Refresher on Statistics

BS3033 Data Science for Biologists

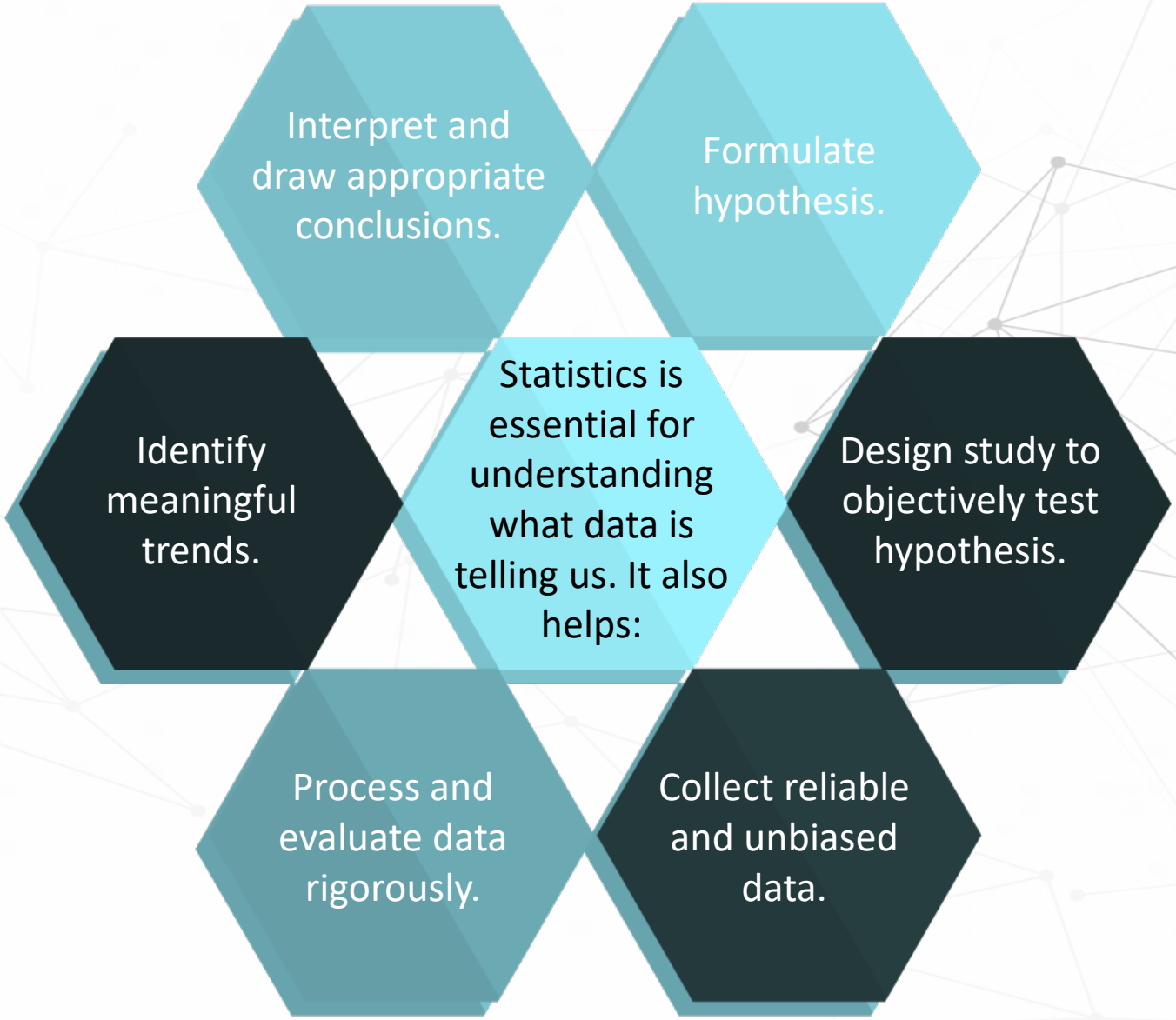
Dr Wilson Goh
School of Biological Sciences

Learning Objectives

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

- Describe the differences between inferential and descriptive statistics.
- Describe the various types of data/variables.
- Describe and know when to use the various measures of centrality and dispersion.
- Describe the two ways of estimating population values.
- Describe the steps of hypothesis testing.
- Distinguish between one-tailed and two tailed tests.
- Distinguish type I and II errors.
- Distinguish the mechanics of the paired and unpaired t-test.
- Describe regression and correlation, and their relationship.

Why (bio)statistics?



NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Elements of Statistics

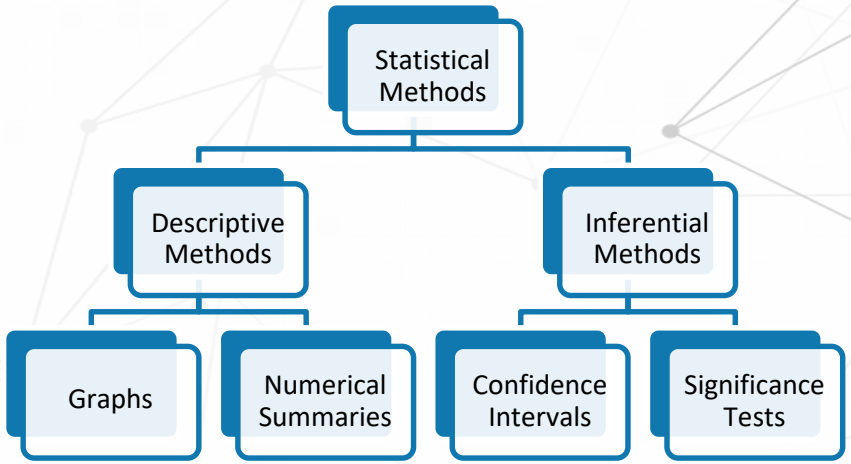
BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

Type of Statistical Methods

Descriptive:

- Summarising existing set of data
- Examples: Mean, Median, Standard Deviation, Coefficient of Variation



Inferential:

- Deducing population properties from existing sample data
- Examples: Hypothesis Testing, Central Limit Theorem, Confidence Interval

Descriptive Statistics

Descriptive statistical methods are used to make sense of the data.

Raw data have to be processed and summarised before one can make sense of data.

Summary can take the form of:

- Numerical Indices (Arithmetic Mean, Median, Standard Deviation, Coefficient of Variation);
- Tables; and
- Graphs/ Diagrams.

Inferential Statistics

Inferential statistical methods use a sample to produce statistical inferences about a population.

It is required to take population and variation into account.

The sample may not always be a good reflection of the population.

Descriptive Statistics vs Inferential Statistics

Descriptive statistics describe, show and summarise data currently being analysed. It does not go beyond the data.

Inferential statistics estimates the true population parameter based on a summary statistics. It goes beyond the collected data.

For example, looking at the height of a class of 10 students,

- The mean height of the class is 171 cm (descriptive).
- The mean height of all the students in the university is 171 cm (inferential → using a sample data to infer about whole population).

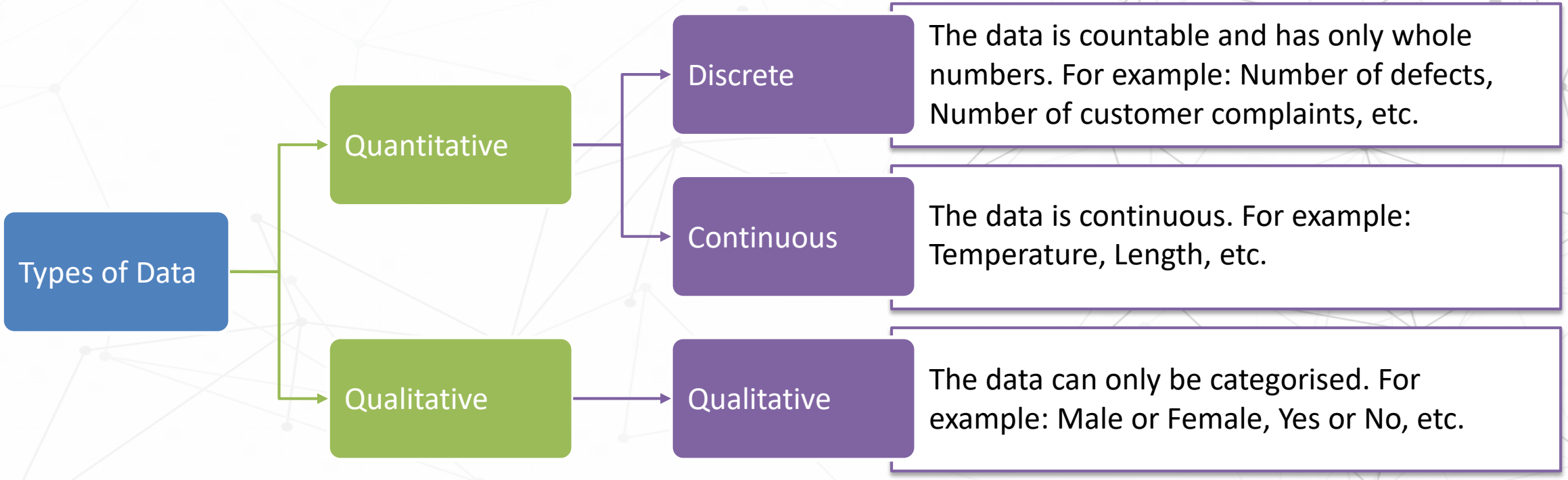
NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Descriptive Statistics

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

Types of Data/Variables



Source: http://www.whatissixsigma.net/basic-statistics/

Summarising Categorical/Qualitative Data

Patient	Gender	Status
1	Male	Alive
2	Female	Alive
3	Male	Dead
4	Female	Alive
etc.	etc.	etc.

	Dead	Alive	Total
Female	12	25	**37**
Male	23	26	**49**
Total	**35**	**51**	**86**

Proportion is a fraction and the **numerator** is a **subset** of the denominator:

- Proportion Dead = 35/86 = 0.41

Odds are fractions where the **numerator** is **not part** of the denominator:

- Odds in Favour of Death = 35/51 = 0.69

Summarising Categorical/Qualitative Data

Patient	Gender	Status
1	Male	Alive
2	Female	Alive
3	Male	Dead
4	Female	Alive
etc.	etc.	etc.

	Dead	Alive	Total
Female	12	25	**37**
Male	23	26	**49**
Total	**35**	**51**	**86**

Ratio is a comparison of two numbers:

- Ratio of Dead:Alive = 35:51

Odds Ratio is commonly used in **case-control studies**:

- Odds in Favour of Death for Females = 12/25 = 0.48;
- Odds in Favour of Death for Males = 23/26 = 0.88;
- **Odds Ratio** = 0.88/0.48 = 1.84

Summarising Quantitative Data

Methods of summarising quantitative data:

Distribution Patterns:

- Symmetrical (bell-shaped) distribution, e.g. normal distribution
- Skewed distribution
- Bimodal and multimodal distribution (i.e. multiple peaks)

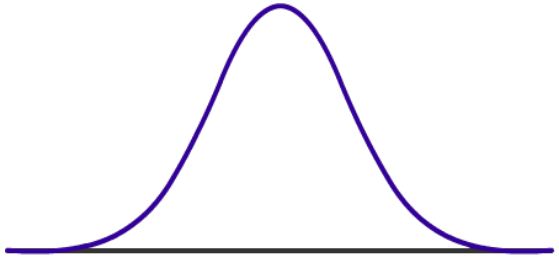
Indices of Central Tendency:

- Mean
- Median
- Quantiles
- Mode

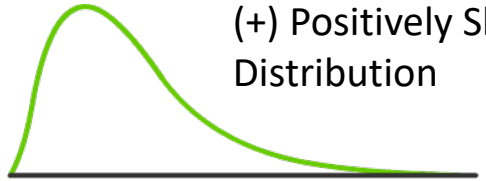
Indices of Dispersion:

- Summarises dispersion from a central value, such as the arithmetic mean
- Variance, standard deviation, coefficient of variation

Examples of Distribution Patterns

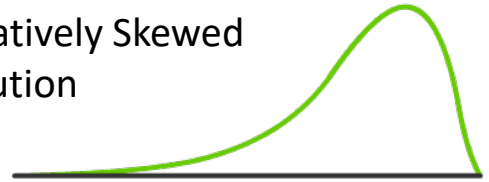


Symmetrical Distribution



(+) Positively Skewed Distribution

(-) Negatively Skewed Distribution



Skewed Distribution

Bimodal

Multimodal

Indices of Central Tendency

Arithmetic Mean is the average of a set of values. Mean is sensitive to extreme values, for example blood pressure reading.

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

x1	87	87
x2	95	95
x3	98	98
x4	101	101
x5	105.0	1050
Mean	97.2	286.2

Robust Measure of Central Tendency

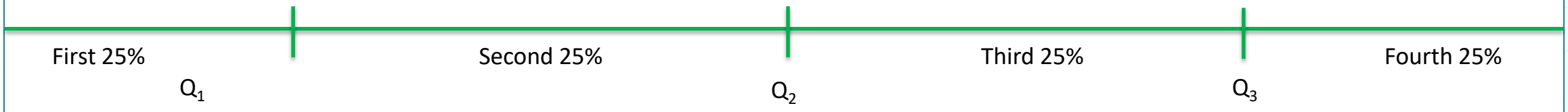
Median is the value separating the first half of a ranked sample, or a population, from the second half. Median is less sensitive to extreme values.

	x1	87	87
	x2	95	95
Median is unchanged	x3	98	98
	x4	101	101
	x5	105.0	1050

Indices of Central Tendency: Quantiles

Quantiles are formed by dividing the distribution of ordered values into equal-sized parts. Here are some types of quantiles:

- Quartiles: 4 equal parts
- Deciles: 10 equal parts
- Percentiles: 100 equal parts



Q_1: First Quartile
Q_2: Second Quartile = Median
Q_3: Third Quartile

Indices of Dispersion: Variance

Variance is the average of squares of deviation from the mean. Population variance: divide by sample size, n:

$$\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}$$

Variance of a sample is usually obtained by subtracting 1 from the denominator, n or the degree of freedom.

$$\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

Effective sample size, also called the degree of freedom.

This results in an awkward unit of measurement since the values are squared.

Indices of Dispersion: Standard Deviation

Standard Deviation (s.d.) is the square root of the variance. It provides solution to the problem of squared values of variance. Population standard deviation (σ): divide by sample size, n:

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{n}}$$

Sample standard deviation (s): divide by (n - 1), or the degrees of freedom

$$s = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n - 1}}$$

Indices of Dispersion: Standard Deviation

Standard deviations can be misleading when comparing between samples/ populations with different orders of magnitude.

Weights of Newborn Elephants (kg)	
929	853
878	939
895	972
937	841
801	826

$\overline{n}$ = 10, x = 887.1, sd = 56.50

Weights of Newborn Mice (kg)	
0.72	0.42
0.63	0.31
0.59	0.38
0.79	0.96
1.06	0.89

$\overline{n}$ = 10, x = 0.68, sd = 0.255

It is incorrect to say that Elephants show greater variation for birth-weights than Mice because of higher standard deviation.

Indices of Dispersion: Coefficient of Variance

Coefficient of Variance (cv) expresses standard deviation relative to its mean.

$$cv = \frac{s}{\overline{X}}$$

A standardised index of comparison:

Weights of Newborn Elephants (kg)	
929	853
878	939
895	972
937	841
801	826

n = 10, $\overline{x}$ = 887.1, sd = 56.50,
cv = 0.0637

Weights of Newborn Mice (kg)	
0.72	0.42
0.63	0.31
0.59	0.38
0.79	0.96
1.06	0.89

n = 10, $\overline{x}$ = 0.68, sd = 0.255,
cv = 0.375

Mice show greater birth-weight variation.

Indices of Dispersion: Coefficient of Variance

When to use cv?

When comparison groups have **very different means** (cv is suitable as it expresses the standard deviation relative to its corresponding mean).

When **different units of measurements** are involved, e.g. group 1 unit is mm, and group 2 unit is mg (cv is suitable for comparison as it is unit free).

In cases such as above, standard deviation should not be used for comparison.

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Inferential Statistics

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

Uses of Inferential Statistics

Statistical Estimation

- Estimating population parameters using sample data.
- Utilising the “Confidence Interval” approach.

Hypothesis Testing

- Checking the validity of hypotheses (on the population) by calculating the probability of the expected outcome occurring in the sample, assuming the assumption holds true.
- Utilising the “Test for Statistical Significance” approach .

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Inferential Statistics
Part 1: Statistical Estimation

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

Statistical Estimation

Two ways to estimate population values from sample values:

Point Estimation

- Using the parameter of a single sample as an estimate for the population parameter.
- Ignores the sampling error (or sample variance).

Interval Estimation or Confidence Interval (CI)

- Using a sample parameter to estimate a population parameter by defining an interval within which the population can be found in a defined probability.
- Takes into account the sampling error (or sample variance).

The main difference between the two approaches lie in their treatment of the sampling error.

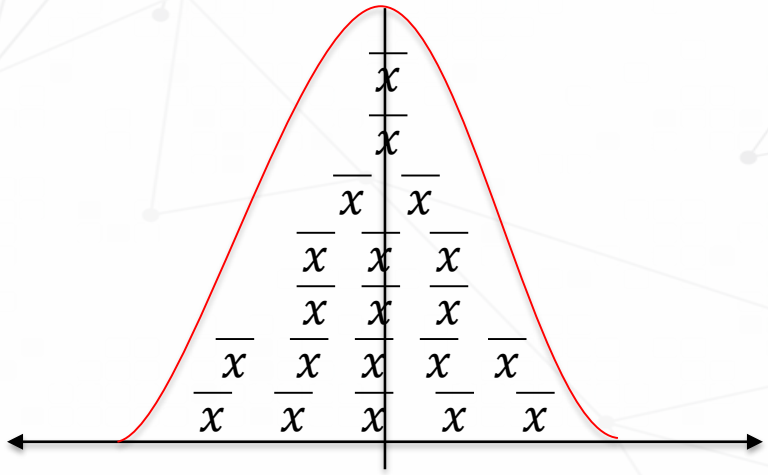
Interval Estimation – Central Limit Theorem

Central Limit Theorem suggests:

- With repeated sampling, the mean of the distribution of sample means is equal to the true population mean, μ.

Central Limit Theorem assumptions:

- Large and constant sample size
- Repeated sampling with replacement
- Samples are randomly taken
- Samples are independent of each other



Interval Estimation – Standard Error

In reality, we are usually unable to take sufficient samples to apply the Central Limit Theorem. However, the Central Limit Theorem allows us to calculate the **Standard Error** (S.E.) or the standard deviation of the sampling distribution.

$$S.E. = \frac{S}{\sqrt{n}}$$

Interval Estimation – Confidence Interval

Through the Confidence Interval, sampling error is taken into account by modifying the sample mean with the product of the Standard Error and the Z-value according to the level of confidence. Thus, at 95% level of confidence, the CI is defined as:

$$\bar{X} \pm 1.96 \frac{S}{\sqrt{n}}$$

Standard Error

In other words, there is a **95% chance that the population mean, μ, can be found within the range**.

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Inferential Statistics
Part 2: Hypothesis Testing

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

Hypothesis Testing – The Situation

Hypothesis testing revolves around two statements:

The Null Hypothesis (H0)

- The neutral statement.
- E.g. There is no difference between NUS and NTU students.

The Alternative Hypothesis (H1)

- Essentially the scientific statement you want to prove.
- E.g. There is a difference between NUS and NTU students.

Hypothesis Testing tells us whether we can reject the null hypothesis, given the data gathered.

General Steps of Hypothesis Testing

1

1.Obtain a **sample** that is representative of population.

2

1.Write down the **null** and **alternative hypotheses.**

3

1.Determine if it is a **one or two-tailed test** & select the **level of significance (α).**

4

1.Choose a test statistic based on the nature of the data collected.

- The test statistic is a numerical value that summarises the sample information
- E.g. Z-score, T-score, sum of positive/negative ranks (non-parametric)

General Steps of Hypothesis Testing

5

1.Set up decision rule.

- The decision rule is a statement that tells under what circumstances to reject the null hypothesis.
- E.g. if test statistic is smaller/bigger than level of significance, we can reject H_0.

6

1.Compute test statistic.

7

1.Make a Conclusion.

- Compare test statistic against predetermined decision rule
- Two conclusions:
  - Reject H_0 (because it is very unlikely to observe the sample data if the null hypothesis is true).
  - Do not reject H_0 (because the sample data is still likely to be observed if the null hypothesis is true).

Test of Significance: An Example

Question: A random sample of 100 male live births delivered at NUH gave a sample mean weight of 3.5kg with an SD of 0.9kg. What is the likelihood that the mean birth weight from the sample population is the same as the mean birth weight of all male live births in Singapore?

Null Hypothesis (H_0): $\mu_{pop} = \sigma_{pop}$

$\overline{X} = 3.5$ kg, SD = 0.9 kg,

$\mu_{pop} = 3.0$ kg, $\sigma_{pop} = 1.8$ kg

Test of Significance makes use of the normal distribution properties of the sampling distribution of the mean.

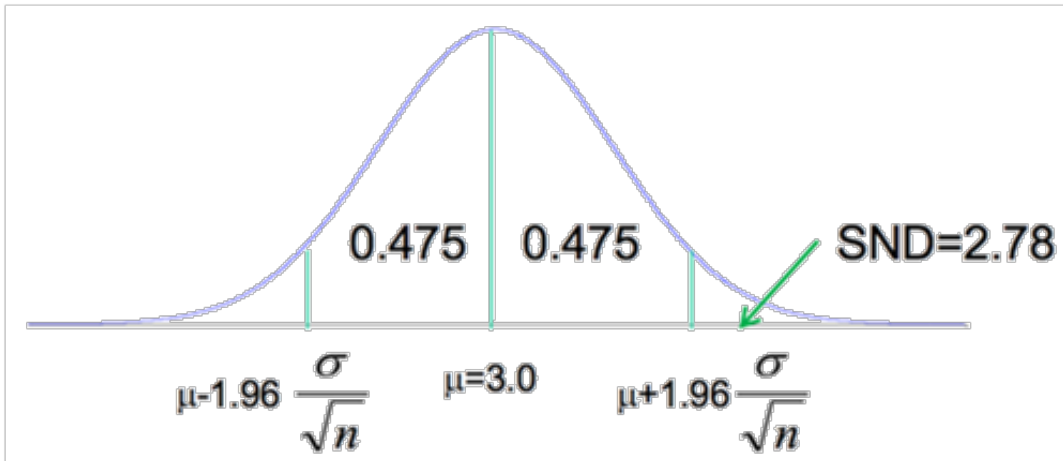
Test of Significance: z-test

Z-score can be computed by:

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Also known as Standard Normal Deviate (SND). For example:

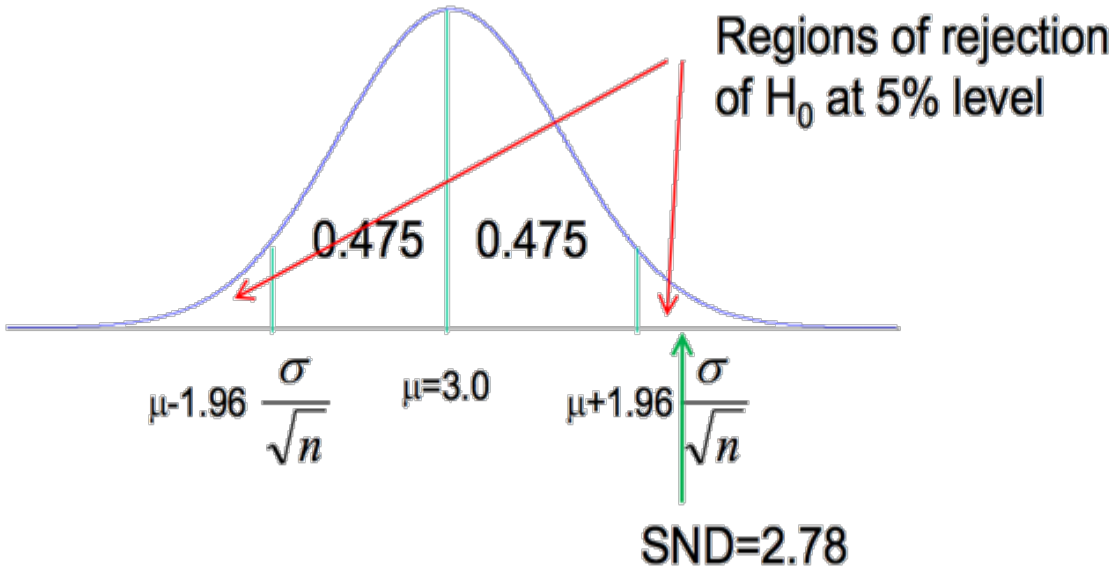
$$\frac{3.5 - 3.0}{1.8/\sqrt{100}} = 2.78$$



Test of Significance: z-test

If H_0 is **rejected**:

- There is less than 5% chance (i.e. ***very low***) that the population of male babies' weights in NUH is equivalent to the population of male babies' weights in Singapore.
- Any difference in weight between the male babies in NUH and the population of male babies in Singapore **should not** be due to chance alone.



Test of Significance: z-test

NORMAL CURVE AREAS

Entries in the body of the Table give the area under the Standard Normal Curve from 0 to z

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0596	.0636	.0675	.0714	.0753
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

SND(z) = 2.78

Pr (two-tailed)
= 2 x (0.5 -0.4973)
= 2 x 0.0027
= 0.0054

Test of Significance: z-test

From the test statistic, we are able to infer its corresponding **p-value**, which is the probability of attaining the observed, or more extreme, results if we assume the null hypothesis, H_0, to be true.

Hence, in this example, a Z-value of 2.78 gives the p-value of 0.0054.

One-tailed vs Two-tailed

One-tailed Tests:

- Used when we can anticipate the direction of difference, usually through scientific evidence.
- E.g. the glucose level in urine of diabetic vs non-diabetic patients.

Two-tailed Tests:

- Used when we do not know the direction of difference (which is usually the case).
- Difference occurs in both sides of the standard normal distribution.

One-tailed vs Two-tailed

However, given the same test statistic, the **p-value in One-tailed Tests will be half** of that in Two-tailed Tests, since the results that are more extreme than that observed can only occur in one direction.

Hence, when using One-tailed Tests, there is a **higher chance of rejecting a true null hypothesis (Type I Error)**!

Type I and Type II Errors

There are two types of wrong conclusions:

- **Type I error:** Wrongly rejecting the null hypothesis .
- **Type 2 Error:** Not rejecting the null hypothesis when you should reject.

		Actual Condition	
		Difference exists (H_0 is incorrect)	**No difference (H_0 is correct)**
Predicted Condition	**Difference exists (reject H_0)**	Correct action (power or 1-)	Type I or error
	No difference (Accept H_0)	Type II or error	Correct action

Note on Clinical Significance

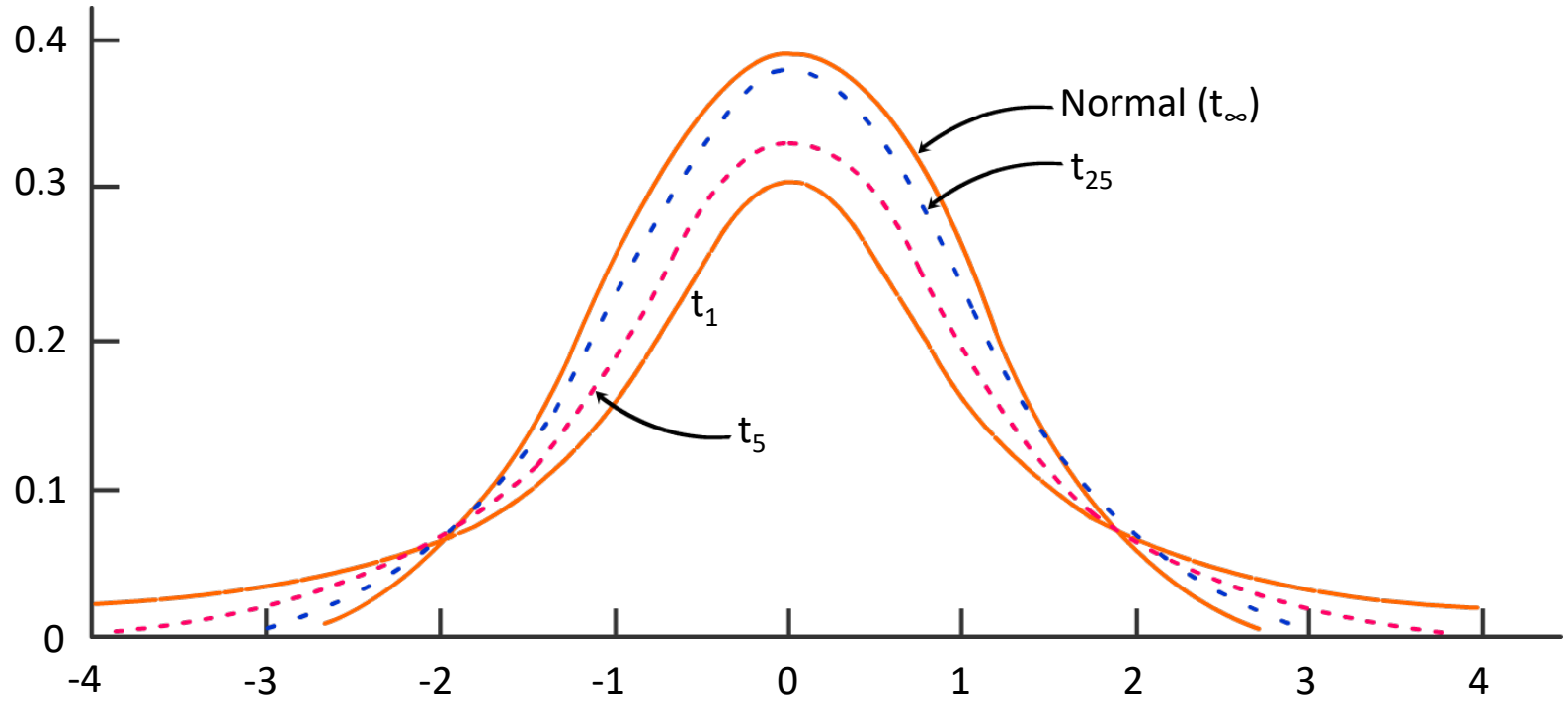
- A statistically significant result in biological research may not be clinically significant.
- E.g. when sample size of drug-testing subjects is big, the standard error of the mean becomes smaller.
- Easier to reject null hypothesis (distribution is narrower).
- Any statistical significance may not mean that the drug is effective in bringing significant therapeutic effects.

Test of Significance: t-test

- When sample size is large, use **Z-test**.
- When sample size is small, use **T-test**.
  - Assumption that sampling distribution is normally distributed is not true for small samples.
  - Smaller samples has a symmetrical distribution but with a wider spread than a normal distribution (larger standard error of mean) → t-distribution.
  - As sample size increases, spread becomes smaller → approaches normal distribution at sample size = infinity.

Family of t-distributions

Varies with sample size; larger sample size = narrower, tails are “lower”.



t-test: Unpaired

- For two independent samples that cannot be paired.
- Examples:
  - Observations on two different groups of patients (control + variable) → data are not collected from the same person.
  - Comparison of data sampled from different areas/ regions.

Blood Pb Concentrations

	Battery Workers (Occupationally Exposed)	Control (Not Occupationally Exposed)
	0.082	0.040
	0.080	0.035
	0.079	0.036
	0.069	0.039
	0.085	0.040
	0.09	0.046
mean	0.086	0.040
std dev	**0.08157**	0.03943
	0.0067047	0.0035523

t-test: Unpaired

Example question:

- For the two independent groups (control and battery workers), what is the probability that the difference in sample mean blood Pb concentrations is due to chance alone?
- Take into consideration the two sample variance/ s.d.

Blood Pb Concentrations

Battery Workers (Occupationally Exposed)	Control (Not Occupationally Exposed)
0.082	0.040
0.080	0.035
0.079	0.036
0.069	0.039
0.085	0.040
0.09	0.046
mean 0.086	0.040
std dev **0.08157**	0.03943
0.0067047	0.0035523

$$\overline{X}_1 - \overline{X}_2 = 0.08157 - 0.03943 = 0.04$$

t-test: Unpaired

Example question:

- We suspect that the battery workers have different mean blood Pb level than control group due to exposure at work → H_1
- H_0: No difference in mean blood Pb level between control and battery workers, i.e. $\mu_{control} = \mu_{battery}$

T-score for unpaired, independent samples:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{SE_{(\bar{X}1 - \bar{X}2)}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \frac{(n_1 - 1){s_1}^2 + (n_2 - 1){s_2}^2}{n_1 + n_2 - 2}}}$$

with a degree of freedom of $(n_1 + n_2 - 2)$

t-test: Unpaired

Example question:

- In this example,

$$t = \frac{0.08157 - 0.03943}{0.002868}$$
$$= 14.7 \; with \; 12 \; d.f.$$

- The p-value for this t-score test statistic is < 0.001, therefore reject null hypothesis.
- Conclusion: There is some evidence, based on the data, that battery workers have higher mean blood Pb levels than the control group.

t-test: Unpaired

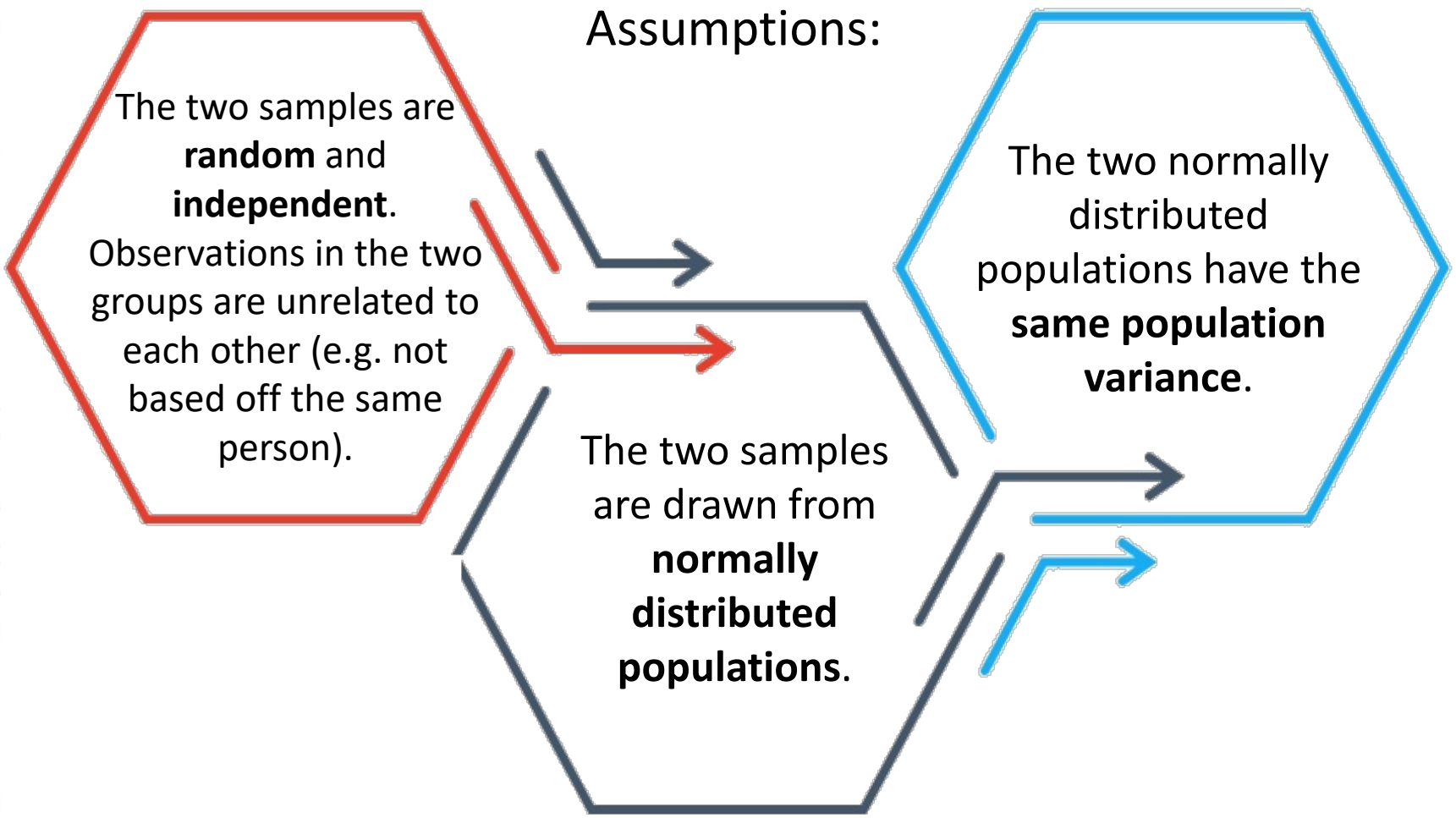
t-table

From our example:
t=14.7 with 12 d.f.

Value far exceeds 4.318, the critical value for statistical significance at the Pr=0.001 (0.1%) level when df=12 i.e. Pr < 0.001

	Probability			
df	.05	.02	.01	.001
1	12.706	31.821	63.657	636.619
2	4.303	6.965	9.925	31.598
3	3.182	4.541	5.841	12.924
4	2.776	3.747	4.604	8.610
5	2.571	3.365	4.032	6.869
6	2.447	3.143	3.707	5.959
7	2.365	2.998	3.499	5.408
8	2.306	2.896	3.355	5.041
9	2.262	2.821	3.250	4.781
10	2.228	2.764	3.169	4.587
11	2.201	2.718	3.106	4.437
12	2.179	2.681	3.055	4.318
13	2.160	2.650	3.012	4.221
14	2.145	2.624	2.977	4.140
15	2.131	2.602	2.947	4.073
16	2.120	2.583	2.921	4.015
17	2.110	2.567	2.898	3.965
18	2.101	2.552	2.878	3.922
19	2.093	2.539	2.861	3.883
25	2.060	2.485	2.787	3.725
26	2.056	2.479	2.779	3.707
27	2.052	2.473	2.771	3.690
28	2.048	2.467	2.763	3.674
29	2.045	2.462	2.756	3.659
30	2.042	2.457	2.750	3.646
40	2.021	2.423	2.704	3.551
60	2.000	2.390	2.660	3.460
120	1.980	2.358	2.617	3.373
α	1.960	2.326	2.576	3.291

t-test: Unpaired



t-test: Paired

- Used in cases when the two samples are paired.
- Examples:
  - **Before-and-after** observations on the **same subjects**.
  - Comparison of **two different methods of measurement** or two different treatments where the measurements/treatments are applied to the same subjects.

Patient	Fasting Cholesterol	Postprandial Cholesterol
1	198	202
2	192	188
3	241	238
4	229	226
5	185	174
6	303	315

Study involves 6 subjects acting as their own control (best match).

t-test: Paired

Null hypothesis: No difference in mean cholesterol levels between fasting and postprandial states ($\mu_{fasting} = \mu_{postprandial}$)

Patient	Fasting Cholesterol	Postprandial Cholesterol	Difference (d)
1	198	202	-4
2	192	188	+4
3	241	238	+3
4	229	226	+3
5	185	174	+11
6	303	315	-12

$\overline{d} = 0.833$

$s_d = 7.885$

$n = 6$

t-test: Paired

Computing the t-score:

$$t = \frac{\bar{d}}{SE_{\bar{d}}} = \frac{\bar{d}}{s_d/\sqrt{n}}$$

$$= \frac{0.833}{3.219} = 0.259$$

df: n-1 (where n is the number of pairs)

Patient	Fasting Cholesterol	Postprandial Cholesterol	Difference (d)
1	198	202	-4
2	192	188	+4
3	241	238	+3
4	229	226	+3
5	185	174	+11
6	303	315	-12

$\bar{d}$ = 0.833
s_d = 7.885
n = 6

t-test: Paired

t-table

From our example:
t=0.259 with 5 d.f.

Value is very much lower than 2.571, the critical value for statistical significance at the Pr=0.05 (5%) level when df=5
i.e. Pr > 0.05

	Probability			
df	.05	.02	.01	.001
1	12.706	31.821	63.657	636.619
2	4.303	6.965	9.925	31.598
3	3.182	4.541	5.841	12.924
4	2.776	3.747	4.604	8.610
5	2.571	3.365	4.032	6.869
6	2.447	3.143	3.707	5.959
7	2.365	2.998	3.499	5.408
8	2.306	2.896	3.355	5.041
9	2.262	2.821	3.250	4.781
10	2.228	2.764	3.169	4.587
11	2.201	2.718	3.106	4.437
12	2.179	2.681	3.055	4.318
13	2.160	2.650	3.012	4.221
14	2.145	2.624	2.977	4.140
15	2.131	2.602	2.947	4.073
16	2.120	2.583	2.921	4.015
17	2.110	2.567	2.898	3.965
18	2.101	2.552	2.878	3.922
19	2.093	2.539	2.861	3.883
. . .	. . .	. . .	. . .	. . .
. . .	. . .	. . .	. . .	. . .
. . .	. . .	. . .	. . .	. . .
25	2.060	2.485	2.787	3.725
26	2.056	2.479	2.779	3.707
27	2.052	2.473	2.771	3.690
28	2.048	2.467	2.763	3.674
29	2.045	2.462	2.756	3.659
30	2.042	2.457	2.750	3.646
40	2.021	2.423	2.704	3.551
60	2.000	2.390	2.660	3.460
120	1.980	2.358	2.617	3.373
α	1.960	2.326	2.576	3.291

t-test: Paired

- Null hypothesis is **not rejected**.
- Conclusion: Insufficient evidence from the data to suggest that postprandial cholesterol levels are on average, higher than fasting cholesterol levels.

Patient	Fasting Cholesterol	Postprandial Cholesterol
1	198	202
2	192	188
3	241	238
4	229	226
5	185	174
6	303	315

t-test: Common Errors

Failure to recognise assumptions:

- Population must not be multimodal.
- Population should be symmetrical.

Failure to distinguish situations that require paired or unpaired tests:

- The conclusion will be affected due to differences in calculating the test statistic and the degrees of freedom

Non-parametric Tests

Parametric tests require assumptions of the distribution of the study variables.

In biology, many situations involve variables that cannot follow a normal or t-distribution, such as:

- # of hospital admissions per person per year
- # of surgical operations per person

In these instances, non-parametric tests are conducted.

Non-parametric vs Parametric: The Advantages

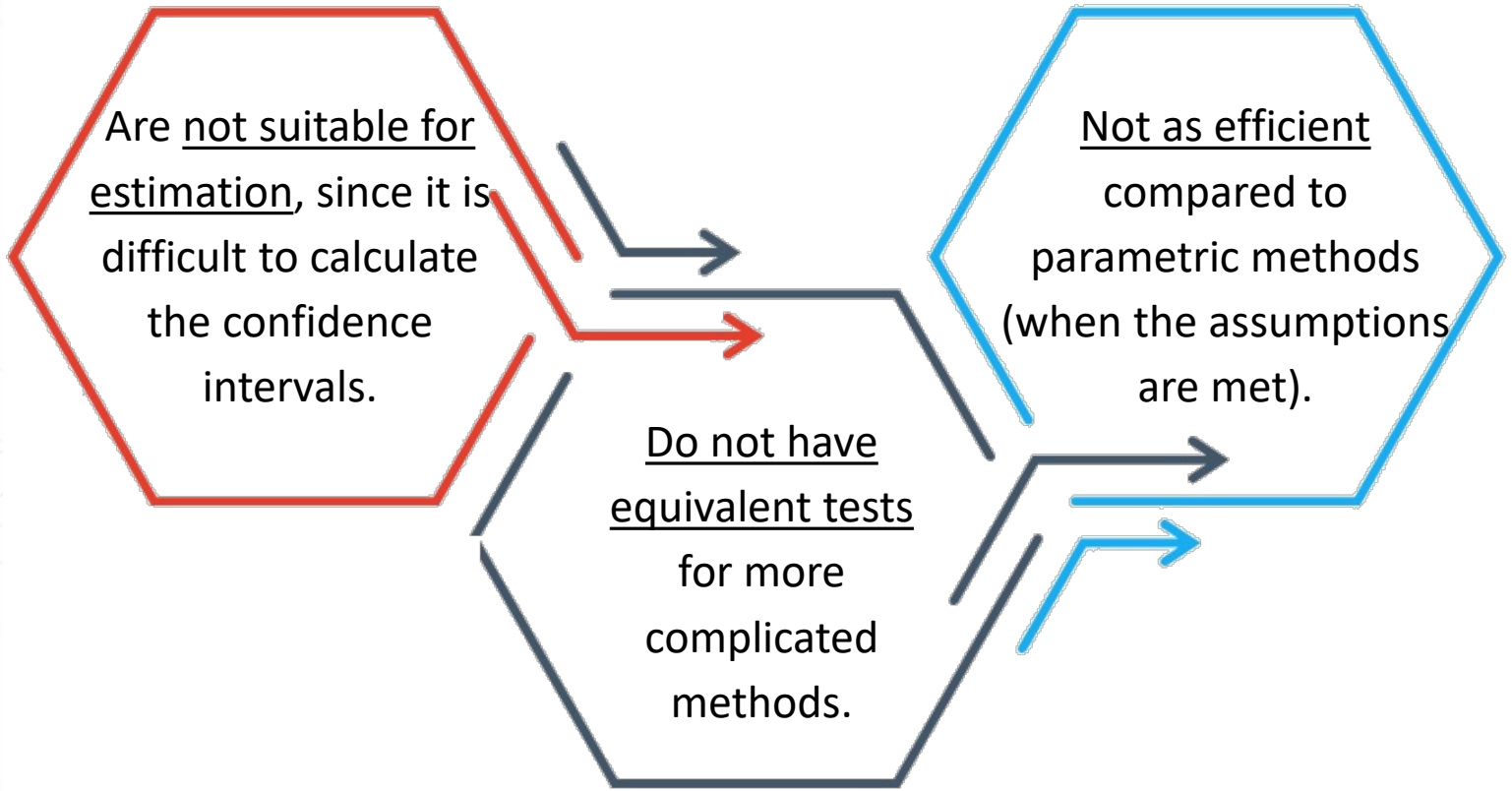
Non-parametric tests can be used for data which are:

- Markedly skewed
- Generated from small sample sizes
- Scores (measured on ordinal scale)

Non-parametric tests are also quick and easy to apply but compare quite well with parametric methods.

Non-parametric vs Parametric: Disadvantages

Non-parametric tests:



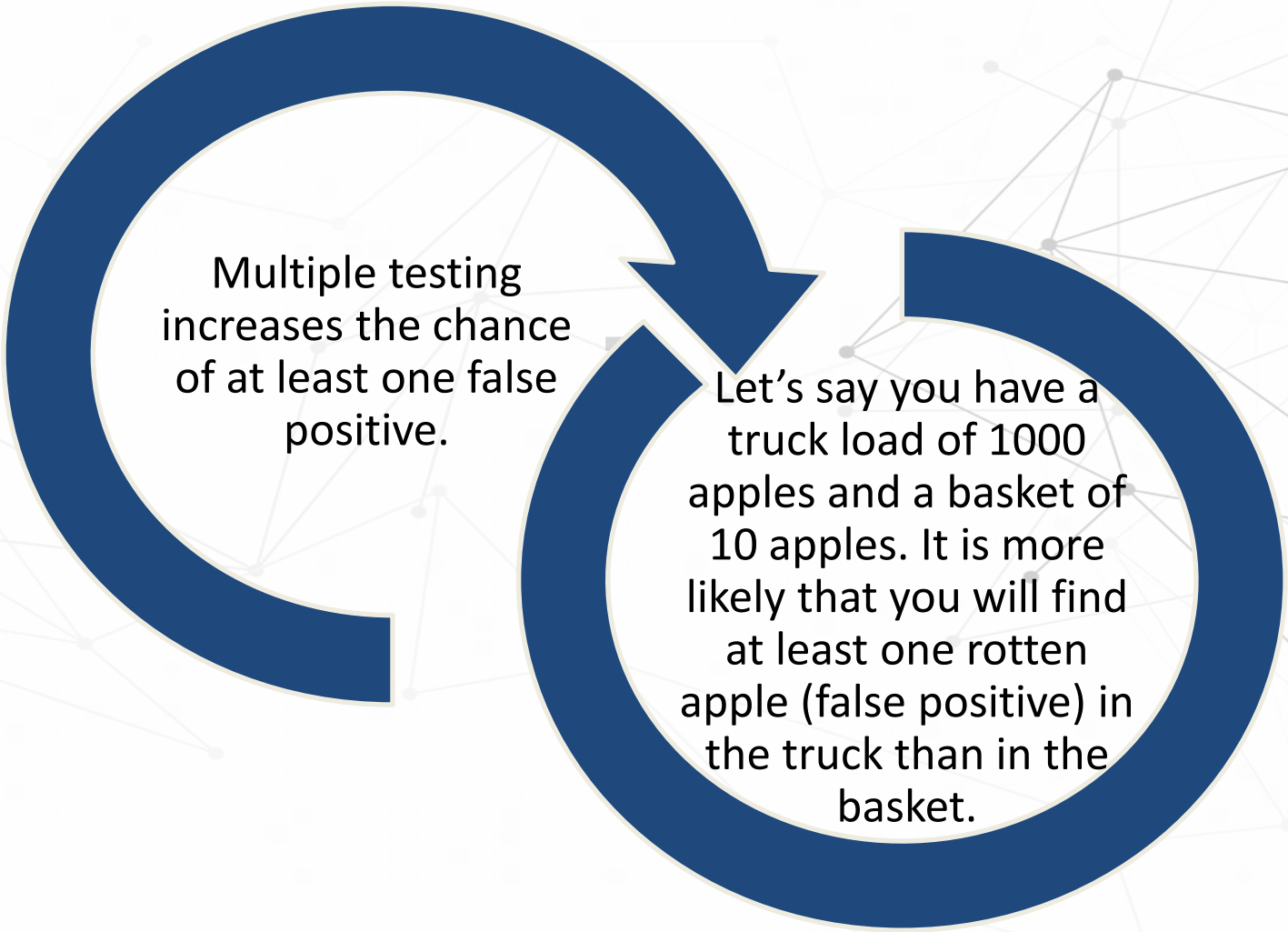
NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Multiple Testing (MT) and Multiple Testing Correction (MTC)

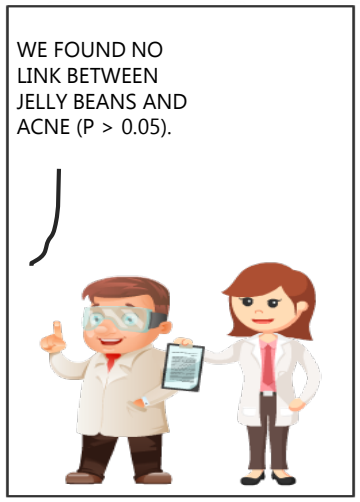
BS3033 Data Science for Biologists

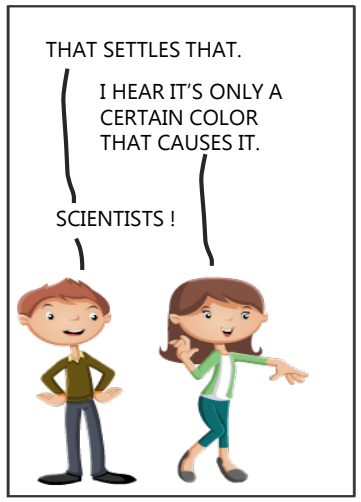
Dr Wilson Goh
School of Biological Sciences

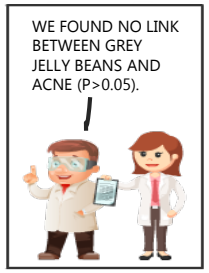
Problems with Multiple Testing

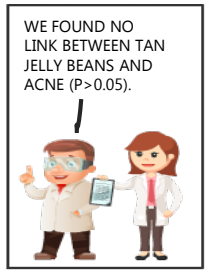


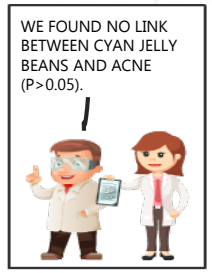
Problems with Multiple Testing

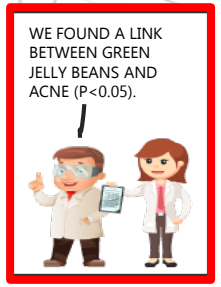


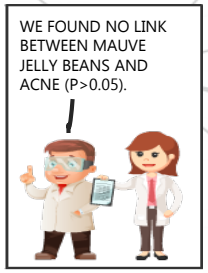


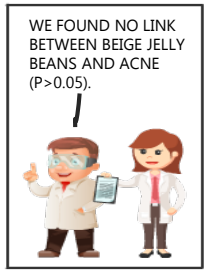


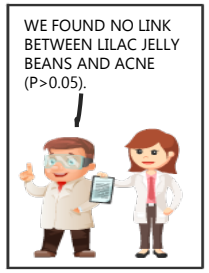


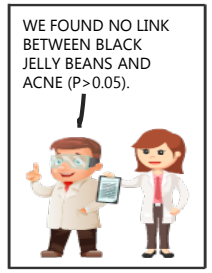


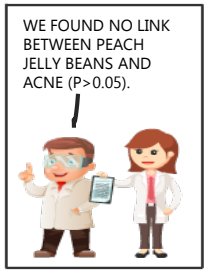


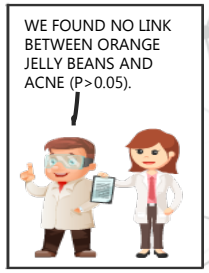


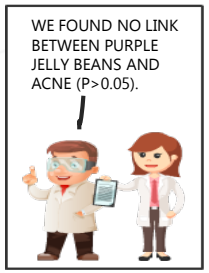


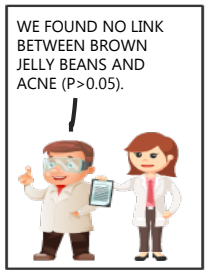


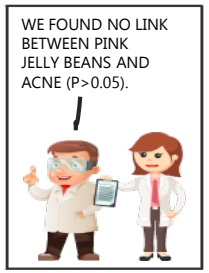


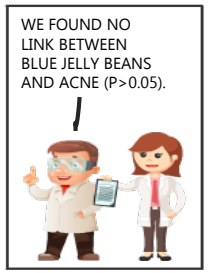


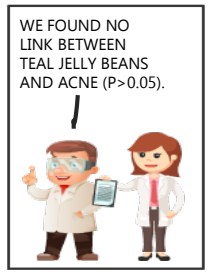


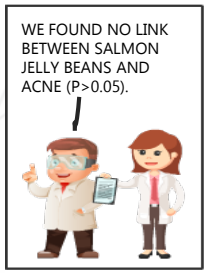


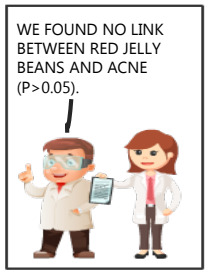


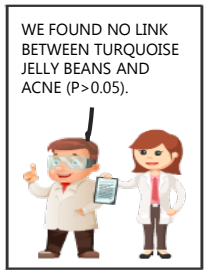


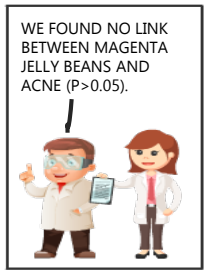


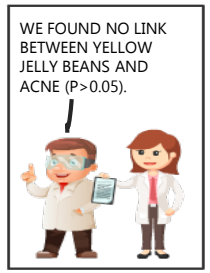


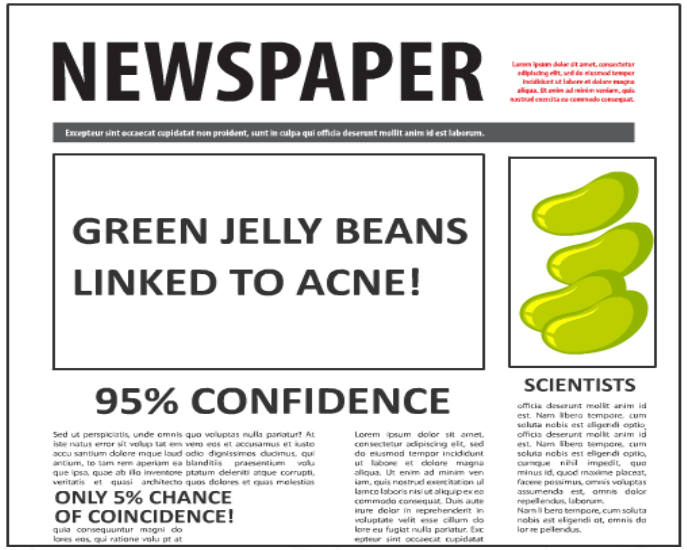












Problems with Multiple Testing: A Mathematical Explanation

- Recall: α (alpha) is the probability of observing a false positive result in a test (e.g. α = 0.01)
- Two ways to explain the problem of multiple testing:

Expected number of false positives:

- **Number of tests × α**
- With 10 tests, E(FP) = 10 × 0.01 = 0.1 → less than 1, unlikely
- With 100 tests, E(FP) = 100 × 0.01 = 1

Probability of observing at least one false positive:

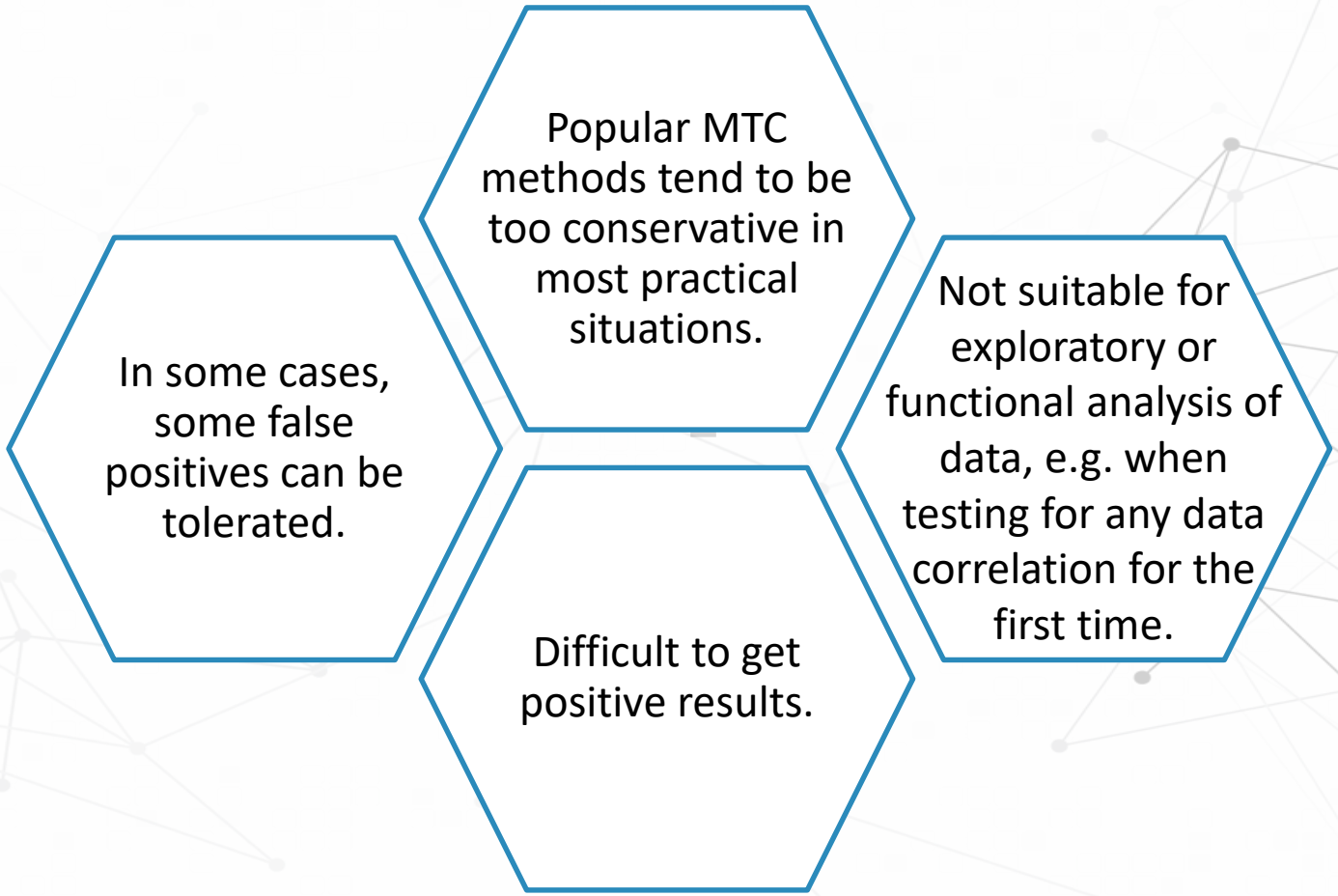
- **1 - (1 – alpha)$^{\text{number of tests}}$**
- For 10 tests, P(at least 1 FP) = 1 – (1 - 0.01)^{10} = 0.09 (almost 10%)
- For 100 tests, P(at least 1 FP) = 1 – (1 - 0.01)^{100} = 0.63 (63%!)

Multiple Testing Corrections

Multiple testing corrections (MTC) takes into consideration this increasing false positive rate when doing multiple tests. MTC methods are:

- Bonferroni (FWEB)
- Holm
- Benjamini-Hochberg

Multiple Testing Corrections: Disadvantages



Only use MTC when you absolutely need the results to be correct (e.g. when confirming a suspected relationship).

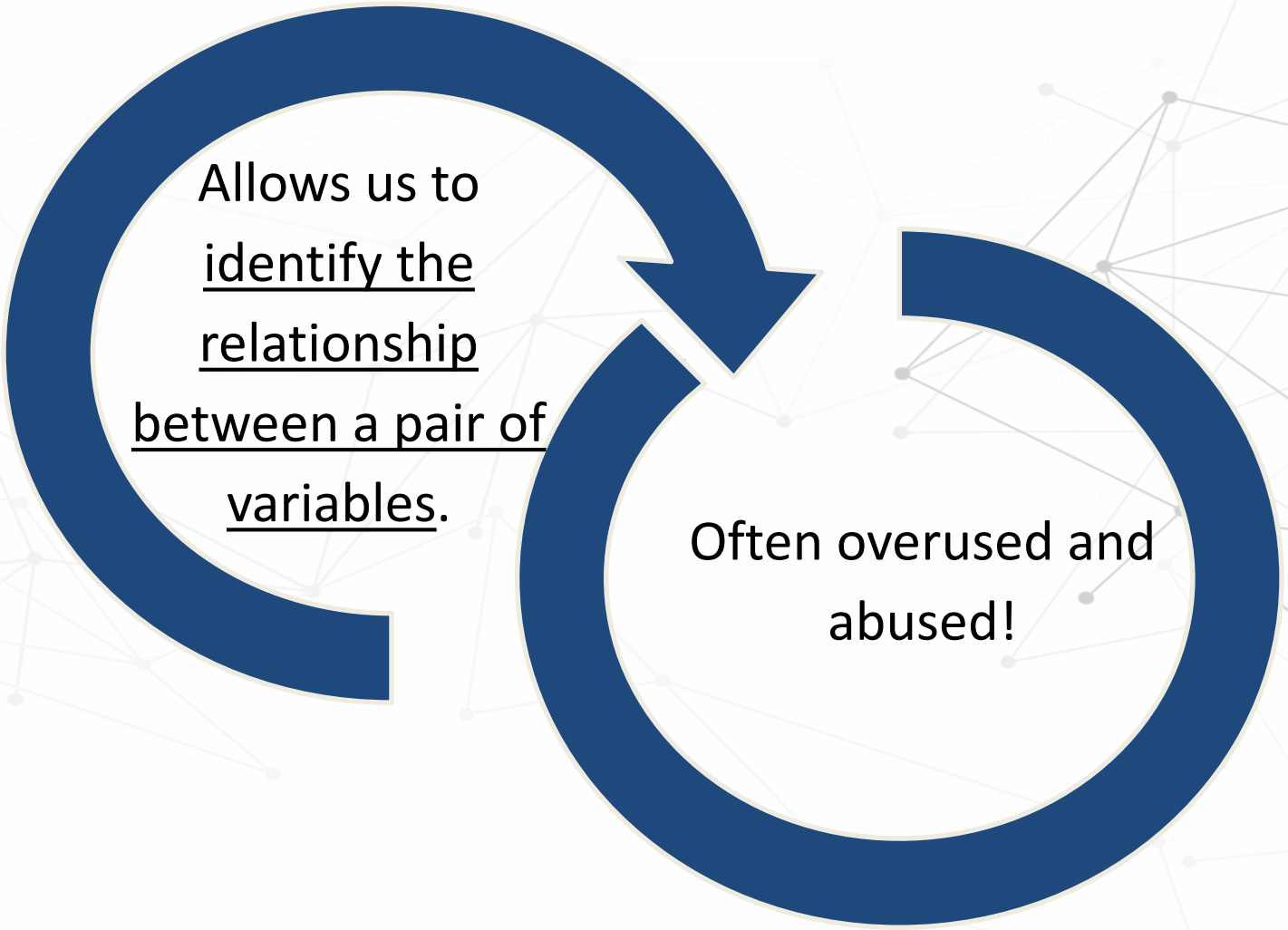
NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Identifying Trends Correlation and Regression

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

Correlation



Covariance – Measuring Correlation

Covariance is the mean value of the product of the deviations of two variables from their respective means, also expressed in equation as:

$$cov\,(x, y) = \frac{\sum_{i=1}^{n}(xi - \bar{X})\,(yi - \bar{Y})}{n - 1}$$

Numerical value can be both positive or negative!

Understanding Covariance

The numerical value of covariance can be:

- cov(X,Y) **> 0** → X & Y are **positively** correlated
- cov(X,Y) **< 0** → X & Y are **negatively** correlated
- cov(X,Y) **= 0** → X & Y are **not** correlated (independent)

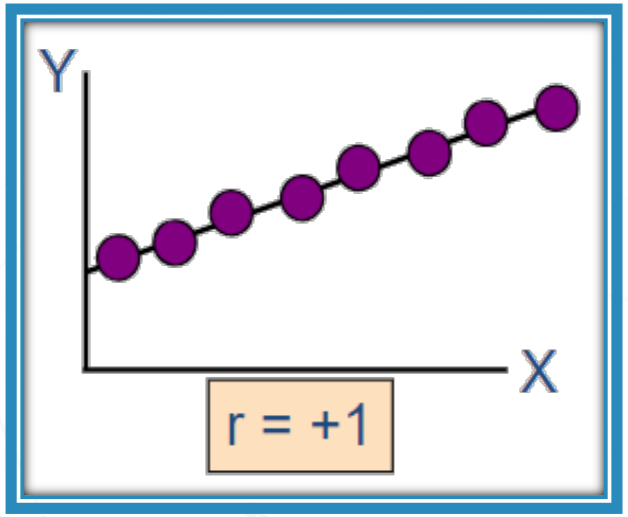
However, covariance does not have a defined range, causing it to be difficult to evaluate the extent of correlation. Hence, we typically standardise the covariance through the Pearson Product Correlation, which introduces fixed boundaries.

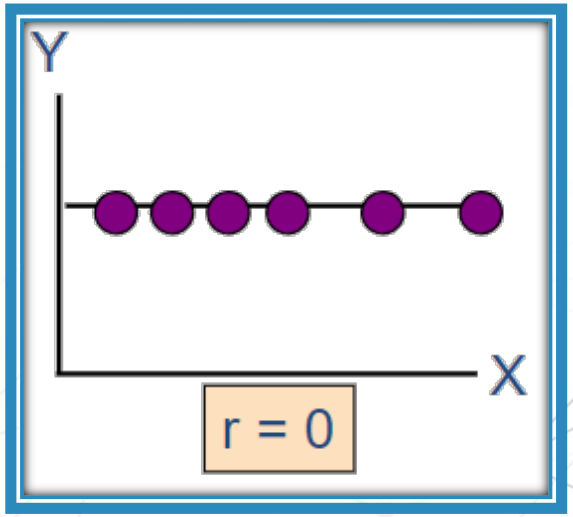
Pearson Product Coefficient

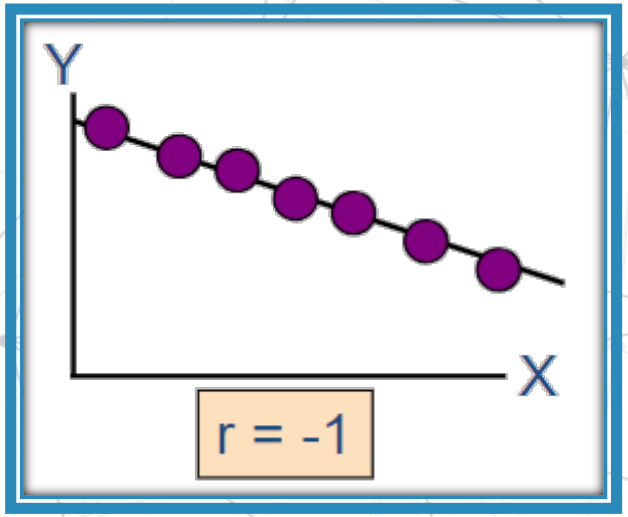
- A **standardised form** of the **covariance** such that its values are **bound between -1 and 1**.
- **No** units (therefore universal standard)
- Value nearer to -1: **Negative** correlation
- Value nearer to +1: **Positive** correlation
- Value is 0: Variables have **no relationship**

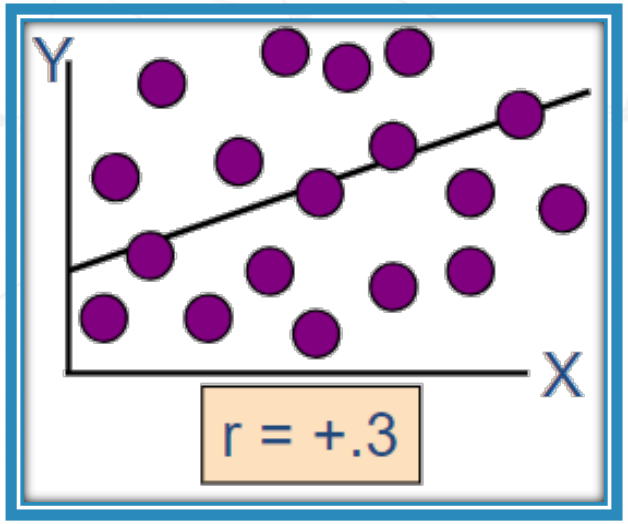
$$r = \frac{covariance\ (x, y)}{\sqrt{var\ x}\ \sqrt{var\ y}}$$

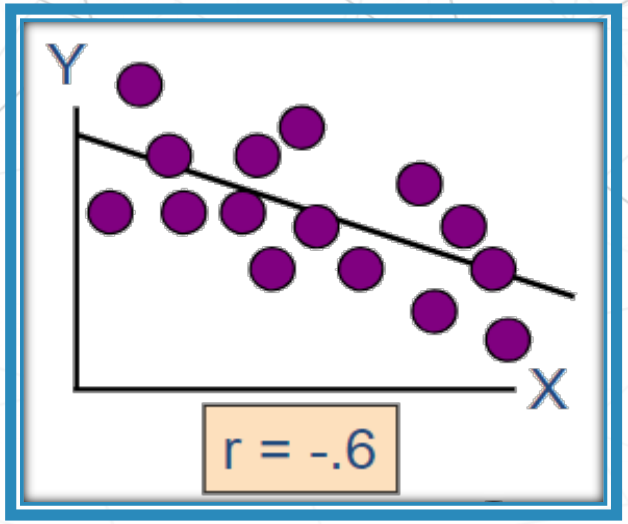
Visual Representation of Correlation





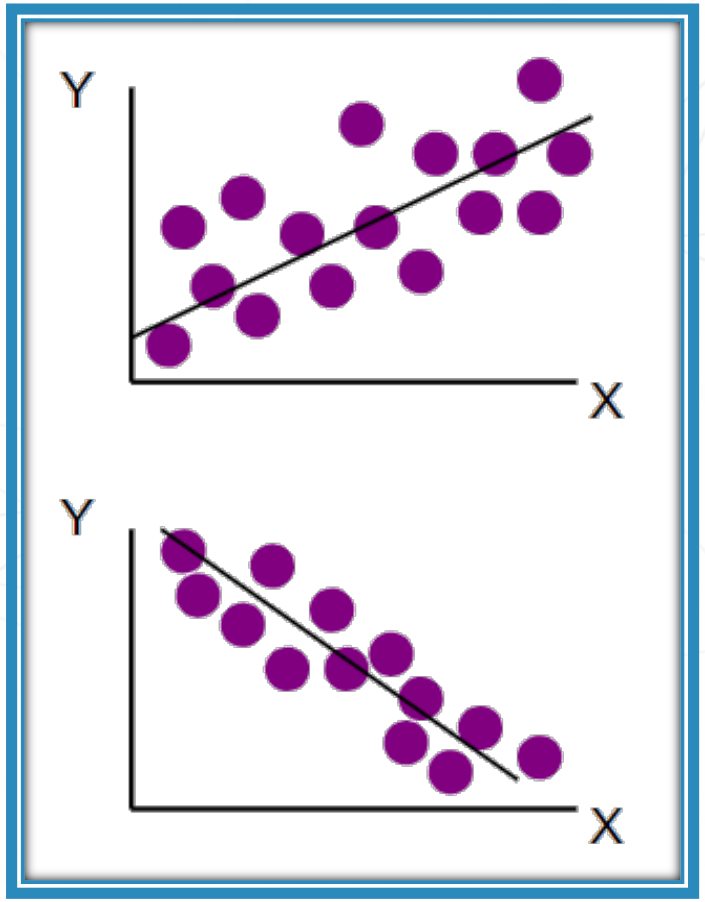




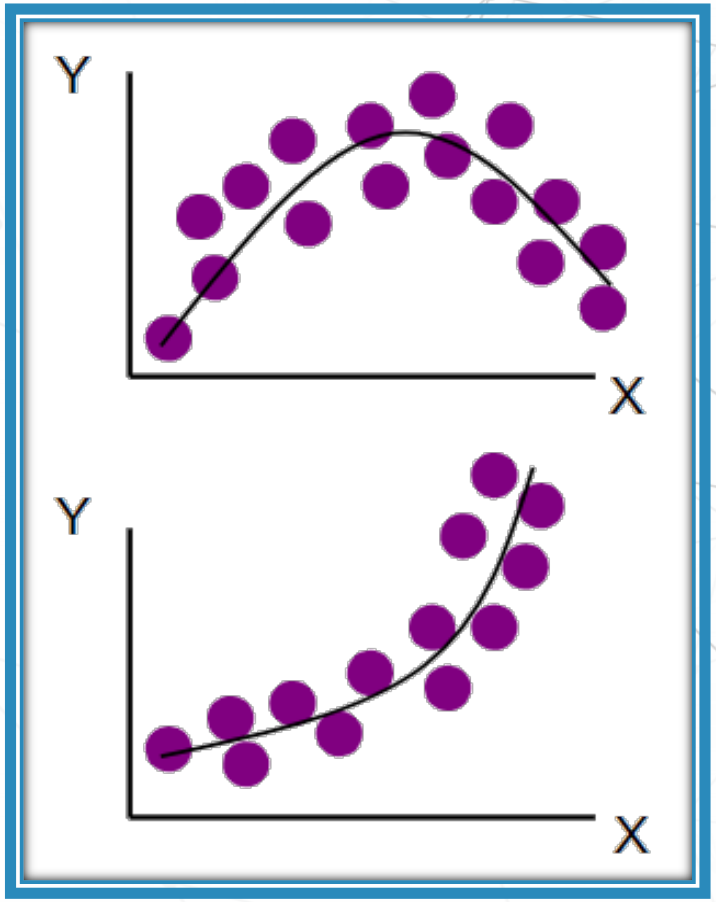


Linear and Non-linear Representation of Correlation

Linear Relationships

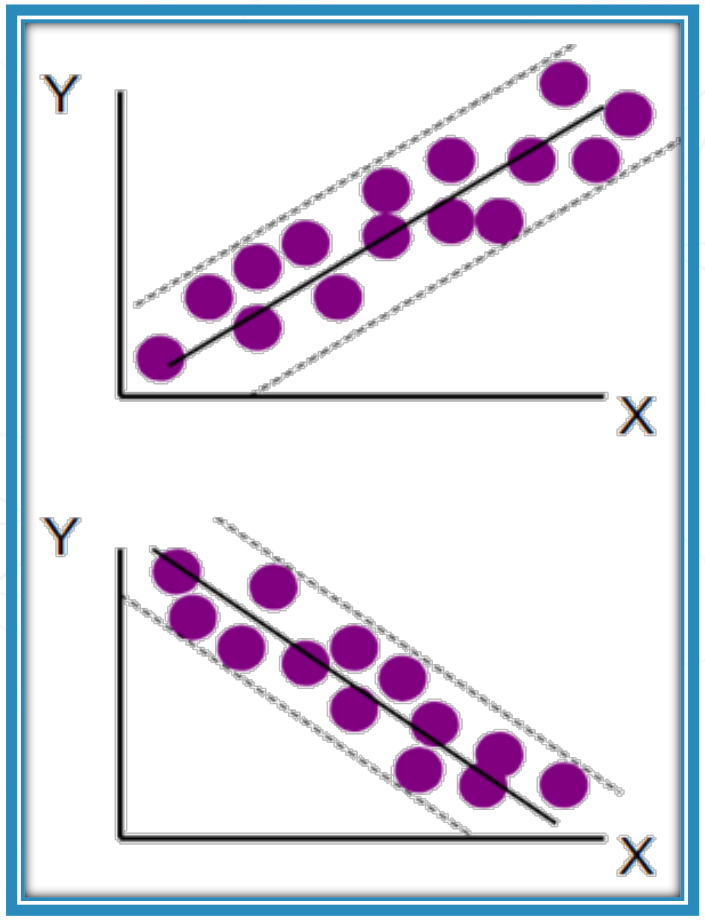


Non-linear Relationships

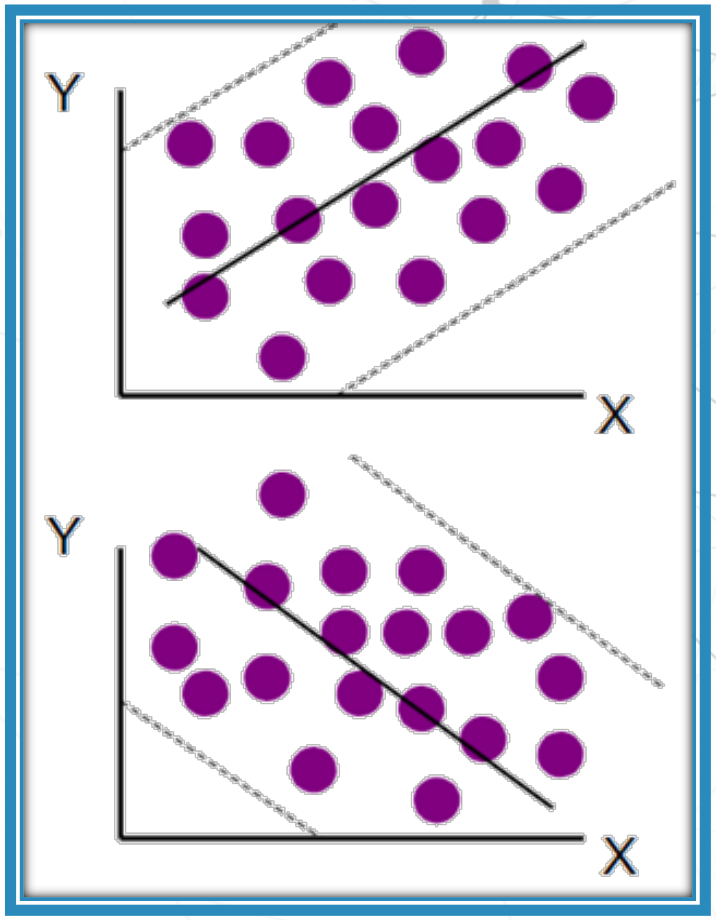


Strong and Weak Relationships

Strong Relationships



Weak Relationships



For strong relationships, we can predict the value of Y given X with little error.

Pearson Product Coefficient

$$\hat{r} = \frac{covariance\ (x,y)}{\sqrt{var\ x}\sqrt{var\ y}}$$

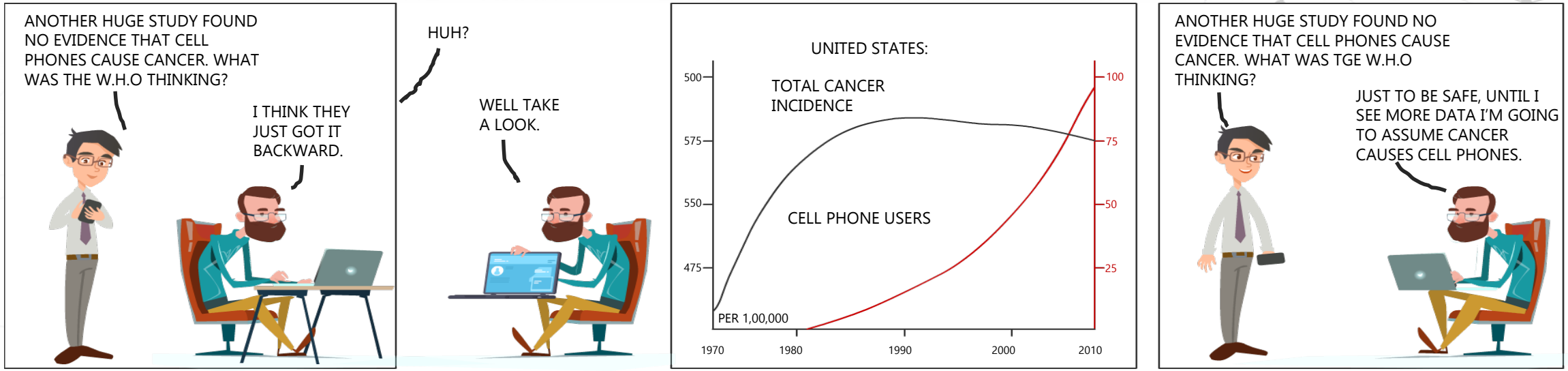
$$\downarrow$$

$$\hat{r} = \frac{\dfrac{\sum_{i=1}^{n}(xi - \overline{x})\,(yi - \overline{y})}{\cancel{n-1}}}{\sqrt{\dfrac{\sum_{i=1}^{n}(xi - \overline{x})^2}{\cancel{n-1}}}\sqrt{\dfrac{\sum_{i=1}^{n}(yi - \overline{y})^2}{\cancel{n-1}}}}$$

$$\downarrow$$

$$\hat{r} = \frac{SS_{xy}}{\sqrt{SS_x SS_y}}$$

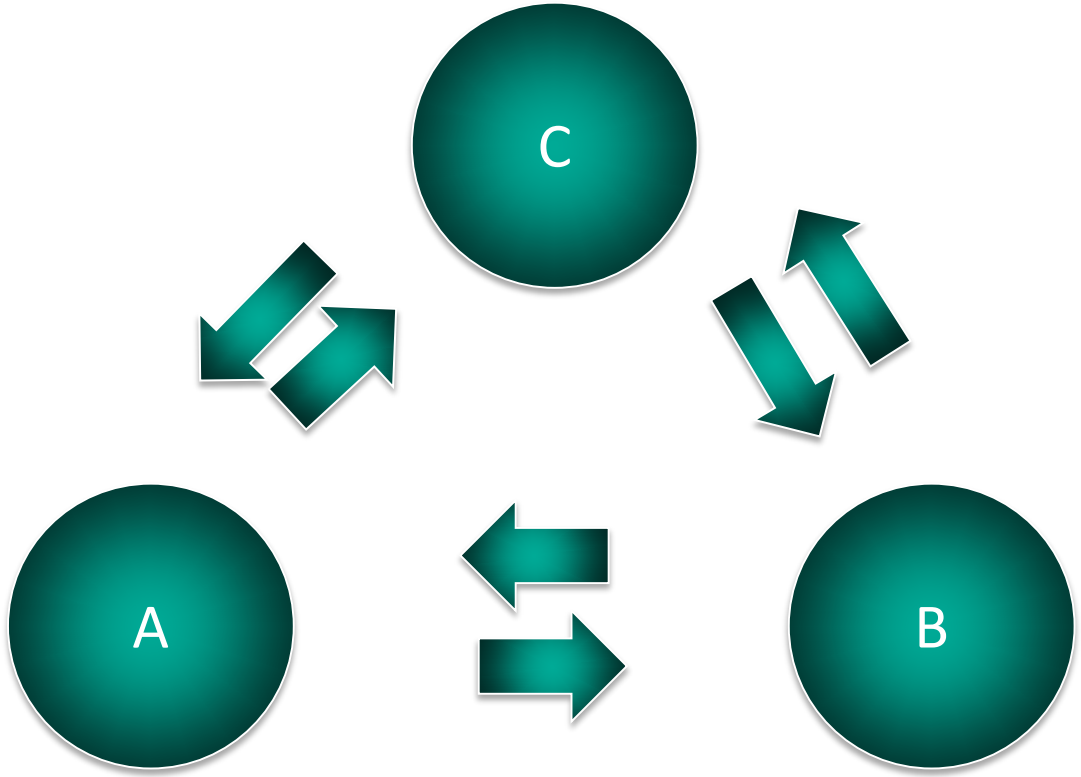
Correlation ≠ Causation!



In many cases, observed correlations are merely coincidental!

Correlation vs Causation!

When two variables A, B are correlated, there are at least 6 possibilities:



Correlation vs Regression

These two terms are frequently confused in biology!

Correlation:

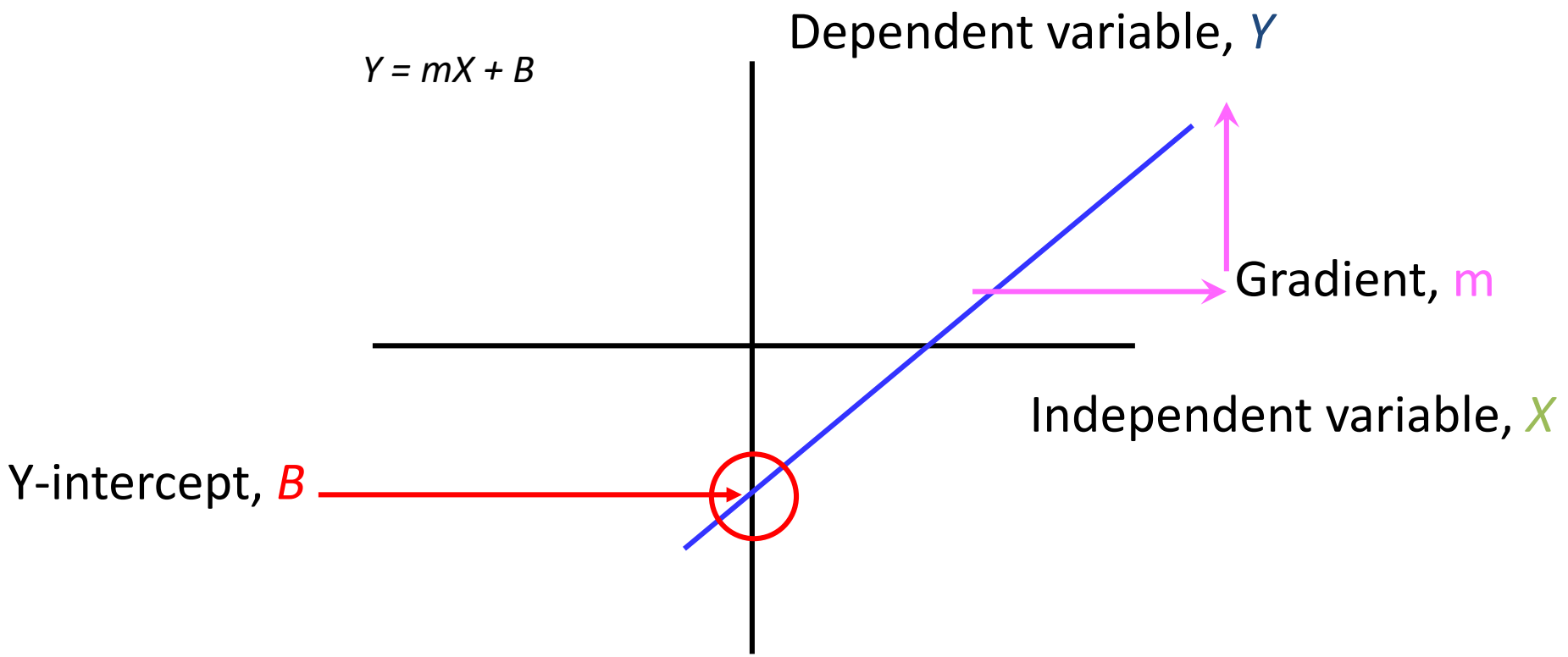
- Shows the general relationship between two variables.
- Variables are treated as independent of each other.

Regression

- Provides a model for the relationship between two variables.
- They are assumed to have a cause and effect relationship, where one variable is independent (predictor) while the other is dependent (outcome). They are therefore non-independent variables.

Linear Regression

A linear equation can be expressed in four components:



Linear Regression – Gradient

The Gradient of the Linear Regression is given by:

$$M = \frac{Y_2 - Y_1}{X_2 - X_1}$$

using any pair of points.

With the y-intercept, we can then come up with the linear equation to represent the relationship between 2 variables.

Linear Regression – Uncertainty

In reality, we also need to account for any uncertainty, which may cause the predicted value to vary from the actual value.

Hence, linear regressions usually appear as:

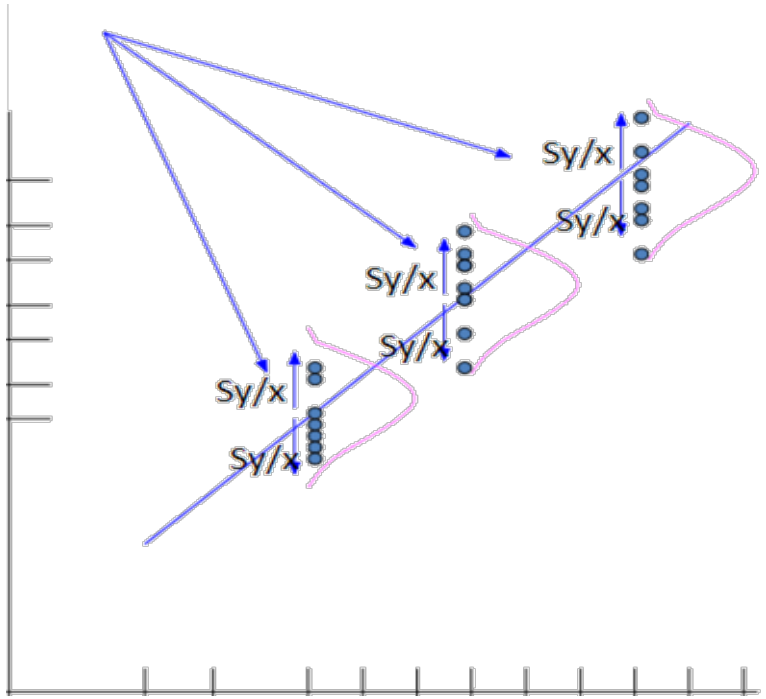
$$\hat{y}_I = \alpha + \beta x_I + \text{Random Error}_I$$

Assumptions of Linear Regression

1. Relationship between X and Y is **linear**.
2. Y is **distributed normally** at each value of X.
3. The **variance** of Y at every value of X is the **same** (homogeneity of variances).
4. Each observation is **independent** of the other.

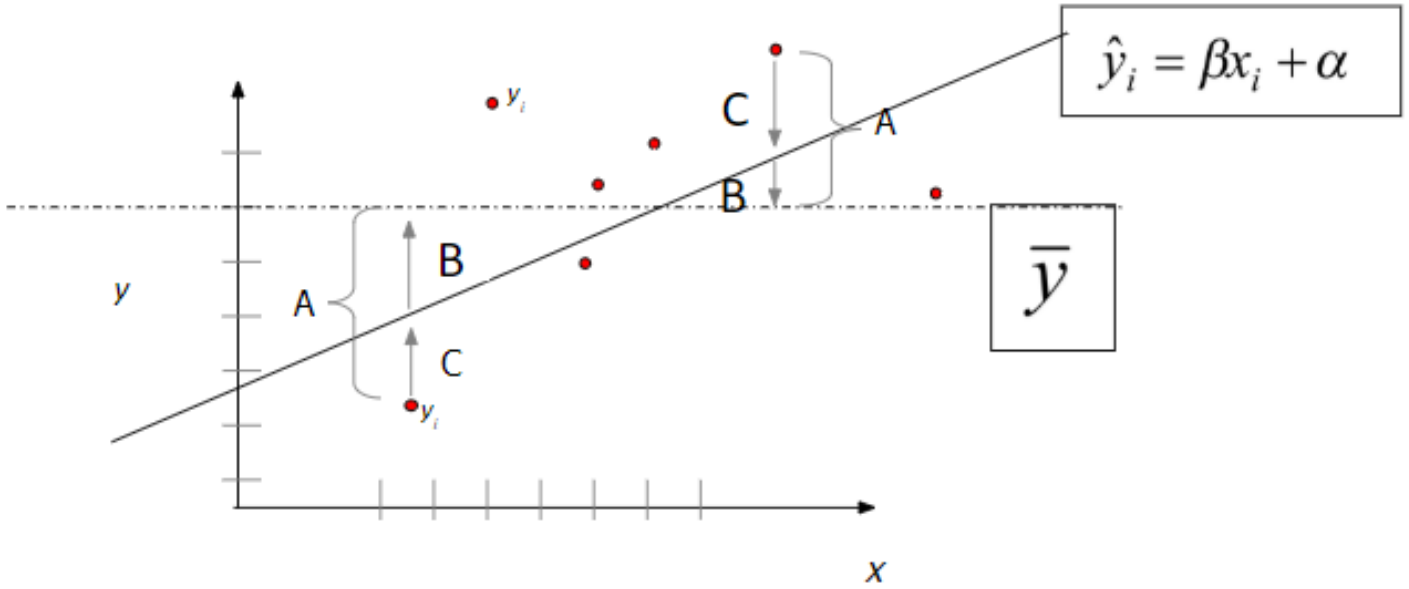
Assumptions of Linear Regression

Standard error of Y given X. It is the average variability around the regression line at any given value of X. It is assumed to be equal at all values of X.



The R^2

Explains how well the regression fits the data and is bound between 0 to 1.



$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

A^2 B^2 C^2

The R^2

Explains how well the regression fits the data and is bound between 0 to 1.

SS_{total}
Total squared distance of observations from naïve mean of y *Total variation.*

SS_{reg}
Distance from regression line to naïve mean of y. Variability due to x (regression).

$SS_{residual}$
Variance around the regression line. Additional variability not explained by x—what least squares method aims to minimise.

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

A^2 B^2 C^2

Estimating Intercept and Slope

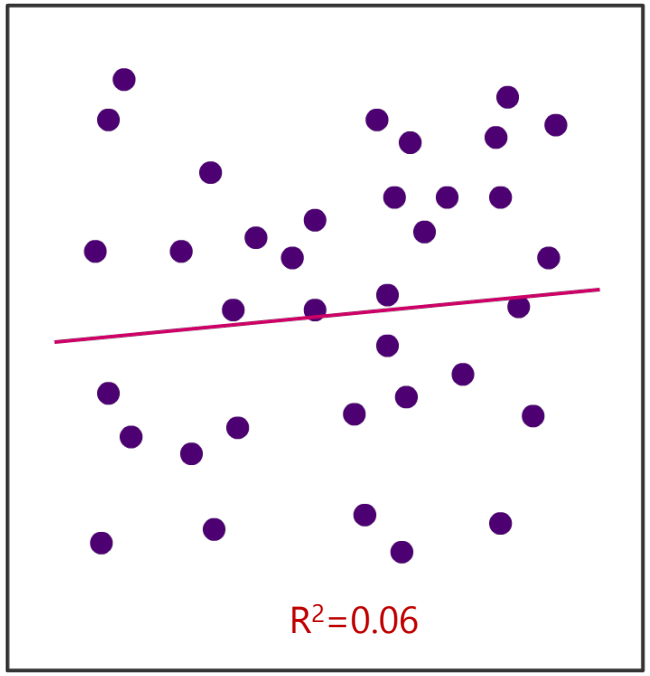
Slope (beta coefficient):

$$\hat{\beta} = \frac{cov\ (x, y)}{var\ (x)}$$

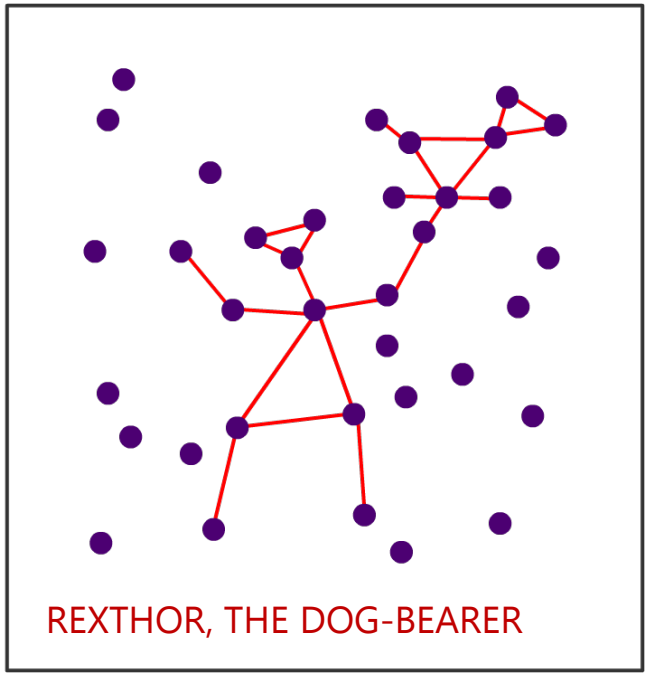
Now that we have beta, we can solve for the y intercept:

$$Calculate\colon\ \hat{\alpha} = \bar{y}\ -\ \hat{\beta}\bar{X}$$

Problems with Regression: Over-interpretation



R^2=0.06

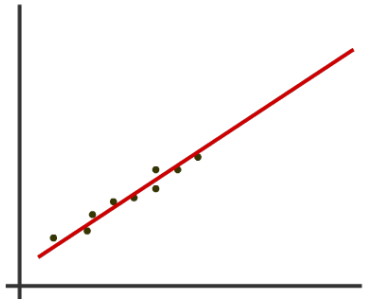


REXTHOR, THE DOG-BEARER

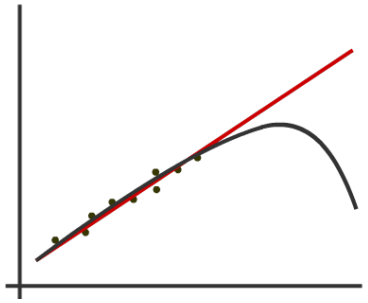
I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

Even when $R^2 \neq 0$, a relationship may not be present (especially if R^2 is small). You can't simply fit a line to everything.

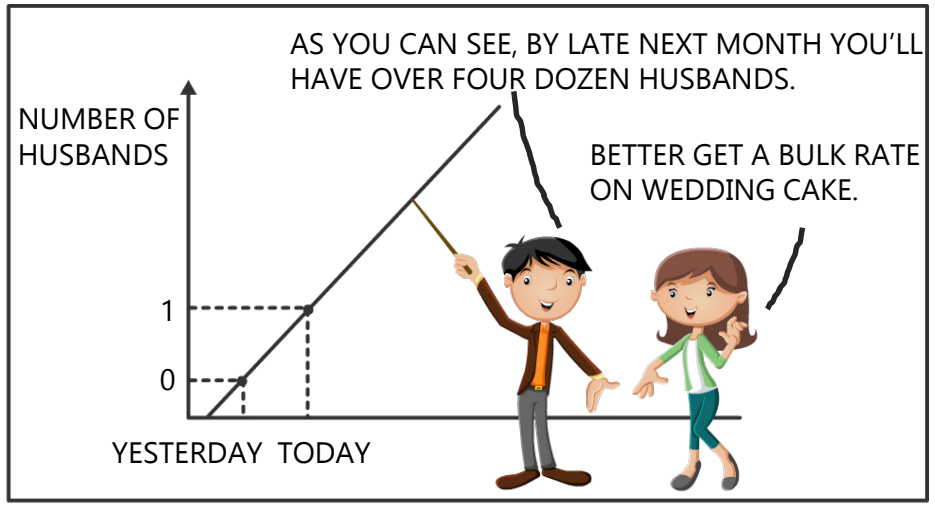
Problems with Regression: Extrapolation



a) Beware of extrapolation past the end of the data.



b) Extrapolated line is red, actual response curve is black.



Any linear relationship only holds true within the data range. Outside the plotted range, the relationship may be different and may lead to seriously biased estimates.

Relationship between Correlation and Regression

Correlation and regression can be interconverted. SD of x and y needs to be known.

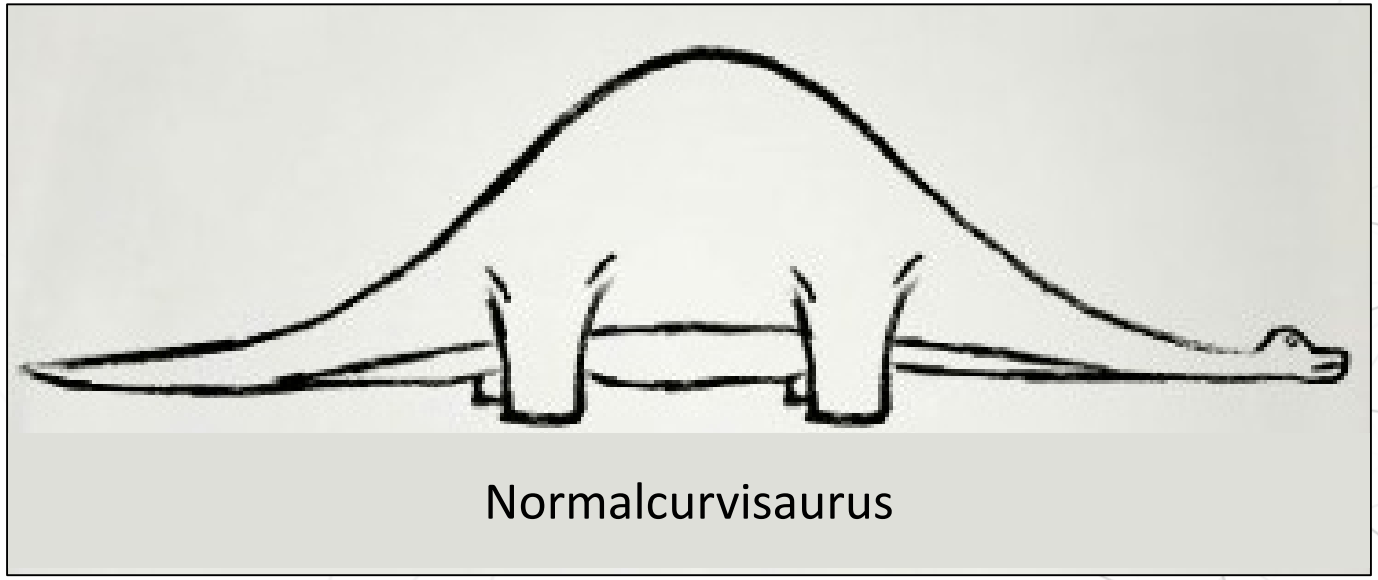
Derivation:

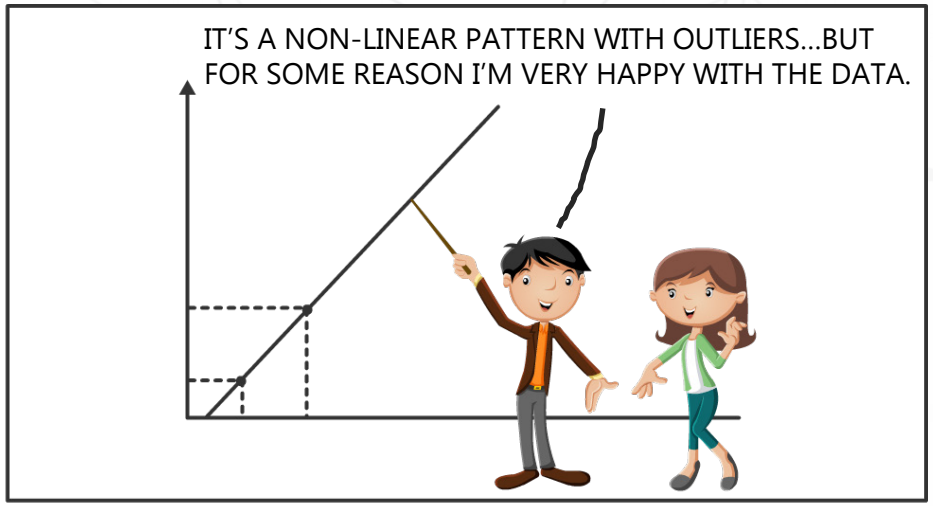
since $\hat{r} = \dfrac{cov\,(x, y)}{\sqrt{var\,(x)}\sqrt{var\,(y)}}$ and $\hat{\beta} = \dfrac{cov\,(x, y)}{var\,(x)} = \dfrac{cov\,(x, y)}{\left(\sqrt{var\,(x)}\right)^2}$,

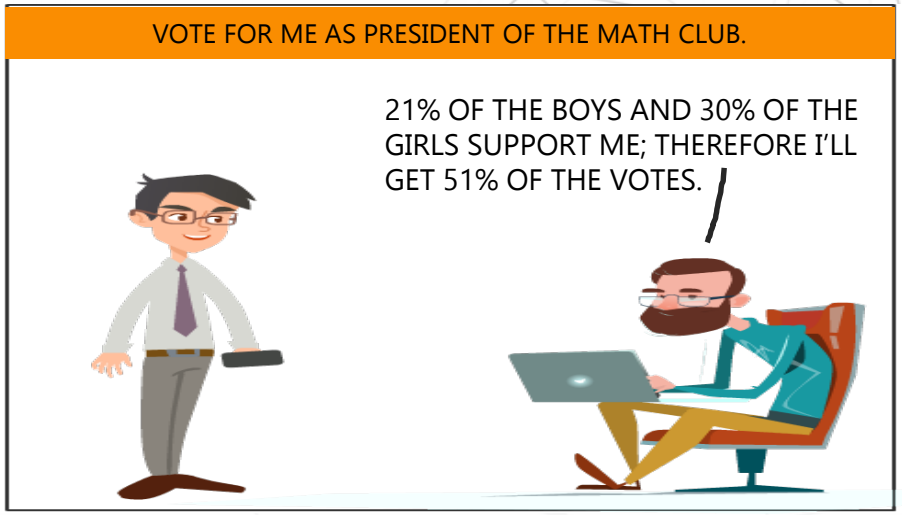
thus $\hat{r} = \hat{\beta}\,\dfrac{\sqrt{var\,(x)}}{\sqrt{var\,(y)}} = \hat{\beta}\,\dfrac{SD_x}{SD_y}$

Relationship between Correlation and Regression

Relationship between Correlation and Regression







NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Summary

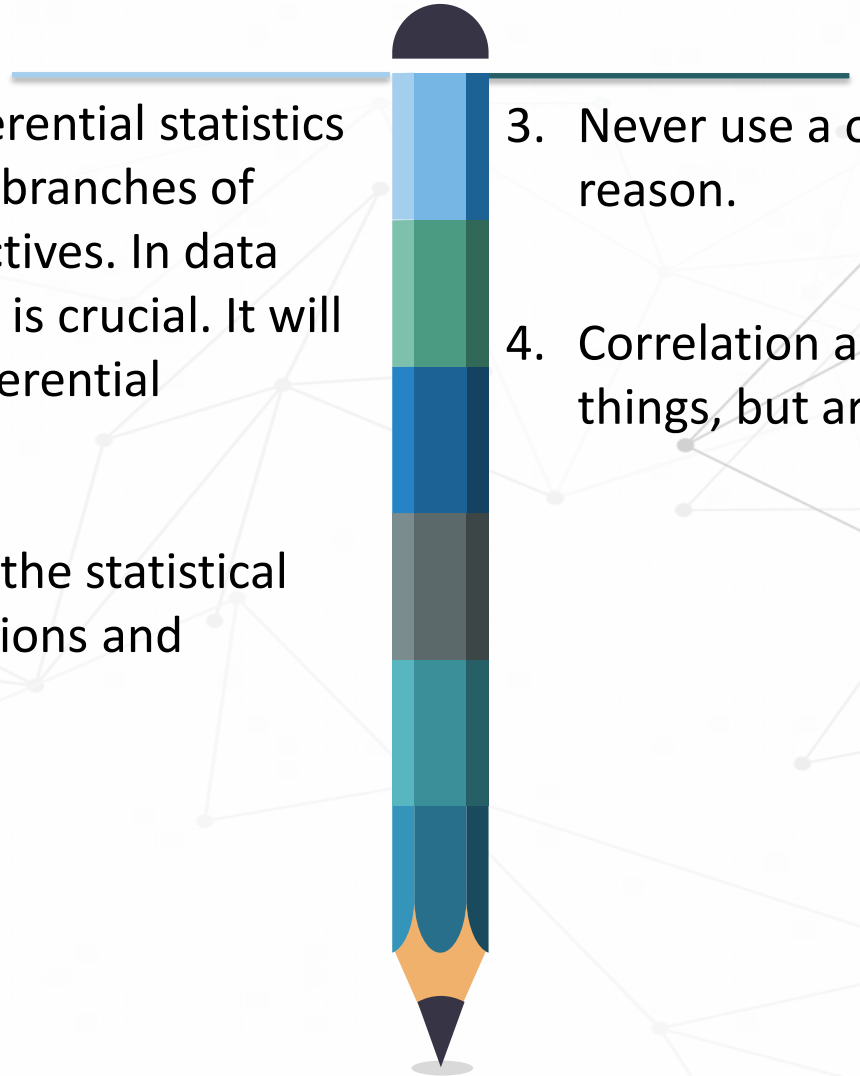
BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

Key Takeaways from this Topic

1. Descriptive statistics and inferential statistics are often taught as separate branches of statistics with different objectives. In data science, descriptive statistics is crucial. It will essentially determine the inferential strategies we will use later.

2. Be careful with the set up of the statistical tests. Be aware of the limitations and assumptions.



3. Never use a one-sided test without good reason.

4. Correlation and regression are not the same things, but are closely associated.