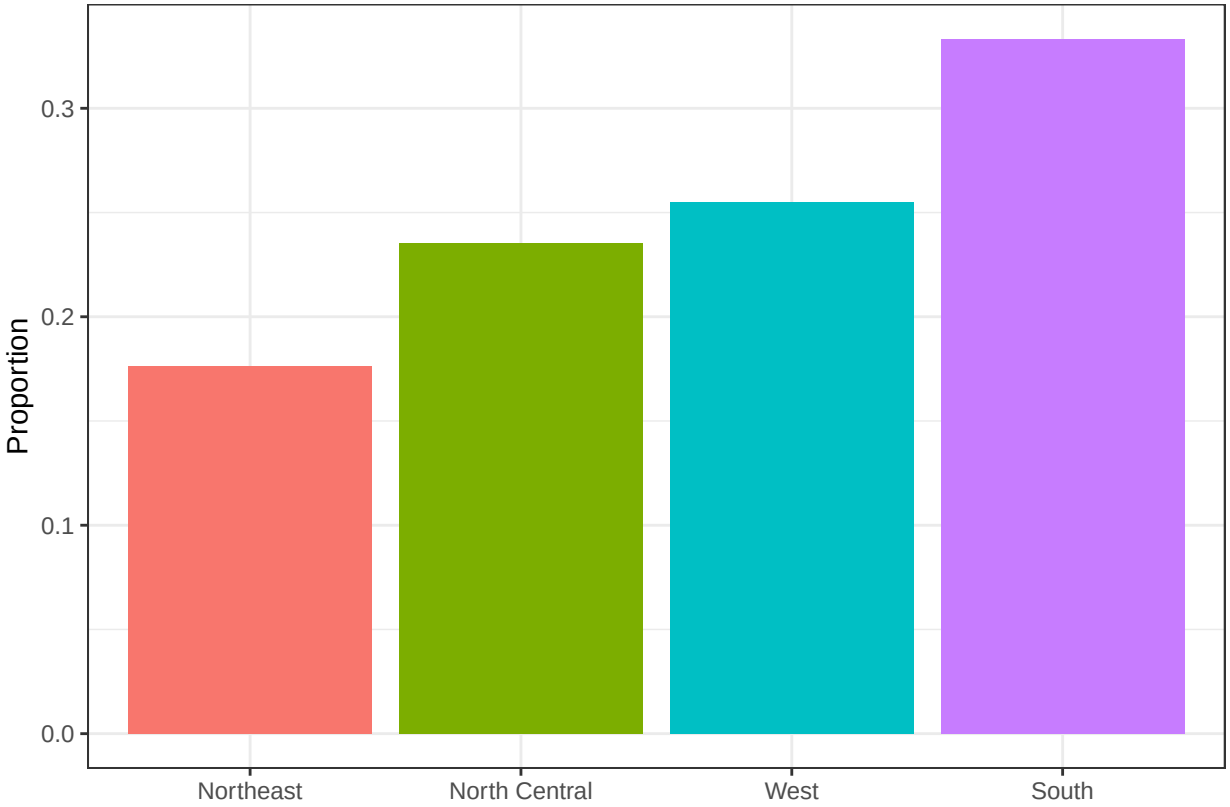
Additional `ggplot2` Exercises

RICC International Young Investigator Training

11/15/2023

1. In the `murders` dataset, the region is a categorical variable and the following is its distribution:

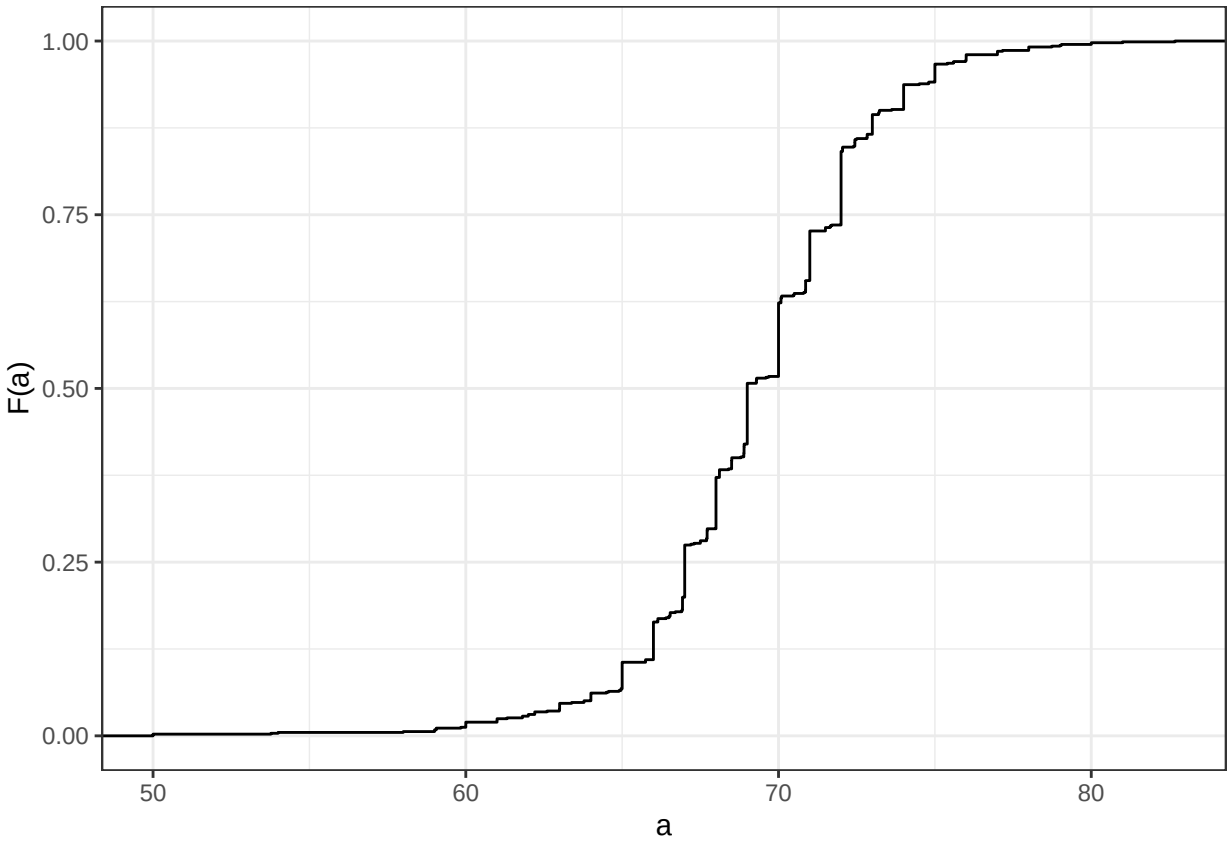


To the closest 5%, what proportion of the states are in the North Central region?

2. Which of the following is true:

   a. The graph above is a histogram.
   b. The graph above shows only four numbers with a bar plot.
   c. Categories are not numbers, so it does not make sense to graph the distribution.
   d. The colors, not the height of the bars, describe the distribution.

3. The plot below shows the eCDF for male heights:



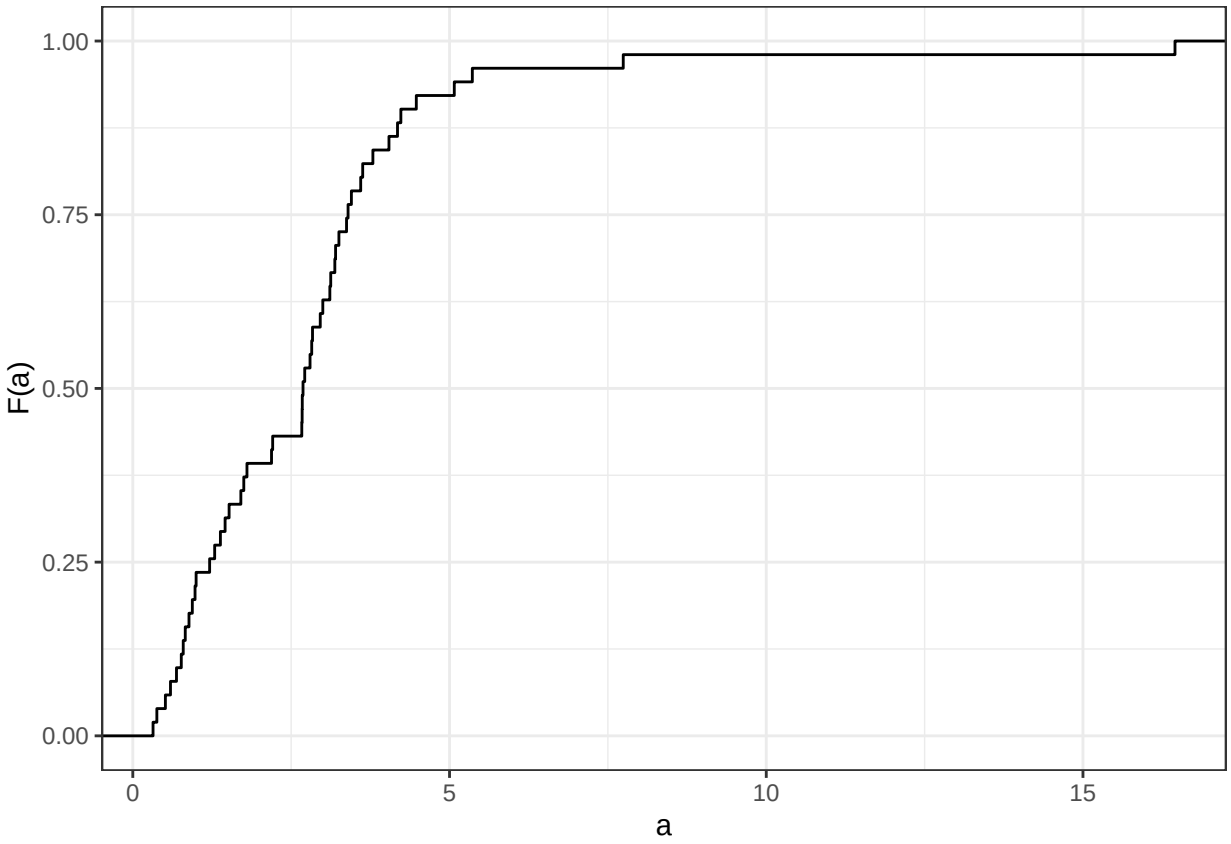
Based on the plot, what percentage of males are shorter than 75 inches?

a. 100%
b. 95%
c. 80%
d. 72 inches

4. To the closest inch, what height `m` has the property that 1/2 of the male students are taller than `m` and 1/2 are shorter?

a. 61 inches
b. 64 inches
c. 69 inches
d. 74 inches

5. Here is an eCDF of the murder rates across states:



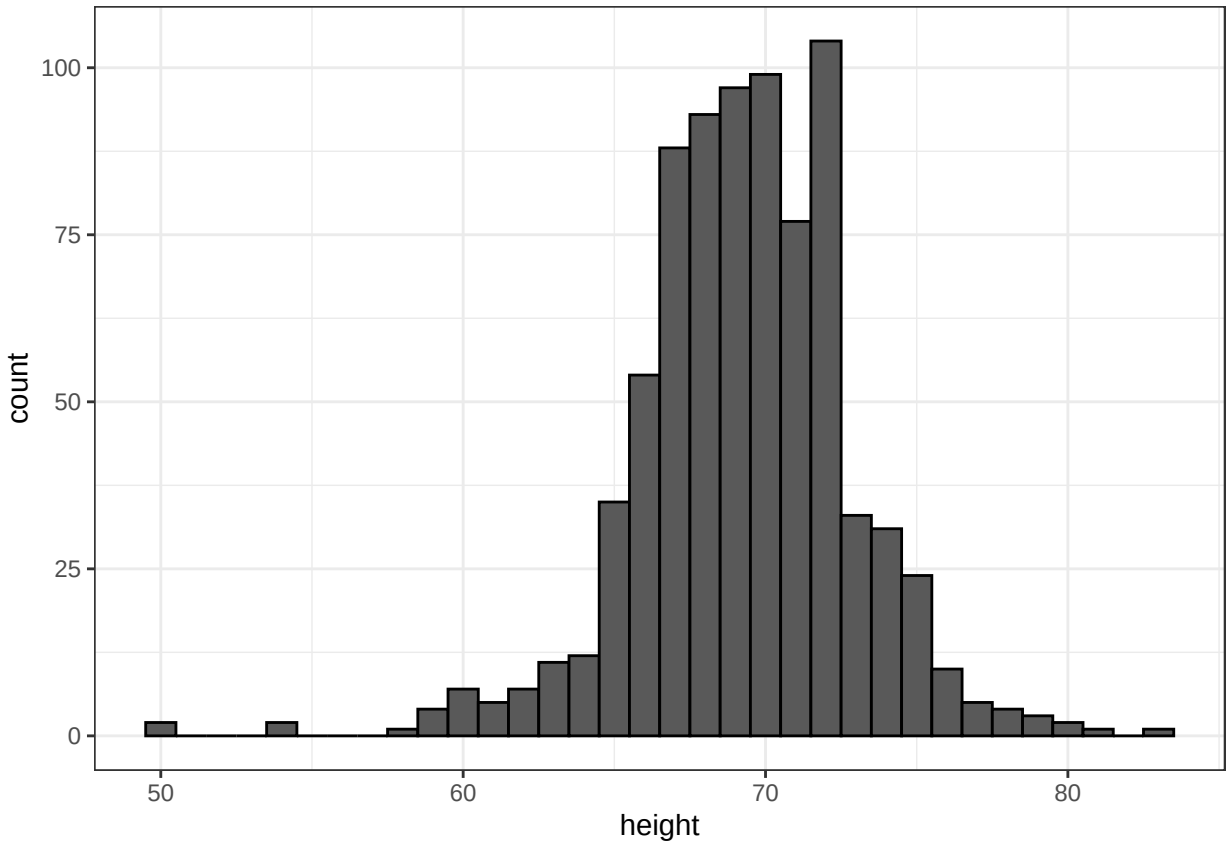
Knowing that there are 51 states (counting DC) and based on this plot, how many states have murder rates larger than 10 per 100,000 people?

a. 1
b. 5
c. 10
d. 50

6. Based on the eCDF above, which of the following statements are true:

a. About half the states have murder rates above 7 per 100,000 and the other half below.
b. Most states have murder rates below 2 per 100,000.
c. All the states have murder rates above 2 per 100,000.
d. With the exception of 4 states, the murder rates are below 5 per 100,000.

7. Below is a histogram of male heights in our `heights` dataset:



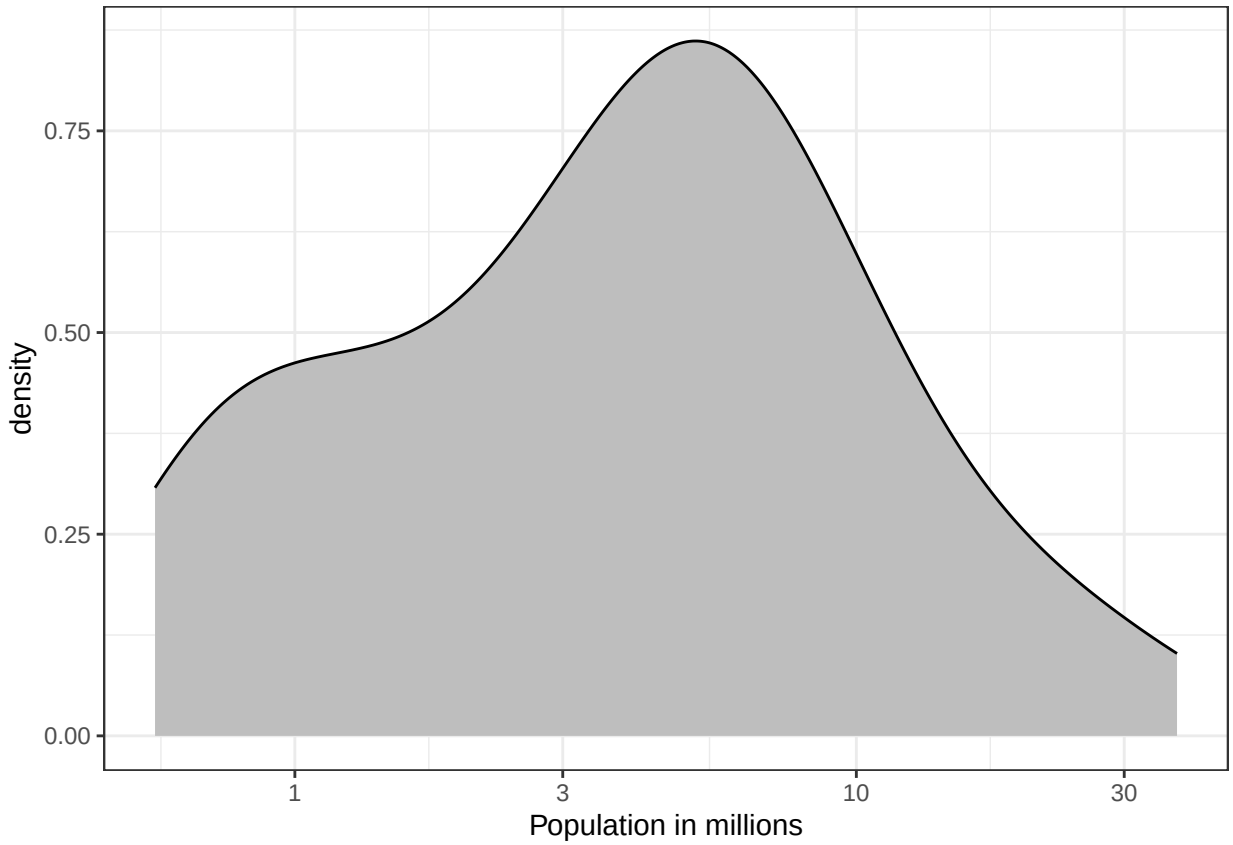
Based on this plot, how many males are between 63.5 and 65.5?

a. 10
b. 24
c. 34
d. 100

8. About what **percentage** are shorter than 60 inches?

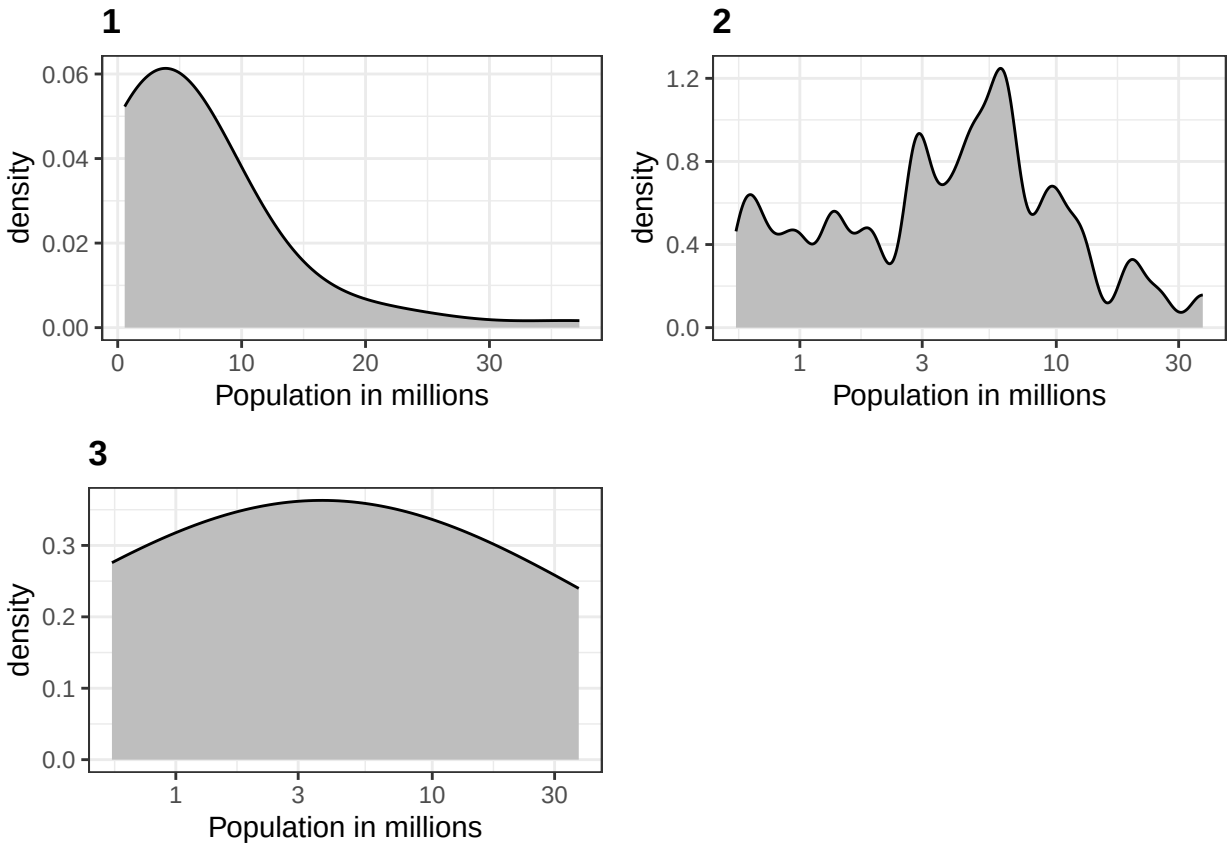
a. 1%
b. 10%
c. 25%
d. 50%

9. Based on the density plot below, about what proportion of US states have populations larger than 10 million?



a. 0.02
b. 0.15
c. 0.50
d. 0.55

10. Below are three density plots. Is it possible that they are from the same dataset?



Which of the following statements is true:

a. It is impossible that they are from the same dataset.
b. They are from the same dataset, but the plots are different due to code errors.
c. They are the same dataset, but the first and second plot undersmooth and the third oversmooths.
d. They are the same dataset, but the first is not in the log scale, the second undersmooths, and the third oversmooths.

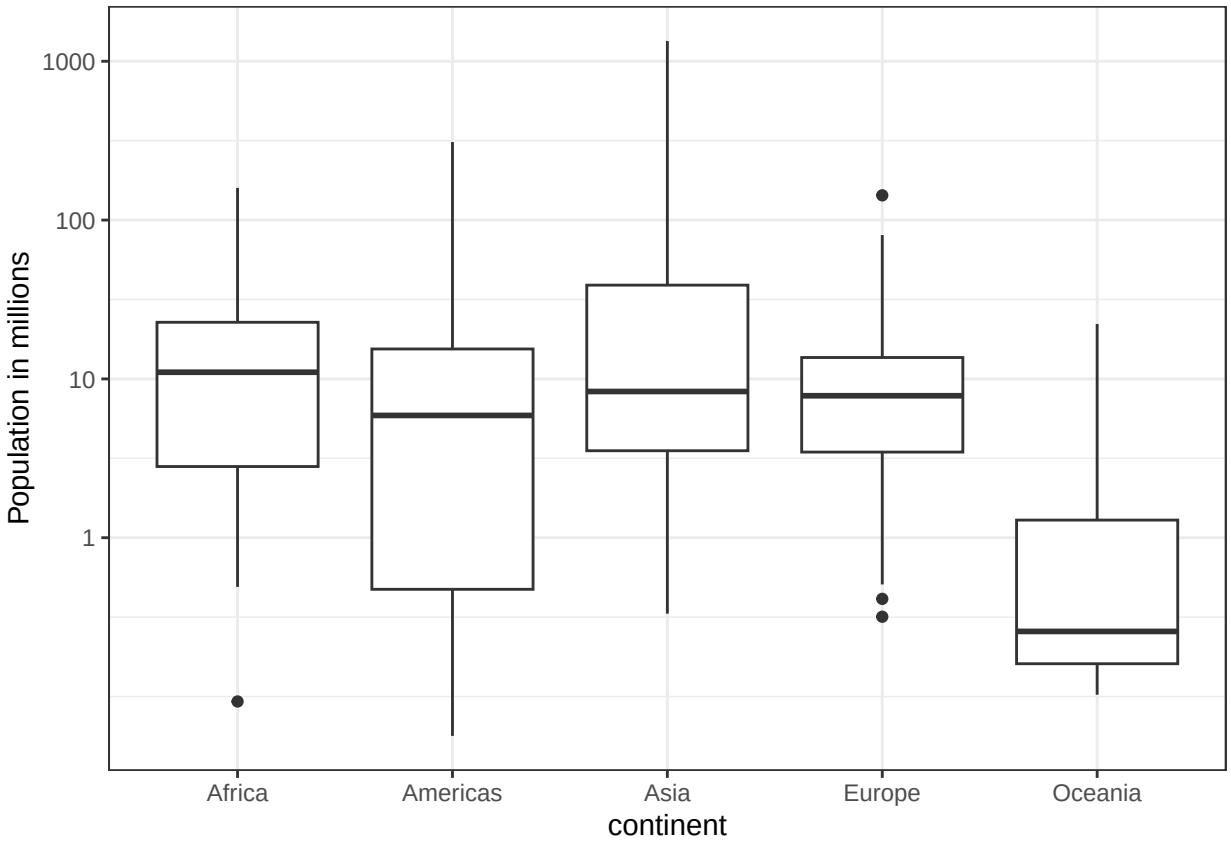
11. Define variables containing the heights of males and females like this:

```
library(dslabs)
data(heights)
male <- heights$height[heights$sex == "Male"]
female <- heights$height[heights$sex == "Female"]
```

How many measurements do we have for each?

12. Suppose we can't make a plot and want to compare the distributions side by side. We can't just list all the numbers. Instead, we will look at the percentiles. Create a five row table showing `female_percentiles` and `male_percentiles` with the 10th, 30th, 50th, 70th, & 90th percentiles for each sex. Then create a data frame with these two as columns.

13. Study the following boxplots showing population sizes by country:



Which continent has the country with the biggest population size?

14. What continent has the largest median population size?

15. What is median population size for Africa to the nearest million?

16. What proportion of countries in Europe have populations below 14 million?

    a. 0.99
    b. 0.75
    c. 0.50
    d. 0.25

17. If we use a log transformation, which continent shown above has the largest interquartile range?

18. Load the height data set and create a vector `x` with just the male heights:

```
library(dslabs)
data(heights)
x <- heights$height[heights$sex=="Male"]
```

What proportion of the data is between 69 and 72 inches (taller than 69, but shorter or equal to 72)? Hint: use a logical operator and `mean`.

19. Suppose all you know about the data is the average and the standard deviation. Use the normal approximation to estimate the proportion you just calculated. Hint: start by computing the average and standard deviation. Then use the `pnorm` function to predict the proportions.

20. Notice that the approximation calculated in question nine is very close to the exact calculation in the first question. Now perform the same task for more extreme values. Compare the exact calculation and the normal approximation for the interval (79,81]. How many times bigger is the actual proportion than the approximation?

21. Approximate the distribution of adult men in the world as normally distributed with an average of 69 inches and a standard deviation of 3 inches. Using this approximation, estimate the proportion of adult men that are 7 feet tall or taller, referred to as *seven footers*. Hint: use the `pnorm` function.

22. There are about 1 billion men between the ages of 18 and 40 in the world. Use your answer to the previous question to estimate how many of these men (18-40 year olds) are seven feet tall or taller in the world?

23. There are about 10 National Basketball Association (NBA) players that are 7 feet tall or higher. Using the answer to the previous two questions, what proportion of the world's 18-to-40-year-old *seven footers* are in the NBA?

24. Repeat the calculations performed in the previous question for Lebron James' height: 6 feet 8 inches. There are about 150 players that are at least that tall.

25. In answering the previous questions, we found that it is not at all rare for a seven footer to become an NBA player. What would be a fair critique of our calculations:

a. Practice and talent are what make a great basketball player, not height.
b. The normal approximation is not appropriate for heights.
c. As seen in question 10, the normal approximation tends to underestimate the extreme values. It's possible that there are more seven footers than we predicted.
d. As seen in question 10, the normal approximation tends to overestimate the extreme values. It's possible that there are fewer seven footers than we predicted.