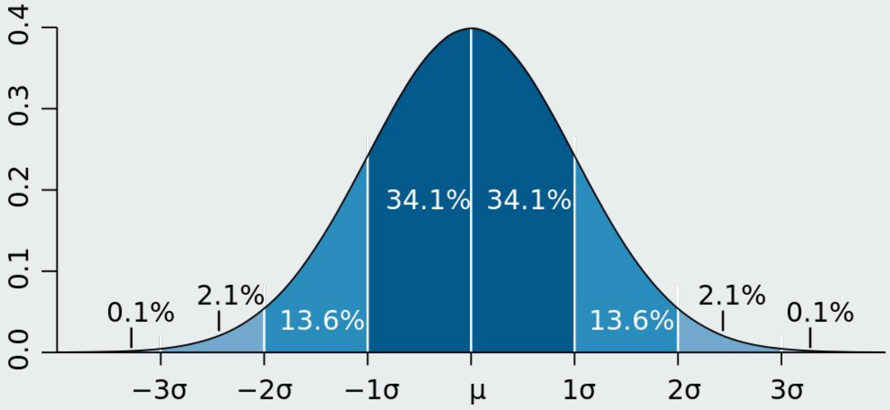
Statistical Inference



Maureen Pittman
PUBS Tech Talk
9 October 2018

Overview

- Statistics and statistical inference
- Hypothesis testing
    - Parametric
    - Non-parametric
- Regression analysis
- Data visualization
- Resources

Descriptive vs Inferential Statistics

Descriptive

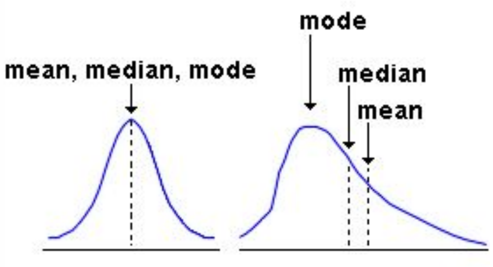
concerned with the properties of observed data

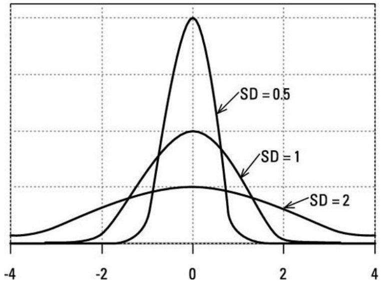
Central tendency

- mean
- median
- mode

Dispersion

- variance
- standard deviation





Inferential

comparing/deducing properties from a sample

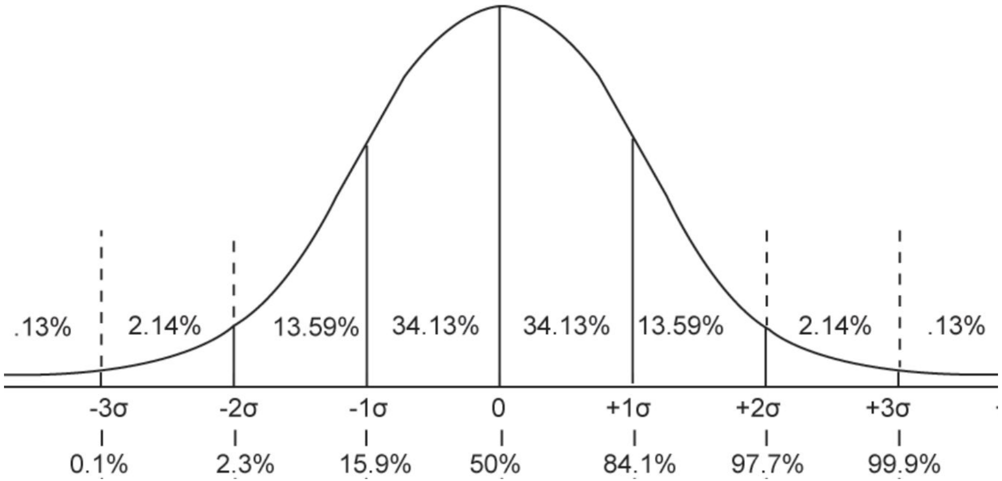
Examples:

- Hypothesis testing
  - Is the mean of Group A significantly different from Group B?
  - Is the distribution of Group A significantly different from Group B?
- Regression analysis
  - What is the statistical relationship between two variables?

Hypothesis Testing

H_0 : there is no relationship between the two variables

H_1 : the variables are associated



P-value definitions:

- the probability of seeing a result as extreme or more extreme than the one observed (if H_0 were true)
- the probability of rejecting H_0 when it is true.

P-value cutoff (also called α, often set to 0.05): the level of uncertainty acceptable to reject H_0

Parametric vs Non-Parametric Tests

Parametric

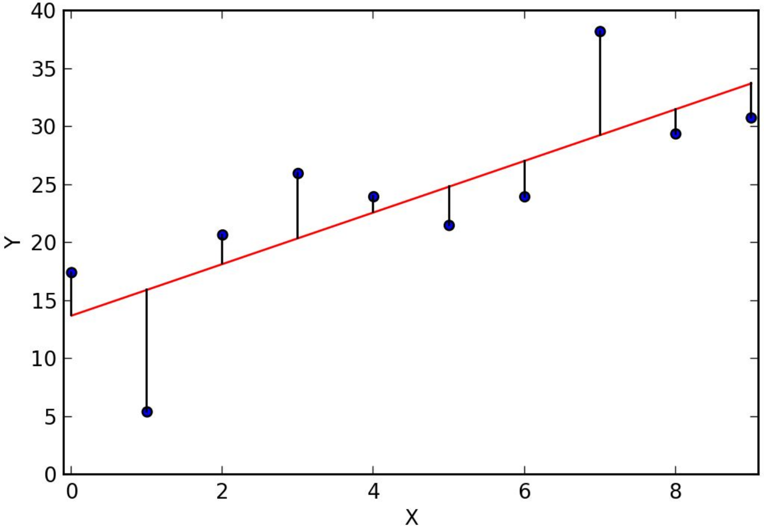
- Make assumptions about the underlying properties of the data
- Examples:
    - T-test/Z-test (assumption: normality)
    - Pearson Correlation (assumption: linear)
    - ANOVA (assumption: F-distribution)

Non-Parametric

- No assumptions about the underlying properties of the data
- Examples:
    - Mann-Whitney-U
    - Spearman’s Correlation
    - Kruskal-Wallis

Regression Analysis

- Examine the relationship between two variables of interest
- Linear (least-squares) regression
  - R-squared value: how well the model fits the data
  - Assumptions
  - Transformations

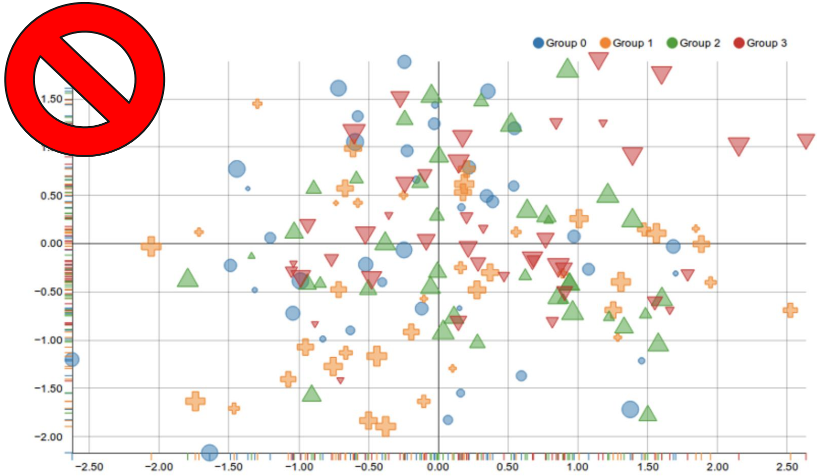


Y = mX + b

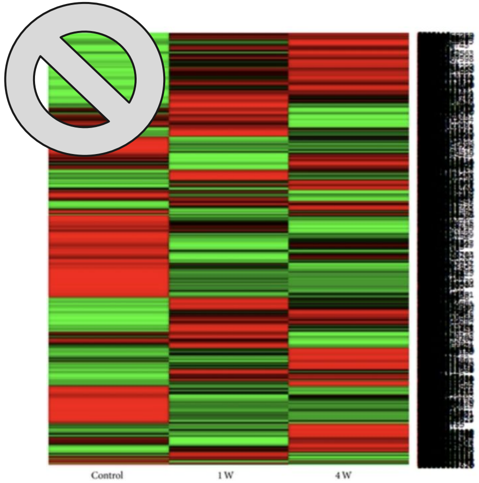
R^2 = Explained variation / Total variation

Data Visualization - Basics

- Clearly label plots, axes, and legends
- Avoid making plots too busy
- Use colorblind-friendly palettes



https://ldld.samizdat.cc/2016/scatter-plot/

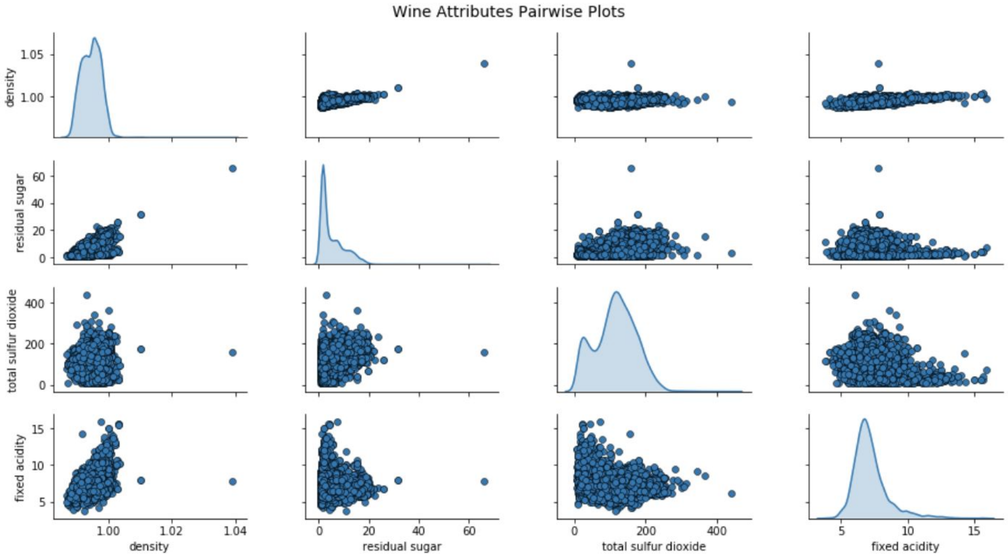


Nahm et al, 2015. BioMed research international.

Data Visualization - High Dimensional Data

Pairwise scatterplots:

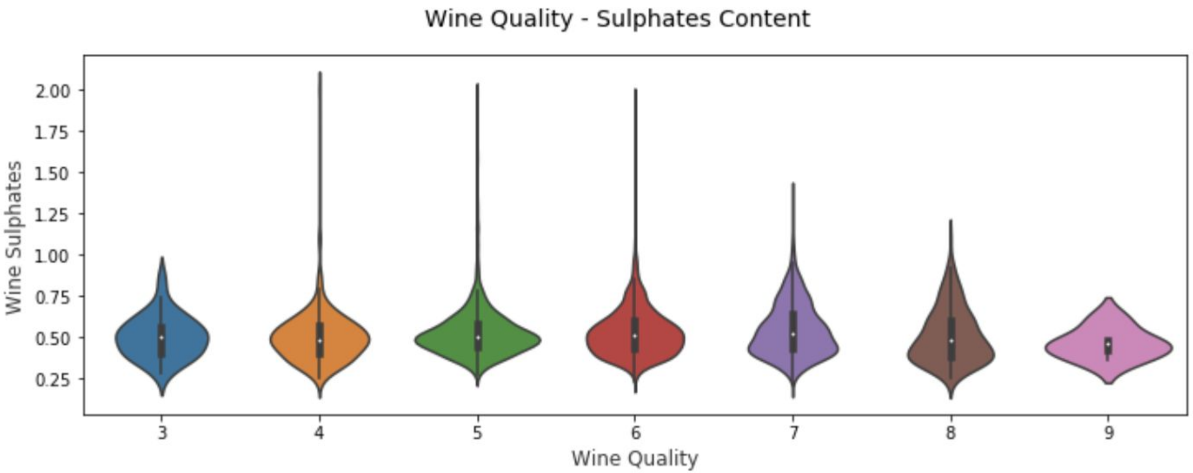
examine the relationships between each possible pairwise combination of variables



Visualizing two-dimensional data with pair-wise scatter plots

https://towardsdatascience.com/the-art-of-effective-visualization-of-multi-dimensional-data-6c7202990c57

Data Visualization - High Dimensional Data



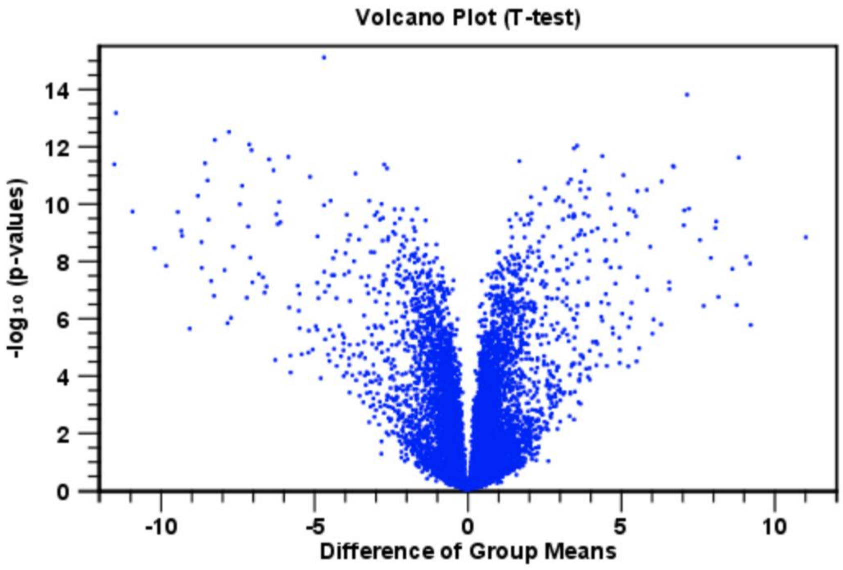
Violin Plots as an effective representation of two-dimensional mixed attributes

Violin plots:

examine the probability density of a continuous variable at different categorical values.

https://towardsdatascience.com/the-art-of-effective-visualization-of-multi-dimensional-data-6c7202990c57

Data Visualization - High Dimensional Data



Volcano plots:

visualize the magnitude and p-value significance of a change or difference between two groups

http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/650/index.php?manual=Transcriptomics.html

Resources

UCLA Institute for Digital Research and Education

- What statistical analysis should I use?
- Choosing the correct statistical test

Cross-Validated
https://stats.stackexchange.com/

UCSF Introduction to Biostatistics by David Quigley