Lecture 3

Part 1 Descriptive Statics, Summary Measures, and Data Visualization

Measures of Relative Standing

- A **measure of relative standing** measures the location of a particular value relative to the rest of the distribution of your data.
- Suppose you order your data from smallest to largest. A **quantile** is the value below which a certain proportion (say p) of your data lie and above which one minus that proportion (i.e., $1 - p$) of your data lie.

Examples of Quantiles

- **Quartiles** divide the data into quarters, and report the observation at the boundaries of the subsets.
  - For example, the first quartile (Q_1) is the value below which 25% of your data lie, the second quartile (Q_2) is the value below which 50% of your data lie, etc.
- **Percentiles** divide the data into hundredths, and report the observation at the boundaries of the subsets.
  - For example, the 91st percentile is the value below which 91% of your data lie (and the value above which 9% of your data lie).

Calculating Percentiles

- If we have n observations, the location of the pth percentile is given by:

$$L_p = (n+1)\frac{p}{100}$$

- Find the 31st percentile of the quiz marks from tutorial session A:

A	5	5	6	7	7	8	8	8

Calculating Percentiles

A	5	5	6	7	7	8	8	8

- The location of the 31st percentile is:

$$L_p = (8+1)\frac{31}{100} = 2.79$$

- So the 31st percentile lies 0.79 of the distance between the 2nd and 3rd observations, which are 5 and 6, respectively.
- Therefore, the 31st percentile is equal to:

$$5 + (0.79) \times (6 - 5) = 5.79$$

Interquartile Range

- The interquartile range (IQR) of a data set is defined to be:

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

A	5	6	5	7	8	7	8	8
B	9	5	6	7	7	6	5	

- Tutorial A: The IQR is $79.25 - 70.75 = 8.5$.
- Tutorial B: The IQR is $98 - 52 = 46$.
- IQR is more stable than the range.

Histograms

- **Histograms** plot the frequency of observations falling into defined intervals.
- Are one of the most useful ways to graphically present continuous data.
- Can give us information about the distribution of our data, including:
  - The approximate location of the mean and median.
  - The amount of spread/variability in our data.
  - The shape of the distribution, e.g., skewness.

Histograms

Constructing a histogram:

- Determine the range of the data, i.e., the smallest and largest values of the variable.
- Divide the range into an appropriate number of intervals or "classes".
- Count the number of observations that fall into each interval.
- Plot the intervals on the x-axis and the frequency (count) for each interval on the y-axis.

Histograms

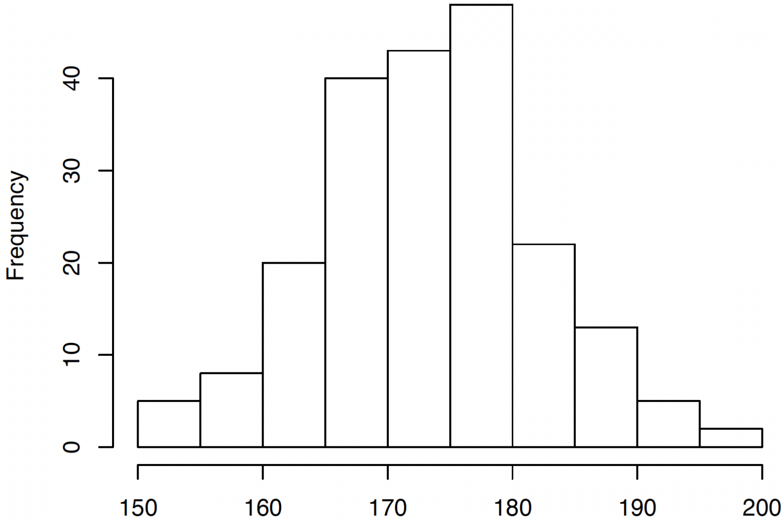
- How many intervals should we use?
- Depends on the number of observations.
- Typically, more observations means we should use more intervals to have a nice-looking histogram.
- Plot the intervals on the x-axis and the frequency (count) for each interval on the y-axis.

Histograms

- Back to favorite color example: Suppose while asking about favorite color, we also measured the height of each of student.
- Some descriptive statistics:

n	$\bar{X}$	s	Min	Q_1	Median	Q_3	Max
206	173.1	8.96	151.2	167.3	172.4	178.5	203.6

Histograms

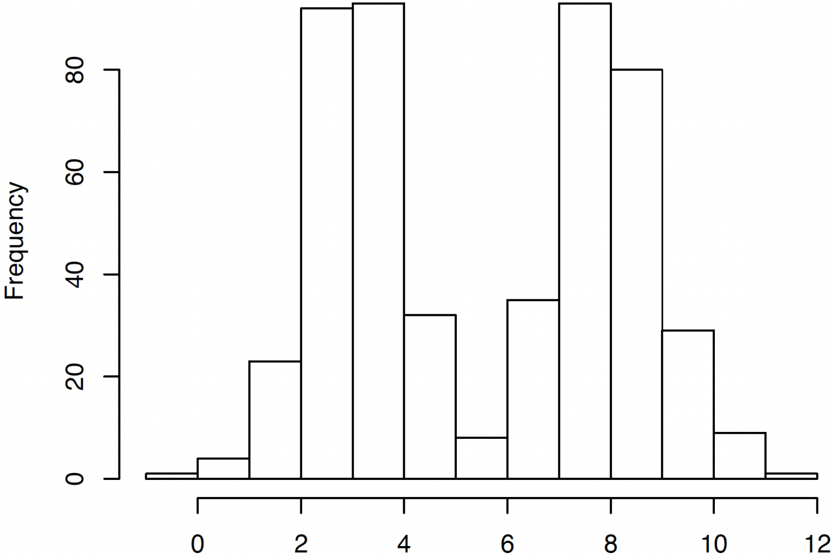


Histograms

From a histogram, we can easily determine:

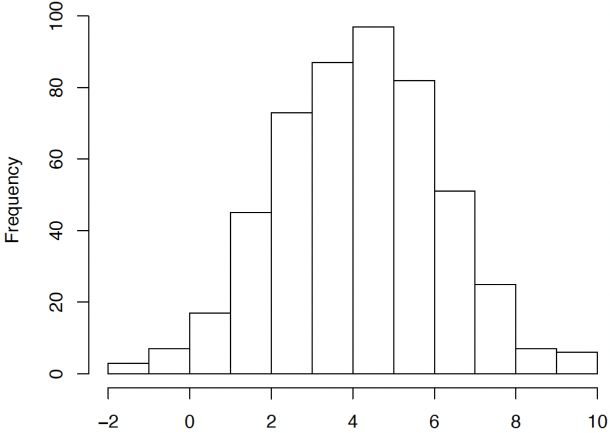
- The modal class, which is the class with the most observations.
- Whether the distribution of the data is uni-modal (one main peak in histogram) or multi-modal (multiple distinct peaks in histogram).
- Whether the distribution of the data is **skewed** or **symmetric**.

Bimodal Distribution



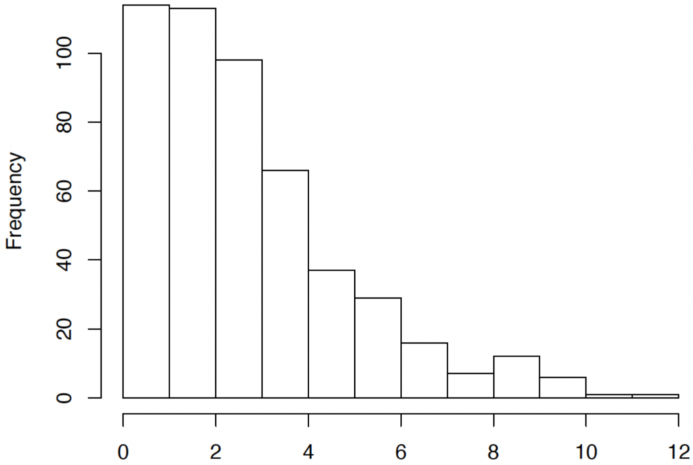
Symmetric Distribution

- **Skewness** is a measure of asymmetry in a given distribution.
- If a distribution is symmetric, then the skewness is 0 and the mean is equal to the median.



Positively Skewed Distribution

- If a distribution is positively skewed (or skewed to the right), then the mean is usually bigger than the median and the histogram will have a long tail to the right.

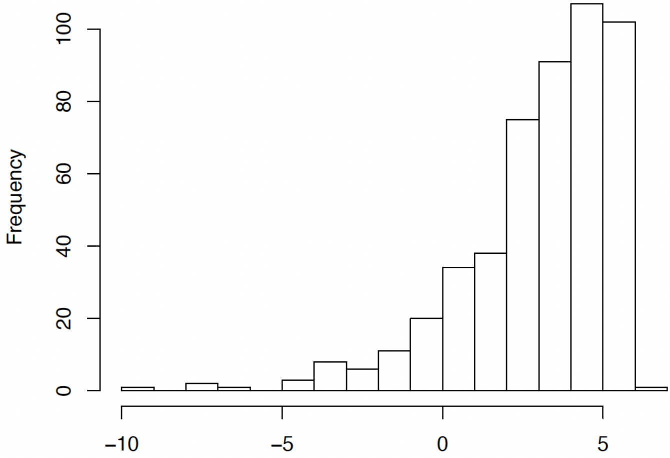


Positively Skewed Distribution: Examples

- Income Distribution: In many countries, a small number of people have very high incomes compared to the majority. This creates a long tail on the right side of the income distribution curve.
- Real Estate Prices: Housing prices can be positively skewed, with most homes priced in a lower range, but a few very expensive properties stretching the price distribution to the right.
- Time to Complete a Task: If most people complete a task relatively quickly but a few take significantly longer, the distribution of completion times might be positively skewed.

Negatively Skewed Distribution

- If a distribution is negatively skewed (or skewed to the left), then the mean is usually smaller than the median and the histogram will have a long tail to the left.



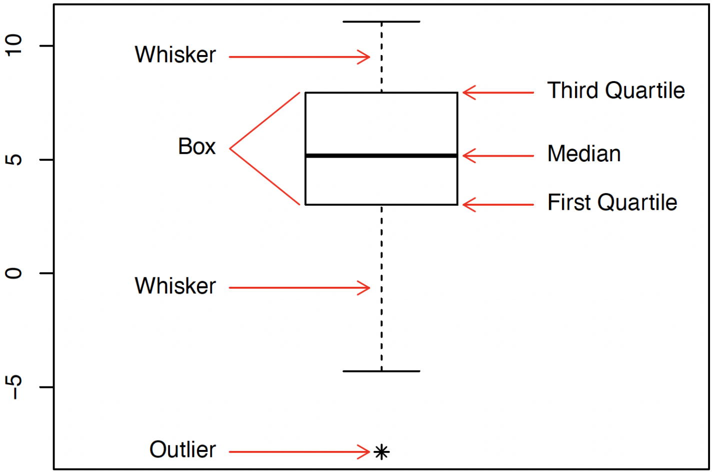
Negatively Skewed Distribution: Examples

- Survival Time of Cancer Patients: For a certain type of cancer, most patients may survive for many years after diagnosis (e.g., 5-10 years), but some patients may pass away shortly after diagnosis. This results in a negatively skewed distribution of survival time.
- Debt Repayment Time: Among debtors, most people may repay their debt close to or by the due date, but a small portion might repay their debt much earlier, leading to a negatively skewed distribution of repayment time.

Boxplots

- Boxplots are another way to display continuous data.
- They can be useful, but do not provide as much information as a histogram.
- Based on five main statistics:
  - The minimum and maximum observations.
  - The first quartile Q_1, the median Q_2 and the third quartile Q_3.
- And they look like...

Boxplots

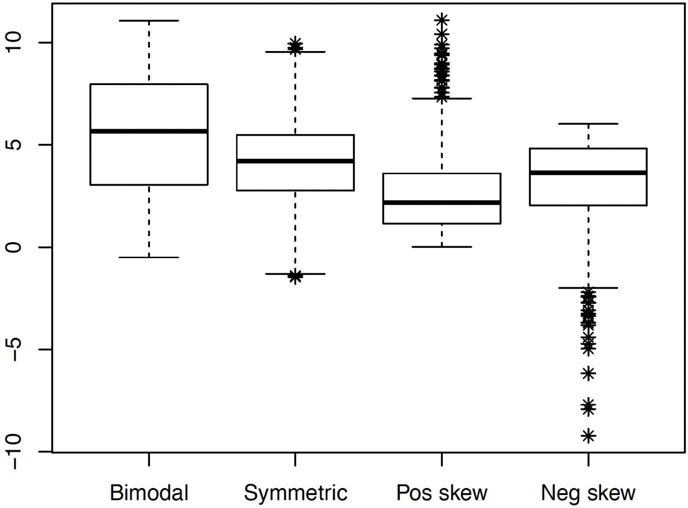

10
5
0
−5
Whisker
Third Quartile
Box
Median
First Quartile
Whisker
Outlier

Boxplots

Constructing a boxplot:

- Draw three lines at Q_1, the median and Q_3. These three lines will form the box.
- Calculate the interquartile range (IQR).
- Draw a line that extends from the first quartile to either the smallest observation or 1.5 times the IQR, whichever distance is shorter, and from the third quartile to either the largest observation or 1.5 times the IQR, whichever distance is shorter.
- These two lines drawn then form the whiskers.
- Any observations which lie further outside these whiskers are marked with stars and are called outliers.

Boxplots


Bimodal
Symmetric
Pos skew
Neg skew