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.

Numeric Variables

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
```

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.

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

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

```
  # The pop2010 variable doesn't have
  # any missing data, so we can skip the
  # na.rm = TRUE command. test
```

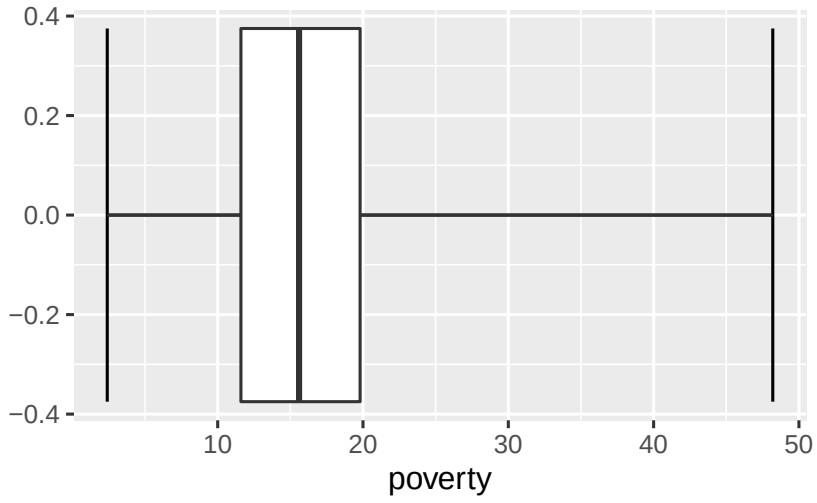
```
summary(county$poverty)
```

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

```
  # 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)
```



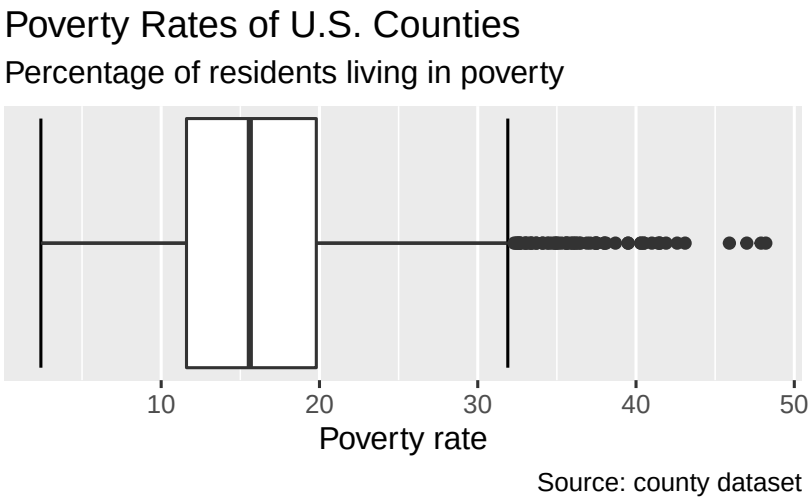
Traditionally, the boxplot includes vertical lines for each of the five numbers in the summary.

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)
```



Also notice I left off `coef = 10` this time. Most of the highest-poverty counties are considered outliers, as they are more that 1.5 times the IQR away from the median. The `coef` variable changes 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.