BST281: Genomic Data Manipulation, Spring 2019

Monday 09: Descriptive statistics

Descriptive statistics summarize information that is already present in a data collection

Plots (box plots, histograms, etc.), measures like average, standard deviation, medians, etc.

Inferential statistics use a sample of data to make predictions about larger populations or about unobserved/future trends

Measurements in the presence of noise, generalizations from a sample to a population, comparisons between datasets: correlations, regression, etc.

A statistic is any single value that summarizes an entire dataset

Parametric summary statistics describe “well-behaved” data, often normally-distributed

Mean, standard deviation, z-scores, etc.

Nonparametric summary statistics can be used to describe data regardless of distribution

Median, percentiles, quartiles, interquartile range, etc.

Summary statistics of paired data allow comparison between datasets

Distances quantify differences between datasets; larger means less similar

Euclidean (L2 norm), Manhattan (L1 norm)

Correlations quantify the similarity of datasets; larger means more similar

Pearson, Cosine (parametric), Spearman (nonparametric)

Beware of summary statistics: Understand (and visualize) your data before summarizing them!

Anscombe’s quartet: a four manually-crafted pairs of datasets with equal means, standard deviations, and Pearson correlations, yet completely different relationships

Basics of probability

Thinking in terms of the probability of sets of outcomes (events) occurring as subsets of the set of all possible outcomes (sample space)

Kolmogorov axioms: one definition of probability that matches reality

Bayes’ theorem: Calculating a conditional probability based on the inverse of its condition

# Textbooks

Basic definitions: Pagano and Gauvreau, Chapters 2.1, 2.3-4

Mean, median, etc.: Pagano and Gauvreau, Chapters 3.1-2

Correlation: Pagano and Gauvreau, Chapters 17.1-3

Probability: Pagano and Gauvreau, Chapters 6.1-3

# Literature

[Tackling the widespread and critical impact of batch effects in high-throughput data. Leek, Nature Rev. Genetics 2010](https://www.ncbi.nlm.nih.gov/pubmed/20838408)

[Ten Simple Rules for Effective Statistical Practice. Kass, PLoS CB 2016](https://www.ncbi.nlm.nih.gov/pubmed/27281180)