ECS7024 Statistics for Artificial Intelligence and Data Science

Topic 11: Confidence Intervals and Hypothesis Testing

William Marsh

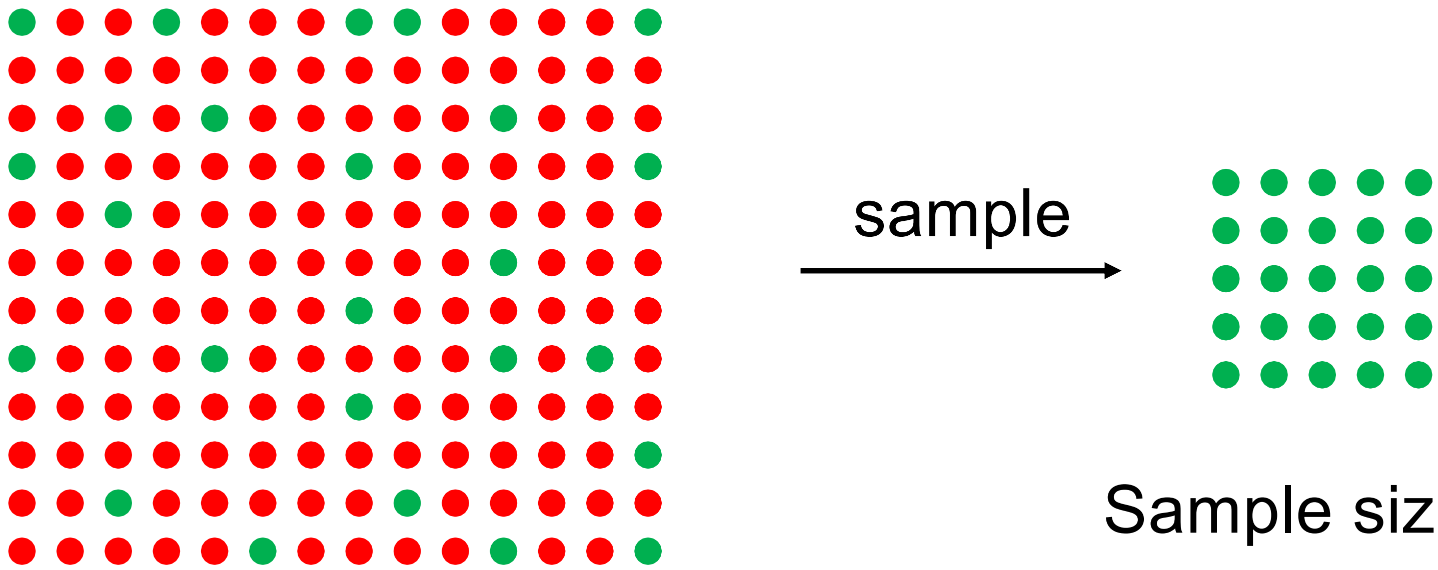
Outline

- Recap of Sampling
  - Estimation and uncertainty
  - Ways to estimate
- Estimating the mean and variance of a Normal
- Sampling distributions
  - Standard error
- Confidence intervals
- Hypothesis testing
  - Null hypothesis
  - Type I & II errors
  - Example: Student's t-test
- Issues with CI and p-values

Sampling

Recap sampling

Population and Sample



Population size N

Sample size n

- Sample from a population
- Measure the sample (e.g. political preference)
- Statistical inference about population

Sampling Introduces Variation

- When you toss a coin
  - P(Heads) = P(Tails) = 0.5
  - BUT you do not always get 5 heads in 10 tosses

- When we generate data (i.e. sample) from an unknown distribution
  - We can calculate the statistics (parameters) of the sample
  - BUT this only gives an estimate of the true parameters (of the distribution)

Statistical Inference: Two Problems

- Estimate a parameter from a sample
  - The mean and variance
  - A rate (probability in a binomial)
  - A regression coefficient

- Say how certain we can be that the estimate is near the true value (in the population)

Population has
Normal Distribution

Different Ways to Estimate

- No unique way to estimate a parameter from some data

- Maximum likelihood estimation (MLE)
  - Choose θ to maximise p(Data | θ)

- Unbiased estimator
  - Average of repeated estimation is θ

- MLE is sometimes unbiased, but not always

Estimating the Mean and Variance

A ‘sample statistic’ (e.g. a mean) is calculated from the sample

Estimate the Mean

- $\bar{x}$ is an estimate of μ
- N is sample size

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- This estimate is unbiased
- ... and an MLE

Estimates of the Variance

- Sample variance s^2 estimates $\boldsymbol{\sigma}^2$
  - Estimate uses $\bar{x}$

- Unbiased estimate

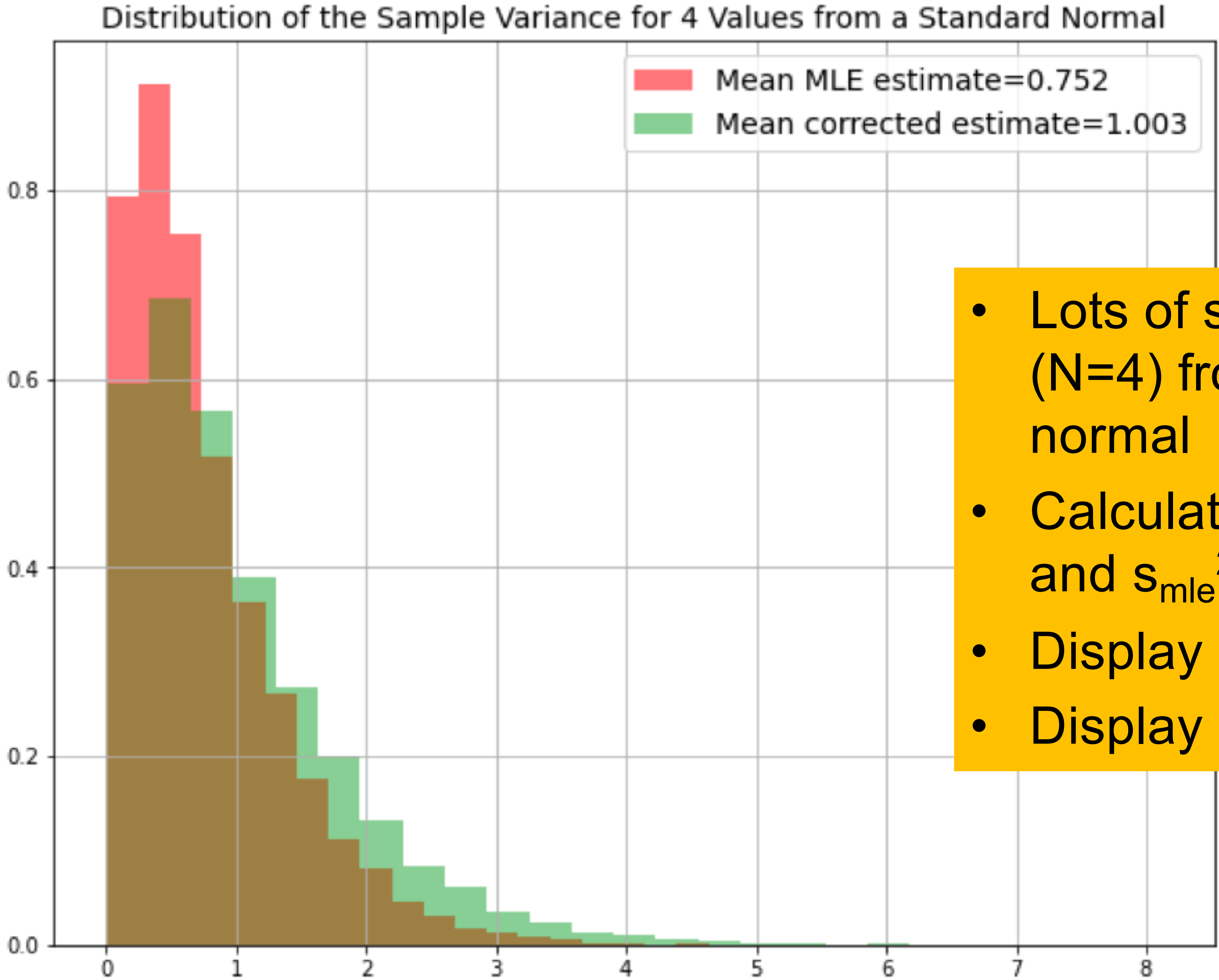
$$s^2 = \frac{\sum_1^N (x_i - \bar{x})^2}{N - 1}$$

Bessel's correction
(degrees of freedom)

- MLE estimate

$$s_{mle}^2 = \frac{\sum_1^N (x_i - \bar{x})^2}{N}$$

Simulation

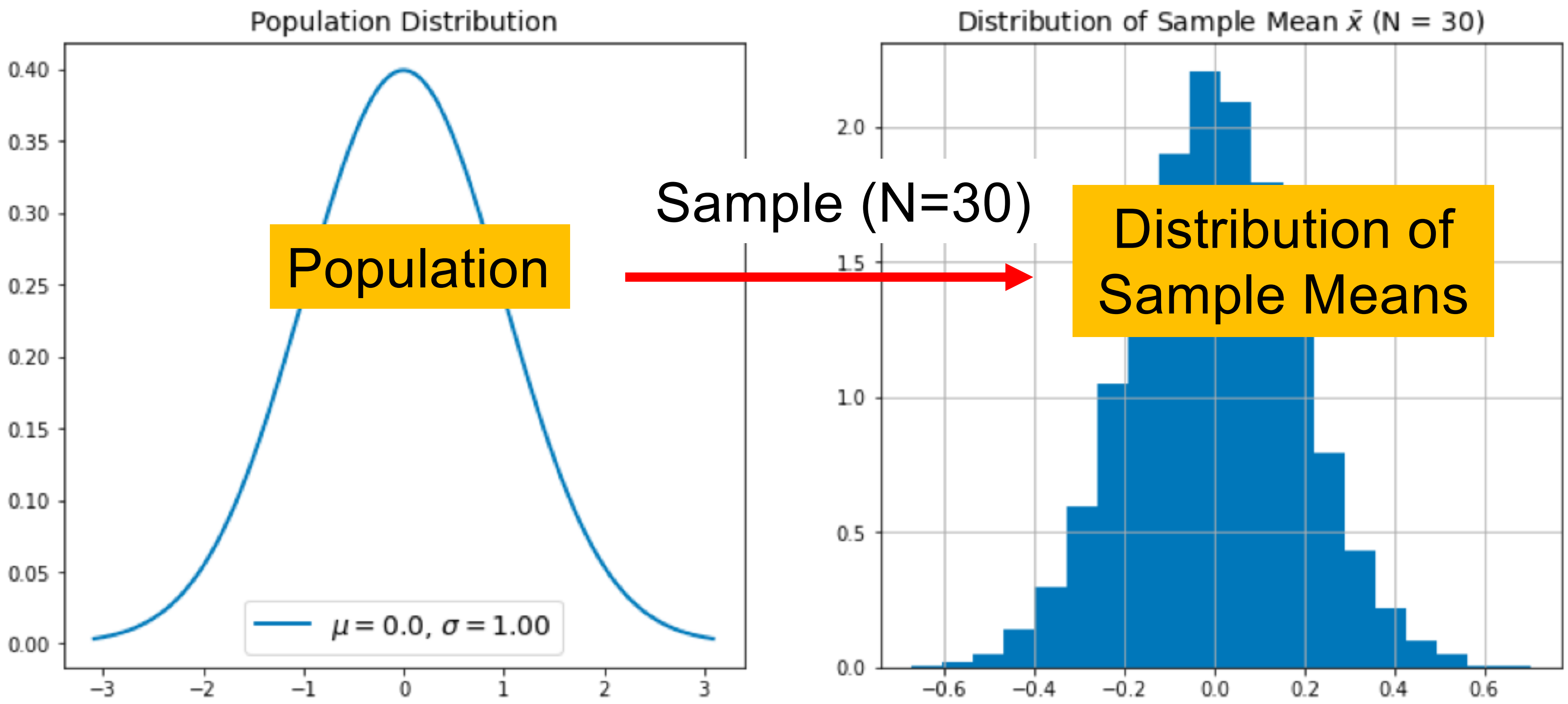


- Lots of samples (N=4) from standard normal
- Calculate both s^2 and s_{mle}^2
- Display distributions
- Display means

Sampling Distribution

Sampling Distribution

- Distribution of the statistic estimated from the sample



Standard Error

- How wide is sampling distribution?

- Standard error of the mean
  - Standard deviation of the sampling distribution

Population deviation

Sample size

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$$

- Estimate by:

$$s_{\bar{x}} = \frac{s}{\sqrt{N}}$$

Approach

- Work out the form of the sample distribution (mathematics)
  - Often normal (central limits theorem)
- Use sample estimates ($\bar{x}$ and s^2) since $\boldsymbol{\mu}$ and σ^2 not known
- Look at width of distribution
  - Use standard error to ...
  - .... estimate ‘confidence interval’ (error)

Quiz 1

Confidence Intervals

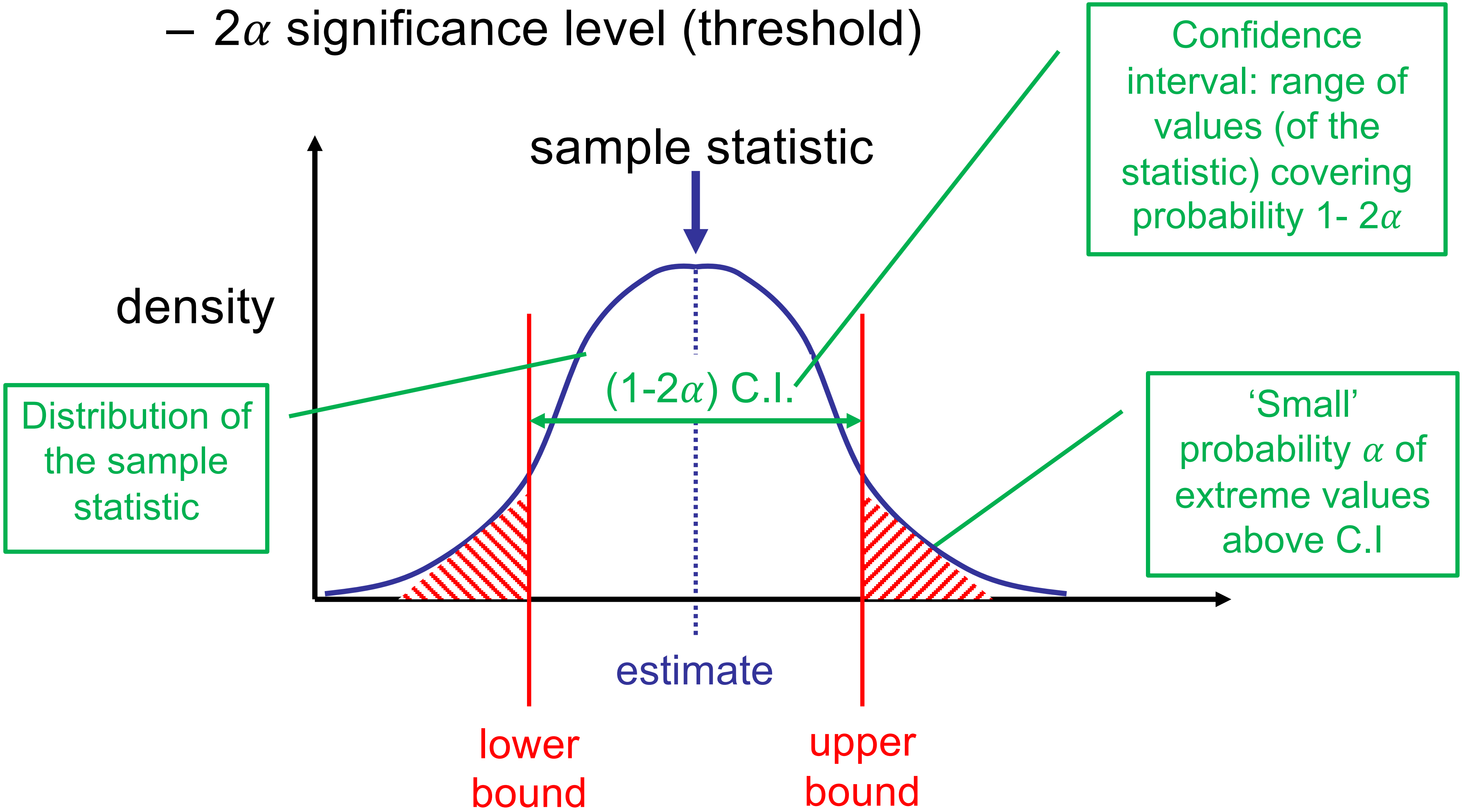
Summary So Far

- We know the sample statistic comes from a distribution
  - Given a population, many samples possible

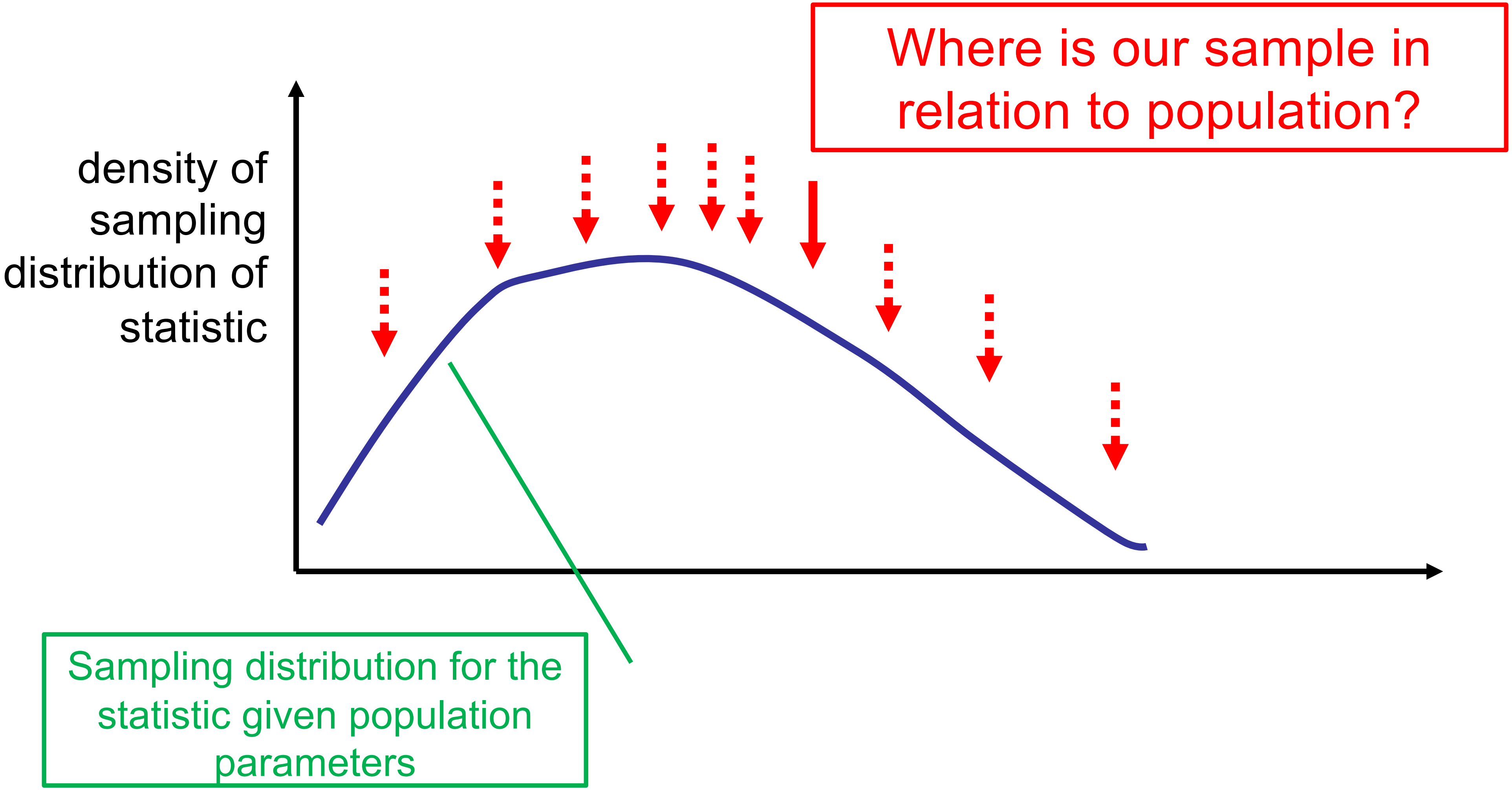
- *We have only one sample value*
  - *Where in the sampling distribution?*
  - *How can a confidence interval be based on a single sample?*

Confidence Intervals

- A statistic, a sampling distribution and probability threshold
  - 2α significance level (threshold)

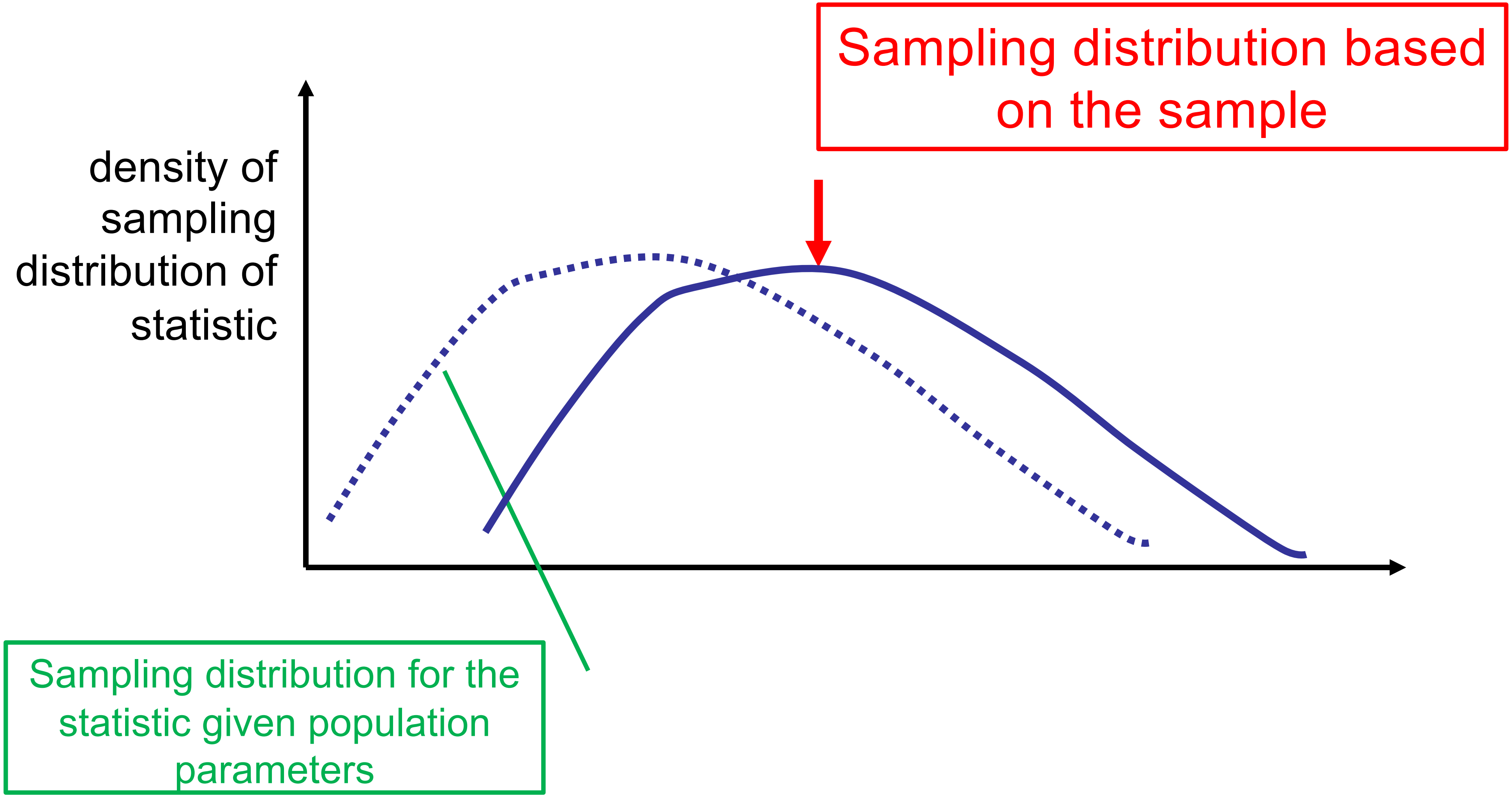


Problem: True Sampling Distribution Depends on Population



Solution: Sampling Distribution based on Sample

- *Let's cheat*



Summary So Far II

- We know the sample statistic comes from a distribution
  - Given a population, many samples possible

- Use the sample statistics to locate the sampling distribution
- Confidence interval
  - Range of a value covering ‘most’ of the area under the sampling distribution

Normal Approximation

z score

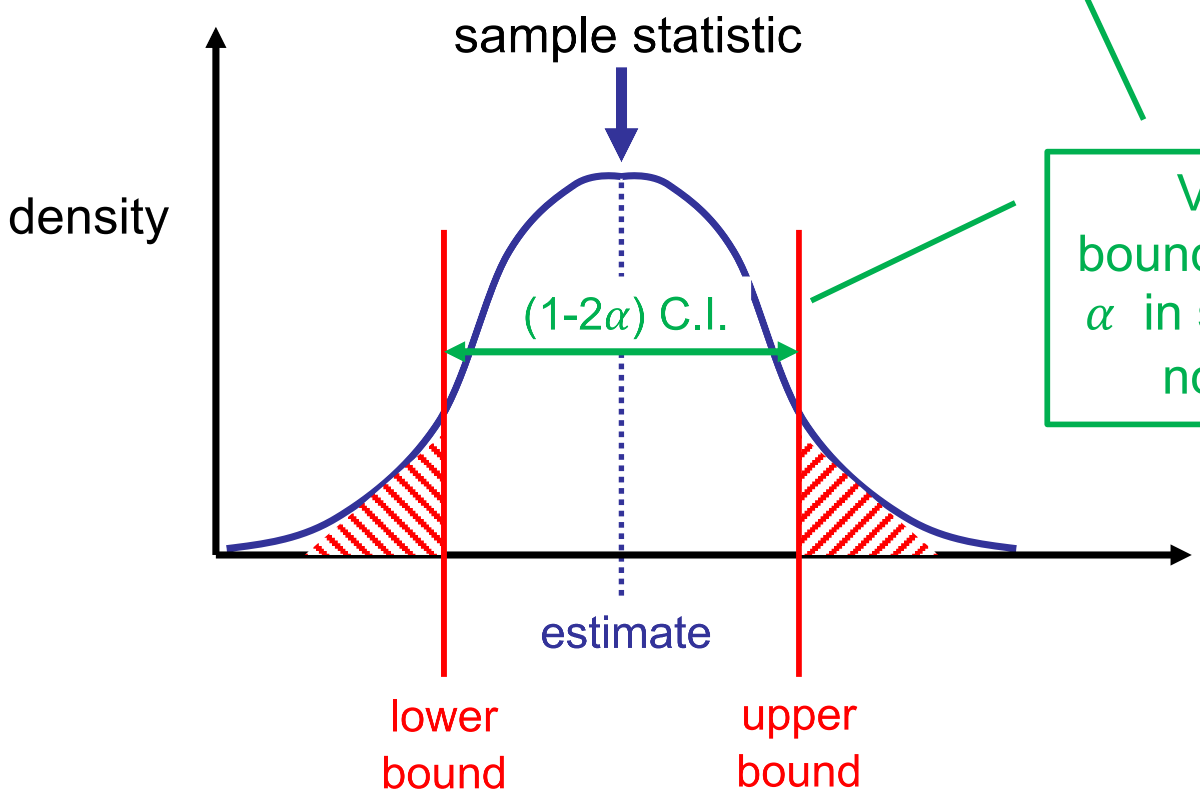
Using Normal to Estimate CI

- Central limits theorem
- Interval

$$\bar{x} - z^* \frac{s}{\sqrt{N}} \text{ to } \bar{x} + z^* \frac{s}{\sqrt{N}}$$

Standard error

Value bounding area α in standard normal



Experiment

- Random data sample (N=30) from Normal
- Calculate 95% CI
- Look at whether the CI contains the true mean (zero)
- Repeat

```
95% CI does NOT contain true value in 6.27% of samples
```

```
95% CI does NOT contain true value in 5.77% of samples
```

This is the meaning of a CI:

In 95% of samples, the 95% CI will contain the true mean

Quiz 2

Student's t-Distribution

Sampling distribution similar to normal,
but variance unknown

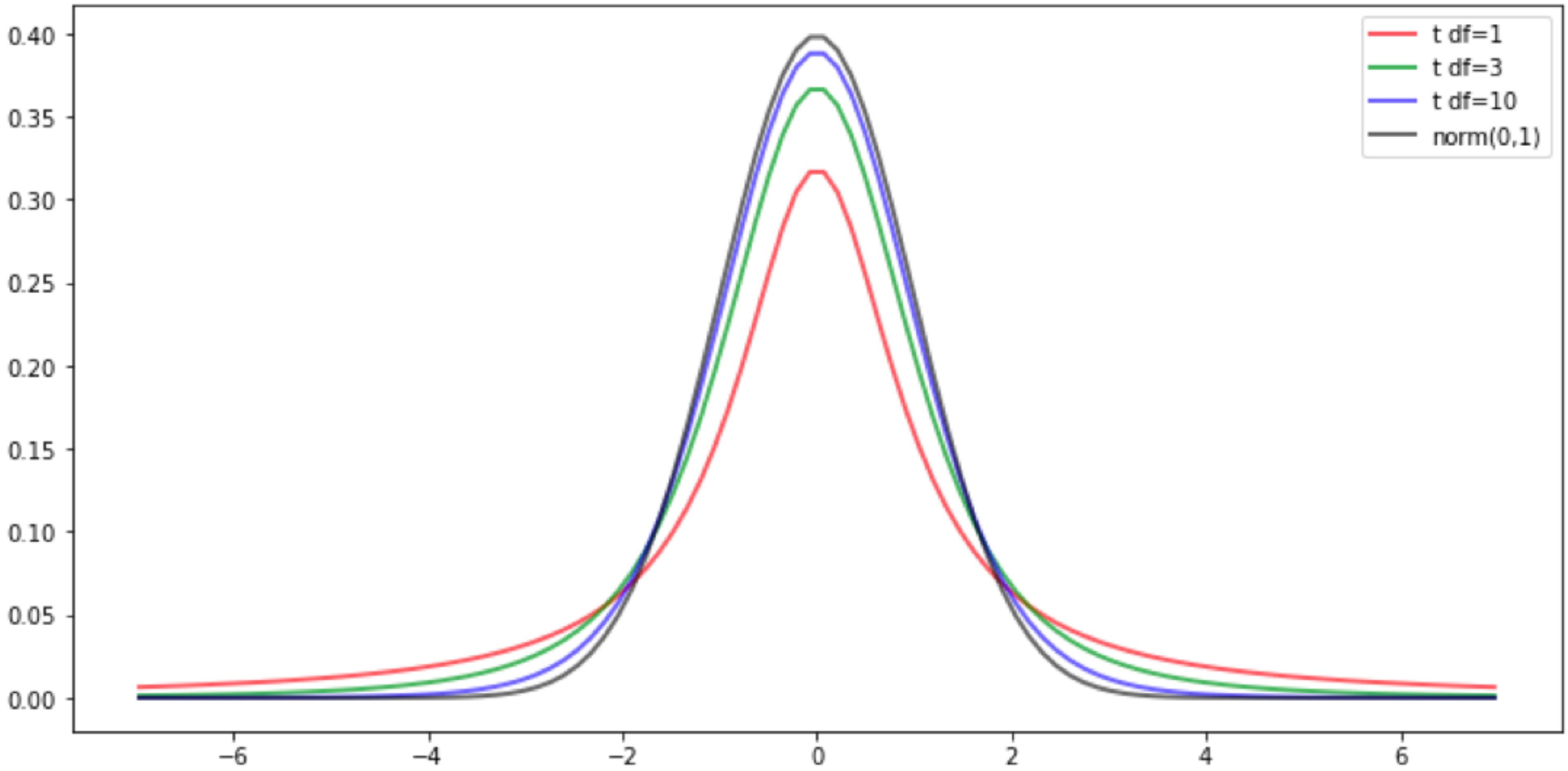
Student’s t-Distribution

- Normal is only approximate when the variance is unknown (the usual case)
  - Good approximation for sample size >= 30
- Student’s t
  - Symmetric
  - Fat tails
  - Parameter ‘degrees of freedom’ (equal to N-1)

- *See notebook for comparison of CI using Normal and Student’s t*

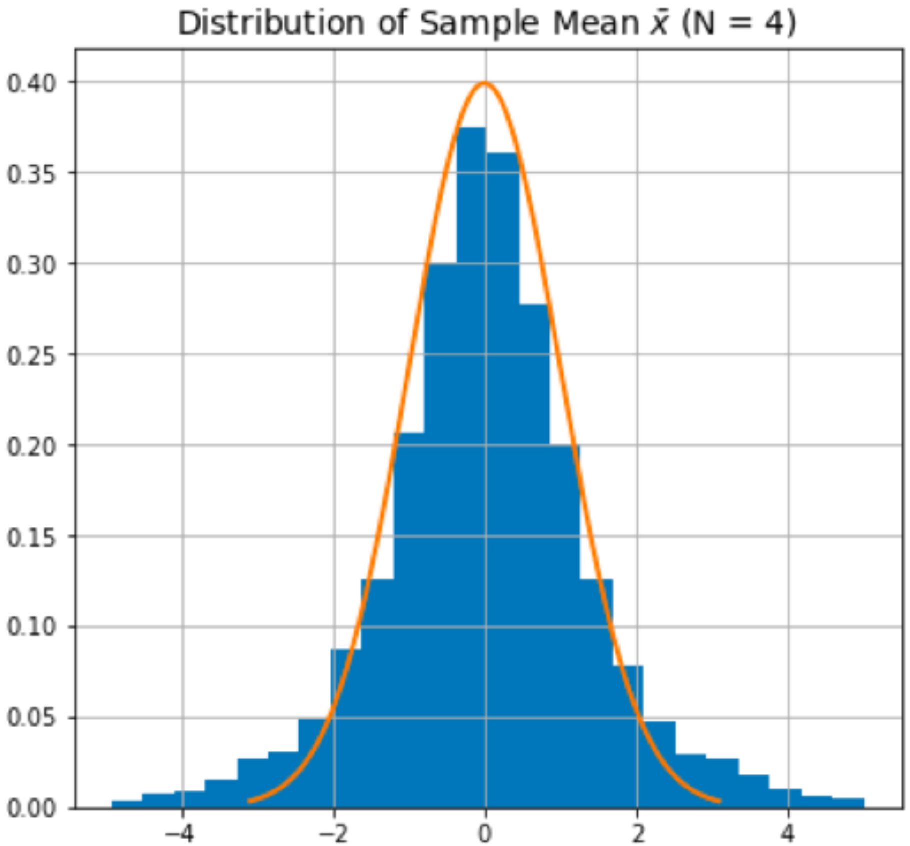
Student's t-Distribution

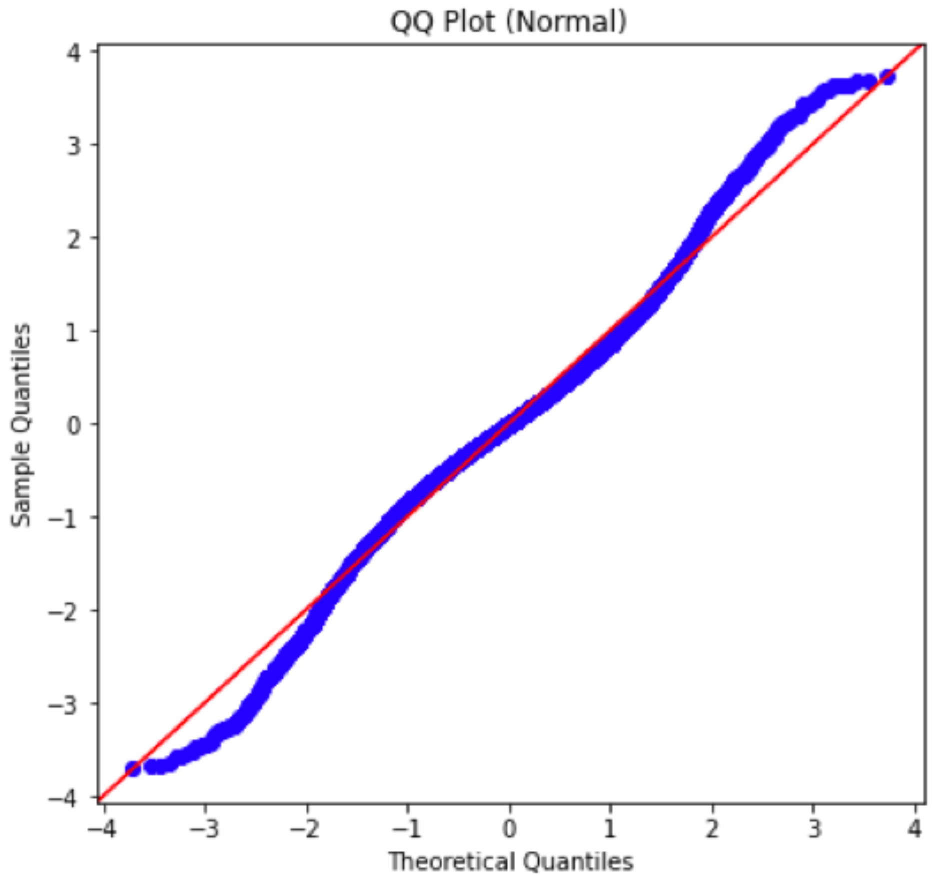
- Parameter: 'degrees of freedom' df ≥ 0
  - Shift or scaled with mean and standard deviation
- Normal with 'fat tails'



QQPlot of Estimated Z-Scores

- Compare distribution of sampled values to standard normal





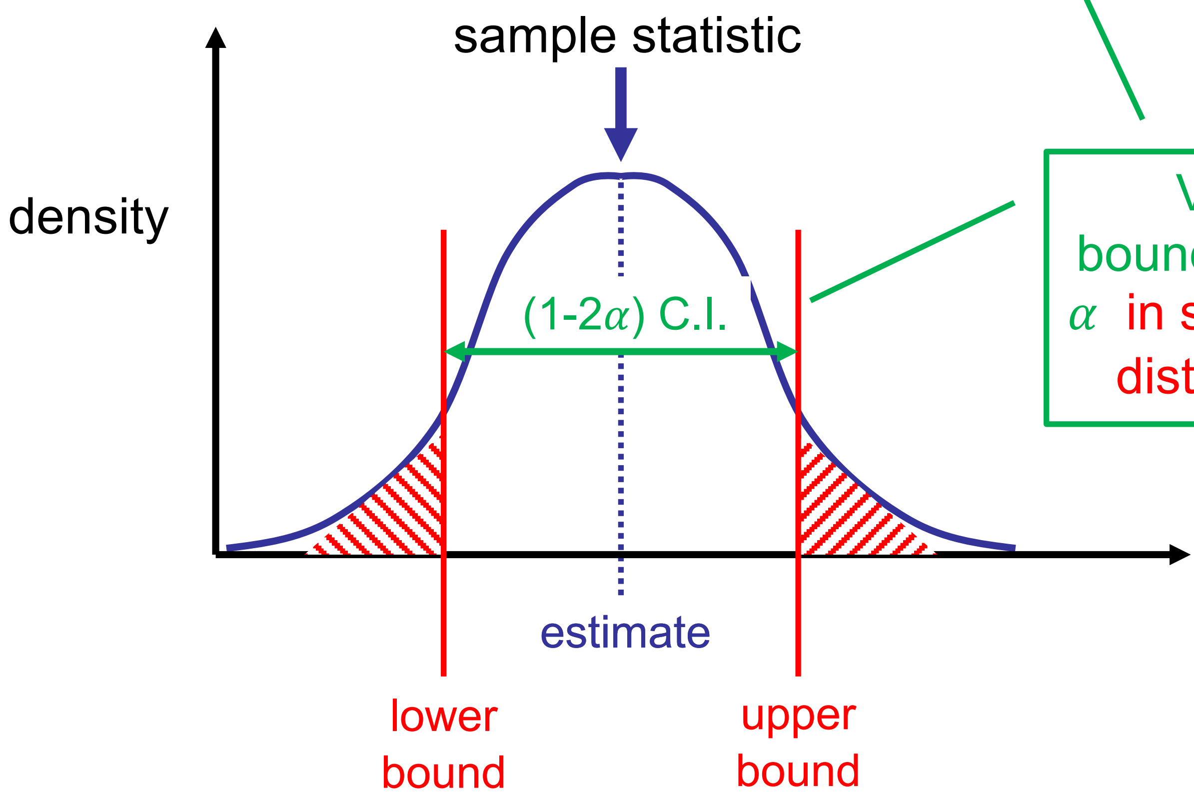
Using Student's t to Estimate CI

- Central limits theorem
- Interval

$$\bar{x} - t^* \frac{s}{\sqrt{N}} \text{ to } \bar{x} + t^* \frac{s}{\sqrt{N}}$$

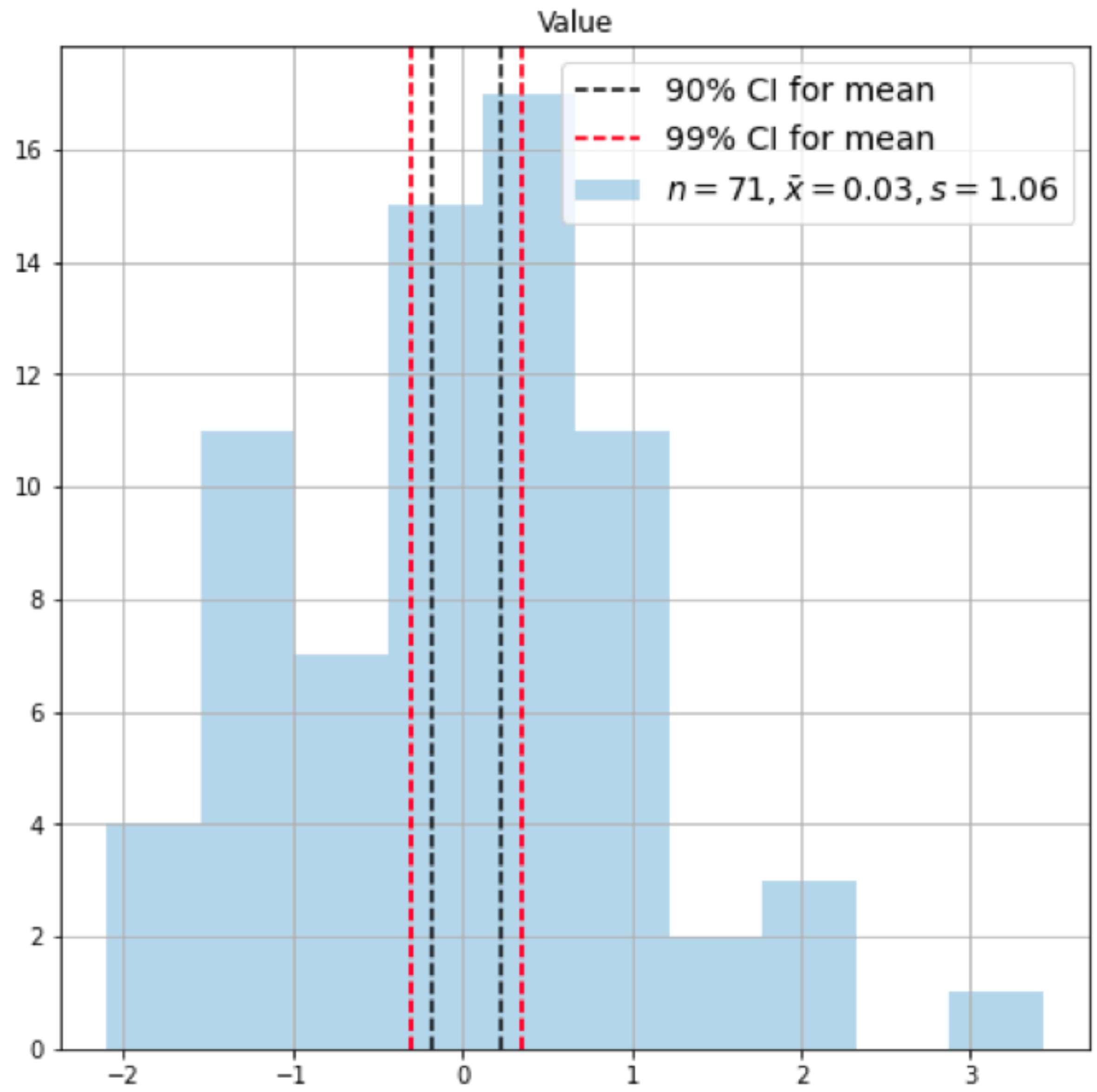
Standard error

Value bounding area α in student's t distribution



Confidence Intervals for a Mean

- Sample of data
  - From a normal
- Sample statistics
  - Mean
  - Standard deviation
- CI from
  - Student's t-distribution
  - Required p-value



Sampling Distribution (of a Statistic)

- We looked at the distribution of the samples
  - Often Normal, at least for large samples

- Some sampling distributions

Distribution	Application
Student's t-distribution	Example: difference between two sample means. Sampling distribution normal but variance unknown
Chi-squared	Difference between frequencies in a contingency tables

Preview of C/W 1

C/W 1: Underground Exits

- Working with a data frame in Pandas
- Summary statistics and histograms
- Interpretation of data and plots
- Simple model and evaluation

- Use any code from notebooks
- Sprint week should cover everything
- All code should be relevant (NO JUNK)
- You can ask questions

Two Key Issues

Readable Document

- Write the notebook as a ‘document’
  - Imagine all the code hidden
  - Reader is a ‘domain’ expert (not a data scientist, not a programmer)
- Guide: it should look like notebooks 1-3
  - Title and section headings
  - Short code cells alternating with markdown
- Write about data manipulation, not code

Running the Notebook

- Problem of order
- Solution: rerun notebook
  - Restart the kernel
  - Rerun everything
- Expect zero if your code does not run
  - You may get more

Hypothesis Testing

Making a decision: are differences real?

Student's t-test

Hypothesis Testing

- Null hypothesis (H_0)
  - Chance is to blame
  - Differences in the sample are not *real*
- Alternative hypothesis (H_1)
  - The difference is not due to chance

- Idea: assume ‘null hypothesis’ and ask ‘is this data likely’?
  - If sufficiently unlikely, we reject the null hypothesis
  - If not, all we can say is ‘no evidence’

Significance and Errors

- ‘Sufficiently unlikely’ – significance
  - Choose a threshold ‘p-value’
- Type 1 error:
  - Mistakenly reject the null hypothesis
  - False positive
- Type 2 error
  - Mistakenly accept the null hypothesis
  - Effect not proven – more data needed
- Increase threshold to reduce type 1 errors

t-Tests

- For statistics that would be normally distributed except that variance also estimated from data
- Examples
  - Could the mean be equal to a given value?
  - Are means the same? Paired data:
    - E.g. same students taking programming and stats
  - Are means the same? Unpaired data:
    - E.g. stats results for men and women

Testing Difference Between Two Means

- Two groups A (size N_A) and B (size N_B), unpaired
  - E.g. those with / without heart disease
  - Samples X_A, X_B of e.g. cholesterol

Type	H0: Null	H1: Alternative
One tailed	A-mean does not exceed B-mean	A-mean does exceed the B-mean
Two tailed	A-mean and B-mean are equal	A-mean and B-mean are not equal

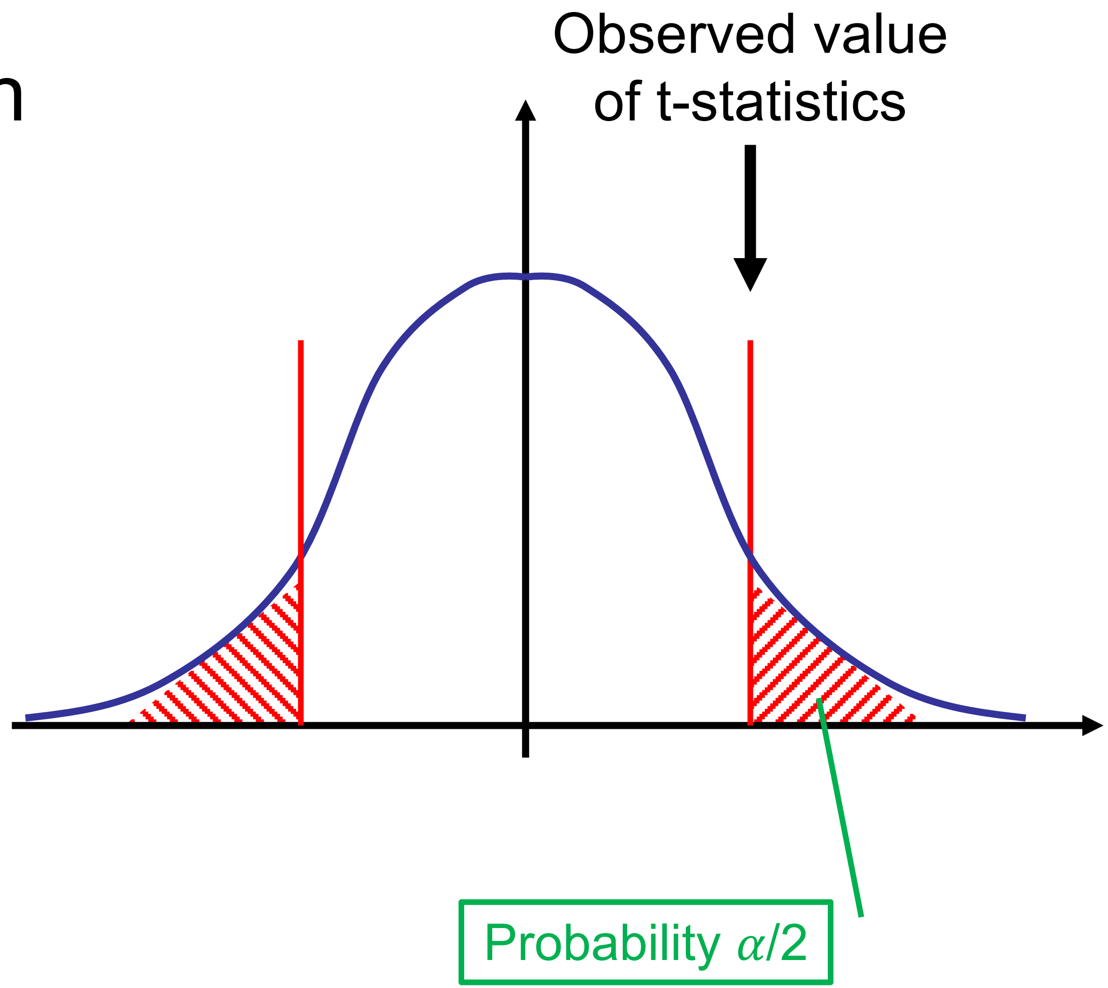
Test statistic: $t = \dfrac{\overline{X_A} - \overline{X_B}}{\sqrt{\dfrac{s_A^2}{N_A} + \dfrac{s_B^2}{N_B}}}$

Difference of means: t increases as $\overline{X_A} > \overline{X_B}$

from t-distribution with degrees of freedom depending on N_A - 1 and N_B - 1

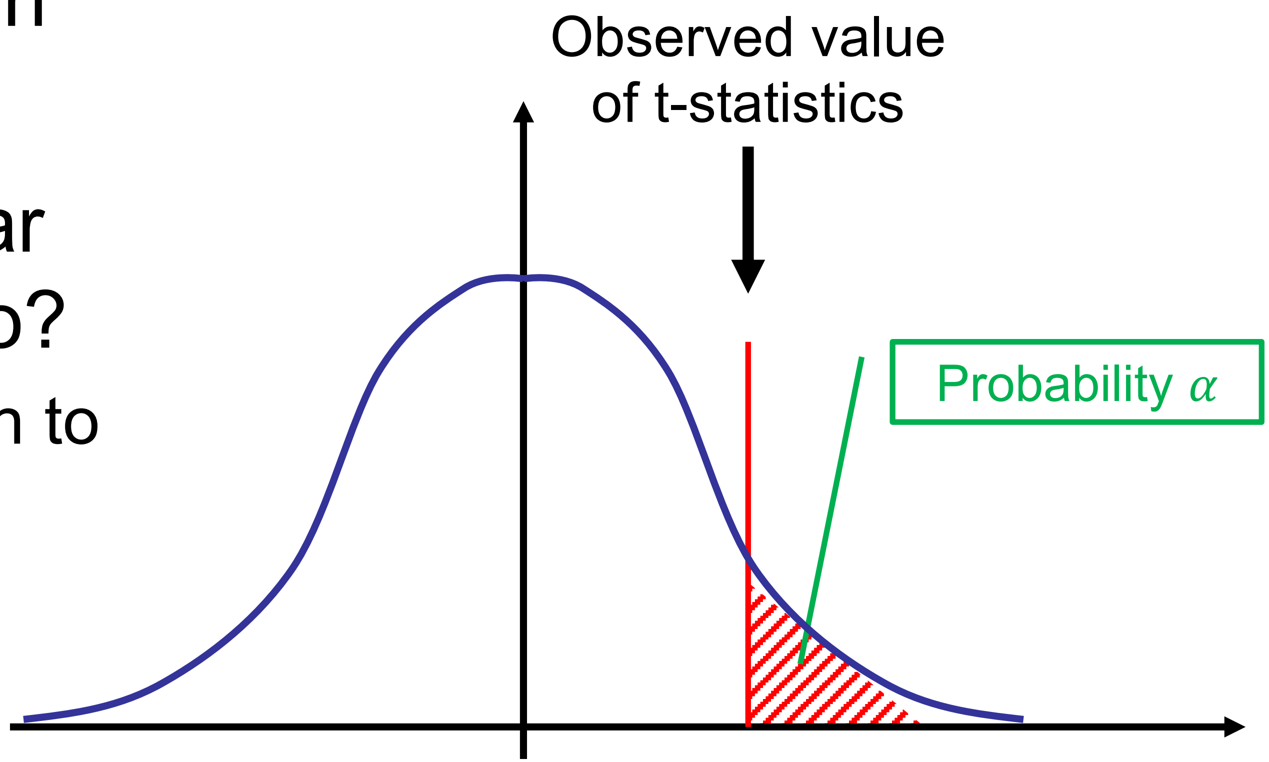
Two-Tailed t-Test

- |t-statistic| increases as A-mean > B-mean (or as A-mean < B-mean)
- Is the t-statistic far enough from zero?
  - Calculate α
  - Is α small enough to reject H_0?
  - Threshold 99% then $\alpha \leq 1\%$



One-Tailed t-Test

- Only look at A-mean > B-mean
- Is the t-statistic far enough from zero?
  - Is α small enough to reject H_0?



Software: given the data for A and B, calculate the statistic and return α <u>usually for 2-tailed test</u>

Relationship between Hypothesis Testing and CI

- These two are connected
  - Confidence interval for $\overline{X_A} - \overline{X_B}$
  - Hypothesis test comparing μ_A and μ_B
- If CI includes zero then means may be equal

- Catch
  - CI of difference in means is not the same as difference between CIs of means

Quiz 3

Issues with Hypothesis Testing, C.I.s and p-Values

Often misinterpreted

Confidence Intervals

- Good idea to have interval estimation
  - Not just point estimates

- Problem: Not the probability that you think it is!

There is 95% probability that the mean is in the CI range ✗

If we repeat the same sampling process many times, the true (population) statistic with be inside the CI range 95% of the time

Hypothesis Testing

- *Not very relevant to data science*

- Problem 1: no account of effect size
  - Effect: e.g. the increase in mean survival
  - A ‘statistically significant’ effect can be **insignificant**
  - ... especially with large datasets
- Problem 2: interpretation
  - Failure to reject the null hypothesis does not imply it is true
  - ... maybe too little data
- Problem 3: p-values (again)
  - At 95% confidence, the conclusion is wrong 5% of the time

t-test Assumes Independent Samples from Normal

- Population is not Normal
  - Skewness
  - Outliers (e.g. errors)
- Sample not independent
  - E.g. time series

- Difficulty: how much non-normality is too much
  - t-test generally ok when sample large

Summary I

- When sampling, difference may arise by chance
- CI: in what range will the sample statistic usually fall?
- Hypothesis test: assume ‘null hypothesis’ (no difference)
- Hypothesis testing and p-values can be dangerously misleading
  - Especially with large datasets: ‘statistical significance’ may be true when difference trivial

Summary II

- What makes this topic difficult?
  - Details of sampling statistic and sampling distributions
  - Assumptions under which different tests are reliable
- Consequence
  - Formulaic approach
  - Poor interpretation of results

- *Computational approaching: bootstrapping*
  - *More uniform*
  - *Focus on interval estimation*