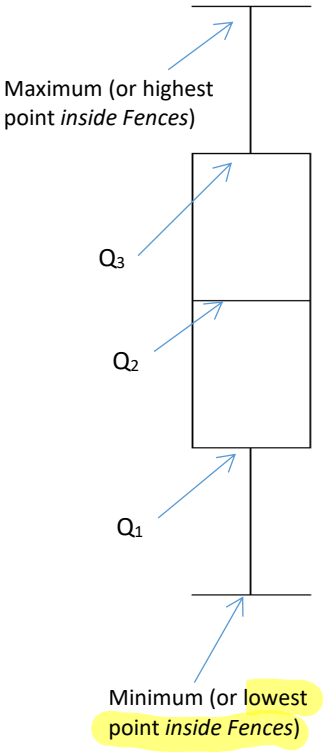
STAT 107 – Fall 2020

Distributions - Shape of the Data

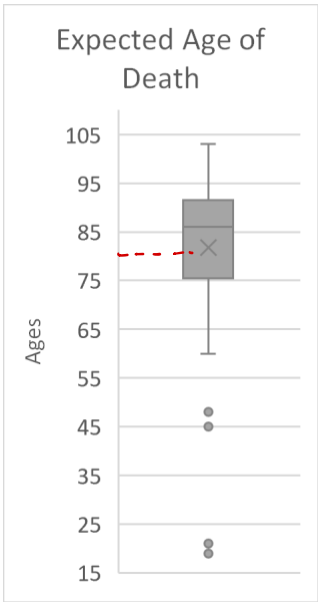
Ha Khanh Nguyen (hknguyen)



Boxplots

- A graphical representation of a data set that uses quartiles
  - 5 vertical lines represent the 5-number summary.
  - The "whiskers" (outside lines) are the minimum and maximum values. **HOWEVER,** when there are extreme values, the whisker line will be the next highest/lowest value that is still inside the Upper/Lower fence
  - Extreme values are denoted by a tiny dots past the first or last whisker.
- By displaying where and how wide each of the 25% marks are, we can quickly see how the data is distributed.

Practice: Consider the boxplot on the right, in which I asked 56 students at what age they expected to die. Let's make some quick observations about the data.



What is the 5-number summary: 18, 75, 86, 93, 104

The IQR is approximately… $Q_3 - Q_1 = 93 - 75 = 18$

The Range is approximately… max - min = 104 - 18 = 96

The Median (Q_2) is approximately… 86

The mean (in this boxplot) is represented with an X. It is approximately… 81

What does the lowest whisker represent?

About 50 % of the data lies between 86 and 104

Is this data positively skewed, negatively skewed, or approximately symmetrical? *Note: seeing skewness in a boxplot can be seen better with this simulation!* https://istats.shinyapps.io/MeanvsMedian/

left-skewed negatively skewed

Practice: The acidity levels (as measured by pH) were determined for 105 samples of rainwater. The mean was 5.43 with a standard deviation of 0.48. The five number summary are (4.33, 5.05, 5.44, 5.79, 6.65). Does this data seem to be positively skewed, negatively skewed, or approximately symmetrical? How might you decide?

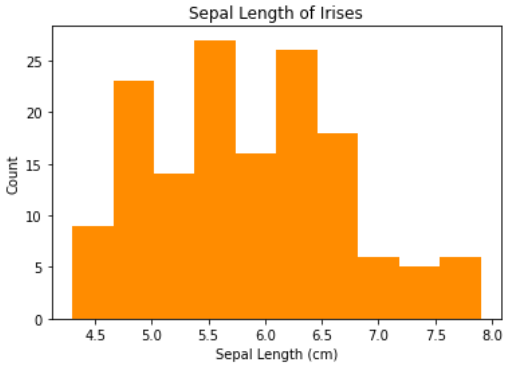
$\bar{x} = 5.43 \approx M = 5.44$

$\Rightarrow$ approx symmetrical

$Q_2 - Q_1 = 0.39$ $\quad$ $Q_3 - Q_2 = 0.35$

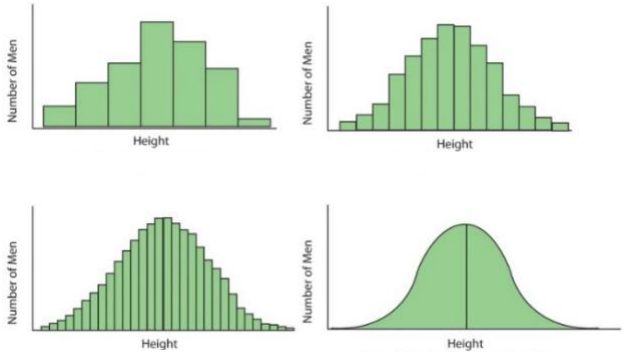
max $- Q_2 = 1.21$ $\quad$ $Q_2 -$ min $= 1.11$

Histograms



- Histograms are a very straightforward representation of distribution for a numerical variable.
- The **variable range is represented in the x axis**, and depending on how many data points lie within a certain range (or "bin"), the taller that bin will be.
- The **width of the bin** is up to the judgment of the graph creator.

Density Curves

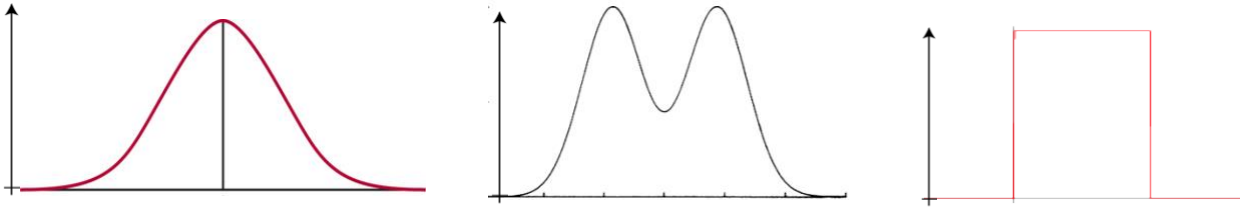


Turley (2012).
http://edensguest.blogspot.com/2012/04/g-is-for-gaussian-curves.html

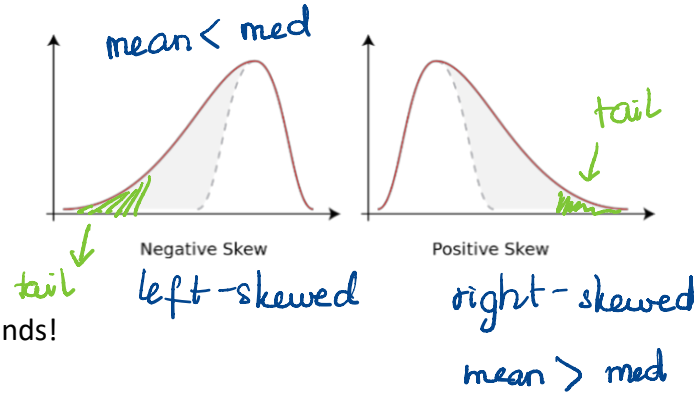
- A density curve is the *smooth* version of a histogram. We use the term **density** because it references how dense or how common it is to find a value within a certain range.
- We often use density curves to *approximate* the distribution of a *population*.
- Density curves represent what a histogram would eventually look like if we had an infinite number of observations and an infinite number of bins.

Characteristics of Distributions

- **Symmetry vs. Skewness**
  - Many distributions are fairly symmetrical, meaning that they balance at some central point. Consider these different examples below of symmetric distributions.



Some distributions might instead be skewed. **Skewed** means that the data stretches or skews out sparsely to one direction while being heavily concentrated on the other side.



To-do:

- Finish Lab 4, commit and push the lab using git commands!
- Complete HW 3 on Prairie Learn!

This lecture notes were adapted from "Course Notes for STAT 100: Statistics" by Kelly Findley & Ha Khanh Nguyen.