

STATISTICS FOR COMPUTATIONAL BIOLOGY PROJECTS

Erica Holdmore

DFCI Department of Data Sciences

June 18, 2025 – 1pm-4pm




<https://github.com/eholdmore/StatsForCompBio>

INTRODUCTION TO STATS AND COMPUTATIONAL BIOLOGY

COMPUTATIONAL BIOLOGY AND ITS APPLICATIONS

Seeks to understand biological systems and their relationships through **data analysis, mathematical modeling,** and other **quantitative tools.**



Often applied to understand:

Population
genomics

Evolutionary
genomics &
proteomics

Regulatory
& metabolic
networks

Gene-
disease
associations

Biomedical
imaging
analysis

Infectious
disease
dynamics

Many more!

STATISTICS IN COMPUTATIONAL BIOLOGY

Computational biology approaches typically generate **large amounts of data**.

Purpose of data analysis is to **identify patterns and trends** in biological data.

Allows us to **rigorously test hypotheses** about biological processes and their relationships.

Helps **estimate parameters, fit models, and validate** models and simulations.

Provides guidance on appropriate **experimental design**.

Can be used to **make predictions and guide future research** directions.

BEGINNING A PROJECT WITH STATISTICS IN MIND

Can provide advice about
aspects of **experimental
design**.

- Sample size
- Replication
- Randomization & Controls
- Batch effects

Important to plan
appropriate analyses
before an experiment
begins.

- Increase power of analyses
- Reduce likelihood of Type II error ("false negatives")

BEGINNING A PROJECT WITH STATISTICS IN MIND

Important to plan appropriate analyses **before** an experiment begins.

A PRIORI ANALYSIS

- “Prospective”, “planned”
- Hypothesis-driven
- Increased power against **Type II error**
- More thoughtful research design
- Not always possible

POST HOC ANALYSIS

- “Posteriori”, “unplanned”
- Exploratory
- Provides insight and generates ideas
- Need to adjust significance value for multiple comparisons
- Interpret with caution
- Not advised for estimating treatment effect in randomized clinical trials

TYPES OF STATISTICAL ERROR

	True	False
Accept H_0	✓	Type II Error “False Negative”
Reject H_0	Type I Error “False Positive”	✓

WHAT IS A SIGNIFICANCE VALUE?

- **p-value** = the probability of observing a particular test statistic value purely by chance given a particular distribution
- Can be interpreted as the probability of type I error (false positive).



WORKSHOP OUTLINE

Experimental Design

Probability Distributions & Data Cleaning

Statistical Inference

Statistical Methods for Genomics

Interpretation & Data Visualization

Ethical Considerations & Challenges

Wrap-up and Q&A

EXPERIMENTAL DESIGN

GENERAL CONSIDERATIONS FOR EXPERIMENTAL DESIGN

Appropriate
controls

Sample size

Replication

Batch
Effects

Control minimize the effects of all variables other than the one(s) being tested.



How do we design an experiment with good controls?

Make observations

Know your study system

Have a clear hypothesis

Select a **specific, measurable** independent variable

Decide on appropriate control groups

Include randomization where appropriate

Monitor controls throughout experiment

APPROPRIATE
CONTROLS

SAMPLE SIZE & REPLICATES

- **Sample** = subset designed to represent the **population** being studied
- **Replicates** = multiple experimental runs under the same treatment
- Proper replication is an essential component of any experiment.
 - Ensures conclusions about experimental treatments are reliable
 - Provides information about natural variability in response variables



Pseudoreplication = "the use of inferential statistics to test for treatment effects with data from experiments where either **treatments are not replicated** (though samples may be) or **replicates are not statistically independent**." (Hurlbert 1984)



Should be avoided or, when unavoidable, statistically accounted for using a **repeated measures** test.



Example: A study participant's blood pressure is taken before being administered a blood pressure lowering drug. Their blood pressure is taken again 10 minutes and 60 minutes after the drug is administered.

PSEUDOREPLICATION & BATCH EFFECTS

POWER & EFFECTS SIZE

- **Power** = the probability that a statistical test will reject a false null hypothesis
- A larger sample size provides more power, meaning your statistical test is more likely to detect an effect.

- **Effect size** = the absolute difference between groups + variability

Cohen (1988) provides cutoffs for effect sizes as:

Small ≥ 0.02 ,

Medium ≥ 0.15

Large ≥ 0.35

ACTIVITY I: POWER ANALYSIS IN R

- The 'pwr' package (Champley 2020) can perform power analyses for a variety of common statistical test.
- Enter three of the four parameters (effect size, sample size, significance level, power) as well as your number of groups and the fourth is calculated.
- On the right is an example power analysis for ANOVA.
- Note that **sample size is per group!**

Input:

```
# load package
#install.packages("pwr")
library(pwr)
# calculate effect size
pwr.anova.test(k=2,
               n=20,
               sig.level=.05,
               power=.8)
```

Output:

Balanced one-way analysis of variance

```
      k = 2
      n = 20
      f = 0.4545483
sig.level = 0.05
power = 0.8
```


ACTIVITY 1: POWER ANALYSIS IN R

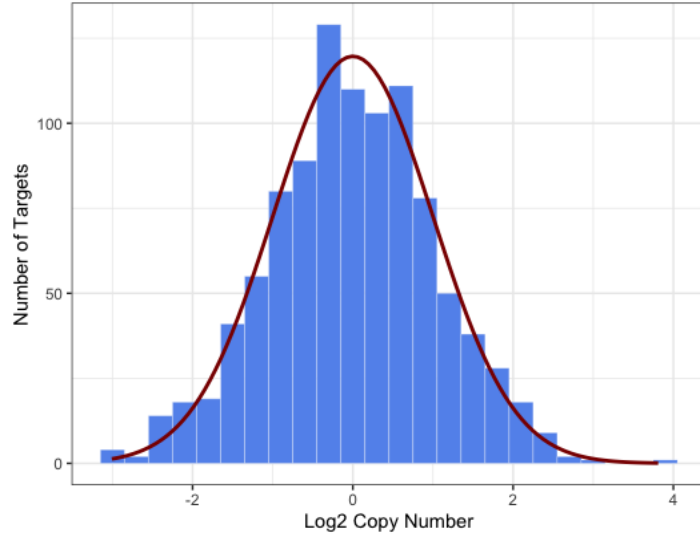
- The 'pwr' package (Champley 2020) can perform power analyses for a variety of common statistical test.
- Enter three of the four parameters (effect size, sample size, significance level, power) and the fourth is calculated.
- **Exercise:** Using the code provided in 'activities.R', determine how many participants you would need in each group (sample size) to have a power of 80% and a moderate effect size of 25% for each of the following tests.
 - **One-way ANOVA**
 - **GLM**
 - **Paired t-test (two tailed)**
 - **Independent t-test (one tailed – “greater”)**
 - **X² test**

PROBABILITY, PROBABILITY DISTRIBUTIONS, AND DATA CLEANING

BASIC PROBABILITY CONCEPTS

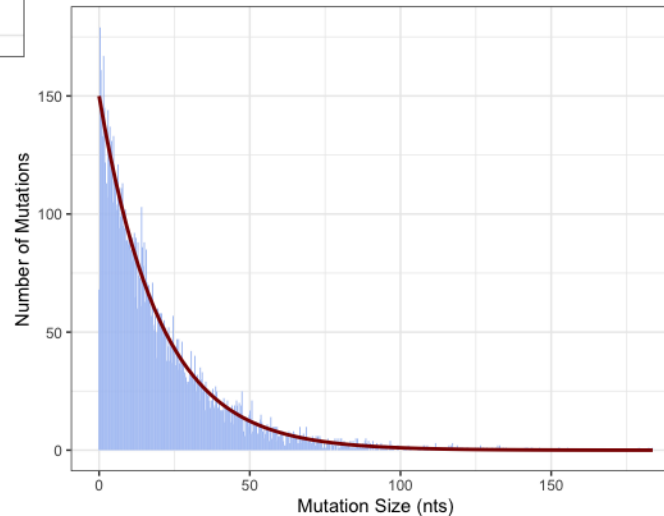
- **Probability** is an area of mathematics that deals with the likelihood of events occurring.
- Many statistical concepts are based upon probability including:
 - Sampling
 - Hypothesis testing
 - Significance values
 - Error & confidence
- Course GitHub includes suggested reading on probability.

PROBABILITY DISTRIBUTIONS



Normal Distribution
used to model log2 copy
number ratios

Exponential Distribution
used to model the size of
mutations



- In statistics, we use **probability distributions** to model "random" variables and quantify uncertainty.
- Fitting observed data to specific distributions allows for **predictions, inferences, and simulations** about populations.

COMMON PROBABILITY DISTRIBUTIONS IN COMPUTATIONAL BIOLOGY

Continuous

Discrete

Distribution

Normal

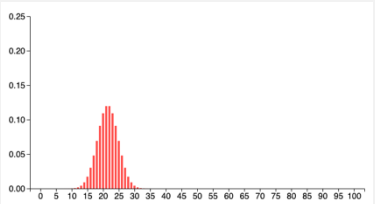
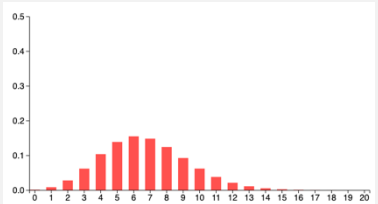
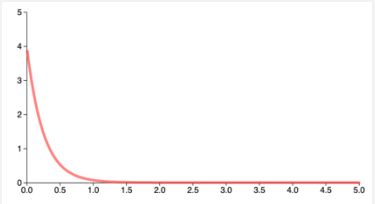
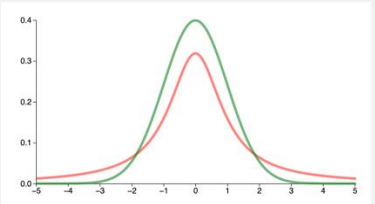
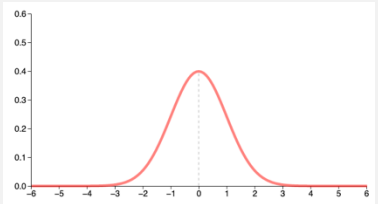
Student's T

Exponential

Poisson

Binomial

Appearance



Characteristics

Bell shape

Shorter, wider

Long left tail

Integer values

Two outcomes

Example Data

Height, weight, test scores

Small sample size

Time between events

Number of events

“Coin toss”

Application

Least squares & uncertainty

Unknown variance

Continuous-time Markov chain

Waiting time between events

Anytime data is binary

ACTIVITY 2: PROBABILITY IN COMPUTATIONAL BIOLOGY

- Interactive simulation
 - <https://seeing-theory.brown.edu/probability-distributions/index.html>
- Additional resource:
 - <https://probstats.org/>

Chapter 3

Probability Distributions

A probability distribution specifies the relative likelihoods of all possible outcomes.

Go to Probability Distributions



Random Variables



Discrete and Continuous

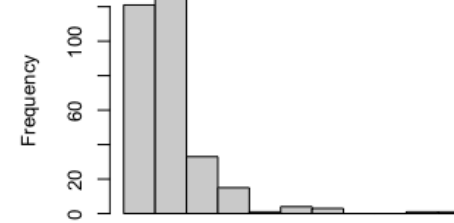


Central Limit Theorem

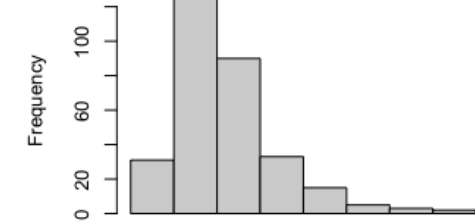
TRANSFORMATIONS

- **Data transformation** = applying the same deterministic function to all data points to facilitate statistical inference and/or interpretation
- For example, many statistical tests assume data is normally distributed.
- Right skewed distributions are the most common in biological data.
- Use tests such as Shapiro-Wilk to rigorously test for deviations for normality.

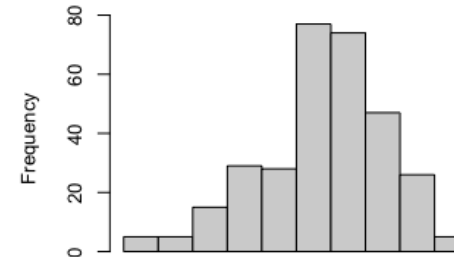
No Transform



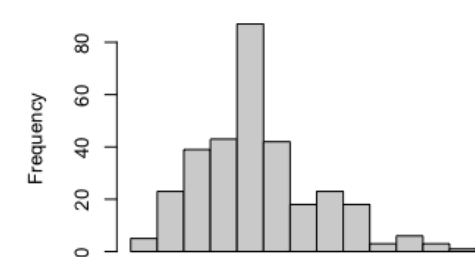
Log



Reciprocal

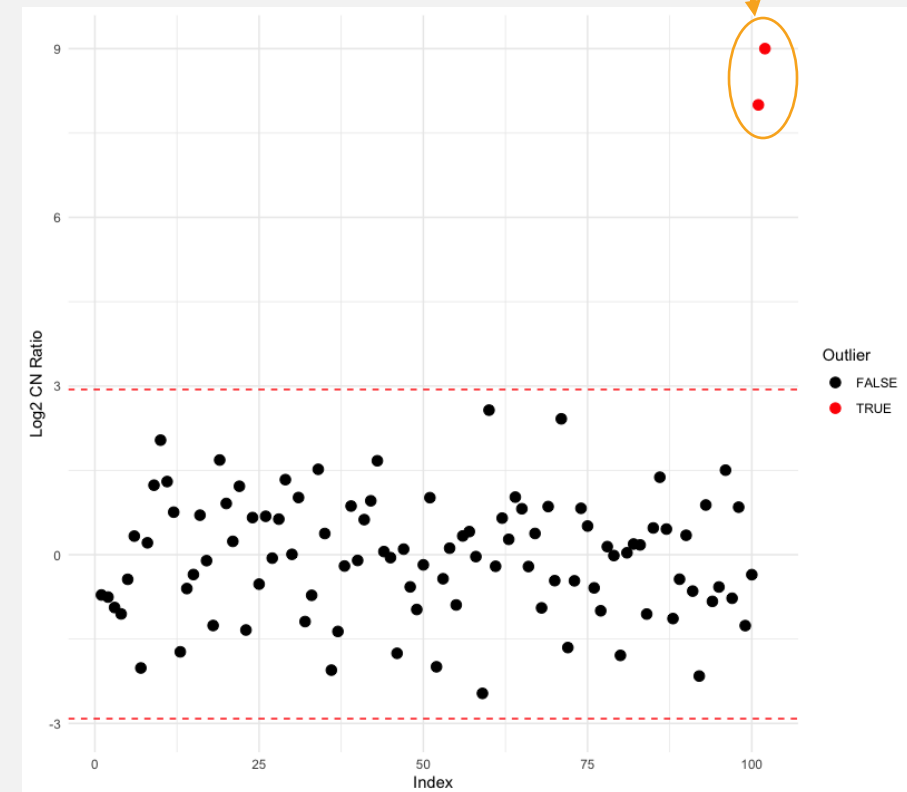
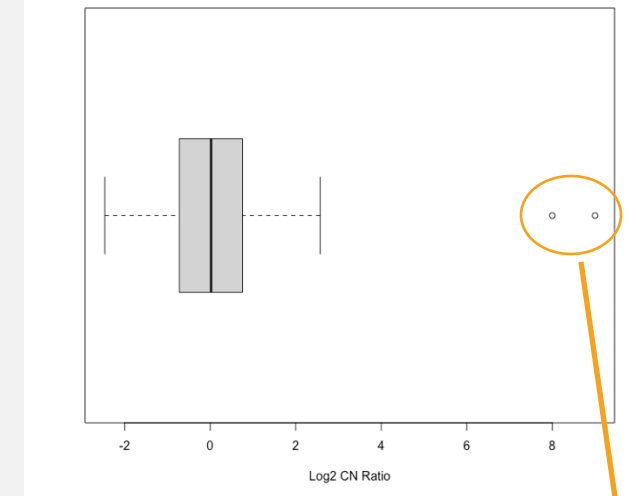


Box-Cox



OUTLIER ANALYSIS

- Outlier analysis seeks to identify and evaluate data points that are unusually far away from the mean of a dataset.
- Outliers may be caused by:
 - Experimental error
 - High variability (noise)
 - Something genuinely biologically interesting!
- Methods for detecting outliers:
 - **Descriptive: interquartile range (IQR), boxplot, Tukey**
 - Significance Tests: Grubbs', Dixon's, Rosner
 - Unsupervised Clustering: DBSCAN



OUTLIER ANALYSIS

- Outlier analysis seeks to identify and evaluate data points that are unusually far away from the mean of a dataset.
- Outliers may be caused by:
 - Experimental error
 - High variability (noise)
 - Something genuinely biologically interesting!
- Methods for detecting outliers:
 - Descriptive: interquartile range (IQR), boxplot, Tukey
 - **Significance Tests: Grubbs', Dixon's, Rosner**
 - Unsupervised Clustering: DBSCAN

Grubbs test for one outlier

```
data: right_outlier
G = 5.63483, U = 0.68252, p-value = 3.586e-08
alternative hypothesis: highest value 9 is an outlier
```

Results of Outlier Test

```
-----
Test Method:                Rosner's Test for Outliers
Hypothesized Distribution:   Normal
Data:                       right_outlier
Sample Size:                102
Test Statistics:             R.1 = 5.634825
                             R.2 = 6.088720
                             R.3 = 2.510327
                             R.4 = 2.454923
                             R.5 = 2.418916
Test Statistic Parameter:   k = 5
Alternative Hypothesis:      Up to 5 observations are not
                             from the same Distribution.
Type I Error:               5%
Number of Outliers Detected: 2
```

	i	Mean.i	SD.i	Value	Obs.Num	R.i+1	lambda.i+1	Outlier
1	0	0.13115401	1.5739345	9.000000	102	5.634825	3.390825	TRUE
2	1	0.04334366	1.3067863	8.000000	101	6.088720	3.387474	TRUE
3	2	-0.03622291	1.0387812	2.571458	60	2.510327	3.384083	FALSE
4	3	-0.06256312	1.0099449	2.416773	71	2.454923	3.380651	FALSE
5	4	-0.08786247	0.9830997	-2.465898	59	2.418916	3.377176	FALSE

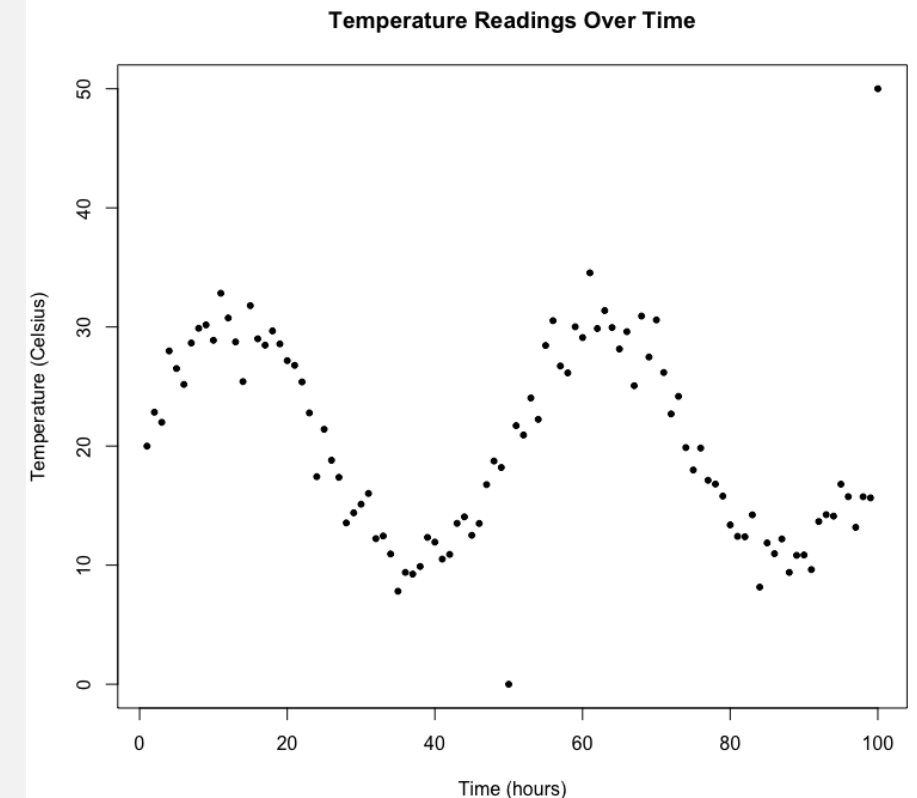
OUTLIER ANALYSIS

- Outlier analysis seeks to identify and evaluate data points that are unusually far away from the mean of a dataset.
- Outliers may be caused by:
 - Experimental error
 - High variability (noise)
 - Something genuinely biologically interesting!
- Methods for detecting outliers:
 - Descriptive: interquartile range (IQR), boxplot, Tukey
 - Significance Tests: Grubbs', Dixon's, Rosner
 - **Unsupervised Clustering: DBSCAN**

```
DBSCAN clustering for 100 objects.  
Parameters: eps = 6, minPts = 4  
Using euclidean distances and borderpoints = TRUE  
The clustering contains 1 cluster(s) and 2 noise points.
```

```
0 1  
2 98
```

```
Available fields: cluster, eps, minPts, dist, borderPoints
```



OUTLIER ANALYSIS

- Quick R demo of Grubbs' test and DBSCAN
- For more info:
 - <https://www.graphpad.com/quickcalcs/grubbs1/>
 - <https://builtin.com/data-science/how-find-outliers-examples>

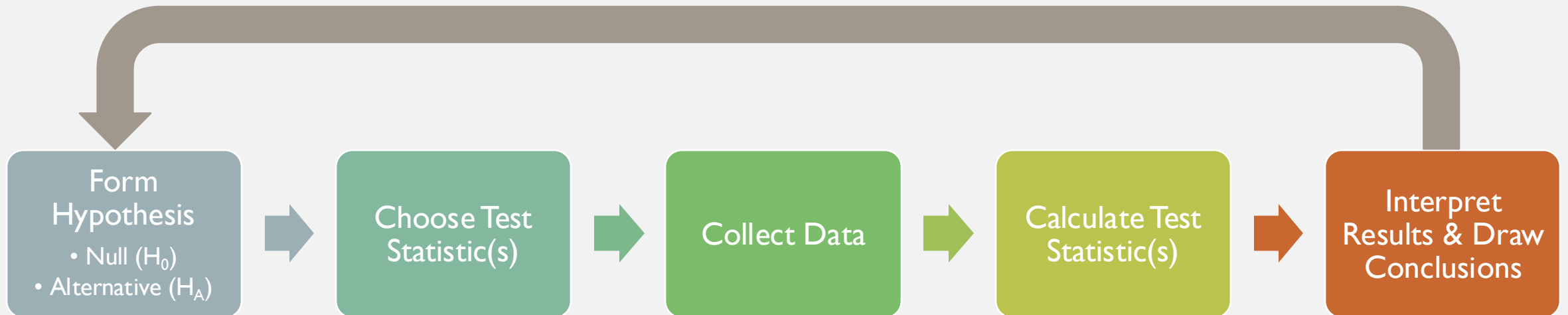
STATISTICAL INFERENCE

TESTING FOR DIFFERENCES

- One of the most common tasks in statistics is comparing 2+ sets of samples.
 - Healthy vs disease
 - Treatment vs placebo
- Simply taking the difference between means does not account for variability.
 - Small differences could be due to chance.
- Is the difference we observe between groups comparable to the difference we would see due purely to sampling (chance)?
- This is the basis for **hypothesis testing**.

BASICS OF HYPOTHESIS TESTING

- **Hypothesis testing** is the fundamental process by which we make inferences about a population based on sample data.



UNCERTAINTY

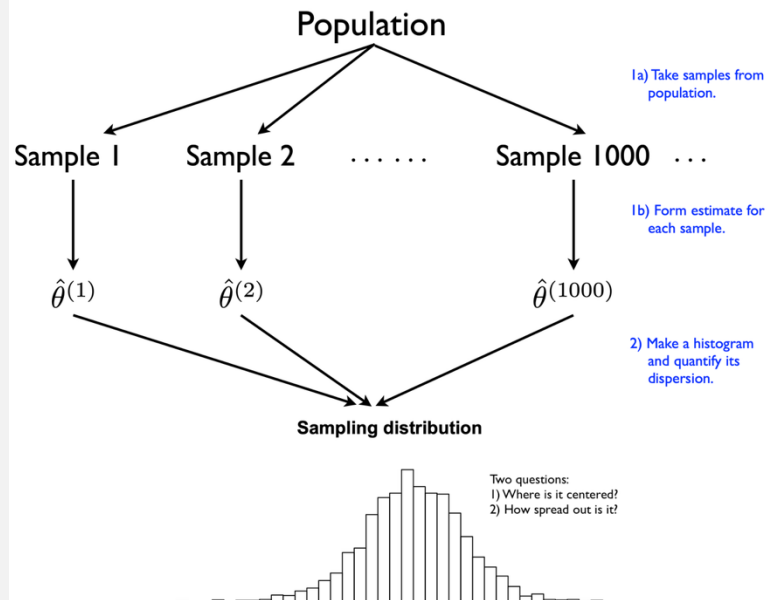


Fig 8.5 from J. Scott (2021)

- **Uncertainty** describes how much an estimate may differ from the true value.
- Fundamentally, it arises from the fact that we must use a **sample** to make inferences about a **population**.
- Mechanistically, uncertainty can arise through measurement/reporting error or variability that is intrinsic to individuals or processes.
- Quick demo:
<https://www.statcrunch.com/applets/type3&samplingdist>

UNCERTAINTY

- How do we **quantify uncertainty**?
 - Variance/Standard deviation/**Standard error**
 - Confidence intervals
 - A 95% confidence interval has a 95% chance of containing the true mean.
 - There is a 95% probability that the true mean lies within the interval.

UNCERTAINTY

- How do we **quantify uncertainty**?
 - Variance/Standard deviation/**Standard error**
 - Confidence intervals
 - **Correct: A 95% confidence interval has a 95% chance of containing the true mean.**
 - Incorrect: There is a 95% probability that the true mean lies within the interval.

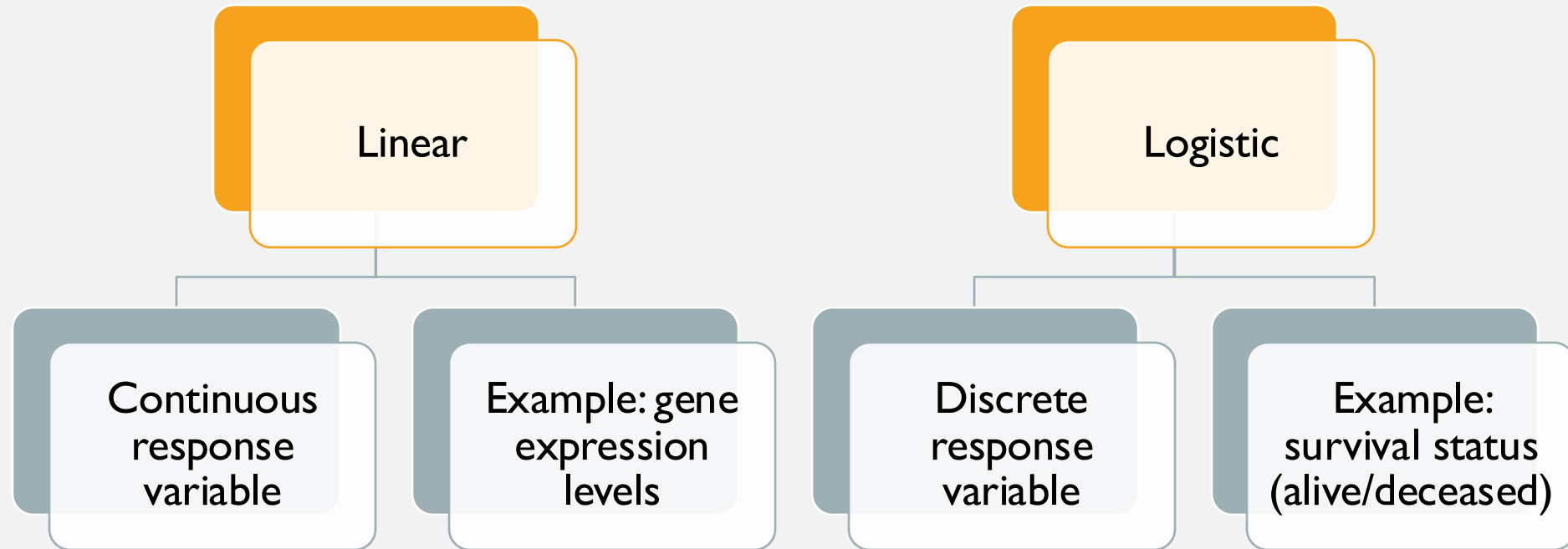
10-MINUTE BREAK

Stretch, use the restroom, grab a snack!

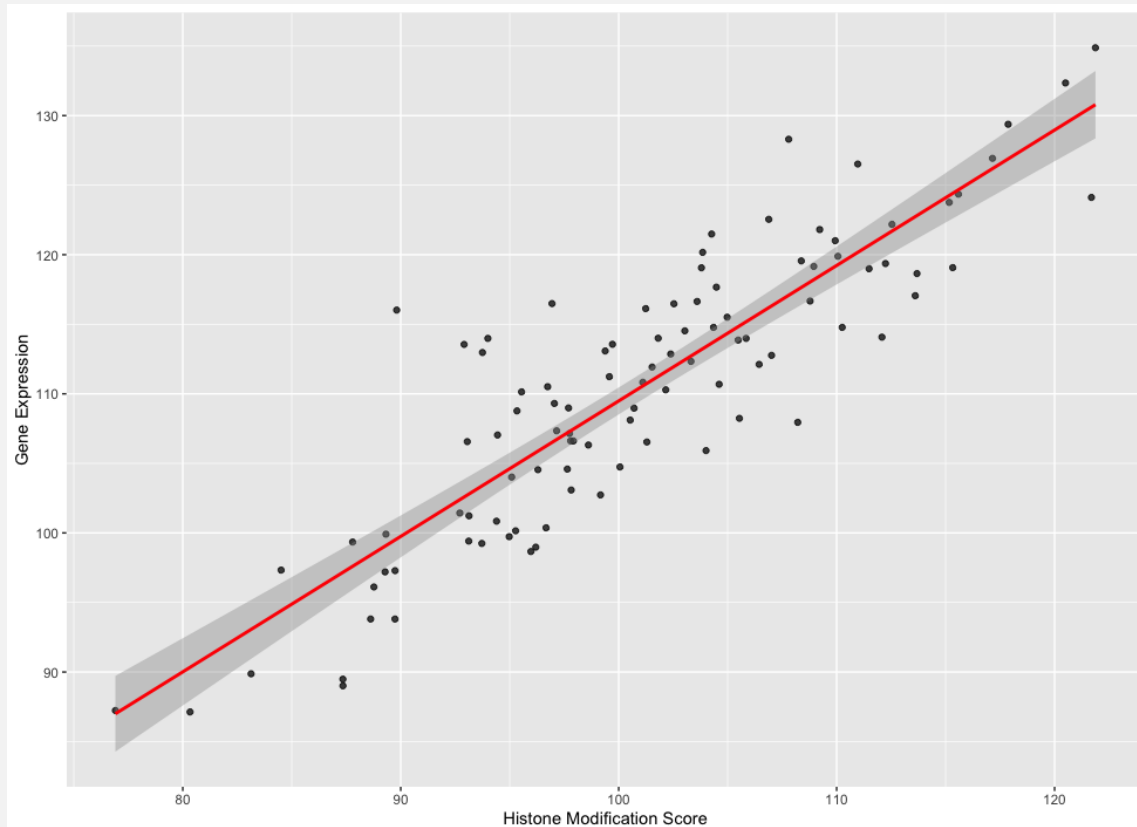
Meeting back here shortly.

STATISTICAL METHODS FOR GENOMICS

COMMON ANALYSIS METHODS: REGRESSION



COMMON ANALYSIS METHODS: LINEAR REGRESSION



Call:

```
lm(formula = normal_data ~ predictor, data = data1)
```

Residuals:

Min	1Q	Median	3Q	Max
-15.5723	-2.3947	0.0966	2.9096	9.8830

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	13.39292	4.82246	2.777	0.00657 **
predictor	0.79292	0.04351	18.222	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.38 on 98 degrees of freedom

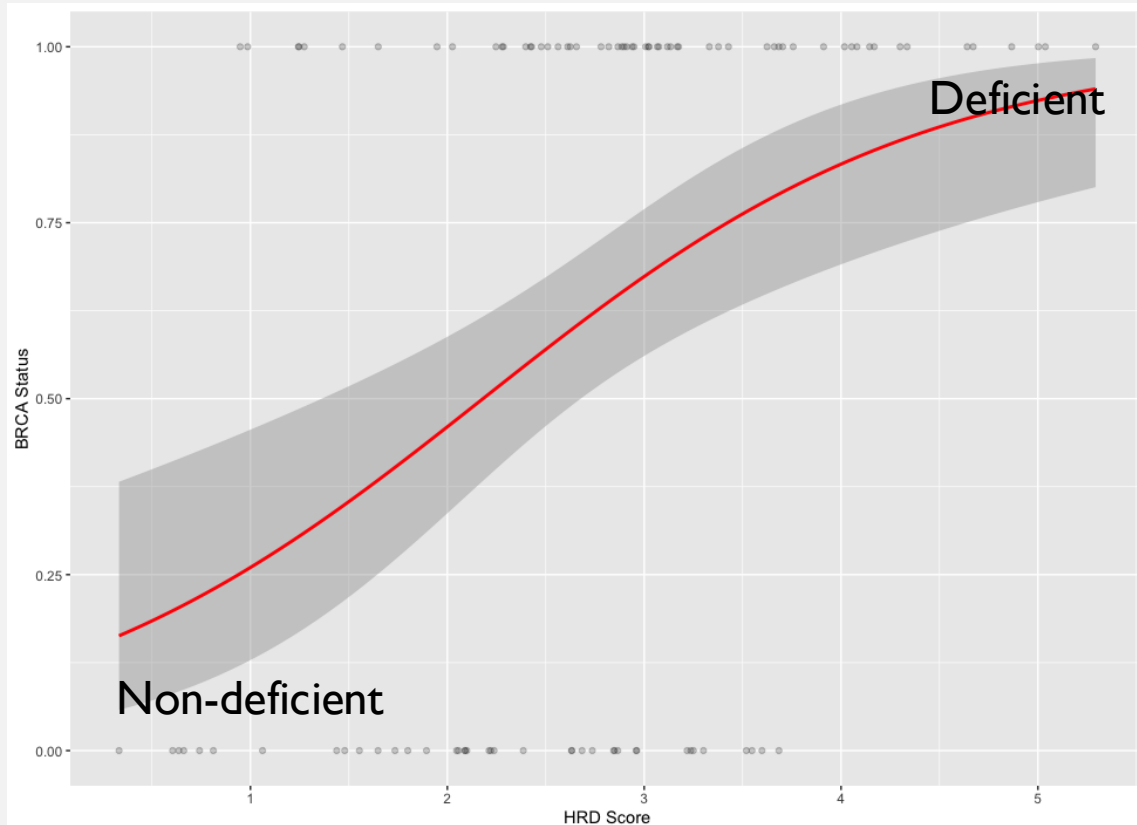
Multiple R-squared: 0.7721, Adjusted R-squared: 0.7698

F-statistic: 332 on 1 and 98 DF, p-value: < 2.2e-16

```
> fit1$coefficients
```

(Intercept)	predictor
13.3929240	0.7929152

COMMON ANALYSIS METHODS: LOGISTIC REGRESSION



Call:

```
glm(formula = binom_data ~ predictor, data = data1)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.05799	0.10925	-0.531	0.597
predictor	0.19520	0.04245	4.598	1.27e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.2014385)

Null deviance: 24.000 on 99 degrees of freedom
Residual deviance: 19.741 on 98 degrees of freedom
AIC: 127.54

Number of Fisher Scoring iterations: 2

```
> fit2$coefficients
```

(Intercept)	predictor
-0.05799028	0.19520065

COMMON ANALYSIS METHODS: GENERALIZED LINEAR MODELS

Gaussian (link = “identity”)

Binomial (link = “logit”)

Poisson (link = “log”)

Negative Binomial (link = “logit”)

Gamma (link = “inverse”)

- Generalized linear models (GLMs)
 - “Non-parametric” in the sense that they do not assume that the response variable is normally distributed.
 - The **link function** relates the distribution of the response variable to the linear predictors of the model.
 - Does still assume:
 - Independence of response variable
 - Relationship between predictor and link-transformed response is linear
 - “Everything is a GLM!”
 - Logistics regression, Poisson regression, ANOVA, t-test, chi-square, and many others are special cases.

GLM STRUCTURE REFRESHER

Response Variable (Y)

- Outcome we're modeling
- e.g., RNAseq counts, methylation values, binary phenotype

Predictors (X)

- Covariates/features used to explain Y
- e.g., gene expression, tumor stage, age

Link Function (g)

- Transforms $E(Y) \rightarrow$ linear predictor
- $g(E[Y]) = X\beta$
- Identity for continuous data
- Log for count data
- Logit for binary outcomes

Distribution of Y

- Choose based on Y
- Exponential family e.g., Poisson, Gaussian, Binomial

GLM BIOLOGICAL EXAMPLES

- **Core Components:**

- Response variable (Y)
- Predictors (X)
- Link function (g)
- Distribution of Y

Predictors	Response Variable	Link Function	Distribution of Response	Example Use
Experimental Design Groups	Raw Count Data	Log	Poisson or Negative Binomial	DE Analysis
Age	Methylation Beta Value (0-1)	Identity or Logit	Gaussian	Epigenetic Age Estimation
Clinical Phenotype	Response to Treatment (Y or N)	Logit	Binomial	Treatment Effect

MODELING RNASEQ WITH GLMS

Why Use Negative Binomial for RNAseq?

- Consists of discrete, non-negative counts
- Counts vary due to both biological variability and technical noise.
- Poisson assumes mean = variance → too strict for real RNAseq data
- Negative Binomial allows for “extra variance” (**overdispersion**) beyond Poisson

Overdispersion in RNAseq

- Observed variance exceeds what Poisson predicts.
- Sources: batch effects, sample heterogeneity, sequencing depth.
- NB models introduce a dispersion parameter to capture this.

Application in DESeq2

- DESeq2 fits a NB GLM for each gene:
 - Link function: log
 - Models mean expression per condition.
- Estimates dispersion per gene, then smooths (**shrinks**) across genes.
- Improves power and stability, especially for low-count genes.

DESIGN MATRIX IN GLMS

What Is the Design Matrix?

- Matrix representation of predictors
- Rows = samples, Columns = variables
- Linear predictor: $\eta = X\beta$ where X is the design matrix and β are the coefficients

How Are Factors and Interactions Encoded?

- Categorical variables become dummy variables (0/1)
- Reference levels determine baseline comparisons
- Interactions: $\sim \text{genotype} * \text{treatment}$ expands to $\sim \text{genotype} + \text{treatment} + \text{genotype:treatment}$

Application in DESeq2

- Common formula: $\sim \text{condition} + \text{batch}$
 - Adjusts for **batch effects** while testing for **condition differences**
- Each term in the formula adds columns to the design matrix.
- Crucial for:
 - Accurate **coefficient interpretation**
 - Setting correct **reference levels**
 - Controlling for **confounding variables**

INTERPRETING GLM COEFFICIENTS

Magnitude & Direction of Effect

- Each coefficient β_i represents the effect of a predictor on the response.
- Sign: Direction of effect (+ = increase, – = decrease)
- Magnitude: Strength of effect (on the link scale)

Why Use Log Fold-Change in RNAseq?

- RNAseq uses a log link
- So, coefficients represent log fold-changes between groups
 - $\beta = 1 \rightarrow 2$ -fold increase
 - $\beta = -1 \rightarrow 2$ -fold decrease
- Log scale stabilizes variance and supports multiplicative interpretation.

• Role of Shrinkage (DESeq2)

- Raw log counts can be noisy, especially for low-count or variable genes.
- DESeq2 applies shrinkage estimators (e.g., ‘apeGLM’, ‘ashr’).
 - Pulls extreme estimates toward zero.
 - Improves reproducibility and ranking of genes.
- Helps control false positives while preserving large, confident effects.

MODEL FIT & DIAGNOSTICS

Residuals

- Residuals = observed — fitted values
- Help identify poor fit, outliers, or heteroscedasticity
- In RNAseq GLMs (e.g., DESeq2), deviance residuals are commonly used

AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion)

- Quantify model quality by balancing fit and complexity
- Lower values = better model (relative comparison)
- Useful when comparing nested or alternative models:
- e.g., $\sim \text{condition} + \text{batch}$ vs. $\sim \text{condition}$

Goodness-of-Fit

- Assesses how well the model explains the data.
- Methods include:
 - Deviance: compares model to a saturated model
 - Pseudo R^2 : adapted for GLMs, not a true R^2
 - In RNAseq, gene-specific fit diagnostics can flag low-quality or inconsistent features.

ACTIVITY 3: GLMS IN ACTION

- **Goal:** Walk through real vs simulated gene expression data using base R or `glm()` to:
 - Fit a linear model (normal)
 - Fit a Poisson model
 - Compare with a negative binomial
 - Interpret coefficients and residuals
 - Relate findings to DESeq2 concepts (without needing the full package)
- **Optional Extension:** Create a mini design matrix from a toy dataset and discuss how to modify it for paired designs, batch effects, etc.

```
> # Fit a Poisson GLM
> pois_mod <- glm(counts ~ group, family = "poisson", data = dat)
> summary(pois_mod)
```

Call:
glm(formula = counts ~ group, family = "poisson", data = dat)

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	2.18979	0.06108	35.85	<2e-16 ***
grouptreatment	0.79252	0.07362	10.76	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 185.10 on 59 degrees of freedom
Residual deviance: 59.97 on 58 degrees of freedom
AIC: 327.44

Number of Fisher Scoring iterations: 4

```
> # Fit a Negative Binomial GLM
> nb_mod <- suppressWarnings(glm.nb(counts ~ group, data = dat)) # might get
  ings; this is okay
> summary(nb_mod)
```

Call:
glm.nb(formula = counts ~ group, data = dat, init.theta = 142771.3383,
link = log)

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	2.18979	0.06109	35.85	<2e-16 ***
grouptreatment	0.79252	0.07363	10.76	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(142771.3) family taken to be 1)

Null deviance: 185.083 on 59 degrees of freedom
Residual deviance: 59.965 on 58 degrees of freedom
AIC: 329.44

Number of Fisher Scoring iterations: 1

Common statistical tests are linear models

Last updated: 02 April, 2019

See worked examples and more details at the accompanying notebook: <https://lindeloef.github.io/tests-as-linear>

	Common name	Built-in function in R	Equivalent linear model in R	Exact?	The linear model in words	Icon
Simple regression: $\text{lm}(y \sim 1 + x)$	y is independent of x P: One-sample t-test N: Wilcoxon signed-rank	t.test(y) wilcox.test(y)	$\text{lm}(y \sim 1)$ $\text{lm}(\text{signed_rank}(y) \sim 1)$	✓ for N > 14	One number (intercept, i.e., the mean) predicts y . - (Same, but it predicts the <i>signed rank</i> of y .)	
	P: Paired-sample t-test N: Wilcoxon matched pairs	t.test(y1, y2, paired=TRUE) wilcox.test(y1, y2, paired=TRUE)	$\text{lm}(y_2 - y_1 \sim 1)$ $\text{lm}(\text{signed_rank}(y_2 - y_1) \sim 1)$	✓ for N > 14	One intercept predicts the pairwise y₂-y₁ differences. - (Same, but it predicts the <i>signed rank</i> of y₂-y₁ .)	
	y ~ continuous x P: Pearson correlation N: Spearman correlation	cor.test(x, y, method='Pearson') cor.test(x, y, method='Spearman')	$\text{lm}(y \sim 1 + x)$ $\text{lm}(\text{rank}(y) \sim 1 + \text{rank}(x))$	✓ for N > 10	One intercept plus x multiplied by a number (slope) predicts y . - (Same, but with <i>ranked x</i> and y)	
	y ~ discrete x P: Two-sample t-test P: Welch's t-test N: Mann-Whitney U	t.test(y1, y2, var.equal=TRUE) t.test(y1, y2, var.equal=FALSE) wilcox.test(y1, y2)	$\text{lm}(y \sim 1 + G_2)^A$ $\text{gls}(y \sim 1 + G_2, \text{weights}=\dots^{B_1})^A$ $\text{lm}(\text{signed_rank}(y) \sim 1 + G_2)^A$	✓ ✓ for N > 11	An intercept for group 1 (plus a difference if group 2) predicts y . - (Same, but with one variance <i>per group</i> instead of one common.) - (Same, but it predicts the <i>signed rank</i> of y .)	
Multiple regression: $\text{lm}(y \sim 1 + x_1 + x_2 + \dots)$	P: One-way ANOVA N: Kruskal-Wallis	aov(y ~ group) kruskal.test(y ~ group)	$\text{lm}(y \sim 1 + G_2 + G_3 + \dots + G_N)^A$ $\text{lm}(\text{rank}(y) \sim 1 + G_2 + G_3 + \dots + G_N)^A$	✓ for N > 11	An intercept for group 1 (plus a difference if group $\neq 1$) predicts y . - (Same, but it predicts the <i>rank</i> of y .)	
	P: One-way ANCOVA	aov(y ~ group + x)	$\text{lm}(y \sim 1 + G_2 + G_3 + \dots + G_N + x)^A$	✓	- (Same, but plus a slope on x .) <i>Note: this is discrete AND continuous. ANCOVAs are ANOVAs with a continuous x.</i>	
	P: Two-way ANOVA	aov(y ~ group * sex)	$\text{lm}(y \sim 1 + G_2 + G_3 + \dots + G_N + S_2 + S_3 + \dots + S_K + G_2 * S_2 + G_3 * S_3 + \dots + G_N * S_K)$	✓	Interaction term: changing sex changes the y ~ group parameters. <i>Note: G_{2 to N} is an indicator (0 or 1) for each non-intercept levels of the group variable. Similarly for S_{2 to K} for sex. The first line (with G_i) is main effect of group, the second (with S_j) for sex and the third is the group * sex interaction. For two levels (e.g. male/female), line 2 would just be "S₂" and line 3 would be S₂ multiplied with each G_i.</i>	[Coming]
	Counts ~ discrete x N: Chi-square test	chisq.test(groupXsex_table)	Equivalent log-linear model $\text{glm}(y \sim 1 + G_2 + G_3 + \dots + G_N + S_2 + S_3 + \dots + S_K + G_2 * S_2 + G_3 * S_3 + \dots + G_N * S_K, \text{family}=\dots)^A$	✓	Interaction term: (Same as Two-way ANOVA.) <i>Note: Run glm using the following arguments: glm(model, family=poisson()) As linear-model, the Chi-square test is $\log(y_i) = \log(N) + \log(\alpha_i) + \log(\beta_j) + \log(\alpha_i \beta_j)$ where α_i and β_j are proportions. See more info in the accompanying notebook.</i>	Same as Two-way ANOVA
	N: Goodness of fit	chisq.test(y)	$\text{glm}(y \sim 1 + G_2 + G_3 + \dots + G_N, \text{family}=\dots)^A$	✓	(Same as One-way ANOVA and see Chi-Square note.)	1W-ANOVA

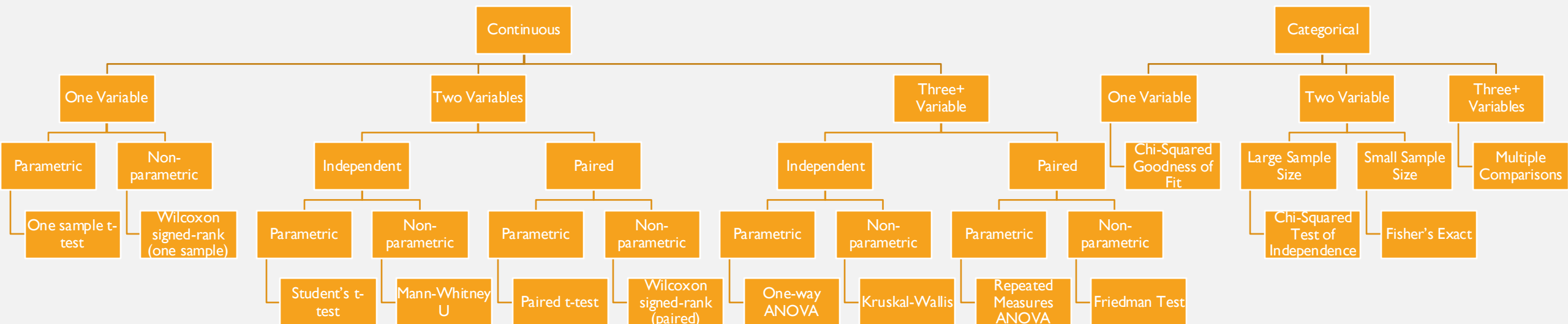
List of common parametric (P) non-parametric (N) tests and equivalent linear models. The notation $y \sim 1 + x$ is R shorthand for $y = 1 \cdot b + a \cdot x$ which most of us learned in school. Models in similar colors are highly similar, but really, notice how similar they *all* are across colors! For non-parametric models, the linear models are reasonable approximations for non-small sample sizes (see "Exact" column and click links to see simulations). Other less accurate approximations exist, e.g., Wilcoxon for the sign test and Goodness-of-fit for the binomial test. The signed rank function is `signed_rank = function(x) sign(x) * rank(abs(x))`. The variables G_i and S_i are "[dummy coded](#)" [indicator variables](#) (either 0 or 1) exploiting the fact that when $\Delta x = 1$ between categories the difference equals the slope. Subscripts (e.g., G_2 or y_1) indicate different columns in data. `lm` requires long-format data for all non-continuous models. All of this is exposed in greater detail and worked examples at <https://lindeloef.github.io/tests-as-linear>.

^A See the note to the two-way ANOVA for explanation of the notation.

^B Same model, but with one variance per group: `gls(value ~ 1 + G2, weights = varIdent(form = ~1|group), method="ML")`.



COMMON ANALYSIS METHODS: COMPARING MEANS



COMMON ANALYSIS METHODS: MULTIPLE COMPARISONS

- Hypothesis testing is not an error-free process.
- More tests on the same data = more type I errors (aka “false positives”)
 - Example: Compare expression of 3000 genes between two mutational types (e.g. BRCA1 & BRCA2)
- To account for this, most methods “adjust” the p-value.

So what can we do in these cases?

Control the familywise error rate

- **Familywise error rate** = proportion of all tests that yield a false positive
- Bonferroni correction: $p\text{-value} \times \text{number of tests}$
- This method is “harsh” in that they increase type II error (false negative) rates

Control the false discovery rate

- **False discovery rate** = proportion of all significant tests that yield a false positive
- Benjamini-Hochberg (BH or FDR)

COMMON ANALYSIS METHODS: MULTIPLE COMPARISONS

Recall that a p-value is the probability of observing a particular value purely by chance (i.e. a false positive).

ACTIVITY 4: MULTIPLE COMPARISONS

- Hands-on multiple comparisons analysis in R

CHOOSING AN APPROPRIATE TEST

Continuous
Predictor

Categorical
Predictor

Multiple
Predictors

Continuous

Counts

Binary

Proportions

Continuous

Counts

Continuous

Categorical

Linear
Regression
Or
GLM w/
Gaussian

GLM w/ Poisson

GLM w/
Binomial

GLM w/
Binomial

T-Test Or
ANOVA

Chi-Square Test

Linear
Regression
Or
GML

Linear
Regression
Or
GML

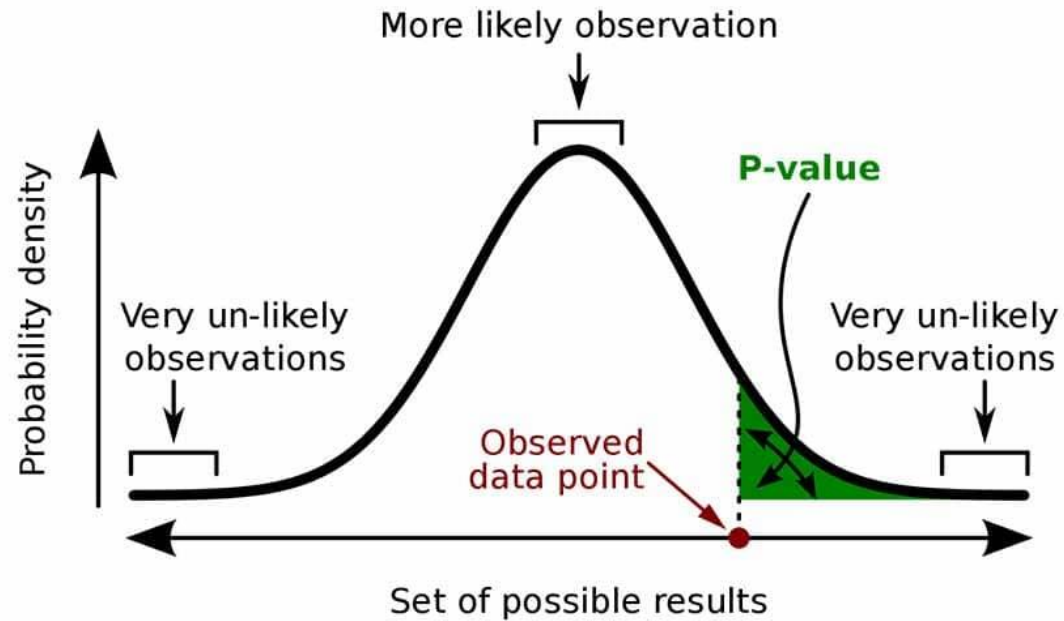
INTERPRETATION AND DATA VISUALIZATION

TYPES OF STATISTICAL ERROR

	True	False
Accept H_0	✓	Type II Error “False Negative”
Reject H_0	Type I Error “False Positive”	✓

WHAT IS A SIGNIFICANCE VALUE?

- **p-value** = the probability of observing a particular test statistic value purely by chance given a particular distribution
- Why is 0.05 commonly used as the threshold for statistical significance?
 - 1/20 chance of false positive
 - Convention



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

MEASURES OF EFFECT: DIRECTION AND MAGNITUDE

- Where **significance** tells us how likely it is that results are due to chance, measures of effect help us understand the **magnitude and direction** of differences.

[J Grad Med Educ](#). 2012 Sep; 4(3): 279–282.

doi: [10.4300/JGME-D-12-00156.1](https://doi.org/10.4300/JGME-D-12-00156.1)

PMCID: PMC3444174

PMID: [23997866](https://pubmed.ncbi.nlm.nih.gov/23997866/)

Using Effect Size—or Why the *P* Value Is Not Enough

[Gail M. Sullivan](#), MD, MPH and [Richard Feinn](#), PhD

▶ [Author information](#) ▶ [Copyright and License information](#) [PMC Disclaimer](#)

Statistical significance is the least interesting thing about the results. You should describe the results in terms of measures of magnitude –not just, does a treatment affect people, but how much does it affect them.

-Gene V. Glass¹

*The primary product of a research inquiry is one or more measures of effect size, not *P* values.*

-Jacob Cohen²

PRINCIPLES OF DATA VISUALIZATION



Color

Monochromatic,
contrasting,
complementary
Colorblind friendly
palettes



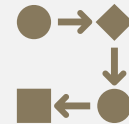
Keep it simple

Don't show too much
data in one figure.



Keep it honest

Avoid confusing scales on
axes.



Be clear

Decide exactly what you
want to show and show
only that.
Remain consistent.
Know your audience.

VISUALIZATION TECHNIQUES FOR GENOMICS & TRANSCRIPTOMICS

Sequence Analysis

- Tracks
 - UCSC Genome Browser/IGV
 - Circular maps

Annotation

Expression Profiles

- Heatmaps

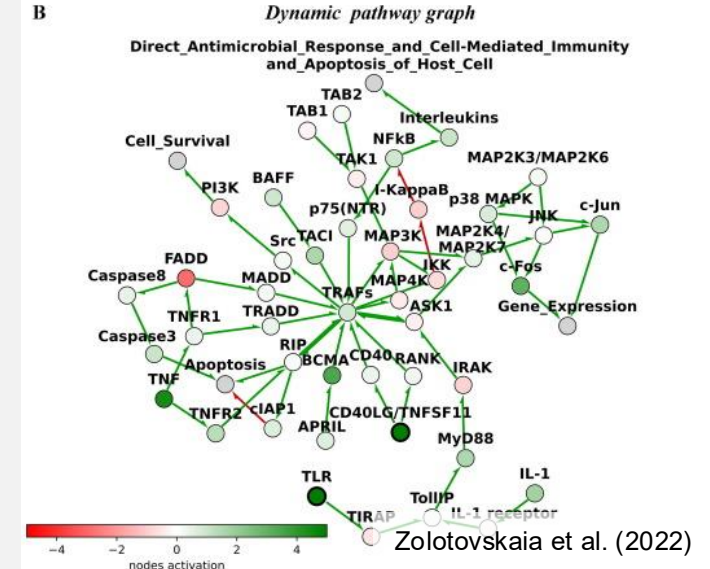
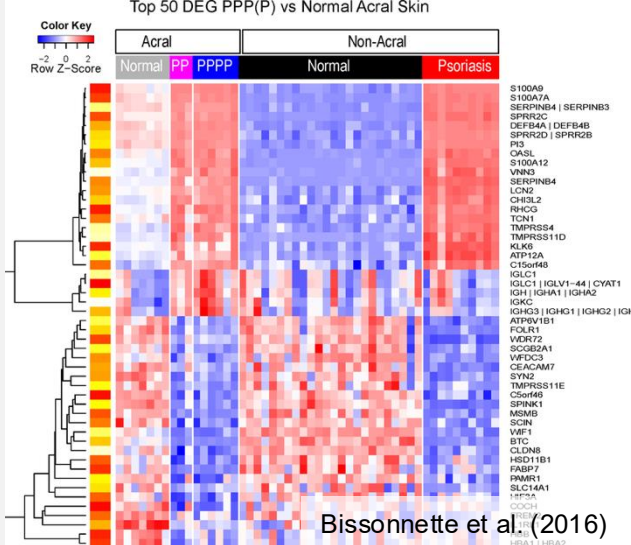
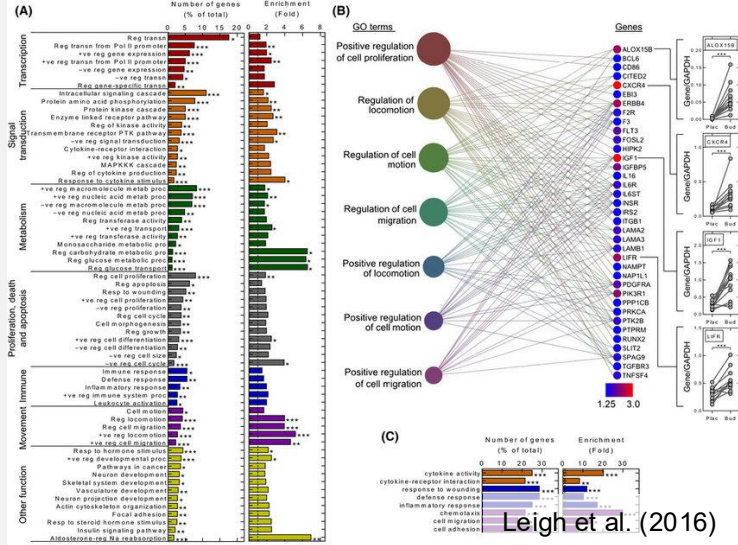
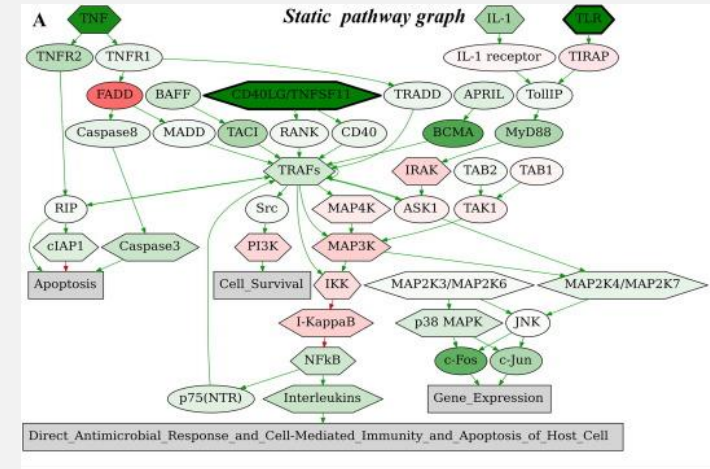
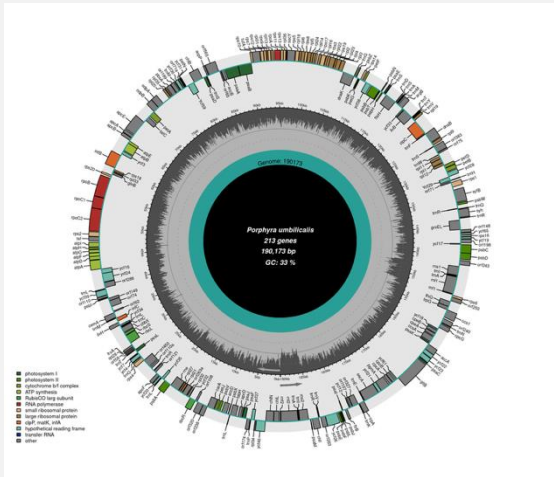
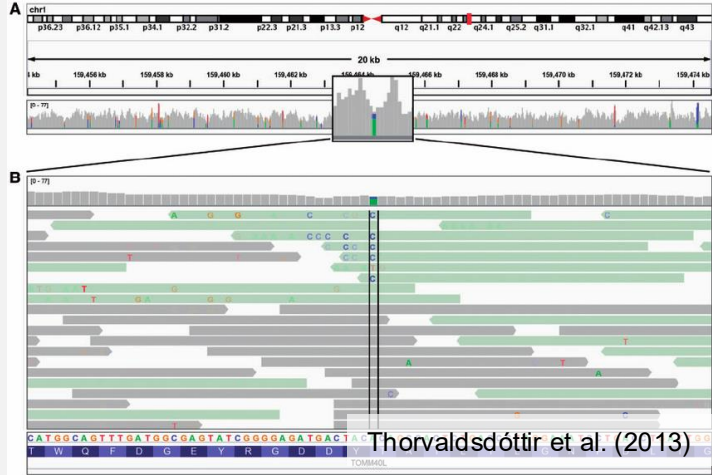
Gene Function & Processes

- Gene Ontology (GO)
- <https://geneontology.org/>

Molecular Pathways

- Network maps

VISUALIZATION TECHNIQUES FOR GENOMICS & TRANSCRIPTOMICS



ACTIVITY 5: DATA VISUALIZATION

- Design and create visualization with real genomic data
- Group discussion on best practices and interpretation

ETHICAL CONSIDERATIONS AND CHALLENGES

CHALLENGES AND LIMITATIONS

- Data Quality
- Interdisciplinary Collaboration
- Development & Optimization
- Education & Training
- Ethics & Communication

ETHICS OF HANDLING AND ANALYZING BIOLOGICAL DATA

- Ethical considerations are always important in science and statistics, especially when the data you are analyzing will be used to make individual and/or public health decisions.

AMERICAN
STATISTICAL
ASSOCIATION
ETHICAL
GUIDELINES

Ethical Guidelines for Statistical Practice

Prepared by the Committee
on Professional Ethics of the
American Statistical Association

Approved by the ASA Board in February 2022

REFLECTION & DISCUSSION: ETHICAL DILEMMAS

- Case studies on ethical dilemmas

WRAP-UP AND Q&A

KEY TAKEAWAYS

- Statistics is one tool that computational biologist use to identify patterns in biological data and rigorously test hypotheses.
- It's important and beneficial to begin a project with statistics in mind.
- Use appropriate techniques to explore and improve the quality of data.
- Similarly, use appropriate statistical tests for the experimental design and data at hand.
- Careful interpretation and visualization are key to good scientific communication.
- As someone who handles data and analyses, you have ethical responsibilities to your colleagues, the scientific community, and society at large.

QUESTIONS?

Thank you!