Lecture 04: Describing data with numbers

Learning objectives for today:

Describing your data:

1. Investigate measures of centrality
   - mean and median, and when they're the same vs. different
2. Investigate measures of spread
   - IQR, standard deviation, and variance
3. Create a visualization of the "five number summary"
   - boxplots using `ggplot`
4. Calculate the variance and standard deviation

Measures of central tendency

Measures of central tendency

- Most common: mean and median

The arithmetic mean

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

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

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

The median

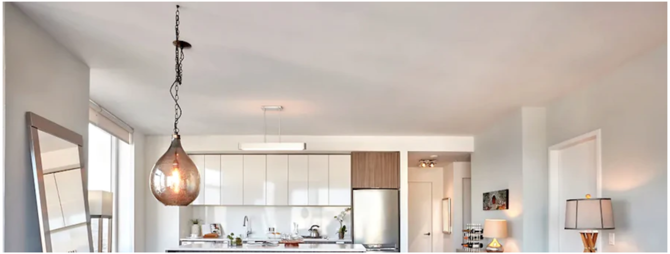
- Half of the measurements are larger and half are smaller.
  - What is the median if there is an odd number of observations?
  - An even number?

Statistics is Everywhere

Bay Area rent

San Francisco

Apartments for rent in San Francisco: What will $3,400 get you?



From Hoodline.com

Bay Area rent


Business > Real Estate • News

Here's how much rent has gone up in the Bay Area since 2010

In San Francisco, tenants are paying 70 percent more than in 2010

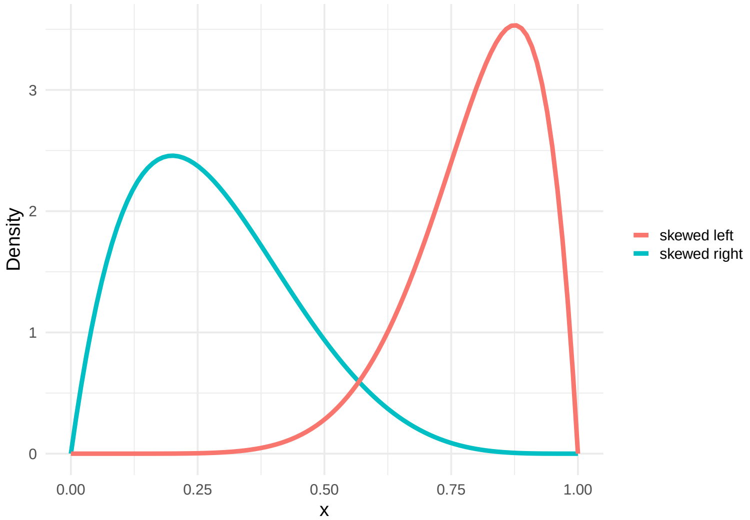
Now sitting at \$3,680, average rent in San Francisco has soared 70 percent since 2010 while home prices climbed an eye-popping 95 percent and median income crept up a comparatively modest 61 percent. Across the bay in Oakland, rent climbed even more — 108 percent. Mercury News article

Discussion

When are these measures approximately equal?

- Answer: When the data has one peak and is roughly symmetric
  - In this case, the mean $\approx$ median, so provide either one in a summary
- Skewed data
  - mean $\neq$ median
  - Right-skewed data will commonly have a ________ mean than median
  - Left-skewed data will commonly have a ________ mean than median
  - Which statistic should we report?

Skewed data



Apartment rent in SF

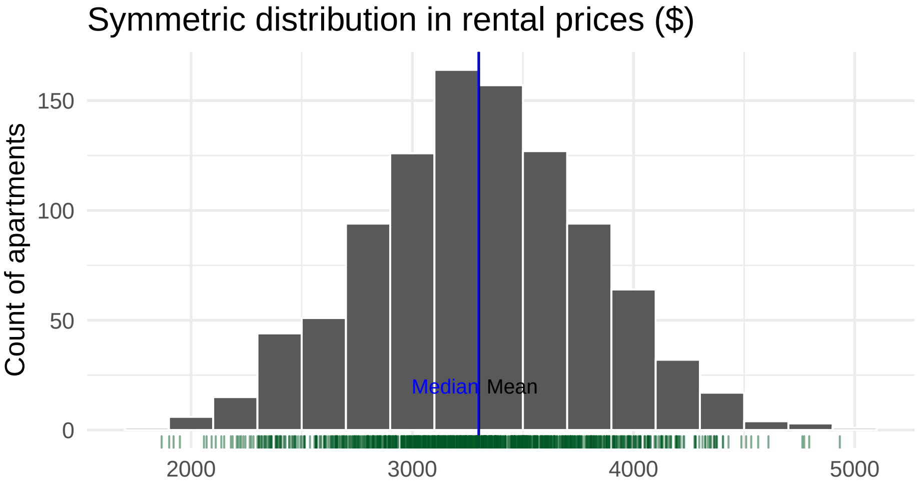
Problem: We want to understand how much it costs for a new resident to rent a 1 bedroom apartment in San Francisco

Plan: Take a sample of 1000 apartment units listed for rent (currently available) and ask the rental price (excluding utilities)

Data: Here I will present data that I simulated in r using a mean value published on rentjungle.com - you will not be expected to do this or be tested on it.

Example: Apartment rent in SF

Suppose that the distribution of rent prices looked like this. The green ticks underneath the histograms shows you the exact rent values that contribute data to each bin.



Example: Apartment rent in SF

From last lecture: We describe this distribution in terms of center, shape and spread:

- Center: Where is the center of the distribution?
- Shape: Is this distribution unimodal or bimodal?
- Spread: How much variability is there between the lowest and highest rent values?

Example: Apartment rent in SF

Summarizing numerically: Center:

```r
# in base R
mean(rent_data[,"sym"])
```

```
## [1] 3301.662
```

```r
median(rent_data[,"sym"])
```

```
## [1] 3298.832
```

```r
# using the summarize function and a pipe operator
rent_data %>% summarize(
  mean=mean(sym),
  median = median(sym))
```

```
##          mean   median
## 1 3301.662 3298.832
```

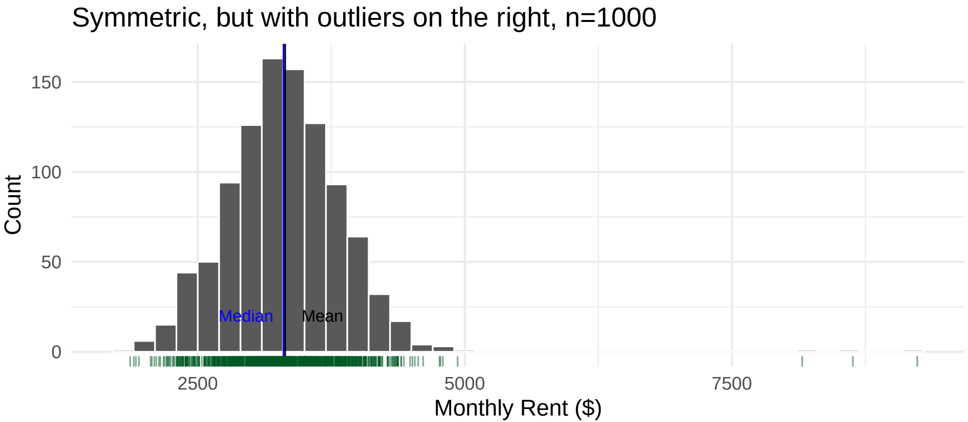
Mean vs Median: Outliers and sample size, skew, shape

When are the mean and median approximately equal?

- If your data has one peak (unimodal), is roughly symmetric, and does not have outliers
  - mean $\approx$ median, so provide either one in a summary

Example: Apartment rent in SF

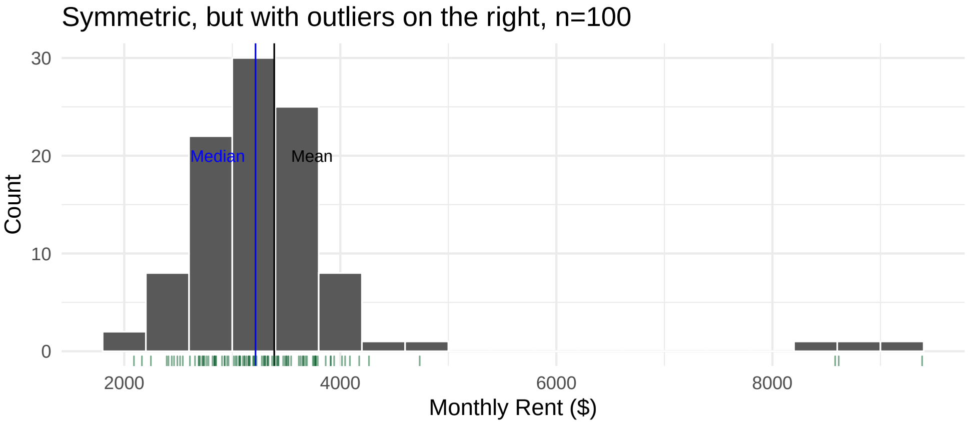
Now suppose that there were three rents within the data set with much larger values than the rest of the distribution. Here is the plot for this updated data.



- With 1000 sampled points the outliers do not have a large effect on the mean

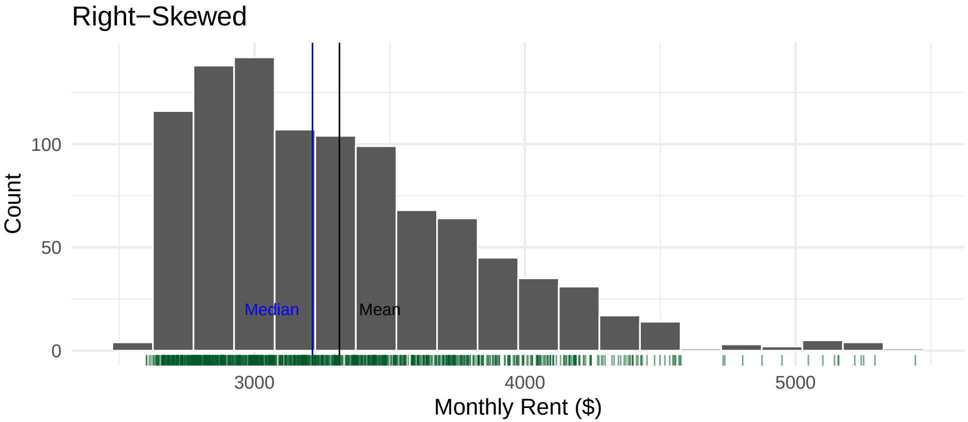
Example: Apartment rent in SF

Imagine instead, there were only 100 sampled points. Here, the outliers have a larger effect on the mean. The mean is not resistant to outliers.



Example: Apartment rent in SF

Consider instead what happens if there are many high-end apartments in the area. Here is the histogram of data for this example:



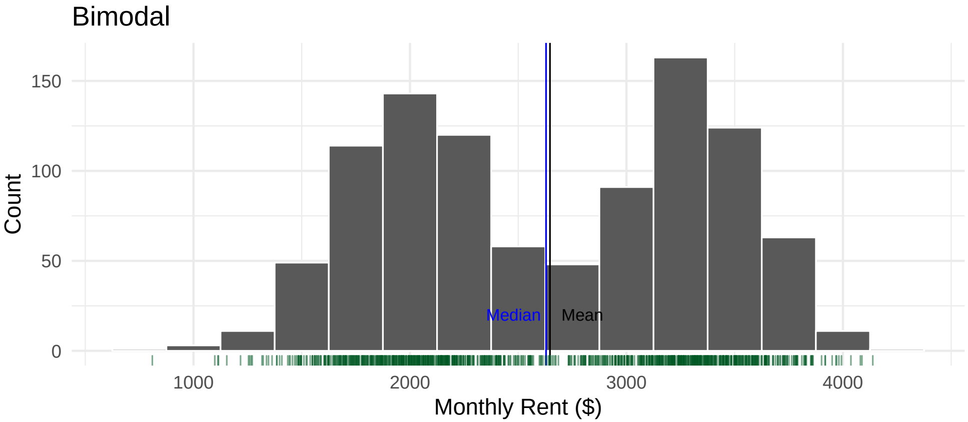
Why is the mean larger than the median in this case?

Skewed data

- mean $\neq$ median
- Data with a long right tail will commonly have a ________ mean than median
- Data with a long left tail will commonly have a ________ mean than median
- Which statistic should we report?

Example: Apartment rent in SF

Now, suppose that the sample of estimates did not look like the distribution in the previous example. Instead, it looked like this:



Describe the distribution. How does it differ from the first plot? Would you want to provide the mean or median for these data?

Summary of measures of central tendency

- The mean and median are similar when the distribution is symmetric
- Outliers affects the mean and pull it towards their values. But they do not have a large effect on the median.
- Skewed distributions also pull the mean out into the tail.
- Measures of central tendency are not very helpful in multi-modal distributions

Measures of spread

The inter-quartile range (IQR)

- Q1 is the 1st quartile/the 25th percentile.
  - 25% of individuals have measurements below Q1.
- Q2 is the 2nd quartile/the 50th percentile/the median.
  - 50% of individuals have measurements below Q2.
- Q3, the 3rd quartile/the 75th percentile.
  - 75% of individuals have measurements below Q3.
- Q1-Q3 is called the inter-quartile range (IQR).
  - What percent of individuals lie in the IQR?
- Know how to find Q1, Q2, and Q3 by hand for small lists of numbers

Quantiles using R

```
quantile(variable, 0.25)
```

```
rent_data %>% summarize(
  Q1 = quantile(sym, 0.25),
  median = median(sym),
  Q3 = quantile(sym, 0.75)
  )
```

```
##          Q1   median       Q3
## 1 2981.445 3298.832 3629.012
```

R's quantile function: Note

- `quantile(variable, 0.25)` will not always give the exact same answer you calculate by hand
- The R function is optimized for its statistical properties and is slightly different than the book's method
- To get the exact same answer as by hand use `quantile(data, 0.25, type = 2)`
- You may use either one in this class. Most commonly, people do not specify `type=2`

Another measure of spread: The (full) range

- The difference between the minimum and maximum value

Concise information about spread and center: The five number summary

- The five number summary (min, Q1,median,Q3, max) is a quick way to communicate a distribution's center and spread.
- Based on the summary you can describe the full range of a dataset, where the middle 50% of the data lie, and the middle value.

dplyr's summarize() to calculate the five number summary

Using our original example of rent data:

```
rent_data %>% summarize(
  min = min(sym),
  Q1 = quantile(sym, 0.25),
  median = median(sym),
  Q3 = quantile(sym, 0.75),
  max = max(sym)
)
```

```
##          min       Q1   median       Q3     max
## 1 1866.829 2981.445 3298.832 3629.012 4932.54
```

Example: Hospital cesarean delivery rates

Example: Hospital cesarean delivery rates

These data were provided by the first author (Kozhimannil) of a manuscript published in the journal *Health Affairs*. link

From the article: Cesarean delivery is the most commonly performed surgical procedure in the United States, and cesarean rates are increasing. In its Healthy People 2020 initiative, the Department of Health and Human Services put forth clear, authoritative public health goals recommending a 10 percent reduction in both primary and repeat cesarean rates, from 26.5 percent to 23.9 percent, and from 90.8 percent to 81.7 percent, respectively.

A targeted approach to achieving such reductions might focus on hospitals with exceptionally high cesarean rates. However, adopting such a strategy requires quantification of hospital-level variation in cesarean delivery rates.

Example: Hospital cesarean delivery rates

Problem: To characterize the variation in cesarean rates between Hospitals in the United States

Plan: Collect existing data from a variety of institutions for one year and compare rates of cesarean delivery. They also looked at cesarean rates among only low risk births at each institution. Why might this be important?

Data: For this article, they worked with 2009 data from 593 US hospitals nationwide

Example: Hospital cesarean delivery rates

We start by importing the data:

```r
library(readxl)
# this library helps with reading xlsx and xls files into R
CS_dat <- read_xlsx("Kozhimannil_Ex_Cesarean.xlsx", sheet = 1)
```

Example: Hospital cesarean delivery rates

```r
head(CS_dat)
```

```
## # A tibble: 6 x 6
##   Births HOSP_BEDSIZE cesarean_rate lowrisk_cesarea~ `Cesarean rate ~
##    <dbl>        <dbl>         <dbl>            <dbl>            <dbl>
## 1    767            1         0.344           0.107              34.4
## 2    183            1         0.454           0.186              45.4
## 3    668            1         0.430           0.195              43.0
## 4    154            1         0.279           0.0844             27.9
## 5    327            1         0.306           0.119              30.6
## 6   2356            1         0.301           0.0662             30.1
## # ... with 1 more variable: `Low Risk Cearean rate*100` <dbl>
```

Example: Hospital cesarean delivery rates

```r
names(CS_dat)
```

```
## [1] "Births"                  "HOSP_BEDSIZE"
## [3] "cesarean_rate"           "lowrisk_cesarean_rate"
## [5] "Cesarean rate *100"      "Low Risk Cearean rate*100"
```

let's take a moment to discuss variable names containing spaces

Sidenote on variable names containing spaces

- Two variables in `CS_dat` contain spaces.
- We generally want to remove spaces from variable names.
- Question: Which `dplyr` function can we use to change the variable names?
- Answer: `rename(new_name = old_name)` can be used. When the old variable name contains spaces, you need to place back ticks around it like this:

```
CS_dat <- CS_dat %>% rename(cs_rate = `Cesarean rate *100`,
                            low_risk_cs_rate = `Low Risk Cearean rate*100`)
```

- See this paper for tips on storing data in Excel for later analysis.

Tidy the data for analysis

For our example, we are only interested in each hospital's cesarean delivery rate, the rate for lower risk pregnancies, and the number of births at the hospital.

```
CS_dat <- CS_dat %>%
  select(Births, cs_rate,low_risk_cs_rate) %>%
  rename(num_births = Births)
```

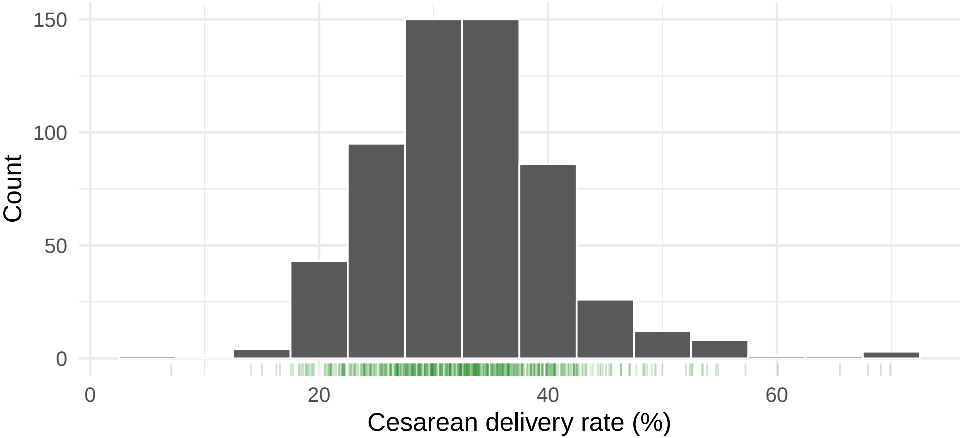
Analysis: Histogram of cesarean delivery rates across US hospitals

```
ggplot(CS_dat, aes(x = cs_rate)) +
geom_histogram(col = "white", binwidth = 5) +
labs( x = "Cesarean delivery rate (%)", y = "Count",

caption = "Data from: Kozhimannil, Law, and Virnig. Health Affairs. 2013;32(3

geom_rug(alpha = 0.2, col = "forest green") + #alpha controls transparency
theme_minimal(base_size = 15)
```

Histogram of cesarean delivery rates across US hospitals



Data from: Kozhimannil, Law, and Virnig. Health Affairs. 2013;32(3):527−35.

Spread of cesarean delivery rates across US hospitals

- What can you say about this distribution? Would you expect so much variation across hospitals in their rates of cesarean delivery?
- Let's describe the spread of these data using the methods from Chapter 2.

Quantiles

```
CS_dat %>% summarize(
  Q1 = quantile(cs_rate, 0.25),
  median = median(cs_rate),
  Q3 = quantile(cs_rate, 0.75)
  )
```

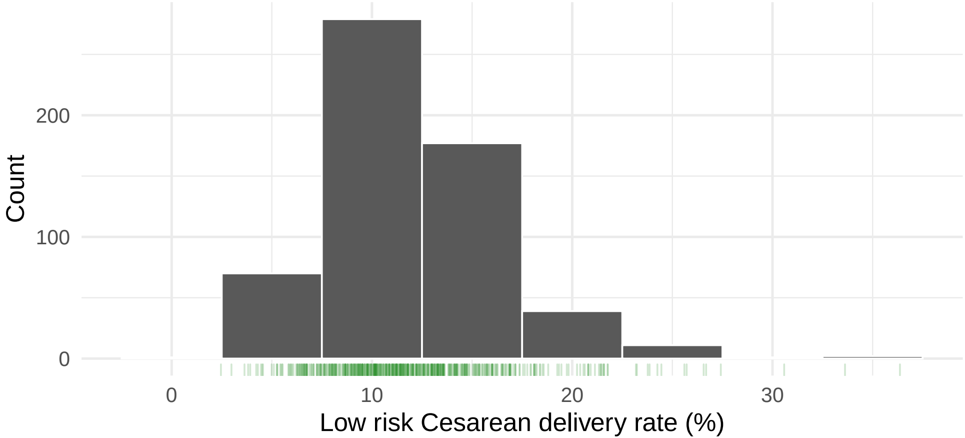
```
## # A tibble: 1 x 3
##      Q1 median    Q3
##   <dbl>  <dbl> <dbl>
## 1  27.6   32.4  37.1
```

dplyr's summarize() to calculate the five number summary

```
CS_dat %>% summarize(
  min = min(cs_rate),
  Q1 = quantile(cs_rate, 0.25),
  median = median(cs_rate),
  Q3 = quantile(cs_rate, 0.75),
  max = max(cs_rate)
)
```

```
## # A tibble: 1 x 5
##     min    Q1 median    Q3   max
##   <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1  7.09  27.6   32.4  37.1  69.9
```

Histogram of low risk cesarean delivery rates across US hospitals



Data from: Kozhimannil, Law, and Virnig. Health Affairs. 2013;32(3):527−35.

dplyr's summarize() to calculate the five number summary

```
CS_dat %>% summarize(
  min = min(low_risk_cs_rate),
  Q1 = quantile(low_risk_cs_rate, 0.25),
  median = median(low_risk_cs_rate),
  Q3 = quantile(low_risk_cs_rate, 0.75),
  max = max(low_risk_cs_rate)
)
```

```
## # A tibble: 1 x 5
##      min    Q1 median    Q3   max
##    <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1   2.46  9.19   11.4  14.2  36.4
```

Sample variance and standard deviation

Sample variance and standard deviation

Let s^2 represent the variance of a sample. Then,

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

$$s^2 = \frac{1}{n-1}\left((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2\right)$$

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

Let s represent the standard deviation of a sample. Then,

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

Sample variance and standard deviation

- Some intuition on why we divide by n-1: link

dplyr's summarize() to calculate the standard deviation and the variance

```
CS_dat %>% summarize(
  cs_sd = sd(cs_rate),
  cs_var = var(cs_rate)
)
```

```
## # A tibble: 1 x 2
##   cs_sd cs_var
##   <dbl>  <dbl>
## 1  8.03   64.5
```

Participation

Example: Hospital cesarean delivery rates

What might we conclude from these data?

Example: Hospital cesarean delivery rates

From the article:

"we found that cesarean rates varied tenfold across hospitals, from 7.1 percent to 69.9 percent. Even for women with lower-risk pregnancies, in which more limited variation might be expected, cesarean rates varied fifteenfold, from 2.4 percent to 36.5 percent. Thus, vast differences in practice patterns are likely to be driving the costly overuse of cesarean delivery in many US hospitals."

Box plots

Box plots provide a nice visual summary of the center and spread

Also called box and whisker plots

The box:

- The centre line is the median
- The top of the box is the Q3
- The bottom of the box is the Q1

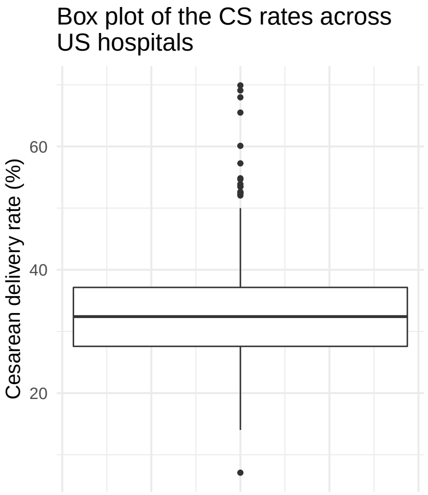
The whiskers - depends:

- The top of the top whisker is either the max value, or equal to the highest point that is below Q3 + 1.5*IQR
- The bottom of the bottom whisker is either min value, or equal to the lowest point that is above Q1 - 1.5*IQR
- In plots where the whiskers are not the min and max, the data points above and below the whiskers are the outliers

Box plots in R

```
ggplot(CS_dat, aes(y = cs_rate)) +
geom_boxplot() +
ylab("Cesarean delivery rate (%)") +
labs(title = "Box plot of the CS rates across US hospitals",
    caption = "Data from: Kozhimannil et al. 2013.") +
theme_minimal(base_size = 15) +
scale_x_continuous(labels = NULL) # removes the labels from the x axis
```

Box plots provide a nice visual summary of the center and spread



Data from: Kozhimannil et al. 2013.

R Recap: What new functions did we use?

1. `quantile(data, 0.25)`, `quantile(data, 0.75)` for Q1 and Q3, respectively
2. `min()` and `max()` for the full range of the data
3. `sd()` and `var()` for sample standard deviation and variance
4. Used the above within `summarize()` to easily output these measures
5. ggplot's `geom_boxplot`

Parting Humor