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 $\widehat{\mu} = 758$ 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 noggranhet, 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 noggranhet 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 $x_1, \ldots, x_n$, som vi kan använda som en (mer eller mindre begåvad) 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

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

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

Här tänker man sig att man, för varje n, har ett stickprov av storleken n, ur vilket man beräknar skattningen $\hat{\theta}$.

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

$$\mathrm{Var}[\hat{\theta}] \to 0 \quad \text{då} \quad n \to \infty$$

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

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

och

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

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.

Faktarutor:

(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.

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. Analoga beteckningar har vi för stickprov nr 2.

(Sida 289) **Fakta om fördelningen för en differens av medelvärden**

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

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

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

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

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

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 $n \to \infty$. 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 $n \to \infty$.

(sida 296) **Konfidensintervall för μ med exakt konfidensgrad $1-\alpha$**

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

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

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

$$\left.\frac{X_{n+1} - \bar{x}}{s\sqrt{1 + 1/n}}\right| x_1, \ldots, x_n \sim \mathrm{t}(n-1)$$

(Sida 298) **Prediktionsintervall** Givet att de n första observationerna är $x_1, \ldots, x_n$, gäller

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

(Sida 303) **Tvåstickprovsfallet**

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 $X_1 \sim \mathrm{N}(\mu_1, \sigma_1)$ och $X_2 \sim \mathrm{N}(\mu_2, \sigma_2)$. Då gäller att

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

där

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

(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 konfidensgraden $1 - \alpha$.

(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

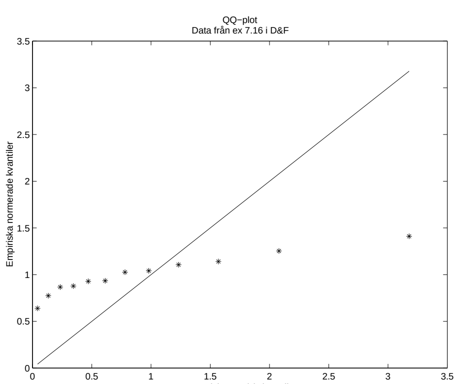
10 502	9 560	11 671	12 825	8 987	7 924
9 508	8 875	14 439	11 320	6 549	10 654

Modell: exponentialfördelning med parameter λ.

Problem: Skatta λ.

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

Kontroll av modellen: Vi plottar det ordnade stickprovet normaliserat mot motsv. kvantiler i Exp(1)-fördelningen och erhåller



ML-metoden

Du har n oberoende observationer $x_1, \ldots, x_n$ 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

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

3) ML-skattningen av ϑ är

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

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

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

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