Packet 2

Todd CadwalladerOlsker

2022-01-27

A .pdf copy of this packet is available at <Packet2.pdf>.

Descriptive Statistics

In order to describe a data set, we need to *summarize* it. The phrase "Exploratory Data Analysis" is used in *Introduction to Modern Statistics*. We can summarize data by visualizing it, describing it numerically, or (even better) doing some of each.

I'll try to keep these packets brief, they won't contain everything there is to say about the material. You should reference the textbooks and keep your own notes as well.

Medians, Quartiles, and the Five-Number Summary

When we look at **numeric** variables, we can look at *median*-based statistics or *mean*-based statistics. The *median* is the value of the middle data point (or when there are an even number of data points, the value halfway between the two middle data points). In R, we can find the median of a data set with:

```
median(county$poverty,na.rm = TRUE)
```

```
## [1] 15.2
```

In this example, we have the median of the *population*, since we have information about every county. (well, except for the NAs, but it's not like we didn't ask for that information.) This is technically a *parameter*, not a *statistic*. If we only take a small sample of the population, then we are calculating the sample median statistic.

Here, the `na.rm = TRUE` tells R to ignore values that have missing data. (Try the command without the `na.rm = TRUE` and see what happens!) You can also just use `median(county$pop2017,TRUE)` for the same result.

Now, a *measure of central tendency* like the median is not worth very much without a *measure of spread* to go along with it. For the median, one way to get a handle on the spread is to also report the maximum, minimum, and 1st and 3rd quartile values. Together, these are called a *five-number summary*.

Recall that the 1st quartile is the "median" of the minimum value and the median, while the 3rd quartile is the "median" of the maximum value and the median.

The measures of spread, more technically, are the *range*, which is the difference of the maximum and minimum, and the *interquartile range*, or IQR, which is the difference of the 3rd and 1st quartiles. The five number summary contains all this information and more.

```
fivenum(county$poverty)
```

```
## [1]  2.4 11.3 15.2 19.4 52.0
```

```
  # The fivenum() function defaults to na.rm = TRUE,
  # so we don't need to add it in.
```

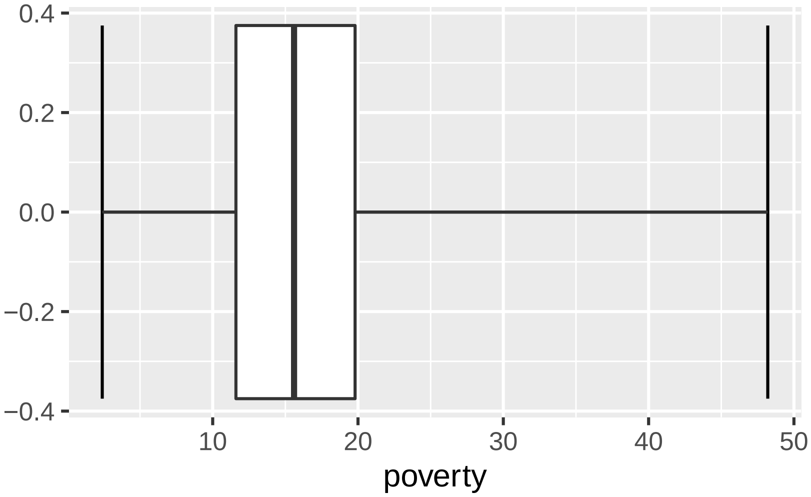
```
summary(county$poverty)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##    2.40   11.30   15.20   15.97   19.40   52.00       2
```

```
  # This is a little friendlier to the eyes,
  # and also includes the mean.
```

We can visualize the five number summary as a plot:

```
county %>% na.omit(poverty) %>% ggplot(aes(x = poverty)) +
  stat_boxplot(coef = 10, geom = "errorbar") +
  geom_boxplot(coef = 10)
```



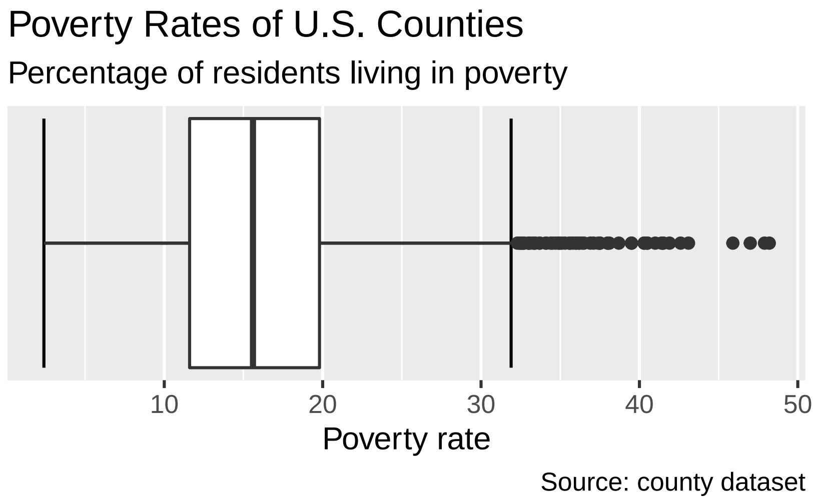
Traditionally, the boxplot includes vertical lines for each of the five numbers in the summary.

A couple of things are going on here: First, notice that we needed to add “error bars” to visualize the whiskers, gglot does not include them as standard.

Second, notice that we set `coef = 10`. What does that mean?

Third, notice that we have values on a vertical axis. These are meaningless, we can suppress those with the following code:

```
county %>% na.omit(poverty) %>% ggplot(aes(x = poverty)) +
  stat_boxplot(geom = "errorbar") +
  geom_boxplot() +
  labs(title = "Poverty Rates of U.S. Counties",
       y = NULL,
       x = "Poverty rate",
       subtitle = "Percentage of residents living in poverty",
       caption = "Source: county dataset") +
  scale_y_continuous(breaks = NULL, labels = NULL)
```


Poverty Rates of U.S. Counties

Percentage of residents living in poverty

Source: county dataset

Also notice I left off `coef = 10` this time. Most of the highest-poverty counties are considered "outliers", as they are more than 1.5 times the IQR away from the median. This may or may not be a good way to think about outliers, depending on the data (in this case, probably not). The `coef` variable overrides this by changing the number of multiples of the IQR we need to be from the median before we call something an outlier. Use `help(geom_boxplot)` for more.

Mean, Variance, and Standard Deviation

In many cases, the *mean* is a more useful measurement than the median. In many other cases the median is more useful – it depends on the situation. The mean is the sum of the values of our data points, divided by the number of data points – exactly how you learned to calculate the average since forever. The calculation is the same whether we want to calculate a population mean or sample mean.

The mean is more useful in the sense that we can do a lot more with it, but it is not as *robust* as the median. It's much more sensitive to outliers and skewed data.

The natural measure of spread to pair with the mean is the *variance*, or the *standard deviation*. The definition of variance starts with a natural idea: let's measure the distance between the value of each data point