snötäckets utbredning i oktober 1980. Önskad konfidens (prediktionsgrad) är 99%.

Sida 303:

Tvåstickprovsvalet

Sats 3 Låt \bar{x}_1, s_1 och \bar{x}_2, s_2 vara medelvärde och standardavvikelse av två oberoende stickprov av storlekarna n_1 resp. n_2 av $X_1 \sim N(\mu_1, \sigma_1)$ och $X_2 \sim N(\mu_2, \sigma_2)$. Då gäller att

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \approx t(\nu)$$

där

$$\nu \approx \frac{s_1^2/n_1 + s_2^2/n_2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 2}}$$

(avrunda nedåt). Härur följer att

$$\mu_1 - \mu_2 = \bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2} \sqrt{s_1^2/n_1 + s_2^2/n_2}$$

är ett konfidensintervall för differensen $\mu_1 - \mu_2$ med den approximativa konfidsgraden $1 - \alpha$.

Sida 305:

Parade (bivariata) data

Låt \bar{d} , s_d beteckna medelvärde och standardavvikelse för ett stickprov bestående av n N-fördelade differenser $D = X - Y$. Då är

$$\mu_d = \bar{d} \pm t_{\alpha/2} s_d / \sqrt{n}$$

ett konfidensintervall för $\mu_d = \mu_X - \mu_Y$ med konfidensgraden $1 - \alpha$. Antalet frihetsgrader i $t_{\alpha/2}$ är $n - 1$.

Exempel 7.16 (s 311) 12 st livstider

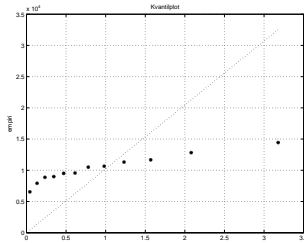
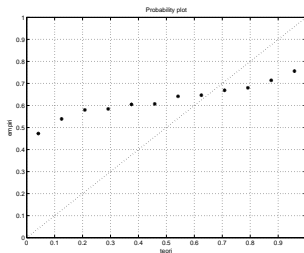
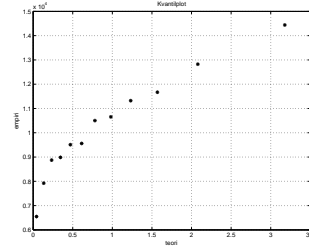
10 502 9 560 11 671 12 825 8 987 7 924
9 508 8 875 14 439 11 320 6 549 10 654

Modell: exponentialfördelning med parameter λ .

Problem: Skatta λ .

Lösning av problemet: Vi har $n = 12$ observationer. Deras medelvärde är $\bar{l} = 10\,234.5$, så ML-skattningen av λ blir $\hat{\lambda} = 1/10\,234.5 \approx 9.771 \times 10^{-5}$.

Kontroll av modellen:

**ML-metoden**

Du har n oberoende observationer x_1, \dots, x_n av en stokastisk variabel X , vars täthet eller massfunktion är $f(x|\theta)$ (θ är en en- eller fler-dimensionell parameter och vår uppgift är att skatta $\vartheta = h(\theta)$).

1) Bilda trolighetsfunktionen

$$L(\theta) = \prod_i f(x_i|\theta)$$

2) Maximera denna, d.v.s. sök

$$\hat{\theta} = \arg \max_{\theta} L(\theta)$$

3) ML-skattningen av ϑ är

$$\hat{\vartheta} = h(\hat{\theta})$$

4) Tips: Låt $\mathcal{L}(\theta) = \ln L(\theta)$. Då gäller att

$$\hat{\theta} = \arg \max_{\theta} \mathcal{L}(\theta)$$

Det är i allmänhet betydligt enklare att maximera $\mathcal{L}(\theta)$.

5) Glöm inte att kontrollera modellen, t.ex. genom att plotta de empiriska kvantilerna mot de teoretiska.

ML-skattningens egenskaper (s 313)

- For large n , the sampling distribution of an MLE is approximately normal, and the estimator is nearly unbiased with a variance smaller than any other estimator.

Exempel 9 Om det vi mäter, x , är Bernoullifördelat med parameter p , så är ML-skattningen av p , lika med $\hat{p} = \sum_i x_i/n = \bar{x}$.

Nyckelord i kursdelen

Punktskattning

Väntevärdesriktighet

Konsistens

Konfidensintervall, konfidensgrad, intervallskattning

Prediktionsintervall, -grad

ML-metoden (trolighetsmetoden)