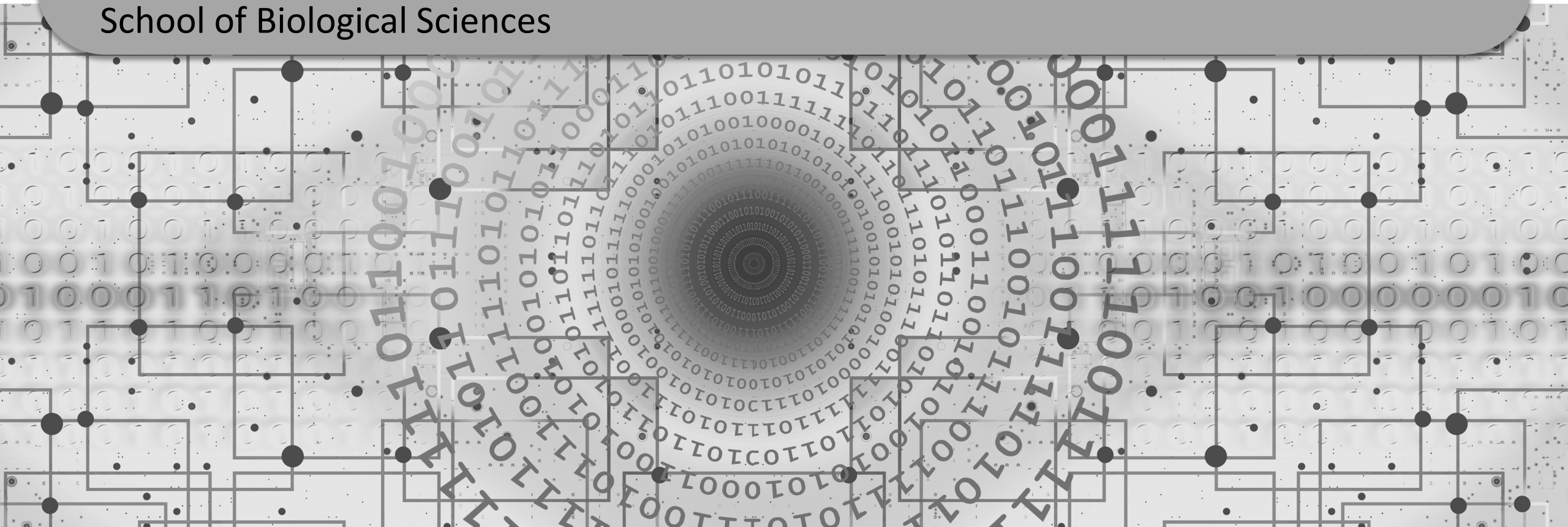
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

How Statistics go Wrong in Data Science

BS0004 Introduction to Data Science

Dr Wilson Goh
School of Biological Sciences



Learning Objectives

By the end of this topic, you should be able to:

- Describe the Anna Karenina principle and the following terms:
  - Random sampling error
  - Subpopulation effects
  - Wrong null distribution
  - Breast cancer biomarkers

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

The Anna Karenina Principle

BS0004 Introduction to Data Science

Dr Wilson Goh
School of Biological Sciences

Anna Karenina Principle

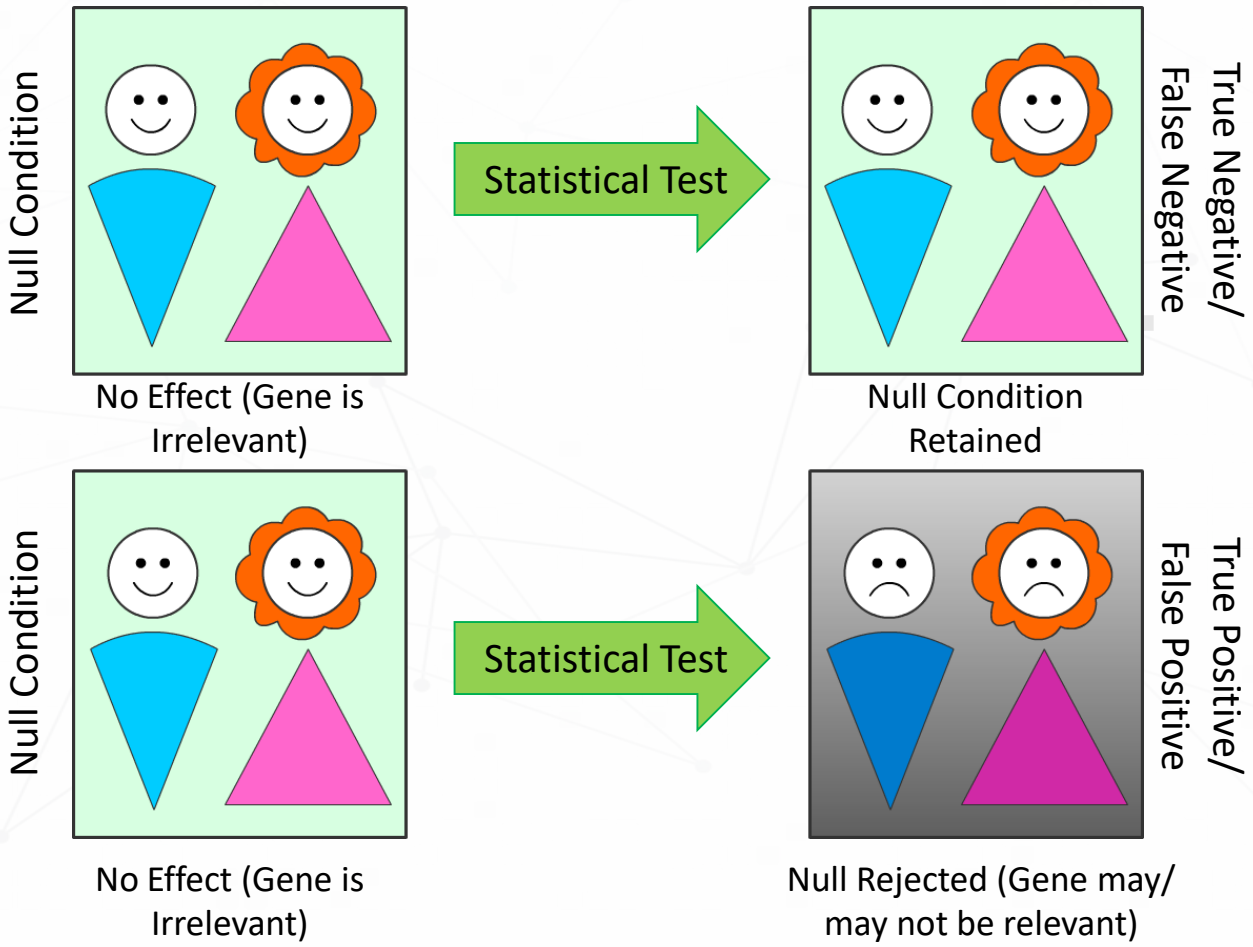
Happy families are all alike; every unhappy family is unhappy in its own way.

~ *Leo Tolstoy*

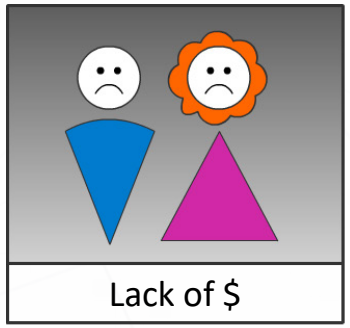
Translation: There are many ways to violate the hull hypothesis but only one way that is truly pertinent to the outcome of interest.

Setup for a Statistical Test

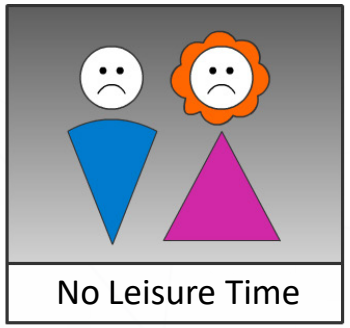
The elements of null hypothesis statistical testing.



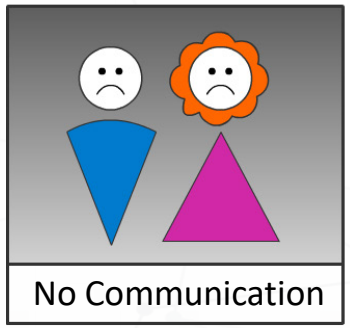
It is in fact, very easy to reject the null hypothesis



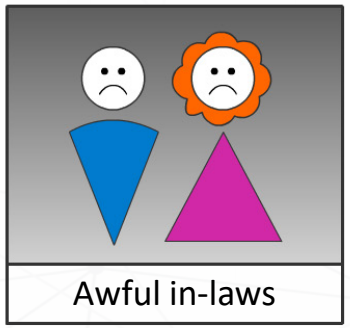
Lack of $



No Leisure Time

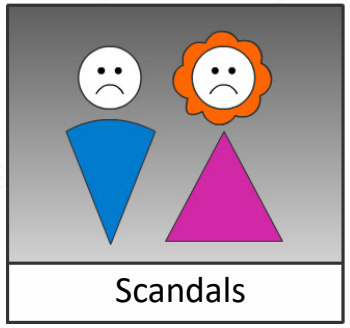


No Communication

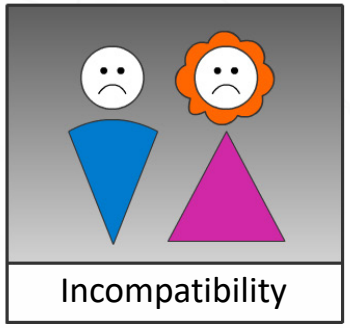


Awful in-laws

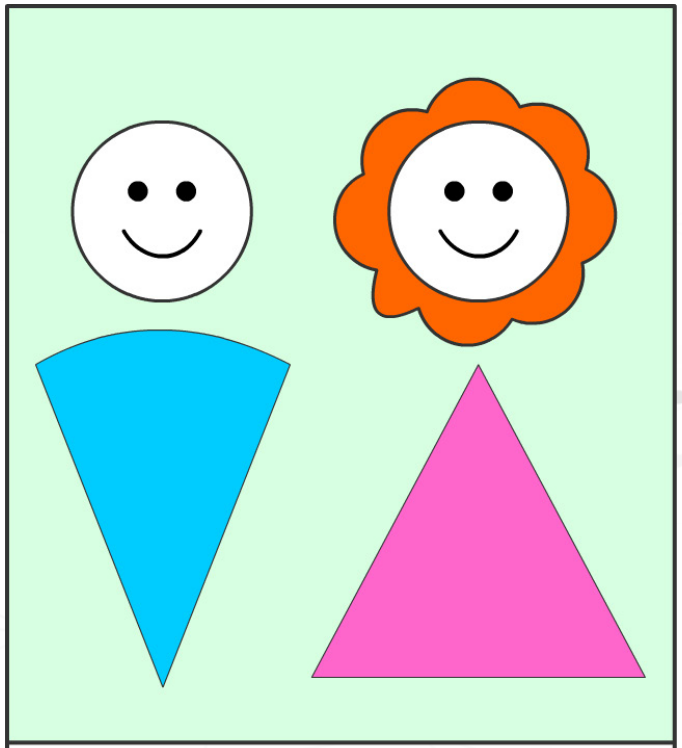
.....



Scandals

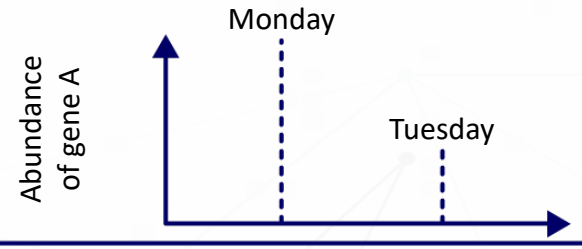
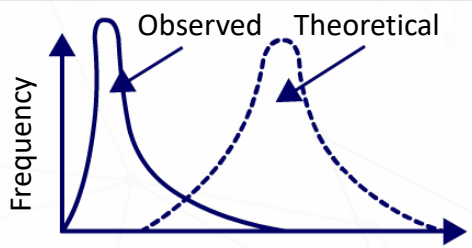
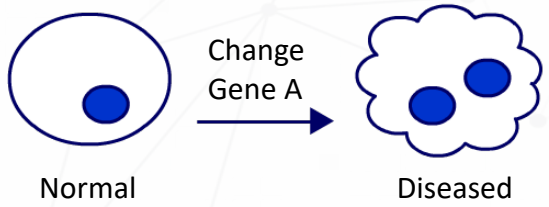
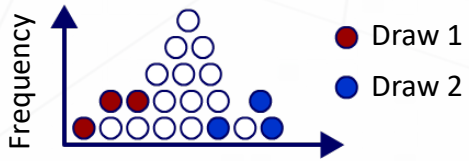
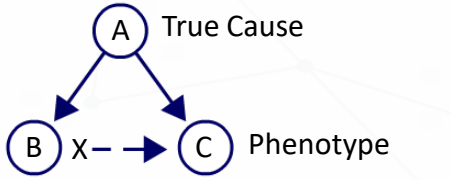


Incompatibility



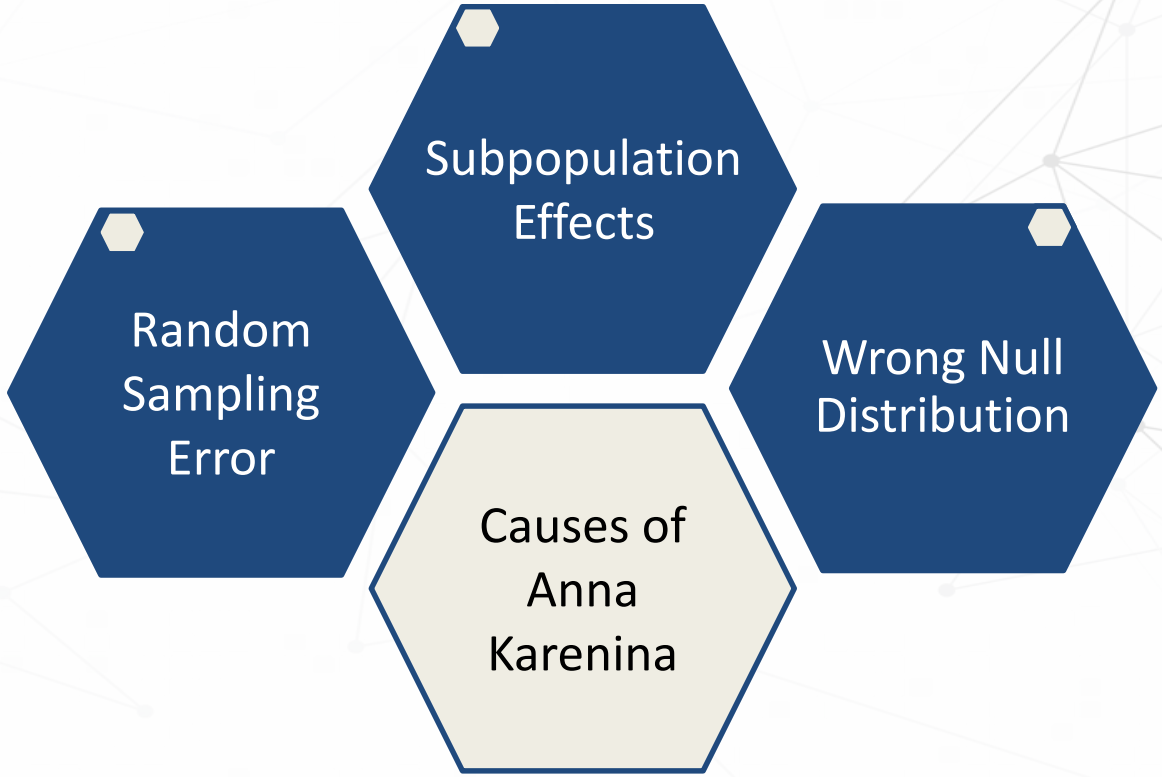
Happiness requires positive fulfillment of all possible categories. Failure in any leads to unhappiness.

How this translates to biology

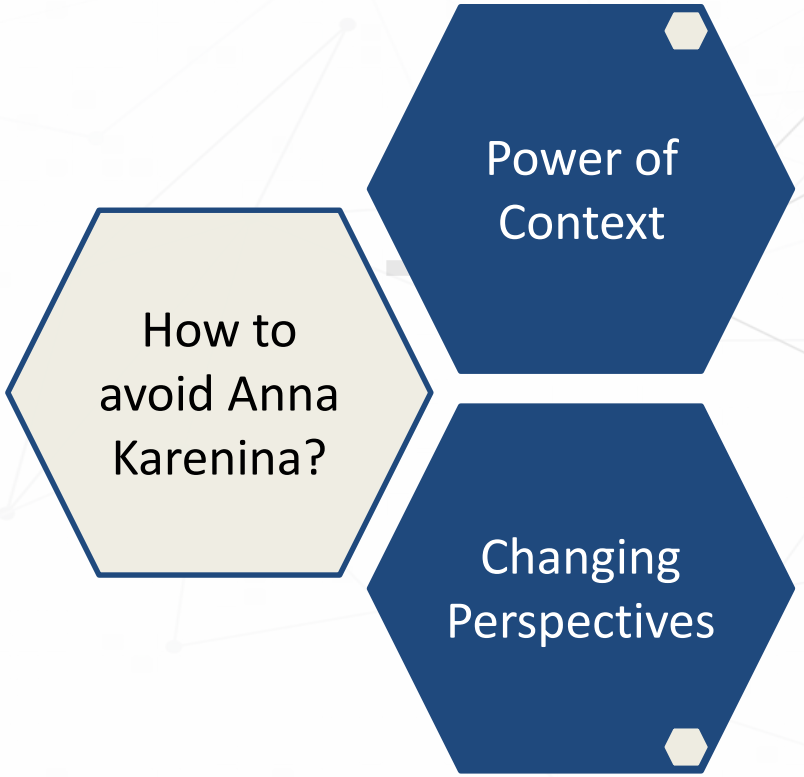
False Dichotomy **Null; Gene does not cause disease** **Alternative: Gene causes disease**	
Wrong Test Construction	Batch Effect
	
Wrong Null Distribution	**Gene is Relevant**
	 B is reported but is merely correlated to A
Chance Association	Non-causal Association

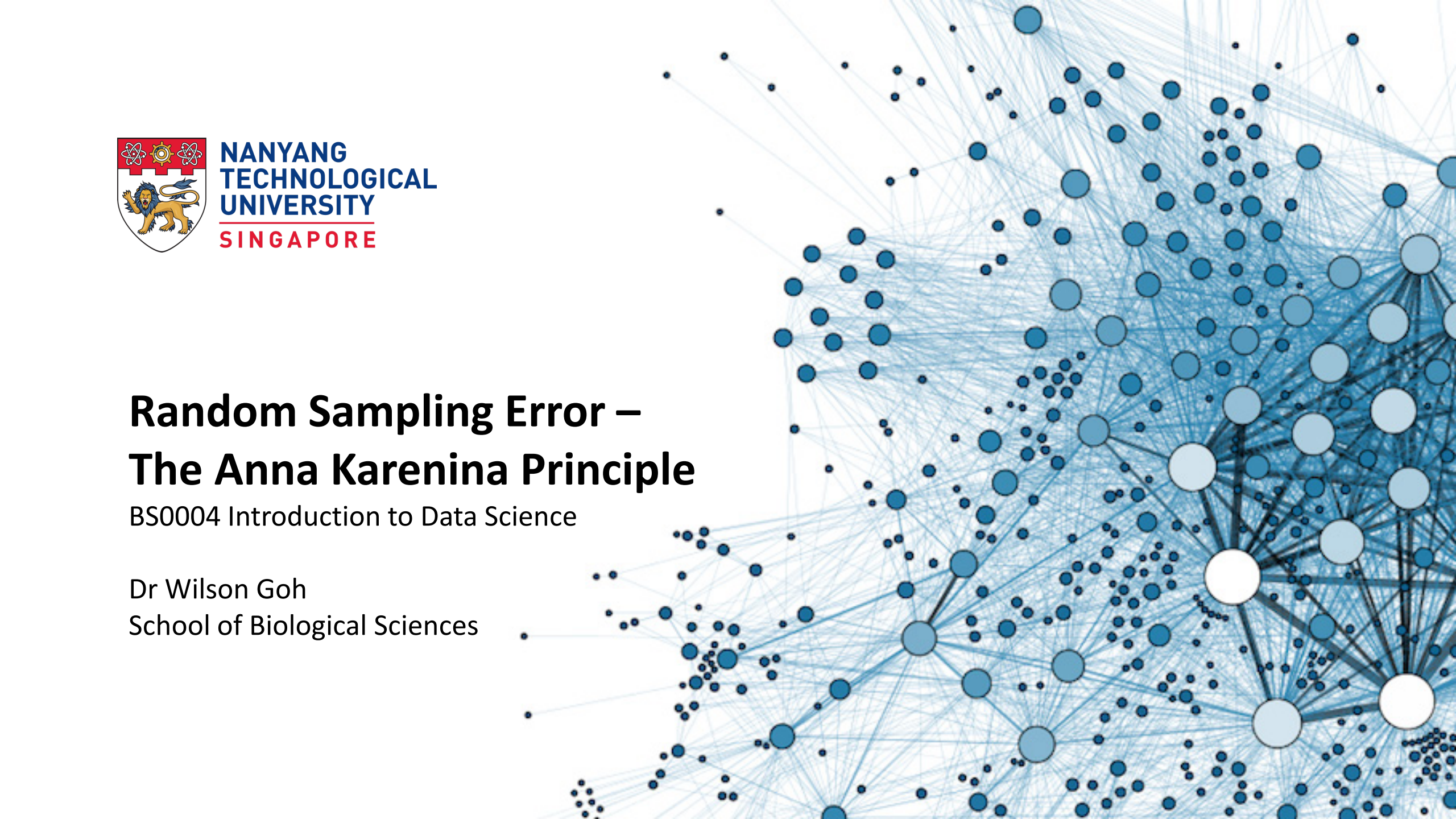
Only **1** of the causes for null hypothesis rejection is the one we want.

Causes of Anna Karenina



Avoid Anna Karenina





NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Random Sampling Error – The Anna Karenina Principle

BS0004 Introduction to Data Science

Dr Wilson Goh
School of Biological Sciences

Random Sampling Error

Consider a gene rs123 with two alleles, A and G.

Original Null: rs123 alleles are identically distributed in the two populations.

Original Alternative: rs123 alleles are non-identically distributed in the two populations.

rs123 chi-square p-value = 4.78E-21

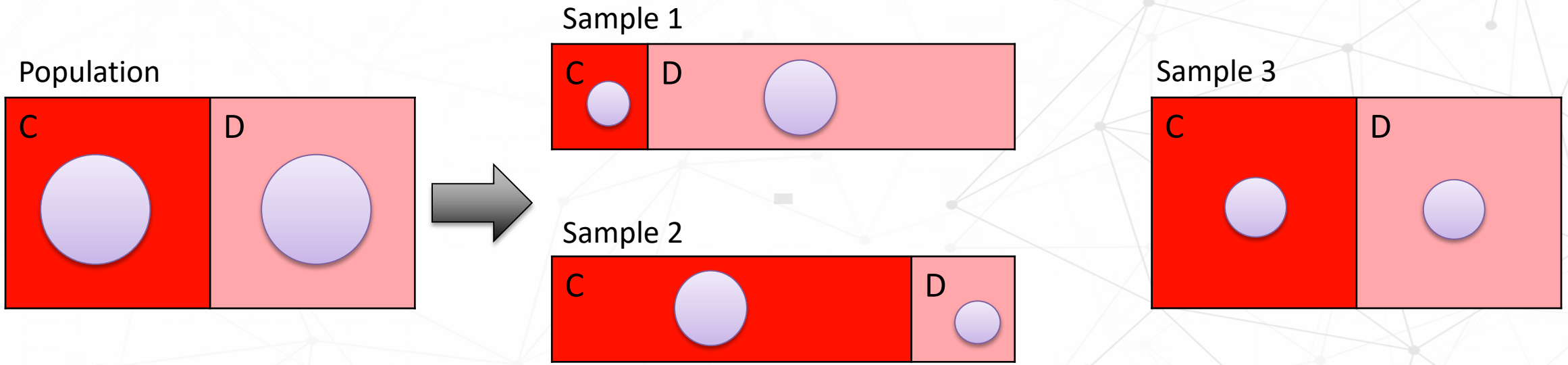
Genotypes	**Controls [n(%)]**	**Disease [n(%)]**
AA	1 (0.9%)	0 (0%)
AG	38 (35.2%)	79 (97.5%)
GG	69 (63.9%)	2 (2.5%)

Is this significant?

But is it true significance?

Sample from a Population

Consider what happens when we sample from a population:



If the sample does not reflect the population, then the sampling bias will cause the statistical test to be significant.

So what's happening here?

So…what can we do?

Let’s try rewriting the null hypothesis statements:

Refined Null: Distributions of rs123 alleles in the samples are reflective of their respective reference populations AND rs123 alleles are identically distributed in the two populations.

Refined Alternative: Distributions of rs123 alleles in the sample are different from their reference populations OR rs123 alleles are non-identically distributed in the two populations.

In other words, if the first statement is satisfied, then rejection of the null must be because rs123 are non-identically distributed in the two populations.

But problem is, how do we know we have sampling bias?

Inferring the Population without Touching the Population

Can we measure all people on earth? Too expensive? Impossible. So does it mean I cannot confirm I have sampling bias?

Let's look at our table again:

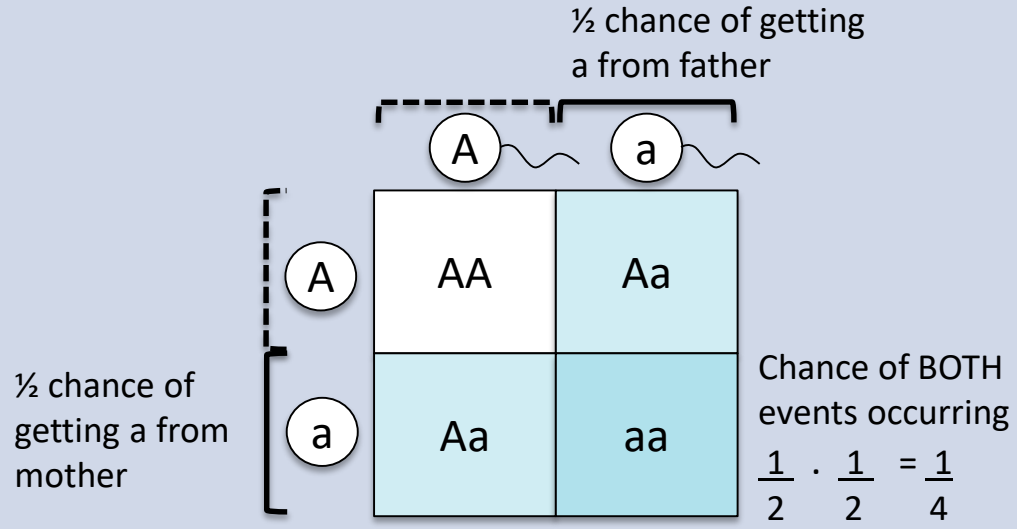
rs123 chi-square p-value = 4.78E-21

Genotypes	Controls [n(%)]	Disease [n(%)]	
			N= 189
AA	1 (0.9%)	0 (0%)	1/189 (<1%)
AG	38 (35.2%)	79 (97.5%)	117/189 (62%)
GG	69 (63.9%)	2 (2.5%)	71/189 (37.9%)

So what can we do with what we know?

So what can we do with what we know?

Let's use what we know about simple human genetics.



Basic rule of human genetics

Let's calculate backwards.

- 62% of our samples are AG.
- So let's say, the probability of a mother and a father both being AG is 0.62 * 0.62 = 0.38.
- And the probability of them having a child that is AA is 0.25 * 0.62 * 0.62 = 0.09 (9%).

Inferring Sampling Bias without the Population

Let's look at our table again.

rs123 chi-square p-value = 4.78E-21

	Genotypes	Controls [n(%)]	Disease [n(%)]	
<1% AA	AA	1 (0.9%)	0 (0%)	1/189
62% AG	AG	38 (35.2%)	79 (97.5%)	117/189
38% GG	GG	69 (63.9%)	2 (2.5%)	71/189

N= 189

We expect 9%. But our data says AA is only < 1%. So unless AA is lethal, our samples do not reflect expectation. Therefore, we conclude that our samples are biased. And therefore, if we reject the null, we need to be careful of Anna Karenina.

Correlation and Causality

Refined H0

- Distributions of rs123 alleles in the two samples are identical to the two populations; and
- rs123 alleles are identically distributed in the two populations.

Refined H1

- Distributions of rs123 alleles in the two samples are different from the two populations; or
- rs123 alleles are differently distributed in the two populations.

Suppose distributions of rs123 alleles in the samples are identical to the populations and the test is significant. Can we say rs123 mutation causes the disease?

Three types of Reasoning

1

Induction

Socrates is a man.

Socrates is mortal.

All men are mortal (provided there is no counter example).

2

Abduction

All men are mortal.

Socrates is mortal.

Socrates is a man (provided there is no other explanation of Socrates' mortality).

3

Deduction

All men are mortal.

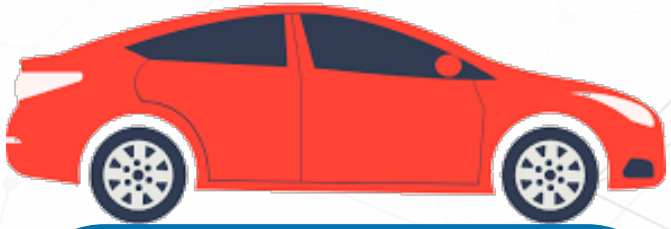
Socrates is a man.

Socrates is mortal.

Which of the following are examples of each reasoning type?

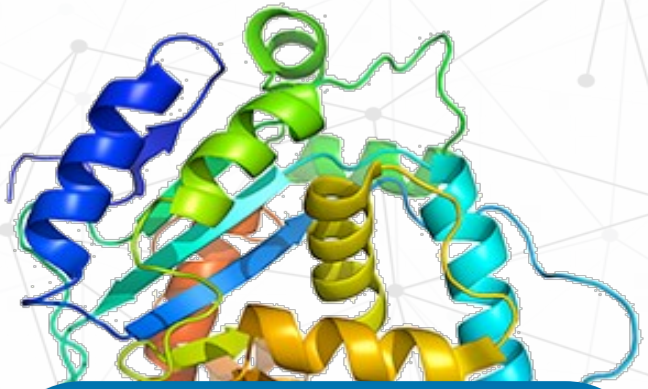
Induction

Gene A performs function X; Gene B is sequentially similar to Gene A. Therefore, Gene B also performs function X.



Abduction

An apple is red, a car is red, so therefore a car is red.



Deduction

A class of proteins, C, performs function X. Protein Z is a member of C, so C must therefore perform function X.

Abduction in Action

Hypothesis: If rs123 mutation causes disease, the statistical test is significant.

Observation: Statistical test is significant

Conclusion by abduction: rs123 mutation causes disease and **provided there is no other explanation for the test to be significant.**

That is, as long as this observation cannot be refuted, it may become a rule.

Group					
SNP	Genotypes	Controls [n(%)]	Cases [n(%)]	X^2	P-value
rs123	AA	1 0.9%	0 0.0%		4.78E-21[b]
	AG	38 35.2%	79 97.5%		
	GG	69 63.9%	2 2.5%		
SNP: Single Nucleotide Polymorphism					

Correlation and Causality

Hypothesis: If rs123 mutation causes disease, the statistical test is significant.

Observation: Statistical test is significant

Conclusion by abduction: rs123 mutation causes disease and **provided there is no other explanation for the test to be significant.**

How to incorporate “provided there is no other explanation” into the analysis?

Group					
SNP	Genotypes	Controls [n(%)]	Cases [n(%)]	X^2	P-value
rs123	AA	1 0.9%	0 0.0%		4.78E-21[b]
	AG	38 35.2%	79 97.5%		
	GG	69 63.9%	2 2.5%		
SNP: Single Nucleotide Polymorphism					

How about this?

H0 - In some stratification:

- Distributions of rs123 alleles in the two samples are identical to the two populations; and
- rs123 alleles are identically distributed in the two populations.

H1 - In every stratification:

- Distributions of rs123 alleles in the two samples are different from the two populations; or
- rs123 alleles are differently distributed in the two populations.

This basically says there is “no exception”. It does not say there is “no other explanation”.

How about this?

- Choose a sample of Cases and a sample of Controls such that for each stratification p1/p2, the distribution of p1/p2 in Cases is same as the distribution of p1/p2 in Controls i.e. equalise/ control for other factors.
- Then test:

H0: X's alleles are identically distributed in the two samples.

H1: X's alleles are differently distributed in the two samples.

- This makes the significance of the test independent of other explanations.
- It still does not say "no other explanation".

Or this?

Look for another gene X such that:

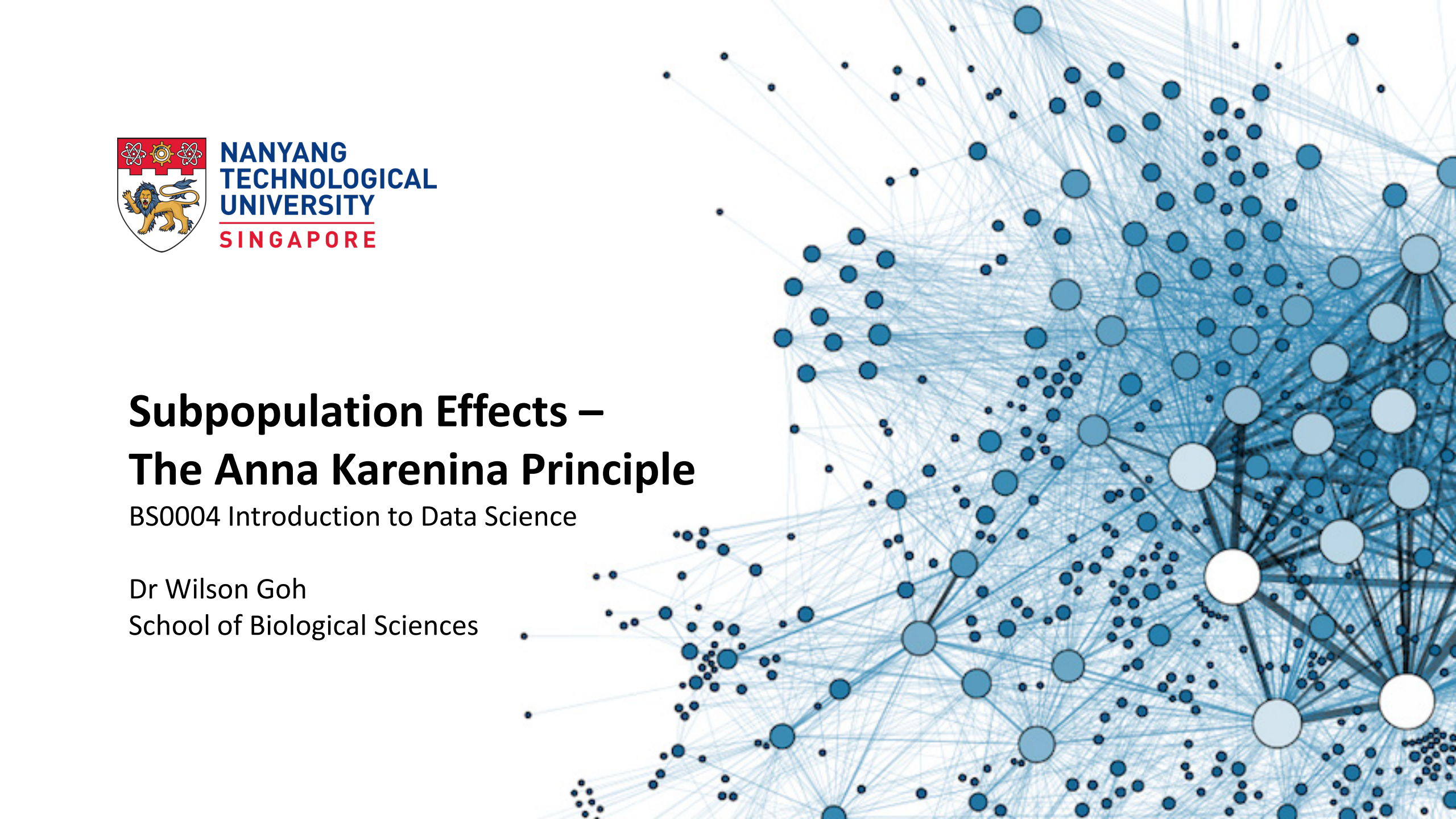
H0:

- Distributions of X's alleles in the two samples are identical to the two populations; and
- X's alleles are identically distributed in the two populations.

H1:

- Distributions of X's alleles in the two samples are different from the two populations; or
- X's alleles are differently distributed in the two populations.

- When the red part of H1 is false, this implies gene X mutation is an alternative explanation for the significance of rs123 mutation and thus the disease.
- In this case, rs123 is clearly not a cause. But has to be considered in light of its relationship with X.



NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Subpopulation Effects – The Anna Karenina Principle

BS0004 Introduction to Data Science

Dr Wilson Goh
School of Biological Sciences

A Seemingly Obvious Conclusion

Overall		
	A	B
Lived	60	65
Died	100	165
	0.60	0.39

Women		
	A	B
Lived	40	15
Died	20	5
	2	3

Men		
	A	B
Lived	20	50
Died	80	160
	0.25	0.31

Looks like treatment A is better.

But splitting the data by gender results in a reversal. Looks like treatment B is better.

In this case, the trouble arises because the proportion of men and women are not equalised the two samples.

Careless Null Hypothesis

“Effective” H0: Treatments are identically distributed in the two samples.

Assumption: All other factors are equalised in the two samples.

Apparent H0: Treatments are identically distributed in the two populations.

Apparent H1: Treatments are differently distributed in the two populations.

Refined Null Hypothesis

Refined H0:	Refined H1:
All other factors are equalised in the two samples; and	Some factors are not equalised in the two samples; or
Treatments are identically distributed in the two samples.	Treatments are differently distributed in the two populations.

Any other thing missing?

A/B sample not equalised in other attributes, viz. sex

Overall		
	A	B
Lived	60	65
Died	100	165

Women		
	A	B
Lived	40	15
Died	20	5

Men		
	A	B
Lived	20	50
Died	80	160

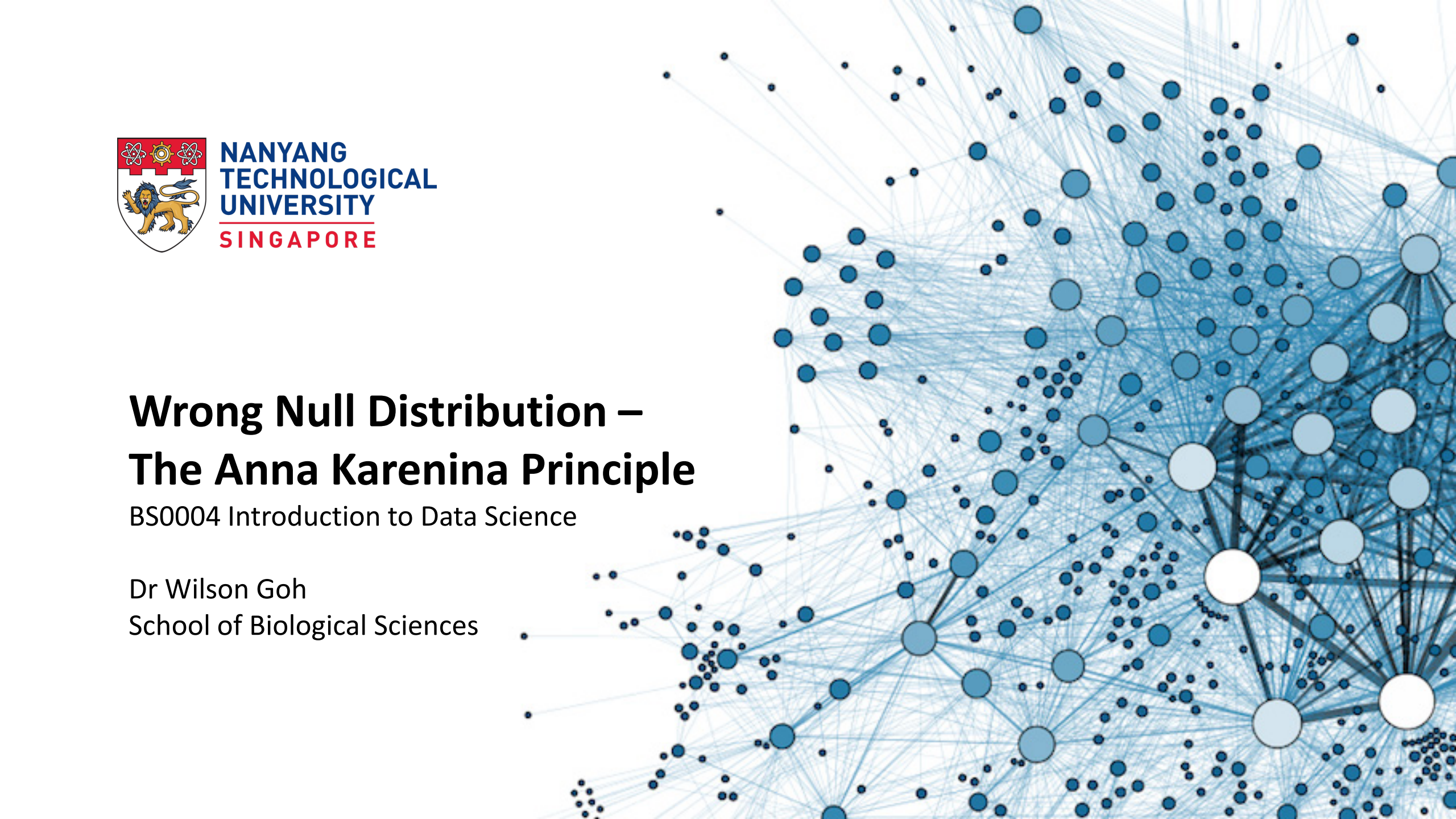
Taking A

- Men = 100 (63%)
- Women = 60 (37%)

Taking B

- Men = 210 (91%)
- Women = 20 (9%)

The differences in proportion in A and B between the two genders is contributing to false effects. The simplest way to deal with this is to simply ensure that the gender proportion is the same in both A and B.



NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

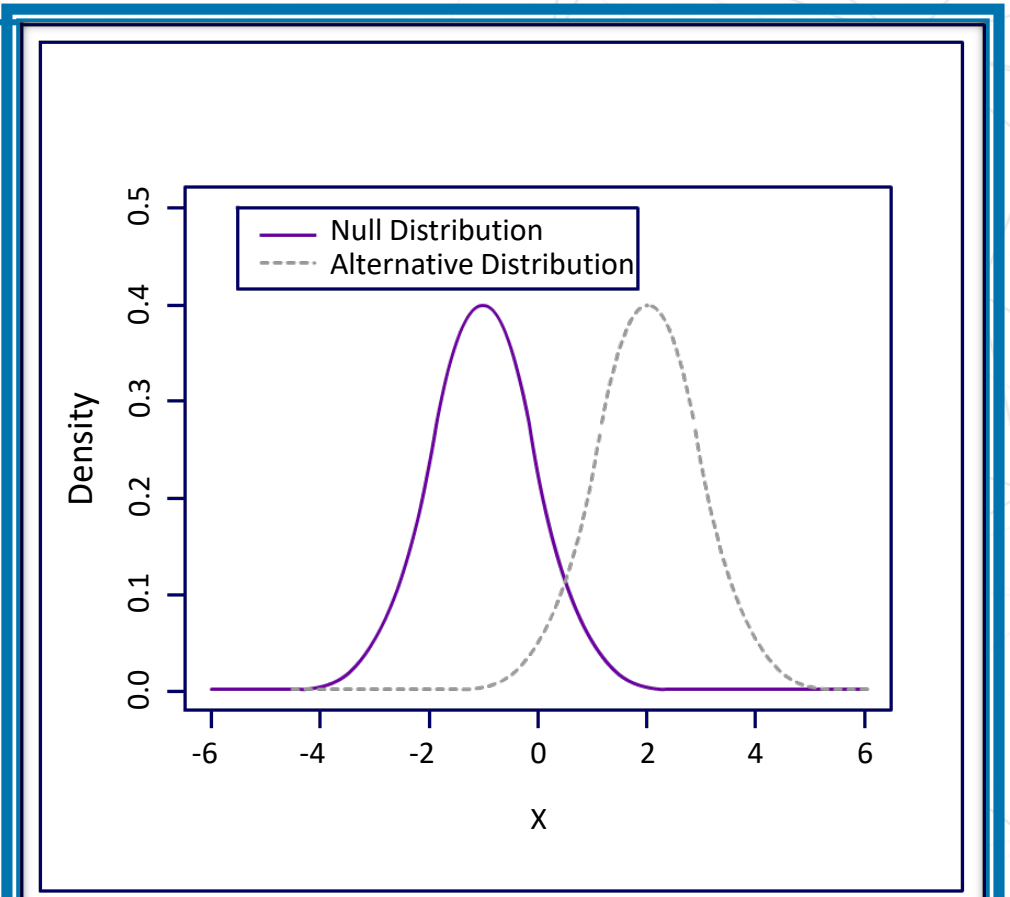
Wrong Null Distribution – The Anna Karenina Principle

BS0004 Introduction to Data Science

Dr Wilson Goh
School of Biological Sciences

Null Distribution

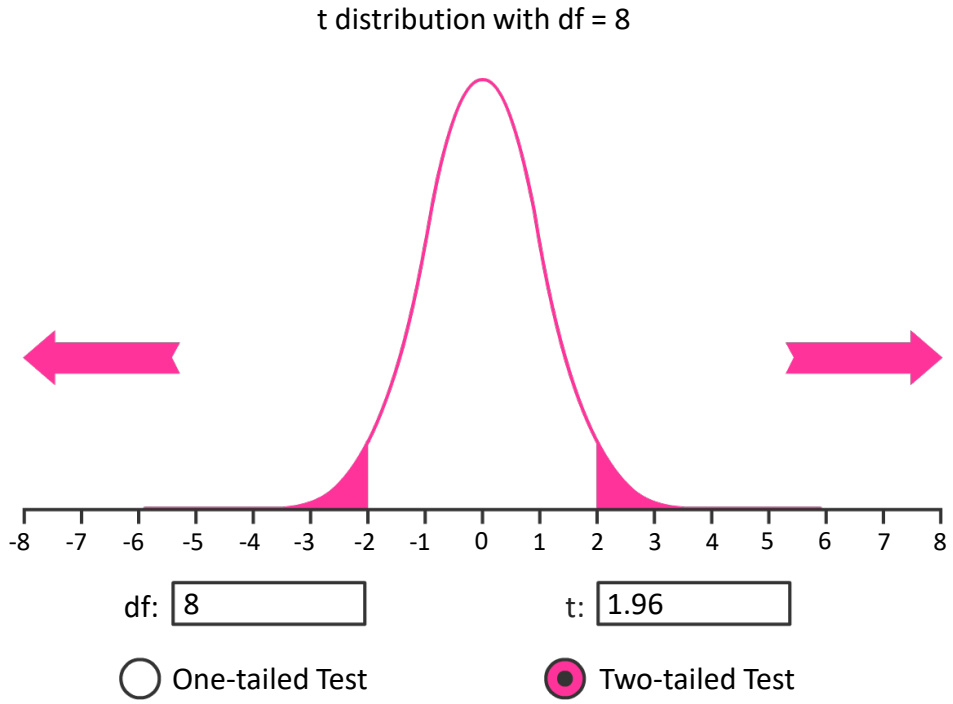
In statistical hypothesis testing, the **null distribution** is the probability **distribution** of the test statistic when the **null** hypothesis is true. For example, in an F-test, the **null distribution** is an F-**distribution**.



Null and Alternative Distribution

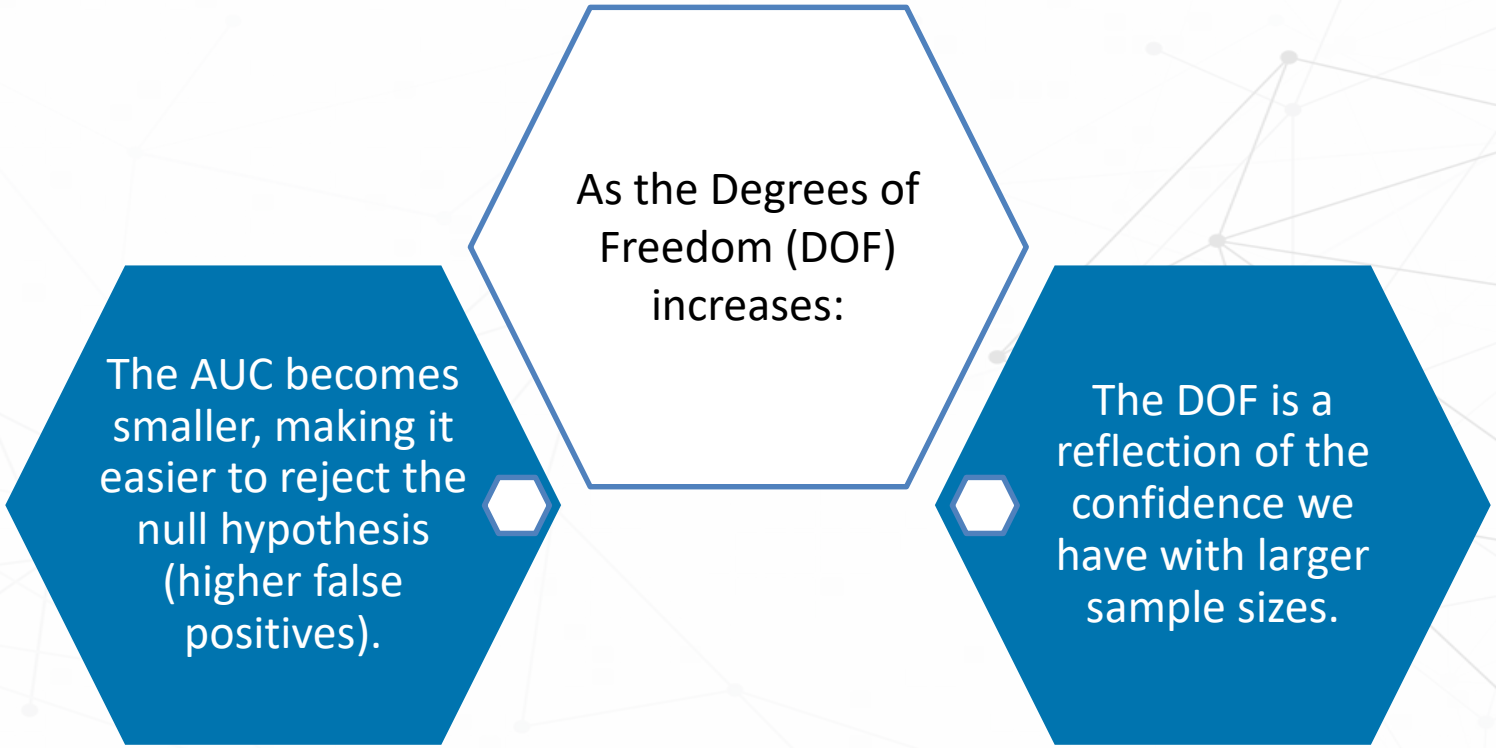
Appropriateness of the Null Distribution is Important

df	$t_{0.1}$	$t_{0.05}$	$t_{0.025}$	$t_{0.01}$	$t_{0.005}$
2	1.89	2.92	4.3	6.96	9.92
3	1.64	2.35	3.18	4.54	5.84
⋮					
10	1.37	1.81	2.23	2.76	3.17
⋮					
20	1.33	1.72	2.09	2.53	2.85
⋮					
30	1.31	1.7	2.04	2.46	2.75
⋮					
100	1.29	1.66	1.98	2.36	2.63
⋮					
400	1.28	1.65	1.97	2.34	2.59
⋮					
∞	$z_{0.1}$	$z_{0.05}$	$z_{0.025}$	$z_{0.01}$	$z_{0.005}$
	1.28	1.645	1.96	2.33	2.58



Shaded Area: 0.0857

Degrees of Freedom (DOF)



Suitable Null Distribution is Important

The **smaller sample size is, the lower the DOF**, and the flatter the t-distribution becomes, making it harder to reject the null hypothesis.

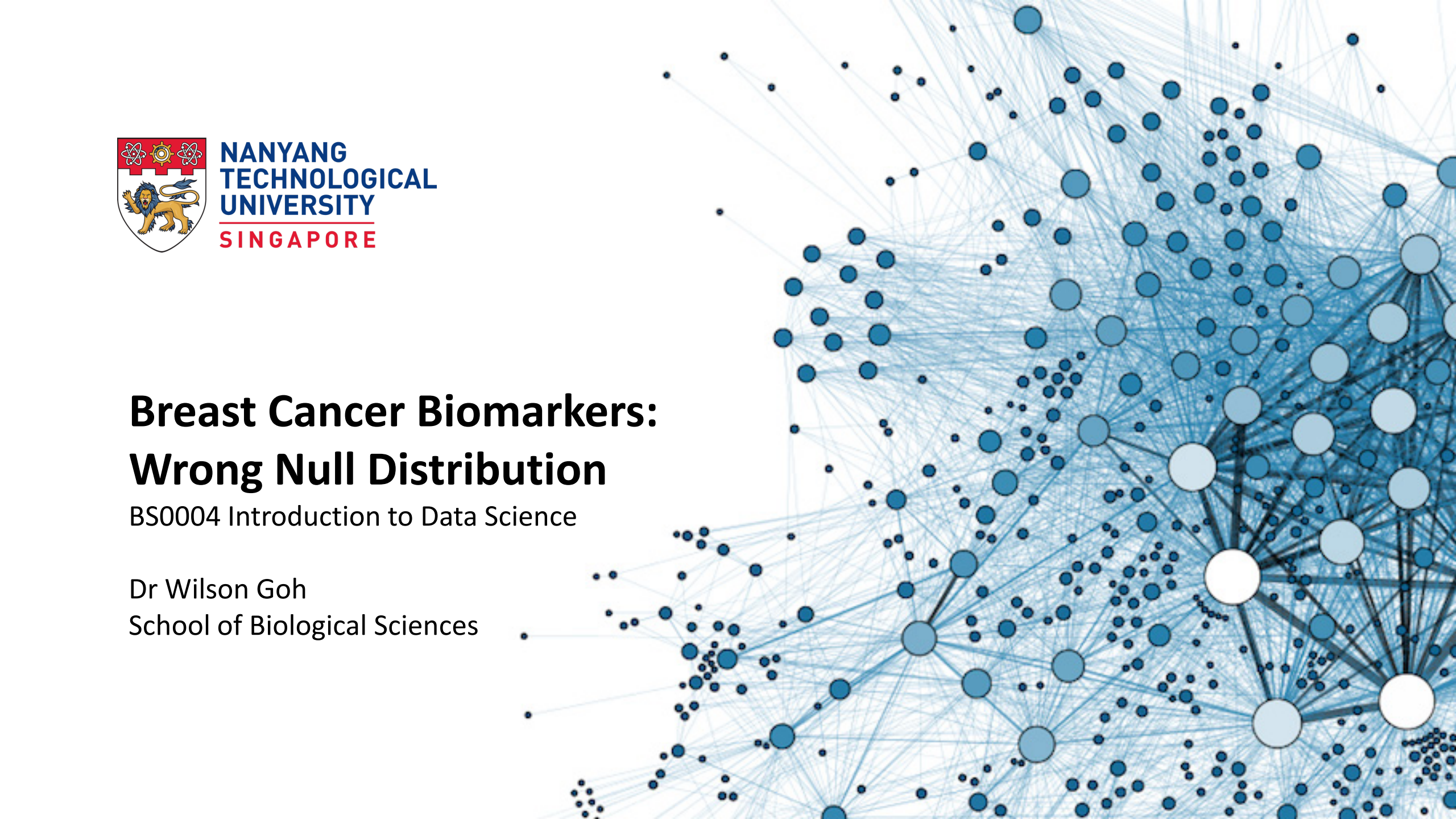
This dynamic adaption of the null distribution to small sample size is important: when sample size is small, we make less reliable estimates of population parameters from sample; a flatter t-distribution means that given this increased uncertainty, we do not reject the null hypothesis as easily.

Types of Null Distributions

Null-model fitting may be broadly divided into parametric (e.g. when the data distribution approximates a bell curve) or non-parametric (e.g. when the data distribution does not approximate a bell curve).

In both scenarios, there are extensive criteria to fulfill: just because the data is not bell-curve like, does not mean it is compatible for use with non-parametric methods (e.g. balanced design with sufficient sample size; and similar distribution shapes between both populations).

*On what basis can I claim that to be true?



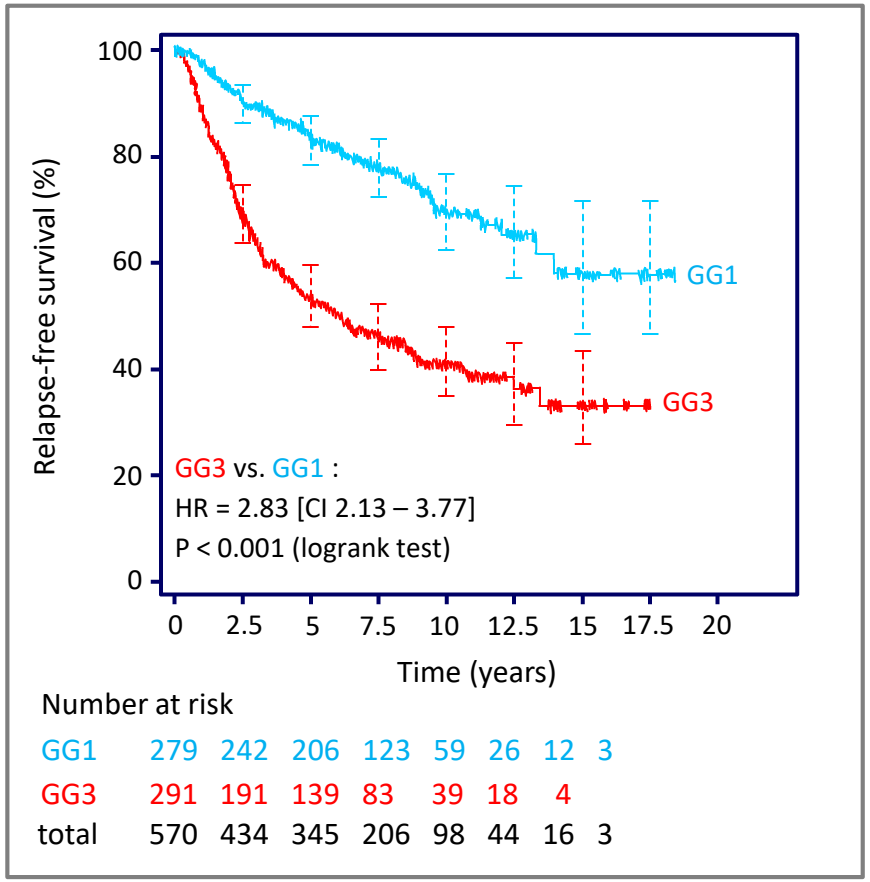
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Breast Cancer Biomarkers: Wrong Null Distribution

BS0004 Introduction to Data Science

Dr Wilson Goh
School of Biological Sciences

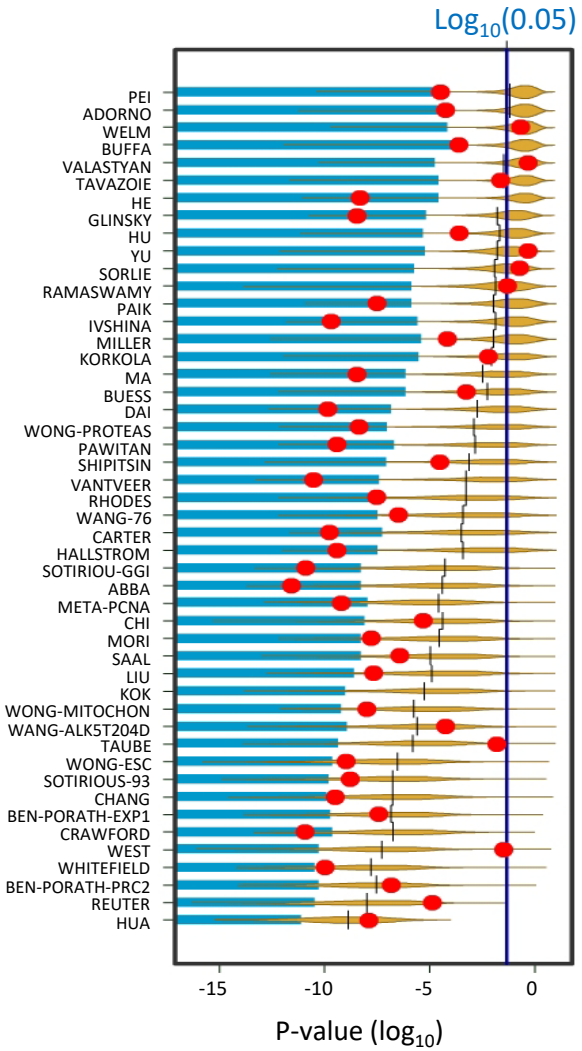
A seemingly Obvious Conclusion



A multi-gene signature is claimed as a good biomarker for breast cancer survival - Cox's survival model p-value << 0.05.

A straightforward Cox's proportional hazard analysis. Anything more/wrong?

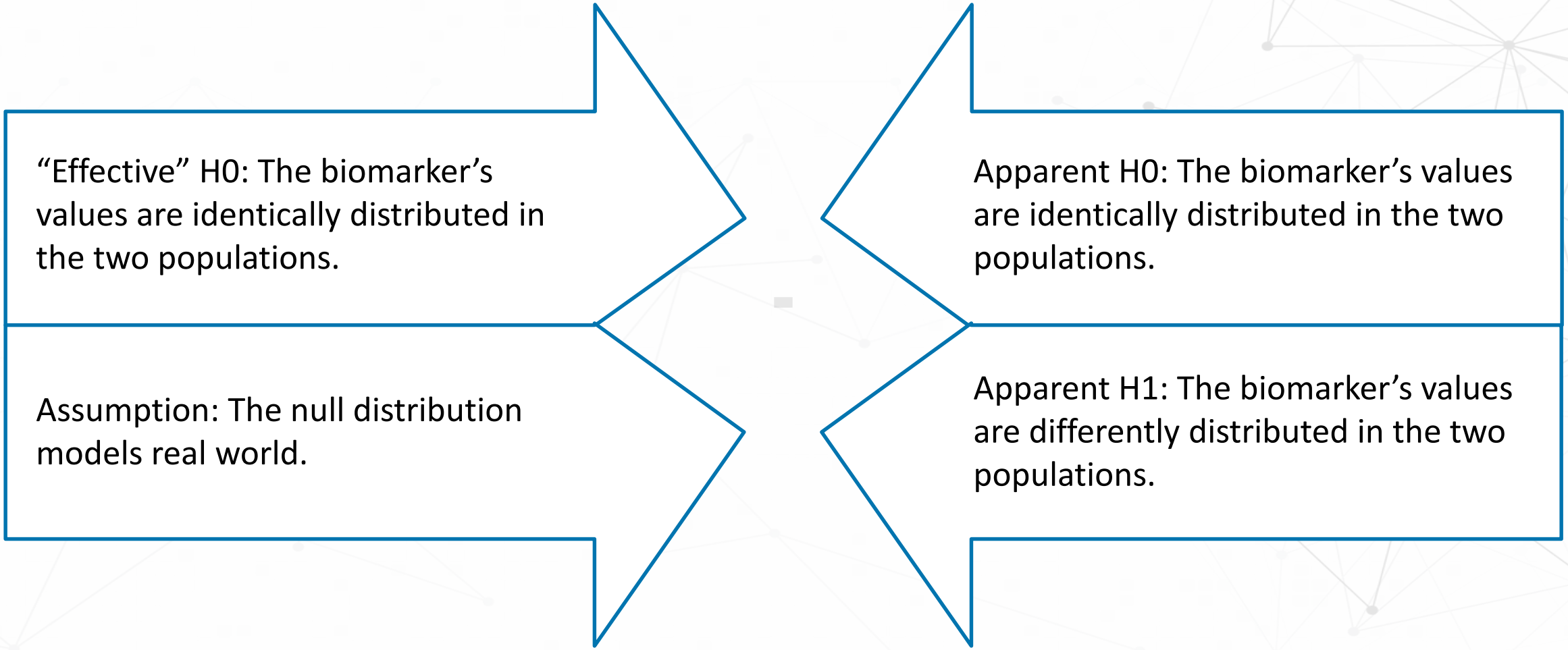
Almost all Random Signatures also have p-value < 0.05



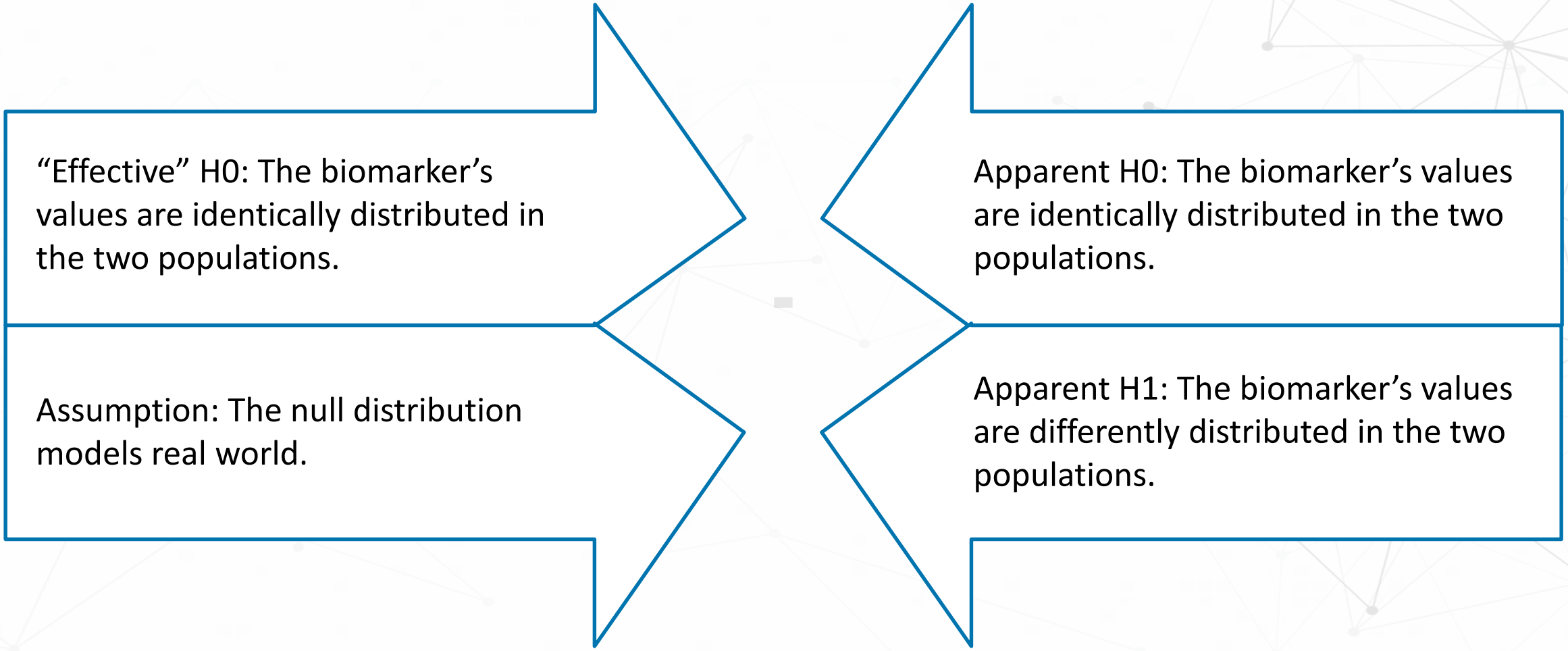
Venet et al., PLOS Comput Biol, 2011

- Theoretical null distribution used in Cox's proportion hazard analysis does not match the empirical null distribution.
- What can we do about this?

Wrong Null Distribution



Wrong Null Distribution



The apparent null / alternative hypothesis is carelessly stated. Why? How to fix this?

Refined Null Hypothesis

Refined H0:

- The biomarker's values are identically distributed in the two populations; and
- The null distribution models real world.

Refined:

- The biomarker's values are differently distributed in the two populations; or
- The null distribution does not model real world.

But how to model the null?

One option exists, in the form of **Permutation Tests** (PT) where the sampling distribution is constructed by resampling the observed data, subject to a crucial assumption of exchangeability of the samples under the null hypothesis.

That is, the reference distribution is constructed by observed data itself, and in a manner that is consistent with the null hypothesis.

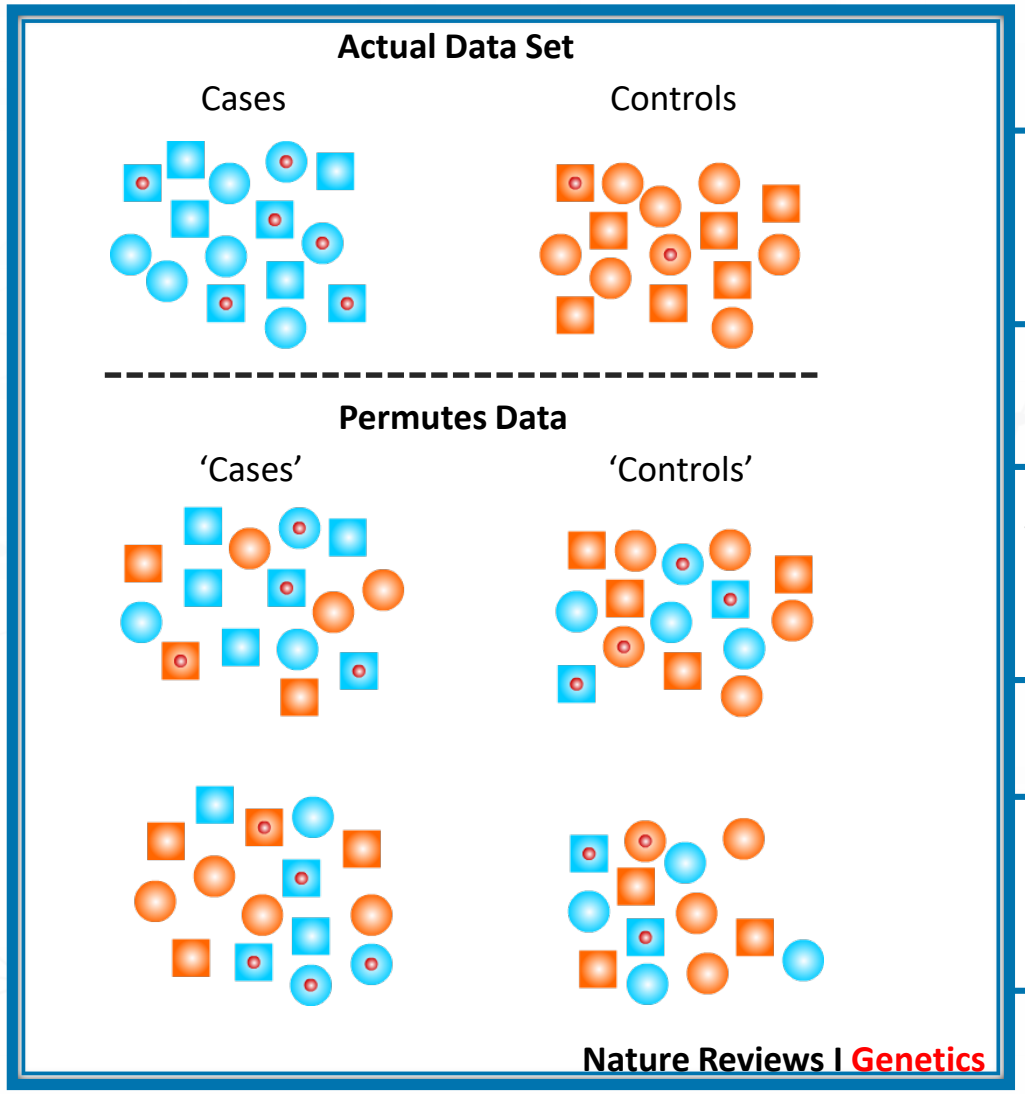
This is called the **empirical distribution** (as opposed to the null distribution, which is inferred independently and theoretically).

Note that, by construction, this empirical distribution is appropriate for the issue at hand only when the null hypothesis itself is appropriate.

Permutation Tests

- In PT, data are randomly re-assigned a class label so that an exact p-value is calculated based on the permutated data (empirical-based resampling).
- A crucial assumption that should not be overlooked when using this kind of test is the **assumption of exchangeability** of the samples under the null hypothesis.
- The null hypothesis has to permit class labels to be swapped.

Permutation Tests



Calculate test statistics of interest in actual data set.

To obtain significance of best actual test statistic compare with distribution of best permuted statistics.

Calculate same test statistics in each permuted data set and record best result for each permutation

Source: https://www.ncbi.nlm.nih.gov/pubmed/15716906

Permutation Test

Randomisation exact test is a test procedure in which data are randomly re-assigned so that an exact p-value is calculated based on the permutated data.

Original Scores of two groups

Web-based		Text-based	
Subject	**Scores**	**Subject**	**Scores**
Jody	99	Alex	87
Sandy	90	Andy	89
Barb	93	Candy	97
More subjects…	More scores…	More subjects…	More scores…

Let's look at the above example. Assume that in an experiment comparing web-based and text-based instructional methods, subjects obtained the given scores:

Permutation Test

Let’s say we do a two-sample t-test, the test returns a t-score of 1.55.

In parametric statistics, we check the t-score against the critical value in the t-distribution to determine whether the group difference is significant.

In resampling statistics, instead of checking the theoretical t-distribution, we can reframe analysis into a "what-if" question.

Maybe It may just happen that Jody, the over-achiever, takes the Web-based version by chance, and Alex, the under-achiever, takes the text-based version by chance, too. What if their positions are swapped?"

We can reframe this question by swapping the class labels (web-based and text-based). Let’s see what the new table will look like.

Permutation Test

We can do this to get all possible rearrangements of the data. This re-sample by random swapping is called “permutated data”.

Permutated Scores of two groups

Web-based		Text-based	
Subject	**Scores**	**Subject**	**Scores**
Alex	87	Jody	99
Sandy	90	Andy	89
Barb	93	Candy	97
More subjects…	More scores…	More subjects…	More scores…

Note that in permutations tests, the order don’t really matter. So they really are all about combinations! (The name is misleading).

Permutation Test

We compute the permutated data and obtains another t-value of -0.64. If we keep swapping observations across the two groups, many more t-values will be returned.

The purpose of this procedure is to artificially simulate "chance". Sometimes the t is large, but other times it is small. After exhausting every possibility, say 100, the inquirer can put these t-scores together to plot an empirical distribution curve, which is built on the empirical sample data.

When the t-score of 1.55 (observed t-score) is exceeded by permutated t-statistics 5 times out of 100 times, the researcher can conclude that the exact p-value (the probability that this difference happens by chances alone) is 0.05.

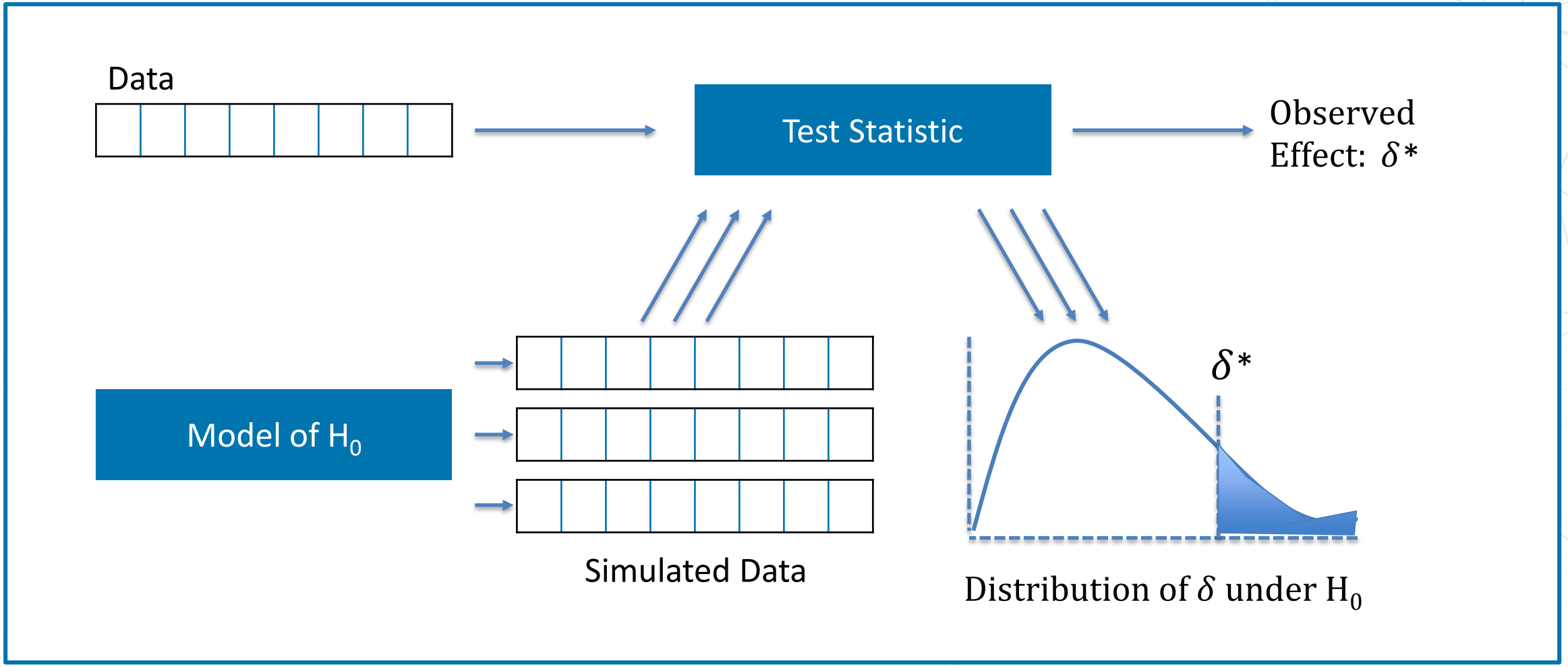
Since we compares the observed t-score with the empirical t-distribution, the latter becomes the reference set.

Other types of resampling are based on the same principle: repeated experiments within the same dataset.

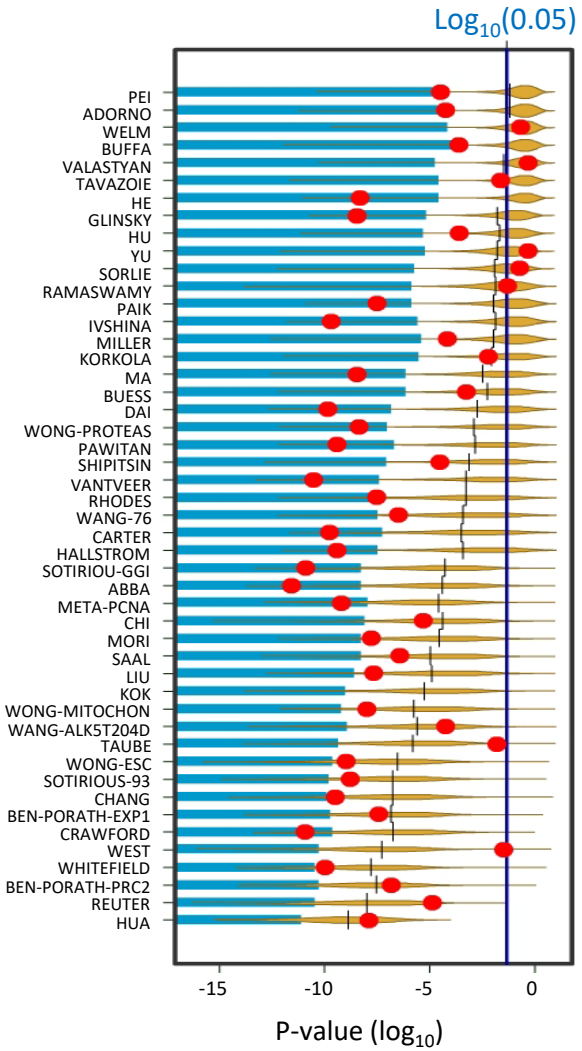
Please note that the underlying principles of this randomisation test and a parametric t-test are closely related because the two are equivalent asymptotically (we are using the same test-statistic but the reference distribution is generated via permutation).

Is this still a parametric test?

Permutation Test



Back to the Venet Example



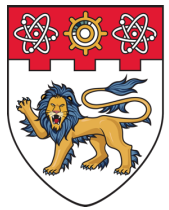
Venet et al., PLOS Comput Biol, 2011

- Green lines are the 5% most significant random signatures.
- Define away the problem… is that a valid solution?

Exchangeability Problem

Recall earlier we said that the class labels must be exchangeable under the null hypothesis?

Obviously H_0' is not implied by H_0; i.e. Venet et al.'s null samples are invalid null samples for generating a null distribution for analysis under H_0. To generate null samples that are exchangeable with the observed sample under H_0, we need to do the equivalent of class-label permutations. The class label in this case is the survival period of the subjects. Each null sample is formed by permuting the survival period of the subjects in the original dataset. We repeat this many times to get many null samples (each null sample is a set of subjects with permuted survival periods). The signature is fixed, but its score computed for each null sample provides the null distribution.



NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Summary

BS0004 Introduction to Data Science

Dr Wilson Goh
School of Biological Sciences

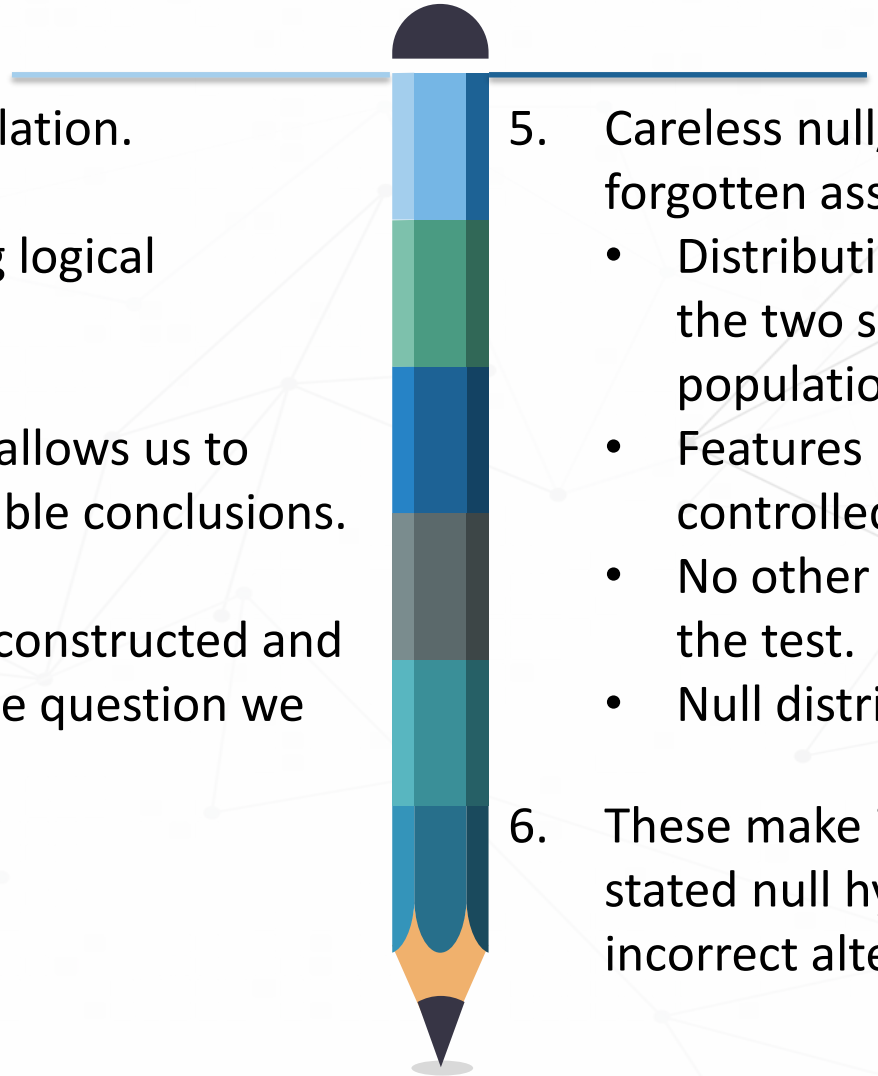
Key Takeaways from this Topic

1. Statistics is only simple calculation.

2. Using statistics without using logical reasoning is dangerous.

3. Statistics + logical reasoning allows us to arrive at much more reasonable conclusions.

4. Any statistical test can be deconstructed and reconstructed to better fit the question we want to answer.



5. Careless null/ alternative hypothesis due to forgotten assumptions:
   - Distributions of the feature of interest in the two samples are identical to the two populations.
   - Features not of interest are equalised/ controlled for in the two samples.
   - No other explanation for significance of the test.
   - Null distribution models the real world.

6. These make it easy to reject the carelessly stated null hypothesis and accept an incorrect alternative hypothesis.