

Correlation Power analysis Multiple hypothesis testing



Biostatistics Course 2023
Lecture 4
Thursday, 27 July 2023
1:00pm - 3:00pm

Correlation

Example: lipids and insulin sensitivity

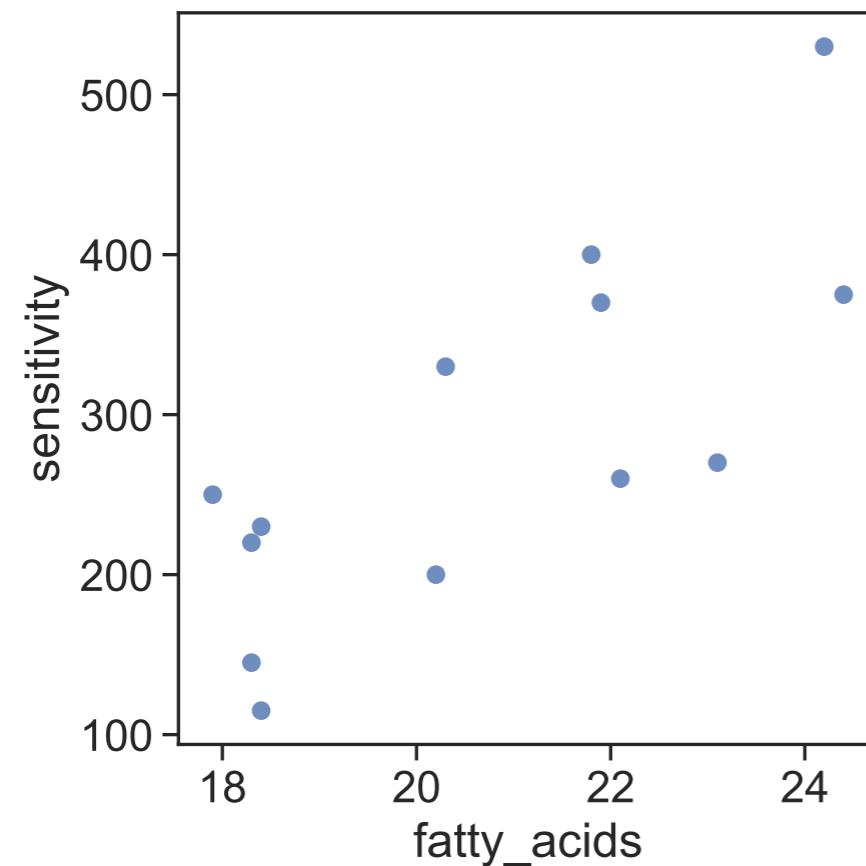
sensitivity	fatty_acid
250	17.9
220	18.3
145	18.3
115	18.4
230	18.4
200	20.2
330	20.3
400	21.8
370	21.9
260	22.1
270	23.1
530	24.2
375	24

Borkman et al. (1993) wanted to understand why insulin sensitivity varies so much among individuals. They hypothesized that the lipid composition of the cell membranes of skeletal muscle affects the sensitivity of the muscle for insulin.

They determined the insulin sensitivity of $N = 13$ healthy men by infusing insulin at a standard rate (adjusting for size differences) and quantifying how much glucose they needed to infuse to maintain a constant a blood glucose level...

They also took a small muscle biopsy from each subject and measured its fatty acid composition. We'll focus on the fraction of of polyunsaturated fatty acids that have between 20 and 22 carbon atoms ("fatty_acid").

Correlation is used to describe relationships between real-numbered variables



summary statistics

pearson	
N	13
r	0.77
95% CI	[0.38, 0.93]
r^2	0.593
P-val	0.00207701

Covariance and correlation are estimated from data in the familiar manner

The formula for variance is

$$\widehat{\text{var}}(x) = \sigma_x^2 = \frac{1}{N-1} \sum_i (x_i - \hat{\mu}_x)^2$$

Covariance is estimated in a manner similar to variance

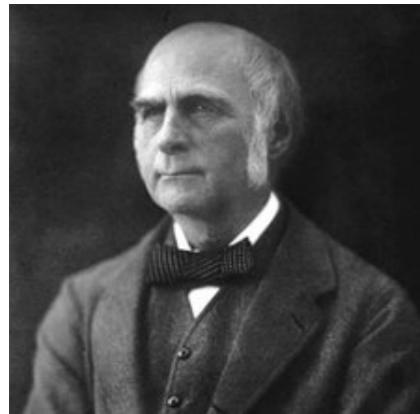
$$\widehat{\text{cov}}(x, y) = \frac{1}{N-1} \sum_i (x_i - \hat{\mu}_x)(y_i - \hat{\mu}_y)$$

The corresponding “correlation coefficient” is

$$r = \frac{\widehat{\text{cov}}(x, y)}{\hat{\sigma}_x \hat{\sigma}_y}$$

The correlation coefficient, as used today, was developed by Karl Pearson and Francis Galton in the 1880s

$$r = \frac{\widehat{\text{cov}}(x, y)}{\hat{\sigma}_x \hat{\sigma}_y}$$
 is often called “Pearson correlation”



Francis Galton
(1822-1911)

- Discovered the correlation coefficient in 1888 (independent of Auguste Bravis, in 1844) and proposed the use of “ r ”.
- Also invented linear regression & idea of “regression toward the mean”
- Invented the term “eugenics” in 1883



Karl Pearson
(1857-1936)

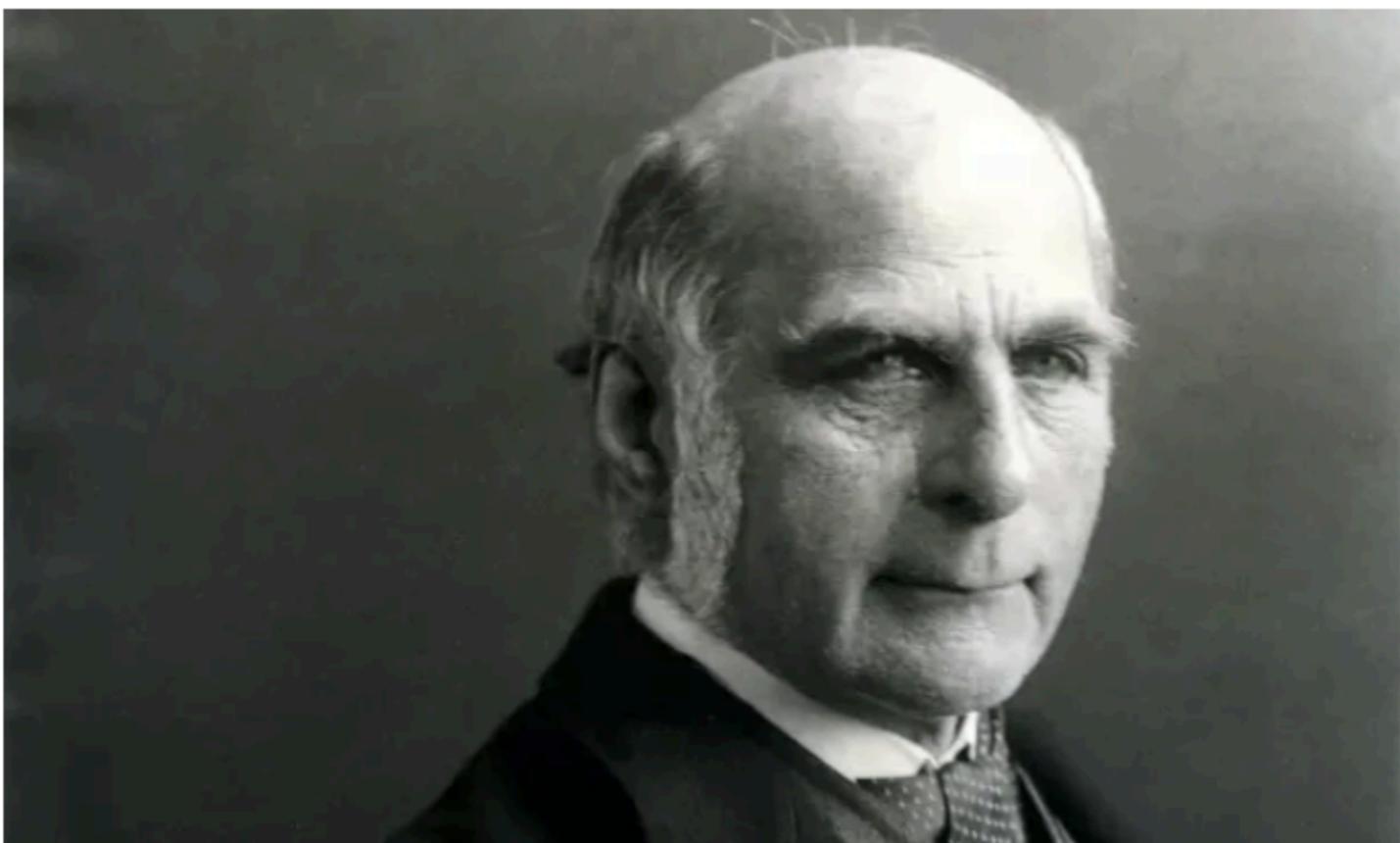
- Is credited with founding the discipline of mathematical statistics.
- Further developed the correlation coefficient proposed by Galton
- Other contributions include P-values, χ^2 tests, and principal component analysis (PCA)
- Founded world’s first university statistics department at University College, London
- Was a vocal and influential social Darwinist and advocate of race wars.
- Founded the “Annals of Eugenics”, which is now “Annals of Human Genetics”

UCL
(University
College
London)

• This article is more than 1 month old

UCL renames three facilities that honoured prominent eugenicists

London university removes names of Francis Galton and Karl Pearson from two lecture theatres and a building



▲ Francis Galton coined the term eugenics in 1883 and endowed UCL with his personal collection and archive.
Photograph: Corbis via Getty Images

UCL has renamed two lecture theatres and a building that honoured the prominent eugenicists Francis Galton and Karl Pearson.

 Replay

 Learn More

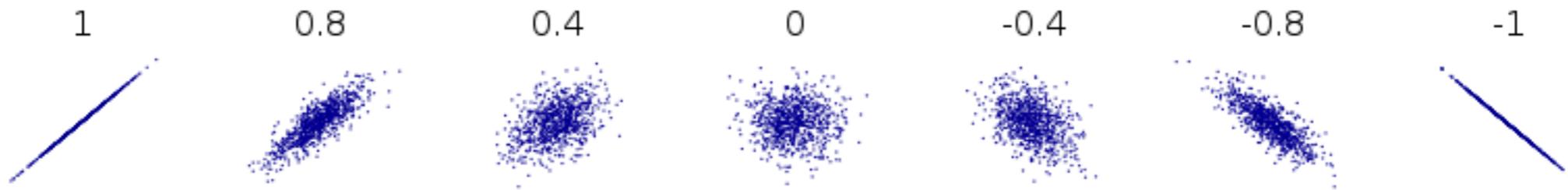
Sponsored Video

Advertisement by Advertising Partner

Watch to learn more

 SEE MORE

This is what the correlation coefficient looks like



Pearson's r ranges from -1 to 1.

$r = 0$ when the two variables are independent, i.e. $p(x, y) = p(x) \cdot p(y)$.

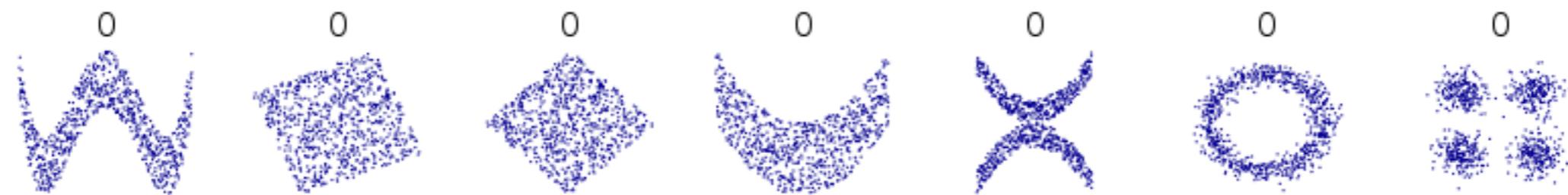
$r = \pm 1$ when the two variables share a deterministic linear relationship.

Adding a constant to all x or all y , or a multiplicative rescaling of all x or all y , do not change r .

This is what the correlation coefficient looks like



In the deterministic case, r is unaffected by the magnitude of the slope relating two variables, while the sign of r is equal to the sign of the slope.



Sometimes $r = 0$ when two variables have a nonlinear relationship.

The coefficient of determination another name for r^2

The coefficient of determination is simply r^2 , which is also often written as R^2 .

r^2 is always between 0 and 1 (inclusive)

Remember that $r^2 \leq |r|$, so beware of people reporting r instead of r^2 to make a correlation seem stronger.

r^2 is commonly interpreted as the fraction of variance in y explained by x (or the other way around).

P-values correspond to the null hypothesis of no correlation in the underlying distribution

The p-value reported alongside values of r or r^2 corresponds to the null hypothesis that the underlying correlation is zero.

If the underlying correlation is zero, then the quantity

$$t = r \sqrt{\frac{N - 2}{1 - r^2}}$$

follows a t-distribution with $N - 2$ degrees of freedom.

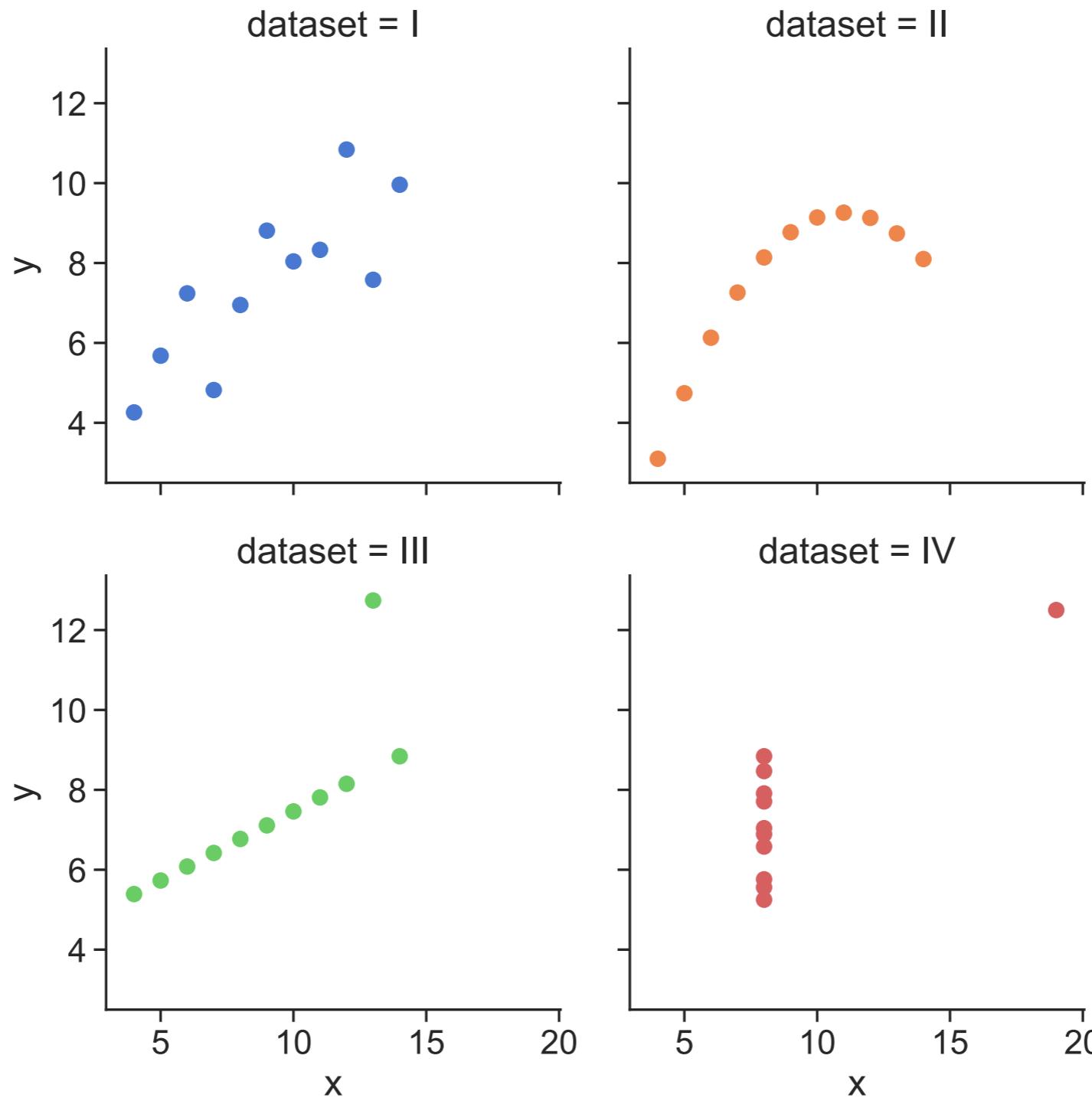
The inverse relationship

$$r = \frac{t}{\sqrt{N - 2 + t^2}}$$

is used to compute values for the 95% confidence interval on r .

Lots of different-looking datasets will have the same value for r .

“Anscombe’s quartet”: $r = 0.816$ for all 4 datasets



Assumptions underlying correlation

Interpreting the correlation coefficient r , and especially the associated P-value, requires multiple assumptions:

- Each data point (x, y) is independently sampled from a 2D Gaussian distribution.
- In particular, x and y each follow a 1D Gaussian distribution
- All covariation between x and y is linear, with perfect concordance disrupted only by Gaussian noise.

There are usually many explanations for why two variables might correlate

Possible reasons for a correlation between lipid levels and insulin sensitivity:

- The lipid content of membranes affects insulin sensitivity
- The insulin sensitivity affects membrane lipid content
- Both insulin sensitivity and lipid content are under the control of some third factor, such as a hormone.
- Lipid content, insulin sensitivity, and other factors are all part of a complex molecular/biochemical/physiological network, perhaps with positive and/or negative feedback components. The correlation observed is just a peak at a much more complex set of interdependent relationships.
- Membrane lipid content and insulin sensitivity don't actually correlate at all; the result is just a coincidence.

Welcome to GraphPad Prism

XY tables: Each point is defined by an X and Y coordinate

NEW TABLE & GRAPH

XY 

Column

Group

Contingency

Survival

Parts of Whole

Multiple variables

Nested

EXISTING FILE

Open a File

LabArchives

Clone a Graph

Graph Portfolio

GraphPad Prism Version 8.2.1 (279)

Data table:

Enter or import data into a new table

Start with sample data to follow a tutorial

Options:

X: Numbers

Numbers with error values to plot horizontal error bars

Dates

Elapsed times

Y: Enter and plot a single Y value for each point

Enter **3**  replicate values in side-by-side subcolumns

Enter and plot error values already calculated elsewhere

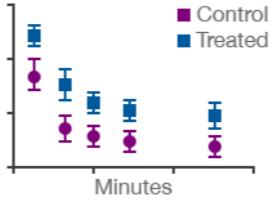
Enter: **Mean, SD, N** 

Prism Tips

Cancel

Create 

	X	A			B		
	Minutes	Control			Treated		
	X	A:Y1	A:Y2	A:Y3	B:Y1	B:Y2	B:Y3
1	Title						
2	Title						
3	Title						



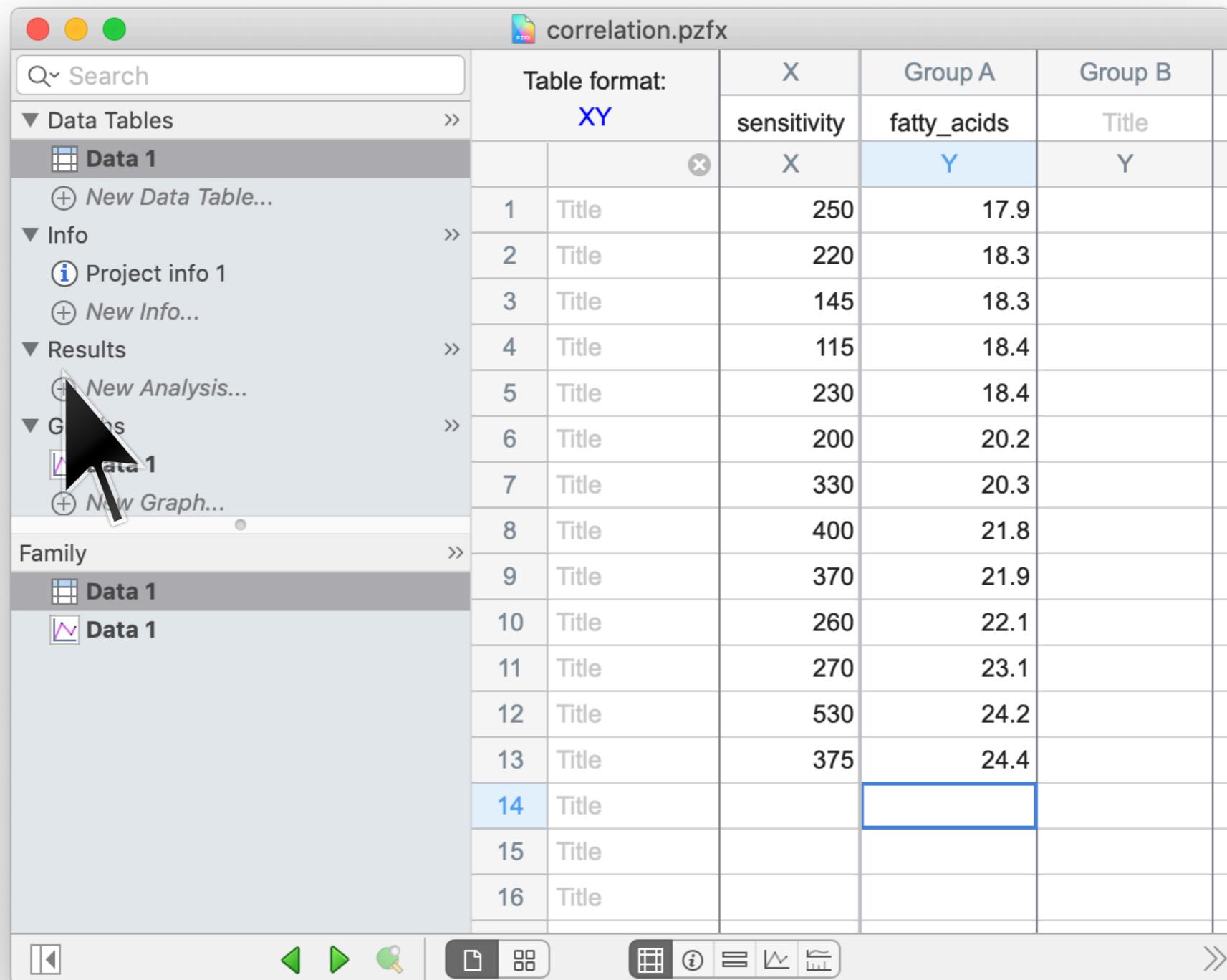
Minutes

Control

Treated

?

Learn more



Create New Analysis

Data to analyze

Table: Data 1

Type of analysis

Which analysis?

▼ Transform, Normalize...

- Transform
- Transform concentrations (X)
- Normalize
- Prune rows
- Remove baseline and column math
- Transpose X and Y
- Fraction of Total

▼ XY analyses

- Nonlinear regression (curve fit)
- Linear regression
- Fit spline/LOWESS
- Smooth, differentiate or integrate curve
- Area under curve
- Deming (Model II) linear regression
- Row means with SD or SEM

Correlation

- Interpolate a standard curve

► Column analyses

► Grouped analyses

► Contingency table analyses

► Survival analyses

Analyze which data sets?

A:fatty_acids

When you analyze tables or graphs with more than one data set, use this space to select which data set(s) to analyze.

Select All

Deselect All

?

Cancel

OK

Parameters: Correlation

Compute correlation between which pairs of columns?

Compute r for every pair of Y data sets (Correlation matrix)

Compute r for X vs. every Y data set:
X: sensitivity

Compute r between two selected data sets:
X: sensitivity
A: fatty_acids

Assume data are sampled from Gaussian distributions?

Yes. Compute Pearson correlation coefficients

No. Compute nonparametric Spearman correlation

Options

P value: One-tailed Two-tailed

Confidence interval: 95%

Output

Show this many significant digits (for everything except P values): 4

P Value Style: GP: 0.1234 (ns), 0.0332 (*), 0.0021 (**),... N= 6

Graphing

Create a heatmap of the correlation matrix

Make these choices the default for future analyses



The screenshot shows the PAST 3.24 software interface with the following details:

File: correlation.pzfx — Edited

Search: Search bar at the top left.

Data Tables: A section under the menu bar containing "Data 1" and "New Data Table...".

Info: A section under the menu bar containing "Project info 1" and "New Info...".

Results: A section under the menu bar containing "Correlation of Data 1" (highlighted in grey), "New Analysis...", and "Graphs".

Graphs: A section under the menu bar containing "Data 1" (highlighted in grey).

Family: A section under the menu bar containing "Data 1" and "Correlation" (highlighted in grey).

Correlation Results: The main content area displays the following results:

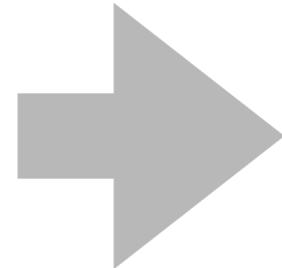
	A	B
1	sensitivity vs. fatty_acids	Title
2	Pearson r	Y
3	r	0.7700
4	95% confidence interval	0.3804 to 0.9275
5	R squared	0.5929
6	P value	
7	P (two-tailed)	0.0021
8	P value summary	**
9	Significant? (alpha = 0.05)	Yes
10	Number of XY Pairs	13
11		
12		
13		
14		

Toolbar: A row of icons at the bottom: back, forward, search, file, data, info, table, and plot.

Spearman's rank correlation is a non-parametric measure of dependence

Spearman's ρ is just Pearson's r computed on the ranks of the x and y values

x	y
17.9	250
18.3	220
18.3	145
18.4	115
18.4	230
20.2	200
20.3	330
21.8	400
21.9	370
22.1	260
23.1	270
24.2	530
24.4	375



x rank	y rank
1.0	6.0
2.5	4.0
2.5	2.0
4.5	1.0
4.5	5.0
6.0	3.0
7.0	9.0
8.0	12.0
9.0	10.0
10.0	7.0
11.0	8.0
12.0	13.0
13.0	11.0

Parameters: Correlation

Compute correlation between which pairs of columns?

Compute r for every pair of Y data sets (Correlation matrix)

Compute r for X vs. every Y data set:
X: sensitivity

Compute r between two selected data sets:
X: sensitivity
A: fatty_acids

Assume data are sampled from Gaussian distributions?

Yes. Compute Pearson correlation coefficients

No. Compute nonparametric Spearman correlation

Options

P value: One-tailed Two-tailed

Confidence interval: 95%

Output

Show this many significant digits (for everything except P values): 4

P Value Style: GP: 0.1234 (ns), 0.0332 (*), 0.0021 (**),... N= 6

Graphing

Create a heatmap of the correlation matrix

Make these choices the default for future analyses

Mutual information is a universal measure of dependence that plays a fundamental role in information theory

Mutual information is symmetric:

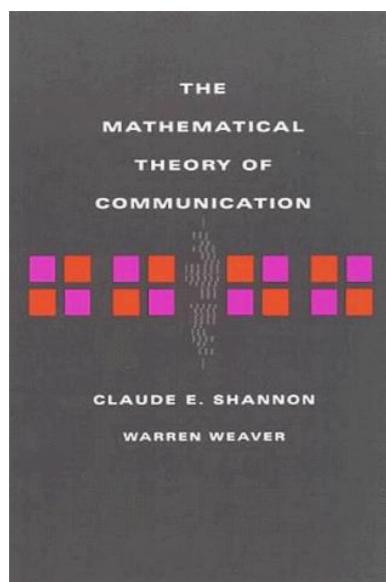
$$I[x; y] = I[y; x]$$



Claude Shannon

Mutual information can range from zero to infinity:

$$0 \leq I[x; y] \leq \infty$$

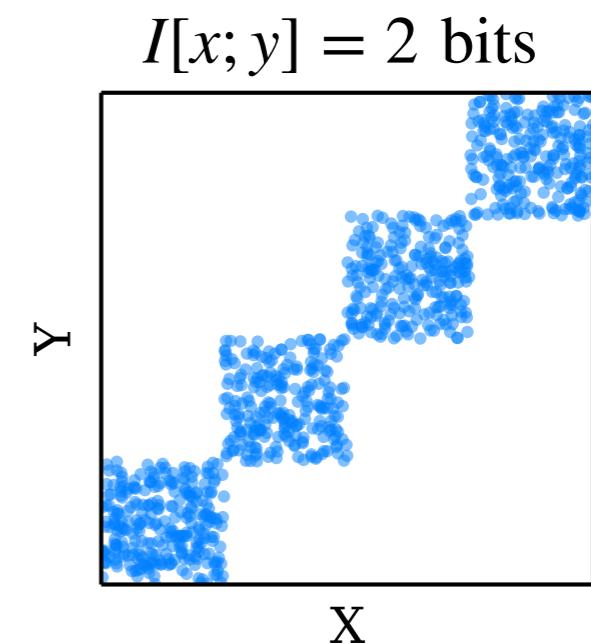
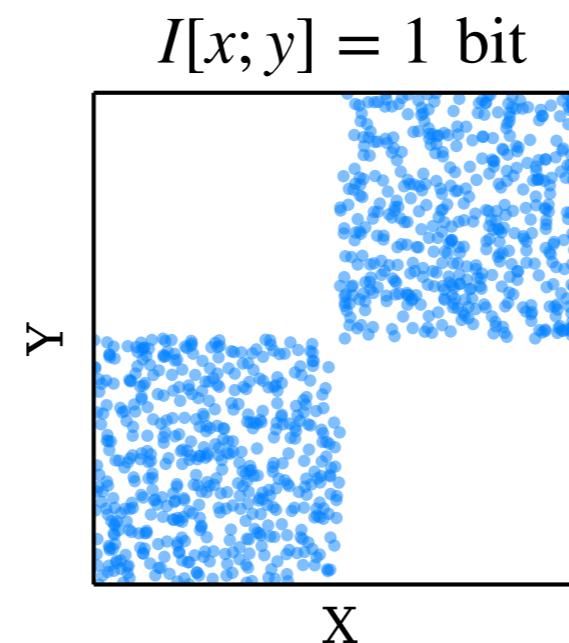
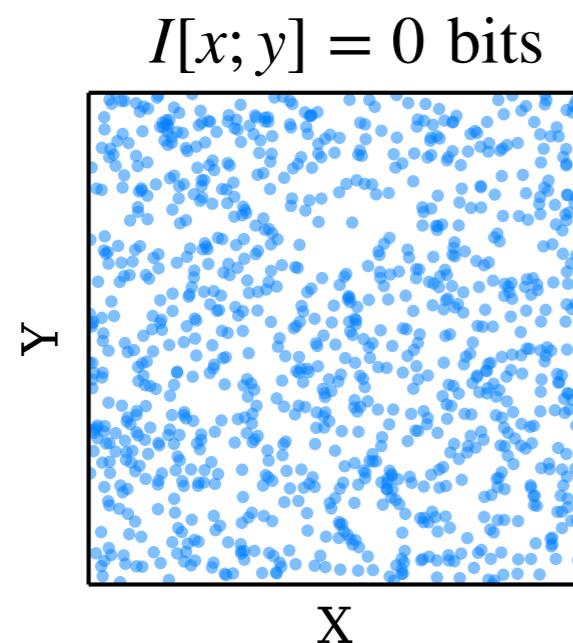


Shannon, 1948

Mutual information is zero only when the two variables are independent:

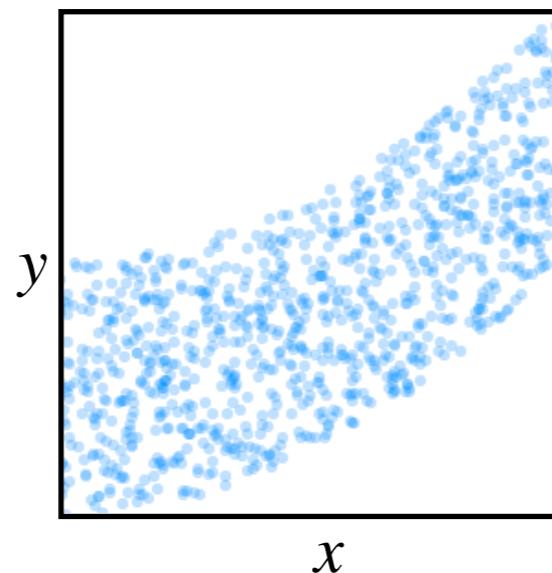
$$I[x; y] = 0 \Leftrightarrow p(x, y) = p(x) p(y)$$

Mutual information is the amount of information in “bits” that knowing one variable tells you about the value of another variable

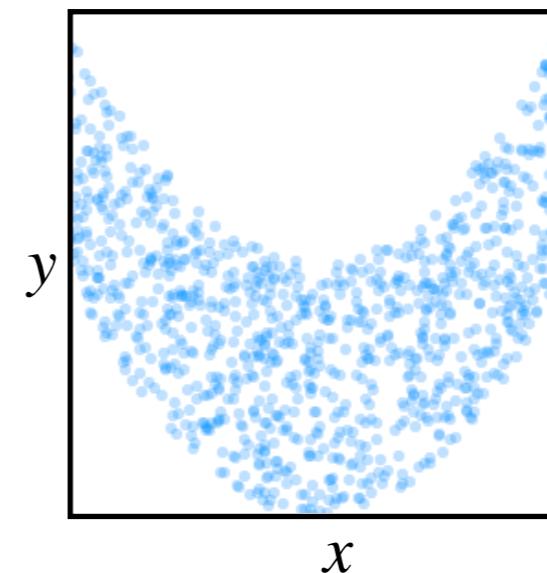


Mutual information, unlike Pearson correlation, quantifies nonlinear and non monotonic relationships in a meaningful way

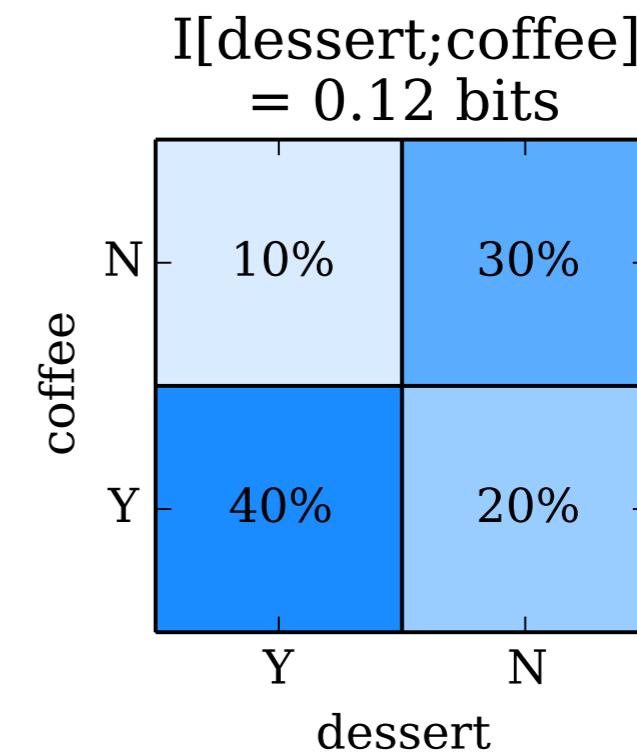
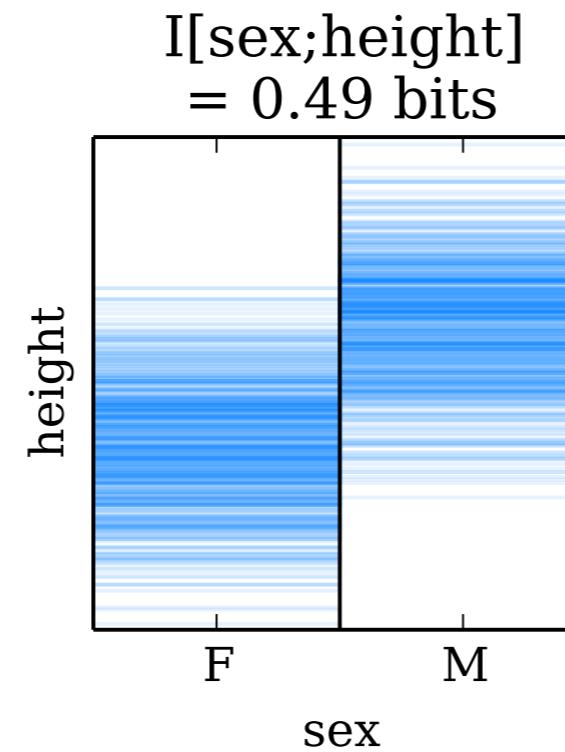
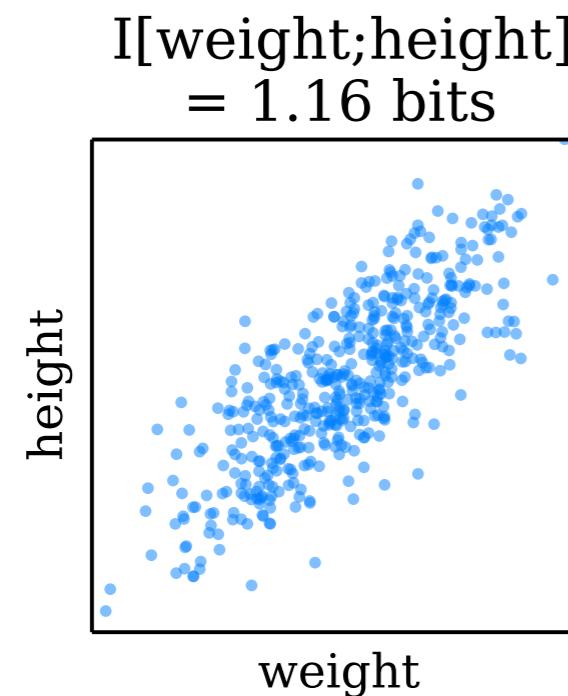
$$I[x; y] = 0.72 \text{ bits}$$
$$r^2 = 0.48$$



$$I[x; y] = 0.72 \text{ bits}$$
$$r^2 = 0$$



Mutual information information can be evaluated between any two types of variables.



Unfortunately, there is no simple plug-in formula for computing mutual information from data.

Mutual information is not commonly used in biological data analysis.

Power analysis

Statistical power is the probability of detecting an effect that actually does exist.

power:

The probability of getting a statistically significant result if the null hypothesis actually is actually false.

power analysis:

The process of assigning and/or computing four quantities (sometimes more) that describe one's experiment:

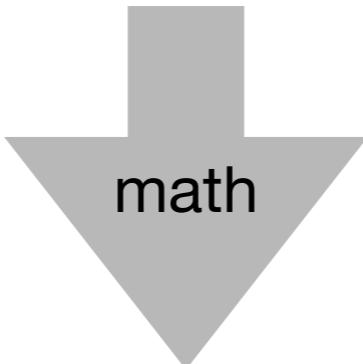
1. The sample size N
2. The false positive probability α (confidence = $1 - \alpha$)
3. The false negative probability β (power = $1 - \beta$)
4. The anticipated effect size

Example: sex ratio

1. Confidence level: $1 - \alpha = 95\%$
2. Number of birth records: $N = 19500$
3. Hypothesized effect size: $|p(\text{boy}) - p(\text{girl})| = 2\%$

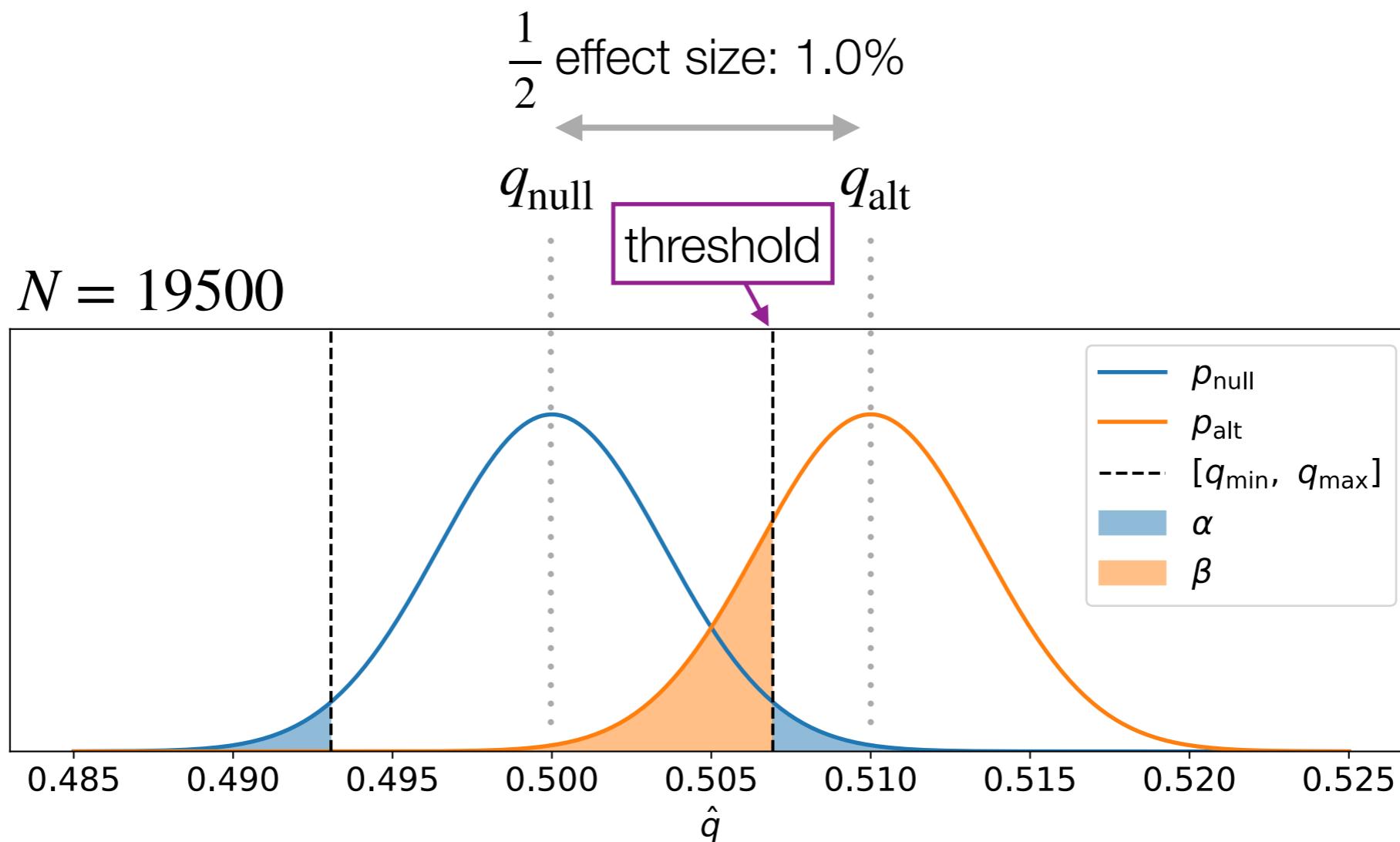
The key parameter is $q = p(\text{boy})$, so we use

$$q_{\text{null}} = 50\%, \quad q_{\text{alt}} = 51\%$$



4. We compute a statistical power of: $1 - \beta = 80\%$

Statistical power example: sex ratio data



False Positive Probability: $\alpha = 0.05$

False Negative Probability: $\beta = 0.20$
(or 80% power)

Power analysis claims come in different forms

There are four relevant parameters: N , α , β , and effect size.

Power analysis involves assuming values for any three parameters and computing the value of the forth

“Controlling the false positive rate at $\alpha = 5\%$, the statistical power at $1 - \beta = 80\%$, and assuming an effect size of 2% , our study will require using $N = 19500$ birth records.”

“Using $N = 19500$ birth records, controlling the false positive rate at $\alpha = 5\%$, and assuming a 2% effect size, our study will have $1 - \beta = 80\%$ power.”

“Controlling the false positive rate at $\alpha = 5\%$, the statistical power at $1 - \beta = 80\%$, and using $N = 19500$ birth records, our study will be sensitive to an effect size of 2% .”

"Using $N = 19500$ birth records, assuming an effect size of 2% , and holding the statistical power to $1 - \beta = 80\%$, our study will be able to hold the false positive rate to $\alpha = 5\%$.”

You will most likely do one of these two things:

You are supposed to do this:

1. Assume a false positive rate of $\alpha = 5\%$ (standard)
2. Assume a power of $1 - \beta = 80\%$ (standard)
3. Assume what you consider to be a biologically significant effect size
4. Compute & use the required sample size N .

You'll actually probably do this:

1. Assume a false positive rate of $\alpha = 5\%$ (standard).
2. Assume a power of $1 - \beta = 80\%$ (standard)
3. Assume a reasonable / affordable sample size N
4. Compute & report the detectable effect size.

If the
detectable
effect size
is too small



Power analysis example: body temperature

1. Assume a false positive rate of $\alpha = 5\%$ (standard).
2. Assume a power of $1 - \beta = 80\%$ (standard)
3. Assume what you consider to be a biologically significant effect size: $\Delta\mu = 0.1\text{ F}$. $\Delta\mu = 0.2\text{ F}$

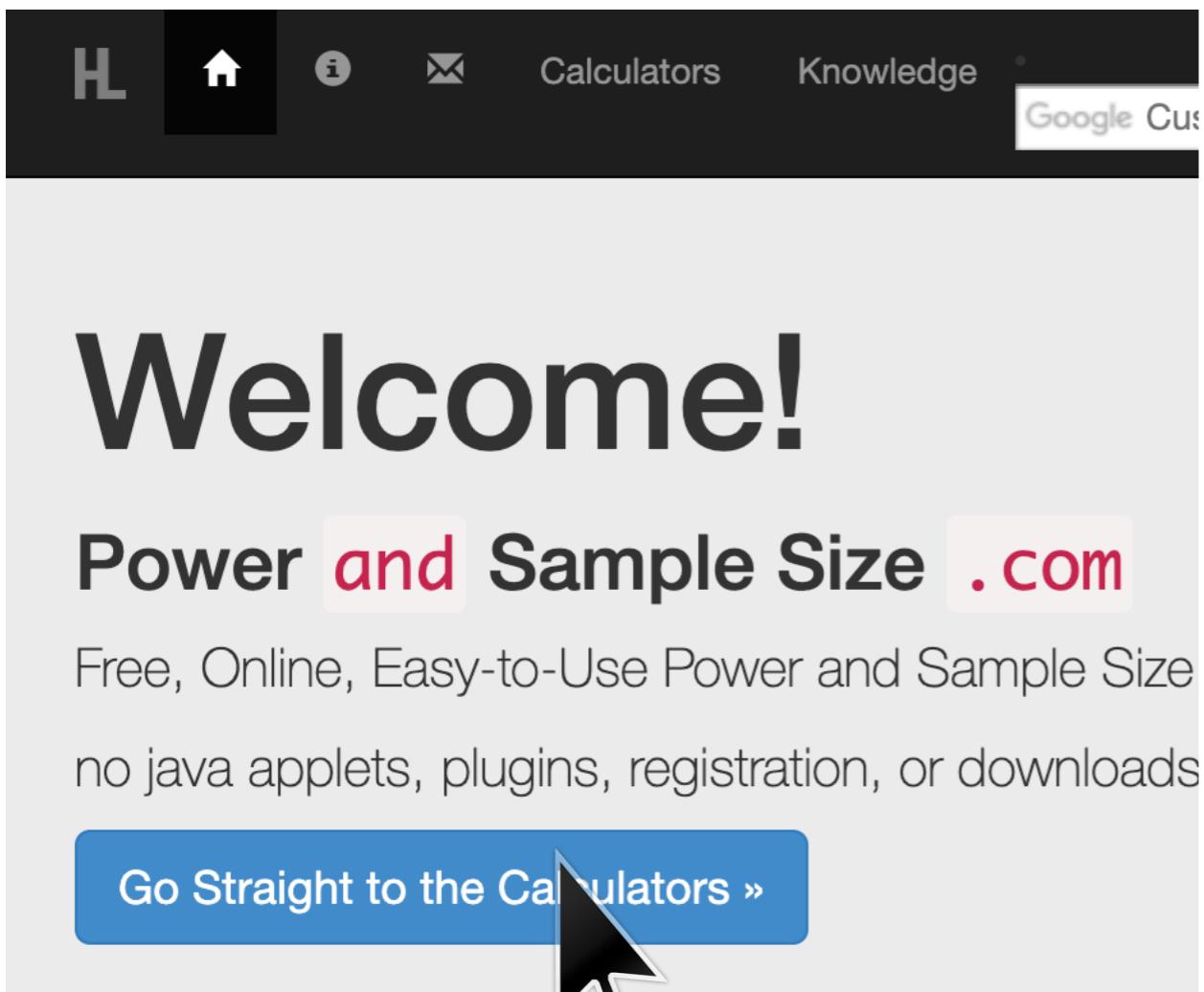
The key parameter is the “normalized effect size”:
$$\frac{\Delta\mu}{\sigma}$$

From preliminary data, we know $\sigma \approx 0.7\text{ F}$

4. Compute the required sample size: $\cancel{N = 1540}$ $N = 386$
Too big! OK.

There are a number of online power analysis calculators

<http://powerandsamplesize.com/>

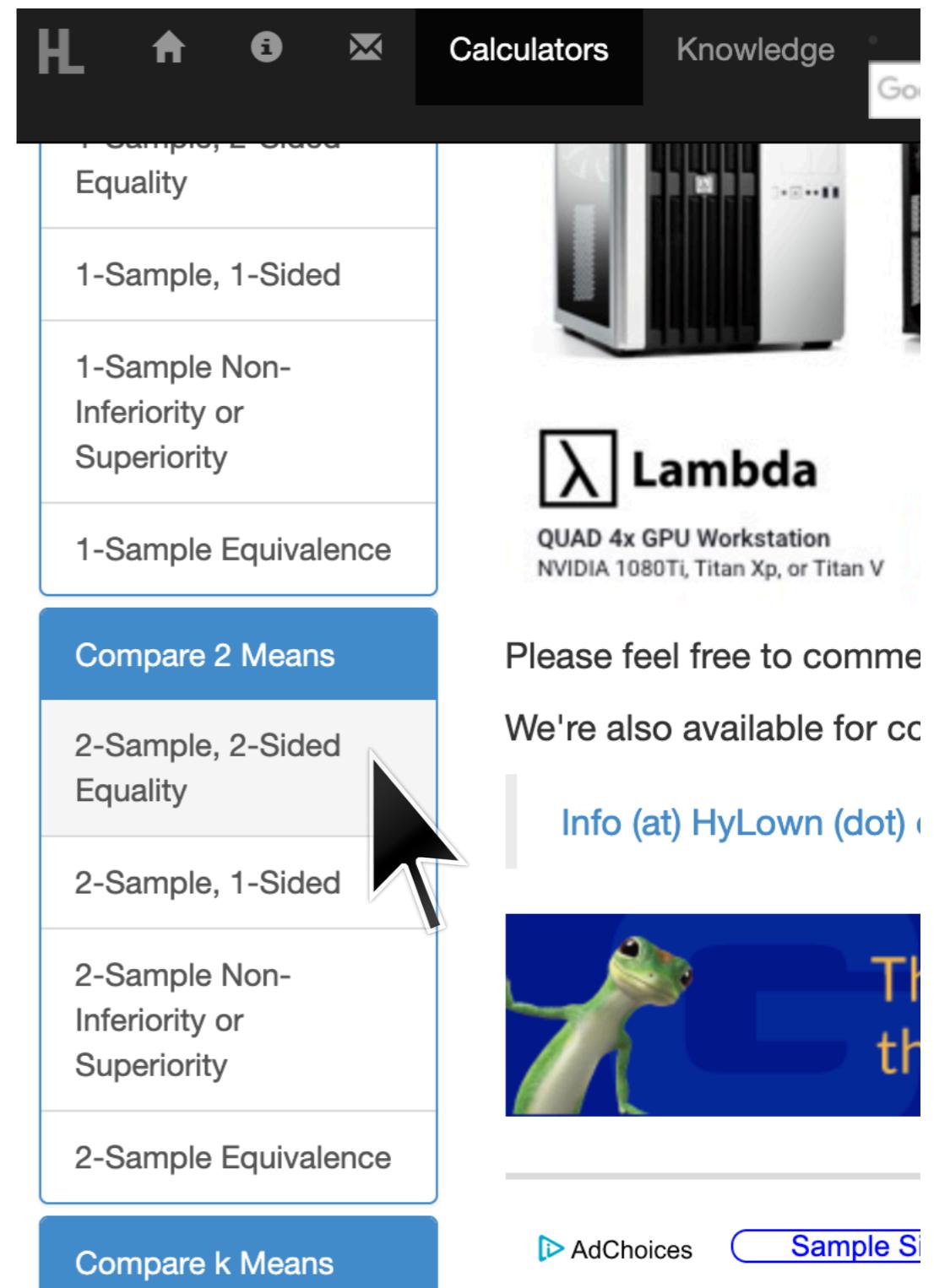


Welcome!

Power and Sample Size .com

Free, Online, Easy-to-Use Power and Sample Size calculators. No Java applets, plugins, registration, or downloads required.

Go Straight to the Calculators »



Equality

1-Sample, 1-Sided

1-Sample Non-Inferiority or Superiority

1-Sample Equivalence

Compare 2 Means

2-Sample, 2-Sided Equality

2-Sample, 1-Sided

2-Sample Non-Inferiority or Superiority

2-Sample Equivalence

Compare k Means

λ Lambda

QUAD 4x GPU Workstation
NVIDIA 1080Ti, Titan Xp, or Titan V

Please feel free to comment or ask questions. We're also available for consulting and training.

Info (at) HyLown (dot) com

Sample Size

Calculate: Sample Size

Sample Size, n_B
192

Power, $1 - \beta$
0.80

Type I error rate, α
5%

98.1 Group 'A' mean, μ_A

98.3 Group 'B' mean, μ_B

0.7 Standard Deviation, σ

1 Sampling Ratio, $\kappa = n_A/n_B$

Calculate

Calculate: Power

Sample Size, n_B 250

Power, $1 - \beta$ 0.892

Type I error rate, α 5%

98.1 Group 'A' mean, μ_A

98.3 Group 'B' mean, μ_B

0.7 Standard Deviation, σ

1 Sampling Ratio, $\kappa = n_A/n_B$

Calculate

98.1

98.3

0.7

1

Group 'A' mean, μ_A

Group 'B' mean, μ_B

Standard Deviation, σ

Sampling Ratio, $\kappa = n_A/n_B$

Calculate

Multiple hypothesis testing

The problem of multiple subgroups



The family-wise error rate increases rapidly with the number of tests performed

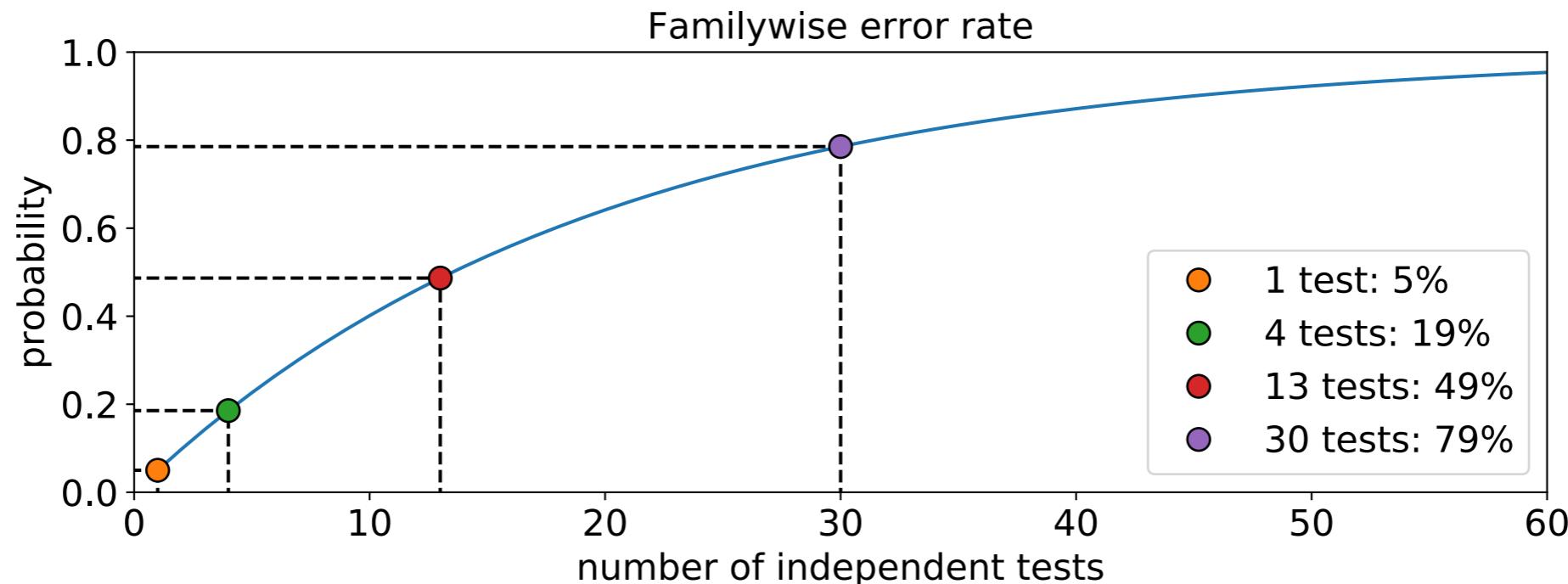
Scenario:

we perform null hypothesis tests on K independent datasets, for each of which the null hypothesis is true.

Family-wise error rate:

the probability of at least one false positive (FP)

$$p(\text{FP} \geq 1 \mid \text{null hypothesis}) = 1 - \text{confidence}^K$$



Summary of multiple hypothesis correction techniques

Approach	What you control	Expression
No correction	α : if all null hypotheses are true, the <u>fraction of tests</u> that produce a significant result	$\alpha = \frac{\text{FP}}{\text{FP} + \text{TN}}$
Bonferroni / Dunn-Sidak	α : if all null hypotheses are true, the <u>chance of obtaining one or more</u> significant results	$\alpha = p(\#\text{FP} > 0)$
False discovery rate (FDR)	Q : the fraction of all discoveries for which the null hypothesis is actually true	$Q = \frac{\text{FP}}{\text{FP} + \text{TP}}$

Simple ways to counteract the multiple hypothesis problem

Bonferroni correction:

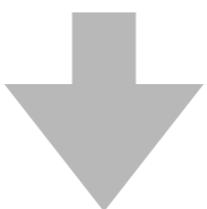
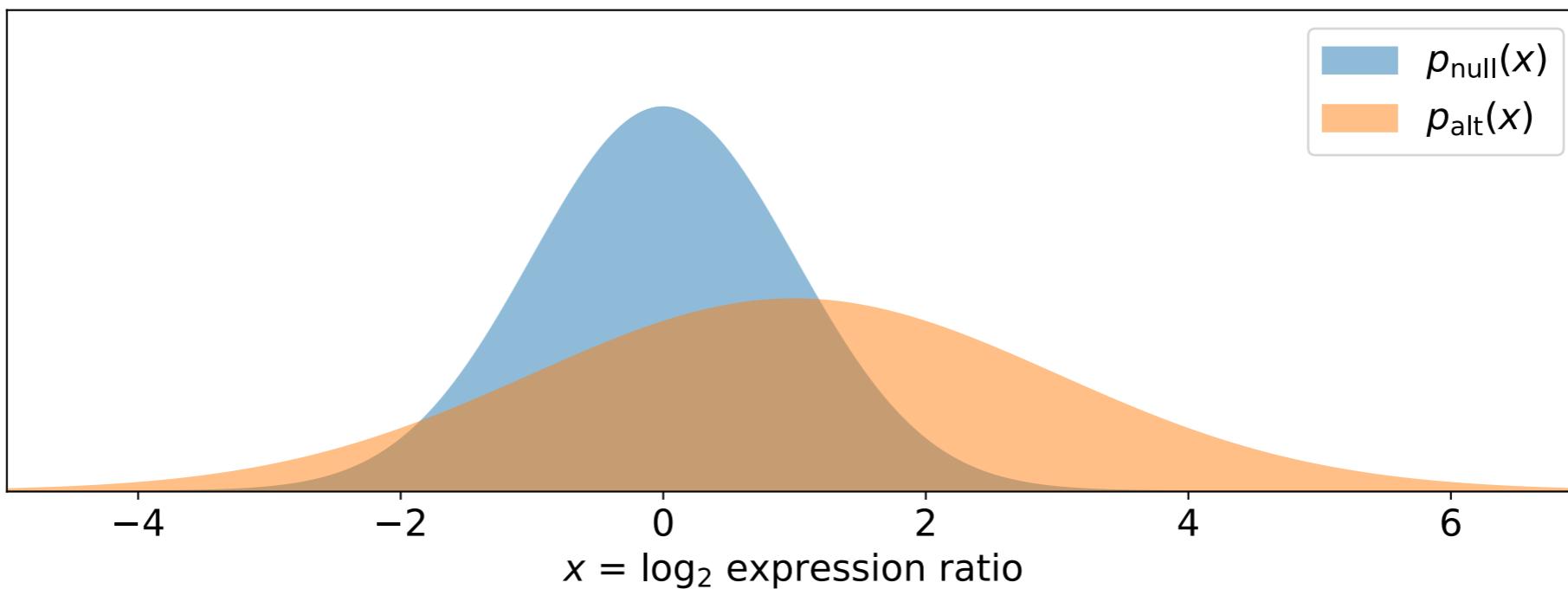
$$\alpha_{\text{Bonferroni}} = \frac{\alpha}{K}$$

Dunn-Sidak correction:

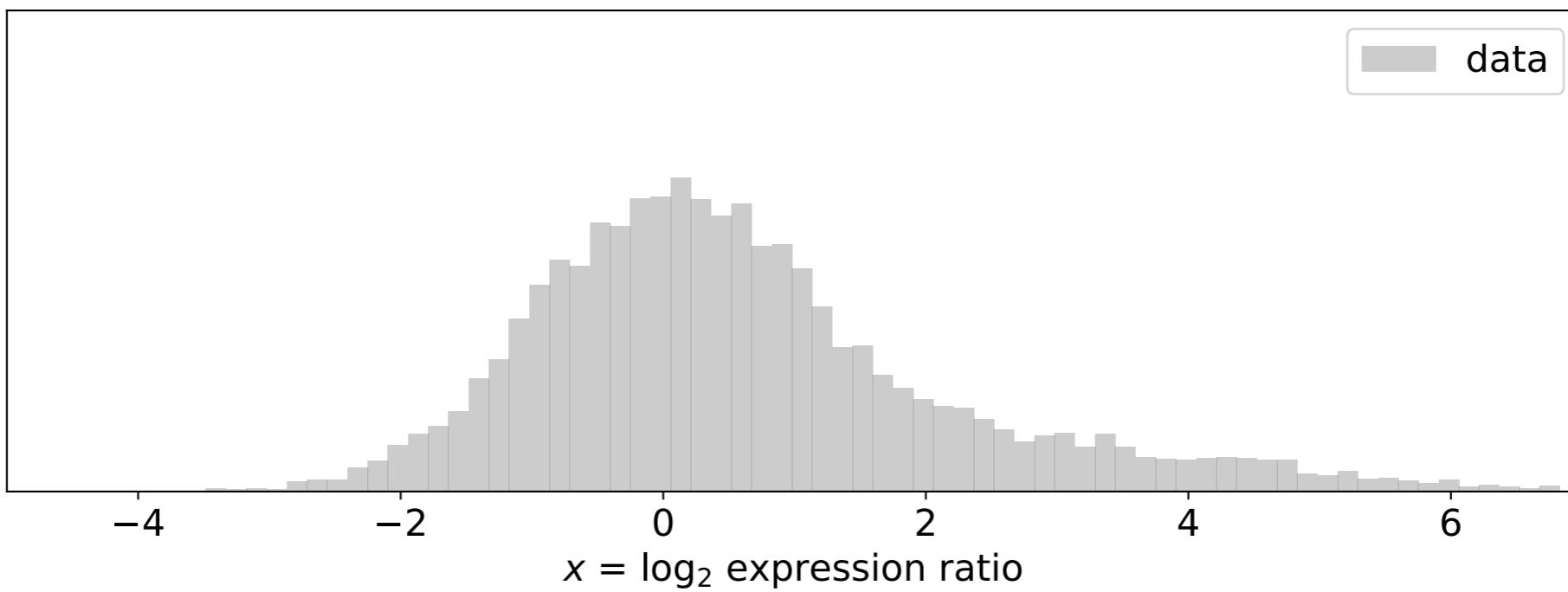
$$\alpha_{DS} = 1 - (1 - \alpha)^{1/K}$$

Dunn-Sidak is the exact solution; Bonferroni is an approximation

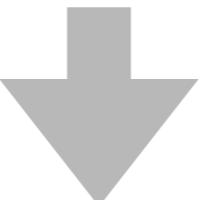
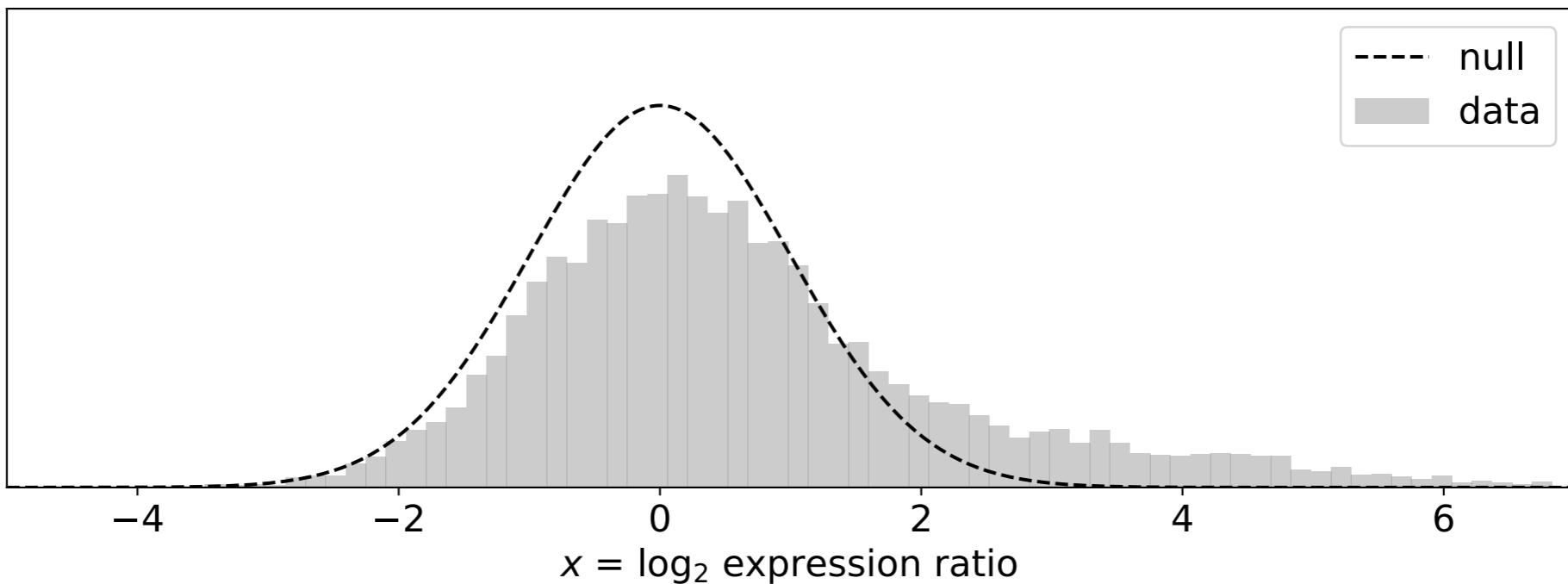
Example: differential expression (simulation)



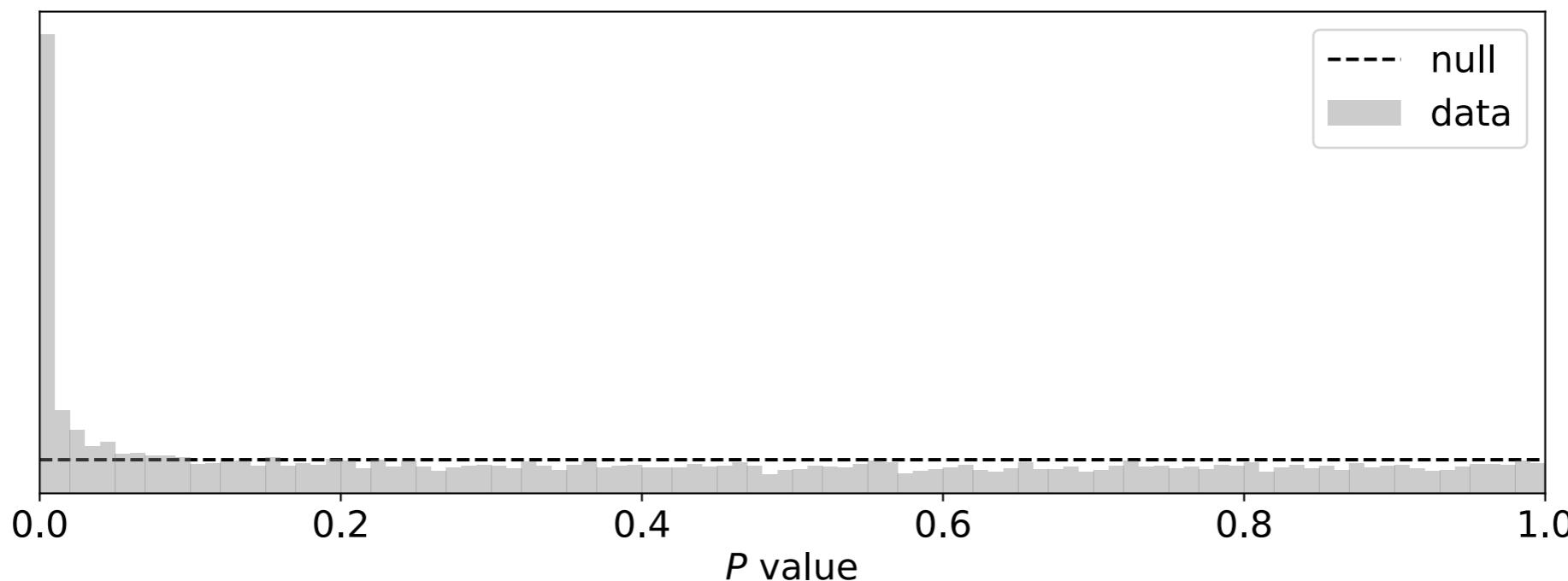
7,000 x s from $p_{\text{null}}(x)$
+ 3,000 x s from $p_{\text{alt}}(x)$



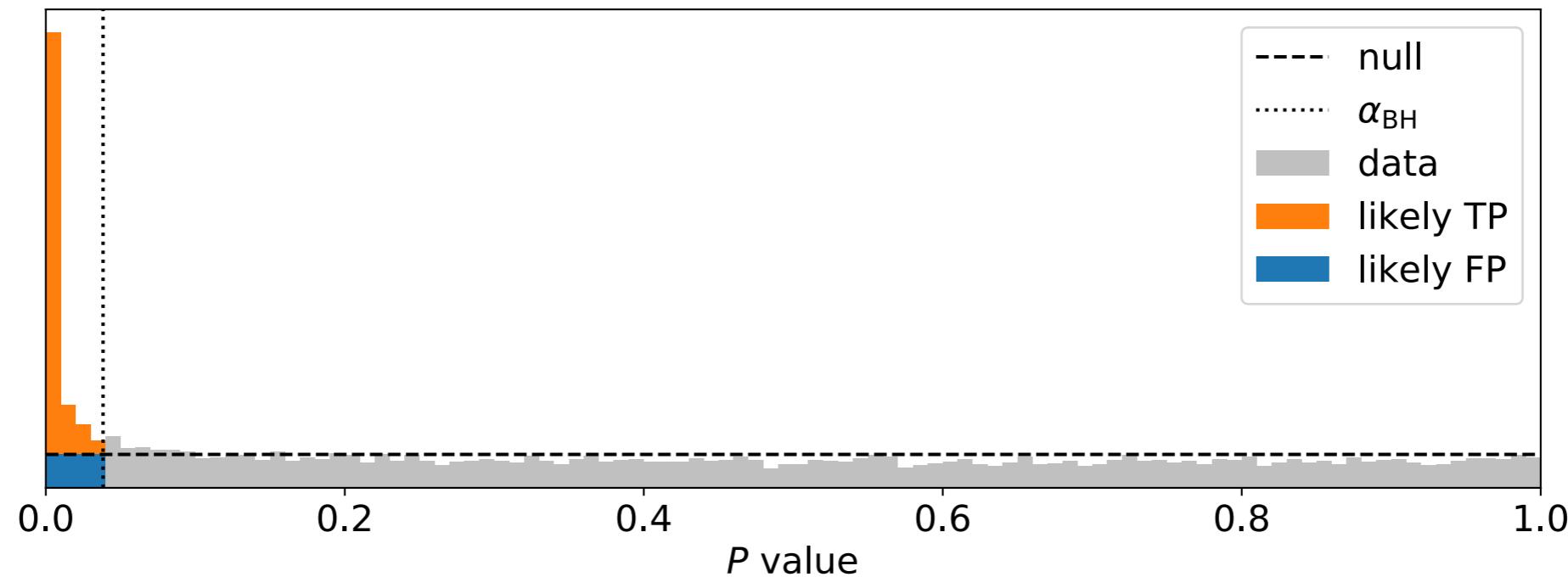
First, convert data to p-values



use knowledge of $p_{\text{null}}(x)$ to
compute a p-value for each datapoint



Benjamini–Hochberg procedure



Choose α_{BH} such to match the target False Discovery Rate (10% here):

$$FDR = Q = \frac{FP}{TP + FP} = \frac{\text{█}}{\text{█} + \text{█}}$$

Declare all P-values below α_{BH} as “discoveries”.

Multiple comparisons are ubiquitous and insidious

“Most scientists are oblivious to the problems of multiplicities. Yet they are everywhere. In one or more of its forms, multiplicities are present in every statistical application. They may be out in the open or hidden. And even if they are out in the open, recognizing them is but the first step in a difficult process of inference. Problems of multiplicities are the most difficult that we statisticians face. They threaten the validity of every statistical conclusion.”

Multiple comparisons arise in many many contexts

multiple subgroups:

You perform tests on multiple subgroups of your data.

multiple ways to dichotomize:

You do pairwise comparisons between different combinations of subgroups.

multiple sample sizes:

You keep collecting data until you find $P < 0.05$.

DO NOT DO THIS.

multiple ways to preprocess the data:

You analyze data preprocessed in multiple different ways.

multiple statistical tests:

You use different statistical tests on the same data before finding $P < 0.05$.

Multiple comparisons arise in many, many contexts

multiple ways to select relevant variables:

You try to model your data using different subsets of possible variables.

multiple ways to analyze your data (“garden of forking paths”):

You try lots of qualitatively different analysis strategies.

outcome switching:

You change the quantity you care about after you've looked at the data.

multiple geographic areas:

E.g., you investigate a “cancer cluster” you hear about in the news.

Correcting for multiple comparisons is not always needed

Scenario 1:

If readers can be reasonably expected to account for multiple comparisons on their own.

Scenario 2:

Before looking at the data, you have clearly defined one outcome as primary and others as secondary.

Scenario 3:

You make only a few planned comparisons and your P-values are not marginal.

Scenario 4:

A large fraction the tests you perform are significant.

Practical advice of avoiding multiple hypothesis pitfalls

Raise your standards: use $\alpha = 0.01$, not $\alpha = 0.05$.

Separate exploratory data analysis from confirmatory data analysis.

Distinguish critical p-values from ancillary p-values.

Don't spend too much time analyzing a small dataset.

When generating small expensive datasets (e.g. mice), blind your experiments as best you can, and plan your analysis ahead of time

When in doubt, double-check your hypothesis with new data

Don't worry about informal multiple hypothesis testing when $P < 10^{-4}$.

10:00a - 12:02p Just slightly over. Good length I think.