Describing Distributions with Numbers (Chapter 2 in book)

Instructors: Alan Hubbard and Tomer Altman

Sept 6, 2024

Learning objectives for today:

1. Investigate measures of centrality
   - mean and median, and when they're the same vs. different
2. Investigate measures of spread
   - Inter-quartile range (IQR), standard deviation, and variance
3. Create boxplots using `ggplot`

Measures of central tendency

- The most common measures of central tendency are the **mean** and the **median**

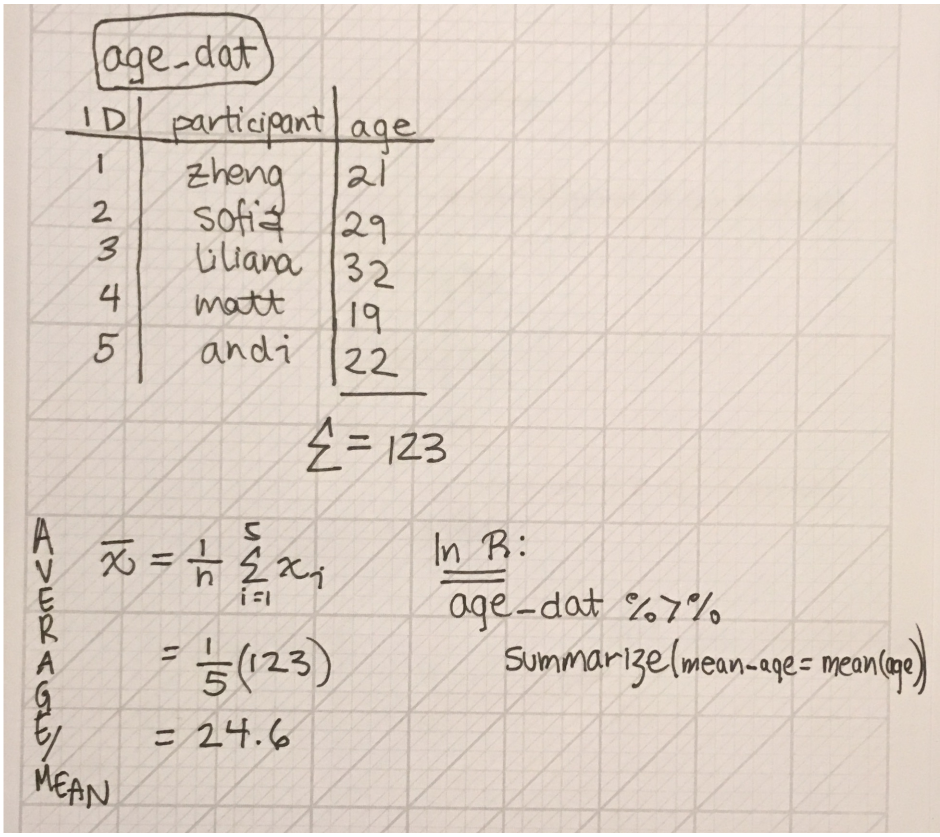
The arithmetic mean

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

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

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

Mean calculation



age_dat

ID	participant	age
1	zheng	21
2	sofia	29
3	liliana	32
4	matt	19
5	andi	22

$\Sigma = 123$

AVERAGE/ MEAN

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{5} x_i$$

$$= \frac{1}{5}(123)$$

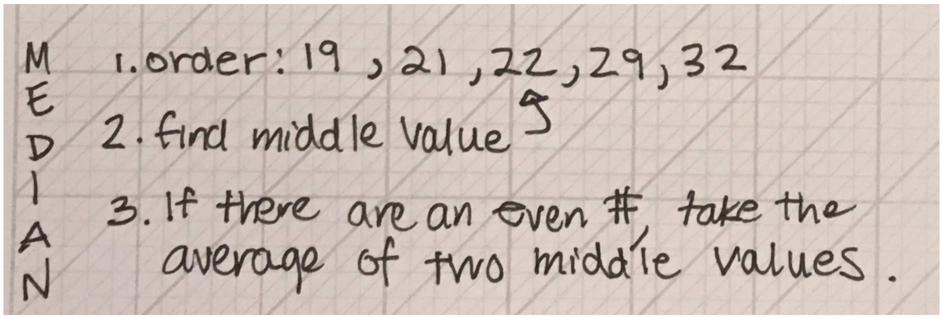
$$= 24.6$$

In R:

```
age_dat %>%
    summarize(mean_age = mean(age))
```

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?



MEDIAN

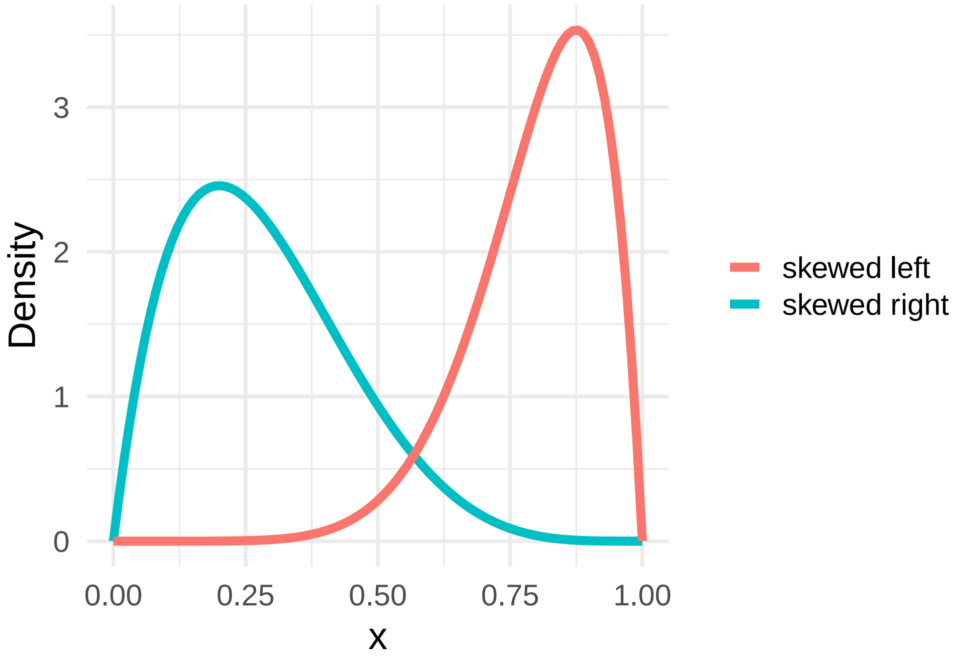
1. order: 19, 21, 22, 29, 32
2. find middle value
3. If there are an even #, take the average of two middle values.

When are these measures approximately equal?

- Answer: When the data has one peak and is roughly **symmetric**
  - In this case, the mean ≈ median, so provide either one in a summary
- **Skewed** data
  - mean ≠ median

  - Right-skewed data will commonly have a higher mean than median
  - Left-skewed data will commonly have a lower mean than median
  - Which statistic should we report? It depends, the median gives a more typical value because 50% of measures are above and below, but the average is important when resource planning.‘

Skewed data



Example: Apartment rent in SF

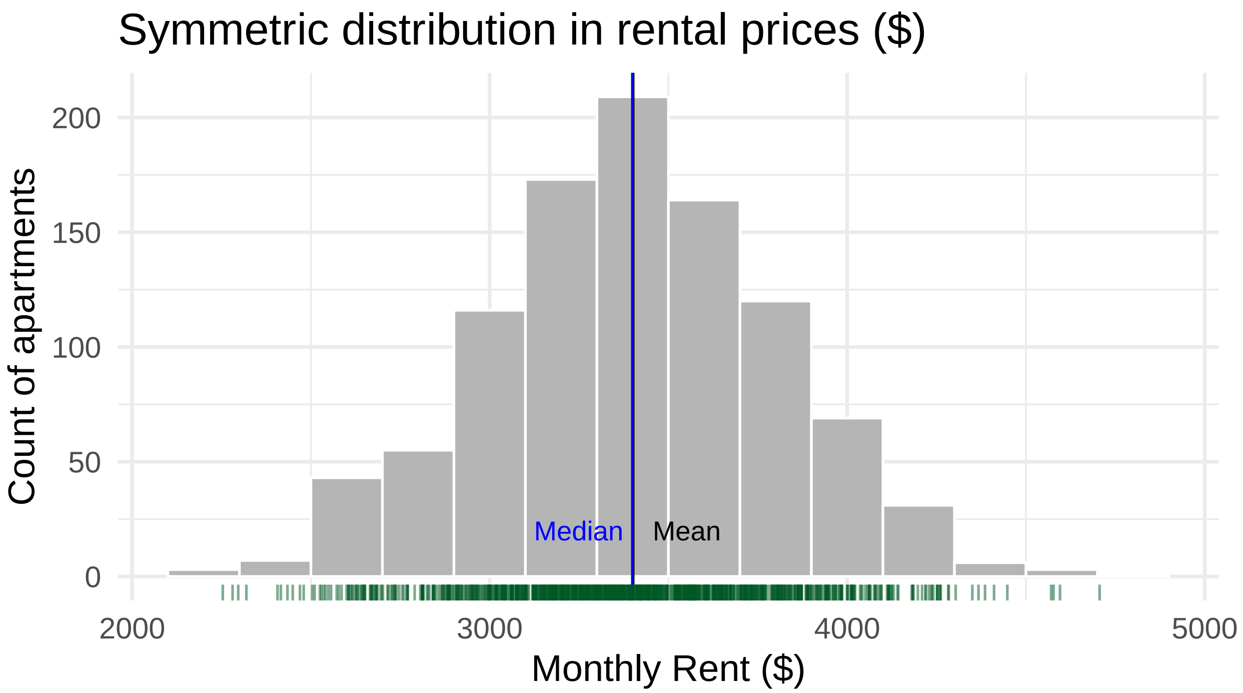
- Let’s imagine that we sampled 1000 apartment units across San Francisco and asked for the monthly rental price (before utilities).

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.

```
## Warning in geom_text(aes(x = rent_summs$sym_mean + 150, y = 20), label = "Mean", : All aesthetics ha
## i Please consider using ‘annotate()‘ or provide this layer with data containing
##   a single row.
```

```
## Warning in geom_text(aes(x = rent_summs$sym_median - 150, y = 20), label = "Median", : All aesthetics
## i Please consider using ‘annotate()‘ or provide this layer with data containing
##   a single row.
```



- For those wondering, here is an easier way to pick colors for plotting: Hex color picker

Example: Apartment rent in SF

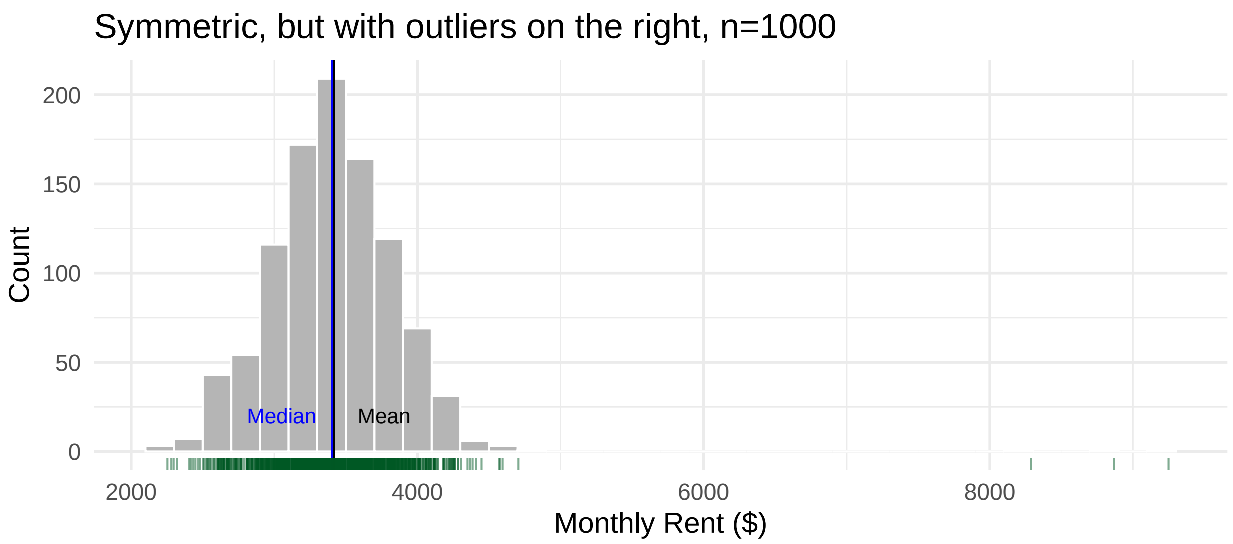
- Is this distribution unimodal or bimodal?
- What do you notice about the estimates of the mean and median in this case?

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.

```
## Warning in geom_text(aes(x = rent_summs$sym_out_right_mean + 350, y = 20), : All aesthetics have leng
## i Please consider using ‘annotate()‘ or provide this layer with data containing
##   a single row.
```

```
## Warning in geom_text(aes(x = rent_summs$sym_out_right_median - 350, y = 20), : All aesthetics have le
## i Please consider using ‘annotate()‘ or provide this layer with data containing
##   a single row.
```



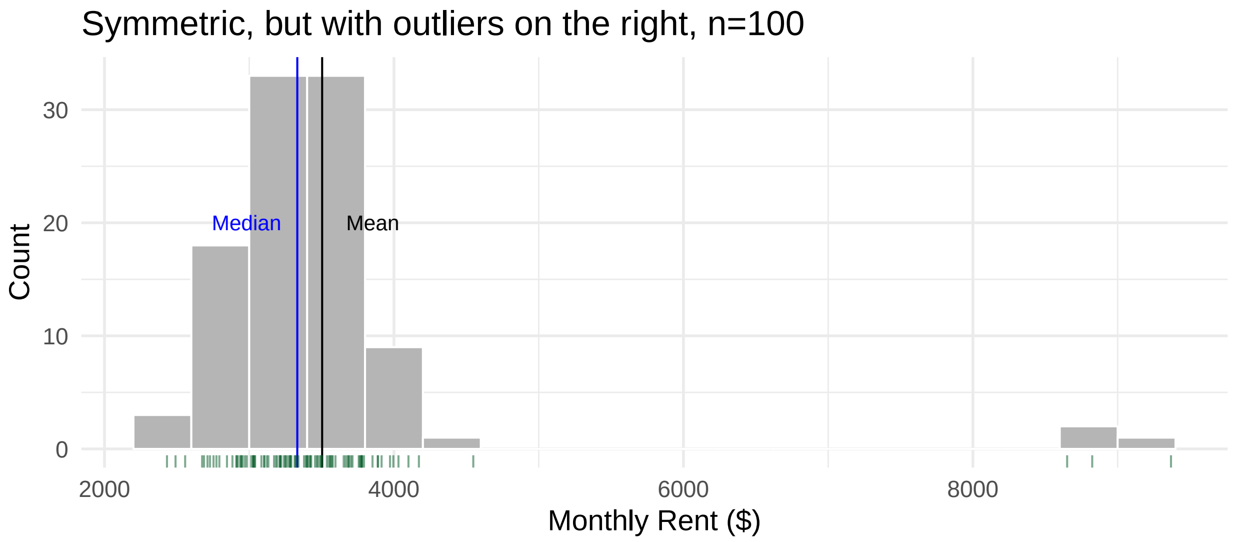
- With 1000 sampled points the outliers do not have a large effect on the mean

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.**

```
## Warning in geom_text(aes(x = rent_summs_small$sym_out_right_mean + 350, : All aesthetics have length
## i Please consider using ‘annotate()‘ or provide this layer with data containing
##   a single row.
```

```
## Warning in geom_text(aes(x = rent_summs_small$sym_out_right_median - 350, : All aesthetics have leng
## i Please consider using ‘annotate()‘ or provide this layer with data containing
##   a single row.
```

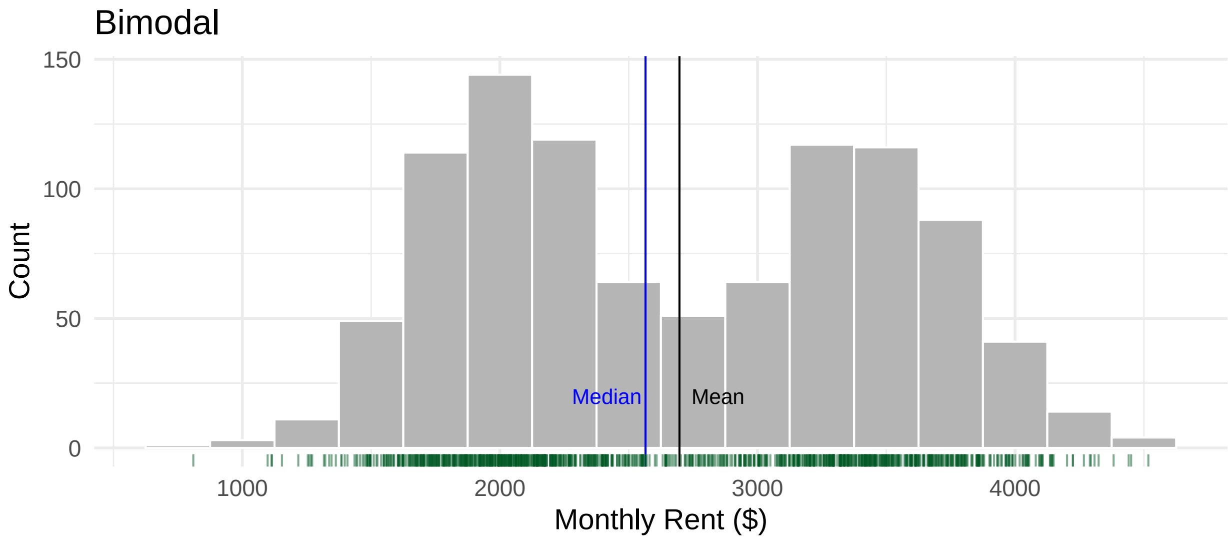


Example: Apartment rent in SF

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

```
## Warning in geom_text(aes(x = rent_summs$bimodal_mean + 150, y = 20), label = "Mean", : All aesthetics
## i Please consider using `annotate()` or provide this layer with data containing
##   a single row.
```

```
## Warning in geom_text(aes(x = rent_summs$bimodal_median - 150, y = 20), label = "Median", : All aesthe
## i Please consider using `annotate()` or provide this layer with data containing
##   a single row.
```



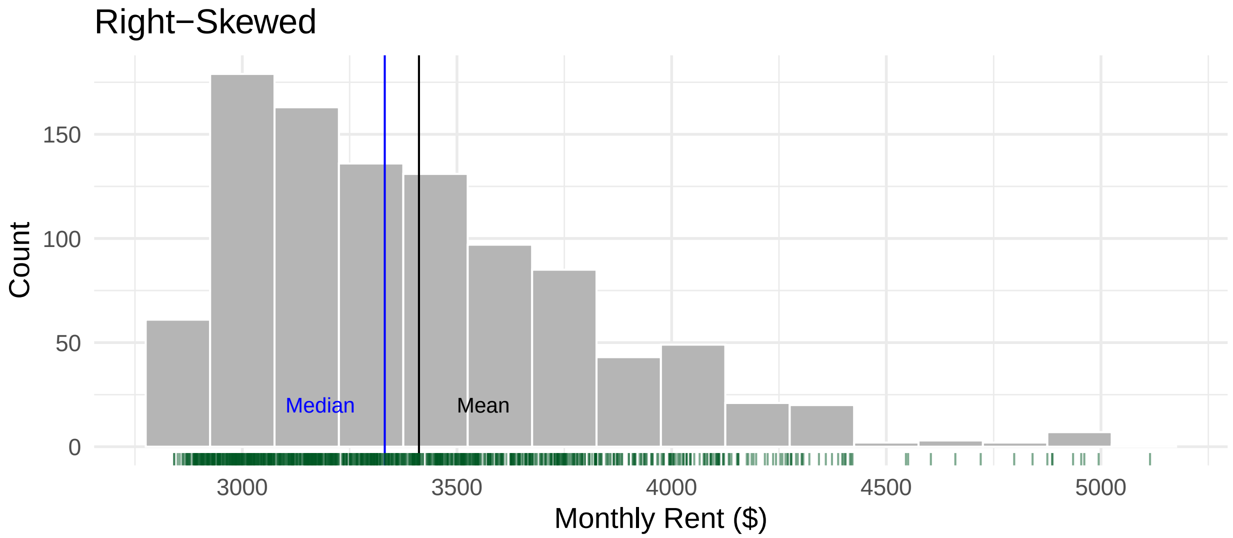
Describe the distribution. How does it differ from the first plot? Would you want to provide the mean or median for these data?

Example: Apartment rent in SF

Consider instead that the data were skewed right, maybe because there are many expensive apartments in the area. Here is the histogram of data for this example:

```
## Warning in geom_text(aes(x = rent_summs$right_skew_mean + 150, y = 20), : All aesthetics have length
## i Please consider using `annotate()` or provide this layer with data containing
##   a single row.
```

```
## Warning in geom_text(aes(x = rent_summs$right_skew_median - 150, y = 20), : All aesthetics have leng
## i Please consider using `annotate()` or provide this layer with data containing
##   a single row.
```



Why is the mean larger than the median when the data is skewed to the right?

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 (or variation)

To motivate this section, we import data on hospital cesarean delivery rates. These data were provided by the first author (Kozhimannil) of a manuscript published in the journal *Health Affairs*.

- Let's first open the spreadsheet in Excel to see what we are dealing with.
- Notice the hidden columns and unhide them. Look at the variable names.
- Question: How do we import data from Excel that is saved as .xls or .xlsx?

Measures of spread (or variation)

- Question: How do we import data from Excel that is saved as .xls or .xlsx?
- Answer: Use a function from the `readxl` library.

```
library(readxl) # this library helps with reading xlsx and xls files into R

CS_dat <- read_xlsx("./data/Ch02_Kozhimannil_Ex_Cesarean.xlsx", sheet = 1)
```

```
## New names:
## * `` -> `...5`
```

```
head(CS_dat)
```

```
## # A tibble: 6 x 7
##   Births HOSP_BEDSIZE cesarean_rate lowrisk_cesarean_rate ...5
##    <dbl>        <dbl>         <dbl>                 <dbl> <lgl>
## 1    767            1         0.344                0.107  NA
## 2    183            1         0.454                0.186  NA
## 3    668            1         0.430                0.195  NA
## 4    154            1         0.279                0.0844 NA
## 5    327            1         0.306                0.119  NA
## 6   2356            1         0.301                0.0662 NA
## # i 2 more variables: ‘Cesarean rate *100‘ <dbl>,
## #   ‘Low Risk Cearean rate*100‘ <dbl>
```

```
names(CS_dat)
```

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

- Notice that some variable names contain spaces. This is a general coding “no-no”.
- Question: Which `dplyr` function can we use to change the variable names?

Sidenote on variable names containing spaces

- 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.

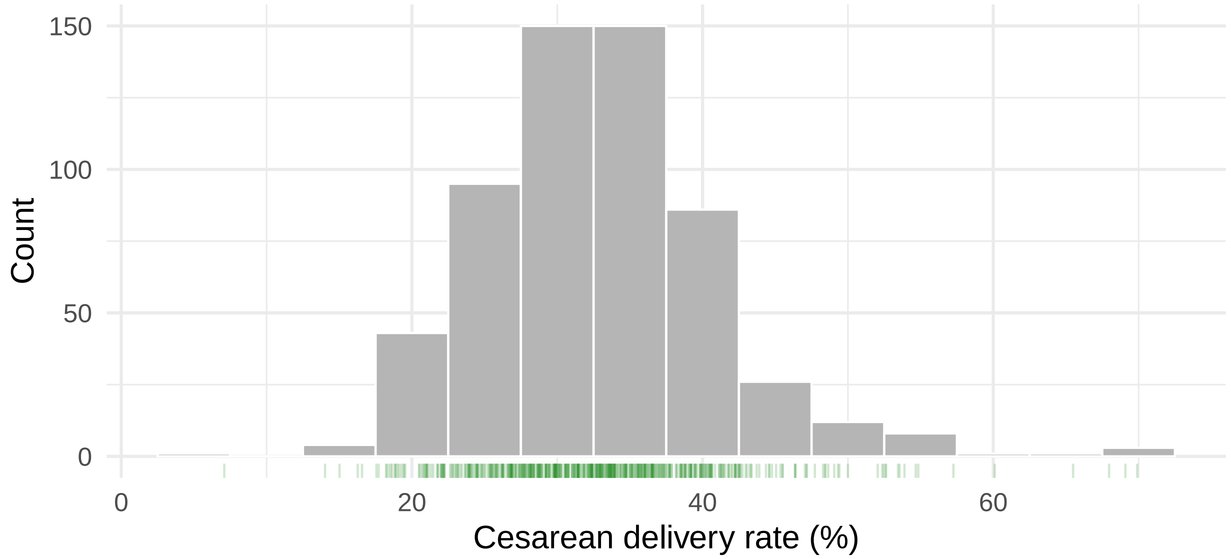
Tidy the data for analysis

For our example, we are only interested in each hospital’s cesarean delivery rate and the number of births at the hospital.

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

Histogram of cesarean delivery rates across US hospitals

```r
ggplot(CS_dat, aes(x = cs_rate)) +
  geom_histogram(col = "white", fill = "grey71", binwidth = 5) +
  labs( x = "Cesarean delivery rate (%)", y = "Count",
    caption = "Data from: Kozhimannil, Law, and Virnig. Health Affairs. 2013;32(3):527-35.") +
  geom_rug(alpha = 0.2, col = "forest green") + #alpha controls transparency
  theme_minimal(base_size = 15)
```



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.

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?

Calculating Q1 and Q3 by hand

Read this on your own time:

If there are an even number of observations, split the ordered list in half and find the **median** of each half of the data. The lower median is the Q1 estimate, and the upper median is the Q2 estimate.

If there are an odd number of observations, then split the data with the lower half being the ordered values below (but not including) the median, and the upper half being the irdered values above (but not including) the median. Take the median for each half of the list to calculate Q1 (lower half) and Q3 (upper half).

You should know how to find Q1, Q2, and Q3 **using a calculator** for small lists of numbers.

Quartiles using R

- Note that the function is called `quantile()` not `quartile()`. This is because `quantile()` is a general term, but `quartile()` corresponds specifically to the four groups of data with partitions at the 25th, 50th, and 75th quantiles.

- That is, you can compute a quantile at the 10th percentile but *not* a quartile at the 10th percentile!

```
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
```

R's `quantile` function

- `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 from the simple 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
- `min()` and `max()` are the relevant R functions

The five number summary

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

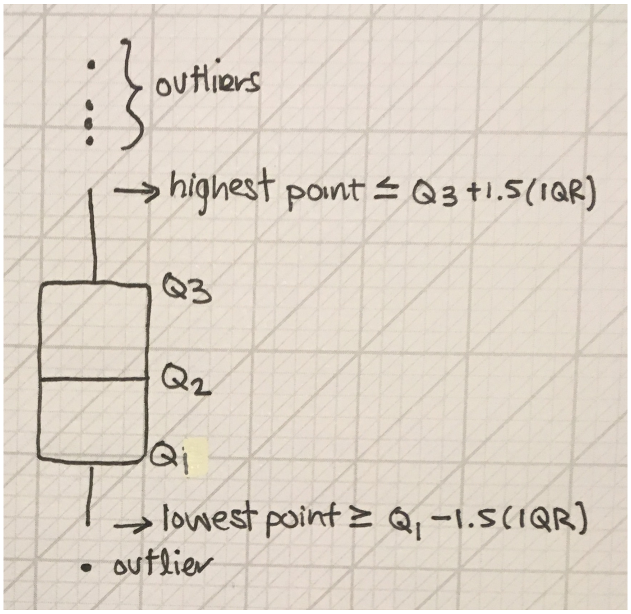
dplyr's summarize() to calculate the five number summary

```
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
```

Box (and whisker) plot

- A box plot is a visualization of the quartiles, plus outliers
- These outliers are found as shown in the figure:

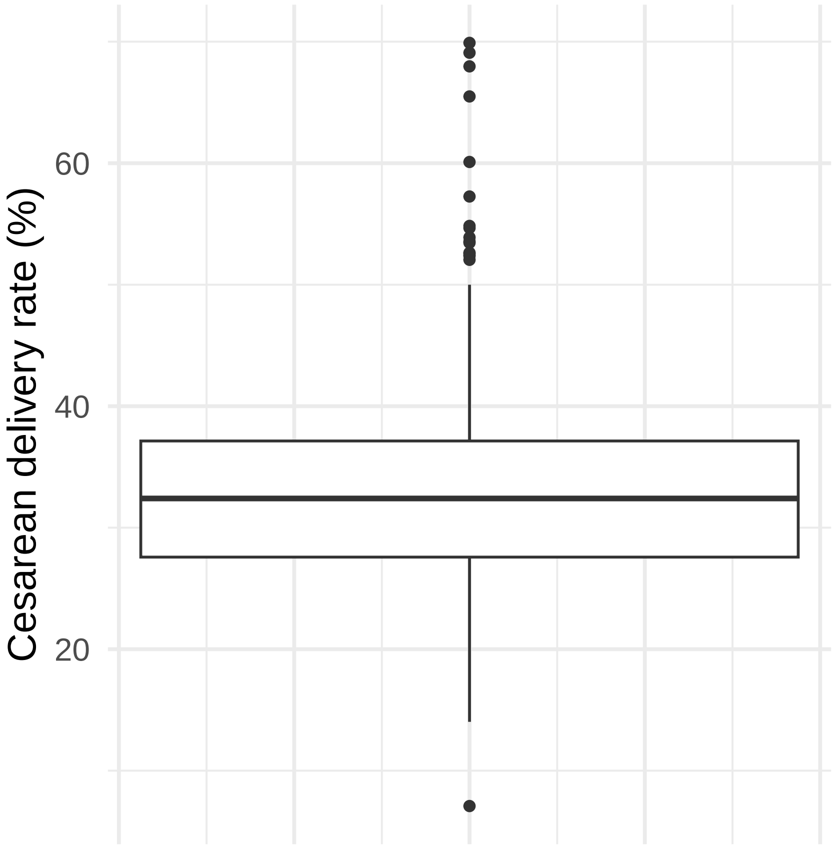


Box plots provide a nice visual summary of the center and spread

Here is the box plot for the cesarean data:

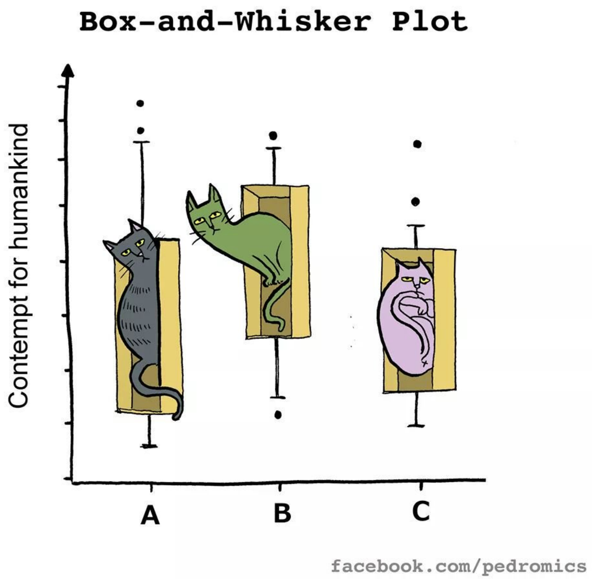
```
ggplot(CS_dat, aes(y = cs_rate)) +
  geom_boxplot() +
  labs(y = "Cesarean delivery rate (%)",
       title = "Box plot of the CS rates across \nUS hospitals",
       caption = "Data from: Kozhimannil et al. 2013.") +
  theme_minimal(base_size = 15) +
  scale_x_continuous(labels = NULL) # removes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

Box plot of the CS rates across US hospitals



Data from: Kozhimannil et al. 2013.

My favorite box plot



Boxplots

- The center line is the ______________.
- The top of the box is the ______________.
- The bottom of the box is the ______________.
- The top of the top whisker is equal to ____________________.
- The bottom of the bottom whisker is equal to ____________________.
- The data points above and below the whiskers are the _________________________.

Sample variance

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

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

- If the denominator was n rather than $n - 1$, then the variance would directly be the average of the squared distances between each observation and the sample's mean.
- Thus, variance is a measure of the average distance and the sample's mean.
- A large variance would indicate a larger spread in the data. A smaller variance would indicate a smaller spread in the data.
- This video provides intuition of why we divide by $n - 1$ rather than n.

Sample variance

We can rewrite sample variance this way:

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

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

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

Sample standard deviation

The sample standard deviation is the square root of the sample variance.

From the last slide, the sample variance is:

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

Thus, the sample standard deviation s is:

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

Why report standard deviation rather than variance?

- The units of variance is squared (e.g., cm^2 or grams^2)
- The units of standard deviation are the same as the observations. That is, if you measured distance between maternal residence and the closest hospital, the variance would be reported in miles squared, while the standard deviation is reported in miles.

`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
```

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`