Bookmark this page

Statistics Package Exercise: Creating Side-by-Side Boxplots

Learning Objective: Compare and contrast distributions (of quantitative data) from two or more groups, and produce a brief summary, interpreting your findings in context.

The objectives of this activity are:

- to teach you how to use to produce side-by-side boxplots and the relevant descriptive statistics,
- to let you practice comparing and contrasting distributions, and
- to help you gain more intuition about variability through the interpretation of your results in context.

The percentage of each entering Freshman class that graduated on time was recorded for each of six colleges at a major university over a period of several years. (Source: This data is distributed with the software package, Data Desk. (1993). Ithaca, NY: Data Description, Inc., and appears in http://lib.stat.cmu.edu/DASL/)

In order to compare the graduation rates among the different colleges, we will create side-by-side boxplots (graduation rate by college), and supplement the graph with numerical measures. Follow the instructions, and then answer the questions based on the output you got.

- R | StatCrunch | TI Calculator | Minitab | Excel

R Instructions

To open R with the dataset preloaded, right-click here and choose "Save Target As" to download

the file to your computer. Then find the downloaded file and double-click it to open it in R.

The data have been loaded into the data frame

```
grad_data
```

. Enter the command

```
grad_data
```

to see the data. There are 6 variables (column titles) in the data frame

```
grad_data
```

:

```
College.A
```

,

```
College.B
```

,

```
College.C
```

,

```
College.D
```

,

```
College.E
```

, and

```
College.F
```

.

You should see graduation data for six colleges over the past eight years. Copy the next command

to see a summary of the data for each college:

```
summary(grad_data)
```

By using the

```
summary()
```

command on the data frame instead of an individual variable the summary statement of the five number summary and mean are provided for each variable in the data frame.

Finally, copy the next command to see side-by-side boxplots of the graduation data for the six colleges:

```
boxplot(grad_data)
```

Just as we did with the histogram we can add x-axis and y-axis labels and titles using the same additional parameters in the

```
boxplot()
```

command,

```
xlab=
```

,

```
ylab=
```

, and

```
main=
```

. For example:

```
boxplot(grad_data, xlab="Colleges",ylab ="Graduation Rates",
main="Comparison of Graduation Rates")
```

We can also modify the direction of the boxes by adding another parameter

```
horizontal=TRUE
```

. For example:

```
boxplot(grad_data, horizontal=TRUE, ylab="Colleges",xlab
="Graduation Rates", main="Comparison of Graduation Rates")
```

Notice that you must switch the x and y-axis labels when you make a horizontal boxplot.

Note: Using R-Use the mouse to grab the corner of the graph window and change its shape. If you make the window wider, you see a label for each boxplot. While the graph window is the active window, try clicking the **File** menu. If you hold the mouse over "Copy to Clipboard," you see two ways that you can copy a graph for pasting into another document. This is how you use R to create data graphs for reports.

R coding Note for reference only: The above command works because all six variables are listed in separate columns. Data organization often plays a role in how you structure an R command. For example, if the data was instead organized into two columns

```
GradRate
```

and

```
College
```

in a new data frame called

```
grad2
```

, where

```
GradRate
```

contained all the numeric responses and

```
College
```

contained the labels *A, B, C,* etc., then the code would be as follows:

```
boxplot(grad2$GradRate~grad2$College)
```

Answer the following questions:

Learn By Doing (1/1 point)

Compare and contrast the distributions of the graduation rates at the different colleges. Be sure to address center, spread and outliers.

Your Answer:

College D has the best because their range of possible values based on the 5 key values is much lower. Some in F have actually gotten higher than most students from D; but D's performance is still more consistent. so COllege F would be a nice 2nd place.
College A and E have the worst. A's whiskers are lowest and lowest Q1. E's IQR is pretty small compared to the others.

Our Answer:

RStatCrunchTI CalculatorMinitab Excel R Center: Of the six colleges, college D has the highest median graduation rate (M = 79), followed by colleges F (M = 72), B (M = 70.15), C (M = 67.65), A (M = 63.75) and college E (M = 59.15). Spread: College B has the smallest variation in graduation rates over the years (range = 9.4%, IQR = 3.5%). College D's graduation rates are also pretty consistent over the years (range = 10.5%, IQR = 4.45%). A larger variation in graduation rates is found in colleges E and C, and the least consistency in graduation rates (i.e., largest variation) is found in college A (range = 30.6%, IQR = 19.55%) and college F (range = 29.7%, IQR = 16.23%). None of the graduation rates distributions have outliers. StatCrunch Center: Of the six colleges, college D has the highest median graduation rate (M = 79), followed by colleges F (M = 72), B (M = 70.15), C (M = 67.65), A (M = 63.75) and college E (M = 59.15). Spread: College B has the smallest variation in graduation rates over the years (range = 9.4%, IQR = 4%). College D's graduation rates are also pretty consistent over the years (range = 10.5%, IQR = 6.1%). A larger variation in graduation rates is found in colleges E and C, and the least consistency in graduation rates (i.e., largest variation) is found in college A (range = 30.6%, IQR = 20.6%) and college F (range = 29.7%, IQR = 19.15%). None of the graduation rates distributions have outliers. TI Here are the descriptive statistics and the graph: Center: Of the six colleges, college D has the highest median graduation rate (M = 79), followed by colleges F (M = 72), B (M = 70.15), C (M = 67.65), A (M = 63.75) and college E (M = 59.15). Spread: College B has the smallest variation in graduation rates over the years (range = 9.4%, IQR = 3.5%). College D's graduation rates are also pretty consistent over the years (range = 10.5%, IQR = 4.45%). A larger variation in graduation rates is found in colleges E and C, and the least consistency in graduation rates (i.e., largest variation) is found in college A (range = 30.6%, IQR = 19.55%) and college F (range = 29.7%, IQR = 16.23%). None of the graduation rates distributions have outliers. Minitab Center: Of the six colleges, college D has the highest median graduation rate (M = 79), followed by colleges F (M = 72), B (M = 70.15), C (M = 67.65), A (M = 63.75) and college E (M = 59.15). Spread: College B has the smallest variation in graduation rates over the years (range = 9.4%, IQR = 4.5%). College D's graduation rates are also pretty consistent over the years (range = 10.5%, IQR = 7.75%). A larger variation in graduation rates is found in colleges E and C, and the least consistency in graduation rates (i.e., largest variation)is found in college A (range = 30.6%, IQR = 21.65%) and college F (range = 29.7%, IQR = 22.07%). None of the graduation rates distributions have outliers. Excel Here are

the descriptive statistics and the graph. Note that these might look a little different than the ones you generated depending on the version of Excel you are using, but the values should be the same. Center: Of the six colleges, college D has the highest median graduation rate (M = 79), followed by colleges F (M = 72), B (M = 70.15), C (M = 67.65), A (M = 63.75) and college E (M = 59.15). Spread: College B has the smallest variation in graduation rates over the years (range = 9.4%, IQR = 3.5%). College D's graduation rates are also pretty consistent over the years (range = 10.5%, IQR = 4.45%). A larger variation in graduation rates is found in colleges E and C, and the least consistency in graduation rates (i.e., largest variation) is found in college A (range = 30.6%, IQR = 19.55%) and college F (range = 29.7%, IQR = 16.23%). None of the graduation rates distributions have outliers.

Resubmit Reset

Learn By Doing (1/1 point)

If you had to choose one college among the six colleges based on this data, which college would it be? Explain your reasoning.

Your Answer:

D. Because statistically speaking, i will most likely get good grades as long as I stuck to the same efforts as the others.

Our Answer:
If I had to choose one college based on the graduation rates, I would choose college D. Not only does this college have the largest median graduation rates, but it also has the smallest variation in graduation rates over the years. This means that even in years when college D has a relatively small graduation rate, it is not MUCH smaller than the median (min = 74.1%, Median = 79%), and is still higher than most graduation rates at the other colleges. In particular, the smallest graduation rate that occurred in college D (74.1%), is still higher than: • the highest graduation rate at colleges A and E, • the third quartile of the distribution of graduation rates at colleges B and C, and • the median graduation rate at college F.

Resubmit Reset

Learn By Doing (1/1 point)

If you were debating between colleges B and F only, which one would you choose based on this data? Explain your reasoning.

Your Answer:

F. Because although B is safer, in a way, I'm confident in my own skills so I'll probably be in Q3 or above based on the boxplot. Outlier that's too high if possible.

Our Answer:

While Colleges B and F have about the same median graduation rate (B: 70.15%, F: 72%), there is a big difference in the variation. College F has a very large variation, and therefore a much less consistent gradation rate over the years (it can get as low as 57.7%). College B, on the other hand, has a much more consistent graduation rate (small variability), and in that sense, college B is less "risky." Since I am not a "risk taker," and since both colleges have approximately the same median graduation rate, my choice between the two would be college B.

Resubmit Reset