Department of Biology
Penn
University of Pennsylvania

ITMAT
Bioinformatics

Introduction to Bioinformatics

Topic 15
Non-Parametric Methods

Fall, 2023

Professor
Gregory R. Grant

Teaching Assistants
Chetan Vadali

Gregory R. Grant
Genetics Department
ggrant@pennmedicine.upenn.edu

ITMAT Bioinformatics Laboratory
University of Pennsylvania

Hypothesis Testing

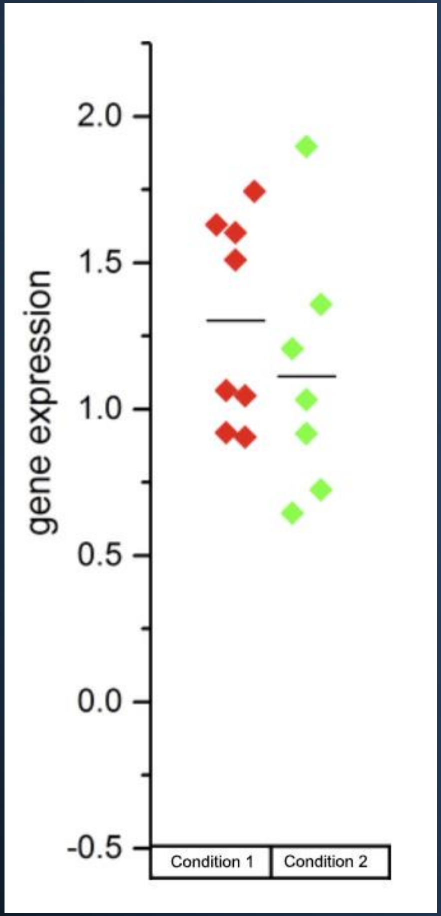
- Consider the most basic statistical problem.

- Two experimental conditions and one quantity of interest.

- For example:
  - WT versus Mutant (the two conditions)
  - Expression of Gene X (the one quantity of interest)

- Question: Is the (unknown) mean of the measurement in condition one different from the (unknown) mean of the measurement in condition two.

- We want a way to answer this with control of the false-positive and false-negative rates.



Significance Testing

- People have literally thought up hundreds of ways to do this.

- Solutions have all kinds of advantages and disadvantages.

- One general approach is to calculate

  $$P(\text{data} \mid H_0)$$

  – H_0 is the null hypothesis that the means are equal.

  – Gives a p-value, which we can use to decide if believing whether H_0 is true or not.

Assumptions

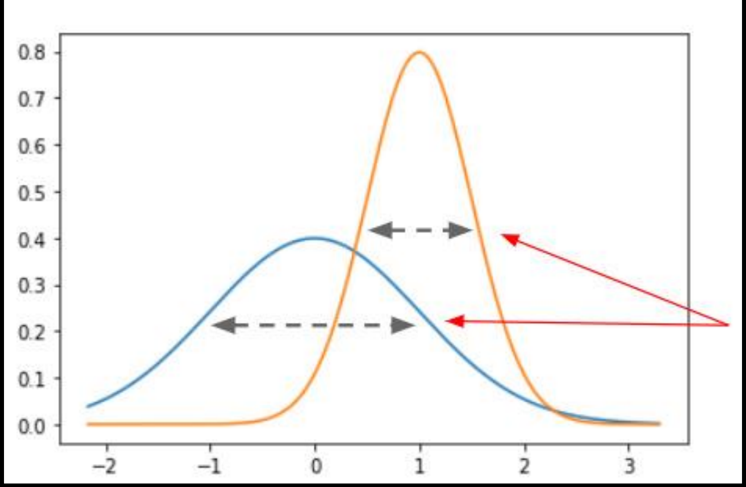
- One of the major disadvantages of many methods is that they make simplifying assumptions about the data.

- For example, the *T*-test assumes:
  - Individual observations are drawn from normal distributions.
  - If the means are different between the two conditions, then the means are the only thing that's different, not the variances.

- These are powerfully strong assumptions that are rarely true.

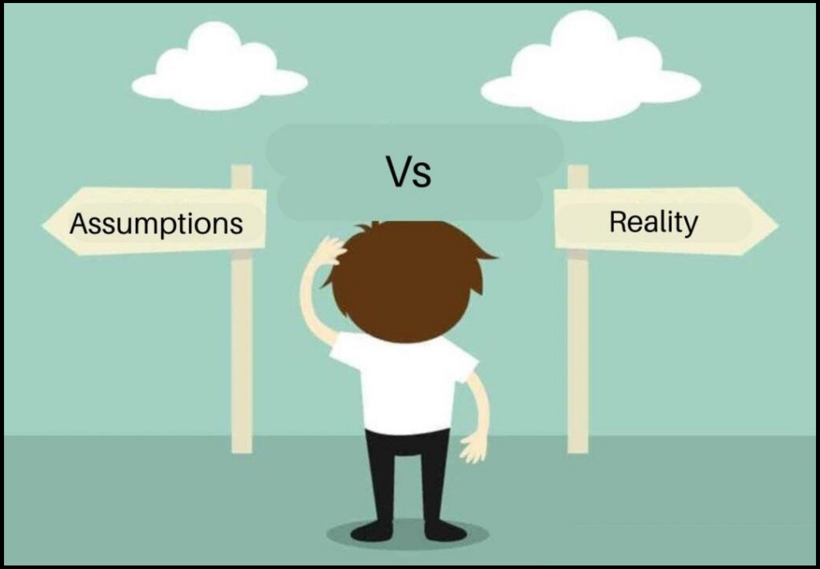
Not all assumptions are the same.

- Sometimes a wrong assumption will ***not*** cause problems.
- By "wrong" here we mean anti-conservative.
  - In other words, Type I error is *higher* than we think it is (more false positives).
- For example, in a *T*-test you can safely ignore the normality assumption *if you have enough replicates.*
  - 8-10 replicates per condition is usually enough to overcome the normality assumption.
  - The central limit theorem says so.
- While you cannot safely assume equal variance in both groups when doing a *T*-test.
  - No matter how many replicates you have.



Unequal means and variances.

If assumptions drive us wrong, why make them?



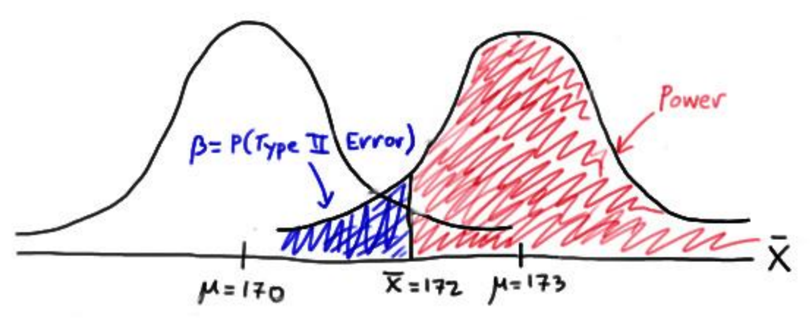
Two main considerations:

1. We have no (known) alternatives.
2. There are alternatives, but they cost us so much power that they're not worth it.

This is called being "fast and loose," an unfortunate reality of significance testing in biology.

Sometimes you simply must shoot from the hip. But that's still better than shooting blind.

Power



- A test with lower false-negative rate is called “more powerful”.
- Strictly speaking the power is one minus the false-negative error rate.
  - A test with low false-negative rate has high power, and conversely.

Type I vs Type II Error

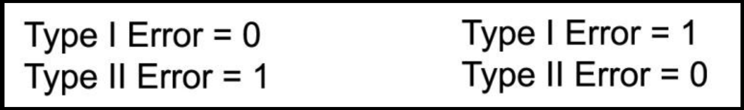
- Type I = False Positive Rate
- Type II = False Negative Rate

- You can always get the Type I error rate down to zero.
  - Simply ***never*** reject the null hypothesis, and you can never be wrong.
  - But this strategy has increased the Type II error rate to 1.0.

- Similarly, you can always get the Type II error rate down to zero.
  - Simply ***always*** reject the null hypothesis, and you can never miss anything.
  - But this strategy has increased the Type I error rate to 1.0.

Trade Offs

- The previous slide represents two extremes.



Type I Error = 0
Type II Error = 1

Type I Error = 1
Type II Error = 0

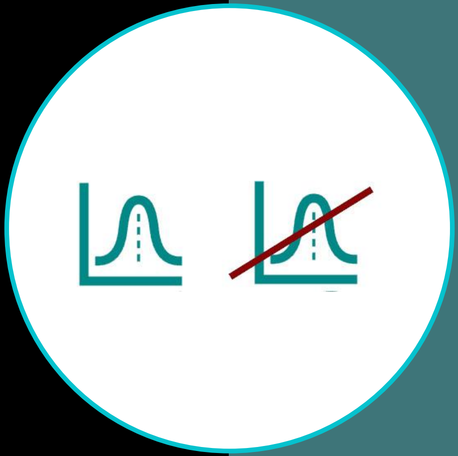
- But everywhere in between, Type I error is always in tension with Type II.
- Whenever you lower one, you raise the other.
- *Unless you can increase the number of replicates indefinitely. Then you can decrease both at the same time.*

Parametric Methods

- A *T*-test assumes normal distributions.

- This is known as a “parametric” assumption.
  - Whenever we assume we know the form of a distribution, we’re in “parametric” territory.

- It is possible to derive tests that do not make such assumptions.
- But without any such assumptions, the data need to speak more for itself.
- And that invariably costs us power.

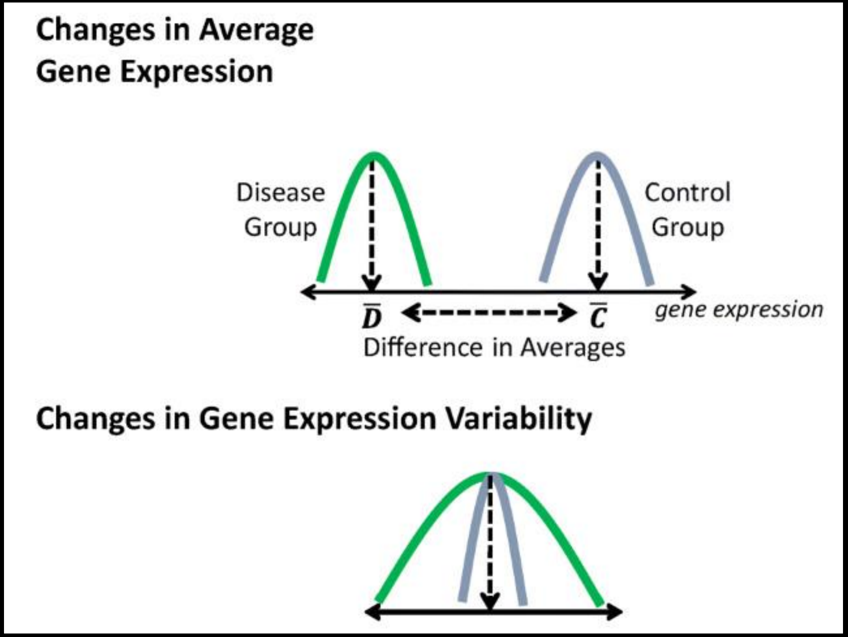


The Null Hypothesis

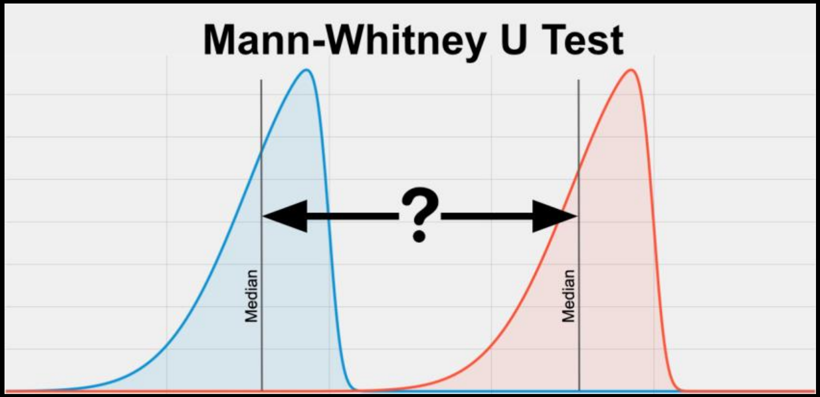
- There are two ways to formulate the null hypothesis for differential effects.

- **First way:**
  - H_0: The *means* in the two conditions are equal.
- **Second way:**
  - H_0: The *distributions* in the two conditions are equal.

- It's the first formulation that we care about.
  - Yet non-parametric methods usually only test for the later.
  - The math makes us do this.



The Mann-Whitney Test



- This test of makes no assumptions about distributions. They can be any shape.
- It ultimately tests for difference in *distributions* not just means.
- Distributions can be different in many ways.
- But because of the particulars of the test, most of the time it is a difference of means driving significant Mann-Whitney *p*-values.
    - Most of the time, but not always, we will see a counter-example.

The Mann-Whitney Test
- non-parametric two-sample test -

Does not assume anything about the underlying distributions.

Based instead on the following concept:

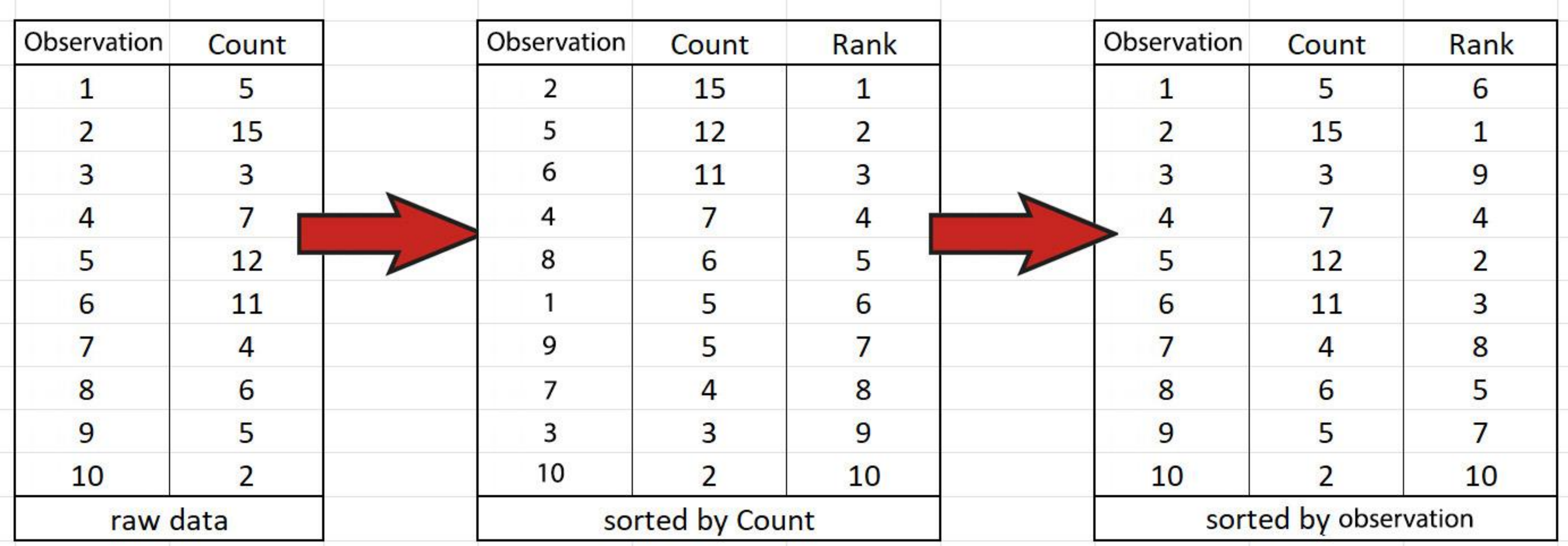
If they're not differential, then if we choose an observation from C_1 at random and another from C_2, then it's a 50/50 chance C_1 is larger.

Replace all data points with their ranks, irrespective of which condition they came from.

Ranks are explained on the next slide

Ranks

- A common approach to avoid parametric assumptions is to work with ranks.
- The counts for each observation are replaced by their ranks.



Observation	Count
1	5
2	15
3	3
4	7
5	12
6	11
7	4
8	6
9	5
10	2

raw data

Observation	Count	Rank
2	15	1
5	12	2
6	11	3
4	7	4
8	6	5
1	5	6
9	5	7
7	4	8
3	3	9
10	2	10

sorted by Count

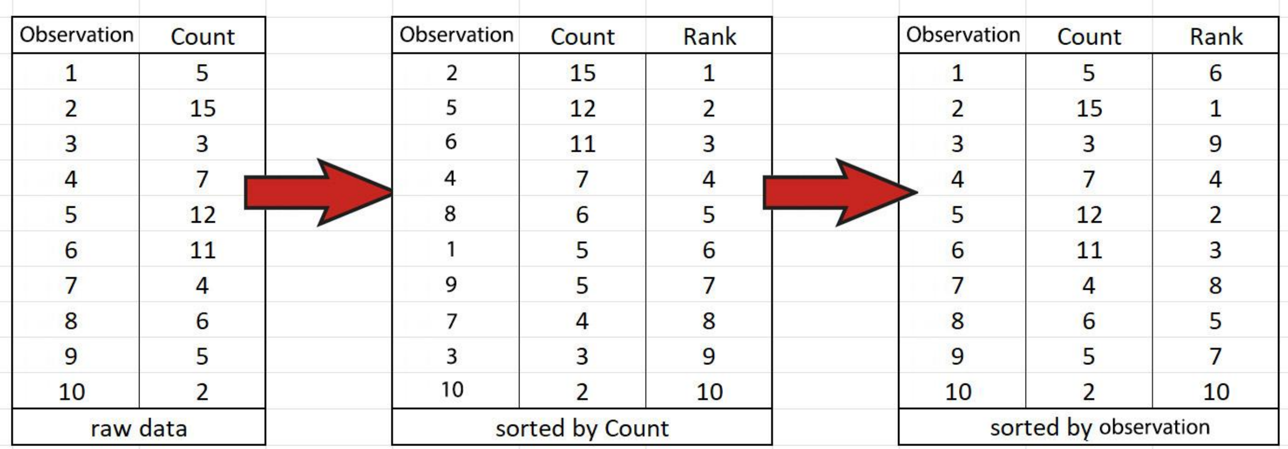
Observation	Count	Rank
1	5	6
2	15	1
3	3	9
4	7	4
5	12	2
6	11	3
7	4	8
8	6	5
9	5	7
10	2	10

sorted by observation

- Statistical analysis then proceeds on the ranks rather than the original counts.
  - Ranks are, after all, numbers, so they can be used to do statistics.
- We'll see this will allows us to calculate *p*-values without making any assumptions about distributions.

Ranks and Outliers

- Ranks tend to be blind to outliers.
- The bottom data has two outliers, but the ranks are the same.



Observation	Count
1	5
2	15
3	3
4	7
5	12
6	11
7	4
8	6
9	5
10	2

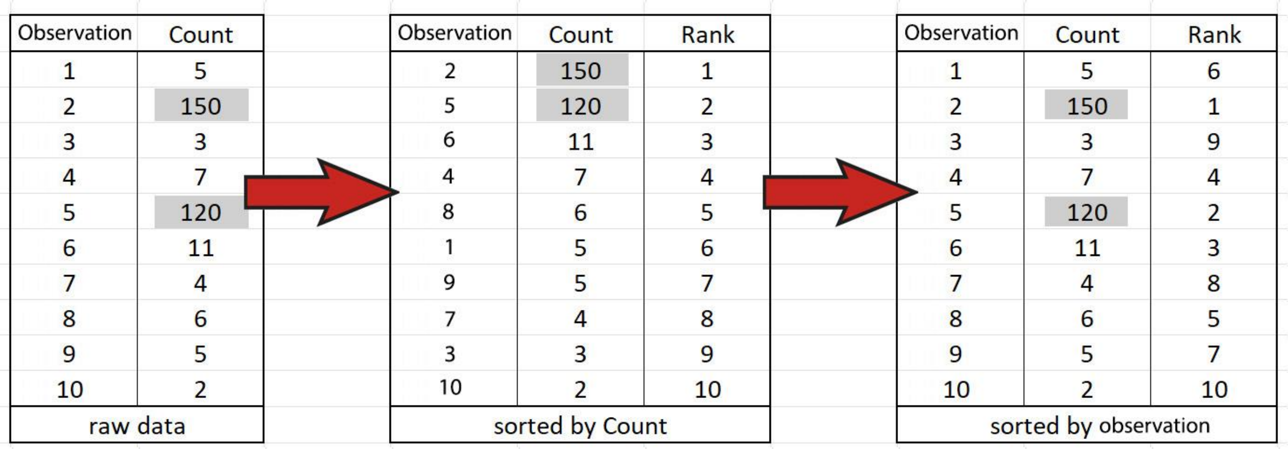
raw data

Observation	Count	Rank
2	15	1
5	12	2
6	11	3
4	7	4
8	6	5
1	5	6
9	5	7
7	4	8
3	3	9
10	2	10

sorted by Count

Observation	Count	Rank
1	5	6
2	15	1
3	3	9
4	7	4
5	12	2
6	11	3
7	4	8
8	6	5
9	5	7
10	2	10

sorted by observation



Observation	Count
1	5
2	150
3	3
4	7
5	120
6	11
7	4
8	6
9	5
10	2

raw data

Observation	Count	Rank
2	150	1
5	120	2
6	11	3
4	7	4
8	6	5
1	5	6
9	5	7
7	4	8
3	3	9
10	2	10

sorted by Count

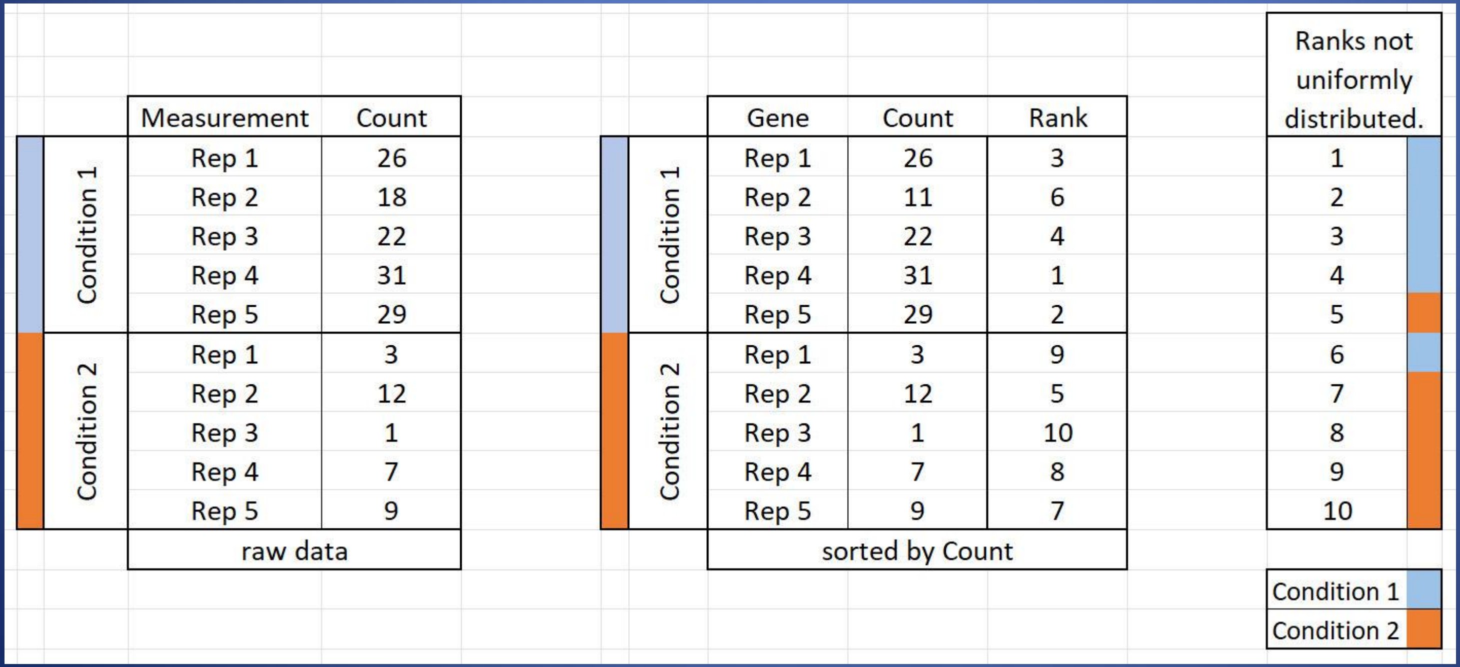
Observation	Count	Rank
1	5	6
2	150	1
3	3	9
4	7	4
5	120	2
6	11	3
7	4	8
8	6	5
9	5	7
10	2	10

sorted by observation

Gene 2 and Gene 5 same ranks as above

Two Groups

- Now suppose we need to test two conditions for difference.
- Rank all values regardless of condition and see how each condition's ranks distribute on the list of all ranks.
  - The example will make it clearer.
- This is the basis of the Mann-Whitney test.



		Measurement	Count
🟦	Condition 1	Rep 1	26
		Rep 2	18
		Rep 3	22
		Rep 4	31
		Rep 5	29
🟧	Condition 2	Rep 1	3
		Rep 2	12
		Rep 3	1
		Rep 4	7
		Rep 5	9
		raw data	

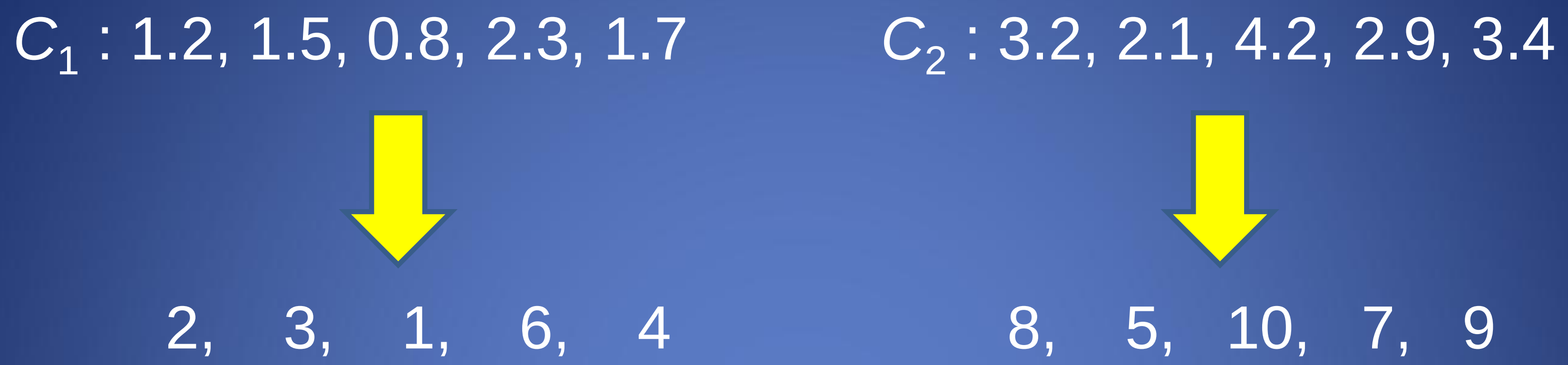
		Gene	Count	Rank
🟦	Condition 1	Rep 1	26	3
		Rep 2	11	6
		Rep 3	22	4
		Rep 4	31	1
		Rep 5	29	2
🟧	Condition 2	Rep 1	3	9
		Rep 2	12	5
		Rep 3	1	10
		Rep 4	7	8
		Rep 5	9	7
		sorted by Count		

Ranks not uniformly distributed.	
1	🟦
2	🟦
3	🟦
4	🟦
5	🟧
6	🟦
7	🟧
8	🟧
9	🟧
10	🟧

Condition 1	🟦
Condition 2	🟧

The blue ones from condition1 bunch up at the top of the ranked list.

Mann-Whitney



C_1 : 1.2, 1.5, 0.8, 2.3, 1.7

2, 3, 1, 6, 4

C_2 : 3.2, 2.1, 4.2, 2.9, 3.4

8, 5, 10, 7, 9

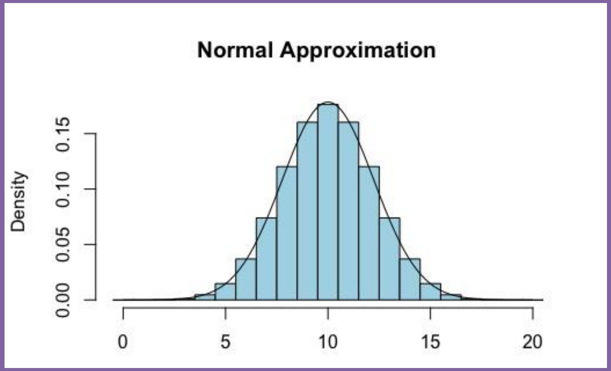
- Sum the ranks from condition C_1

  $R = 2 + 3 + 1 + 6 + 4 = 16$

- It indicates differential effect if *R* is particularly small or particularly large.
  - We're impressed if R=1+2+3+4+5=15 or R=6+7+8+9+10=40 or anything "significantly" close to these extremes.
- The null hypothesis is that the ranks from either condition are like rolling a *fair* 10-sided die.

The Null Hypothesis

- The distribution of *R* is easy to calculate (under the null hypothesis).
- For small numbers of replicates in each condition (up to about 20) computers can list out all possible ways to rank them and calculate *R* by brute force.
- We'll illustrate the concept on the next slide with a 2 replicates versus 2 replicates comparison.
- When there are more than about 20 replicates, the *p*-values can be approximated by a normal.



- To illustrate we'll do the calculation when there are two replicates in each group.
  - If $M = N = 2$ then the ranks can happen in only six possible ways.
  - R is the sum of the ranks in condition C1

C_1	1,2	1,3	1,4	2,3	2,4	3,4
C_2	3,4	2,4	2,3	1,4	1,3	1,2
R	3	4	5	5	6	7

- Each of these six possibilities are equally likely.
- But two of them give R=5, so $R=5$ is twice as likely as the other values..

R	3	4	5	6	7
***P*(*R*)**	1/6	1/6	1/3	1/6	1/6

Probability R=5 is 1/6+1/6 since it can happen in two ways.

- Then use the probabilities of R to calculate p-values (tail probabilities).
- Since R is only interesting if it is very small *or* very large, the proper p-value is two-sided.

R	3	4	5	6	7
$P(R)$	1/6	1/6	1/3	1/6	1/6

- The p-value for $R=3$ or $R=7$ is $1/6 + 1/6 = 1/3$
- The p-value for $R=4$ or $R=6$ *is* $1/6+1/6+1/6+1/6 = 5/6$
- And the p-value for $R=5$ is 1.
- None of these are significant.

Power of Mann-Whitney

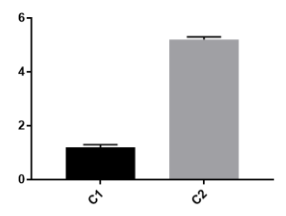
- A 2-vs-2 comparison can never be significant by Mann-Whitney because there simply is not enough information.
  - The minimum viable design would be 3-vs-3.

- Compare two examples:
  1. If the two values in one condition are 2 and 3 and the two values in the other condition are 12,000 and 12,001, it’s still not significant by Mann-Whitney.
  2. Mann-Whitney can’t tell the difference between case 1 above and 2.9 and 3 in one condition, versus 3.1 and 3.2 in the other, because they translate into exactly the same ranks.
     1. And nobody would call this data differential with just two replicates.

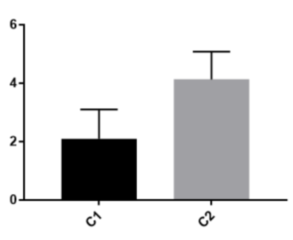
- In contrast, a parametric *T*-test *would* call the first one significant.
  - Because it would be unlikely to obseve four such disparate observations from the same normal distribution.

Mann-Whitney vs. *T*-Test 3-vs-3 Example

- C_1: 1.1, 1.2, 1.3
- C_2: 5.1, 5.2, 5.3



  - Mann-Whitney *p*-value = 0.05
  - *T*-Test *p*-value < 0.0001
  - The *T*-Test is far more powerful, if the assumptions hold
- C_1: 1.1, 2.1, 3.1
- C_2: 3.2, 4.1, 5.1



  - Mann-Whitney *p*-value = 0.05
  - *T*-Test *p*-value = 0.0631
  - Mann-Whitney is more powerful in this case.
- Conclusion: Life is complicated.

Interpretation

In the Mann-Whitney test, the null hypothesis is that the two *distributions* are equal.

In a *T*-test the null hypothesis is that the *means* of the distributions are equal.

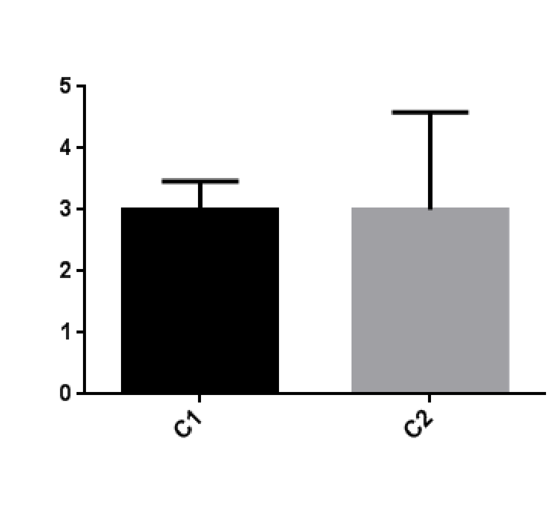
Testing specifically about means is something we lose when we do non-parametric testing.

There are more ways for two distributions to be different besides their means.

Variance, skewness, kurtosis, 5th moment, 6th moment, etc, there are infinitely many moments.

Example of Significant Mann-Whitney Test with Equal Means

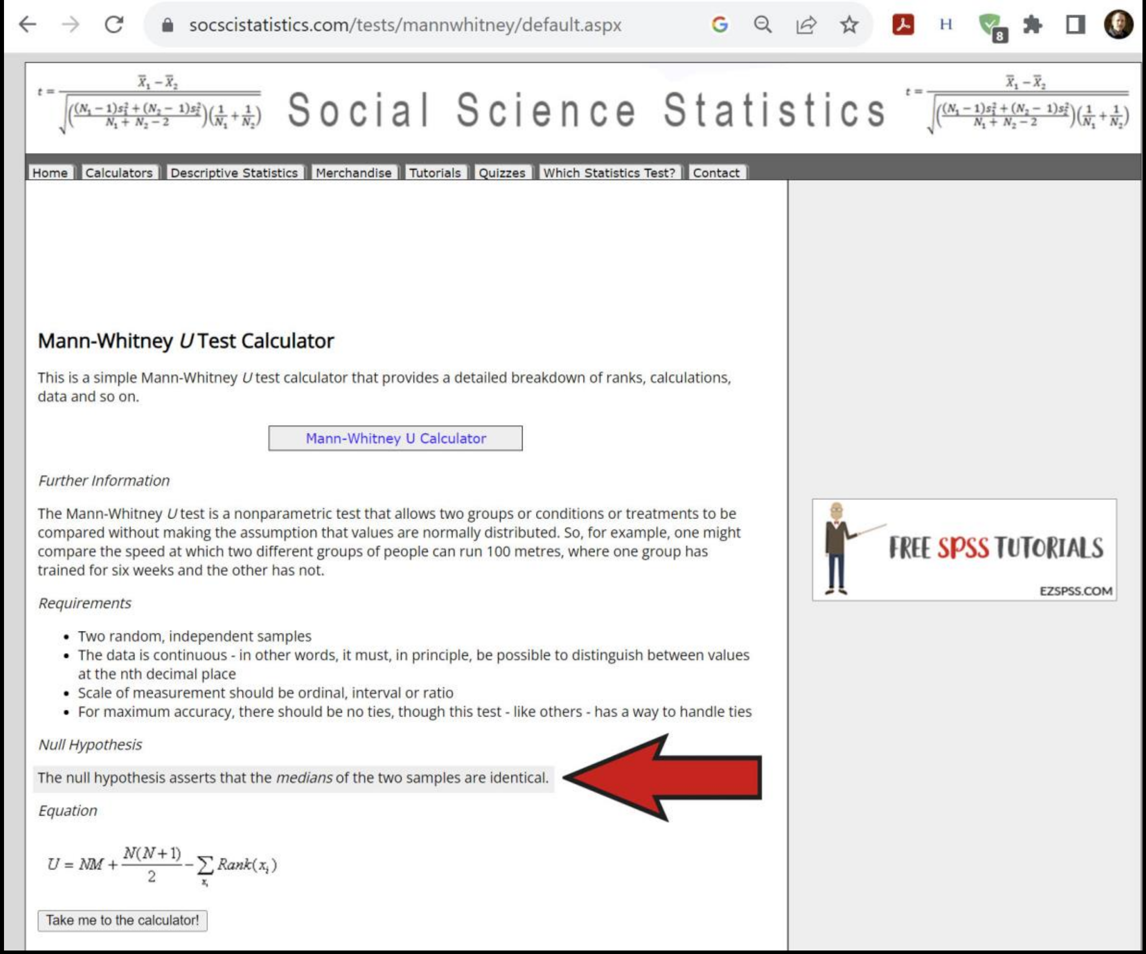
- C_1: 1,1,1,1,1,1,1,1,1,2,2,2,5,5,5,5,5,5,6,6,6
  - median=2, mean = 3
- C_2: 0,0,0,0,0,0,0,0,0,0,2,3,3,3,3,3,3,3,3,3,34
  - median=2, mean = 3



- The *two-sided* Mann-Whitney *p*-value for these data is 0.03788.

Widespread Misconception

Even this online Mann-Whitney server gets it wrong.


socscistatistics.com/tests/mannwhitney/default.aspx
$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{(N_1-1)s_1^2+(N_2-1)s_2^2}{N_1+N_2-2}\right)\left(\frac{1}{N_1}+\frac{1}{N_2}\right)}}$
Social Science Statistics
Home
Calculators
Descriptive Statistics
Merchandise
Tutorials
Quizzes
Which Statistics Test?
Contact
FREE SPSS TUTORIALS
EZSPSS.COM

Mann-Whitney *U* Test Calculator

This is a simple Mann-Whitney *U* test calculator that provides a detailed breakdown of ranks, calculations, data and so on.

Mann-Whitney U Calculator

Further Information

The Mann-Whitney *U* test is a nonparametric test that allows two groups or conditions or treatments to be compared without making the assumption that values are normally distributed. So, for example, one might compare the speed at which two different groups of people can run 100 metres, where one group has trained for six weeks and the other has not.

Requirements

- Two random, independent samples
- The data is continuous - in other words, it must, in principle, be possible to distinguish between values at the nth decimal place
- Scale of measurement should be ordinal, interval or ratio
- For maximum accuracy, there should be no ties, though this test - like others - has a way to handle ties

Null Hypothesis

The null hypothesis asserts that the *medians* of the two samples are identical.

Equation

$$U = NM + \frac{N(N+1)}{2} - \sum_{x_i} Rank(x_i)$$

Interpretation

Don’t interpret a significant Mann-Whitney *p*-value as a difference of means without looking at the data.

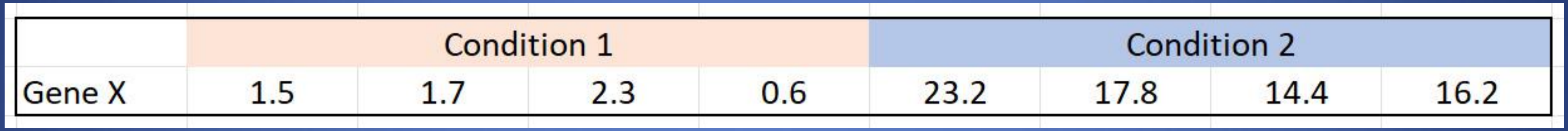
Inspect the graph to make sure what’s different about them is what you expect.

If it is just a difference in variance or skewness, that’s not necessarily uninteresting.

But it needs to be interpreted as such.

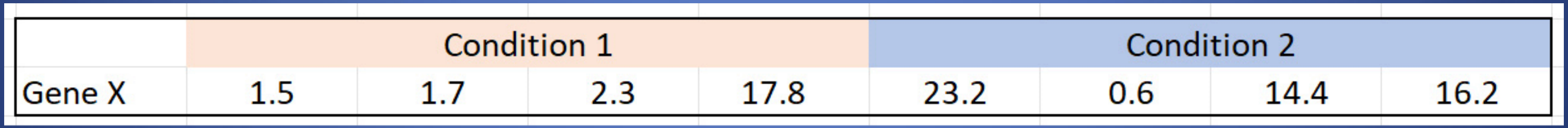
Permutations

- A permutation is where we randomly swap values between conditions.
- Suppose the original (unpermuted) data looks like this.



	Condition 1				Condition 2			
Gene X	1.5	1.7	2.3	0.6	23.2	17.8	14.4	16.2

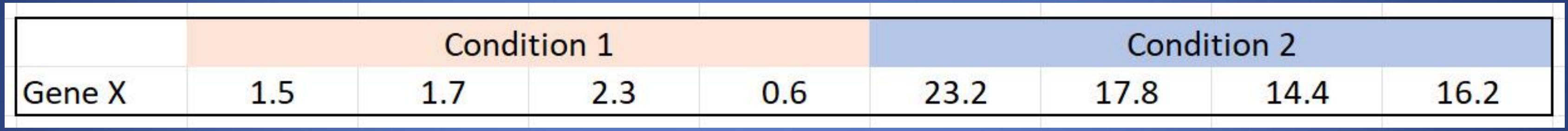
- Swap pairs like this a bunch of times to thoroughly randomize them. Such a randomization then constitutes *one* permutation.



	Condition 1				Condition 2			
Gene X	1.5	1.7	2.3	17.8	23.2	0.6	14.4	16.2

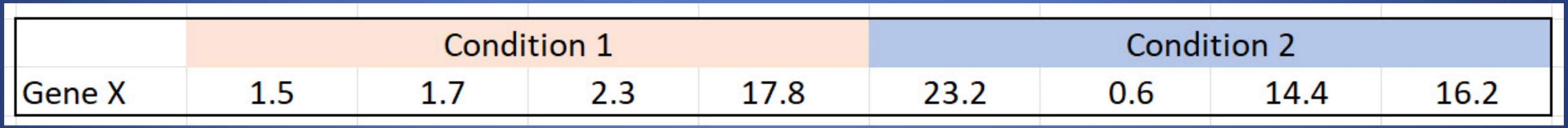
Permutations

- A permutation is where we randomly swap values between conditions.
- Suppose the original (unpermuted) data looks like this.

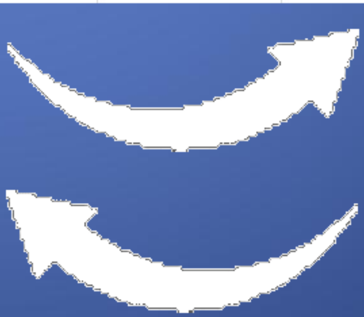


	Condition 1				Condition 2			
Gene X	1.5	1.7	2.3	0.6	23.2	17.8	14.4	16.2

- Swap pairs like this a bunch of times to thoroughly randomize them. Such a randomization then constitutes *one* permutation.



	Condition 1				Condition 2			
Gene X	1.5	1.7	2.3	17.8	23.2	0.6	14.4	16.2



Number of Permutations

- Suppose C_1 has *N* replicates and C_2 has *M* replicates.
  - There are *N+M* observations in total.
  - A permutation consists of choosing *N* of them and putting them in group one, and putting the remaining *M* in group two.
  - Thus, each way we can choose *N* things from *N+M* things gives a distinct permutation.
  - The number of ways to choose *N* things from *N+M* things is given by the formula:

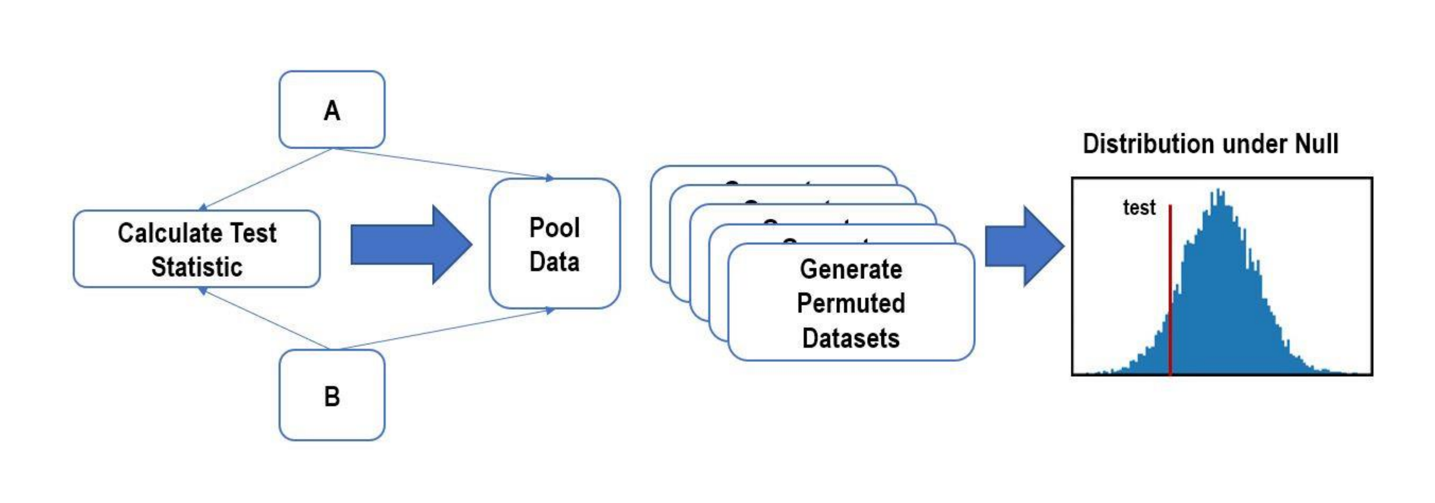
$$\binom{N+M}{N} = \frac{(N+M)!}{M!\,N!}$$

- *N=2, M=2*: 6 permutations
- *N=3, M=3*: 20 permutations
- *N=4, M=4*: 70 permutations
- *N=5, M=5*: 252 permutations
- …
- *N=10, M=10*: 184,756 permutations

Note: The unpermuted data counts as one permutation, you could by chance shuffle them back into place.

Permutation Tests

- A “permutation test” is a general approach to significance testing.
- It can be used almost anywhere, and it allows you to avoid all the difficult issues of determining the distribution of a random variable.
- It is like duct tape, a multipurpose down and dirty tool in every bioinformatician’s toolbox.



Permutation Tests When/Where/Why

If permutation tests have so many advantages, why not always do things that way?

  - – Because sometimes it's just better to hang a picture on the wall with a nail and a hammer.

1) Power and Efficiency Issues

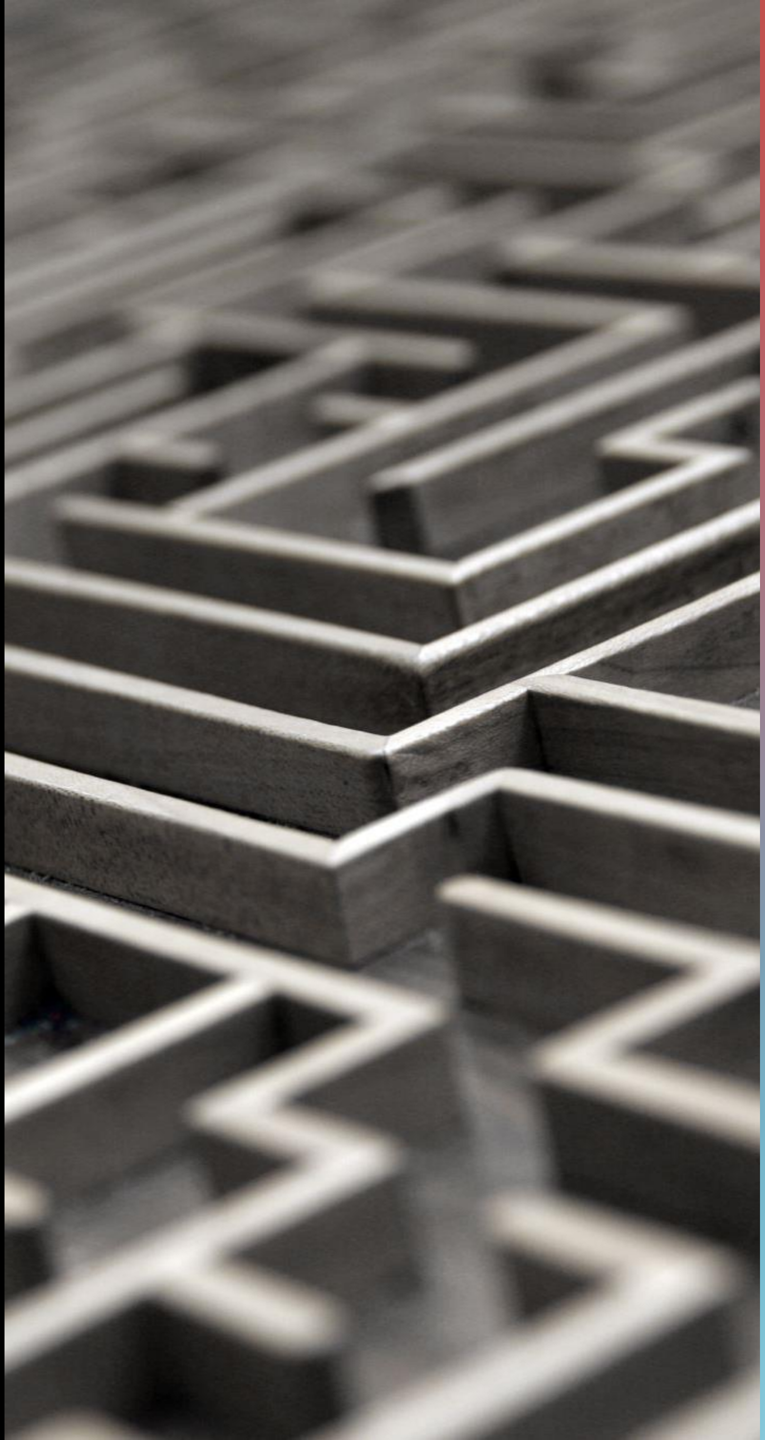
- If you have a closed form parametric solution to a problem, it's almost always better in terms of power and efficiency.
  - – For example, the parametric *T*-test is more powerful than any non-parametric approach, when its assumptions are not excessively violated.

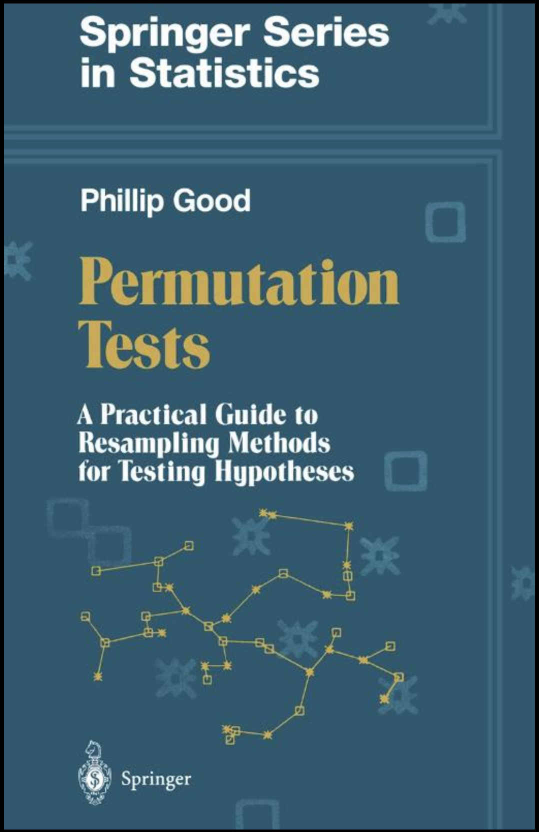
2) Null Hypothesis Issues

- Permutation tests like most non-parametric tests rely on the stronger null hypothesis of equal distributions.

3) Implementation Issues

- Figuring out what to permute is not always straightforward and can involve simplifying assumptions.





Springer Series in Statistics

Phillip Good

Permutation Tests

A Practical Guide to Resampling Methods for Testing Hypotheses

Springer

Permutation Tests

- A permutation test allows us to design a *non-parametric* test using any statistic.

- We'll use the *T*-Statistic.
  - Just because we use the *T*-Statistic does not mean we're doing a parametric *T*-test.

- As with the Mann-Whitney and most non-parametric tests, we again assume the stronger null distribution of equal distributions.

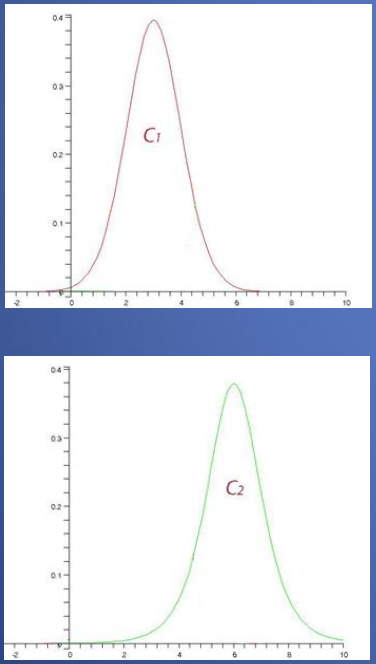
Implications

- As with Mann-Whitney, the null is “equal distributions” not “equal means”.
  - Therefore, the null could be rejected when the means are equal.
  - There’d have to be something else different about the distributions, such as their spread (variance).
- But it’s much less likely to happen with a permutation test than a Mann-Whitney, because of how the permutation test utilizes the *T*-statistic.
  - The *T*-statistic is severely underpowered for detecting anything but a difference in means.
- So, a rejection tends to be about means, even if it could in theory be about variance in edge cases.

$$T(X,Y)=\frac{\bar{X}-\bar{Y}}{S\sqrt{1/N+1/M}}$$

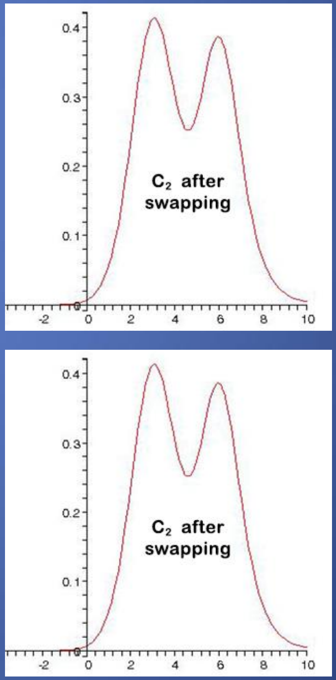
Permutations when null is false

- Permutation tests rely on the following principle:
  - If the groups have the different distribution, then a permutation should tend to equalize them.



Before permuting both groups are different (the null is false)

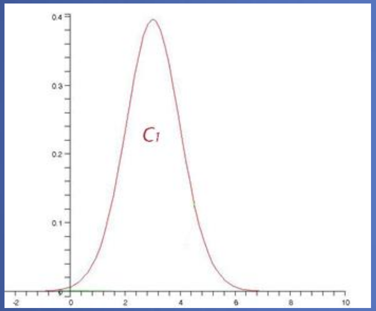
Randomly swap values between the two groups

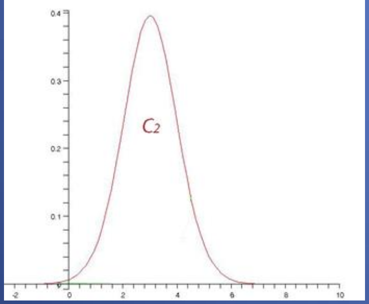


After permuting both groups look the same (like this)

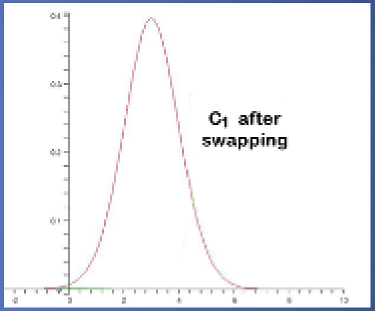
Permutation when null true

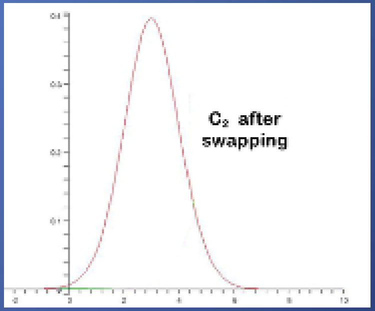
- If they were not different before permuting, then they will tend to not be different after permuting.





Randomly swap values between the two groups

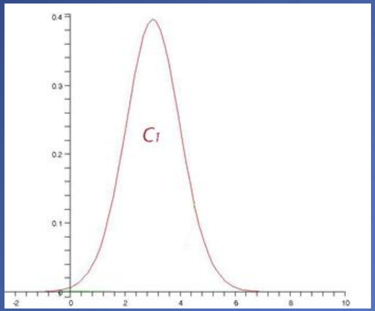


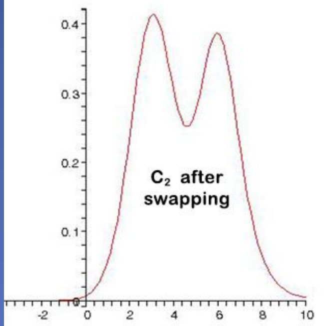


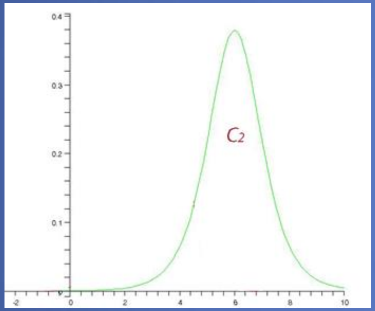
Before permuting both groups are the same (the null is true)

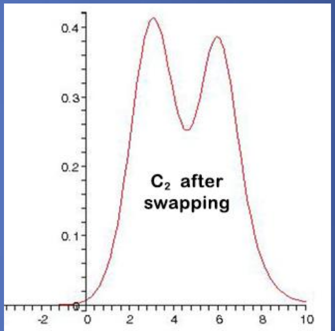
After permuting both groups still look the same

Permutations





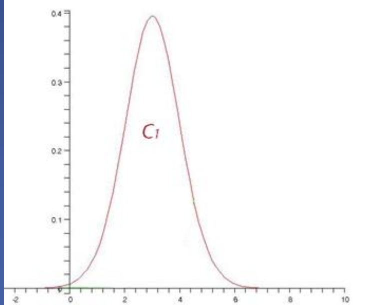


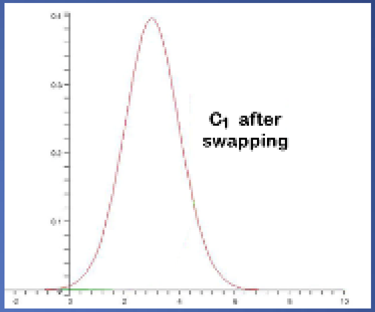


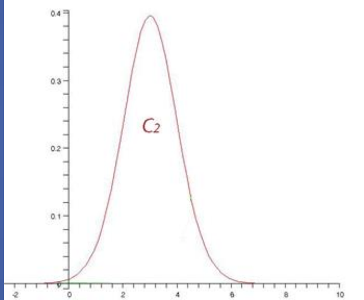
The *T*-statistic for data drawn from the distributions on the left (the unpermuted data) will be much more extreme (greater mean) than the *T*-statistic for data drawn from the distributions on the right (one permutation of the data)

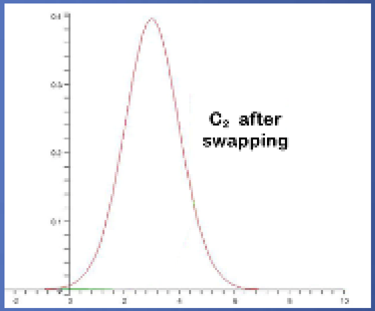
This is because the means are different on the left and the close to the same on the right.

Permutations



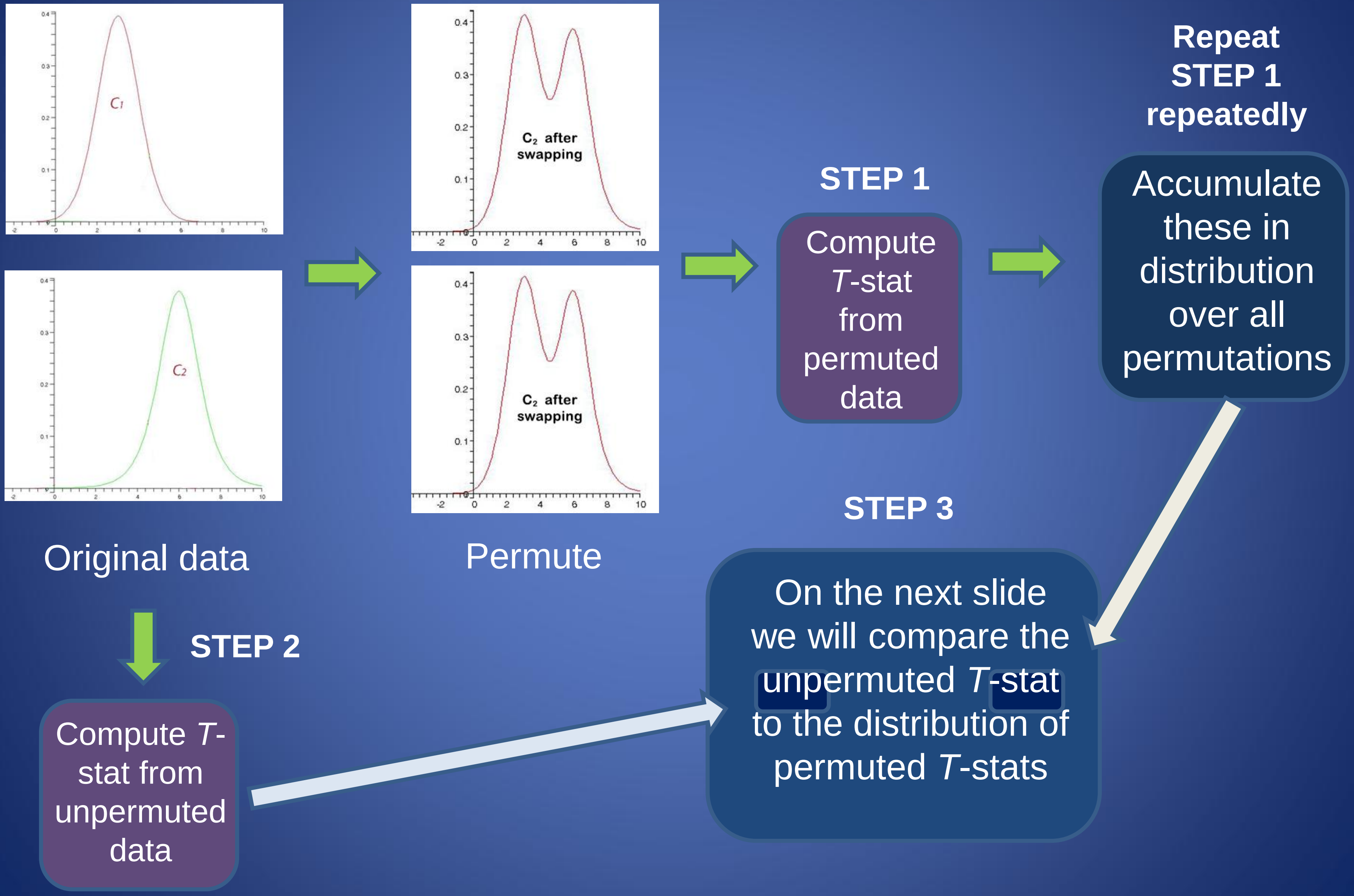






If C_1 and C_2 have the same distribution, then the T-statistic would look the same on unpermuted data is it does on permuted.

Permutation *P*-values



Original data

Permute

STEP 1

Compute *T*-stat from permuted data

Repeat STEP 1 repeatedly

Accumulate these in distribution over all permutations

STEP 2

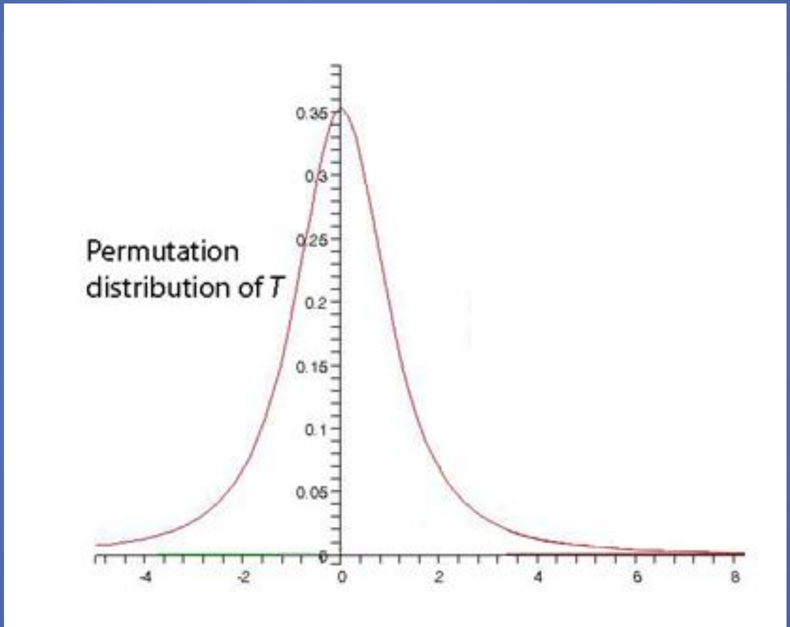
Compute *T*-stat from unpermuted data

STEP 3

On the next slide we will compare the unpermuted *T*-stat to the distribution of permuted *T*-stats

The Permutation Distribution

- The permutation distribution is obtained by plotting the *T*-Statistic *over all possible permutations*. It will look something like this:

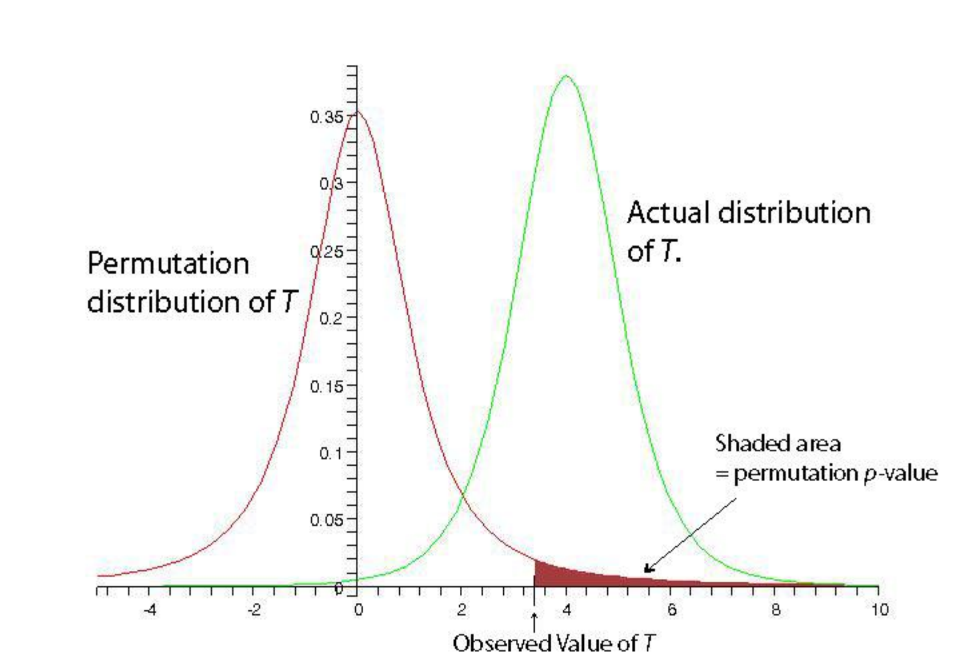


- If the Null Hypothesis is false, then the *T*-Statistic on the *unpermuted* data should tend to be extreme with respect to this distribution.

The Permutation *p*-value

– The area under the permutation distribution to the right of the observed *T*-Stat (calculated from the unpermuted data) is called the *permutation p-value* (shaded area).

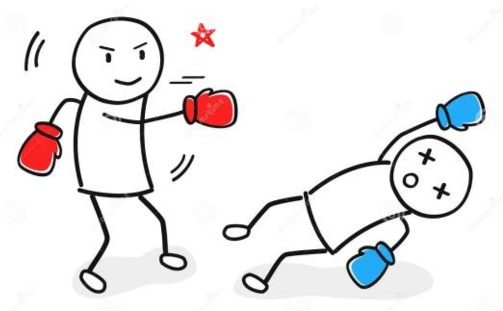
– Rejecting the null hypothesis if the permutation *p*-value is less than α controls the Type I error at level α.



Practical Calculation of the Permutation *p*-value

- Let T_0 be the *T*-Statistic on the unpermuted data.
- Count *m* = number of permutations for which the permuted *T*-Statistic is greater than T_0.
- Let *k* = total number of permutations.
- The permutation *p*-value = *m/k.*

Mann-Whitney vs. Permutation *T*-Test



- Let's revisit the problematic example where means were equal but Mann-Whitney *p*-value is significant. Let's see what happens with the permutation test on the same data.

- C_1: 1,1,1,1,1,1,1,1,1,2,2,2,5,5,5,5,5,5,6,6,6
  - median=2, mean = 3
- C_2: 0,0,0,0,0,0,0,0,0,0,2,3,3,3,3,3,3,3,3,3,34
  - median=2, mean = 3

- The Mann-Whitney *p*-value for these data is 0.0374.
- The permutation *p*-value ≈ 0.5011

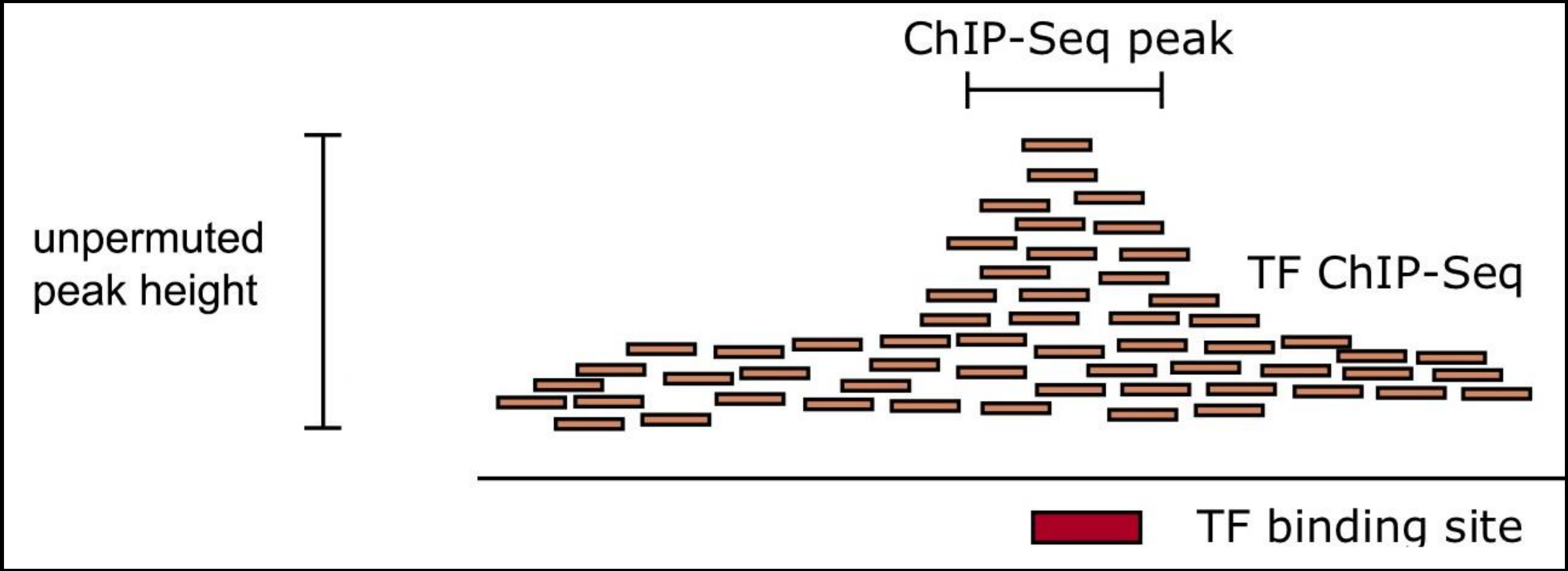
- **Conclusion:** *In contrast to Mann-Whitney, the T-statistic permutation p-value is less sensitive to differences other than the mean.*

Advantages of the Permutation Approach

- Permutation methods are employed widely.
  - Not just in differential expression.
- Whenever you need to calculate *p*-values, you can often solve the problem with permutations.
  - Sometimes there's no other obvious way to do it.
- The great advantage of the permutation approach is there's no need to calculate any distributions of any type.
  - One simply permutes, and counts.
- The trick is to find a good statistic make sense of the appropriate permutations.
  - It's not always as straightforward as it was for the DE problem.

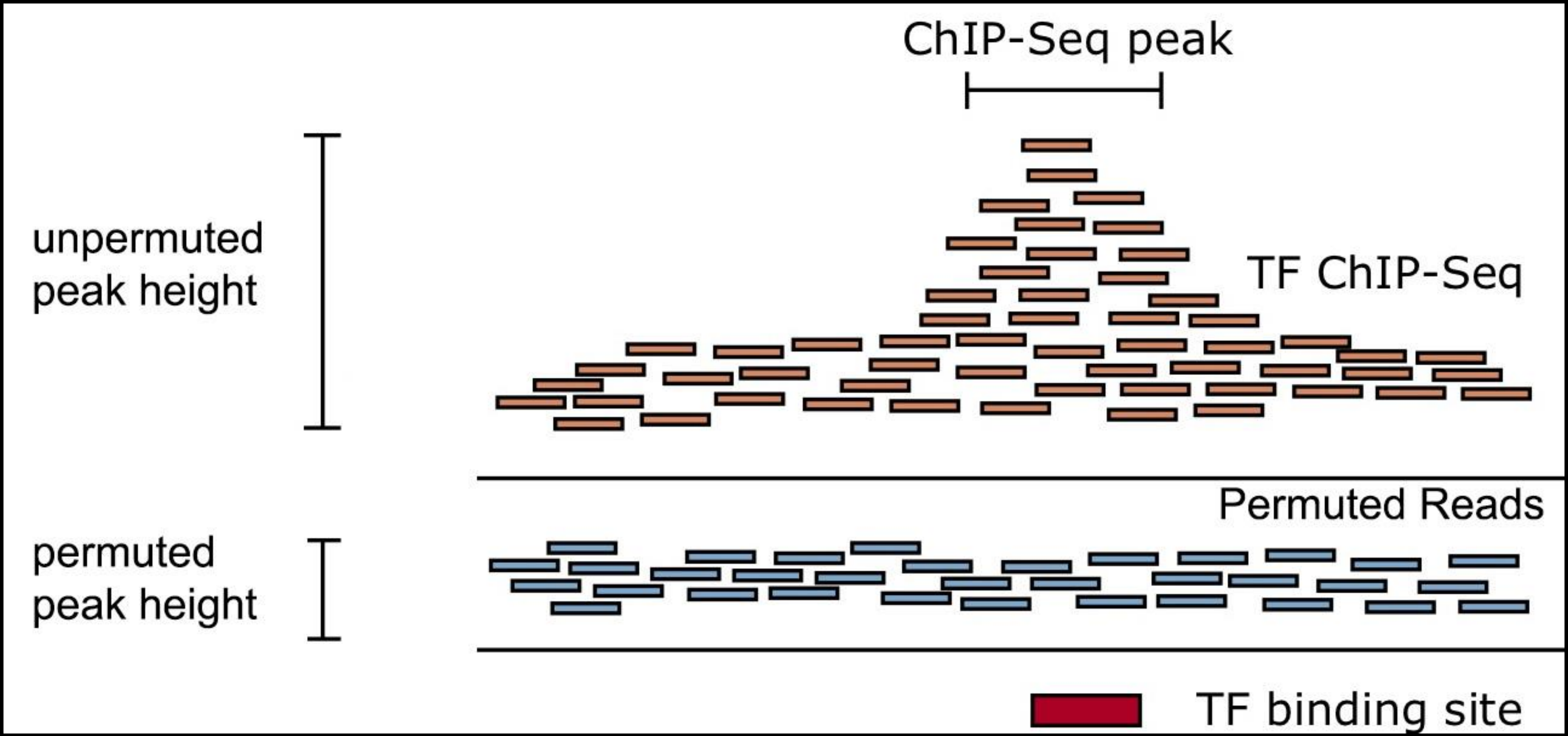
Example Peak Finding

- Suppose you want to test for significance of a ChIP-seq peak height.
- The statistic is “peak height”.
- We want to know if reads are accumulating in one spot just by chance.
- What do you permute?



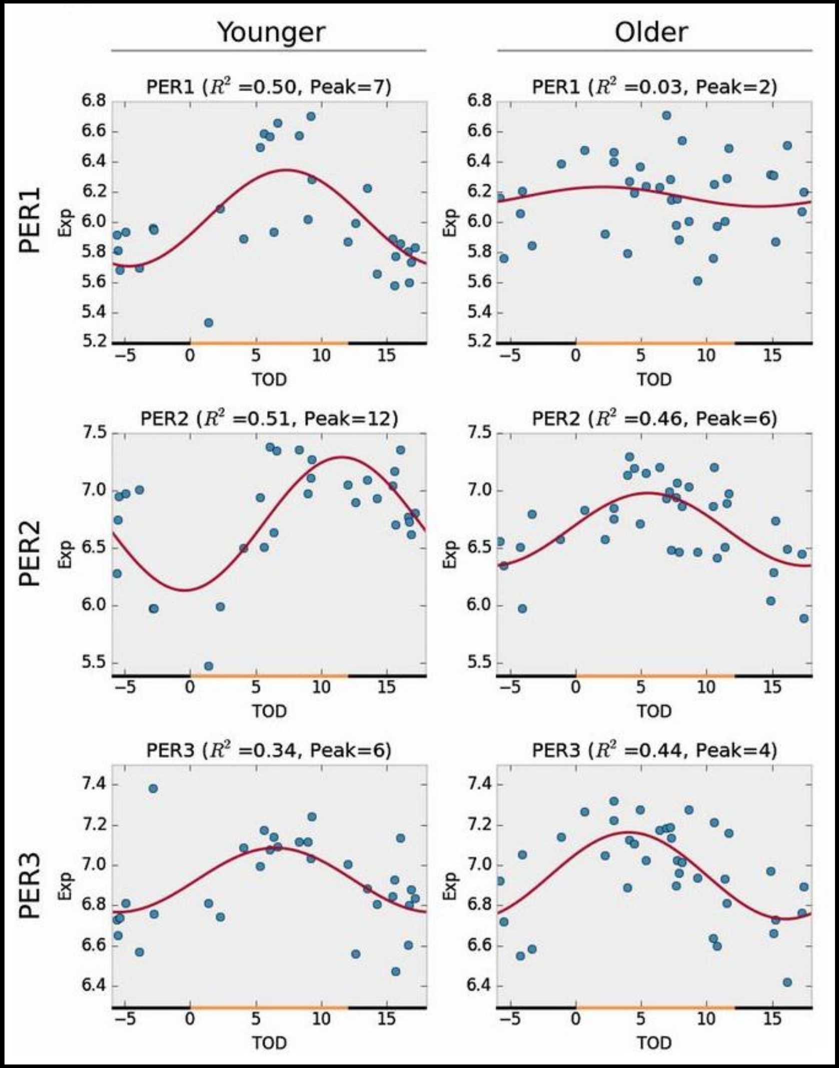
Example Peak Finding

- Permute the alignment location of reads.
- Move them to a random location.
- Calculate the *maximum* peak height in the permuted data.
- Repeat many times to get permutation distribution



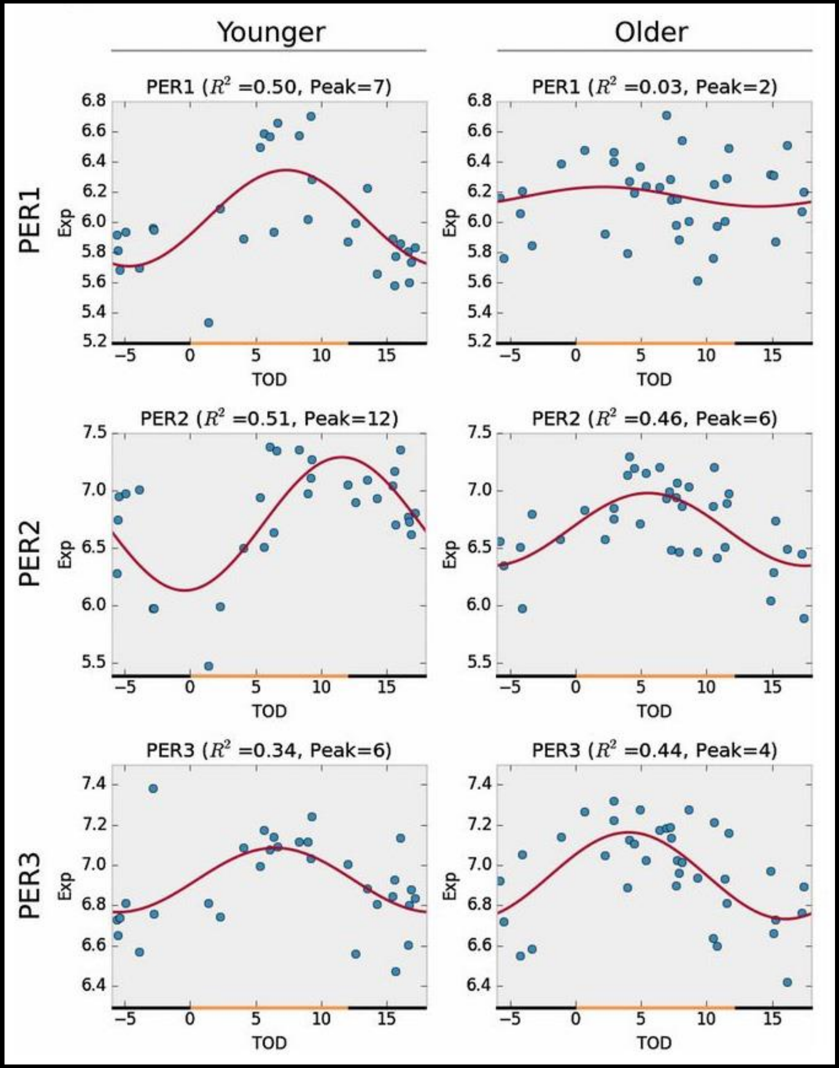
Example Circadian Genes

- We want to test genes for circadian behavior across time.
- What should we permute?
- What's a good statistic?



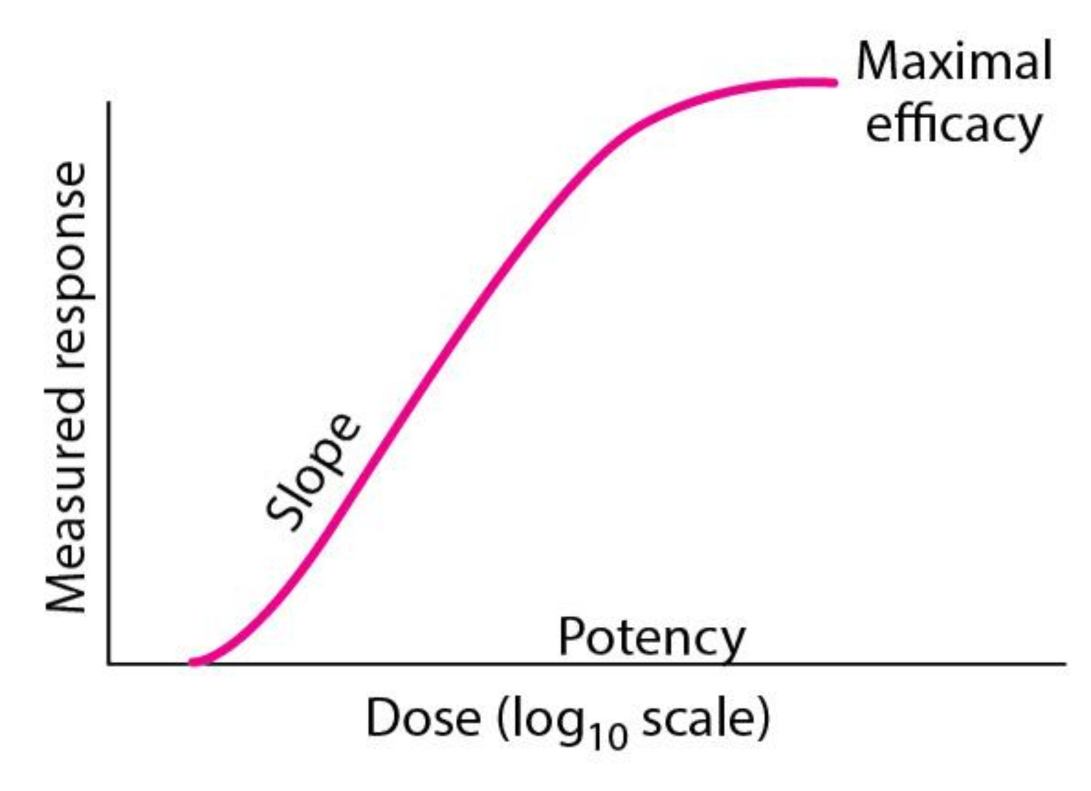
Example
Circadian Genes

- What should we permute?
  - Time points.
- What's a good statistic?
  - How well a cosine curve explains the variance.



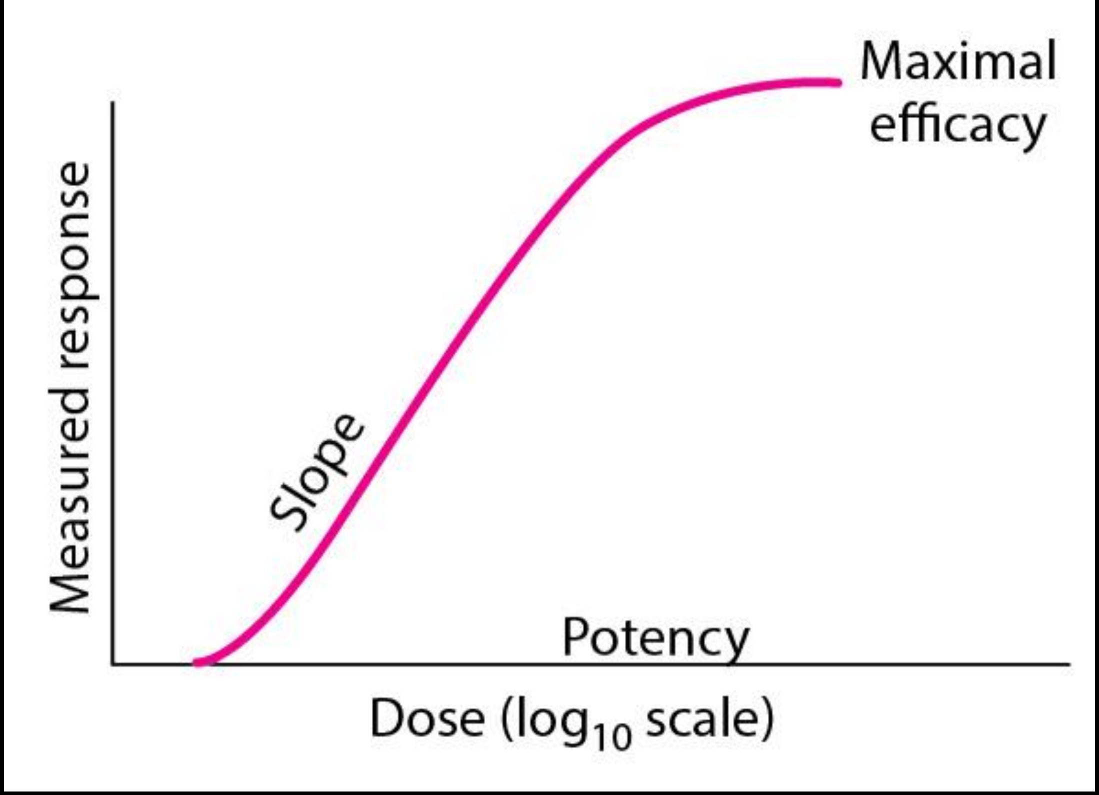
Example Dose/Response Curve

- Give increasing amounts of a drug and test for increasing effects.
- What should we permute?
- What's a good statistic?



Example Dose/Response Curve

- Give increasing amounts of a drug and test for increasing effects.
- What should we permute?
  - Dosage
- What's a good statistic?
  - Slope

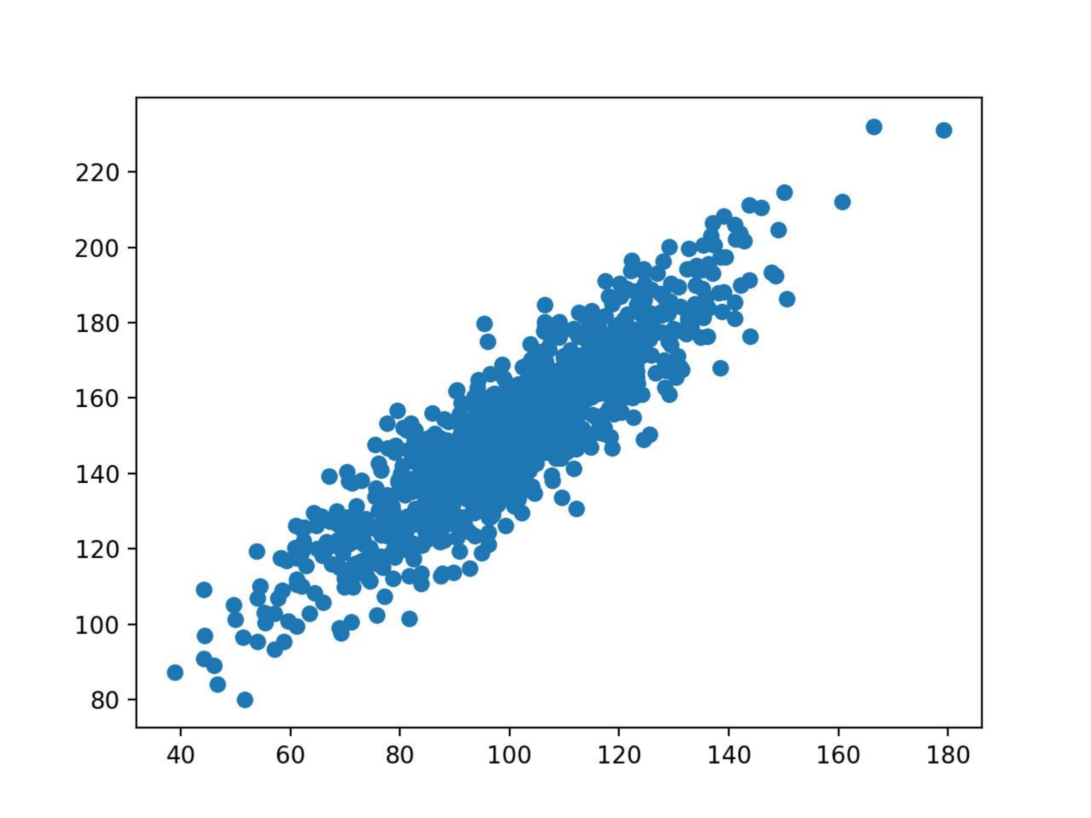


Example Correlation

- Figure shows correlated variables.
- For example, is gait associated with lifespan?
  - Data: Gait and lifespan on 100,000 individuals.
  - The U.K. Biobank has data like this.

- What's a good statistic?

- What should be permuted?

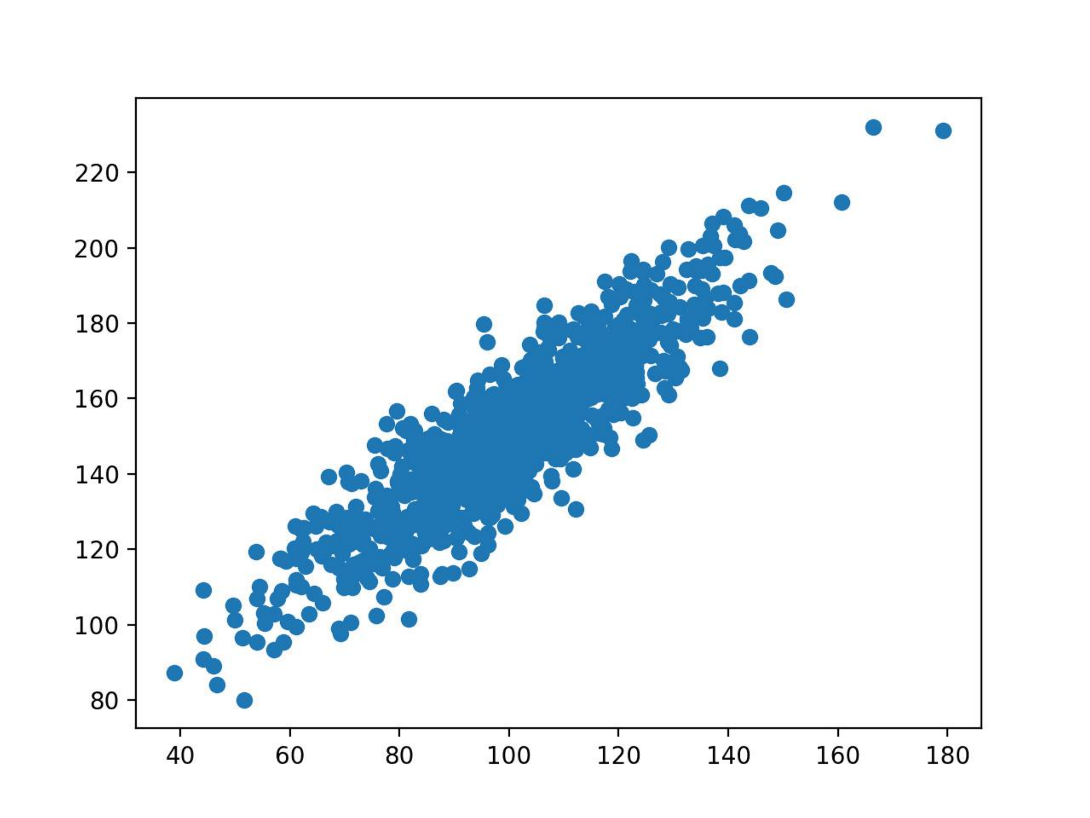


Example Correlation

- Figure shows correlated variables.
- For example, is gait associated with lifespan?
  - Data: Gait and lifespan on 100,000 individuals.
  - The U.K. Biobank has data like this.

- What's a good statistic?
  - Correlation

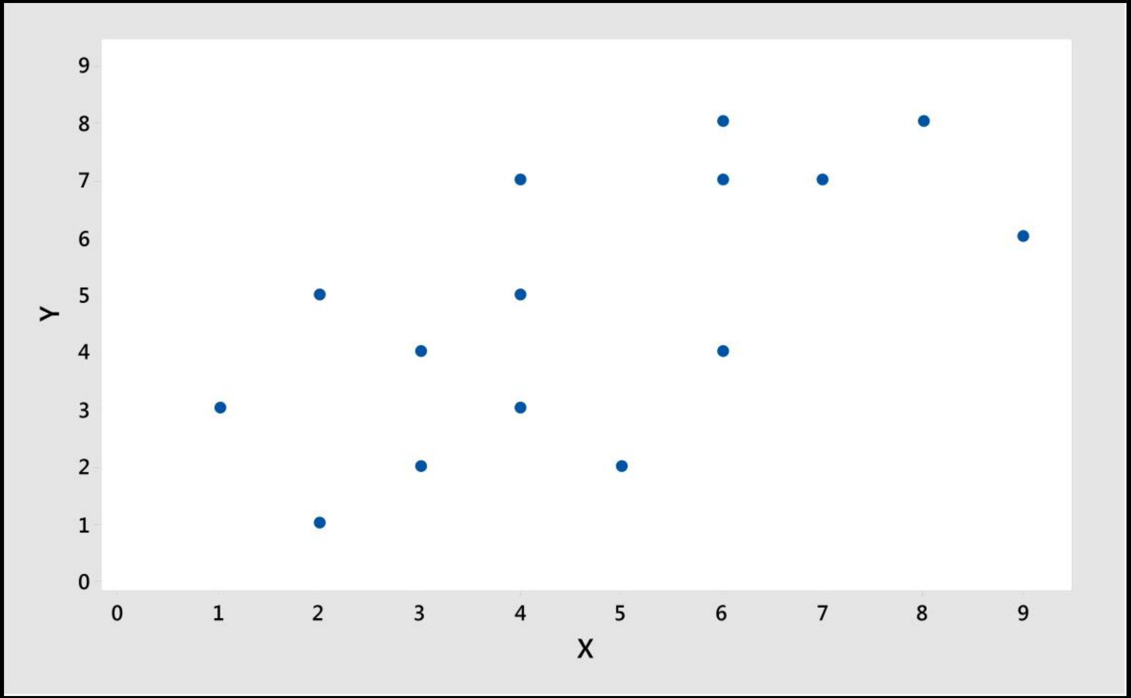
- What should be permuted?
  - The x-axis values of each point.
  - Shuffle around the first column of data while leaving the second column fixed.



Example Correlation

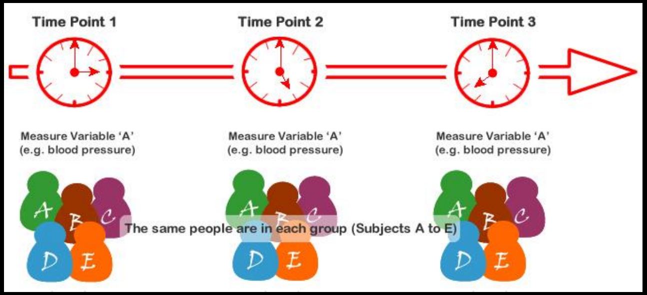
- Figure shows less obvious correlation.
- It’s also possible to have many points but for the correlation to be very weak.

- Both cases call for testing.



This data is less obvious and a rigorous test is required to draw conclusions.

Paired Data


Time Point 1
Time Point 2
Time Point 3
Measure Variable 'A' (e.g. blood pressure)
Measure Variable 'A' (e.g. blood pressure)
Measure Variable 'A' (e.g. blood pressure)
A B C D E
A B C D E
A B C D E
The same people are in each group (Subjects A to E)

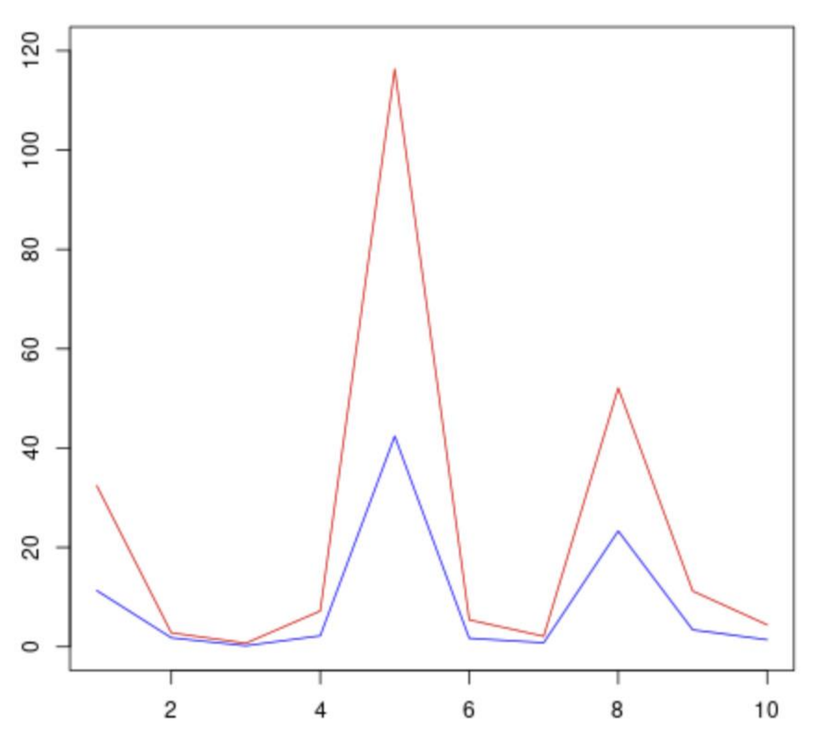
- Suppose we want to know if blood pressure is higher, or lower, in the evening than in the morning.

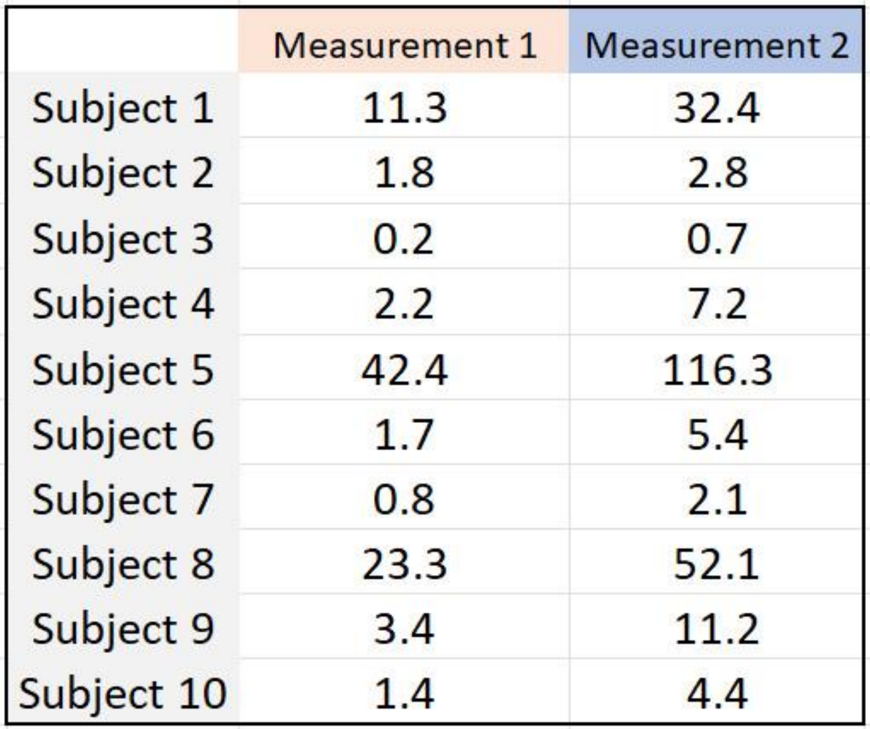
- **Study Design 1:** We could take blood pressure measurements from 100 individuals in the morning and 100 different individuals in the evening.

- **Study Design 2:** Or we could take two blood pressure measurements, one in the morning and one in the evening, in the same 100 people.
  - This is called a "repeated measurement design".

Repeated Measures

- Repeated measures design controls for extreme individual variability.
- In this data Measurement 2 was generally two to three times higher from baseline (Measurement 1).
- But baseline varies greatly.

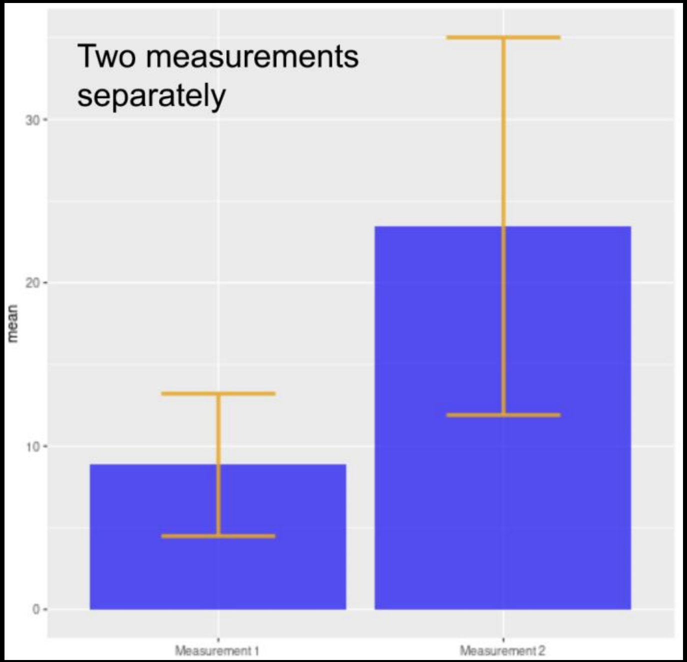




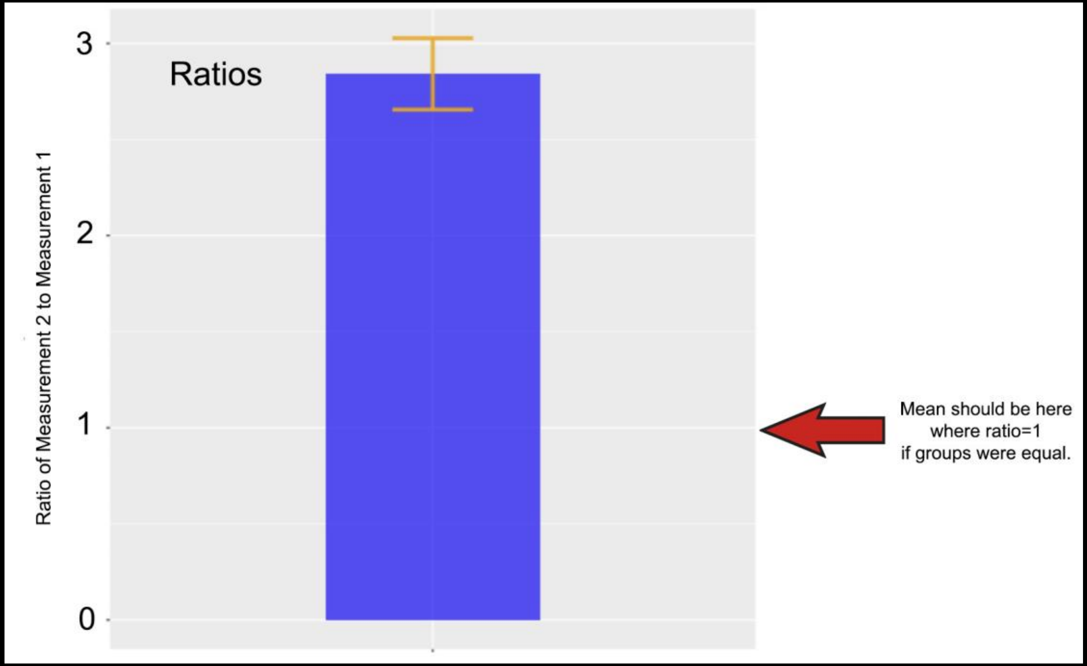
	Measurement 1	Measurement 2
Subject 1	11.3	32.4
Subject 2	1.8	2.8
Subject 3	0.2	0.7
Subject 4	2.2	7.2
Subject 5	42.4	116.3
Subject 6	1.7	5.4
Subject 7	0.8	2.1
Subject 8	23.3	52.1
Subject 9	3.4	11.2
Subject 10	1.4	4.4

Repeated Measures

- Consider the graph of the two measurements without taking repeated measures into account.
  - Graph on the left.
- Compare that to the graph of the ratios of Measurement 2 to Measurement 1.
  - Graph on the right.



Groups not clearly separated



Mean should be here where ratio=1 if groups were equal.

Ratio is more clearly separated from Y=1

Implication

- There are both parametric and non-parametric tests for repeated measures (paired) data.

- You cannot just apply a regular *T*-test to repeated measures data because it assumes all observations are independent.
  - In paired data the two measurements from the same individual are not independent.
  - The test needs to account for the dependence.

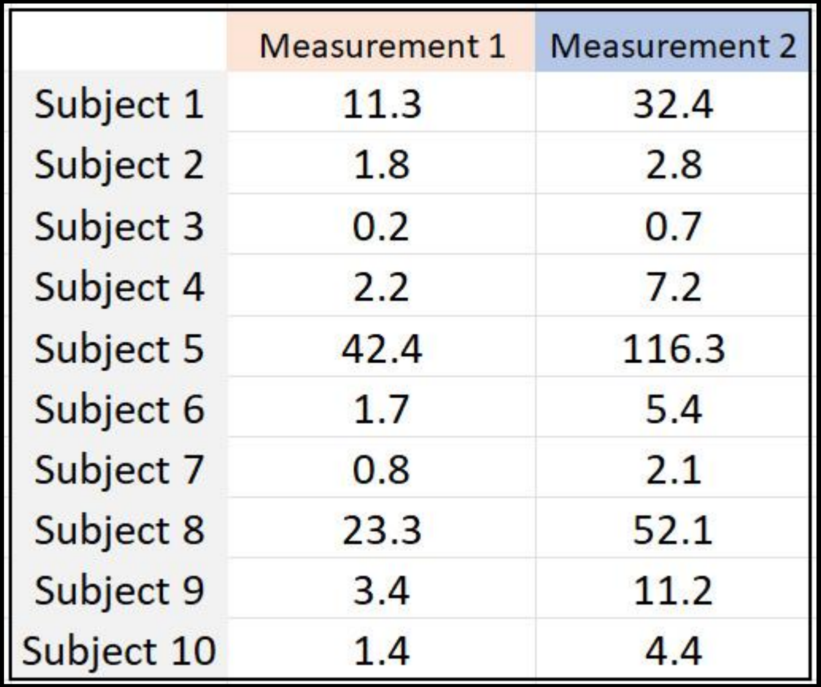
- One way to do this is to work with ratios:
  - E.g., evening/morning and using the null hypothesis

  H_0: mean ratio = 1

Permutation Test for Paired Data

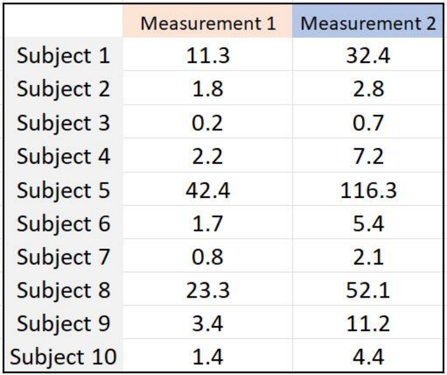
- What are the appropriate permutations in the context of repeated measures?

- When they aren't repeated measures, we just swap anything in column 1 with anything in column 2.

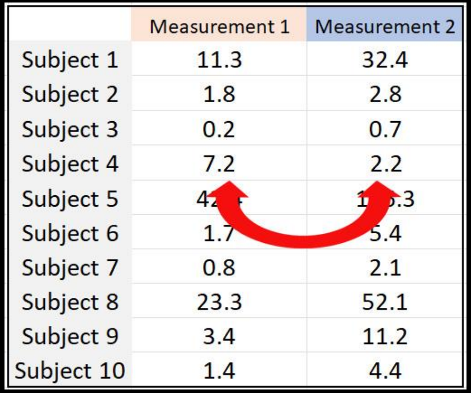


	Measurement 1	Measurement 2
Subject 1	11.3	32.4
Subject 2	1.8	2.8
Subject 3	0.2	0.7
Subject 4	2.2	7.2
Subject 5	42.4	116.3
Subject 6	1.7	5.4
Subject 7	0.8	2.1
Subject 8	23.3	52.1
Subject 9	3.4	11.2
Subject 10	1.4	4.4

Permute Within Subject



	Measurement 1	Measurement 2
Subject 1	11.3	32.4
Subject 2	1.8	2.8
Subject 3	0.2	0.7
Subject 4	2.2	7.2
Subject 5	42.4	116.3
Subject 6	1.7	5.4
Subject 7	0.8	2.1
Subject 8	23.3	52.1
Subject 9	3.4	11.2
Subject 10	1.4	4.4



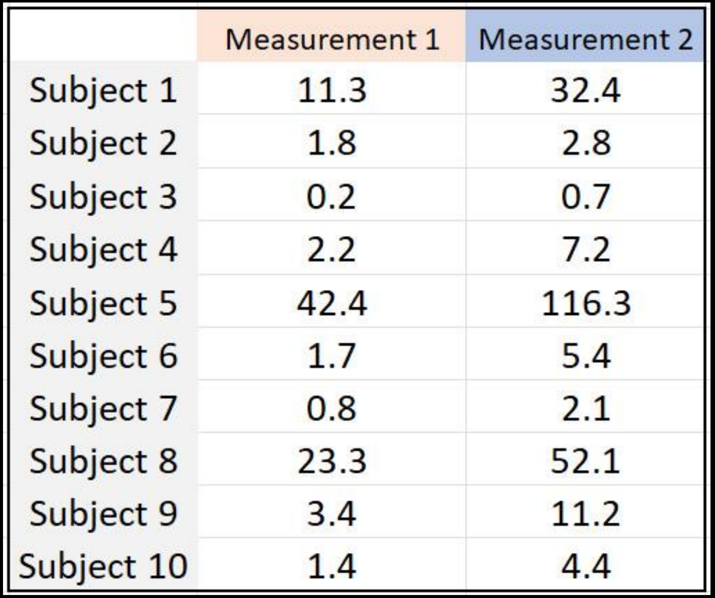
	Measurement 1	Measurement 2
Subject 1	11.3	32.4
Subject 2	1.8	2.8
Subject 3	0.2	0.7
Subject 4	7.2	2.2
Subject 5	42.4	116.3
Subject 6	1.7	5.4
Subject 7	0.8	2.1
Subject 8	23.3	52.1
Subject 9	3.4	11.2
Subject 10	1.4	4.4

- With repeated measures (paired) data we only swap the measurements from *the same subject*.

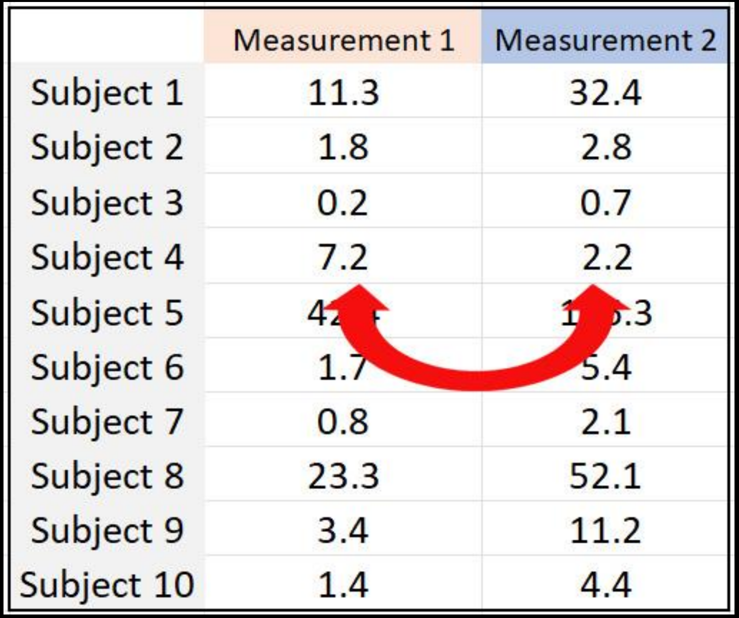
- We'll do this swap a bunch of times until it's well randomized.
  - That then constitutes *one* permutation.

Number of Paired Permutations

- With unpaired data there's $\binom{20}{10} = 184{,}756$ permutations.
- With paired data there are $2^{10} = 1{,}024$ permutations.
  - So far fewer. That means the smallest possible *p*-value is $\frac{1}{1024} = 0.00097$



	Measurement 1	Measurement 2
Subject 1	11.3	32.4
Subject 2	1.8	2.8
Subject 3	0.2	0.7
Subject 4	2.2	7.2
Subject 5	42.4	116.3
Subject 6	1.7	5.4
Subject 7	0.8	2.1
Subject 8	23.3	52.1
Subject 9	3.4	11.2
Subject 10	1.4	4.4



	Measurement 1	Measurement 2
Subject 1	11.3	32.4
Subject 2	1.8	2.8
Subject 3	0.2	0.7
Subject 4	7.2	2.2
Subject 5	42.4	116.3
Subject 6	1.7	5.4
Subject 7	0.8	2.1
Subject 8	23.3	52.1
Subject 9	3.4	11.2
Subject 10	1.4	4.4

If you are comparing measurements at two times, why not always use repeated measures?

- For several reasons.
- For one, you may have to sacrifice the animal to make the measurement.
    - Blood, skin, hair, even adipose can be obtained from live animals.
    - And various biometrics like blood pressure.
    - Most other tissues and many biometrics require sacrifice.
        - E.g. brain tissue or total weight of brown adipose fat in a mouse.

Variability Also Matters

- Doing a paired test can lower the variability of the measurements (by working with ratios).

- But what if there is no subject-to-subject variability to begin with?
  - In that case working with ratios has not decreased the variance.
  - But now we have less permutations and less powerful tests in general.

- *In a nutshell, paired tests help you when there's a lot of subject level variance and hurt you when there isn't.*

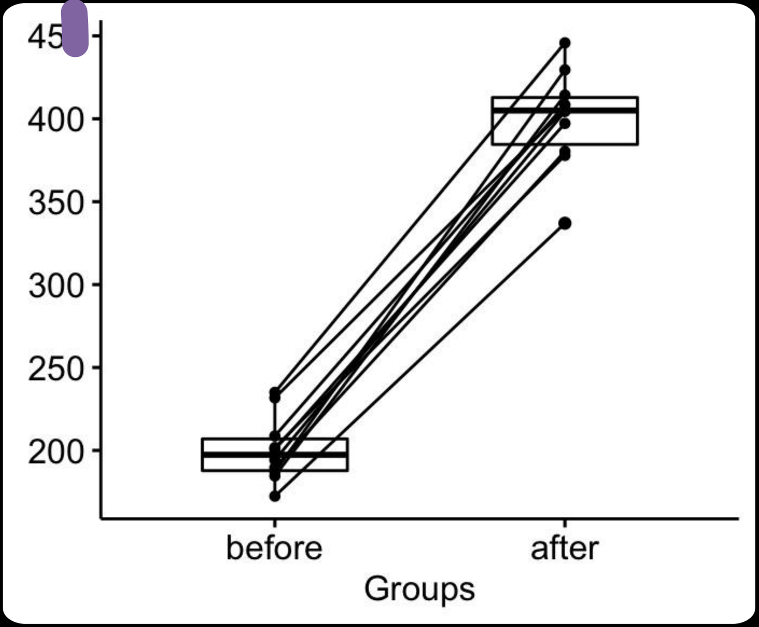
The Wilcoxon Signed-Rank Test

- There's also a *rank-based* non-parametric method *for paired data,* called the Wilcoxon Signed-Rank Test.
  - – Mann-Whitney: Unpaired
  - – Wilcoxon: Paired (repeated measures)
- Call measurement 1 m_1 and measurement 2 m_2.
- There are *N* subjects and these two measurements made for each.
- Wilcoxon works with the list of *N* differences

$$m_1 - m_2$$

- These are then sorted and given ranks.

- Under the null, half should be positive and half negative.



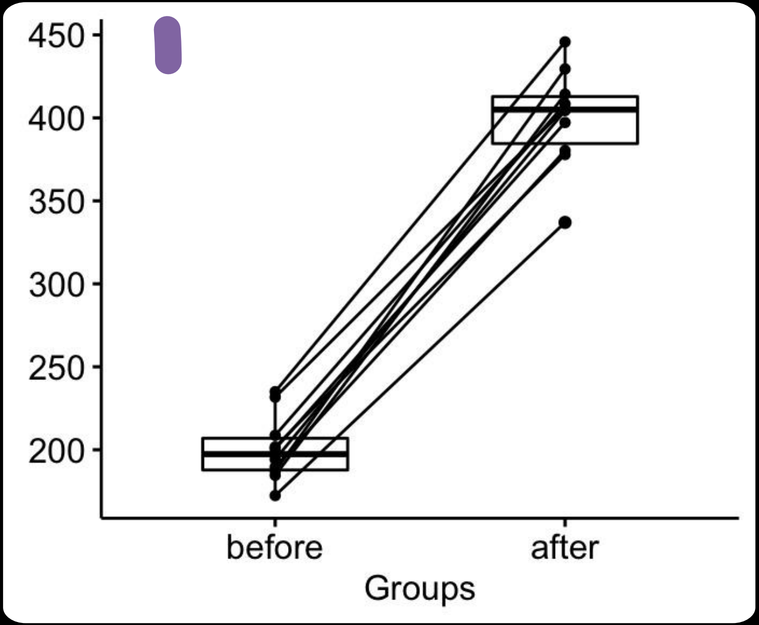
The Wilcoxon Signed-Rank Test

- Under the null we expect this difference to be positive in about half the subjects and negative in the other half.

- A quantity *T* is obtained by *adding* the ranks for each subject for which the difference $m_1 - m_2 > 0$ and *subtracting* the rank for each subject for which the difference $m_1 - m_2 < 0$

- *T* is another quantity like *R* in Mann-Whitney for which it is relatively easy to calculate its distribution under the null hypothesis, without making any parametric assumptions.

- We're skipping the details, for time.


before
after
Groups