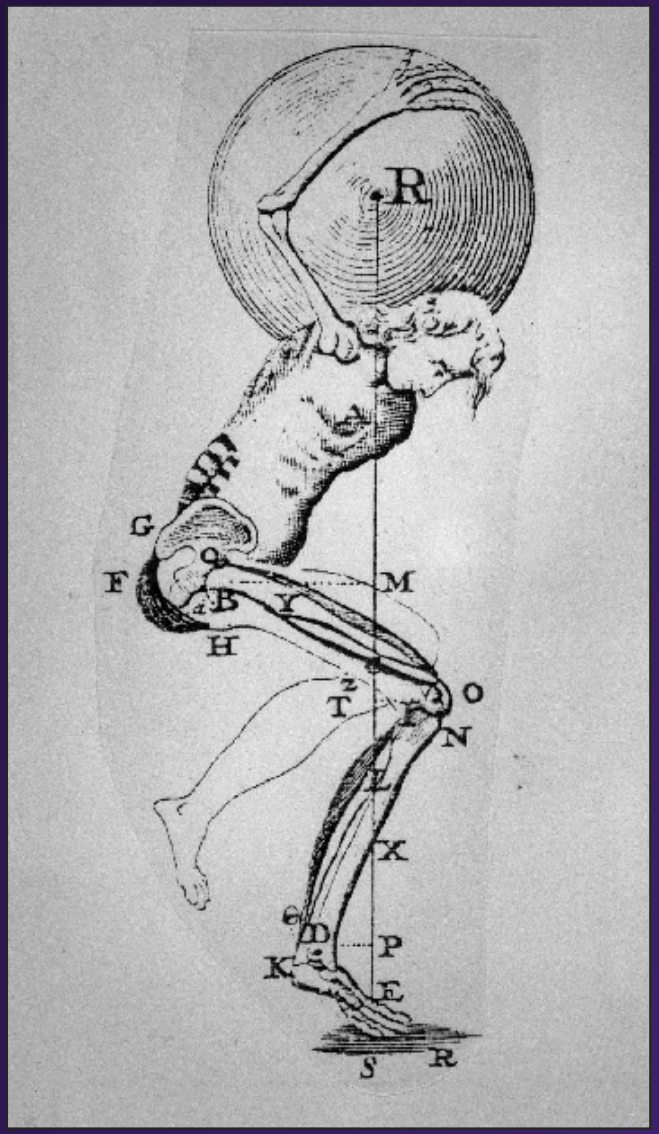


Musculoskeletal Biomechanics

BIOEN 520 | ME 527

Session 16B

Intro to Biostatistics

Statistical Overview for Biomechanical Engineering

Jane Shofer, MS

Department of Psychiatry and Behavioral Medicine and CSDE, UW

RR&D, VAPSH

Pithy opening quote

- All models are wrong

Pithy opening quote

- All models are wrong
- Some models are useful

George Box

Example

- 2 groups of patients who had ankle osteoarthritis
- Group 1 had a coronal plane deformity in the affected limb, n=48
- Group 2, neutral alignment, n=64
- Main outcome: MFA—high score means poor functioning

Example

- Question: Is mean MFA different between those with a coronal deformity vs. those neutrally aligned?

Graphing the data

Dynamite plot

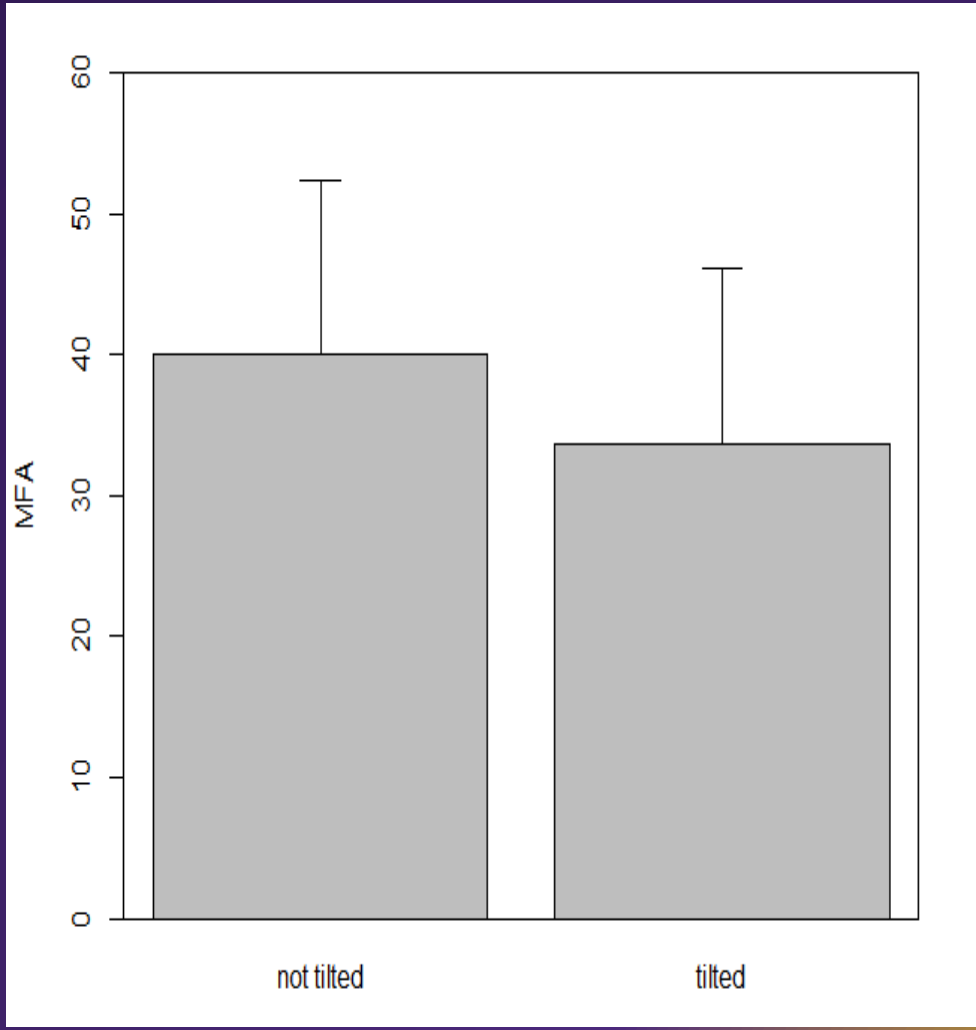
Problems:

1. A lot of space to summarize 4 numbers

2. Error bar in one direction

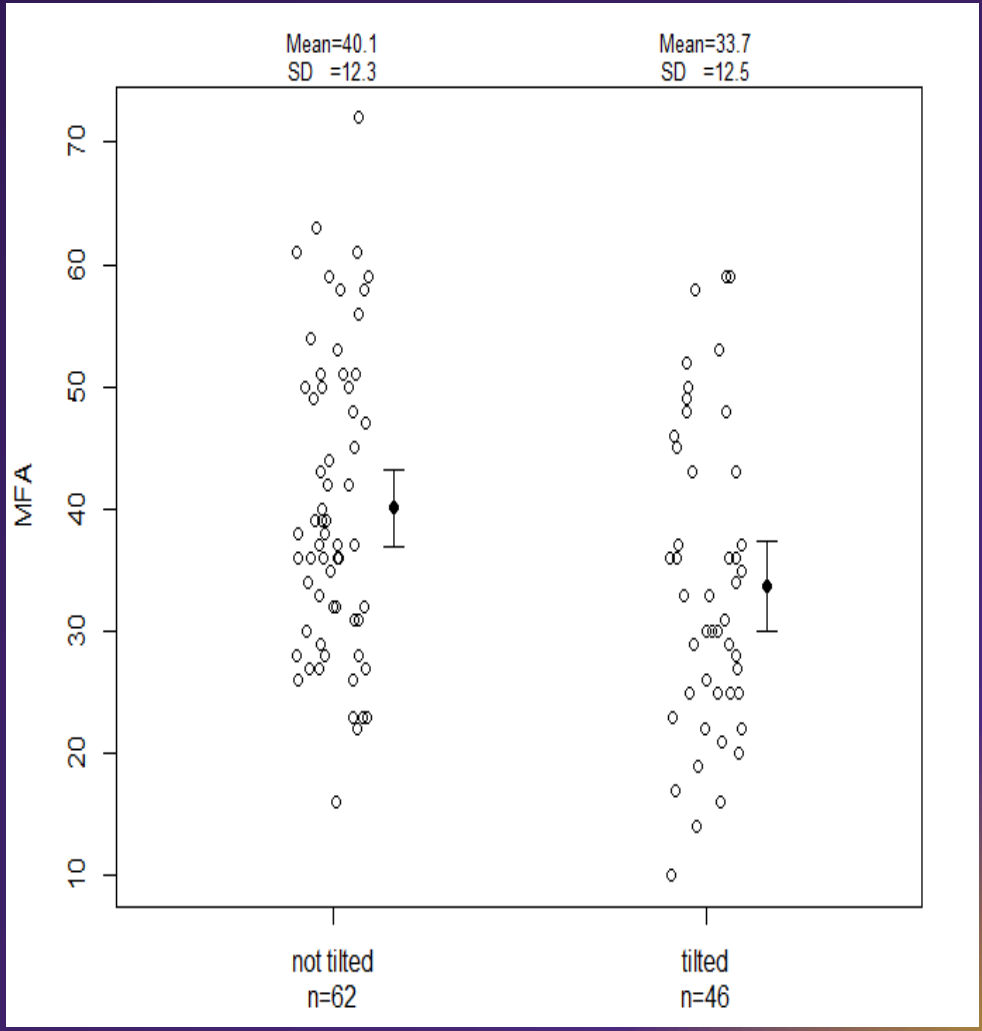
3. No information as to the shape of distribution of MFA between groups

4. No info about potential outliers



Graphing the data

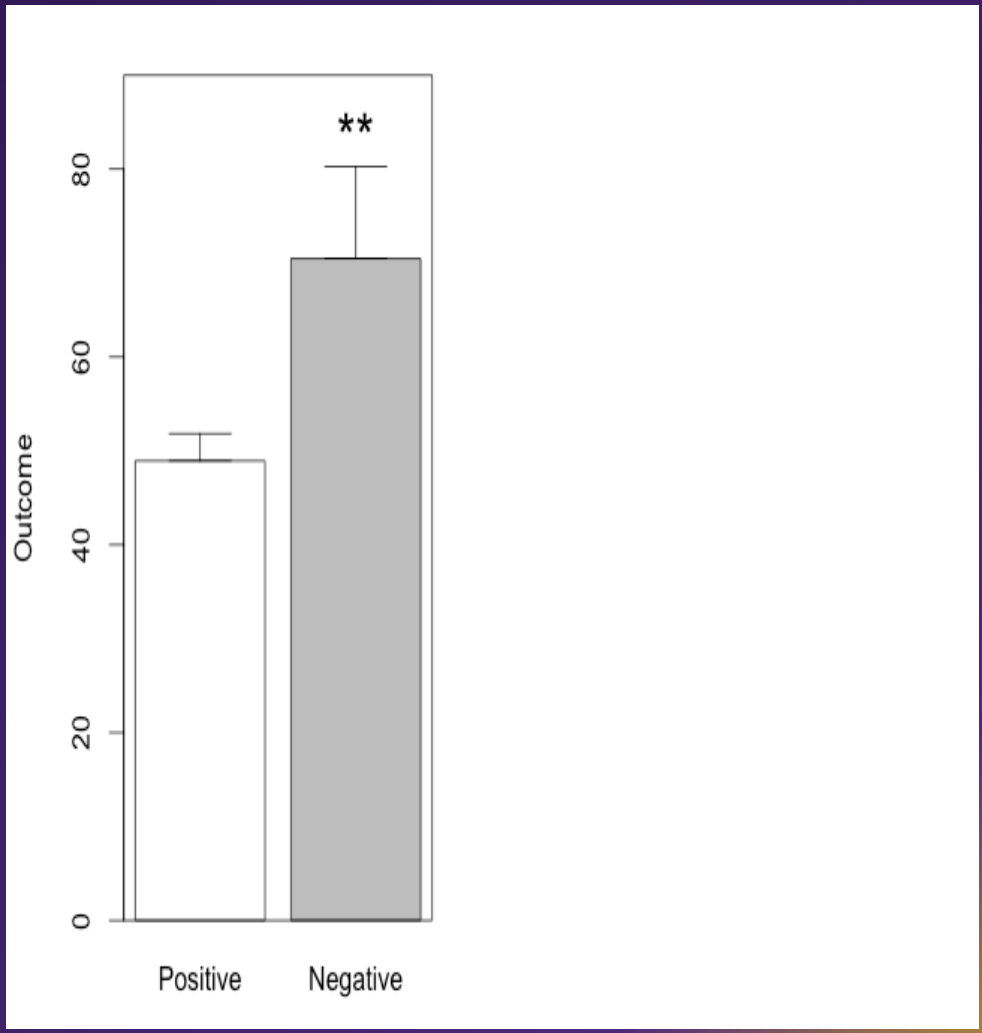
Strip plot with means and 95% confidence intervals



Graphing the data

Another example

The means of these 2 groups were determined to be statistically different

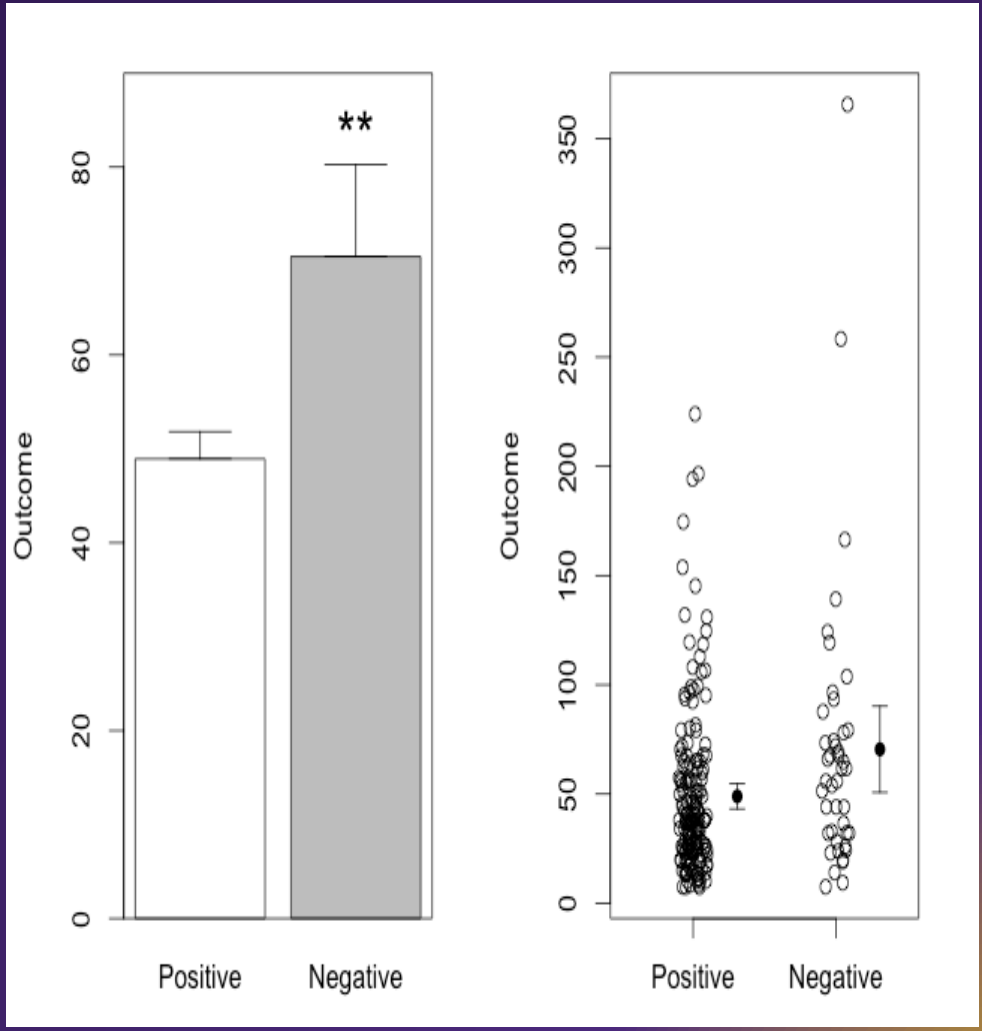


Graphing the data

Another example

The means of these 2 groups were determined to be statistically different

However, the mean for the Negative group was strongly influenced by an outlier



Why do we use statistics?

- Separating the signal from the noise
- Using data from a sample to generalize to a larger population (in this case generalizing to the population of ankle OA patients in the US.)

2-sample t-test

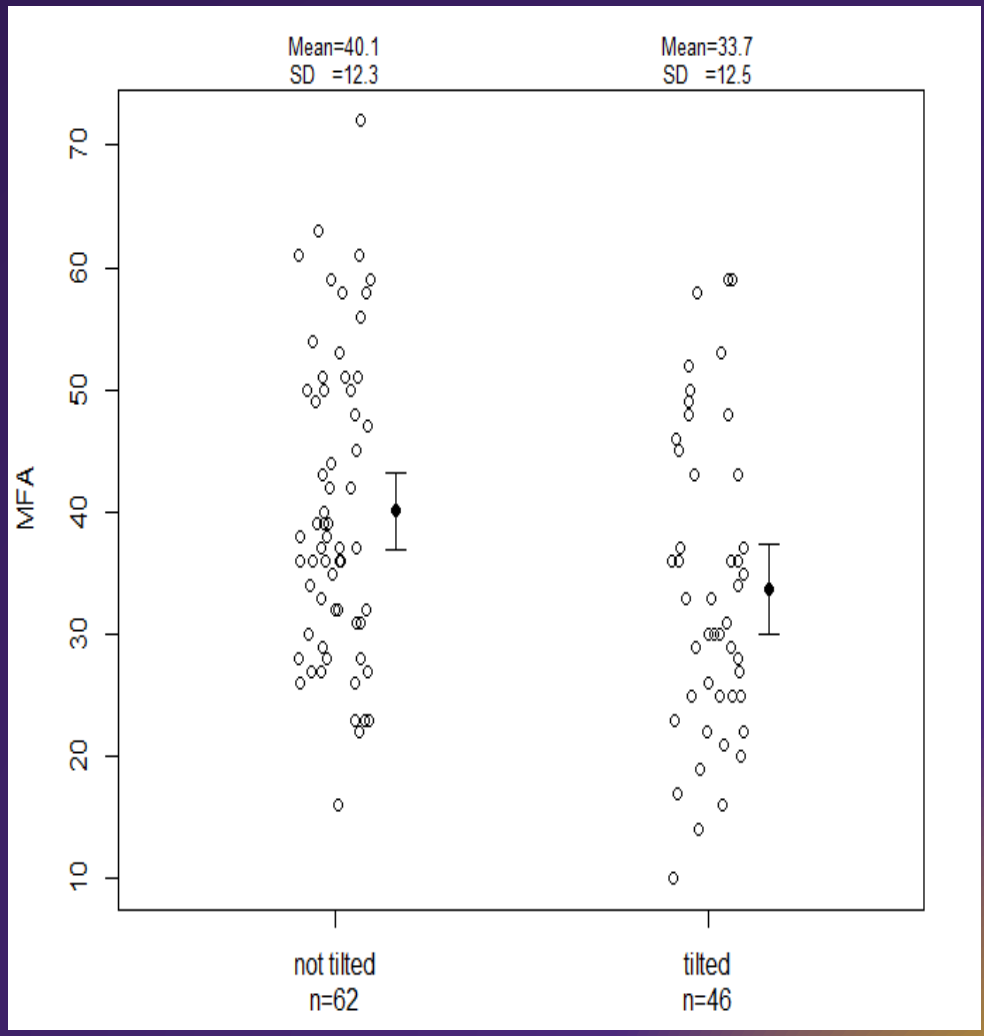
- Signal divided by noise (in this case the difference between the means divided by the standard deviation of the difference)
- The larger the ratio, the stronger the signal
- We assess the strength of the signal by assigning a probability to its occurrence.

Back to our example

The difference in the means (neutral minus tilted) = 6.4;

Standard deviation of the difference (also known as the standard error of the mean) = 2.4

Signal/noise = mean/SE = 2.6



Hypothesis testing

- Structure your research question as a “null” hypothesis, H0, vs. an “alternative” hypothesis, H1
- Null hypothesis- no difference in mean MFA between those with a coronal deformity and those neutrally aligned, i.e., no association between MFA and coronal deformity
- Alternative-mean MFA for those with coronal deformity differs from the mean MFA for those neutrally aligned, i.e., association between MFA and coronal deformity

Just a little bit of theory

- Why do we set up our hypotheses this way?
- Most statistical theory based on the distribution of the null hypothesis
- We assume a distribution for the null hypothesis and assign probability of getting a particular outcome based on the null hypothesis

The Normal Distribution

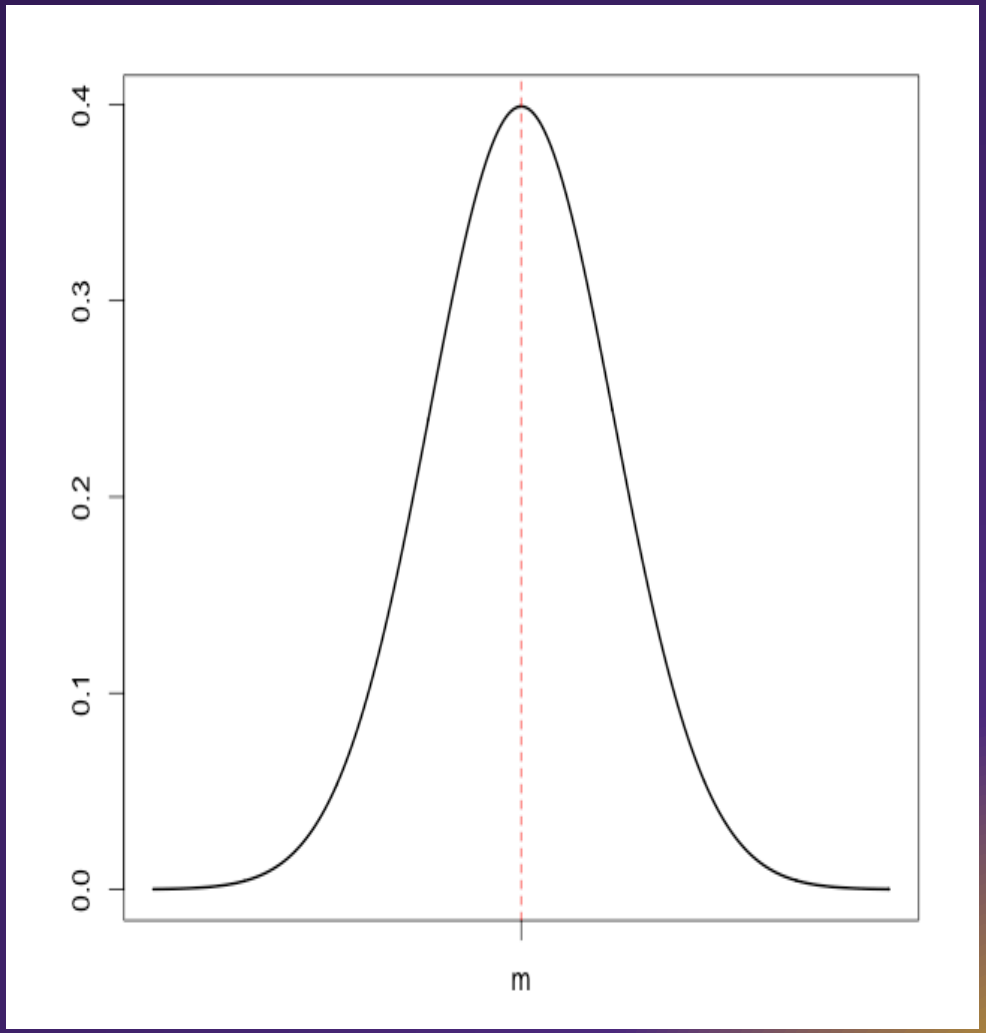
The “bell” curve

Used to assign probability to the occurrence of a result

Average (mean) is at the peak

Symmetric around the mean

Central limit theorem: if you have a large enough sample, the mean of that sample will come from an approximate normal distribution, regardless of the distribution of the data in the sample



The Normal Distribution

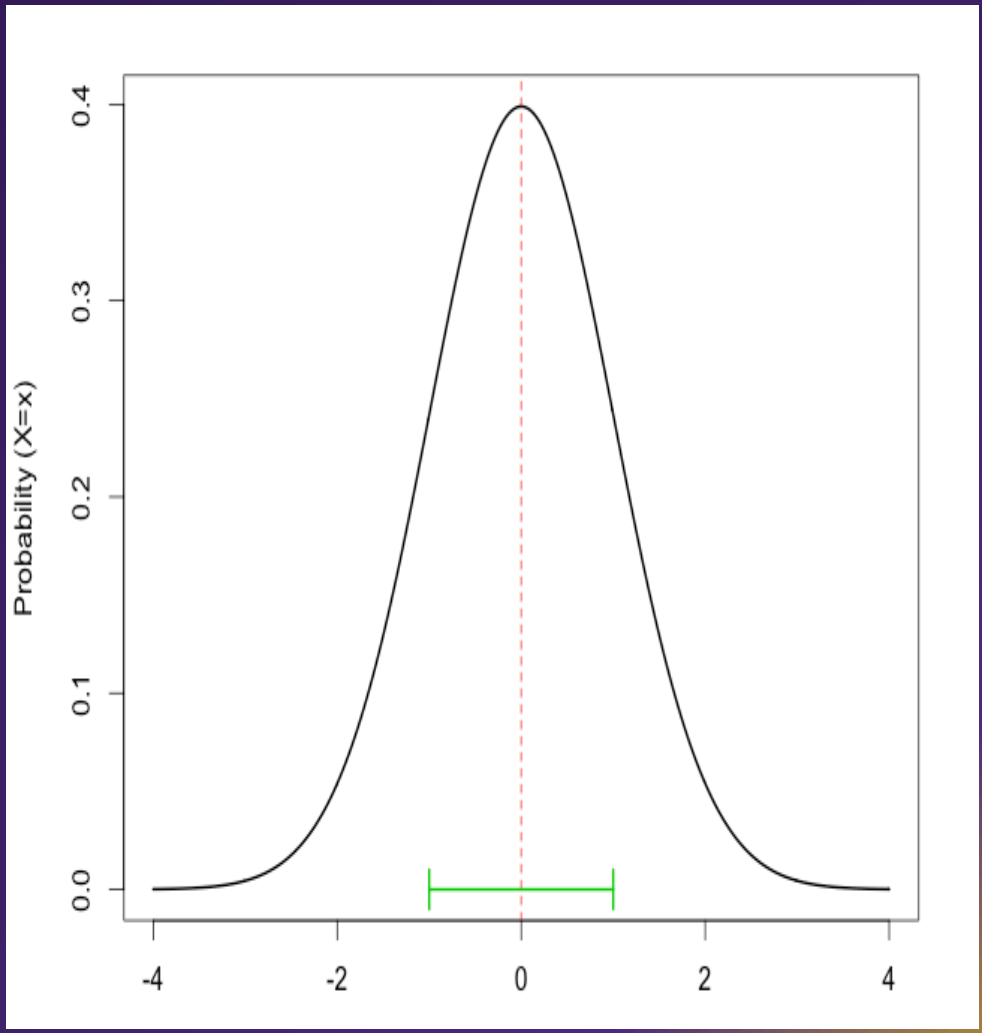
The **Standard** Normal Distribution

Mean = 0

SD = 1

Any data can be “standardized” by subtracting the mean and dividing by the SD

Basis for many statistical tests

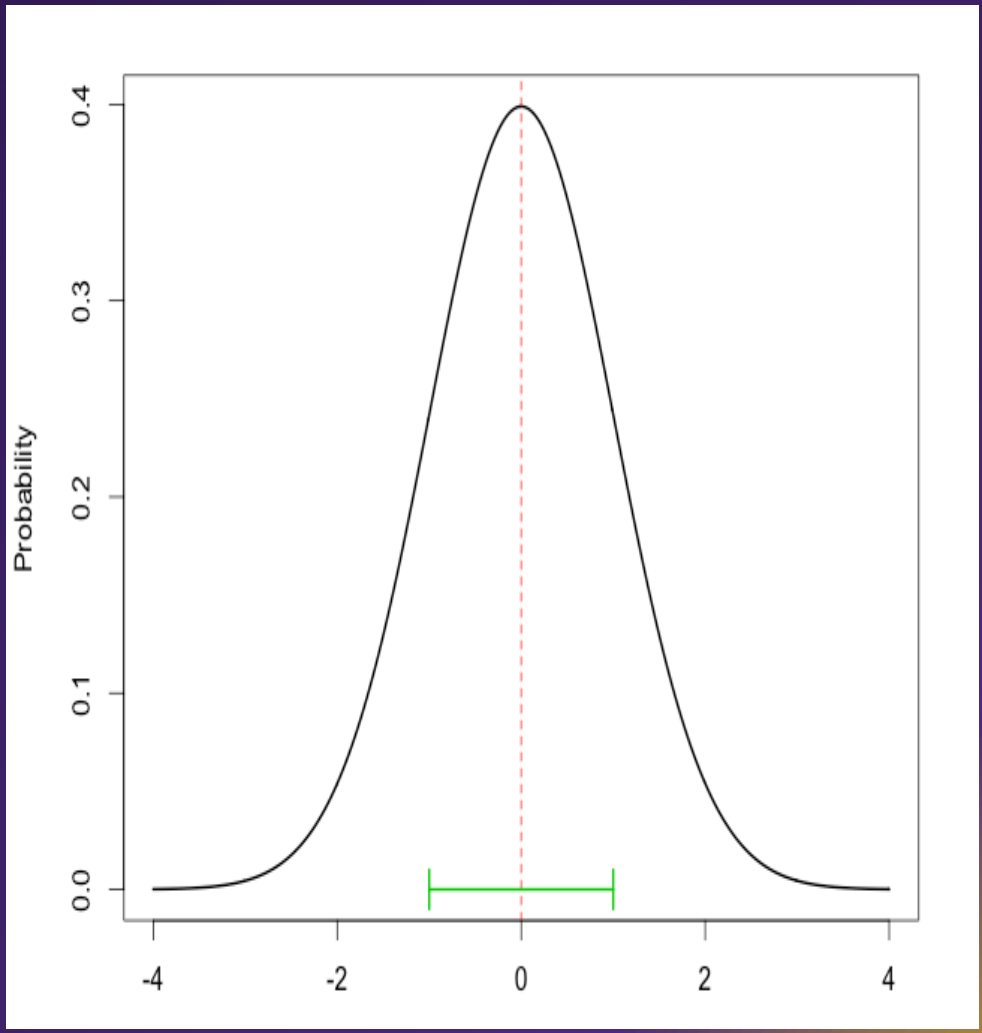


Hypothesis testing

From our example, under our null hypothesis, the difference in the mean MFA between those with coronal deformity and those neutrally aligned is zero.

We assume the standardized difference comes from a normal distribution with mean zero and SD 1—this is the null hypothesis

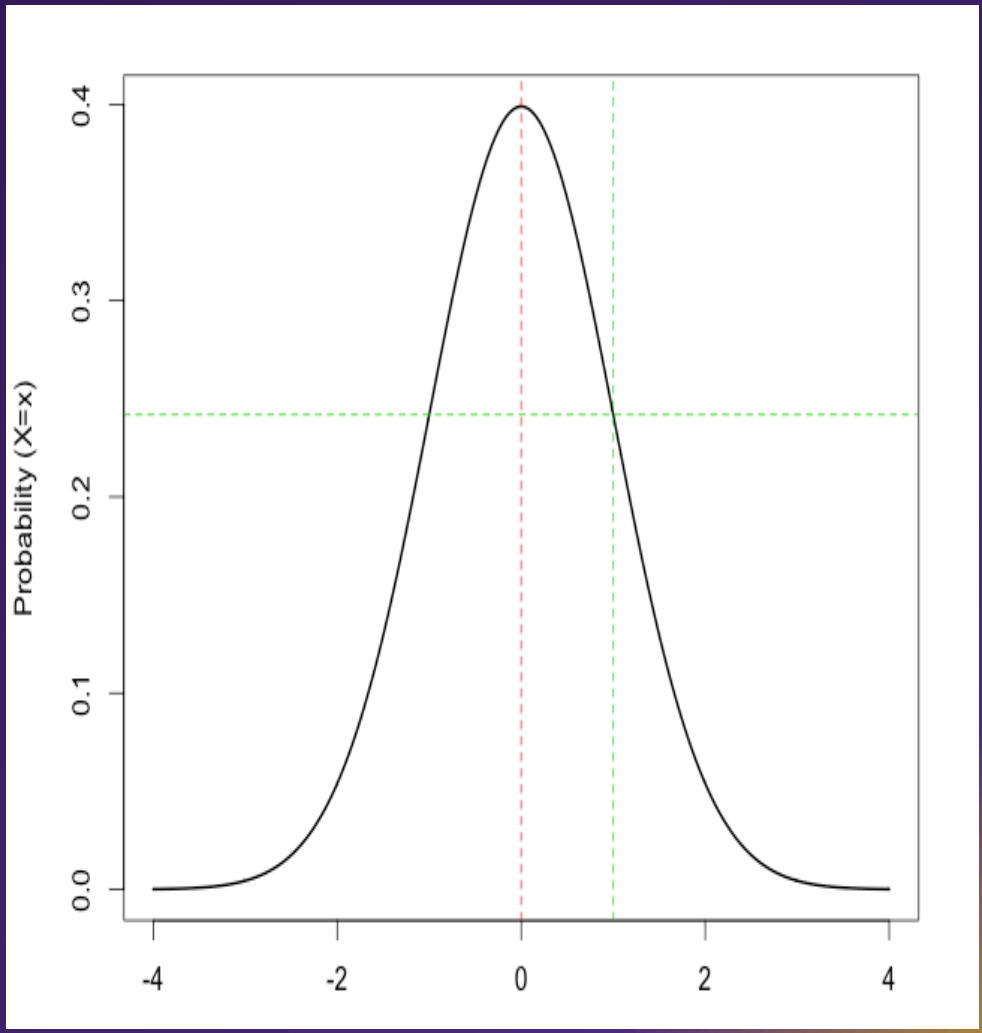
Alternative hypothesis: mean difference is not zero.



Hypothesis testing

Suppose our standardized mean difference=1

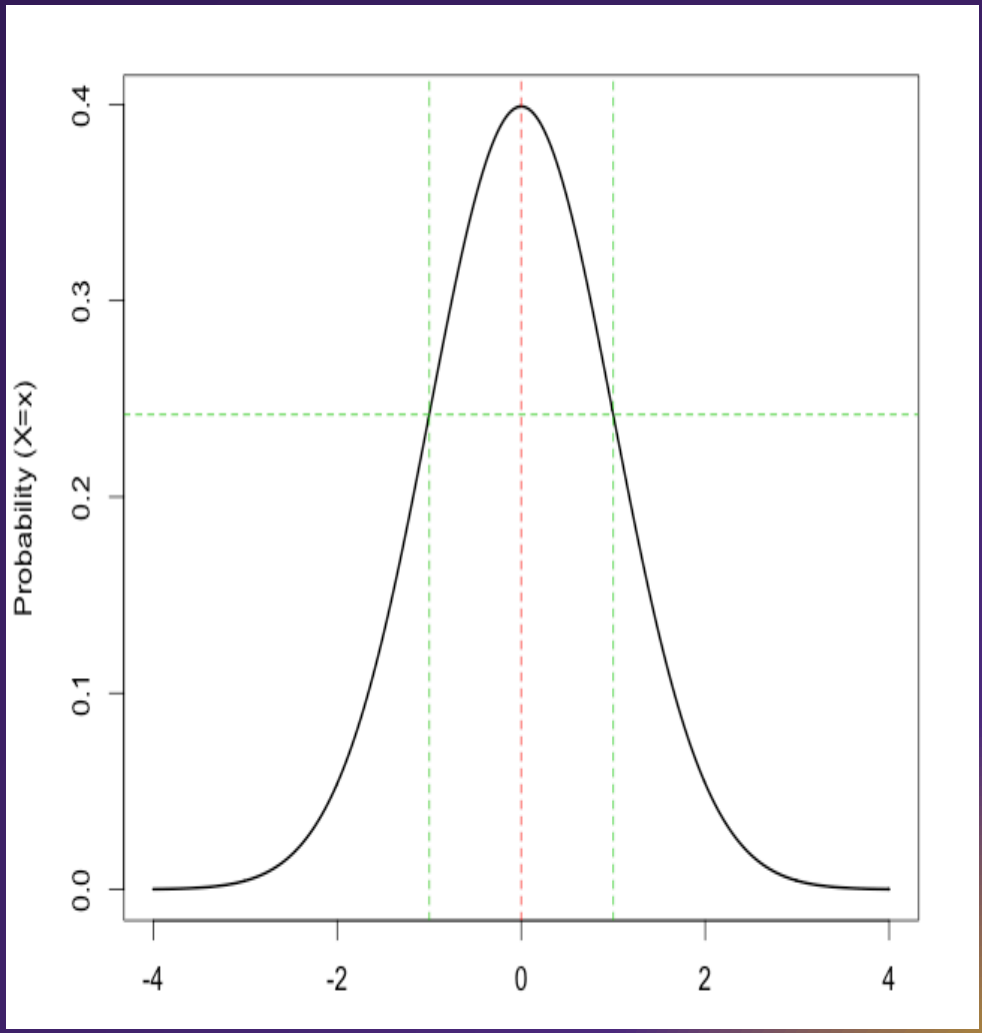
Based on the null hypothesis that the true mean difference is zero, the probability of getting a difference of 1=0.24



Hypothesis testing

Note that the probability is the same if the difference = -1

i.e., symmetry around zero



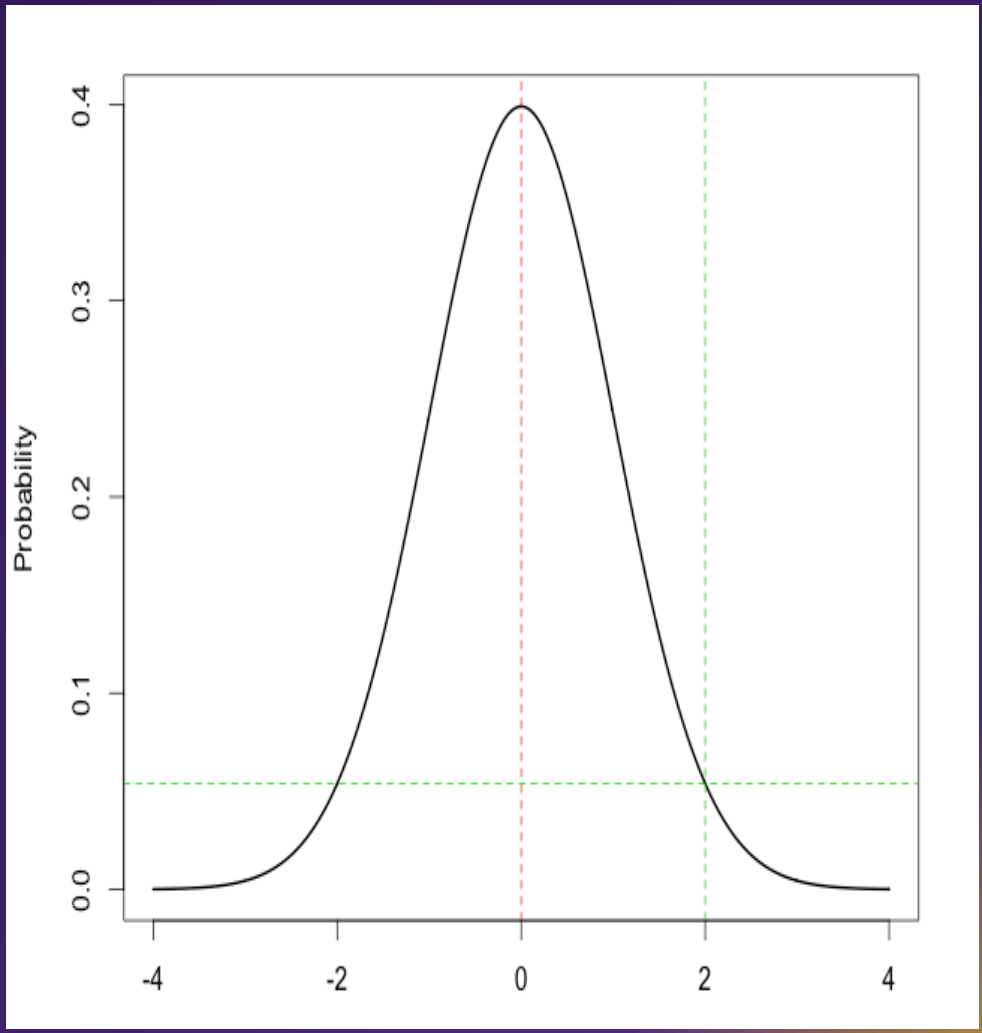
Hypothesis testing

Suppose the difference between groups = 2

Probability that this difference comes from a standard normal distribution with mean zero = 0.05

The farther the difference is from zero, the less likely that the difference is zero.

i.e., the more likely the "alternative" hypothesis is true

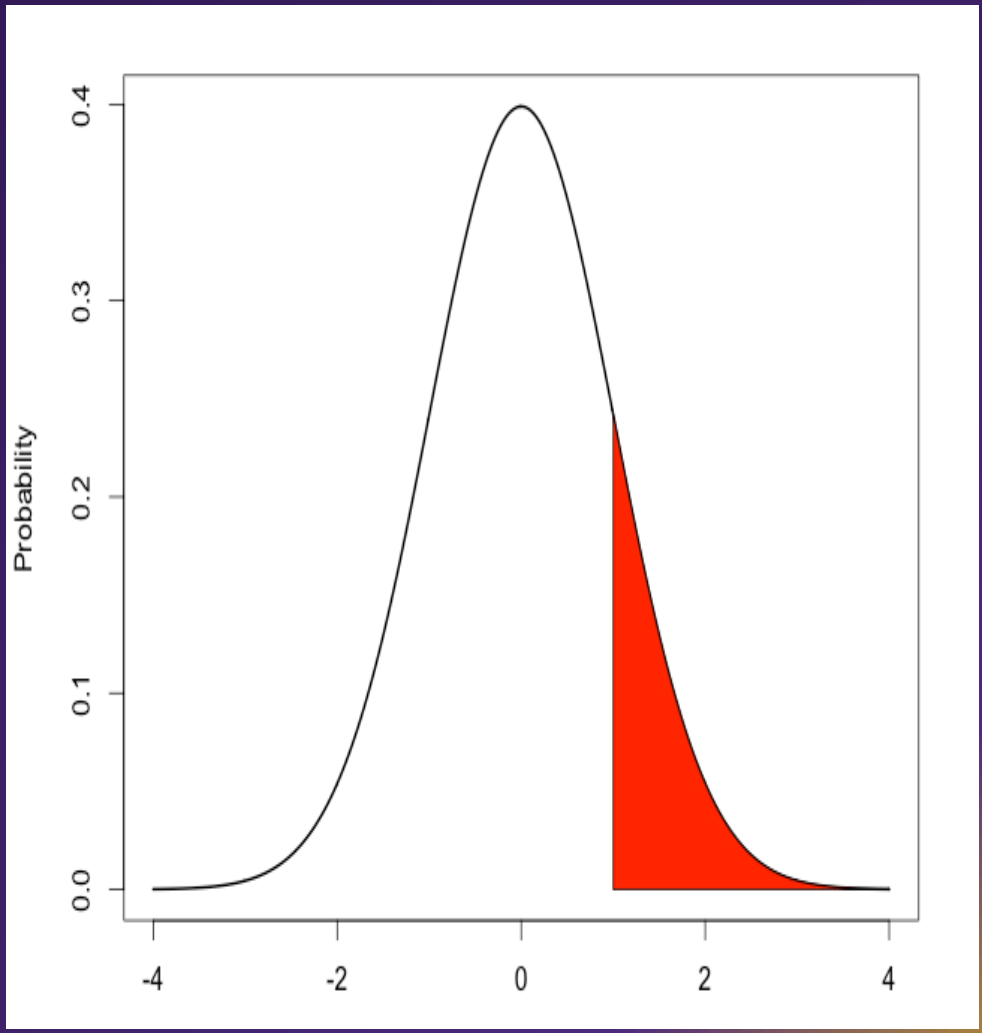


Hypothesis testing

In most cases, we are interested in whether the difference in means **is greater** than a certain value, not whether the difference **equals** a certain value.

We obtain this probability using the area under the curve for the standard normal where total AUC= 1.0

The area shaded in red represents the probability that the difference between means is >=1 = 0.16



Hypothesis testing

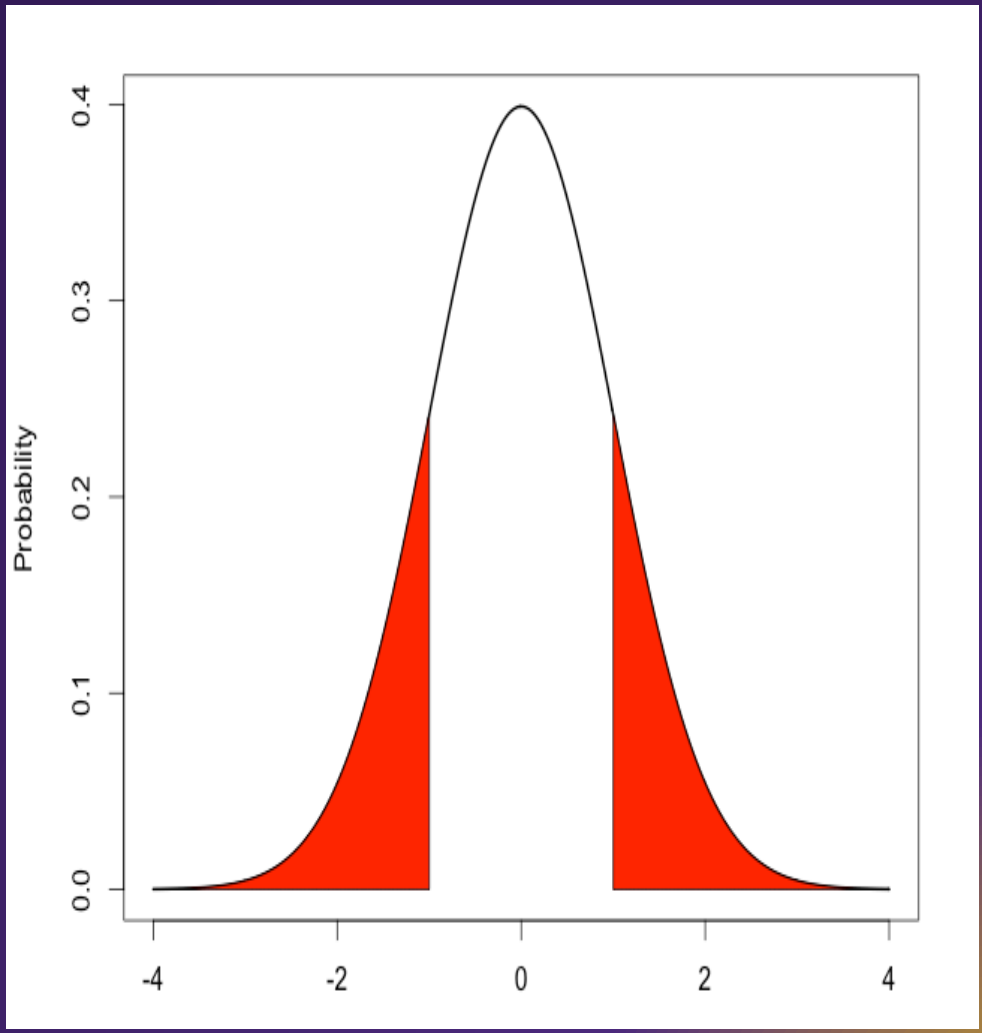
Most times we don't want to assume the difference is only in one direction.

Thus we carry out a "two-tailed" test.

The shaded area corresponds to the probability that the difference between means **in either direction** is >= 1 (or that the absolute value of the difference is >=1)

= 0.32

This probability is known as the **p-value**

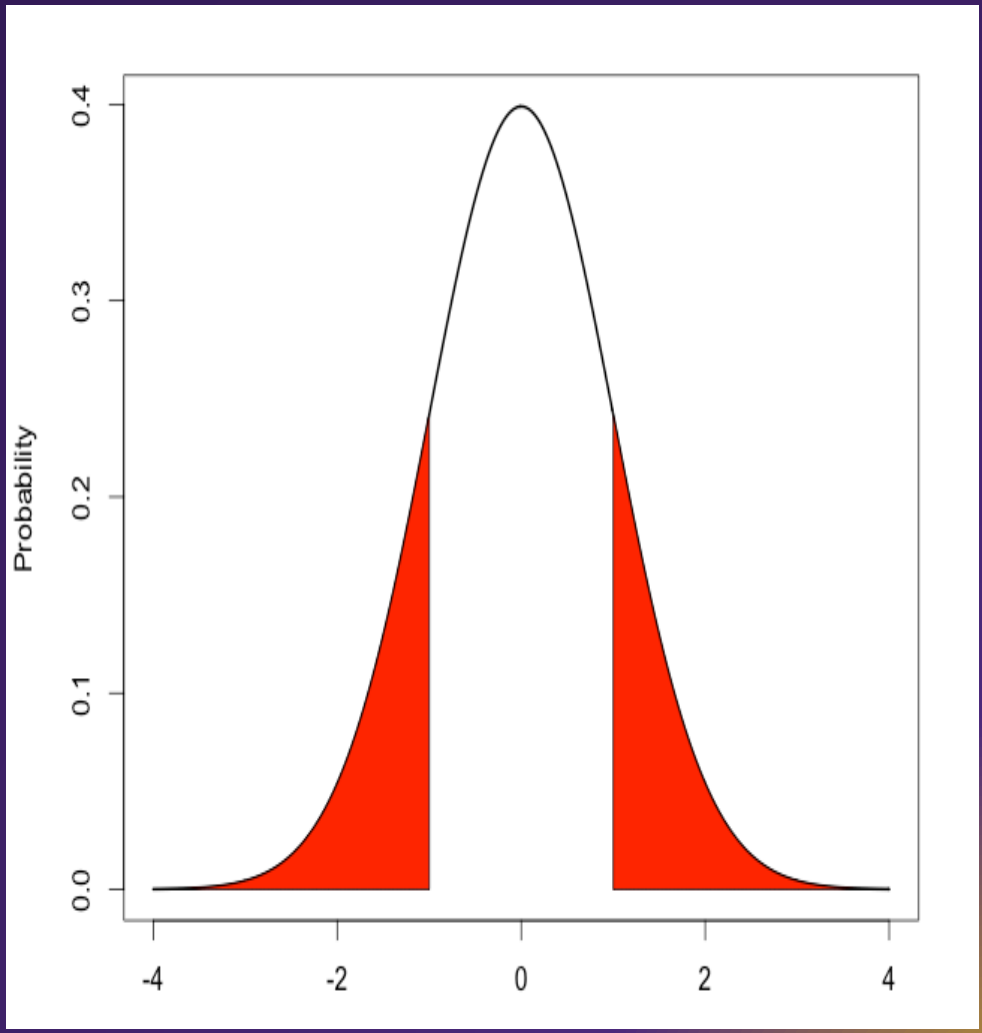


Hypothesis testing

The p-value =

The probability, given the null hypothesis of no differences between means, that you obtain a difference at least as large as (in this case) a difference of 1.

The smaller the p-value, the more likely that our alternative hypothesis is true.



Back to our example

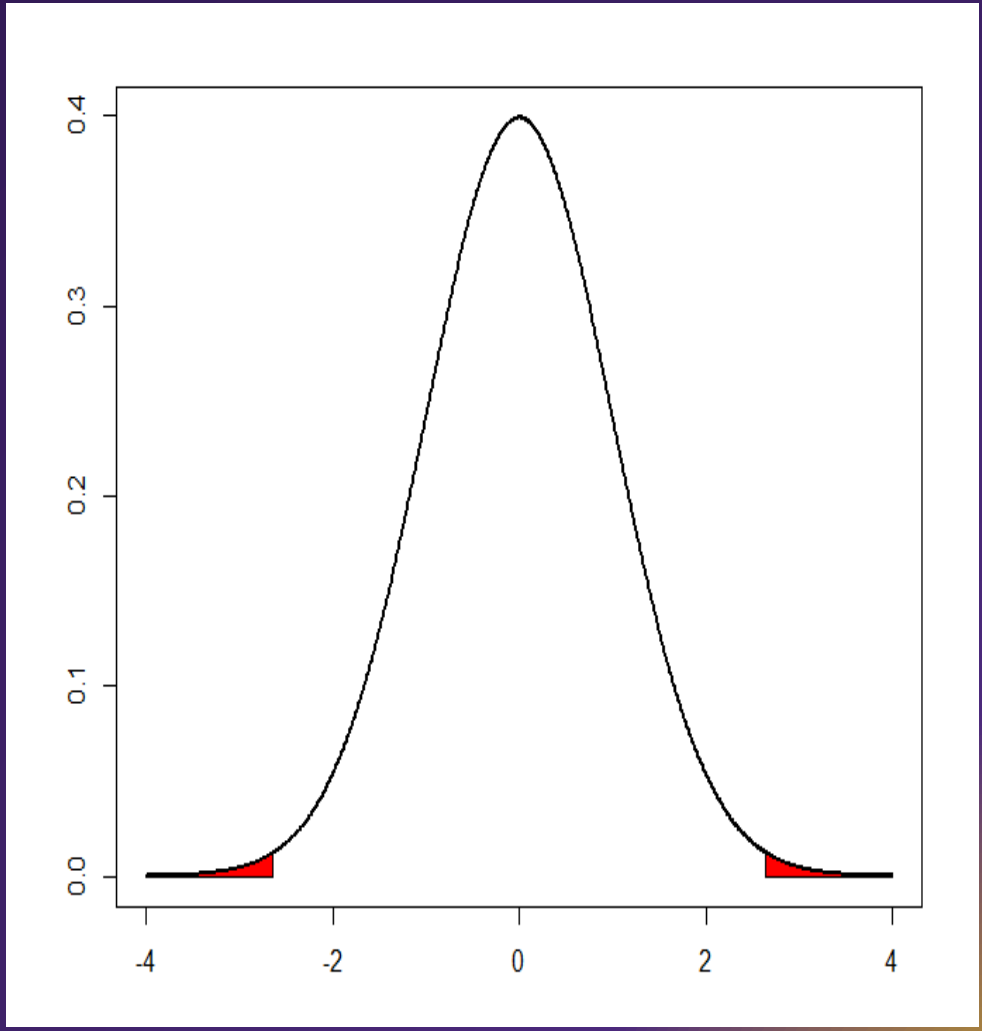
The absolute value difference in the means = 6.4;

Standard deviation of the difference (also known as the standard error of the mean) = 2.4

Noise/signal = 2.6

p-value = 0.009

i.e., the probability of finding a difference equal to or larger than 6.4 in either direction, given no difference is 0.009



The famous 0.05 criteria

- Traditionally, the criteria for a **significant** difference is p<=0.05. This criteria is known as the “Type 1” error.
- Since the p-value for the difference between means is <=0.05, we **reject** the null hypothesis of no difference.
- We conclude that the difference between the 2 means is **significant** at p=0.009

What does this result mean?

- That, **on average**, those with a coronal deformity will have an MFA 6.4 less than those with neutral alignment
- It does not mean that every person with a coronal deformity will have a lower MFA than those with neutral alignment
- Statistical significance does not necessarily imply biological significance—e.g., is 6.4 a meaningful difference?
- In our example, the difference in the direction opposite of what we would expect

95% confidence intervals

- The interval which contains the true mean with probability 0.95
- 95% CI for the mean difference of 6.4 is (1.6, 11.2)
- If the mean difference is significant at p=0.05 , the 95% CI will not include 0.

One-tailed vs. two tailed tests

There are times when we are only interested in differences in one direction.

H0: no difference in means

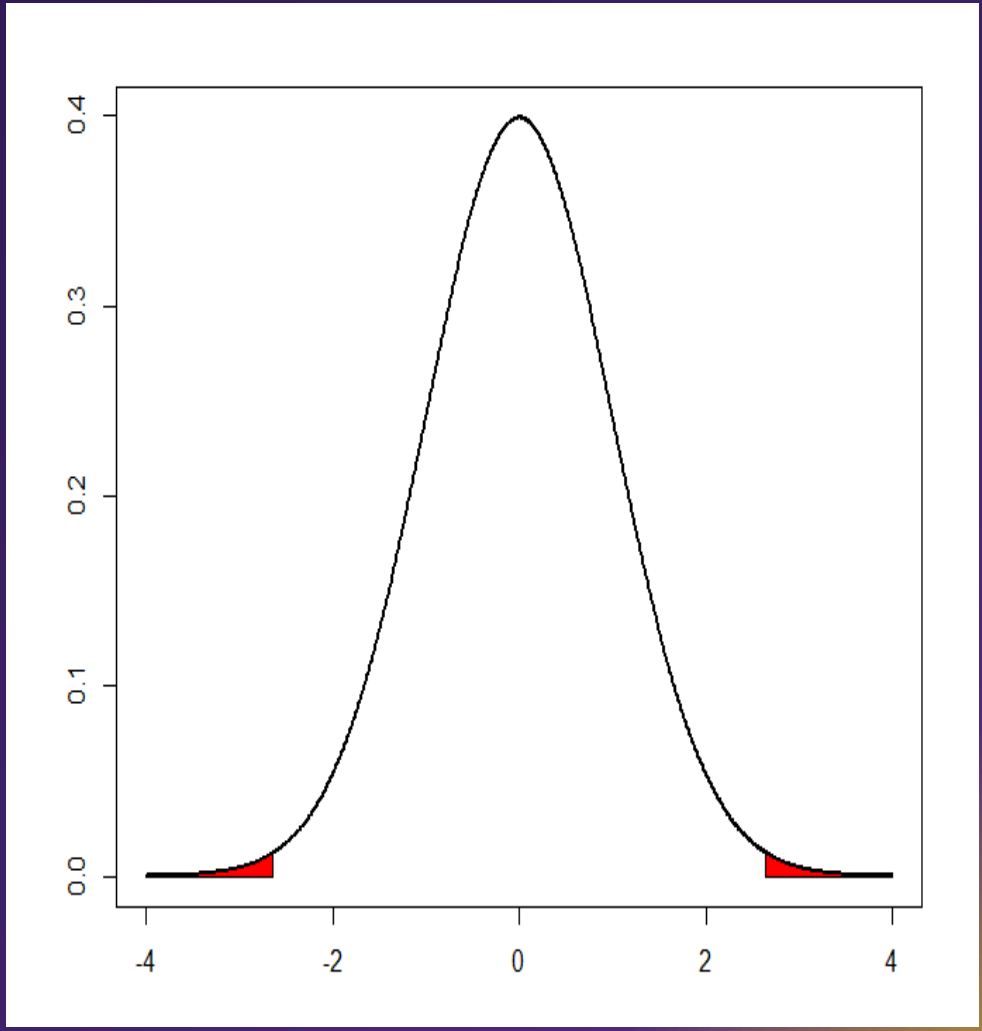
H1: mean for MFA worse (higher) for the tilted group

Prob(difference>=1 if there is no difference)

1-tailed test: p=0.16

2-tailed test: p=0.32

One-tailed tests mostly produce lower (more significant) p-values then 2 tailed tests.



One-tailed vs. two tailed tests

There are times when we are only interested in differences in one direction.

H0: no difference in means

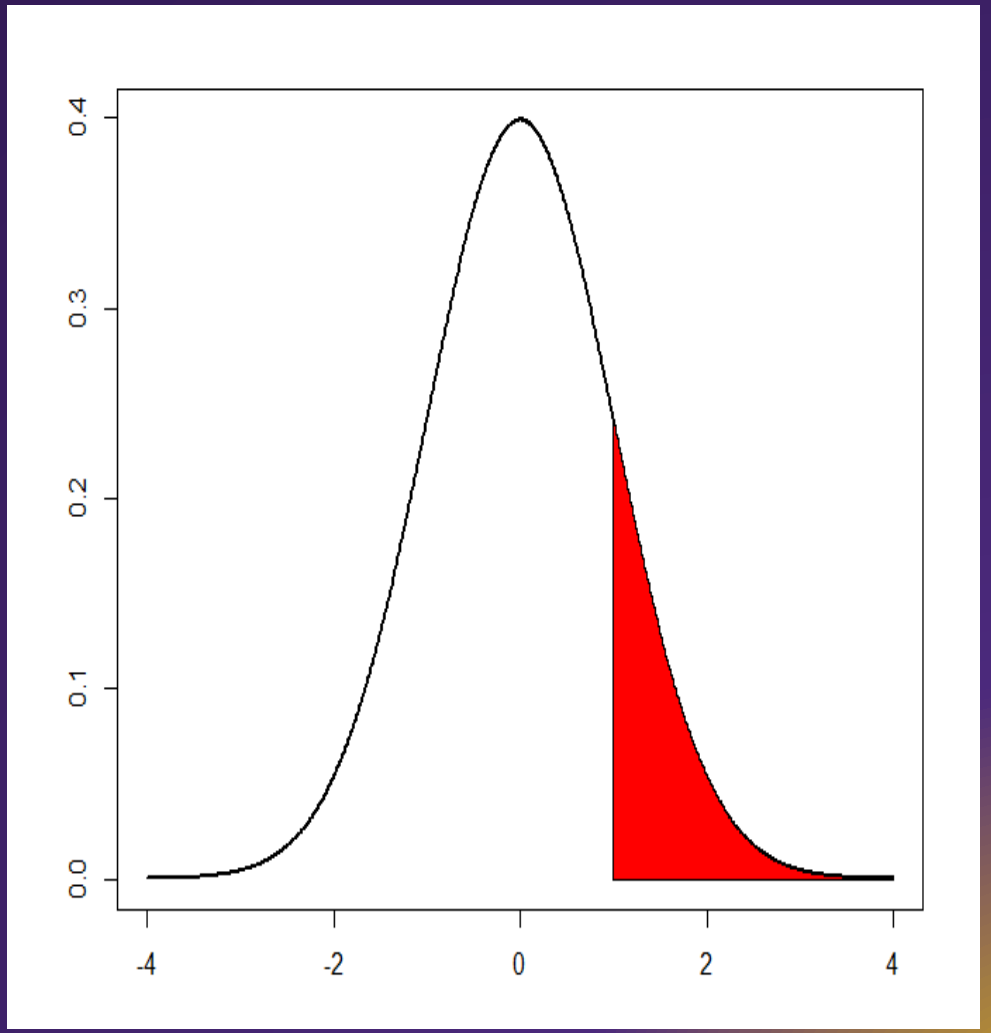
H1: mean for MFA worse (higher) for the tilted group

Prob(difference>=1 if there is no difference)

1-tailed test: p=0.16

2-tailed test: p=0.32

One-tailed tests mostly produce lower (more significant) p-values then 2 tailed tests.



One-tailed vs. two tailed tests

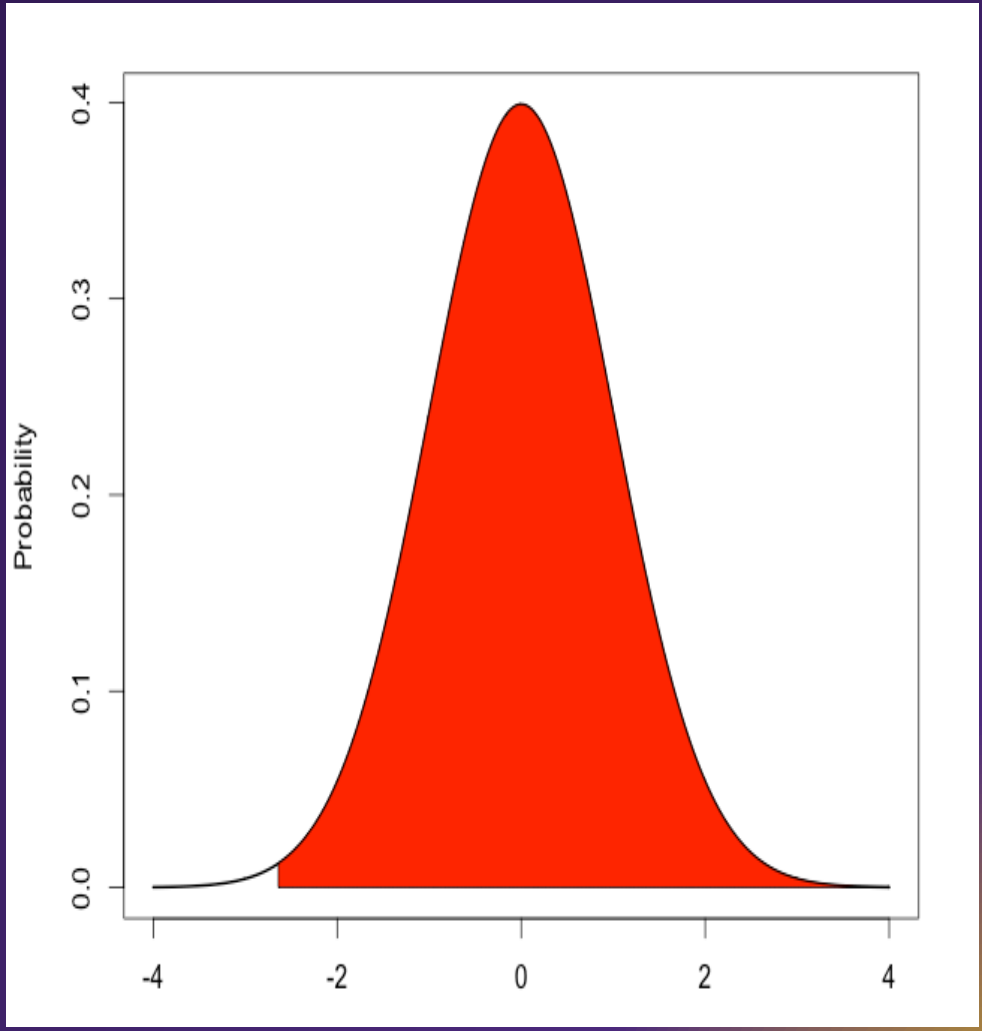
Suppose in our example we are only interested if MFA for a coronal deformity is worse (higher) than the MFA for neutral alignment.

In this case whether tilted minus neutral is higher than a certain value.

Here we have tilted minus neutral = -6.4, standardized difference -2.6

Prob(MFA difference >-2.6) =0.99

We accept the null hypothesis of no differences in mean MFA.

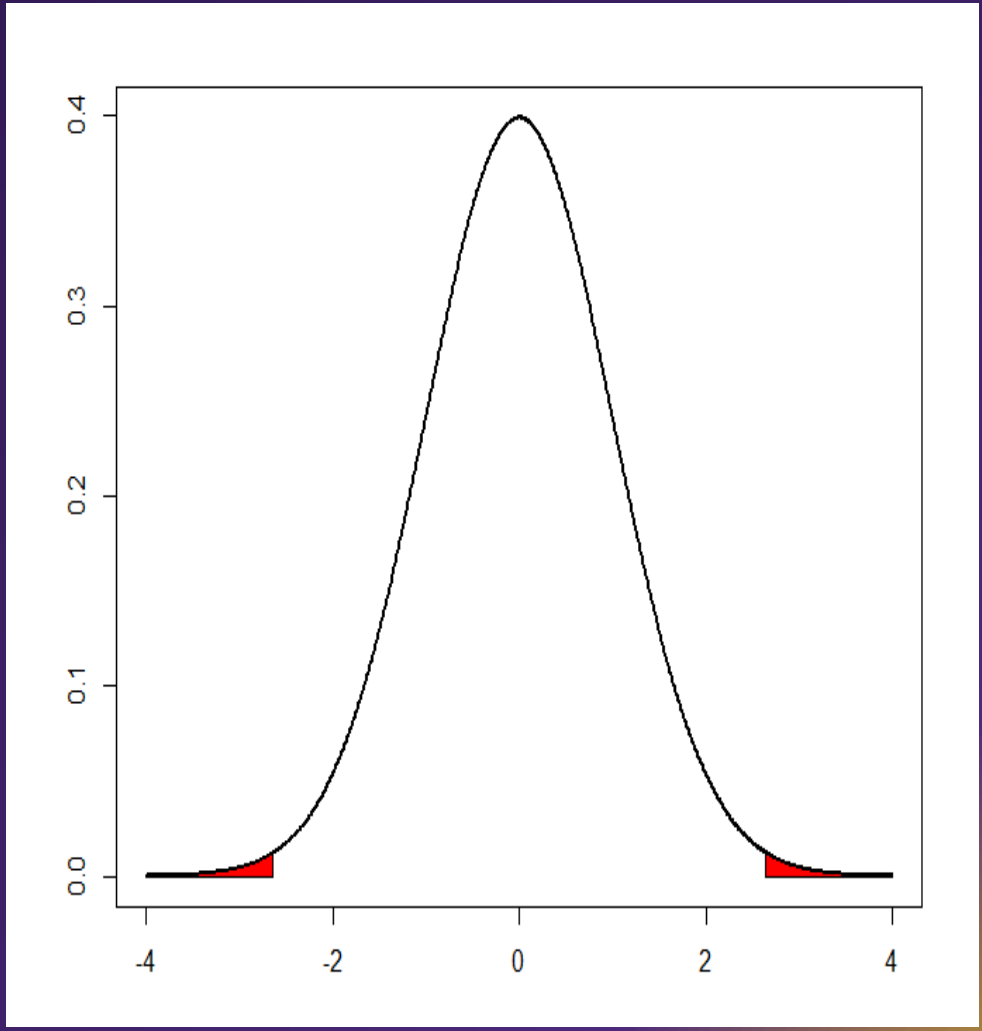


One-tailed vs. two tailed tests

This is a DIFFERENT finding from the 2-tailed test where we rejected H0 for no differences.

We would miss the fact that the difference could be in another direction, which may or may not have biomechanical implications.

Moral: (almost) never do a 1-sided test



Errors in Statistics

Type 1 error = the probability of rejecting the null hypothesis when the null hypothesis is true.

Prob(choose H1|H0 is true)

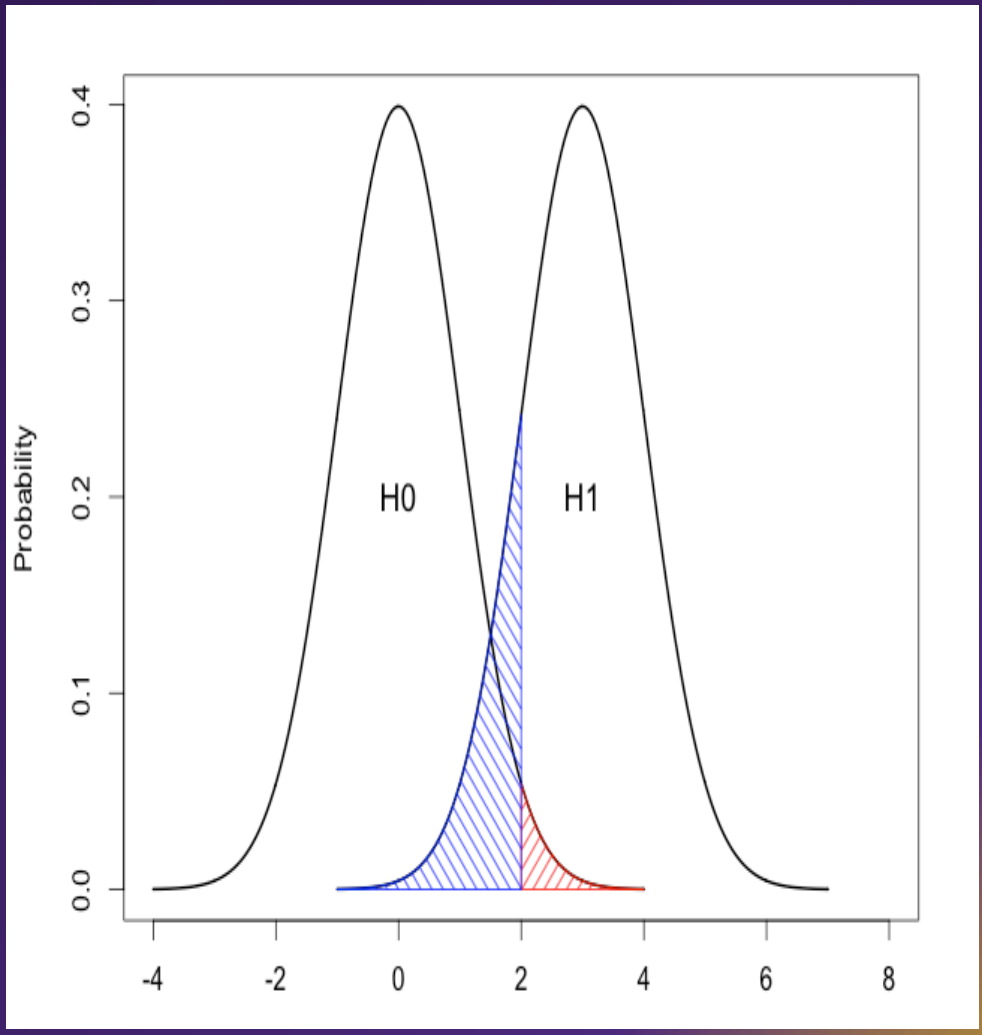
Type 2 error = the probability of accepting the null hypothesis when the alternative is true

Prob(choose H0|H1 is true)

Suppose we define our hypothesis test that any standardized difference in our means greater than 2 we reject H0 in favor of H1, and any difference less than 2 we accept H0.

Type 1 error is in red = 0.05

Type 2 error is in blue = 0.24

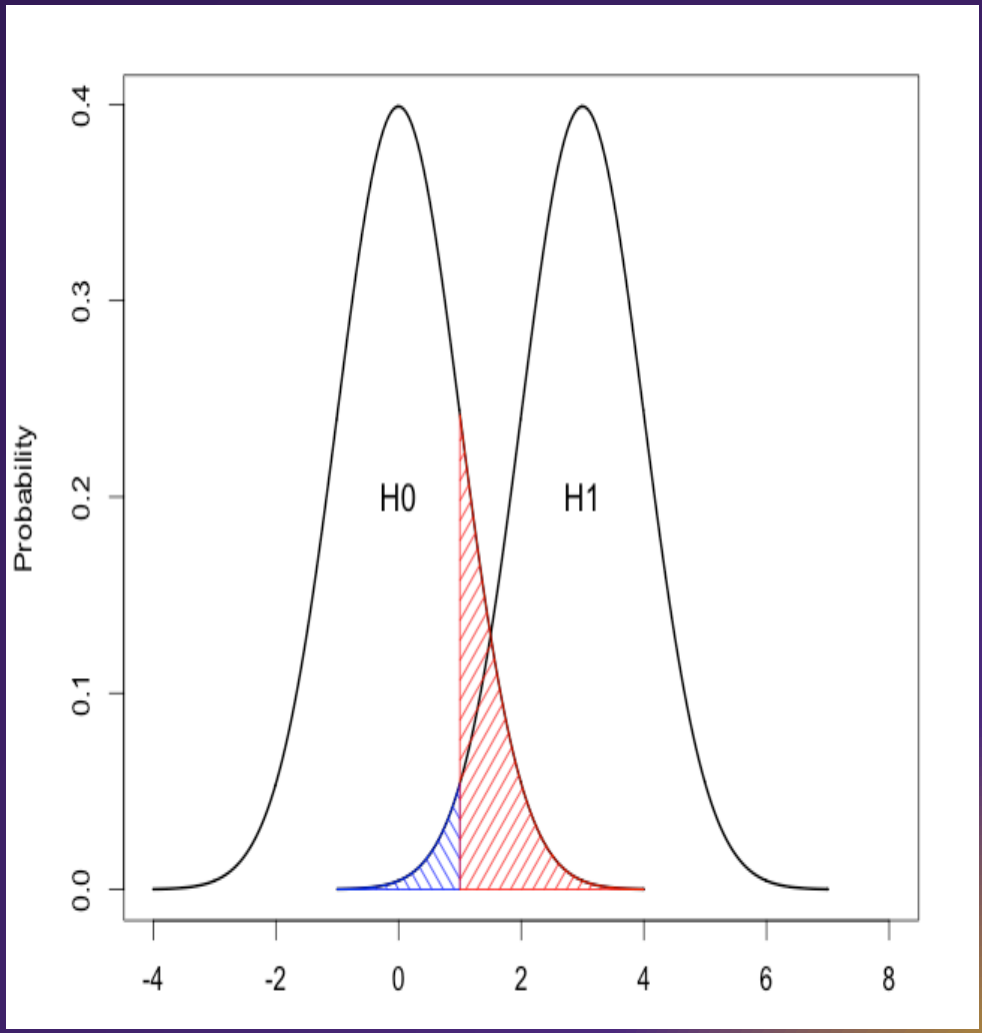


Errors in Statistics

Type 1 error is in red

Type 2 error is in blue

Given these distributions for H0 and H1, a decrease in type 2 error results in an increase in type 1 error.



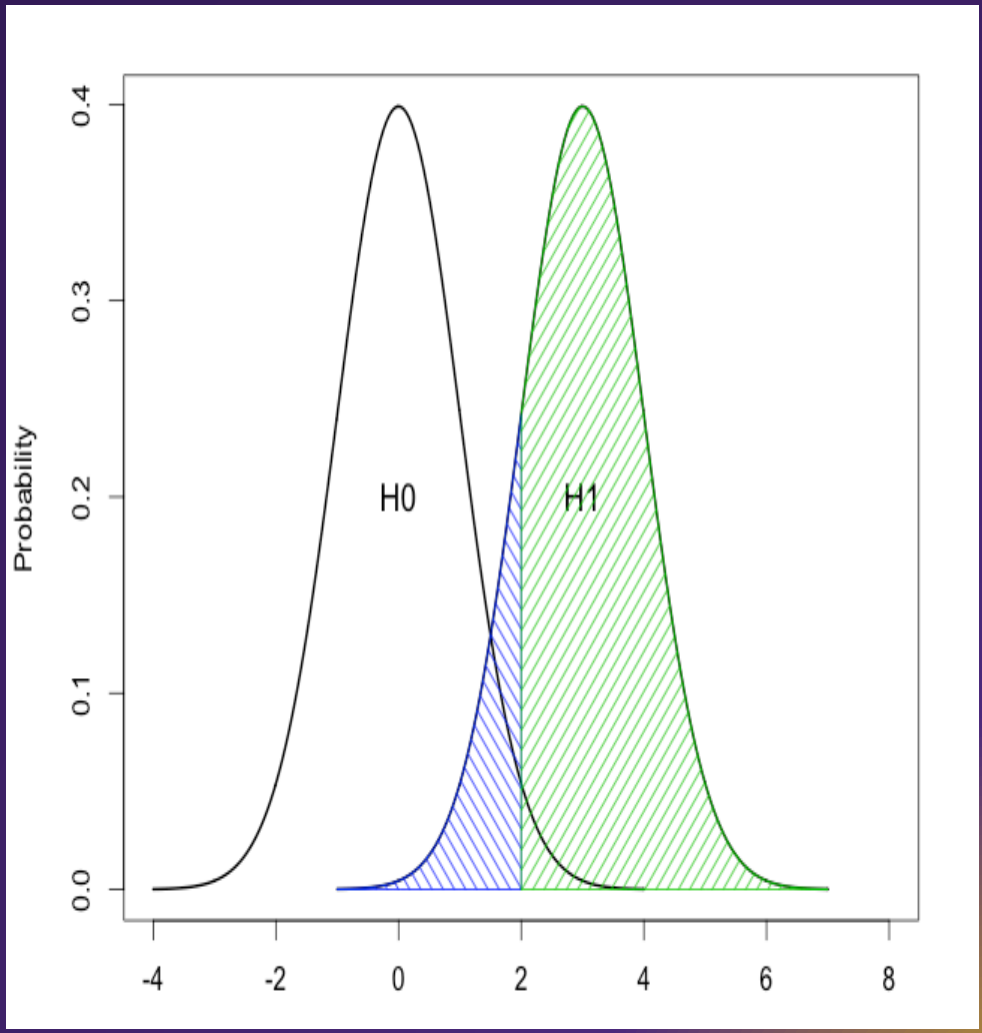
Power

Power = 1- type 2 error

= the Prob(choose H1|H1 is true)

Power is used when designing studies.

***You want to make sure that you have adequate power to detect differences of biological/ clinical interest



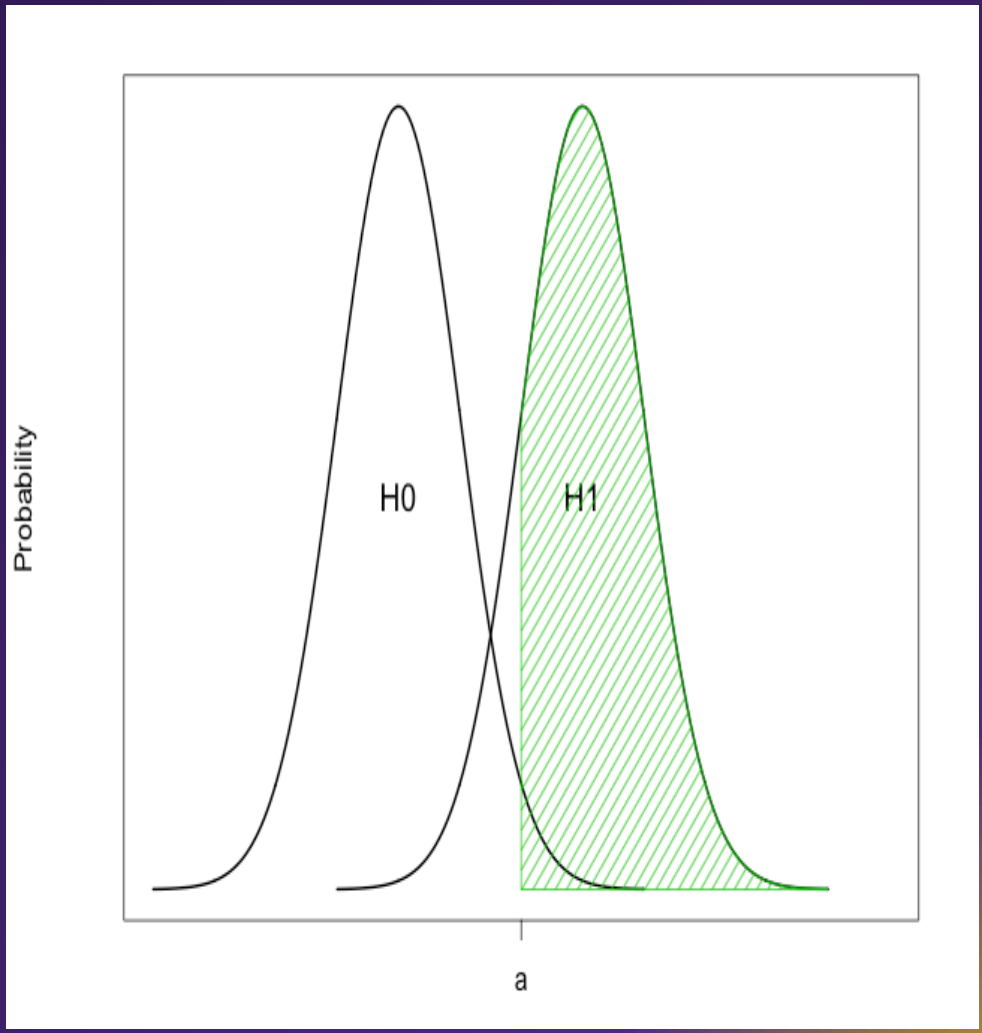
Power

For example, suppose prior research has established that a meaningful difference in standardized mean MFA is "a" points.

Power = the shaded portion of the plot

Power in this example = 0.76

To increase power, increase sample size

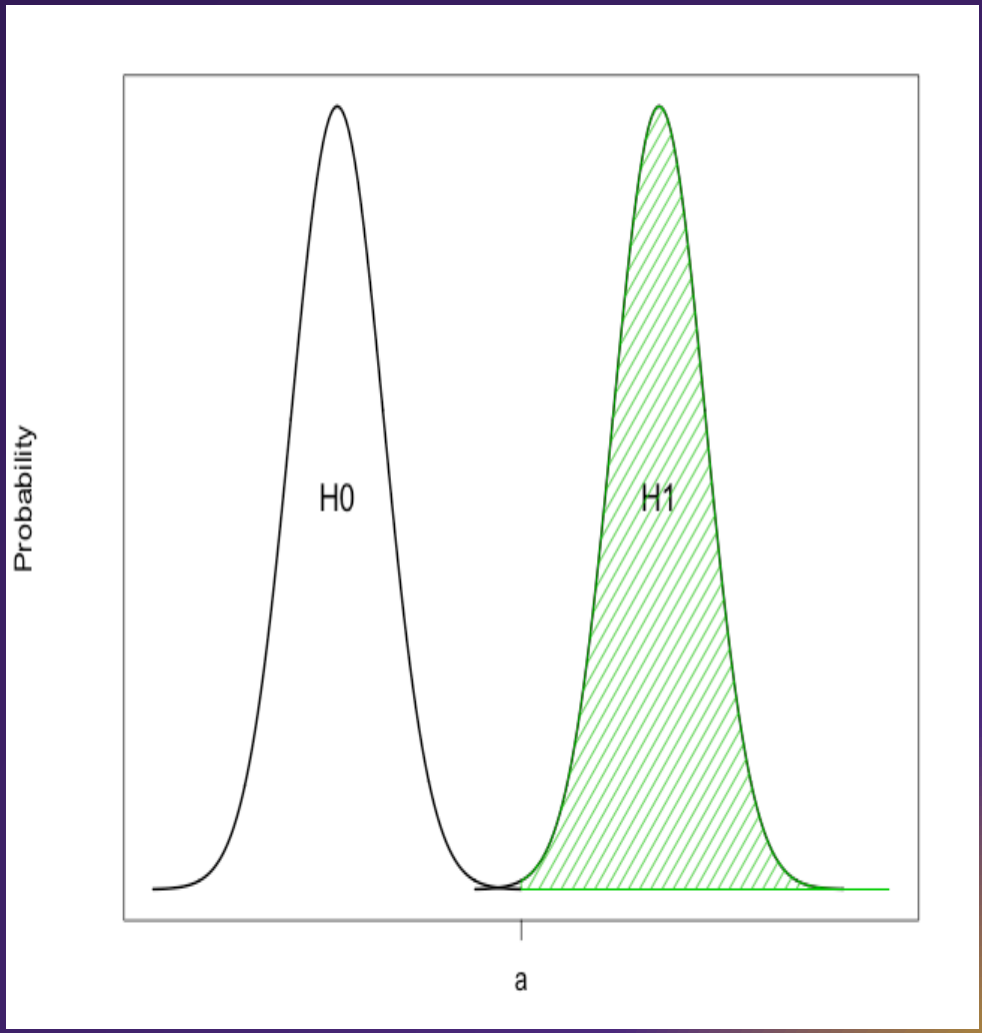


Power

Based on the “law of large numbers”, increasing your sample size increases the precision of the estimate of the sample mean

→ “skinnier” distributions for the mean under H0 and H1

→ increase in power



Power

- Only to be used when DESIGNING studies

- **Do not** carry out post-hoc power analyses.

Confounding

- How do we know that the difference between groups is not due to some other characteristic, e.g., age? (i.e., if the mean age of the tilted group was older, then the decreased functioning could be due to age)
- This phenomenon is called confounding
- Can adjust for confounding using statistics. In this example we use linear regression

Linear regression

- The 2-sample t-test described above can be generalized to a linear regression
- Model: Y = b0 + b1*X + error
- In our example: MFA = b0 + b1*group (group=1 if tilted; =0 if not) = 40.1 – 6.4*group

Linear regression

- To adjust for confounding due to age, just add age to the Model:
- MFA = b0 + b1*group + b2*age
- 53.1 – 6.2*group - 0.2*age
- Assumptions:
  - Linear association between age and MFA
  - The difference in means by coronal deformity is consistent across age

Linear regression

- 53.1 – 6.2*group - 0.2*age
- Adjusted for age, the difference between group means is 6.2.
- This difference is associated with a p-value of 0.011
- Still statistically significant
- Age not a confounder.

“Normalization”

- Control confounding by dividing the outcome by the confounder
- Results in a “ratio” variable (outcome/ confounder)

“Normalization”

- Hypothesis testing using ratio variables can be problematic
  - Can result in extreme values
  - Hypothesis test results can be dependent on the units of the divisor
  - Assumes a linear relationship between numerator and denominator which may not be true

Back to our example

All subjects in our study received surgery for ankle arthritis

They were followed at 6 mo, 1, 2 and 3 years

Main questions:

1. Did subjects improve after surgery?
2. Did improvement differ by coronal deformity?

Back to our example

Specific hypotheses

1. Did subjects improve from preop to 1, 2 or 3 years?
2. Did subjects improve from 1 to 2, or 2 to 3 years?
3. Did subjects without a coronal deformity improve from preop to 1, 2 or 3 years?
4. Did subjects without a coronal deformity improve from 1 to 2, or 2 to 3 years?
5. Did subjects with a coronal deformity improve from preop to 1, 2 or 3 years?
6. Did subjects with a coronal deformity improve from 1 to 2, or 2 to 3 years?
7. Did improvement from pre-op to 1 year differ by coronal deformity
8. Did improvement from pre-op to 2 year differ by coronal deformity
9. Did improvement from pre-op to 3 year differ by coronal deformity

Etc.

Constructing hypotheses in complicated study designs

Problem with this approach– too many pair-wise comparisons!!

The greater the number of tests, the more likely you will reject H0 when H0 is true—i.e. increase in type 1 error

Carrying out separate tests does not take advantage of the full data.

Constructing hypotheses in complicated study designs

Better approach– carry out 2 models

To address

1. Did subjects improve after surgery?

a. Combine data across all time-points

b. Construct a model to test if, MFA changes over time — the "omnibus test"

c. If significant, then you can carry out pair-wise comparisons (still need to correct for the increase in type 1 error due to carrying out multiple comparisons)

Constructing hypotheses in complicated study designs

2nd model

To address

2. Did improvement differ by coronal deformity?

a. Combine data across all time-points

b. Carry out omnibus test for whether the pattern of change in MFA across study time differs by coronal deformity—in a regression context this is known as a time by deformity interaction term

c. If omnibus test significant, then do pair-wise comparisons of interest.

Independent vs. dependent data

Independent: occurrence one observation has no bearing on any other observations in a set of data

In our example, one subject's MFA at baseline has no influence on another subject's MFA at baseline—the MFA data at baseline are independent

Many of the standard statistical models (e.g. linear regression) assume independent data

Independent vs. dependent data

Dependent: occurrence of one observation could potentially influence another observation in a dataset differentially from other observations in that dataset

In our example, one subject's MFA at baseline may be related to that subject's MFA at any follow-up, but will not influence the MFA of any other subject. Here the MFA data are considered dependent.

Another way of describing this data: repeated measures

Examples of repeated measure data

- Repeated measures across time (our MFA example)
- Data on two feet per person (e.g., measuring pressure under the foot when walking for a sample of subjects
- Data on multiple sites within a foot (comparing pressure under the heel vs. pressure at each of the 5 metatarsals) for a sample of subjects
- Multiple trials per subject—looking at speed of walking in amputees comparing different prosthetics
- Comparison of multiple procedures carried out on a single specimen—e.g. simulating different surgeries to correct foot deformities in a sample of foot cadavers—first you simulate the deformity, then you simulate the correction

Why repeated measures?

- Often you are more interested in within subject differences. E.g., you may be more interested if ankle OA surgery improves your walking speed, as opposed to the surgery improving walking speed for a population
- Within subject differences are usually measured with more accuracy than between subject differences

Why repeated measures?

Caveat:

It does you no good to have lots of repeated measures per subject/specimen, if you only have a few subjects. (e.g., 1000 repeated measures on 3 subjects)

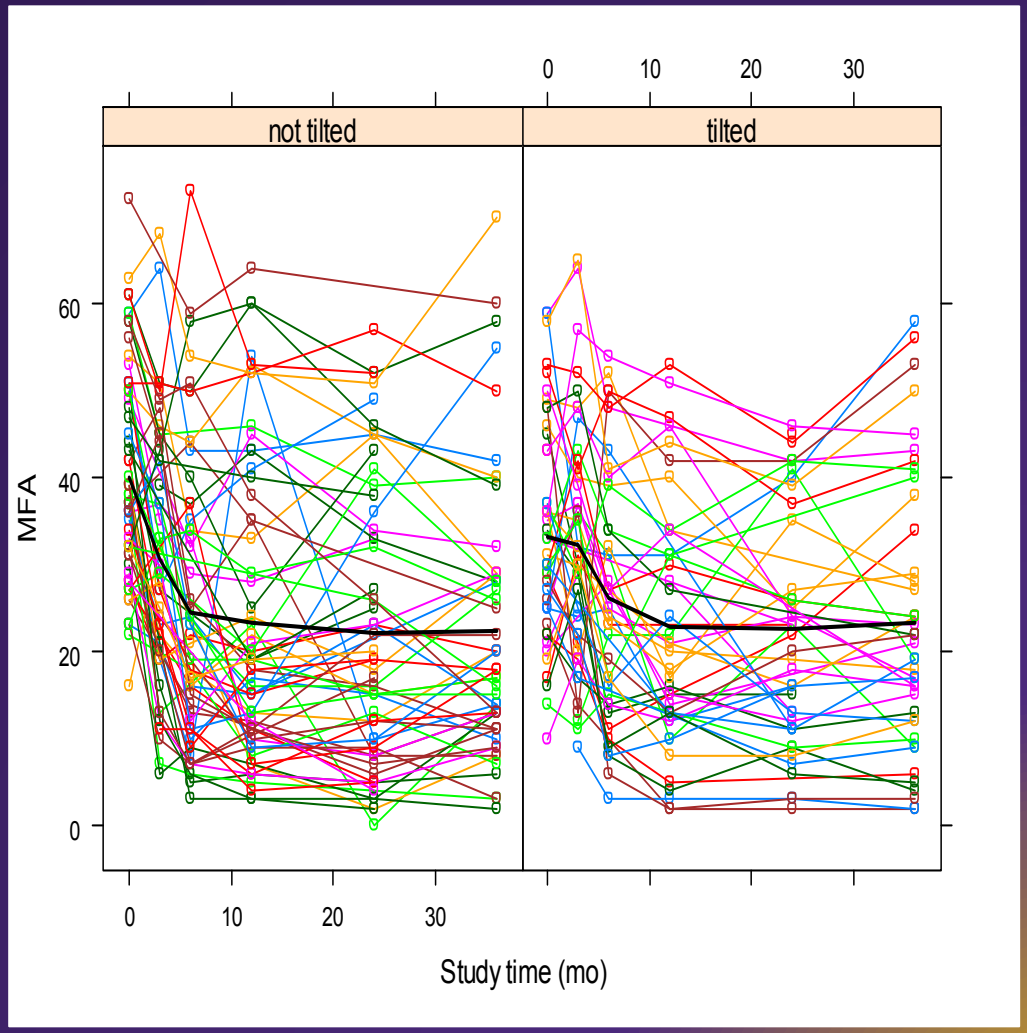
Very problematic in generalizing your results to a larger population

How to analyze repeated measures?

Spaghetti plots

For our example, each colored line represents an individual patient MFA trajectory

The thick black line represent the average MFA at each visit



How to analyze repeated measures data?

Hypothesis testing: Linear mixed effects regression

- Separate out error into between and within subject.
- Observations within subject are considered "independent" of other observations within subject
- Observations between subjects are considered "independent" of other observations between subjects

How to analyze repeated measures data?

Hypothesis testing: Linear mixed effects regression

Linear regression: Y = b0 + b1*X + error

Linear mixed effects regression:

Y = b0 + b1*X + error(between) + error (within)

b0 and b1=fixed effects

error(b) and error(w)= random effects

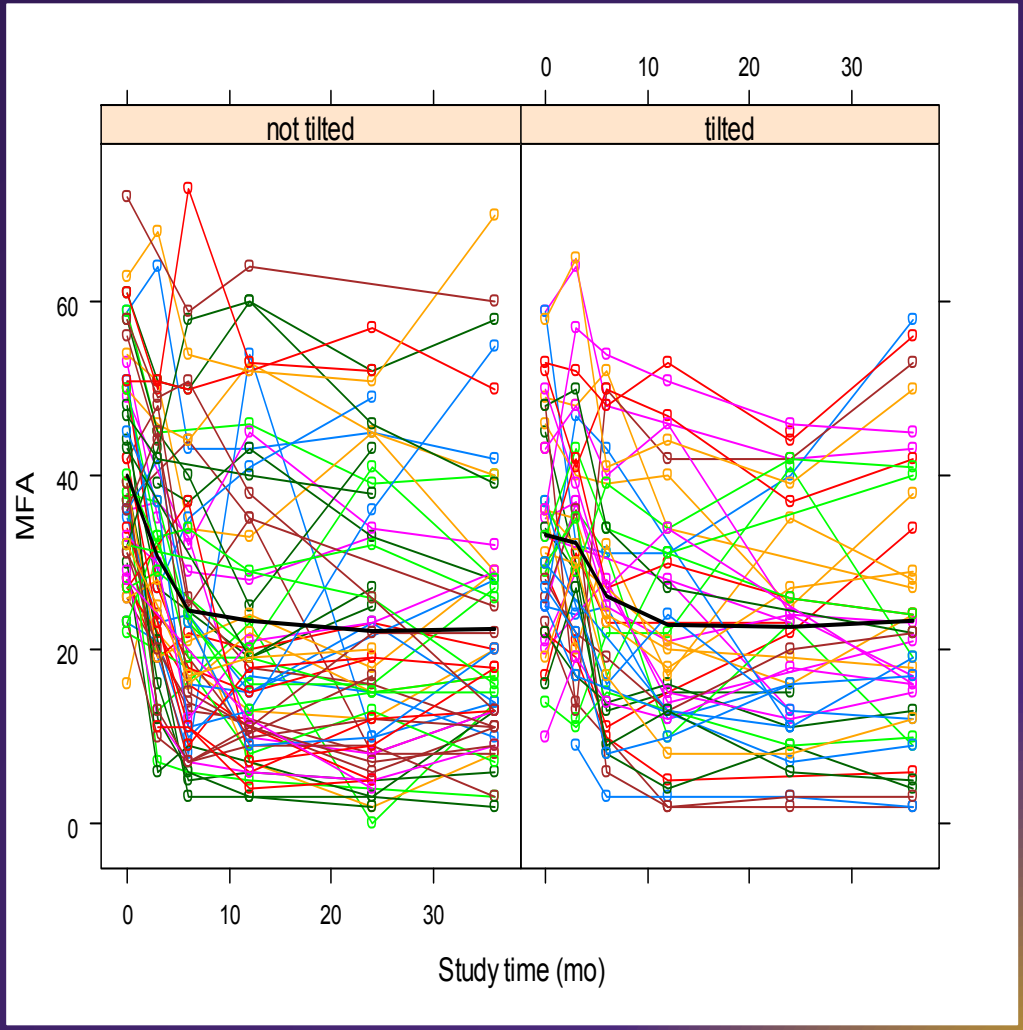
Back to our example

Omnibus test for change in MFA across study time: p<0.0001

MFA = 37.1 – 11.9*m6 – 14.1*y1 – 14.8*y2 – 14.3*y3

Between subject SD: 12.4

Within subject SD: 8.0



Back to our example

Omnibus test for differences in the pattern of change by coronal deformity: p= 0.0008

Neutral group

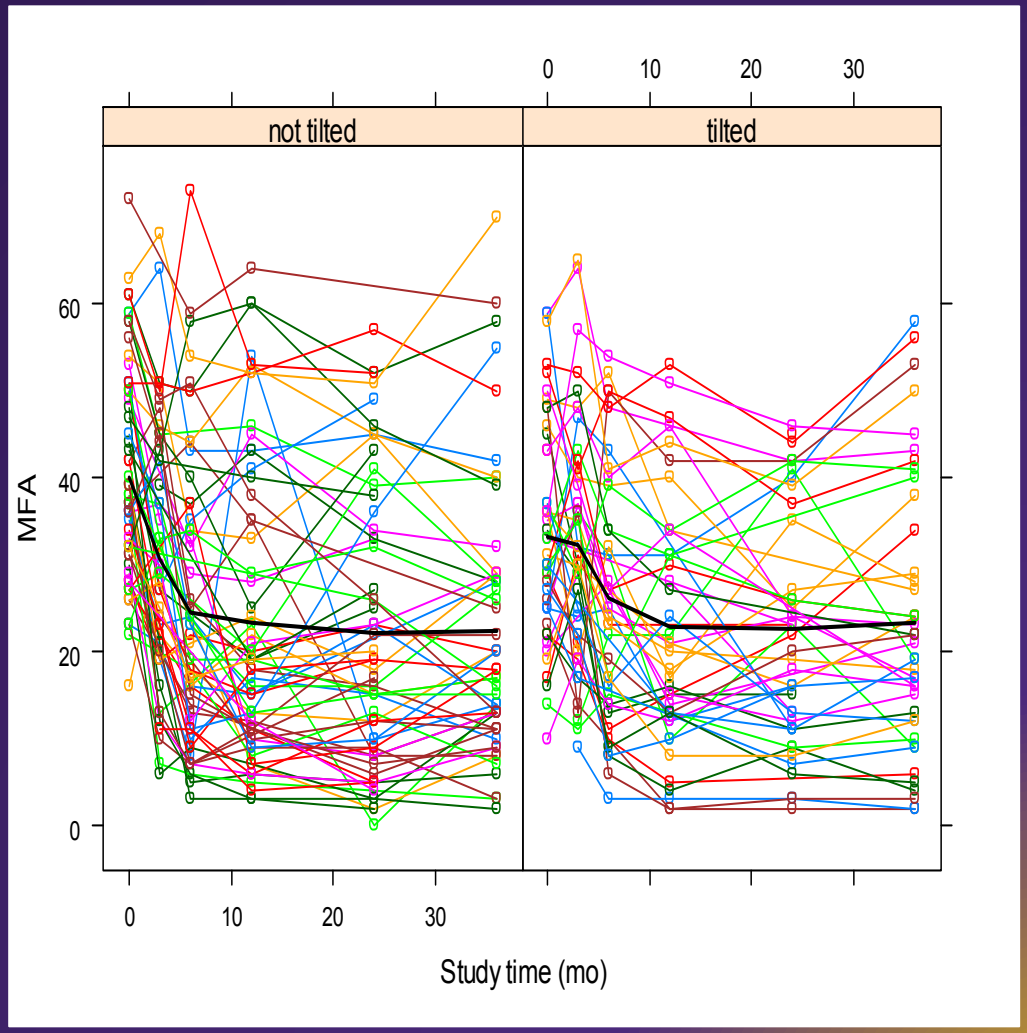
MFA = 40.0 – 15.5*m6 – 16.7*y1 – 17.8*y2 – 17.5*y3

Tilted group

MFA = 33.2 – 7.1*m6 – 10.5*y1 – 10.6*y2 – 9.9*y3

Between subject SD: 12.7

Within subject SD: 7.6



LME vs. Repeated Measures ANOVA

Advantages of LME

- Can use with missing data
- Can use with unequal subjects per group or unequal number of repeated measures per subject
- Can use with time dependent covariates

Summary of basic concepts

- Goal of statistics: separating signal from noise
- Visualize your data with appropriate graphics
- Two types of errors: type 1 and type 2. Control type 1 error in hypothesis testing. Control type 2 error with good study design.
- Statistical significance does not necessarily imply biological or clinical relevance
- Consider confounding in your analysis
- Distinguish independent and dependent data