Section 2.4:

Boxplots and Quantitative / Categorical Relationships

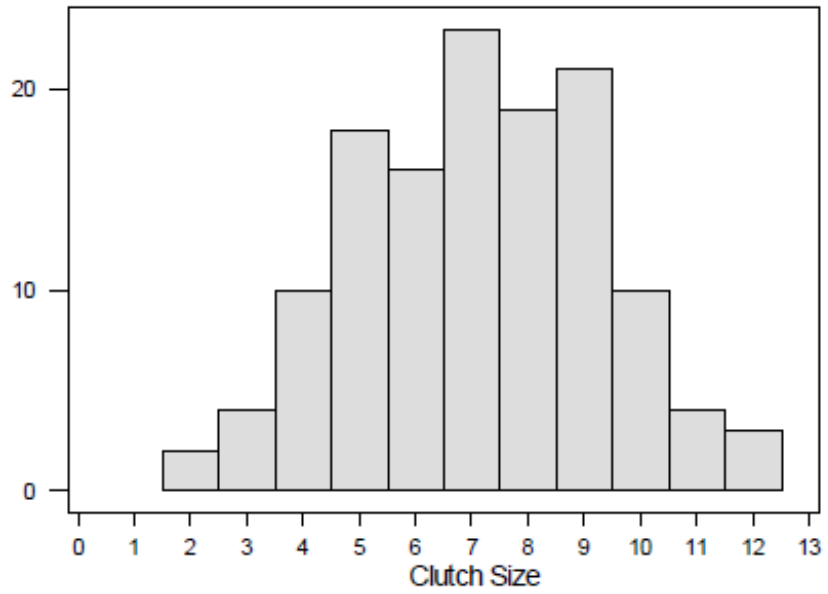
A naturalist counts the number of baby birds, or *clutch size*, in 130 different nests. The standard deviation of clutch sizes is closest to…

A. 1

B. 2

C. 3

D. 4



Outliers

- Outliers can be informally identified by looking at a plot, but one rule of thumb for identifying outliers is data values more than 1.5 of the IQR beyond the quartiles

- A data value is an **outlier** if it is

Smaller than Q1 – 1.5(IQR)

or

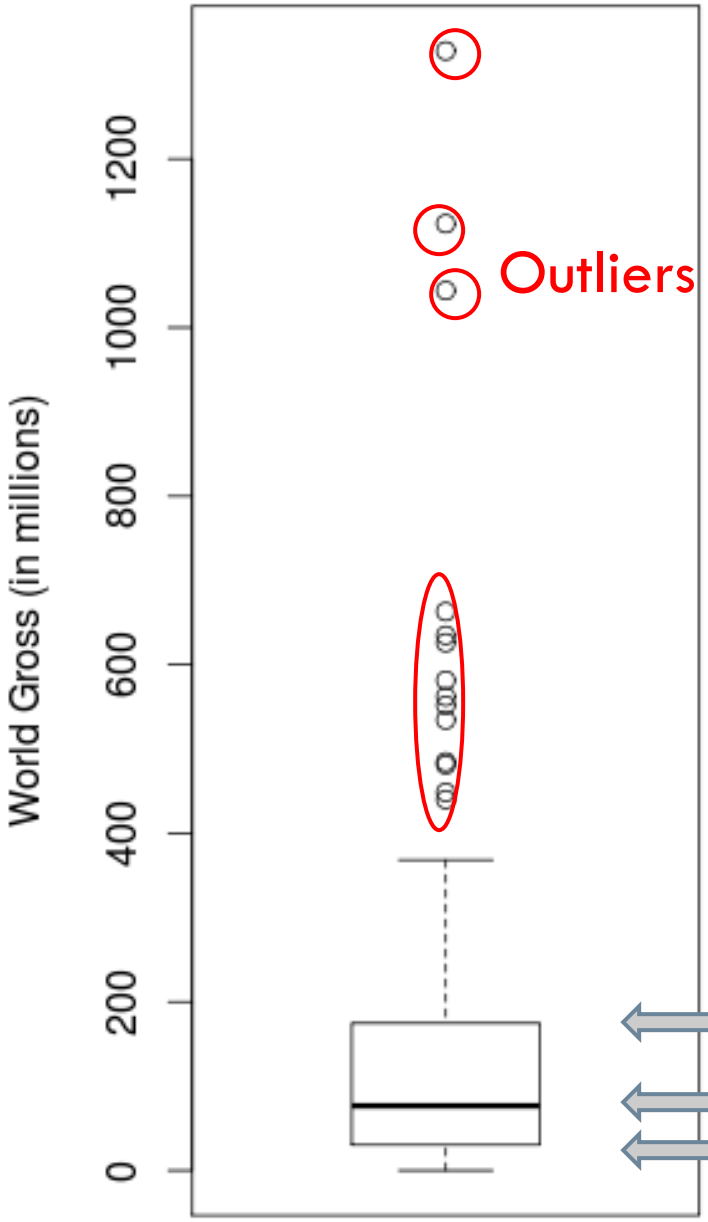
Larger than Q3 + 1.5(IQR)

Boxplot

To draw a ***boxplot***:

- Draw a numerical scale appropriate for the data
- Draw a box stretching from Q_1 to Q_3
- Divide the box with a line at the median
- Draw a line from each quartile to the most extreme data value **that is *not* an outlier**
- Identify each outlier individually by plotting with a symbol such as an asterisk or dot

Boxplot



- Lines ("whiskers") extend from each quartile to the most extreme value that is not an outlier

Example: Infection in Dialysis Patients

- The table on the next slide gives data showing the time to infection, at the point of insertion of the catheter, for kidney patients using portable dialysis equipment. There are 38 patients, and the data give the first observation for each patient. The five number summary for the data is (2, 15, 46, 149, 536).

McGilchrist, C. and Aisbett, C., "Regression with frailty in survival analysis," *Biometrics*, 1991; 47: 461-466

Example: Infection in Dialysis Patients

2	5	6	7	7	8	12	13	15
15	17	22	22	23	24	27	30	34
39	53	54	63	96	113	119	130	132
141	149	152	152	185	190	292	402	447
511	536							

a) Identify any outliers in the data. Justify your answer.

b) Draw the boxplot.

c) Describe the shape of the data.

Example: Infection in Dialysis Patients

a) Identify any outliers in the data. Justify your answer.

The five number summary for the data is (2, 15, 46, 149, 536).

$Q_1 - 1.5(IQR) =$

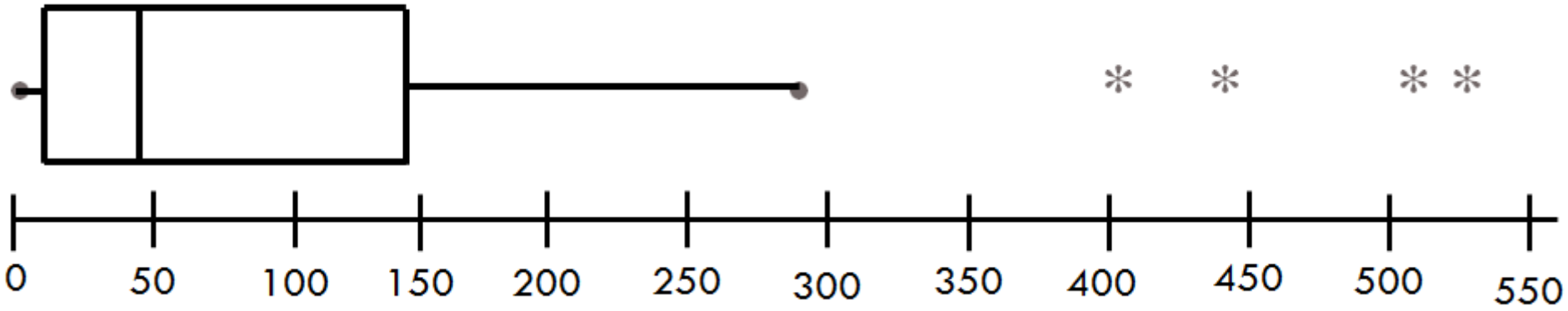
$Q_3 + 1.5(IQR) =$

Outliers:

Example: Infection in Dialysis Patients

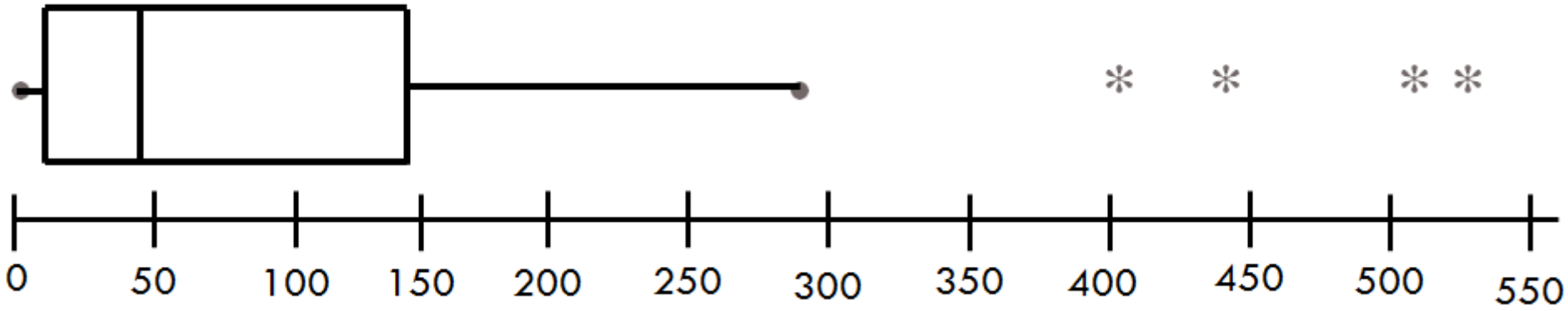
b) Draw the boxplot.

c) Describe the shape of the data from the *Infection in Dialysis Patients* dataset.



A. Skewed right
B. Skewed left
C. Approximately symmetric
D. None of these

d) Give a rough approximation for the mean of the *Infection in Dialysis Patients* dataset.



A. 15

B. 45

C. 100

D. 175

Summary: One Quantitative Variable

- Summary Statistics
  - Center: mean, median
  - Spread: standard deviation, range, IQR
  - Measures of Location: z-scores, Percentiles, Quartiles

- Visualization
  - Dotplot
  - Histogram
  - *Boxplot*

- Other concepts
  - Shape: symmetric, skewed, bell-shaped
  - Resistance
  - *Outliers*

Quantitative and Categorical Relationships

- In this case, we are interested in breaking down a quantitative variable by categorical groups

Tea and the Immune System

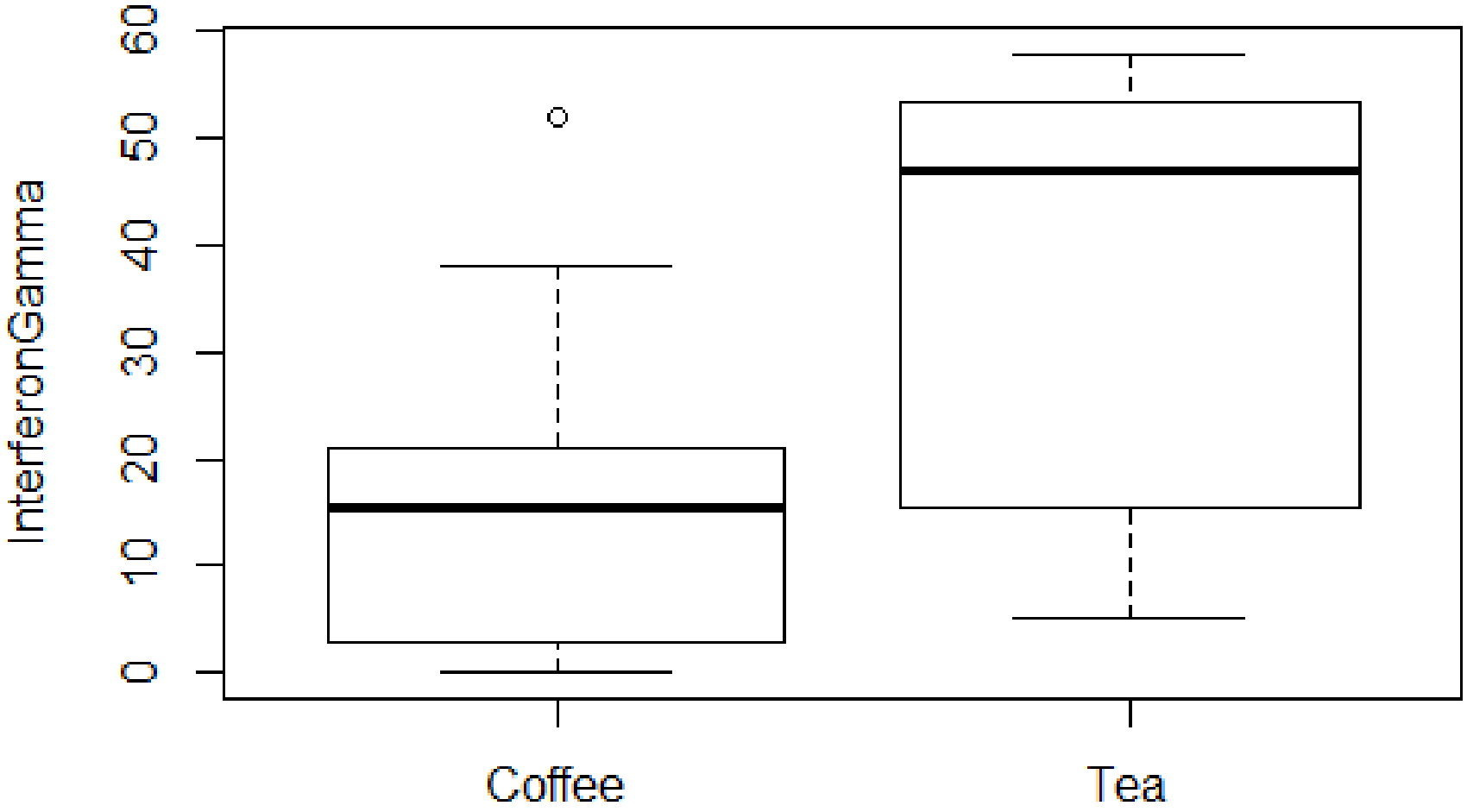
- Participants were randomized to drink five or six cups of either tea or coffee every day for two weeks (both drinks have caffeine but only tea has L-theanine)
- After two weeks, blood samples were exposed to an antigen, and production of interferon gamma (immune system response) was measured
- Explanatory variable: tea or coffee
- Response variable: measure of interferon gamma

Mednick, Cai, Kanady, and Drummond (2008). "Comparing the benefits of caffeine, naps and placebo on verbal, motor and perceptual memory," *Behavioral Brain Research,* 193, 79-86.

If the tea drinkers have significantly higher levels of interferon gamma, can we conclude that drinking tea rather than coffee *caused* an increase in this aspect of the immune response?

A. Yes
B. No

Visualization for One Categorical and One Quantitative Variable: Side-by-Side Boxplots



Quantitative Statistics by a Categorical Variable

- Any of the statistics we use for a quantitative variable can be looked at separately for each level of a categorical variable

- Mean level of interferon gamma by drink:

Tea	Coffee
$\bar{x}_T = 34.82$	$\bar{x}_C = 17.70$

Summary Statistic for One Categorical and One Quantitative Variable: Difference in Means

- Often, when comparing a quantitative variable across two categories, we compute the ***difference in means.***

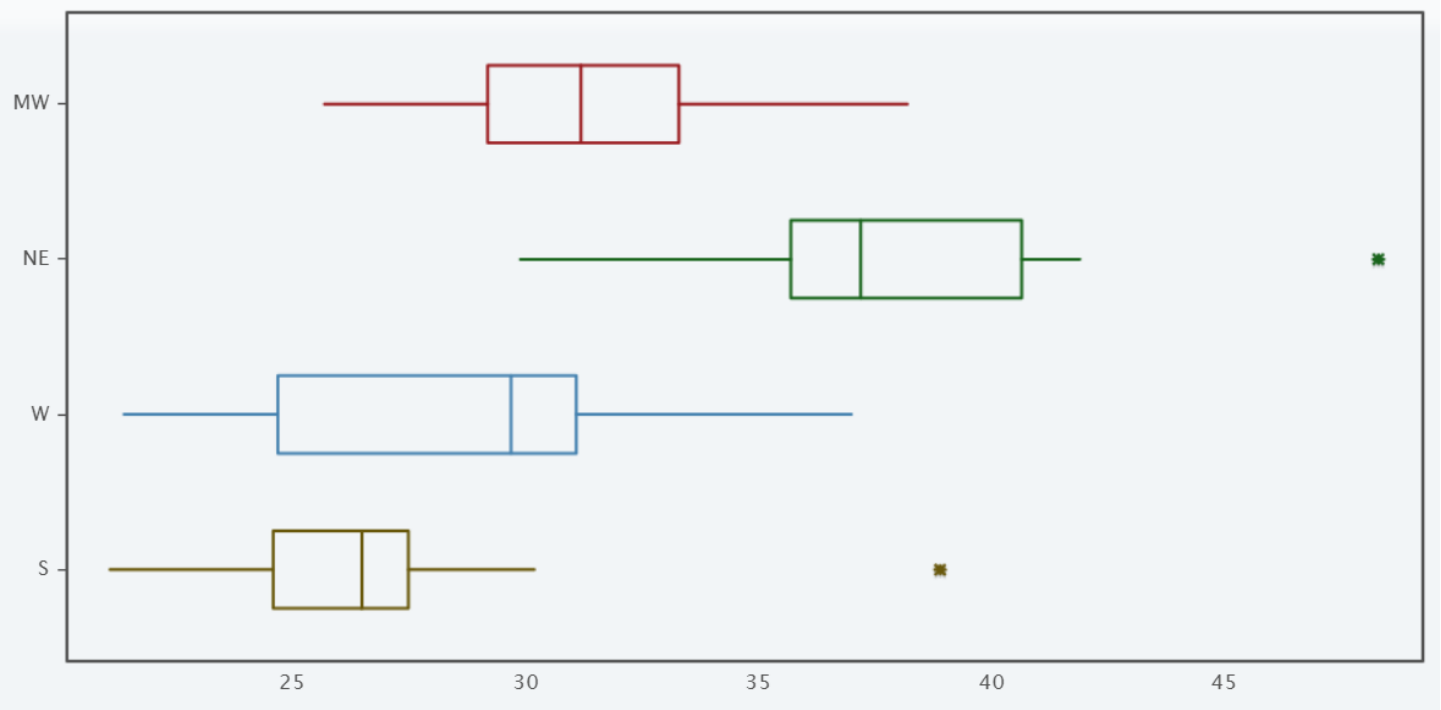
- $\bar{x}_T - \bar{x}_C = 34.82 - 17.70 = 17.12$

Percent of College Graduates by Region of the US

- The dataset **USStates** includes information on the percent of the population to graduate from *college* (of those age 25-34) for each US state. Use *StatKey* to obtain side-by-side boxplots for percent of college graduates by region of the country (Midwest, Northeast, South, and West.) Then answer the following questions.

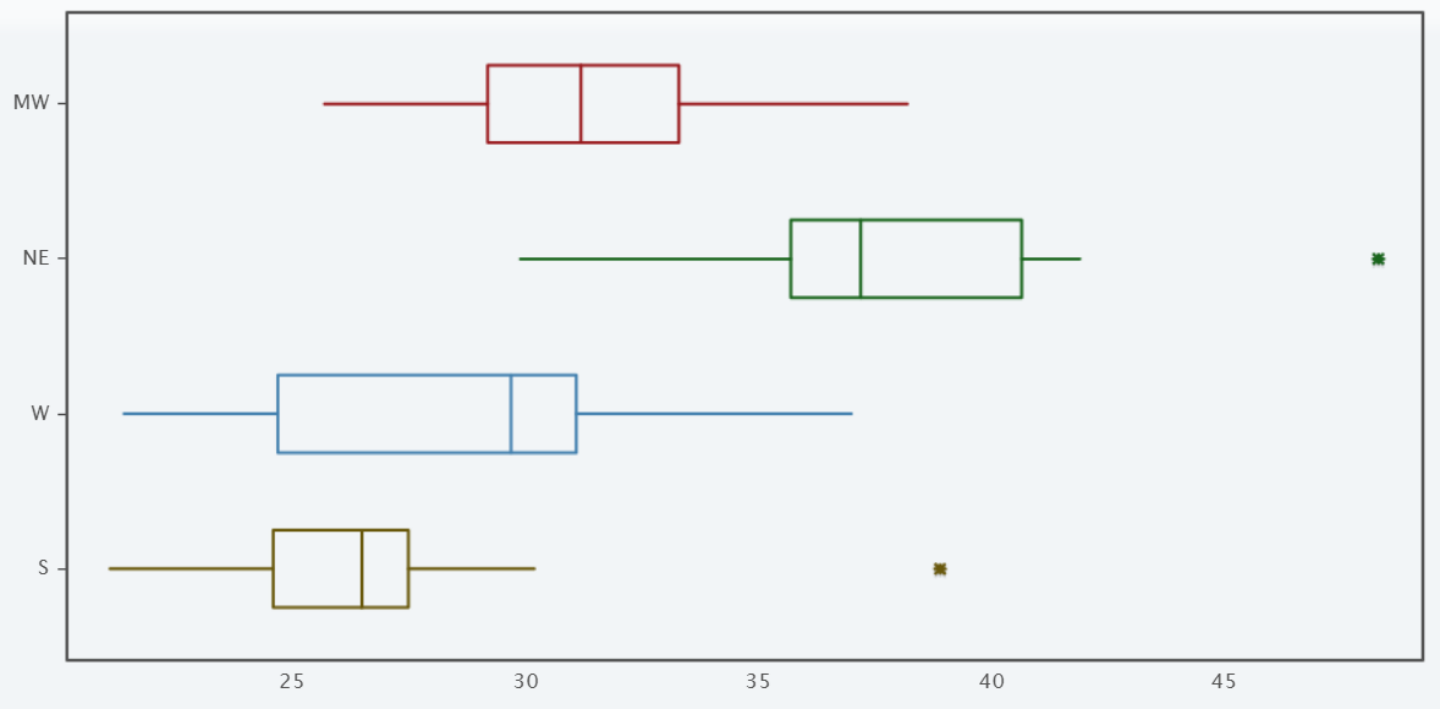
Various online sources, mostly at www.census.gov

Which region has the highest percent of college graduates?



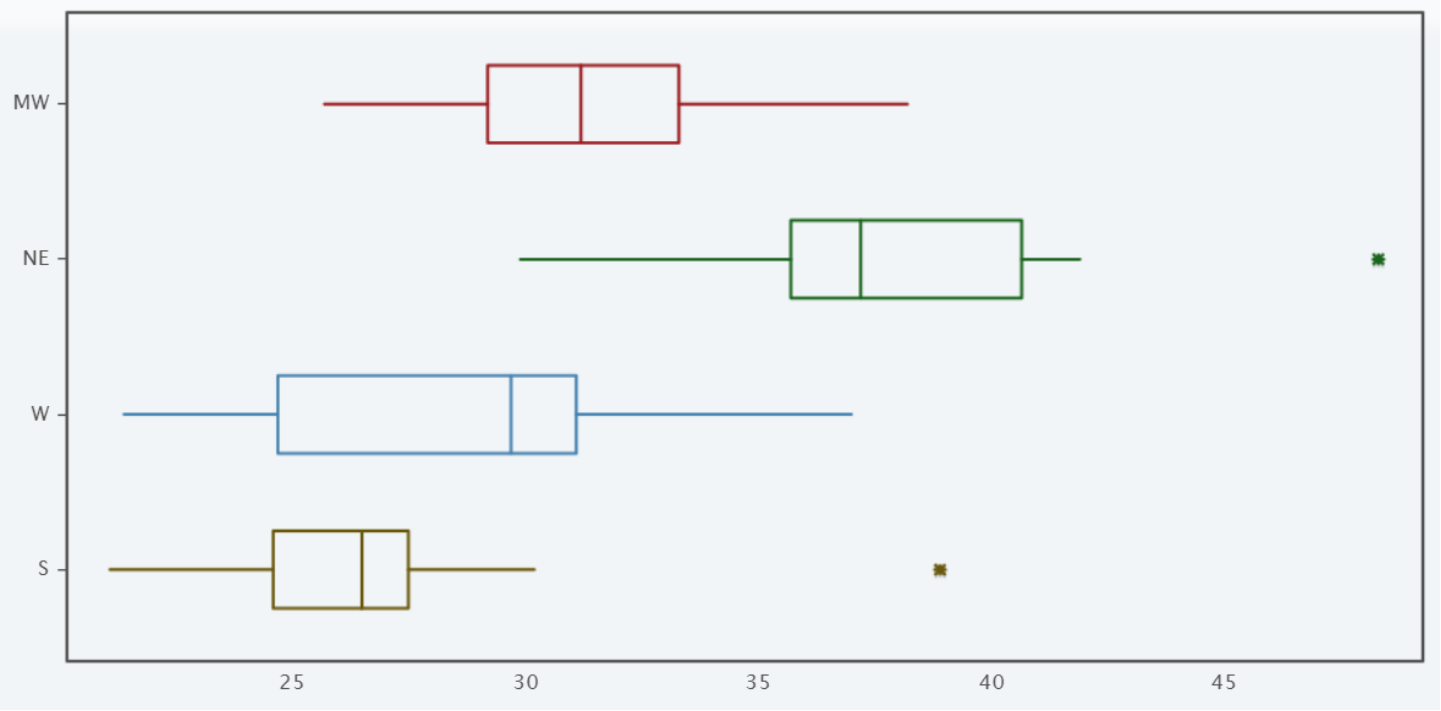
A. Midwest
B. Northeast
C. South
D. West

Which region has the lowest percent of college graduates?



A. Midwest
B. Northeast
C. South
D. West

How many outliers are there?



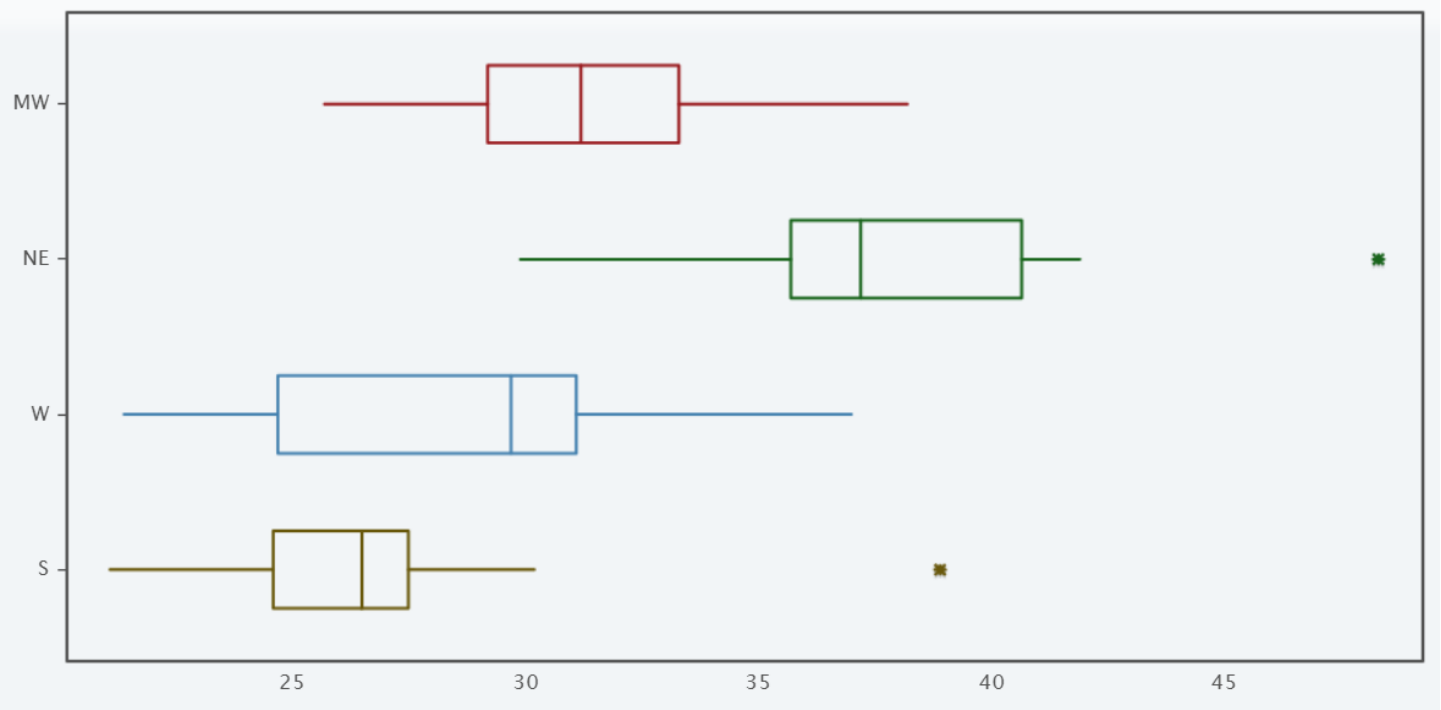
A. 2

B. 4

C. 6

D. 8

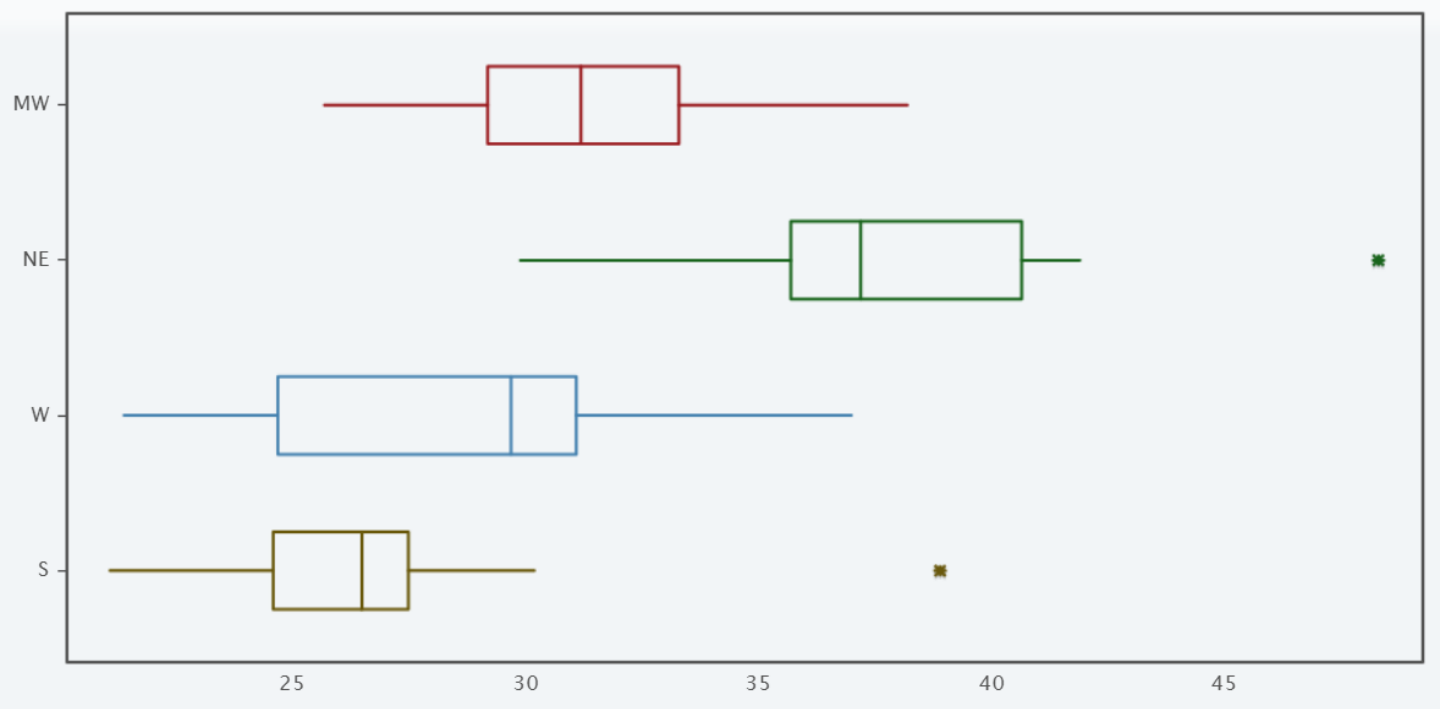
Does there appear to be an association between these two variables?



A. Yes

B. No

Can we conclude that the *Region* and *College* are causally related?



A. Yes

B. No

Summary: One Quantitative and One Categorical

- Summary Statistics
  - Any summary statistics for quantitative variables, broken down by groups
  - Difference in means
- Visualization
  - Side-by-side boxplots