CSCI 3022-002 Intro to Data Science
Visual Exploratory Data Analysis

Opening Zoom Example: *Calculate the Mean and Standard Deviation of the data set:*
Data (units in dollars): 2,4,3,5,6,4.

Opening Sol'n

we estimated

Example: *Calculation of the SD*

Data (units in dollars): 2,4,3,5,6,4.

$$\bar{X} = \frac{\sum_{i=1}^{6} X_i}{6} = \frac{2+4+3+5+6+4}{6} = \frac{24}{6} = \boxed{4}$$

$$s^2 = \frac{\sum_{i=1}^{6} (X_i - \boxed{4})^2}{6-1} = \frac{(2-4)^2+(4-4)^2+(3-4)^2+(5-4)^2+(6-4)^2+(4-4)^2}{5}$$

$$= \frac{2^2+1^2+2^2+1^2}{5} = \frac{10}{5} = 2$$

$$s = \sqrt{s^2} = \sqrt{2}$$

Opening Sol'n

Example: *Calculation of the SD*
Data (units in dollars): 2,4,3,5,6,4.

Since we mean business, we need the average first.

$$\bar{X} = \frac{2+4+3+5+6+4}{6} = \frac{24}{6} = 4$$

Opening Sol'n

Example: *Calculation of the SD*

Data (units in dollars): 2,4,3,5,6,4.

Since we mean business, we need the average first.

$$\bar{X} = \frac{2+4+3+5+6+4}{6} = \frac{24}{6} = 4$$

Now let's compute the deviations...

$$\overbrace{[(X-\bar{X})^2]}^{\text{vectorized deviations}} = [(2-4)^2, (4-4)^2, (3-4)^2, (5-4)^2, (6-4)^2, (4-4)^2]$$

Opening Sol'n

Example: *Calculation of the SD*

Data (units in dollars): 2,4,3,5,6,4.

Since we mean business, we need the average first.

$$\bar{X} = \frac{2+4+3+5+6+4}{6} = \frac{24}{6} = 4$$

Now let's compute the deviations...

$$\overbrace{[(X-\bar{X})^2]}^{\text{vectorized deviations}} = [(2-4)^2, (4-4)^2, (3-4)^2, (5-4)^2, (6-4)^2, (4-4)^2]$$

and sum and "average" those!

$$s^2 = \frac{4+0+1+1+4+0}{5} = 2$$

Announcements and To-Dos

Announcements:

1. HW 1 Posted, due Monday!
2. Another nb day this Friday!

Last time we learned:

1. Loading, beginning to manipulate data in Python.

To do:

1. Start that HW! ~~Ensure you can load the data~~ and work with it. Practice your TeX/markdowns!

Weekly recap: common questions mostly were coding based: lambda and axis

axis = 0 "rows"

axis=1 "columns"

def funct.

Last Time Recap:

We talked about two big types of measures for a data set $X_1 \dots X_n$: centrality and dispersion.

Measures	Stat	Calculation	Advantages
Centrality	Mean	$\frac{\sum_{i=1}^n X_i}{n}$	Uses all data, behaves "nicely"
	Median	middle value	not pulled by single outliers
	Mode	most common value	'indicative' of true data value
Dispersion	Variance	$\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$	Squared distance can be nice, no radical
	SD	$\sqrt{s^2}$	*units* same as data
	Range	Max minus min	shows extremes
	IQR	Q_3-Q_1	like median: avoids outliers

"interquartile range"

Means and Medians:

One way we conceptualize the mean and the median as data scientists is we ask the question: "what single number is **closest** to our data." This requires us choose a definition of distance: squared or absolute?

We prove: Show that the *sample mean* of data $X_1, X_2, \ldots X_n$ is the unique minimizer c of the function

$$f(c) = \sum_{i=1}^{n} (X_i - c)^2$$

squared distance from "c" to each datum

NB: The *median* of data $X_1, X_2, \ldots X_n$ is the possibly non-unique minimizer c of the function

$$f(c) = \sum_{i=1}^{n} |X_i - c|$$

absolute distance from "c" to each datum.

$X_1 = 0$
$x_2 = 1$

choose $c = .5$
dist: $.5 + .5 = 1$
choose $c = .1$
dist: $.1 + .9 = 1$

Proof: Differentiating yields

Opening

$$f(c) = \sum_{i=1}^{n} (x_i - c)^2$$

we input/choose "c"

$\uparrow d/dc.$

$$\frac{d}{dc} f(c) = \frac{d}{dc} \sum_{i=1}^{n} (x_i - c)^2$$

$$= \sum_{i=1}^{n} -2(x_i - c)$$

$$\frac{d}{dx}\left[(2-x)^2 + (1-x)^2 + (0-x)^2\right]$$

$$\frac{d}{dx}\left((2-x)^{\boxed{2}}\right) = 2(2-x) \cdot \frac{d}{dx}(2-x)$$

$$= 2(2-x)(-1)$$

$-2(\Sigma x_i)$ $\quad 2cn$

$$\begin{aligned} &-2x_1 + 2c \\ &+ -2x_2 + 2c \\ &+ -2x_3 + 2c \\ &\vdots \\ &+ -2x_n + 2c \end{aligned}$$

set $= 0$

$$0 = \sum_{i=1}^{n} (-2x_i + 2c)$$

Proof: Differentiating yields

$$f'(c) = \frac{df}{dc} \sum_{i=1}^{n} (X_i - c)^2 = \sum_{i=1}^{n} -2(X_i - c).$$

Setting $f'(c) = 0$ gives

at: "find the value of c so that"

$$0 = \cancel{-2}\left(\sum_{i=1}^{n} X_i\right) + \cancel{2}nc$$

divide 2

$$\sum_{i=1}^{n} X_i = nc \Rightarrow c = \frac{\sum_{i=1}^{n} X_i}{n} = \overline{X}.$$

Proof: Differentiating yields

$$f'(c) = \frac{df}{dc} \sum_{i=1}^{n} (X_i - c)^2 = \sum_{i=1}^{n} -2(X_i - c).$$

Setting $f'(c) = 0$ gives

$$\begin{aligned} 0 &= \sum_{i=1}^{n} -2(X_i - c) \\ &= 2nc - 2\sum_{i=1}^{n} X_i \\ \implies \qquad c &= \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X} \end{aligned}$$

The Interquartile Range

The interquartile range is defined to be the difference between the upper and lower quartiles:

$$IQR = Q_3 - Q_1$$

+4.5

It's a spread measure standardly used in *box plots*, which we introduce formally next time.

Tukey's Five Number Summary

option 1: (mean, sd) or $(\bar{x}, s)$

John Tukey, father of modern EDA, advocated summarizing data sets with 5 values:

1. Min value
2. Lower quartile
3. Median
4. Upper quartile
5. Max value

range

IQR

Centrality

Advantages:

- gives the center of the data
- gives the spread of the data (range in IQR)
- gives an idea of skewness (compare how far away Q1 and Q3 are from median!)

See the PD.DESCRIBE() method from pandas!

Histograms

Definition: A *histogram* is a graphical representation of the distribution of numerical data.

To construct a histogram:

"Bin" the measured values of the VoI. (The bins are typically consecutive, non-overlapping, and are usually equal size.)

Frequency histogram: count how many values fall into each bin/interval and draw accordingly.

y-axis: "Count" or "Frequency"

Density histogram: count how many values fall into each bin, and adjust the height such that the sum of the area of all bins equals 1. Equivalently: construct a Frequency histogram and divide the y axis by the total data count.

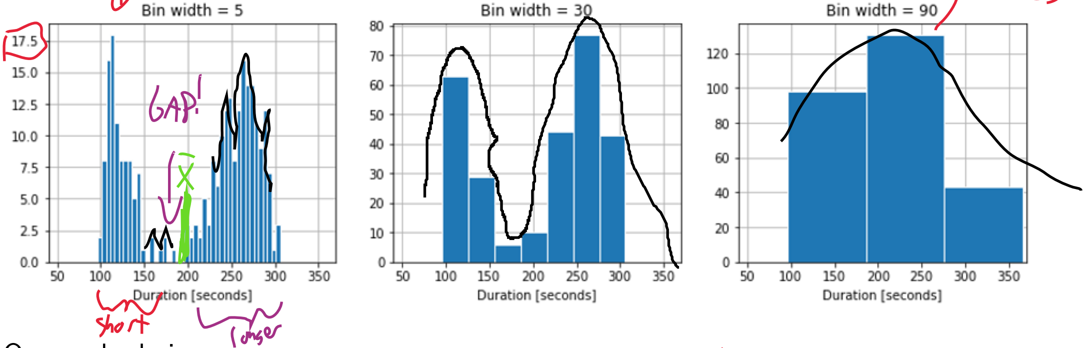
Old Faithful Histogram

count: 2 + 8 + 6 + 18 ....

"2 modes"

total count = 125 + 98 + 41

The number of bins chosen may lead to very different pictures of the data!



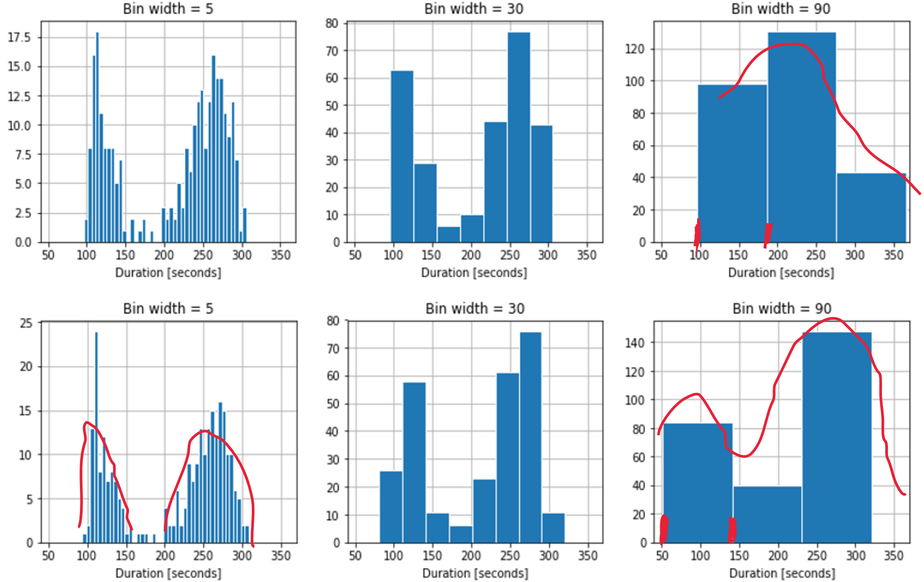
One such choice:

Friedman-Diaconis: bin width $= 2\frac{IQR}{\sqrt[3]{n}} = 2\frac{Q_3-Q_1}{\boxed{n^{1/3}}}$ → python

→ more data → finer bins

Where bins begin and end may also matter!

choose # bins so it doesn't matter



How many bins?

A lot of statisticians advise different rules or sanity checks for histogram bins.

Textbook:

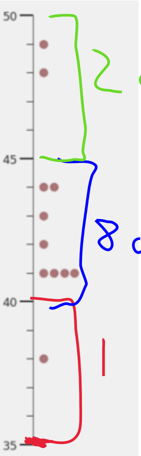
$$n_{bins} = 1 + 3.3\log_{10}(n)$$
$$w_{bins} = \frac{3.49s}{n^{1/3}}$$

Don't memorize these. My heuristic for binning: start with "too many" bins at first if you have to, and slowly expand the bin size to ensure:

1. The data starts to "smooth" out a little... but
2. We don't smooth over what appear to be distinct multiple modes

Histogram Example

Find the frequency histogram with bin width 5 of the data on left, with left-most bin edge at 35.



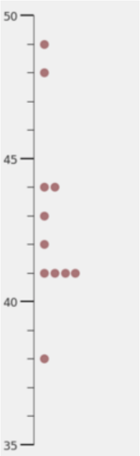
2 of 'em

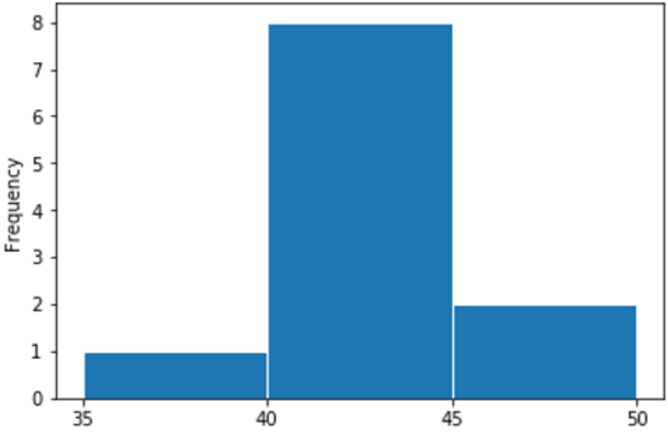
8 counts

1

Histogram Example

Find the frequency histogram with bin width 5 of the data on left, with left-most bin edge at 35.





Histogram Descriptives

Histograms come in a variety of shapes.

Negative Skew

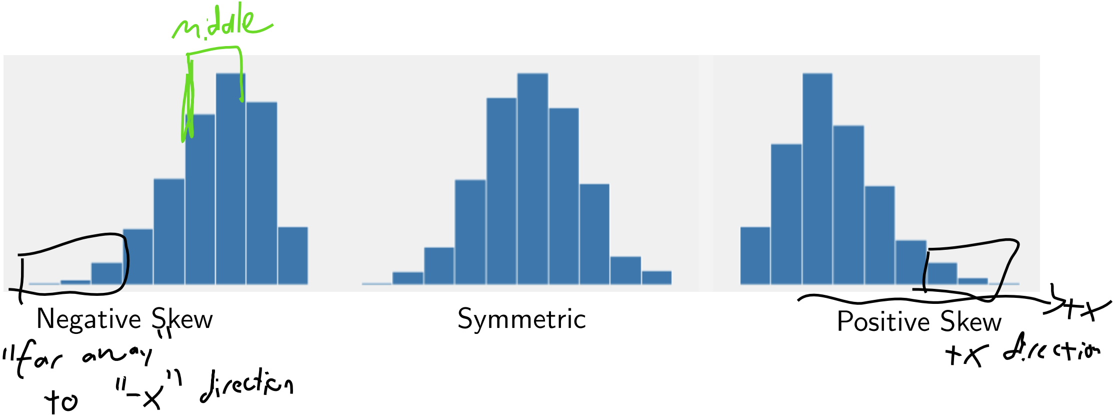
Symmetric

Positive Skew

Histogram Descriptives

Skewness

Histograms come in a variety of shapes.



Histogram Descriptives

Histograms come in a variety of shapes.

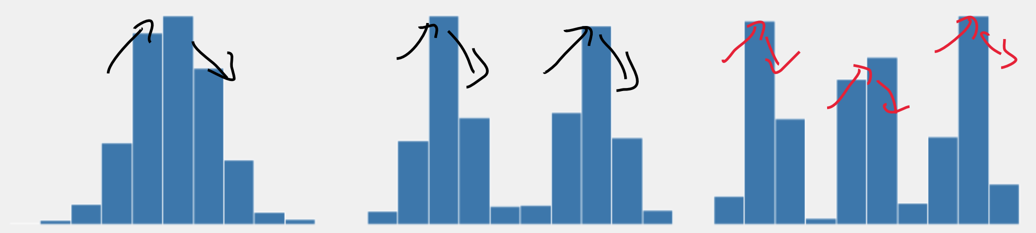
Unimodal

Bimodal

Multimodal

Histogram Descriptives

Histograms come in a variety of shapes.



Unimodal | Bimodal | Multimodal

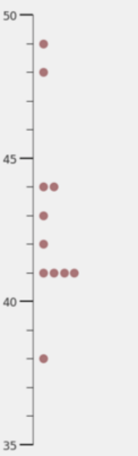
3 or more

Quartiles, Day 2

Compute the Quartiles and the IQR of the data to the left, with

$$x = [38, 41, 41, 41, 41, 42, 43, 44, 44, 48, 49]$$

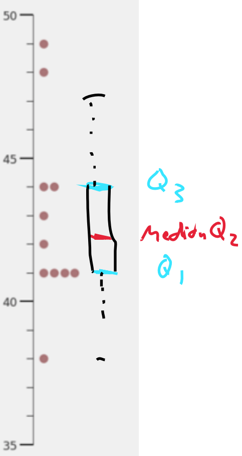
median!



Quartiles, Day 2

Compute the Quartiles and the IQR of the data to the left, with

$$x = [38, 41, 41, 41, 41, 42, 43, 44, 44, 48, 49]$$



$n = 11$ is odd, so Q_2 or the median is the 6th sorted value of 42. Then 41 and 44 divide the the halves in half, and are the 3rd and 9th sorted data points.

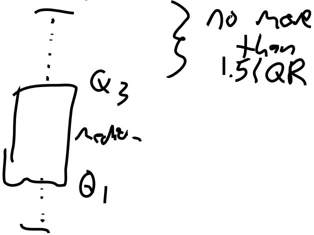
$$x = [38, 41, 41, 41, 41, 42, 43, 44, 44, 48, 49]$$

This makes the $IQR = 44 - 41 = 3$

Boxplots

A boxplot is a convenient way of graphically depicting groups of numerical data through the five number summary: minimum, first quartile, median, third quartile, and maximum.

1. The box extends from Q1 to Q3
2. The median line displays the median
3. The whiskers extend to farthest data point within $1.5 \times IQR$ of each quartile
4. The fliers or outliers are any points outside of the whiskers
5. The width of the box is unimportant
6. Can be horizontally or vertically oriented



Boxplots

Why do we use box plots?

1. They depict centrality via the median.
2. They depict dispersion through both the range and the IQR
3. Major outliers are shown
4. The median's location within the IQR suggests skewness; so too may lopsided whisker lengths or outliers

When might a box-whisker plot be misleading?

-

When might a box-whisker plot be particularly useful?

-

Boxplots

Why do we use box plots?

1. They depict centrality via the median.
2. They depict dispersion through both the range and the IQR
3. Major outliers are shown
4. The median's location within the IQR suggests skewness; so too may lopsided whisker lengths or outliers

When might a box-whisker plot be misleading?

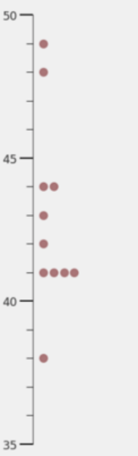
• No indication of how data are dispersed (is there "no-man's land"?)

When might a box-whisker plot be particularly useful?

• Comparing medium numbers of variables or columns quickly (say, 3-10); and much easier than histograms

Boxplot Example

Draw the box-whisker plot for the data to the left.



Title

Today we learned

1. How to represent data with histograms and box-whisker plots (boxplots)
2. Some strengths and weaknesses of each

Moving forward:

- ~~No class Monday for Labor Day.~~
- Next notebook day: making some histograms, boxplots, and playing around with data frames.

Next time in lecture:

- We probably talk about probability!