STATISTICAL CONCEPTS FOR BIOLOGISTS

BGGN 213: FOUNDATIONS OF BIOINFORMATICS

UNIVERSITY OF CALIFORNIA, SAN DIEGO

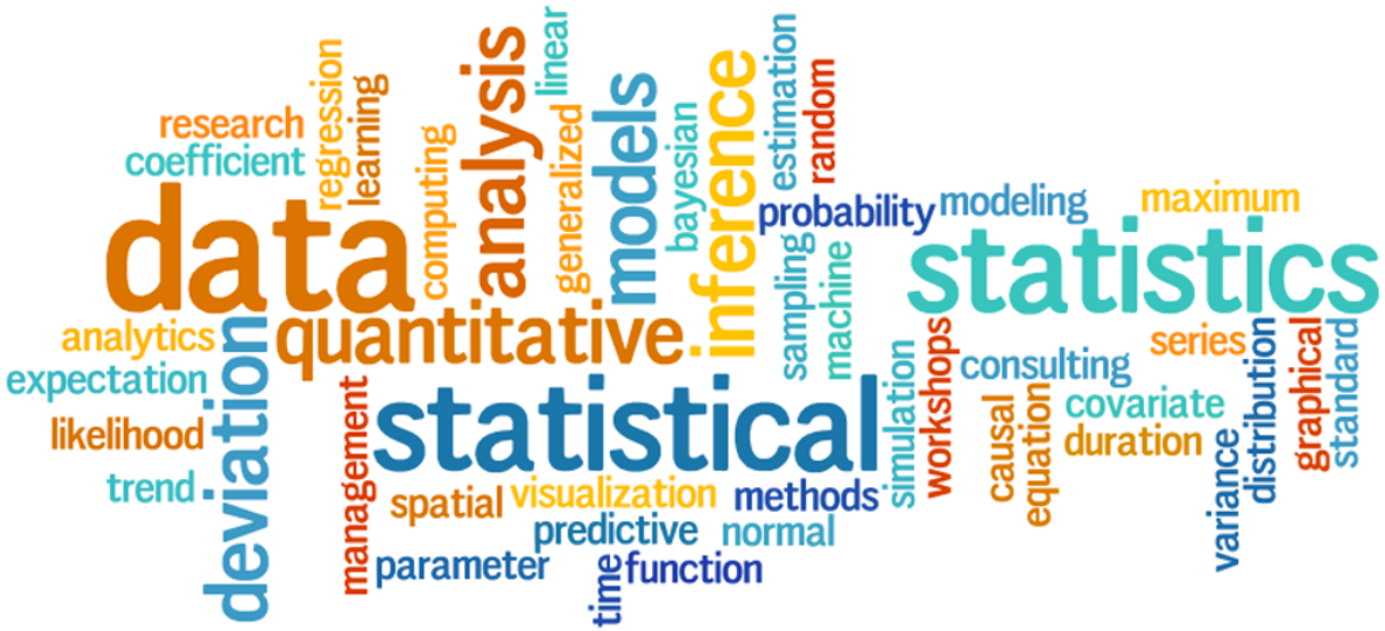
ILEENA MITRA

FALL 2017

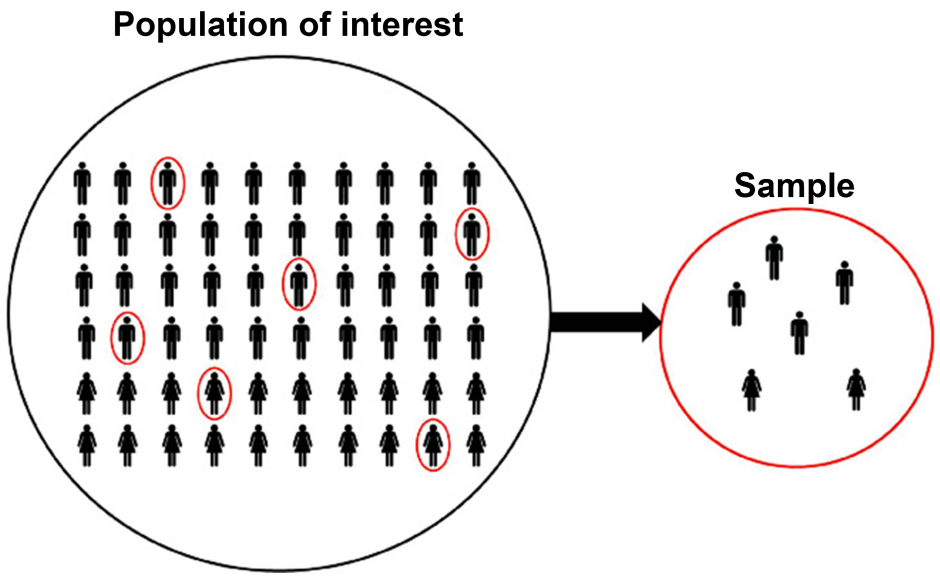
"Data don't make any sense,
we will have to resort to statistics."

OVERVIEW

1. Data summary statistics
2. Inferential statistics
3. Significance testing
4. Two sample T-test
5. Power analysis
6. Chi-square Test
7. Multiple testing correction
8. Correlation
9. Simple linear regression



DATA SUMMARY STATISTICS



- **Sample mean** = $\bar{x}$
  - mean(data)
  - summary(data) results in minimum, 1st quantile, median, mean, 3rd quantile, and max values
- **Sample standard deviation** = s

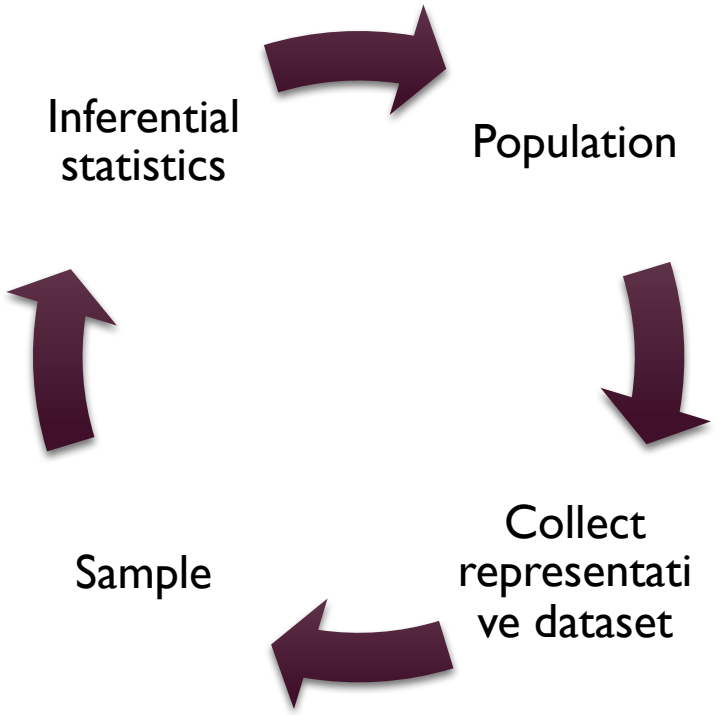
$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}$$

  - sd(data)

INFERENTIAL STATISTICS

Inferential statistics methods are used to make a generalization, estimate, prediction or decision about a population based on a sample.

- Hypothesis Testing
- Confidence Intervals
- Compare distributions
- Comparison of means
- Regression Analysis / Linear Regression



LET'S TRY THIS OUT IN R!

SIGNIFICANCE TESTING

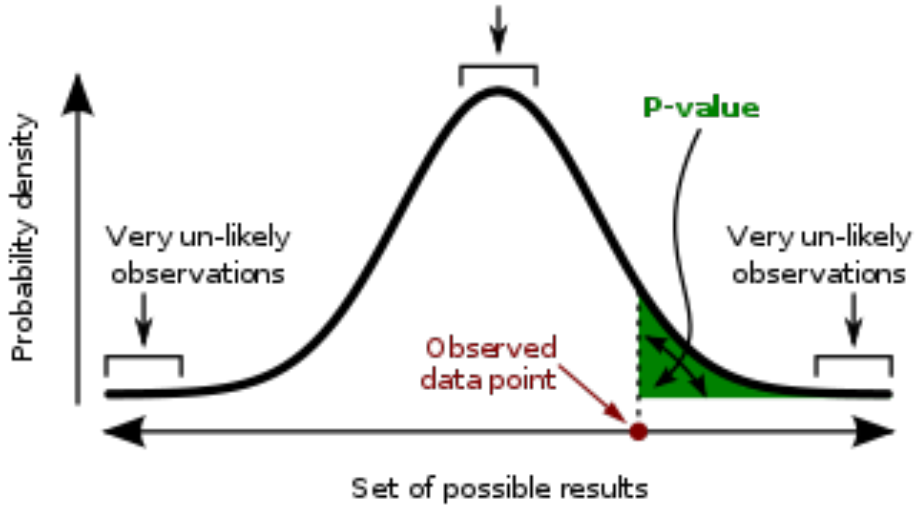
- The general idea of **hypothesis testing** involves:
  - Making an initial assumption (**null hypothesis**).
  - Collecting evidence (data).
  - Based on the available evidence (data), deciding whether to reject or not reject the initial assumption.
- The **null hypothesis (H_0)** is the default hypothesis that there is no significant difference between specified populations, any observed difference being due to sampling or experimental error.
  - A **null hypothesis** is stated, such as: "There is no difference in signal intensity for the gene expression measurements in normal and diseased samples." The alternative hypothesis is that there is a difference.

HYPOTHESIS TESTING PROCEDURE

1. State hypothesis (H_0 and H_A)
2. Set a significance level (α)
3. Determine appropriate statistical test based on assumptions met.
4. Perform calculations (R or other statistical software)
5. Draw conclusion and determine statistical significance (p-value < α)

DETERMINING SIGNIFICANCE

- We use a **test statistic** to decide whether to accept or reject the null hypothesis. For many applications, we set the **significance level** to $\alpha = 0.05$.
- We reject the **null hypothesis** and determine our results are statistically significant if the **p-value** is less than or equal to a predefined significance threshold.
- The **p-value** is the probability of obtaining a result (a test statistic) that is at least as extreme as the one observed, assuming that the null hypothesis is true.



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

TWO SAMPLE T-TEST

- Two Sample T-test: Used to determine if two population means are equal.
  - **Null Hypothesis**: The two population means are equal.
  - **Alternative Hypothesis**: The two population means are **not** equal.
- T-test test statistic:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

- t.test(x, y)

POWER ANALYSIS

- The following **four quantities** have an intimate relationship:
  - sample size
  - effect size
  - significance level (α)
  - power
- Given any three, we can determine the fourth.
- **Power** is the fraction of true positives that will be detected. It is a value between 0 and 1. The larger the sample size, the larger the power.
- R packages pwr or power.t.test(n=10, delta=1, sig.level=0.05)

CHI-SQUARE TEST

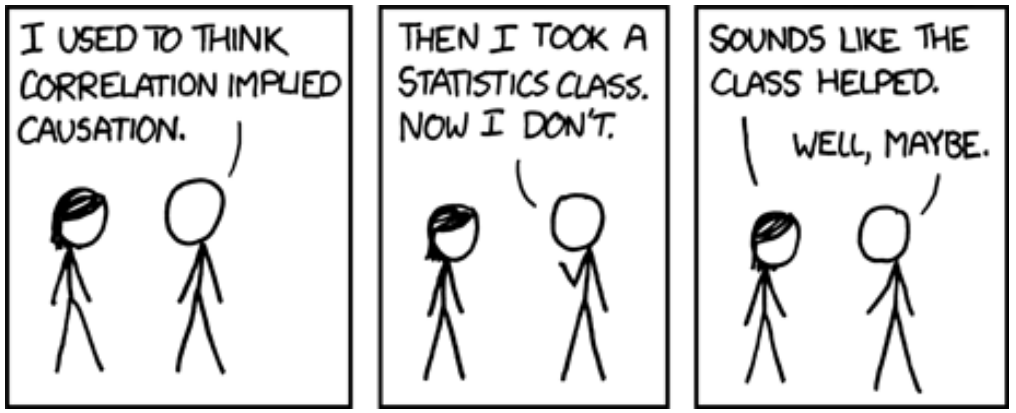
- Chi-Square Test of Independence: Used test the independence of two categorical variables.
  - **Null Hypothesis**: The two categorical variables are independent.
  - **Alternative Hypothesis**: The two categorical variables are dependent.
- Chi-square test statistic:

$$X^2 = \sum_{i=1}^{g} \frac{(O_i - E_i)^2}{E_i}$$

- chisq.test(data)

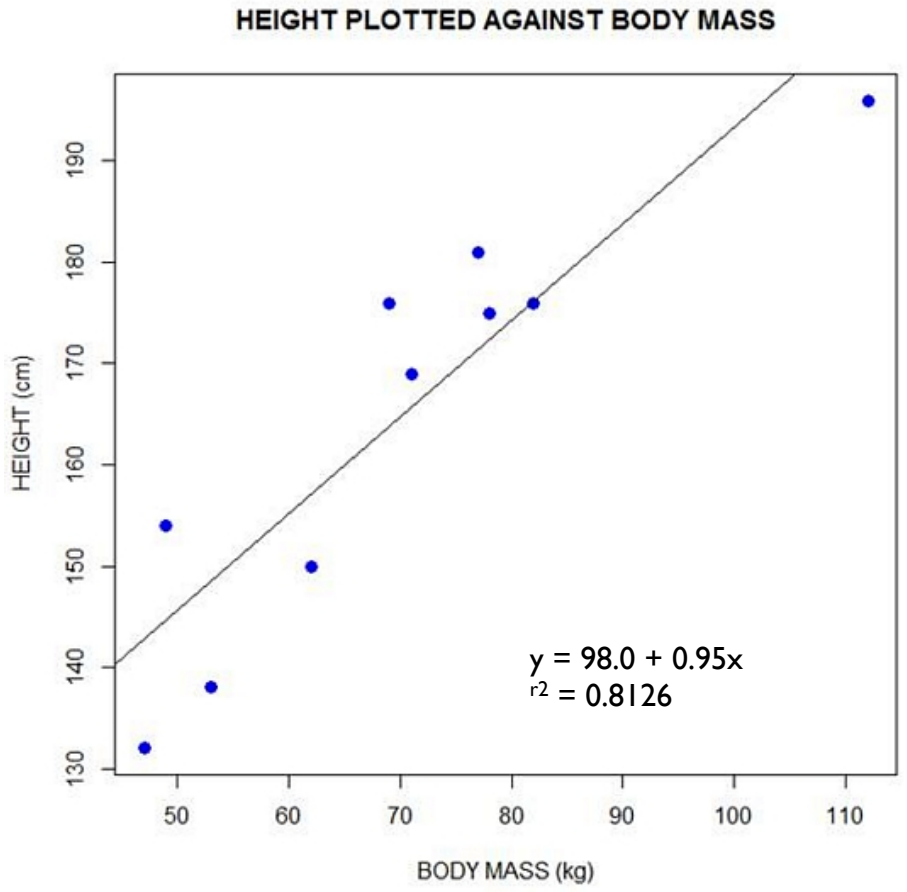
CORRELATION

- **Correlation (r)** describes the strength of association between two quantitative variables (x and y).
- **r** can range from −1 (a perfect negative correlation) to 1 (a perfect positive correlation), or can be 0 in the case of no correlation.
- cor(x, y)
- Test correlation p - value: What is the probability that random chance resulted in a correlation coefficient as far from zero as the one observed?
- cor.test(x, y)



SIMPLE LINEAR REGRESSION

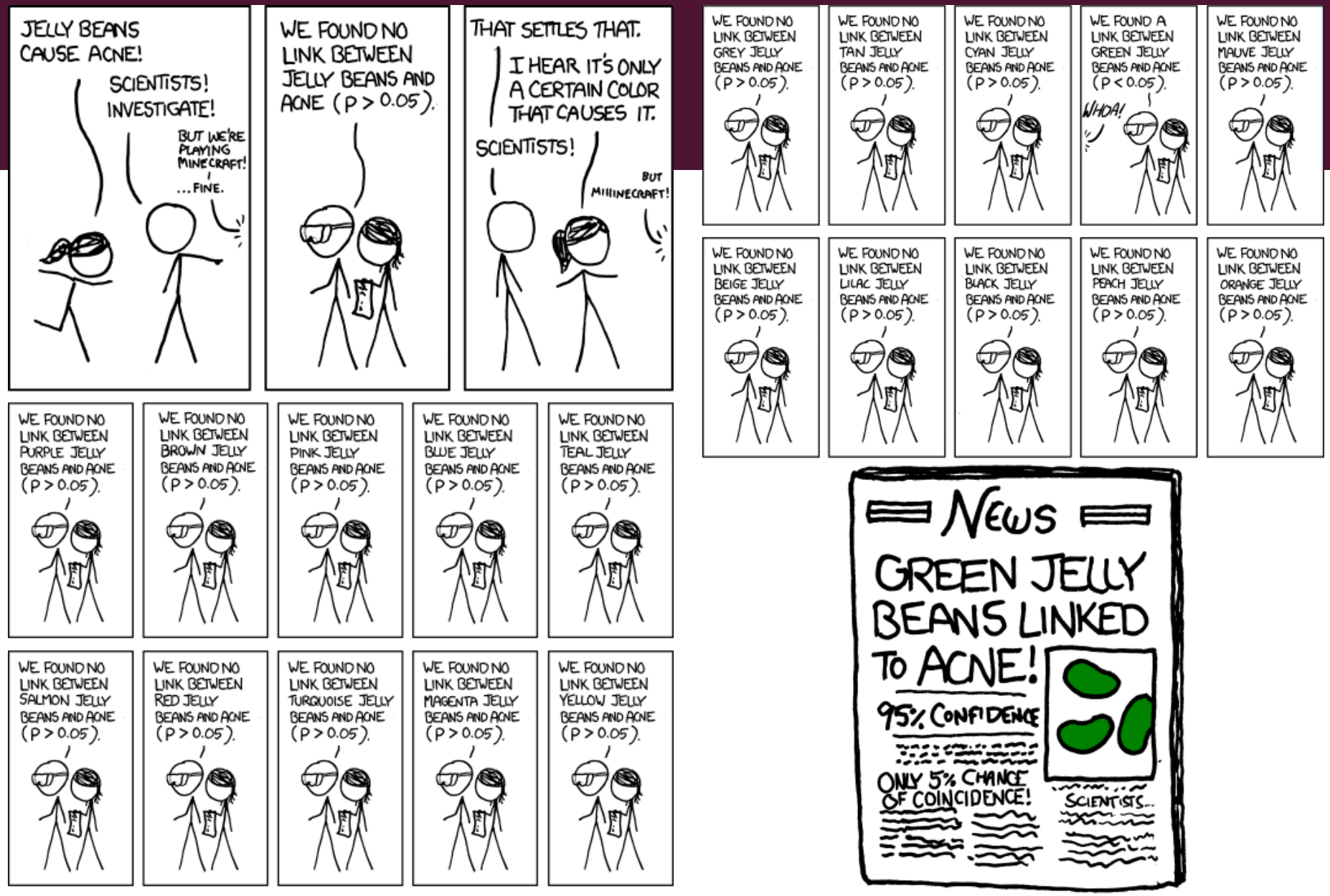
- **Linear regression** indicates the relationship between one quantitative response variable (y) and one predictor variable (x).
- lm(formula = y ~ x)
- **Linear least squares method**: The basic idea of this method is to find a straight line that best represents the trend indicated by the data, such that a roughly equal proportion of data points is observed above and below the line.



COMMON STATISTICAL TESTS

Type of Test:	Use:
Correlational	These tests look for an association between variables
Pearson correlation	Tests for the strength of the association between two continuous variables
Spearman correlation	Tests for the strength of the association between two ordinal variables (does not rely on the assumption of normal distributed data)
Chi-square	Tests for the strength of the association between two categorical variables
Comparison of Means:	***look for the difference between the means of variables***
Paired T-test	Tests for difference between two related variables
Independent T-test	Tests for difference between two independent variables
ANOVA	Tests the difference between group means after any other variance in the outcome variable is accounted for
Regression:	***assess if change in one variable predicts change in another variable***
Simple regression	Tests how change in the predictor variable predicts the level of change in the outcome variable
Multiple regression	Tests how change in the combination of two or more predictor variables predict the level of change in the outcome variable
Non-parametric:	***are used when the data does not meet assumptions required for parametric tests***
Wilcoxon rank-sum test	Tests for difference between two independent variables - takes into account magnitude and direction of difference
Wilcoxon sign-rank test	Tests for difference between two related variables - takes into account magnitude and direction of difference
Sign test	Tests if two related variables are different – ignores magnitude of change, only takes into account direction

https://cyfar.org/types-statistical-tests


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

α/n

MULTIPLE TESTING

- The **multiple testing issue** is when a large number of statistical tests are performed simultaneously on the dataset, and therefore, a number of false positive results will occur by random chance.
- **Bonferroni correction:** The simplest and most conservative approach, which sets a more stringent significance threshold for the entire set of comparisons by taking the initial significance threshold (α) an dividing by the number of tests performed (n).
  - α/n
- **False discovery rate correction**: A more complicated and less stringent method, which controls the probability of *at least one* "false discovery".
- p.adjust(data_pvals, method = p.adjust.methods)
- p.adjust.methods are one of c("holm", "hochberg", "hommel", "bonferroni", "BH", "BY", "fdr", "none")

FOLLOW UP

- **Introduction to Data:** https://www.datacamp.com/courses/introduction-to-data
- **Correlation and Regression** https://www.datacamp.com/courses/correlation-and-regression
- **Statistical Modeling in R (Part 1)** https://www.datacamp.com/courses/statistical-modeling-in-r-part-1
- **Statistical Modeling in R (Part 2)** https://www.datacamp.com/courses/statistical-modeling-in-r-part-2
- **Data Visualization with ggplot2 (Part 2)** https://www.datacamp.com/courses/data-visualization-with-ggplot2-2