Research Design

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 ‘forgotten assumptions’ in research design.
- Describe the ‘overlooked information’ in research design.
- Describe the various sampling techniques.
- Describe the various normalisation techniques.
- Describe reproducibility and independent corroboration.
- Describe and distinguish meta and mega analyses.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Forgotten Assumptions: Assumptions on Distribution

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

Normal Distribution and Central Limit Theorem

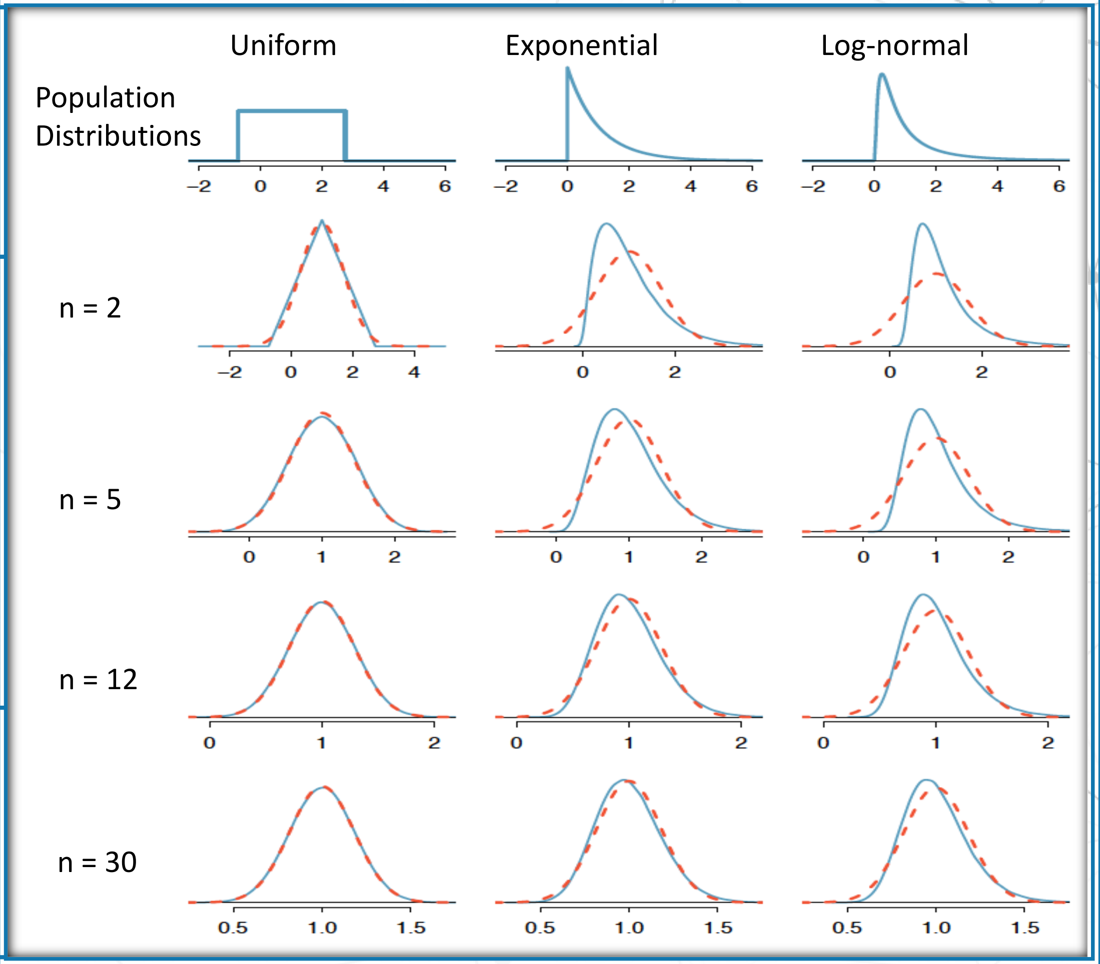
The Central Limit Theorem (CLT) says:

If you sample enough randomly, the distribution of the samples (each with its own mean and variance), will be approximately normal, **regardless of the underlying distribution**.

This means, repeated sampling will produce a normally distributed distribution from which we may estimate the population parameters.

Central Limit Theorem

When sampling size increases, by the CLT distribution of sample mean becomes more symmetrical, and better approximates true mean.



Sampling distributions for the mean at different sample sizes and for three different distributions. The dashed red lines show normal distributions.

Source: LibreTexts Libraries | Creative Commons Attribution-Noncommercial-Share Alike 3.0 United States License.

Central Limit Theorem

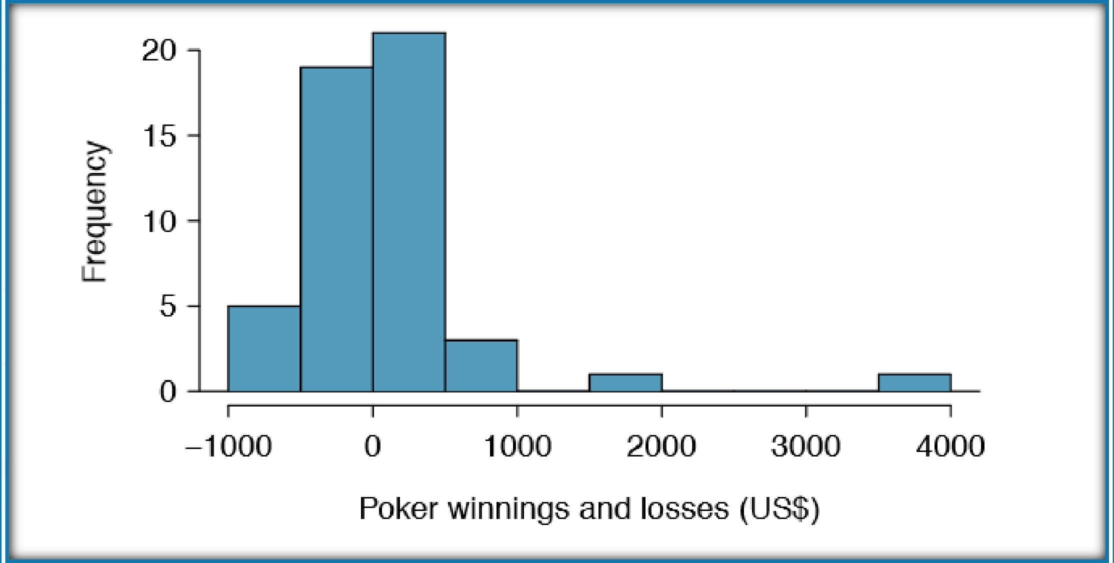
What do you notice about the normal approximation for each sampling distribution as the sample size becomes larger?

Would the normal approximation be good in all applications where the sample size is at least 30?

Not necessarily. For example, the normal approximation for the log-normal example is questionable for a sample size of 30. Generally, the more skewed a population distribution or the more common the frequency of outliers, the larger the sample required to guarantee the distribution of the sample mean is nearly normal.

Central Limit Theorem

Here's a histogram of 50 observations. These represent winnings and losses from 50 consecutive days of a professional poker player. Can the normal approximation be applied to the sample mean?



Sample distribution of poker winnings. These data include some very clear outliers. These are problematic when considering the normality of the sample mean. For example, outliers are often an indicator of very strong skew.

1. These are referred to as time series data, because the data arrived in a particular sequence. If the player wins on one day, it may influence how she plays the next. No evidence was found to indicate the observations are not independent.
2. The sample size is 50, satisfying the sample size condition.
3. There are two outliers, one very extreme, which suggests the data are very strongly skewed or very distant outliers may be common for this type of data. **Outliers can play an important role and affect the distribution of the sample mean and the estimate of the standard error.**

Central Limit Theorem

Caution: Watch out for strong skew and outliers.

Strong skew is often identified by the presence of clear outliers.

If a data set has prominent outliers, or such observations are somewhat common for the type of data under study, then it is useful to collect a sample with many more than 30 observations if the normal model will be used for sample mean.

There are no simple guidelines for what sample size is big enough for all situations, so proceed with caution when working in the presence of strong skew or more extreme outliers.

Forgotten Assumptions:

Independent and Identically Distributed (IID)

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

Independent and Identically Distributed (IID)

The condition of IID states that every sample has equal chance of being selected (**identically distributed**). The selection of one sample does not influence the chance of another being selected (**independent**). This is a common assumption used in many statistical models but...

Does IID reflect reality?

Consider the following scenarios.

Which of the following violate IID and why?

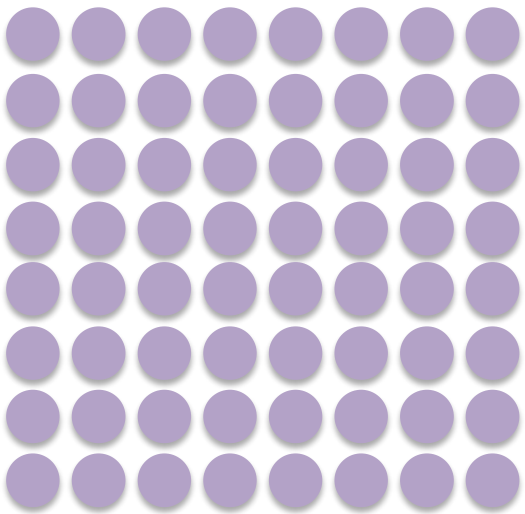
Bringing your friends and family with you to a poll.

Asking students in SBS about student life in NTU.

Doing a Bonferroni correction on a high-throughput study involving 20,000 genes (Hint: Remember what is the assumption of the Bonferroni?).

Does IID reflect reality?

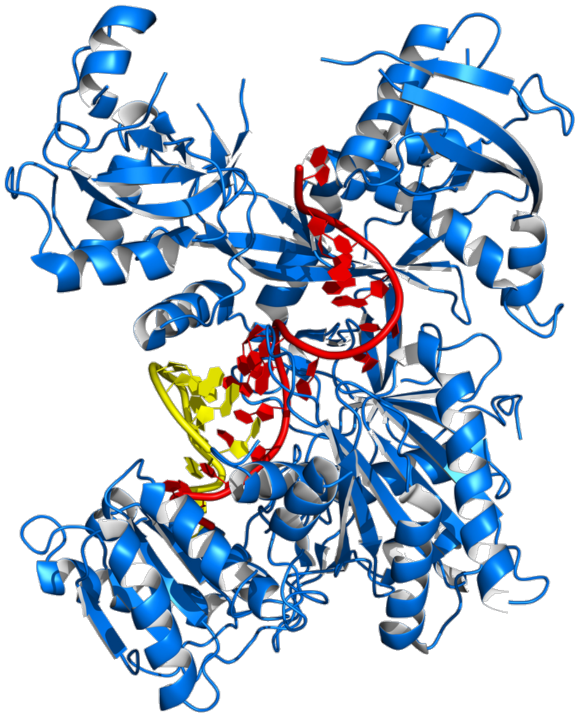
Assumption



All genes behave independently. All genes have equal probability of being sampled/detected.

Statistical assumptions often do not reflect biological reality.

Reality



Genes do not behave independently. High abundance genes are easier to detect.

Proper Design of Experiment: Inclusion Criteria

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

Inclusion Criteria

In clinical testing, we carefully choose the sample to ensure the test is valid.

- Independent: Patients are not related
- Identical: Similar # of male/female, young/old, in cases and controls (apples to apples)

Inclusion Criteria

In big data analysis, and in many datamining works, people sometimes do not set inclusion criteria.

This is not sound as it leads to the generation of hidden confounders.

However, setting very stringent inclusion criteria may limit our ability to generalise (limited scope).

Proper Design of Experiment: Simpsons’ Paradox

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

Simpsons' Paradox

Watch:
https://ed.ted.com/lessons/how-statistics-can-be-misleading-mark-liddell

The presence of lurking variables leads to a reversal of findings once the data has been split by the lurking variable (e.g. male and female).

Best Practice:
Beware anytime data is aggregated. Try to keep dataset balanced across any split by sub-variables (very hard to do). Check that the findings are consistent despite splitting by each potential variable.

Simpsons' Paradox

Looks like A is better		
Overall		
	A	B
Lived	60	65
Died	100	165

Looks like B is better						
Women				**Men**		
	A	B			A	B
Lived	40	15		Lived	20	50
Died	20	5		Died	80	160

Looks like A is better						
History of heart disease				**No history of heart disease**		
	A	B			A	B
Lived	10	55		Lived	10	45
Died	70	50		Died	10	110

Simpsons' Paradox

Looks like A is better

Overall

	A	B
Lived	60	65
Died	100	165

Taking A:
- Men = 100 (63%)
- Women = 60 (37%)

Looks like B is better

Women

	A	B
Lived	40	15
Died	20	5

Men

	A	B
Lived	20	50
Died	80	160

Taking B:
- Men = 210 (91%)
- Women = 20 (9%)

Looks like A is better

History of heart disease

	A	B
Lived	10	55
Died	70	50

No history of heart disease

	A	B
Lived	10	45
Died	10	110

Men taking A:
- History = 80 (80%)
- No history = 20 (20%)

Men taking B:
- History = 55 (26%)
- No history = 155 (74%)

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Proper Design of Experiment: Bias and Fallacies

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

Bias

Axiom:

- An unfair/tainted perspective.
- "The mind sees what it chooses to see." --- *Robert Langdon, The da Vinci Code*

Commonly encountered as follows:

- You see your favorite gene X turn up in a screen, you jump for joy.
- You believe gene X causes disease Y. You only look for evidence in support of your belief.

How to avoid bias?

- Consider evidences objectively.
- Weigh-in/check your thinking with others to derive more fair-handed interpretations.

Sampling/ Ascertainment Bias

Sample is collected such that it is non-representative of the actual population. Estimation of the population parameter from this sample is thus biased.

It can arise from :

- Self-selection
- Pre-screening (or advertising)

Sampling/ Ascertainment Bias

In 1936 a postal survey was conducted to predict the next president of the USA.

The survey was comprised of readers of the American Literary Digest magazine, with additional responses from registered car and phone owners.

The survey predicted Alf Landon, the Republican candidate, would easily win. The actual election was an easy victory for Franklin Roosevelt.

What happened?

Sampling/ Ascertainment Bias

The people surveyed were not randomly chosen and were not a statistically representative sample of the American population.

They were disproportionately rich, when compared to the average voter, and more likely to vote Republican.

Cherry Picking

The act of only considering individual cases or data that confirms a particular position, while ignoring a significant portion of related cases or data that may contradict that position.

If I flipped a fair coin 100 times and I withheld half the data, I can convince you the coin has two heads.

Publication Bias

A type of bias occurring in published academic research. Publication bias is of interest because literature reviews of claims about support for a hypothesis or values for a parameter will themselves be biased if the original literature is contaminated by publication bias.

Publication Bias

In science, we only see the good stuff. But we never see what fails.

A positive study is 3x more likely to be published. So does this mean that scientists are smart people and always succeed in their projects? (you know this is not true!)

But what is more dangerous is that a commonly held but erroneous assertion is held to be truth, and only subsequent works that supports it are publishable, while works that do not support it are assumed to be due to be mistakes (or incompetence).

Insensitivity to Sample Size

Insensitivity to sample size is a cognitive bias that occurs when the probability of obtaining a sample statistic is judged without respect to the sample size.

Insensitivity to Sample Size

People tend to deploy "thinking shortcuts" or heuristics.

Heuristics are economical (reduce thinking effort) and pretty effective usually, but they can also lead to systematic and predictable errors.

Insensitivity to sample size stems from the "representativeness heuristic" where people compare an event to another which is largely similar in characteristics, but neglect consideration of other factors (e.g. sample size).

Fallacies

Axiom:

- An error in reasoning.
- "Having observed 99 heads, the next coin flip must be a tail." --- *Compulsive Anonymous Gambler*

Commonly encountered as follows:

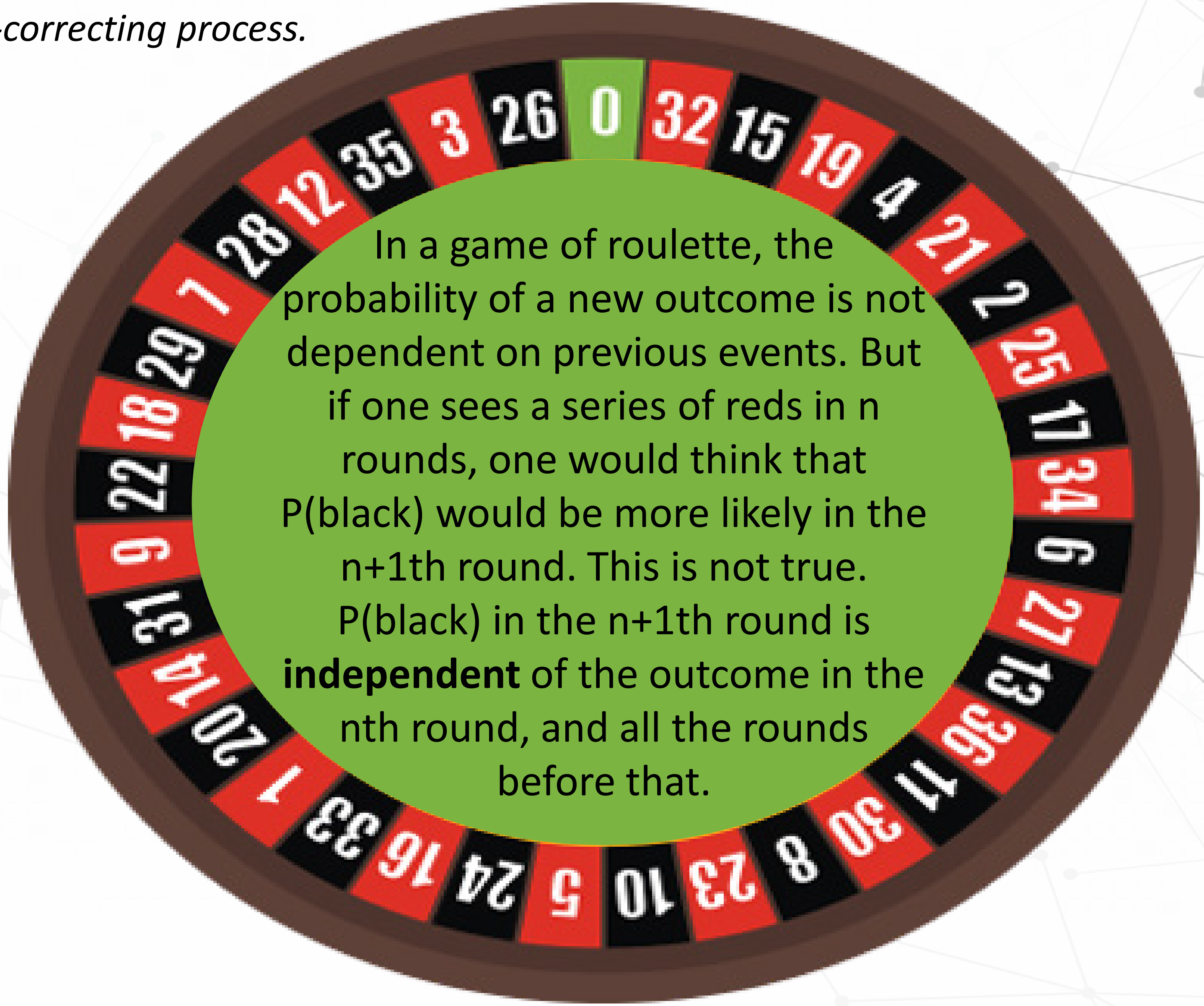
- Gene X is significantly up-regulated in Disease Y, you claim X causes Y.
- When predicting who will come out of the men's bathroom next, you assume equal probabilities between men and women.

How to avoid fallacies?

- Check your reasoning often.
- Write out your logic flow and look for gaps/flaws.
- Check with others and see if you can argue it through.

Gambler’s Fallacy

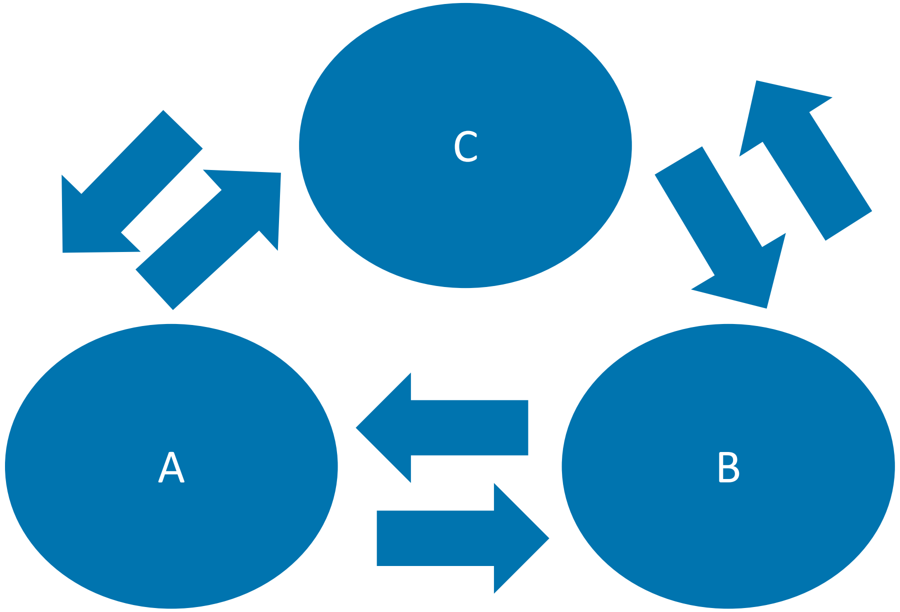
Chance is a not a self-correcting process.



In a game of roulette, the probability of a new outcome is not dependent on previous events. But if one sees a series of reds in n rounds, one would think that P(black) would be more likely in the n+1th round. This is not true. P(black) in the n+1th round is **independent** of the outcome in the nth round, and all the rounds before that.

Correlation-causation

When two variables A, B are correlated, there are at least 6 possibilities: A causes B, B causes A, A and B are controlled by C, A causes C which causes B, B causes C which causes A.



There are also other possibilities: A and B are simply correlated by chance alone.

Ludic

Use of inappropriate model to represent real life. Assuming flawless statistical models apply to situations where they actually don't. Consider the following conversation/ example:

Jason: Since about half the people in the world are female, the chances of the next person to walk out that door being female is about 50/50.

Sarah: Do you realise that is the door to Dr. Chao, the gynecologist?

Proper Design of Experiment: Batch Effects

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

Batch Effects

Batch effects are sub-groups of measurements that have exhibit different behavior across conditions and are unrelated to the biological or scientific variables in a study.

If not properly dealt with, these effects can have a particularly strong and pervasive impact. This can lead to selection of wrong variables from data.

Some Simple Examples

Oven A tends to overheat. Oven B has uneven heating issues. You bake 5 cookies in each oven set to the same temperature. They turn out differently.

Baking

Two people split 10 samples equally between them on a western blot. Person A tends to press down harder on average. Person B tends to press lighter. Blots by person A turn out darker generally.

Pipetting

A More Complex Example

Transcriptomics

You have 2 phenotypes, A and B, with 2 samples each. You split these into 2 runs, 1 and 2 and analyse their gene profiles (A1 B1 and A2 B2). You find that samples tends to cluster by run rather than phenotype.

Question

If you run the samples as A1 A1 and B2 B2, what will happen?

Two Ways of Dealing with Batch Effects

Batch Correction Algorithms

Advantages

- Maintains the "scale" of the data while removing batch-correlated variation.

Disadvantages

- Difficult to use.
- Many different types (need to know how the algorithm works).
- Can affect data integrity (create false positives).

Re-normalise the Data

Advantages

- Simple to use and understand.
- Does not adversely affect data integrity.
- Does not require prior knowledge of batch factors.

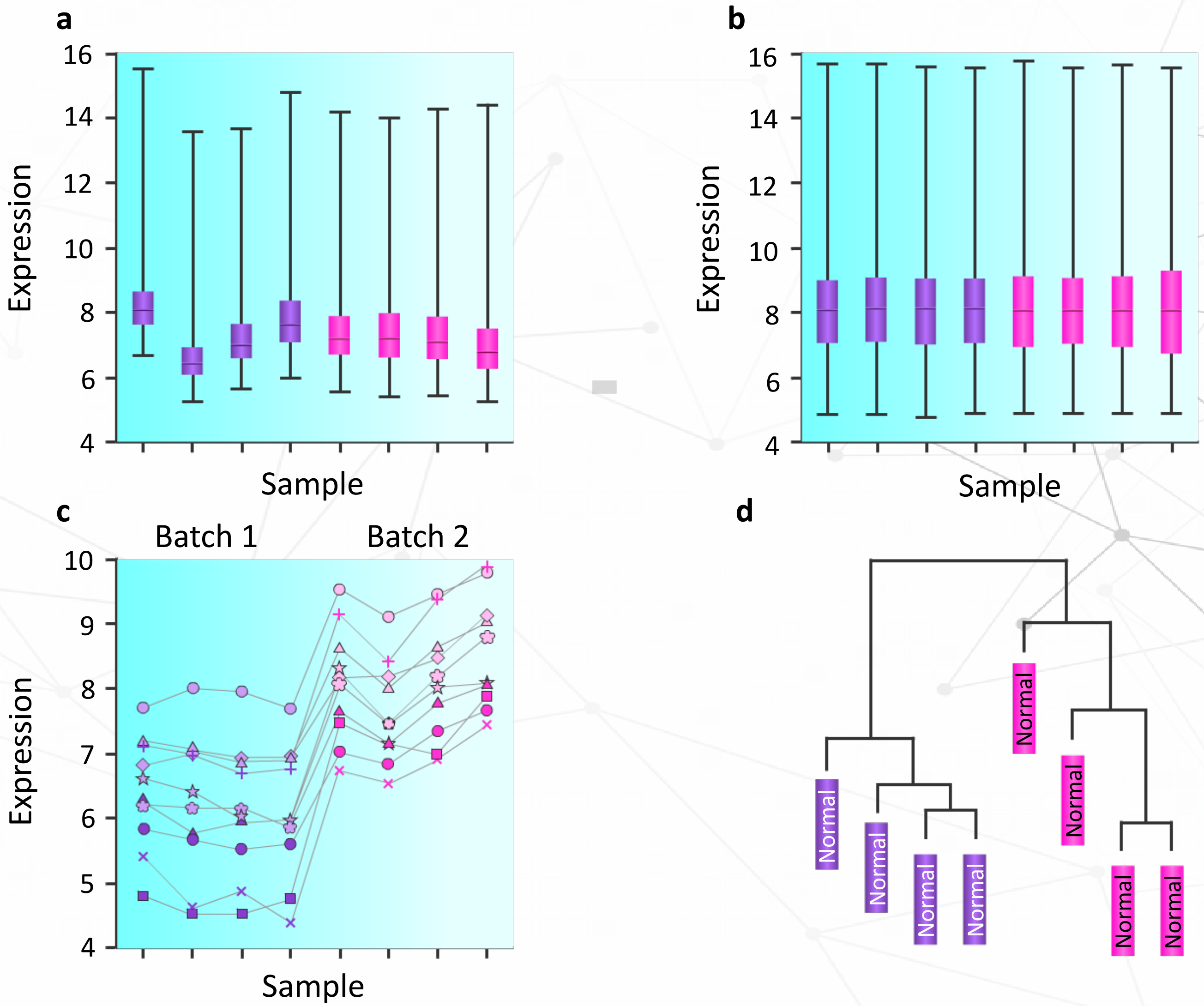
Disadvantages

- Changes the "scale" of the data e.g. in z-norm, you lose information on actual data magnitude.
- Limited efficacy.

Source: Goh WWB et al, Trends in Biotechnology, 2017

Two Ways of Dealing with Batch Effects

Simple normalisation does not guarantee batch effect removal.



Source: Leek et al, Nature Reviews Genetics, 2010

Two Ways of Dealing with Batch Effects

Exploratory Analyses

Hierarchically cluster the samples and label them with biological variables and batch surrogates (such as laboratory and processing time).

Plot individual features versus biological variables and batch surrogates.

Calculate principal components of the high-throughput data and identify components that correlate with batch surrogates.

Downstream Analyses

Do you believe that measured batch surrogates (processing time, Laboratory, etc.) represent the only potential artefacts in the data?

Yes

Use measured technical variables as surrogates for batch and other technical artefacts.

No

Estimate artefacts from the high-throughput data directly using surrogate variable analysis (SVA).

Perform downstream analyses, such as regressions, t-tests or clustering, and adjust for surrogate or estimated batch effects. The estimated/ surrogate variables should be treated as standard covariates, such as sex or age, in subsequent analyses or adjusted for use with tools such as ComBat.

Diagnostic Analyses

Use of SVA and ComBat does not guarantee that batch effects have been addressed. After fitting models, including processing time and date or surrogate variables estimated with SVA, re-cluster the data to ensure that the clusters are not still driven by batch effects.

Source: Leek et al, Nature Reviews Genetics, 2010

Forgotten Assumptions: Domain-specific Laws

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

Domain-specific Laws

Laws of genetics gives us an expectation on genotype distribution frequencies.

Why do you think the data on the right looks suspicious?

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%)

Domain-specific Laws

Laws of genetics gives us an expectation on genotype distribution frequencies.

Why do you think the data on the right looks suspicious?

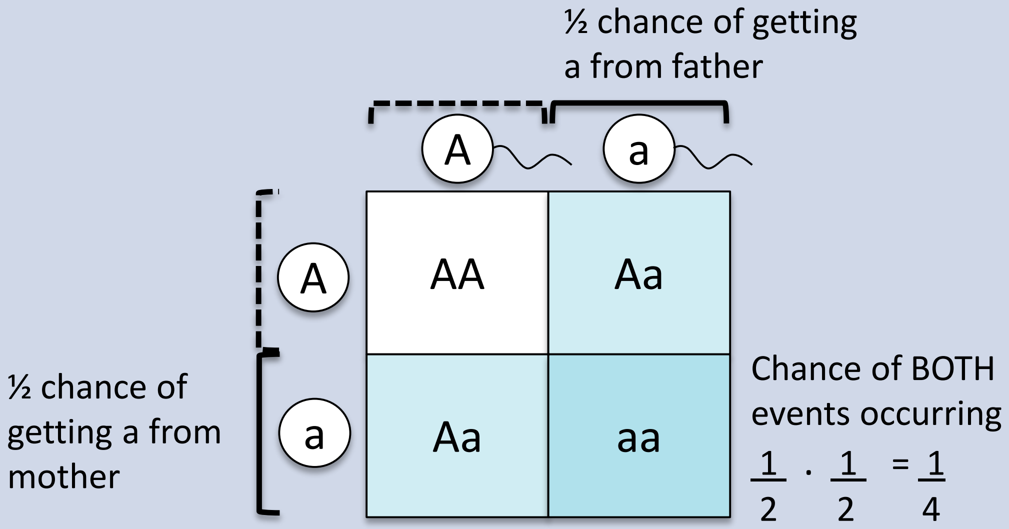
rs123 chi-square p-value = 4.78E-21

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%)

Domain-specific Laws

Laws of genetics gives us an expectation on genotype distribution frequencies.

Let’s use what we know about simple 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%).

Domain-specific Laws

Laws of genetics gives us an expectation on genotype distribution frequencies

Let's look at our table again.

We expect 9%. But our data says AA is only < 1%. So unless AA is lethal, our samples do not reflect expectation.

Therefore, via the use of domain-specific laws (in this, mendelian segregation proportion) we infer that our samples could be biased.

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

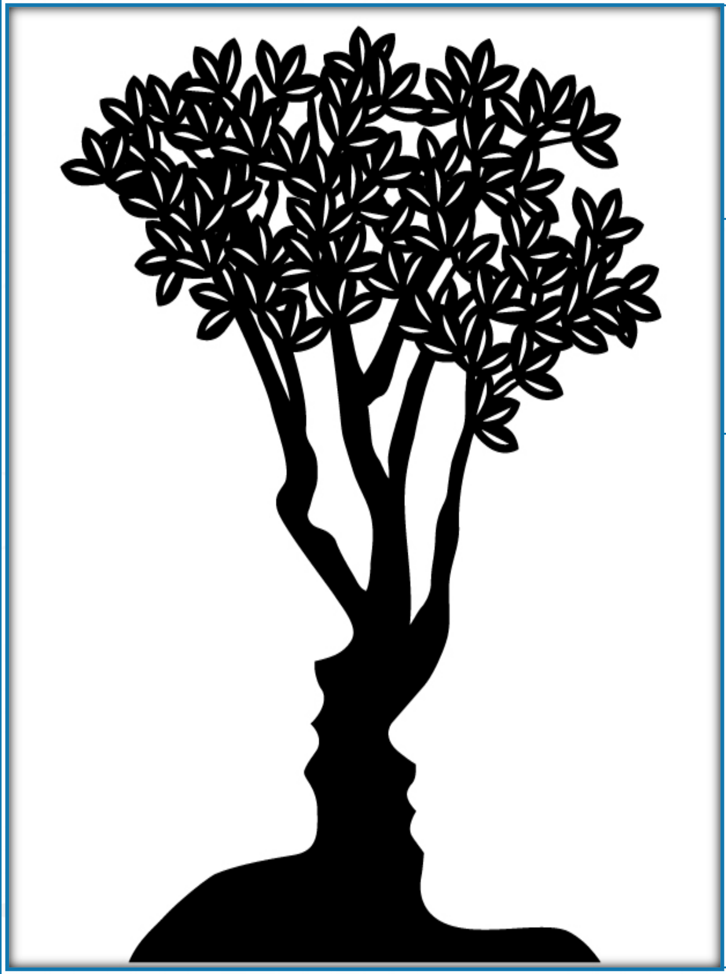
N= 189

Overlooked Information: Non-association

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

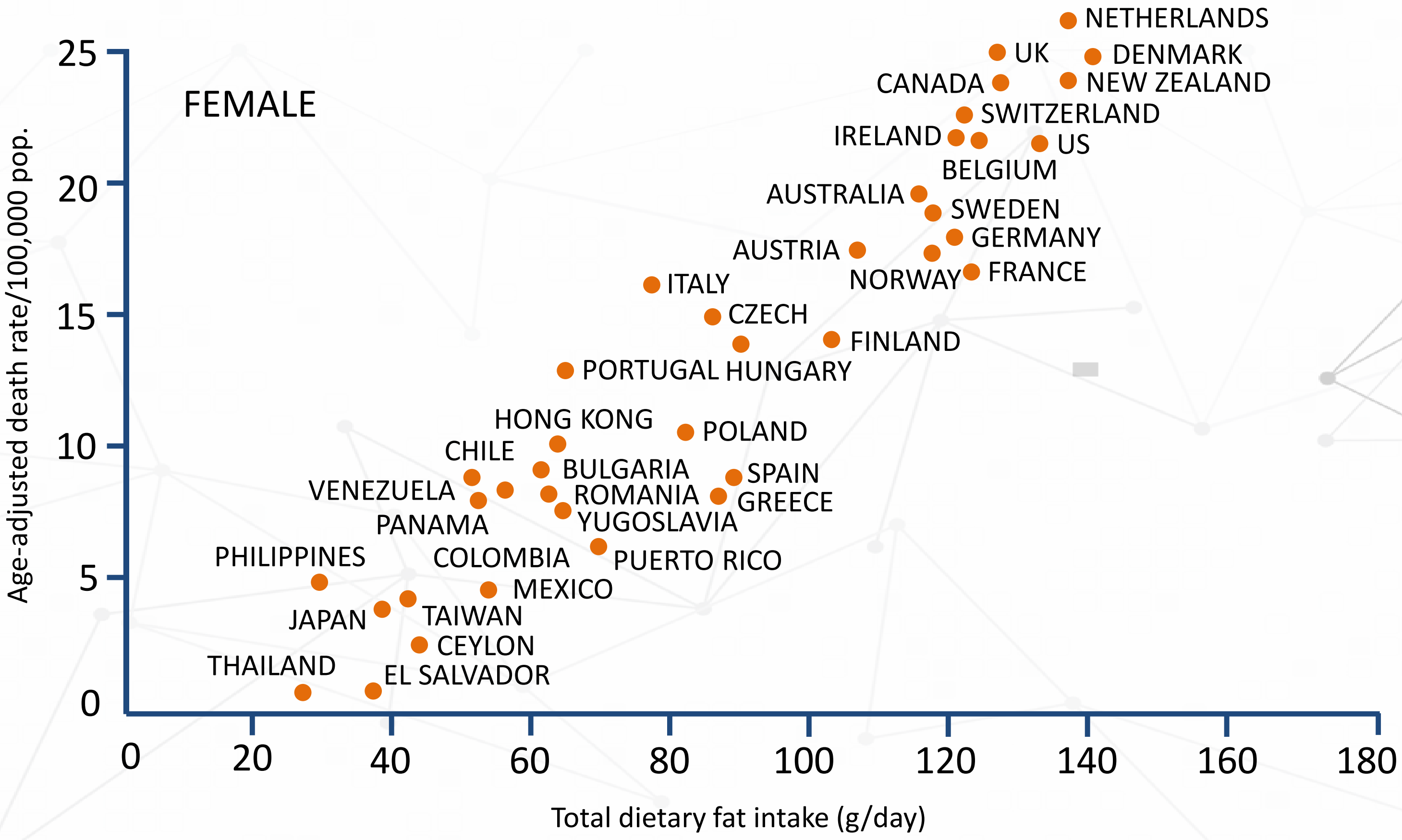
Positive vs. Negative Space



- What is positive to you?
- What is negative to you?
- In the image here, which one do you think is more important?

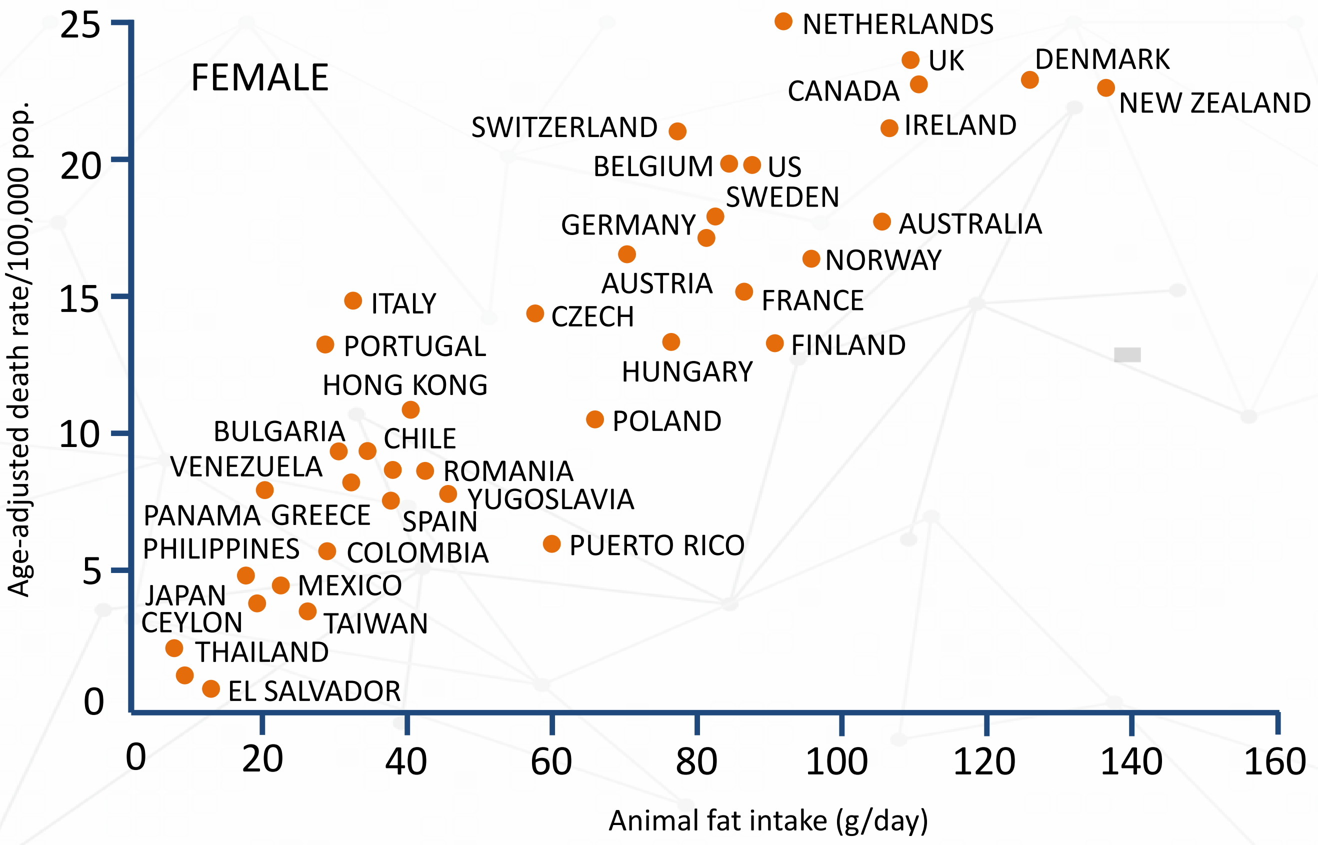
- We have many methods to look for associations and correlations (positive space), for example statistical test.
- We tend to ignore non-associations (negative space).
  - We think they are not interesting/ informative.
  - There are too many of them.
- We also tend to ignore relationship between associations (aka multi-collinearity).

We love to find correlations like this…



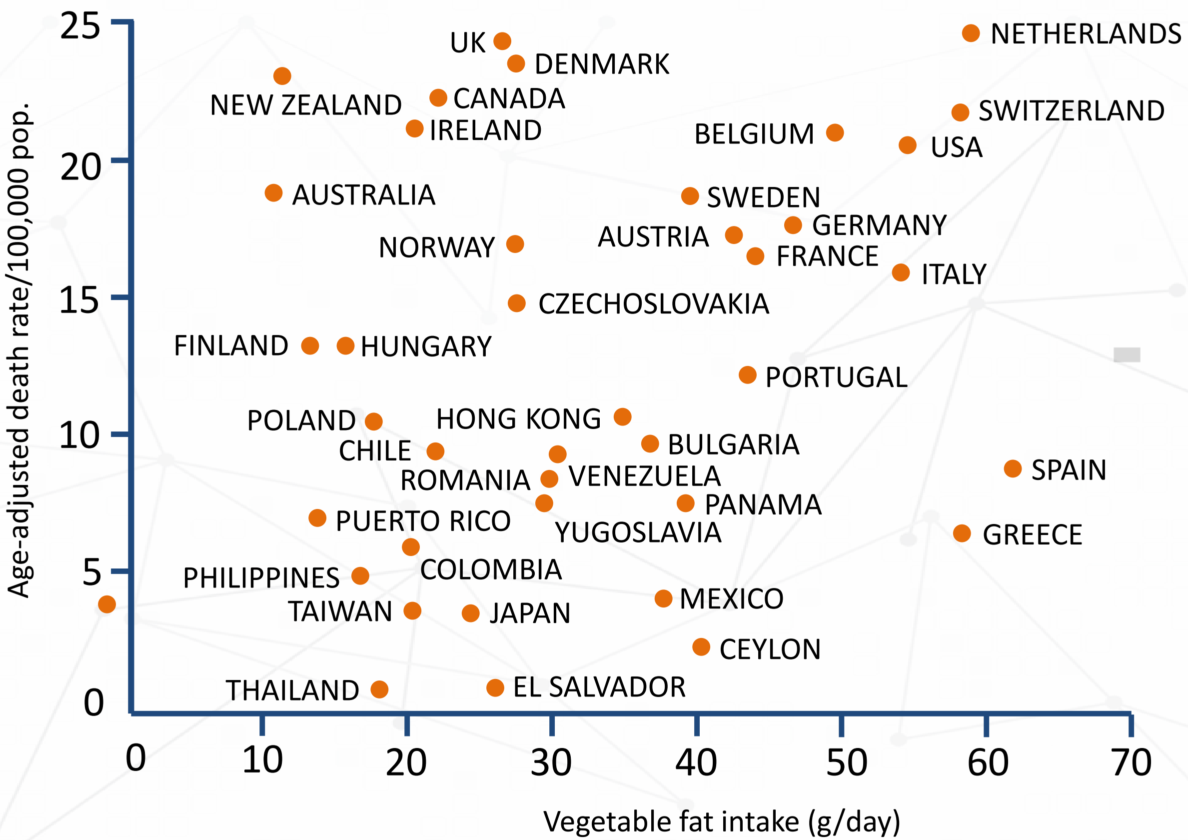
Dietary fat intake correlates with breast cancer.

And like this…(positive)



Animal fat intake correlates with breast cancer.

But not this…(negative)



Plant fat intake doesn't correlate with breast cancer.

But there is much to be gained…

A: Dietary fat intake correlates with breast cancer.

B: Animal fat intake correlates with breast cancer.

C: Plant fat intake doesn’t correlate with breast cancer.

- Given C, we can eliminate A from consideration, and focus on B!
- **You may also conclude that not all fats are bad, and that you may quite liberally eat plant fat.**

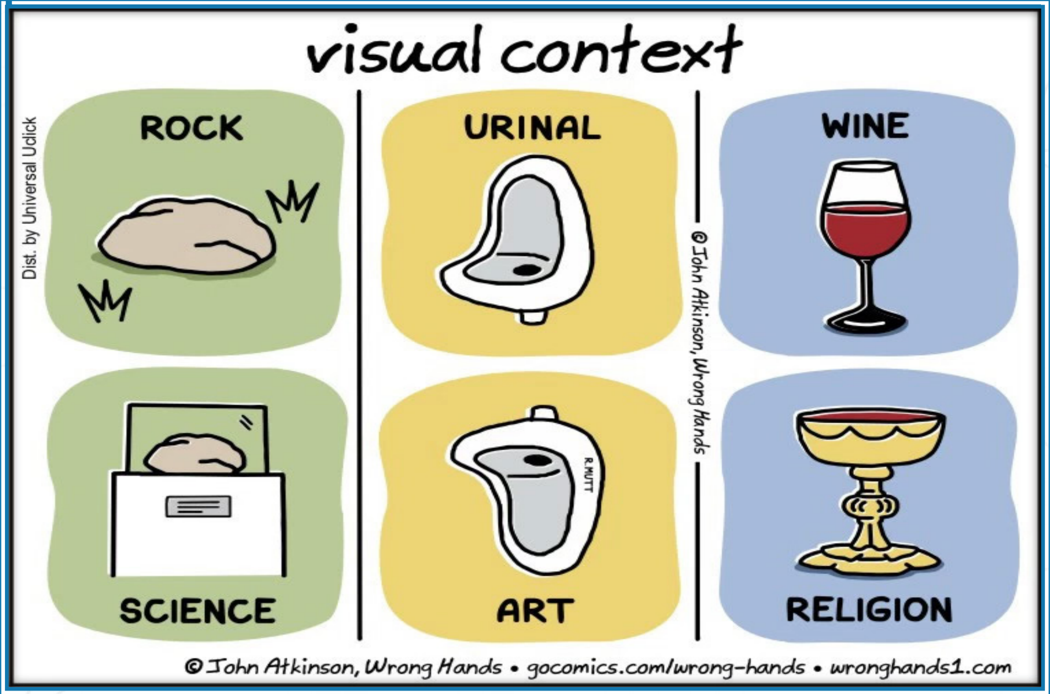
Overlooked Information: Context

BS3033 Data Science for Biologists

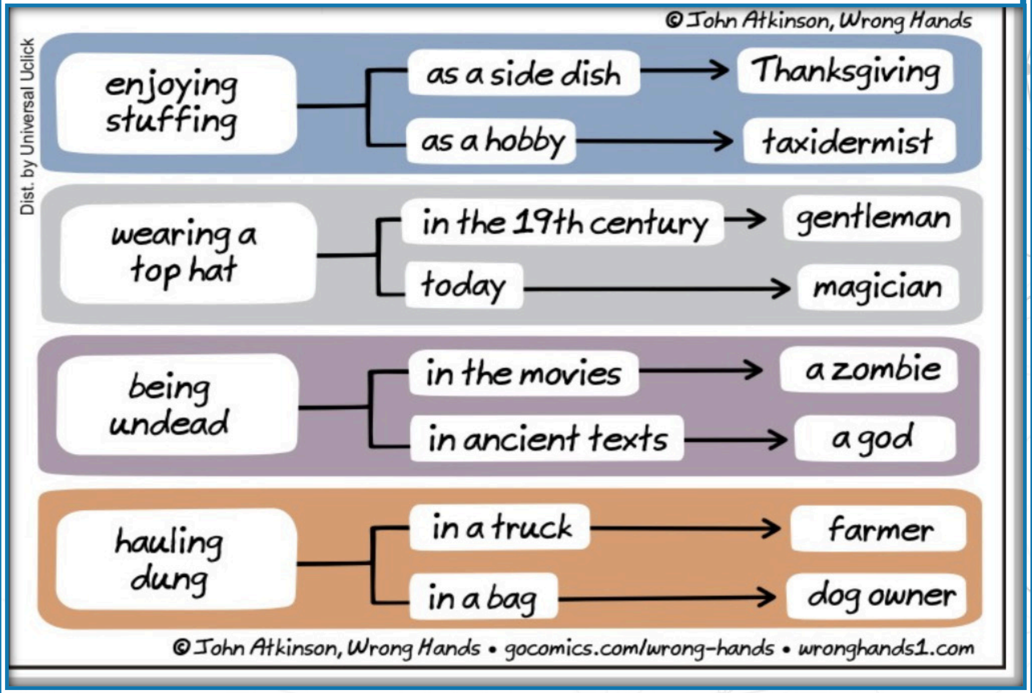
Dr Wilson Goh
School of Biological Sciences

Context

The term ‘context’ is a noun. It is the circumstances that form the setting for an event, statement, or idea, and in terms of which it can be fully understood.



Source: Creative Common License
https://wronghands1.files.wordpress.com/2017/07/visual-context.jpg



Source: Creative Common License
https://wronghands1.files.wordpress.com/2017/02/contextual.jpg

Why is context important in biology?

Gene isoform switching:

- Same gene, but produces different isoforms (splice variants) in different tissues, i.e. a gene functions differently in different parts of the body.
- Refer to lecture notes for link to website for further information.

Human behavior:

- In our typical environment, we are generally well-behaved, well-adjusted individuals.
- In an alternative environment with new rules (e.g. Stanford Prison Experiment), people can behave in extreme ways.

Context

Gene networks:

- Genes do not function independently of each other but rather in pathways and networks.
- When several components of a single pathway are affected, we can generally deduce that this pathway (including the unobserved components) as a whole is important to the phenotype.

Evolution:

- Interplay between genetics and environment (via natural selection).
- In Galapogos, finches varied from island to island (their beaks adapted to the type of food they ate; filling different niches on the Galapagos Islands).
- Refer to lecture notes for link to website for further information.

Context (Biological Complexes)

Postulate: The chance of a protein complex being present in a sample is proportional to the fraction of its constituent proteins being correctly reported in the sample. Suppose proteomics screen has 75% reliability; a complex comprises proteins A, B, C, D, E; and screen reports A, B, C, D only but not E.

- Complex has 60% (= 0.75 * 4 / 5) chance to be present.
- The unreported protein E also has ≥ 60% chance to be present, as presence of the complex implies presence of all its constituents **(improving coverage and recover missing proteins).**
- Each of the reported proteins (A, B, C, and D) individually has 90% (= 100% * 0.6 + 75% * 0.4) chance of being true positive, whereas a reported protein that is isolated has a lower 75% chance of being true positive **(removing noise)**.

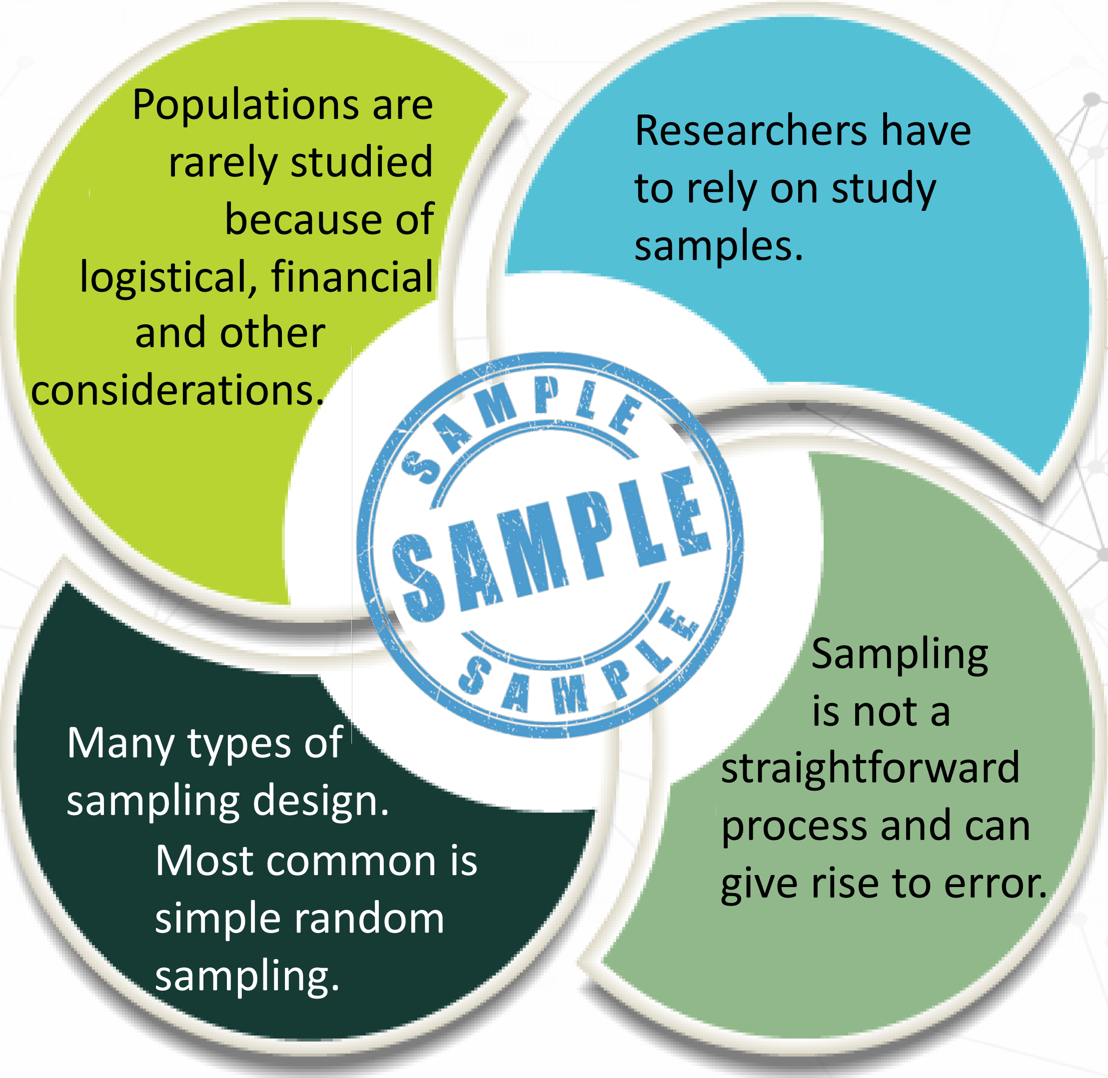
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Sampling Techniques

BS3033 Data Science for Biologists

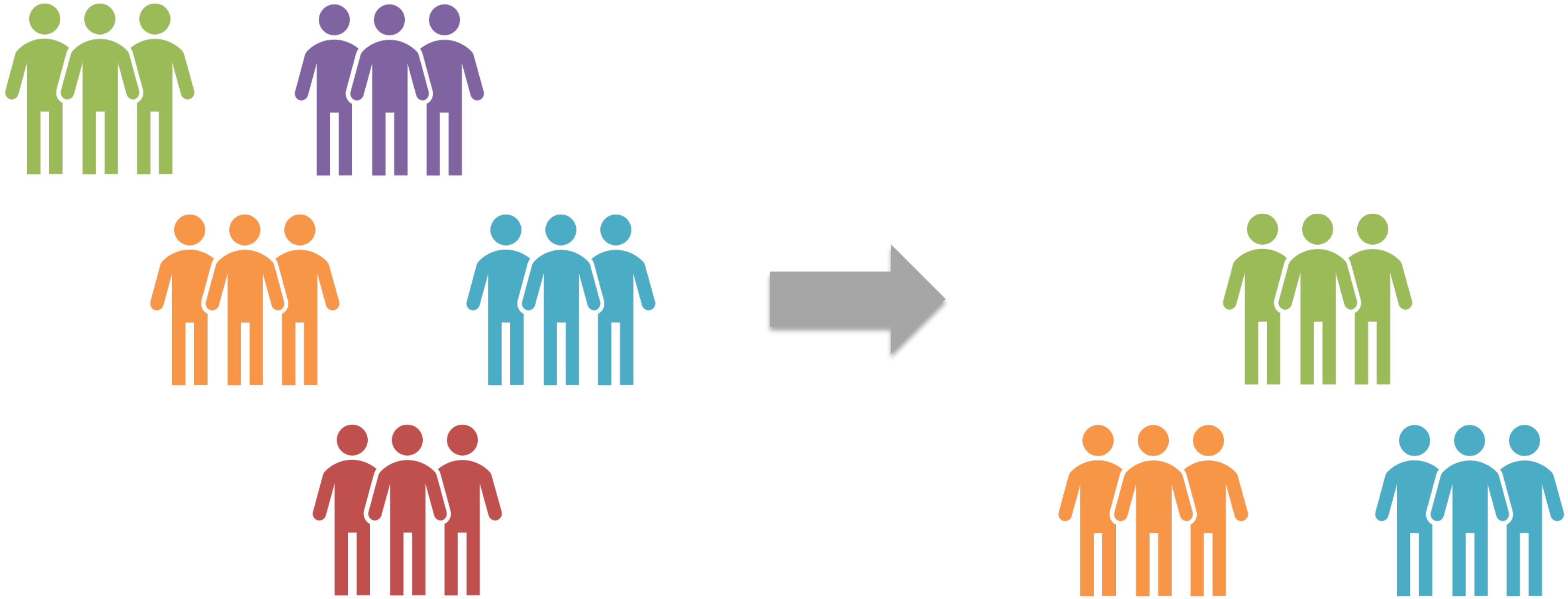
Dr Wilson Goh
School of Biological Sciences

Why do we want to sample?



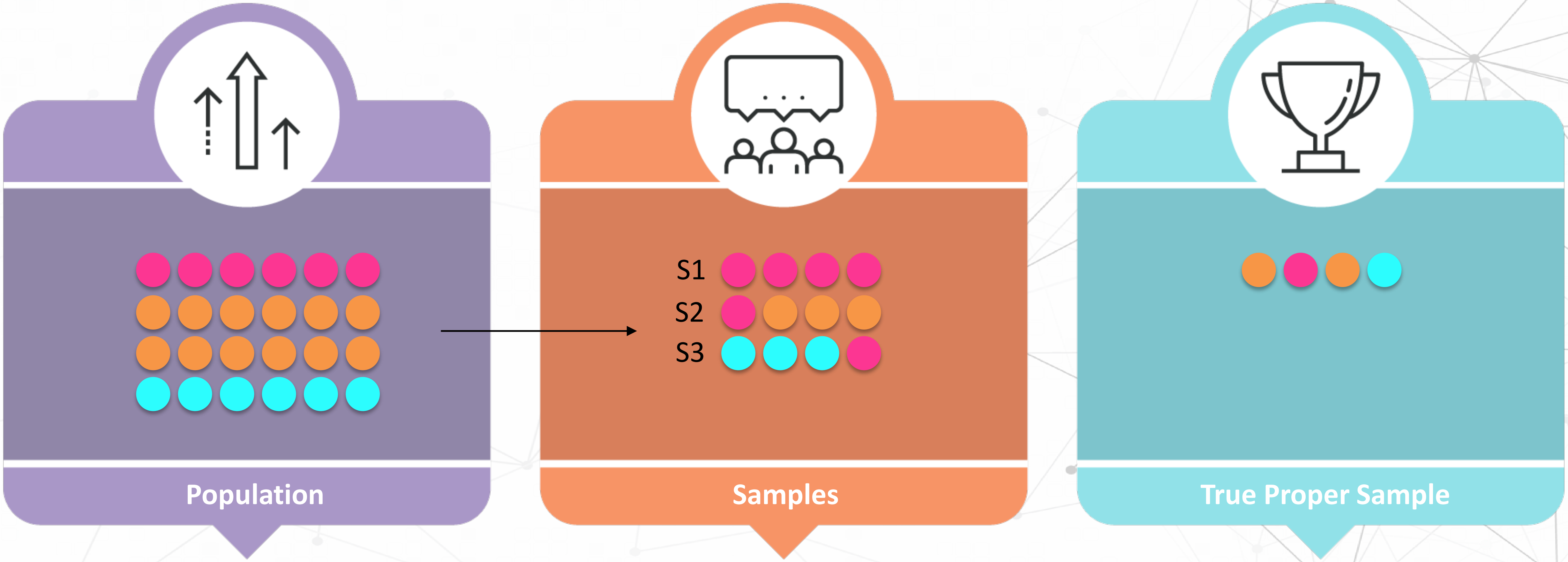
An Example of Sampling Error

An example of sampling (systematic) error.



Javier needs 9 participants for his study. But he is too lazy to collect 9. So, he calls 3 of his friends, and asks them to include their parents so that he can easily get 9. Is this sufficiently random? What kind of problems do you think this can cause?

Random Error in Sampling



Here we have 24 people, of which ¼ are Indian, ¼ are Taiwanese and ½ are Chinese. Suppose if I randomly sample 4 people 3 times each. Do my samples represent the population? They don't because by random chance, we may observe samplings that have a different distribution to the population.

Sampling Methods

Simple Random Sampling:

Each subject in the population is equally likely to be selected.

Stratified Sampling:

First divide the population into homogenous strata (subjects within each stratum are similar, across strata are different), then randomly sample from within each strata.

Sampling must take into account the various groups that need to be included in order to better resemble the population.

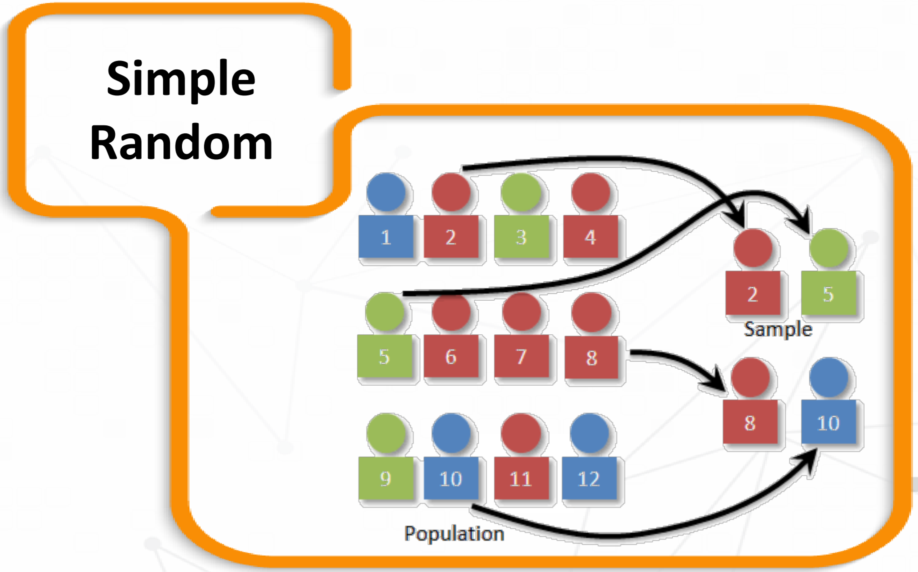
Cluster Sampling:

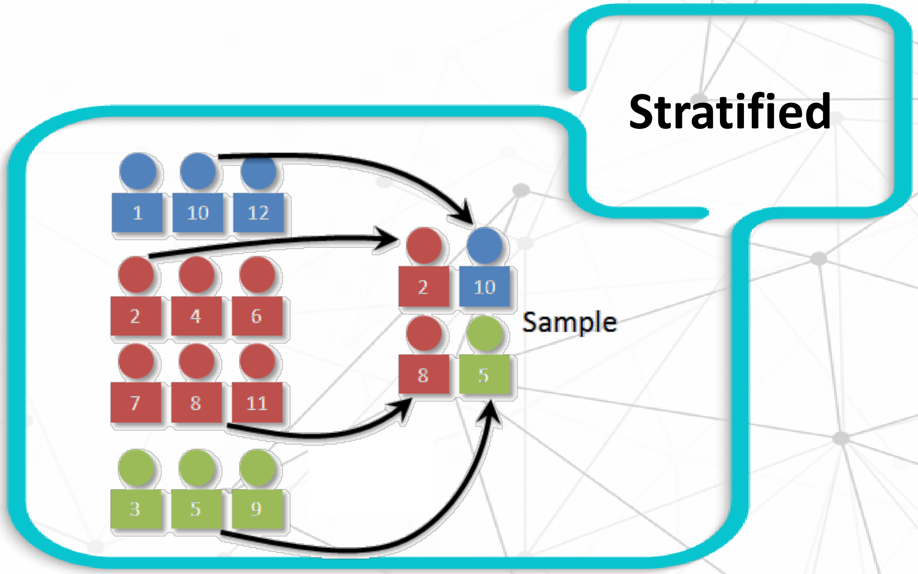
First divide the population into clusters (subjects within each cluster are non-homogenous, but clusters are similar to each other), then randomly sample a few clusters, and then randomly sample from within each cluster.

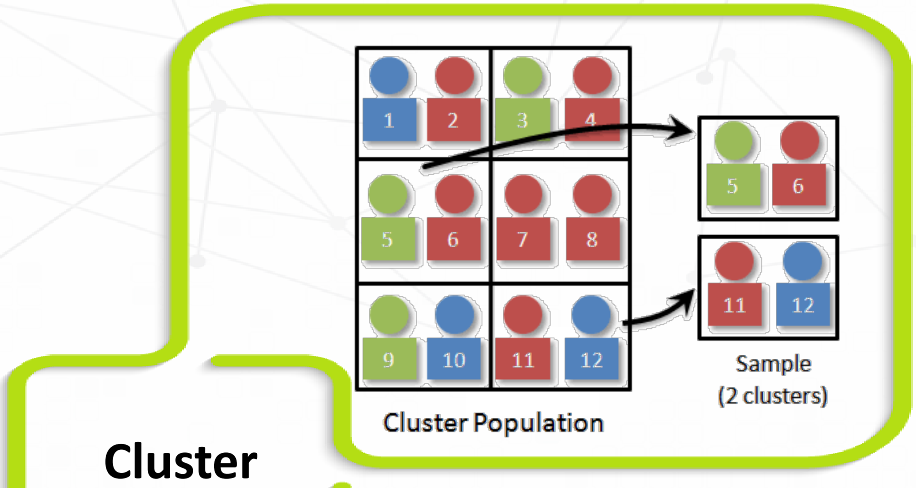
Systematic Sampling:

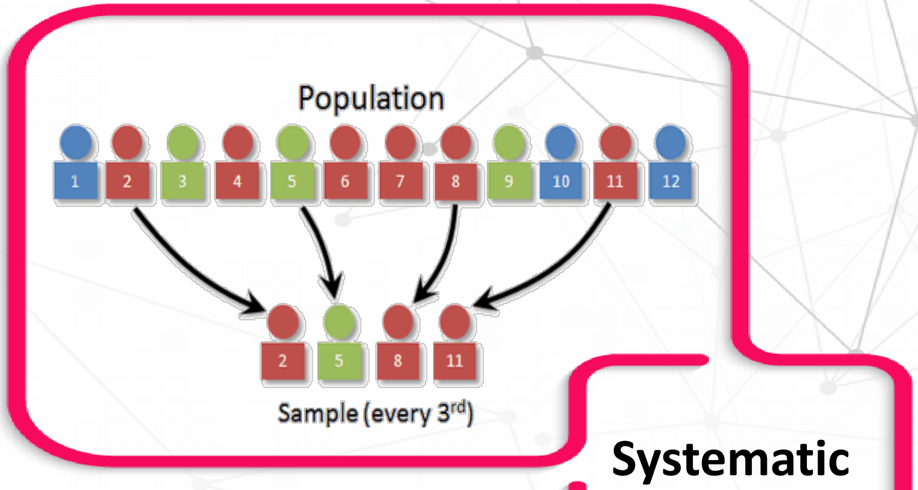
Every kth individual is selected.

Sampling Methods









Refer to online resources on how to implement these in R:

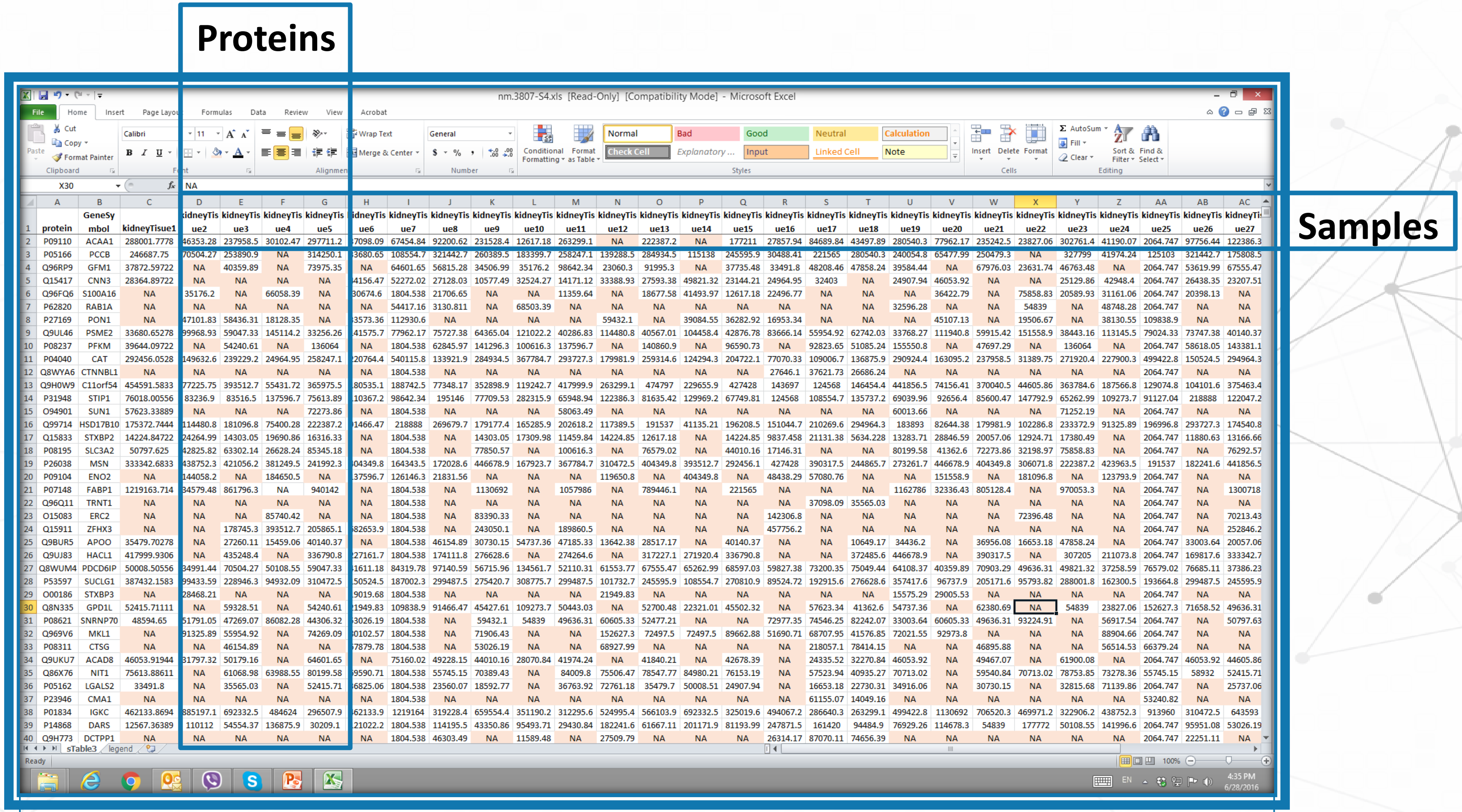
http://faculty.elgin.edu/dkernler

Massaging the Data: Normalisation

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

An Example of High-dimensionality (Multivariate Data)



So where do we begin with analysing such big and complex data?

Normalisation

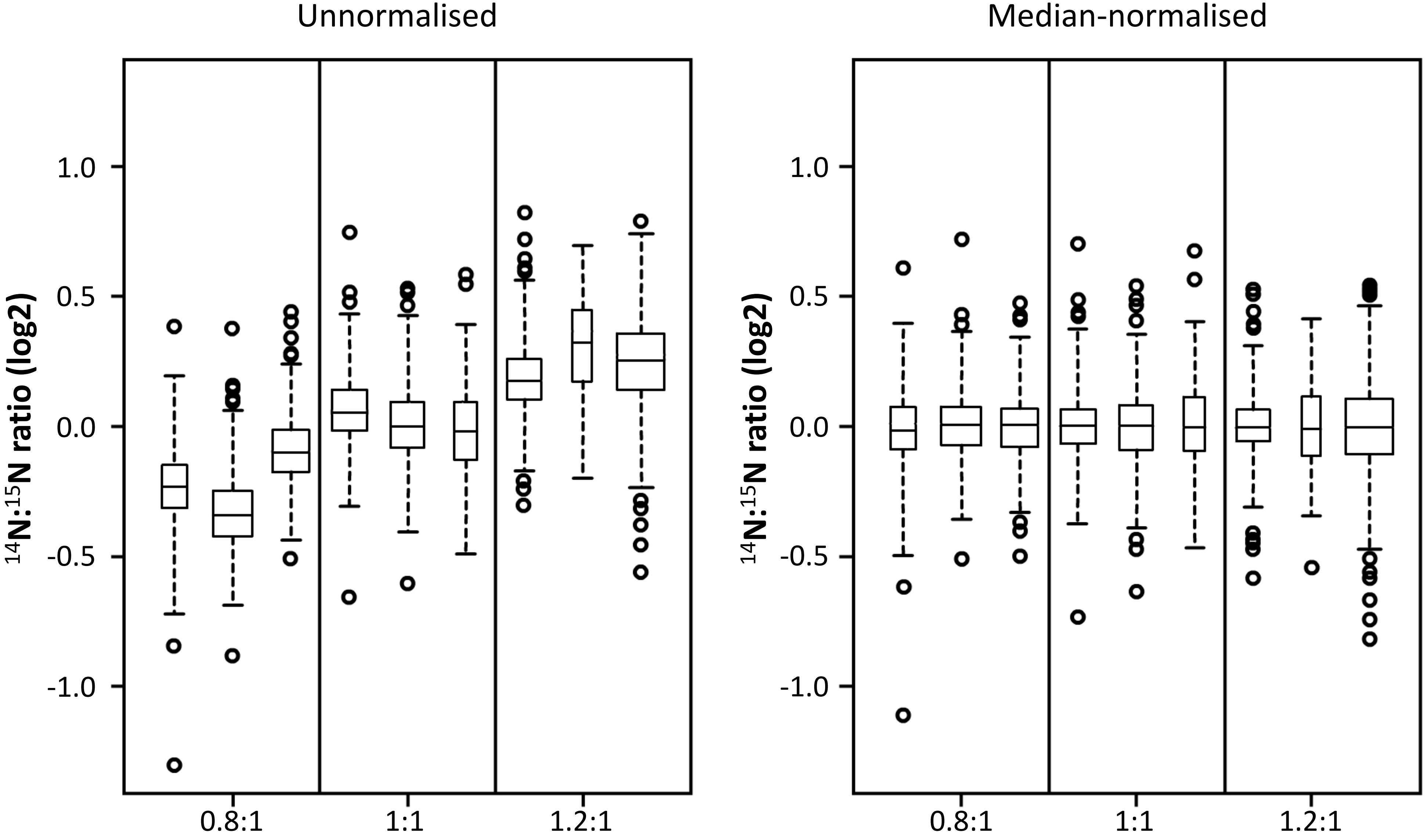
Normalisation means adjusting values measured on different scales to a notionally common scale e.g. resetting values to between 0 to 1.

Normalisation can also mean to bring different probability distributions into alignment with each other (e.g. making two skewed distributions more similar in shape to each other).

Why do it? Suppose if one variable is 100 times larger than another (on average), then our model may be better behaved if you normalise/ standardise the variables to be approximately equivalent. It also prevents variables with high values from dominating the model.

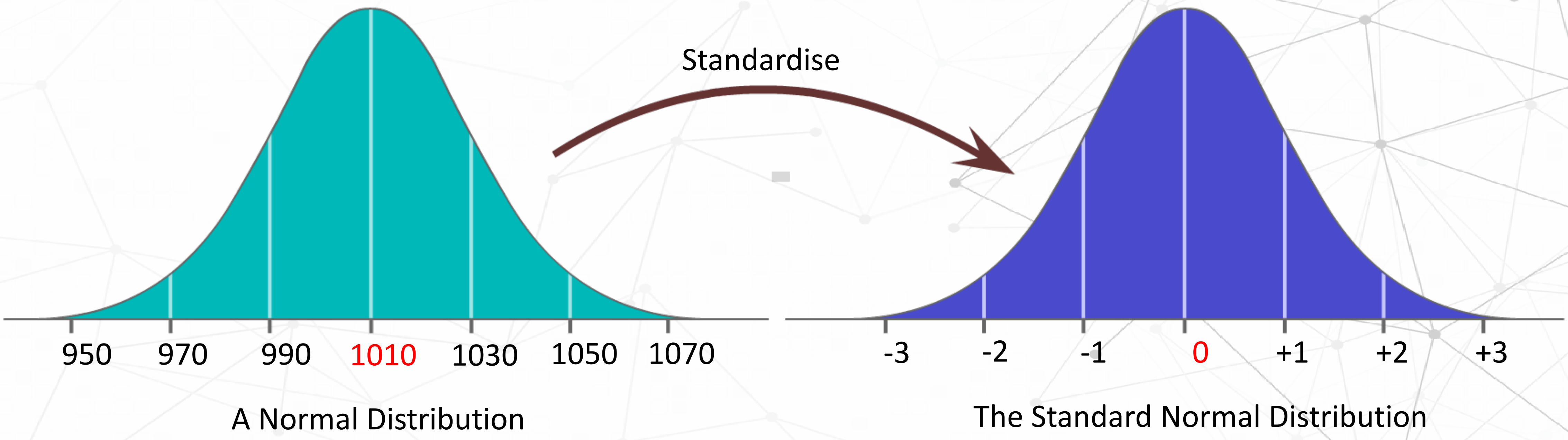
Visualising Normalisation

What is normalisation?



Z-normalisation

The z-score is the most common way of normalising multivariate data. Recall (for one variable):



$$z = \frac{x - \mu}{\sigma}$$

z is the "z-score" (Standard Score)
x is the value to be standardised
μ is the mean
σ is the standard deviation

Z-normalisation

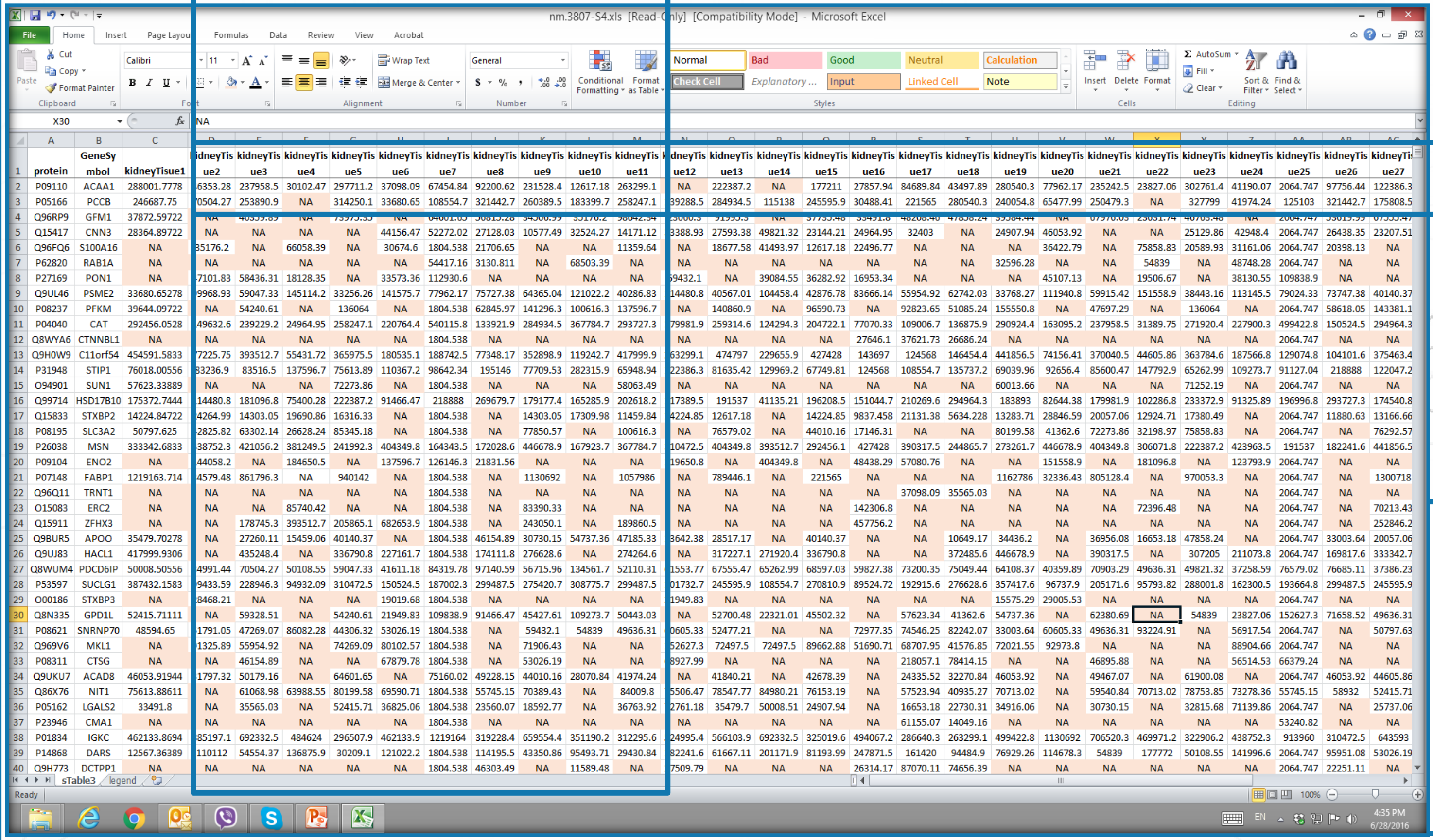
We may convert all observations across n variables into z-scores with a mean of 0, and s.d. of 1.

The z-score for each observation represents how many s.d. from the mean it lies away from.

We have a problem though. Do we normalise each observation based on the mean and s.d. of each gene (genewise), or do we normalise each observation based on the mean and s.d. of each sample (samplewise)?

Z-normalisation

Normalise by samples?



Normalise by variables/genes?

You should not normalise by genes first because we do not have assurance that the data distributions between different samples are comparable. So, we should normalise by samples first.

Which do you think is more correct?

Z-normalisation

Use when variables are normally distributed.

If variables are not normally distributed, then z-score conversion can mislead.

Good statistical properties and compatible with many parametric type tests.

Linear Scaling

For the entire dataset, find the minimum value X_{min}, and the maximum value, X_{max}.

For all observations, subtract by Xmin and divide by the delta of $X_{max} - X_{min}$.

This conversation will bound the data values between 0 to 1.

It shifts all data points by a fixed magnitude but does not change the data distribution, hence "linear" scaling.

$$X_{i,\,0\text{ to }1} = \frac{X_i - X_{Min}}{X_{Max} - X_{Min}}$$

Where:

X_i = Each data point i

X_{Min} = Minima among all the data points

X_{Max} = Maxima among all the data points

$X_{i,\,0\text{ to }1}$ = Data point i normalised between 0 and 1

Linear Scaling

Linear scaling can be modified to obtain a more "centralised" dataset, with 0 as the center point.

Subtract the mean of X_{min} and X_{max} from each observation.

And divide by its delta/2.

$$X_{i,\,-1\text{ to }1} = \frac{X_i - \left[\dfrac{X_{Max} + X_{Min}}{2}\right]}{\left[\dfrac{X_{Max} - X_{Min}}{2}\right]}$$

Where:

X_i = Each data point i

X_{Min} = Minima among all the data points

X_{Max} = Maxima among all the data points

$X_{i,\,-1\text{ to }1}$ = Data point i normalised between -1 and 1

Linear Scaling

Use when data do not meet normality assumption, for example heavily skewed.

Use when you want to standardise data to a common interval, for example [0,1].

Use when variable distributions are more or less similar to each other.

Rank-normalisation

It is “quietly incorporated” in many non-parametric tests. Absolute values are converted into ranks by assigning values of 1 to R to observations (where R is total sample size). We can further convert the ranks into standardised values of zero to 1.

$$x = \frac{r - 1}{R - 1}$$

Original Index	-2	-1	0	1	2	= i
Converted to Rank	1	2	3	4	5	= r R = max (r) = 5
Normalised Rank	0	¼	½	¾	1	$x = \frac{r-1}{R-1}$

Where x is the normalised rank, r is the assigned rank, and R is the highest rank value. This approach towards rank normalisation assumes that rank can be normalised as a quantitative variable. We may use the rank-normalised values as quantitative values for use with other statistical tests or calculation of distance metrics.

Quantile Normalisation

Quantile normalisation is a technique for making two distributions identical in statistical properties.

Raw data

2	4	4	5
5	14	4	7
4	8	6	9
3	8	5	8
3	9	3	5

Order values within each sample (or column)

2	4	3	5
3	8	4	5
3	8	4	7
4	9	5	8
5	14	6	9

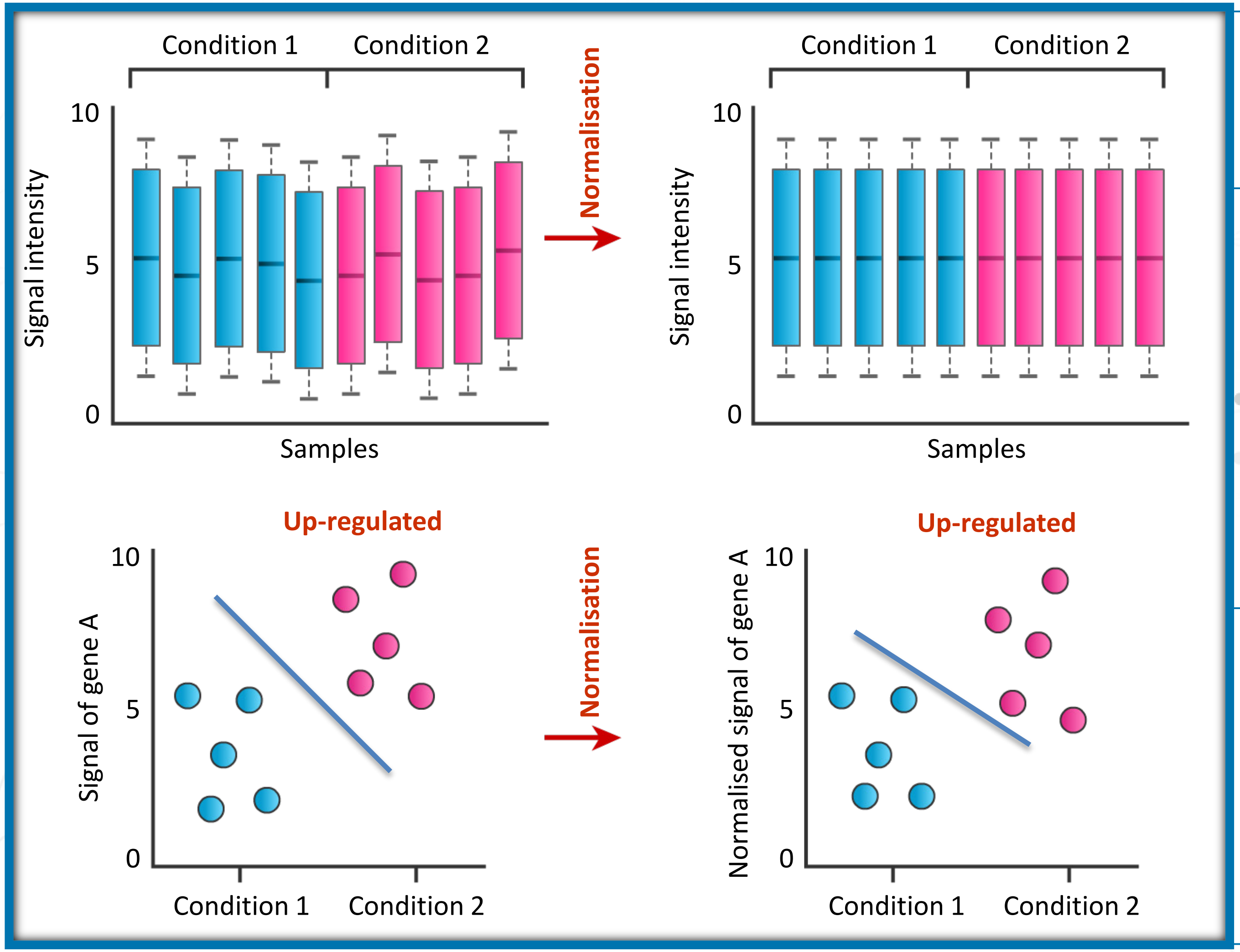
Average across rows and substitute value with average

3.5	3.5	3.5	3.5
5.0	5.0	5.0	5.0
5.5	5.5	5.5	5.5
6.5	6.5	6.5	6.5
8.5	8.5	8.5	8.5

Re-order averaged values in original order

3.5	3.5	5.0	5.0
8.5	8.5	5.5	5.5
6.5	5.0	8.5	8.5
5.0	5.5	6.5	6.5
5.5	6.5	3.5	3.5

Normalisation



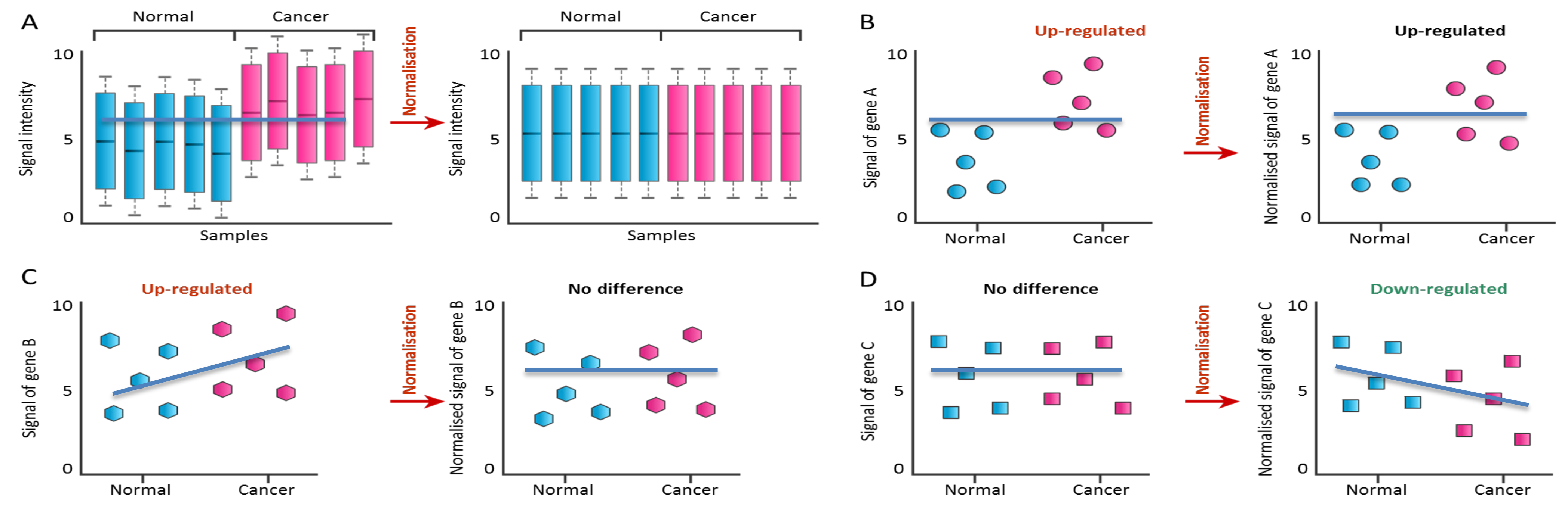
Normalisation works well if two sets of distributions are not too different from each other.

The common assumptions for normalisation are reasonable if similar global signal distributions are seen in the different conditions. In such cases, normalisation has little influence on the interpretation of expression data.

Source: Wu et al. 2014

Normalisation

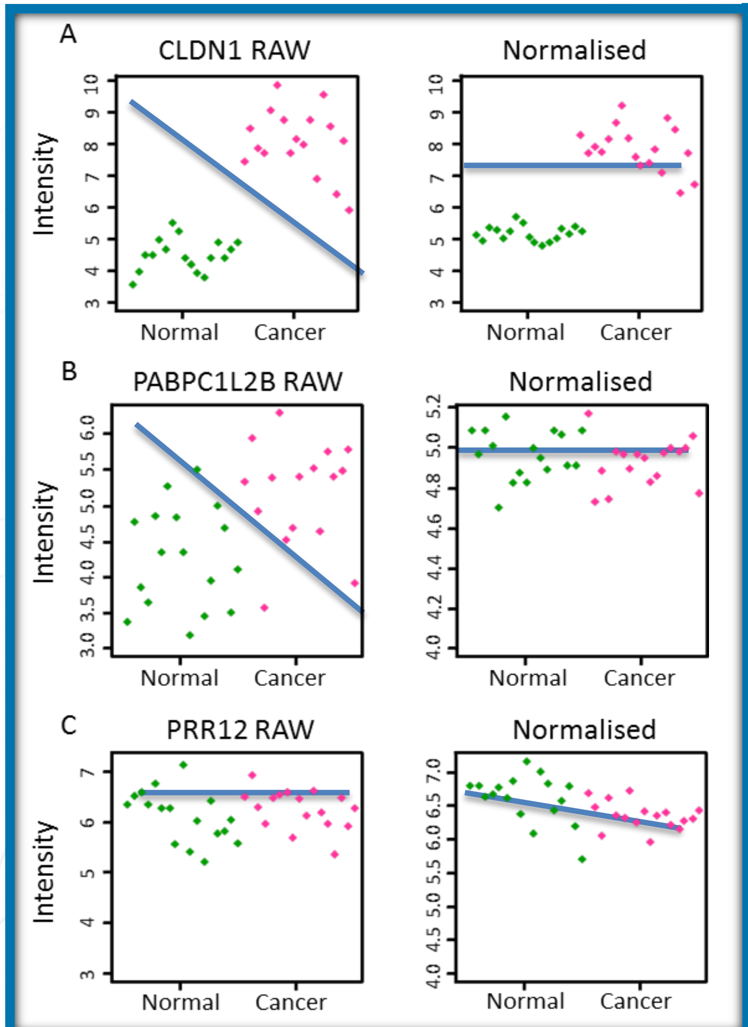
Normalisation does not work well if two sets of distributions are very different from each other.



(A) The yellow and blue samples represent cancer samples and normal samples with large differences in signal patterns. The signal intensities were normalised across all arrays to have the same distribution. (B) A gene shows strong up-regulation in cancer samples in the raw signals. Though normalisation may reduce the size of the difference, this gene could be still selected as a differential up-regulated gene after normalisation. (C) A gene shows moderate upregulation in cancer samples in the raw signals. After normalisation, it cannot be identified as a differentially expressed gene. (D) A gene shows little difference in expression between cancer samples and normal samples in the raw signals. After normalisation, it may be identified as a differential downregulated gene.

Source: Wu et al. 2014

What happens to each of these genes after normalisation?

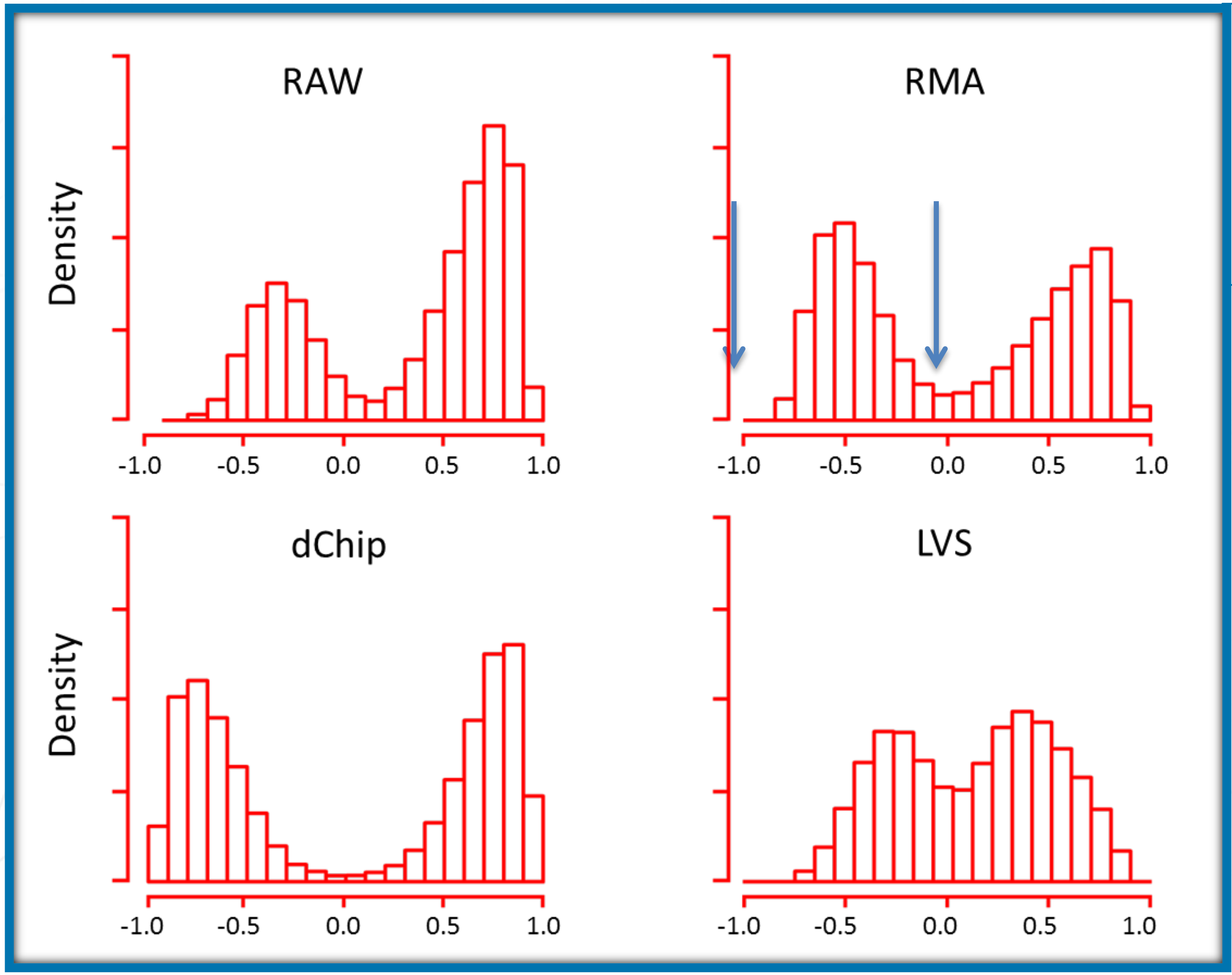


Effect of RMA normalisation on expression directions in the mRNA colon34 dataset.

Colon 34 is a pair-matched dataset in which the normal samples were taken from the same subjects as the cancer samples.

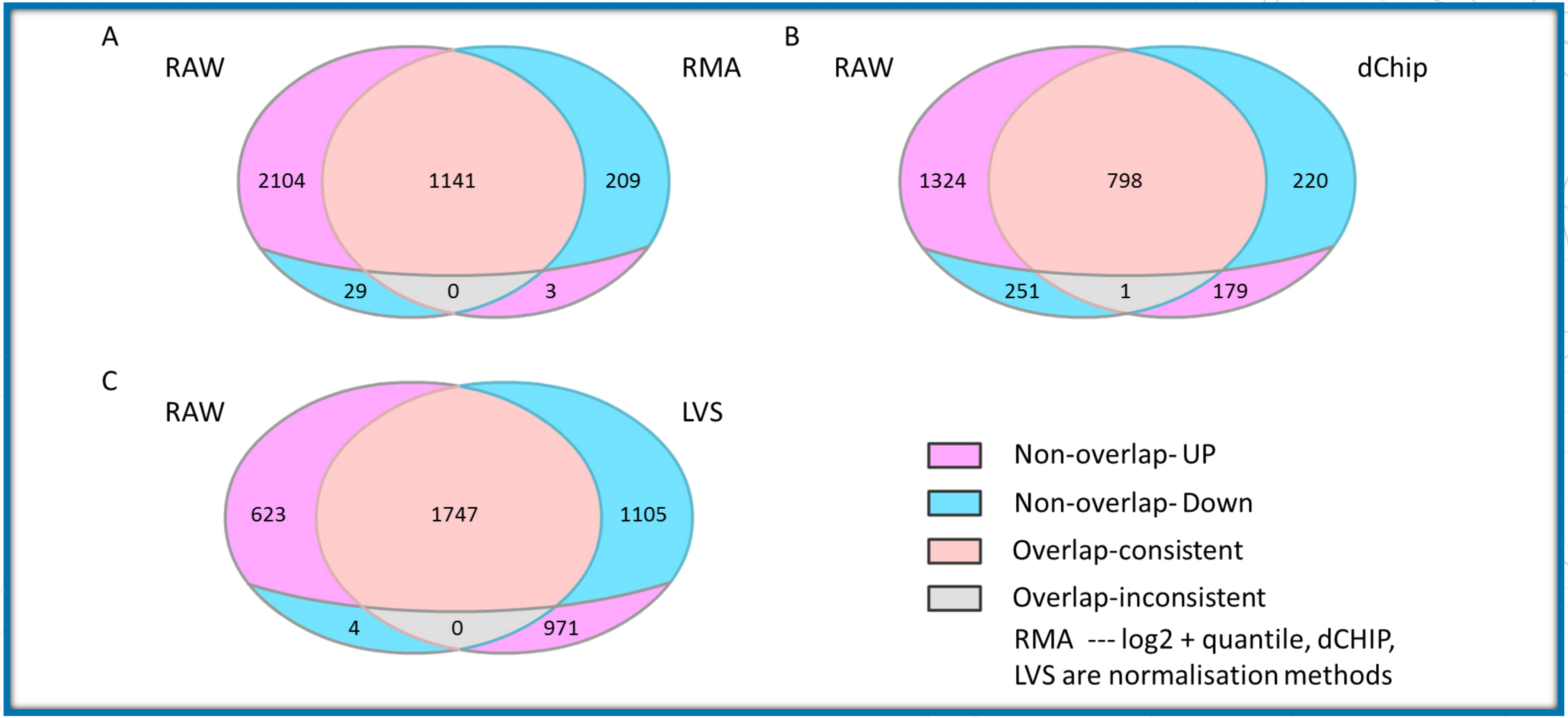
Source: Wu et al. 2014

Normalisation Leads to Erosion of Signal



The density distributions of pair-wise Pearson correlation coefficients before and after normalisation of the mRNA colon34 dataset.

Normalisation Can Lead to Disagreements on the DEGs



Source: Wu et al. 2014

NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

Reproducibility and Independent Corroboration

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

Why are these important?

Reproducibility

"Now you see it, now you don't".

Do the same experiment twice, you expect to see the same results.

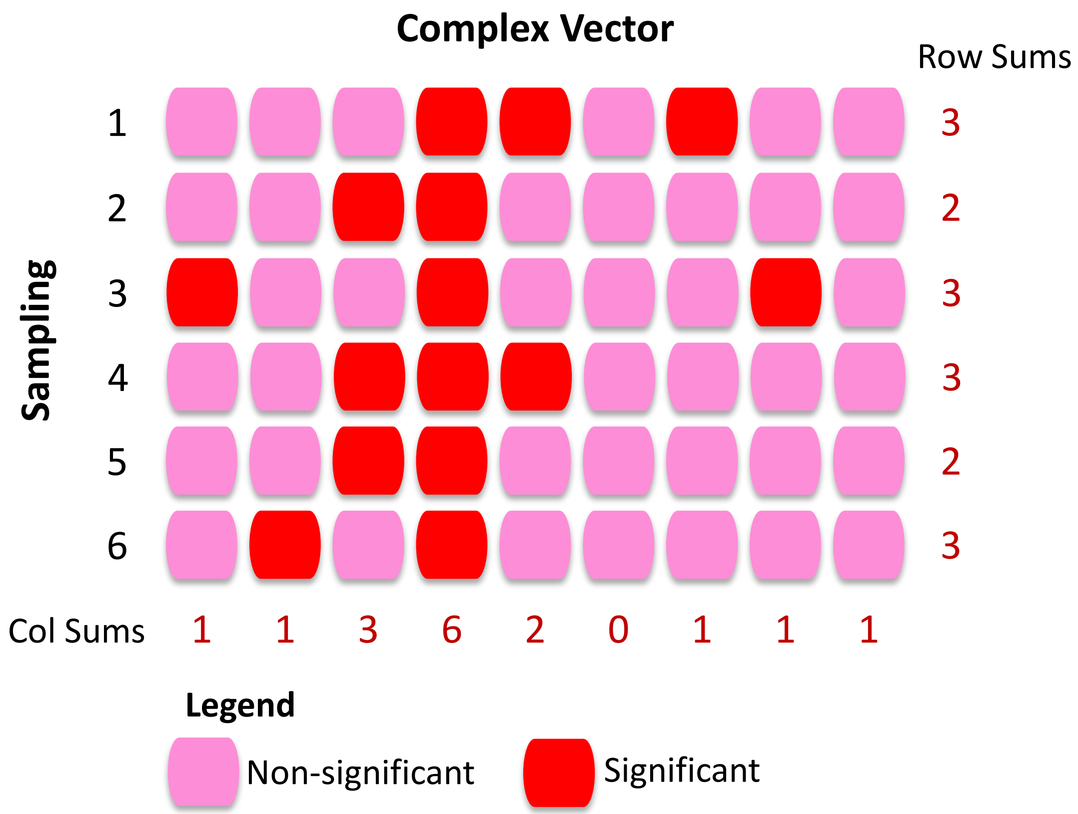
Independent Corroboration

Use multiple threads of independent evidences (each imperfect on their own), to derive increased confidence.

"Confirm, double confirm, triple confirm…".

Check Reproducibility using Resampling

The binary matrix is useful for comparing stability and consistency of significant features produced by some feature-selection method.



Rows represent samplings and columns represent complexes/genes/proteins.

Red are significant features (1) while pink are non-significant (0).

Source: Goh & Wong, Design principles for clinical network-based proteomics. Drug Discovery Today, 2016

Independent Corroboration

Statement: Gene X causes Disease Y		
Experiment	**Result**	**Support?**
Genomics	Gene is reattached to a more active promoter. But we do not know if the gene is expressed.	Maybe
Transcriptomics	mRNA X is high. Many copies of mRNA, but many different splice forms.	Maybe
Proteomics	Protein X is up-regulated in Y. But only one unique peptide.	Maybe
Each evidence is imperfect. But together, they give us more confidence.		

Meta and Mega Analyses

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

Big and Small Data

Data science isn’t necessarily concerned only with big data. Small data is also important. But what’s the difference?		
	Big Data	**Small Data**
Data Condition	Usually unstructured, not ready for analysis	Usually structured, ready for analysis
Location	Cloud, Offshore, SQLServer, etc.	Database, Local PC
Size	Over 50k variables, over 50k individuals, random samples, unstructured.	File that is in a spreadsheet, that can be viewed on a few sheets of paper.
Purpose	No intended purpose.	Intended purpose for data collection.

Meta-analysis

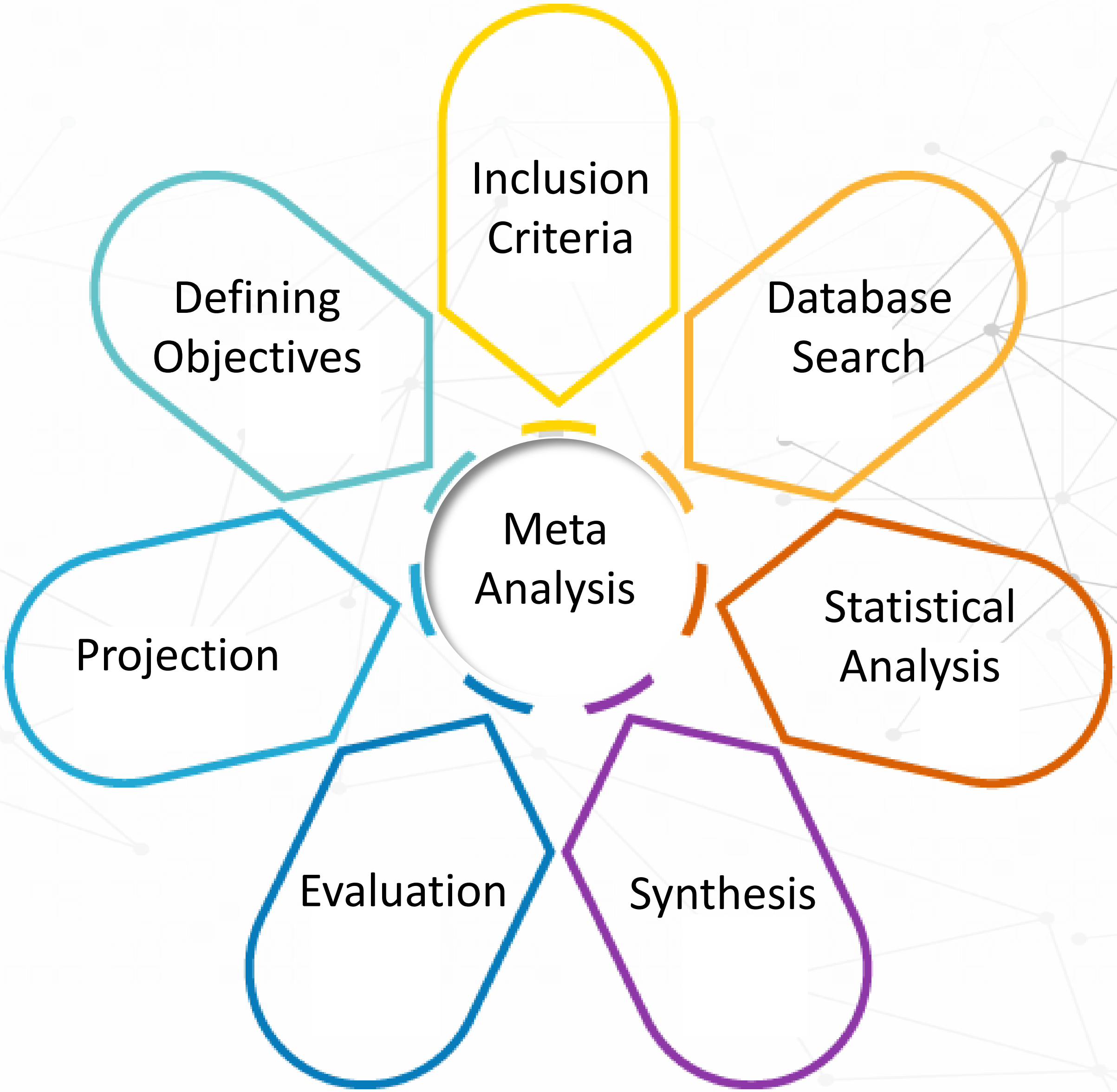
Meta-analysis is a statistical procedure that integrates the results of several independent studies.

It can be a very useful method to summarise data across many studies, but requires careful thought, planning and implementation.

A meta-analysis goes beyond a literature review.

Is this equivalent to big data?

Considerations for Meta Analysis



Between Small and Large Data

D1 → Standard/Small Data Analysis

In series:

D1 D2 D3 → Big Data (Mega) Analysis

In parallel:

D1 → Small Data Analysis

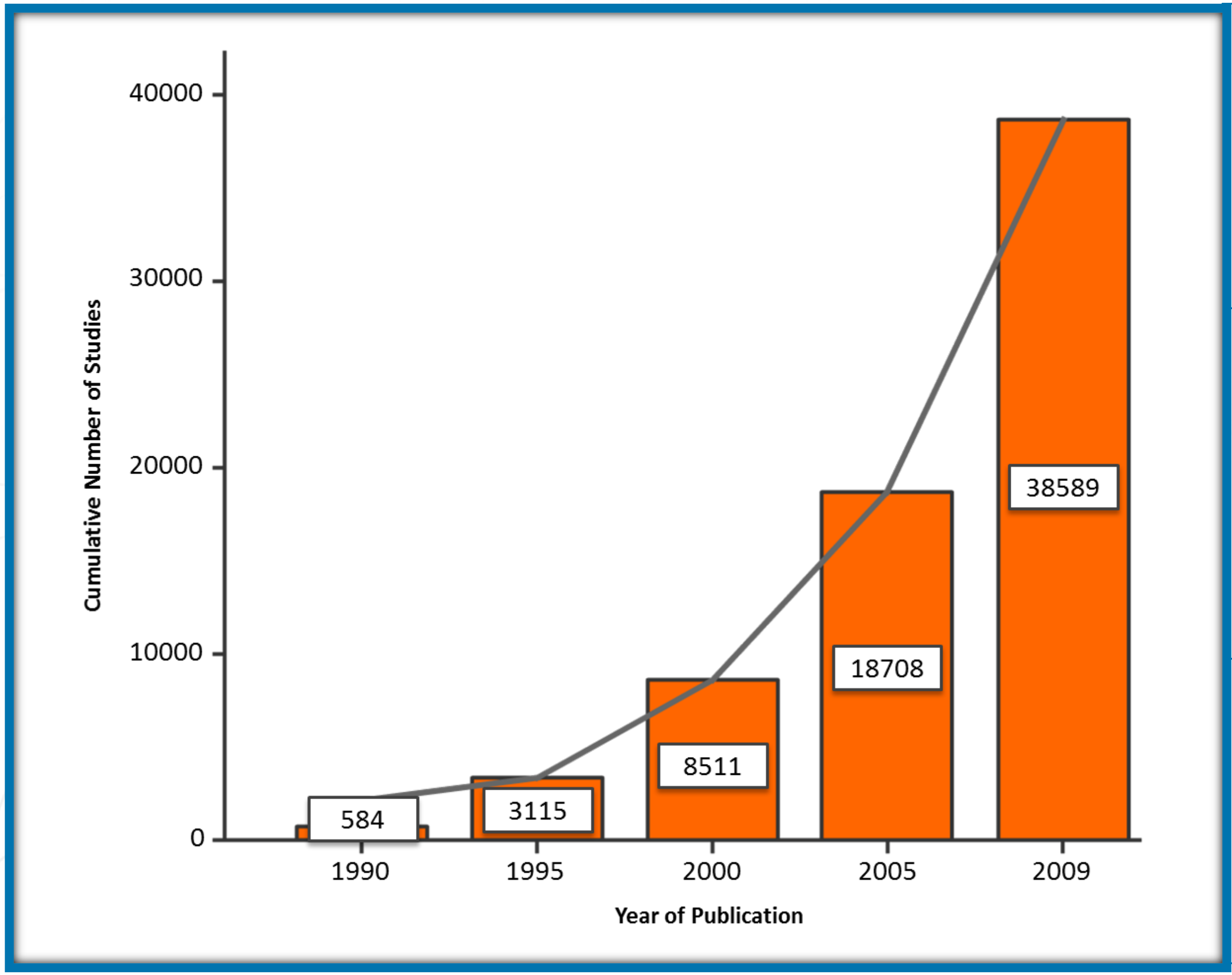
D2 → Small Data Analysis

D3 → Small Data Analysis

Integration

“Meta-analysis”

Meta-analysis is Increasingly Common



Cumulative number of publications about meta-analysis over time, until 17 December 2009 (results from Medline search using text "meta-analysis").

This upward trend is also partly because of the larger amount of existing data available to us. And not simply because meta is necessarily seen as more important.

Source: Haldich, Hippokratia. 2010

Papers Discussing Meta-analysis

Papers for discussion (feel free to add more):

- Berman and Parker, Meta-analysis: Neither quick nor easy, BMC Medical Research Meth, 2002.
- Haidich, Meta-analysis in medical research, Hippokratia, 2010.
- Nakagawa et al, Meta-evaluation of meta-analysis: ten appraisal questions for biologists, BMC Biology, 2017.

Questions for thought:

- Do the various papers agree with each other?
- What are some simple examples of finding consensus amongst the individual datasets?
- “Meta-analysis” is less powerful. Do you agree?

Example of Mega-analysis (aka big data analysis or data pooling)

Papers for discussion (feel free to add more):

- Hess et al. Transcriptome-wide mega-analyses reveal joint dysregulation of immunologic genes and transcription regulators in brain and blood in schizophrenia, Schizophr Res, 2016.
- This paper puts together 9-11 datasets to generate pooled data for deriving markers for schizophrenia.

Questions for thought:

- Do you foresee any problems? Comment on their methodology and critique their findings.
- You may also relate what Hess et al did and whether they should also have performed a meta-analysis as well. What should they expect to see?
- How would you have designed the analysis?

Relating Meta-analysis and Big Data Analysis

	Meta-analysis	Big Data
Addresses	Heterogeneity	Power
What it is	Systematic review with synthesis of findings	Integration-based knowledge discovery
How to do it?	No set protocol	No set protocol
Relies on	Consistency	Strength of larger sample size (pooling)
Uses	Many datasets (in parallel)	Many datasets (in series)
Achilles heel	Data selection bias; not being "expansive" enough; many conflicting results; false negatives	Not addressing dataset; heterogeneity issues; false positives

Summary

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

Final Word

The collected data can be sufficient and representative or not . . .

The statistical calculations can be correct or incorrect. . . .

But even when the data are good and the calculations are correct . . the numbers are open to different interpretation . . . hence should not be taken as undeniable "gospel truth".

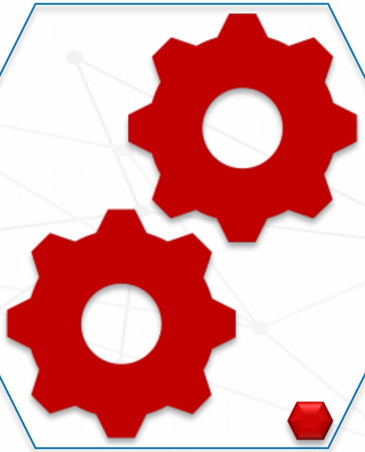
Wilko Dijkhuis

It is so easy to make bad inferences with data… there's a creative part of understanding quantitative data that requires a sort of artistic or creative approach to research.

Nate Bolt

What have we learnt?

Be wary of erroneous preconceived notions.



Mechanical application of statistical and data mining techniques often does not work.

Understand the statistical and data mining tools, and the problem domain. Know how to logically exploit both.

Key Takeaways from this Topic

1. Normal distribution, CLT, IID, Proper design of experiment (Inclusion Criteria, Simpson's Paradox, Bias and Fallacies and Batch Effects), and Domain-specific laws are the common forgotten assumptions in research design.

2. Non-associations and Context are the commonly overlooked information in research design.

3. Sampling must take into account the various groups that need to be included in order to better resemble the population. Simple random sampling, Stratified sampling, Cluster sampling, Systematic sampling are some of the sampling techniques used in research design.

4. In statistics and applications of statistics, normalisation can have a range of meanings. In the simplest cases, normalisation of ratings means adjusting values measured on different scales to a notionally common scale, often prior to averaging.

5. Reproducibility is the closeness of the agreement between the results of measurements of the same measurand carried out under changed conditions of measurement. Independent corroboration is evidence that supports a proposition already supported by initial evidence, therefore confirming the original proposition.

6. Meta analysis is a statistical method of combining the results of independent studies. It uses summary data from groups of people rather than data from individual subjects. In contrast, mega analysis refers to a technique of summarising the results of independent studies using data from the individual subjects.

References (Recommended)

Context

- Goh WWB, Wong LS. Integrating networks and proteomics: moving forward. Trends in Biotechnology, 34(12):951-959, Dec 2016.

Batch Effects

- Goh WWB, Wang W, Wong LS. Why batch effects matter in omics data, and how to avoid them. Trends in Biotechnology, S0167-7799(17)30036-7, Mar 2017.
- Goh WWB, Wong LS. Protein complex-based analysis is resistant to the obfuscating consequences of batch effects --- A case study in clinical proteomics. BMC Genomics, 18(Suppl 2):142, Mar 2017.