Inferential Statistics
Hypothesis Testing (made easy?)

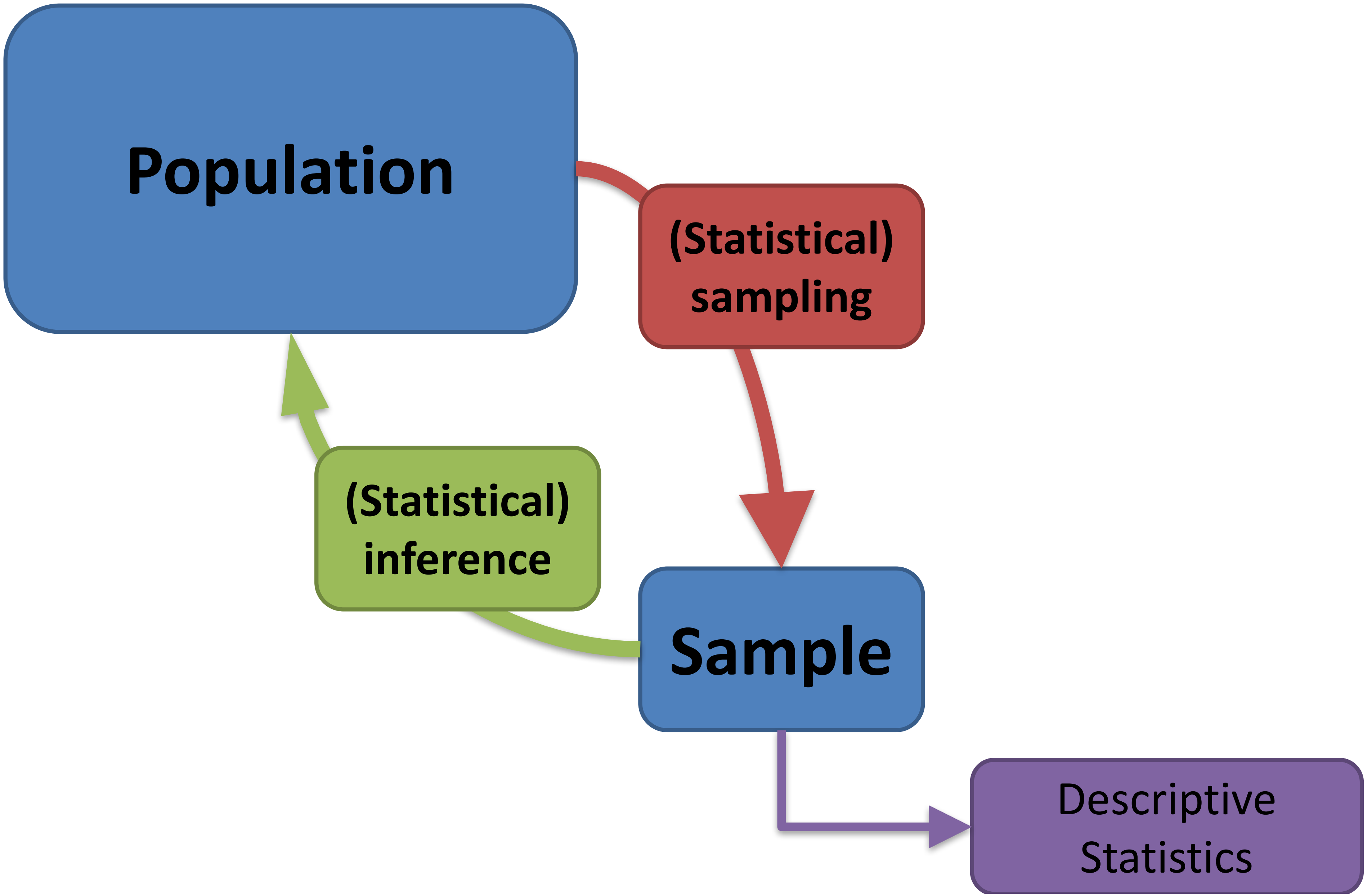
Prof. Christian Wallraven

wallraven@korea.ac.kr

EXAM!

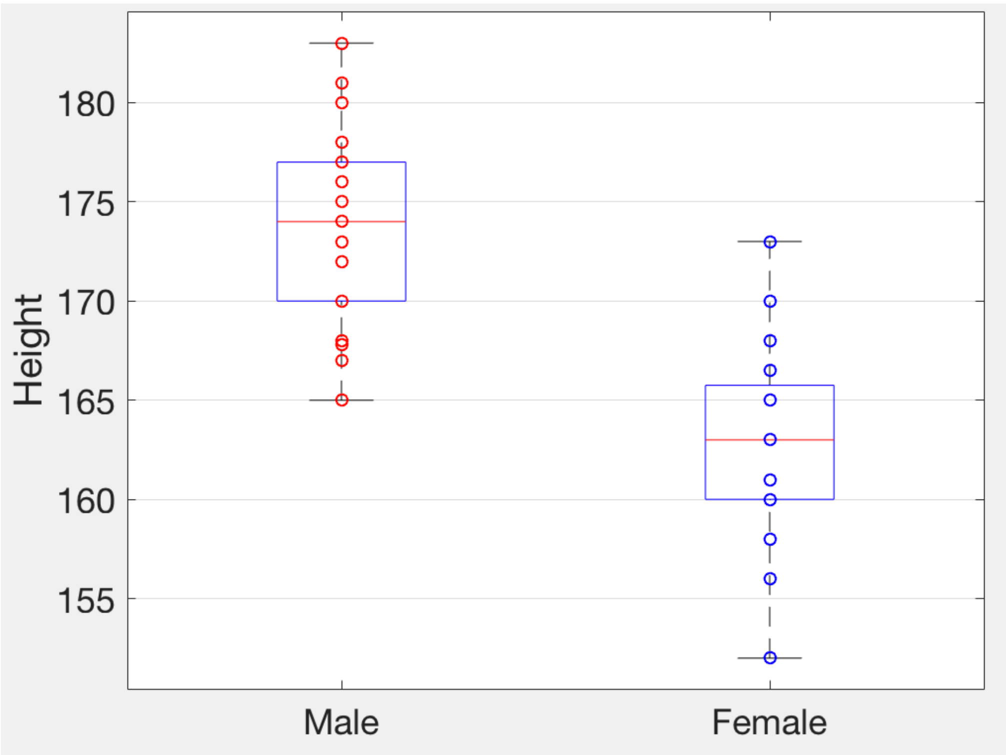
- The exam will be on TUESDAY 16th of October from 3PM-11PM as a take-home exam!!!

The Big Picture



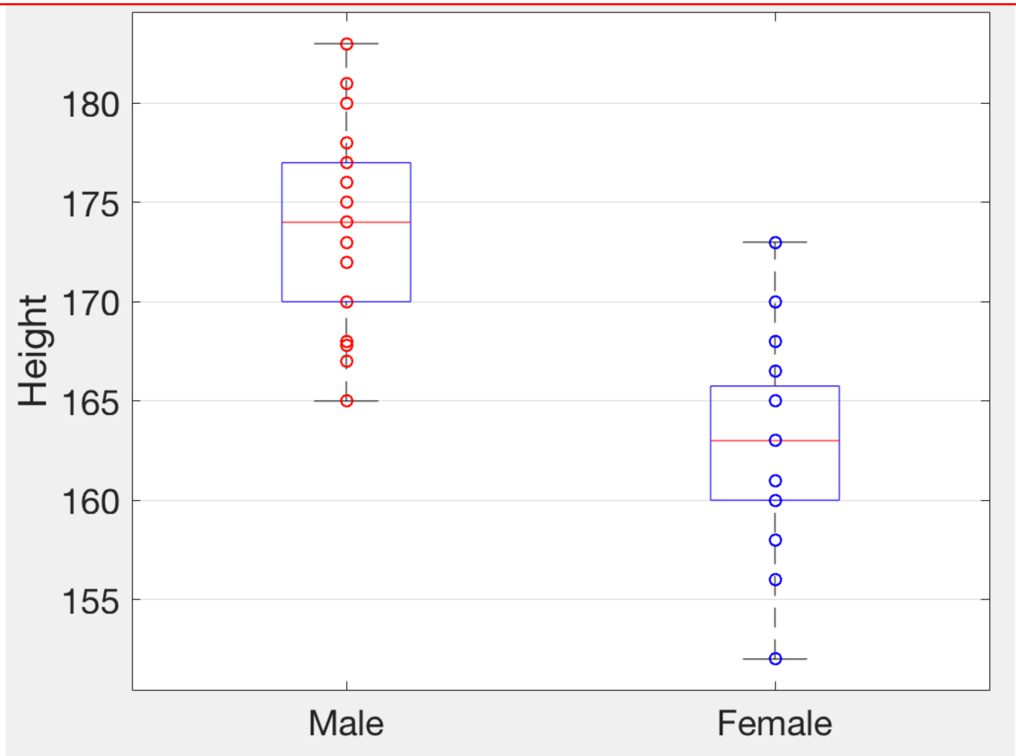
Remember our data?

- Somehow we feel that the two distributions for height for male and female students are different on average
- Most tests in inferential statistics work by estimating the probability that they are the **same**



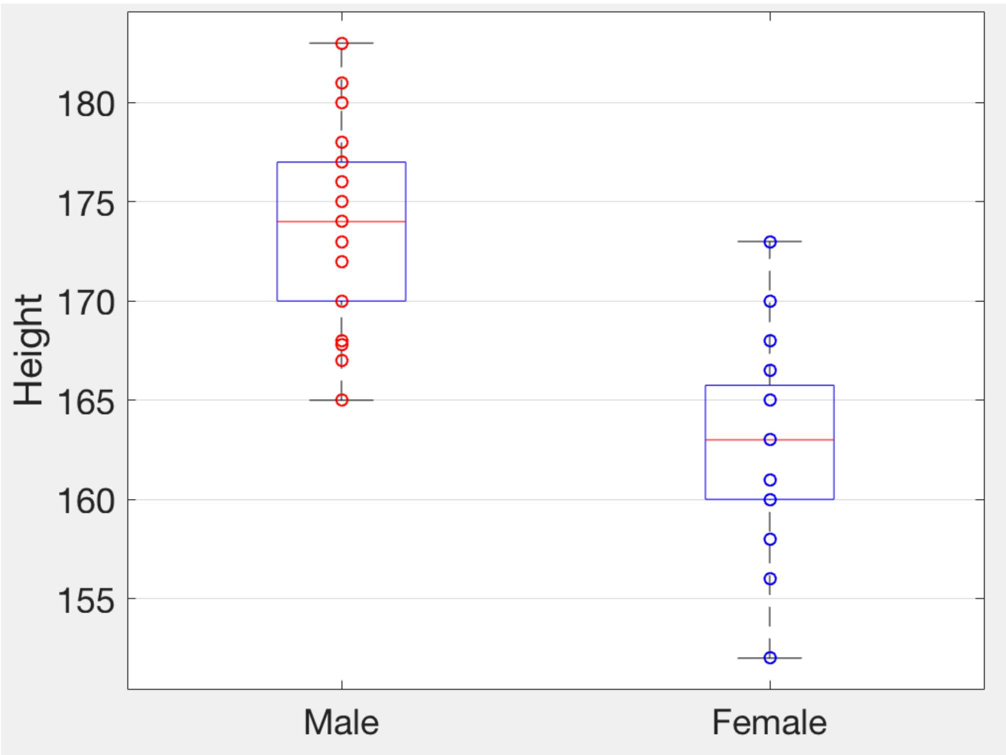
Remember our data?

- If the probability p that the two averages are really the same **is high**, then there is no meaningful difference between male and female heights in our data



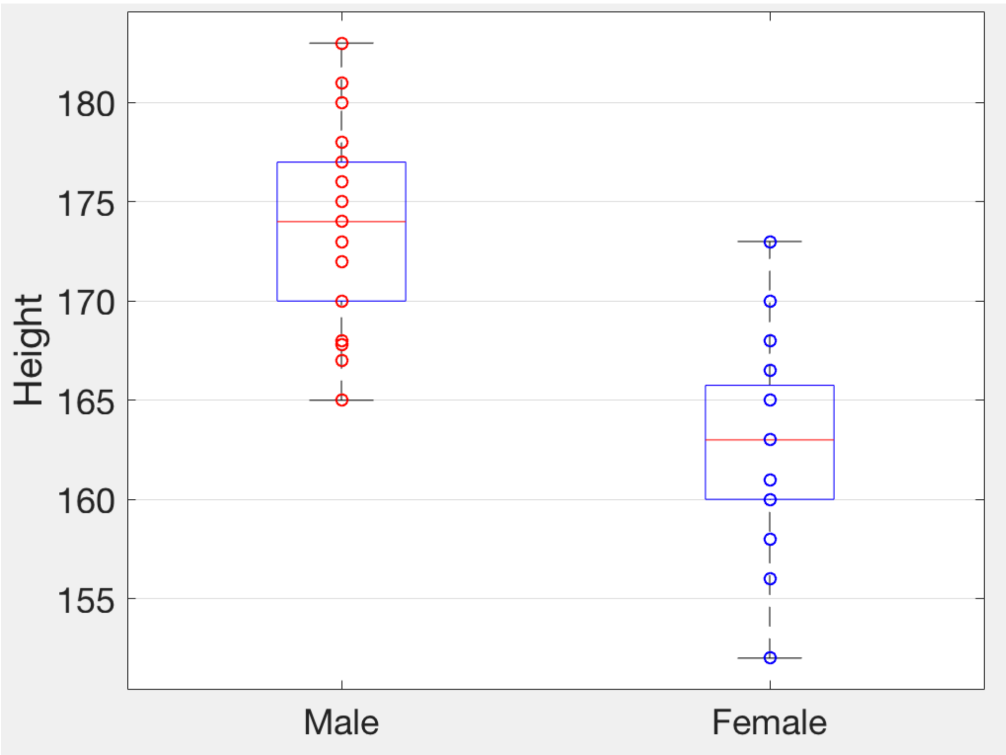
Doing stats yourself

- Before we go on to describing actual statistical tests that can do this, we can try to do some of this by hand
  - well, with the computer



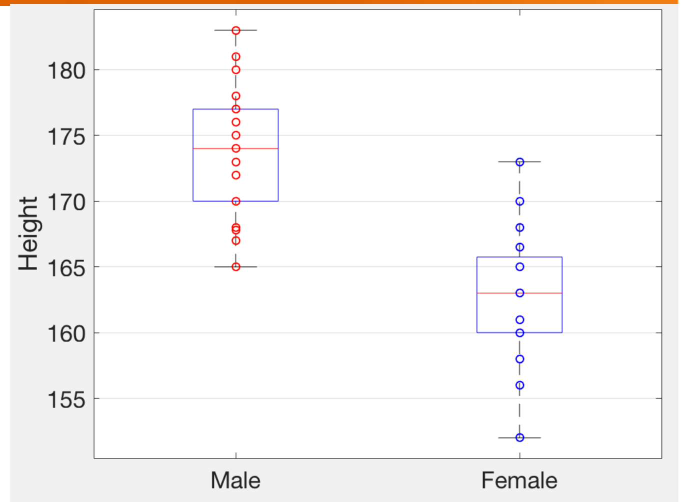
Differences in means

- We have two groups, and we feel that they have different mean (median)
  - Male mean = 174.17cm
  - Female mean = 163.10cm → Difference = 11.07cm

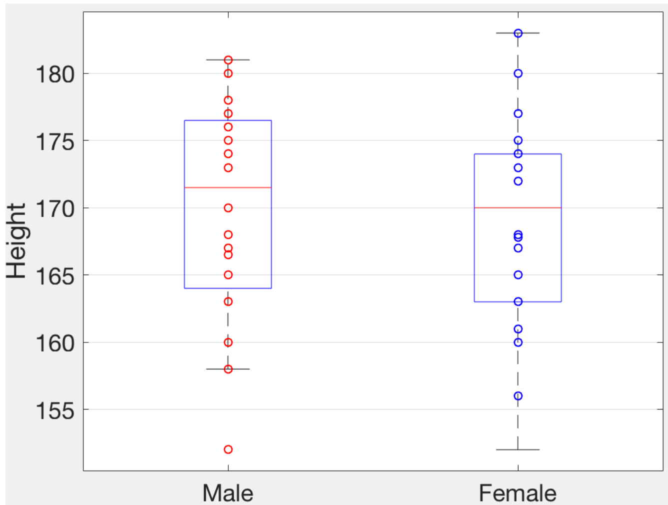


Shuffling!

- Basic idea:
  - we take all of our data and put it together
  - then we are going to assign each person in this data randomly to either the “female” or the “male” group
  - we are going to calculate the new means and the new difference
  - for this example:
    - Male mean = 169.75cm
    - Female mean = 169.37cm
    - Difference = 0.38cm

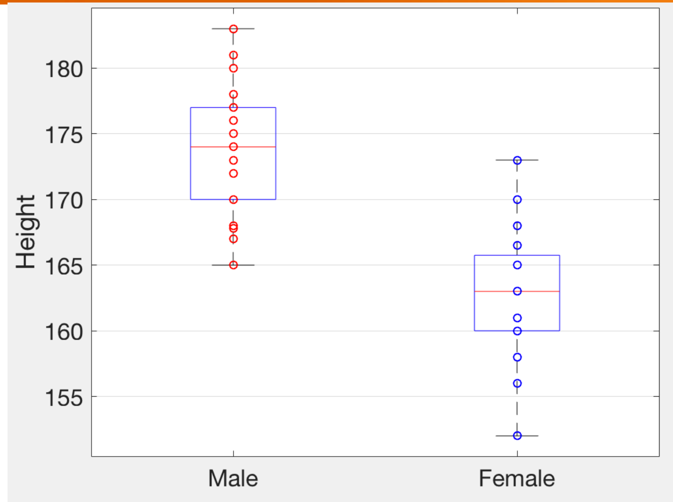


Random shuffling

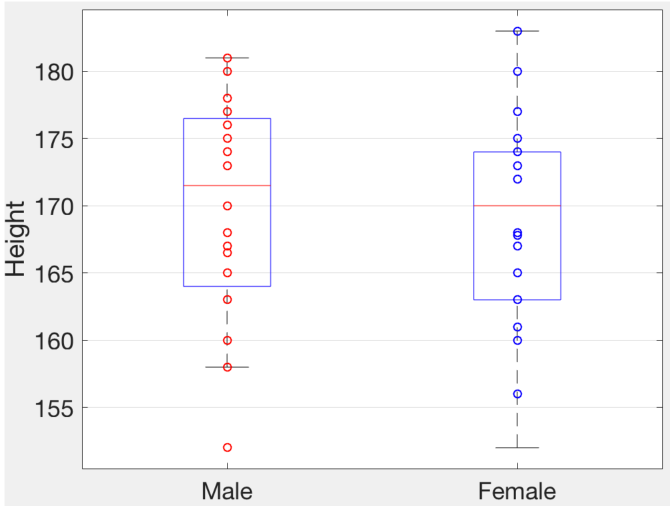


Shuffling!

- Basic idea:
  - for this example:
    - Male mean = 169.75cm
    - Female mean = 169.37cm
    - Difference = 0.38cm
  - we repeat this procedure many times and count how many times we observe a value that's larger than this

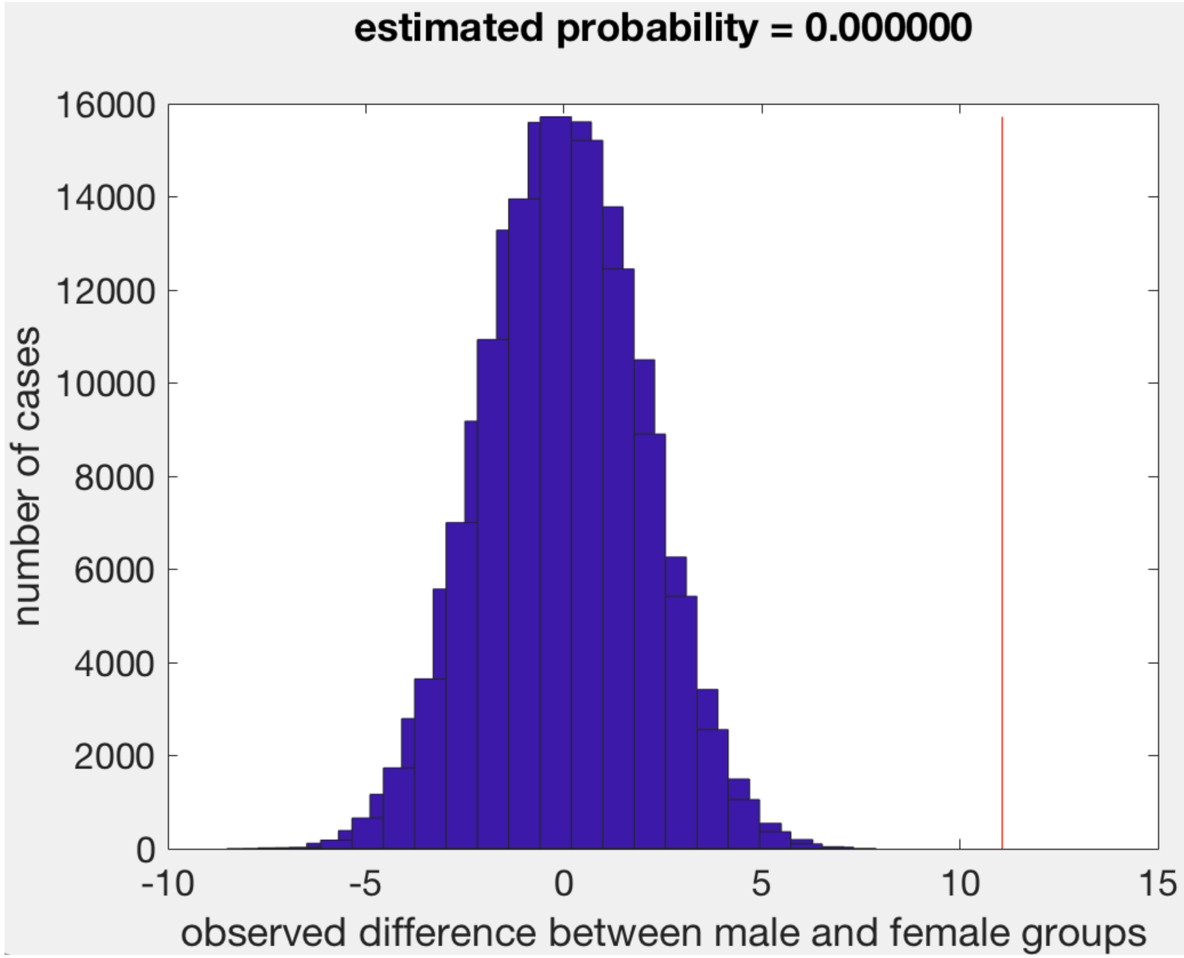


Random shuffling



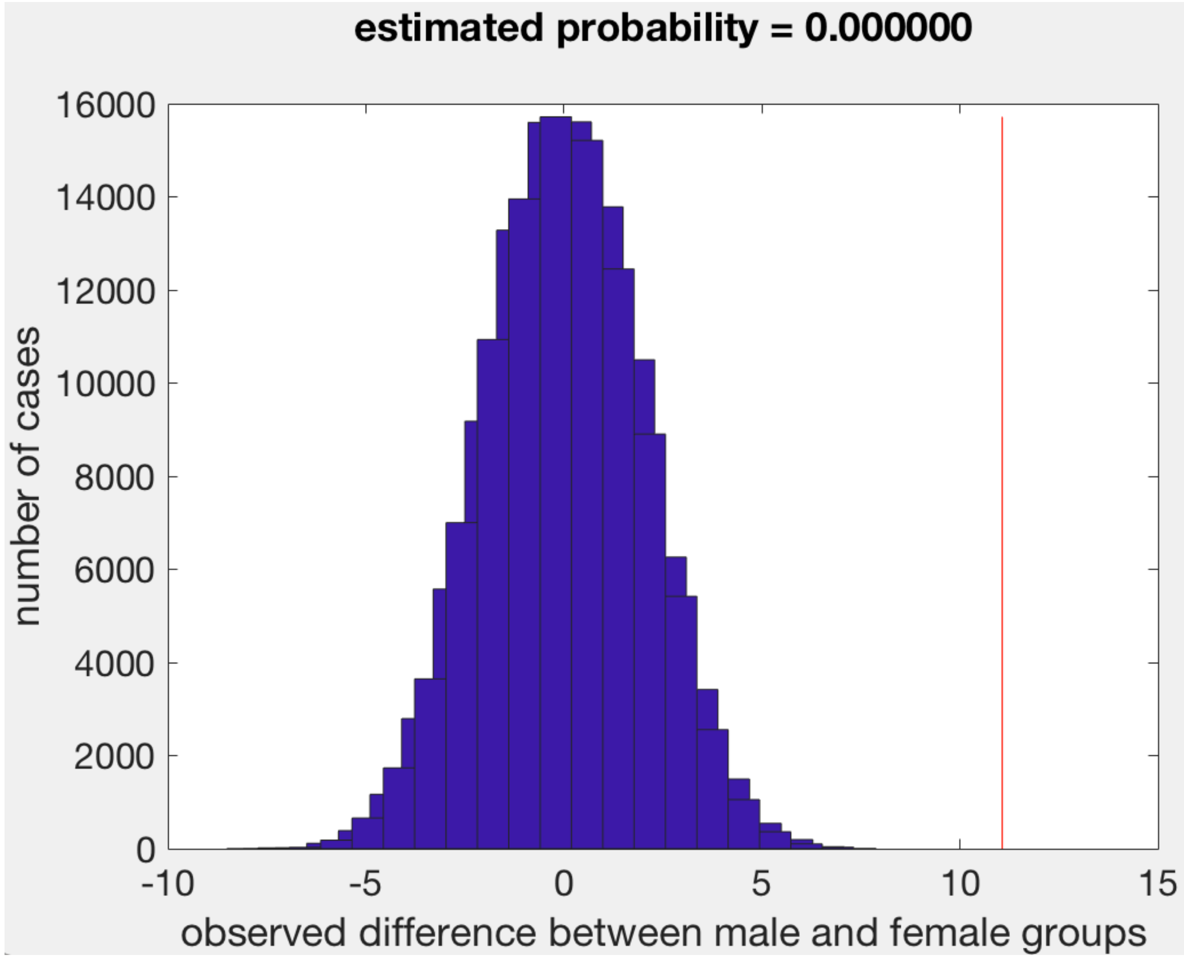
Example

- For our two groups, if I do 100000 "virtual" groups, I get the following distribution of "random" differences
  - no random difference was larger than the original one!!
- **It seems virtually impossible to achieve this 11.07cm difference by chance**



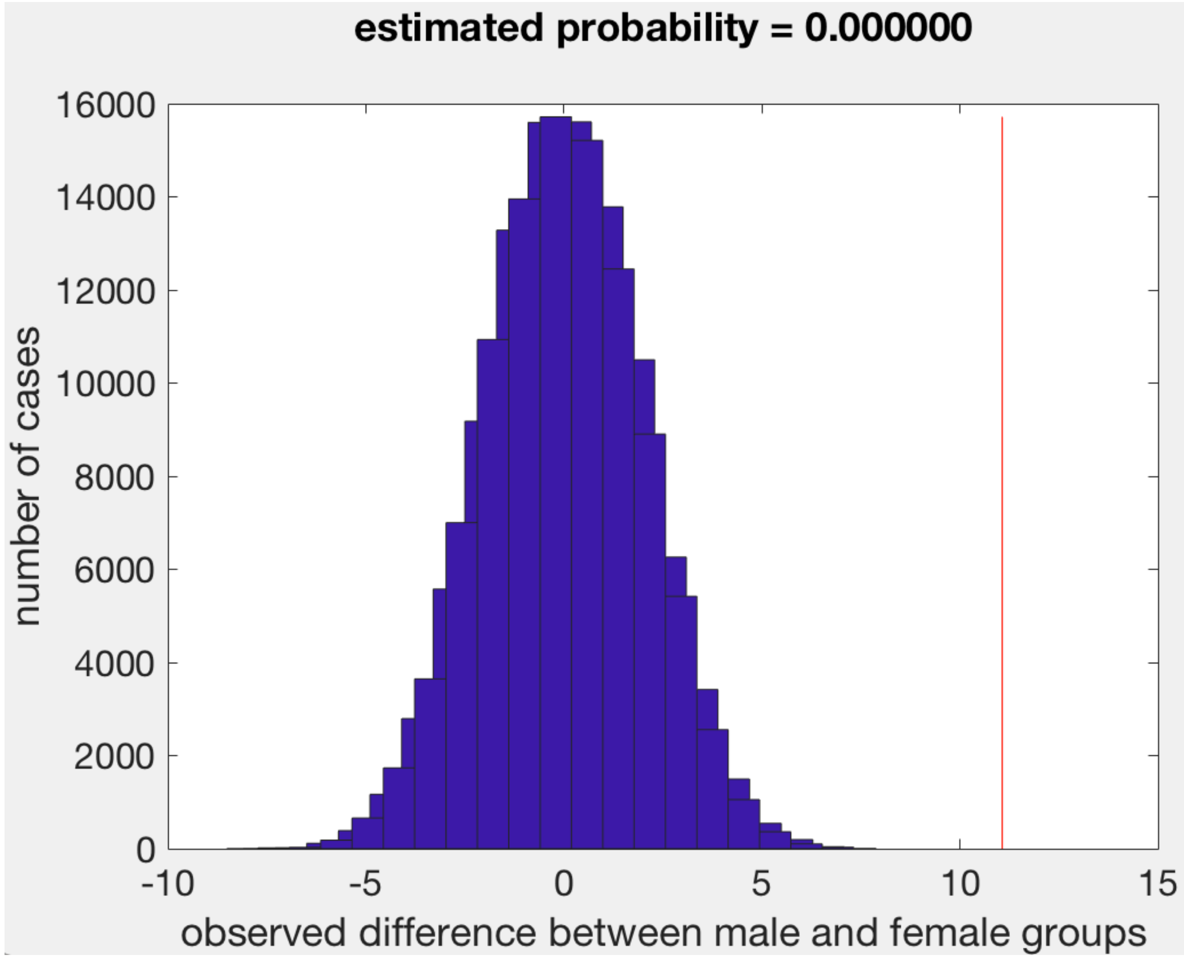
Example

- The blue distribution is an estimate of how two groups of the same sample size as our data differ on average, when they are **randomly created**



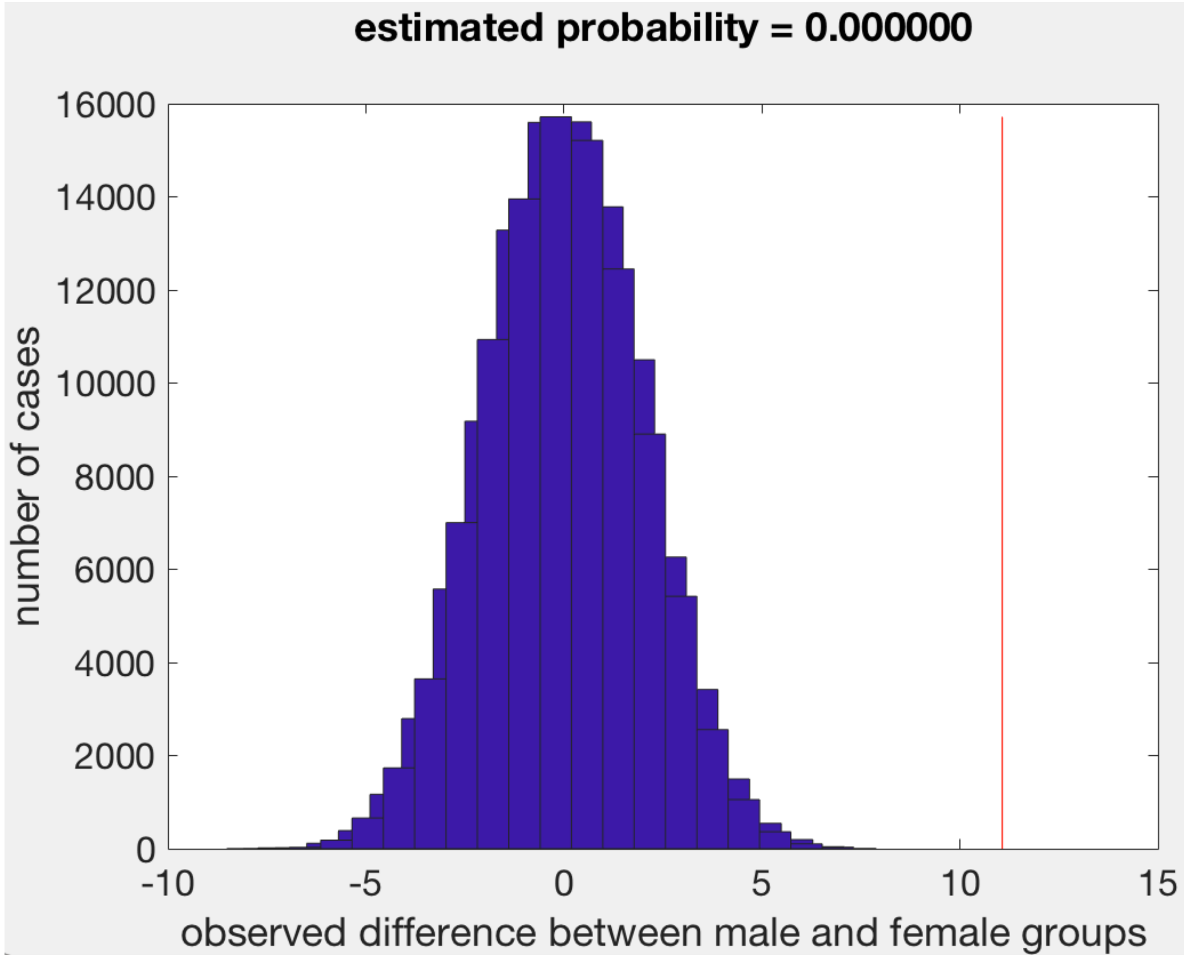
Example

- The red line is our data – since the probability that it is part of the blue distribution is **virtually 0**, it seems unlikely that our data is the result of a random observation



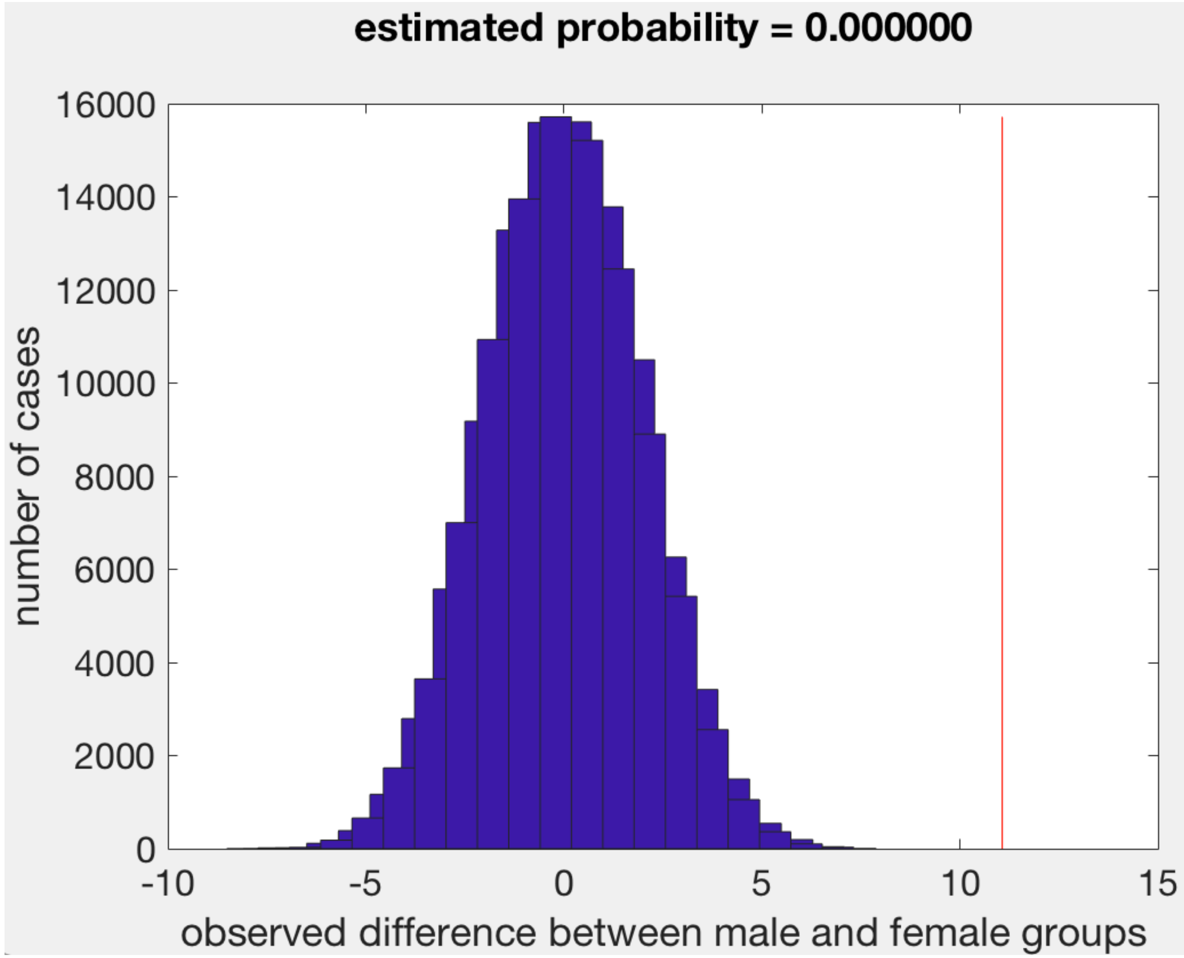
Example

- Therefore, we conclude that the 11.07cm is a meaningful, statistically significant difference



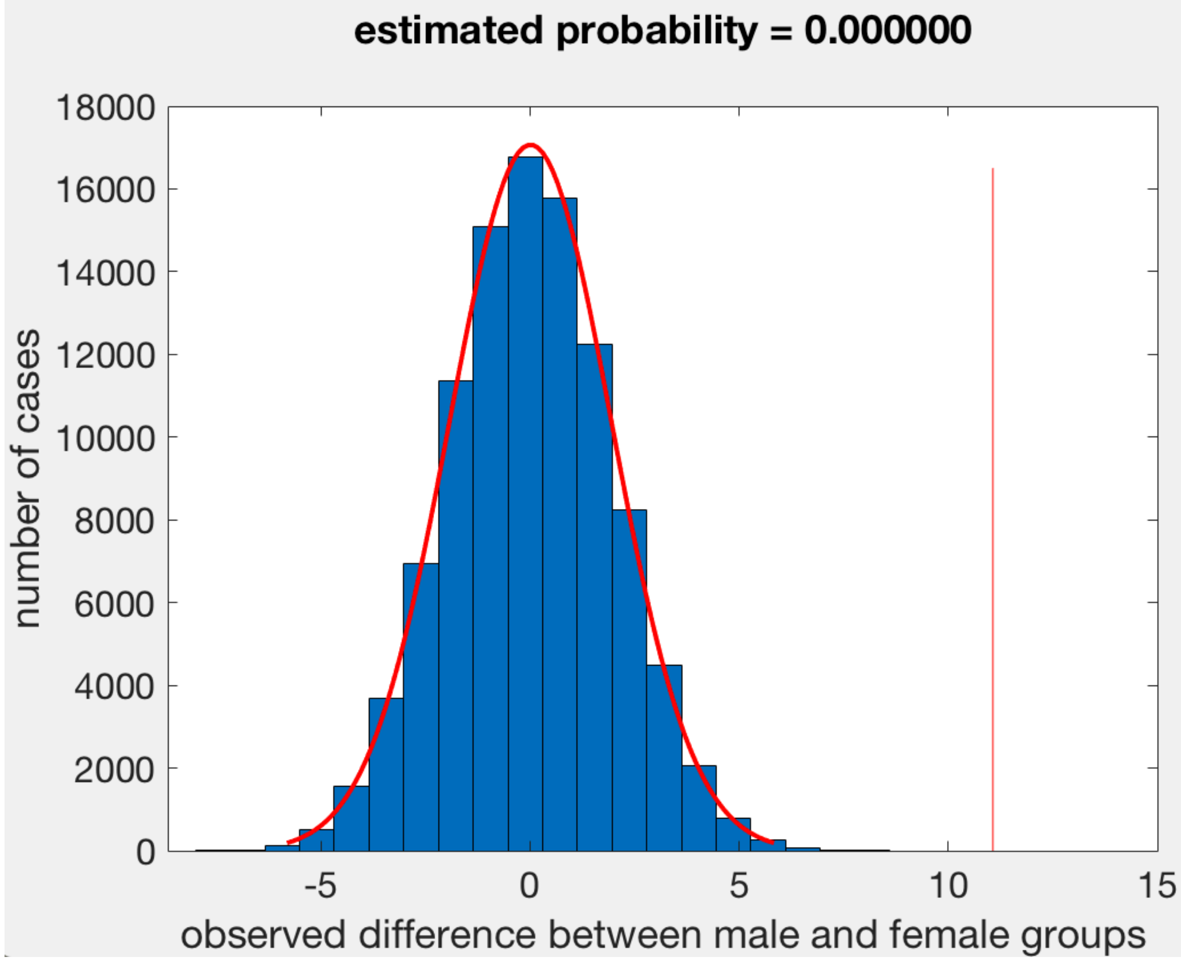
Example

- We say: **the difference between male and female heights is statistically significant with p=0.000**
- Notice again, that the p-value tells you the probability that the difference is ZERO on average!



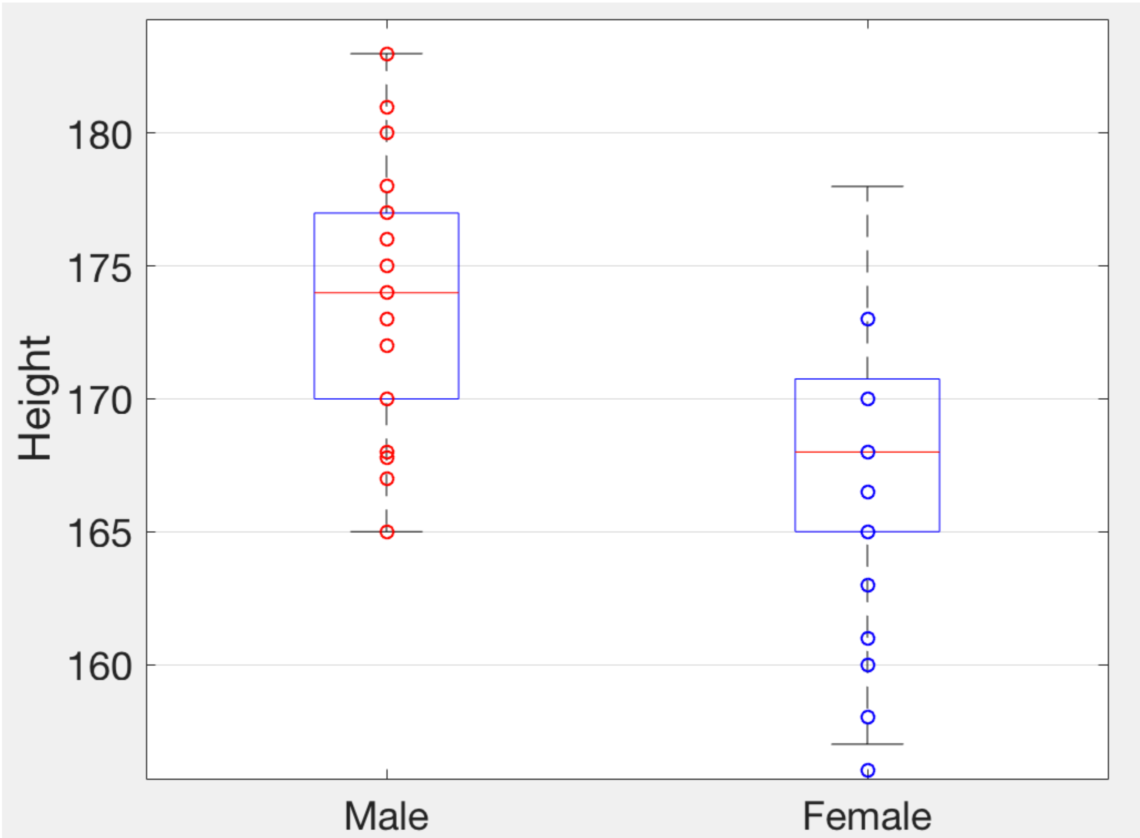
Example

- Also: **we can see that this distribution looks like a normal distribution!**
- So: if there is a way to calculate this distribution from the data directly, we don't need to do the costly simulation
  - there is, but we don't cover it here…



Making the difference smaller

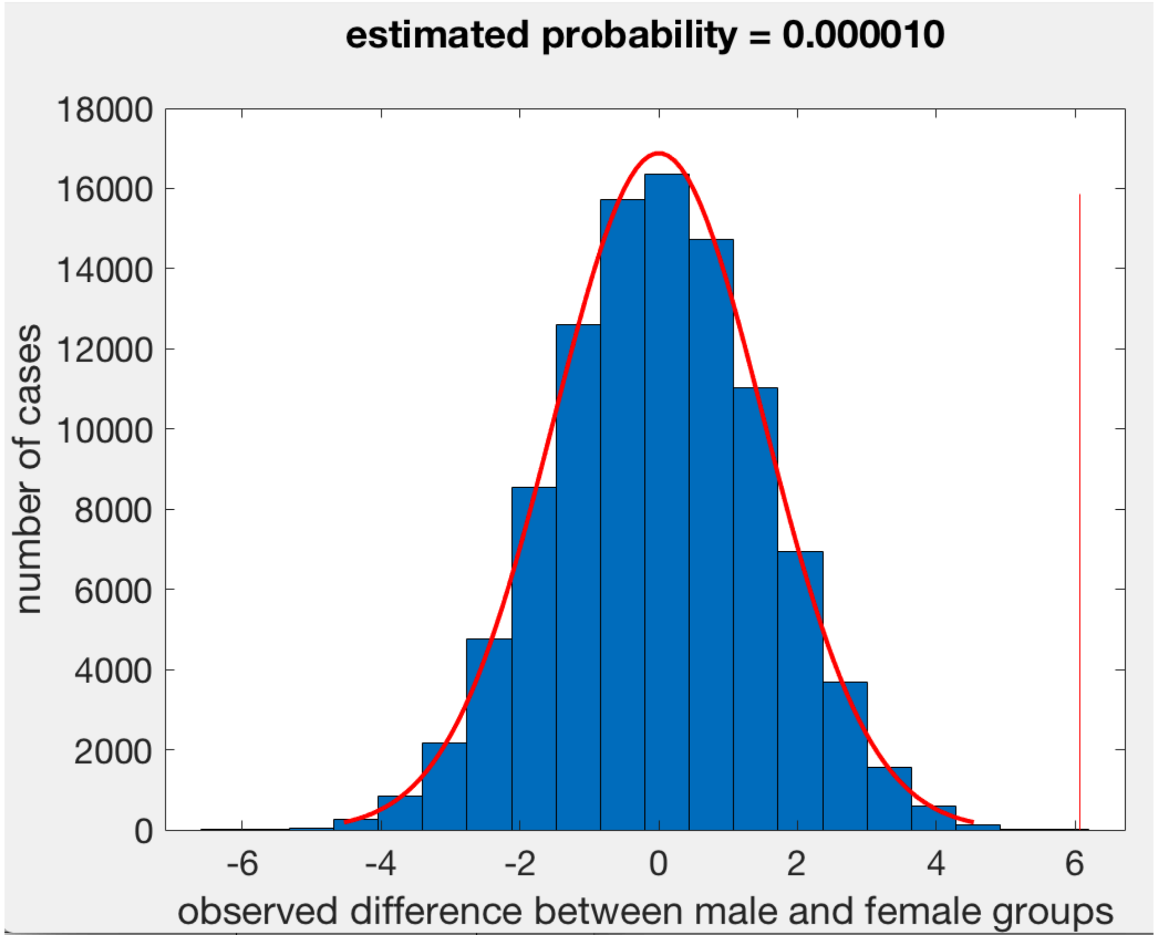
- What about a new dataset, in which females are a little taller?
  - let's make you taller by 5cm → difference = 6.07cm



Example

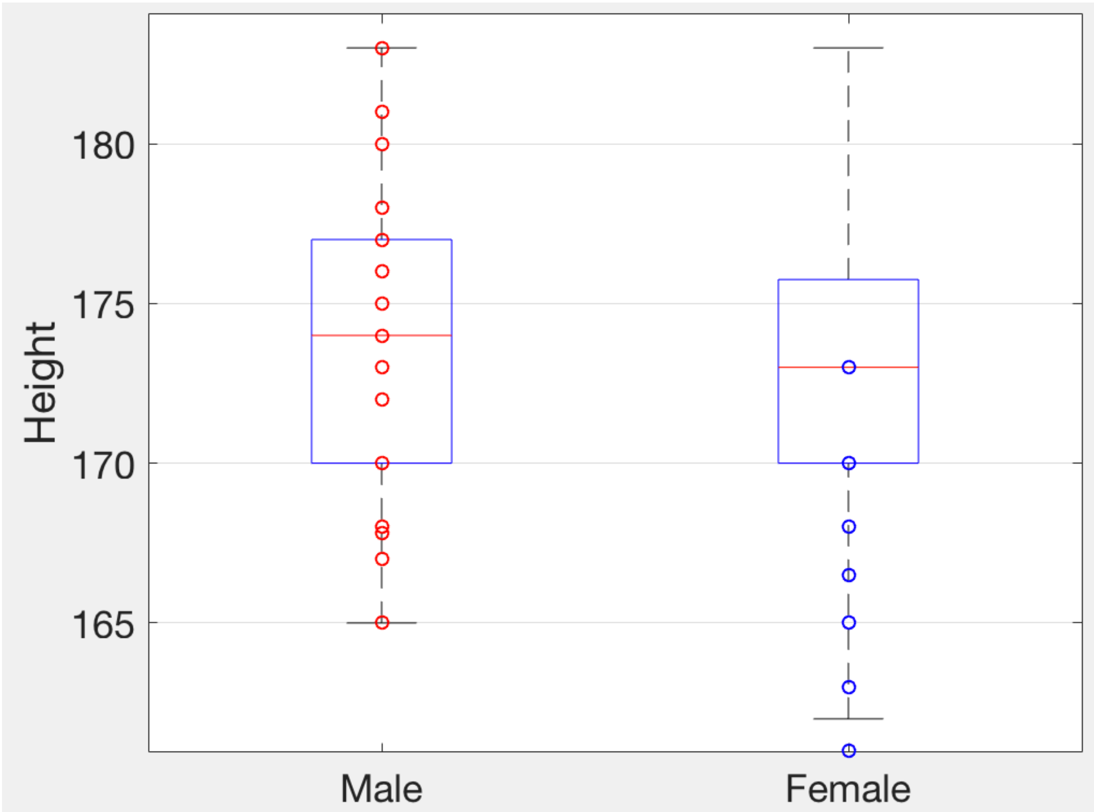
CognitiveSystems

- For our two groups, if I do 100000 “virtual” groups, I get the following distribution of “random” differences
  - 1 random difference was larger than the original one!!
  - p~0
- **Still 0% probability to achieve 6.07cm difference by chance**



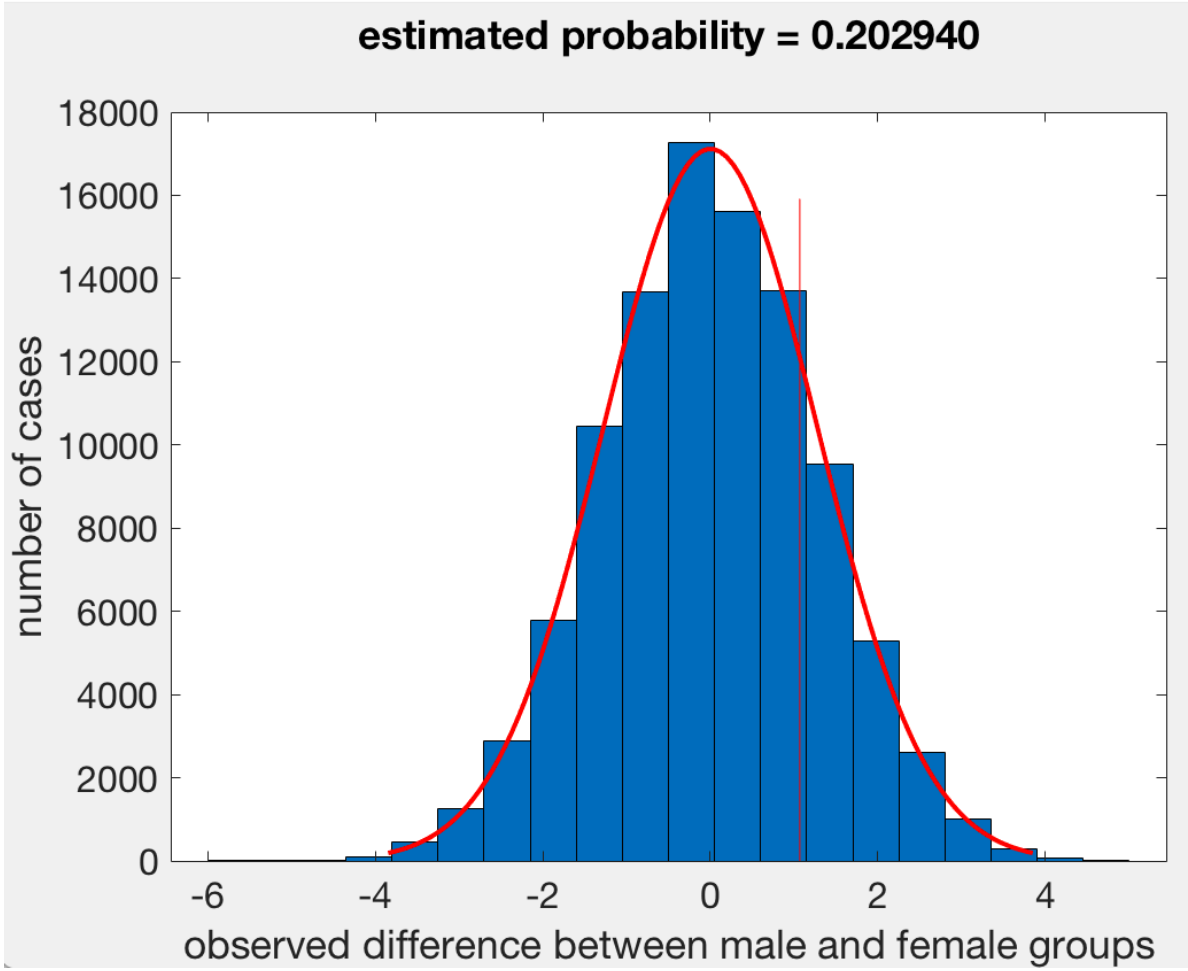
Making the difference smaller

- What about a new dataset, in which females are even taller?
  – let's make you taller by 10cm → difference = 1.07cm



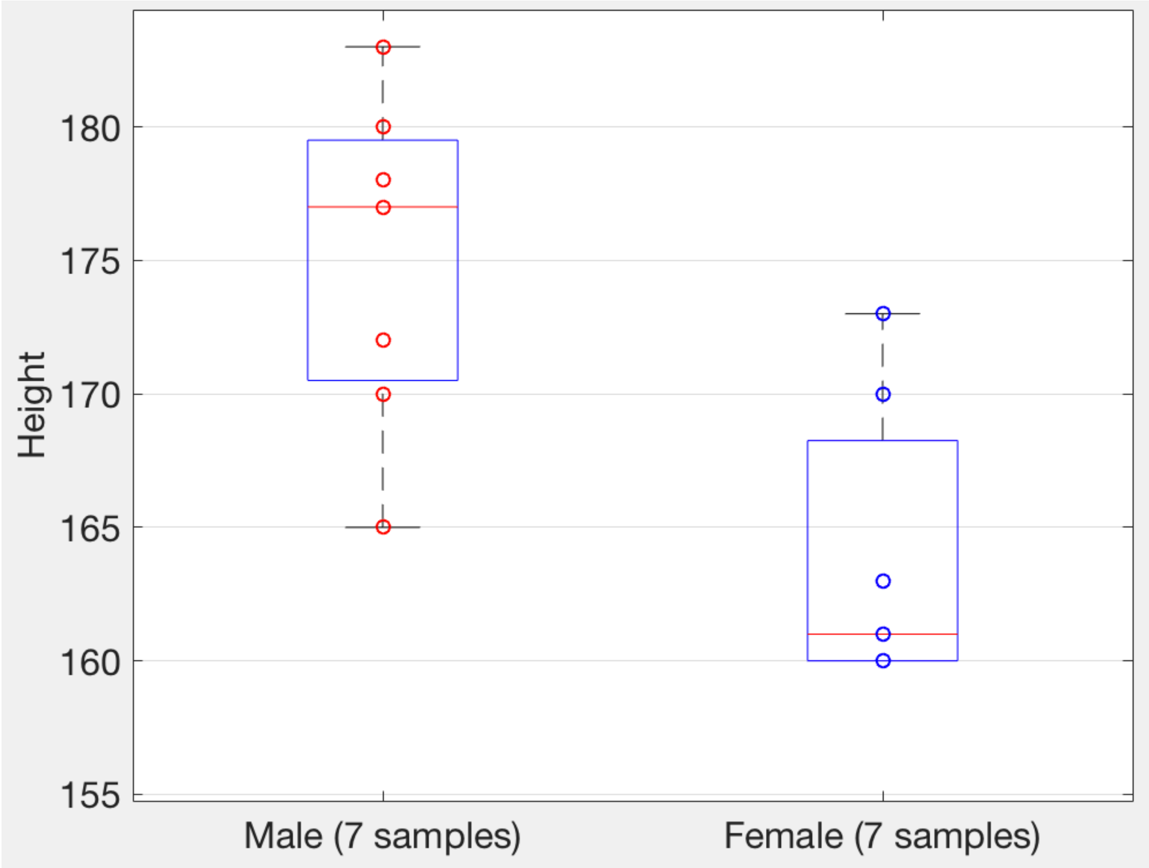
Example

- For our two groups, if I do 100000 “virtual” groups, I get the following distribution of “random” differences
  – 20294 random differences were larger than the original one!!
  – p=0.203!!
- **20% probability to achieve 1.07cm taller difference by chance**



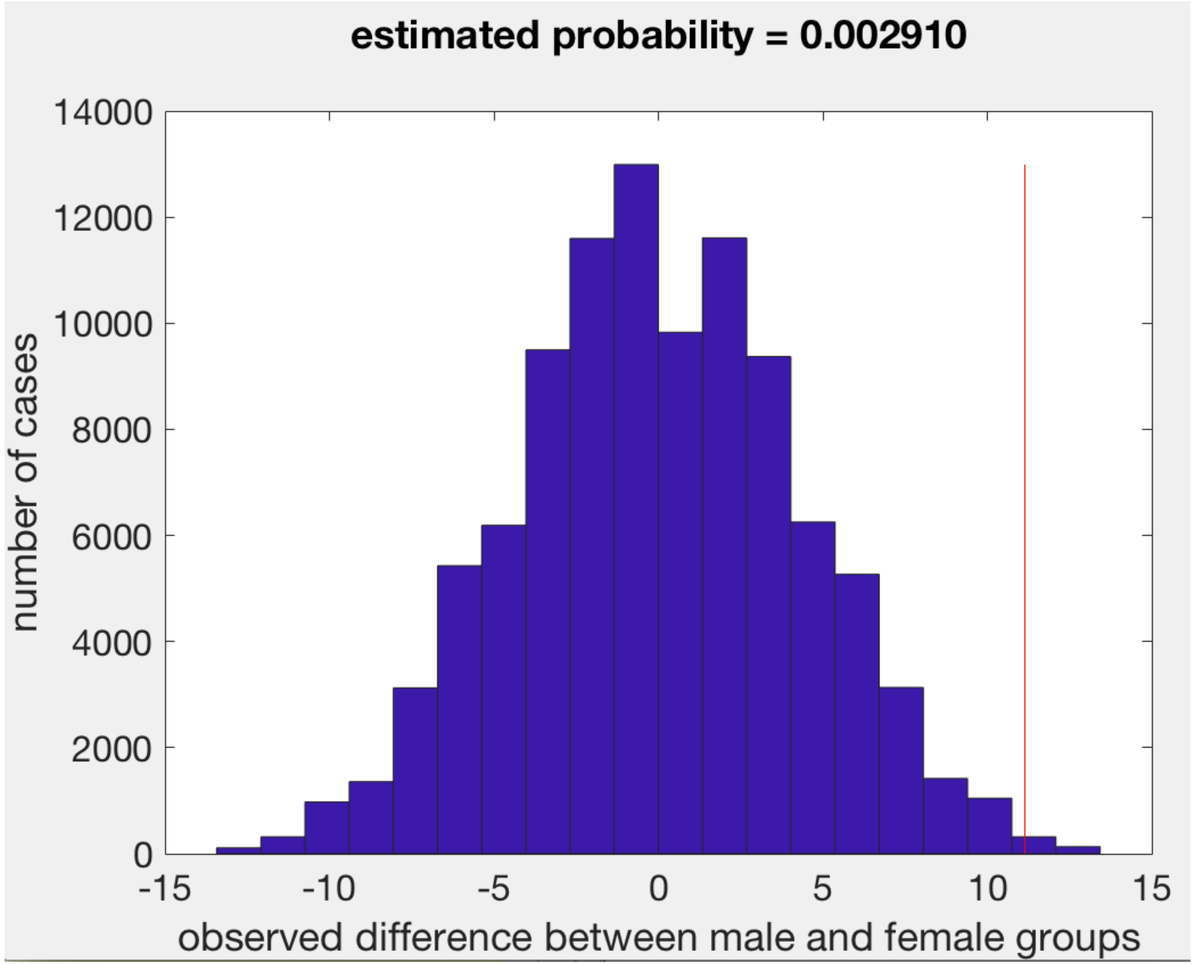
Reducing the number of samples

- Alright, but what about the influence of the sample size? We had 63 people so far with a difference of 11.07cm, so let’s try to reduce the number of samples
  - n=7 male
  - n=7 female
  - difference=11.1cm



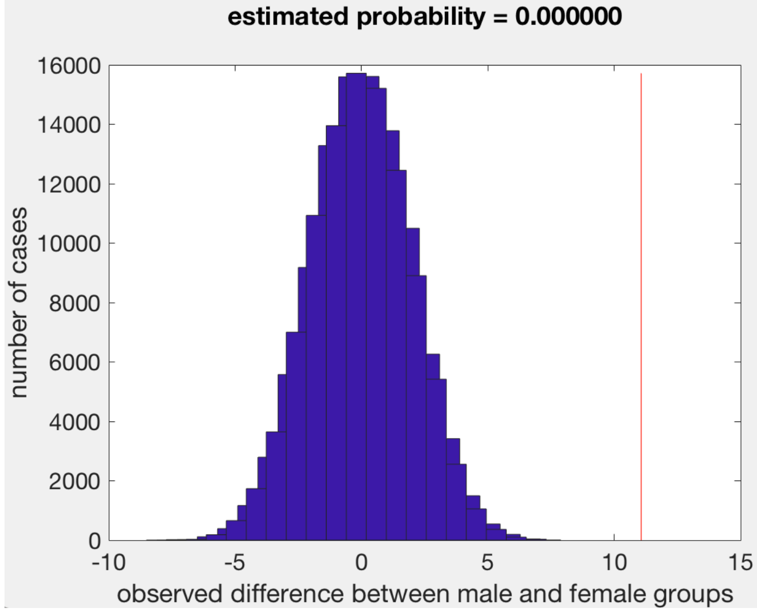
Example

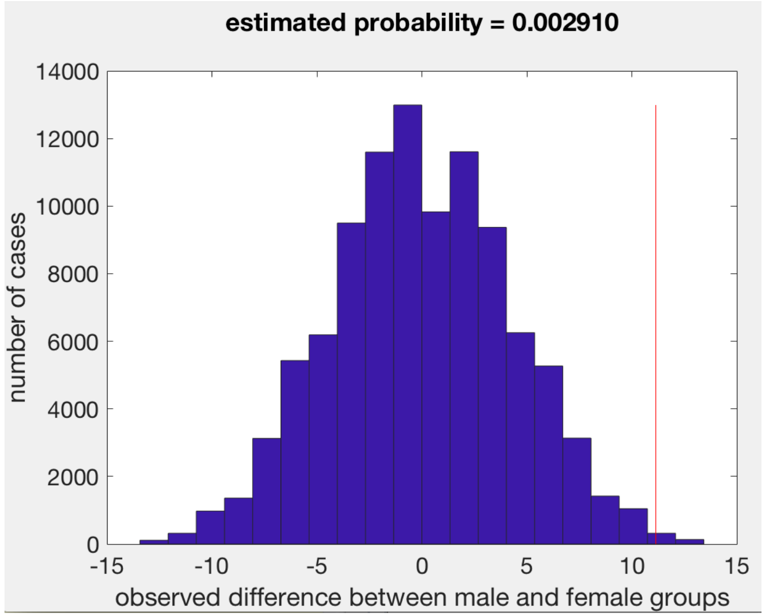
- For our two groups, if I do 100000 "virtual" groups, I get the following distribution of "random" differences
  - 291 random differences were larger than the original one!!
  - p=0.003
- **0.3% probability to achieve 11.1cm difference by chance**



Example

- Even though our chances for the smaller sample were also very low, the estimated random distribution looked different from our first case with the full sample!
- **The full-sample distribution had values between -5 and 5**
- **The small-sample distribution had values between -12 and 12**
- **Both look normal**





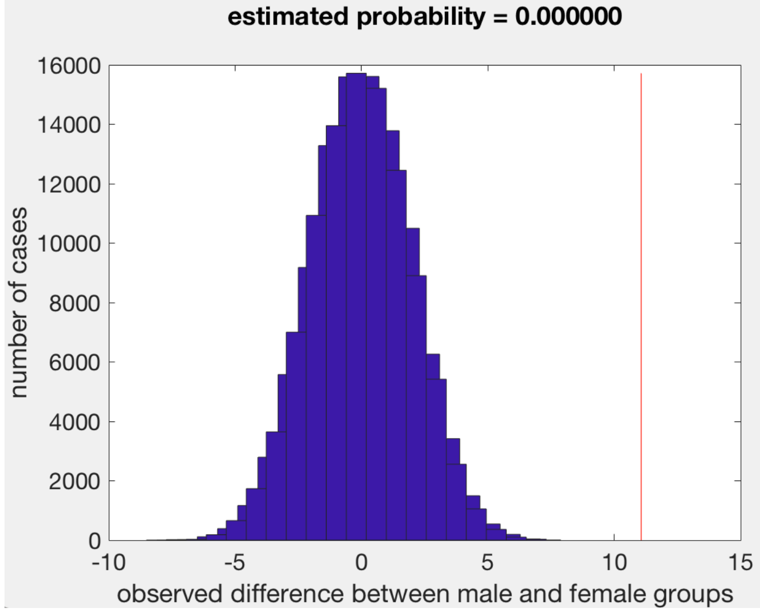
Example

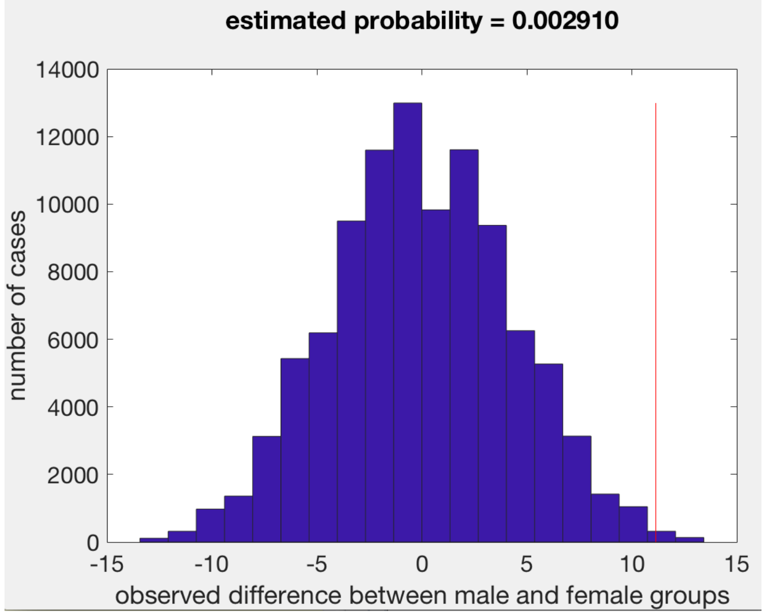
- And we have exactly seen this before in our examples showcasing the influence of sample size on the width of the sample distribution!!

Central Limit Theorem

- If the sample size (n) is large enough, $\overline{X}$ has a normal distribution with
- mean = $\mu_{\bar{x}} = \mu$
- and
- standard deviation = $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$

- **regardless of the population distribution**



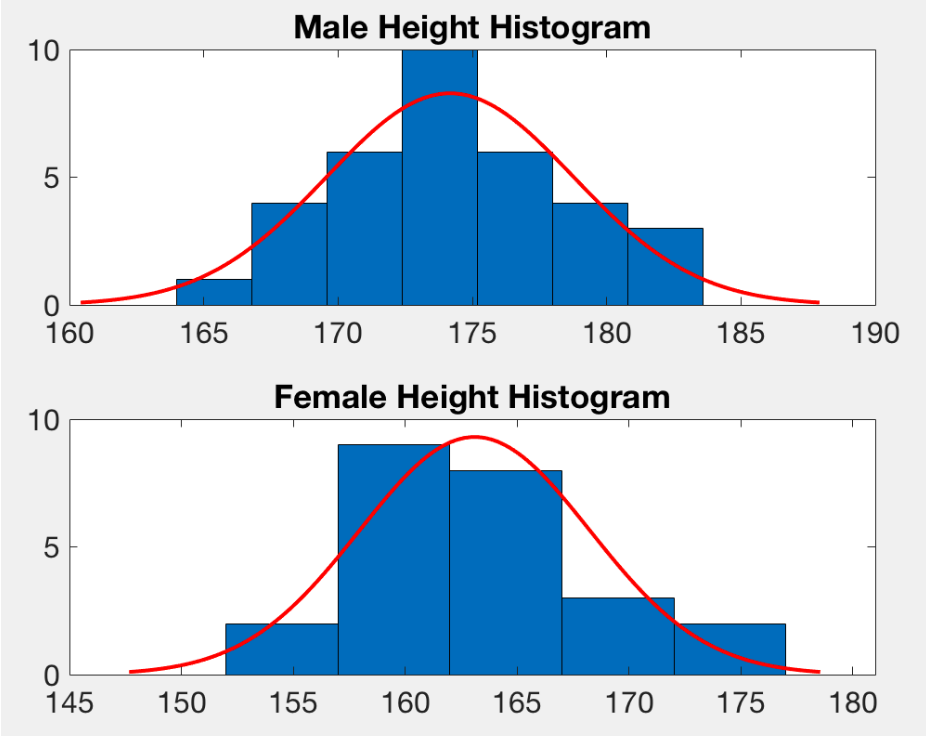


Influence on test

- If we are looking for differences between the means of two groups, the probability of having obtained the results by chance is influenced by **the size of the difference between means** and **the sample size**
- More pronounced differences will be more powerful
- The fewer samples you have, the more likely it is that any observed difference between the means is due to chance!

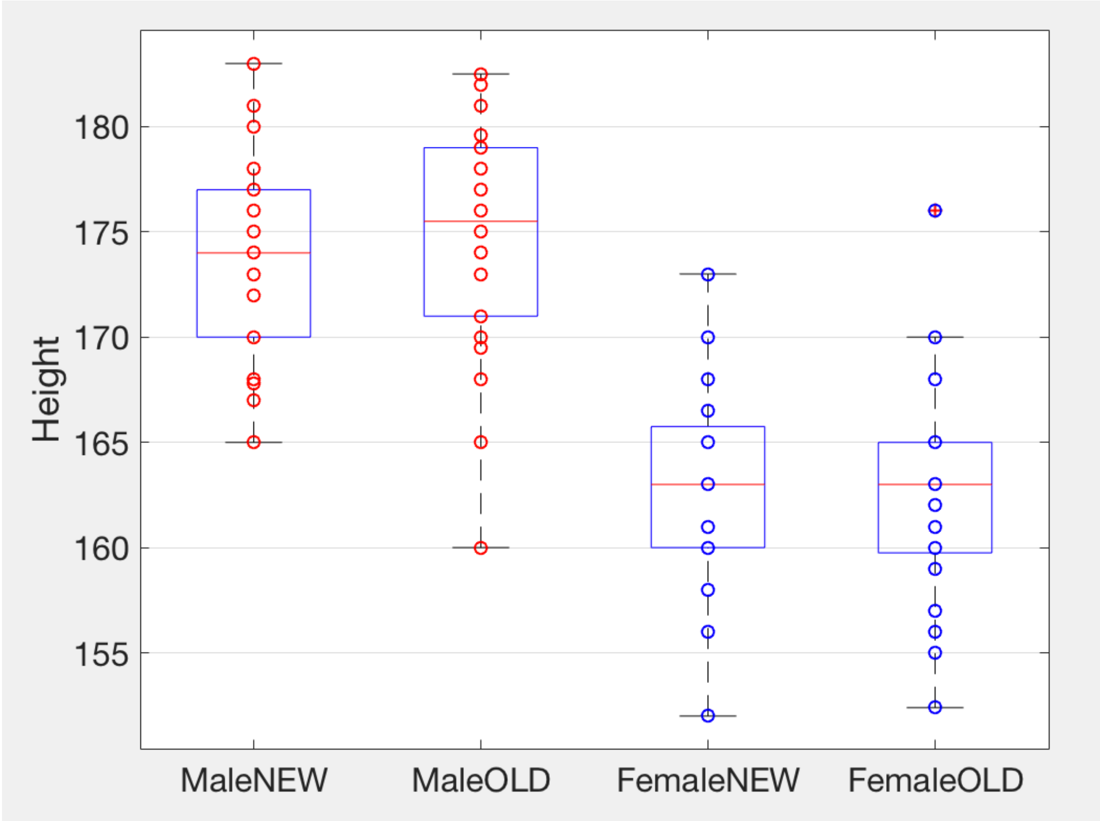
Our data

- For female heights, the means would predict
  - Female mean = 163.10cm
  - sem = std/sqrt(24) = 1.05cm
- For male heights, the means would predict
  - Male mean = 174.17cm
  - sem = std/sqrt(34) = 0.79cm
- Again, we are talking about the distribution of means
- as you can see, the distribution of heights may be very different…



Replicability

- The whole point of doing experiments is that we can repeat them
- I have done this exact poll in previous classes as well and for the last class I found this data

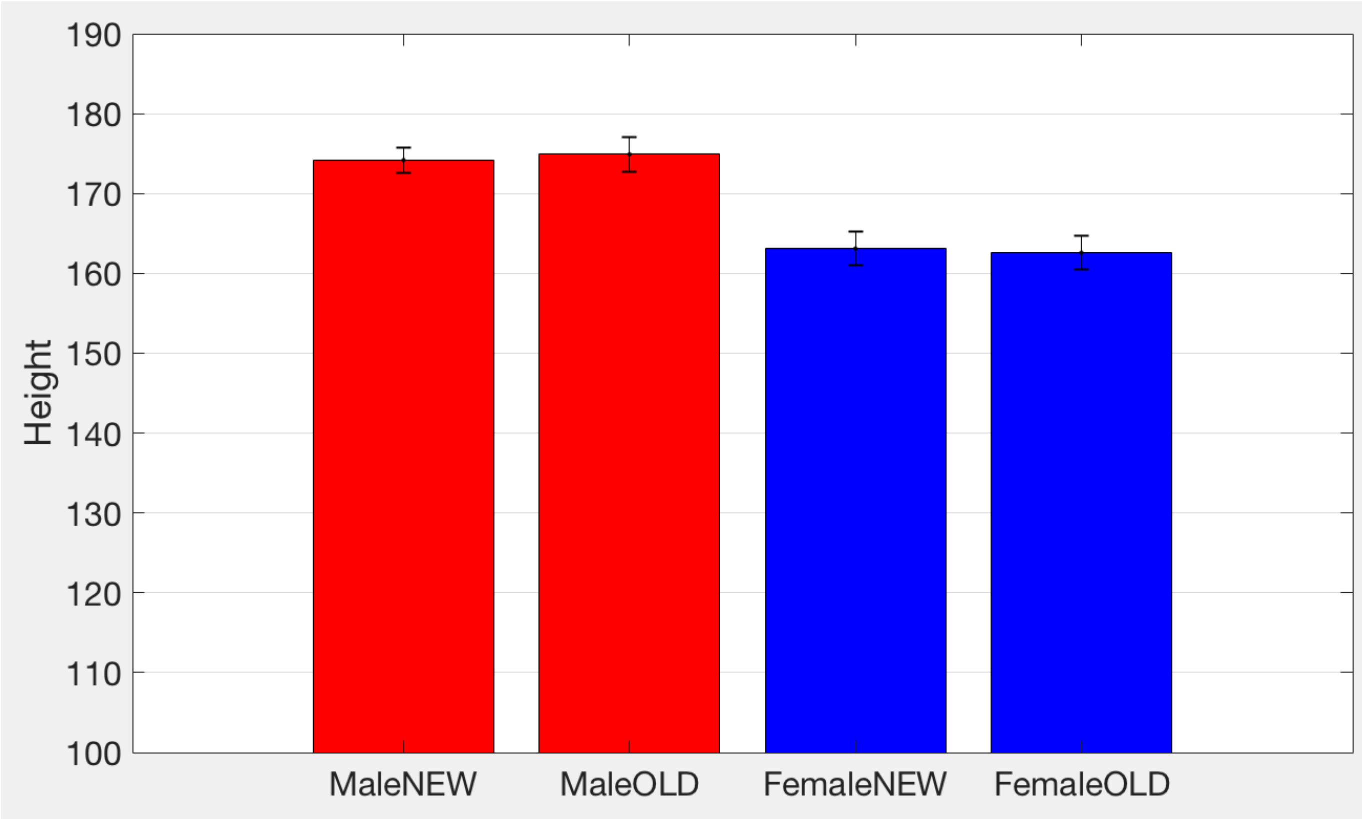


Replicability

- Remember confidence intervals? For our data, 95% confidence intervals would be ~2*SEM, so for last class:
  - Female mean = 162.58+/-2.1cm
  - Male mean = 174.91+/-2.2cm
- And for this class:
  - Female mean = 163.10cm+/-2.1cm
  - Male mean = 174.17cm+/-1.6cm

Replicability

- Those intervals clearly overlap WITHIN male and female datasets, but they do not overlap ACROSS male and female datasets



Key concepts

- With this random distribution, we can then estimate the probability that the actually observed difference between the two sample means is random
- We have just executed a real statistical test “by hand”!!

Hypothesis testing

Hypothesis tests

- Test should begin with a set of specific, testable hypotheses that can be tested using data:
  - Not a meaningful hypothesis – Was safety improved by improvements to roadway?
  - Meaningful hypothesis – Were speeds reduced when traffic calming was introduced?
- We need to formulate a hypothesis based on differences in
- Hypothesis testing is a decision-making tool.

Hypothesis Step 1

- Provide one working hypothesis – the null hypothesis – and an alternative
- The **null or nil hypothesis** convention is generally that nothing happened
  - speeds were not reduced after traffic calming – **Null Hypothesis H0**
  - speeds were reduced after traffic calming – **Alternative Hypothesis HA**
- When stating the hypothesis, the analyst must think of the impact of the potential error.

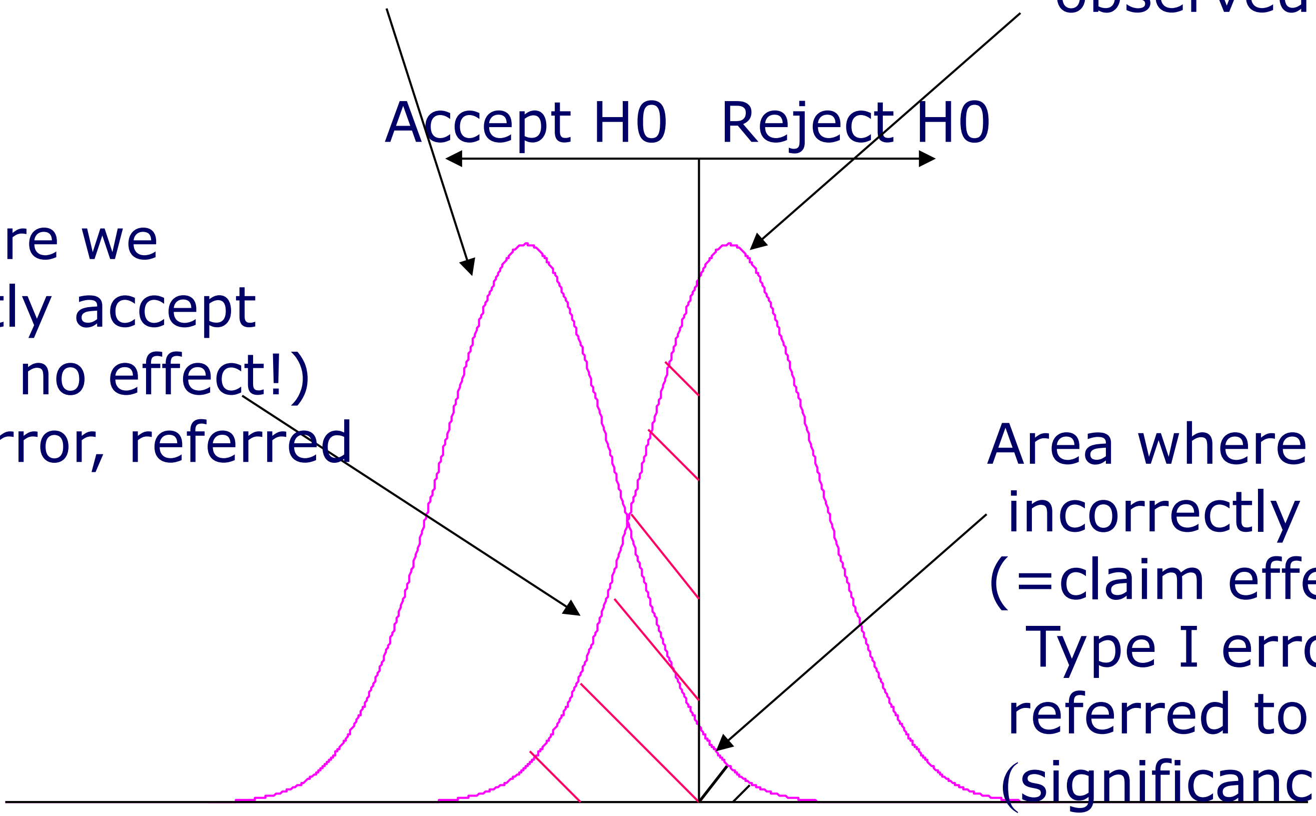
Step 2, select appropriate statistical test

- The analyst may wish to test
  - Changes in the mean of events
  - Changes in the variation of events
  - Changes in the distribution of events
- Any of these will allow you to look for differences, but they all describe different general concepts statistically

Step 3, Formulate decision rules and set levels for the probability of error

Distribution of
no effect

Distribution of
observed "effect"



Area where we
incorrectly accept
(= claim no effect!)
Type II error, referred
to as β

Area where we
incorrectly reject
(=claim effect!)
Type I error,
referred to as α
(significance level)

Type I and II errors

	True state = no effect	True state = effect
Accept H0 = no effect	All good	Type II error (False reject)
Reject H0 = effect	Type I error (False alarm)	All good

Step 4 Check statistical assumption

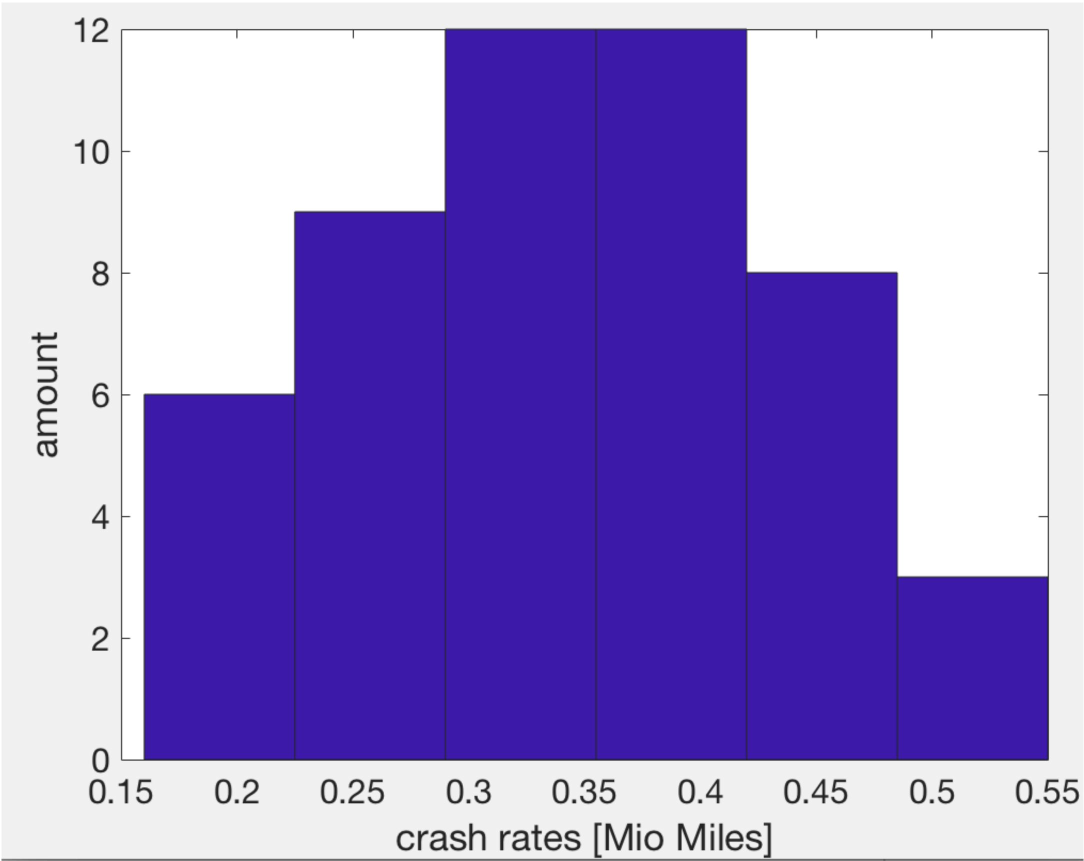
- Draw samples to check answer
- Check the following assumption
  - Are data continuous or discrete
  - Plot data
  - Inspect to make sure that data meets assumptions
    - For example, the normal distribution assumes that mean = median
  - Inspect results for reasonableness

Step 5 Make decision

- Now, we have a value from a statistical test (usually a p-value) – if this value is smaller than α, we say that
  - the result is significant, or equivalently
  - we have enough evidence to reject the null hypothesis, and we therefore may accept the alternate hypothesis
- Next, we need to interpret the finding in the context of the real world
  - if the speed reduction was significant, how LARGE is it?
  - this relates again to the EFFECT SIZE

Transportation Example

- Crash rates, in 100 million vehicle miles, were calculated for 50, 20 mile long segments of interstate highway during 2002
  - mean = 0.345
  - std = 0.095

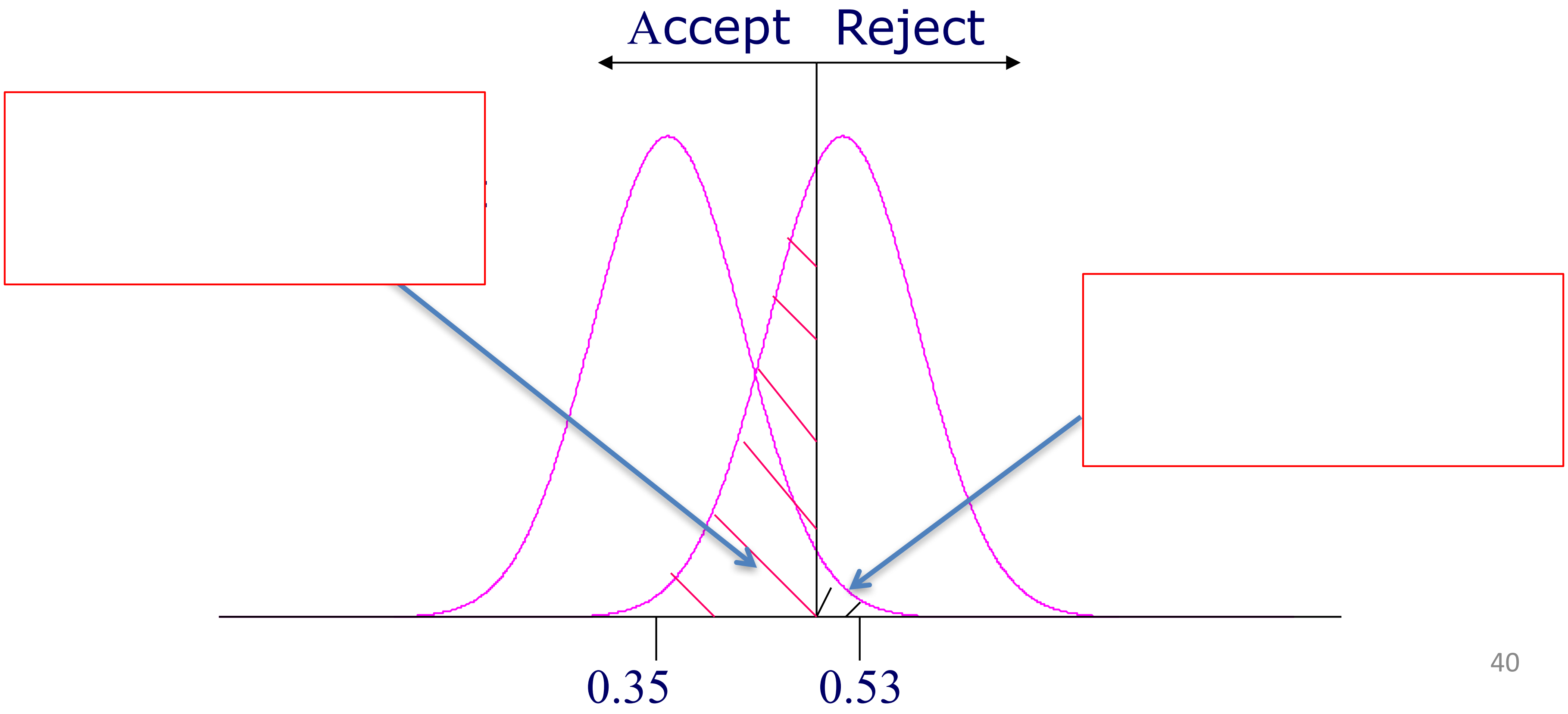


Example continued

- Crash rates were collected from non-interstate system highways built to slightly **lower** design standards.
- We again measure crash rates and find a greater average value (0.53).
- We assume both means have the same standard deviation (0.095)
- The question is: do we arrive at the **same** accident rate with both facilities?
- Our null-hypothesis is that both have the same means mf = mnf
- Can we accept or reject our hypothesis?

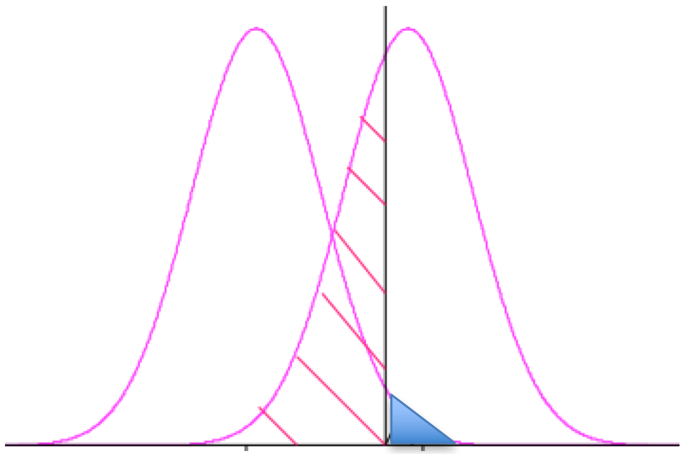
Example continued

Is this part of the crash rate distribution
for interstate highways



Example continued

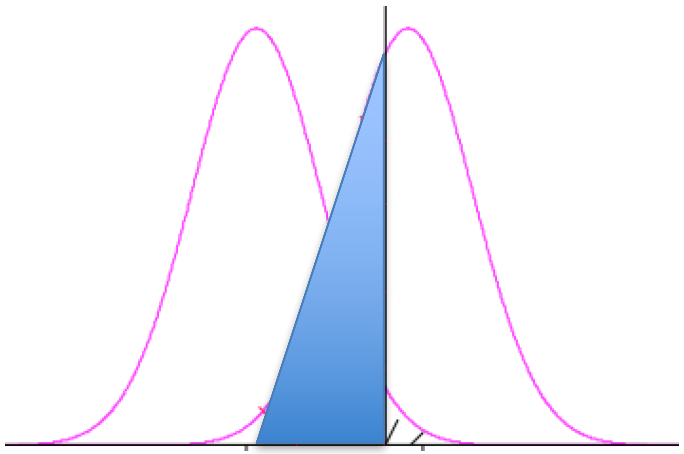
- Lets set the probability of a Type I error at 5%
- Now we need to find the value for which the area (blue triangle) is only 5% of this normal distribution
- For this, we use the z-score to transform our distribution into the standard normal distribution
  - set (upper boundary - 0.35)/ 0.095 = 1.645 (area under normal distribution corresponding to 95%)
  - Upper boundary = 0.51 < 0.53!
- Therefore, we reject the hypothesis H0



$$Z = \frac{x - \mu}{\sigma} = \frac{value - mean}{stdev}$$

Example continued

- What's the probability of a Type II error?
- Now we need to find the z-score for the other distribution!
  - (0.51 – 0.53)/0.095 = -0.21
  - This corresponds to a probability of: 41.7%
- There is a 41.7% chance of what??
  - when repeating this experiment, we have a 41.7% chance that the highways with the lower design standards will result in a mean that will not be rejected!
  - so, we would claiming that there is no effect, when in fact, they do result in higher crash rates on average



Example: Grade inflation?

Population of
5 million college
students

Is the average
GPA 2.7?

Sample of
100 college students

How likely is it that
100 students would
have an average
GPA as large as 2.9
if the population
average was 2.7?

Example continued

H_0: $\mu = 2.7$

H_A: $\mu > 2.7$

Random sample of students $\rightarrow$

Data

$n = 100$

$s = 0.6$

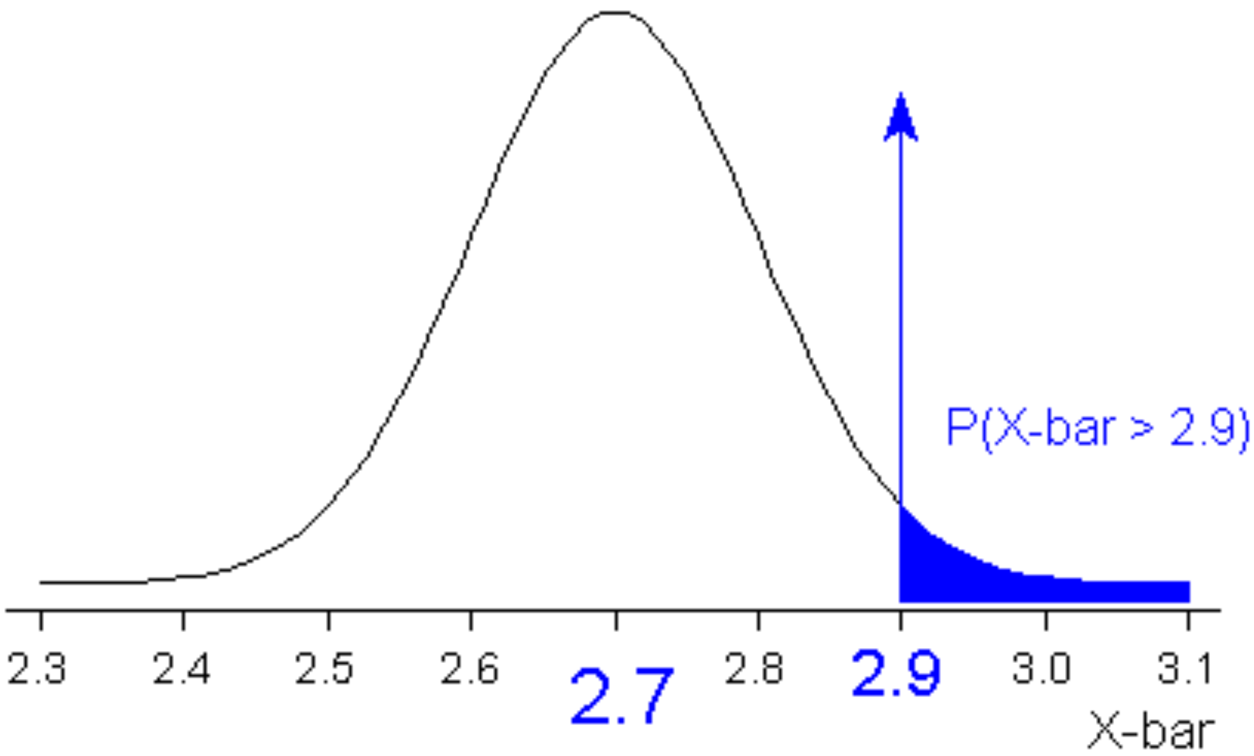
and $\overline{X}$

Decision Rule

Set significance level $\alpha = 0.05$.

If p-value<0.05, reject null hypothesis.

The p-value illustrated



How likely is it that 100 students would have an average GPA as large as 2.9 if the population average was 2.7?

Determining the p-value

H_0: μ = average population GPA = 2.7
H_A: μ = average population GPA > 2.7

If 100 students have average GPA of 2.9 with standard deviation of 0.6, the P-value is:

$$P(\bar{X} > 2.9) = P[Z > (2.9 - 2.7)/(0.6/\sqrt{100})]$$
$$= P[Z > 3.33] = 0.0004$$

Making the decision

- The p-value is "small." It is unlikely that we would get a sample as large as 2.9 if the average GPA of the population was 2.7.
- Reject H_0. There is sufficient evidence to conclude that the average GPA is greater than 2.7.

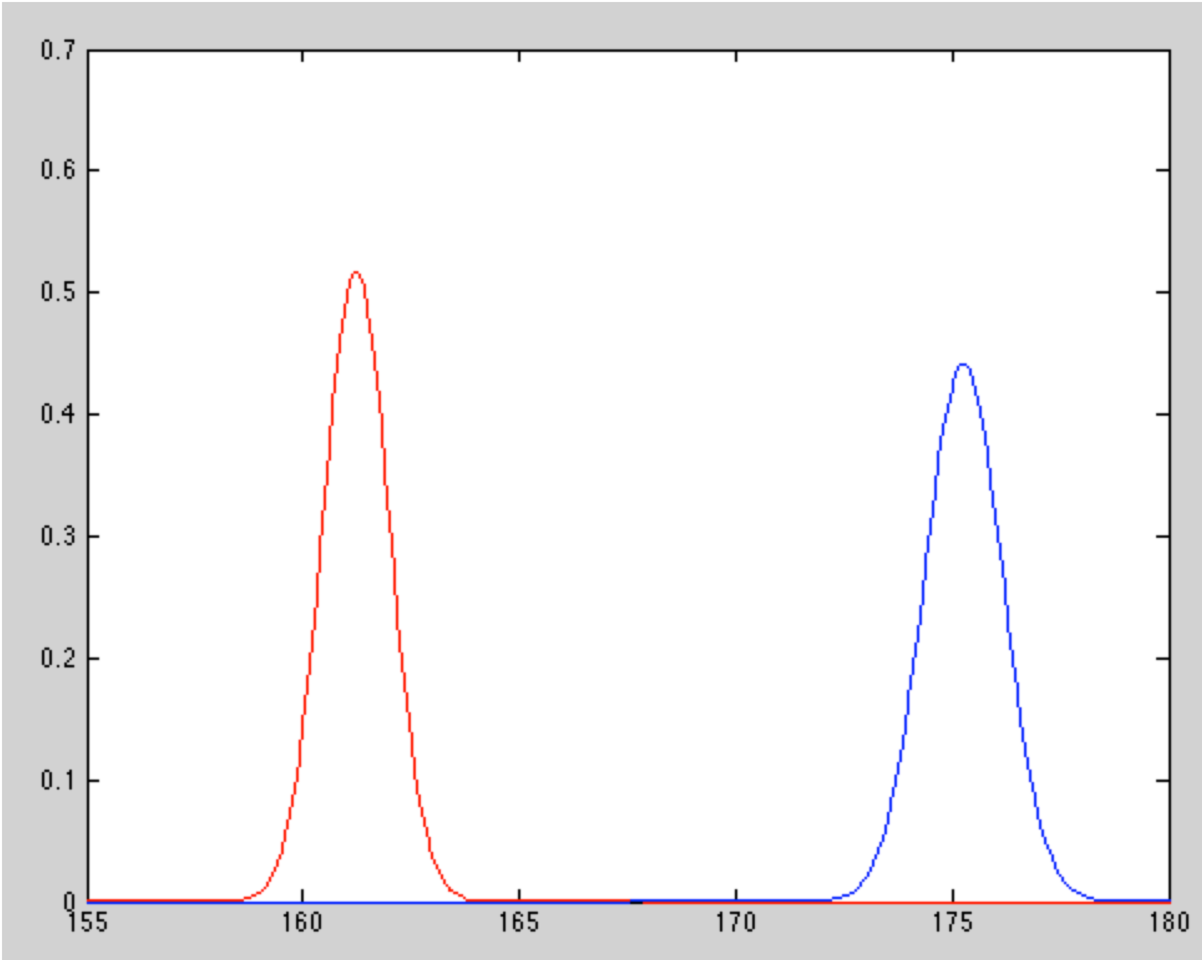
Terminology

- H_0: $\mu = 2.7$ versus H_A: $\mu > 2.7$ is called a "**right-tailed**" or a "**one-sided**" hypothesis test, since the p-value is in the right tail.
- $Z = 3.33$ is called the "**test statistic**".
- If we think our p-value small if it is less than 0.05, then the probability that we make a Type I error is 0.05. This is called the "**significance level**" of the test. We say, **α=0.05**, where α is "alpha".

Our data

- With the mean and sem values from before, the two distributions for female and male heights look like this
- No overlap!!

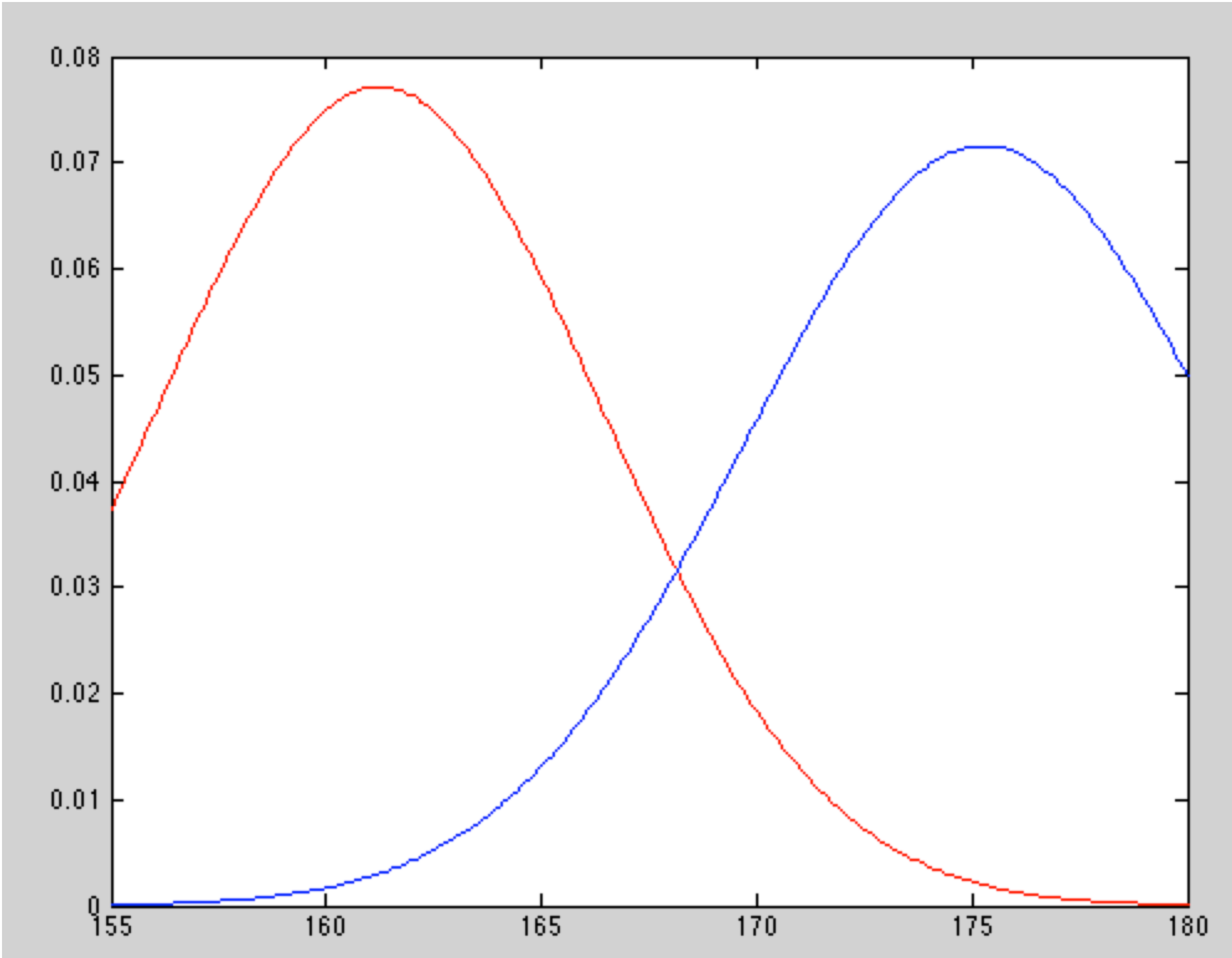
→ p=0.00000



Our data

CognitiveSystems

- Again, this is different from looking at the (fitted normal) sample distributions of your data!!
- These are much broader!



Minimize chance of Type I error...

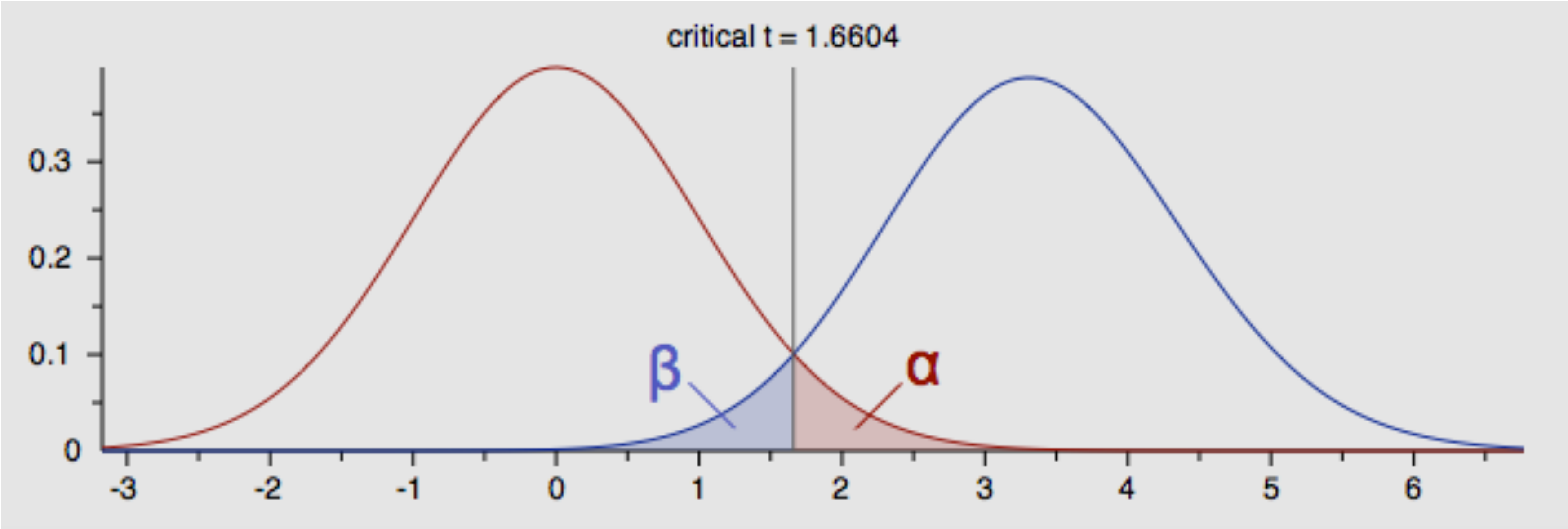
- ... by **making significance level α small**.
- Common values are α = 0.01, 0.05, or 0.10.
- "How small" depends on seriousness of Type I error.
- This decision is **not a statistical one but a practical one**
  - alpha should be small for safety analysis, drug tests
  - alpha can be larger for analysis of traffic congestion

Type II Error and Power

- "**Power**" of a test is the probability of rejecting null when alternative is true.
- **"Power" = 1 - P(Type II error=β)**
- To minimize the P(Type II error), we equivalently want to maximize power.
- But power depends on the value under the alternative hypothesis ...

Type II Error and Power

- Power = 1-β= 0.95



Factors affecting power...

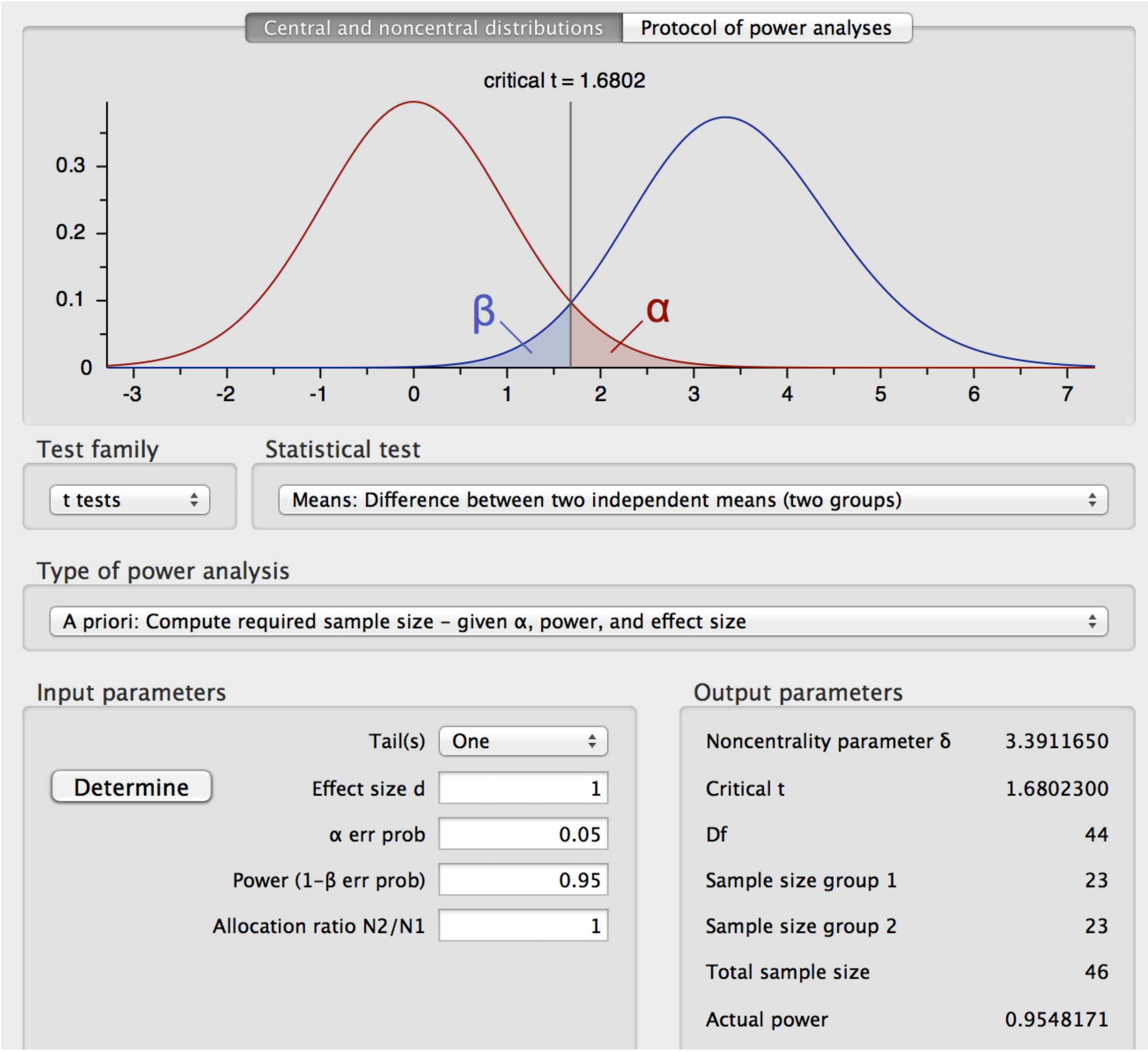
- Difference between value under the null and the actual value
- P(Type I error) = α
- Standard deviation
- Sample size

Strategy for designing a good hypothesis test

- Use pilot study to estimate std. deviation.
- Specify α
  - Typically 0.01 to 0.10.
- Decide what a meaningful difference would be between the mean in the null and the actual mean.
  - Look for small, medium, large effect (sizes)
- Decide power
  - Typically 0.80 to 0.99.
- Chose the appropriate statistical test
- Use software to determine sample size!

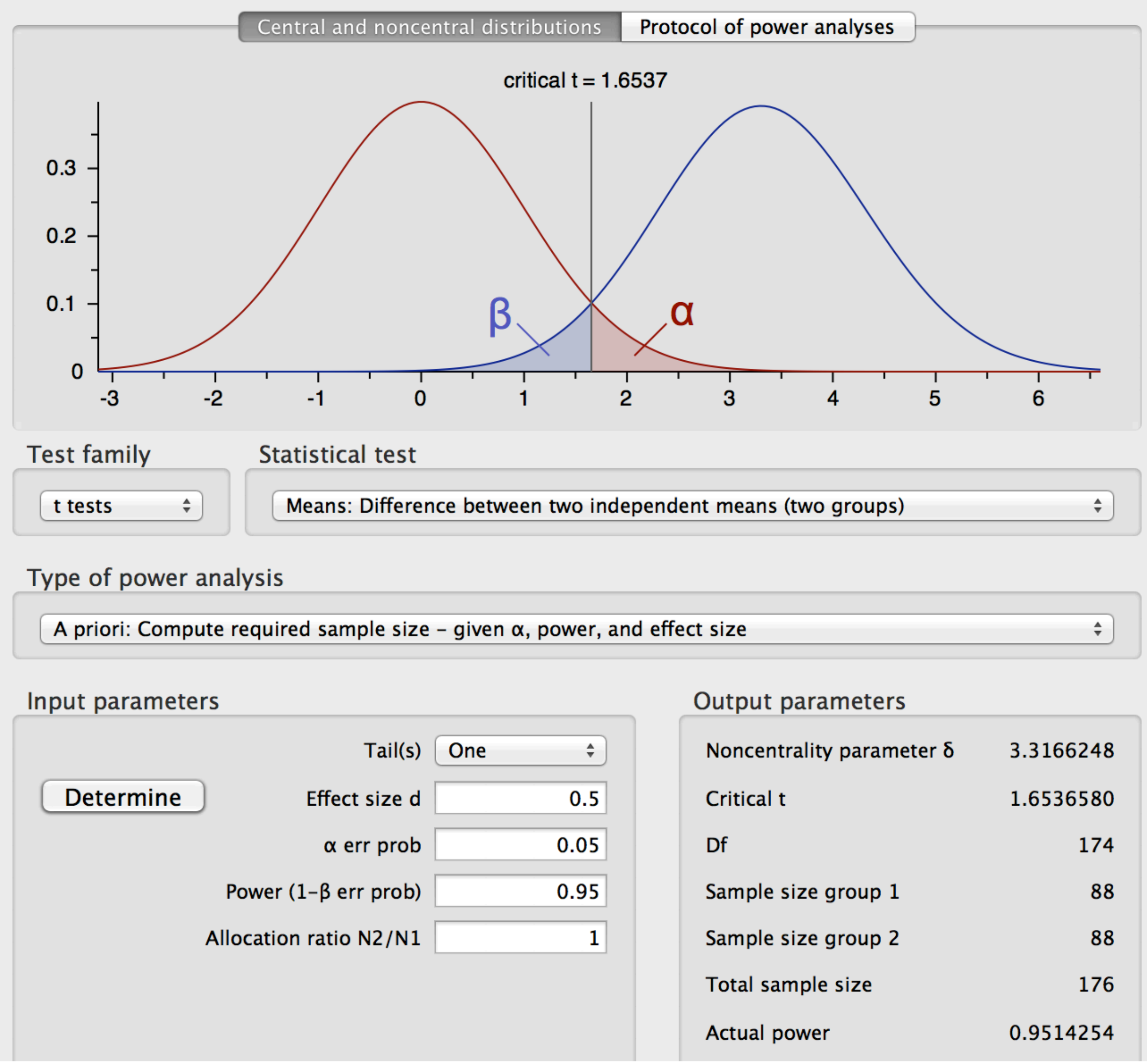
Strategy for designing a good hypothesis test

- When I want to look for a large difference between two groups with α=0.05 and 1-β=0.95 I find:
  - 46 people needed



Strategy for designing a good hypothesis test

- When I want to look for a smaller difference between two groups with α=0.05 and 1-β=0.95 I find:
  - 176 people needed!!



If sample is too small ...

- ... the power can be too low to identify even large meaningful differences between the null and alternative values.
  - Determine sample size in advance of conducting study.
  - Don't believe the "fail-to-reject-results" of a study based on a small sample.

If sample is really large ...

- … the power can be extremely high for identifying even meaningless differences between the null and alternative values.
  - In addition to performing hypothesis tests, use a confidence interval to estimate the actual population value.
  - If a study reports a “reject result,” ask how much different?

What about α and β??

- Often β is not considered in the development of the test – we usually simply set $\alpha=0.05$
  - However, in general there is a trade-off between α and β!!
- In science today, over-emphasis is placed on the level of significance of the test (p-value bias) hence avoiding false-alarms!
- The level of α should be appropriate for the decision that is being made.
  - Small values for decisions where errors cannot be tolerated and β errors are less likely
  - Larger values where type I errors can be more easily tolerated

Typical misconceptions

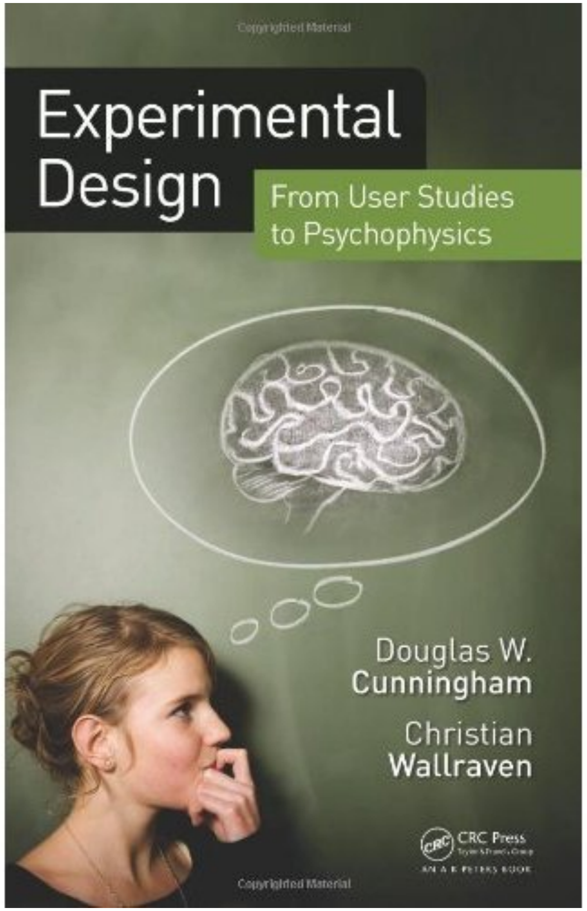
- α is the most important error
  - β is important, too!
- Hypothesis tests are unconditional
  - They do not provide evidence that the working hypothesis is true!!
  - For example, let's do 300 experiments:
    - We set $\alpha = 0.05$ and $\beta = 0.10$
    - Among the 300 experiments, we reject our hypothesis 100 times
    - So that means, we get 0.05 x 100 = 5 Type I errors [we claimed we had a result, when we didn't have one]
    - And we get 0.1 x 200 = 20 Type II errors [we claimed we didn't have a result, when we actually had one]
    - **And a total of 25 times out of 300 our test results led to the wrong conclusions!!**
  - **Hypothesis testing is NOT 100% GUARANTEED**

Key concepts

- No decision we make can prove the null hypothesis or the alternative hypothesis.
- We can only say, that given the assumptions about the sample and the population, there is enough evidence to conclude one way or the other.
- No matter what decision you make, there is **always a chance you have made an error!**

Key concepts for review

- The following slides list some important concepts that will be topics in the exam
- It should be completely fine to ace the exam if you attended class and carefully review the lecture notes
- If you would like to study more in-depth, please consult the textbook by Cunningham, Wallraven
  - other basic stats books are also fine, but will most likely be way too complicated…



Key concepts for review

- Experimental Design
  - What is a scientific theory?
  - Main types of research methods
    - pros and cons for each method!
    - dangers of correlation
  - Experiment as approximating unknown function
    - specificity versus generalization
    - within- and between-participant noise
    - repeated measures
  - Representative participants

Key concepts for review

- Experimental Analysis
  - Population versus Sample
  - Probability Distributions
    - The normal distribution and its importance
  - Central Limit Theorem
    - What does it state?
    - Why is this important?

Key concepts for review

CognitiveSystems

- Experimental Analysis – Descriptive Statistics
  - Three types of variables
  - Visualizing and summarizing categorical and quantitative variables
    - e.g., difference between histogram and barplot
    - shape of histograms
  - Measures of central tendency
    - you will need to be able to calculate those for very simple datasets
  - Measures of spread
    - you will not need to calculate these, but should definitely know the concepts and properties!
  - Concept of outliers and resistance of statistics to outliers
  - Properties of linear correlation!

Key concepts for review

- Experimental Analysis – Inferential Statistics
  - How to do tests “by hand” (only concept!)
  - Distribution of sample measures (e.g., mean)
    - You need to know standard error of the mean equation
    - Central Limit Theorem
  - Steps in Hypothesis Testing
  - Type 1 and Type 2 errors
    - meaning!
    - identify these in plots of distributions
    - you will not need to calculate those!

Final Caveat

- Start to review early!
- If you have any question about the course contents, please contact me as soon as possible!!