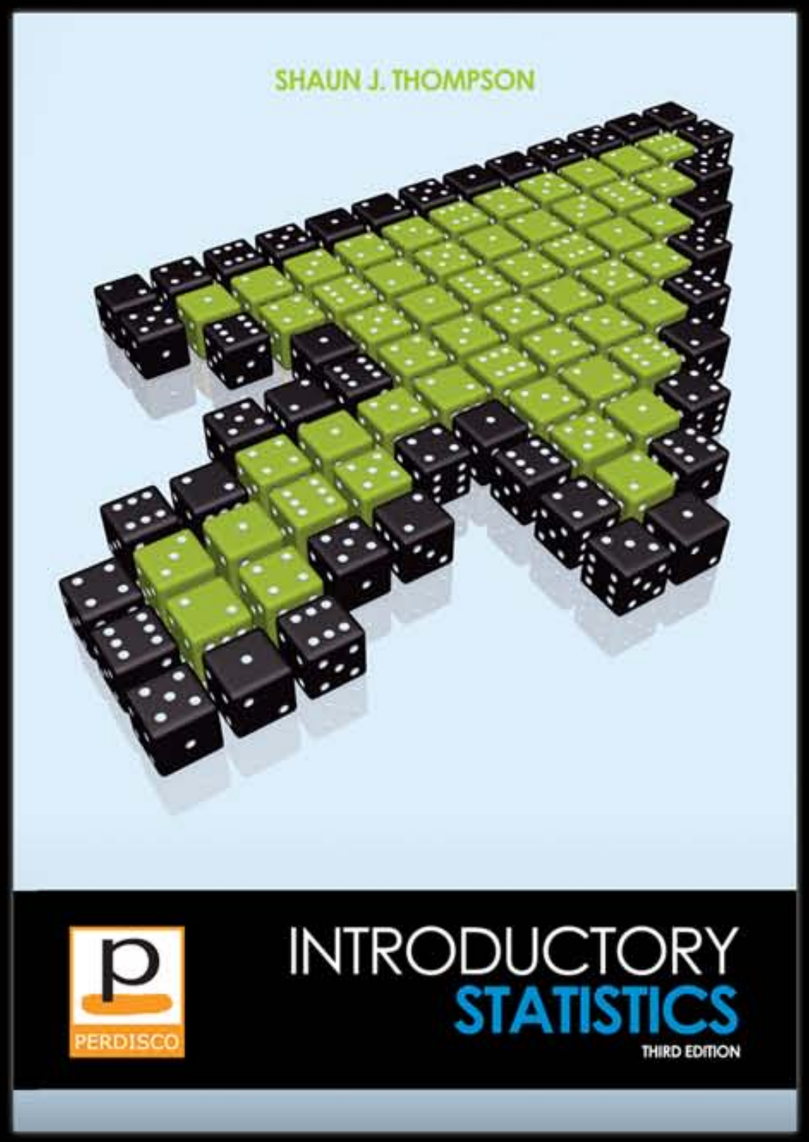


These lecture slides are designed to accompany:

Introductory Statistics, Third Edition

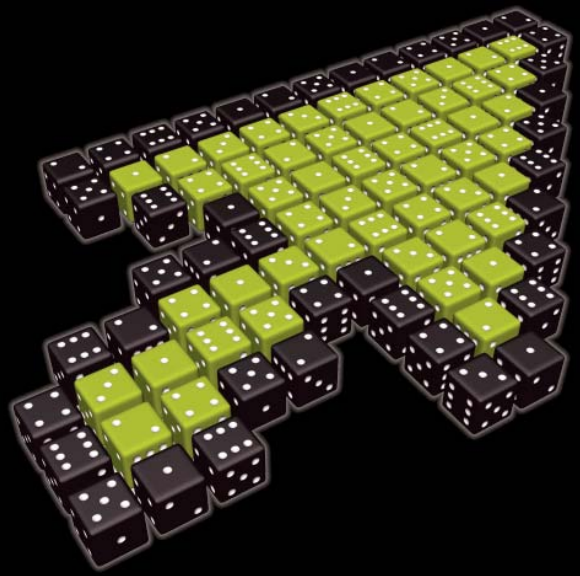
Other features include:

- Chapter summary videos
- MP3 audio podcasts
- VirtualTutor e-learning
- AntiCheat and AutoGrade homework
- Detailed instructor resources

To find out more, visit: perdisco.com/stats

Chapter 2

Presenting data



Presenting categorical data

- Categorical data start out as a list of observations
- Example: a survey asks 200 people which of three election candidates they will vote for
- Data are list of 200 responses
- This is long and unwieldy – we need to summarize
- This is what presenting data is all about: putting raw data into a form that is ‘easy to read’!

Survey answers
Candidate B
Candidate A
Candidate B
Candidate C
.
.
.

Counting

- We want to read data so that we can answer questions about it
  - e.g. Which candidate got the most votes?
- First step is to count the data – count the number of times each category is observed
- Example: Candidate A got 54 votes, Candidate B got 104 votes, Candidate C got 42 votes
- Then you can answer questions about the data

Frequency table

- A **frequency table** is a table that shows the data we've counted

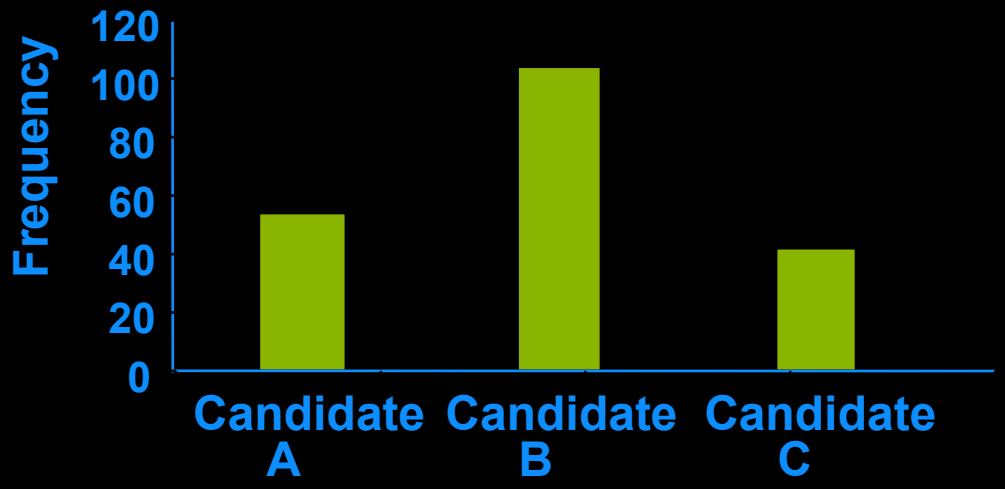
Candidate	Frequency
A	54
B	104
C	42

- A **relative frequency table** shows the **proportion** of observed values for each category

Candidate	Relative frequency (%)
A	27
B	52
C	21

Bar chart

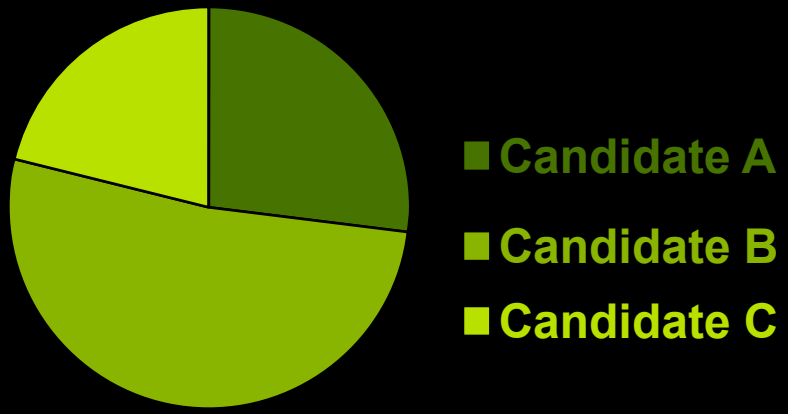
- Used to give a more ‘graphical’ depiction of data
- A bar is drawn for each category
- Height represents observed frequency in data



- Can be used to detect trends in data

Pie chart

- Another graphical depiction of data
- The ‘pie’ is a circle, divided up into slices
- Each slice represents one category
- The size of each slice shows the relative proportion of observed values in that category



Presenting numerical data

- Same principles apply to numerical data
- Major difference: numerical variables often have many more values

- Example: measure heart rate in beats/minute (bpm) for 200 athletes
- This data could take many different values!
- Compare this to the election survey, which had only three different values

Frequency distribution table

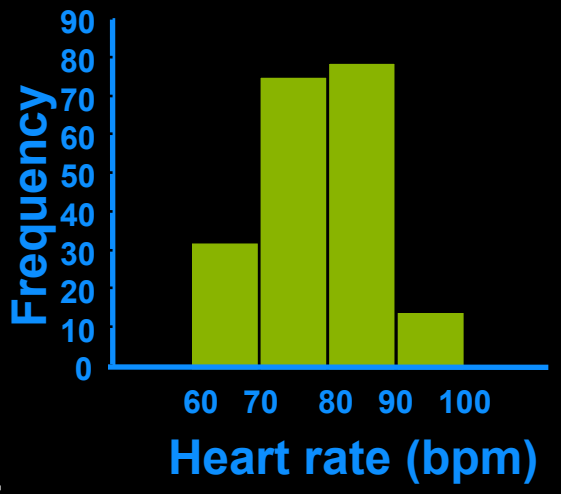
- So group values together into **classes**
- Record the number (or percentage) of observations in each class:

Class (bpm)	Frequency	Relative frequency (%)
60 to 69	32	16
70 to 79	75	37.5
80 to 89	79	39.5
90 to 99	14	7

- This is a **frequency distribution table** – and the column on the right shows the relative frequency

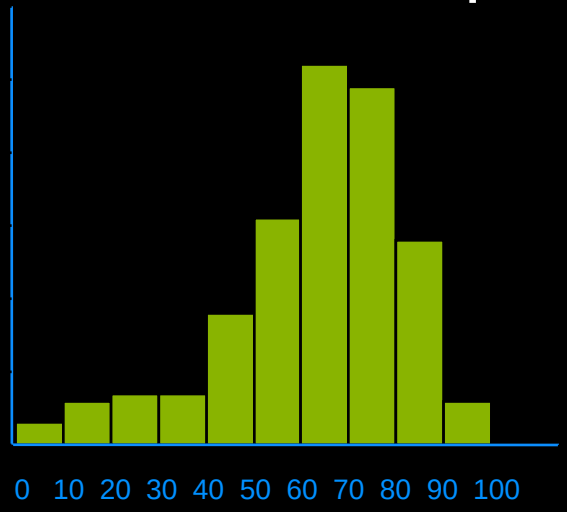
Histogram

- Similar to a bar chart
- However, there is no horizontal space between the bars here
- Like for a bar chart, the histogram can be used to detect trends in data
- However, we can typically say a lot more about trends in numerical data than in categorical data!



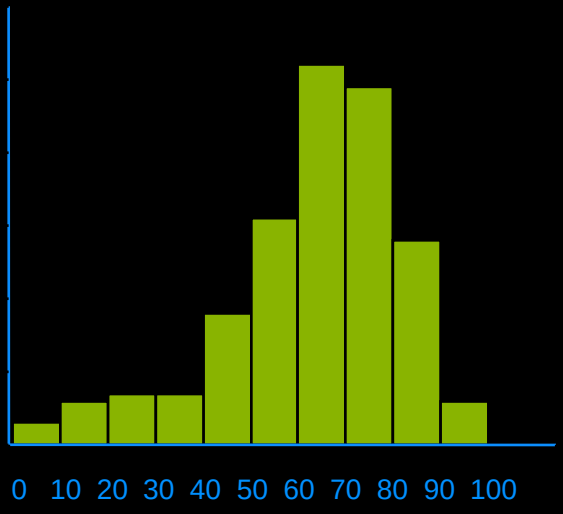
Reading the histogram: middle of the data

- A histogram can go deeper than a bar chart in describing trends in data
- Example: it can be used to estimate the ‘middle’ of a set of numerical data
- The middle is the ‘balance point’ of a histogram



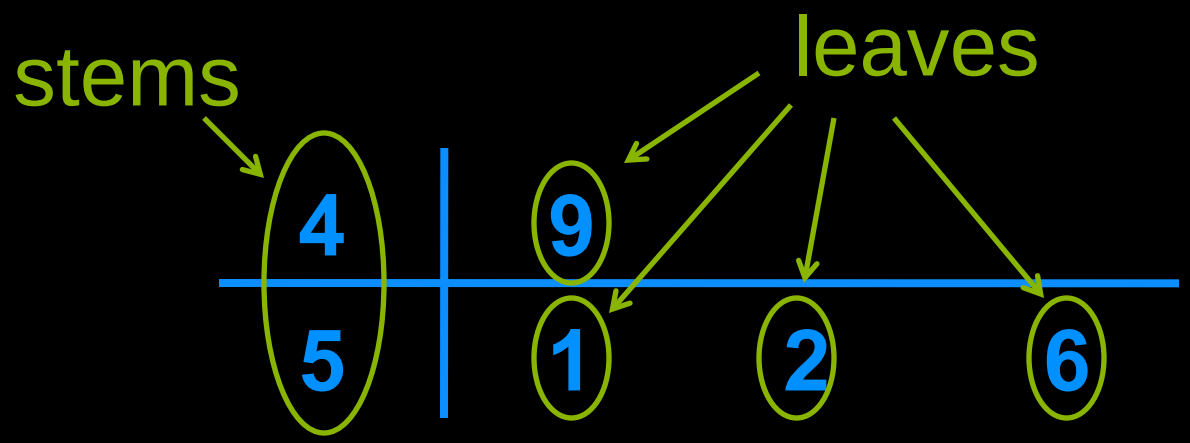
Reading the histogram: symmetry & skew

- The ‘middle’ of this histogram is around 60 to 65
- Why? Aren’t there more values above 60 to 65 than below it?
- Yes, but the low values are very low
- The histogram is not **symmetric**
- It is **skewed**
- That is, the values to the left are more spread out than the values to the right



Stem-and-leaf plot

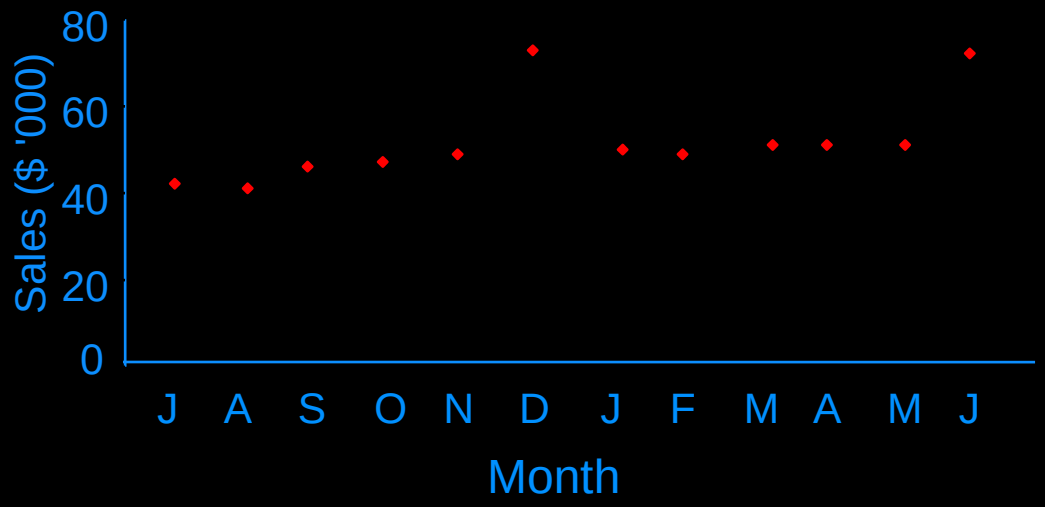
- Used when there is a small number of data values
- Example: For 4 data values 49, 51, 52, 56



- More info than in a histogram – the plot shows every value
- However, only practical for small number of values

Time plot

- Sometimes numerical data are collected over time
  - e.g. sales figures collected every month for a year
- Can use a **time plot** to present this data



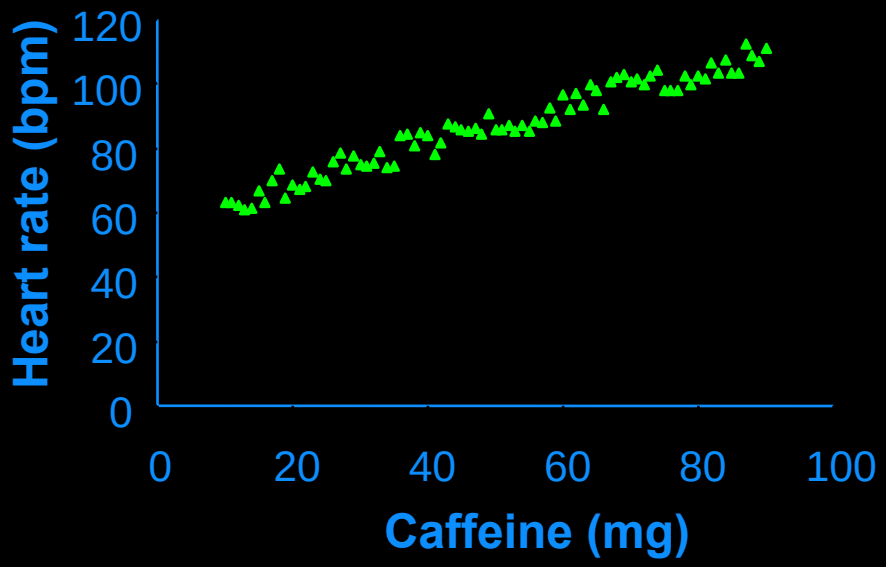
- Can detect trends over time

Presenting relationships

- Often want to know if values of one variable will tend to coincide with values of another variable
- Example: Does drinking more coffee coincide with higher heart rate?
- The manner in which we study a relationship depends on the type of variables involved:
  - two numerical variables?
  - two categorical variables?
  - one of each?
- In any case, raw data consists of pairs of values

Scatterplot

- A plot that displays pairs of data values when both variables are numerical
- Points coincide with data-value pairs



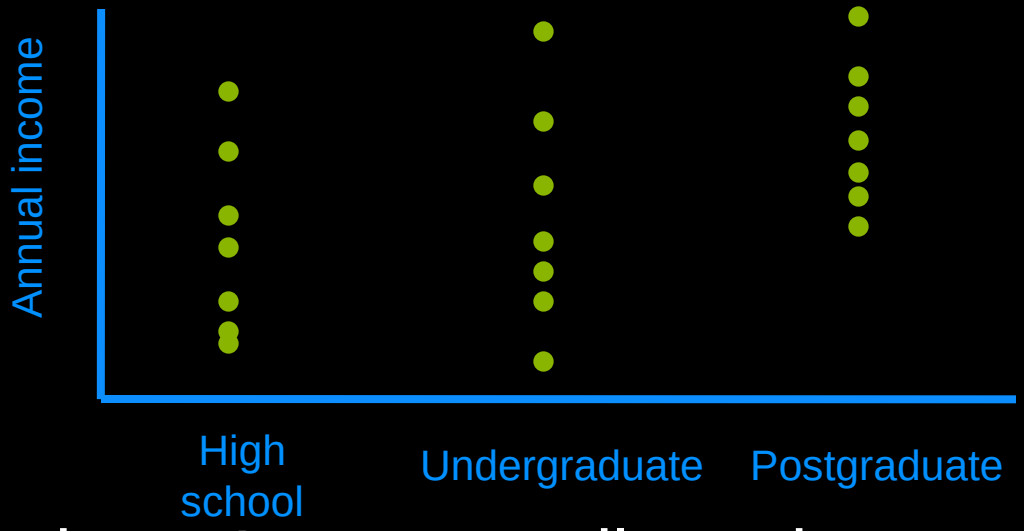
- Can use the plot to get a feel for the relationship

Studying scatterplots

- We use a scatterplot to study the type and strength of a relationship
- A common type of relationship is the **linear** relationship (i.e. straight line)
- A relationship is strong if the points don't scatter much
- A weak relationship means lots of scatter

Other scatterplots

- Scatterplots can also be used to study a relationship between a numerical and categorical variable
  - e.g. level of annual income versus level of education



- But it doesn't necessarily make sense to talk about type or strength of relationship here

Relationships for two categorical variables

- If both variables in a relationship are categorical, a scatterplot won't be useful
- Example: Voter preference versus gender of voter
- Ask 1000 men and 1000 women who they'll vote for out of three political candidates (A, B or C)

Men

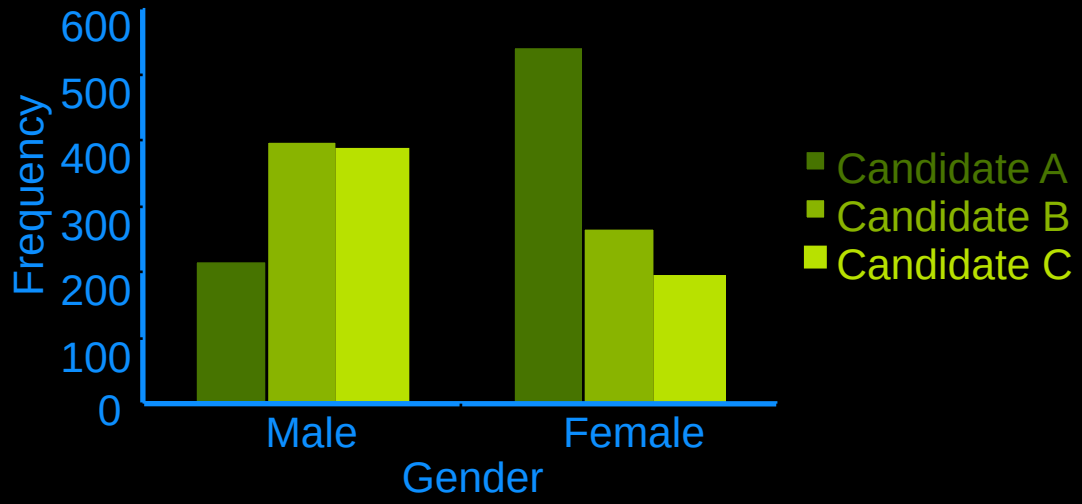
Candidate	Frequency
A	217
B	398
C	385

Women

Candidate	Frequency
A	541
B	265
C	194

Side-by-side bar chart

- These two frequency tables can be presented in a **side-by-side bar chart**
- This is effectively two bar charts in one



- Can use this to study trends in the data
  - e.g. Candidate A more popular among women than men

Contingency table

- Another way of presenting such a relationship is to combine the two frequency tables into one:

	Men	Women
Candidate A	217	541
Candidate B	398	265
Candidate C	385	194

- This is known as a **contingency table**
- This information can be presented in a **comparative bar chart**

Comparative bar chart

- A bar is presented for each gender
- Each bar is then divided up proportionally according to how popular each candidate is within that gender