Recap

• (Ch 1) Visualizing and summarizing data
  • Tables, bar charts, histograms, conditional histograms
  • Mean, standard deviation and variance

Today

• (Ch 1) Visualizing and summarizing data
  • Standardizing data to look at its shape
  • Median, interquartile range, box plots and outliers
• (Ch 2) Visualizing and summarizing relationships in 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:

\[
\hat{x}_i = \frac{x_i - \text{mean}(\{x\})}{\text{std}(\{x\})}
\]

- We say \( \{\hat{x}\} \) is in standard coordinates
Median

• If there are an odd number of items,

  median = middle item when sorted

• If there are an even number of items,

  median = mean of middle 2 items when sorted

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

• Scaling data scales the median

\[ \text{median}(\{kx_i\}) = k \text{median}(\{x_i\}) \]

• Translating data translates the median

\[ \text{median}(\{x_i + c\}) = \text{median}(\{x_i\}) + c \]
Interquartile range

\[ \text{iqr} = (75\text{th percentile}) - (25\text{th percentile}) \]

- Scaling data scales the interquartile range

\[ \text{iqr}({kx_1}) = |k|\text{iqr}({x_i}) \]

- Translating the data does **not** change the interquartile range

\[ \text{iqr}({x_i + c}) = \text{iqr}({x_i}) \]
Box plots are useful in identifying outliers, but you should think about how 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
Looking at relationships in data

Jupyter notebook 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 \( (x_1, y_1), \ldots, (x_N, y_N) \)

- 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? positive
- Body temperature and height? zero
- GPA and hours spent playing video games? negative
- Earnings and happiness? who knows
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?