

Matematisk Statistik

7. Forelæsning 23.02.2021

- ☐ Anvendelser af CLT
- ☐ Bootstrap metoden
- ☐ (hvis der er tid: start med estimation i parametrisk statistik)

- ▶ En **stikprøvefunktion** T er en funktion på stikprøver $\mathbf{x} = (x_1, \dots, x_n)$.

Eksempel:

- ▶ T gennemsnit:

$$T(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n x_i =: \bar{x}.$$

- ▶ T stikprøvevarians:

$$T(\mathbf{x}) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

- ▶ Ved anvendelse på tilfældige stikprøver $\mathbf{X} = (X_1, \dots, X_n)$ får vi en **stokastisk variabel** $T = T(\mathbf{X})$.
- ▶ Fordelingen af $S(\mathbf{X})$ kaldes **sampling distribution** (stikprøvefordeling), og standardafvigelsen kaldes **standard error**, eller **middelfejl**.
- ▶ Stikprøvefordelingen **afhænger** som regel **af populationen**, og af måden vi samler på.
- ▶ I nogle tilfælde hvor populationen er givet ved en fordelingsfunktion kan vi beregne stikprøvefordelingen. Eksempel: sum af Poisson fordelte variabler.
- ▶ Stikprøvefordelingen af gennemsnittet følger *som regel* den **centrale grænseværdisætning**.

Formål:

Vi vil gerne beregne sandsynligheder vedrørende **stikprøvegennemsnit** \bar{X} eller **-sum** S fra en stikprøve af n *ikke* normalfordelte stokastiske variable X_1, \dots, X_n .

To mulige situationer:

- ▶ Vi kender middelværdi $\mu = E[X]$ og varians: $\sigma^2 = \text{Var}[X]$, eller
- ▶ Middelværdi og varians estimeres fra en stikprøve.

Princippet:

- ▶ Brug henholdsvis transformationen

$$Z = \frac{\bar{X} - \mu}{\sqrt{\sigma/n}} \quad \text{eller} \quad Z = \frac{S - n\mu}{\sqrt{n\sigma^2}}$$

og udtryk hændelserne ved hjælp af Z frem for X eller S .

- ▶ Beregn sandsynlighederne ved at antage, at Z har en standard normalfordeling, $N(0, 1)$.

Vi kigger på et simulationsexperiment:

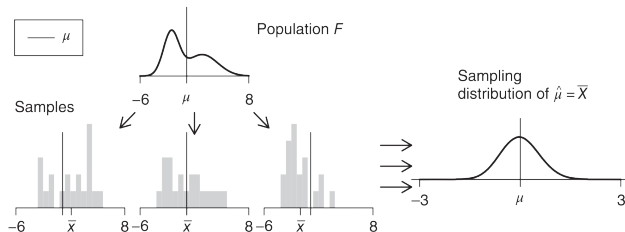
estimerer middelværdi og median af exponentialfordelingen ved gennemsnit og stikprøvemedian.

Hvor godt passer normalapproximation til stikprøvefordelingen?

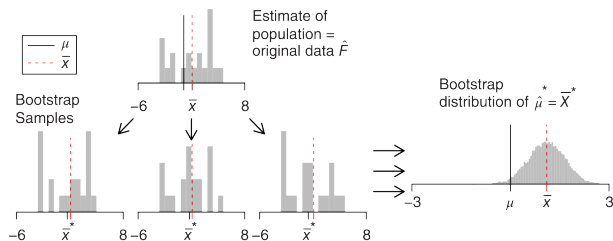
Idé: Erstat den ukendte fordelingsfunktion F i populationen med ecdf \hat{F} fra observationen.

Beregn (simuler) sampling fordeling ved at antage at \hat{F} er den sande fordeling.

Sampling distribution af $\hat{\mu} = \bar{X}$



Bootstrapfordeling af $\hat{\mu} = \bar{X}$



Hvordan simuleres fra bootstrapfordelingen?

Lad x indeholde data vektor (x_1, \dots, x_n) fra stikprøven

☞ `R sample(x, replace = TRUE)` giver bootstrap sample

sammenlign permutationsfordeling: `sample(x, replace = FALSE)`.

For de fleste estimatorer og populationer fanger bootstrapfordelingen lavet fra én observerede stikprøve de følgende egenskaber af stikprøvefordelingen:

Position	✗	$E[\hat{\theta}^*] \neq \theta$
Bias	✓	$E[\hat{\theta}^*] - \theta^* \approx E\hat{\theta} - \theta,$ hvor θ^* er den tilsvarende parameter af X^* 's fordeling, oftest $\theta^* = \hat{\theta}_{\text{obs}}$
Spredning	✓	$\widehat{SE}[\hat{\theta}^*] \approx SE[\hat{\theta}]$
Skævhed	✓	Skævhed af $\hat{\theta}^*$'s fordeling \approx skævhed af $\hat{\theta}$'s fordeling.

Disse påstande kan vises under relativt generelle antagelser om fordelingen, men matematikken bag bootstrap metoden er ikke trivial.


Bootstrap Percentile Confidence Intervals

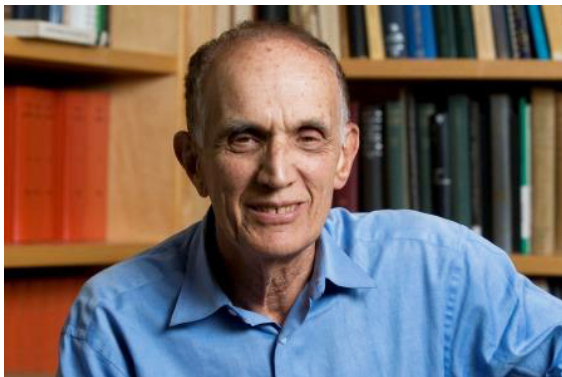
The interval between the 2.5 and 97.5 percentiles of the bootstrap distribution of a statistic is a 95% *bootstrap percentile confidence interval* for the corresponding parameter.

R kommando: `quantile(x, probs = c(0.025, 0.975))`

`x` holder værdier af bootstrapped stikprøvefunktion

Bootstrap percentil interval for θ

- ▶ giver en idé om præcisionen af estimatet $\hat{\theta}$.
- ▶ Mangler matematisk baggrund som *konfidensinterval*
 Skal bruges med forbehold!
- ▶ OK ved stikprøvestørrelse $n > 50$ og symmetriske fordelinger.
- ▶ Der findes teoretisk funderede metoder til konfidensintervaller (MSRR 7.5).



Bootstrap Methods: Another Look at the Jackknife,
1979, The Annals of Statistics

Fra Wikipedia:

Efron has been president of the American Statistical Association (2004) and of the Institute of Mathematical Statistics (1987–1988). He is a past editor (for theory and methods) of the Journal of the American Statistical Association, and he is the founding editor of the Annals of Applied Statistics. Efron is also the recipient of many awards.

Eksempel 5.3: Vandforurening i Bangladesh

Bootstrap af stikprøvegennemsnittet.

Data, $n = 271$

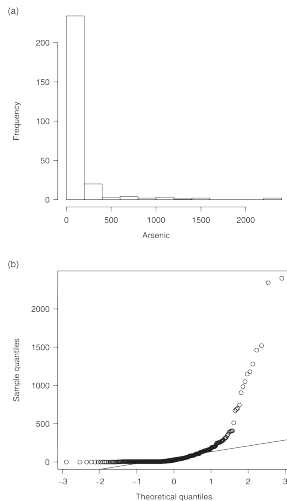


Figure 5.6 (a) Histogram and (b) QQ plot of arsenic levels in 271 wells in Bangladesh.

Bootstrap, \bar{X}

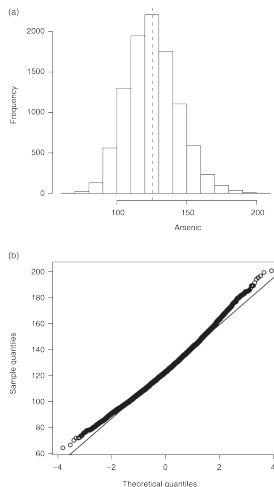
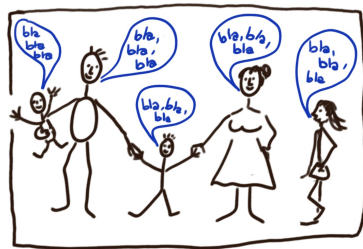


Figure 5.7 (a) Histogram and (b) QQ plot of the bootstrap distribution for mean arsenic concentration.

Teoretisk udgangspunkt:

- ▶ Observerede data $\mathbf{x} = (x_1, \dots, x_n)$ er en realisering af en vektor $\mathbf{X} = (X_1, \dots, X_n)$.
- ▶ X_1, \dots, X_n er uafhængige og identisk fordelt.

- ▶ Fordelingen af $X_i, i = 1, \dots, n$ er medlem i en **fordelingsfamilie**, i.e., en mængde af sandsynlighedsmål.
- ▶ Fordelinger identificeres ved en parameter $\theta \in \Theta$, hvor Θ er **parameterrummet**.
- ▶ Klassisk inferens:
find den fordeling i familien, der synes mest passende som ophavsmand af de observerede data.



$$\mathcal{H} = \{ \text{b}, \text{f}, \text{d}, \text{r}, \text{e} \}$$

$$\underline{X} = (\text{bla}_1, \text{bla}_2, \text{bla}_3)$$

$$\underline{x}_i = (\text{"op"}, \text{"endelig"}, \text{"ryd"})$$

Familier til kontinuerte data:

- ▶ Normalfordelinger $\{N(\mu, \sigma^2) : \underbrace{(\mu, \sigma)}_{\theta} \in \underbrace{\mathbb{R} \times \mathbb{R}^+}_{\Theta}\}$
Tæthedsfunktion $f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-\mu)^2}{2\sigma^2}),$
udfaldsrum: $x \in \mathbb{R}.$
- ▶ Exponentialfordelinger $\{\text{Exp}(\lambda) : \lambda \in \mathbb{R}^+\}$
Tæthedsfunktion $f(x; \lambda) = \lambda \exp(-\lambda x),$
udfaldsrum: $x \in \mathbb{R}^+.$
- ▶ kender I flere?

Familier til diskrete data:

- ▶ Binomialfordelinger $\{\text{Binom}(n, p) : n \in \mathbb{N}, p \in [0, 1]\}$
sandsynlighedsfunktion: $f(x; n, p) = \binom{n}{x} p^x (1 - p)^{n-x}$,
udfaldsrum: $x \in \mathbb{N}_0$.

I de fleste anvendelser er antalsparameteren n kendt, så vi arbejder effektivt med familien $\{\text{Binom}(n, p) : p \in [0, 1]\}$, med kendt n .

- ▶ kender I flere?

Hvad gør den “klassiske statistiker”, når den skal analysere data?

1. Bestemmer en fordelingsfamilie der generelt passer til problemet / data. Det er modellen.
2. Fitter modellen = at finder den passende fordeling i familien ved at estimere parameteren.

Definition (igen):

En **estimator** $\hat{\theta}$ er en statistik, som mapper udfaldsrummet af (X_1, \dots, X_n) på parameterummet Θ .

Betragt en fordelingsfamilie $\{F(.,\theta), \theta \in \Theta\}$, og lad r være dimensionen af Θ .

Vi ønsker at fitte modellen til stikprøven (x_1, \dots, x_n) .

Momentestimatoren $\hat{\theta}(x_1, \dots, x_n) \in \Theta$ er den værdi, der opfylder, at, for $X \sim F(.,\theta)$,

$$EX = \frac{1}{n} \sum_{i=1}^n x_i, \quad EX^2 = \frac{1}{n} \sum_{i=1}^n x_i^2, \quad \dots \quad EX^r = \frac{1}{n} \sum_{i=1}^n x_i^r.$$

Husk, ved kontinuerte fordelinger på \mathbb{R}

$$EX^k = \int_{-\infty}^{\infty} x^k f(x; \theta) dx,$$

og ved diskrete fordelinger på \mathbb{Z}

$$EX^k = \sum_{x \in \mathbb{Z}} x^k f(x; \theta).$$

Vi regner lidt på tavlen / seddel