<InfPALS/>

for students, by students

Skills - Mondays 16:00-17:00, room 7.14
Wednesdays 14:00-15:00, room 7.14

Projects - Thursday 10:00-11:00, room 7.14

All sessions are in Appleton Tower!

Come along to learn useful skills like **Git**, **Command Line**, **and LaTeX** and build projects with **Pygame** and **AI Models**

infpals.github.io

Inf2 - Foundations of Data Science: Estimation - Principle of confidence intervals

THE UNIVERSITY of EDINBURGH
informatics

FOUNDATIONS OF DATA SCIENCE

Announcements

InfPALS - especially LaTeX

Workshop

Lab

Project ideas - due Friday!

Office Hour - 10 on Wednesday

Last Lecture

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

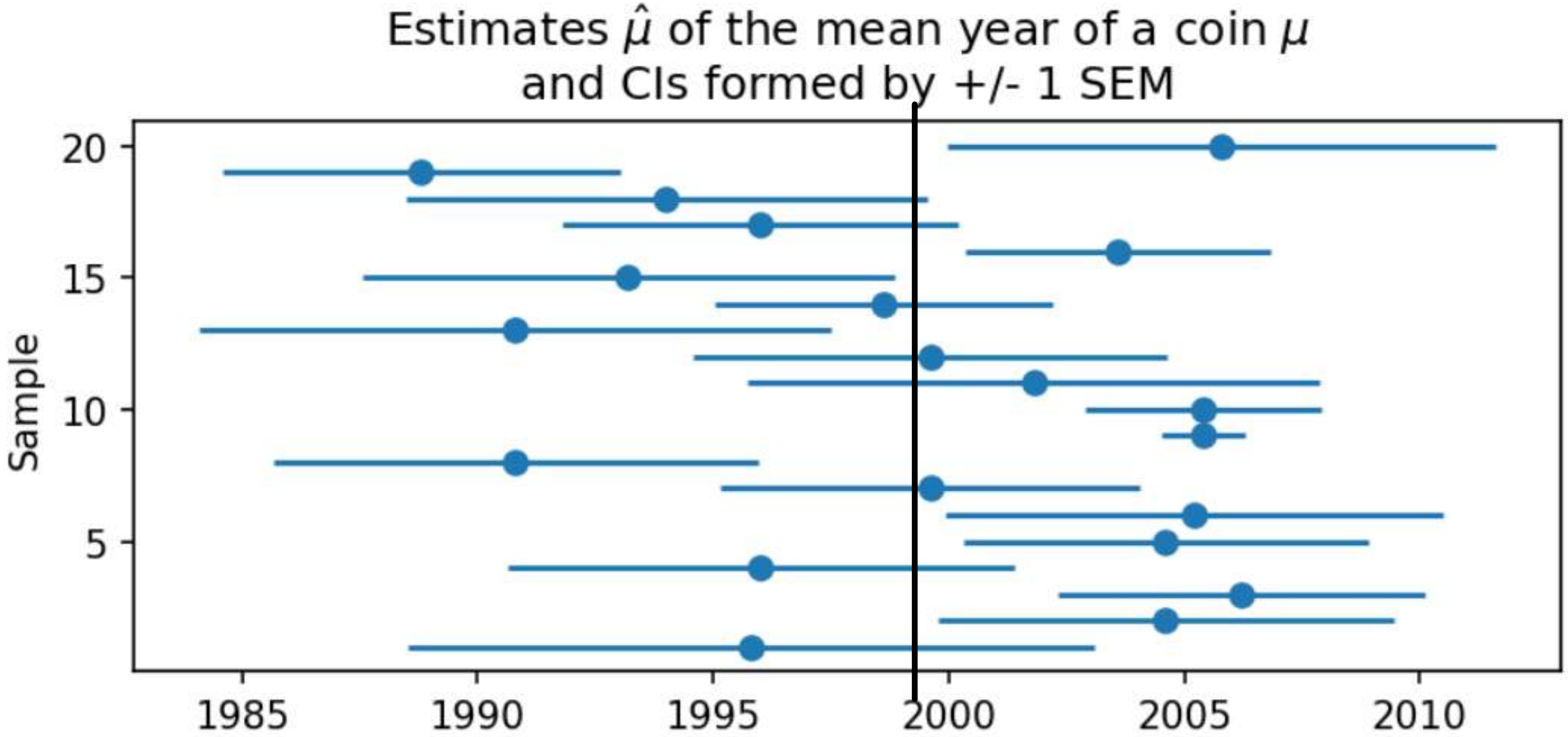
2. Point estimator
   - Method of converting sample into estimate of paramater
   - E.g. Mean of sample ($\bar{x}$) estimates mean of population μ

     $\hat{\mu}$

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. Idea of confidence intervals for estimator
   - based on sample standard error of estimator

Confidence intervals from different samples



How do we calibrate the width of the confidence interval so that there is a specified chance that it encloses the true value?

Today

1. How to convert inferred sampling distribution of estimator into a confidence interval with a specified chance of enclosing true value

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

3. Choosing confidence levels and how much data to collect

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

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

Theory reminder:
Standard error of mean for known distribution variance σ

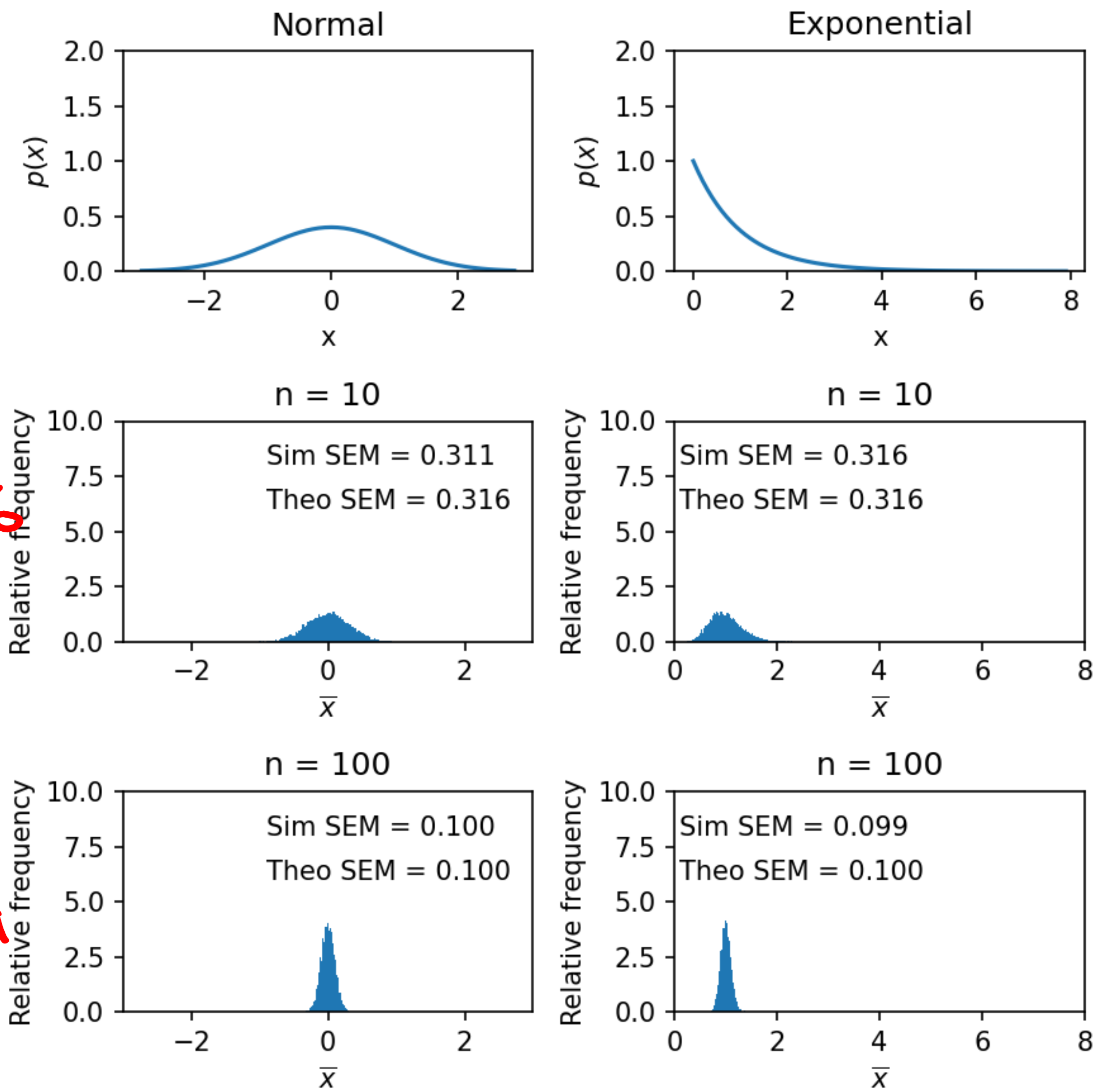
Standard deviation

$\sigma = 1$

SEM

$\sigma_{\hat{\mu}} = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{10}} = 0.316$

$\sigma_{\bar{x}} = \frac{1}{\sqrt{100}} = \frac{1}{10} = 0.1$



Estimated standard error for distribution with unknown variance σ

What if we don't know σ ?

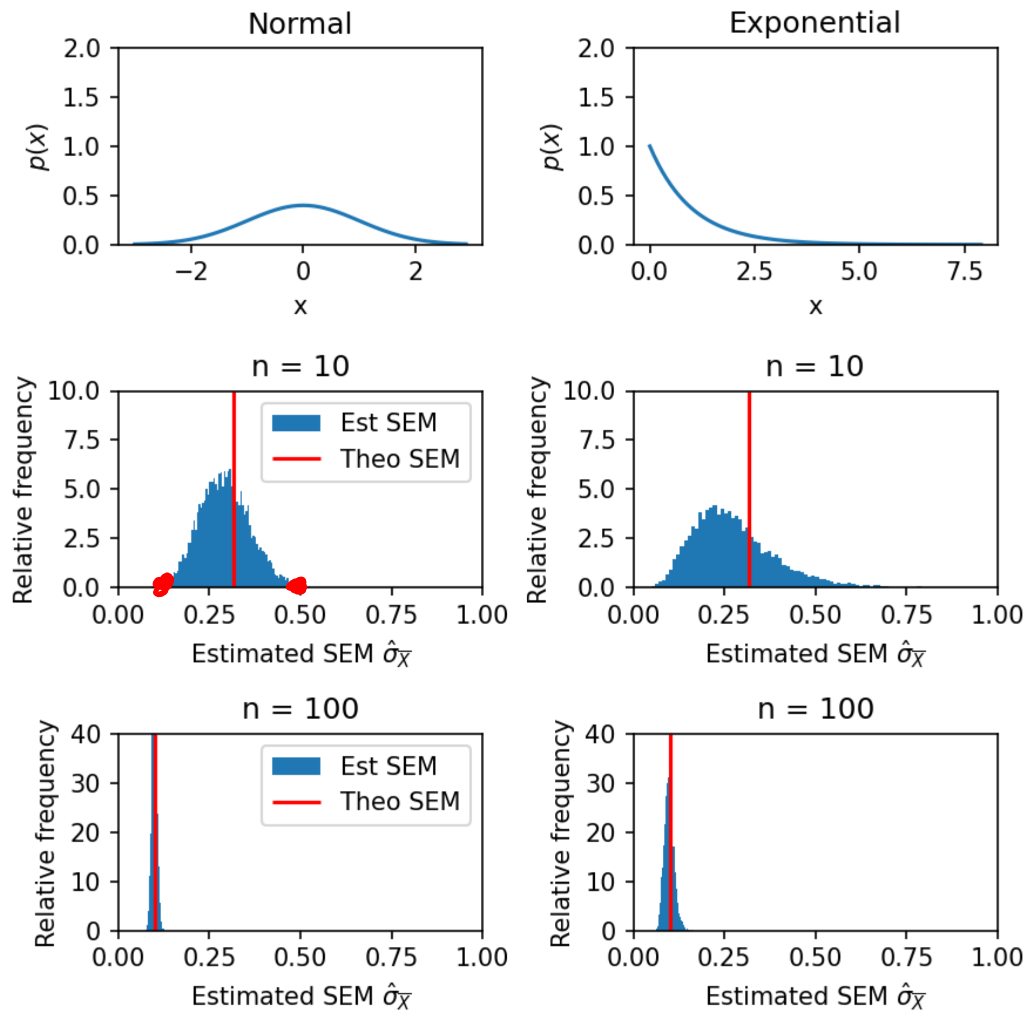
Estimated S.E.

of estimator $\hat{\sigma}_{\hat{\theta}}$

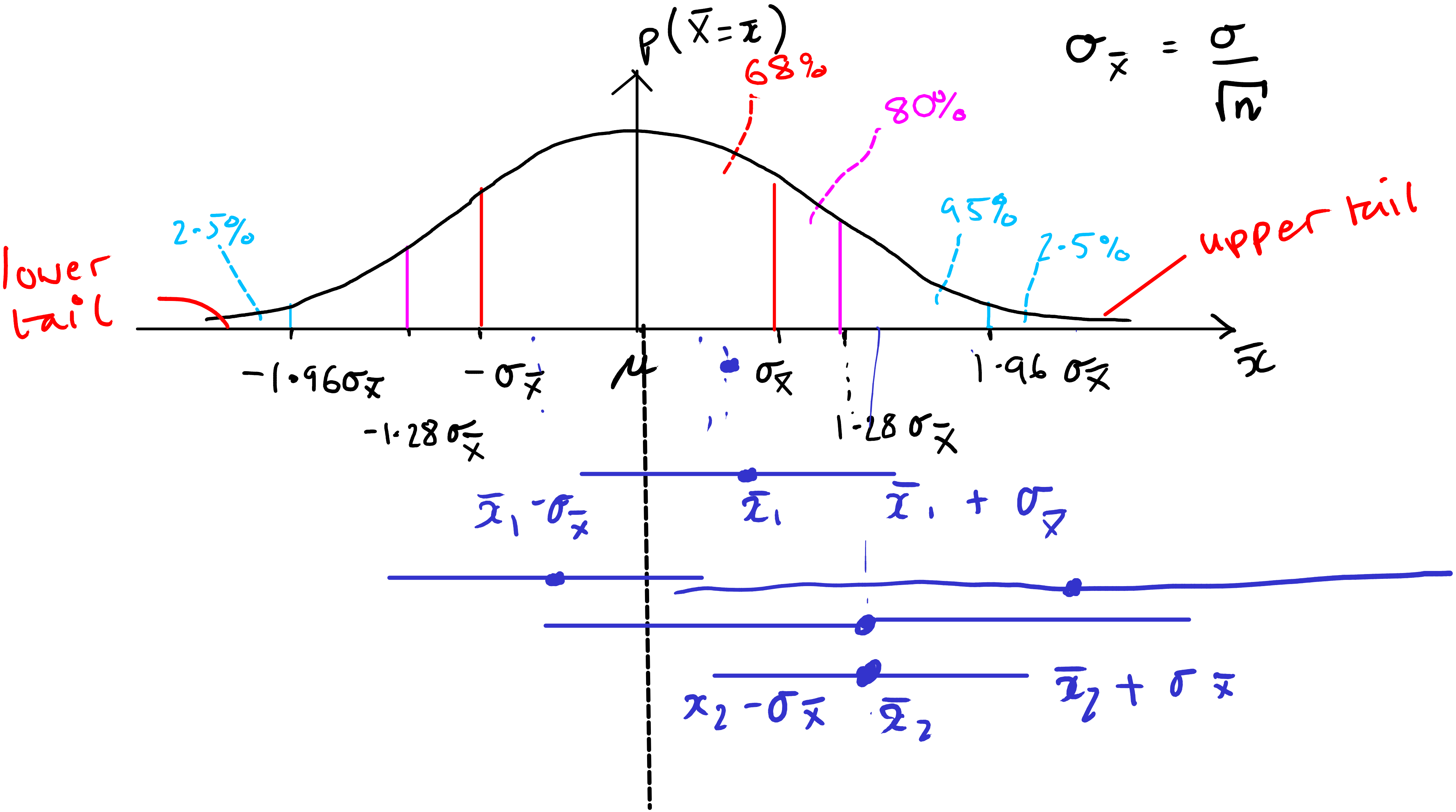
$$\hat{\sigma}_{\hat{\mu}} = \frac{S}{\sqrt{n}} \leftarrow \text{r.v.}$$

$n \to \infty$

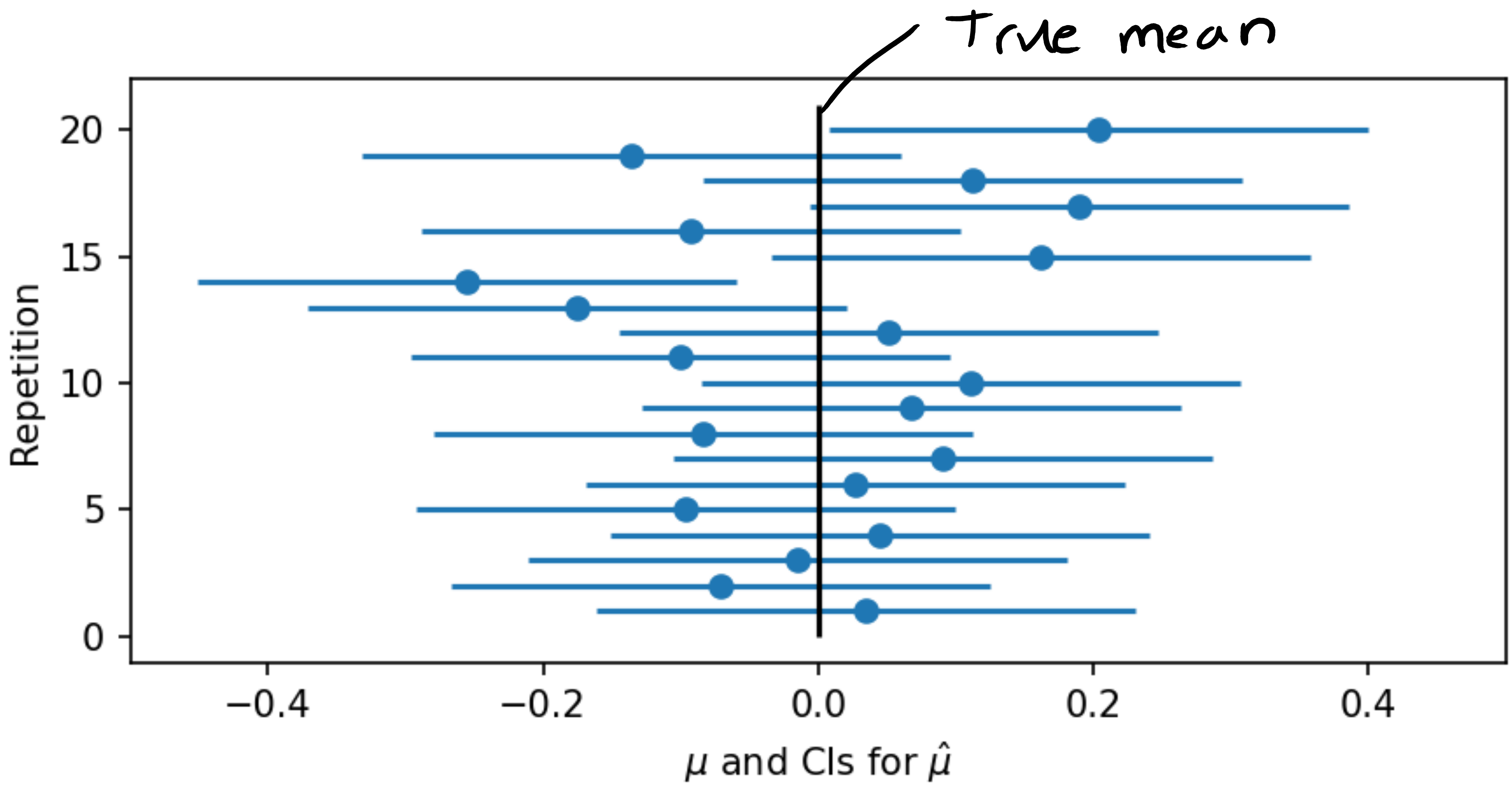
n large $\qquad \hat{\sigma}_{\hat{\mu}} = \sigma_{\hat{\mu}}$



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



E.g.: 95% Confidence intervals of mean of 100 samples from normal distribution with mean 0 and variance 1



Inf2 - Foundations of Data Science: Estimation - Definition of a confidence interval

THE UNIVERSITY *of* EDINBURGH
informatics

FOUNDATIONS OF DATA SCIENCE

Definition of a confidence interval

Confidence interval: An interval for a parameter ϑ

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

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

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

$$P\left(\hat{\vartheta} - a\hat{\sigma}_{\hat{\vartheta}} < \overset{\text{True}}{\vartheta} < \hat{\vartheta} + \underset{a}{b}\hat{\sigma}_{\hat{\vartheta}}\right) = 1 - \alpha$$

Often the interval is symmetric, ie $a = b$.

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

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

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

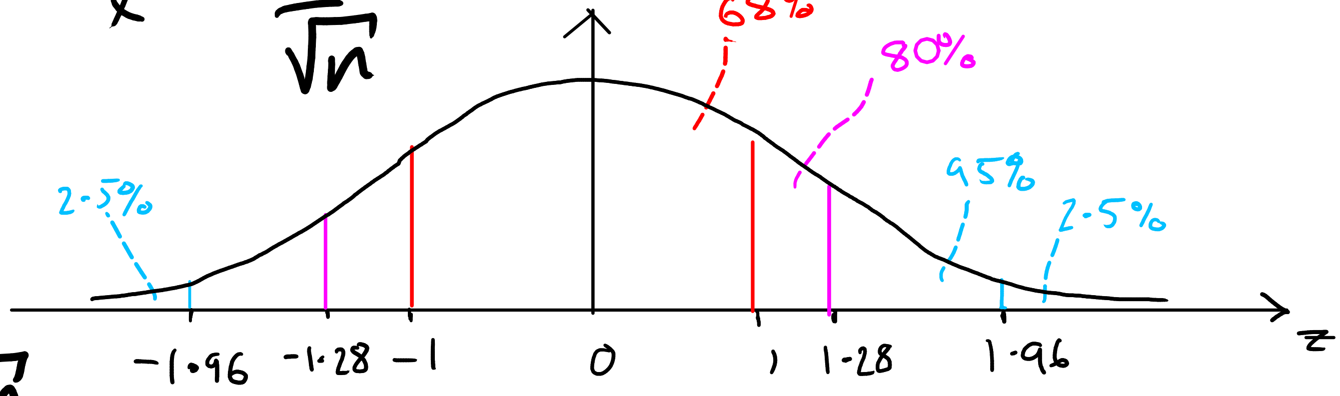
$$\Rightarrow \quad P\left(-b < \boxed{\frac{\hat{\vartheta} - \vartheta}{\hat{\sigma}_{\hat{\vartheta}}}} < a\right) = 1 - \alpha$$

random variable

The distribution of the standardised sample mean of a large sample - the z-distribution

$\vartheta = \mu \quad \hat{\vartheta} = \bar{X} \quad \hat{\sigma}_{\hat{\vartheta}} = \hat{\sigma}_{\bar{X}} = \dfrac{s}{\sqrt{n}} \approx \sigma \quad p(Z = z)$

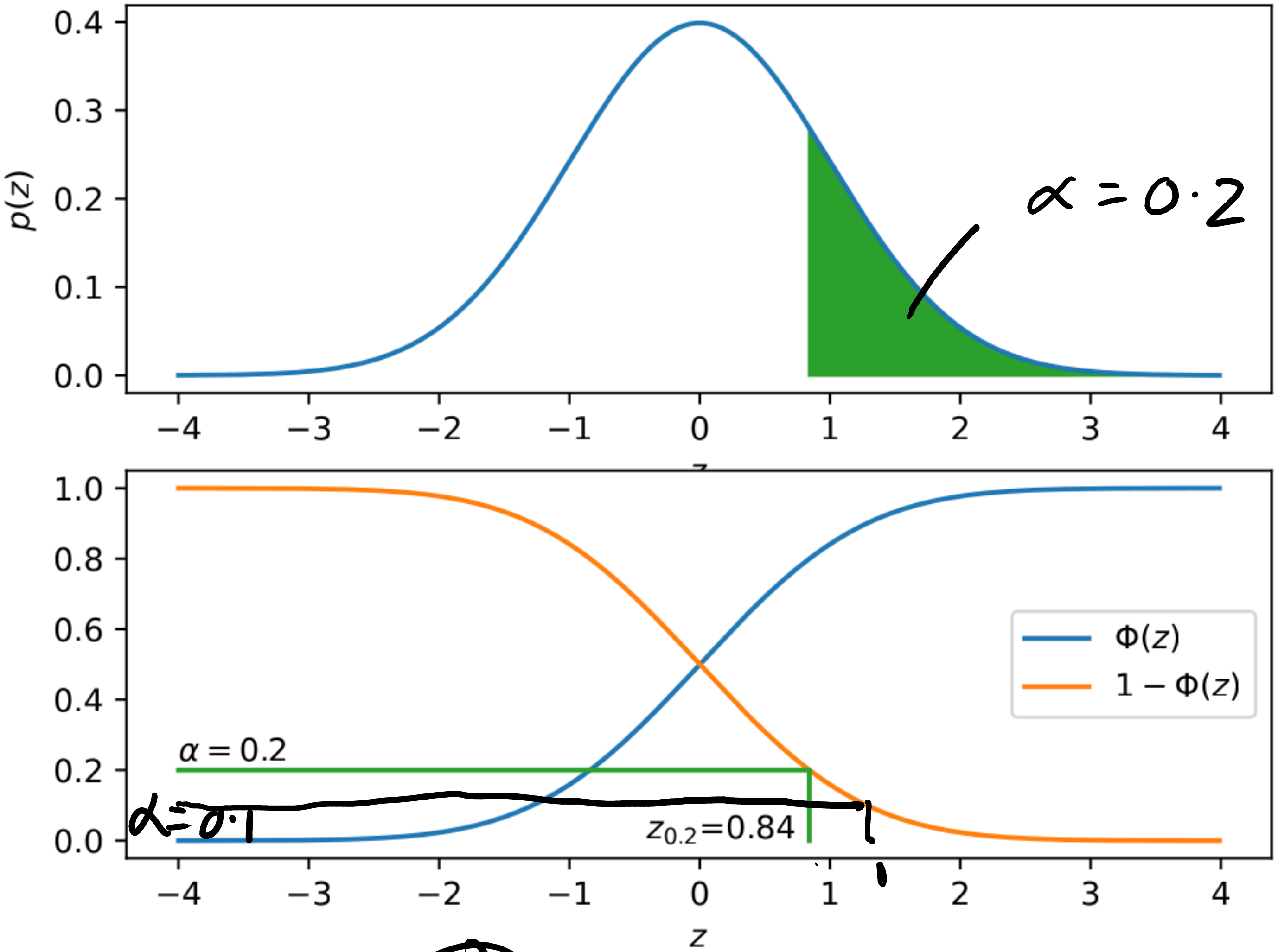
$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$



Normal dist $\mu = 0$, $\sigma^2 = 1$

$$\hat{\sigma}_{\bar{X}} \simeq \frac{\sigma}{\sqrt{n}}$$

z-critical values



α	0.100	0.050	0.025	0.010	0.005	0.001
ν						
1	3.078	6.314	12.706	31.821	63.657	318.309
2	1.886	2.920	4.303	6.965	9.925	22.327
3	1.638	2.353	3.182	4.541	5.841	10.215
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
20	1.325	1.725	2.086	2.528	2.845	3.552
30	1.310	1.697	2.042	2.457	2.750	3.385
40	1.303	1.684	2.021	2.423	2.704	3.307
∞	1.282	1.645	1.960	2.326	2.576	3.090

α	0.100	0.050	0.025	0.010	0.005	0.001
∞	1.282	1.645	1.960	2.326	2.576	3.090

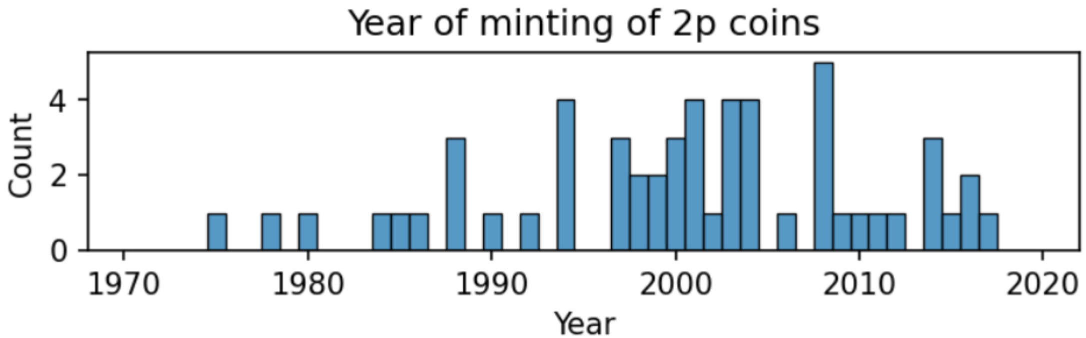
Inf2 - Foundations of Data Science: Estimation - Theoretical method of estimating the confidence interval of the mean of a large sample

THE UNIVERSITY of EDINBURGH informatics

FOUNDATIONS OF DATA SCIENCE

Confidence interval for the year of a 2p coin

$n = 56$



In a sample of 56 2p coins, the mean year of minting of the 2p coins is 2000.8 and the sample standard deviation is 10.4. Give a 95% confidence interval for the mean year of minting in the population of all 2p coins.

Practice more in this week's workshop sheet

Solution

$n = 56$

Mean age $\bar{x} = 2000.8$ years

Sample st. dev $s = 10.4$ years

S.E.M $\hat{\sigma}_{\bar{x}} = \dfrac{s}{\sqrt{n}} = \dfrac{10.4}{\sqrt{56}} = 1.390$

Large sample $(n > 40) \Rightarrow Z = \dfrac{\bar{x} - \mu}{\hat{\sigma}_{\bar{x}}}$ is z-distributed

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

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

$(\bar{x} - z_{\alpha/2}\hat{\sigma}_{\bar{x}},\ \bar{x} + z_{\alpha/2}\hat{\sigma}_{\bar{x}})$

$= (2000.8 - 1.96 \times 1.390,\ 2000.8 + 1.96 \times 1.390)$

$= \underline{(1998.1,\ 2003.5) \text{ is } 95\% \text{ CI}}$

Reporting confidence intervals

(1998 , 2004)

$M = 2001$, CI = 1998 – 2004 (95% CI)

$\hat{\mu} = 2001 \pm 3$ (95 % CI)

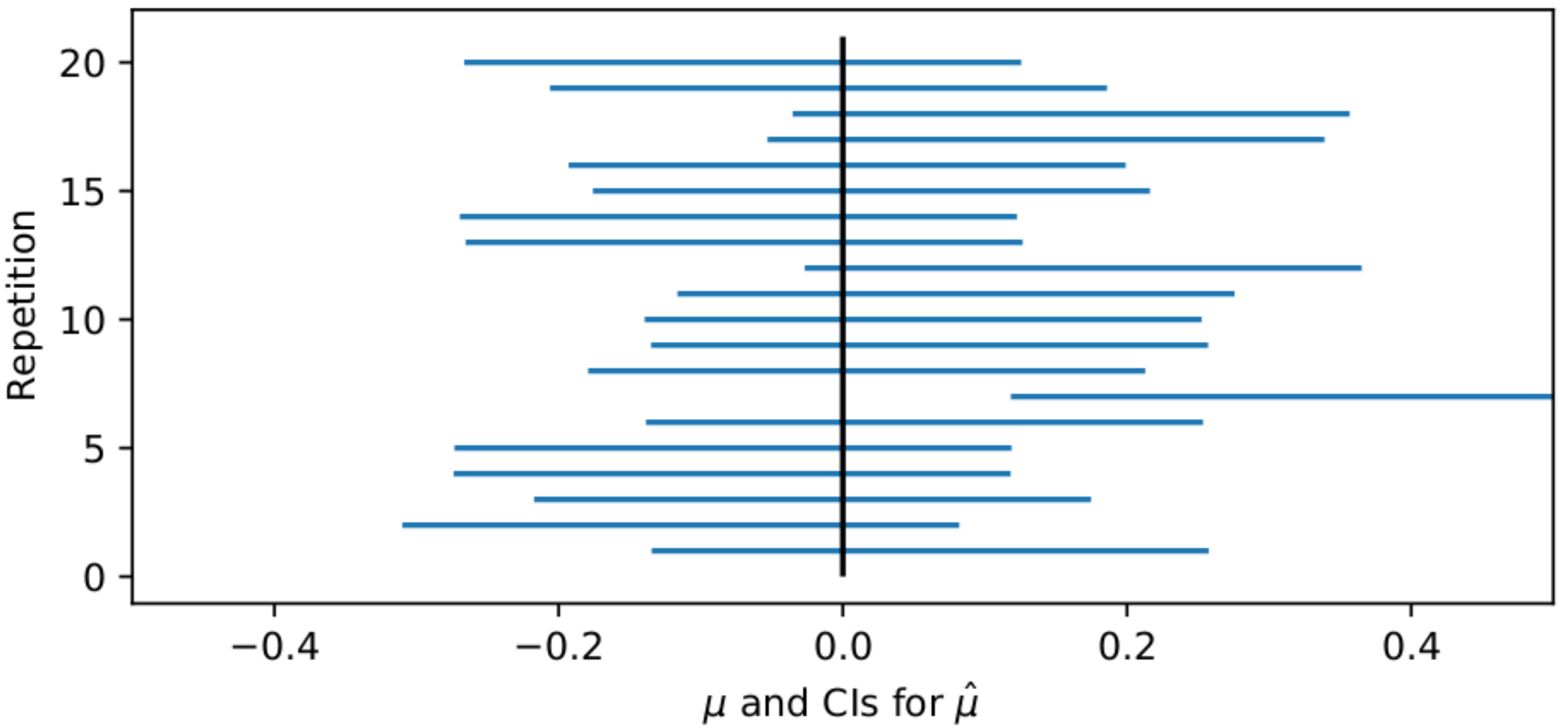
$\hat{\mu} = 2006.1 \pm 1.4$ (Mean ± 1. SEM)

Inf2 - Foundations of Data Science: Estimation - Interpretation of confidence intervals

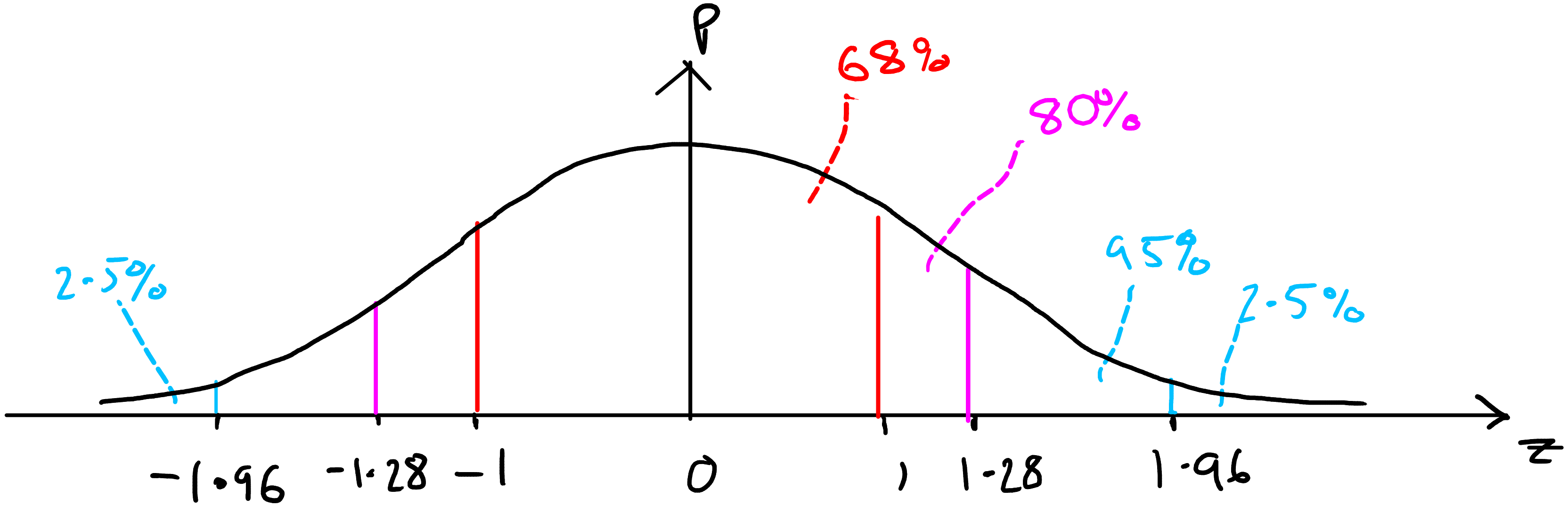
THE UNIVERSITY of EDINBURGH
informatics

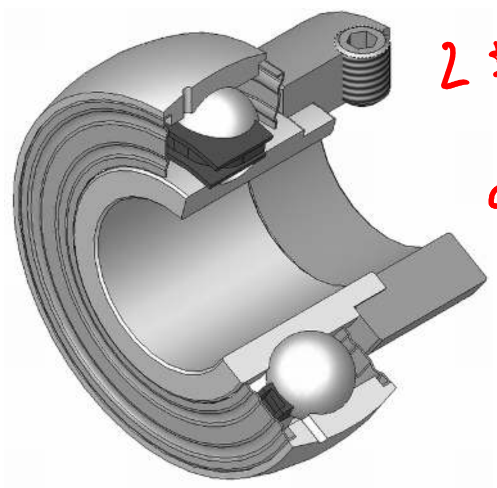
FOUNDATIONS OF DATA SCIENCE

Confidence intervals are a random interval



What level of confidence should we choose?





2 ± 0.00001 mm

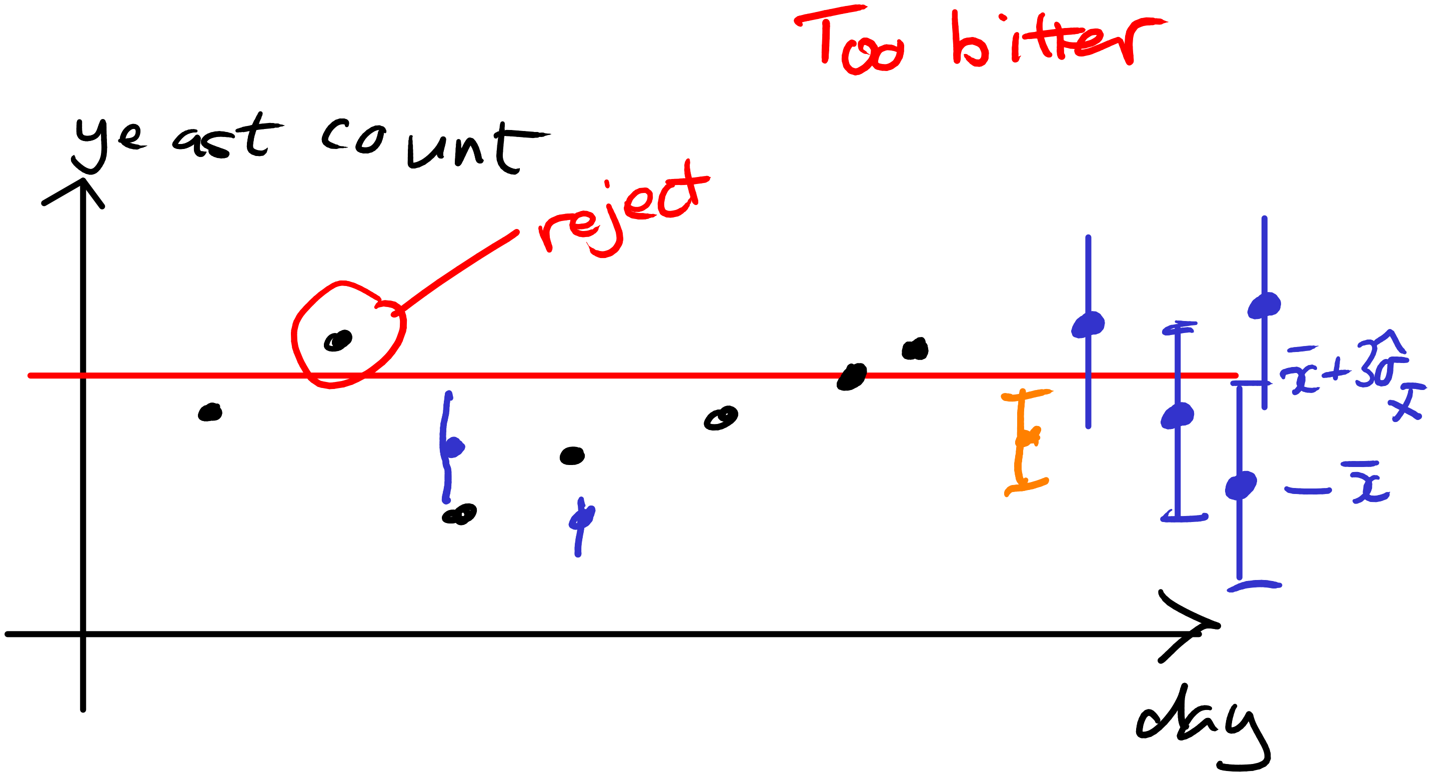
99.999 %

Wikimedia commons, Silberwolf, CC BY 2.5

Wikimedia commons, Kiefer, CC BY SA 2.0

How much data do we collect?

A question inspired by the work of "STUDENT" (aka W. S. Gosset) in a brewery



Suppose we only want 1% of beer to be too bitter
What level of confidence should we have?
How many samples per day should we make?

Inf2 - Foundations of Data Science: Estimation - Bootstrapping

THE UNIVERSITY *of* EDINBURGH
informatics

FOUNDATIONS OF DATA SCIENCE

Principle of bootstrapping

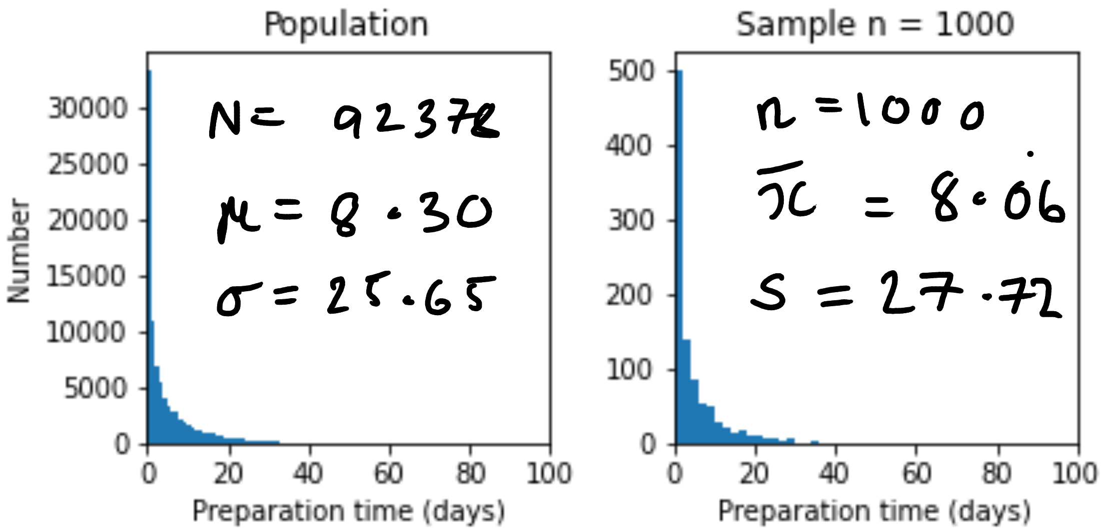
Baron Münchhausen - Wikipedia

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

Related lab on the bootstrap

E.g. Japanese restaurant reservation times

2016



	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

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

Bootstrap confidence interval for the mean

$n = 1000 \qquad x$

\# Bootstrap samples

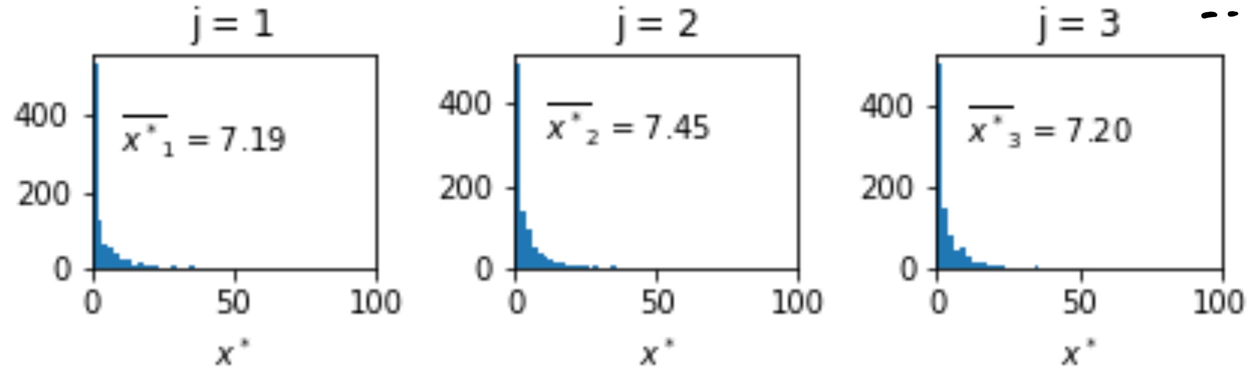
for $j = 1, \ldots, B$

x^* of size n

from x

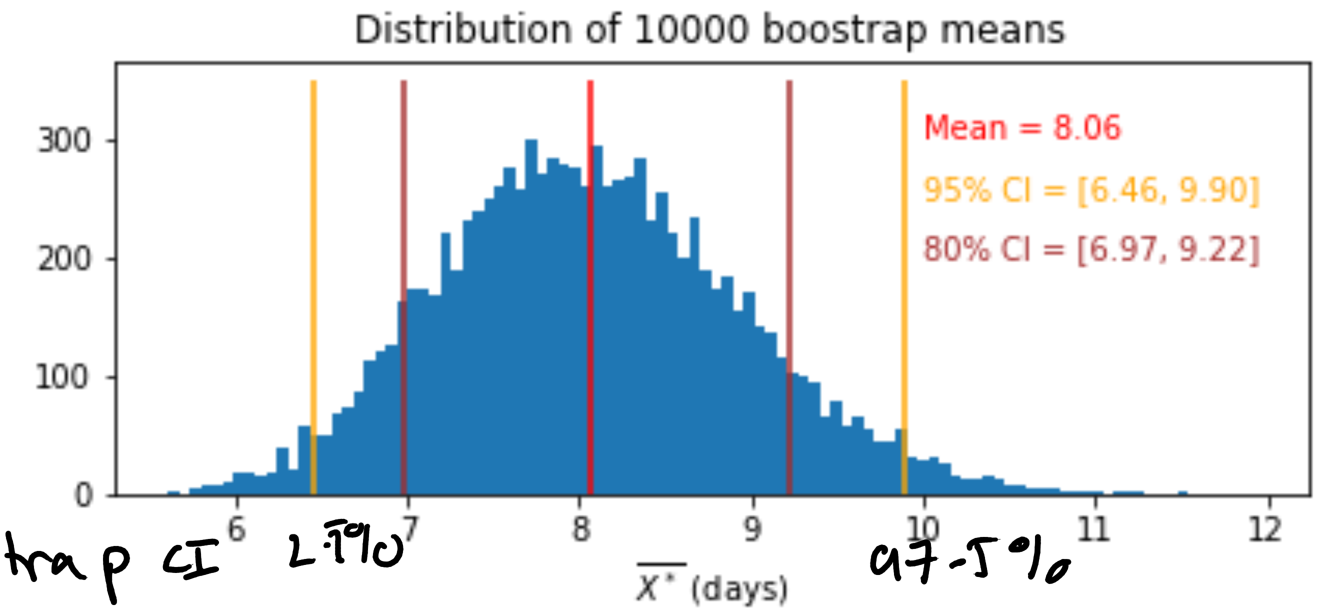
with replacement

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



... B

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



2.5% 97.5%

(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

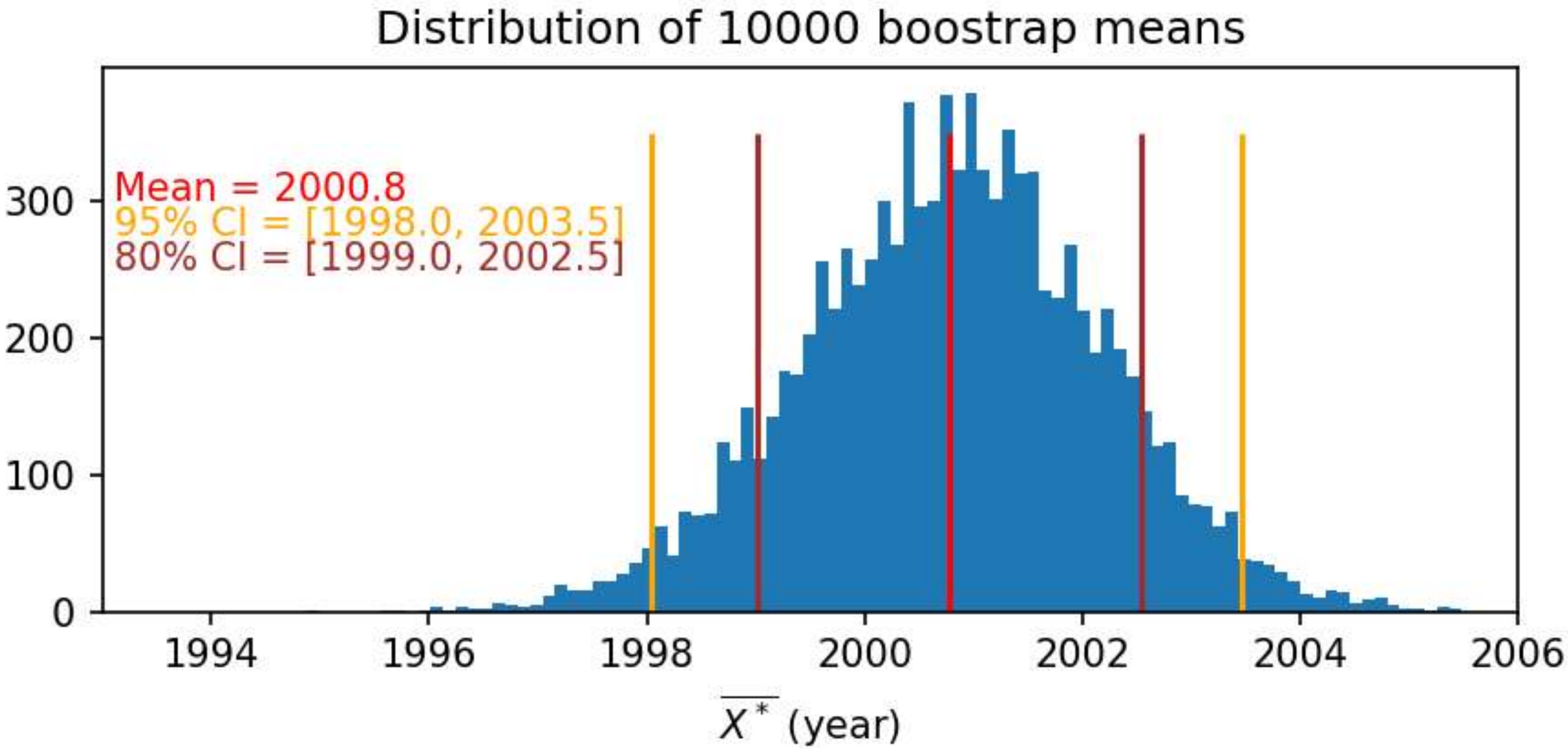
$$s^2_{boot} = \sum_{j=1}^{B} \frac{(\hat{\vartheta}_j^* - \hat{\vartheta})^2}{B-1}$$

- Find CI from Bootstrap dist.

✓ Centrality – median mean

× Extremes – max or min

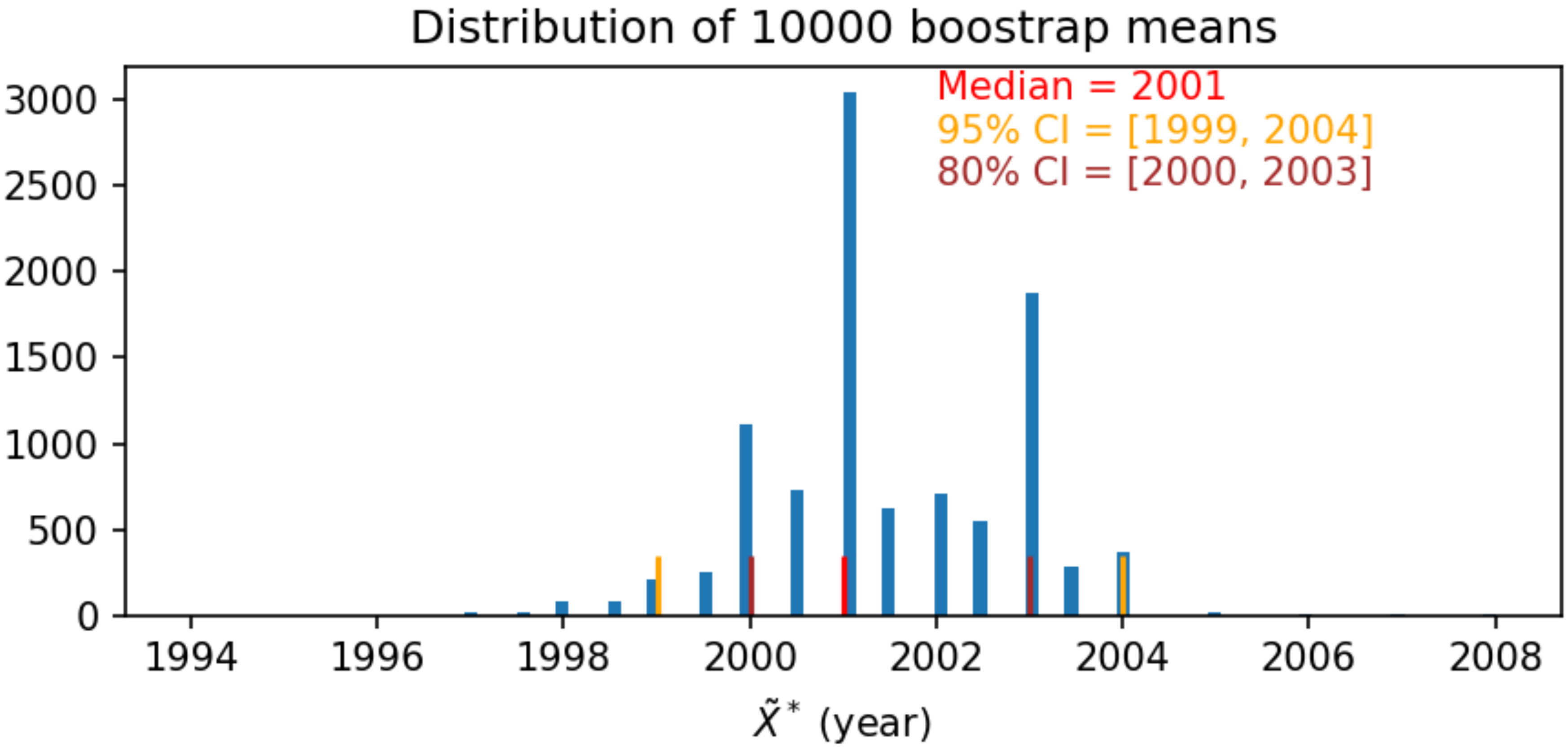
Bootstrap mean coin year



C.f theoretical estimate

(1998.1, 2003.5)

Bootstrap median coin year

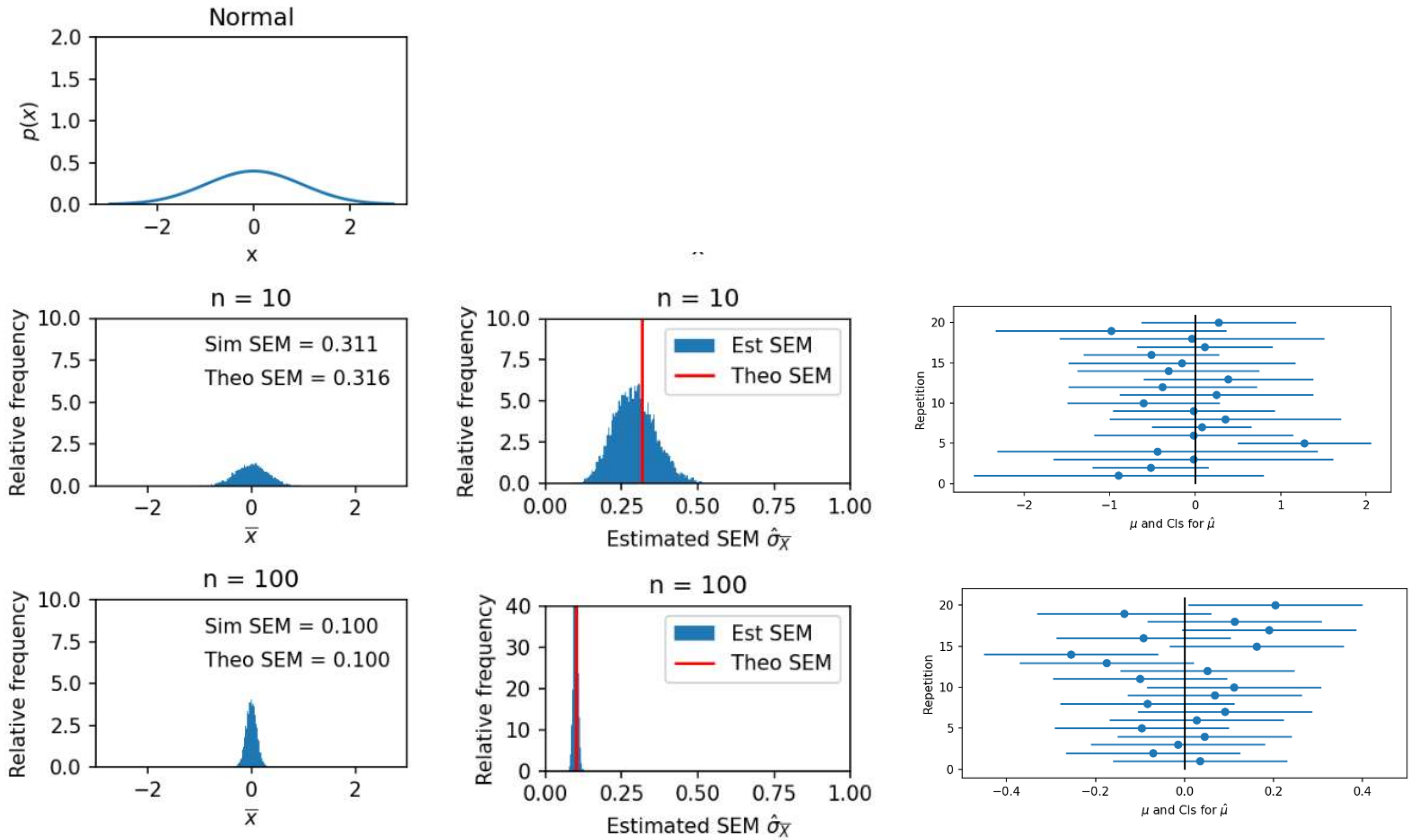


Inf2 - Foundations of Data Science: Estimation - Confidence intervals for the mean for small samples

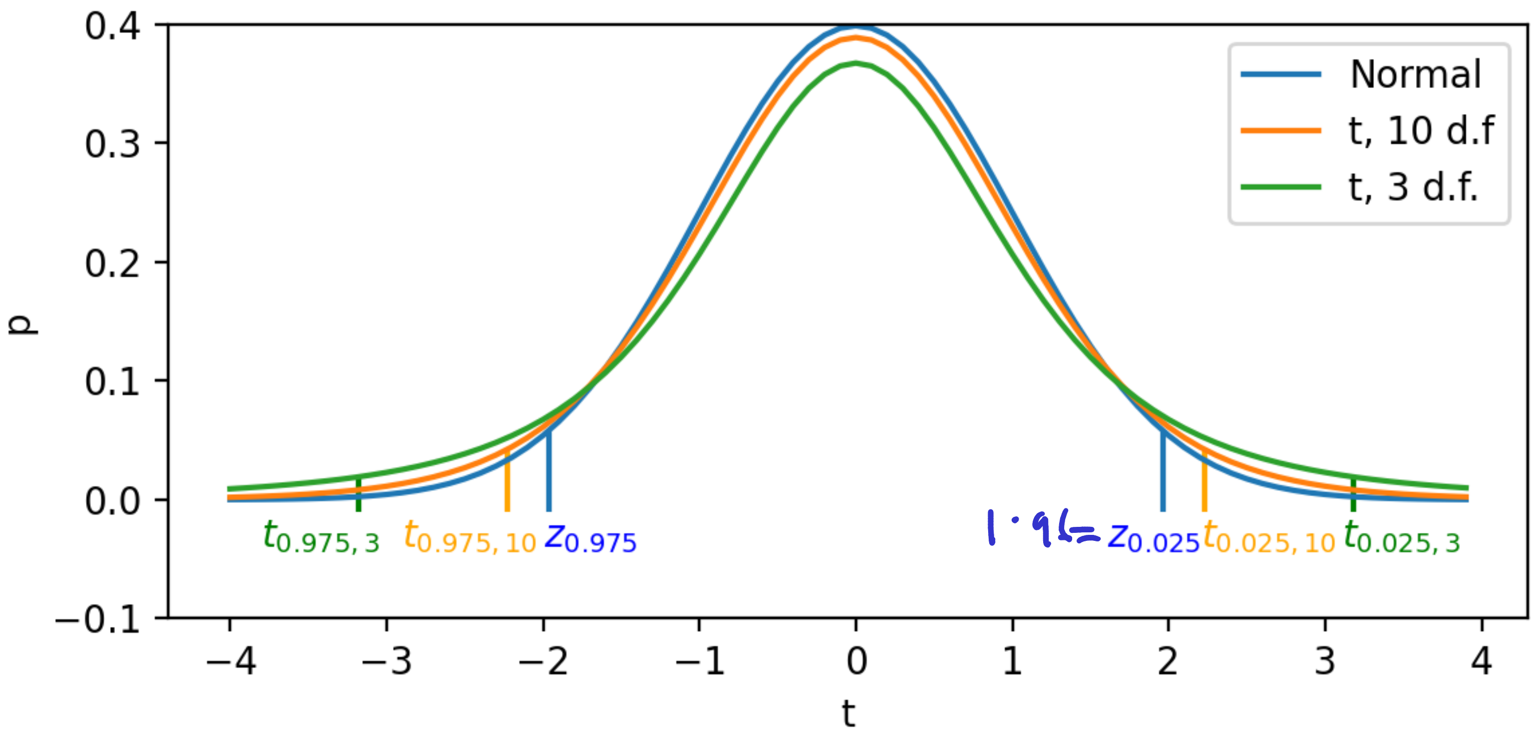
THE UNIVERSITY *of* EDINBURGH
informatics

FOUNDATIONS OF DATA SCIENCE

Small samples



The t-distribution



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

<u>t-critical value</u> $t_{\alpha,\nu}$: value of t at which, in a t-distribution with ν degrees of freedom has an area under the curve of α to its right.

Small samples

$n \leq 40$ Number of degrees of freedom $\nu = n-1$

Example

$n = 29$ coins

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

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

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

<u>t - statistic</u> $T = \frac{\bar{X} - \mu}{\hat{\sigma}_{\bar{x}}}$

Using the t-distribution to calculate a confidence interval

95% C.I $\Rightarrow$ $\alpha = 0\cdot 05$

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

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

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

$n = 29$, $\alpha = 0\cdot 05$ $\Rightarrow$ $t_{0\cdot 025,\, 29-1} = t_{0\cdot 025,\, 28}$

$= 2\cdot 281$

$t_{0\cdot 025,\, 28}\, \hat{\sigma}_{\bar{x}} = 2\cdot 281 \times 2\cdot 125 = 4\cdot 871$ 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

Japanese restaurant confidence interval calculation

$N = 92378$ $\qquad$ $n = 1000$

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

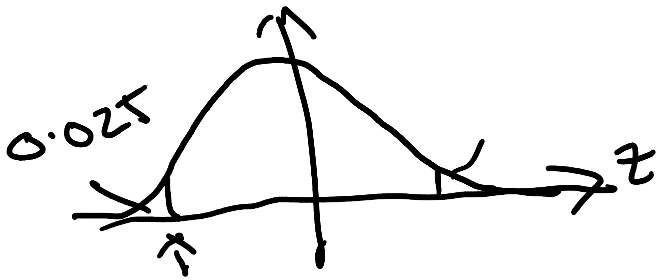
$\sigma = 25.65$ $\qquad$ $s = 27.72$

Estimated SEM $\hat{\sigma}_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{27.72}{\sqrt{1000}} = 0.88$ days

Large sample $\Rightarrow$ Normal distribution of sample mean $\Rightarrow$ "z" distribution

$95\% \Rightarrow \alpha = 0.05$

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



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

Reporting confidence intervals

$\underline{(6\cdot 34, 9\cdot 78)}$

$M = 8\cdot 06$, $CI = 6\cdot 34 - 9\cdot 78$ (95% CI)

$\hat{\mu} = 8\cdot 06 \pm 1\cdot 72$ (95 % CI)

$z_{0\cdot 025}\,\hat{\sigma}_{\bar{x}} = 1\cdot 96 \times 0\cdot 88$

$\hat{\mu} = 8\cdot 06 \pm 0\cdot 88$ (Mean ± 1. SEM)