Workshop #3.5 - Advanced Data Analysis

Presented by the SLB Outreach Committee

SLB Introductions!

Some PSAs:

- Safety screening deadline is Dec 31st!
  - Fair updates are on the website
- Follow us on instagram @gsdsef for updates, new workshops, etc
- Follow our youtube channel for workshop replays and safety/screening/applying tutorials

Agenda

- What is Data Analysis? Recap
- Analyses & when to use them:
  - Error Bars, Box & Whiskers
  - Linear Regression (Line of best fit)
  - Statistical Tests
    - Intro to statistical tests
    - R-squared (Student's T-test)
    - Chi-squared
  - Confusion Matrices & ROC curves

Findings: Analyzing the Data

1. **Organize** the data/results in charts, tables, and graphs
2. **Review and Interpret** the Data/Results – Do the Math!
3. **Summarize/Discuss** the Data/Results

Findings: Organize the Results

- Raw, numerical data→ tables
- Qualitative data - organized w/ captions(e.g. photos, description)
- All data goes into the appendix

Tip: Don't leave anything out or skip any information. Some of the best science discoveries come from our "mistakes."

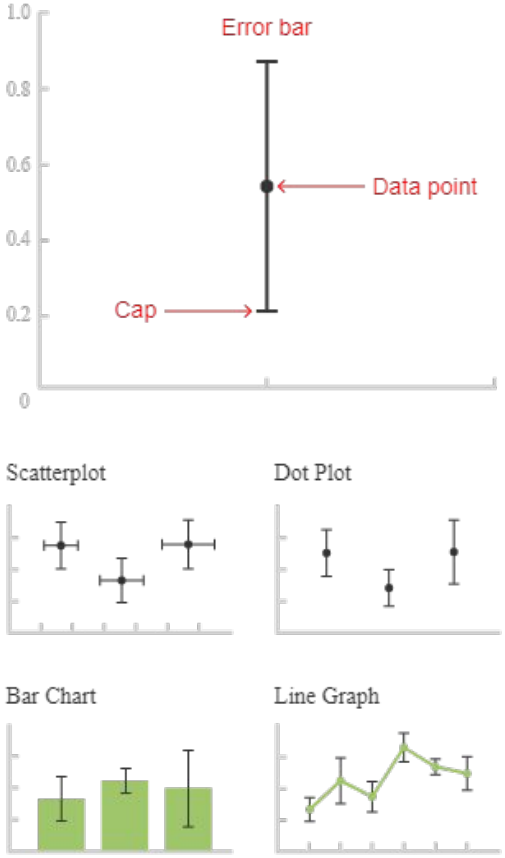
What analysis, when?

- Not all analyses are meant for every project
- Make sure to check your sources - use the methods you find in papers
- If you're unsure, ask us!

Analyses- When to Use What

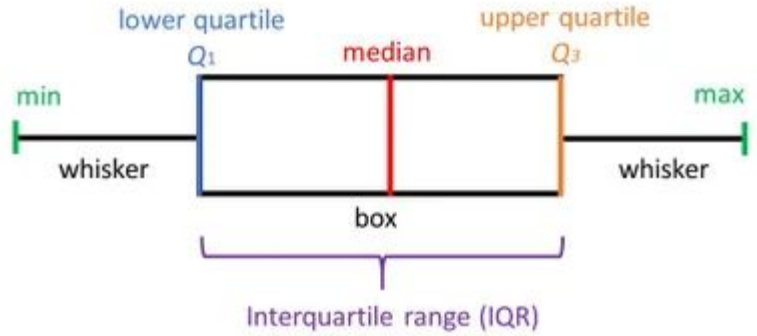
Error Bars

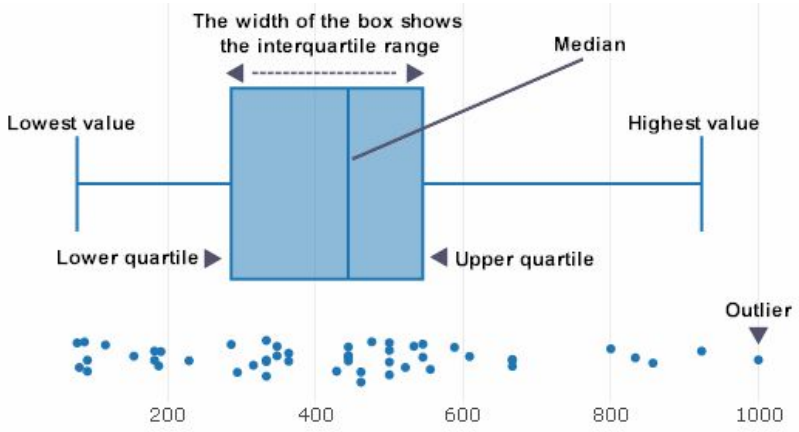
- Every data point has some error, every average value represents a range
- An error bar can help show how reliable or precise your data is
- Smaller error → more reliable or precise
- Standard deviation, or a set error value



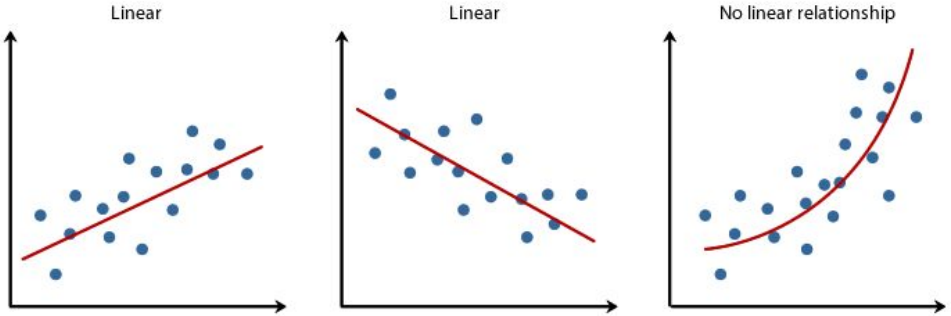
Box & Whisker Plots

- Representing a distribution
- Contains:
  - Min & Max points
  - Median (middle of data)
  - Quartiles (25% - 75% of data)
- The length of the box and distances of min/max show how spread out the data is

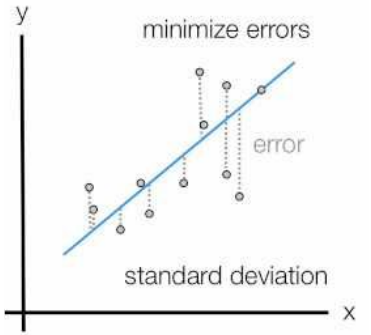




Linear Regression (line of best fit)



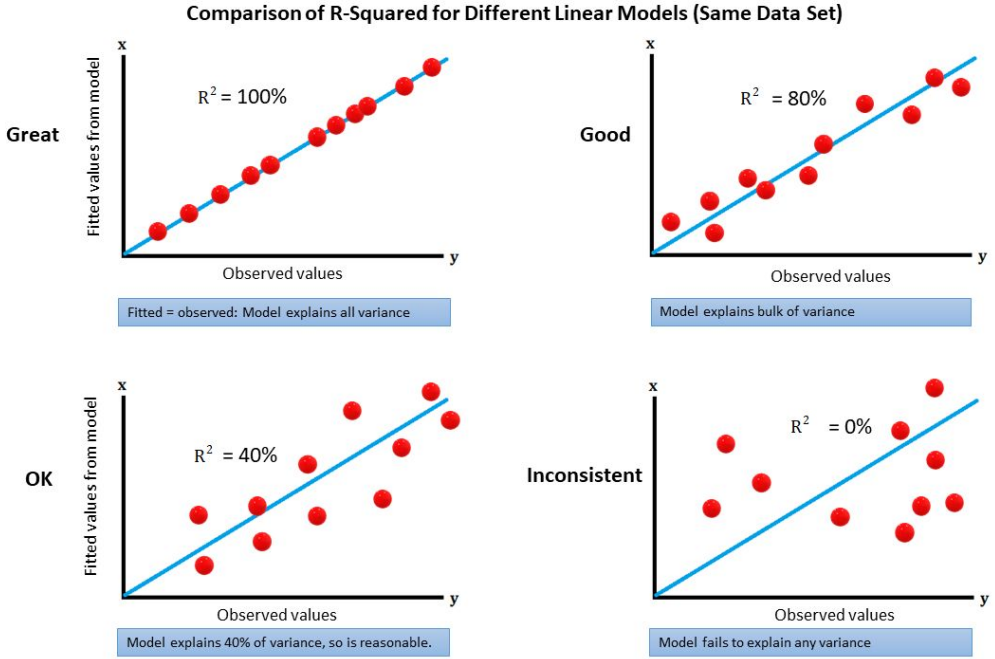
- For scatter plots
- Is there a linear relationship between x & y?
- Can be used to extrapolate new values from data



- If X↑ does y↑? (positive correlation) or
- If X↓ does y↑? (negative or inverse correlation)
- Reasoning: trying to minimize distance(error) from each point to the predicted line

R^2 (coefficient of determination) - How good is the line of best fit?

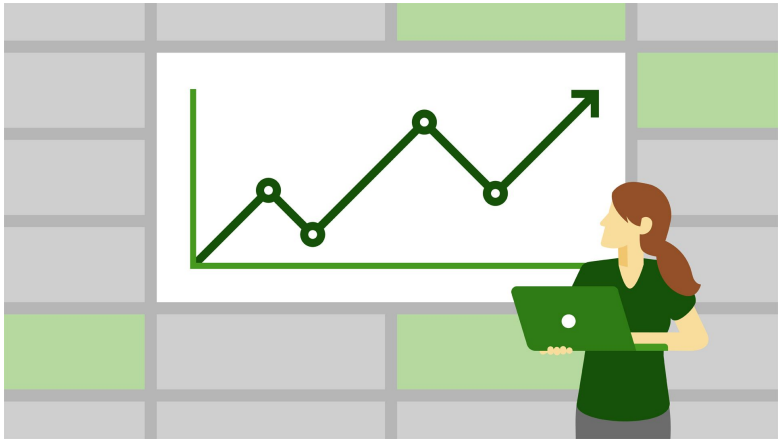
- A value to tell how well your model represents variance in the data
- R^2 of 1 means that your line goes through every point



Statistical Tests

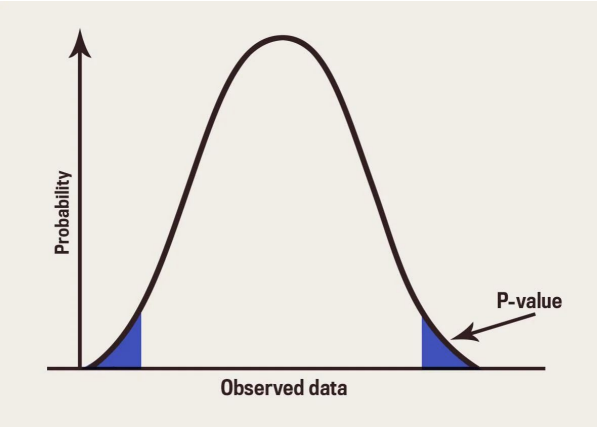
What is a Statistical Test?

- Is your result significant?
- Are you actually seeing a trend?
- Is your data real? Or is it all just noise?
- Examples of these include:
  - Probability values (t-test, chi-squared, z-test)
  - Correlation coefficient
  - Matthew's coefficient



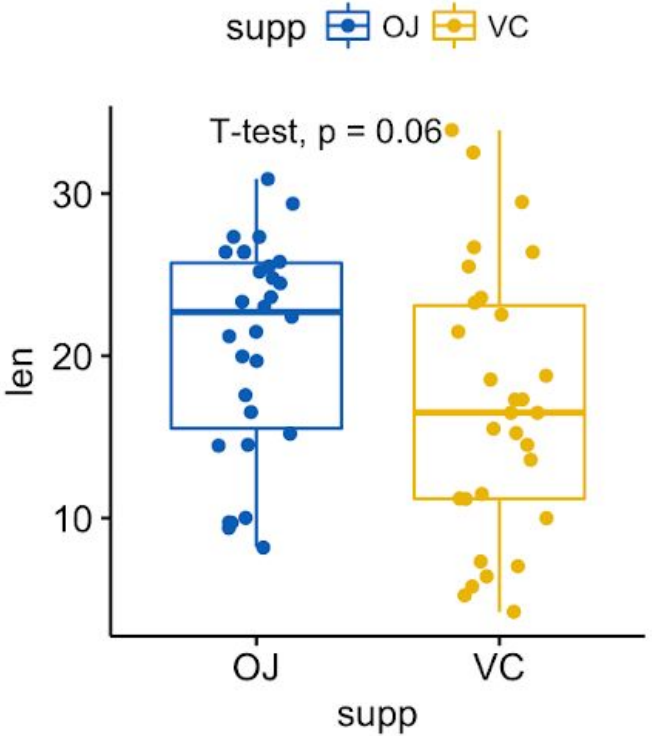
Some vocabulary:

- Null hypothesis - the default hypothesis that there is no difference between the two variables
- Alternate hypothesis - the hypothesis that one variable affects the other
- P-value - probability value, the probability that the results happened randomly



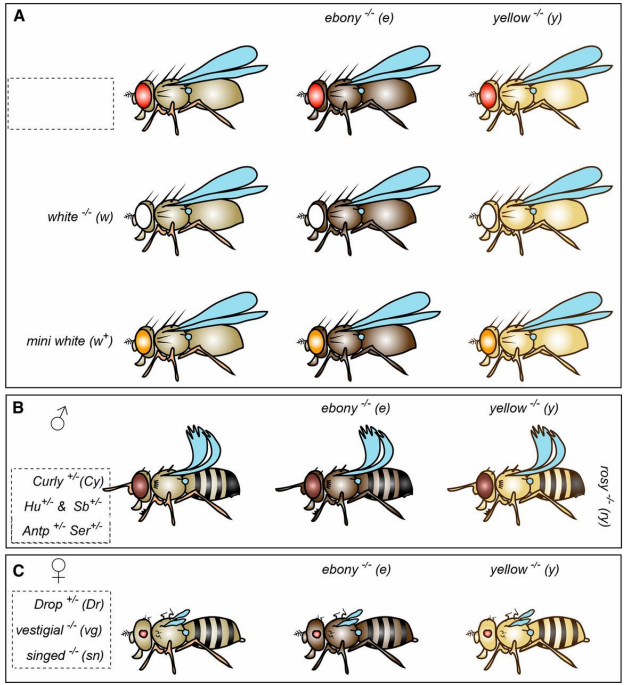
Student’s T-test

- Common test for producing p-values
- Determining probability of result being random
- Testing numerical data for categories
- A decimal between 0-1
    - Critical value @ 0.05

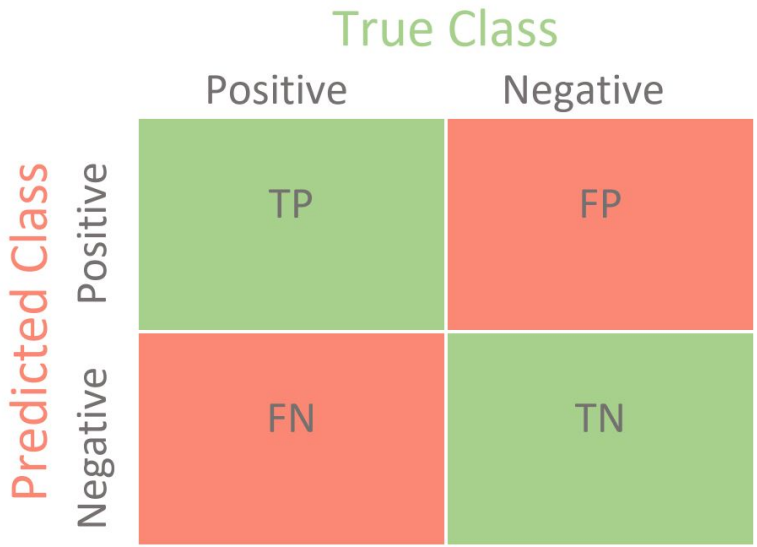


Chi-squared

- Used with purely categorical data, large sample size (30+)
- Typically for genomic data, confirming genotype predictions
- Comparing observations to expected value
- Same critical value of 5%



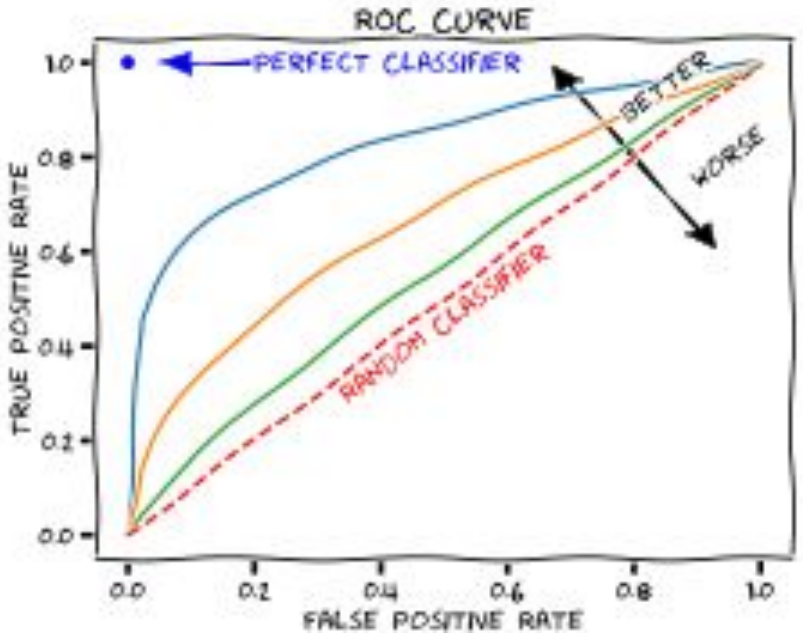
Confusion Matrix



- Machine learning & classifier projects
- Separating out prediction types for a model
- Can calculate:
  - Sensitivity (TP rate)
  - Specificity (TN rate)
  - Accuracy (True rate)
  - Precision

ROC curves

- Varying the threshold to determine optimal
- True positive rate vs. false positive rate
- Area under the curve tells how good of a classifier it is



Thanks for your time!

Next Workshop:
Workshop #4 - Abstract, Screening, Slides
(Jan. 16, 9-10 am)

Breakout Rooms (please rename yourself with the number, e.g. 3 Tony Stark, or select the breakout room):

1 - Statistical Tests in Microsoft Excel (Jessie)
2 - Statistical Tests in Google Sheets (Andrew)
3 - Life sciences (Eleanor)
4 - Physical Sciences (Emily)