Department of Linguistics and Translation
香港城市大學
City University of Hong Kong

Fundamentals of Statistics for Language Sciences LT2206

Jixing Li

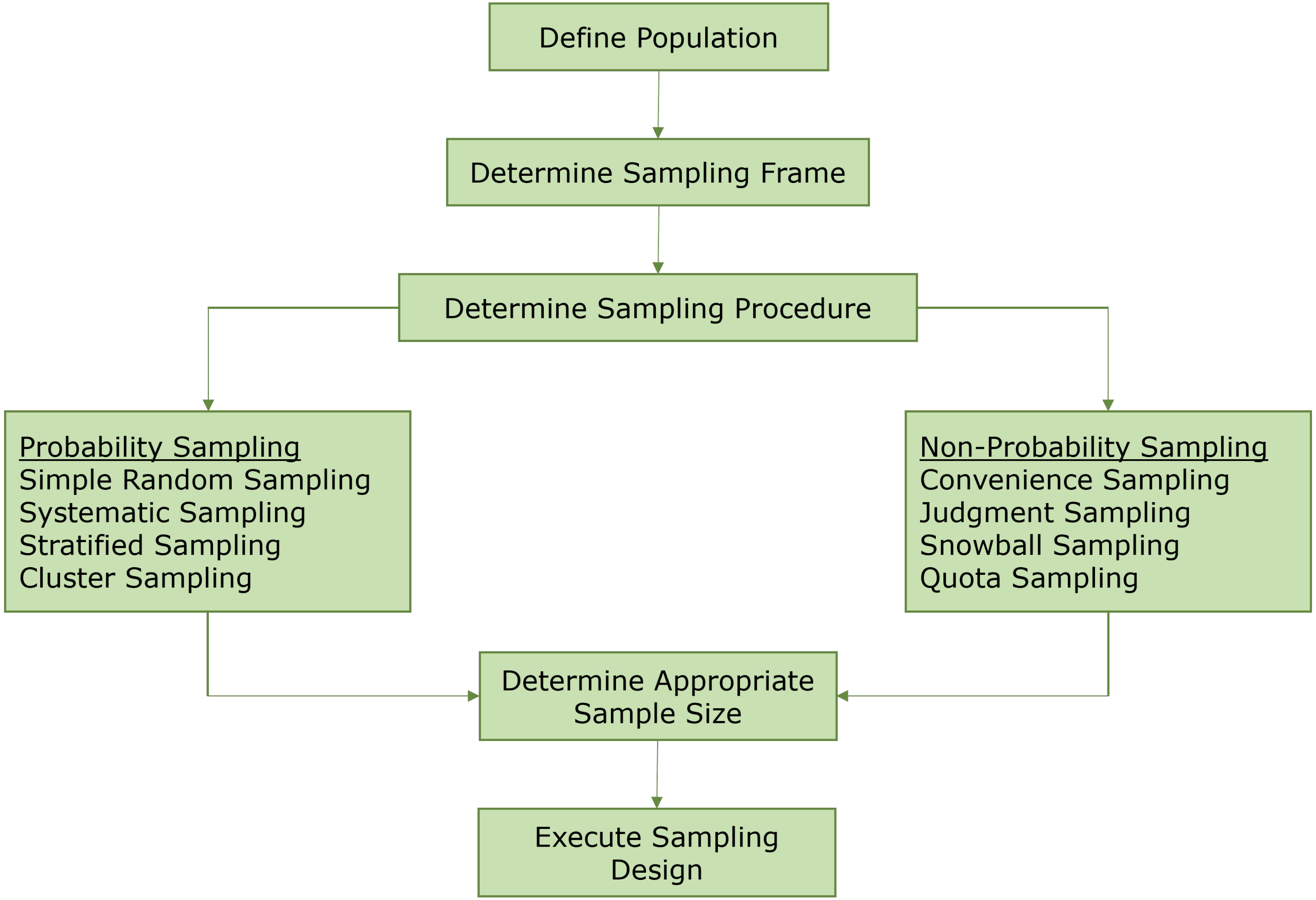
Lecture 3: Descriptive statistics

Slides adapted from Cecilia Earls

Lecture plan

- Review on sampling
- Descriptive statistics
- Short break (15 mins)
- Hands-on exercises

Sampling design process



Descriptive statistics

Basic goal: Understanding your data.

- Summary & description
- Look for peculiarities (unusual data values)

Should always be the first step in any data analysis!

Example: Survivorship on the Titanic

Goal: Describe survival patterns for the ill-fated passengers of the Titanic.

The Titanic dataset:
n = 891 passengers
Variables

- age (in years)
- gender (male, female)
- class (1,2,or 3)
- survived (yes=1, no=0)

Data types

Qualitative variables: categorical; vary in “level”, but lack specific units of measure

- **Nominal:** survived (yes/no), gender (male/female)
- **Ordinal:** passenger class (first, second, or third)

Quantitative variables: numerical; vary in magnitude with specific units of measure

- Passenger age (in years)

Graphical methods

Visually summarize the data to gain an understanding of the composition of the data (in this case, the Titanic passengers)

Categorical variables:

- pie charts
- bar charts

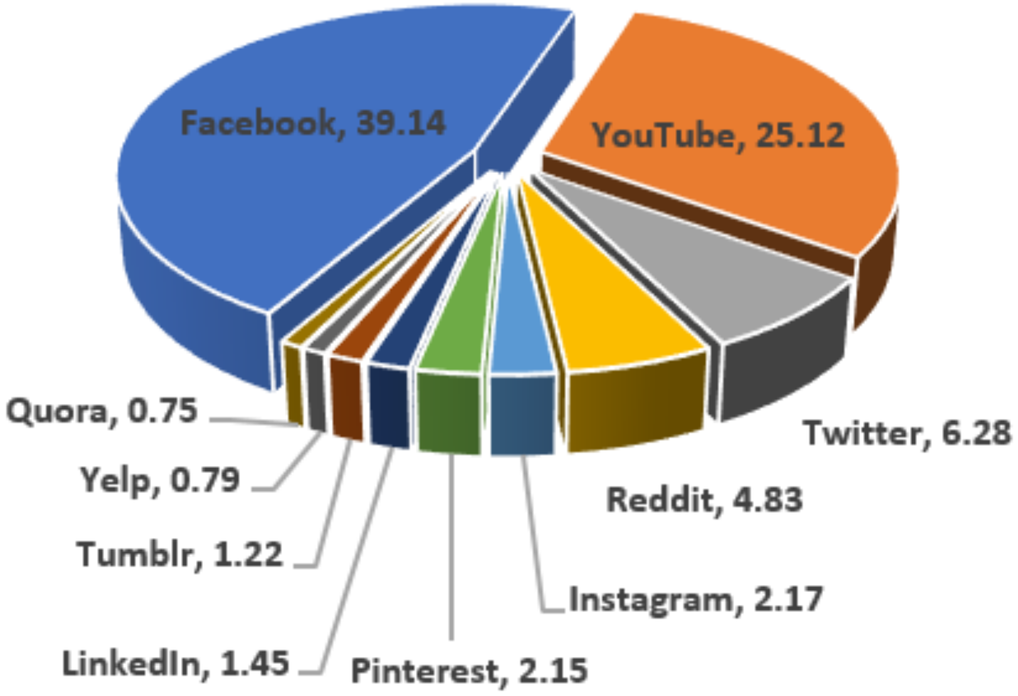
Quantitative variables:

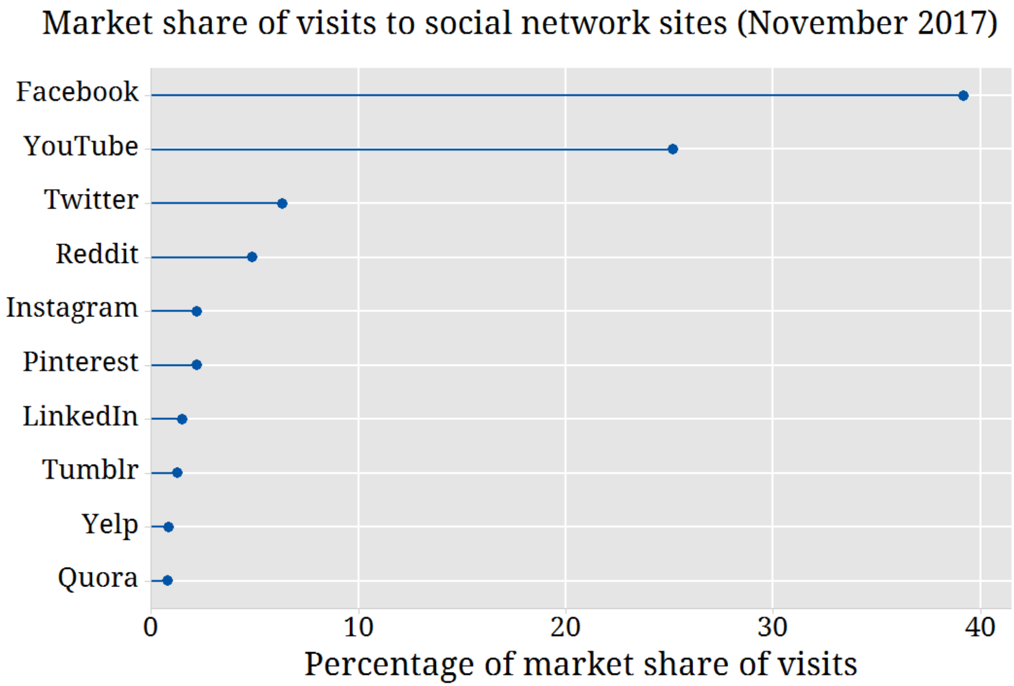
- histograms
- box plots

Categorical variables: Pie charts

Pie charts along with other area-based charts (e.g. donut chart) are not recommended!

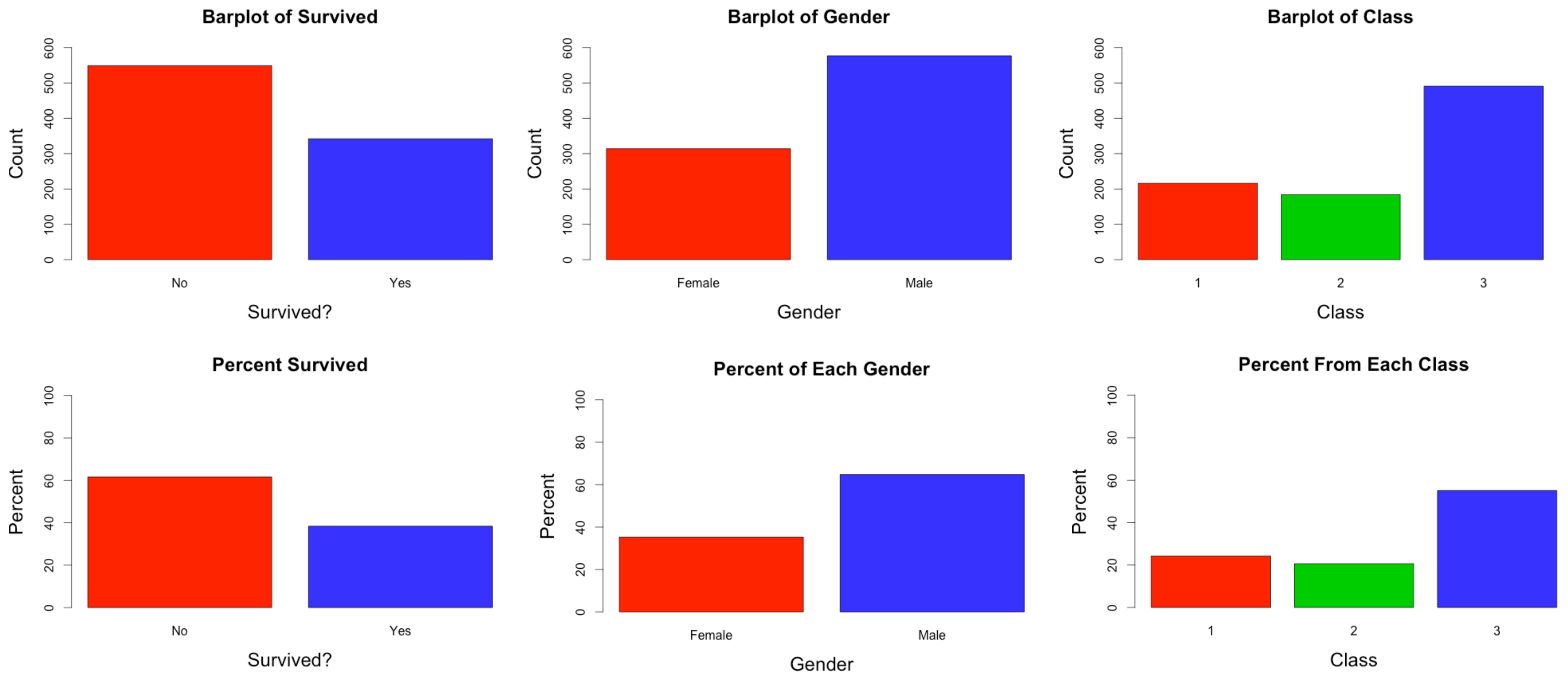
- Difficult to decode the information in the data.
- Completely defeats the purpose of including a chart instead of a table.





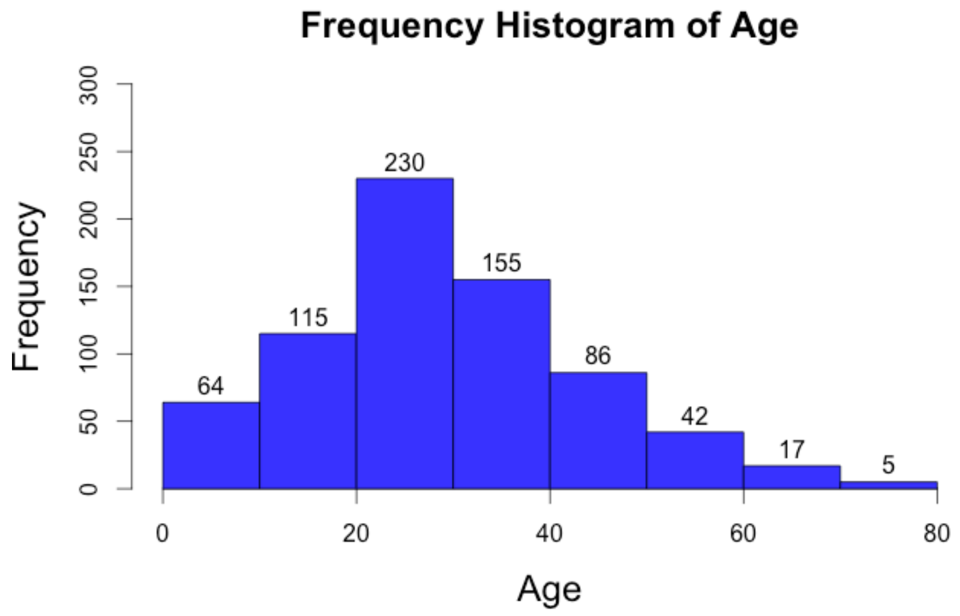
Categorical variables: Bar charts

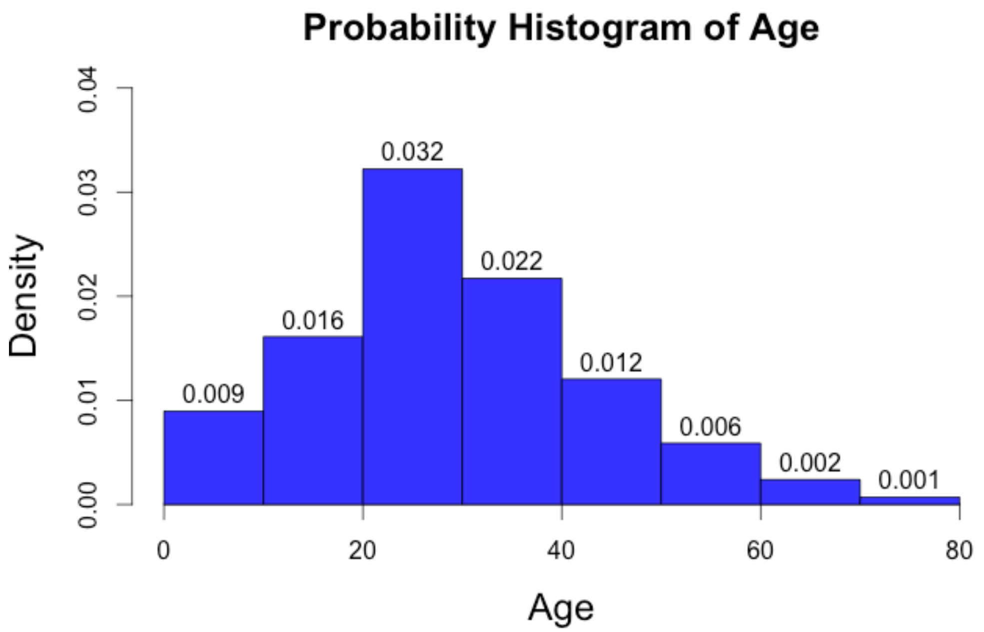
Displays the total number or percent of observations falling in each category



Quantitative variables: Histograms

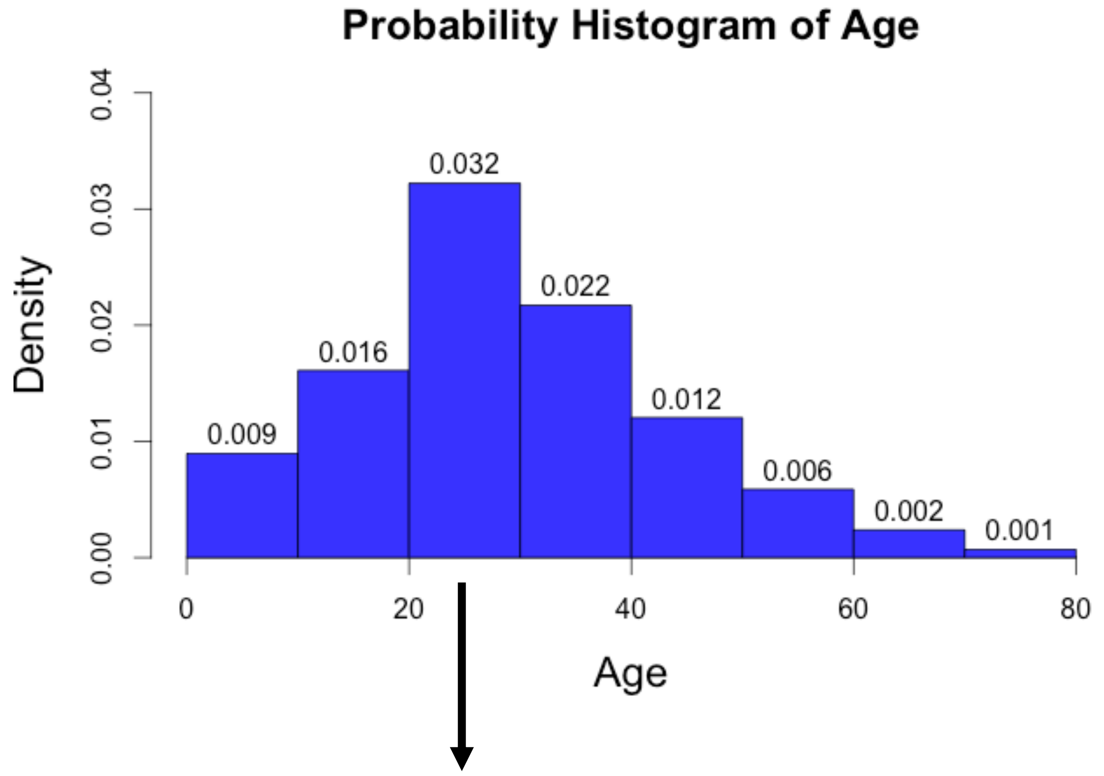
- Class-specific **counts** or **relative frequencies** are summarized in a bar-type plot
- Used to summarize the shape of the distribution, assess spread, and look for "extreme" values
- Particularly useful for "large" datasets (>30 observations)

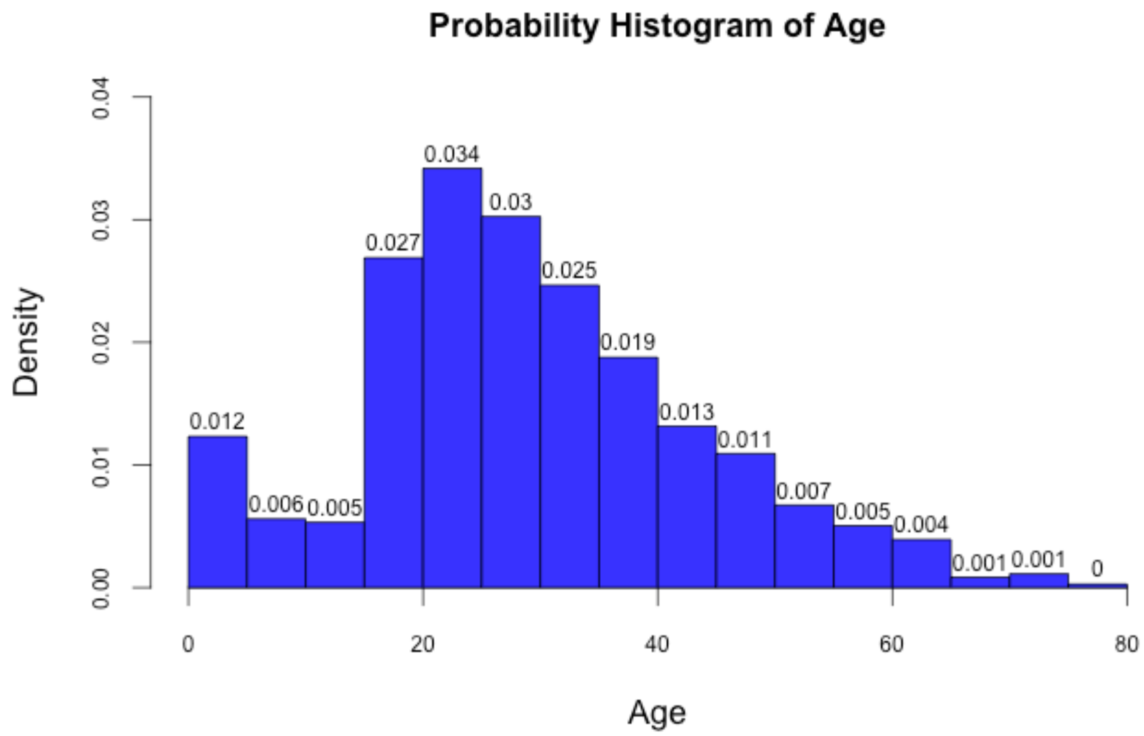




Probability histograms

Height of each bar = Probability per unit age for that group.



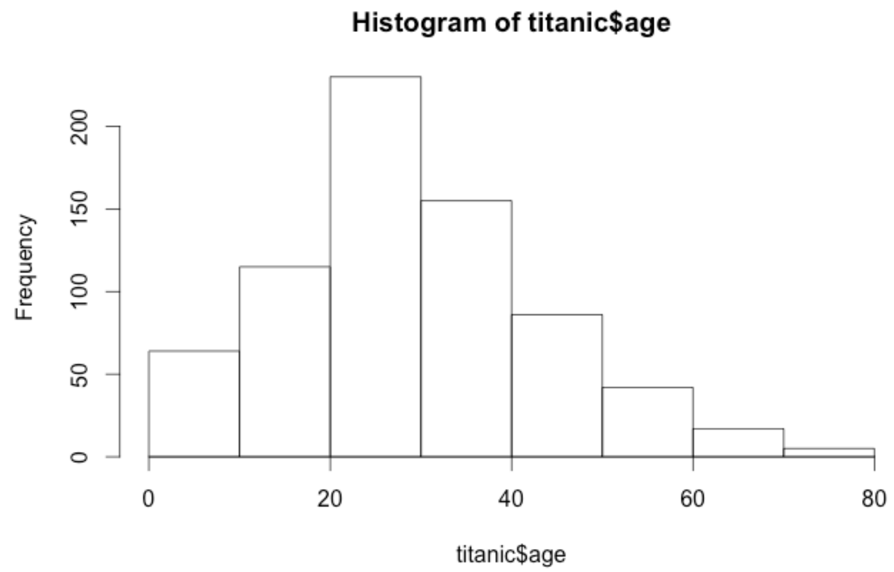


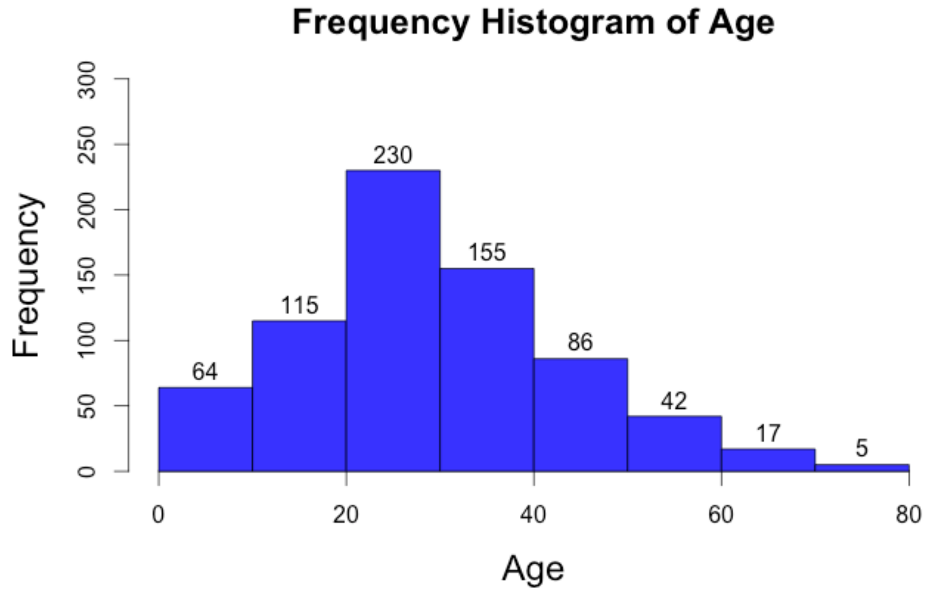
P(age between 20 and 30) = 0.032 x 10 = 0.32

Histograms in R

```
titanic = na.omit(titanic)
hist(titanic$Age)

hist(titanic$Age,
ylim=c(0,300),
col='blue',
cex.lab=1.5,
cex.main=1.5,
xlab='Age',
main='Frequency
Histogram of Age',
labels=TRUE)
```

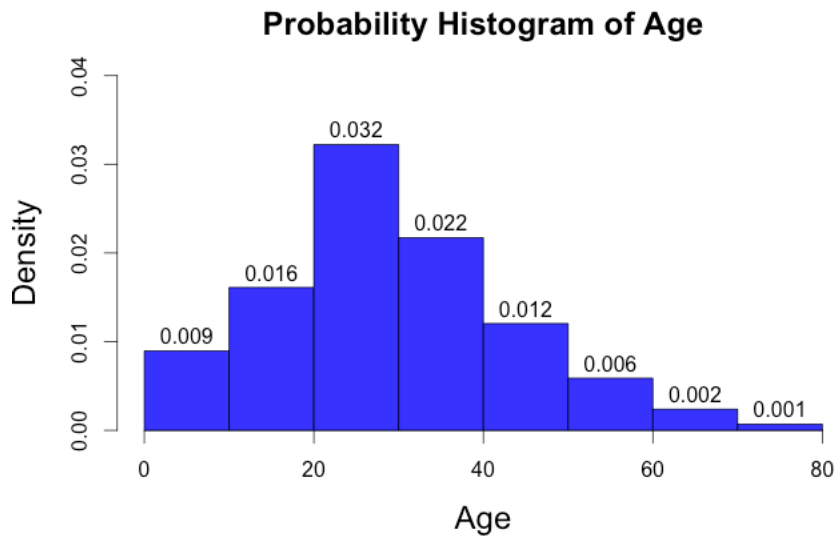


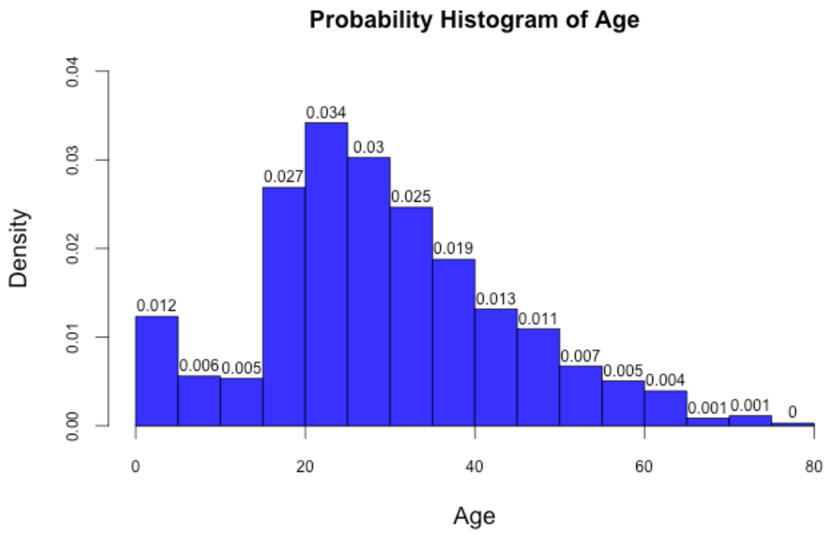


Histograms in R

```
hist(titanic$Age,
freq=FALSE,
ylim=c(0,0.04),
col='blue',
cex.lab=1.5,
cex.main=1.5,
xlab='Age',
main='Probability Histogram of Age',
labels=TRUE)
```

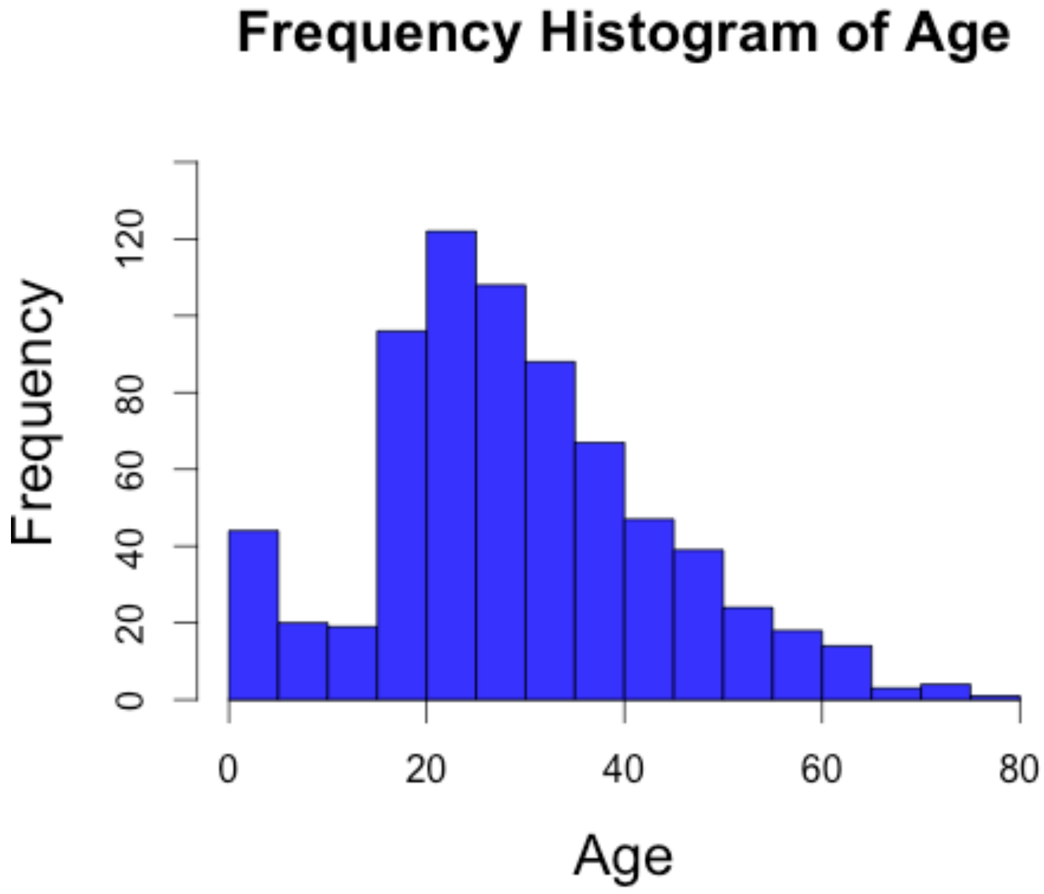
```
hist(titanic$Age,
freq=FALSE,
breaks=15,
ylim=c(0,0.04),
col='blue',
cex.lab=1.5,
cex.main=1.5,
xlab='Age',
main='Probability Histogram of Age',
labels=TRUE)
```





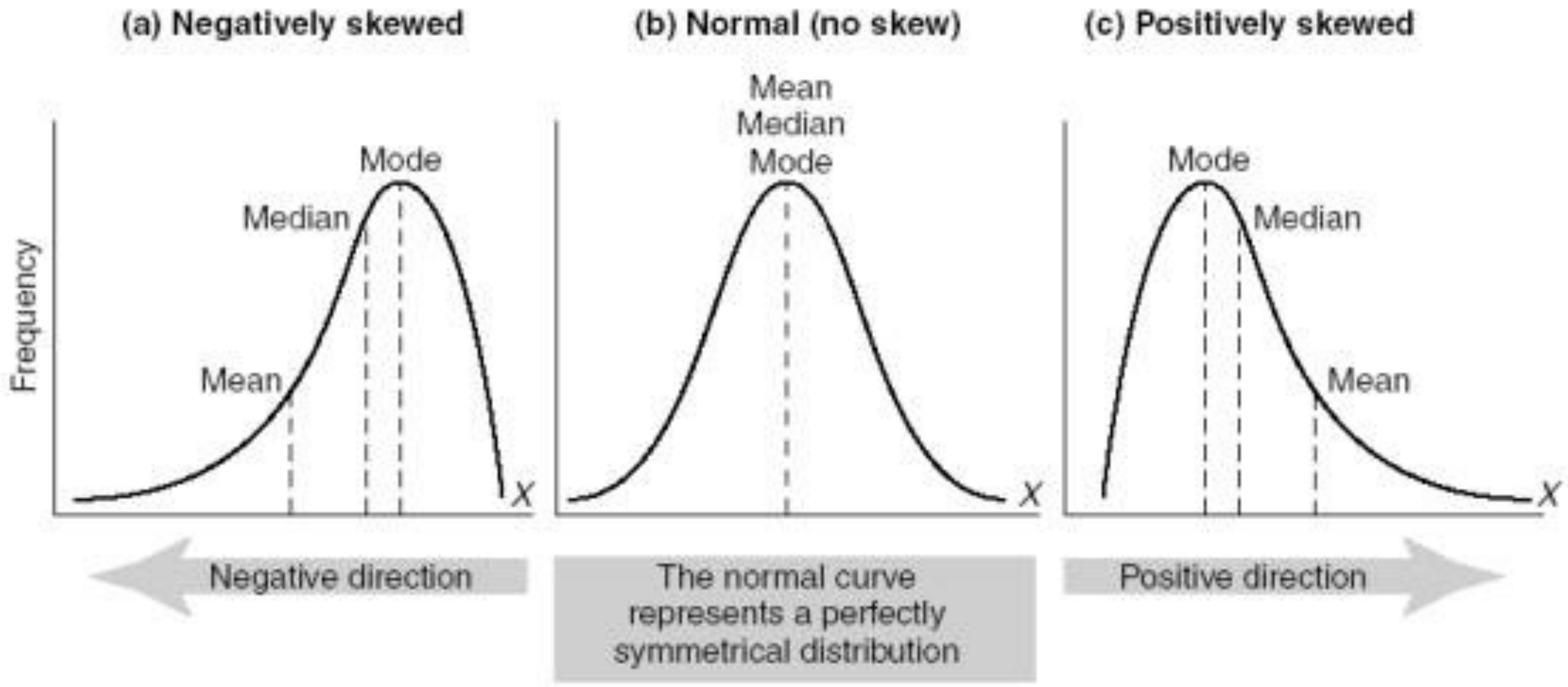
Common descriptive features

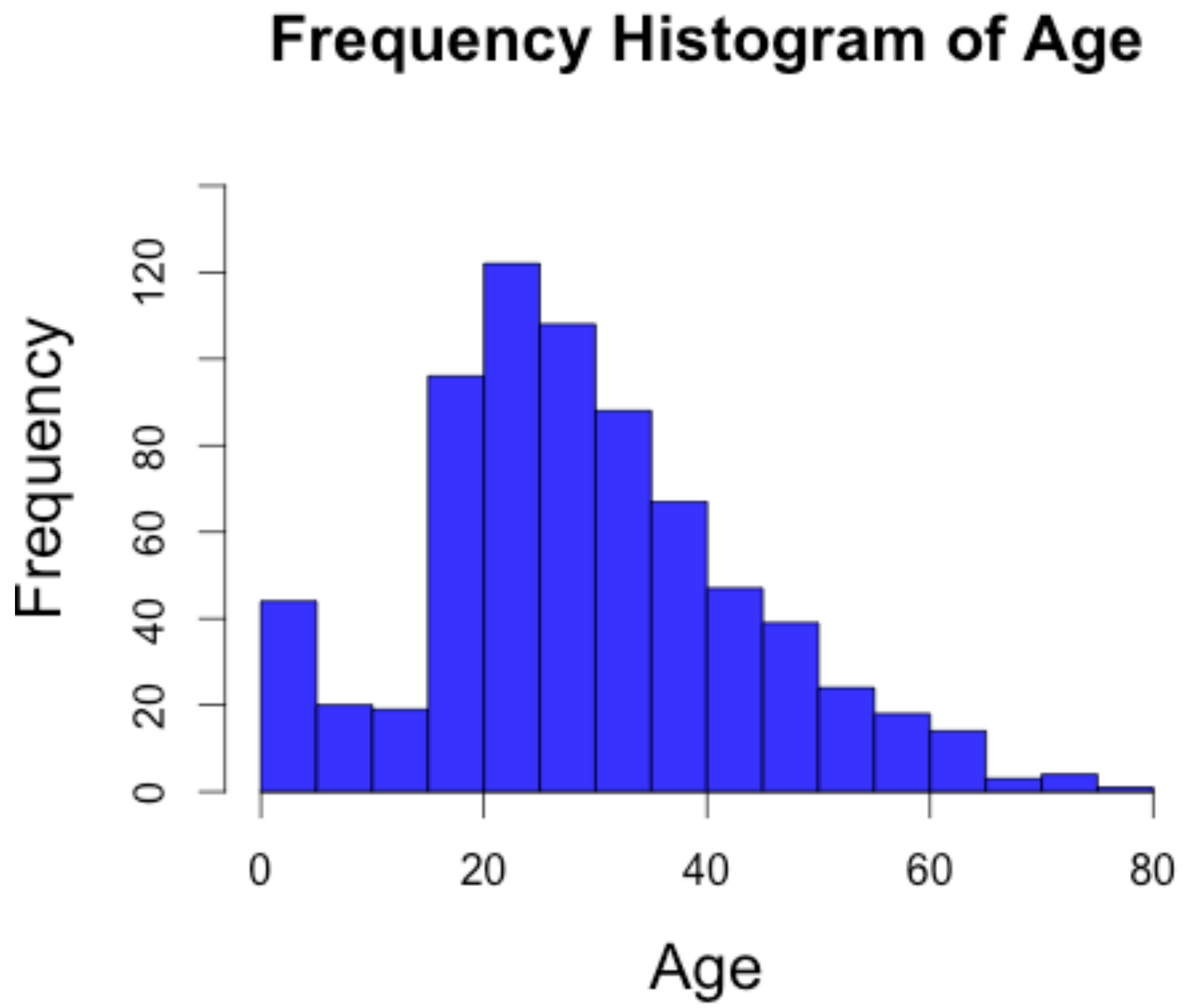
- **Center:** Where is the "middle"?
- **Spread:** How much individual to individual variation exists?
- **Clustering** (number of modes):
  - No bumps: uniform
  - 1 bump: unimodal
  - 2 bumps: bimodal
- **Skewness:** Symmetry? Or is one "tail" longer than the other "tail"?
- **Outliers:** Are there extremes that stand out in the data?



Unimodal distributions

left-skewed symmetrical right-skewed





Slightly skewed to the right
Mean = 29.70 years
Median = 28.0 years
Mode = 24 years

Numerical methods

To extract meaningful information for purposes of description and comparison; to reduce quantitative data to a few “talking points”.

Statistic: A numerical summary computed from a sample of data on a quantitative variable.

- Population version of these numerical summaries is generally unknown
- One of the goals of the field of statistics is to **estimate** the population numerical summaries

Numerical measures of central tendency

- **Mean:**
  - Average value
- **Median:**
  - "Midpoint" of the data when values are ordered from smallest to largest (50th percentile)
- **Mode** (meaningful with categorical data):
  - Measurement that occurs most often (most "popular")
  - Multiple modes are possible

Mean is sensitive to the magnitude of all data values, but the median is not. Median may be a more useful statistic for skewed data

Mathematical notation for the sample mean

In general, we can represent a generic sample of ***n*** data points as an indexed list (order irrelevant):

$$x_1, x_2, x_3, \ldots, x_n$$

Sample mean = arithmetic average:

$$\bar{x} = \frac{(x_1 + x_2 + \cdots + x_n)}{n}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Sensitive to the magnitude of each data value.

Numerical measures of sample dispersion, spread and variability

- **Range:**
  - maximum data value - minimum data value

- **Rth percentile** (also called **quantiles**):
  - Sort data values from smallest to largest, find the value that has at most R% of measurements below it and at most (100 – R)% above it ($0 \leq R \leq 100$).

- **Quartiles:**
  - 1st, 2nd and 3rd are the 25th, 50th, and 75th percentiles, respectively
  - **IQR (interquartile range)** = 3rd quartile – 1st quartile

Sample variance and standard deviation

Measures dispersion of individual data points about the average

Sample variance: expressed in squared units of x.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

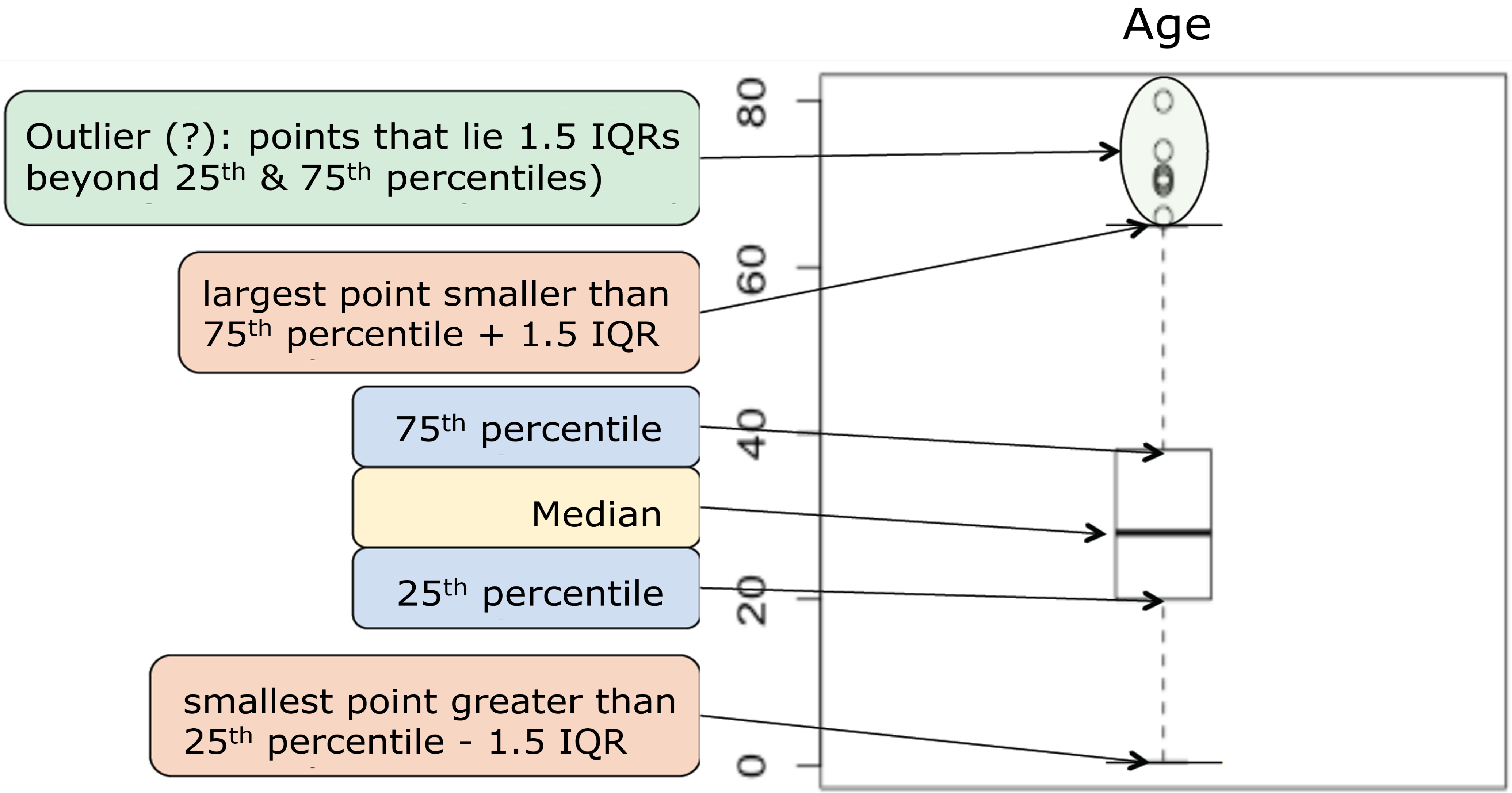
Sample standard deviation: expressed in the same units as x.

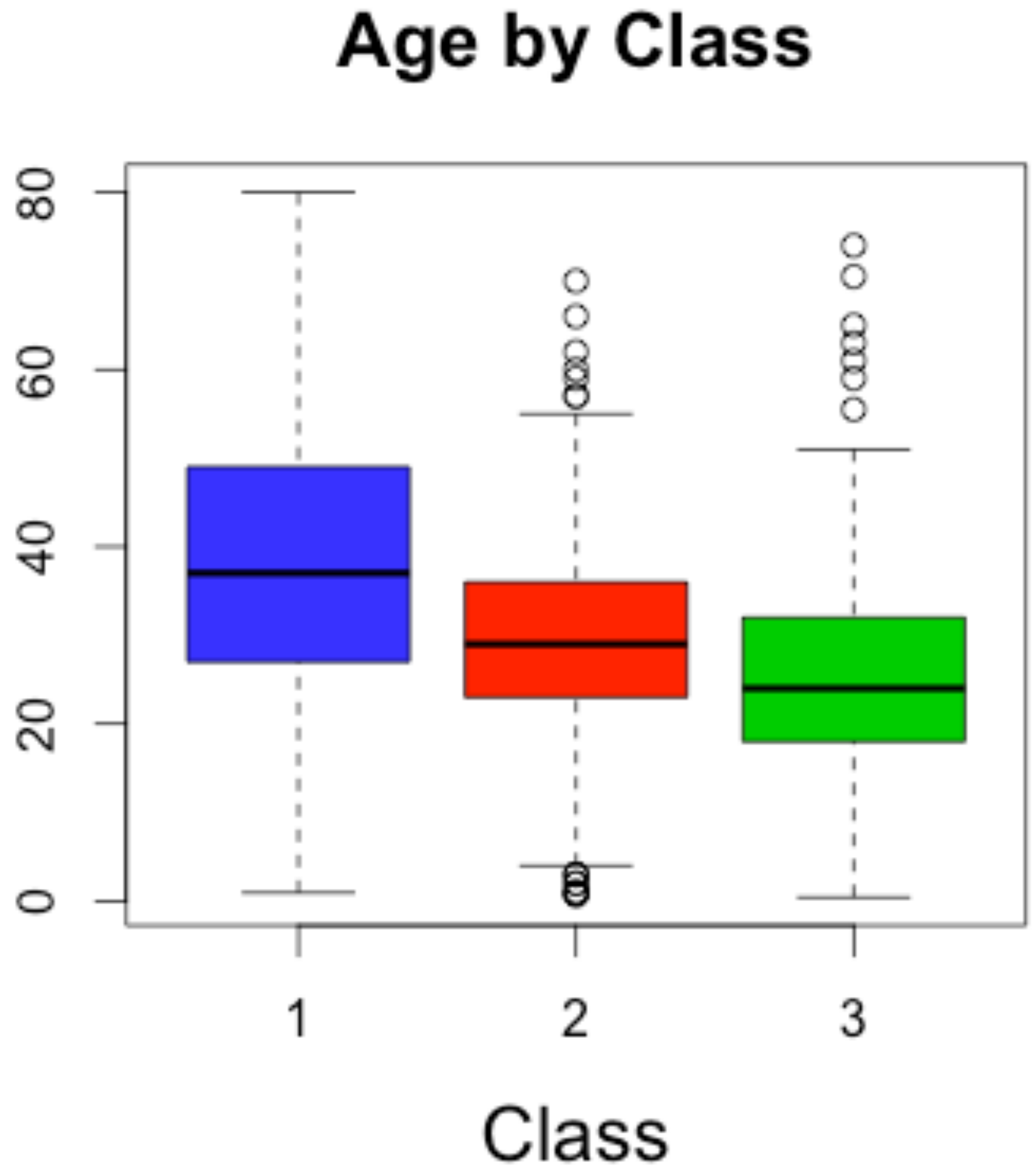
$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Sample variance and standard deviation

- Sensitive to the magnitude of each element.
- Used frequently, but may not be very informative about shape in cases where data are highly skewed.
- Both are always non-negative.
- Both equal zero if, and only if, all data values are exactly the same.

Quantitative variables: Boxplots





- Median age is noticeably higher in 1st class, but much closer for 2nd and 3rd class.
- Outliers appear to be present in the 2nd and 3rd classes.

Relationships between two or more variables

Quantitative vs. Qualitative

- Side-by-side boxplots (used for concrete data)

- **Quantitative vs. Quantitative**
  - Scatterplots
- **Quantitative-Quantitative-Qualitative**
  - Coded scatterplots
- **Qualitative vs. Qualitative**
  - Stacked / unstacked bar charts, contingency tables

Example: Canadian Prestige Data

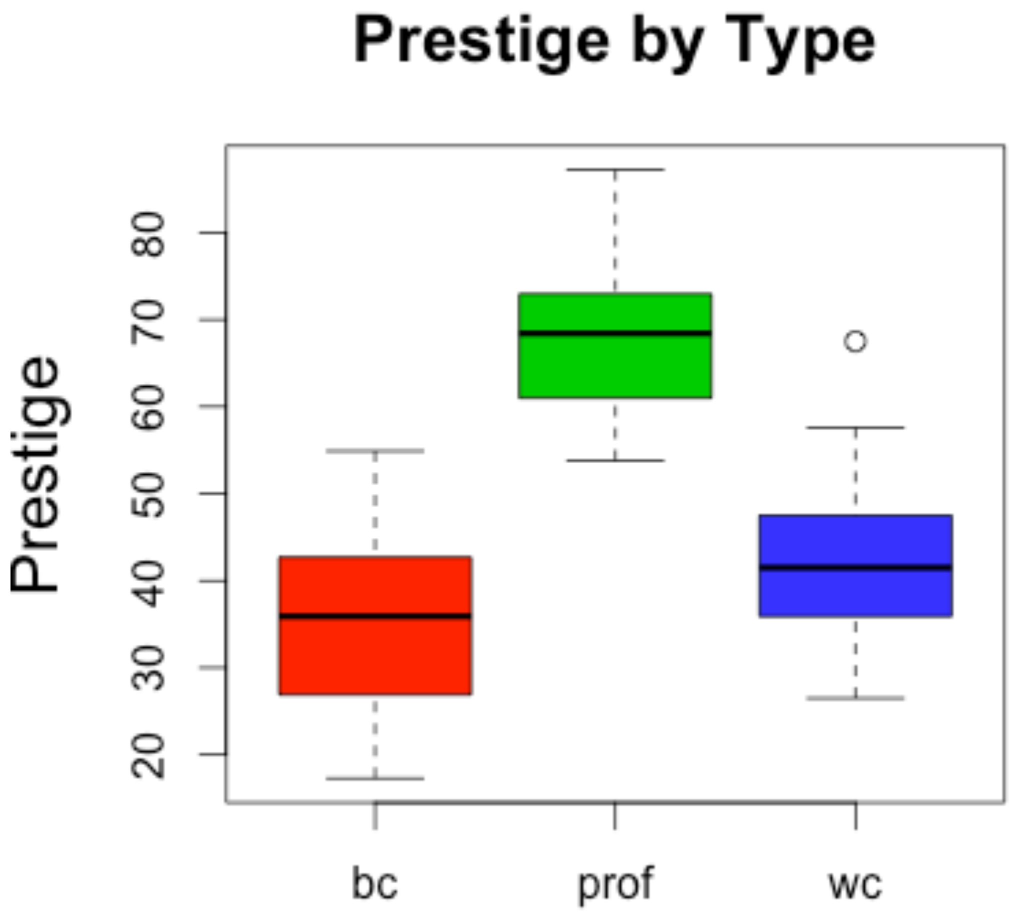
Data obtained on Canadian workers from 98 different occupations in 1971.

- **Education:** Average education of subjects working in a given occupation in 1971 (years after grade 4)
- **Income:** Average income (1971 Canadian dollars)
- **Women:** % of women in a given occupation
- **Prestige:** Occupational prestige score
- **Occupation class:** Blue Collar (bc), Professional, Managerial, and Technical (prof), White Collar (wc)

Quantitative vs. Qualitative

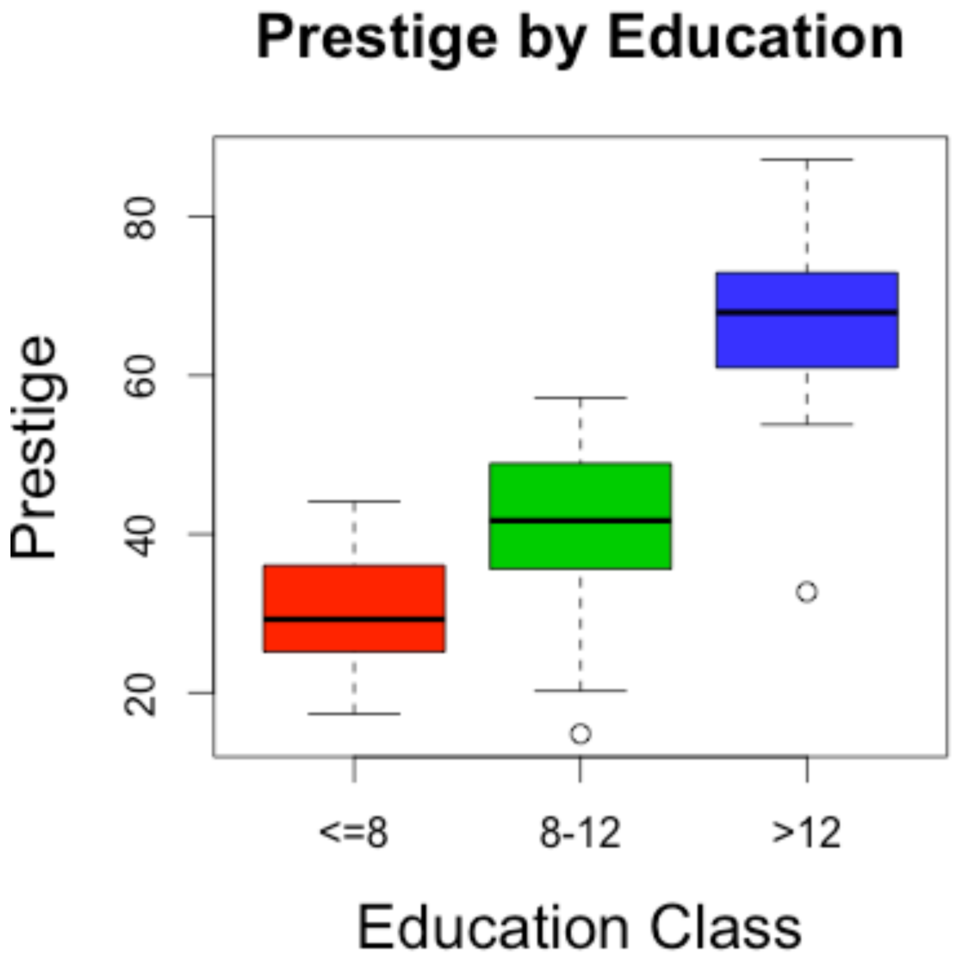
Side-by-side boxplots allow us to quickly compare medians, variability and general shape of a quantitative variable for different levels of a qualitative variable.

- Professionals enjoy higher prestige ratings than blue or white collar workers.
- White collar workers may have slightly higher prestige than blue collar workers.
- Variability in prestige looks similar across occupation classes



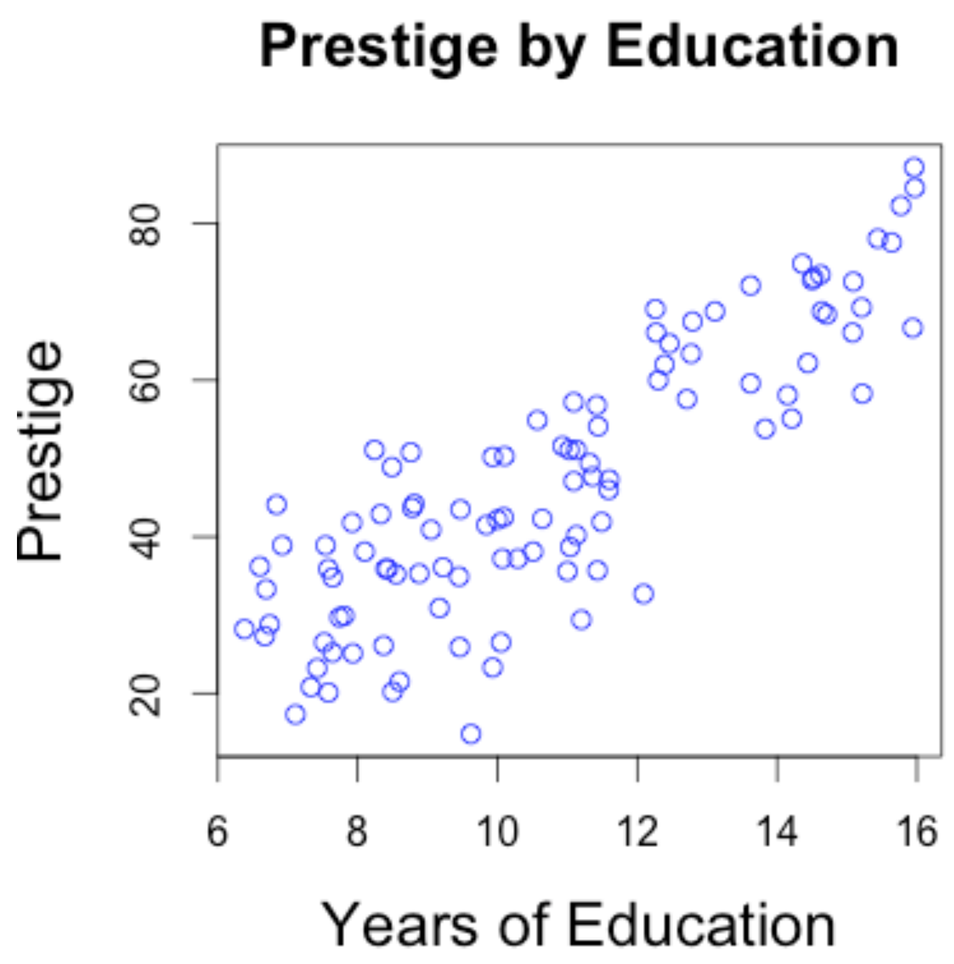
Boxplots of prestige vs. education can be constructed by categorizing education.

- Prestige increases with education.
- The greatest difference is between those who have education beyond high school and those who do not.

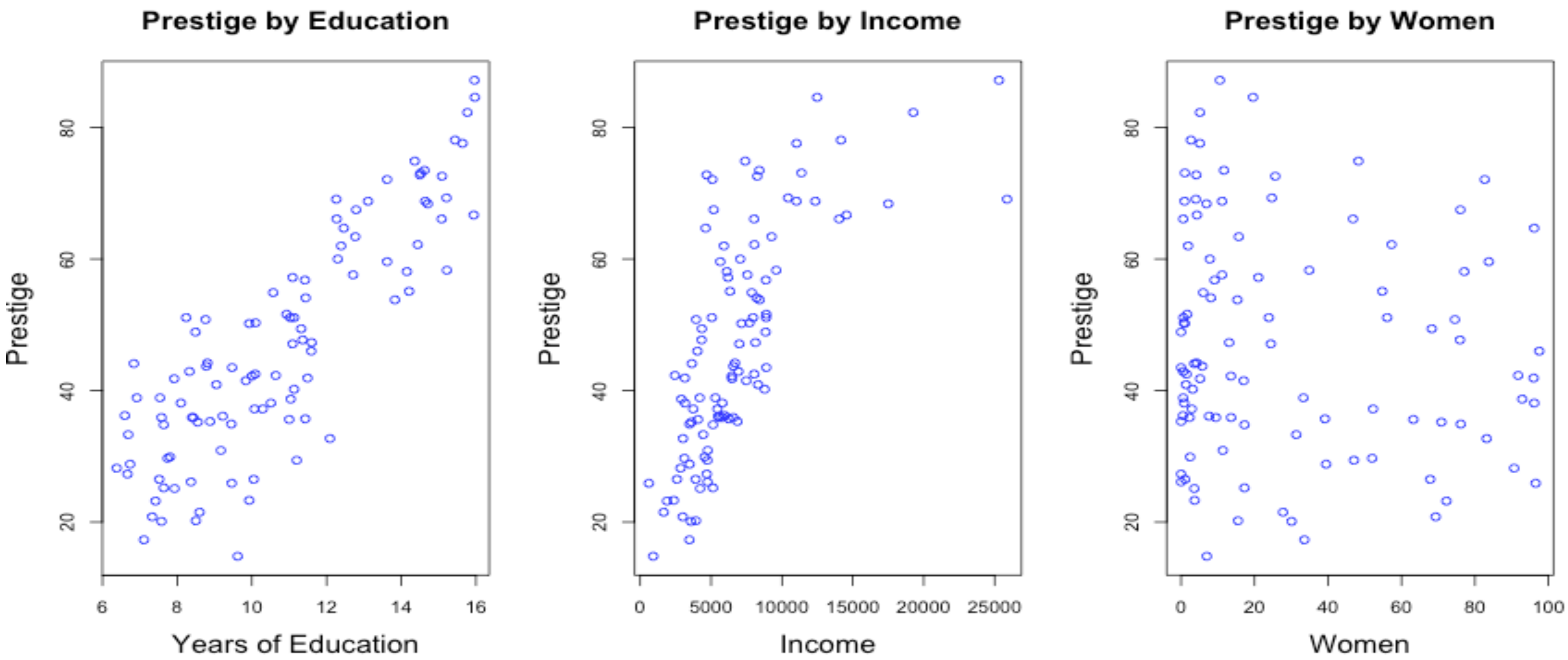


Quantitative vs. Quantitative

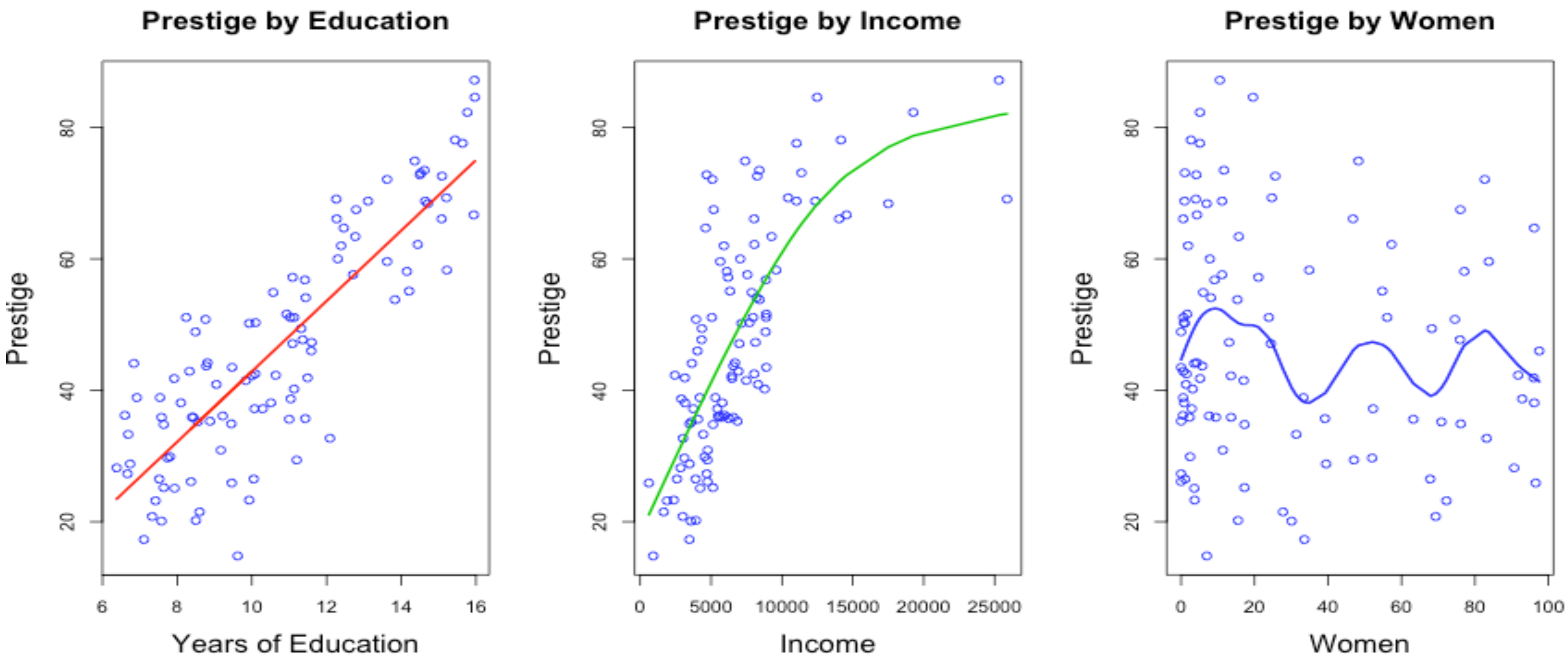
Scatterplots are the best way to visualize the relationship between two quantitative variables.



- **x-axis**
  - Predictor
  - Explanatory variable
  - Independent variable
- **y-axis**
  - Response
  - Dependent variable

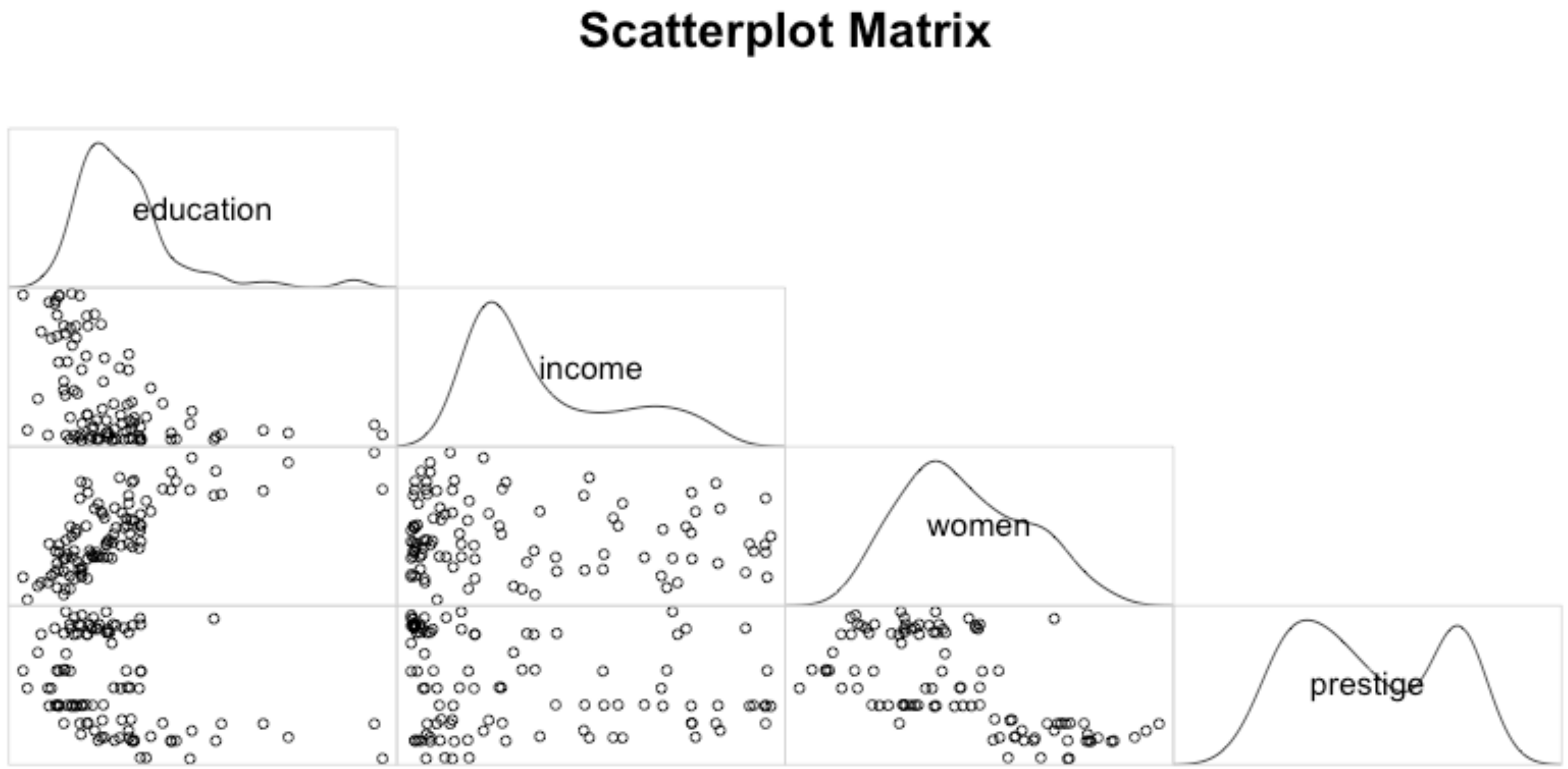


Direction: Does “Y” increase or decrease with “X”?
Trend: Linear?
Strength of association: Amount and width of scatter?
Outliers: Any unusual or extreme observations?



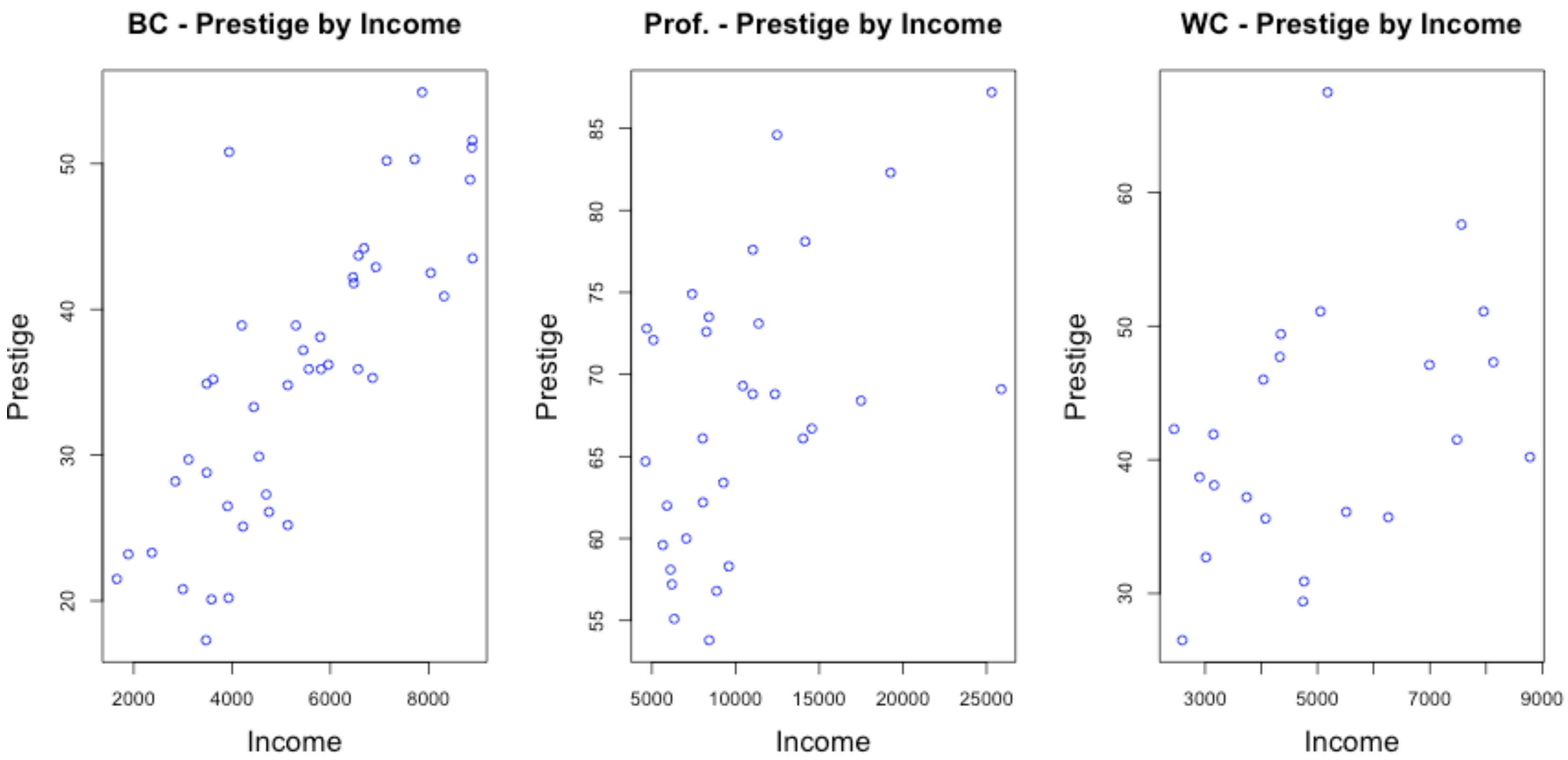
Smooth curves can be fit to scatterplots to help visualize trends.
The most common “smooth” curve is a line (**linear regression**).

Scatterplot matrices can be used to look at all pairwise relationships between quantitative variables.



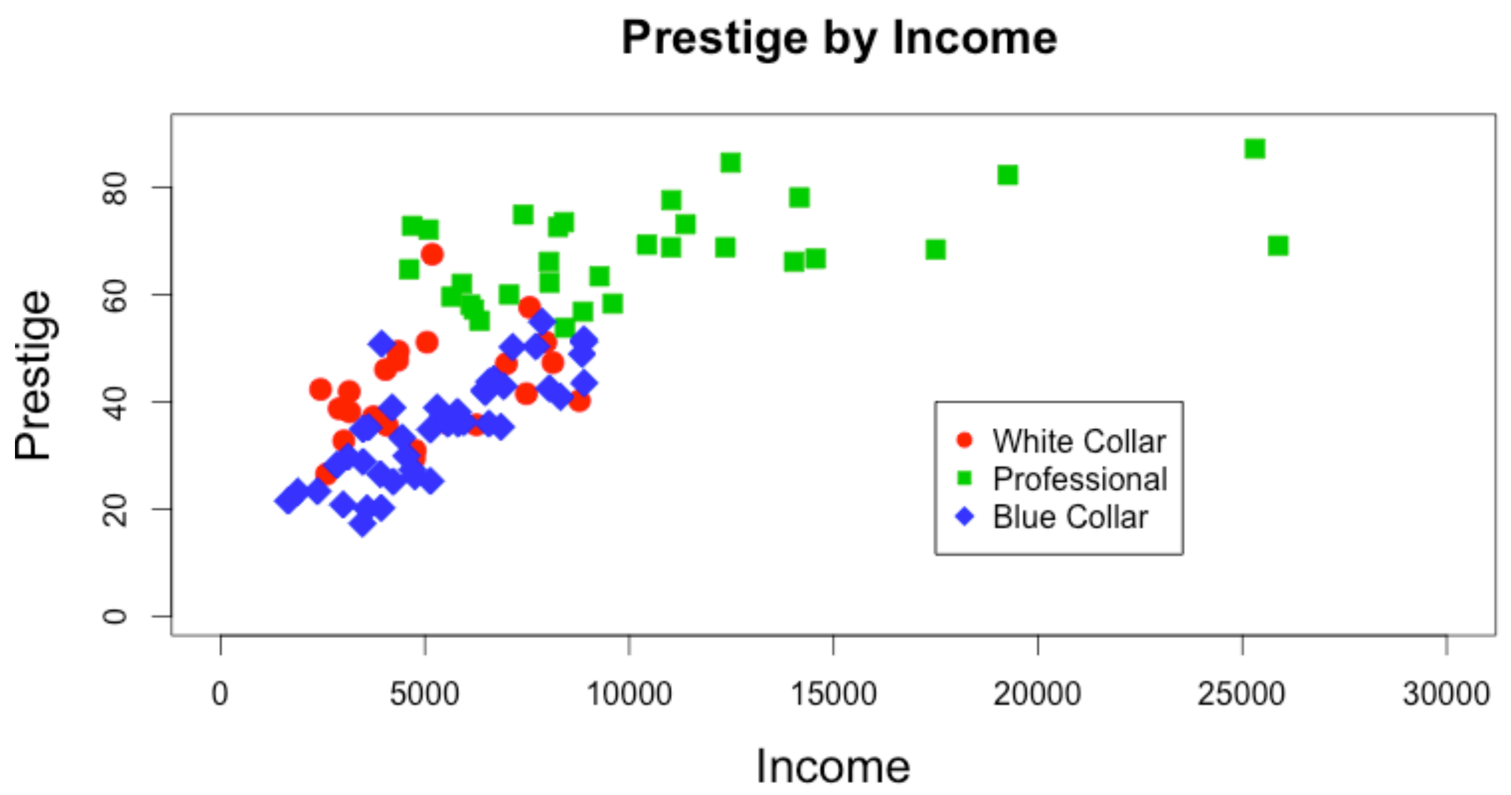
Quantitative-Quantitative-Qualitative

Question: Does the relationship between prestige and income depend on the type of profession?



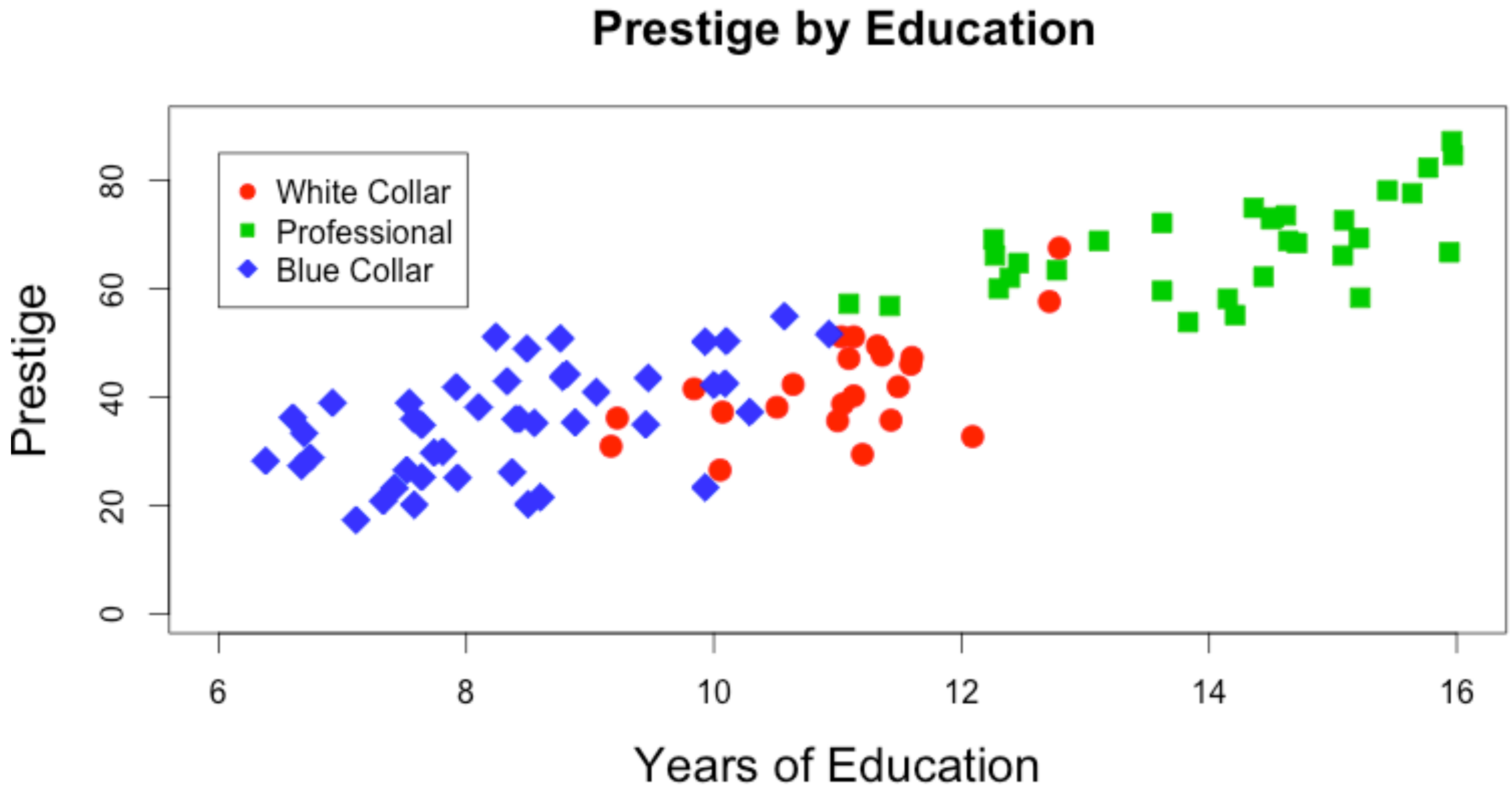
Not easy to tell from this set of plots!

Coded scatterplot: Label the income & prestige pairs by type of occupation.



- The relationship between prestige and income is similar for blue collar and white collar workers.
- Range of prestige and income values is different for professionals and the relationship is "flatter".

Relationship between prestige and education also appears to depend on occupation type.



To do

- Finish Lab 3
- Read: This lecture: Textbook Ch3
- Next lecture: Textbook Ch4