Summarizing Numerical Data

Seeing the forest for the trees.

2/1/23

PDF

- Man feigns madness, contemplates life and death, and seeks revenge.
- Son avenges his father, and it only takes four hours.
- A tragedy written by the English playwright around 1600.
- 29,551 words on a page.

You may recognize each of these as summaries of the play, "Hamlet". None of these are wrong, per se, but they do focus on very different aspects of the work. Summarizing something as rich and complex as Hamlet invariably involves a large degree of omission; we're reducing a document of 29,551 words down to a single sentence, after all. But summarization also involves important choices around what to include.

The same considerations of omission and inclusion come into play when developing a numerical or graphical summary of a data set. Some guidance to bear in mind:

What should I include?

- Qualities relevant to the question you're answering or claim you're making
- Features that are aligned with the interest of your audience

What should I omit?

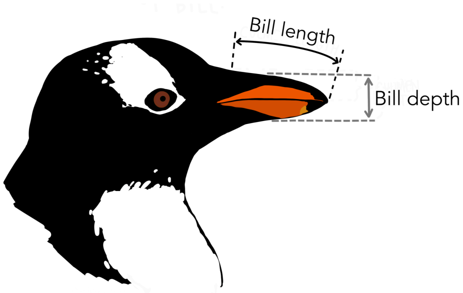
- Qualities that are irrelevant, distracting, or deceptive

- Replicated or assumed information

In these notes we'll keep this guidance in mind as we discuss how to summarize numerical data with graphics, in words, and with statistics.

> **Code along**
>
> As you read through these notes, you'll find figures and tables that were produced by running R code. If you'd like to peek into that code, click the gray arrow on the left side of the code. There's no need to learn all of these functions just yet, but it might be helpful to start reasoning through what you think the different functions are doing. When you learn these functions in the coming weeks, these notes can serve as a source of helpful examples.

Constructing Graphical Summaries



Let's turn to an example admittedly less complex than Hamlet: the Palmer penguins. One of the numerical variables Dr. Gorman recorded was the length of the bill in millimeters. The values of the first 16 penguins are:

```
library(tidyverse)
```

```
-- Attaching packages --------------------------------------- tidyverse 1.3.2 --
v ggplot2 3.4.1     v purrr   1.0.1
v tibble  3.2.1     v dplyr   1.1.0
v tidyr   1.3.0     v stringr 1.5.0
v readr   2.1.3     v forcats 0.5.1
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

```
library(stat20data)
bill_16 <- penguins %>%
  select(bill_length_mm) %>%
  slice(1:16)
bill_16
```

```
# A tibble: 16 x 1
   bill_length_mm
            <dbl>
 1           39.1
 2           39.5
 3           40.3
 4           36.7
 5           39.3
 6           38.9
 7           39.2
 8           41.1
 9           38.6
10           34.6
11           36.6
12           38.7
13           42.5
14           34.4
15           46  
16           37.8
```

We have many options for different plot types that we could use to summarize this data graphically. To understand the differences, it's helpful to lay out the criterion that we hold for a summary to be a success. Let's call those criteria the *desiderata*, a word meaning "that which is desired or needed".

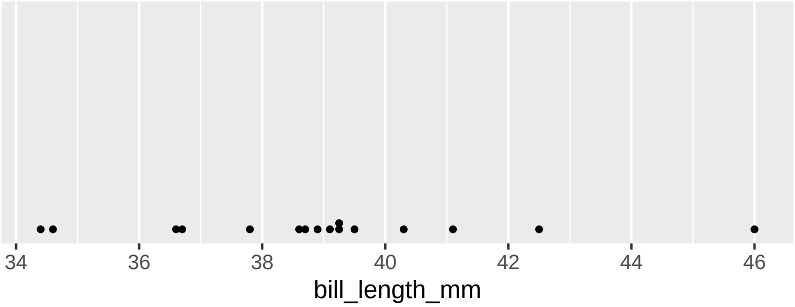
For our first graphic, let's set a high bar.

Desiderata

- All information must be preserved.

The most commonly used graphic that fulfills this criterion is the **dot plot**.

```
bill_16 %>%
  ggplot(aes(x = bill_length_mm)) +
  geom_dotplot(binwidth = .1) +
  scale_y_continuous(NULL, breaks = NULL)
```



The dot plot is, in effect, a one-dimensional scatter plot. Each observation shows up as a dot and its value corresponds to its location along the x-axis. Importantly, it fulfills our desiderata: given this graphic, one can recreate the original data perfectly. There was no information loss.

As the number of observations grow, however, this sort of graphical summary becomes unwieldy. Instead of focusing on the value of each observation, it becomes more practical to focus on the general shape of the distribution. Let's consider a broader goal for our graphic.

Desiderata

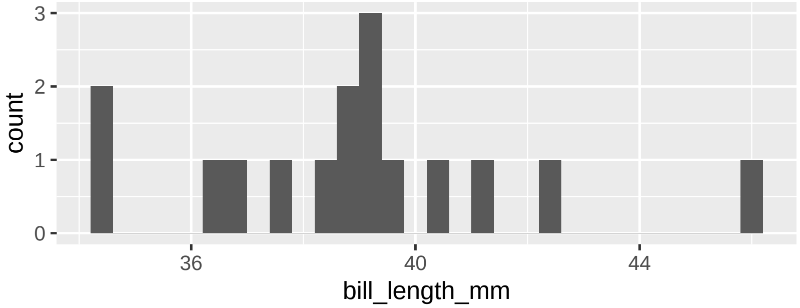
- Balance depiction of the general characteristics of the distribution with a fidelity to the individual observations.

There are several types of graphics that meet this criterion: the histogram, the density plot, and the violin plot.

Histograms

```r
bill_16 %>%
  ggplot(aes(x = bill_length_mm)) +
  geom_histogram()
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



At first glance, a histogram looks like deceptively like bar chart. There are bars arranged along the x-axis according to their values with heights that correspond to the count of each value found in the data set. So how is this not a bar chart?

A histogram involves *aggregation*. The first step in creating a histogram is to divide the range of x into bins of equal size. The second step is to count up the number of observations that occur in each bin. In this way, some observations will have their own bar (every bar with a count of one) but others will be aggregated into the same bar: the tallest bar, with a count of 3, corresponds to all observations from 39.09 to 39.30: 39.1, 39.2, and 39.3.

The degree of aggregation performed by the histogram is determined by the *binwidth*. Most software will automatically select the binwidth[1], but it can be useful to tinker with different values to see the distribution at different levels of aggregation.

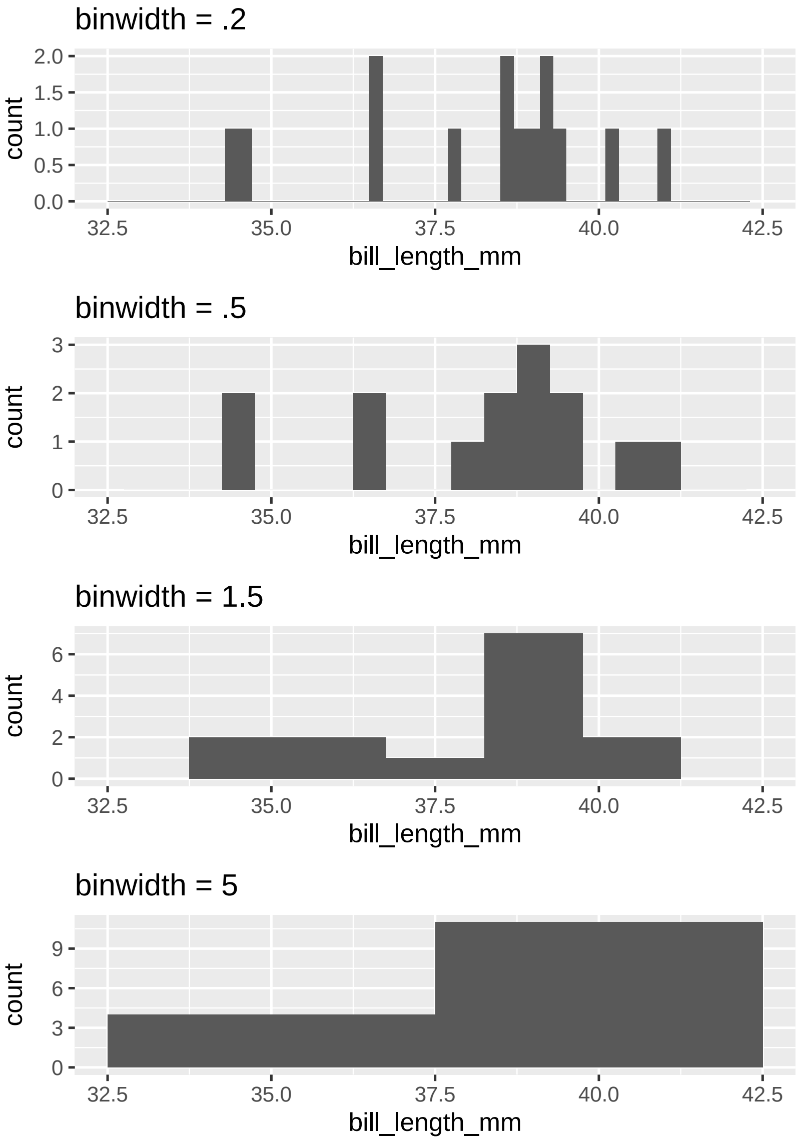
[1] The `ggplot2` package in R defaults to 30 bins across the range of the data. That's a very rough rule-of-thumb, so tinkering is always a good idea.

Here are four histograms of the same data that use four different binwidths.

```r
library(patchwork)
p1 <- bill_16 %>%
  ggplot(aes(x = bill_length_mm)) +
  geom_histogram(binwidth = .2) +
  labs(title = "binwidth = .2") +
  lims(x = c(32.5, 42.5))
p2 <- bill_16 %>%
  ggplot(aes(x = bill_length_mm)) +
  geom_histogram(binwidth = .5) +
  labs(title = "binwidth = .5") +
  lims(x = c(32.5, 42.5))
p3 <- bill_16 %>%
  ggplot(aes(x = bill_length_mm)) +
  geom_histogram(binwidth = 1.5) +
  labs(title = "binwidth = 1.5") +
  lims(x = c(32.5, 42.5))
p4 <- bill_16 %>%
  ggplot(aes(x = bill_length_mm)) +
  geom_histogram(binwidth = 5) +
  labs(title = "binwidth = 5") +
  lims(x = c(32.5, 42.5))
p1 / p2 / p3 / p4
```

```
Warning: Removed 1 rows containing non-finite values (`stat_bin()`).

Warning: Removed 2 rows containing missing values (`geom_bar()`).

Warning: Removed 1 rows containing non-finite values (`stat_bin()`).

Warning: Removed 2 rows containing missing values (`geom_bar()`).

Warning: Removed 1 rows containing non-finite values (`stat_bin()`).

Warning: Removed 2 rows containing missing values (`geom_bar()`).
```

```
Warning: Removed 1 rows containing non-finite values (`stat_bin()`).
```



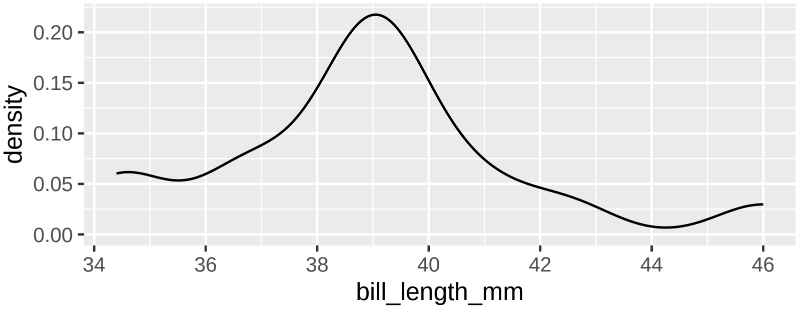
If you are interested in only the coarsest structure in the distribution, best to use the larger binwidths. If you want to see more detailed structure, a smaller binwidth is better.

There is a saying that warns about times when you, “can’t see the forest for the trees”, being overwhelmed by small details (the trees) and unable to see the bigger picture (the forest). The histogram, as a graphical tool for summarizing the distribution of a numerical variable, offers a way out. Through your choice of binwidth, you can determine how much to focus on the forest (large bindwidth) or the trees (small binwidth).

Density plots

Imagine that you build a histogram and place a cooked piece of spaghetti over the top of it. The curve created by the pasta is a form of a density plot.

```
bill_16 %>%
  ggplot(aes(x = bill_length_mm)) +
  geom_density()
```



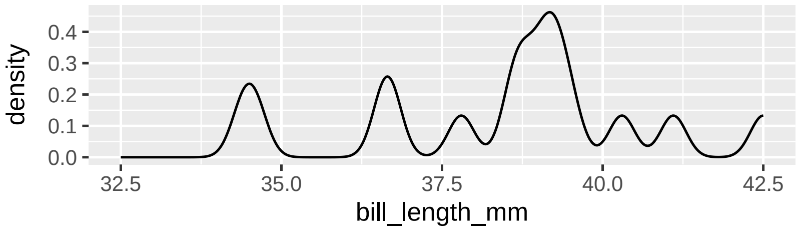
Besides the shift from bars to a smooth line, the density plot also changes the y-axis to feature a quantity called "density". We will return to define this term later in the course, but it's sufficient to know that the values on the y-axis of a density plot are rarely useful. The important information is relative: an area of the curve with twice the density as another area has roughly twice the number of observations.

The density plot, like the histogram, offers the ability to balance fidelity to the individual observations against a more general shape of the distribution. You can tip the balance to feature what you find most interesting but adjusting the *bandwidth* of the density plot.

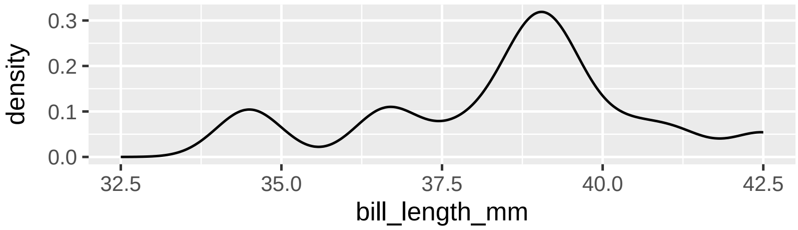
```
p1 <- bill_16 %>%
  ggplot(aes(x = bill_length_mm)) +
  geom_density(bw = .2) +
  labs(title = "bandwidth = .2") +
  lims(x = c(32.5, 42.5))
p2 <- bill_16 %>%
  ggplot(aes(x = bill_length_mm)) +
  geom_density(bw = .5) +
  labs(title = "bandwidth = .5") +
  lims(x = c(32.5, 42.5))
p3 <- bill_16 %>%
  ggplot(aes(x = bill_length_mm)) +
  geom_density(bw = 1.5) +
  labs(title = "bandwidth = 1.5") +
  lims(x = c(32.5, 42.5))
p4 <- bill_16 %>%
  ggplot(aes(x = bill_length_mm)) +
  geom_density(bw = 5) +
  labs(title = "bandwidth = 5") +
  lims(x = c(32.5, 42.5))
p1 / p2 / p3 / p4
```

```
Warning: Removed 1 rows containing non-finite values (`stat_density()`).
Removed 1 rows containing non-finite values (`stat_density()`).
Removed 1 rows containing non-finite values (`stat_density()`).
Removed 1 rows containing non-finite values (`stat_density()`).
```

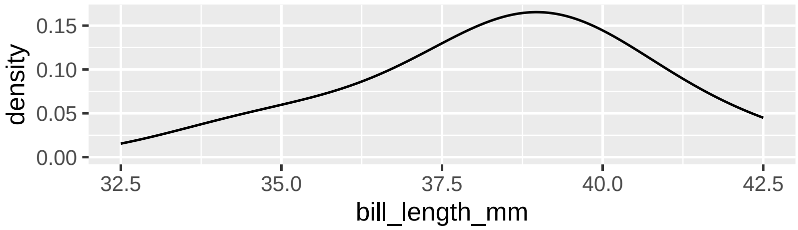
bandwidth = .2



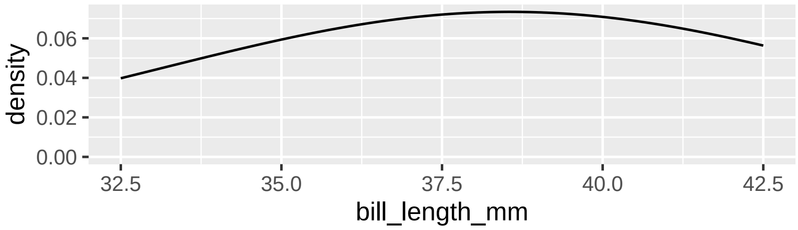
bandwidth = .5



bandwidth = 1.5



bandwidth = 5



A density curve tends to convey the overall shape of a distribution more quickly than does a histogram, but be sure to experiment with different bandwidths. Strange but important features of a distribution can be hidden behind a density curve that is too smooth.

Violin plots

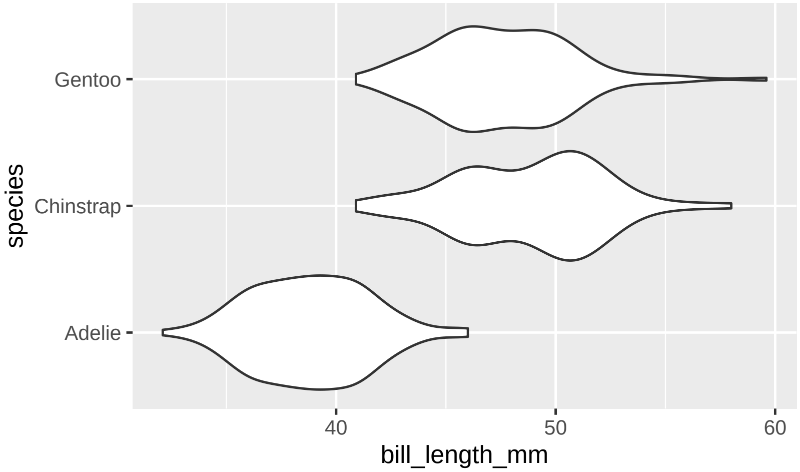
Often we're interested not in the distribution of a single variable, but in the way the distribution of that variable changes from one group of observational units to another. Let's add this item to our list of criteria for a statistical graphic.

Desiderata

- Balance depiction of the general characteristics of the distribution with a fidelity to the individual observations.
- Allow for easy comparisons between groups.

There are several different ways to compare the distribution of a variable across two or more groups, but one of the most useful is the violin plot. Here is a violin plot of the distribution of bill length across the three species of penguins.

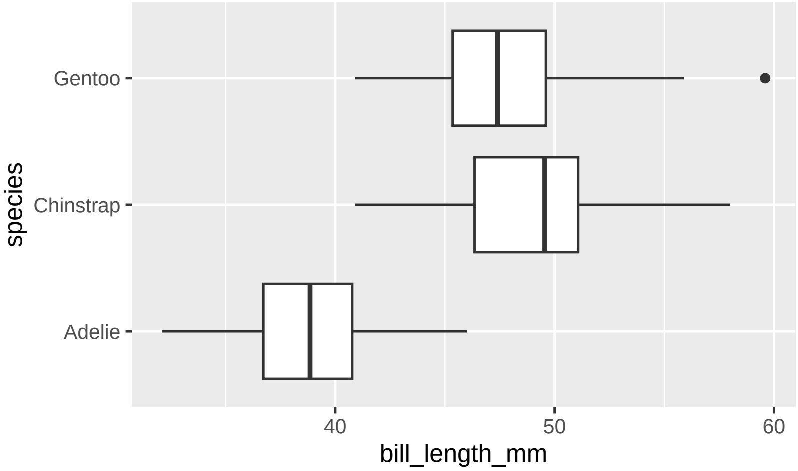
```
penguins %>%
  ggplot(aes(y = species,
             x = bill_length_mm)) +
  geom_violin()
```



The distribution of bill length in each species is represented by a shape that often looks like a violin but is in fact a simple density curve reflected about its x-axis. This means that you can tinker with a violin plot the same as a density plot, but changing the bandwidth.

If this plot type looks familiar, you may have seen its cousin, the box plot.

```
penguins %>%
  ggplot(aes(y = species,
             x = bill_length_mm)) +
  geom_boxplot()
```



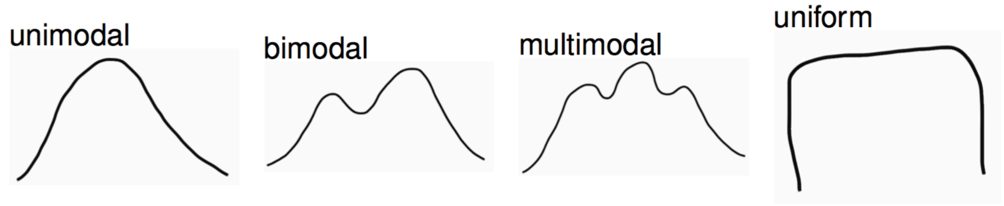
The box plot conveys a similar story to the violin plot: Adelies have shorter bills than Chinstraps and Gentoos. Box plots have the advantage of requiring very little computation to construct[2], but in a world of powerful computers, that is no longer remarkable. What they lack is a "smootness-knob" that you can turn to perform more or less smoothing. For this reason, violins tend to be a more flexible alternative to box plots.

[2] To read more about one common way to construct a box plot, see the ggplot2 documentation.

Describing Distributions

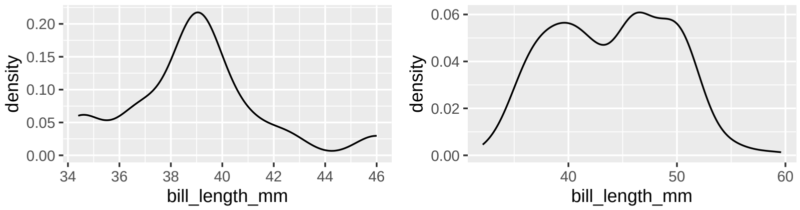
The desideratum that we used to construct the histogram and the violin plot include the ability to “depict general characteristics of the distribution”. The most important characteristics of a distribution are its *shape*, *center*, and *spread*.

When describing the shape of a distribution in words, pay attention to its *modality* and *skew*. The modality of a distribution captures the number of distinct peaks (or modes) that are present.



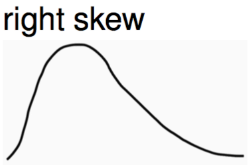
A good example of a distribution that would be described as unimodal is the original density plot of bill lengths of 16 Adelie penguins (below left). There is one distinct peak around 39. Although there is another peak around 34, it is not prominent enough to be considered a distinct mode. The distribution of the bill lengths of all 344 penguins (below right), however, can be described as bimodal.

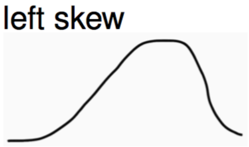
```
p1 <- bill_16 %>%
  ggplot(aes(x = bill_length_mm)) +
  geom_density()
p2 <- penguins %>%
    ggplot(aes(x = bill_length_mm)) +
    geom_density()
p1 + p2
```

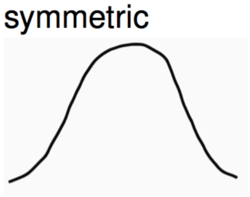


Multiple modes are often a hint that there is something more going on. In the plot to the right above, Chinstraps and Gentoo penguins, which are larger, are clumped under the right mode while the smaller Adelie penguins are dominant in the left mode.

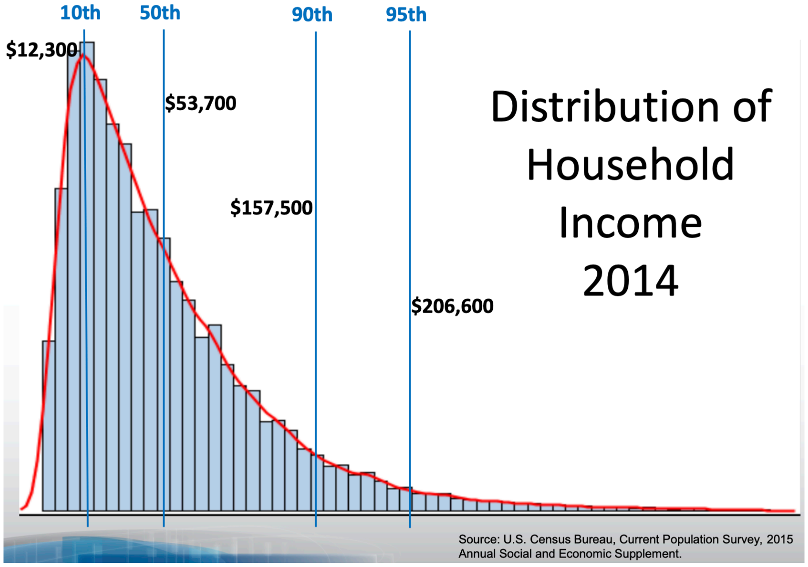
The other important characteristic of the shape of a distribution is its skew.







The skew of a distribution describes the behavior of its tails: whether the right trail stretches out (right skew), the left tail stretches out (left skew), or if both tails are of similar length (symmetric). An example of a persistently right skewed distribution is household income in the United States:



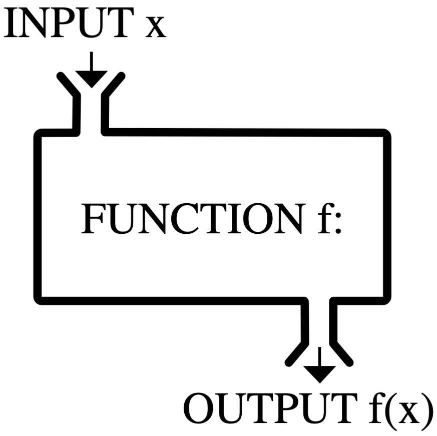
Source: U.S. Census Bureau, Current Population Survey, 2015 Annual Social and Economic Supplement.

In the US, the richest households have much much higher incomes than most, while the poorest households have incomes that are only a bit lower than most.

When translating a graphical summary of a distribution into words, some degree of judgement must be used. When is a second peak a mode and when is it just a bump in the distribution? When is one of the tails of a distribution long enough to tip the description from being symmetric to being right skewed? You'll hone your judgement in part through repeated practice: looking at lots of distributions and readings lots of descriptions. You can also let the questions of inclusion and omission be your guide. Is the feature a characteristic relevant to the question you're answering and the phenomenon you're studying? Or is it just a distraction from the bigger picture?

Modality and skew capture the shape of the distribution, but how do we describe its center and spread? "Eyeballing it" each of these by looking at a graphic is an option. A more precise option, though, is to calculate a statistic.

Constructing Numerical Summaries



Statistics is both an area of academic study and the object of that study. Any numerical summary of a data set - a mean or median, a count or proportion - is a *statistic.* A statistic is, fundamentally, a mathematical function where the data is the input and the output is the observed statistic.

Statisticians don't just study statistics, though, they *construct* them. A statistician gets to decide the form of f and, as with graphics, they construct it to fulfill particular needs: the desiderata.

To examine the properties of common statistics, let's move to an even simpler data set: a vector called `x` that holds 11 integers.

$$8, 11, 7, 7, 8, 11, 9, 6, 10, 7, 9$$

Measures of Center

The mean, the median, and the mode are the three standard statistics used to measure the center of a distribution. Despite their ubiquity, these three are not carved somewhere on a stone tablet. They're common because they're useful and they're useful because they were constructed very thoughtfully.

Let's start by layout some possible criteria for a measure of center.

Desiderata

- Synthesizes the magnitudes of all n observations.
- As close as possible to all of the observations.

The (arithmetic) **mean** fulfills all three of these needs.

The function in R: `mean()`

$$\frac{8+11+7+7+8+11+9+6+10+7+9}{11} = \frac{93}{11} = 8.45$$

The mean synthesizes the magnitudes by taking their sum, then keeps that sum from getting larger than any of the observations by dividing by n. In order to express this function more generally, we use the following notation

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

where x_1 is the first observation, x_2 is the second observation, and so on till the n^{th} observation, x_n; and $\bar{x}$ is said "x bar".

The mean is the most commonly used measure of center, but has one notable drawback. What if one of our observations is an *outlier*, that is, has a value far more extreme than the rest of the data? Let's change the 6 to -200 and see what happens.

$$\frac{8+11+7+7+8+11+9-200+10+7+9}{11} = \frac{-113}{11} = -10.27$$

The mean has plummeted to -10.27, dragged down by this very low outlier. While it is doing it's best to stay "as close as possible to all of the observations", it isn't doing a very good job of representing 10 of the 11 values.

With this in mind, let's alter the first criterion to inspire a different statistic.

Desiderata

- Synthesize the *order* of all n observations.
- As close as possible to all of the observations.

If we put the numbers in order from smallest to largest, then the number that is as close as possible to all observations will be the middle number, the **median**.

$$6 \quad 7 \quad 7 \quad 7 \quad 8 \quad 8 \quad 9 \quad 9 \quad 10 \quad 11 \quad 11$$

The function in R: `median()`

As measured by the median, the center of this distribution is 8 (recall the mean measured 8.45). If there were an even number of observations, the convention is to take the mean of the middle two values.

The median has the desirable property of being resistant (or "robust") to the presence of outliers. See how it responds to the inclusion of -200.

$$-200 \quad 7 \quad 7 \quad 7 \quad 8 \quad 8 \quad 9 \quad 9 \quad 10 \quad 11 \quad 11$$

With this outlier, the median remains at 8 while the mean had dropped to -10.27. This property makes the median the favored statistic for capturing the center of a skewed distribution.

What if we took a stricter notion of "closeness"?

Desiderata

- Is identical to as many observations as possible.

That leads us to the measure of the **mode**, or the most common observation. For our example data set, the mode is 7.

$$6 \quad 7 \quad 7 \quad 7 \quad 8 \quad 8 \quad 9 \quad 9 \quad 10 \quad 11 \quad 11$$

While using the mode for this data set is sensible, it is common in numerical data for each value to be unique[3]. Is that case, there are no repeated values, and no identifiable mode. For that reason, it is unusual to calculate the mode to describe the center of a numerical variable.

For categorical data, however, the mode is very useful. The mode of the species variable among the Palmer penguins is “Adelie”. Trying to compute the mean or the median species will leave you empty-handed. This is one of the lingering lessons of the Taxonomy of Data: the type of variable guides how it is analyzed.

Measures of Spread

There are many different ways to capture spread or dispersion of data. Two commonly-used statistics are the (sample) **variance**, s^2, and the (sample) **standard deviation**, s.

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

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

Although the formulations might look dense, there are good reasons for each decision that was made in constructing these statistics. We will discuss those decisions, as well as several other ways one could measure spread, next class.

Summary

A summary of a summaries…this better be brief! Summaries of numerical data - graphical and numerical - often involve choices of what information to include and what information to omit. These choices involve a degree of judgement and knowledge of

[3] That is, unless you aggregate! The aggregation performed by a histogram or a density plot is what allows us to describe numerical data as unimodel or bimodal.

the criteria that were used to construct the commonly used statistics and graphics.