Lecture 6: Visualizing Numerical and Categorical Data

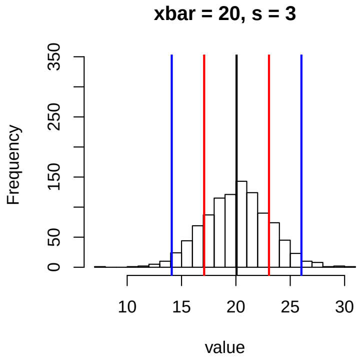
Chapter 1.6+1.7

Goals for Today

- Rule of thumb for standard deviations
- Population vs sample mean/variance/standard deviations
- Percentiles and Quartiles
- Boxplots
- Piecharts, barplots, mosaicplots

Rule of Thumb for Standard Deviations

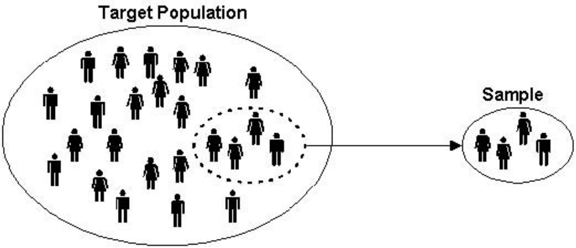
Example



- black line is mean $\overline{x}$
- red lines mark about $\frac{2}{3}$: $[\overline{x} - s, \overline{x} + s] = [20 - 3, 20 + 3] = [17, 23]$.
- blue lines mark about 95%: $[\overline{x} - 2s, \overline{x} + 2s] = [20 - 6, 20 + 6] = [14, 26]$.

Population vs Sample Mean/Variance/Standard Deviation

Recall the notion of taking a representative sample from a study/target population. Say we are interested in the income of the individuals.



Population vs Sample Mean/Variance/Standard Deviation

- The sample mean $\overline{x}$ is the mean income of the 4 sampled people.

Population vs Sample Mean/Variance/Standard Deviation

- The sample mean $\overline{x}$ is the mean income of the 4 sampled people.
- The population mean μ is the mean income of all 24 people in the target population.

Population vs Sample Mean/Variance/Standard Deviation

- The sample mean $\overline{x}$ is the mean income of the 4 sampled people.
- The population mean μ is the mean income of all 24 people in the target population.
- We say $\overline{x}$ estimates μ. If the sample is representative, then $\overline{x}$ estimates μ with high accuracy i.e. it is unbiased.

Population vs Sample Mean/Variance/Standard Deviation

	True Population Value	Sample Value
Mean	μ	$\overline{x}$
Variance	σ^2	s^2
Standard Deviation	σ	s

Population vs Sample Mean/Variance/Standard Deviation

	True Population Value	Sample Value
Mean	μ	$\overline{x}$
Variance	σ^2	s^2
Standard Deviation	σ	s

The sample value is used to estimate the (true) population value.

Percentiles

A percentile (%'ile) indicates the value below which a given %'age of observations fall.

Percentiles

A percentile (%'ile) indicates the value below which a given %'age of observations fall.

SAT Scores from 2012
http://media.collegeboard.com/digitalServices/pdf/research/SAT-Percentile-Ranks-2012.pdf

Percentiles

A percentile (%'ile) indicates the value below which a given %'age of observations fall.

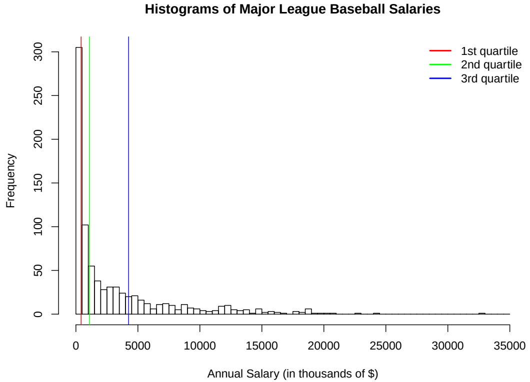
SAT Scores from 2012
http://media.collegeboard.com/digitalServices/pdf/research/SAT-Percentile-Ranks-2012.pdf

So for example, if you scored 700 in critical reading, 95% of college-bound seniors who took the test did worse.

Quartiles and IQR

MLB Data Quartiles

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
400.0	418.3	1094.0	3282.0	4250.0	33000.0



The IQR is (3rd Quartile - 1st Quartile) = 4250.0 - 418.3 = 3831.7
i.e the distance between the red and blue line.

Robust Statistics (Chapter 1.6.6)

Robust estimates are statistics where extreme observations (outliers) have less effect on their values, i.e. are more resistant to their effect. The median and IQR are two examples.

Robust Statistics (Chapter 1.6.6)

Robust estimates are statistics where extreme observations (outliers) have less effect on their values, i.e. are more resistant to their effect. The median and IQR are two examples.

Example: Old scoring system in figure skating: drop the highest & lowest scores and then take the average.

Robust Statistics (Chapter 1.6.6)

Robust estimates are statistics where extreme observations (outliers) have less effect on their values, i.e. are more resistant to their effect. The median and IQR are two examples.

Example: Old scoring system in figure skating: drop the highest & lowest scores and then take the average.

Say we have a figure skater who gets judged by countries V-Z:

Country	V	W	X	Y	Z
Score	4.0	5.2	5.2	5.3	6.0

Robust Statistics (Chapter 1.6.6)

Robust estimates are statistics where extreme observations (outliers) have less effect on their values, i.e. are more resistant to their effect. The median and IQR are two examples.

Example: Old scoring system in figure skating: drop the highest & lowest scores and then take the average.

Say we have a figure skater who gets judged by countries V-Z:

Country	V	W	X	Y	Z
Score	4.0	5.2	5.2	5.3	6.0

Drop the 4.0 and 6.0, then the final score is: $\frac{5.2+5.2+5.3}{3} = 5.23$

Boxplots

Boxplots are visual summaries of a sample $x_1, \ldots, x_n$ that bring to light unusual values (potential outliers):

Boxplots

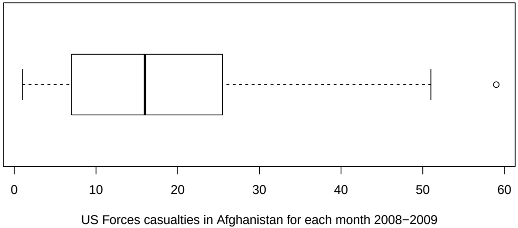
Boxplots are visual summaries of a sample $x_1, \ldots, x_n$ that bring to light unusual values (potential outliers):

Example: # US Forces casualties in the war in Afghanistan for each month from 2008-2009:

7, 1, 7, 5, 16, 28, 20, 22, 27, 16, 1, 3, 14, 15, 13, 6, 12, 24, 44, 51, 37, 59, 17, 17

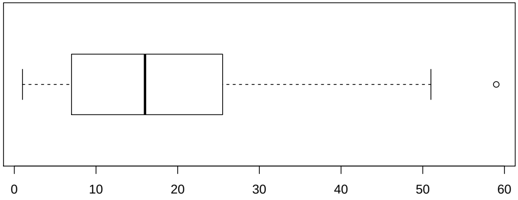
Boxplots

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.00	7.00	16.00	19.25	24.75	59.00



Boxplots

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.00    7.00   16.00   19.25   24.75   59.00
```



US Forces casualties in Afghanistan for each month 2008−2009

Page 29 of text describes the length of the whiskers: they capture data that is no more than $1.5 \times IQR$ of both ends of the box.

Outliers Are Relatively Extreme

An outlier is an observation that appears extreme relative to the rest of the data.

Outliers Are Relatively Extreme

An outlier is an observation that appears extreme relative to the rest of the data.

Why it is important to look for outliers? Examination of data for possible outliers serves many useful purposes, including

Outliers Are Relatively Extreme

An outlier is an observation that appears extreme relative to the rest of the data.

Why it is important to look for outliers? Examination of data for possible outliers serves many useful purposes, including

- Identifying strong skew in the distribution.

Outliers Are Relatively Extreme

An outlier is an observation that appears extreme relative to the rest of the data.

Why it is important to look for outliers? Examination of data for possible outliers serves many useful purposes, including

- Identifying strong skew in the distribution.
- Identifying data collection or entry errors.

Outliers Are Relatively Extreme

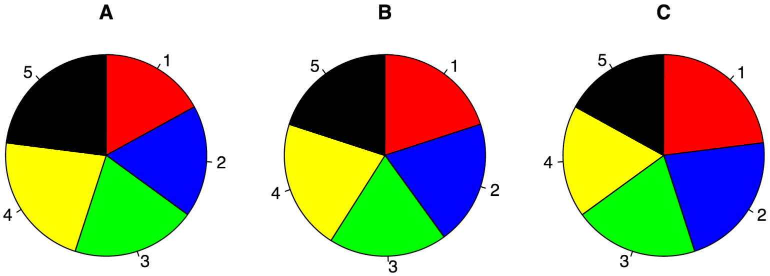
An outlier is an observation that appears extreme relative to the rest of the data.

Why it is important to look for outliers? Examination of data for possible outliers serves many useful purposes, including

- Identifying strong skew in the distribution.
- Identifying data collection or entry errors.
- Providing insight into interesting properties of the data.

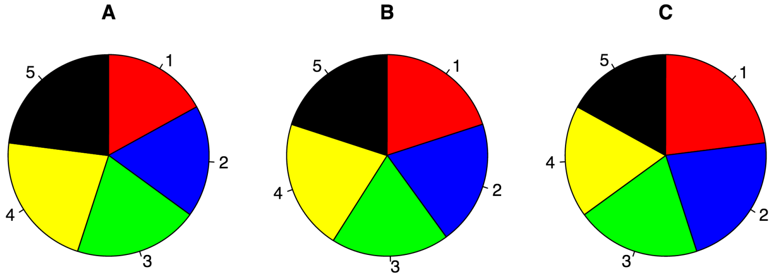
Piecharts

Say we have the following piecharts represent the polling from a local election with five candidates (1-5) at three different time points A, B, an C:



Piecharts

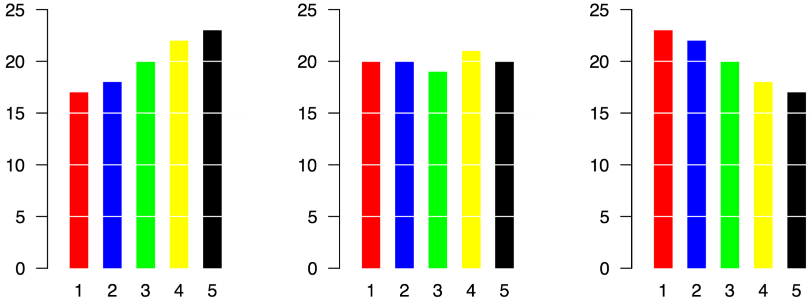
Say we have the following piecharts represent the polling from a local election with five candidates (1-5) at three different time points A, B, an C:



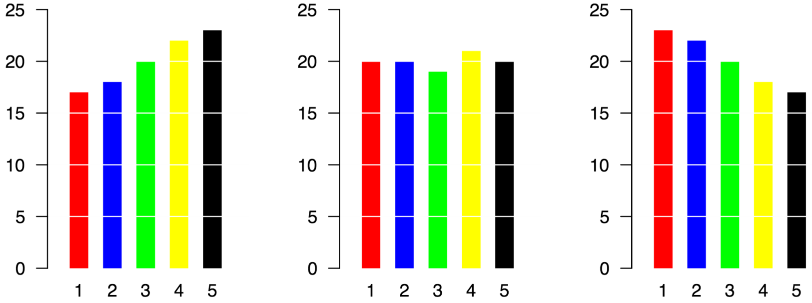
Answer the following questions:

- In the first race, is candidate 5 doing better than candidate 4?
- Who did better between time A and time B, candidate 2 or candidate 4?

Barplots Instead



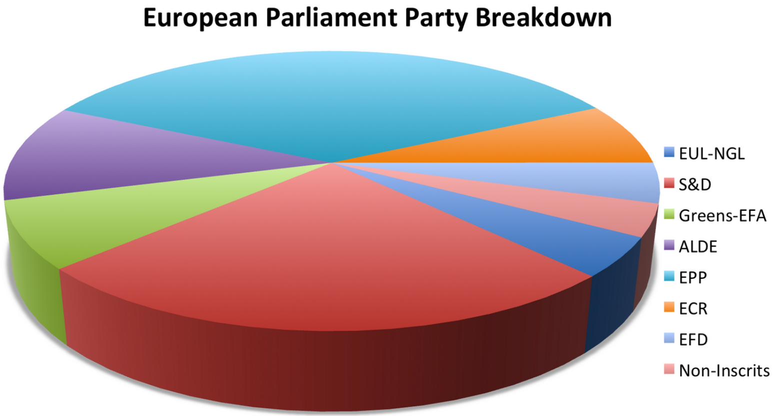
Barplots Instead



Answers:

- Candidate 5 is doing better than 4
- Between A and B, candidate 2 went from about 17% to 20% while candidate went from about 22% to 21%. So candidate 2 did better

3D Piecharts Can Be Deceiving



EEP (teal) has 266 seats, whereas S&D (red) has 190 seats.

Titanic Survival Data

Typing `data(Titanic)` in R loads the survival and death counts, split by each of the following categories:

Titanic Survival Data

Typing `data(Titanic)` in R loads the survival and death counts, split by each of the following categories:

- Class: 1st, 2nd, 3rd, or crew (4 levels)
- Gender (2 levels)
- Age: Child or adult (2 levels)

i.e. $4 \times 2 \times 2 = 16$ possible groups to consider.

Questions

Titanic Survival Data

Typing `data(Titanic)` in R loads the survival and death counts, split by each of the following categories:

- Class: 1st, 2nd, 3rd, or crew (4 levels)
- Gender (2 levels)
- Age: Child or adult (2 levels)

i.e. $4 \times 2 \times 2 = 16$ possible groups to consider.

Questions

- What was the effect of class (1st, 2nd, 3rd, crew) on your chances of survival?

Titanic Survival Data

Typing `data(Titanic)` in R loads the survival and death counts, split by each of the following categories:

- Class: 1st, 2nd, 3rd, or crew (4 levels)
- Gender (2 levels)
- Age: Child or adult (2 levels)

i.e. $4 \times 2 \times 2 = 16$ possible groups to consider.

Questions

- What was the effect of class (1st, 2nd, 3rd, crew) on your chances of survival?
- Did the "women and children" first lifeboat policy hold?

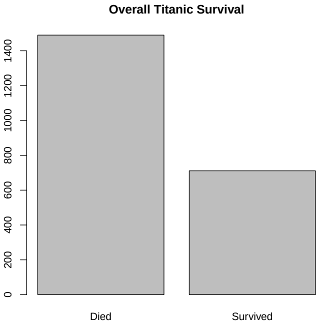
Frequency Table

A table summarizing a single categorial variable is called a frequency table. Overall:

Died	1490
Survived	711
Total	2201

Barplot

Barplots are ways to display categorial variables:



Contingency Table

A table that cross-classifies two categorical variables is a contingency table. Now let's split survival by class: 1st, 2nd, 3rd, and crew.

Contingency Table

A table that cross-classifies two categorical variables is a contingency table. Now let's split survival by class: 1st, 2nd, 3rd, and crew.

Before:

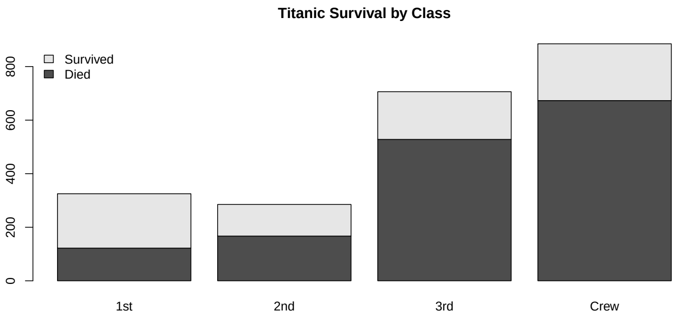
Died	1490
Survived	711
Total	2201

After:

	1st	2nd	3rd	Crew	Total
Died	122	167	528	673	1490
Survived	203	118	178	212	711
Total	325	285	706	885	2201

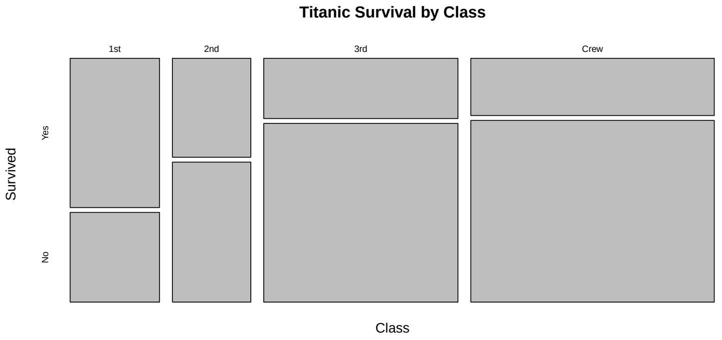
Stacked Barplot

Stacked barplots are one way to display values from a contingency table:



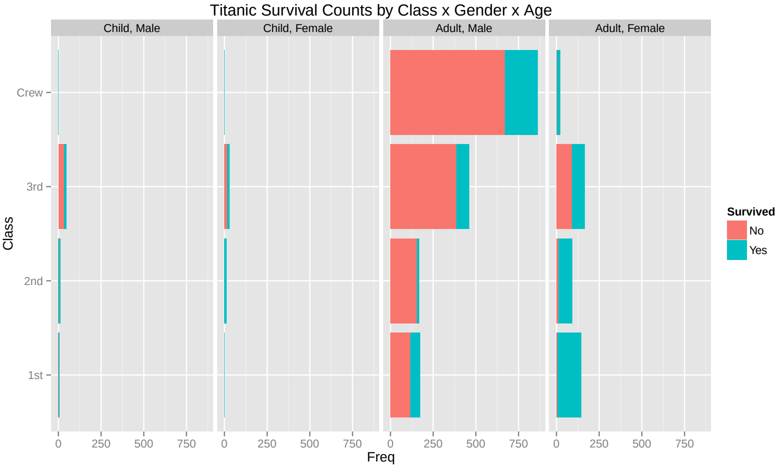
Mosaic Plots

Mosaic plots are similar, but the widths of the bars now reflect proportions:



Stacked Barplots

Using the ggplot2 package, we can plot survivals by class, age, and gender all at once.



Standardized/Normalized Stacked Barplots

Instead of raw counts, we can expand each bar to reflect proportions (i.e. standardize/normalize them).