Computer Science

CSC380: Principles of Data Science

Statistics 2

Xinchen Yu

Outline

- Interval estimation

- Hypothesis testing

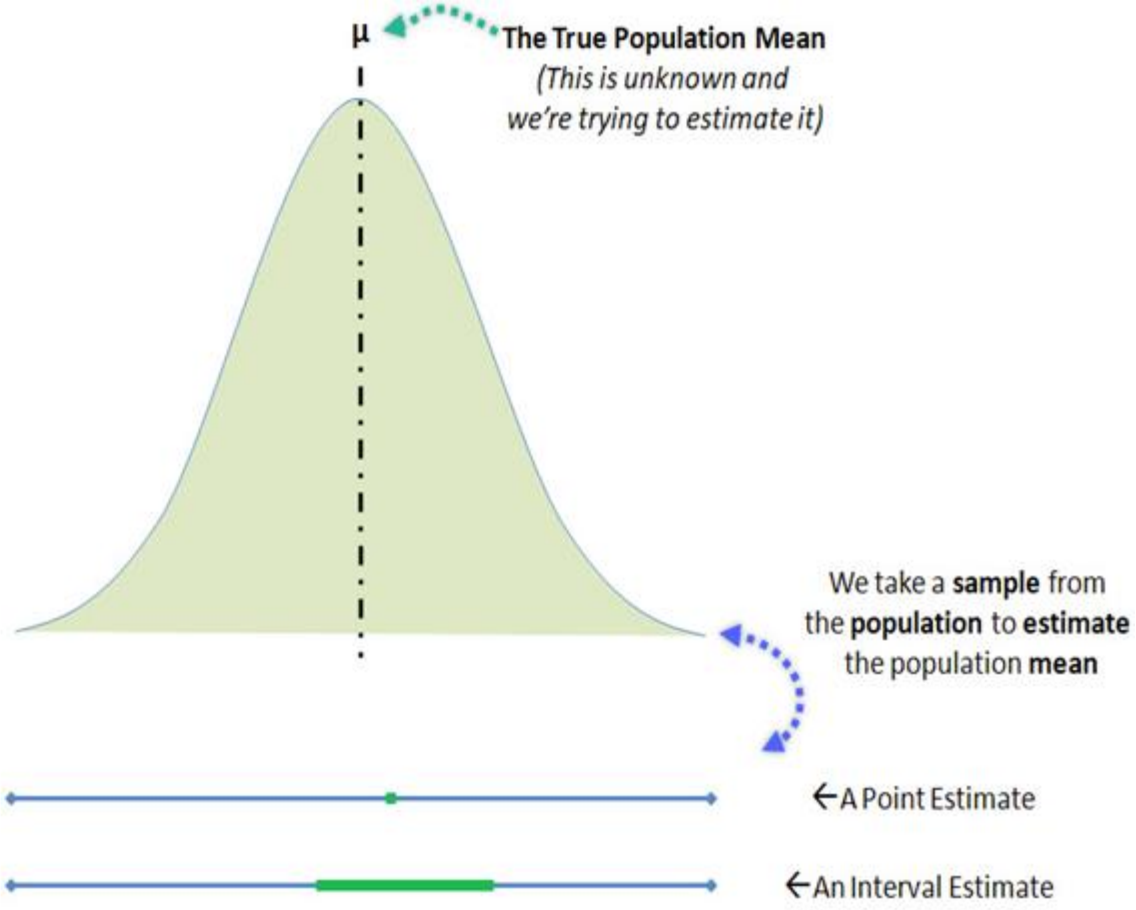
Interval estimation

Motivation

- Point estimation:
  - "Given the data, I estimate the bias of the coin to be 0.73"
  - "Given the data, I estimate the mean height of UA students to be 172cm"

- In many applications, we'd like to make statements with uncertainty quantifications
  - "Given the data, I estimate the bias of the coin to be 0.73 $\pm$ 0.05"
  - "Given the data, I estimate the mean height of UA students to be 172 $\pm$ 2cm"

- This is called *interval estimation*



Interval Estimation: basic setup

$$\theta \ \rightarrow X_1, \dots, X_n \ \rightarrow I_n = [\hat{\theta}_n \pm b_n]$$

data generation process

Confidence Interval (CI) for θ

Examples

Coin toss: $\theta = p, \ X_1, \dots, X_n \sim \text{Bernoulli}(p)$

Student height: $\theta = \mu, X_1, \dots, X_n \sim N(\mu, 8^2)$

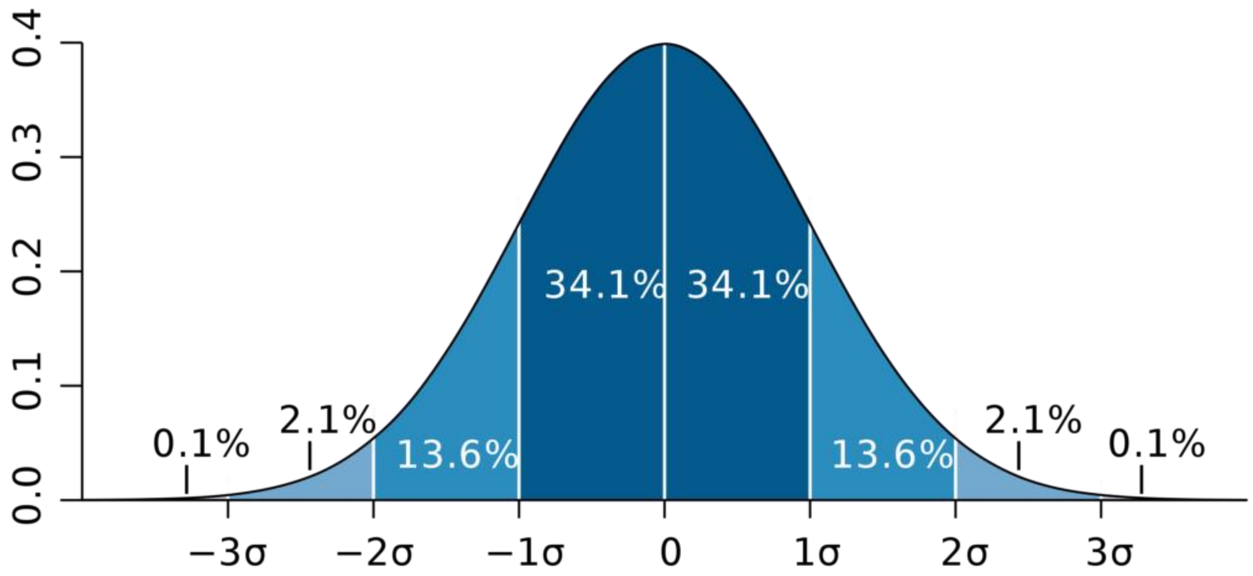
Goal: construct I_n using data, such that with 95% confidence (say),

$$\theta \in I_n$$

We will mostly focus on estimating θ = population mean, and will take $\hat{\theta}_n$ = sample mean.

How to choose b_n? **uncertainty of our estimate**

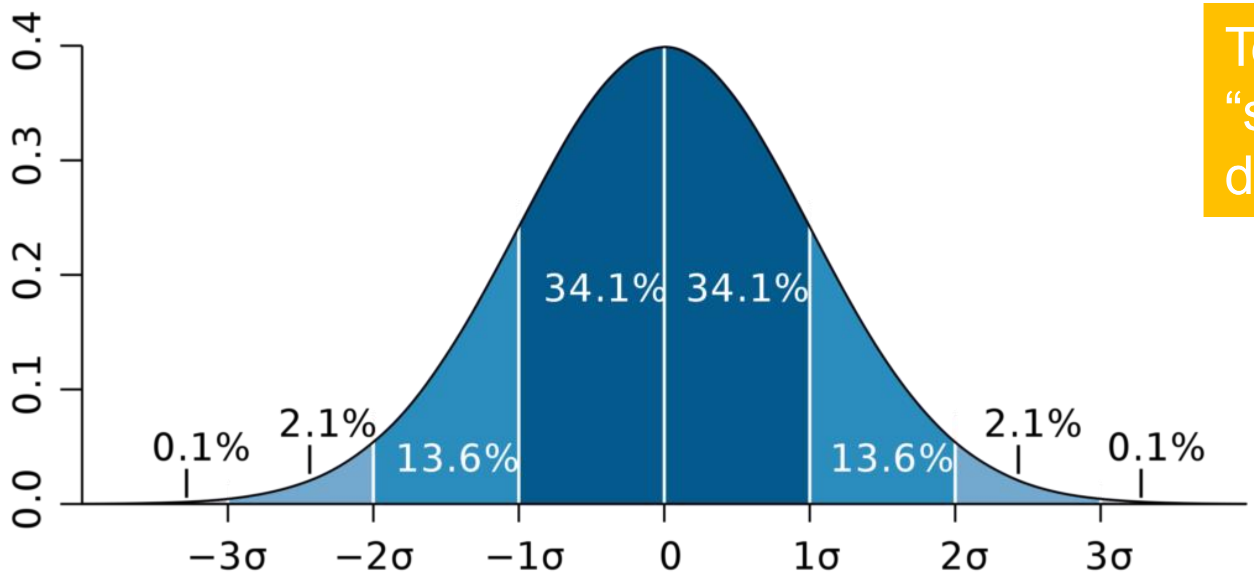
Recall: Normal distribution



For $X \sim N(\mu, \sigma^2)$, we can transform it into $X - \mu \sim N(0, \sigma^2)$

- the area under a normal distribution curve (PDF) represents probability.
- the total area under the curve is equal to 100%.
- the area within a certain range of values corresponds to the probability of a random variable falling within that range.

Recall: Normal distribution



Terminology: "standard" normal distribution := $N(0,1)$

Fact If $X \sim N(\mu, \sigma^2)$ or $X - \mu \sim N(0, \sigma^2)$, then

$$P(-1.96\sigma \leq X - \mu \leq 1.96\sigma) = 0.95$$

In words, with 95% confidence, X falls within 1.96 standard deviation of μ

$$P(X - 1.96\sigma \leq \mu \leq X + 1.96\sigma) = 0.95$$

i.e, with 95% confidence, μ falls within 1.96 standard deviation of X

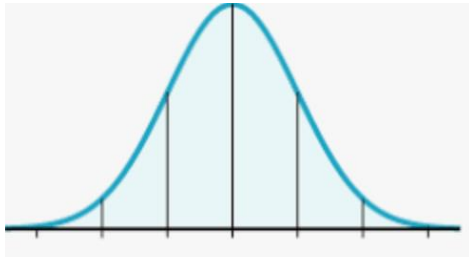
$[X - 1.96\sigma, X + 1.96\sigma]$ is a 95% confidence interval for μ

Constructing confidence interval

- **We know if $X \sim N(\mu, \sigma^2)$, then $[X - 1.96\sigma, X + 1.96\sigma]$ is a 95% CI for μ**

- **Fact**: Let $X_1, \ldots, X_n$ be iid with mean μ and variance σ^2. Then for large n, the sample mean $\bar{X}_n$ roughly follow a normal distribution:

$$\bar{X}_n \approx N\left(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right)$$



Corollary with 95% confidence, μ lies within $1.96\frac{\sigma}{\sqrt{n}}$ of $\bar{X}_n$

Our confidence interval for μ: $I_n = [\bar{X}_n - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X}_n + 1.96\frac{\sigma}{\sqrt{n}}]$

Example: UA student height

Example Assume that UA students' heights (in centimeters) follow $N(\mu, 8^2)$, and we observe 4 students' heights:

163, 171, 179, 167

Find a 95% confidence interval for μ

Solution

Sample mean

population stddev

sample size

our CI for μ: $I_n = [\bar{X}_n - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X}_n + 1.96 \frac{\sigma}{\sqrt{n}}]$

$= 170$ $\quad \sigma = 8$ $\quad$ n=4

Plugging in all values, $I_n = [170 \pm 7.84] = [162.1, 177.8]$

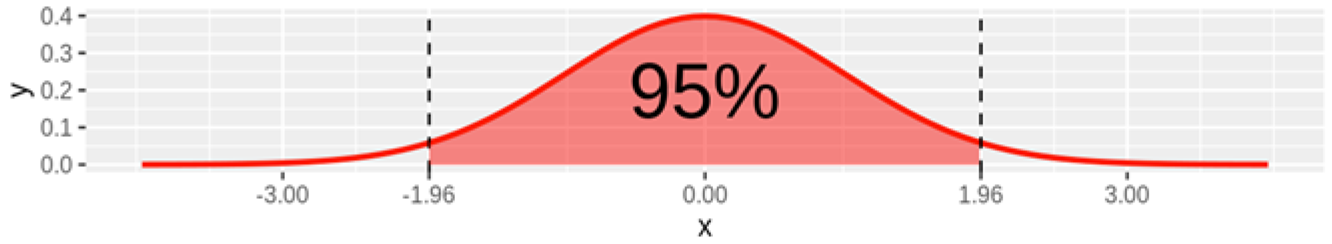
Confidence intervals: extensions

Given if $X \sim N(\mu, \sigma^2)$ or $X - \mu \sim N(0, \sigma^2)$, then

$$P(-1.96\sigma \leq X - \mu \leq 1.96\sigma) = 0.95$$

Where does the 1.96 come from?

st.norm.ppf(0.975) gives 1.96



Fact If $X \sim N(\mu, \sigma^2)$, then

Φ: standard normal CDF

$$P(-k\,\sigma \leq X - \mu \leq k\,\sigma) = 2\Phi(k) - 1 = p$$

$$2\Phi(k) - 1 = 0.95 \Rightarrow \; k = \Phi^{-1}\left(\frac{0.95+1}{2}\right) = \Phi^{-1}(0.975) = 1.96$$

k: $\left(\frac{1+p}{2}\right)$-quantile of *the standard normal distribution*

CI for μ: $I_n = [\bar{X}_n - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X}_n + 1.96\frac{\sigma}{\sqrt{n}}]$

Confidence intervals: extensions

- What if we'd like to find 99% confidence interval? 99.9%? 90%?

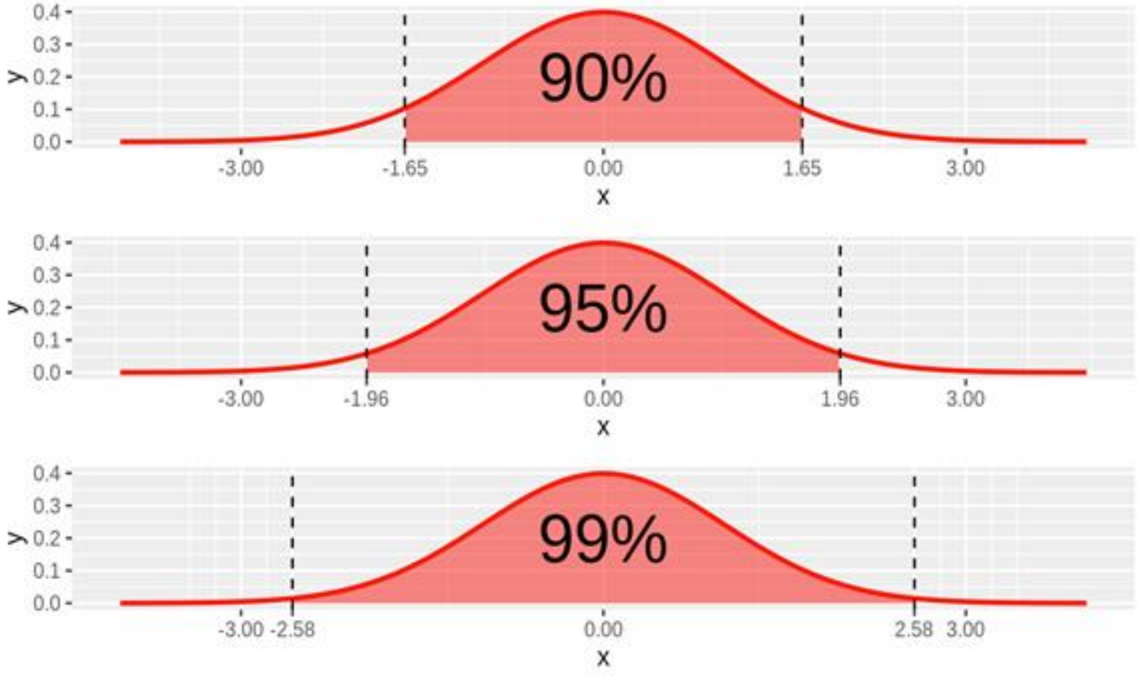
Fact If $X \sim N(\mu, \sigma^2)$, then

$$P(-k\,\sigma \le X - \mu \le k\,\sigma) = 2\Phi(k) - 1 = p$$

Our p confidence interval for μ:

$$I_n = [\bar{X}_n \pm \Phi^{-1}\left(\frac{p+1}{2}\right)\frac{\sigma}{\sqrt{n}}] = [\bar{X}_n \pm k\,\frac{\sigma}{\sqrt{n}}]$$

p	$\mathrm{k} = \Phi^{-1}\left(\frac{p+1}{2}\right)$
0.95	1.96
0.99	2.58
0.999	3.29



Example: UA student height

Example Assume that UA students' heights (in centimeters) follow $N(\mu, 8^2)$, and we observe 4 students' heights:

163, 171, 179, 167

Find 99%, 99.9% confidence intervals for μ

Solution

our p-CI for μ: $I_n = [\bar{X}_n \pm \Phi^{-1}\left(\frac{p+1}{2}\right)\frac{\sigma}{\sqrt{n}}]$

$p = 0.99 \Rightarrow [159.7, 180.3]$

$p = 0.999 \Rightarrow [156.9, 183.1]$

p	$\Phi^{-1}\left(\frac{p+1}{2}\right)$
0.95	1.96
0.99	2.58
0.999	3.29

Confidence interval: observations

$$p\text{-CI for } \mu\text{: } I_n = [\bar{X}_n \pm \Phi^{-1}\left(\frac{p+1}{2}\right)\frac{\sigma}{\sqrt{n}}]$$

$p = 0.95 \Rightarrow [162.1, 177.8]$
$p = 0.99 \Rightarrow [159.7, 180.3]$
$p = 0.999 \Rightarrow [156.9, 183.1]$

The center is always at $\bar{X}_n$

The width of the interval depends on:

- Sample size n: width smaller when n larger
- Confidence level p: width larger when p closer to 1
- Population stddev σ: width larger when σ large (more noise)

What if σ is unknown?

- We will address this soon..

Is confidence = probability?

Example Assume that UA students' heights (in centimeters) follow $N(\mu, 8^2)$, and we observe 4 students' heights:

163, 171, 179, 167

we found that a 95% CI for μ is $[162.1, 177.8]$

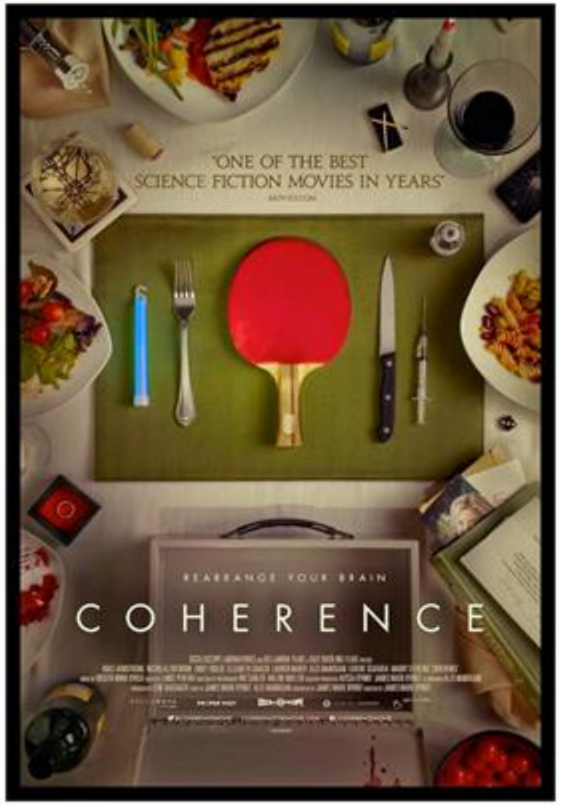
Can we say "with probability 95%, the population mean height μ lies in interval $[162.1, 177.8]$"?

No! This is a common misinterpretation

- μ is deterministic, and $[162.1, 177.8]$ is deterministic,
- Proposition $\mu \in [162.1, 177.8]$ is either true or false!

Then, what does "95% probability" mean?

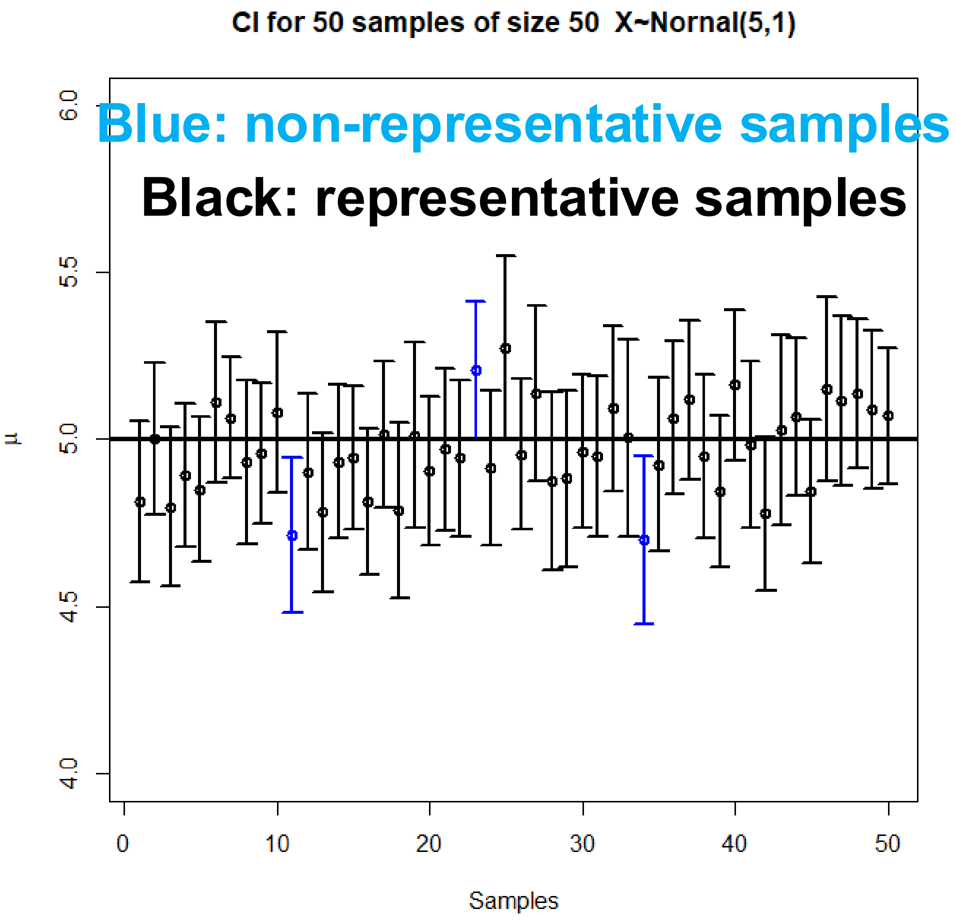
Interpreting CI (think of parallel universe…)



Multiple different universes…

Caveat: interpreting confidence intervals

Recommended point of view:



universe 1: get sample 1, and confidence interval 1

universe 2: get sample 2, and confidence interval 2

……

universe 50: get confidence interval 50

True: With probability 0.95 *over the draw of a sample*, $[\bar{X}_n \pm 1.96\frac{\sigma}{\sqrt{n}}]$ contains μ

Confidence interval: interpretation

Example Assume that UA students' heights (in centimeters) follow $N(\mu, 8^2)$, and we observe 4 students' heights:

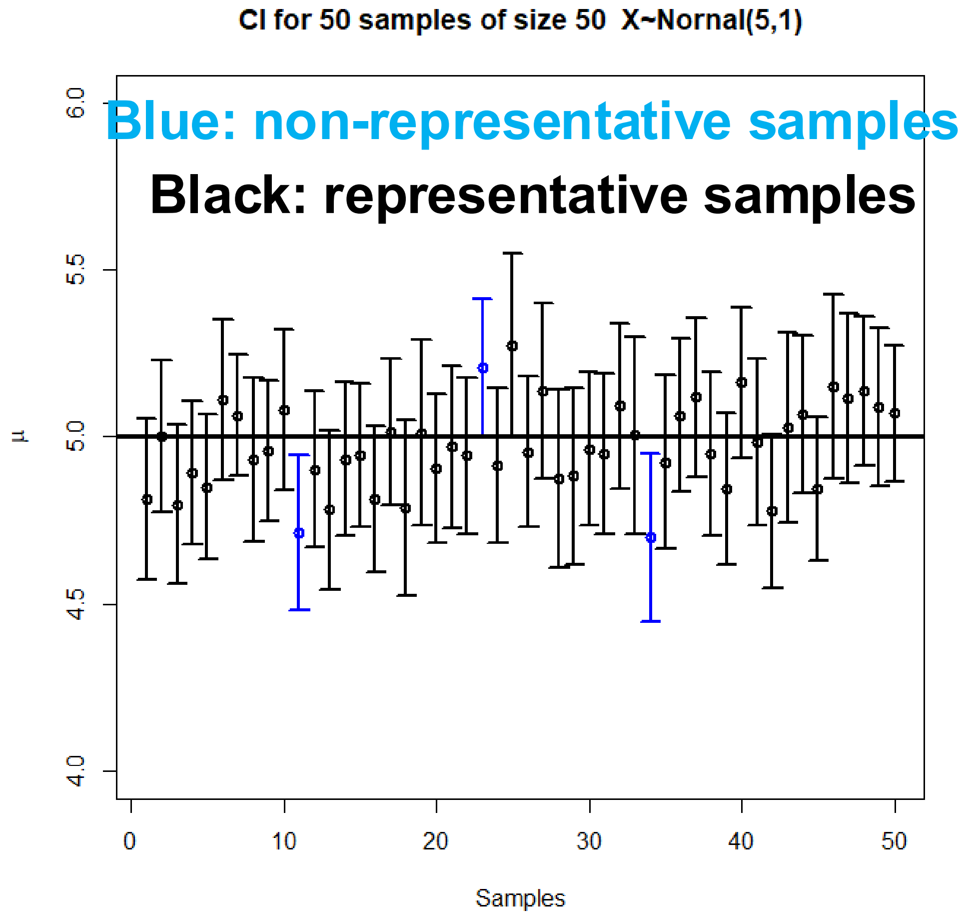
163, 171, 179, 167

True: With probability 0.95 *over the draw of a sample*, $[\bar{X}_n \pm 1.96 \frac{\sigma}{\sqrt{n}}]$ contains μ

50 draws of samples

$\Rightarrow$ 50 CIs

$\Rightarrow$ expect 50$\times$ $95\% = 47.5$ CI's to contain μ



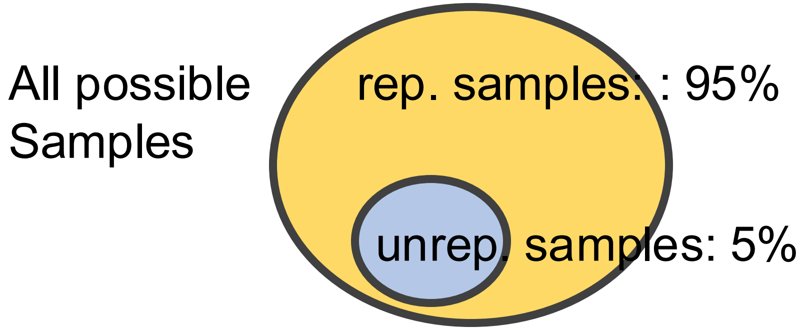
Confidence interval: interpretation

Example Assume that UA students' heights (in centimeters) follow $N(\mu, 8^2)$, and we observe 4 students' heights:

163, 171, 179, 167

True: With probability 0.95 *over the draw of a sample*, $[\bar{X}_n \pm 7.84]$ contains μ

As long as we are not extremely unlucky / our sample is mildly representative, my CI contains μ



Confidence interval for mean with unknown variance

Example Assume that UA students' weights (in kgs) follow $N(\mu, \sigma^2)$, and we observe 4 students' weights:

60, 65, 70, 75

Find a 95% confidence interval for μ

Note The CI construction before $[\bar{X}_n - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X}_n + 1.96\frac{\sigma}{\sqrt{n}}]$ no longer works, since σ is *unknown*

How to fix this?

The student-t distribution

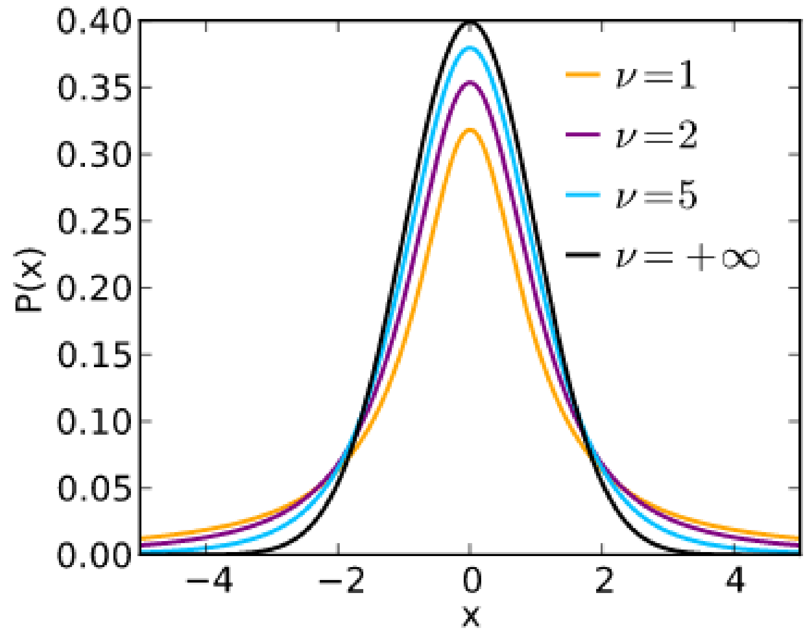
- $[\bar{X}_n - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X}_n + 1.96\frac{\sigma}{\sqrt{n}}]$ no longer works: σ is unknown

Fact $X_1, \ldots, X_n$ is an iid sample with unknown μ & σ^2.

Let $sample\ stddev$: $\hat{\sigma}_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}$. Then, approximately:

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\hat{\sigma}_n} \sim \text{student-t}(n-1)$$

t-statistic   degree of freedom

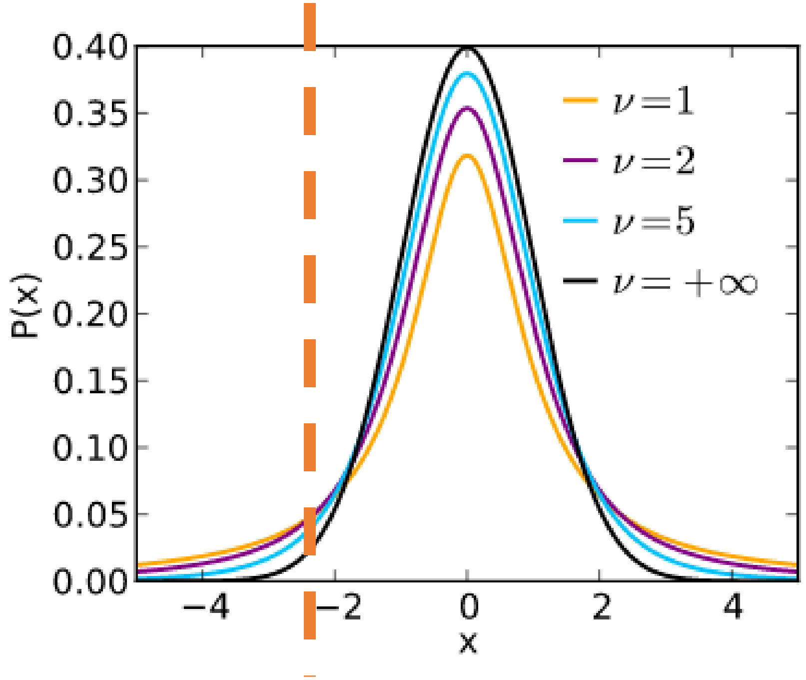


student-t(ν) is a family of distributions

The student-t distribution

student-t(ν) distribution family

- goes to Gaussian when ν is large
- generally has heavier tail than Gaussian




```
import scipy.stats as st

st.t.ppf(0.975,df=3)
=> 3.18

st.t.ppf(0.975,df=5)
=> 2.57

st.t.ppf(0.975,df=10)
=> 2.23

st.t.ppf(0.975,df=100)
⇒ 1.98
```

Recall:
st.norm.ppf(0.975) gives 1.96

CI: $\left[\bar{X}_n - w\frac{\hat{\sigma}_n}{\sqrt{n}}, \bar{X}_n + w\frac{\hat{\sigma}_n}{\sqrt{n}}\right]$, w: $\left(\frac{1+p}{2}\right)$−quantile of the $t(n-1)$ distribution

Confidence interval for mean with unknown variance

Example Assume that UA students' weights (in kgs) follow $N(\mu, \sigma^2)$, and we observe 4 students' weights:

$$60, 65, 70, 75$$

st.t.ppf(0.975,df=3)
=> 3.18

Find a 95% confidence interval for μ

Solution With 95% confidence,

$$\Rightarrow \mu \in \left[\underbrace{\bar{X}_4}_{=67.5} - 3.18\overbrace{\frac{\hat{\sigma}_4}{\sqrt{4}}}^{=6.45}, \bar{X}_4 + 3.18\frac{\hat{\sigma}_4}{\sqrt{4}}\right]$$

Plugging data,

our CI is $[67.5 - 10.3, 67.5 + 10.3] = [57.2, 77.8]$ **Our confidence interval**

Confidence interval for mean with unknown variance

General result given a sample $X_1, \dots, X_n$ drawn from a distribution with mean μ, a p-confidence interval (e.g. p=95%) is

$$\left[\bar{X}_n - w\frac{\hat{\sigma}_n}{\sqrt{n}}, \bar{X}_n + w\frac{\hat{\sigma}_n}{\sqrt{n}}\right],$$

where w is the $\left(\frac{1+p}{2}\right)$-quantile of the $t(n-1)$ distribution

Example p=0.95, n=4 $\Rightarrow w = 3.18$

```
st.t.ppf(0.975,df=3)
=> 3.18
```

p=0.99, n=4 $\Rightarrow$ $w = 5.84$

p=0.99, n=9 $\Rightarrow w = 3.35$

Confidence interval: closing remarks

How to construct confidence intervals for μ?

- When σ is known
  - $\boldsymbol{CI}$: $\left[\bar{X}_n - k\frac{\sigma}{\sqrt{n}}, \bar{X}_n + k\frac{\sigma}{\sqrt{n}}\right]$, k: $\left(\frac{1+p}{2}\right)$−quantile of the standard normal distribution

    st.norm.ppf((1+p)/2)

- When σ is unknown
  - $\boldsymbol{CI}$: $\left[\bar{X}_n - w\frac{\hat{\sigma}_n}{\sqrt{n}}, \bar{X}_n + w\frac{\hat{\sigma}_n}{\sqrt{n}}\right]$, w: $\left(\frac{1+p}{2}\right)$−quantile of the $t(n-1)$ distribution

    st.t.ppf((1+p)/2,df=n-1)