8-30 Visual EDA
Friday, August 30, 2019 11:47 AM

8-30 Visual EDA

8-30: Exploratory Data Analysis (EDA)
CSCI ~~X~~3022 Fall 19

→ Do quiz+et00 right this second if you haven't (due 1:00pm).

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

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

Opening Example Sol'n

$$f(c) = \sum_{i=1}^{n} (x_i - c)^2 = (x_1 - c)^2 + (x_2 - c)^2 + \cdots + (x_n - c)^2$$

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

Set $=0 \Rightarrow$

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

$$0 = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} c$$

note: $\overline{X} = \frac{\sum x_i}{n}$

$$= n\overline{X} - n \cdot c \Rightarrow c = \overline{X}$$

Opening Example Sol'n

Differentiating yields

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

Setting $f'(c) = 0$ gives

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

Announcements and Reminders

- Homework 1 posted soon, probably over the weekend
- Wednesday is a notebook day!
- Last time: numerical measures for centrality and dispersion

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.

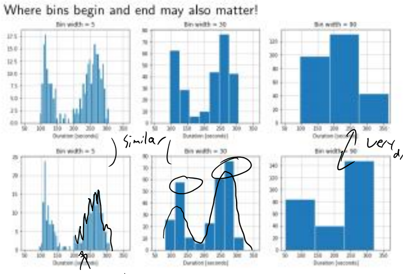
Density histogram: count how many values fall into each bin, and adjust the height such that the sum of the area of all bins equals 1. Equivalently: construct a Frequency histogram and divide the y axis by the total data count.

Old Faithful Histogram

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

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

Where bins begin and end may also matter!



1.4, 2.4, 3.4

bins width 2.5

(0, 2.5) (2.5, 5)

2 1

(-1, 1.5) (1.5, 4) (4, 5.5)

1 2

How many bins?

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

Textbook:

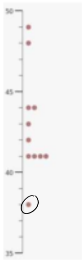
$$n_{bins} = 1 + 3.3\log_{10}(n)$$
$$w_{bins} = \frac{3.49s}{n^{1/3}}$$

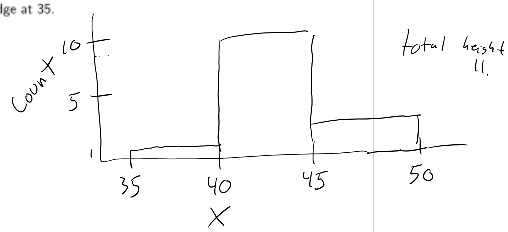
Don't memorize these. My heuristic for binning: start with "too many" bins at first if you have to, and slowly expand the bin size to ensure:

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

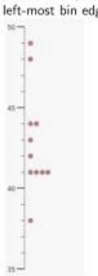
Find the frequency histogram with bin width 5 of the data on left, with left-most bin edge at 35.

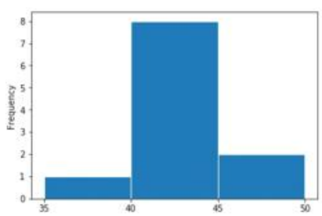




Histogram Example

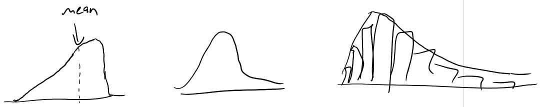
Find the frequency histogram with bin width 5 of the data on left, with left-most bin edge at 35.





Histogram Descriptives

Histograms come in a variety of shapes.



Negative Skew | Symmetric | Positive Skew

Histogram Descriptives

Histograms come in a variety of shapes.

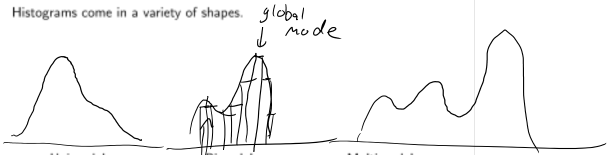
Negative Skew | Symmetric | Positive Skew

Histogram Descriptives

mode: local maximum

Histograms come in a variety of shapes.

global mode



Unimodal Bimodal Multimodal

Secondary mode

Histogram Descriptives

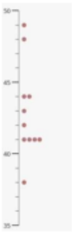
Histograms come in a variety of shapes.

Unimodal Bimodal Multimodal

Quartiles, Day 2

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

$$x = [38, 41, 41, 41, 41, \boxed{42}, 43, 44, 44, 48, 49]$$



median or $Q_2 = 42$

$Q_1 = 41$

$Q_3 = 44$

$IQR = Q_3 - Q_1 = 3$

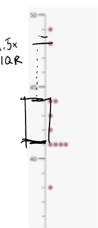
$44 - 41 =$

Quartiles, Day 2

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

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

1.5x IQR



$n = 11$ 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.

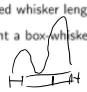
1. The box extends from Q1 to Q3
2. The median line displays the median
3. The whiskers extend to farthest data point within $1.5 \times IQR$ of each quartile
4. The fliers or outliers are any points outside of the whiskers
5. The width of the box is unimportant
6. Can be horizontally or vertically oriented

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?

-  multiple modes not captured

  Old faithful

When might a box-whisker plot be particularly useful?

- unimodal; if there are outliers
  easy to compare

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?

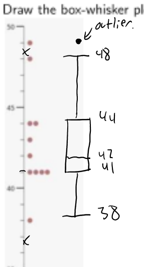
• No indication of how data are dispersed (is there "no-man's land"?)

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.



IQR = 3 from 41 to 44

median 42

Title

Today we learned

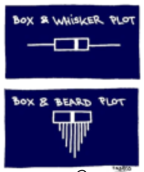
1. How to represent data with histograms and box-whisker plots (boxplots)
2. Some strengths and weaknesses of each

Moving forward:

- No class Monday for Labor Day.
- Notebook day: making some histograms, boxplots, and playing around with data frames.
- 3 of the next 5 course meetings are notebook days.

Next time in lecture (Friday):

- We probably talk about probability!



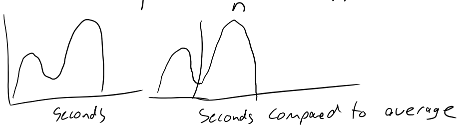
Suppose we have a data set $X_1, X_2 \cdots X_n$.

a) What happens to the mean of X if 3 is subtracted from each data value?

i) It's 3 smaller.

$\bar{X} = \frac{\sum X_i}{n}$; goal find $\bar{Y} = \frac{\sum (X_i - 3)}{n}$

$\bar{Y} = \frac{\sum X_i}{n} - \frac{\sum 3}{n} = \bar{X} - 3$



b) What happens to the Std. dev?

i) No change

$S_x^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$; if we double the data;

$S_x = \sqrt{s^2}$

$S_Y^2 = \frac{\sum (2 \cdot X_i - 2\bar{X})^2}{n-1} = \frac{\sum 2^2 (X_i - \bar{X})^2}{n-1}$

$S_Y^2 = 4 \cdot \frac{\sum (X_i - \bar{X})^2}{n-1} = 4 S_X^2$

$S_Y = \sqrt{4 S_x^2} = 2 S_X$.