Foundations of Data Science: Estimation – Principle of confidence intervals

Last Lecture

1. Parameter
   - value of a statistic (e.g. mean or max) in population
   - parameter in distribution (e.g. mean, variance of normal)

2. Point estimator
   - Method of converting sample into estimate of paramater
   - E.g. Mean of sample ( ) estimates mean of population

3. Point estimator is random variable
   - a different random sample from population =>
     different value of point estimator
   - But we only have one sample, so only one value

4. For mean, standard error of mean gives width of sampling distribution

Today

1. How to convert inferred sampling distribution of estimator into a confidence interval

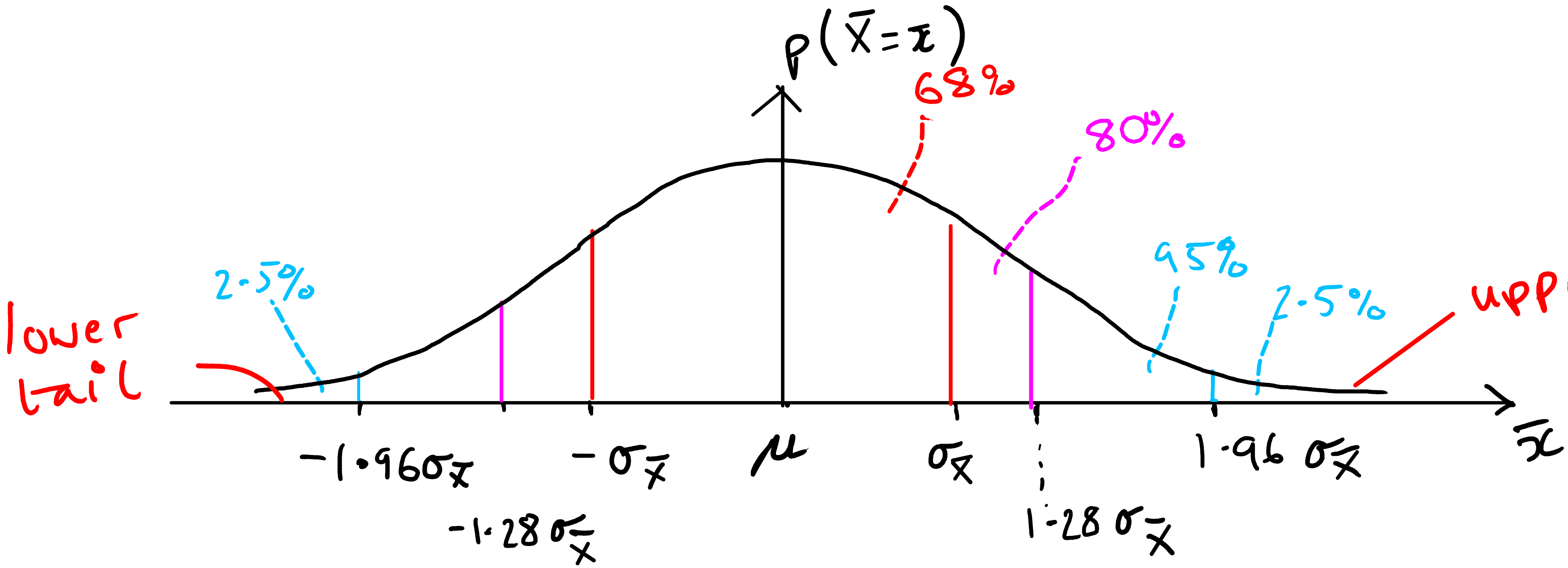
2. How to compute a confidence interval for mean of large sample - z distribution

3. Confidence intervals of parameters other than the mean - Bootstrap

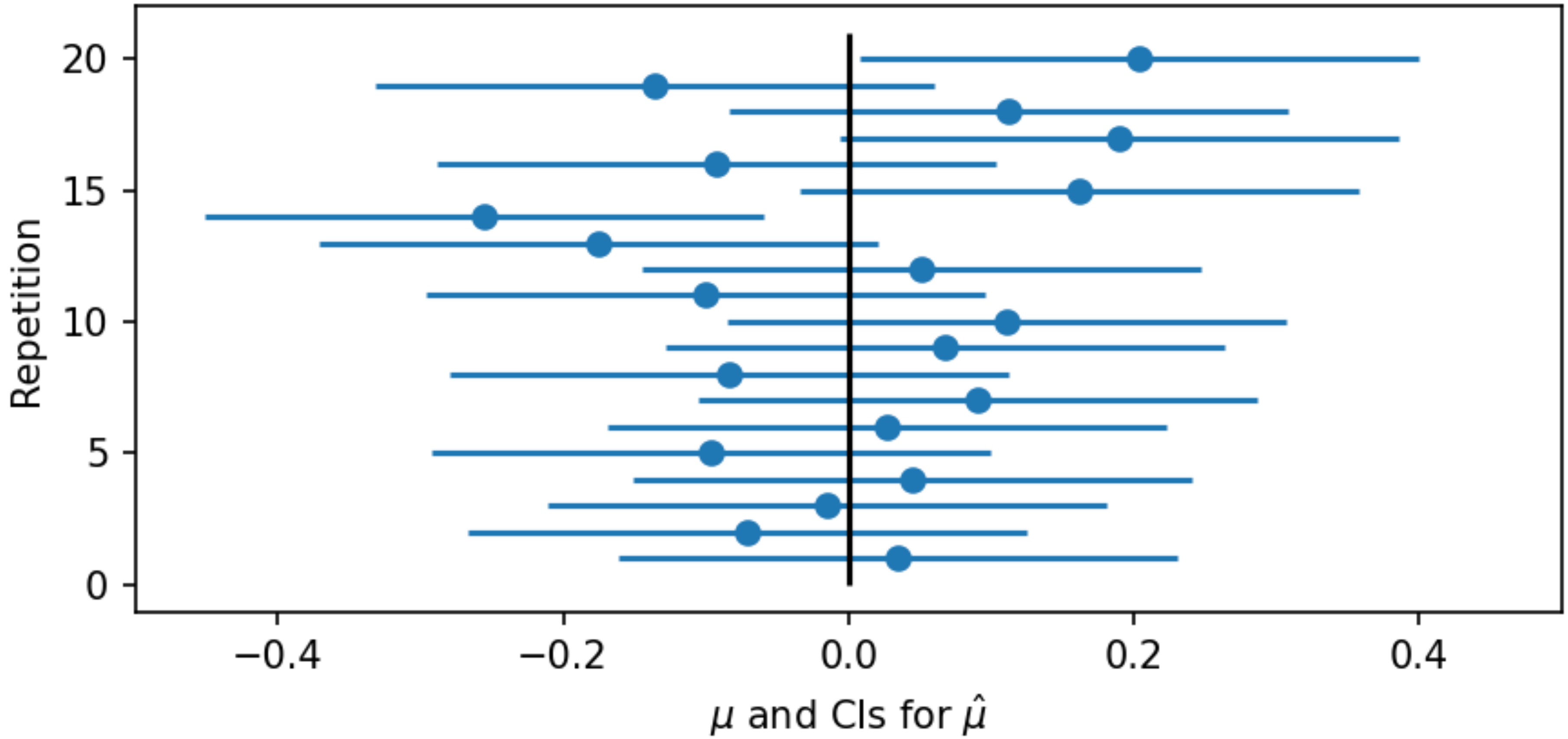
4. How big should a confidence interval be?

5. How to calculate a confidence interval for mean of a small sample - t ditribution

Confidence interval of the mean of a sample from a distribution with unknown mean and known variance



E.g.: Confidence intervals of mean of 100 samples from normal distribution with mean 0 and variance 1



Foundations of Data Science: Estimation – Definition of a confidence interval

Definition of a confidence interval

Confidence interval: An interval

$$(\hat{\vartheta} - a\hat{\sigma}_{\hat{\vartheta}},\ \hat{\vartheta} + b\hat{\sigma}_{\hat{\vartheta}})$$

that has a specified chance $1-\alpha$ of containing the parameter ϑ.

e.g. $\alpha = 0.05 \Rightarrow 1-0.05 = 95\%$ C.I.

$$P(\hat{\vartheta} - a\hat{\sigma}_{\hat{\vartheta}} < \vartheta < \hat{\vartheta} + b\hat{\sigma}_{\hat{\vartheta}}) = 1-\alpha$$

$$P(\hat{\vartheta} - a\hat{\sigma}_{\hat{\vartheta}} < \vartheta < \hat{\vartheta} + b\hat{\sigma}_{\hat{\vartheta}}) = 1-\alpha$$

$$\Rightarrow P(-\hat{\vartheta} + a\hat{\sigma}_{\hat{\vartheta}} > -\vartheta > -\hat{\vartheta} - b\hat{\sigma}_{\hat{\vartheta}}) = 1-\alpha$$

$$\Rightarrow P(a\hat{\sigma}_{\hat{\vartheta}} > \hat{\vartheta} - \vartheta > -b\hat{\sigma}_{\hat{\vartheta}}) = 1-\alpha$$

$$\Rightarrow P\left(a > \frac{\hat{\vartheta} - \vartheta}{\hat{\sigma}_{\hat{\vartheta}}} > -b\right) = 1-\alpha$$

r.v

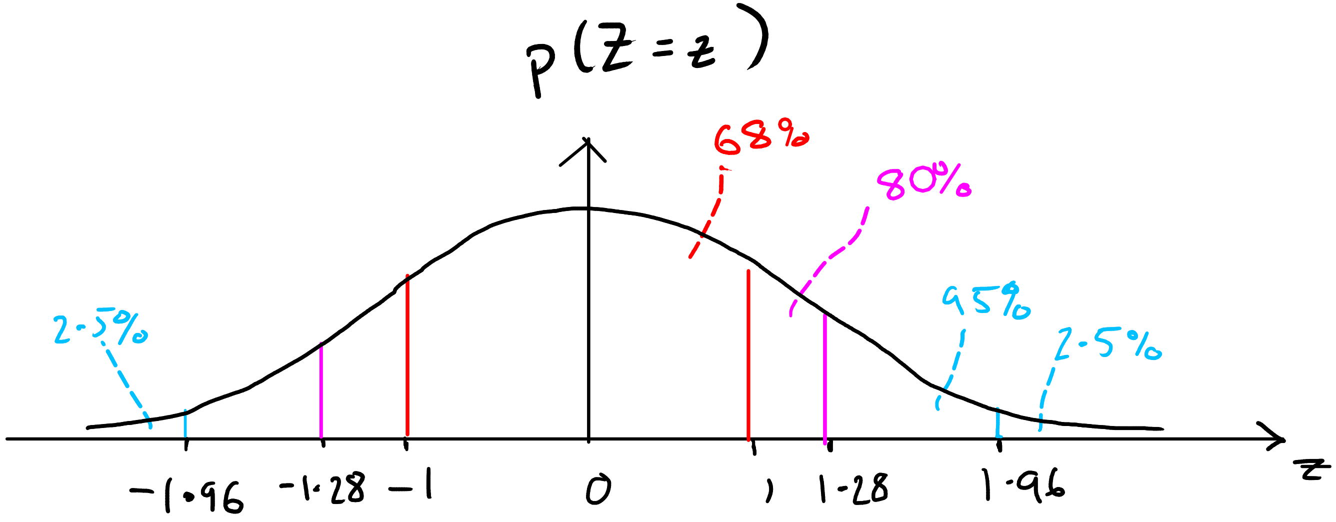
$$P\left(-b < \frac{\hat{\vartheta} - \vartheta}{\hat{\sigma}_{\hat{\vartheta}}} < a\right) = 1-\alpha$$

r.v.

in general not normal

The distribution of the standardised sample mean of a large sample

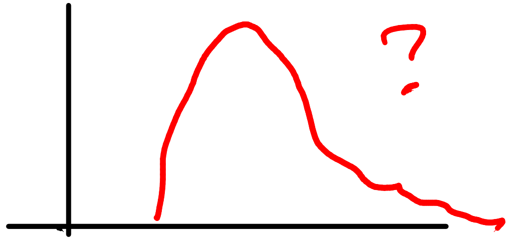
$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$



Foundations of Data Science: Estimation – Method of estimating the confidence interval of the mean of a large sample

Methods of estimating confidence intervals

$$\frac{\hat{\theta} - \theta}{\hat{\sigma}_{\hat{\theta}}}$$



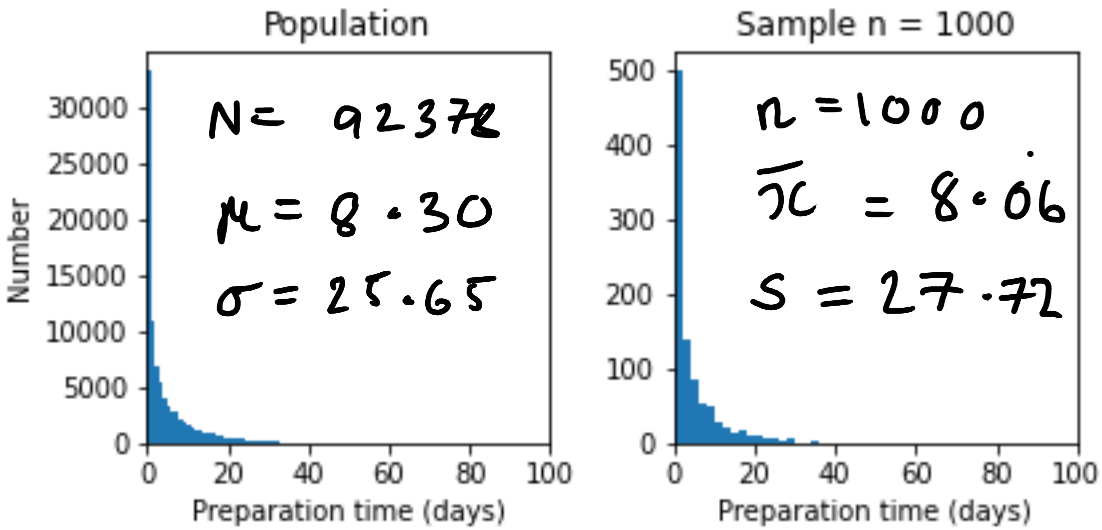
1. Theory: Assumptions about x
   Number of samples
   ⇒ Distributions

2. Bootstrap estimator – statistical
   simulation

E.g. Japanese restaurant reservation times

Mstyslav Chernov, Wikimedia Commons, CC BY SA 3.0

2016



"Preparation time"
= Time of reservation
 − Time reservation made

	Population	Sample
count	92378.00	1000.00
mean	8.30	8.06
std	25.65	27.72
min	0.00	0.00
25%	0.21	0.17
50%	2.08	1.96
75%	7.88	6.92
max	393.12	364.96

$N = 92378$ $n = 1000$

$\mu = 8.30$ $\bar{x} = 8.06$

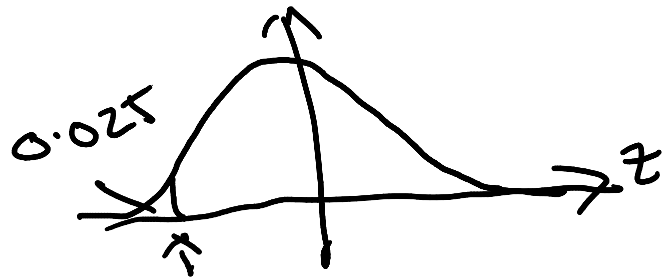
$\sigma = 25.65$ $s = 27.72$

Estimated SEM $= \frac{s}{\sqrt{n}} = \frac{27.72}{\sqrt{1000}} = 0.88$ days

$\hat{\sigma}_{\bar{x}}$

Large sample ⇒ Normal distribution of sample mean ⇒ "z" distribution

$95\% \Rightarrow \alpha = 0.05$



$z_{\alpha/2} = z_{0.025} = 1.96$

$(\bar{x} - z_{0.025}\,\hat{\sigma}_{\bar{x}},\ \bar{x} + z_{0.025}\,\hat{\sigma}_{\bar{x}}) = (6.34, 9.78)$

Reporting confidence intervals

$(6.34, 9.78)$

$M = 8.06$, $CI = 6.34 - 9.78$ (95% CI)

$\hat{\mu} = 8.06 \pm 1.72$ (95 % CI)

$z_{0.025} \hat{\sigma}_{\bar{x}} = 1.96 \times 0.88$

$\hat{\mu} = 8.06 \pm 0.88$ (Mean $\pm$ 1. SEM)

Summary

- confidence intervals for mean of large samples
- General definition of confidence intervals
- Example of theoretical method of computing confidence intervals from sample data.

Summary so far

- Confidence intervals for mean of large samples

- General definition of confidence intervals

- Example of theoretical method of computing confidence intervals from sample data

Foundations of Data Science: Estimation – Bootstrapping

Principle of bootstrapping

Baron Münchhausen - Wikipedia

- Treat the sample like a population
- Resample estimator from it to get sampling distribution
- Sample is similarly to population for a large sample

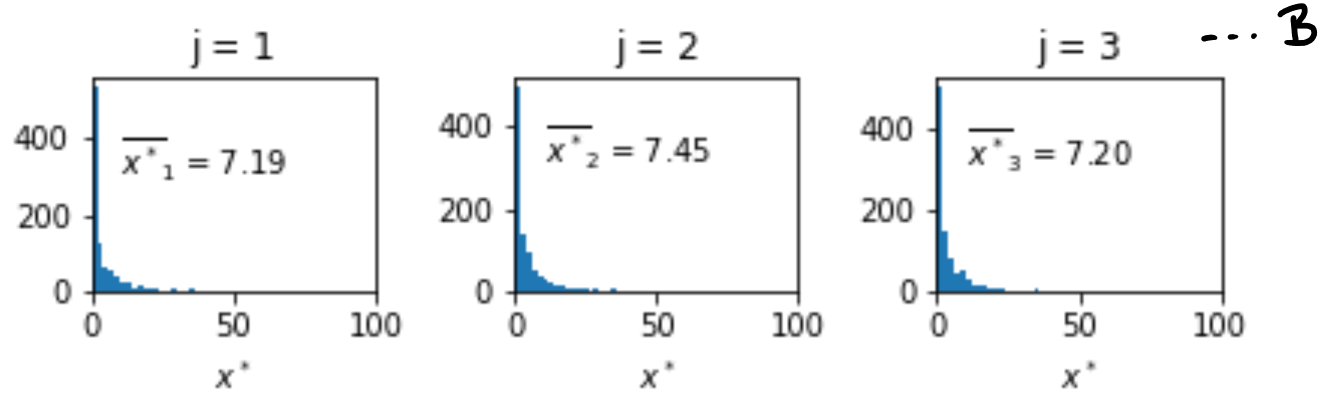
Bootstrap confidence interval for the mean

$n = 1000$ $\quad x$

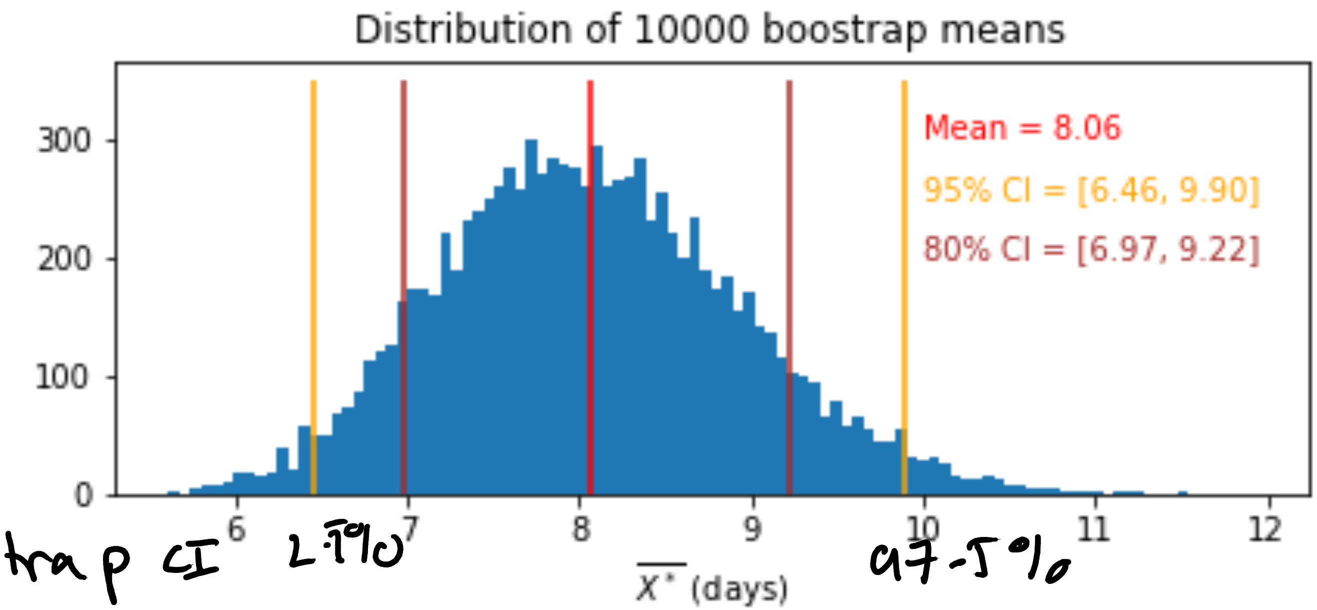
for $j = 1, \ldots, B$ — # Bootstrap samples

x^* of size n from x with replacement

$\bar{x}^*_j \leftarrow$ mean x^*



$$S^2_{boot} = \sum_{j=1}^{B} \frac{(\bar{x}^*_j - \bar{x})^2}{B-1}$$



(6.46, 9.90) – Bootstrap CI

(6.34, 9.78) – Normal approx

General formulation of the bootstrap

Bootstrap CI. $\hat{\vartheta}$ — $\hat{\mu}$, $\hat{\sigma}^2$, $\hat{\beta}$

- For j in 1, ..., B
  - Sample n items from x with replacement
  - Compute sample stat of the new sample $\hat{\vartheta}^*_j$
- Bootstrap estimator of variance of statistic

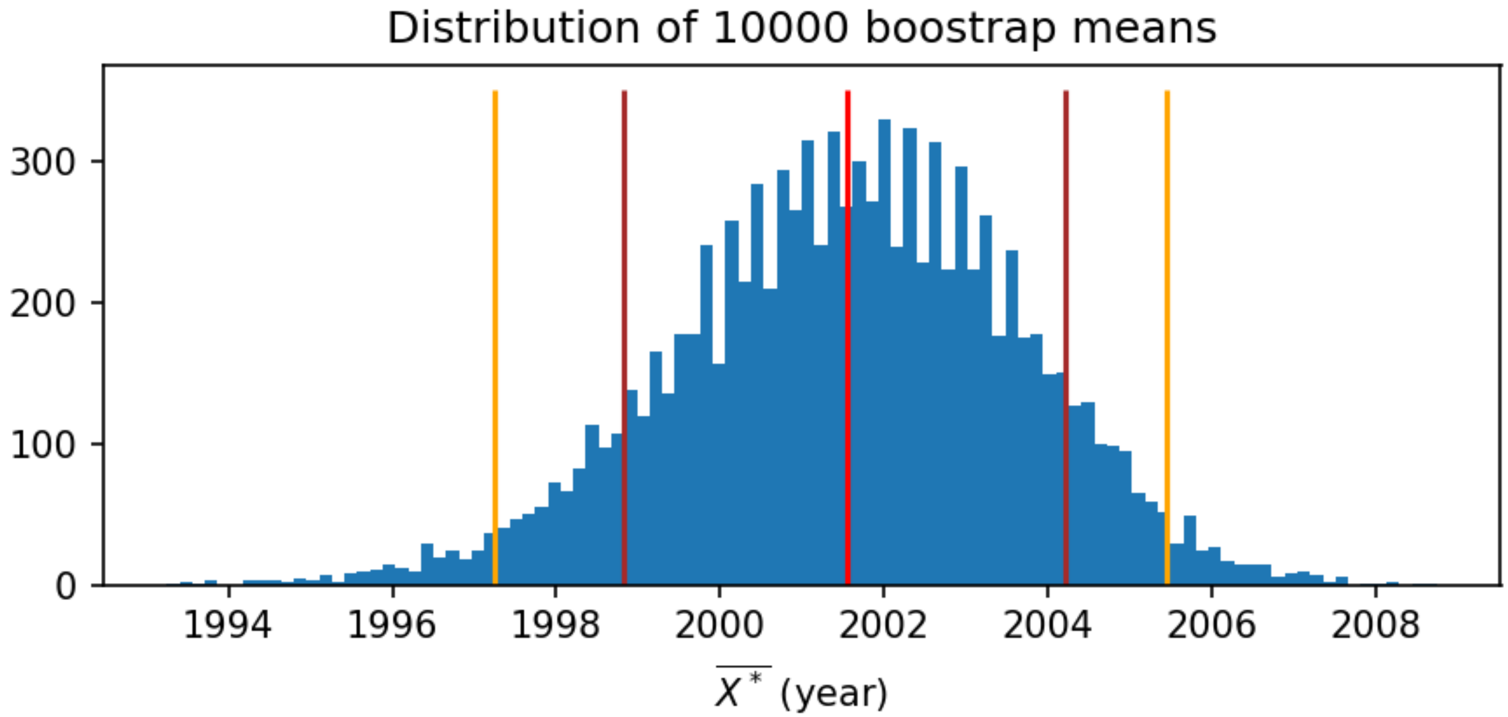
$$s^2_{boot} = \sum_{j=1}^{B} \frac{(\hat{\theta}_j^* - \hat{\vartheta})^2}{B-1}$$

- Find CI from Bootstrap dist.

✓ centrality – median mean

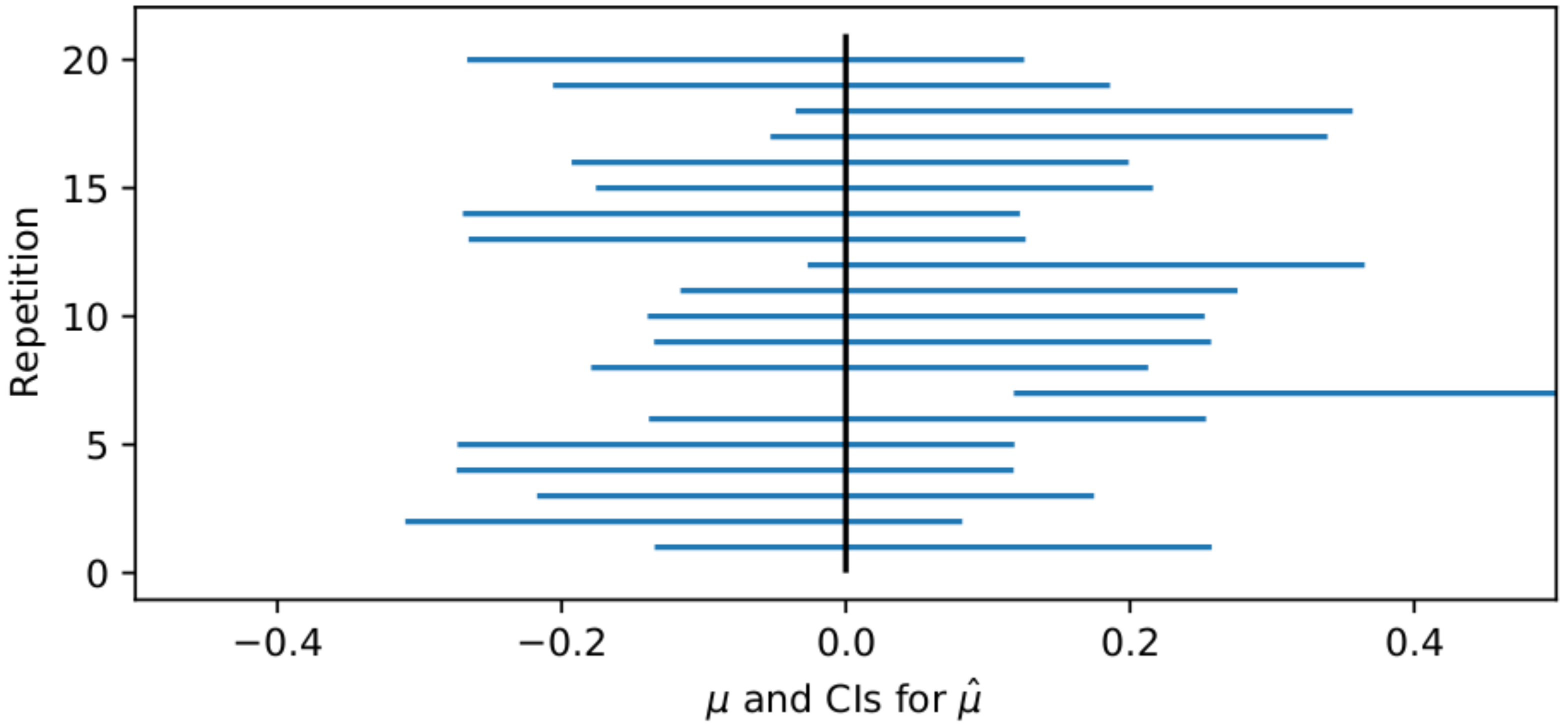
✗ Extremes – max or min

Bootstrap coin year

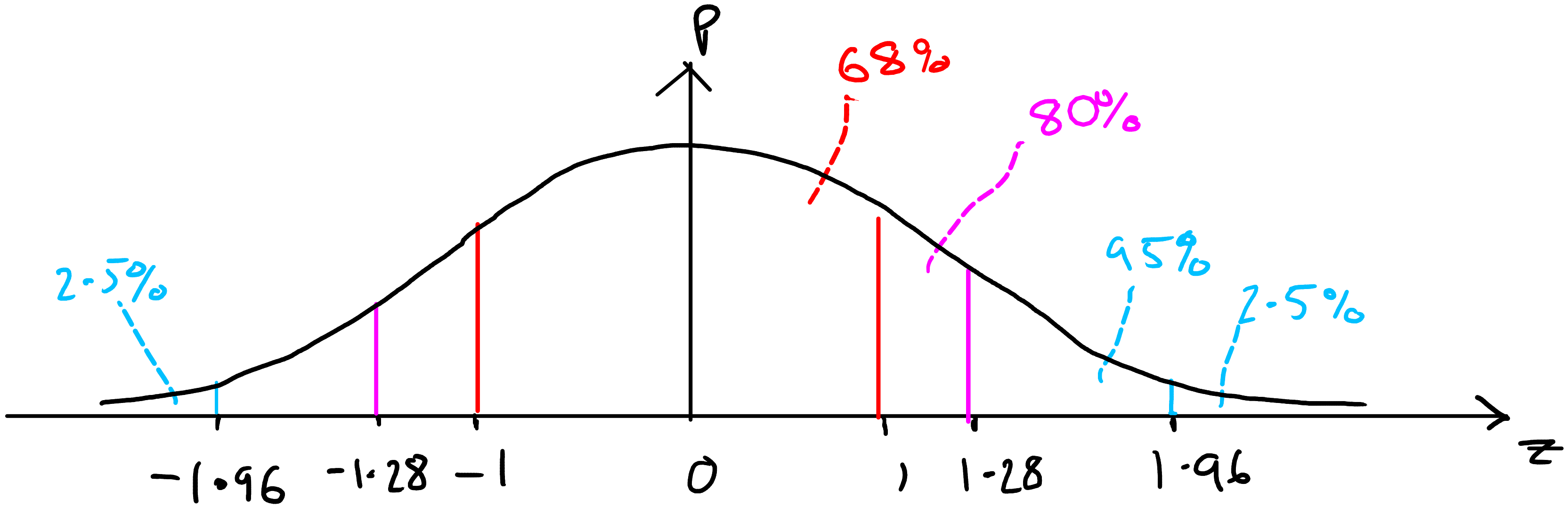


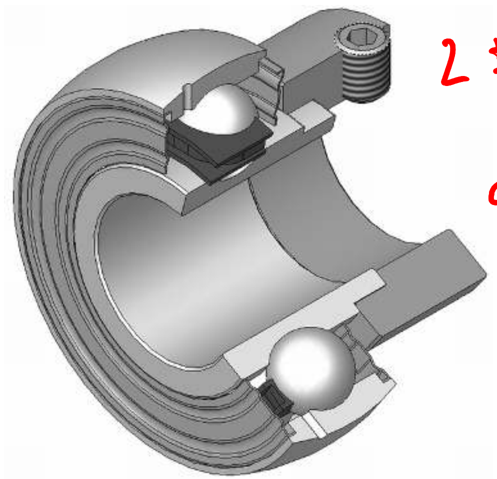
Foundations of Data Science: Estimation – Interpretation of confidence intervals

Confidence intervals are a random interval



How big should a confidence interval be?





2 ± 0.0001 mm

99.999 %

Wikimedia commons, Silberwolf, CC BY 2.5

Wikimedia commons, Kiefer, CC BY SA 2.0

How big should a confidence interval be?

- Say 68% confident of being in 2 years of the true date

- What could go wrong if the estimated date is further away?

Foundations of Data Science: Estimation – Confidence intervals for the mean for small samples

Small samples

$n \leq 40$

$n = 29$ coins

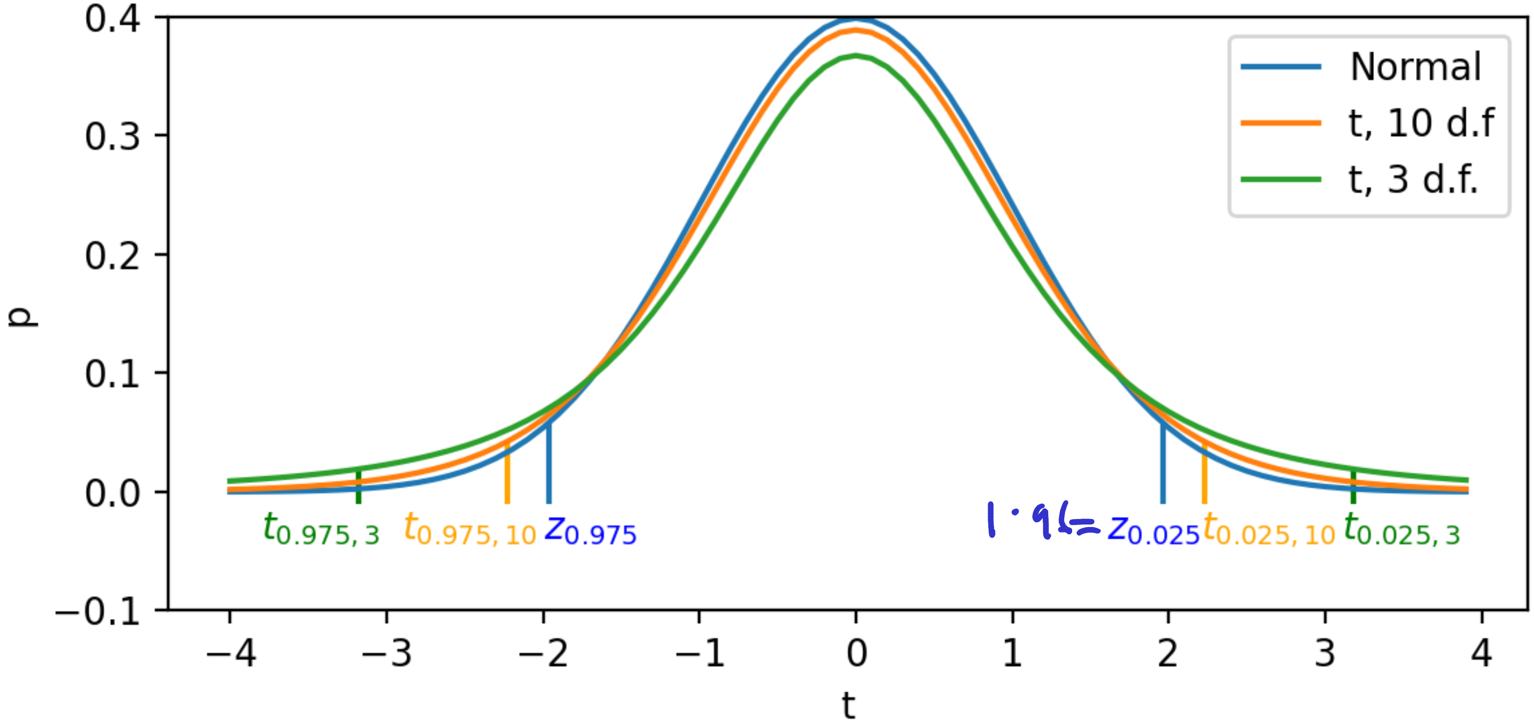
$\bar{x} = 2001.551$ years $\qquad s = 11.444$ years

Estimated SEM, $\sigma_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{11.444}{\sqrt{29}} = 2.125$ years

$\hat{\mu} = \bar{x}$

t - statistic $\quad T = \frac{\bar{X} - \mu}{\hat{\sigma}_{\bar{x}}}$

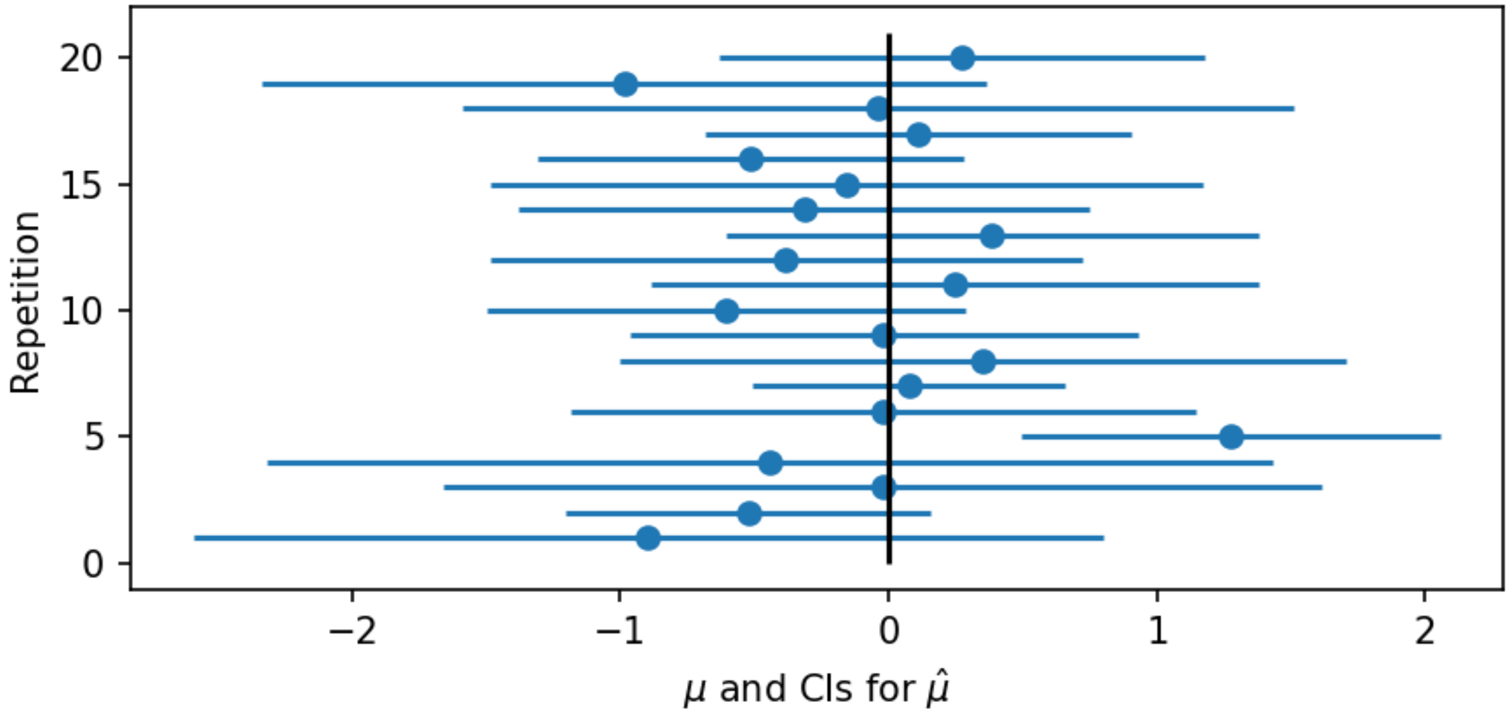
The t-distribution



$$T = \frac{\bar{x} - \mu}{\hat{\sigma}_{\bar{x}}}$$

t- critical value $t_{\alpha,\nu}$: value of t at which, in a t-distribution with ν d.f there area under the curve of α to its right.

Confidence intervals with small samples



Using the t-distribution to calculate a confidence interval

95% C.I $\Rightarrow$ $\alpha = 0.05$

Sample size $n \Rightarrow$ $\nu = n-1$ d.f.

$t_{\alpha/2, \nu} = t_{\frac{\alpha}{2}, n-1}$ t-critical value

$\bar{x} - t_{\frac{\alpha}{2}, n-1} \hat{\sigma}_{\bar{x}}$, $\bar{x} + t_{\frac{\alpha}{2}, n-1} \hat{\sigma}_{\bar{x}}$

$n = 29$, $\alpha = 0.05$ $\Rightarrow$ $t_{0.025, 29-1} = t_{0.025, 28}$

$= 2.281$

$t_{0.025, 28} \hat{\sigma}_{\bar{x}} = 2.281 \times 2.125 = 4.871$ years

$\Rightarrow$ $\hat{\mu} = 2001 \pm 5$ years (95% C.I.)

Summary

1. Principle and meaning of confidence intervals

2. Confidence intervals of the mean of a large samples ($n > 40$)
   computed theoretically
   - z distribution

3. Confidence intervals for more types of estimator
   computed using the bootstrap

4. Confidence intervals of the mean of a small sample ($n < 40$)
   computed theoretically
   - t distribtion