Study design & univariate analysis

Dr Joram M. Posma

Lecturer in Cancer Informatics

Section of Bioinformatics, Division of Systems Medicine, Department of Metabolism, Digestion and Reproduction, Imperial College London, UK

Rutherford Fellow

Health Data Research (HDR) UK, HDR-London, UK

HDRUK Health Data Research UK

MRC Medical Research Council

On the (pre-dinner) menu for today

- Particulars of analytical experimental design for metabolic profiling
  - Collection protocols
  - Preparation protocols
  - Quality controls
- Statistical experimental design
  - Errors in inference
  - Univariate data analysis (a refresher of "*stats 101*")
  - Multiple testing correction
  - Bias
  - Confounding

Shout-outs

By the end of this session, you will be better able to

- Describe some analytical strategies that improve data quality
- Explain the concept of multiple testing
- Distinguish between different methods of error control
- Identify different types of bias
- Discuss the concept of confounding

The importance of following protocols

- When to collect samples?
  - Sampling time could affect measured outcome, e.g. kinetics of effect of drug treatment
- How to collect samples?
  - Always with the same procedure/order/materials
- How to store biological samples?
  - Always in the same way, as soon as possible in -40°C or -80°C
- Contaminants, coagulants, and other chemicals
  - These can be measured with MS and NMR, can affect the data
- How to prepare samples?
  - Different analysts can have different skills (*how accurate do you pipette?*) and speeds
- Freeze-thaw cycles
  - Some metabolites may evaporate over time…

- Common aspects of the above examples?
  - Refer to standard operating protocols (SOPs)

Quality controls for improving data *quality*

- High data quality: accurate, stable and reproducible

- During sample preparation make a ‘pooled sample’
- Pooled sample is also prepared multiple times (technical replicate)
- Sample is run every few samples, across the entire data acquisition period, across all batches/plates
- Allows checking whether performance during acquisition is stable using a sample that is the average of all samples (representative for study)
  - Is there a pattern in QCs over time? Analytical drift, temperature changes, etc.
  - Between batches? Instrument needs tuning
  - Between changing solvents? Different chemical conditions
  - Can be used to correct the data *a posteriori* (to some extent…!)

- Also use a comparative control, a long-term reference used across multiple studies
  - Allows comparison across studies
  - Good for reproduction of results

And then there is randomization order: for collection, preparation and acquisition…

Repeated-measures design

- Mass spectrometry suffers from analytical drift, can create batch effects
- In repeated-measures design: randomize people and run all samples from one person after another
  - Minimize within person variation due to changes in analytical technique
- So simple, yet so effective

Types of errors in statistical inference

		Actual truth based on entire population sample (unknown)	
		True association	**No association**
Test result based on subset of population (measured)	**Evidence found for association**	Correct inference	Wrong inference
	No confidence for an association	Wrong inference	Correct inference

Types of errors in statistical inference

		Actual truth based on entire population sample (unknown)	
		True association	**No association**
Test result based on subset of population (measured)	**Evidence found for association**	True positive	False positive (type I error)
	No confidence for an association	False negative (type II error)	True negative

Types of errors in statistical inference

		Actual truth based on entire population sample (unknown)	
		True association	**No association**
Test result based on subset of population (measured)	**Evidence found for association**	Power = $1 - \beta$	α
	No confidence for an association	β	$1 - \alpha$

HDRUK Health Data Research UK

MRC Medical Research Council

False positive

- Put faith into something that does not deserve it
- Falsehood

- Chance
- Random variation
- Non-representative subset of population
- (False positive paradox: pr(FP) > pr(TP))

False negative

- Failure to believe the truth
- Not realizing there is a causality

- Too stringent criteria for beliefs
- Low sample size $\propto$ low effect size
- Subset does not extrapolate to entire population
- (Wrong assumption about data made…)

Error recap

- Often α is set to 0.05 and β is fixed at 0.2

- What does this mean?

  (in the terms we just discussed)

- Why does this make sense?

  (from a research perspective)

How many samples to collect? Do a power calculation*

* Have a chat with Goncalo Correia

The null hypothesis (H_0)

- *Something* is not true (e.g. there is no association)
- Doing a (frequentist) statistical test to verify this, …can we?

- We get a *p*-value: probability of obtaining (new) data that is ≥ extreme than the data you have, assuming the null hypothesis was true (no association)

- High *p*-value: *maybe* there is no effect
- Low *p*-value: *maybe* there is an effect (this is what we typically focus on)

Choosing a test depends on the *hypothesis* and the *data*

Parametric

- Pearson correlation

- Two-sample t-test
- One-way ANOVA

- Paired t-test
- Two-way ANOVA

- Multiple Linear/Logistic Regression
- *Et al*.

Non-parametric

- Spearman/Kendall rank correlation

- Wilcoxon rank sum test (= Mann-Whitney U-test)
- Kruskal-Wallis test

- Wilcoxon signed rank test
- Friedman test

- Rank regression
- *Et al*.

Three commonly used univariate testing strategies

- Group comparison
- Comparing means/medians
- Case/control
- E.g. t-test
- Assume data is normally distributed

- Correlations
  - But not causation
- Association between continuous variables
- E.g. Pearson correlation
- Result between -1 and 1 (0 is absence)

- Regression
  $Y = \beta_0 + \beta_1 \times X_1$
- The slope of linear regression is directly related to Pearson correlation (intuitively: same *p*-value…)
- Can be used to predict new data

One, two, multiple tests

- One statistical test: what significance level determines ‘enough confidence’?
- Same test, but applied on two variables: what significance level?
- Same test, applied to two (cor)related variables?
- Same test, applied to 10,000 variables: what significance level?

- One size fits all? Why (not)?

Family Wise Error Rate

- Control type I errors, set an acceptable level of false (positive) discoveries

- Probability of making at least one type I error:

$$\mathrm{FWER} = \mathrm{pr}(\#_{FP} > 1) = 1 - (1 - \alpha)^{n_t}$$

Family Wise Error Rate

n_t (at $\alpha = 0.05$)	FWER
1	0.05
2	0.0975
3	0.142625
4	0.18549375
5	0.2262190625
10	0.401263060761621
20	0.641514077591458
100	0.994079470779666
10000	1

- Our data has 1000s of variables
- What do we do to prevent?

Control the Family Wise Error Rate

Change α for α':

- Bonferroni: $\alpha' = {}^{\alpha}/_{n_t}$
- Šidák: $\alpha' = 1 - (1 - \alpha)^{1/n_t}$

These require an ordered list of p-values ($P_{i-1} \leq P_i \leq P_{i+1}$):

- Holm: $\alpha' = \min_i \left\{ P_i > \frac{\alpha}{n_t+1-i} \right\}$
- Simes: $\alpha' = \min_i \left\{ P_i > \frac{\alpha \times i}{n_t} \right\}$
- Hommel: $\alpha' = {}^{\alpha}/_{i} : \min_i \left\{ P_{n_t-i+j} < \frac{\alpha \times j}{i} \right\}$ where $1 \leq j \leq i \leq n_t$
- Hochberg: $\alpha' = {}^{\alpha}/_{n_t+1-i} : \min_i \left\{ P_i > \frac{\alpha}{n_t+1-i} \right\}$

Keeping it simple: most people use Bonferroni… …but why you should be using Šidák instead

Number of tests	Bonferroni α'	Bonferroni FWER	Šidák α'	Šidák FWER
1	$\frac{0.05}{1} = 0.05$	$1-(1-0.05)^{1} = 0.05$	$1-(1-0.05)^{\frac{1}{1}} = 0.05$	$1-(1-0.05)^{1} = 0.05$
2	$\frac{0.05}{2} = 0.025$	$1-(1-0.025)^{2} = 0.0494$	$1-(1-0.05)^{\frac{1}{2}} = 0.0253$	$1-(1-0.0253)^{2} = 0.05$
5	$\frac{0.05}{5} = 0.01$	$1-(1-0.01)^{5} = 0.0490$	$1-(1-0.05)^{\frac{1}{5}} = 0.0102$	$1-(1-0.0102)^{5} = 0.05$
10	$\frac{0.05}{10} = 0.005 = 5\times10^{-3}$	$1-(1-5\times10^{-3})^{10} = 0.0489$	$1-(1-0.05)^{\frac{1}{10}} = 5.12\times10^{-3}$	$1-(1-5.12\times10^{-3})^{10} = 0.05$
100	$\frac{0.05}{100} = 5\times10^{-4}$	$1-(1-5\times10^{-4})^{100} = 0.0488$	$1-(1-0.05)^{\frac{1}{100}} = 5.13\times10^{-4}$	$1-(1-5.13\times10^{-4})^{100} = 0.05$
500	$\frac{0.05}{500} = 1\times10^{-4}$	$1-(1-1\times10^{-4})^{500} = 0.0488$	$1-(1-0.05)^{\frac{1}{500}} = 1.03\times10^{-4}$	$1-(1-1.03\times10^{-4})^{500} = 0.05$
1,000	$\frac{0.05}{1000} = 5\times10^{-5}$	$1-(1-5\times10^{-5})^{1000} = 0.0488$	$1-(1-0.05)^{\frac{1}{1000}} = 5.13\times10^{-5}$	$1-(1-5.13\times10^{-5})^{1000} = 0.05$
10,000	$\frac{0.05}{10000} = 5\times10^{-6}$	$1-(1-5\times10^{-6})^{10000} = 0.0488$	$1-(1-0.05)^{\frac{1}{10000}} = 5.13\times10^{-6}$	$1-(1-5.13\times10^{-6})^{10000} = 0.05$
1,000,000	$\frac{0.05}{1000000} = 5\times10^{-8}$	$1-(1-5\times10^{-8})^{1000000} = 0.0488$	$1-(1-0.05)^{\frac{1}{1000000}} = 5.13\times10^{-8}$	$1-(1-5.13\times10^{-8})^{1000000} = 0.05$

$$\mathrm{FWER} = \mathrm{pr}(\#_{FP} > 1) = 1-(1-\alpha)^{n_t}$$

Control the False Discovery Rate

- These require an ordered list of p-values ($P_{i-1} \leq P_i \leq P_{i+1}$):
- Benjamini-Hochberg: $pFDR_i = \frac{n_t \times P_i}{i}$ and

  $$pFDR_i = pFDR_{i+1} \forall (pFDR_i > pFDR_{i+1})$$
- Benjamini-Yekutieli: $\alpha' = \frac{\alpha \times i}{n_t \times c} : \min_i \left\{ P_i > \frac{\alpha \times i}{n_t \times c} \right\}$ where $c \in \left[1, \sum_{j=1}^{n_t} {}^1/_j\right]$
- Storey-Tibshirani (q-value): $q_i = \hat{\pi}_0(\lambda_{opt}) \times pFDR_i$

  where $\hat{\pi}_0(\lambda_{opt}) = \frac{\sum_{i=1}^{n_t} (P_i > \lambda)}{n_t \times (1-\lambda)}$ and $\lambda_{opt} = \min_{MSE}(\pi_0(\lambda))$

OPEN ACCESS

ESSAY

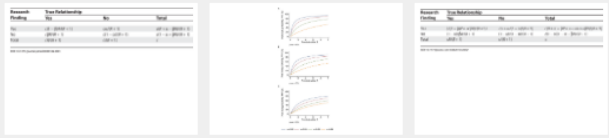
Why Most Published Research Findings Are False

Published: August 30, 2005 • https://doi.org/10.1371/journal.pmed.0020124

Abstract

Summary

Figures



Related PLOS Articles

Why Current Publication Practices May Distort Science
View Page PDF

Why Most Published Research Findings Are False: Author's Reply to Goodman and Greenland
View Page PDF

Why Most Published Research Findings Are False: Problems in the Analysis
View Page PDF

Most Published Research Findings Are False—But a Little Replication Goes a Long Way
View Page PDF

When Should Potentially

Categories of bias

- Accidental bias: introduced in algorithms by ignorance (avoidable!)
- Deliberate bias: intentionally introduced in algorithms (hopefully, avoidable)
- Implicit bias: absorbed by algorithms from (finite) data (avoidable?)

Bias

- Prior knowledge of types of bias?

- Which of these 3 categories are they?
  - accidental
  - deliberate
  - implicit

Bias

- Sampling bias: some parts of total population more likely to be included in study than other parts
- Attrition bias: some people from the sampled population are more likely to drop out than others
- Recall bias: misreporting (self-reporting data)
- Selection bias: samples are not properly randomized in data analysis, some have bigger influence on model than others
- Biased error estimates: most likely a positive bias (lower error), due to inclusion of information in modelling (self-fulfilling)
- Biased predictors: deviation from the true parameter value, i.e. different slope of coefficient
- Bias/variance trade-off: how unbiased should a model be opposed to how accurate/stable (variance) should the predictions be – can be a little biased, if that makes the results more stable
- Reporting bias: positive results are more likely to be reported than negative ones
- Observer bias: deliberately changing or subconsciously influencing analysis to conform to cognitive bias of hypothesis

Bias

- Some form of systematically introduced variation
- A result of errors in collection, measurement and analysis
- Can be avoided!

Bias

- Sampling bias (DoE)
- Attrition bias (DoE)
- Recall bias
- Selection bias (SED)
- Biased error estimates (SED)
- Biased predictors (SED)
- Bias/variance trade-off (SED)
- Reporting bias
- (And the unforgivable sin: observer bias)

Bias

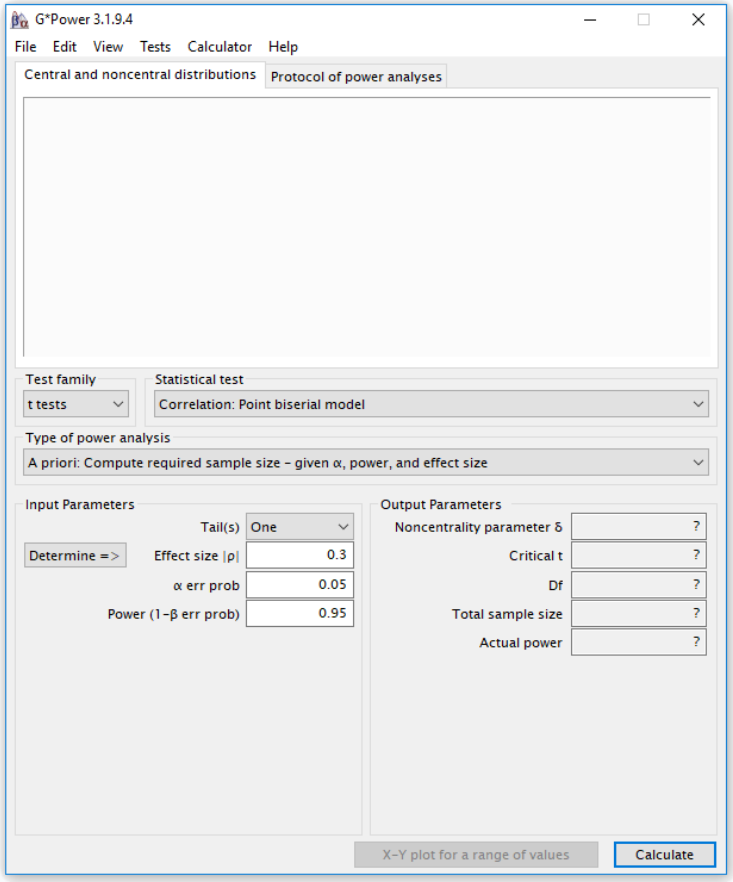
- Sampling bias (DoE)
- Attrition bias (DoE)
- Recall bias
- Selection bias (SED)
- Biased error estimates (SED)
- Biased predictors (SED)
- Bias/variance trade-off (SED)
- Reporting bias
- (And the unforgivable sin: observer bias)

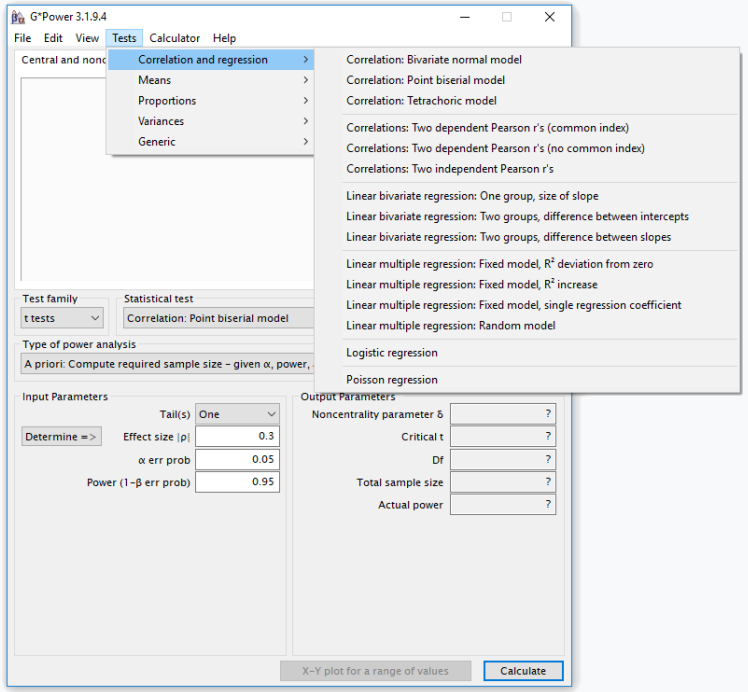
Sampling and attrition bias
Sample size and power calculations

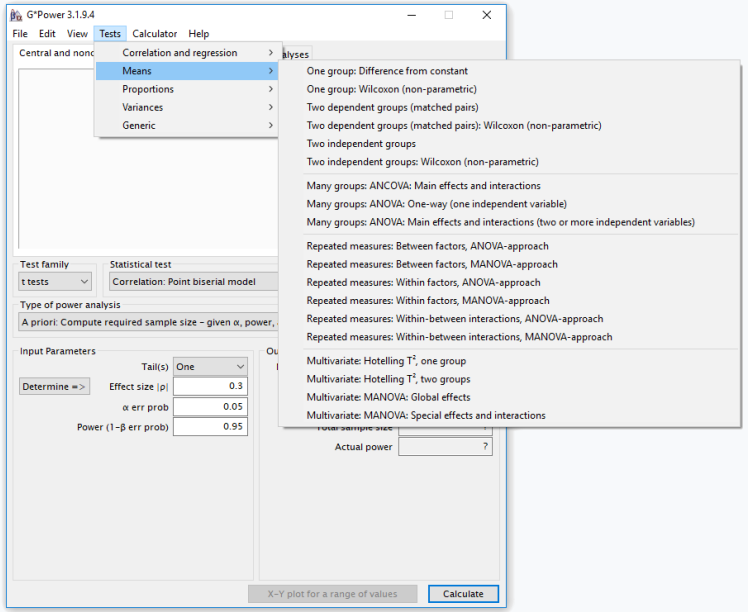
- How many samples do you need to study something?
  - What are you studying?
  - What is known already?
  - How are you testing the (null) hypothesis?

- For univariate tests you can do a power calculation, what you need:
  - What type of statistical test (group comparison, correlation, etc.)
  - Prior knowledge of what you are testing? E.g. what is the expected correlation or expected difference between groups

Power calculation: G*power software



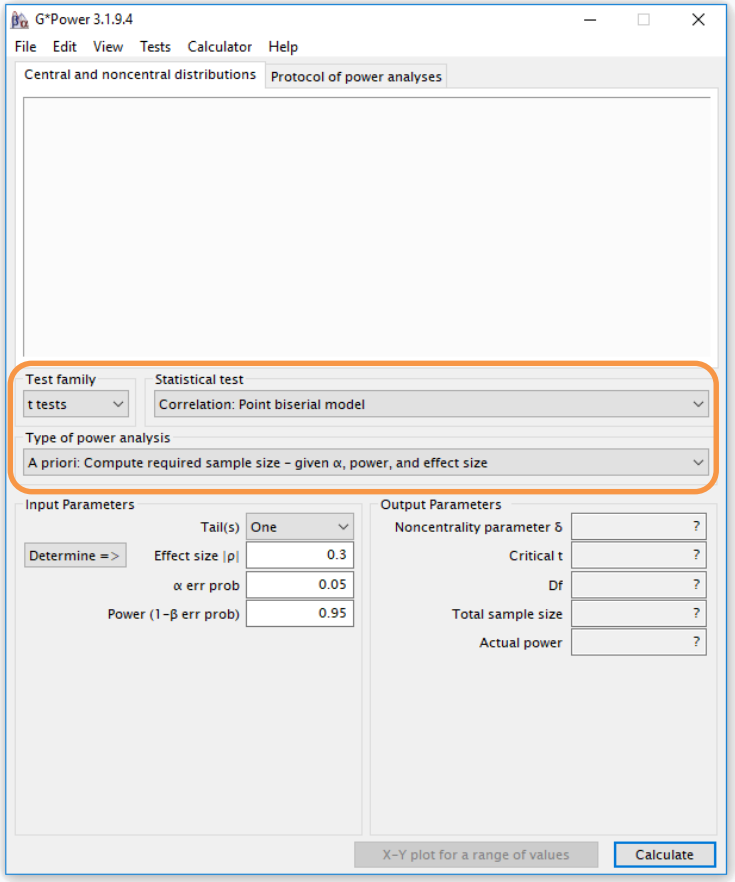


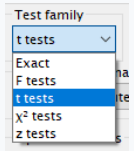


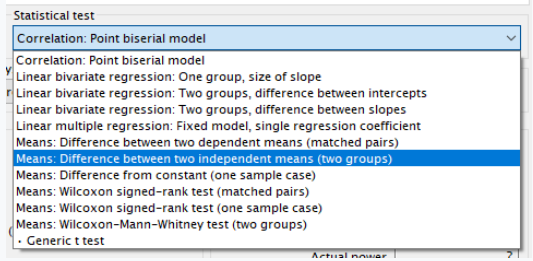
HDRUK Health Data Research UK

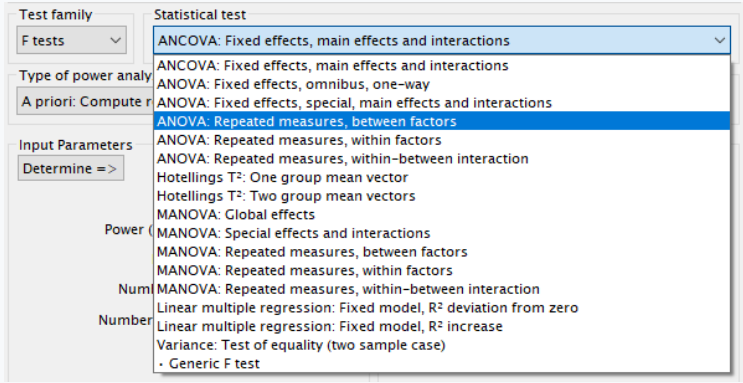
MRC Medical Research Council

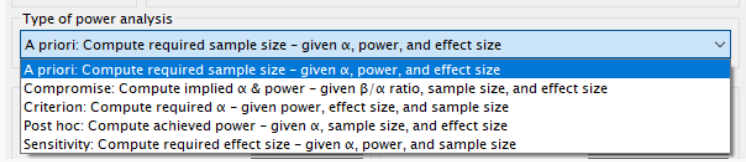
Power calculation: G*power software







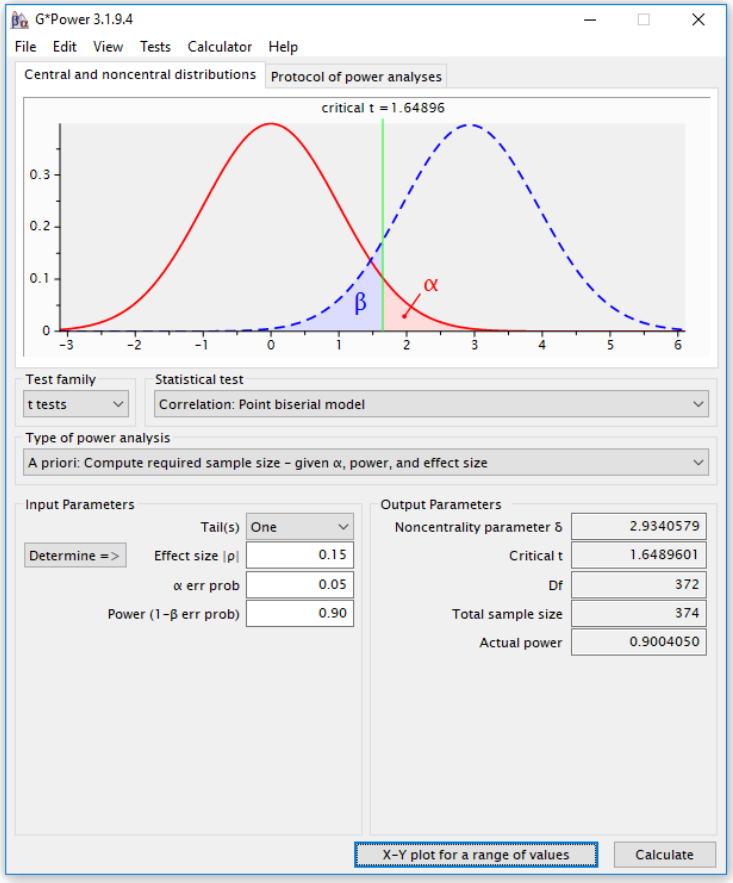


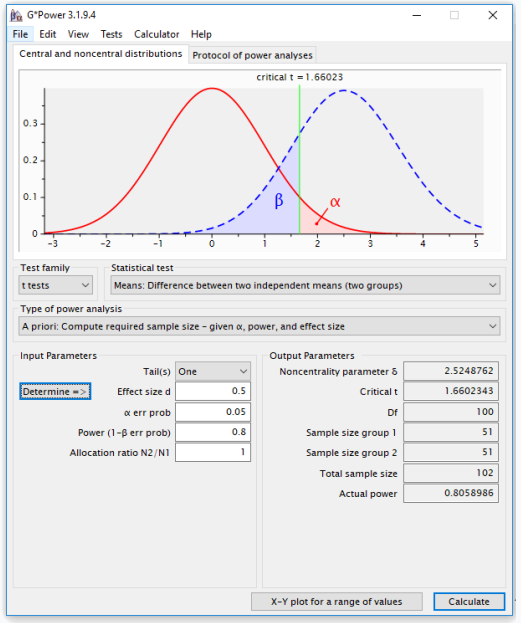


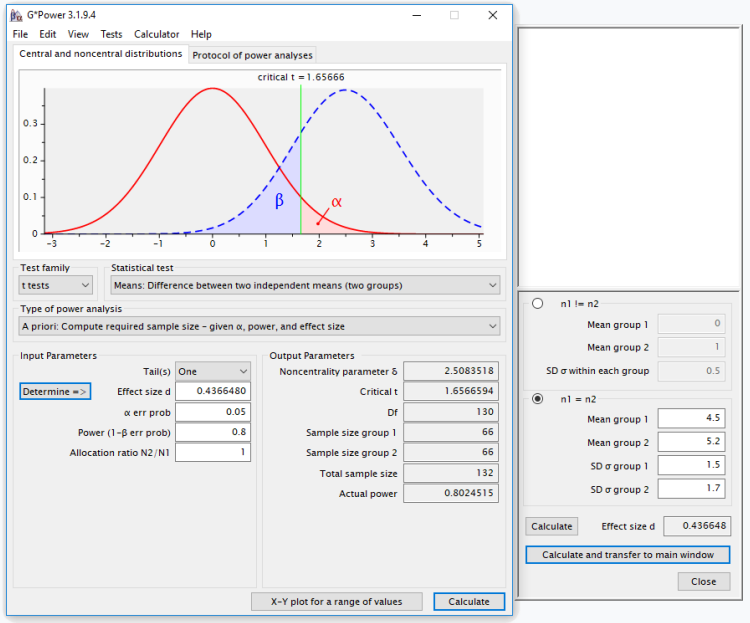
HDRUK Health Data Research UK

MRC Medical Research Council

Power calculation: G*power software







HDRUK Health Data Research UK

MRC Medical Research Council

Study design, recruitment

- Case/control study to investigate differences in circulating fatty acids
- Participants recruited after visiting GP for a blood draw

- Cases: people with type-2 diabetes
- Controls: healthy people without diabetes

Do truly healthy people go to a GP for a blood draw?
What is the definition of healthy?
Are there other diseases associated with the case group (e.g. obesity), or these also present in the control group?

Study design, recruitment

- Case/control study investigating protein levels in CSF
- Recruitment from patients referred to specialist

- 85 Guillain-Barré syndrome patients
- Controls: 15 patients without GBS

Balanced design?
What is likelihood in general population?
Why were controls referred to specialist?

Study design, recruitment

- Recruitment of 100 people to a case/control study to phenotype a disease

- Fanconi syndrome n=50
- Control also n=50

Good for modelling, equal weight to both groups, but outcome must be weighted
We are assuming these conditions have equal likelihood of occurring
Good to avoid FN, but not to avoid FP

Data acquisition

- Samples are analysed using liquid chromatography mass spectrometry (LC-MS)
- Ordered based on label identifiers as sorted by clinicians after sample collection (A-1, A-2, …, A-50, B-1, B-2, …, B-50)

Was it randomized, does not look like it at all?
Run order should be randomized

Confounding

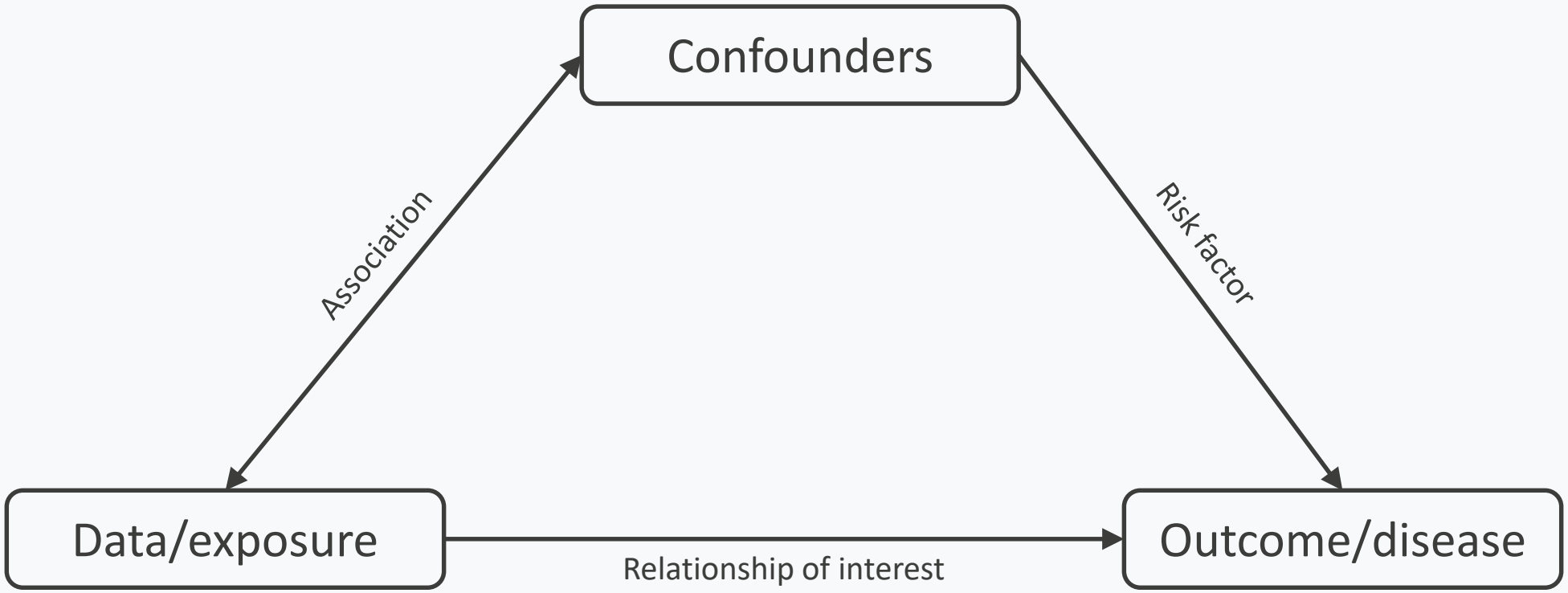
- Misrepresentation of association caused by effect of another factor

- 1) Association between outcome and confounder
- 2) Confounder influences data as well
- 3) Confounder is not a cause of the outcome

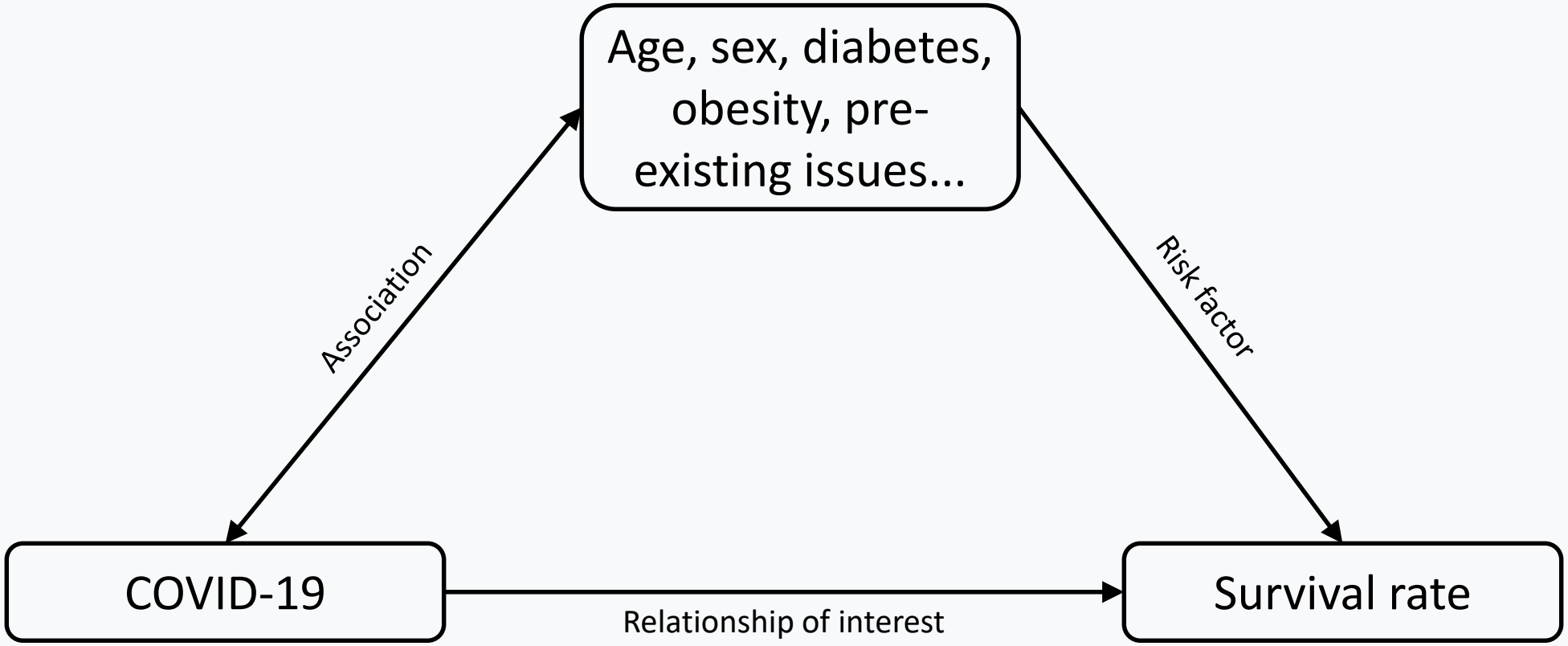
requires knowledge of pathophysiological mechanisms

- Positive confounding: overestimating effects
- Negative confounding: underestimating effects

Confounding



Confounding (in 2020)



Confounding – example

- Students with tutors get lower grades than students without tutors

- Do not get a tutor: you score higher?

(or could there be confounding…?)

Confounding – example

Game	Number of police man hours logged	Number of arrests due to rioting or hooliganism
Fulham – Chelsea	75	12
Crystal Palace – Watford	50	5
Tottenham – Arsenal	750	235
West Ham – Queens Park Rangers	100	19
Arsenal – Chelsea	500	187
Queens Park Rangers – Watford	25	2
West Ham – Crystal Palace	125	25
Tottenham – Fulham	250	51
Watford – Fulham	30	1

- Met Police office: the analysis of the data shows that if we stop sending officers, there will not be any rioting

- Which confounding factor(s) can you identify here?
  (not errors of reasoning)

Positive and negative confounding of creatinine on associations with BMI

Table 2. Association with BMI of a set of metabolites measured by targeted IEC (log_{10} values) in 1880 U.S. INTERMAP participants. Excluding metabolic outliers based on Hotelling's T^2 test ($n = 132$) and participants with doctor-diagnosed diabetes mellitus ($n = 152$). Partial correlation (r) and corresponding P values are listed for each metabolite. Statistical significance based on a Bonferroni threshold of $P \leq 4.55 \times 10^{-4}$ ($P \leq 0.01/22$). n.s., not significant.

Urinary variable	Model 1*			Model 2*†			Model 3*†‡		
	r	*P*	Significance	*r*	*P*	Significance	*r*	*P*	Significance
Taurine	0.02	3.68×10^{-1}	n.s.	0.01	6.09×10^{-1}	n.s.	−0.11	1.13×10^{-6}	
Threonine	0.15	4.44×10^{-11}		0.15	2.34×10^{-10}		−0.02	3.82×10^{-1}	n.s.
Serine	0.05	3.06×10^{-2}	n.s.	0.06	1.58×10^{-2}	n.s.	−0.12	4.41×10^{-7}	
Asparagine	0.01	7.69×10^{-1}	n.s.	0.00	9.97×10^{-1}	n.s.	−0.11	7.93×10^{-7}	
Glutamine	0.10	1.29×10^{-5}		0.12	3.34×10^{-7}		−0.05	4.41×10^{-2}	n.s.
Glycine	−0.02	3.98×10^{-1}	n.s.	0.01	6.00×10^{-1}	n.s.	−0.10	6.08×10^{-6}	
Alanine	0.20	4.85×10^{-18}		0.18	3.47×10^{-15}		0.02	3.65×10^{-1}	n.s.
Valine	0.19	6.87×10^{-16}		0.17	6.51×10^{-14}		0.07	3.87×10^{-3}	n.s.
Cystine	0.33	4.85×10^{-49}		0.30	4.59×10^{-41}		0.18	6.80×10^{-15}	
Methionine	0.06	1.06×10^{-2}	n.s.	0.05	4.32×10^{-2}	n.s.	−0.03	1.57×10^{-1}	n.s.
Isoleucine	0.12	4.23×10^{-7}		0.10	2.11×10^{-5}		0.05	2.73×10^{-2}	n.s.
Leucine	0.08	3.66×10^{-4}		0.06	7.53×10^{-3}	n.s.	0.02	5.08×10^{-1}	n.s.
Tyrosine	0.32	2.02×10^{-45}		0.30	8.28×10^{-41}		0.18	1.58×10^{-14}	
Phenylalanine	0.20	8.35×10^{-19}		0.19	1.42×10^{-16}		0.08	9.99×10^{-4}	n.s.
Ethanolamine	0.26	2.00×10^{-29}		0.27	2.19×10^{-31}		0.14	7.39×10^{-10}	
Lysine	0.27	4.64×10^{-33}		0.26	7.81×10^{-30}		0.12	1.06×10^{-7}	
1-Methylhistidine	0.11	9.47×10^{-7}		0.09	7.20×10^{-5}		−0.05	4.48×10^{-2}	n.s.
Histidine	0.14	3.02×10^{-9}		0.15	6.23×10^{-11}		0.01	6.43×10^{-1}	n.s.
Tryptophan	0.15	4.49×10^{-11}		0.15	1.79×10^{-10}		0.09	4.35×10^{-5}	
3-Methylhistidine	0.42	8.28×10^{-80}		0.38	1.87×10^{-65}		0.12	2.14×10^{-7}	
Carnosine	0.18	7.16×10^{-15}		0.17	2.28×10^{-13}		0.02	2.90×10^{-1}	n.s.
Arginine	0.08	8.71×10^{-4}	n.s.	0.07	2.77×10^{-3}	n.s.	0.03	1.34×10^{-1}	n.s.

*Model 1 is adjusted for age, gender, and sample. †Model 2 is adjusted for all factors in model 1 plus cardiovascular disease history (heart disease/stroke), physical activity, medication for hypertension, prescribed lipid-lowering drugs, NSAID use, dietary supplement use, special diet, smoking, education, and total energy intake per day (kcal/day). ‡Model 3 is adjusted for all factors in models 1 and 2 plus 24-hour urinary creatinine.

Elliott and Posma, *et al*. (2015) *Science Translational Medicine* **7**(285)

HDRUK Health Data Research UK

MRC Medical Research Council

Positive and negative confounding of creatinine on associations with BMI

Table 2. Association with BMI of a set of metabolites measured by targeted IEC (log_{10} values) in 1880 U.S. INTERMAP participants. Excluding metabolic outliers based on Hotelling's T^2 test ($n = 132$) and participants with doctor-diagnosed diabetes mellitus ($n = 152$). Partial correlation (r) and corresponding P values are listed for each metabolite. Statistical significance based on a Bonferroni threshold of $P \leq 4.55 \times 10^{-4}$ ($P \leq 0.01/22$). n.s., not significant.

Urinary variable	Model 1*			Model 2*†			Model 3*†‡		
	r	*P*	Significance	*r*	*P*	Significance	*r*	*P*	Significance
Taurine	0.02	3.68×10^{-1}	n.s.	0.01	6.09×10^{-1}	n.s.	−0.11	1.13×10^{-6}	
Threonine	0.15	4.44×10^{-11}		0.15	2.34×10^{-10}		−0.02	3.82×10^{-1}	n.s.
Serine	0.05	3.06×10^{-2}	n.s.	0.06	1.58×10^{-2}	n.s.	−0.12	4.41×10^{-7}	
Asparagine	0.01	7.69×10^{-1}	n.s.	0.00	9.97×10^{-1}	n.s.	−0.11	7.93×10^{-7}	
Glutamine	0.10	1.29×10^{-5}		0.12	3.34×10^{-7}		−0.05	4.41×10^{-2}	n.s.
Glycine	−0.02	3.98×10^{-1}	n.s.	0.01	6.00×10^{-1}	n.s.	−0.10	6.08×10^{-6}	
Alanine	0.20	4.85×10^{-18}		0.18	3.47×10^{-15}		0.02	3.65×10^{-1}	n.s.
Valine	0.19	6.87×10^{-16}		0.17	6.51×10^{-14}		0.07	3.87×10^{-3}	n.s.
Cystine	0.33	4.85×10^{-49}		0.30	4.59×10^{-41}		0.18	6.80×10^{-15}	

HDRUK Health Data Research UK

Elliott and Posma, *et al*. (2015) *Science Translational Medicine* **7**(285)

MRC Medical Research Council

Adjusting for confounders in univariate analyses: stratification and sub-sampling

- Stratify data according to confounders
- Sample from data to ensure data is matched (e.g. case/control)
- Pros and cons?

Adjusting for confounders in univariate analyses: Univariate regression

- Include confounders in the modelling of variable X_1

$$Y = \beta_0 + \beta_1 \times X_1$$

$$Y = \beta_0 + \beta_1 \times X_1 + \boldsymbol{\beta_2} \times \boldsymbol{C_1} + \cdots$$

- Does it change regression coefficient for X_1 (β_1)?

- If C explains part of Y, then contribution of X_i will be attenuated
- Adjusting Y for C first and then modelling $Y_{adj} \sim X_i$ results in modelling not Y but an adjusted version

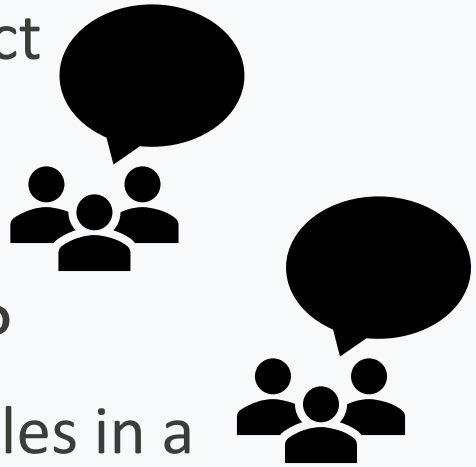
- Linear regression is commutative: *p*-value of $\alpha_1 \equiv$ *p*-value of β_1, but $\alpha_1 \neq \beta_1$)

$$X_1 = \alpha_0 + \alpha_1 \times Y$$

- Partial correlation is also commutative ($r(X_1, Y|C_1) = r(Y, X_1|C_1)$)

Univariate versus multivariate

- Adjusting for confounders in regression/correlation
  - Add confounder to data and model it with variable on Y
  - If *p*-value of $X_{(i)}$ is now higher, then confounder had an effect
  - But how big was the effect? Is the *effect* significant?



- How many can we add? Can we adjust for everything?
  - No, there is a limit, specifically where the number of variables in a regression model is more than the number of samples
  - We need another strategy (que: multivariate*) then, because there is too much data to explain too little (variation): perfect fit?

* For example this morning's session (unsupervised) and Tim's tomorrow (supervised) on multivariate analysis

(And there also is multivariate confounder/covariate adjustment: Posma, *et al*. (2018) *Journal of Proteome Research* **17**(4) doi)

Summary

- Different forms of bias can be introduced in study design, sample collection, data acquisition and data analysis
- Bias cannot be adjusted for
- Cannot adjust a model for confounders arising from poor experimental design
- Everything is carried forward: rubbish in, rubbish out
- In data analysis keep it as free from bias as possible