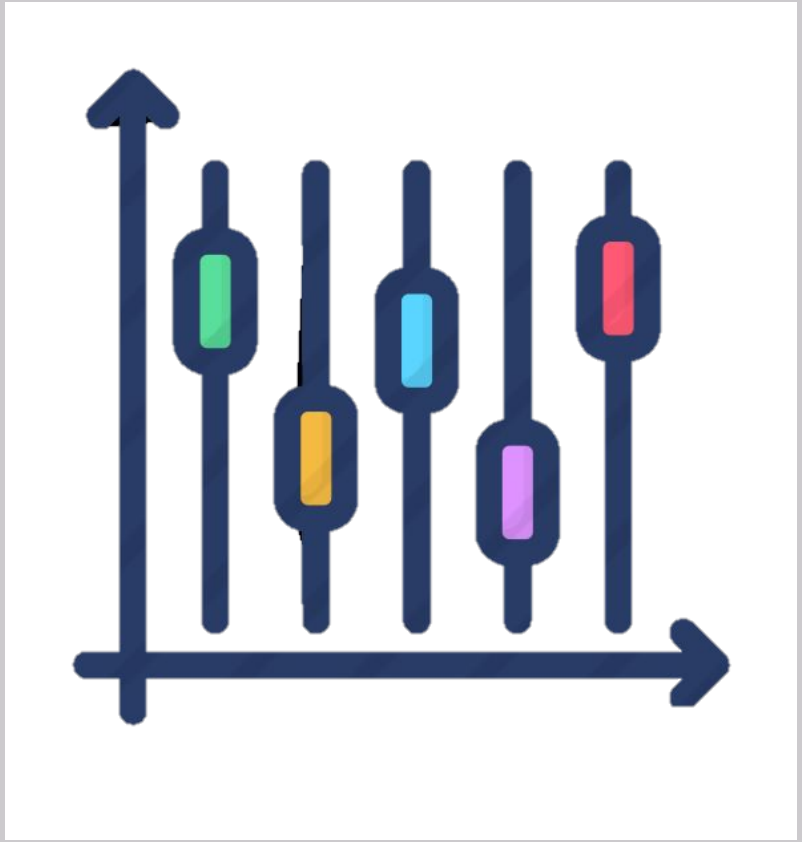


Describing data using tables and graphs

Descriptive statistics

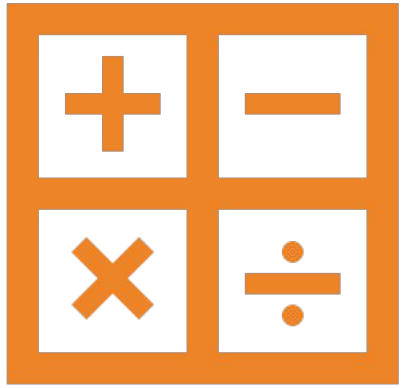
- Concerned with techniques that are used to describe, organize, or summarize data

- We do this with graphs!

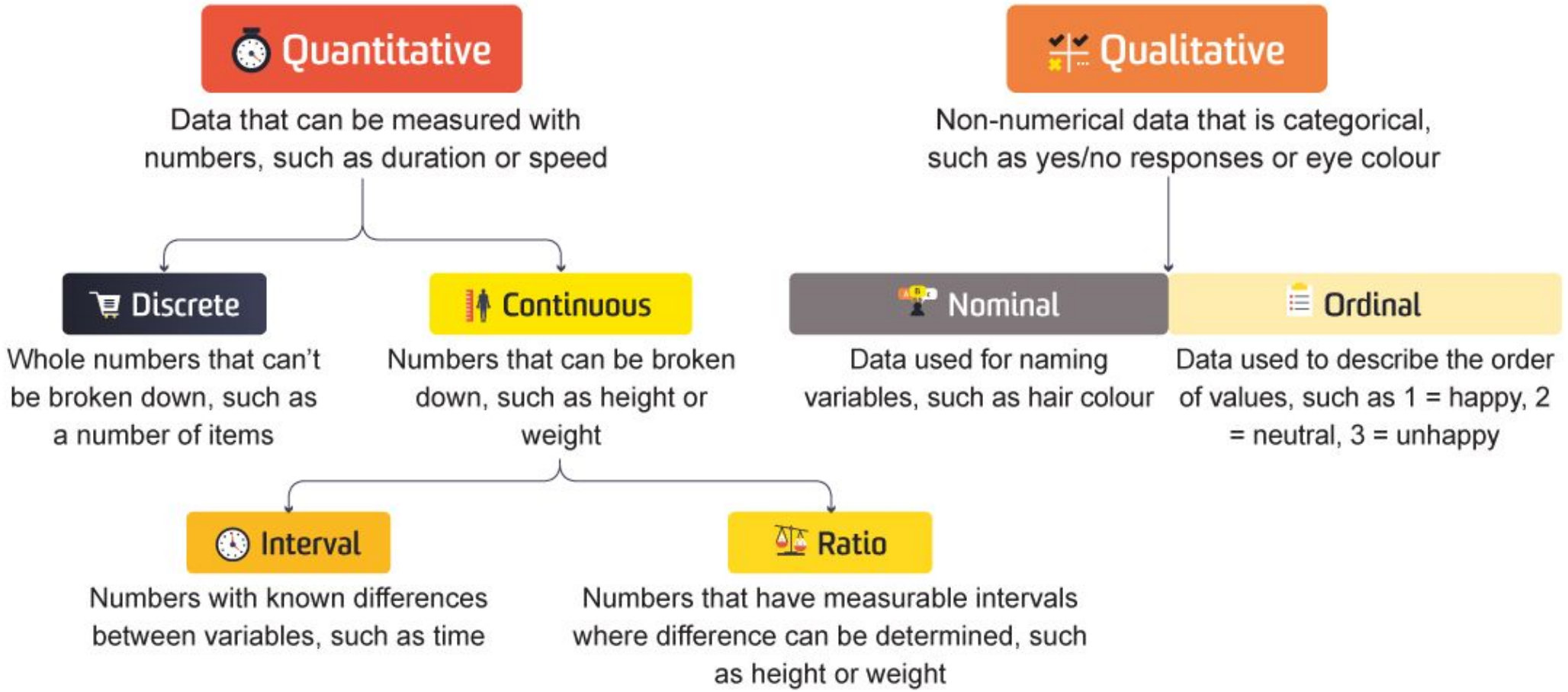
Levels of measurement

Type of scale influence operations

- Nominal and Ordinal – you cannot use addition, subtraction, multiplication, division, or ratios
  - Nominal data are qualitative
  - Ordinal data tell us about magnitude
- Interval - multiplication, division
- Ratio – you can use all the operations!



Types of Data



Likert Scales – the exception

- In order to calculate summary statistics on data, we need them to be on an interval or ratio scale
- We treat Likert scales (any scale that measures attitudes/ratings using labels rather than numbers) as interval

	Not at all offensive			Moderately offensive			Highly offensive
Paying someone else to teach your child social and cultural values	○	○	○	○	○	○	○
Paying someone else for sex	○	○	○	○	○	○	○
Paying someone else to love your friend	○	○	○	○	○	○	○
Paying someone to help you recover from a personal loss	○	○	○	○	○	○	○
Paying someone else to take care of your sick family member	○	○	○	○	○	○	○
Paying someone else to breastfeed your child	○	○	○	○	○	○	○
Paying someone to have a baby whom the buyer subsequently raises	○	○	○	○	○	○	○

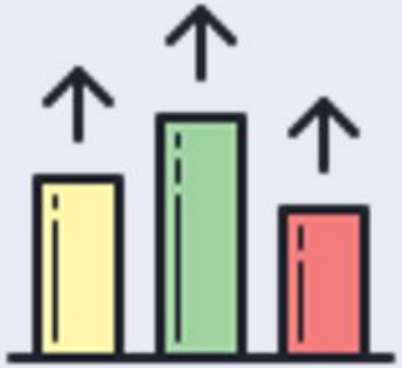
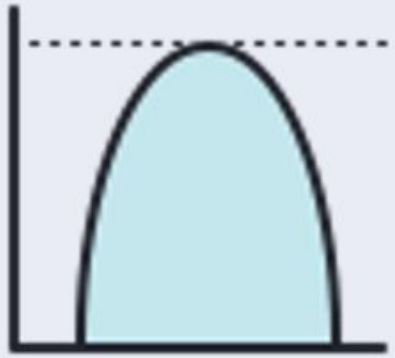
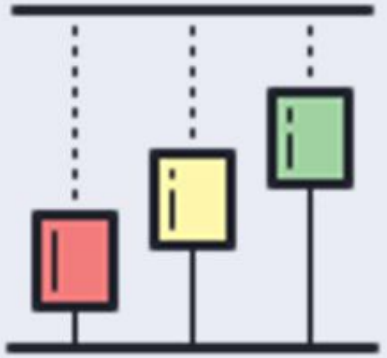
Graphing data

Types of data influence graphs

- Quantitative data (interval, ratio, continuous or discrete)
  - Frequency tables
  - Stem and leaf plots
  - Histograms
  - Box plots

- Qualitative data (nominal, ordinal, discrete variables)
  - Pie charts
  - Bar graphs
  - Line graphs

Types of data influence graphs

Pie Chart	Bar Chart	Histogram/Density Plot	Box Plot
Categorical	Categorical	Numerical	Numerical
			

Frequency distributions

- Ordered list of all values of a variable and their frequencies
- Logical order (usually descending)
- Helps calculate range, most frequent value, at first glanc
- Calculate proportions and percentages

Age	Tally Marks	Number of people
5	𝍸	5
17	𝍸	4
23	𝍸	3
30	𝍸	3
39	𝍸	1
40	𝍸	3
48	𝍸	3
51	𝍸	2
62	𝍸	3
65	𝍸	3
		30

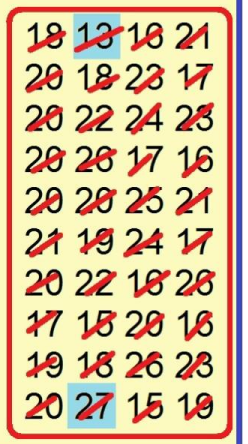
Frequency distributions (continued)

- Can be tables or graphs, but contains two elements

- *f* = Frequency
  - # of times a value of variable occurs
  - $\Sigma f = n$

Frequency Distribution Table for Grouped Data

Class Limits	Frequency
25 - 27	5
22 - 24	7
19 - 21	14
16 - 18	11
13 - 15	3
Total	40



Frequency example 1 ☑

The following set of $N = 20$ scores was obtained from a 10-point statistics quiz. We will organize these scores by constructing a frequency distribution table.

The scores are:

8, 9, 8, 7, 10, 9, 6, 4, 9, 8

7, 8, 10, 9, 8, 6, 9, 7, 8, 8

Frequency example 1

Scores:

8, 9, 8, 7, 10, 9, 6, 4, 9, 8

7, 8, 10, 9, 8, 6, 9, 7, 8, 8

Highest score is X = 10

Lowest score is X = 4

X	f
10	2
9	5
8	7
7	3
6	2
5	0
4	1

Notice that all the possible values between 10 and 4 were used!

Obtaining ΣX from frequency distribution

$\Sigma f = n = 20$

1) $\Sigma X =$

$\Sigma X = 10 + 10 + 9 + 9 + 9 + 9 + 9 + 8 + 8 + \ldots$

$\Sigma X = 158$

2) $\Sigma X^2 = 1288$

Example 1

X	f
10	2
9	5
8	7
7	3
6	2
5	0
4	1

Learning check 1

1) Place the following scores in a frequency distribution table.

2, 3, 1, 2, 5, 4, 5, 5, 1, 4, 2, 2

Learning check 1 - Answer

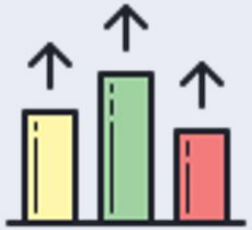
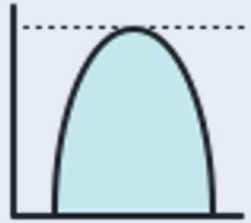
1) Place the following scores in a frequency distribution table.

2, 3, 1, 2, 5, 4, 5, 5, 1, 4, 2, 2

X	f
5	3
4	2
3	1
2	4
1	2

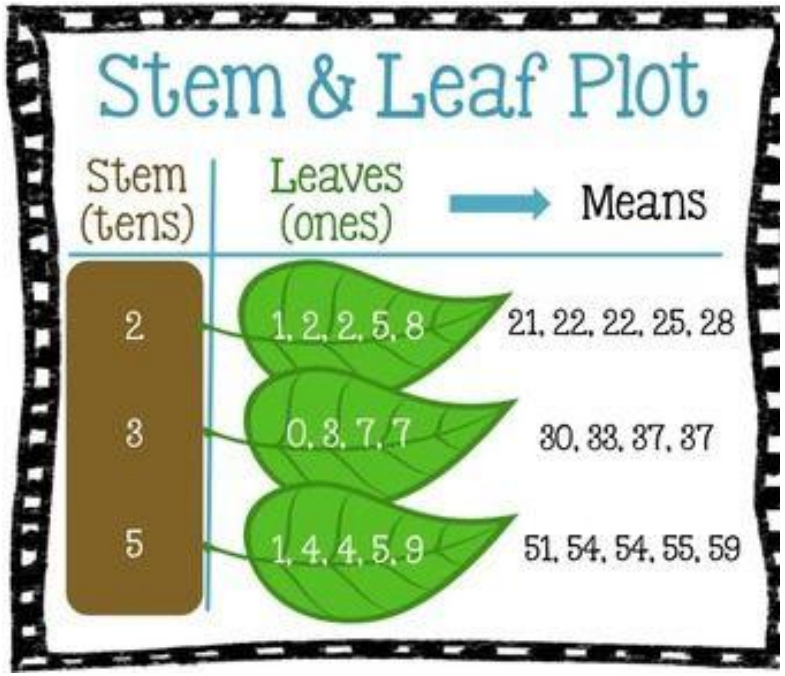
More on graphing quantitative data

- Interval and ratio data can be graphe in the following plots
  - Stem and leaf plots
  - Histograms
  - Box plots
  - Bar charts

Pie Chart	Bar Chart	Histogram/Density Plot	Box Plot
Categorical	Categorical	Numerical	Numerical
			

Stem-and-leaf plots

- Groups data with the same stem
  - All possible stems are listed in a column
  - The leaf for each quantitative measurement is placed in the stem row
  - Leaves with the same stem value are listed in increasing order horizontally



Stem-and-leaf Example 1 ☑

Number of touchdown passes thrown by each of the 31 teams in the National Football League in the 2000 season:

37, 33, 33, 32, 29, 28, 28, 23, 22, 22, 22, 21, 21, 21, 20, 20, 19, 19, 18, 18, 18, 18, 16, 15, 14, 14, 14, 12, 12, 9, 6

```
3 | 2 3 3 7
2 | 0 0 1 1 1 2 2 2 3 3 8 8 9
1 | 2 2 4 4 4 5 6 8 8 8 8 9 9
0 | 6 8
```

OR

```
3 | 7
3 | 2 3 3
2 | 8 8 9
2 | 0 0 1 1 1 2 2 2 3 3
1 | 5 6 8 8 8 8 9 9
1 | 2 2 4 4 4
0 | 6 8
```

Histograms

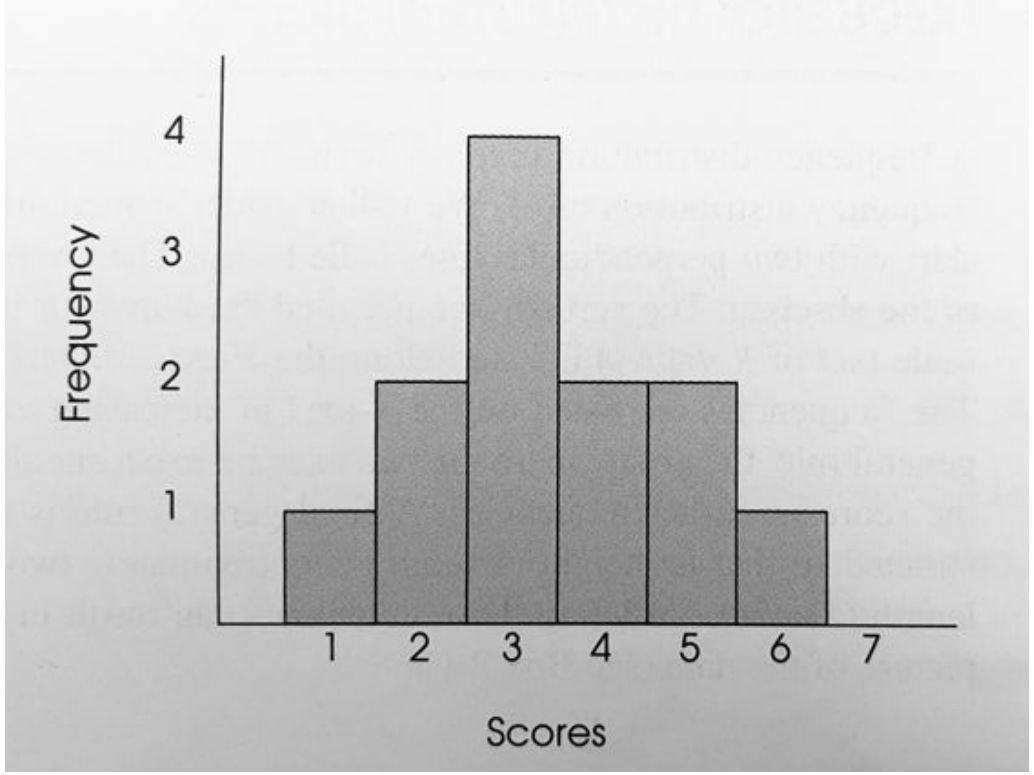
- List of variables and their frequencies
  - *X*-axis – class intervals of variables (same width)
  - *Y*-axis – vertical bar of frequencies (or relative frequencies)

Histogram Example 1

Scores

X	f
6	1
5	2
4	2
3	4
2	2
1	1

Histogram

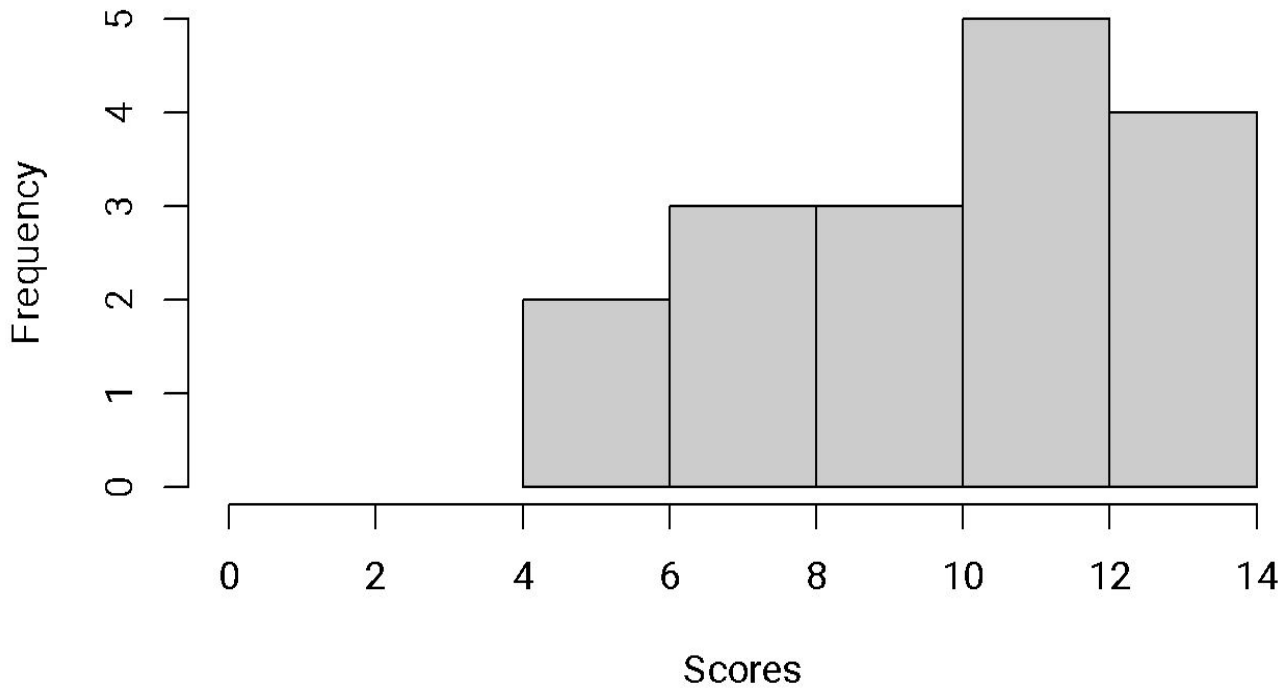


Histogram Example 2

Scores

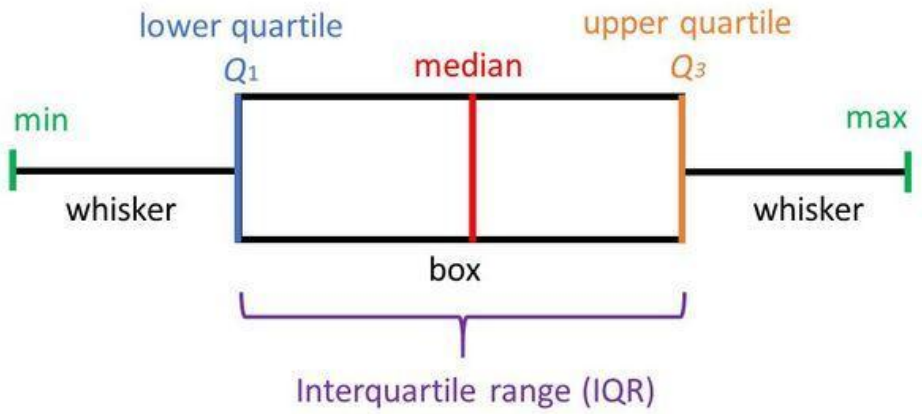
X	f
12-13	4
10-11	5
8-9	3
6-7	3
4-5	2

Histogram



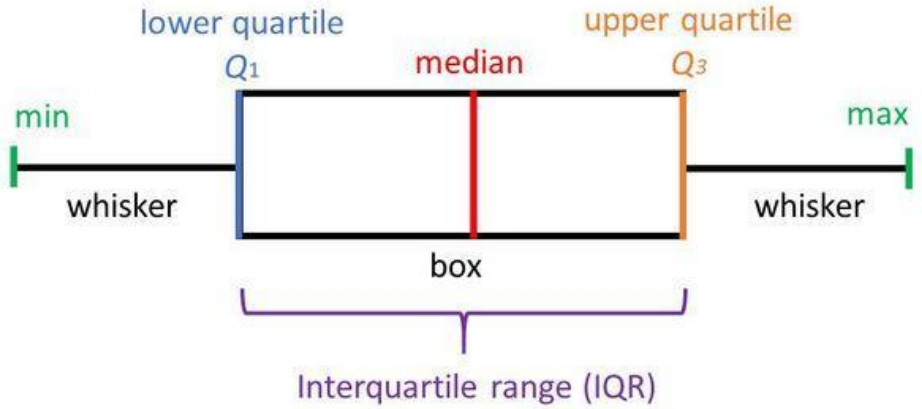
Box plots

- Five-number summary of a set of data: the minimum, first quartile, median, third quartile, and maximum
  - Minimum - the lowest score, excluding outliers
  - Lower Quartile – 25% of scores fall below the lower quartile value
  - Median - mid-point of the data; shown by the line that divides the box into two parts

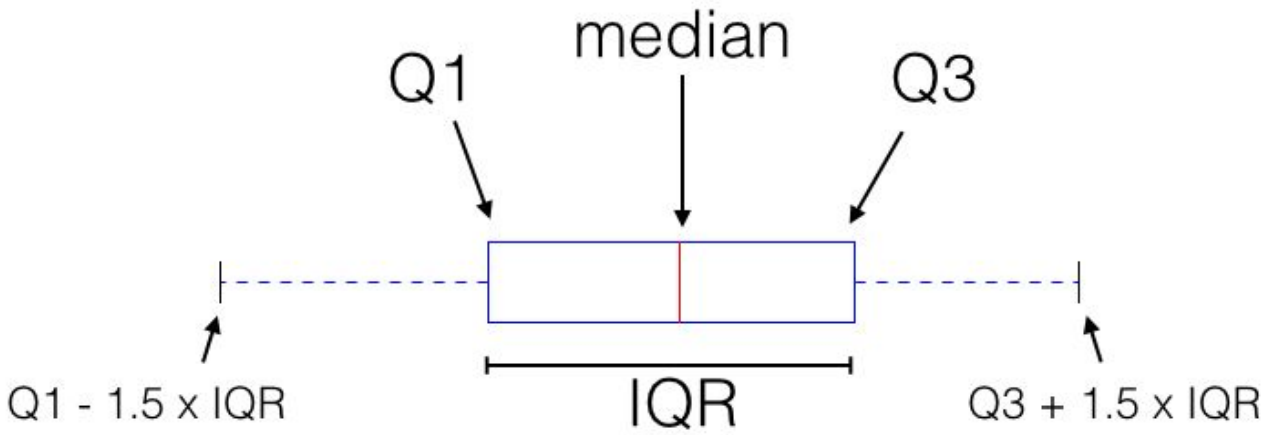


Box plots (continued)

- Upper Quartile – 75% of the scores fall below the upper quartile value
- Maximum – the highest score, excluding outliers
- Whiskers – scores outside the middle 50% (i.e. the lower 25% of scores and the upper 25% of scores)
- The Interquartile Range (IQR) –middle 50% of scores (i.e., the range between the 25th and 75th percentile)



Boxplot summary



Q1 = Quartile 1, or median of the left data subset
after dividing the original data set into 2 subsets via the median
Q3 = Quartile 3, like Q1, median of the "right" data subset
IQR = Interquartile-range, Q3 - Q1

values < *Q1 - 1.5 x IQR*
and
values < *Q3 + 1.5 x IQR*

are considered as outliers

Box plot Example 1 ☑

Find the quartiles of this data set:

6, 47, 49, 15, 43, 41, 7, 39, 43, 41, 36

- Step 1: Arrange data set in increasing order
- Step 2: Find the rank of the median split

  (n + 1) ÷ 2 = (11 + 1) ÷ 2 = 6

Rank of data points

Rank	Value	
1	6	
2	7	
3	15	**Q1**
4	36	
5	39	
6	41	**Median**
7	41	
8	43	**Q3**
9	43	
10	47	
11	49	

Box plot Example 1 ☑

- Step 3: Split the lower half of the data in two again to find the lower quartile. Do the same to the upper half

  (n + 1) ÷ 2 = (5+1)/2 = 6

- Step 4: Calculate IQR

  Q3-Q1 = 43-15 = 28

Rank of data points

Rank	Value	
1	6	
2	7	
3	15	**Q1**
4	36	
5	39	
6	41	**Median**
7	41	
8	43	
9	43	**Q3**
10	47	
11	49	

Creating the box plot

Rank of data points

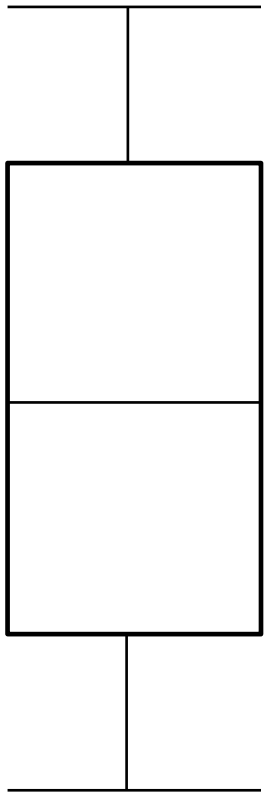
Rank	Value
1	6
2	7
3	15
4	36
5	39
6	41
7	41
8	43
9	43
10	47
11	49

IQR = 28

Q1

Median

Q3



You can use IQR to identify outliers!

Check for low outliers
Q1 – (1.5*IQR) =
15 – (1.5*28) =
15 – 42 = -27

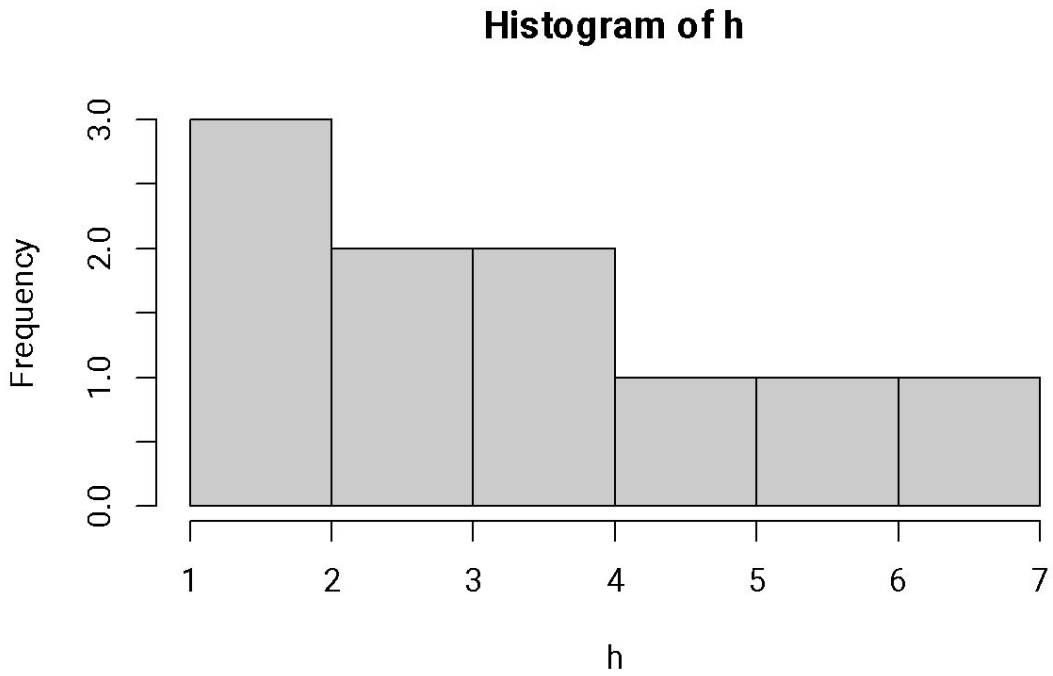
Check for high outliers
Q3 + (1.5*IQR) =
43 + (1.5*28) =
43 + 42 = 85

Learning Check 2

1) Use a histogram to draw the following set of data containing the number of times a group of students watched the Harry Potter series.

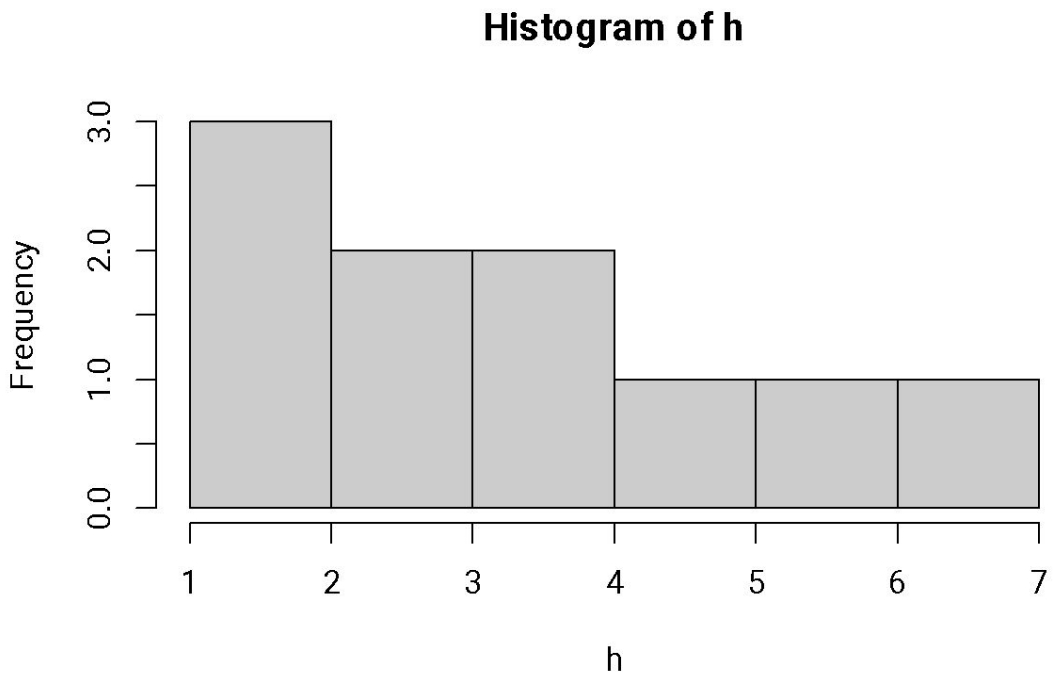
1, 4, 3, 5, 2, 7, 4, 6, 2, 3

Generally, you want to use whole numbers on the y-axis (but I had trouble doing that in my stats program



Learning Check 2 - Answer ☑

Set of data: 1, 4, 3, 5, 2, 7, 4, 6, 2, 3



Generally, you want to use whole numbers on the y-axis (but I had trouble doing that in my stats program)

Learning Check 3

2) Use a stem-and-leaf plot to organize the following scores:

86, 114, 94, 107, 96, 100, 98, 118, 107

132, 106, 127, 124, 108, 112, 119, 125, 115

Reorder:

86, 94, 96, 98, 100, 106, 107, 107, 108

112, 114, 115, 118, 119, 124, 125, 127, 132

8	6
9	4 6 8
10	0 6 7 7 8
11	2 4 5 8 9
12	4 5 7
13	2

Learning Check 3 - Answer

Step 1: reorder data

86, 94, 96, 98, 100, 106, 107, 107, 108

112, 114, 115, 118, 119, 124, 125, 127, 132

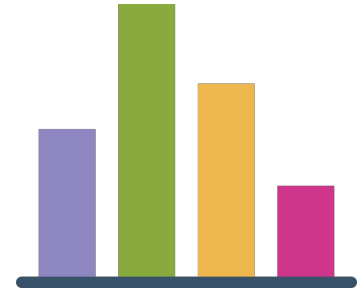
Step 2: Create stem-and-leaf plot

```
 8  | 6
 9  | 4 6 8
10  | 0 6 7 7 8
11  | 2 4 5 8 9
12  | 4 5 7
13  | 2
```

Graphs for nominal or ordinal data

- Values of a qualitative variables can only be classified into categories (classes)

- Graphical methods for describing qualitative data include
  - Pie charts
  - Bar graphs
  - Line graphs



Bar graphs

- A bar graph is essentially the same as a histogram, except there are spaces between adjacent bars
  - Scale consists of separate, distinct categories

- *X*-axis - categories or classes

- *Y*-axis - class frequency, class relative frequency, or class percentage

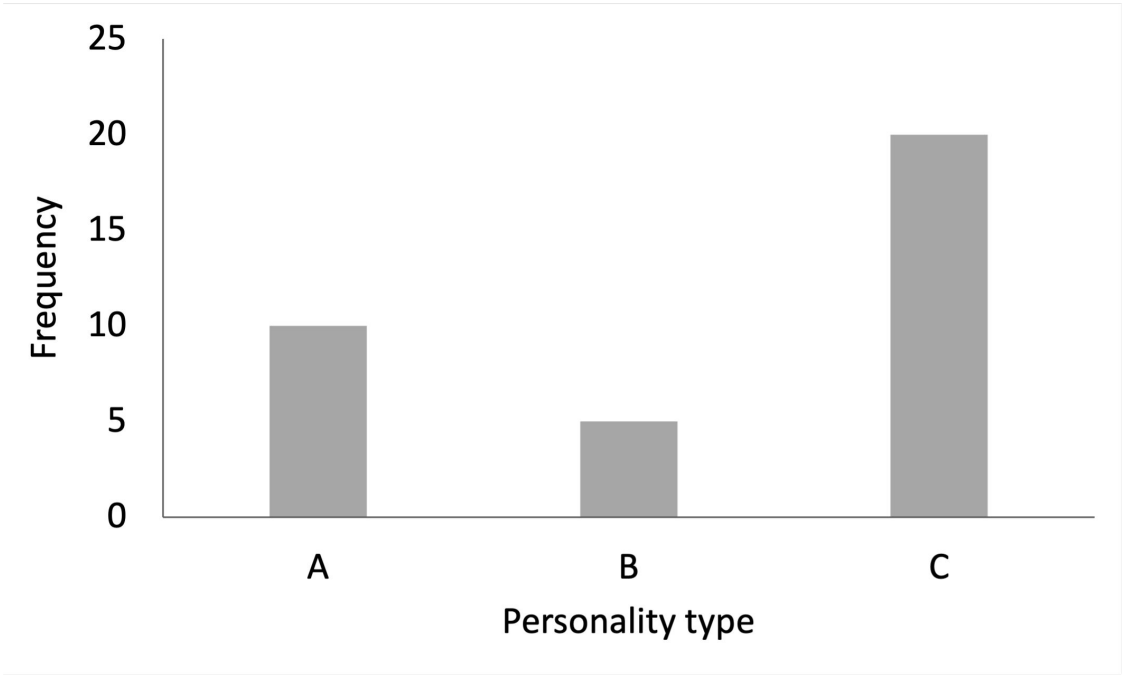
Bar Graph Example 1

Distribution of personality types in a sample of college students

Data

Personality type	f
A	10
B	5
C	20

Bar Graph



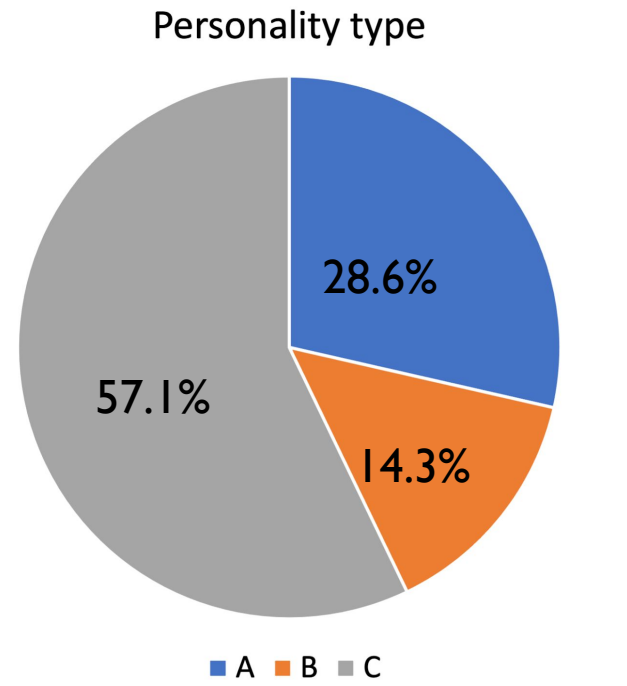
Pie charts

- Uses relative frequency to depict data as slice of pie
- Proportional to responses in each category
- Good for small number of categories

Data

Personality type	f	*Relative frequency*
A	10	10/35 = 28.6 %
B	5	5/35 = 14.3 %
C	20	20/35 = 57.1%

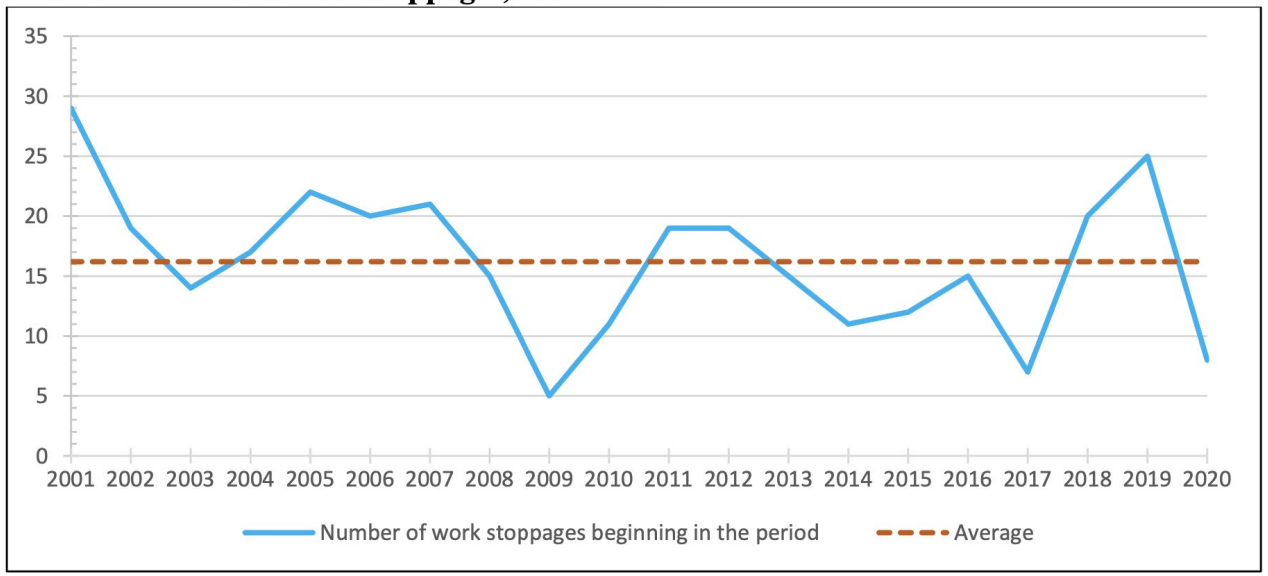
Pie Chart



Line graphs

- Like a bar graph with dots at the top, and dots are connected by lines
- Best for representing changes in time

Chart 1: Number of work stoppages, 2001-2020



Note. BLS issued a data correction from 2018 to 2020 – one work stoppage was not included.

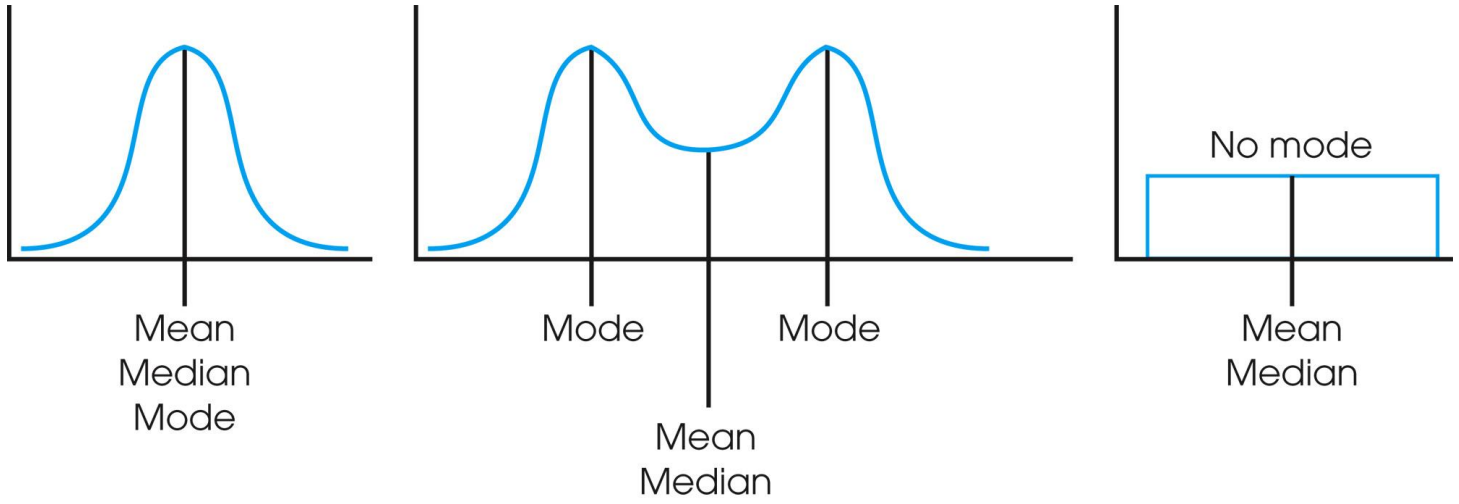
Shapes of distributions

Remember central tendency?

- Central tendency describes center of distribution
  - Mean, median, mode

Central tendency and the shape of the distribution

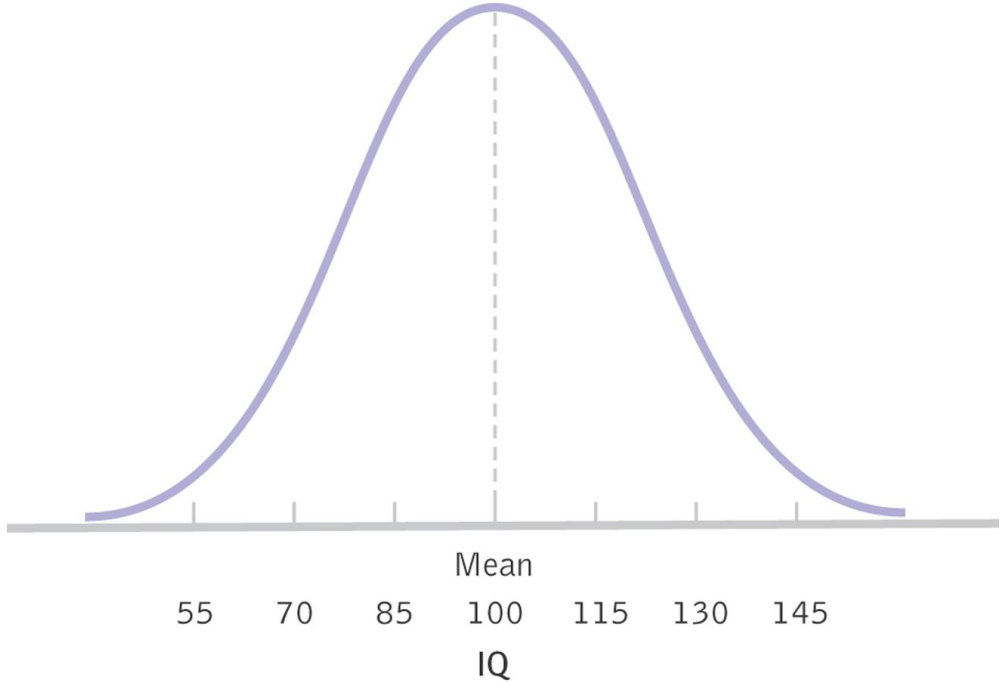
- Symmetrical distributions
  - Mean and median have same value
  - If exactly one mode, it has same value as the mean and the median
  - Distribution may have more than one mode, or no mode at all



The Normal Distribution

- Bell-shaped
  - One mode
  - Symmetric

- Many naturally-occurring variables (like height) are approximately normally distributed

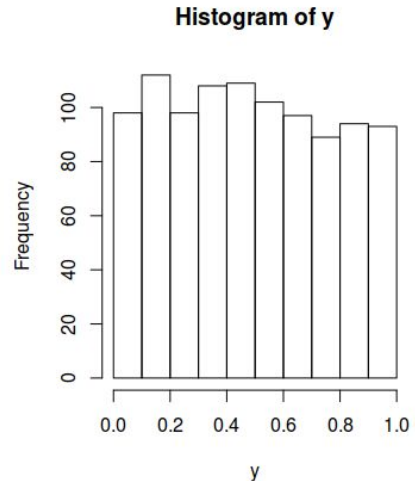


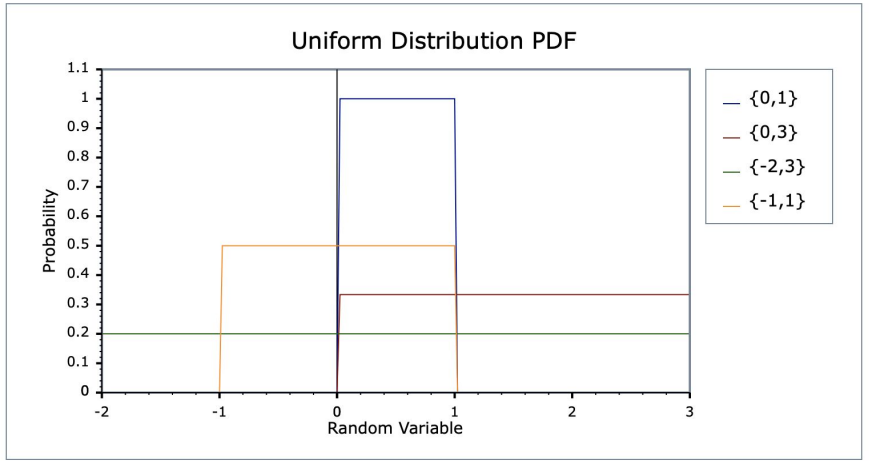
Uniform distributions

- Every outcome is equally likely

- Examples of discrete uniform distributions
  - Probability of hitting heads or tails
  - Probability of landing on one side of a die

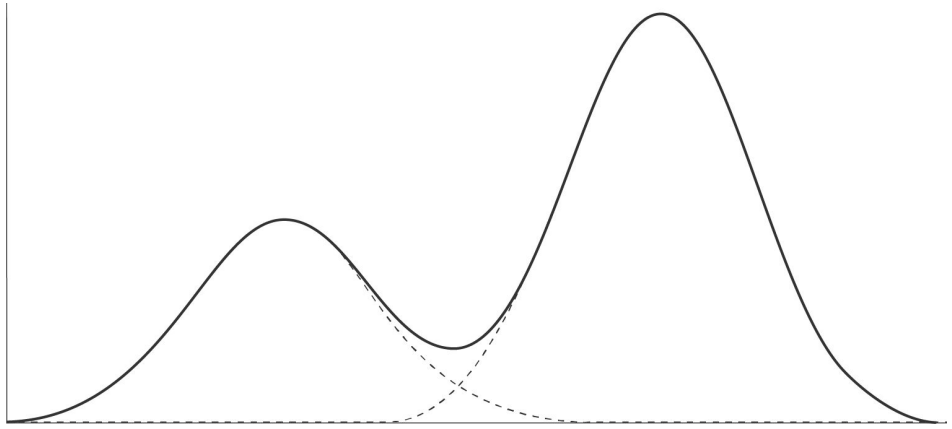
- Examples of continuous uniform distributions
  - Perfect random number generators





Bimodal distributions

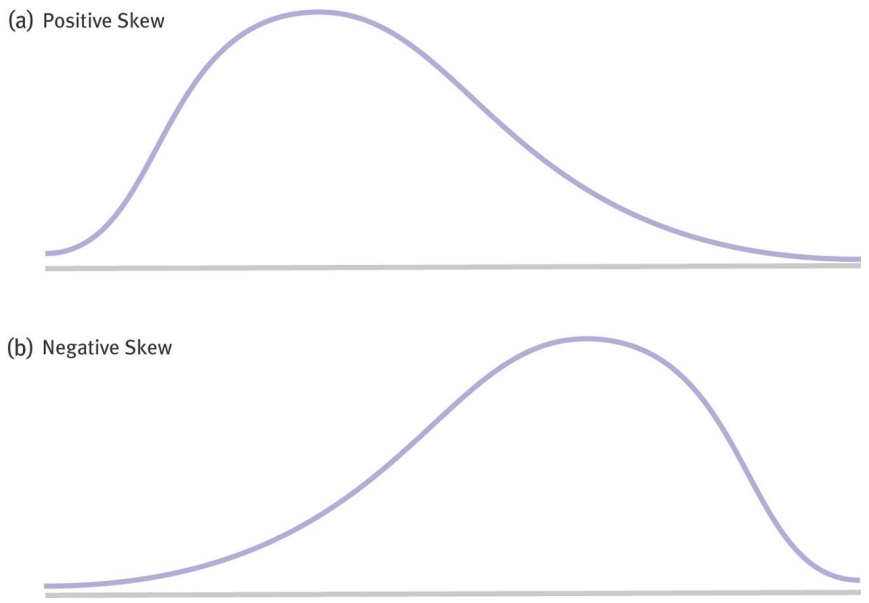
- Bimodal (two peaks) or multi-modal
  - Two most frequently occurring values
  - May indicate relevant subgroups



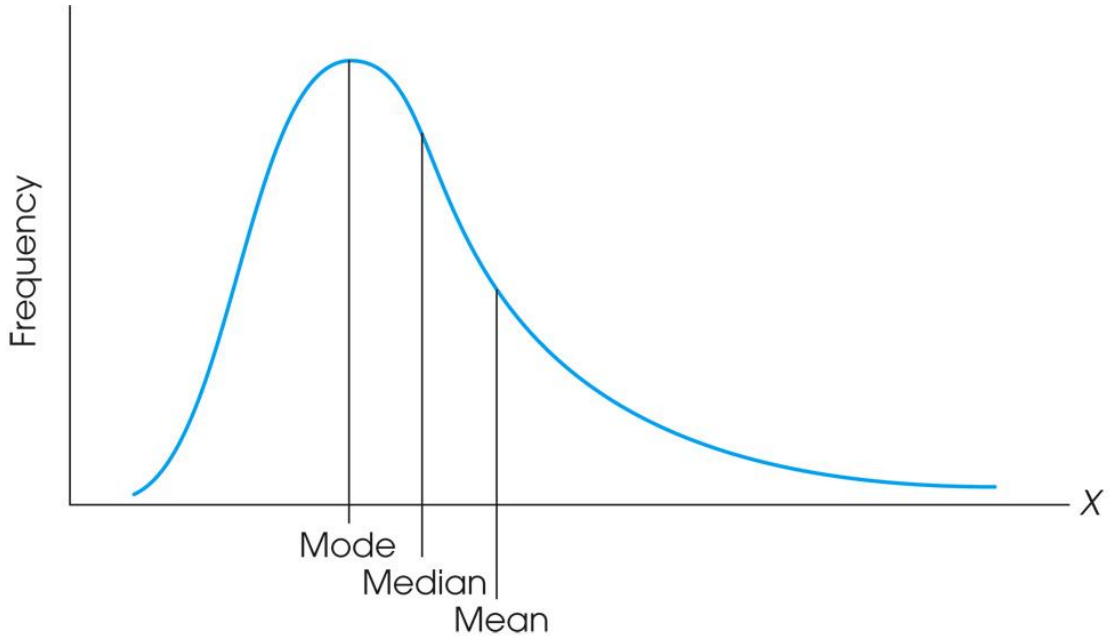
Central tendency in skewed distributions

- Mean, influenced by extreme scores, is found far toward the long tail (positive or negative)
- Median is found toward the long tail, but not as far as the mean

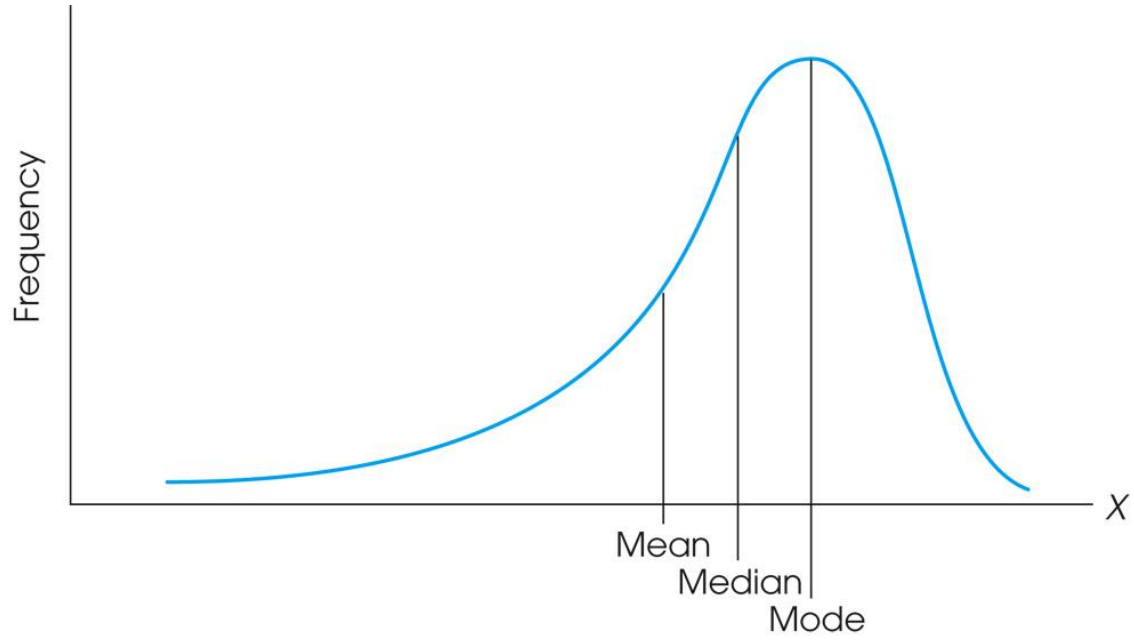
- Mode is found near the short tail
  - Positively skewed
    - If Mean – Median > 0
    - Extreme observations in right tail
  - Negatively skewed
    - If Mean – Median < 0
    - Extreme observations in left tail



Skewed distributions

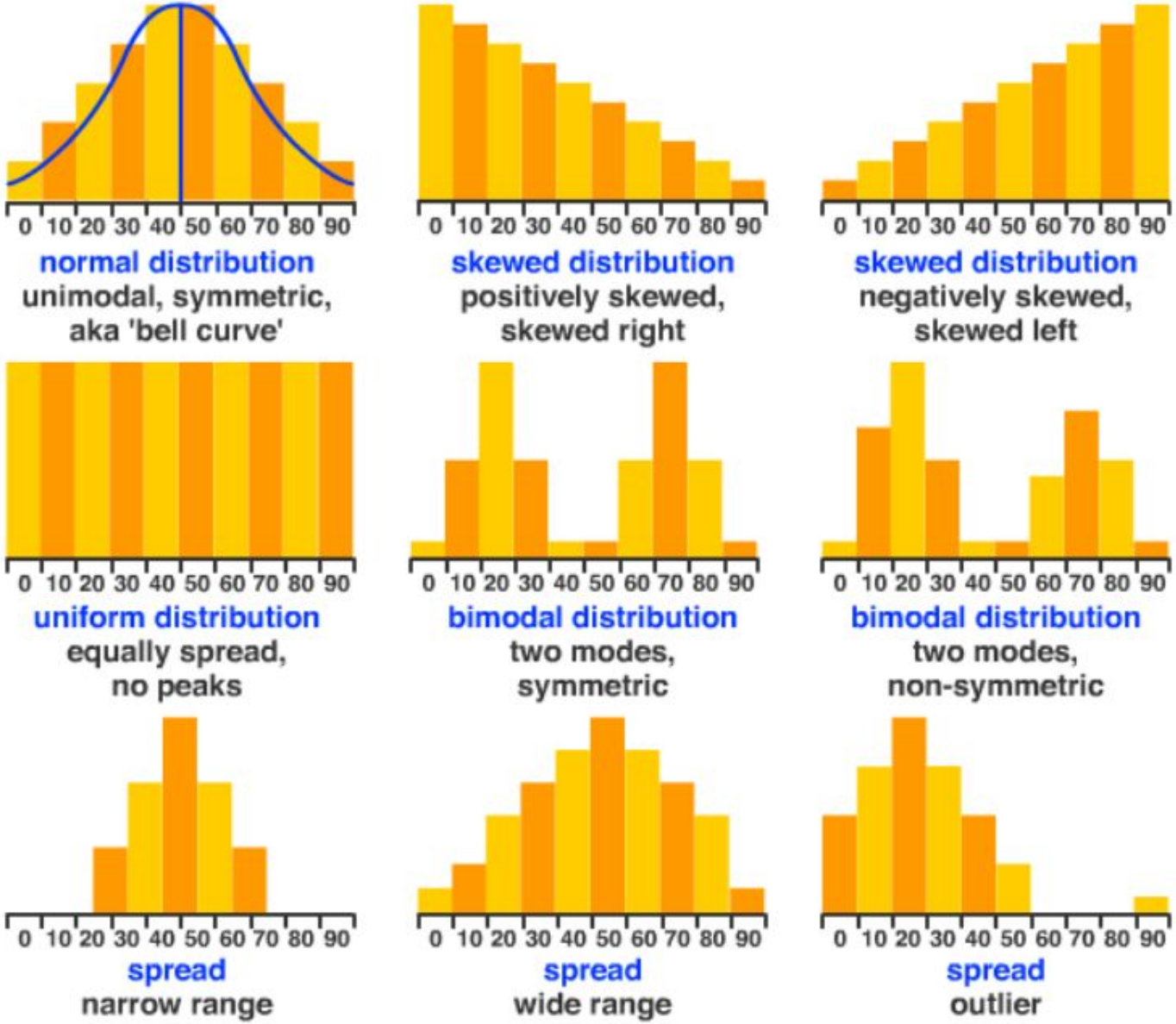


Positive skew / right-tailed



Negative skew / left-tailed

Summary of distributions



Learning objectives

By the end of this lecture, you should be able to:

- Identify methods of graphing qualitative and quantitative data
- Describe shapes of data