ECO220Y

Displaying and Describing Quantitative Data

Readings: Chapter 4 (4.1-4.3) and Chapter 5

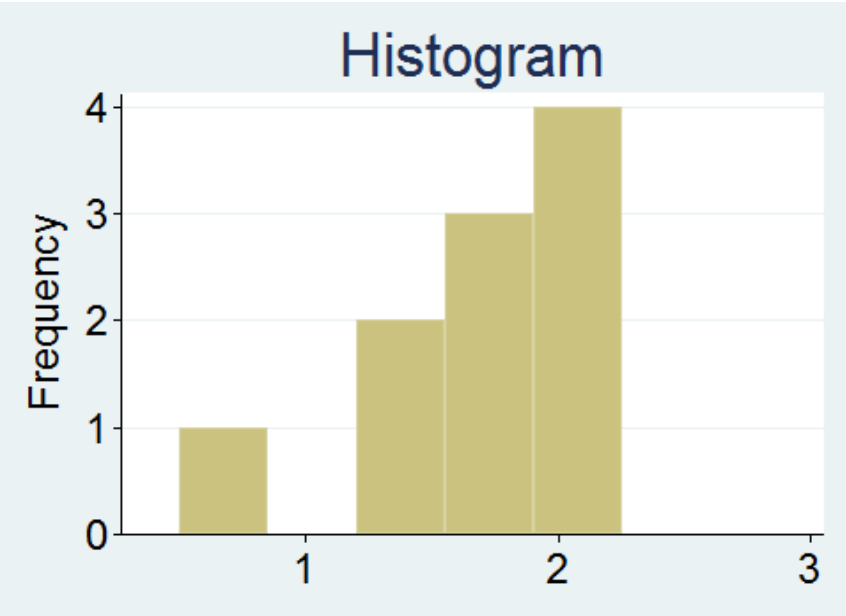
Fall 2011

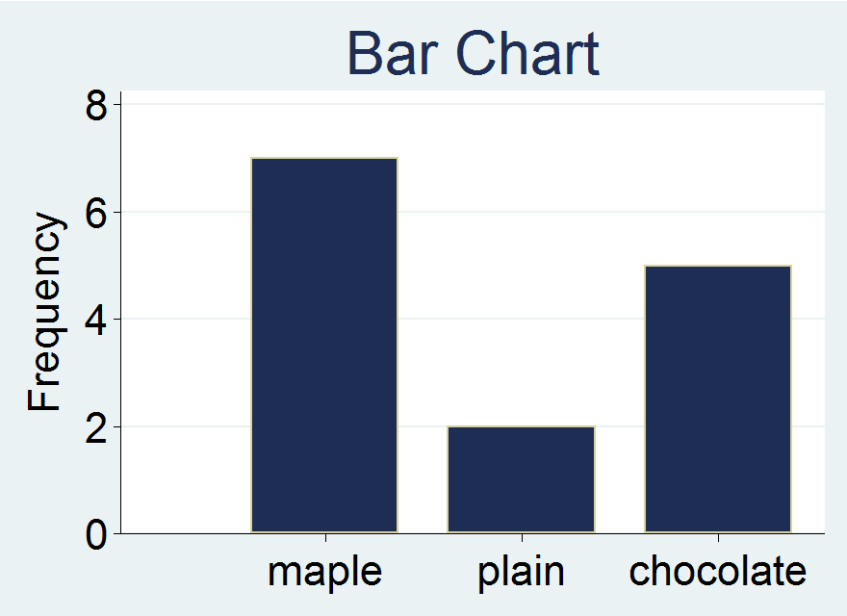
Lecture 2

Displaying Data

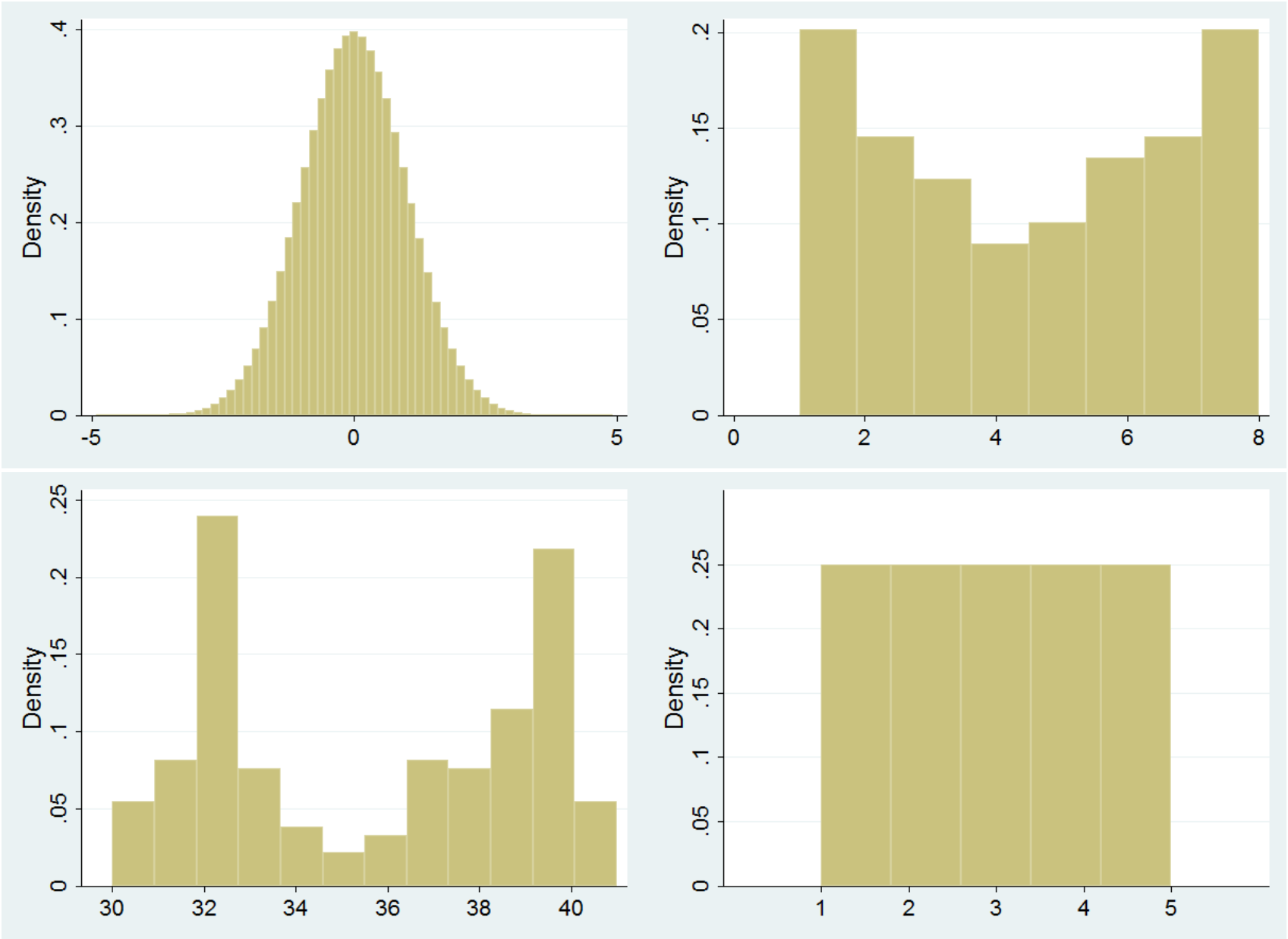
- Quantitative, or Interval Variables
  - Histogram
  - Stem-and-Leaf Display
  - Line Chart (Time-series Data)
- Categorical Data
  - Bar Chart
  - Pie Chart

Histogram vs Bar Chart

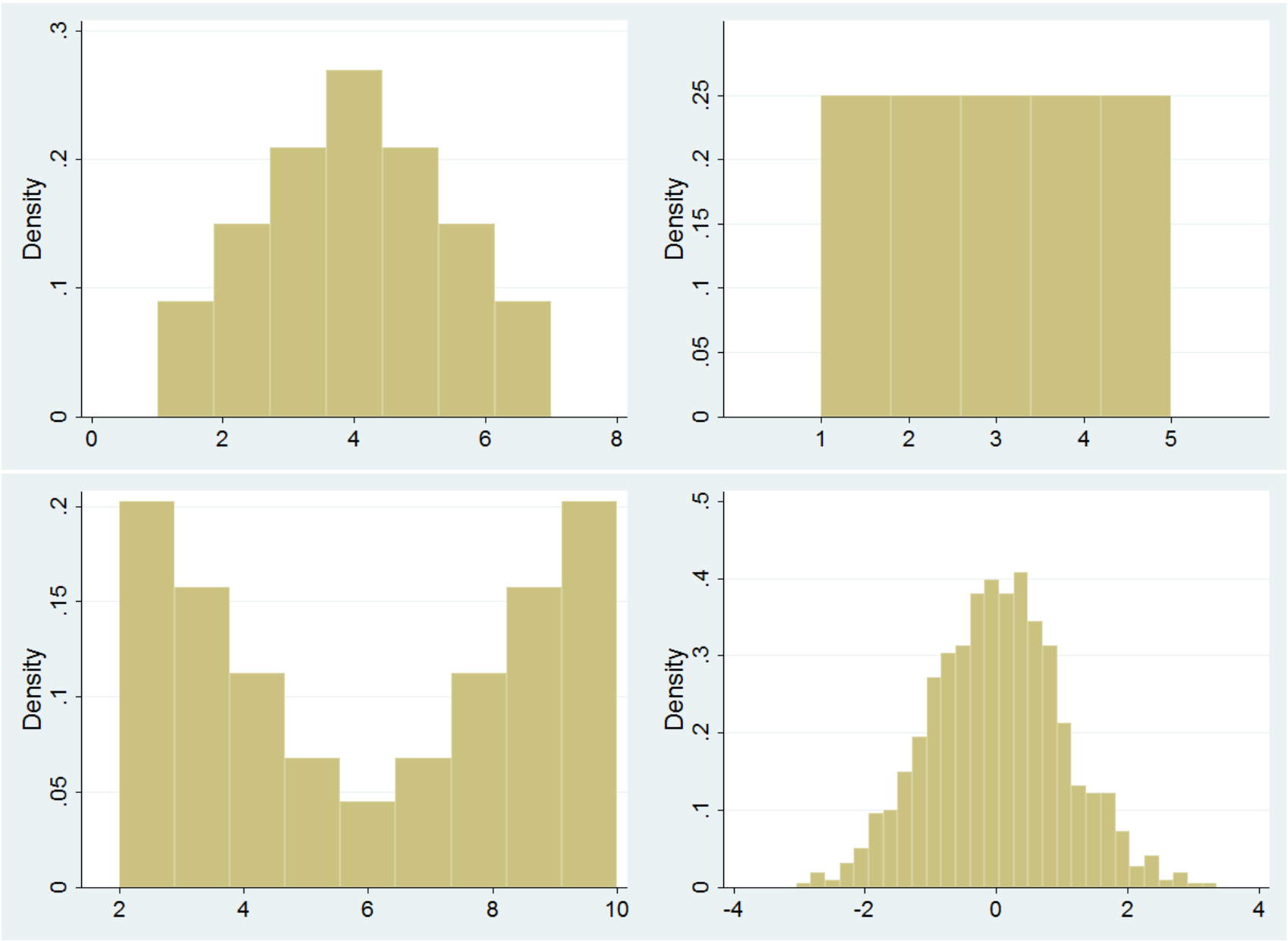




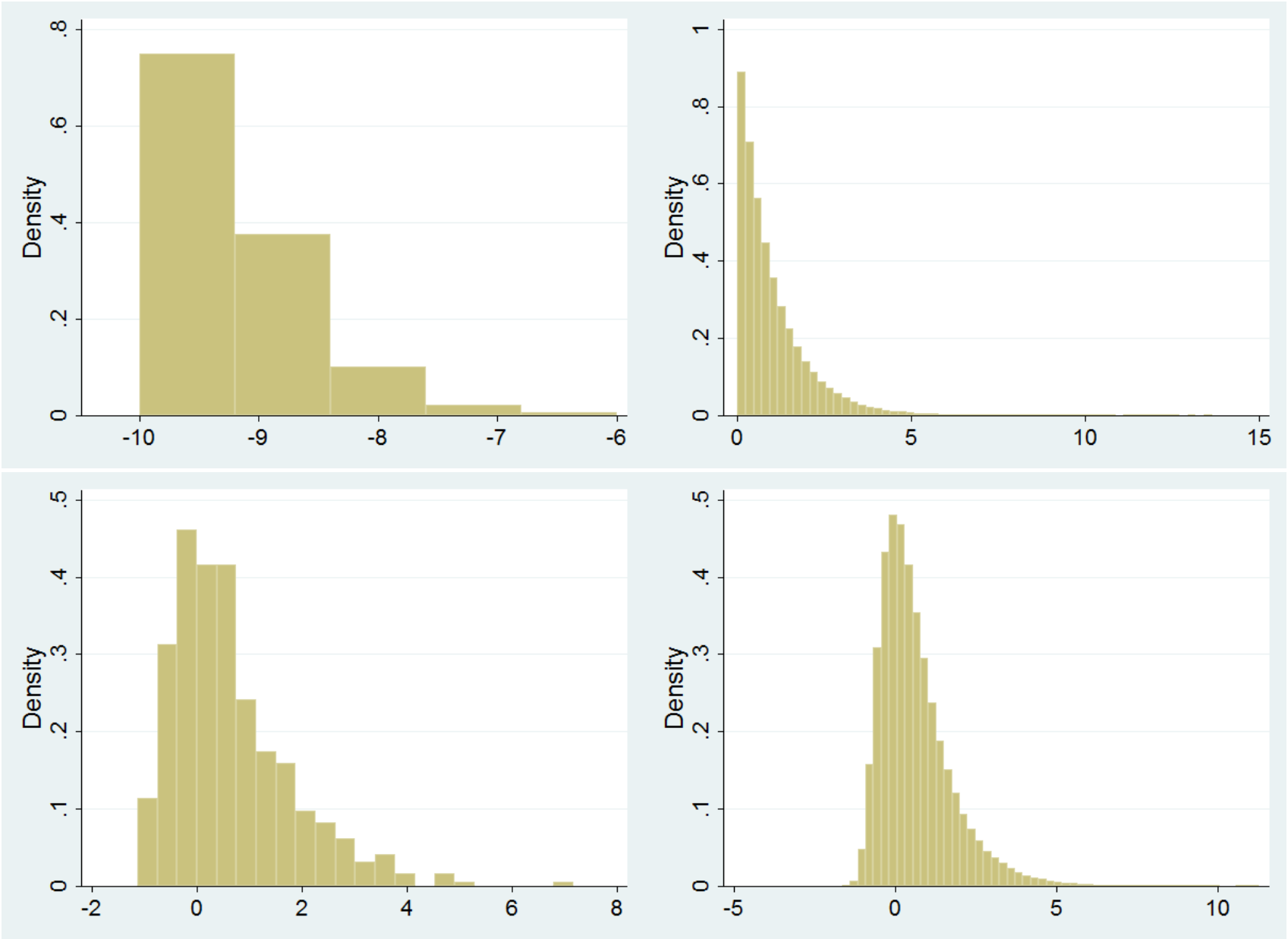
Shape of the Histogram - Modality



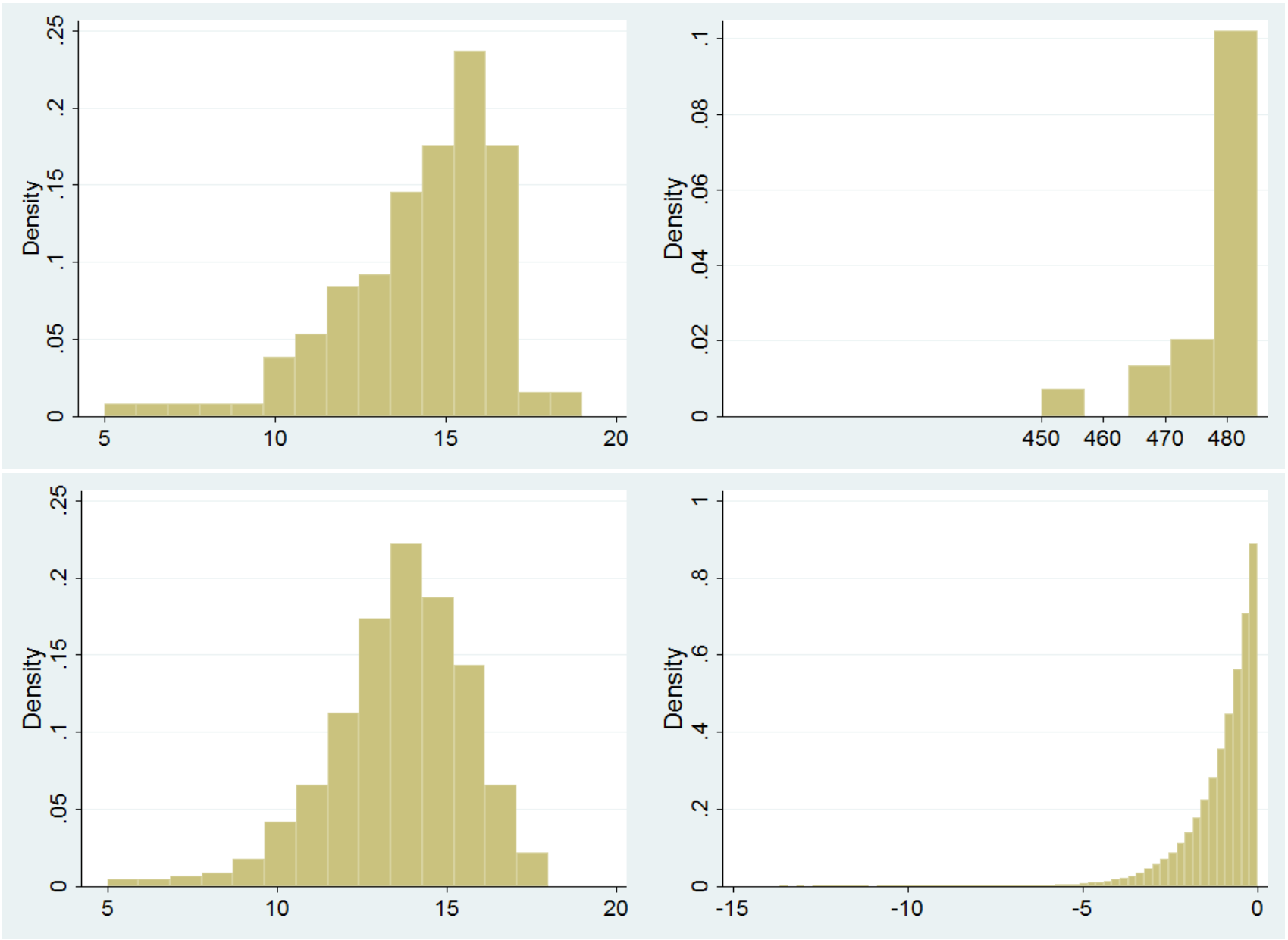
Shape of the Histogram - Symmetry



Shape of the Histogram - Skewness



Shape of the Histogram - Skewness



Summary

- Histograms are useful summaries that convey the following information:
  1. the general shape of the frequency distribution
  2. symmetry of the distribution and whether it is skewed
  3. modality - unimodal, bimodal or multi-modal
- With small sample size the shape of the distribution could have a non-normal shape and still possibly be from a population which has the normal distribution.
- Histogram of a sample conveys information about the population.
  - Be careful when describing histogram of a small sample (sampling error).
    - ★ Large sample size (big n) - we are confident about symmetry, skew, modality and shape.
    - ★ Small sample size (small n) - must interpret graph with caution.
    - ★ Big sample size - smaller sampling error.

Measure of Central Tendency - Mean

	Population	Sample
Number of observations	N	n
Mean	μ ("mu")	$\bar{y}$("y-bar")
How to find	$\frac{\sum_{i=1}^{N} y_i}{N}$	$\frac{\sum_{i=1}^{n} y_i}{n}$

Measures of Central Tendency - Mode and Median

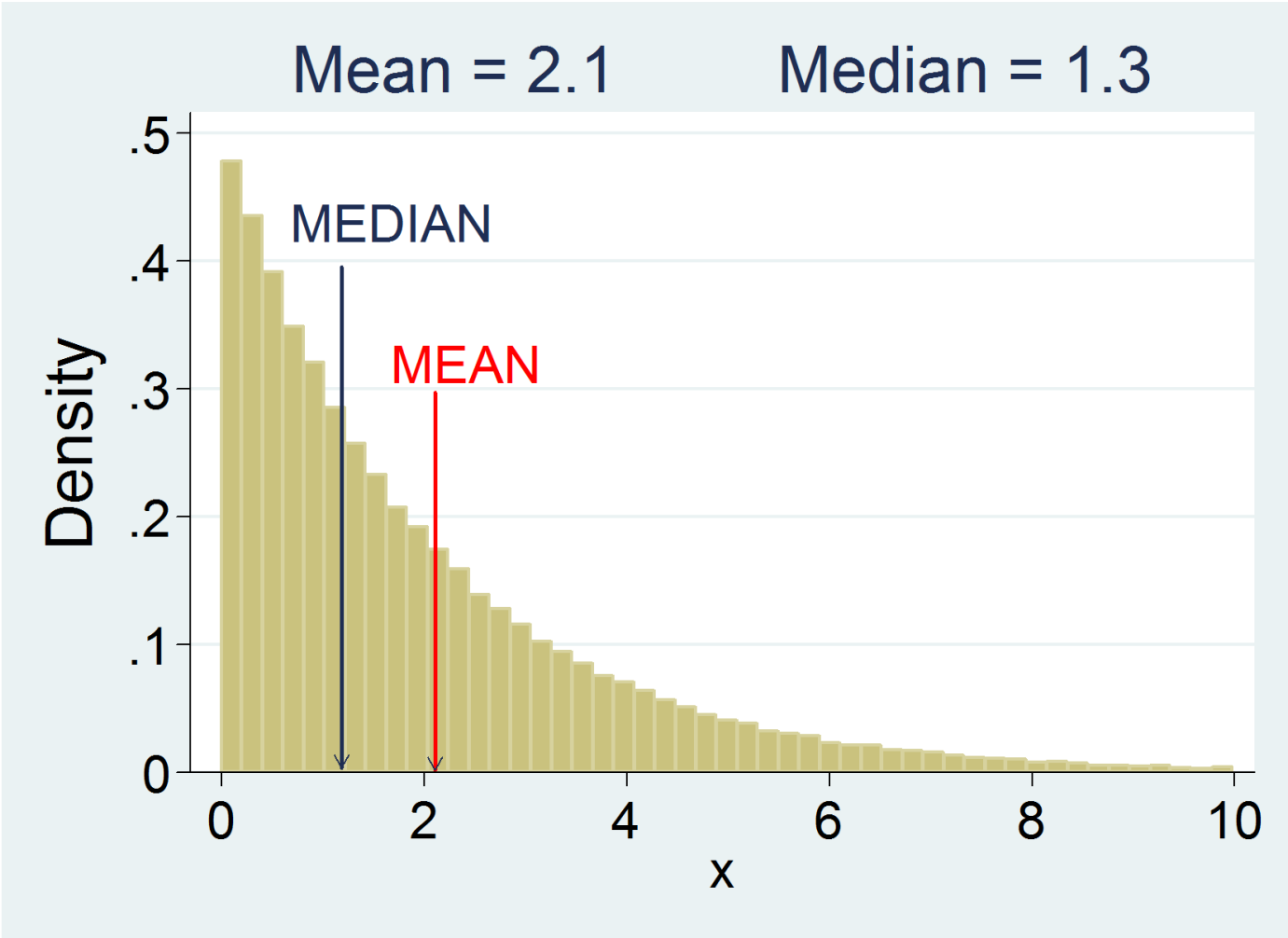
1. Median
   - Order observations from smallest (in value) to the largest
   - Find the middle - that would be the median of your data
2. Mode
   - Observation that occurs more often
   - Also called modal class
   - Not unique (unimodal, bimodal)
   - The mode is seldom used in practice, except to answer very specific questions:
     - ★ What is the most watched TV show?
     - ★ What is the best-selling automobile?
     - ★ What is the most common cause of death?

Measures of Central Tendency - Mode and Median

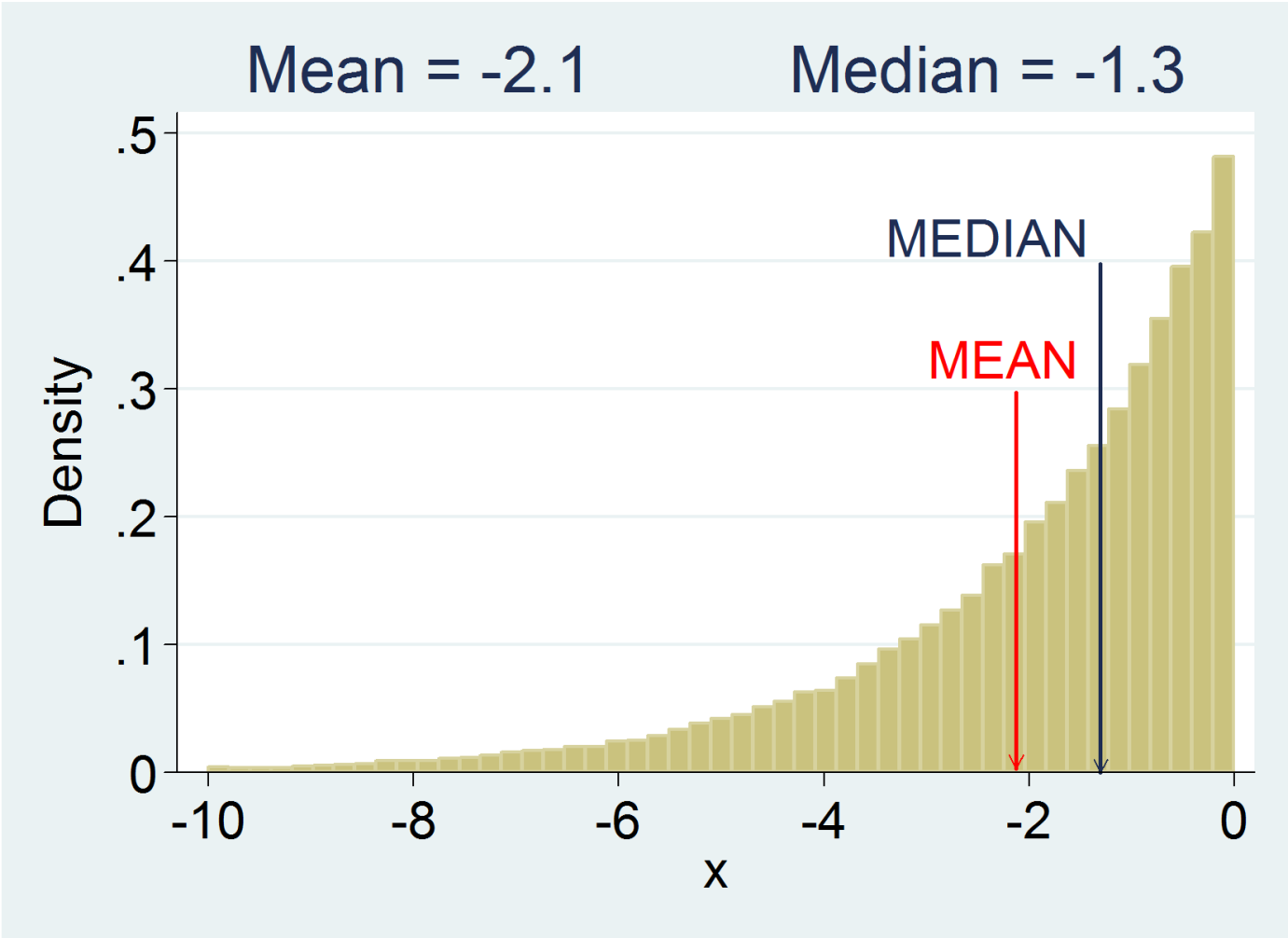
HH id	Children
1	2
2	1
3	1
4	3
5	2
6	1
7	10
8	3
9	1
10	2
11	1
Total	27

$\Longrightarrow$

order	HH id	Children
1	2	1
2	3	1
3	6	1
4	9	1
5	11	1
6	1	2
7	5	2
8	10	2
9	4	3
10	8	3
11	7	10
Total	11	27

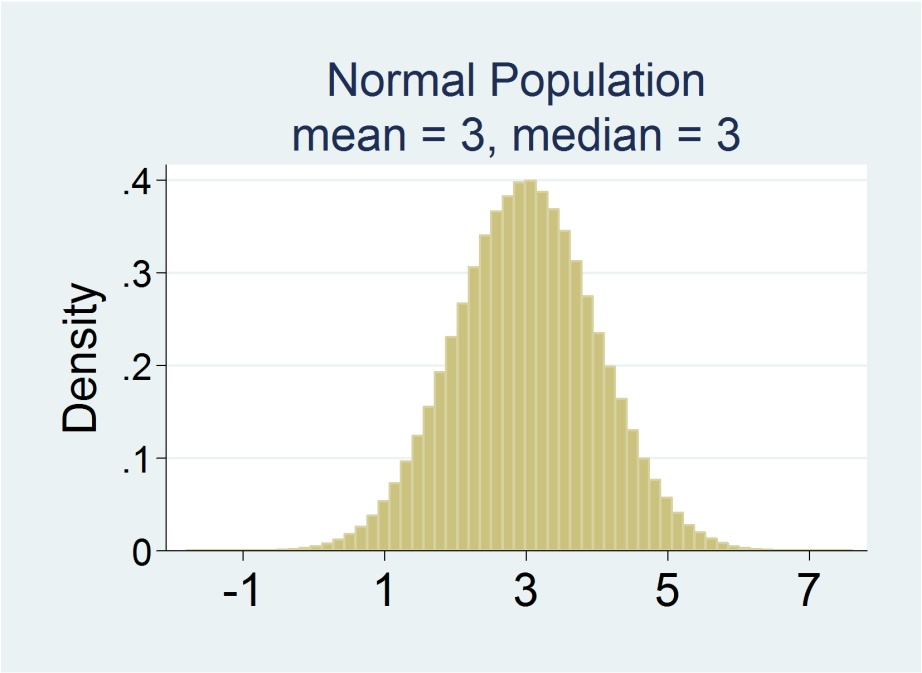


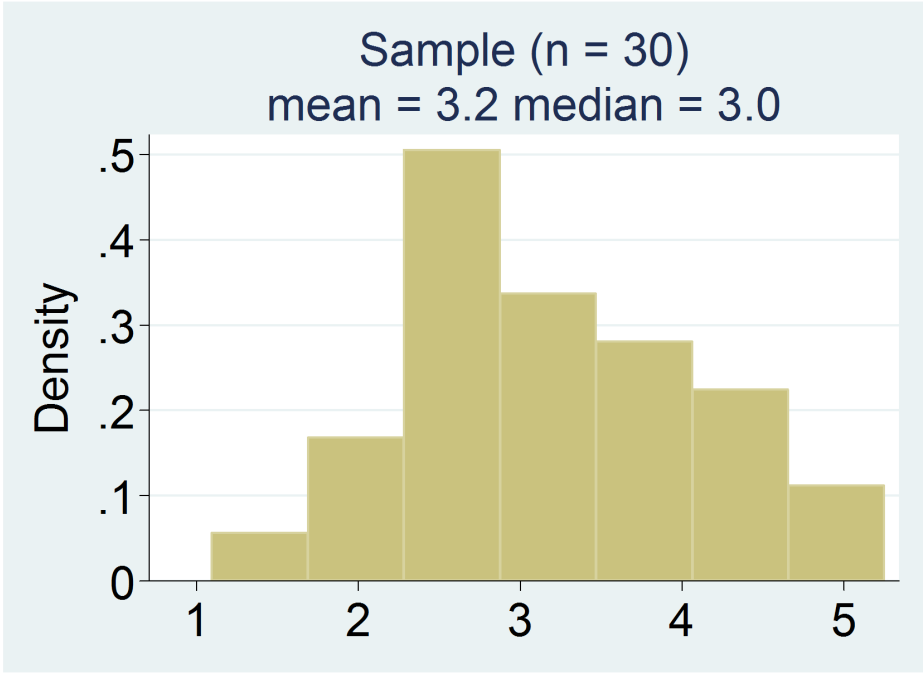
Why median is smaller than the mean?



Why mean is smaller than the median?

Normal Distribution





What are the median and mean of the population?

Why does the population mean equal to the population median?

Why isn't the sample mean equal to the sample median?

How can we explain numerical difference between the sample statistic and the population parameter?

How to Compare Different Measures?

Criteria for comparison:

- Sensitivity to outliers.
  - outlier is an extremely large or small value different from the bulk of the data.
- Retains potentially valuable information.

Robustness to Outliers

Mean	Not robust to outliers	Uses all the information
Median	Robust to outliers	Does not use all the information
Mode	Robust to outliers	Does not use all the information

Note: Robust=Not Sensitive

Summary

- Know what measure of central tendency to use.
- Mean, median and mode can be used with interval variables.
- Median and mode can be used with nominal/ordinal data.
- If histogram is single peaked and perfectly symmetrical, then the mean, mode and median coincide.
- If histogram has more than one peak or is asymmetrical, then the mean, mode and median will not coincide.

Measure of Spread - Range and IQR

- Range - is an absolute difference between the smallest and the largest value in the data
- Interquartile Range
  - Sort your data in ascending order.
  - Divide your data into two equal groups at the median.
  - Find the median of the first, "low" group. This is called Q1, or first quartile.
  - The median of the second, "high" group is the third quartile, Q3.
  - The interquartile range (IQR) is the difference between Q3 and Q1

Measures of Spread - Variance

	Population	Sample
Number of observations	N	n
Variance	σ^2	s^2
How to find	$\frac{\sum_{i=1}^{N}(y_i-\mu)^2}{N}$	$\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}$

$\sum_{i=1}^{n}(y_i - \bar{y})^2$ is called TSS, or Total Sum of Squares, and $n-1$ we call ν, or degrees of freedom.

Measures of Spread - Standard Deviation

There is one problem with variance as a measure of spread: variance (σ^2, s^2) is measured in units **squared**. For instance, variance for the income data is dollars squared, for work hours - hours squared!

The obvious thing to do is to take the square root$\longrightarrow$ standard deviation.

We define population standard deviation $\sigma = \sqrt{\frac{\sum_{i=1}^{N}(y_i-\mu)^2}{N}}$

And sample standard deviation $s = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}}$

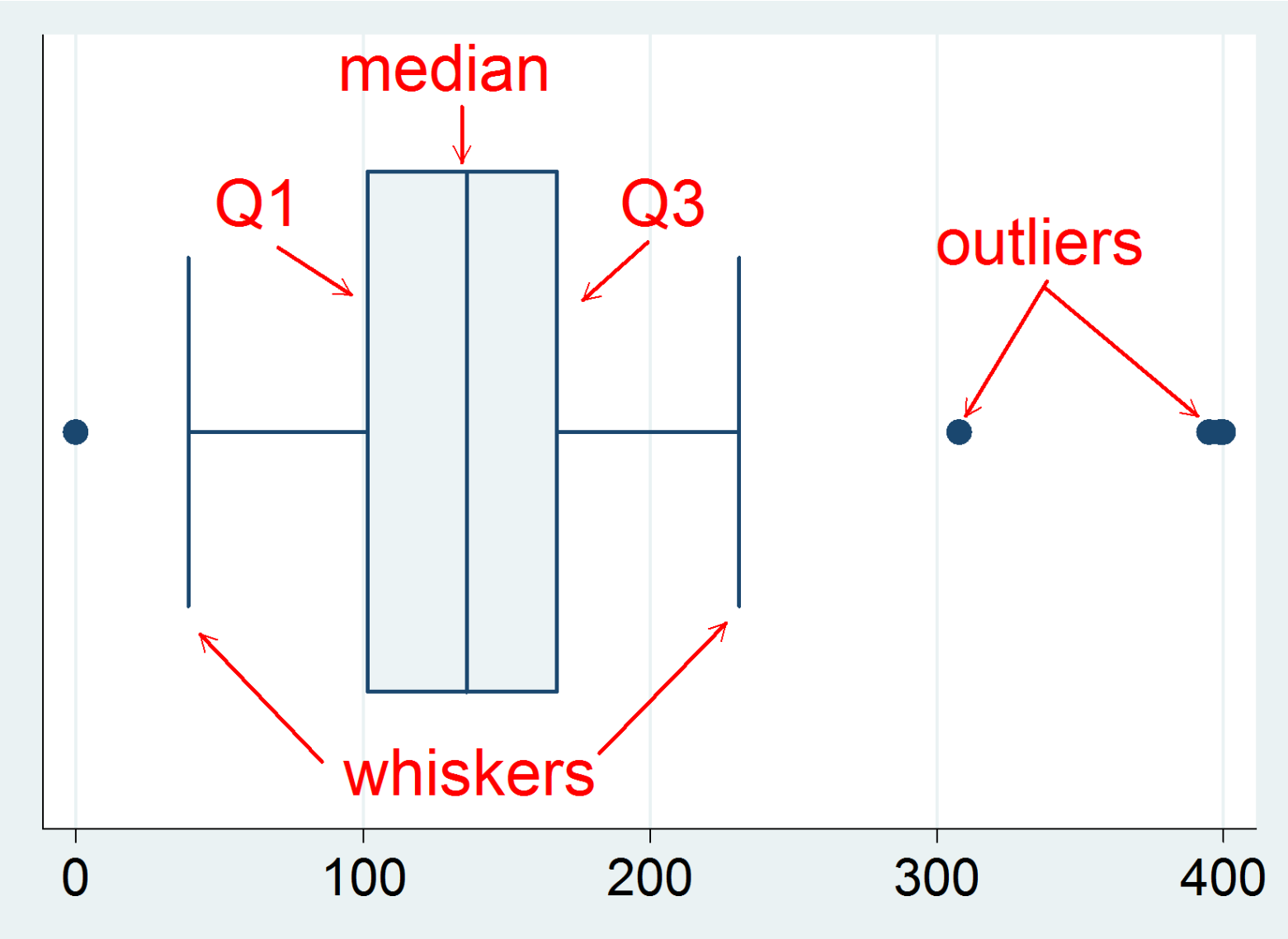
Empirical Rule

- 68.26% of observations are within 1 standard deviation from the mean
- 95.44% of observations are within 2 standard deviation from the mean
- 99.74% of observations are within 3 standard deviations from the mean
- Only for distributions which are approximately normal

Chebyshev's Theorem

- At least 75% of observations are within 2 standard deviations from the mean
- At least 89% of observations are within 3 standard deviations from the mean
- In general, $1 - \frac{1}{k^2}$ of observations are within k standard deviations from the mean
- Chebyshev's Theorem holds for *all* distributions

Boxplot and Five-Number Summary



Measure of Relative Standing - Percentiles

- Percentile is a value at cut-off between the bottom X percent of the data and upper $1 - X$ percent.
- For instance, median is 50^{th} percentile since 50% of the data are below the median and 50% of the data are above.
- First quartile Q1 is 25^{th} percentile, and third quartile, Q3 is 75^{th} percentile.
- Percentiles are useful when we want to compare how an observation fares compared to other observations in a sample.

Measures of comparison/Standardization

1. Coefficient of Variation (CV)
   - CV=Standard deviation/Mean
   - Measures how much variability is in the data compared to the mean
2. Standardization - z-score
   - Z-score is a standardized variable
   - $z = \frac{y-\bar{y}}{s}$
   - Variable z has a mean of 0 and standard deviation equal to 1.
   - Value of z-score indicates how many standard deviations a value is from the mean