Day 05

Replication, Circularity, & Regression to the mean

- Pseudoreplication
- Confounding variables
- Circular analysis
- Regression to the mean & stopping rules

Replication & Pseudoreplication

Science relies on replicate measurements.

- Var(measurement) = External factor variability

  + Natural biological variability

  + Measurement error
- Additional replicates → more accurate & reliable summary statistics.
- Replicates can be used to:
  - Assess & isolate sources of variation in measurements
  - Limit the effect of spurious variation on hypothesis testing & parameter estimation.

Replication & Pseudoreplication

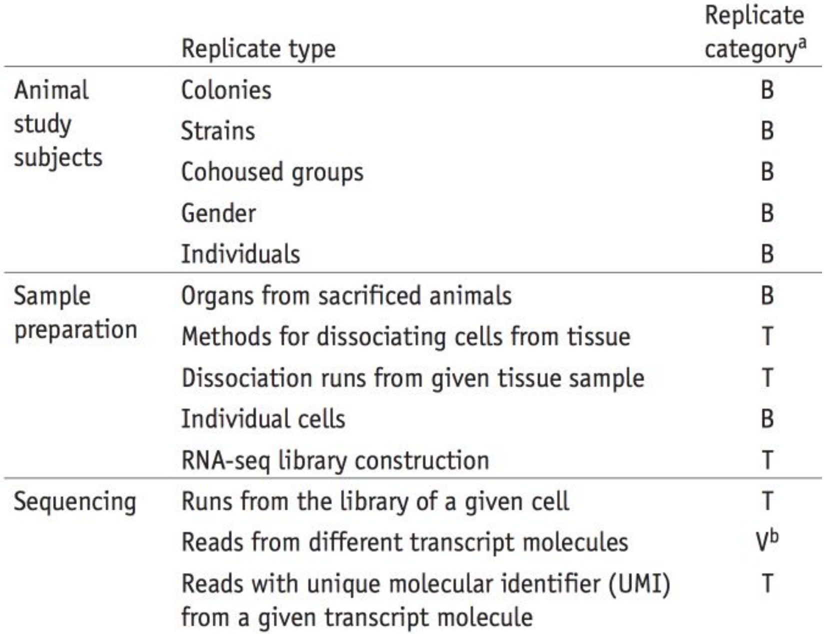
Biological replicates:

- Parallel measurements of biologically distinct samples
- Capture random biological variation (could be subject of study or a noise source).

Technical replicates:

- Repeated measurements of the same sample
- Represent independent measures of the random noise associated with protocols or equipment.

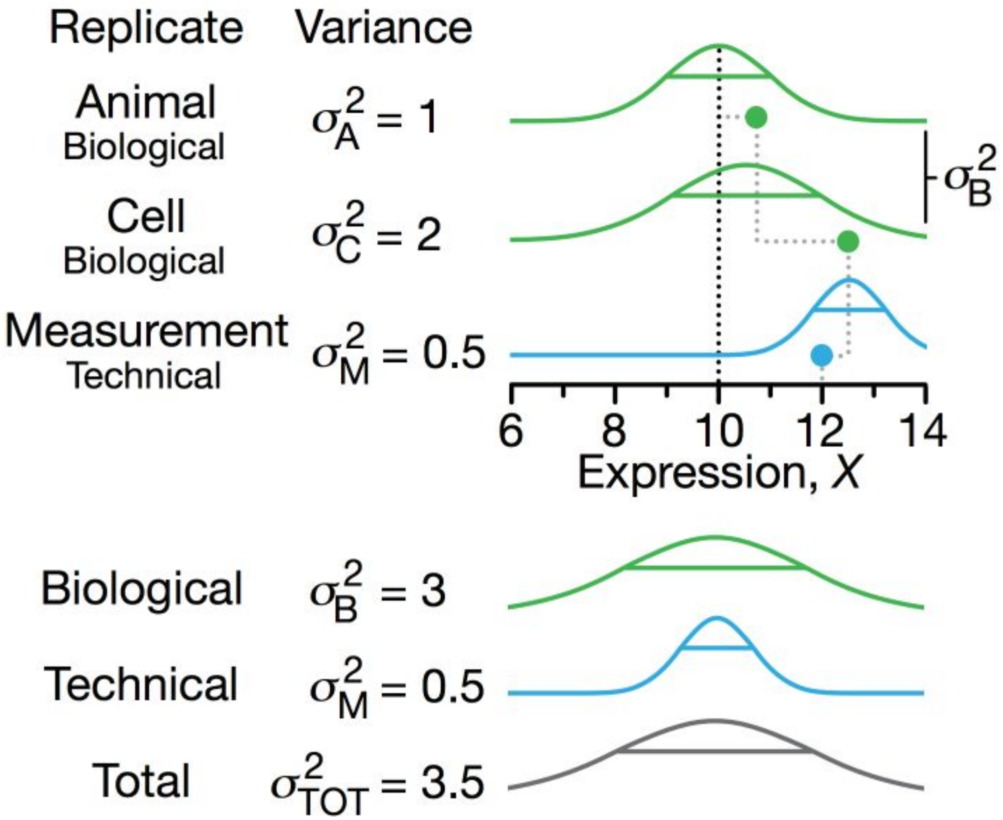
Which sources of variation are being studied & which are considered noise?
B: biological, T: technical



	Replicate type	Replicate category[a]
Animal study subjects	Colonies	B
	Strains	B
	Cohoused groups	B
	Gender	B
	Individuals	B
Sample preparation	Organs from sacrificed animals	B
	Methods for dissociating cells from tissue	T
	Dissociation runs from given tissue sample	T
	Individual cells	B
	RNA-seq library construction	T
Sequencing	Runs from the library of a given cell	T
	Reads from different transcript molecules	V[b]
	Reads with unique molecular identifier (UMI) from a given transcript molecule	T

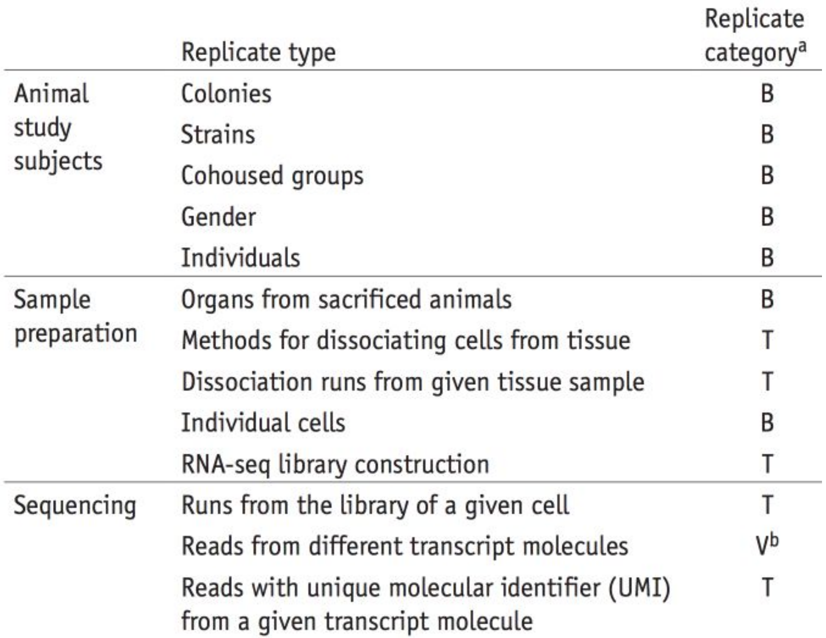
https://www.nature.com/articles/nmeth.3091

Replication & Pseudoreplication



Which sources of variation are being studied & which are considered noise?

B: biological, T: technical



	Replicate type	Replicate category[a]
Animal study subjects	Colonies	B
	Strains	B
	Cohoused groups	B
	Gender	B
	Individuals	B
Sample preparation	Organs from sacrificed animals	B
	Methods for dissociating cells from tissue	T
	Dissociation runs from given tissue sample	T
	Individual cells	B
	RNA-seq library construction	T
Sequencing	Runs from the library of a given cell	T
	Reads from different transcript molecules	V[b]
	Reads with unique molecular identifier (UMI) from a given transcript molecule	T

https://www.nature.com/articles/nmeth.3091

Replication & Pseudoreplication

- Sample size (n)

  $\mathrm{Var}(\overline{X}) = \sigma^2 / n$

- Effective sample size (n_{eff})

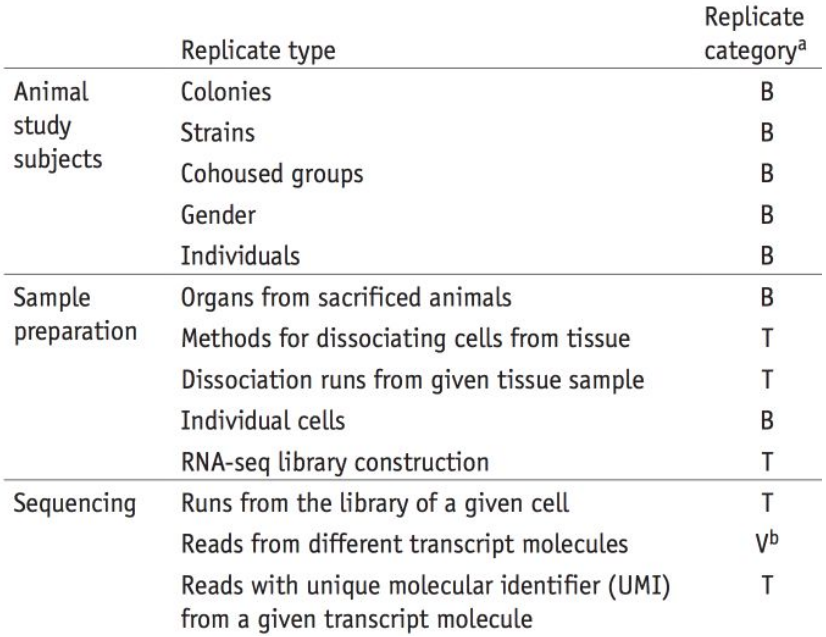
  $\mathrm{Var}(\overline{X}) = \sigma^2 / n_{\mathrm{eff}}$

If ρ is the correlation between samples,

$$n_{\mathrm{eff}} = \frac{n}{1 + (n-1)\rho}$$

$n_{\mathrm{eff}} \neq n$: Pseudoreplication

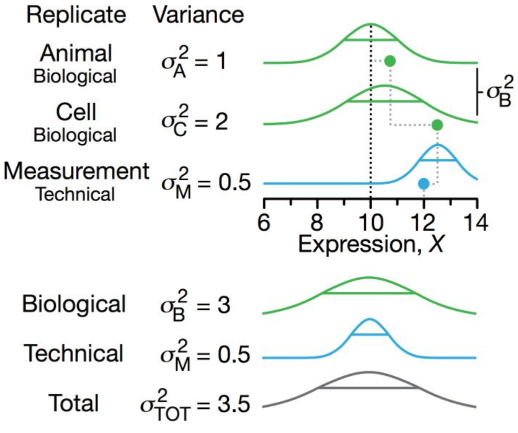
B: biological, T: technical



	Replicate type	Replicate category[a]
Animal study subjects	Colonies	B
	Strains	B
	Cohoused groups	B
	Gender	B
	Individuals	B
Sample preparation	Organs from sacrificed animals	B
	Methods for dissociating cells from tissue	T
	Dissociation runs from given tissue sample	T
	Individual cells	B
	RNA-seq library construction	T
Sequencing	Runs from the library of a given cell	T
	Reads from different transcript molecules	V[b]
	Reads with unique molecular identifier (UMI) from a given transcript molecule	T

https://www.nature.com/articles/nmeth.3091

Replication & Pseudoreplication



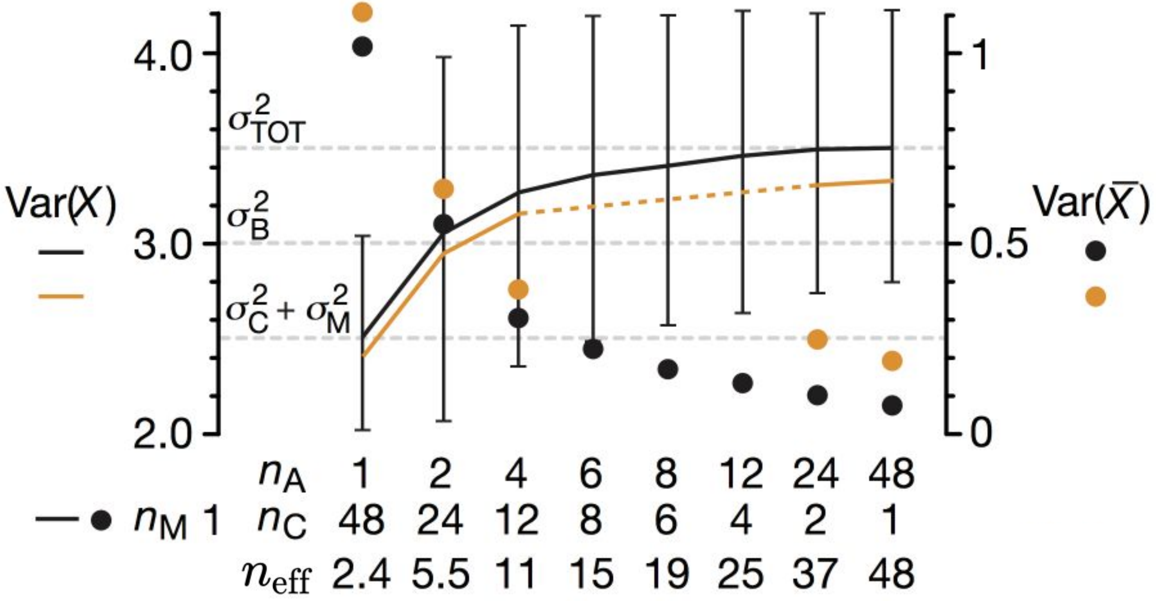
Simulation

$n = n_A n_C n_M = 48$

$n_A = 1{:}48,\ n_C = 1{:}48,\ n_M = 1, 3$

$n_{eff} = 2{:}48 = \mathrm{Var}(X)/\mathrm{Var}(X_{mean})$

Variance of expression & expression mean



$\mathrm{Var}(X_{mean}) = \sigma_A{}^2/n_A + \sigma_C{}^2/n_A n_C + \sigma_M{}^2/n_A n_C n_M$

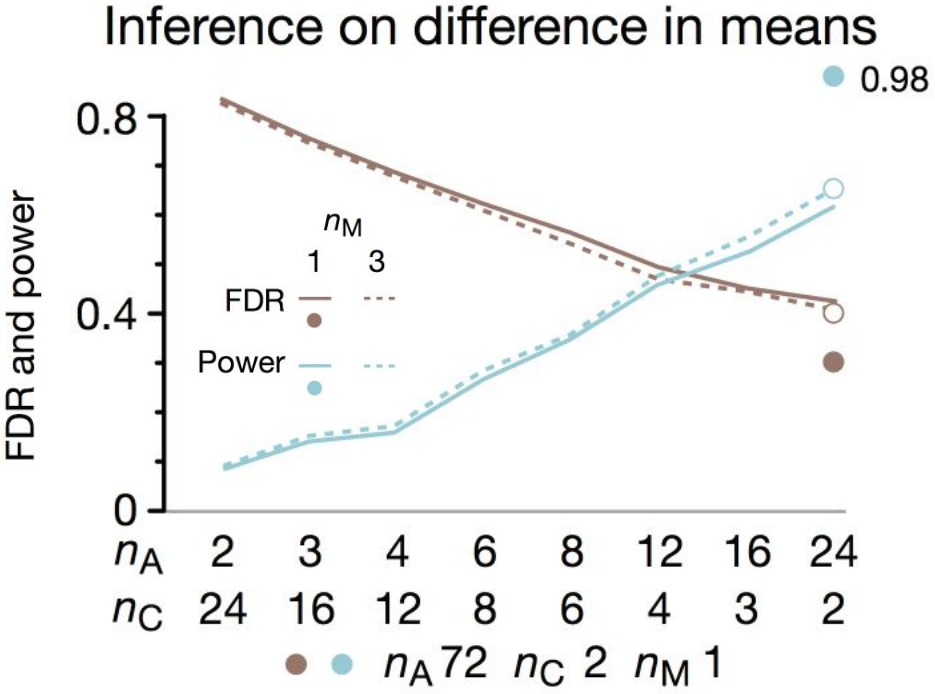
https://www.nature.com/articles/nmeth.3091

Replication & Pseudoreplication

No. replicates has a practical effect on inference errors in analysis of differences of means or variances.

Simulation of 10% effect in mean

- More animals the better.
- (n_A,n_C,n_M) from (24,2,3) to (72,2,1): 50% inc. in power (0.66→0.98).
- Consider cost difference between biological and technical replicates.



https://www.nature.com/articles/nmeth.3091

Replication & Pseudoreplication

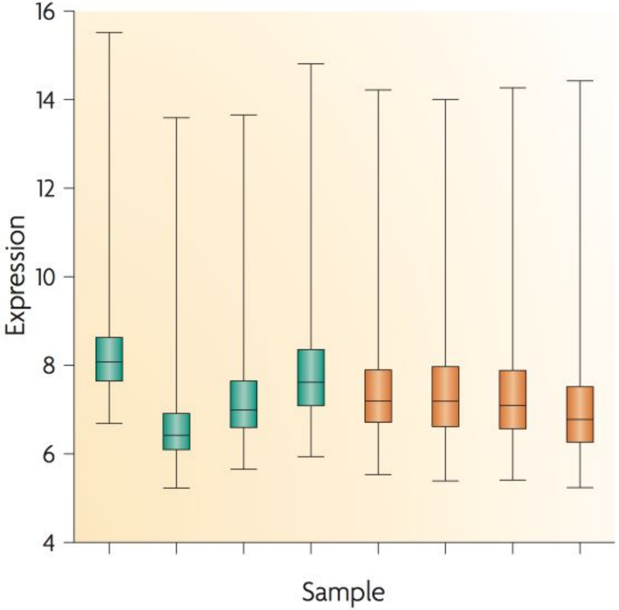
- Typically, biological variability >> technical variability.
  - Commit resources to sampling biologically relevant variables (unless measures of technical variability are themselves of interest).
- Planning for replication:
  1. Identify the question the experiment aims to answer.
  2. Determine proportion of variability induced by each step.
  3. Distribute the capacity for replication of the experiment across steps.
  4. Be aware of the potential for pseudoreplication and aim to design statistically independent replicates.
- As capacity for higher-throughput assays increases: more is not always better.

https://www.nature.com/articles/nmeth.3091

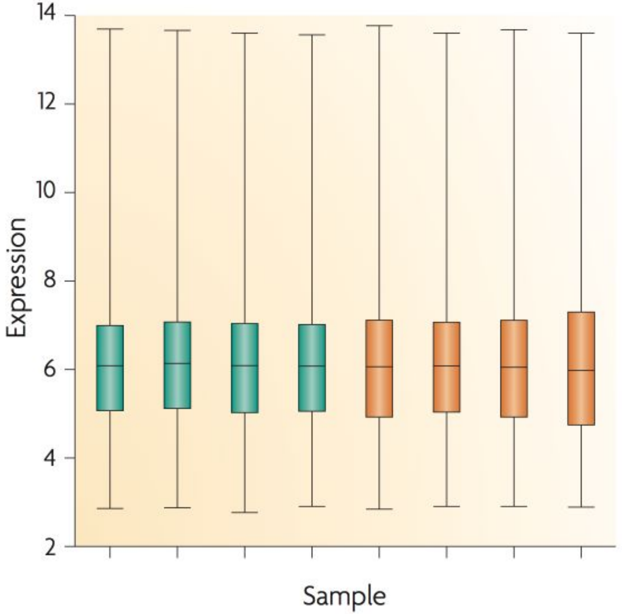
Confounding variables

Extraneous variables (e.g. processing data) can be *confounded* with the outcome of interest (e.g. disease state) when it correlates both with the outcome and with an independent variable of interest (e.g. gene expression).

Before normalization



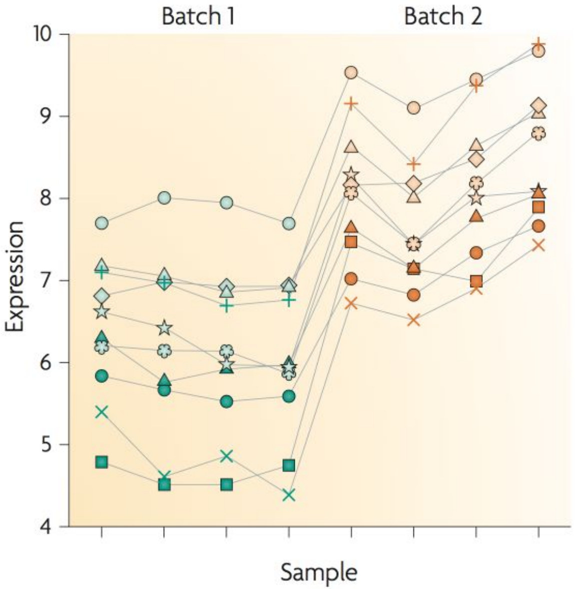
After normalization



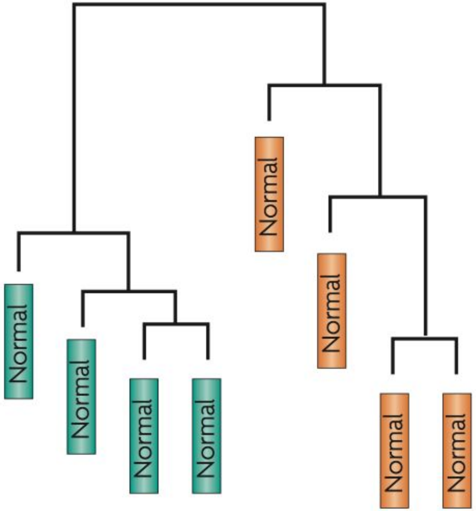
Confounding variables

Extraneous variables (e.g. processing data) can be *confounded* with the outcome of interest (e.g. disease state) when it correlates both with the outcome and with an independent variable of interest (e.g. gene expression).

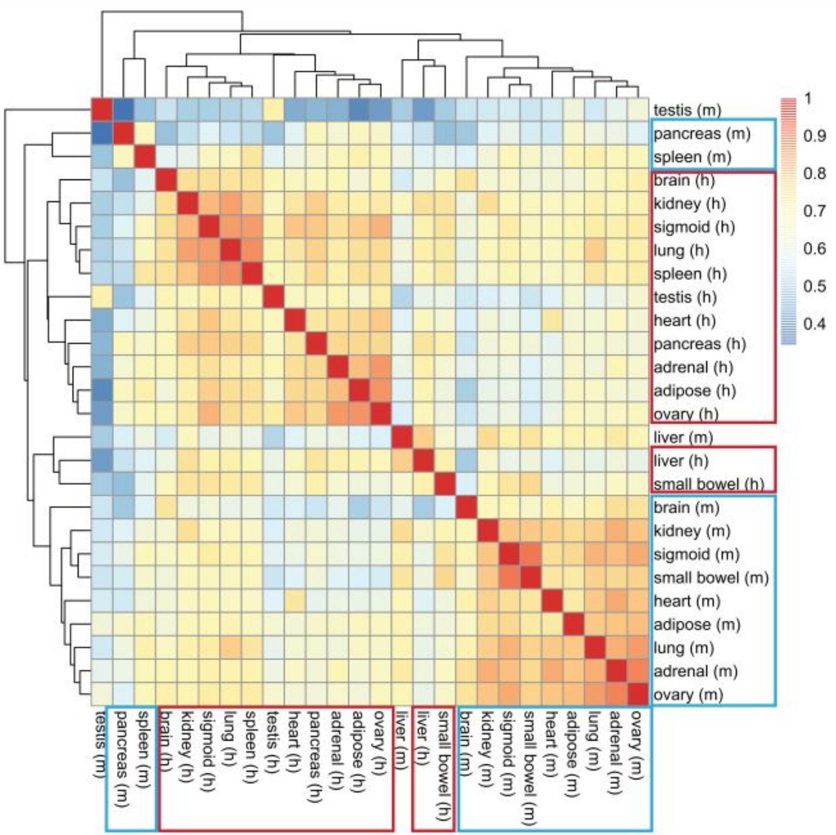
After normalization



Clustering of samples

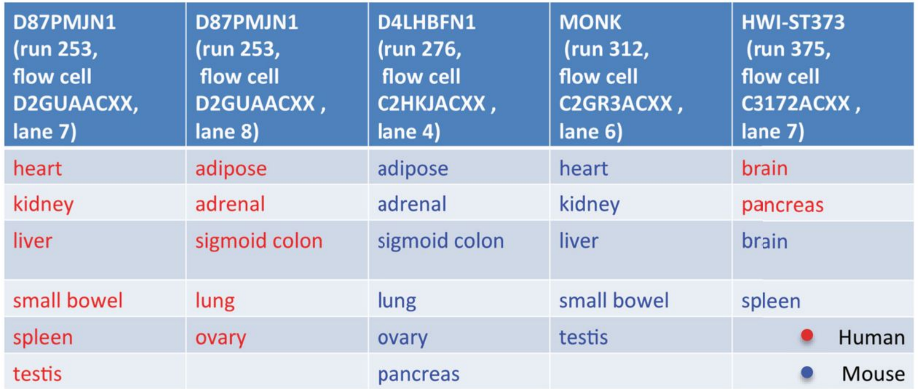


Confounding variables



Mouse ENCODE comparative gene expression data

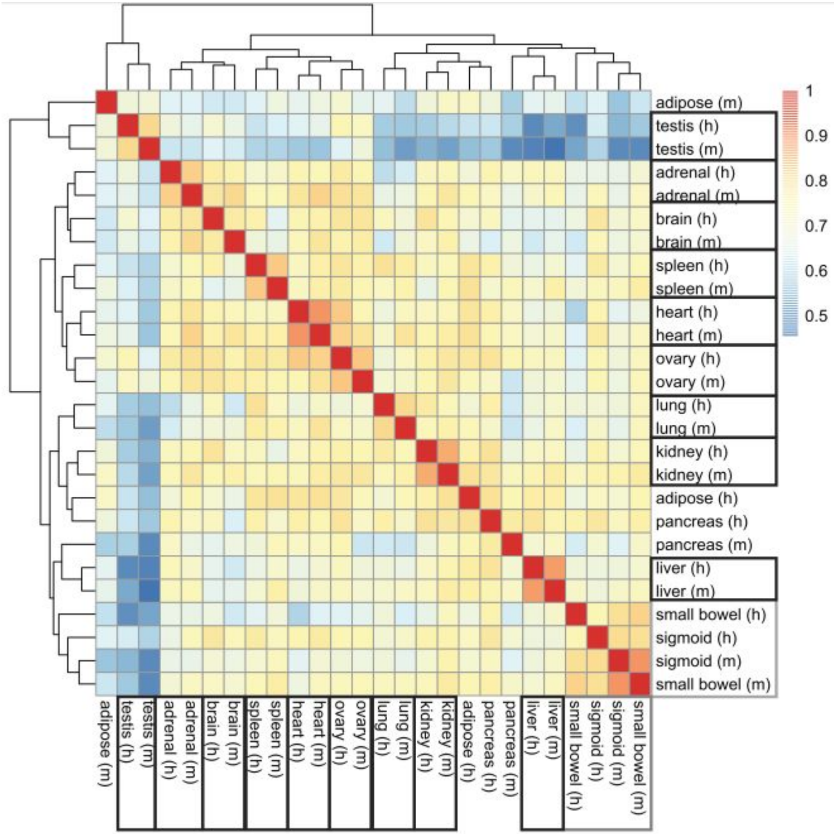
Lin et al. (2011) Comparison of the transcriptional landscapes between human and mouse tissues. PNAS 111:17224.



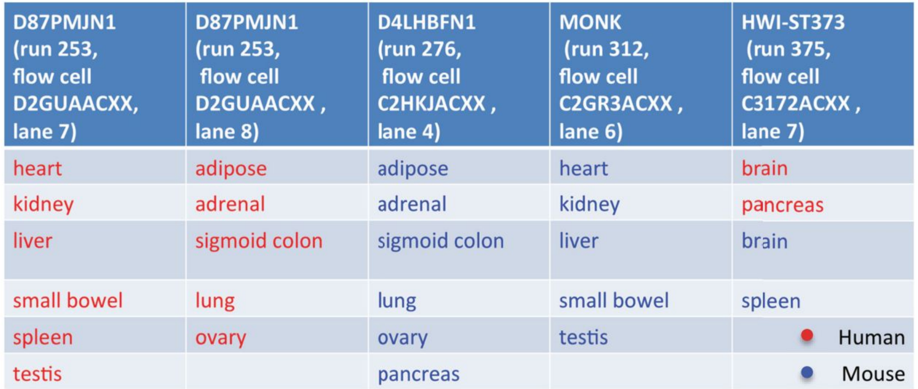
D87PMJN1 (run 253, flow cell D2GUAACXX, lane 7)	D87PMJN1 (run 253, flow cell D2GUAACXX , lane 8)	D4LHBFN1 (run 276, flow cell C2HKJACXX , lane 4)	MONK (run 312, flow cell C2GR3ACXX , lane 6)	HWI-ST373 (run 375, flow cell C3172ACXX , lane 7)
heart	adipose	adipose	heart	brain
kidney	adrenal	adrenal	kidney	pancreas
liver	sigmoid colon	sigmoid colon	liver	brain
small bowel	lung	lung	small bowel	spleen
spleen	ovary	ovary	testis	🔴 Human
testis		pancreas		🔵 Mouse

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4516019/

Confounding variables



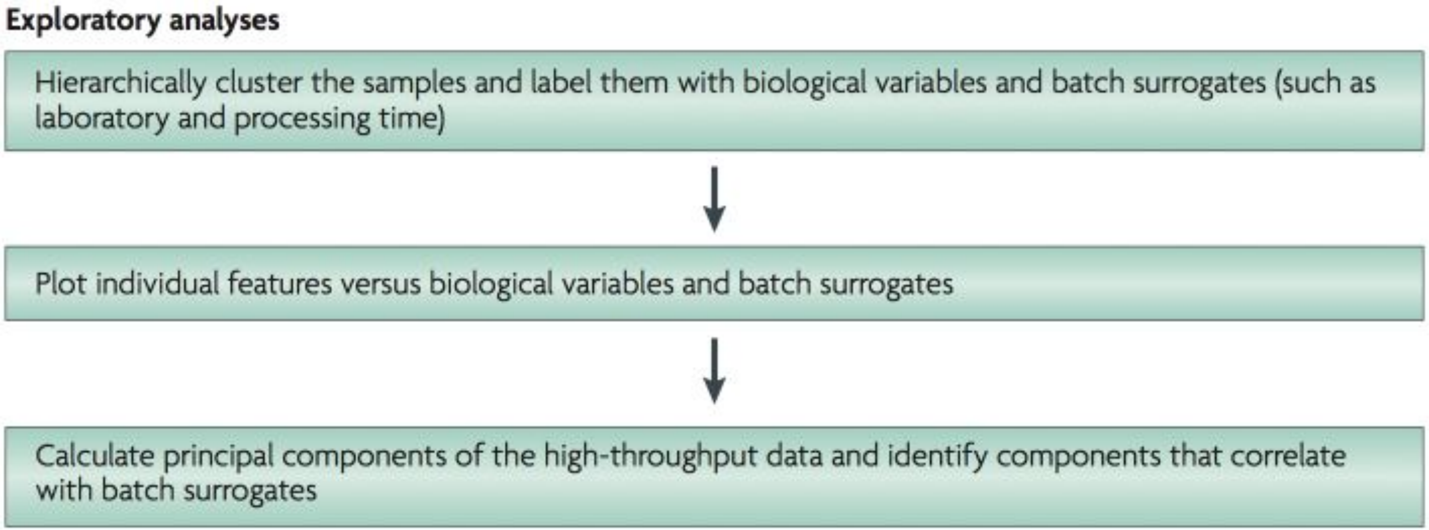
Re-analysis of the data after correcting for batch-effects.



D87PMJN1 (run 253, flow cell D2GUAACXX, lane 7)	D87PMJN1 (run 253, flow cell D2GUAACXX , lane 8)	D4LHBFN1 (run 276, flow cell C2HKJACXX , lane 4)	MONK (run 312, flow cell C2GR3ACXX , lane 6)	HWI-ST373 (run 375, flow cell C3172ACXX , lane 7)
heart	adipose	adipose	heart	brain
kidney	adrenal	adrenal	kidney	pancreas
liver	sigmoid colon	sigmoid colon	liver	brain
small bowel	lung	lung	small bowel	spleen
spleen	ovary	ovary	testis	● Human
testis		pancreas		● Mouse

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4516019/

Confounding variables



Exploratory analyses

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

↓

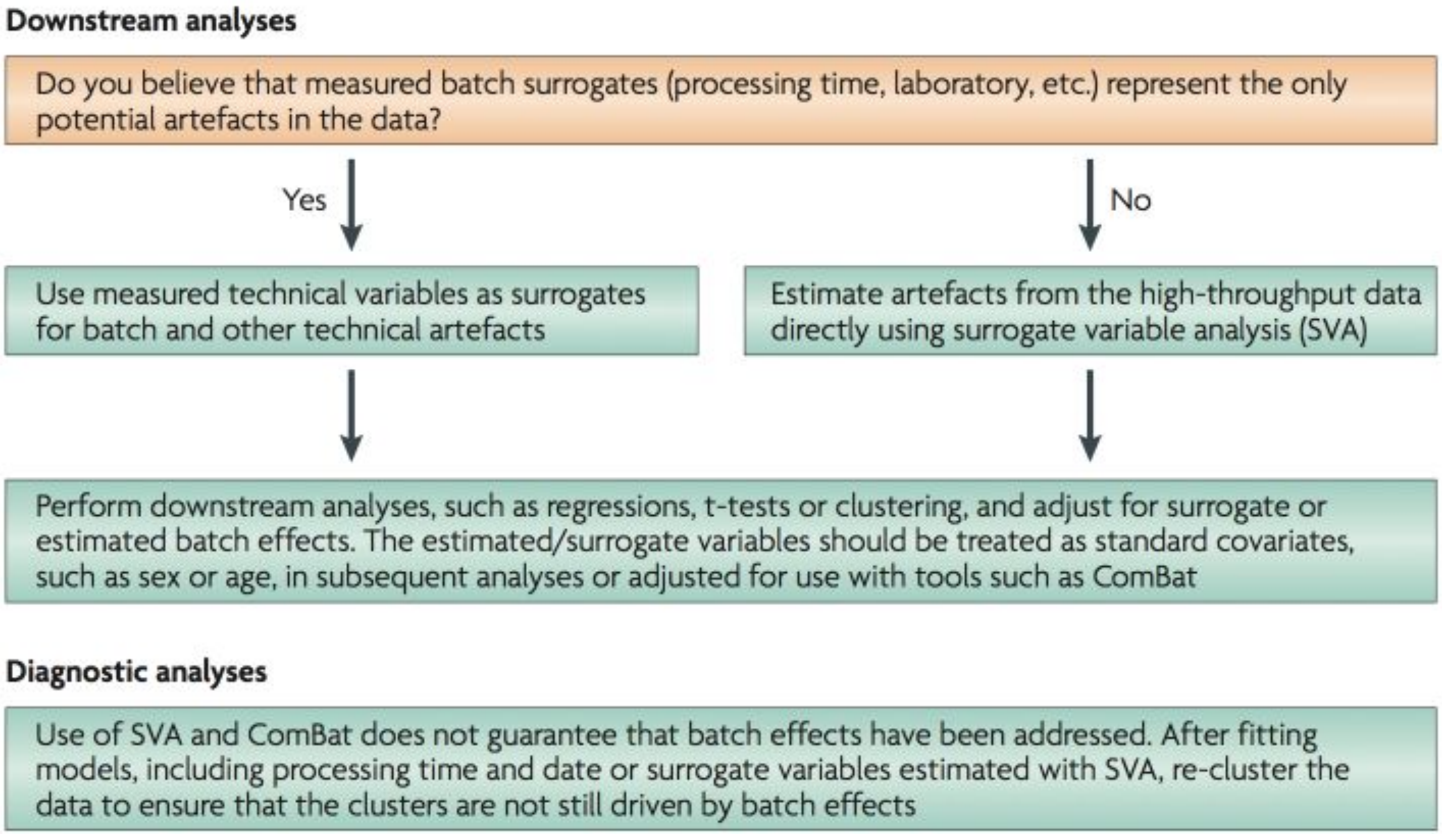
Plot individual features versus biological variables and batch surrogates

↓

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

https://www.nature.com/articles/nrg2825

Confounding variables



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

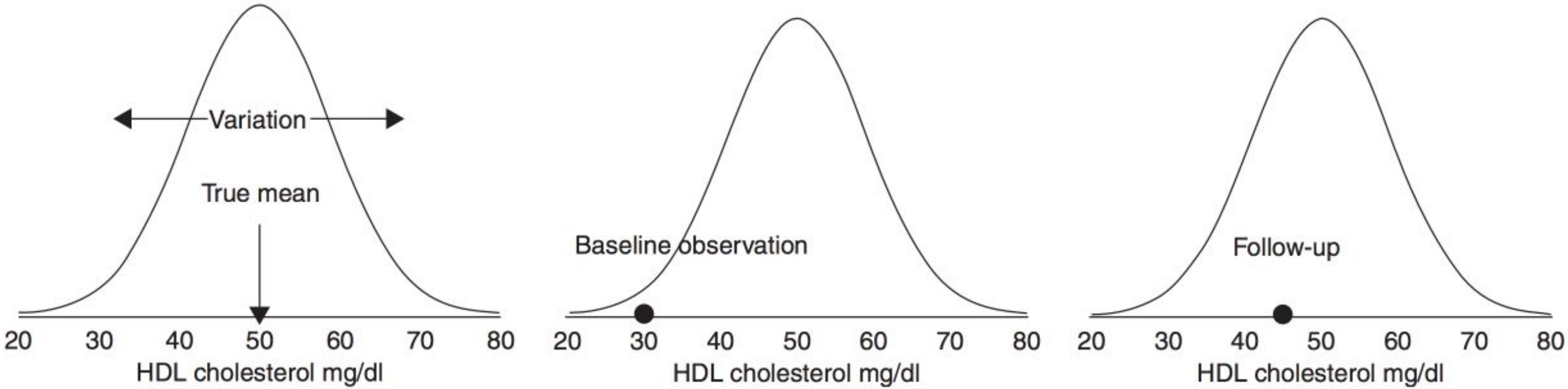
Diagnostic analyses

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

https://www.nature.com/articles/nrg2825

Regression to the mean

Regression to the mean (RTM) is a statistical phenomenon that is characterized by the fact that unusually large or small measurements tend to be followed by measurements that are closer to the mean.

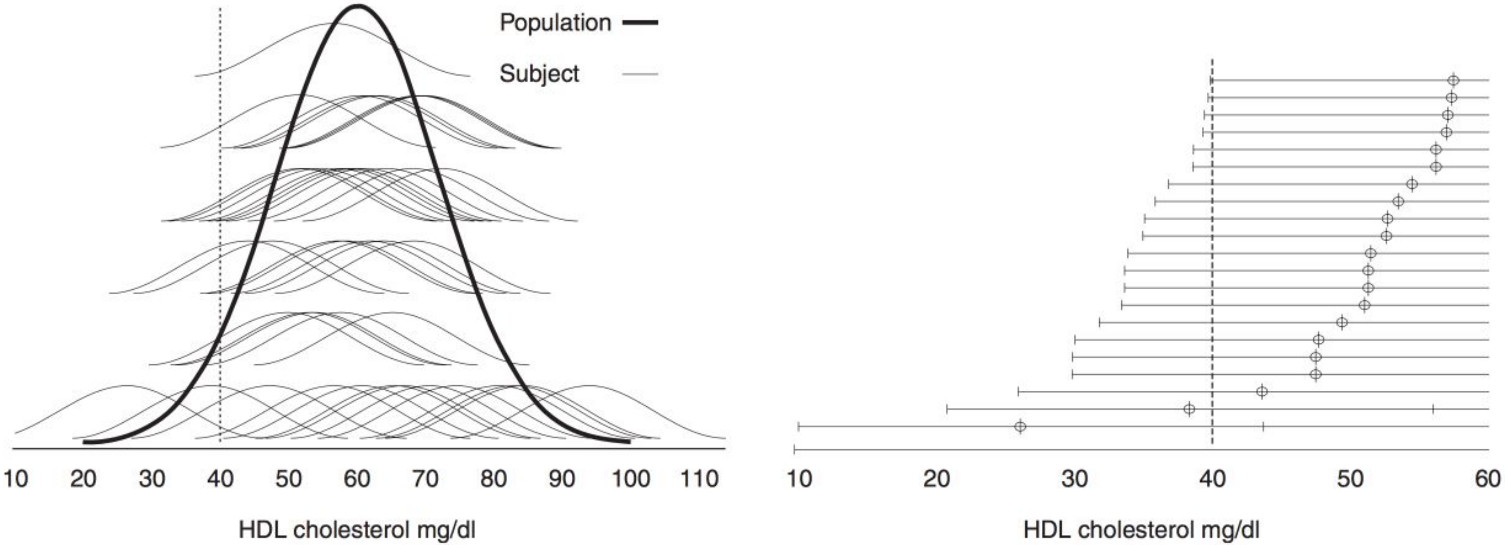


Regression to the mean

Regression to the mean (RTM) is a statistical phenomenon that is characterized by the fact that unusually large or small measurements tend to be followed by measurements that are closer to the mean.

- Occurs when repeated measurements are made on the same subject or unit of observation.
- Can make natural variation in repeated data look like real change.
- Happens because values are observed with random error (non-systematic variation like random measurement error or random fluctuations in a subject).
- There is almost no data without random error → makes RTM a common phenomenon.

Regression to the mean



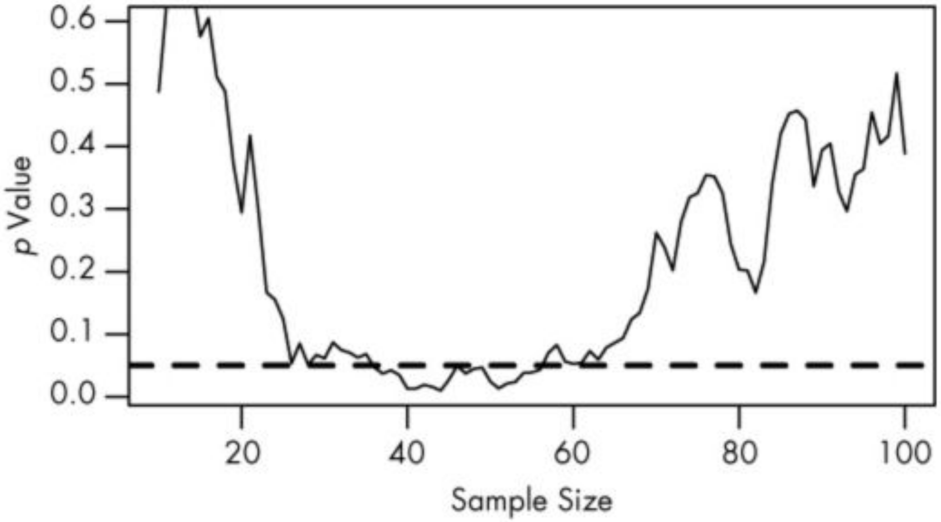
- Effect of RTM is compounded by categorizing subjects into groups based on baseline measurements.
- Variability in individual measurements > variability in the true means → Attenuation of association (regression dilution bias).

https://academic.oup.com/ije/article/34/1/215/638499

Regression to the mean

Longitudinally tracking the effect of drug.

- Terminate the study early if there if it is clear that the drug has an effect.
- In fact, it is *unethical* to withhold the drug from the control group.



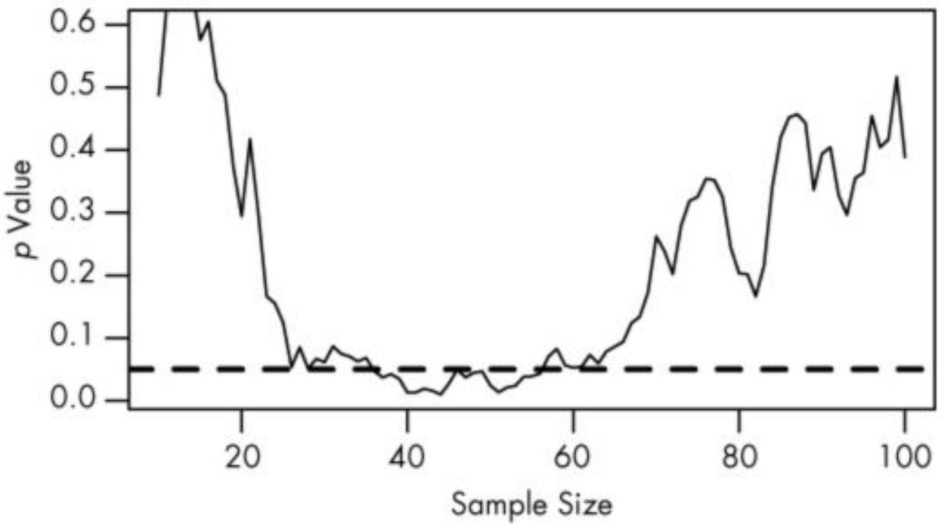
Issues:

- Null hypothesis should take varying group size into account.
- Truth inflation (lucky patients, not brilliant drugs): stopped trials exaggerate their effect by 29% more than trials that run their full course.

Regression to the mean

Longitudinally tracking the effect of drug.

- Terminate the study early if there if it is clear that the drug has an effect.
- In fact, it is *unethical* to withhold the drug from the control group.



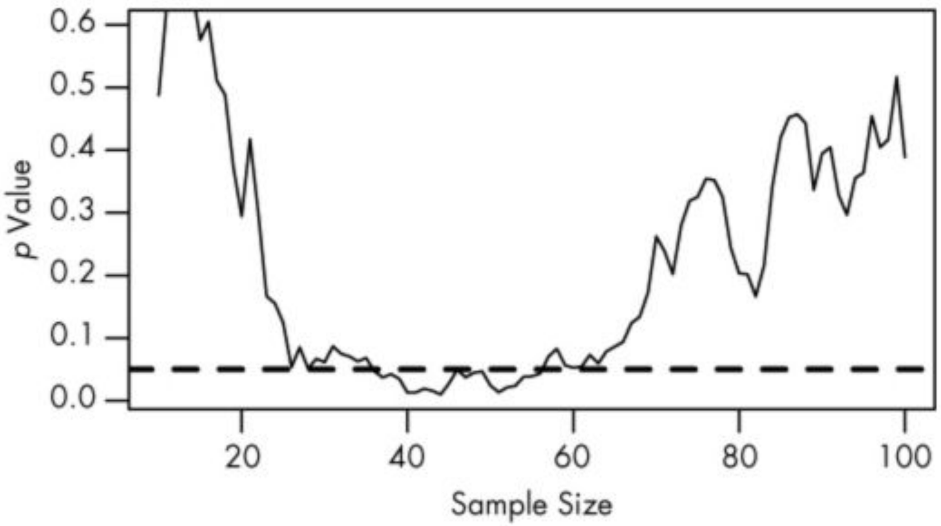
- Many published studies do not publish their:
  - original intended sample size or
  - the stopping rule used to justify terminating the study

Regression to the mean

Longitudinally tracking the effect of drug.

- Terminate the study early if there if it is clear that the drug has an effect.
- In fact, it is *unethical* to withhold the drug from the control group.

- Preregistration!
  - Statistical protocols
  - Few pre-selected evaluation points



Regression to the mean

Assigning a narrative or causal reasoning from observed data is often very hard:

- Galton's observation: Tall parents had (on average) children who were shorter than them, and short parents had (on average) children who were taller than them.
- “Norway had a great first jump; he will be tense, hoping to protect his lead and will probably do worse”; “Sweden had a bad first jump and now he knows he has nothing to lose and will be relaxed, which should help him do better.
- Depressed children treated with an energy drink improve significantly over a three-month period.

Circular analysis & Double-dipping

Statistical analysis is often exploratory: no hypothesis is advance.

- Collect data → Poke around to see if there's something interesting → New hypotheses → Perform new experiments / Collect new data → Test the hypotheses.
- Collect data → Poke around to see if there's something interesting → New hypotheses → Take the subset of the original data that appears to show signal →Test the hypotheses.
  - Double-dipping → truth inflation.
  - Happens all the time in neuroimaging (apparently 40% of the literature), genetics, epidemiology.

Circular analysis & Double-dipping

- Collect data → Poke around to see if there's something interesting → New hypotheses → Take the subset of the original data that appears to show signal → Test the hypotheses.
  - Null hypothesis based on random chance is wrong at the final stage.
  - Only signals with the strongest random noise make it into further analysis.
- Mitigating this problem:
  - Split data in half; Reduces power.
  - Choose hypotheses based on prior knowledge.