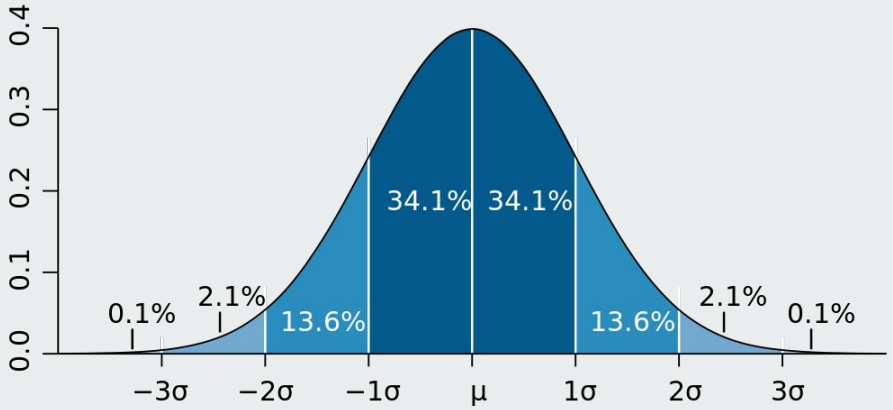
Introduction to Statistics

iPQB Bootcamp Statistics Module
September 4, 2018
Maureen Pittman



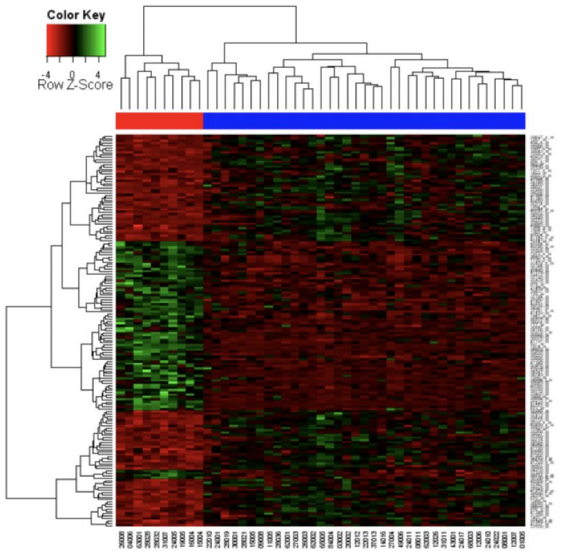
UCSF
University of California
San Francisco

Statistics in Practice

https://en.wikipedia.org/wiki/Statistics ▾

Statistics is a branch of mathematics dealing with the collection, organization, analysis, interpretation and presentation of data.

- How many subjects should I include in my experiment?
- Is there a relationship between the independent and dependent variables?
  - Is the relationship causative or associative?
- Comparing two observed states:
  - Are genes differentially expressed between two groups?
  - Are these genes enriched for some biological process?



https://warwick.ac.uk/fac/sci/moac/people/students/peter_cock/r/heatmap/

Statistical Goals

Abstract question → mathematical equation

- Inference: Is a drug effective? → Is $p < 0.05$?
- Prediction: How effective is a drug? → Central tendency and dispersion of treatment results
- Modeling: How does x process work? → What set of parameters/equations best describe it?
- Estimation: How many people in the US are blond? → What % of a sample are blond?

Essentials of statistical analysis

Probability

Sampling

Variables

Hypothesis Testing

Probability

- Unconditional probability: the likelihood that some event will occur.

$$P(A)$$

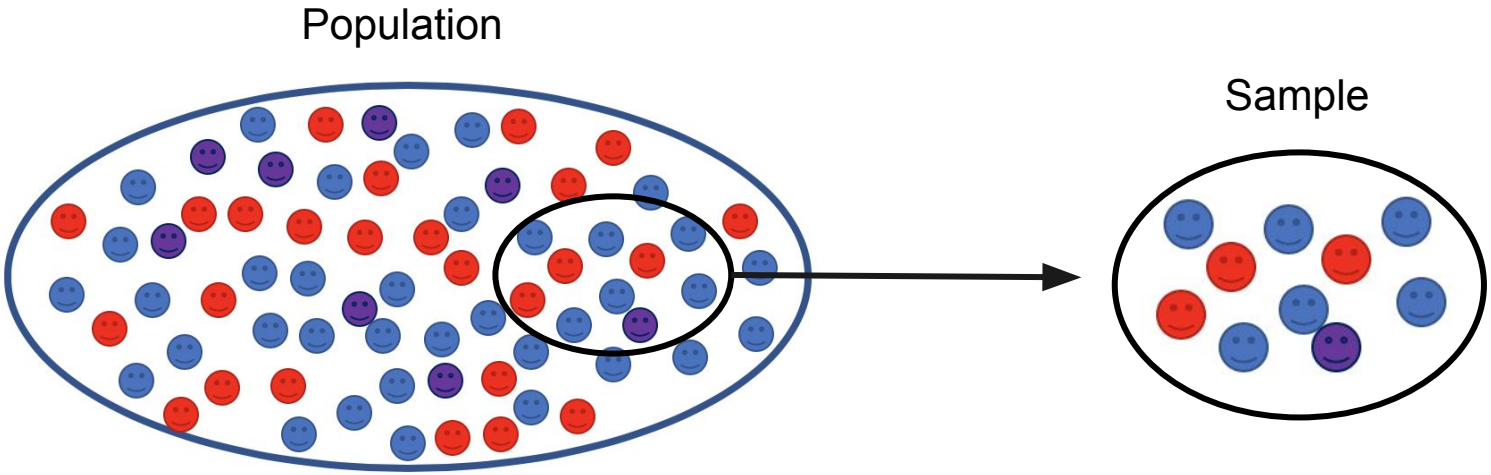
- Conditional probability: the likelihood of event A occurring given that B is true.

$$P(A \mid B)$$

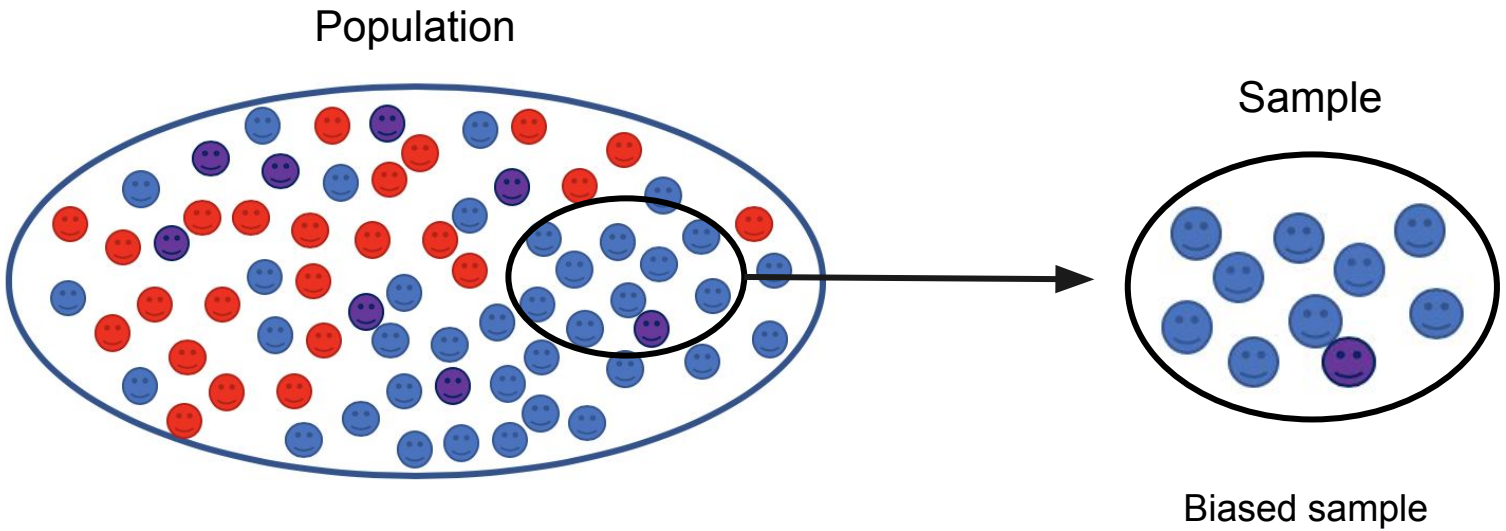
- Bayes Rule:

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

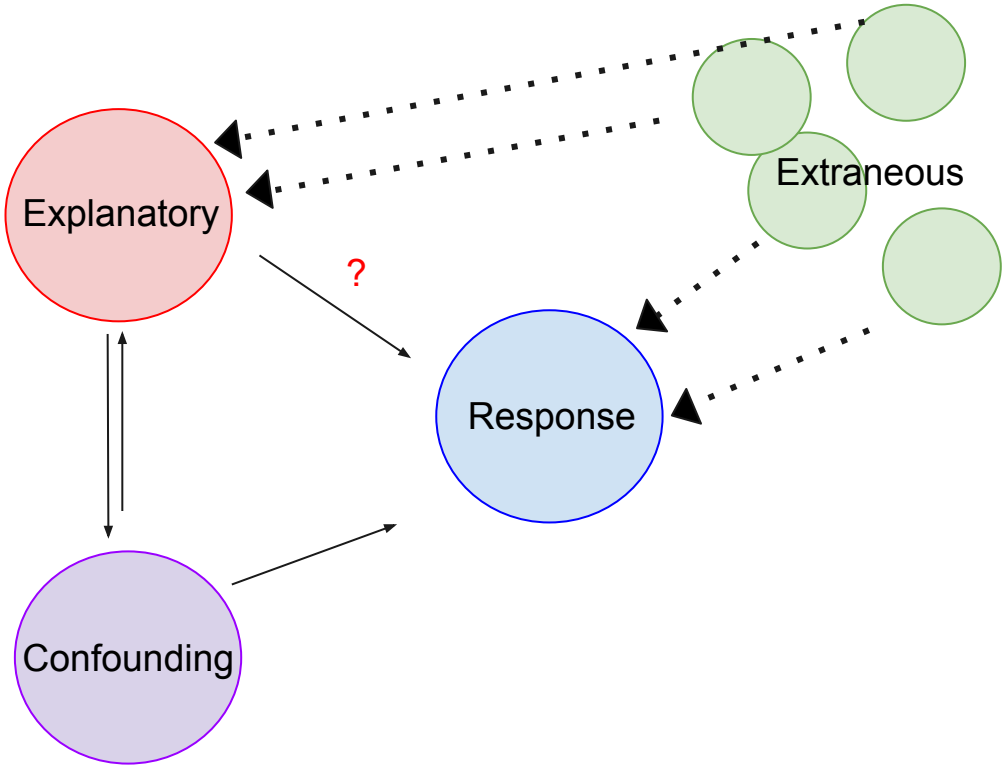
Sampling and Bias



Sampling and Bias



Variables and their interactions



Simpson's Paradox

- Gender bias at UC Berkeley

	Applicants	Admitted
Men	8442	**44%**
Women	4321	35%

Significance: $p \approx 10^{-26}$

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	373	6%	341	**7%**

"Small but significant bias in favor of women in most departments"

Simpson's Paradox

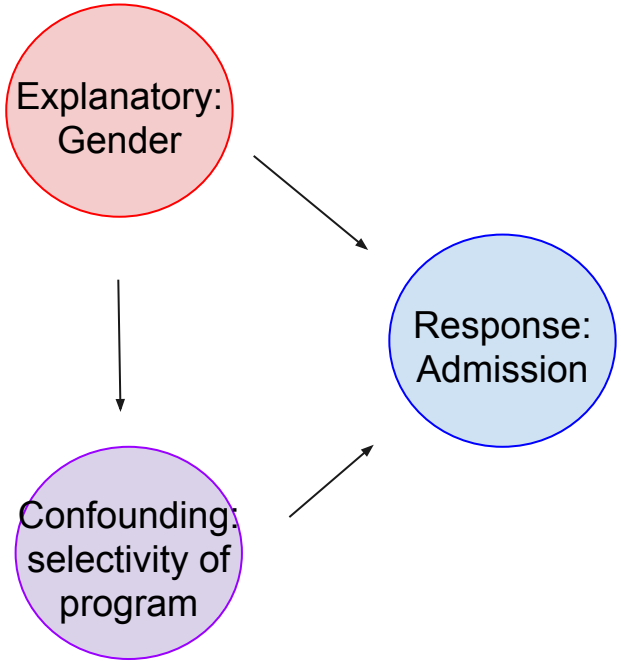
- Gender bias at UC Berkeley

	Applicants	Admitted
Men	8442	**44%**
Women	4321	35%

Significance: $p \approx 10^{-26}$

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	373	6%	341	**7%**

"Small but significant bias in favor of women in most departments"



Variable types

Categorical

- Nominal

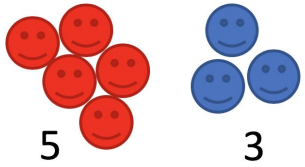
categories with no intrinsic order

- Ordinal

ordered categories

Numerical

- Discrete



e.g. integers, count data

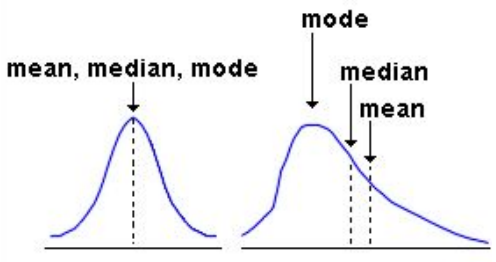
- Continuous

infinite possible values

Data Properties

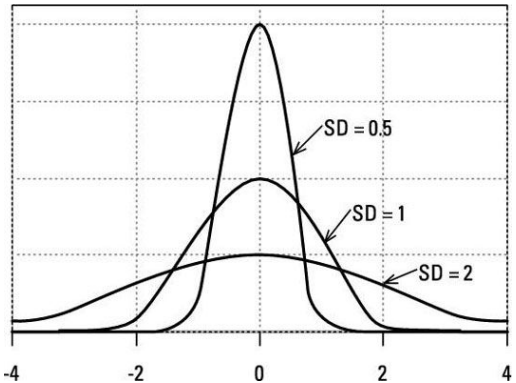
Central Tendency

- Mean
- Median
- Mode



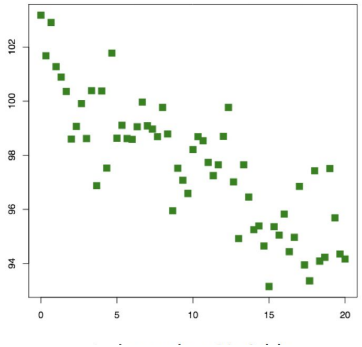
Dispersion

- Variance
- Standard deviation



Multivariate

- Covariance
- Correlation



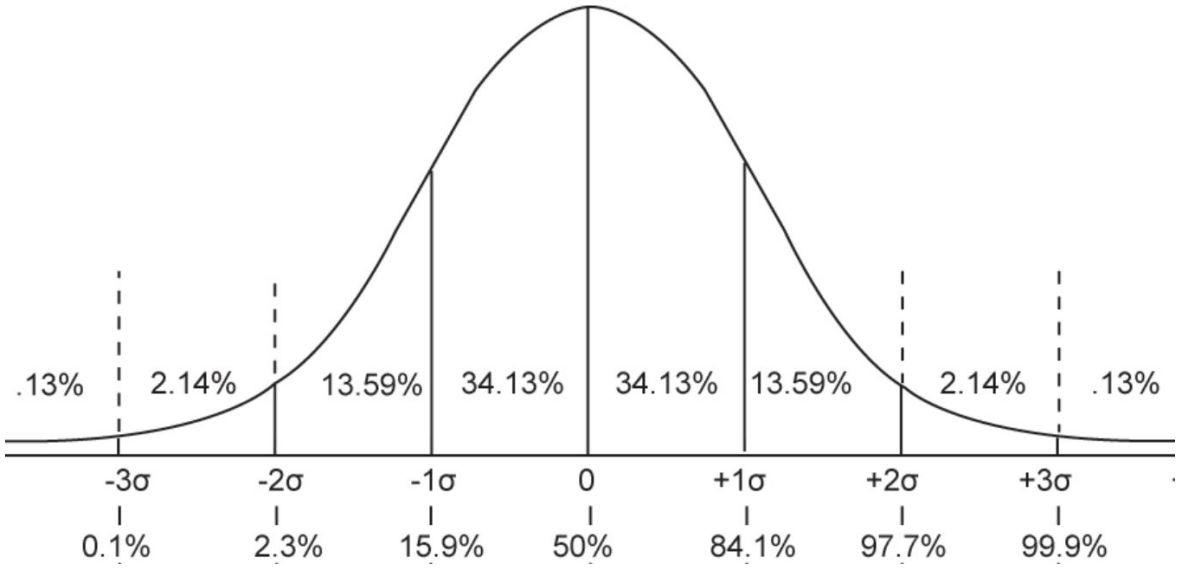
Hypothesis Testing

- H_0 : there is no relationship between the two variables
- H_1 : the variables are associated

A neurologist is testing the effect of a drug on response time by injecting 100 rats with a unit dose of the drug, subjecting each to neurological stimulus, and recording its response time. The neurologist knows that the mean response time for rats not injected with the drug is 1.2 seconds. The mean of the 100 injected rats' response times is 1.05 seconds with a sample standard deviation of 0.5 seconds.

https://www.khanacademy.org/math/statistics-probability/significance-tests-one-sample/idea-of-significance-tests/e/

A neurologist is testing the effect of a drug on response time by injecting 100 rats with a unit dose of the drug, subjecting each to neurological stimulus, and recording its response time. The neurologist knows that the mean response time for rats not injected with the drug is 1.2 seconds. The mean of the 100 injected rats' response times is 1.05 seconds with a sample standard deviation of 0.5 seconds.



Error Types

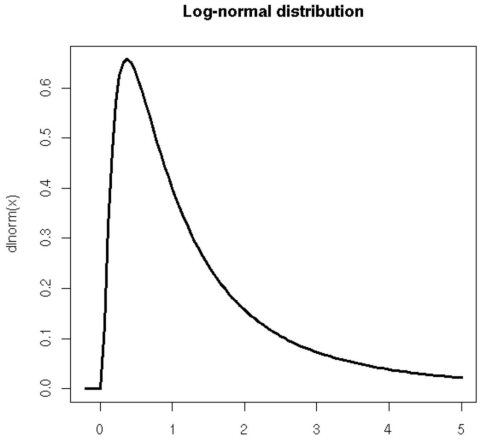
- Type I: the null hypothesis was true, but you rejected it
- Type II: the null hypothesis was false, but you fail to reject it

Table of error types		Null hypothesis (H_0) is	
		True	False
Decision About Null Hypothesis (H_0)	Fail to reject	Correct inference (True Negatives)	Type II error (False Negative)
	Reject	Type I error (False Positive)	Correct inference (True Positives)

Intuitive definitions

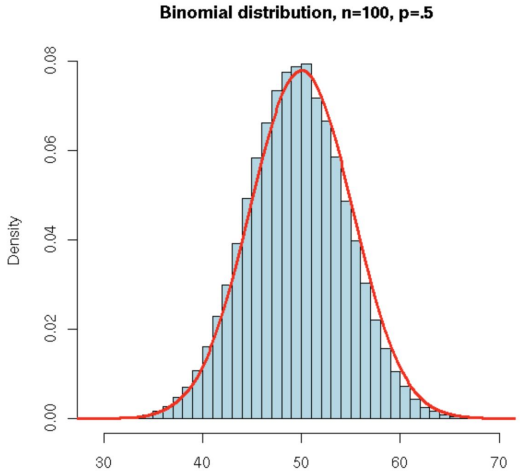
- P-value: the probability of seeing a result as extreme (or more extreme) as the one observed, assuming the null hypothesis is true.
- P-value: the probability of making a Type I error (rejecting the null when it is true).
- P-value cutoff (also called α, often set to 0.05): the level of uncertainty you're willing to accept in order to reject the null.

Other kinds of distributions



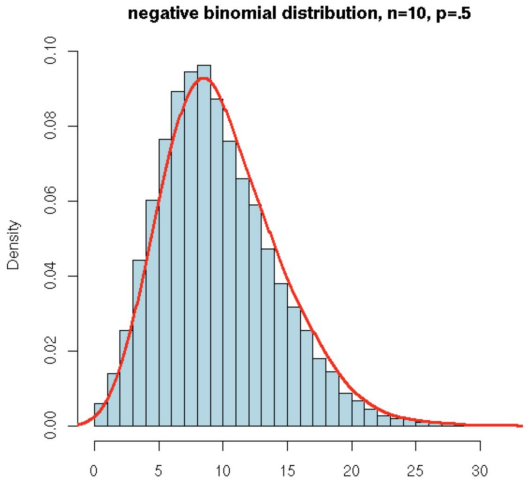
$Y = \ln(X)$ has a normal distribution

Example: Normalized RNA-seq counts



Discrete probability distribution of a series of true/false questions

Example: Inheritance of an allele to the next generation

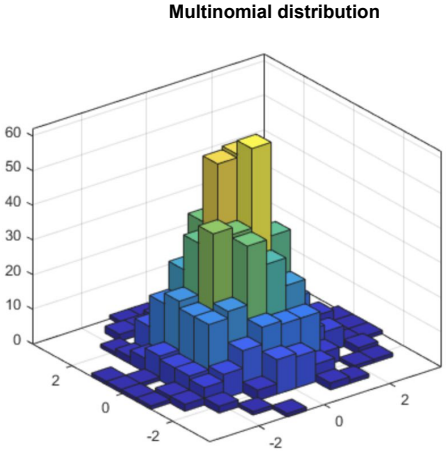


Binomial distribution of the number of successes before a specified number of failures

Example: overdispersed data (data with high variance)

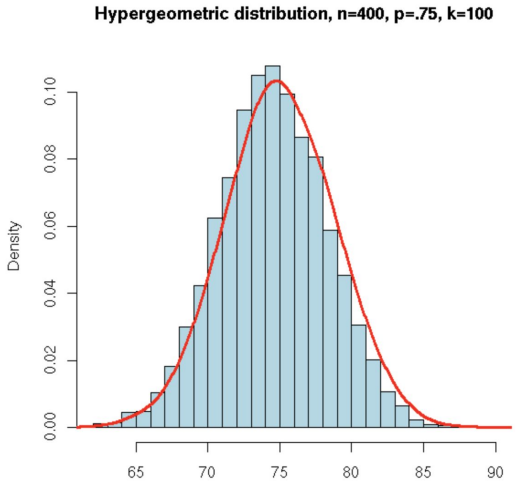
http://zoonek2.free.fr/UNIX/48_R/07.html

Other kinds of distributions



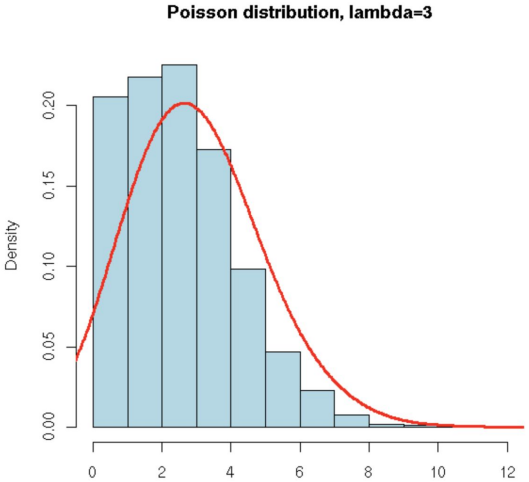
Generalization of the binomial distribution

Example: inheritance of an allele at a triallelic site



Discrete probability of sampling k times from a group of n, without replacement

Example: modeling the number of activated synapses in ganglion



Models how many times an event will occur in an interval given a rate (lambda)

Example: rare events with predictable variance (de novo mutation arising)

http://zoonek2.free.fr/UNIX/48_R/07.html

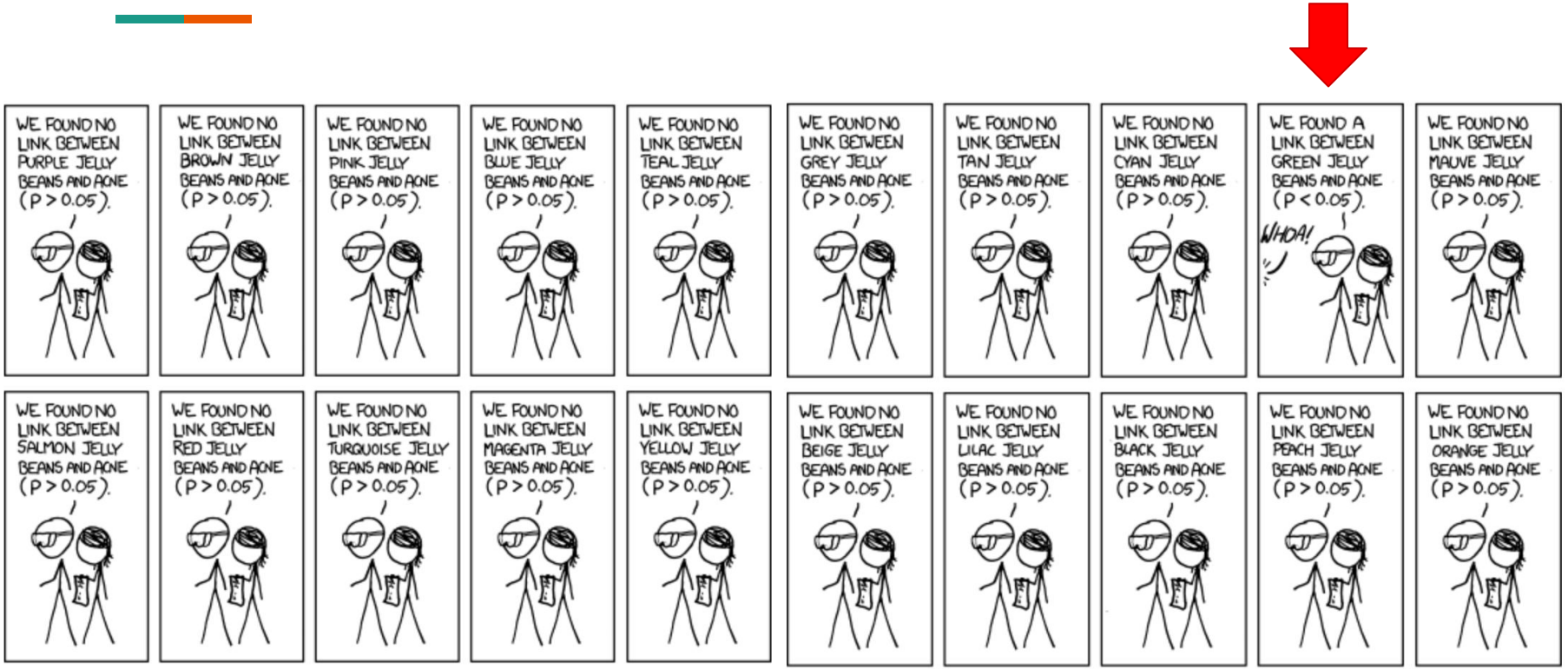
Crimes against data

Confounding variables

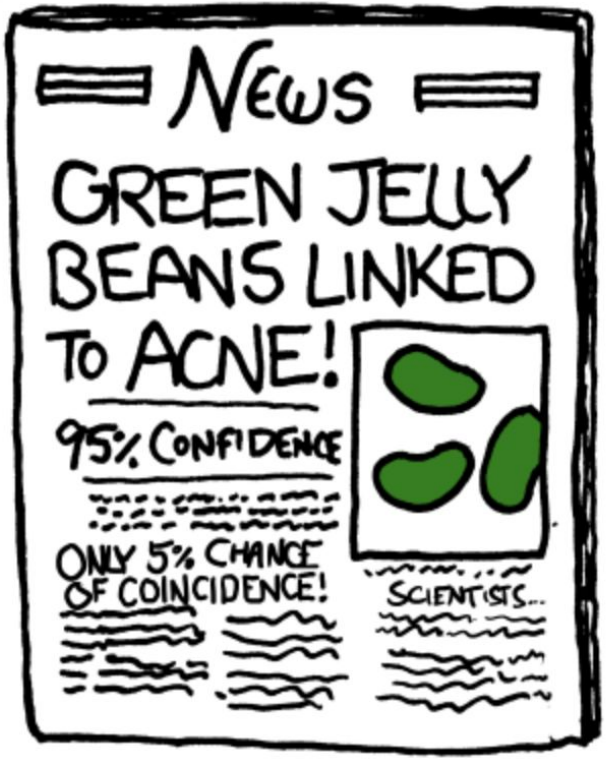
Multiple hypotheses

Publication bias

Multiple Hypotheses

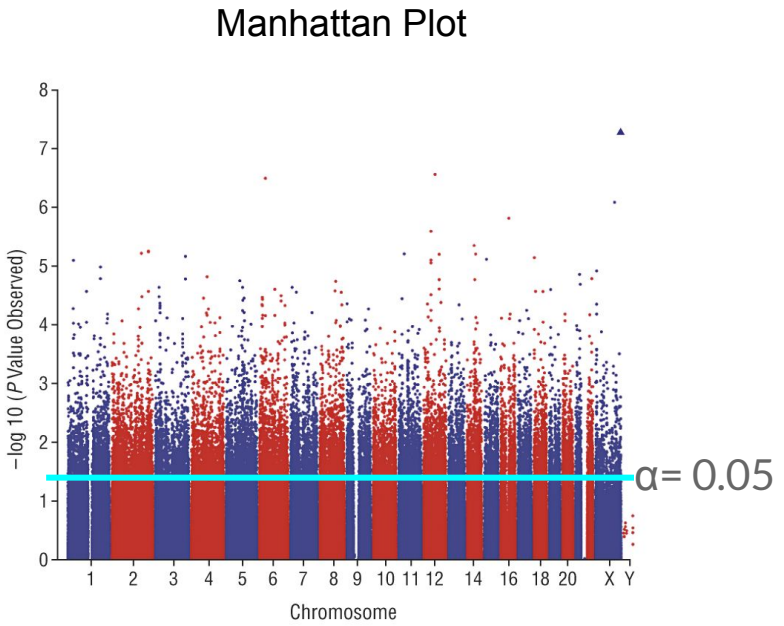


https://xkcd.com/882/


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Correction for Multiple Hypotheses

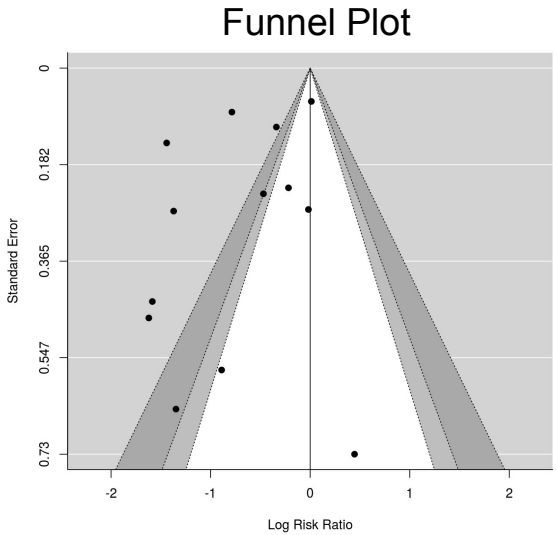
- Bonferroni correction: divide α by the number of hypotheses, so that the new cutoff is α/m
- Sidak correction: $\alpha_{SID} = 1 - (1 - \alpha)^{\frac{1}{m}}$
- Family-Wise Error Rate (FWER)



Publication Bias

- Studies with positive results are more likely to:
  - Be published
  - Be cited by others
  - Produce multiple publications

- Replication crisis



Retraction Watch

Tracking retractions as a window into the scientific process

Summary

- Statistics is math for data analysis
- Populations - sampling and bias
- Variables
  - Roles: Explanatory, Response, Extraneous, Confounding
  - Types: Categorical (nominal & ordinal); Numerical (discrete & continuous)
- Data properties
  - Central tendency
  - Dispersion
  - Correlation/association
- Hypothesis testing
  - What does a p-value actually mean?
- Types of distributions
  - Normal, binomial, negative binomial
  - Multivariate, hypergeometric, Poisson
- Pitfalls
  - Confounding variables
  - Multiple hypotheses
  - Publication bias

Acknowledgements


Kevin Hartmann


Katie Pollard

Summary

- Statistics is math for data analysis
- Populations - sampling and bias
- Variables
  - Roles: Explanatory, Response, Extraneous, Confounding
  - Types: Categorical (nominal & ordinal); Numerical (discrete & continuous)
- Data properties
  - Central tendency
  - Dispersion
  - Correlation/association
- Hypothesis testing
  - What does a p-value actually mean?
- Types of distributions
  - Normal, binomial, negative binomial
  - Multivariate, hypergeometric, Poisson
- Pitfalls
  - Confounding variables
  - Multiple hypotheses
  - Publication bias

Questions?