Recap

• (Ch 1) Visualizing data
  • Tables
  • Bar charts
  • Histograms
  • Conditional histograms

• (Ch 1) Summarizing data
  • Mean
  • Standard deviation
  • Variance
Today

• (Ch 1) More on visualizing and summarizing data
  • Standardizing data to look at its shape
  • Median and interquartile range
  • Box plots and outliers
  • Modes and skew

• (Ch 2) Visualizing and summarizing relationships in data
  • Time series data
  • Scatter plots
  • Correlation coefficient
Standard coordinates

• The mean tells where the data set is, and the standard deviation tells us how spread out it is, but what about its shape?
• Standardizing the data set shifts its mean to 0 and scales its standard deviation to 1.
• Given a data set \{x\}, we standardize it to the data set \{\hat{x}\} as follows:

• We say \{\hat{x}\} is in standard coordinates
Median

• If there are an odd number of items,

• If there are an even number of items,

• The median is also known as the 50th percentile
Properties of the median

• Scaling data scales the median

• Translating data translates the median
Interquartile range

• Scaling data scales the interquartile range

• Translating the data does not change the interquartile range
Box plots

- Boxplots are useful in identifying outliers, but you have to be able to justify discarding those outliers from your data set
Sensitivity of summary statistics to outliers

- Mean and standard deviation are very sensitive to outliers
- Median and interquartile range are not sensitive to outliers
Modes

- Modes are peaks in a histogram
- If there is more than 1 mode, we should be curious as to why
Tails and skew

Symmetric Histogram

mode, median, mean, all on top of one another

Left Skew

Right Skew
Ch 2: Looking at relationships in data

Python example: Stock prices of Fedex and UPS
  • Time series data
  • Standardization
  • Scatter plots
  • The correlation coefficient
Correlation, a rough idea

In a data set \( \{(x, y)\} \) consisting of items \((x_1, y_1), ..., (x_N, y_N)\)

- we say \( x \) and \( y \) have **positive correlation** if
  - small \( x \) and small \( y \) tend to occur together
  - large \( x \) and large \( y \) tend to occur together

- we say \( x \) and \( y \) have **negative correlation** if
  - small \( x \) and large \( y \) tend to occur together
  - large \( x \) and small \( y \) tend to occur together

- we say \( x \) and \( y \) have **zero correlation** if
  - there is no tendency for \( x \) and \( y \) to be small together or large together
Correlation examples

• Lines of code in a codebase and number of bugs?

• Body temperature and height?

• GPA and hours spent playing video games?

• Earnings and happiness?
Correlation does not imply causation!

Sleeping with your shoes on is positively correlated with waking up with a headache.

Does sleeping with your shoes on cause headaches?