Introduction to Statistics

Maria Beatriz Walter Costa
Viral Ecology and Omics (VEO)

Foto: Anne Günther

FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

Outline

- **What is statistics?**
- **Types of variables**
- **Distributions**
- **Choosing tests**
- **Parametric test**
  t-test
- **Non-parametric test**
  Mann-Whitney U test
- **p-values**
- **Linear regression**
- **Correlation**
  Pearson, Spearman
- **Set Enrichment Analysis**
- **How to continue learning**
- **Slides at:**
  https://github.com/waltercostamb/statistics

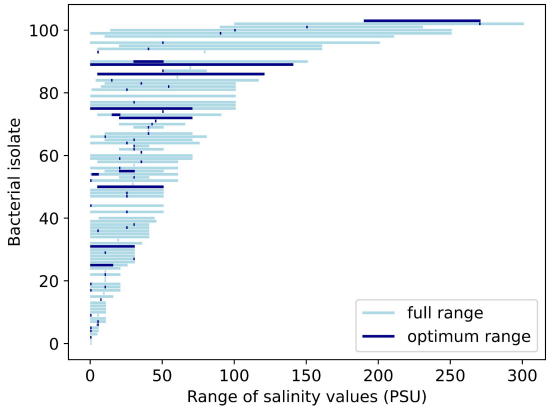
Statistics: analysis and interpretation of data

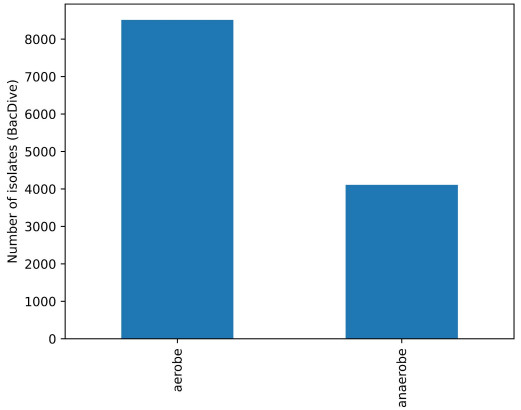
- **Descriptive statistics**
  Mean, median, standard deviation
- **Inferential statistics**
  Conclusions that extend beyond data you have, test ideas/hypothesis
- **Frequentist/classical statistics**
  Testing null hypothesis
- **Bayesian statistics**
  No rejection of hypothesis, rather degrees of belief

- **Shannon diversity index**
  Measures diversity of species in community
- **Kruskal-Wallis rank sum test**
  Tests if samples originate from same distribution
- **Linear regression**
  Model that estimates linear relationship between response and explanatory variables

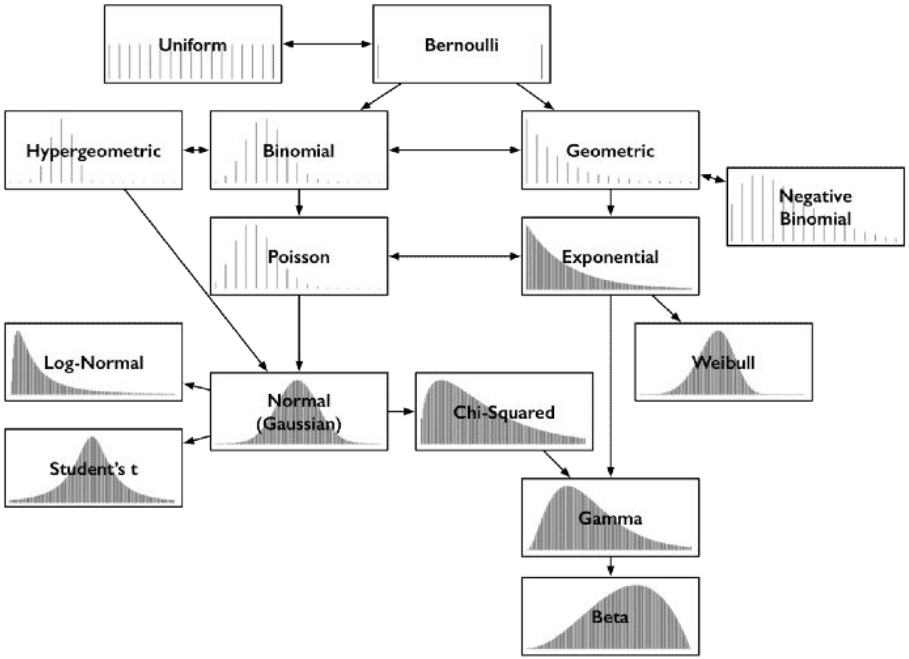
Types of variables

- **Qualitative/categorical**
  Oxygen tolerance; biomes
- **Quantitative**
  - **Discrete**
    Counts of isolates per class
  - **Continuous**
    Salinity tolerances





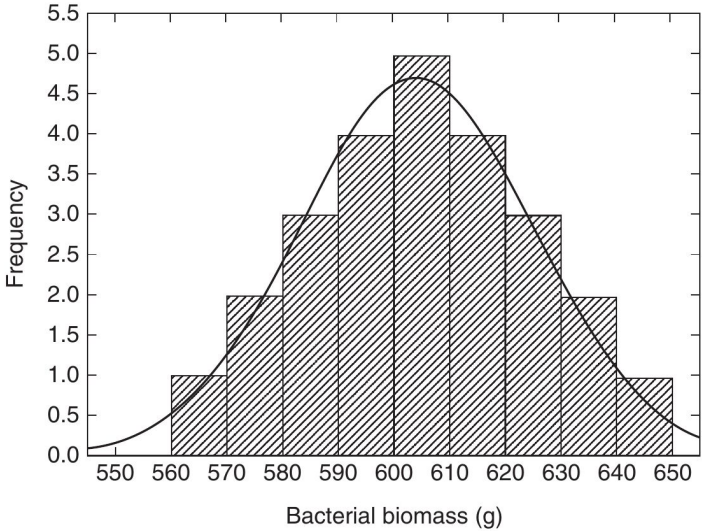
Distributions: how data spreads across a range of values



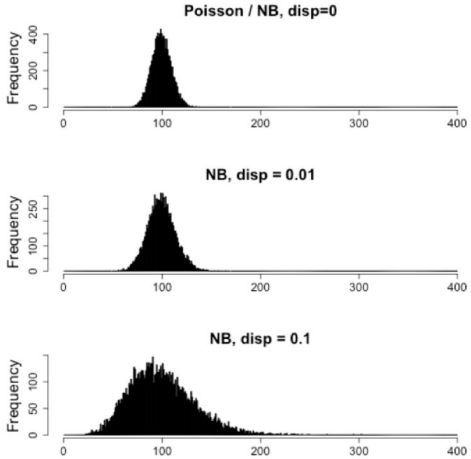
https://www.ztable.net/probability-distributions-types/

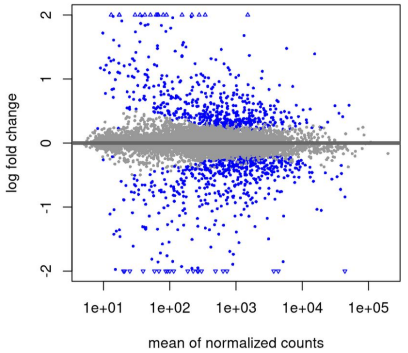
Distributions

- **Normal**
  Mean, st dev; shown by Kolmogorov-Smirnov test



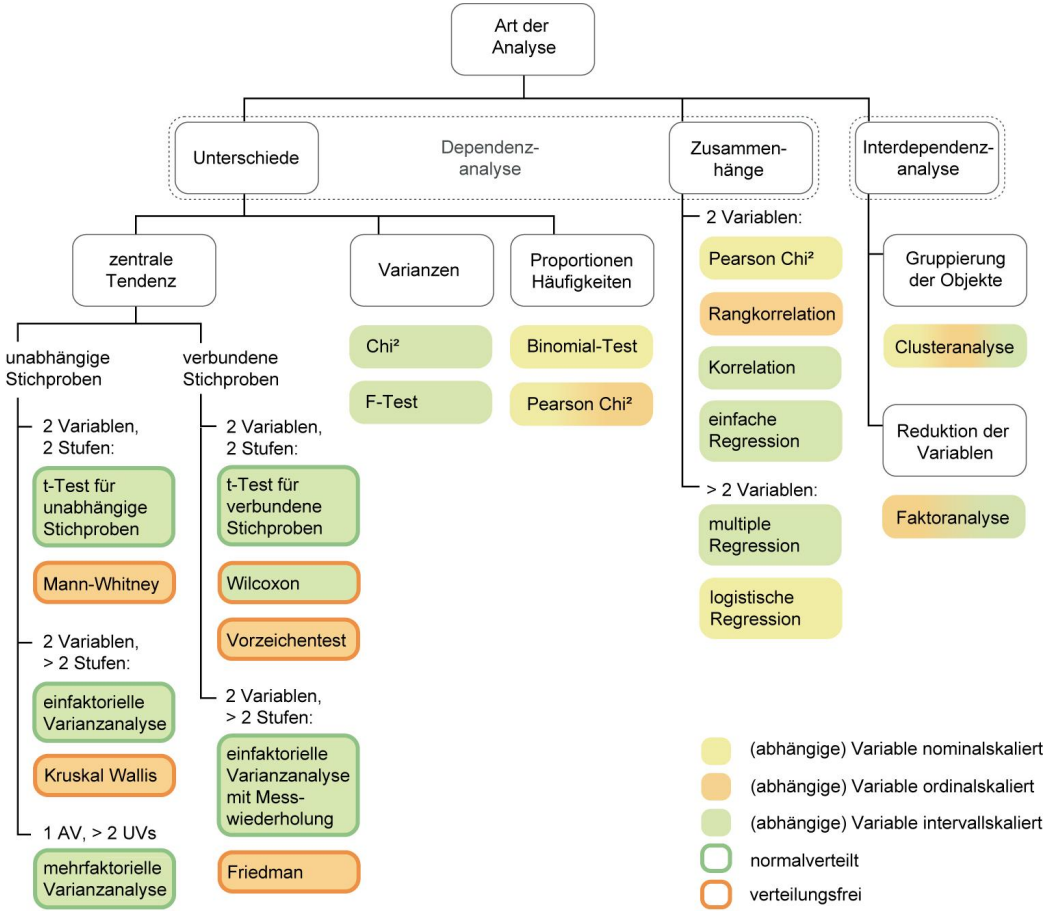
- **Negative binomial**
  - expression (RNAseq);
  - appropriate for counts
  - mean < variance
  - DESeq2 for finding fold changes





Answering questions and choosing tests

Form of the Data	Possible Statistical Procedures: Parametric	Possible Statistical Procedures: Nonparametric
A single observation x	Is x a member of a specific population (**2.5**)?	—
A sample of x values	Construct frequency distribution, calculate x^*, SD, SEM, CI (**2.4**). Is X normally distributed? (**1**)	Mode, median, 95th percentile (**4**)
Two independent samples (x_1, x_2)	Unpaired t test (**3.4**)	Mann–Whitney U test (**4.7**)
Two paired samples ($x_1 - x_2$)	Paired t test (**3.6**)	Wilcoxon signed-rank test (**4.8**)



https://onlinelibrary.wiley.com/doi/epdf/10.1002/9780470905173.app1
https://www.methodenberatung.uzh.ch/de/datenanalyse_spss.html

Parametric test: t-test

- **Inferential test for hypothesis testing**
  - Null-hypothesis: means are the same
  - Balance of signal/noise
  - Input: mean, std dev, sample size
- **Are observed differences meaningful or random?**
  - Compare calculated value against a table of standard ones to get p-value
- **Used when data follows normal distribution, similar variances**
- **Suited for small sample sizes (> 40)**

- **student's t-distribution: gets name from William Sealy Gosset, worked for Guiness Brewery in Dublin**
- **interested in monitoring quality of stout**

https://www.investopedia.com/terms/t/t-test.asp,
YouTube Channel "Bozeman Science", Student's t-test,
Images: wikipedia and www.guiness.com

Non-parametric test: Mann-Whitney U test

- **Counterpart to t-test**
- **Inferential test for hypothesis testing**
  - Null-hypothesis: rank sums are same
  - Input: ranks (sum), sample size
- **Are observed differences meaningful or random?**
  - Compare calculated values against a table of expected ones to get p-value
- **Used when data <u>does not</u> follow normal distribution**

Gender	Reaction time	Rang
female	34	2
female	36	4
female	41	7
female	43	9
female	44	10
female	37	5
male	45	11
male	33	1
male	35	3
male	39	6
male	42	8

Calculation of the rank sums

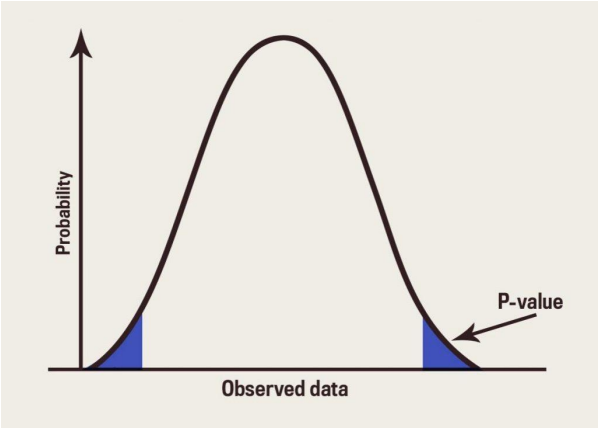
$T_1 = 2 + 4 + 7 + 9 + 10 + 5 = 37$

$T_2 = 11 + 1 + 3 + 6 + 8 = 29$

p-value

- **Probability: [0, 1] indicating confidence of result**
  finding observed or more extreme values for same number of samples (if null hyp. true)
- **Helps decide if we should reject null hypothesis**
- **How close to zero do we have to be to be sufficiently confident?**
- **Rule of thumb: <0.05 to reject null hypothesis**
  by random chance, we could get similar results 5% of times (false-positives!)

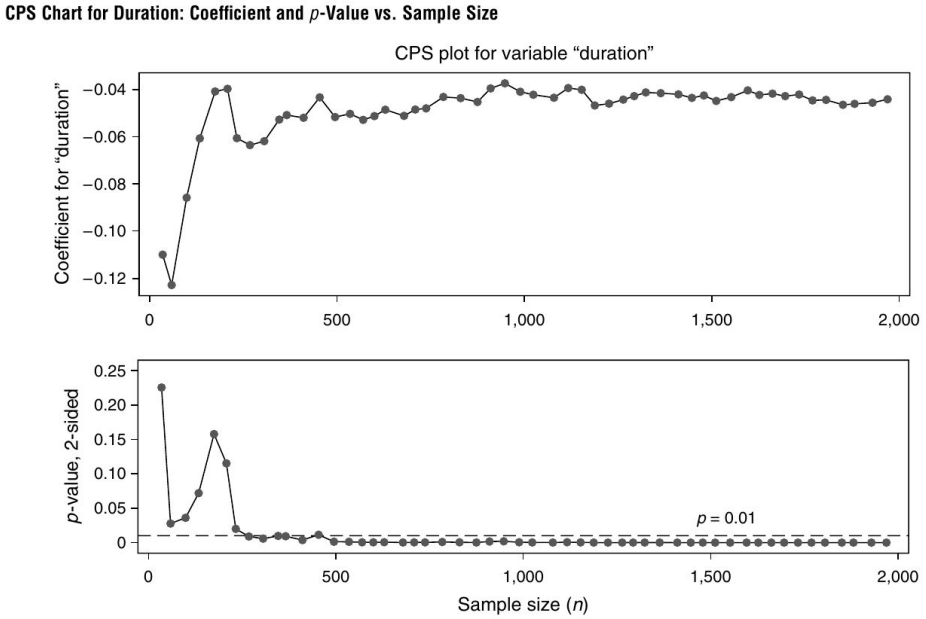
- **p-value does not tell if alternative hypothesis is true**
- **P-value cannot inform on how much difference there is**
  For this, you need to calculate the effect size



YouTube Channel “Bozeman Science”, p-values: What they are and how to interpret them,
Image: https://www.leansixsigmadefinition.com/glossary/p-value/

Effect size is important!

- **Traditional tests are appropriate for small sample sizes**
- **In large samples, even tiny differences will lead to small p-values**
- **What to do?**
  - Present effect sizes
  - Present confidence intervals
  - Use visualization
  - CPS (coefficient, p-value, sample size) charts: sub-sample for different sample sizes and plot
  - Split data into training and "holdout" for validation

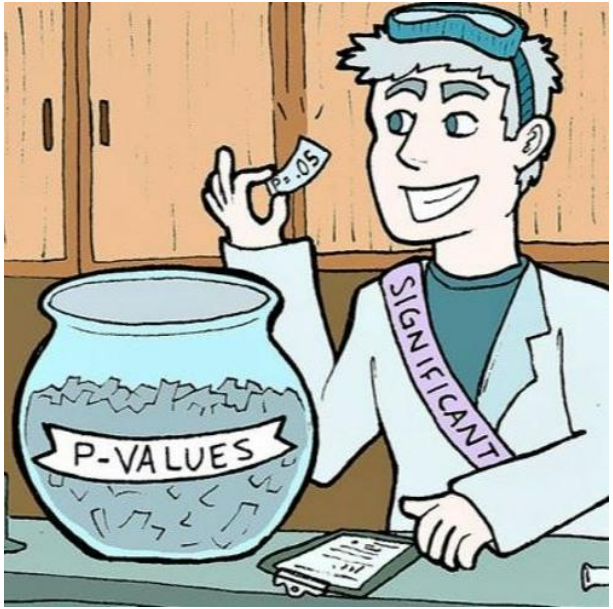


Lin 2013: "Too big to fail: large samples and the p-value problem";
Example shows camera sales on ebay and effect of auction's duration
https://elifesciences.org/articles/48175

Replication crisis and p-hacking

- **What are researchers trying to find?** Causality
- P-values help, but aren't proof
    - Correlation does not imply causation
- **Reproducibility** helps further, but is unfortunately discouraged by how publication works

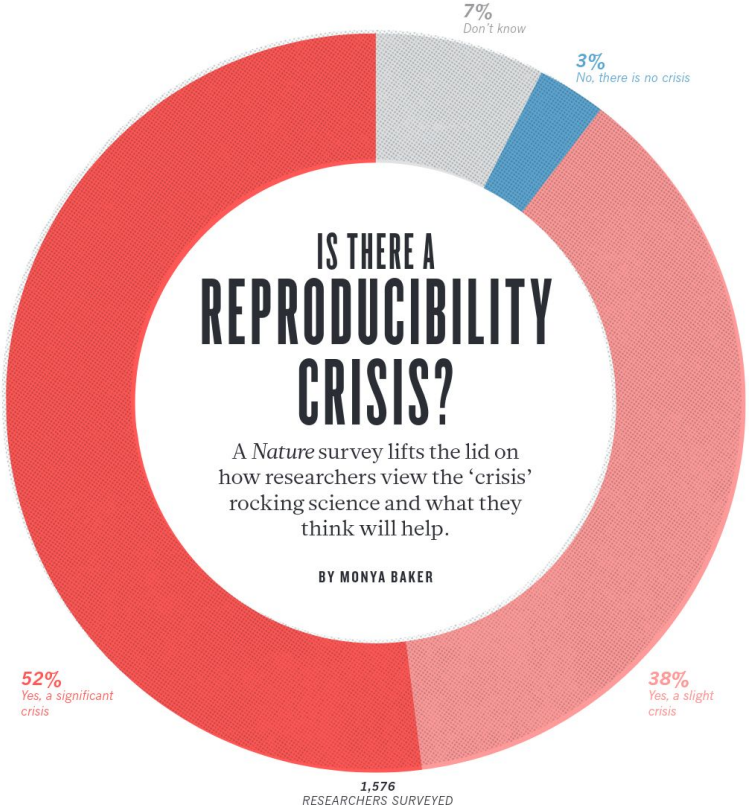
- **P-hacking**: manipulation of data that artificially produces significant results
    - Major player in replication crisis
    - Intention or unintentional
        - Failing to do multiple test correction
- **What to do??**
    - Be transparent
    - ask an expert



https://statisticsbyjim.com/hypothesis-testing/p-hacking/;
Open science collaboration, 2015: https://www.science.org/doi/10.1126/science.aac4716;
Image from: https://soundcloud.com/user-605679240/p-hacking-its-a-big-problem

How many scientific studies are reproducible?

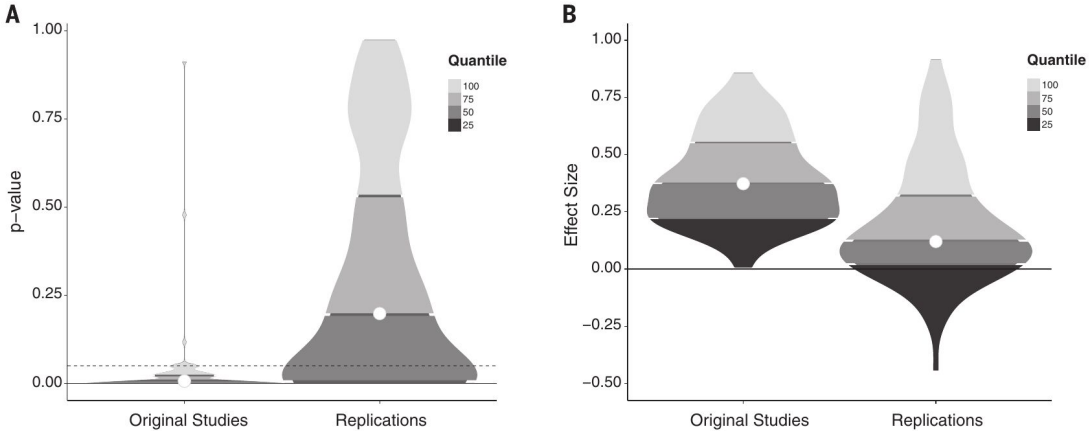
- **“Estimating the reproducibility of psychological science”,** Open science collaboration (300 authors), Science, 2015
- Replicated 100 experimental studies
- Compared: p-values, effect sizes, subjective assessments



Open science collaboration, 2015: https://www.science.org/doi/10.1126/science.aac4716;
Baker 2016 “1,500 scientists lift the lid on reproducibility”, Nature

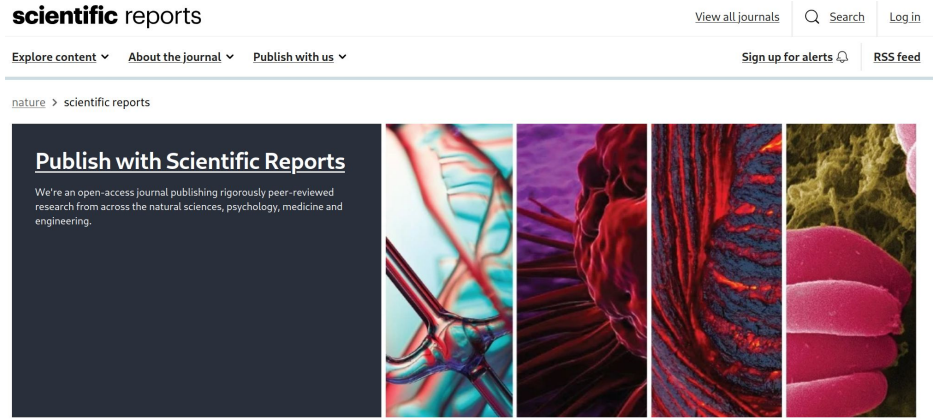
How many scientific studies are reproducible?

- **“Estimating the reproducibility of psychological science”,** Open science collaboration (300 authors), Science, 2015
- Replicated 100 experimental studies
- Compared: p-values, effect sizes, subjective assessments
- 36% of original studies were reproducible
- Effect size of reproduced studies had half the magnitude of original ones



Open science collaboration, “Estimating the reproducibility of psychological science”, Science, 2015: https://www.science.org/doi/10.1126/science.aac4716

Guidelines of Scientific Reports



Statistical guidelines

If your paper contains statistical testing, it should state the name of the statistical test, the n value for each statistical analysis, the comparisons of interest, a justification for the use of that test (including, for example, a discussion of the normality of the data when the test is appropriate only for normal data), the alpha level for all tests, whether the tests were one-tailed or two-tailed, and the actual P value for each test (not merely "significant" or "P < 0.05"). Please make it clear what statistical test was used to generate every P value. Use of the word "significant" should always be accompanied by a P value; otherwise, use "substantial," "considerable," etc.

Data sets should be summarised with descriptive statistics, which should include the n value for each data set, a clearly labelled measure of centre (such as the mean or the median), and a clearly labelled measure of variability (such as standard deviation or range).

Ranges are more appropriate than standard deviations or standard errors for small data sets. Graphs should include clearly labelled error bars. You must state whether a number that follows the ± sign is a standard error (s.e.m.) or a standard deviation (s.d.).

You must justify the use of a particular test and explain whether the data conforms to the assumptions of the tests. Three errors are particularly common:

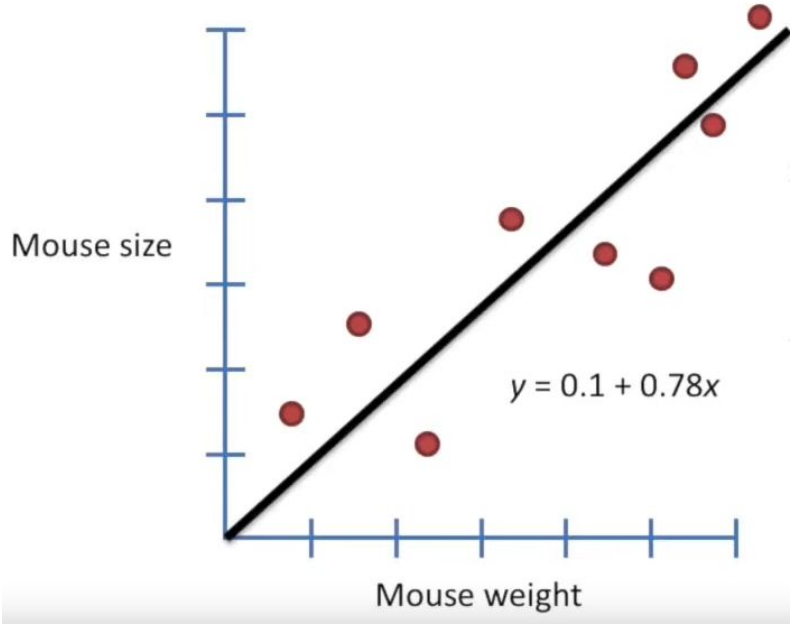
- Multiple comparisons: when making multiple statistical comparisons on a single data set, you should explain how you adjusted the alpha level to avoid an inflated Type I error rate, or you should select statistical tests appropriate for multiple groups (such as ANOVA rather than a series of t-tests).
- Normal distribution: many statistical tests require that the data be approximately normally distributed; when using these tests, you should explain how you tested your data for normality. If the data does not meet the assumptions of the test, you should use a non-parametric alternative instead.
- Small sample size: when the sample size is small (less than about 10), you should use tests appropriate to small samples or justify the use of large-sample tests.

https://www.nature.com/srep/author-instructions/submission-guidelines

Linear regression

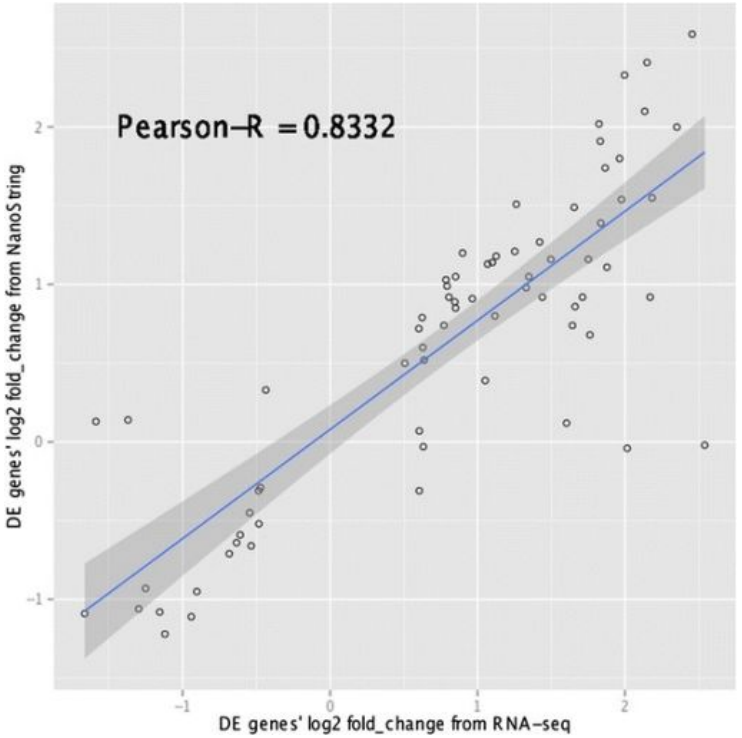
- **Models relationship between independent (x) and dependent (y) variables**

- **f(x) = a + bx**
  - **a = intercept in y axis**
  - **b = slope**
- **Use least squares to fit straight line to data**
- **Calculate R squared**
  - Distance between prediction and actual values (residuals)
  - How much weight explains size
- **Calculate p-value for R squared**
  - prob that random dots result in similarly stronger relationship or higher



Parametric correlation: Pearson

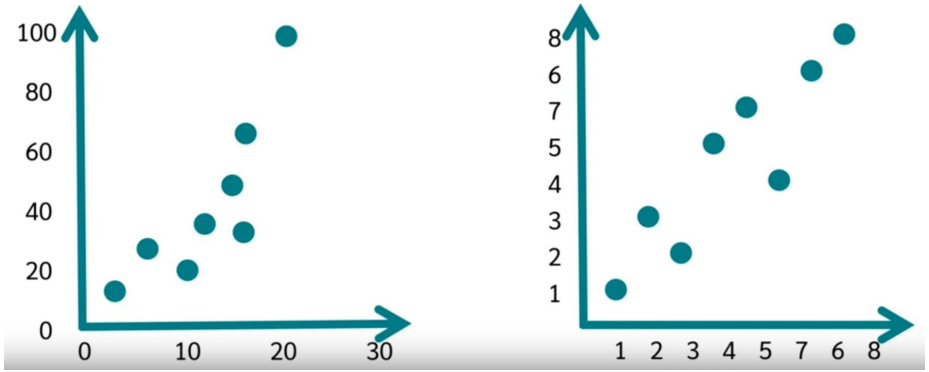
- **Models relationships between two numeric variables**
- **Inferential model, can be used to predict new data based on trend**
    - Measures strength of relationship (strong, moderate, weak)
    - [-1, 0, 1]
- **P-values:** prob that random dots result in similarly stronger relationship or higher; confidence on model
- **Assumes linear relationships**
    - check with visualization
- **<u>NO CAUSAL</u> assumptions**
    - A 3rd factor could cause the variables



YouTube Channel “StatQuest”, Pearson Correlation, Clearly Explained; Zhou et al 2015

Non-parametric correlation: Spearman

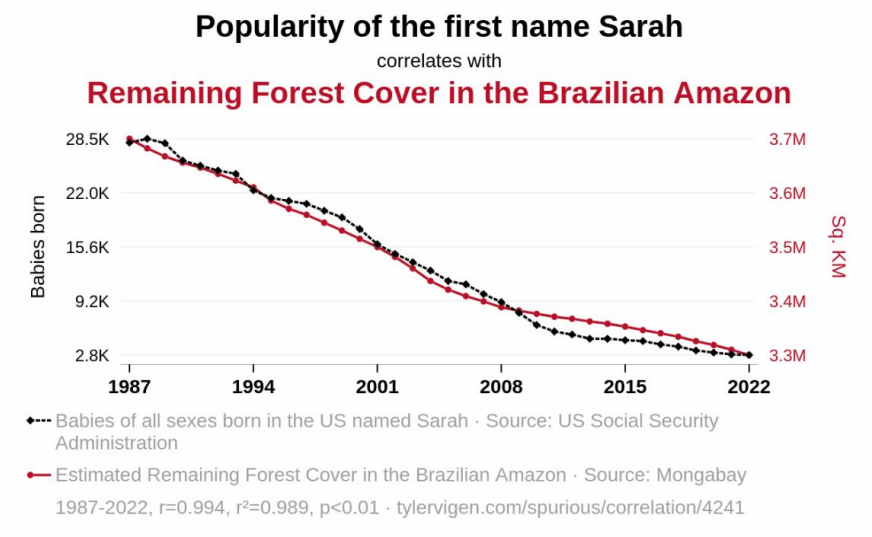
- **Non-param counterpart of Pearson**
- **Applies Pearson on ranks instead of raw values**
- **Inferential model**
  - strength of relationship (strong, moderate, weak)
  - [-1, 0, 1]
- **P-values:** prob that random dots result in similarly stronger relationship or higher
- **<u>NO CAUSAL</u> assumptions**

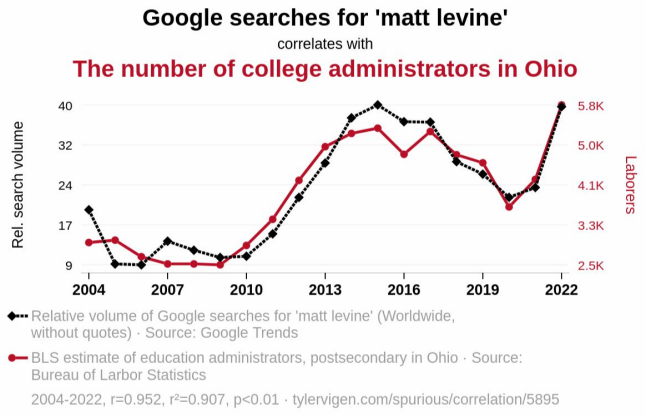


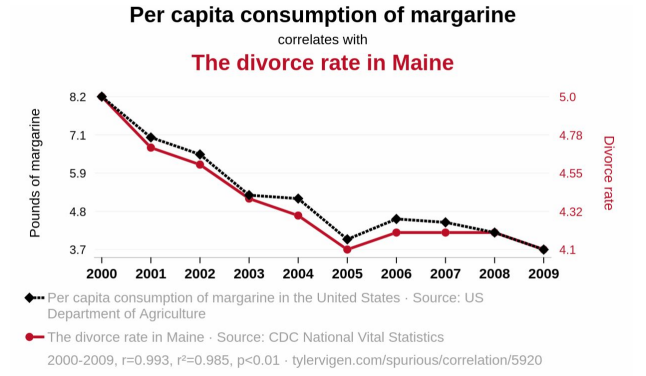
YouTube Channel “StatQuest”, Pearson Correlation, Clearly Explained; Zhou et al 2015

Spurious correlations

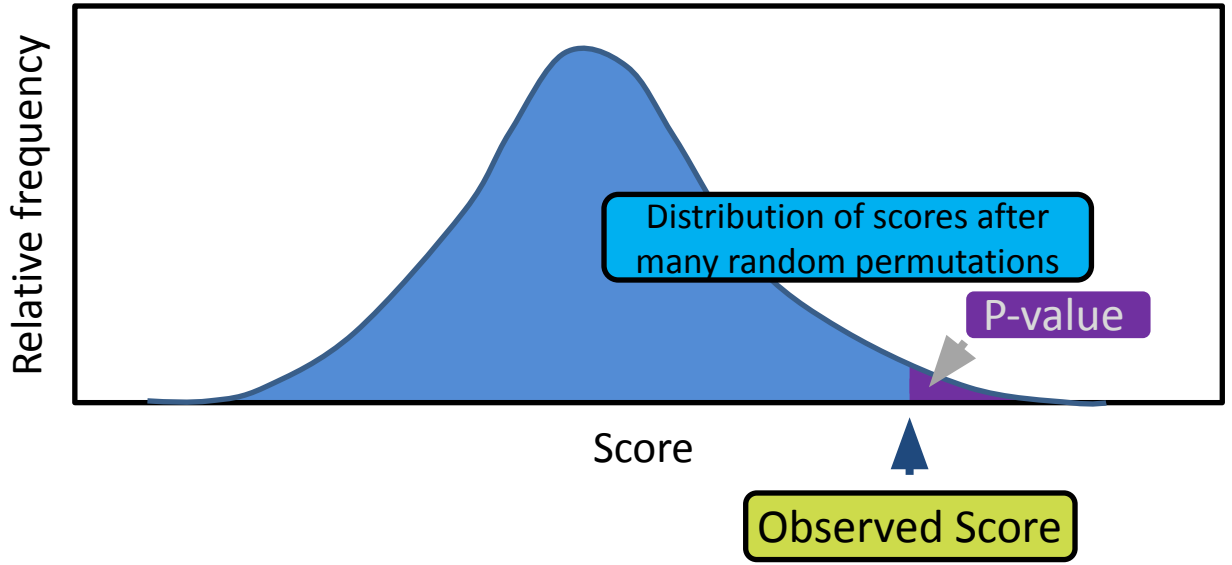
- **Correlation does not imply causation**





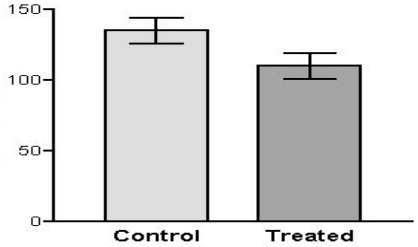


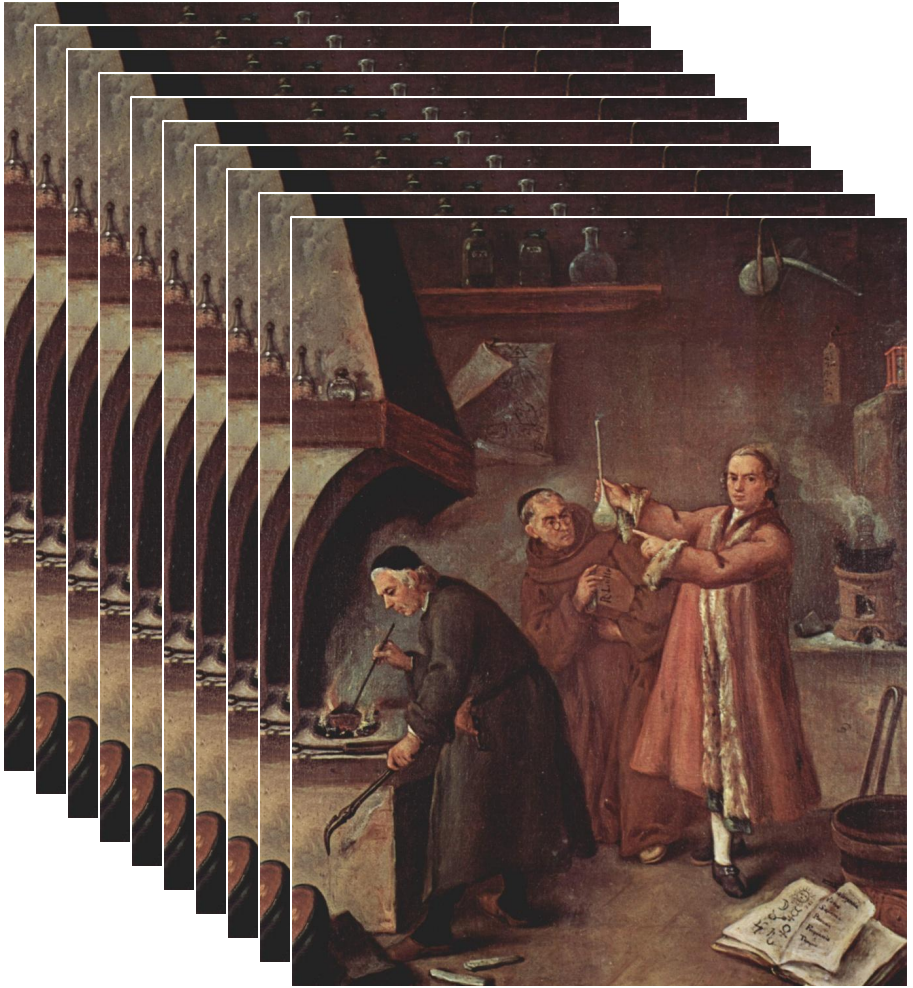
Permutation statistics



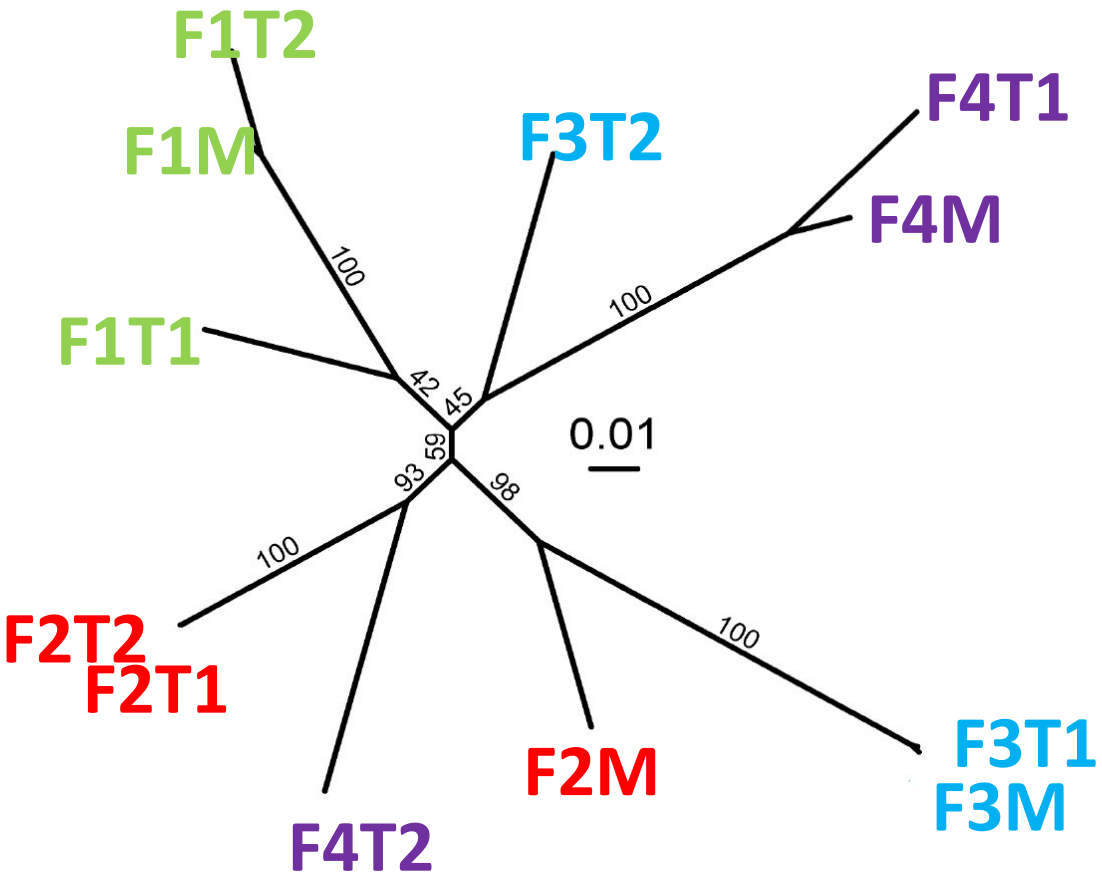
Statistics

- We use statistics to assess confidence in an experimental result
  - We repeat the experiment *N* times and test how <u>robust</u> the result is
  - How much variation is there in the result?



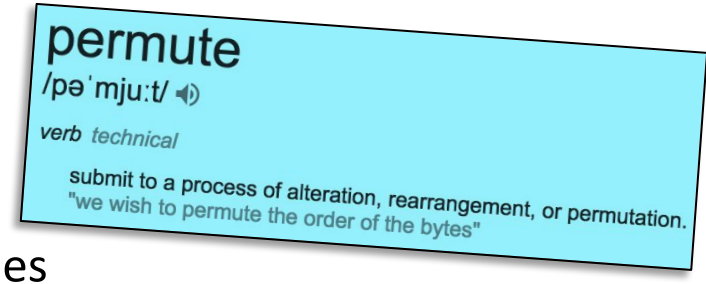


Are the viruses from family members closely related?



A question of statistics

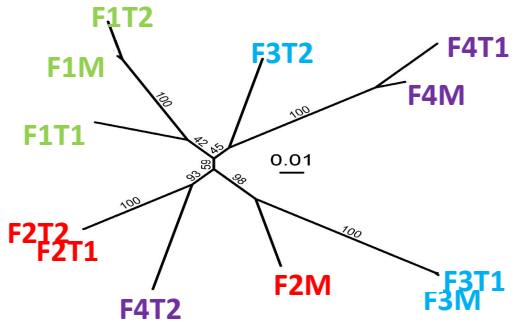
- What do we mean by “closely” related?
  - More closely than expected!
  - This is a question of statistics
- Statistics are not well defined for many bioinformatic analyses
  - For example, we need to consider the number of leaves in the tree, the shape of the tree, the number of families and family members in the analysis, and possibly many other variables



- A simple solution is data permutation:
  - Permute (shuffle) the leaves in the tree 1000* times
  - Register if the family members cluster together better than in the original tree
    - If they often cluster better, this suggests they originally did not cluster so well
    - If they rarely cluster better, this suggests they originally clustered pretty well

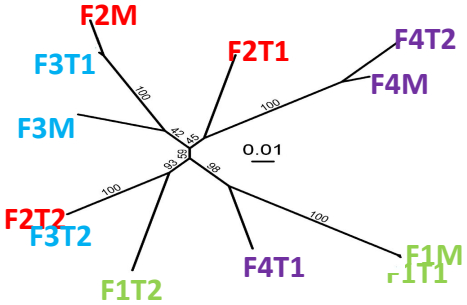
* or another high number

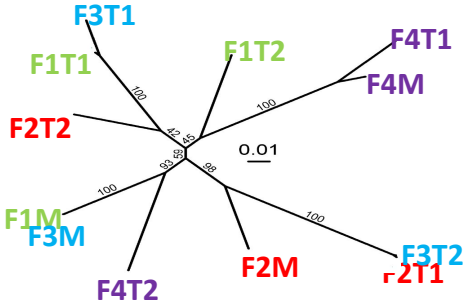
First, we need a “score”

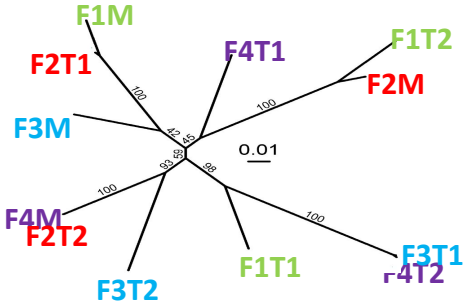


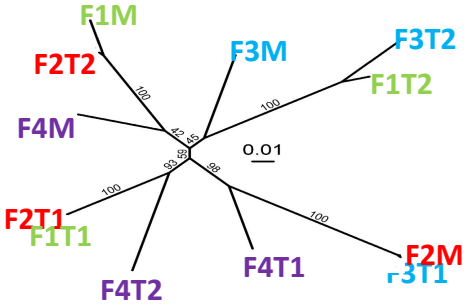
- In our example, the score should register “how well” the family members are clustered
  - For example:
    - The number of leaves whose closest other leaf is from the same family
    - The number of branches that you can break off the tree, that contain only leaves from one family
    - The minimum number of branches you need to switch around before the tree is perfect (reversed)
    - etc…
- This score is known as a “statistic”
  - For some bioinformatic questions, robust statistics have already been developed

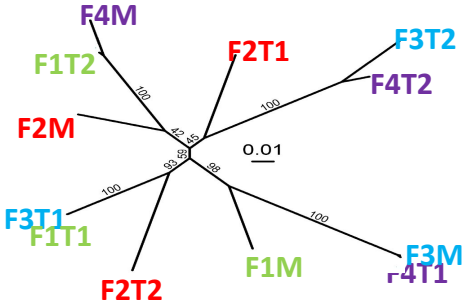
How high is the score by random chance?

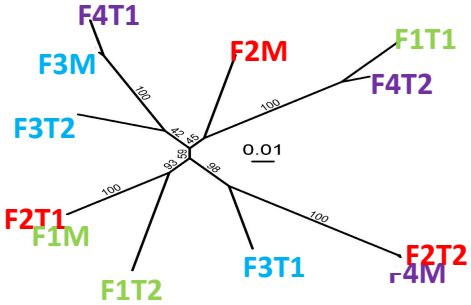


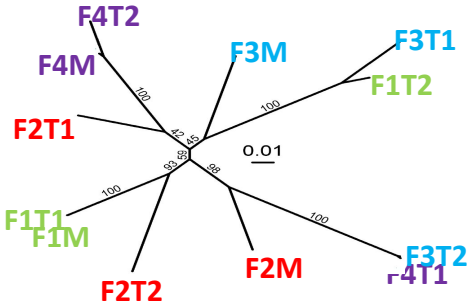


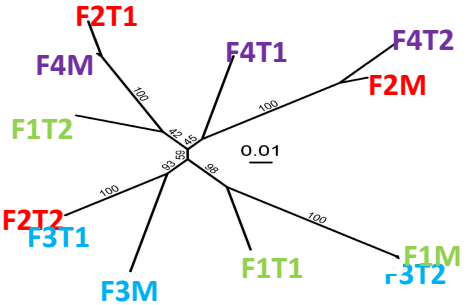


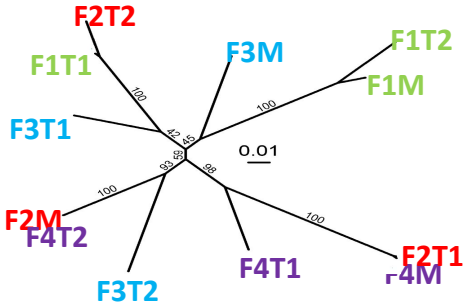






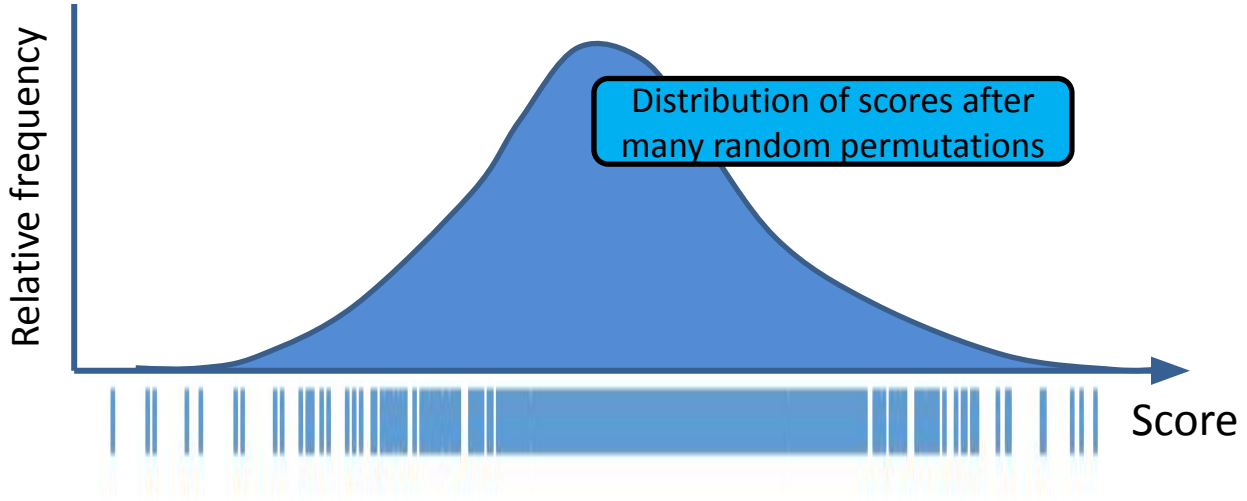






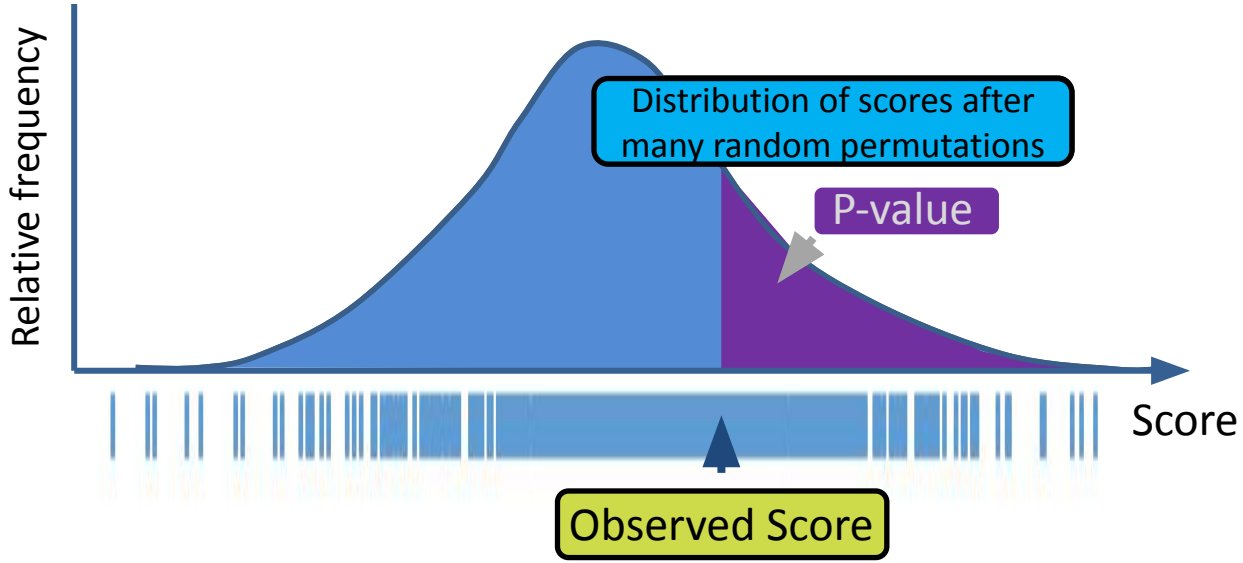
1000x

How often is every score found?



Probability value (P-value)

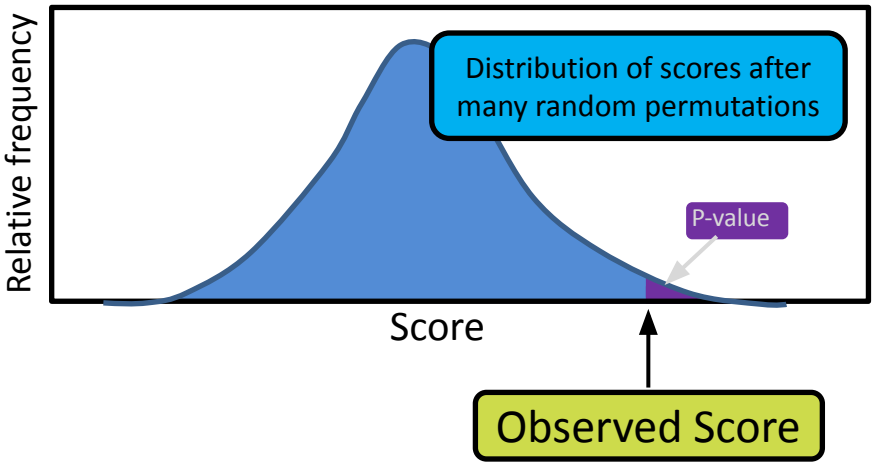
- The P-value is defined as the probability of observing a score as good as, or better than your **Observed Score** by chance



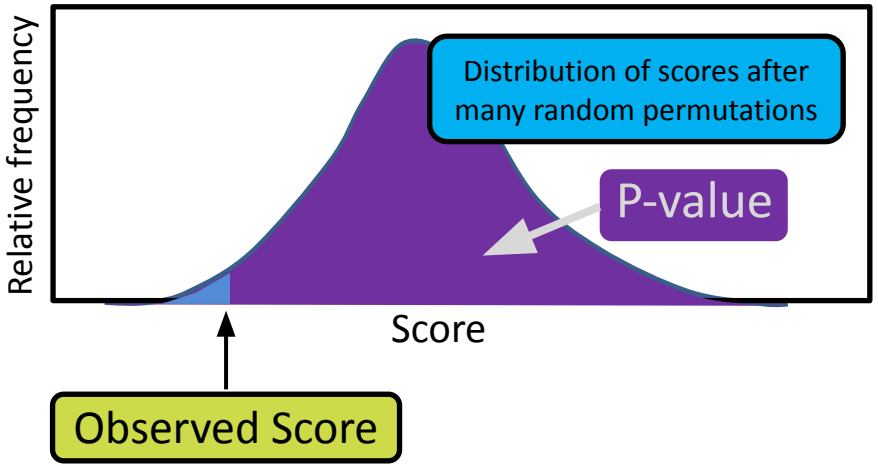
- In permutation statistics, this corresponds to the fraction of times that the permuted score is equal or higher than Your Score

Permutation statistics (shuffling)

- If the randomly permuted data **rarely** has a higher score than your Observed Score, then the P-value is **low**, and your observation is **meaningful (significant)**



- If the randomly permuted data **often** has a higher score than your Observed Score, then the P-value is **high**, and your observation is **not meaningful (not significant)**



Youtube videos

- **CrashCourse, “What is statistics”**

https://www.youtube.com/watch?v=sxQaBpKfDRk

- **Bozeman science, “Student’s t-test”**

https://www.youtube.com/watch?v=pTmLQvMM-1M

- **StatQuest, “p-values: What they are and how to interpret them”**

https://www.youtube.com/watch?v=vemZtEM63GY

- **DATAtab, “Mann-Whitney U test”**

https://www.youtube.com/watch?v=Twk6lBhBI88

- **StatQuest, “Pearson’s correlation”**

https://www.youtube.com/watch?v=xZ_z8KWkhXE

- **DATAtab, “Spearman Rank Correlation”**

https://www.youtube.com/watch?v=XV_W1w4Nwoc

- **StatQuest, “Linear Regression”**

https://www.youtube.com/watch?v=7ArmBVF2dCs

Papers

- Open science collaboration, “**Estimating the reproducibility of psychological science**”, Science, 2015

https://www.science.org/doi/10.1126/science.aac4716

- Baker 2016, “**1,500 scientists lift the lid on reproducibility**”, Nature, 2016

https://www.nature.com/articles/533452a

Other material

- Science fictions links

https://www.sciencefictions.org/p/science-fictions-links-for-october

- Scientific reports, submission guidelines

https://www.nature.com/srep/author-instructions/submission-guidelines

How to continue your learning

- **These slides:** https://github.com/waltercostamb/statistics

- **StatQuest with Josh Starmer YouTube channel**

https://www.youtube.com/@statquest

  - Frequentist statistics
  - Short introduction videos on many different topics

- **Course "Statistical Rethinking", by Richard McElreath from MPI Leipzig**

https://github.com/rmcelreath/stat_rethinking_2024

  - Bayesian statistics
  - hapters 1 - 3: Foundation of Bayesian inference
  - Chapter 4: Linear regression
  - Chapters 5 - 17: Advanced statistics (including generalized linear models)

Image from: https://www.statology.org/7-best-youtube-channels-statistics-free/

Take home message

- **Statistics offers several methods for data analysis and interpretation**
- **Understanding your variables is key**
  - Visualization
  - Filtering
  - Data preparation (e. g. normalization, transformation)
  - Analysis
- **Defining your problem well allows you to choose an appropriate method**
- **p-values do not tell you the whole story**
  - effect sizes are also important

- **Learning more statistics will help you become an even better a scientist**
- **Statistics can be fun!**

Image from: https://medium.com/geekculture/2-most-popular-statistics-concepts-in-data-science-b2b166b779da

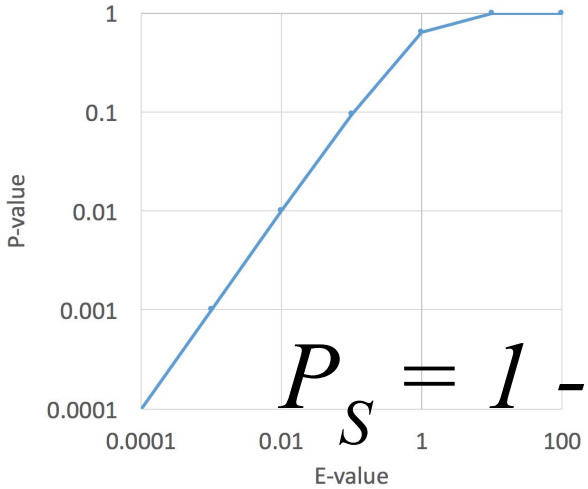
Thank you!

Permutation statistics

- In permutation statistics, the minimum P-value depends on the number of random permutations
  - For 100 permutations, the best P-value is <0.01
  - For 1000 permutations, the best P-value is <0.001
- Data permutation can randomize the signal, while preserving important characteristics of the data, such as:
  - Shape of a phylogenetic tree
  - Number and lengths of sequences
  - Nucleotide or amino acid composition of sequences
  - etc... (basically anything, depending on your analysis)
  - Sometimes it can be challenging to figure out which aspects of the data to randomize, and which aspects to preserve
- Permutation is a form of non-parametric statistics because it makes no assumptions about how the data is distributed
- Permutation statistics are often applied in bioinformatics
  - The computer can easily shuffle a dataset many times

The P-value and the E-value

- How often do you expect to find a score, as good as, or better than your score (≥*S*) by random chance?
  - P-value: probability of observing at least one hit with score ≥*S* by chance
  - E-value: expected number of hits with score ≥*S* by chance



$$P_S = 1 - e^{-E_S}$$