Math 247: Summaries for Skewed Distributions (Sections 3.3, 3.4)

Measuring Center (Finding a "Typical Value" for the Data):

Median: The *median* of a data set is the number that splits the ordered data in half.

For an odd number of data values, the median is the center data value.

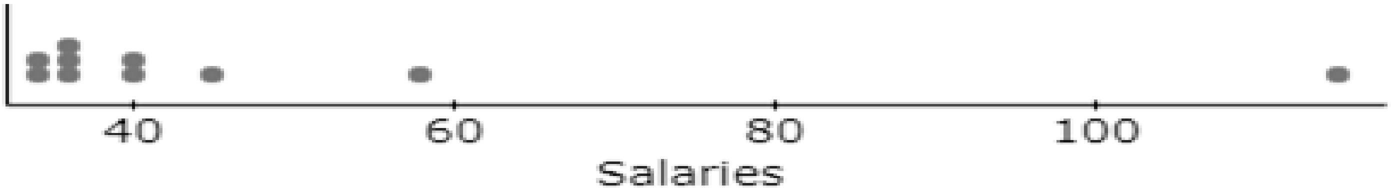
For an even number of data values, the median is the average of the two center data values.

Resistance: A statistic (or parameter) is *resistant* if its value is not much affected by extreme values in the data.

Example: Suppose a small company has annual salaries (in thousands of dollars) as given below with a dotplot of the data. Find, by hand, the **mean**, **median**, and **mode** of the salary data. Show the mean and median on the dotplot.

34, 36, 36, 40, 115, 40, 58, 34, 45, 36,

First step to find the median: ____



Does there appear to be an outlier(s) in the data set?

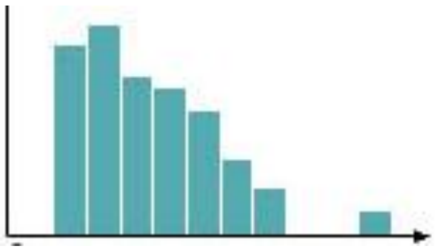
Describe the shape of the distribution of data:

Is the mean or the median more "typical" of the data? Explain.

Which value, the mean or the median, is more resistant to the outlier? Explain.

How skewness and outliers affect the mean and median:

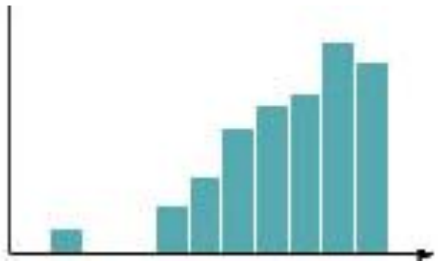
Estimate where the median will be on this graph, then estimate where the mean will be.



If the data is <u>skewed to the right</u> and/or has <u>outliers to the right,</u> the mean is pulled up from the center.

So…the <u>*mean*</u> **will be** _________________the <u>*median*</u>

Estimate where the median will be on this graph, then estimate where the mean will be.



If the data is <u>skewed to the left</u> and/or has <u>outliers to the left</u> the mean is pulled down from the center

So ….the <u>*mean*</u> **will be** _________________ the <u>*median*</u>

Variability with Skewed Data: We already saw with the Texting Data from 3.1 that the standard deviation is inflated by outliers. This means we need a more ***resistant*** method to measure variability.

We can find the "**<u>Inner Quartile Range</u>**" to measure variability of skewed data. The IQR isn't affected by outliers, so it's resistant!

Quartiles: Quartiles cut the data into ______________________________.

Q_1 = ___________ percentile Q_2 = ___________ percentile = _______________ Q_3 = ___________ percentile

The easiest way (not actually done in the real world) to find the quartiles is to split the data in half,

then find the median of the first half of the data for Q_1

and find the median of the second half of the data for Q_3

If the **median** is part of the data set, then **don't include it** when finding Q_1 and Q_3.

Example: Go back to the salary data and find the quartiles by hand. Plot these values on the dotplot.

Note: StatCrunch will provide this information when you find the "summary statistics"

Summary statistics:

Column	n	Mean	Variance	Std. dev.	Std. err.	Median	Range	Min	Max	Q1	Q3
Salaries	10	47.4	616.26667	24.824719	7.8502654	38	81	34	115	36	45

How to measure variability with <u>skewed</u> data: The Inner Quartile Range (IQR)

$$\mathrm{IQR} = Q_3 - Q_1$$

About 50% of the data falls between Q_1 and Q_3

What is the IQR for the salaries?

Summary:

If a data set is **<u>symmetric</u>**, we should use the ________________________ to summarize the center ("typical value")

and the _____________________________________ to summarize the variation (how spread out the data is).

If a data set is **<u>skewed</u>**, we should use the ________________________ to summarize the center ("typical value")

and the _____________________________________ to summarize the variation (how spread out the data is).

When comparing two data sets, **<u>if one or both of the sets are very skewed</u>**, we should use

the _________________ and the ______________________ to compare them.

How do we tell whether a data set is skewed or symmetric?

(1) Look at a graph of the data.

- Histograms or Dotplots will tell you whether the data is skewed or symmetric
- Boxplots are another type of graph we'll learn about in the next section.

(2) If you don't have access to a graph, then the mean and the median can be used to judge whether the data is skewed. (This isn't a perfectly reliable way to judge due to the possible influence of outliers!)

If Mean > Median then the data is likely skewed right

If Mean < Median then the data is likely skewed left.

(3) In advanced statistics, there are more technical ways to determine skewness. We won't study this but it falls under the category of "Goodness of Fit".

Math 247: Using Boxplots to Display Summaries (Section 3.5)

Five-Number Summary: This provides a quick snapshot of how the data is distributed.

Minimum, Q_1, Median, Q_3, Maximum

Example: What is the Five-Number Summary for the Salary data? Find by hand and confirm with StatCrunch table.

Salaries: 34, 34, 36. 36, 36, 40, 40, 45, 58, 115

Five Number Summary:

Outliers: Outliers are unusual data values that lie above or below the outlier boundaries.

Left (Lower) Outlier Limit = $Q_1 - 1.5 \cdot \text{IQR}$ Right (Upper) Outlier Limit = $Q_3 + 1.5 \cdot \text{IQR}$

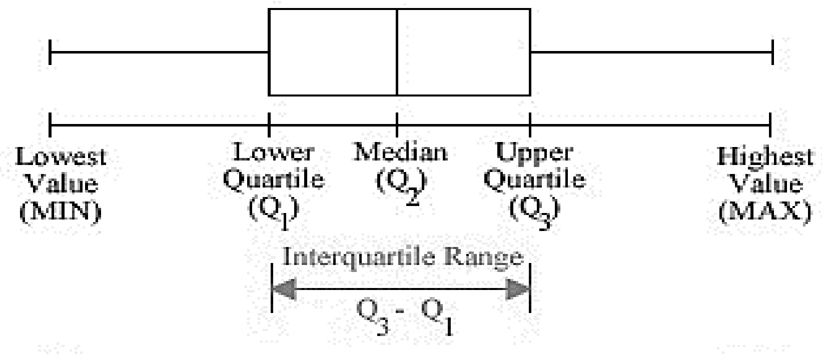
Example: Find the Outlier Limits for the salary data. Then identify the outlier(s).

Example: The Five Number Summary of a data set is 2, 33, 38, 43, 50. Is either the max or min value of this data set an outlier? Explain.

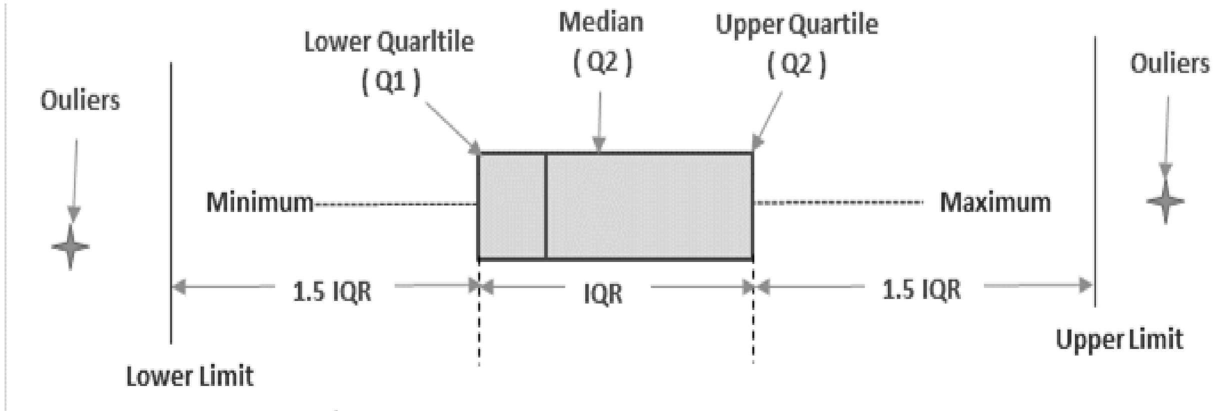
Boxplot: A simple <u>boxplot</u> is a graph of the 5-number summary. It can be **horizontally** or **vertically** oriented.

What it's good for: Boxplots give a quick view of the center and variability of the data. They can also indicate shape (skewness and symmetry) but aren't the best tool for this (a histogram or dotplot is more useful for determining shape).

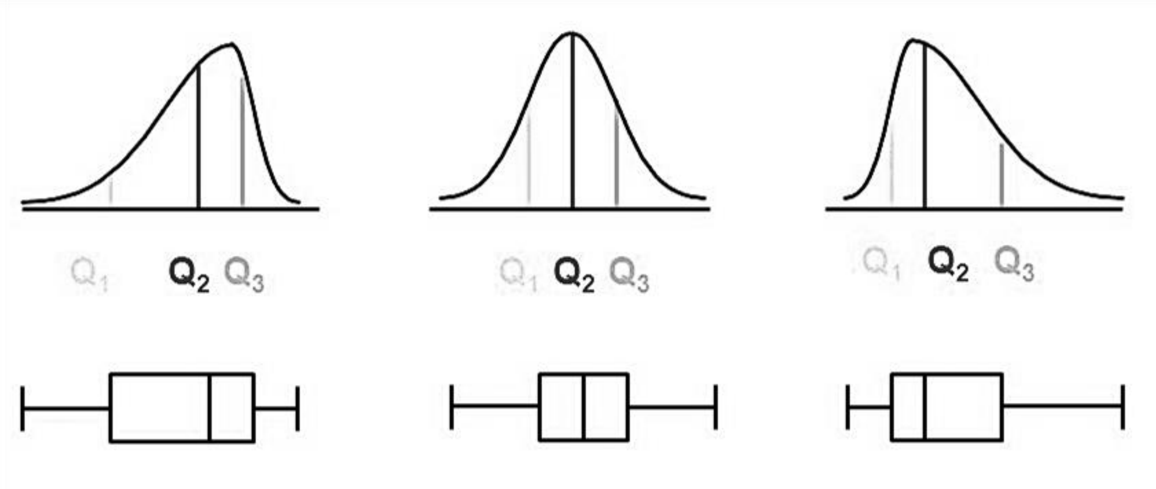
Boxplot (Simple):



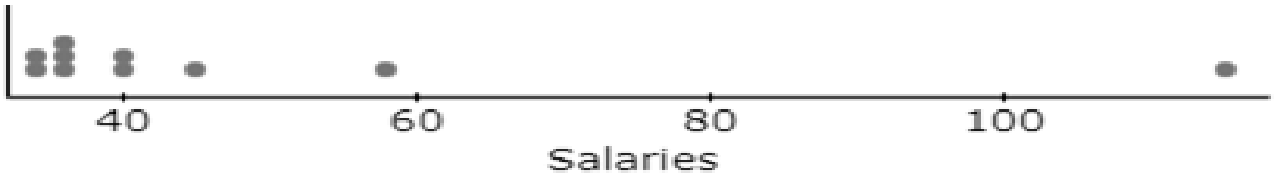
Boxplot (with Fences and Outliers): These boxplots also show outliers which can be very useful.
Fix the Q3 error on this graph!



Shape of Distribution: Both the box <u>symmetry</u> and the <u>whiskers</u> can indicate skewness.

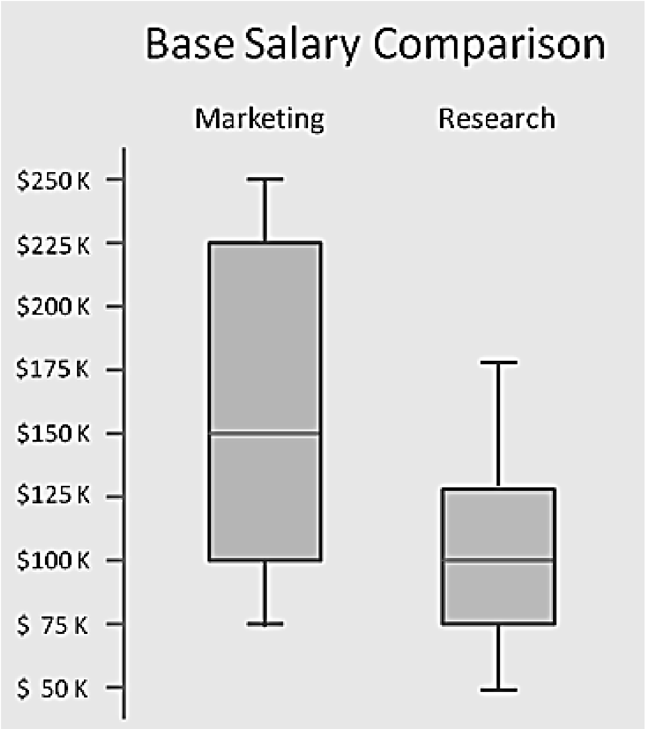


Example: By hand, **c**onstruct a boxplot with fences above the dotplot of the Salary Data:



Comparing Multiple Boxplots: At a glance, boxplots can tell you the **median** value for a group and how much **variability** is in the group. This is best use in comparing data sets.
Note: ***Boxplots can be graphed vertically or horizontally.***

Example: A company has both Marketing and Research Employees. The given boxplots show the distribution of these employees' salaries.



Which group's median salary is the lowest? Give the median for each group.

Which group has the greatest variability in salaries? How can you tell?