Chapter 2: Summarizing Data

Math 140

Based on content in OpenIntro Stats, 4th Ed

Hitchman

Section 2.1
Examining Numerical Data

Starwars Data

- Load the `tidyverse` into your RStudio session to get access to data from galaxies far, far, away...
- The data frame **starwars** has 87 observations and 14 variables:

```
"name" "height" "mass" "hair_color" "skin_color"
"eye_color" "birth_year" "sex" "gender" "homeworld"
"species" "films" "vehicles" "starships"
```

Examining a Numerical Variable

- Large data sets can often be summarized effectively with visual displays and summary statistics.
- Visual displays help to reveal
  - the overall *shape* of a data set
  - patterns within it, and exceptions to these patterns (*outliers*)
- A *summary statistic* is a number summarizing a data set.
- Summary statistics help measure
  - the *center* - what is a "typical element"?
  - the *spread* - how widely do values vary, and/or stray from center?

Common plots of numerical data

Two common ways to picture one numerical variable:

- Histograms
- Boxplots

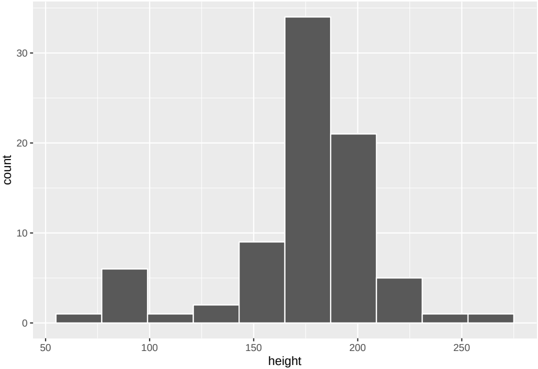
Picturing the relationship between two numerical variables:

- Scatterplot

Histograms

- Histograms provide a view of the range of values and the *data density*.
  - horizontal axis - values obtained, gathered in "bins" by value range that you specify
  - vertical axis - frequency (number of data values falling in that bin)
- Histograms are especially convenient for describing the *shape* of the data distribution.
- The chosen *bin width* can alter the story the histogram is telling.

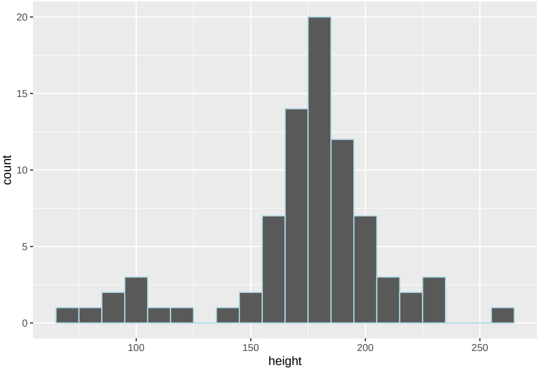
Starwars - height of characters



(`tidyverse` already loaded):

```
ggplot(starwars)+
  geom_histogram(aes(x=height),
                 bins=10,col="white")
```

Specifying bin width

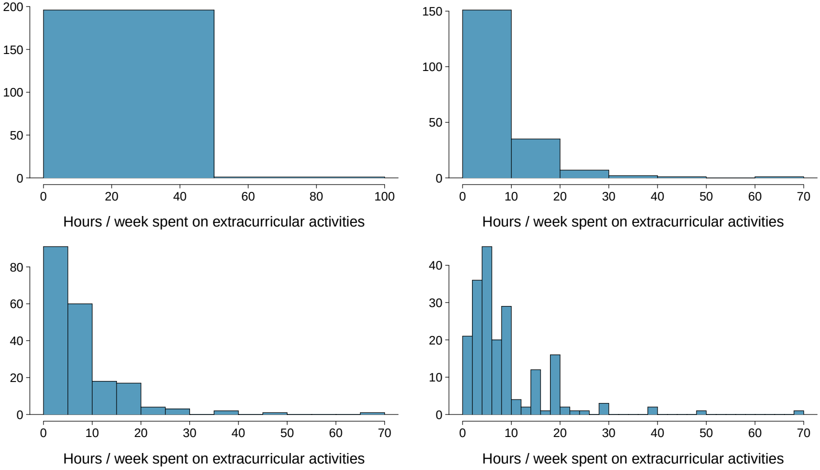


(`tidyverse` already loaded):

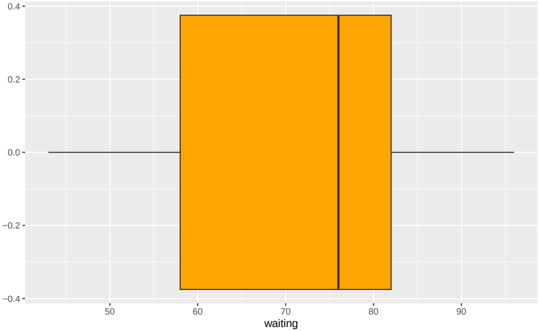
```
ggplot(starwars)+
   geom_histogram(aes(x=height),
                  binwidth=10,col="lightblue")
```

Bin choice matters!

Q: *Which one(s) of these histograms are useful? Which reveal too much about the data? Which hide too much?*



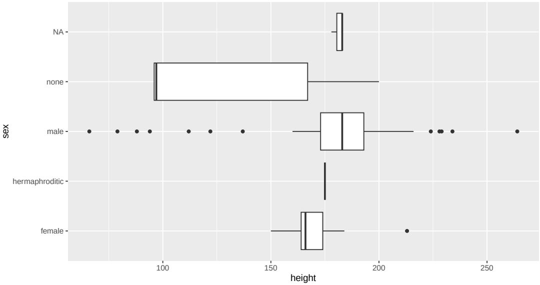
Box Plot - sneak peek



(`tidyverse` already loaded):

```
ggplot(faithful)+
   geom_boxplot(aes(x=waiting),fill="orange")
```

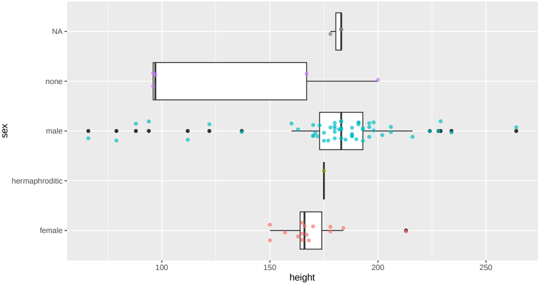
Box Plot - grouping by some categorical variable



(`tidyverse` already loaded):

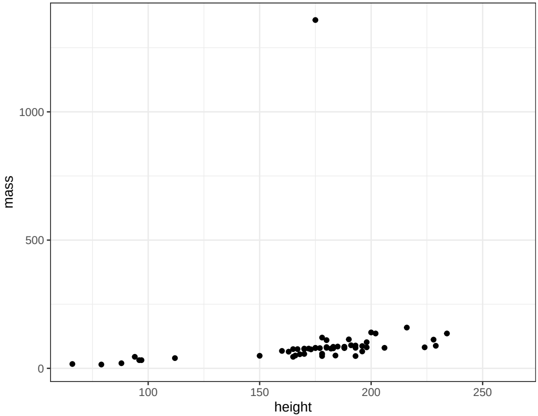
```
ggplot(starwars)+
   geom_boxplot(aes(x=height,y=sex))
```

Box Plot - grouping by some categorical variable



Boxplots by themselves, don't tell you anything about the size of the data set.

Scatter plot

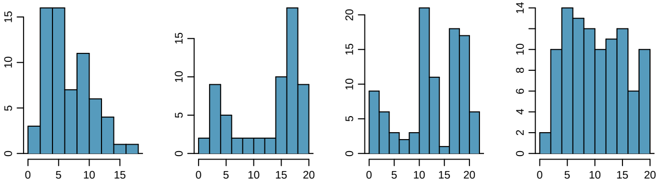


(`tidyverse` already loaded):

```
ggplot(starwars)+
   geom_point(aes(x=height,y=mass))+
   theme_bw()
```

Shape of a distribution: modality

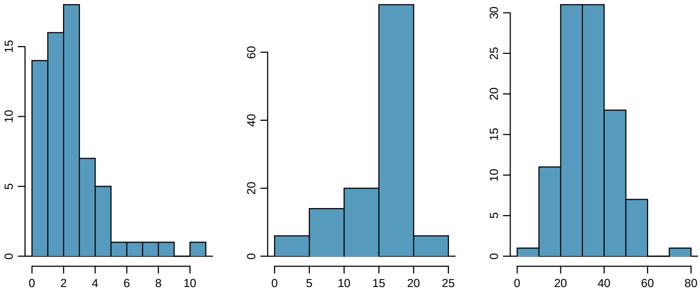
Does the histogram have a single prominent peak (*unimodal*), several prominent peaks (*bimodal/multimodal*), or no apparent peaks (*uniform*)?



From the text: *In order to determine modality, step back and imagine a smooth curve over the histogram – imagine that the bars are wooden blocks and you drop a limp spaghetti over them, the shape the spaghetti would take could be viewed as a smooth curve.*

Shape of a distribution: skewness

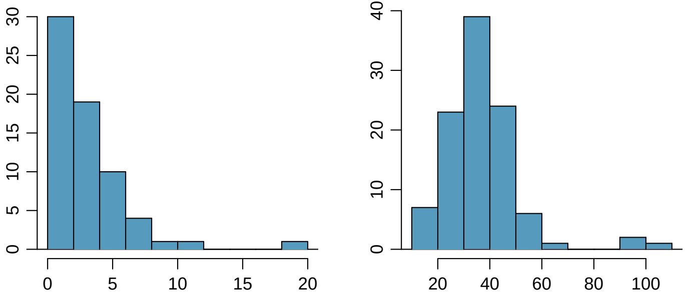
Is the histogram *right skewed*, *left skewed*, or *symmetric*?



Histograms are said to be skewed to the side of the long tail.

Shape of a distribution: unusual observations

Are there any unusual observations or potential *outliers*?

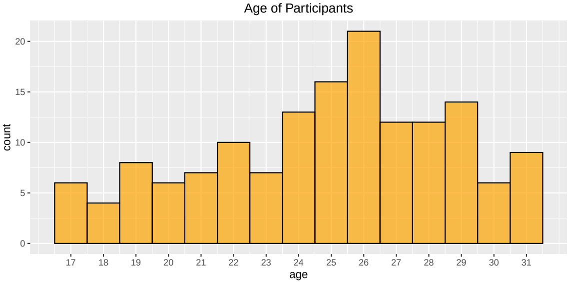


Shape of a distribution

To summarize, when describing the shape of a distribution, it is useful to describe these three features:

- modality
- skewness
- unusual observations

Chopin Competition Contestant Ages

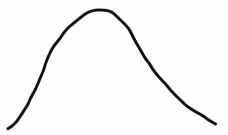


- What is a typical age? What is the shape of the dist'n?
- This distribution looks moderately skewed to the left, unimodal, with no real outliers.

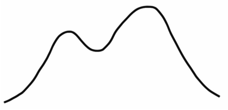
Commonly observed shapes of distributions

- modality

unimodal

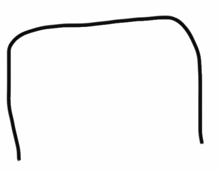


bimodal



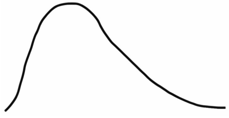
multimodal

uniform

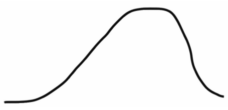


- skewness

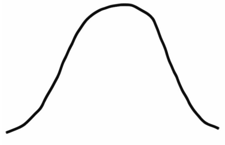
right skew



left skew



symmetric

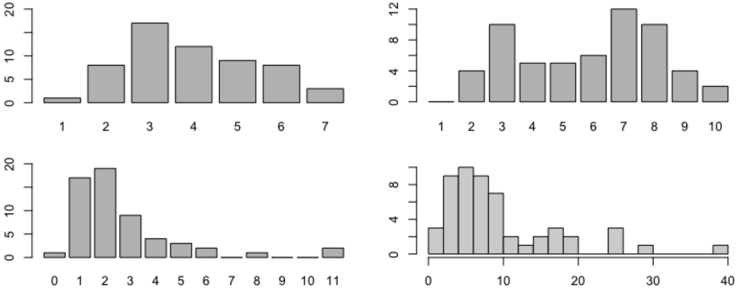


Student Survey

Sketch the expected distributions of the following variables from our student survey:

- Random number between 1 and 10
- Study hours per week
- Countries you've been in
- Daily phone usage in hours (0 to 7 or more)

Student Survey



Mean - One measure of the center of a distribution

- The *sample mean*, denoted as $\bar{x}$, can be calculated as

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

  where $x_1, x_2, \cdots, x_n$ represent the n observed values.
- The sample mean is a *sample statistic*, and serves as a *point estimate* of the population mean (μ.). This estimate may not be perfect, but if the sample is good (representative of the population), it is usually a pretty good estimate.

The average mass of Star Wars characters

In R, we use the `mean()` command to find the mean of a numerical variable.

```
> mean(starwars$mass)
[1] NA
```

wait... what?This means the `mass` column has missing values (NA - 'not available'). To ignore those, use

```
> mean(starwars$mass, na.rm=TRUE)
[1] 97.31186
```

Median - A second measure of center

- The *median* is the value that splits the data in half when ordered from smallest to largest.

$$0, 1, 2, 3, 4$$

- If there are an even number of observations, then the median is the average of the two values in the middle.

$$0, 1, \underline{2, 3}, 4, 5 \rightarrow \frac{2+3}{2} = 2.5$$

- Since the median is the midpoint of the data, 50% of the values are below it, and the median is also called the 50^{th} *percentile.*
- The median mass of Star Wars characters?

```
> median(starwars$mass, na.rm=TRUE)
[1] 79
```

- Why is the mean so much larger than the median for Star Wars masses?

LEGO

Q: *What is the median number of pieces per set in my LEGO collection? I have 7 sets, and these are the piece counts:*

923, 617, 759, 811, 1792, 1015, 739

- *We order the data from smallest to largest to find the one (or ones) in the middle:*

  617, 739, 759, 811, 923, 1015, 1792.

- *The median is* $M = 811$.

Comparing Mean and Median

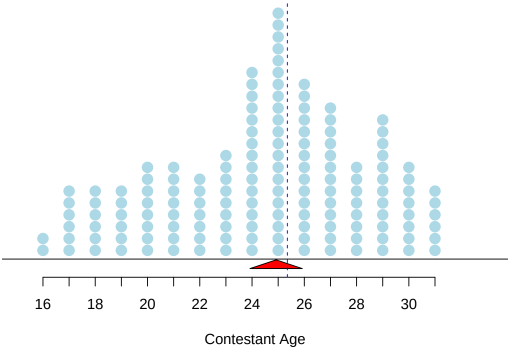
The median is a more *robust* measure of center than the mean - it is less sensitive to extreme values.

Consider these two LEGO collections (piece per set)

Old collection	617	739	759	811	923	1015	1792
New collection	617	739	759	811	923	1015	4092

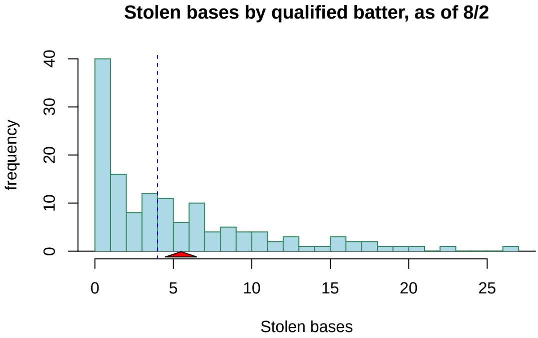
- The median of the new collection is the same as the old - 811.
- The mean, however, changes from 950.9 to 1279.4. Quite a jump!

Mean and Median for Chopin Competition ages



- Median (blue line) splits the data counts into two equal halves
- Mean (red Δ) marks where to put your finger to balance the plot

MLB Stolen bases



- The mean gets pulled toward the longer tail!
- Distribution is skewed right
- mean is 5.5 (red balancing triangle) median is 4 (blue line).

Are you typical?

http://www.youtube.com/watch?v=4B2xOvKFFz4

Q: *How useful are centers alone for conveying the true characteristics of a distribution?*

Measuring the spread of a distribution

1. The *standard deviation* is a single number that captures how far the elements tend to be from the mean.
2. The *five number summary* is a set of 5 numbers that captures the spread and overall range of the data.
3. We will see that the five number summary is a more *robust* measure of spread than the standard deviation - it is less sensitive to extreme values, and it can reveal skewness.

Standard Deviation

The standard deviation of a set of values is a single number that captures how much the values tend to be from the mean.
Here are three data sets, and all of them have the same mean, $\overline{x} = 5$.

1. $[5, 5, 5, 5, 5, 5]$
2. $[4, 4, 5, 5, 6, 6]$
3. $[0, 0, 0, 10, 10, 10]$

- In the first set, none of the values deviates at all from the mean, and it turns out the standard deviation of this set is 0.
- The second data set has modest deviation away from 5 compared to the third data set, so the third data set will have the largest standard deviation!

Variance and Standard Deviation

- The *variance* of a data set with n values $x_1, x_2, \ldots, x_n$ is denoted by the symbol s^2, and is roughly the average squared deviation from the mean:

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}$$

- The *standard deviation* of a data set, denoted by the symbol s, is the square root of the variance:

$$s = \sqrt{s^2}.$$

Example

The data set $[2, 3, 10]$ has mean $\overline{x} = (2 + 3 + 10)/3 = 5$. The standard deviation will be

$$s = \sqrt{\frac{(2-5)^2 + (3-5)^2 + (10-5)^2}{3-1}} = \sqrt{\frac{9+4+25}{2}} = \sqrt{19} \approx 4.36.$$

Five Number Summary

The five number summary consists of the 5 statistics:

$$L \quad Q_1 \quad M \quad Q_3 \quad H$$

- L stands for ‘low’ - it is the minimum value.
- H stands for ‘high’ - it is the maximum value.
- M stands for median, as usual
- Q_1 stands for the first quartile, a number marking the 25% mark.
- Q_3 stands for the third quartile, a number marking the 75%mark.

The InterQuartile Range (IQR)

The InterQuartile Range

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

where Q_1 and Q_3 are the 25th and 75th percentiles, respectively.

The IQR represents the spread of the middle 50% of the data.

Determining the Five Number Summary by Hand

The five number summary: $L \quad Q_1 \quad M \quad Q_3 \quad H$

- L and H are found by inspection (the smallest and largest values).
- We've discussed above how to find the median, M.
- Q_1 is the median of the data less than or equal to M
- Q_3 is the median of the data greater than or equal to M.

Example: 5 number summary

Find the 5 number summary of this data set, which has an **even** number of values.

12 19 21 23 25 26 31 33 34 37

Example: 5 number summary

Find the 5 number summary of this data set, which has an **even** number of values.

$$\begin{array}{cccccccccc} 12 & 19 & 21 & 23 & \color{blue}{25} & \color{blue}{26} & 31 & 33 & 34 & 37 \\ \uparrow & & & & \multicolumn{2}{c}{\uparrow} & & & & \uparrow \\ L & & & & \multicolumn{2}{c}{M} & & & & H \end{array}$$

Example: 5 number summary

Q_1 is the median of the data less than or equal to the Median spot:

$$\underbrace{12 \quad 19 \quad 21 \quad 23 \quad 25}_{\substack{\uparrow \\ Q_1}} \quad 26 \quad 31 \quad 33 \quad 34 \quad 37$$

Example: 5 number summary

Q_3 is the median of the data greater than or equal to the Median spot:

$$12 \quad 19 \quad 21 \quad 23 \quad 25 \quad \underbrace{26 \quad 31 \quad \underset{\substack{\uparrow \\ Q_3}}{33} \quad 34 \quad 37}$$

Example: 5 number summary

All Together Now!

$$\begin{array}{cccccccccc} 12 & 19 & 21 & 23 & 25 & & 26 & 31 & 33 & 34 & 37 \\ \uparrow & & \uparrow & & & \uparrow & & & \uparrow & & \uparrow \\ L & & Q_1 & & & M & & & Q3 & & H \end{array}$$

- The five number summary is 12   21   25.5   33   37.
- The IQR is $33 - 21 = 12$.

Example: 5 number summary

Find the 5 number summary of this data set, which has an **odd** number of values.

2.5 2.5 2.9 3.1 3.1 3.4 3.7 3.9 4.0 4.2 4.3

Example: 5 number summary

Find the 5 number summary of this data set, which has an **odd** number of values.

$$\begin{array}{ccccccccccc} 2.5 & 2.5 & 2.9 & 3.1 & 3.1 & \color{blue}{3.4} & 3.7 & 3.9 & 4.0 & 4.2 & 4.3 \\ \uparrow & & & & & \uparrow & & & & & \uparrow \\ L & & & & & \color{blue}{M} & & & & & H \end{array}$$

Example: 5 number summary

Find the 5 number summary of this data set:
Q_1 is the median of the data less than or equal to the Median spot:

$$\underbrace{2.5 \quad 2.5 \quad 2.9 \quad 3.1 \quad 3.1 \quad 3.4}_{\substack{\uparrow \\ Q_1}} \quad 3.7 \quad 3.9 \quad 4.0 \quad 4.2 \quad 4.3$$

So $Q_1 = (2.9 + 3.1)/2 = 3.0$.

Example: 5 number summary

Find the 5 number summary of this data set:
Q_3 is the median of the data greater than or equal to the Median spot:

$$2.5 \quad 2.5 \quad 2.9 \quad 3.1 \quad 3.1 \quad \underline{3.4 \quad 3.7 \quad 3.9 \quad 4.0 \quad 4.2 \quad 4.3}$$

$$\uparrow$$
$$Q_3$$

So $Q_3 = (3.9 + 4.0)/2 = 3.95$

Example: 5 number summary

All Together Now!

$$\begin{array}{ccccccccccc} 2.5 & 2.5 & 2.9 & 3.1 & 3.1 & 3.4 & 3.7 & 3.9 & 4.0 & 4.2 & 4.3 \\ \uparrow & & \uparrow & & & \uparrow & & \uparrow & & & \uparrow \\ L & & Q_1 & & & M & & Q_3 & & & H \end{array}$$

- The five number summary is 2.5 3.0 3.4 3.95 4.3.
- The IQR is $3.95 - 3.0 = 0.95$.

Five Number Summary in R

- The `fivenum()` function in R returns a five number summary for a data set that matches this approach.
- For the even sample size example:
```
> fivenum(c(12, 19, 21, 23, 25, 26, 31, 33, 34, 37))
[1] 12.0 21.0 25.5 33.0 37.0
```
- For the odd sample size example:
```
> fivenum(c(2.5, 2.5, 2.9, 3.1, 3.1, 3.4, 3.7, 3.9, 4.0,
4.2, 4.3))
[1] 2.50 3.00 3.40 3.95 4.30
```

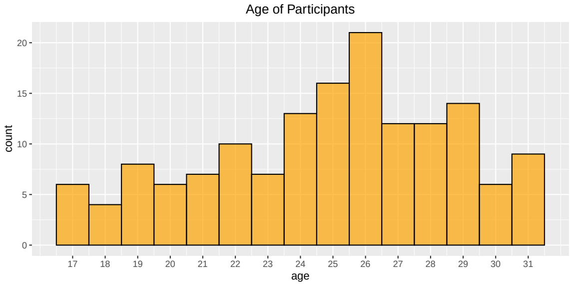
Robustness with measures of spread

- The standard deviation is greatly influenced by outliers.
- The IQR is not.

Example

data set 1 : $4, 6, 6, 7, 7, 8, 8, 10, 11, 12$
data set 2 : $4, 6, 6, 7, 7, 8, 8, 10, 11, 22$
Standard deviations: $s_1 = 2.47$ and $s_2 = 5.02$.
Q_1, and Q_3 are the same for both distributions (as are the medians), so the IQRs will be equal.

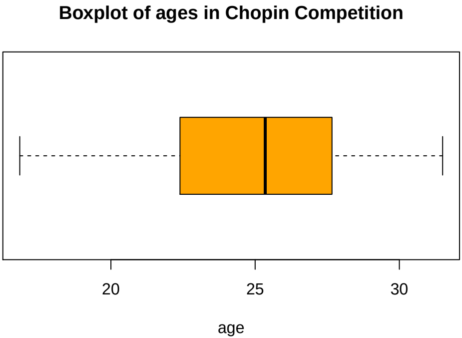
Chopin Competition Contestant Ages



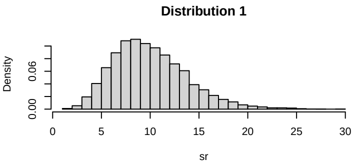
- Standard deviation is $s = 3.794$
- The 5 number summary is 16.84 22.40 25.35 27.67 31.50
- The middle half of the contestants run from 22.4 to 27.7 years old.

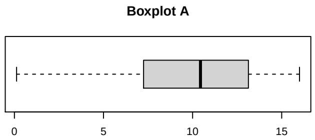
Box Plots

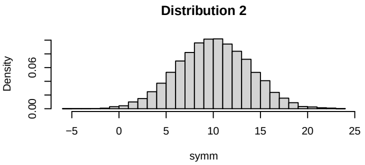
A **box plot** is a pictorial representation related to the 5 number summary. A middle box represents the range from Q_1 to Q_3, with the median M drawn inside the box. Then whiskers run down to L and up to H, unless outliers are taken into account.

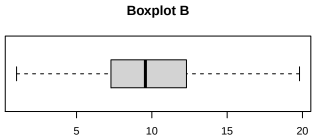


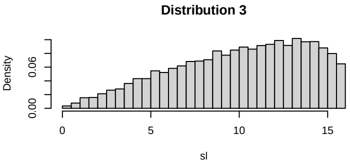
Match the distribution with the box plot

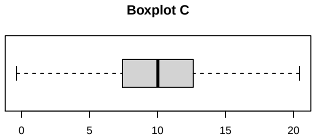








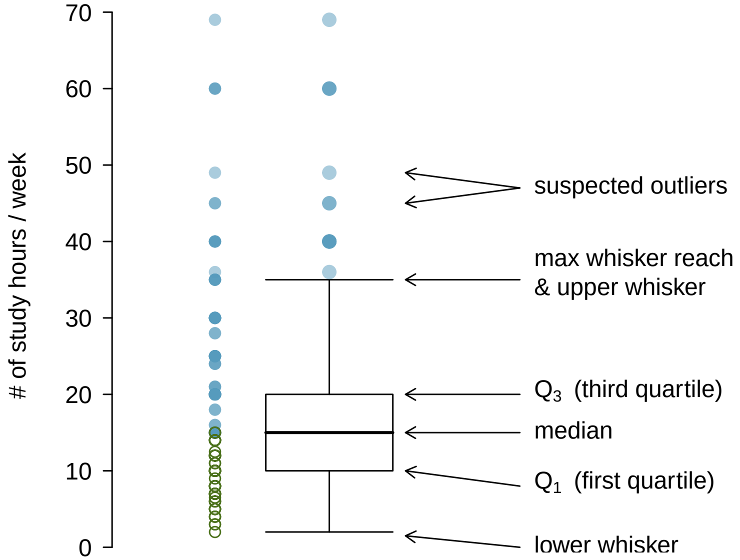




Mean and Standard Deviation vs Five Number Summary

- All three of the distributions in the previous example have mean 10.0 and standard deviation 3.8.
- The mean and standard deviation alone cannot detect skewness and they are also influenced by extreme values
- The five number summary (and corresponding box plot) captures skewness
- A standard box plot done with software will also earmark extreme values (outliers)

Anatomy of a box plot



Whiskers and outliers in RStudio

- *Whiskers* of a box plot can extend up to $1.5 \times IQR$ away from the quartiles.

$$\text{max upper whisker reach} = Q3 + 1.5 \times IQR$$
$$\text{max lower whisker reach} = Q1 - 1.5 \times IQR$$

In the previous slide:

$$\text{IQR} : 20 - 10 = 10$$
$$\text{max upper whisker reach} = 20 + 1.5 \times 10 = 35$$
$$\text{max lower whisker reach} = 10 - 1.5 \times 10 = -5$$

- A potential *outlier* is defined as an observation beyond the maximum reach of the whiskers. It is an observation that appears extreme relative to the rest of the data.

Outliers (cont.)

Q: *Why is it important to look for outliers?*

- *Identify extreme skew in the distribution.*
- *Identify data collection and entry errors.*
- *Provide insight into interesting features of the data.*

Section 2.2
Considering Categorical Data

18th International Chopin Competition

- The Chopin Competition data set includes the following variables most of which are categorical.

```
"player" "sex" "country" "country2" "birthdate"
"birthyear" "birthmonth" "age" "advance" "alpha"
"Perfnum" "perfdate" "order" "prelim1" "prelim2"
"prelim3" "prelim4" "prelim5" "prelim6"
```

- We can summarize a categorical variable with a table of counts or frequencies.
- We can summarize two categorical variables with a *contigency table*.

A table recording counts by country

`country`	Count
China	32
Japan	31
Poland	16
South Korea	14
Chinese Taipei	9
Canada	8
Italy	8
Russia	8
US	6
France	2
Germany	2
UK	2
13 countries*	1
Total	151

* Belarus, Bulgaria, Cuba, Greece, Hungary, Israel, Latvia, Malaysia, Romania, Spain, Thailand, Ukraine, Vietnam

A contingency table

Comparing a player's sex against whether they advanced.

		sex		
		F	M	Total
advance	yes	31	47	78
	no	40	33	73
	Total	71	80	151

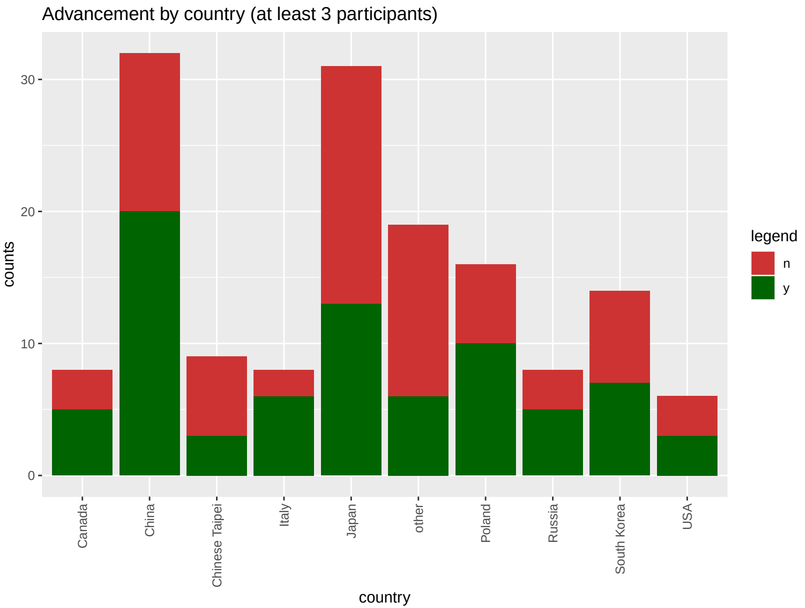
Country vs Advancement

		advance		
		yes	no	Total
country	China	20	12	32
	Japan	13	18	31
	Poland	10	6	16
	South Korea	7	7	14
	Chinese Taipei	3	6	9
	Canada	5	3	8
	Italy	6	2	8
	Russia	5	3	8
	US	3	3	6
	France	0	2	2
	Germany	0	2	2
	UK	1	1	2
	13 countries	5	8	13
	Total	78	73	151

Bar Plots

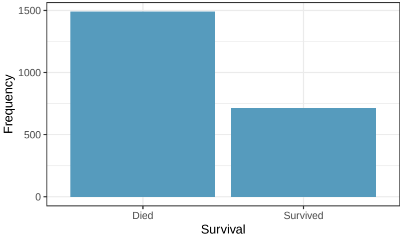
- A *bar plot* is a common way to display the distribution of a single categorical variable.
- We can make *stacked bar plots* to visualize a contingency table.

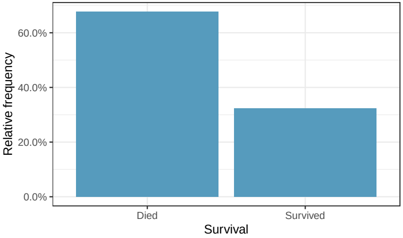
Contestants by Country and Advancement



Bar plots

- A bar plot where proportions instead of frequencies are shown is called a *relative frequency bar plot*.





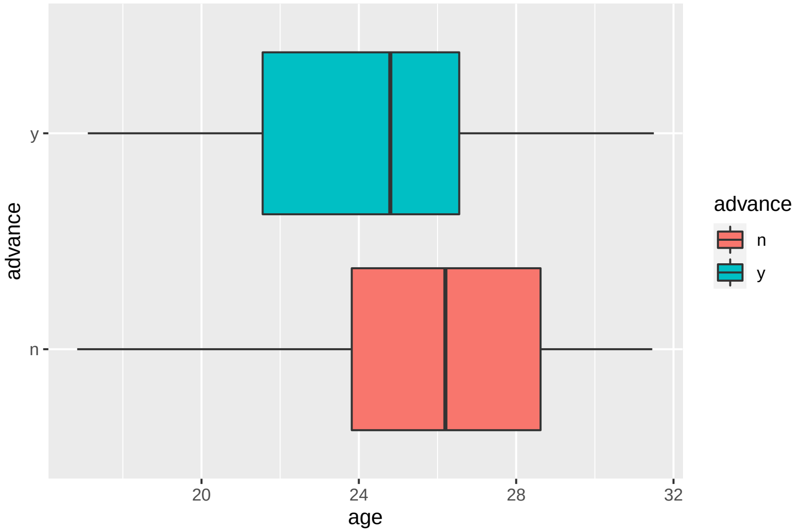
(Deaths and Survivors on the *Titanic*)

Q: *How are bar plots different than histograms?*

Bar plots - displaying ategorical variables, histograms - for numerical variables. The x-axis in a histogram is a number line, so bar order cannot change. In a bar plot, the categories can be listed in any preferred order.

Comparing Numerical Data Across Groups

- Group the data according to whether they advance
- Find and compare summary statistics for age within each group



Comparing Numerical Data Across Groups

Question: What question does this graphic address? What groups are being considered?