# Lab #1 – Describing Data

This lab covers the use of summary statistics and basic graphs to describe data and address research questions. We will use survey data from STA-209-02 and STA-209-04. The lab will consist of three parts. The first part of the lab will be a guided tutorial to various descriptive statistics and graphics in Minitab. The second part will consist of a few review questions ensuring you understand concepts related to these modes of description. The third part will be an opportunity for your group to formulate and research question (of your choosing) and answer it using data.

As you work through the lab there are several questions your group needs to answer. These write-up questions are colored in red. I expect to receive exactly one lab write-up per group (with everyone’s names on it), but I encourage each group member to keep their own copy.

**PLEASE ADDRESS THESE QUESTIONS IN A SEPARATE DOCUMENT**

## PART 1:

Begin by downloading the file “class\_data.csv” from P-Web and loading it in to Minitab. This data contains de-identified responses to survey from the first day of class.

We will begin by considering how these data were collected.

### Question 1

Are the survey data a population or a sample? Do you believe that the collection of these data introduces any bias? Record and briefly explain (1-2 sentences) your responses.

Recall that there are two types of variables, **categorical** and **quantitative**. This lab will focus on ways to describe both types, as well as ways to describe possible relationships between them.

## Summary Statistics for One Categorical Variable:

One way of describing a categorical variable is to use a numeric quantities called **summary statistics**. As the name indicates, a summary statistic summarizes a variable using a numeric quantity(s).

The most basic example of a summary statistic are **frequencies**, which are simply the counts of categorical level. Often we display frequencies using **frequency tables**, for example using the Happy Planet data:

|  |  |
| --- | --- |
| **Region:** | **Frequency:** |
| 1 | 24 |
| 2 | 24 |
| 3 | 16 |
| 4 | 33 |
| 5 | 7 |
| 6 | 12 |
| 7 | 27 |
| Total | 143 |

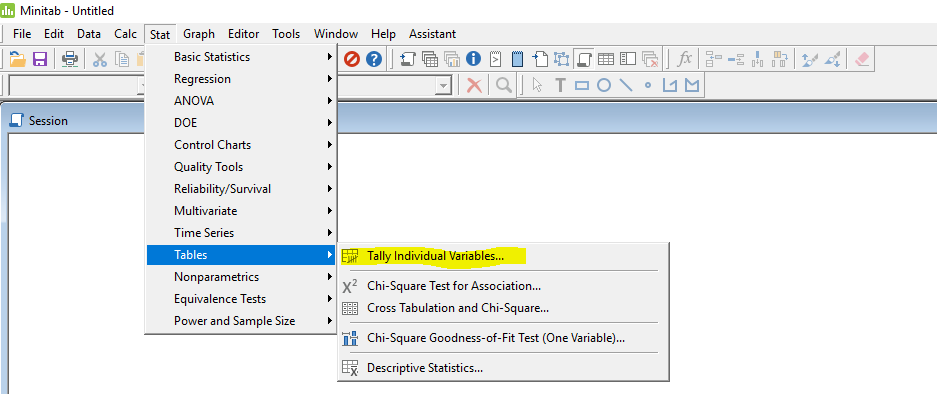
A related summary statistic is the **proportion**, or equivalently the percentage of cases in a category:

For example, the proportion of countries that are “former communist nations” (region 7) is given by: 27/143 = 0.189, or equivalently 18.9%

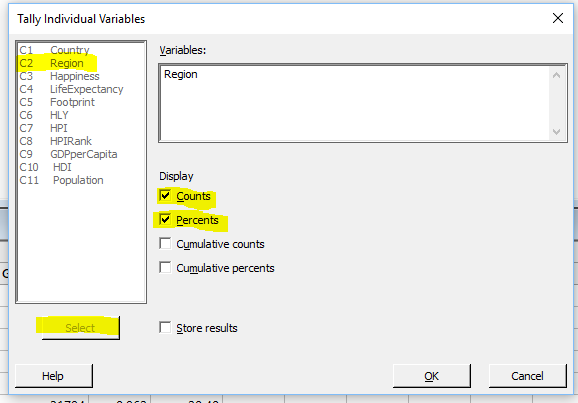
* When describing a *sample* we denote an observed proportion using , for example:
* When describing a *population* we denote a proportion using
* Using hats to denote estimates from a sample is a common convention in statistics

It is very tedious to tally frequencies and proportions by hand, so let’s see how to do it in Minitab:

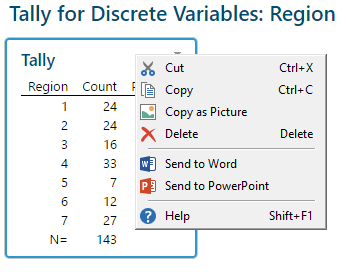
1. Begin by going to the “Stat” menu and selecting “Tally Individual Variables” under the “Tables”



1. Highlight the categorical variable you want to tabulate and click “Select”. Then make sure you’ve selected “Counts” and “Percents” to get frequencies and proportions for each category



1. Results will be output into your “Session” window. You can easily copy or export the table simply by right clicking on the table header

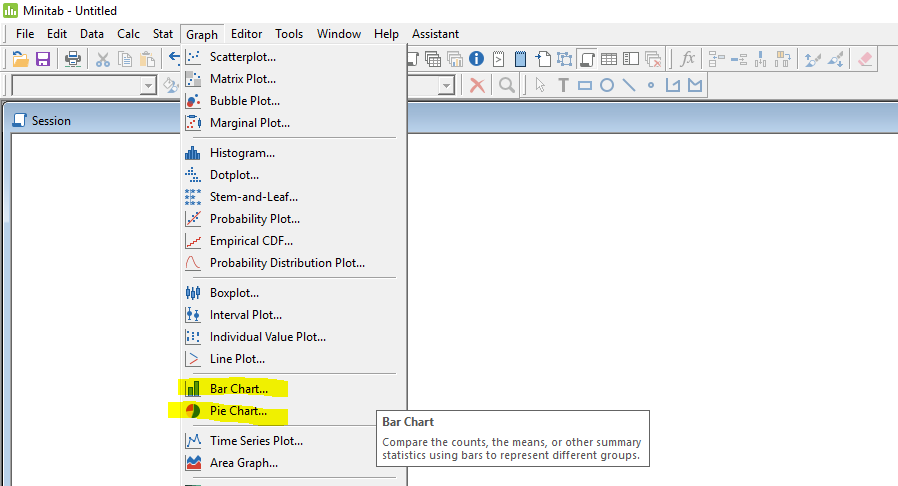


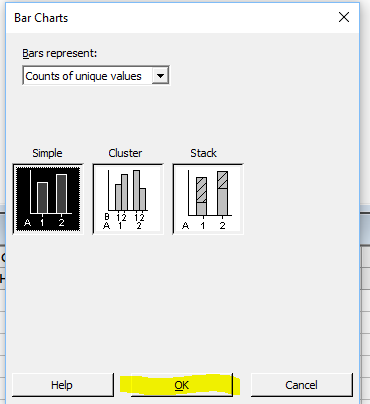
### Question 2

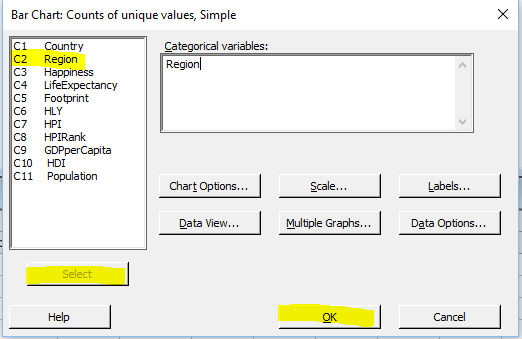
Create a frequency table for the Q3, “Circle your current (or intended) area of study (division)”. Record your output in your group’s lab write-up as an image or word table.

## Visualizations for One Categorical Variable:

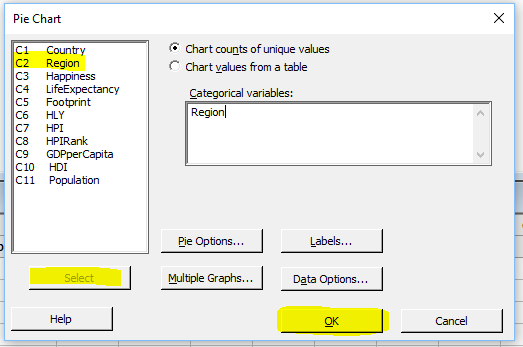
Summary statistics are often complemented by visualizations, which pictorially illustrate the characteristics of a variable. The options for visualizing a single categorical variable are somewhat limited, but two common choices are **bar charts**, which display frequencies for each category, and **pie charts**, which display proportions for each category. These can be created using the “Graph” menu:













### Question 3

Create both a bar chart and a pie chart for Q3 of the class survey data, record your output as an image or word picture.

### Question 4

Which do you feel does a better job conveying the information in the data, the bar chart or the pie chart? (explain in 1-2 complete sentences)

## Summary Statistics for Two Categorical Variables

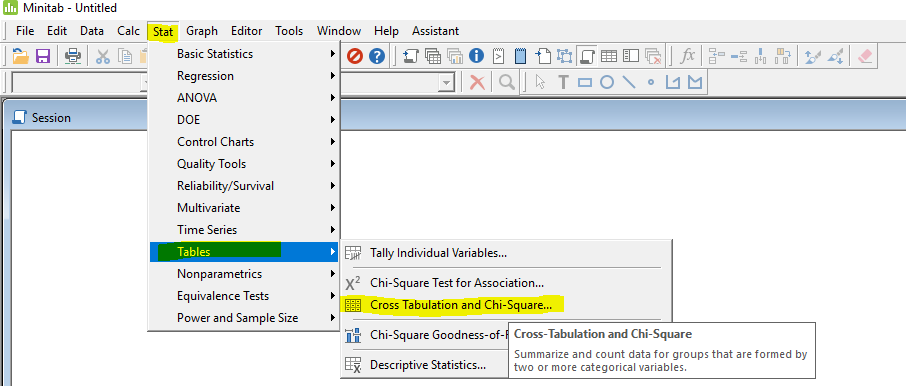
The story that can be told using a single variable is pretty limited. Most often we are interested in the relationships between two or more variables. **Frequencies** and **Proportions** can also be used to simultaneously describe two categorical variables. We’ve already seen an example of a **two way frequency table** in the Florida Courts Racial Bias Case Study:

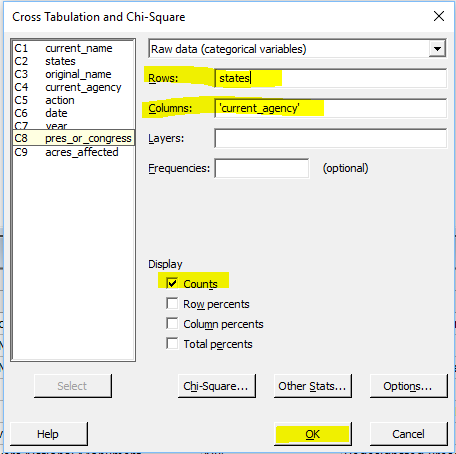
|  |  |  |
| --- | --- | --- |
|  | **Death Penalty Sentence** | **Not Lethal Sentence** |
| **White Offender** | 46 | 152 |
| **Black Offender** | 38 | 142 |

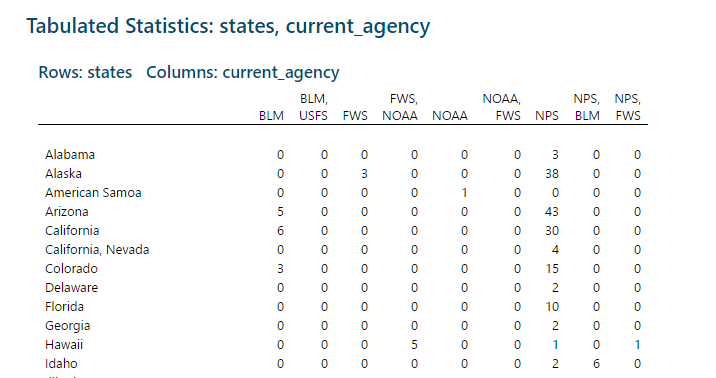
In this case study we also explored the idea of **conditioning,** or restricting our attention to a subset of cases in a particular category. For example, conditional upon an offender being white, what is the proportion of death penalty sentences?

In this example conditioning meant restricting our attention to the category described by a particular row. Similarly we could condition on a category described by a particular column, for example:

Now let’s see how to create two way tables in Minitab using the Happy Planet Data:







### Question 5

Create a two way frequency table for the responses to Q3: “Circle your current (or intended) area of study (division)” and Q7: “Which of the following best describes you (please circle only one)”. Use Q7 as the variable for table’s rows and Q3 for the table’s columns. Record your output as an image or word table.

### Question 6

Use conditional proportions to determine the area of study in which respondents most likely to be night owls. Show your calculation and interpret these results using a complete sentence.

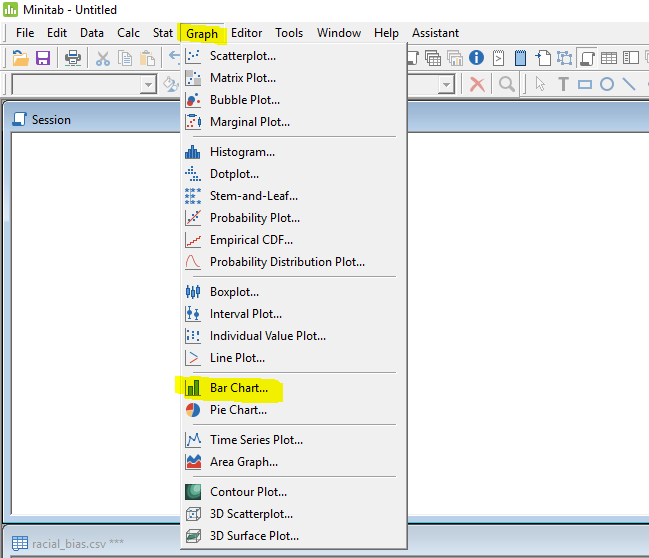
### Question 7

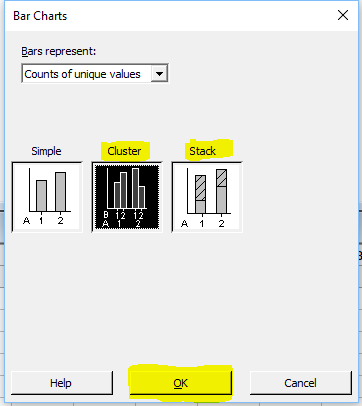
Does your finding in question 6 provide evidence for a *causal association* between these two variables? Explain why or why not in 1-2 sentences.

## Visualizations for Two Categorical Variables

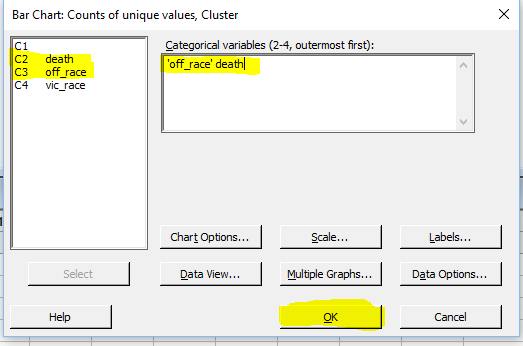
Two popular options for visualizing the relationship between categorical variables are **clustered bar charts**, where bars are grouped into clusters by an outer variable, and **stacked bar charts**, where bars are stacked onto each other for the different levels of an outer variable. Our textbook refers to these as “side-by-side” and “segmented” bar charts.

In Minitab these can be constructed using the “Graph” menu:





You need to select 2 variables for these graphs. The first variable you select will be the outer variable, it will determine the clusters. The second variable will be the inner variable, it will determine what gets displayed for each cluster.







### Question 8

Create both a clustered bar chart and a stacked bar chart to visualize the relationship between Q8: “What is your initial perception of statistics (circle one (1) option)” and Q6: “Have you taken a statistics course before (please circle only one)”. Record your results as an image or word graph. Write 1-2 sentences explaining how these two variables appear to be related (or not related).

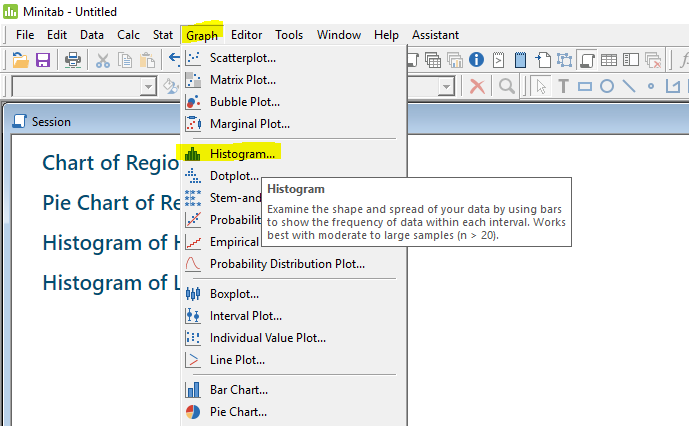
### Question 9

In the previous question, does it make more sense to use Q6 or to use Q8 as the outer variable? Try out both, and write 1-2 sentences explaining your preference.

## Histograms for One Quantitative Variable

While categorical variables are limited in scope by how the categories are defined, quantitative variables can often be more difficult to describe. Statisticians usually think about quantitative variables in terms of their **distribution**, a statistical concept that captures the *shape*, *center*, and *spread* of a variable. The best way to understand distributions is through visualization.

One of the most common visualizations for the distribution of a single quantitative variable is the **histogram**, which groups the data into equally spaced intervals and plots the frequencies of each group. In Minitab histograms can be created using the “Graph” menu. Below is an example for the variable “LifeExpectancy” in the Happy Planet Data:





Each **bin** used to construct the histogram above represents a 2.5 year interval. Minitab labels bins using their midpoint.

We are often interested in the general shape of a distribution, specifically whether is **symmetric** or if it is **skewed**. A symmetric distribution can be folded over a center line and the two sides will closely match each other. A skewed distribution has most of its data piled up on one side and a long tail in the other direction. The variable “LifeExpectancy” depicted above is **Skewed to the Left**, most countries have life expectancies between 70 and 80 years, but there is a long tail of countries with lower life expectancies. The graph of GDP per capita shown below is **Skewed to the Right**, it also contains an **outlier** (Luxembourg) whose per capita GDP of near 60,000 is much larger than any other country.



It is rare that we see *exact symmetry* in a real dataset, but often approximate symmetry is good enough. The histogram of “Happiness” is an example of a distribution that is **approximately symmetric**.



### Question 10

Make a histogram of the results for Q2 “Randomly choose a number between 1 and 20 and write it in the space below”. Record your output as an image or word graph.

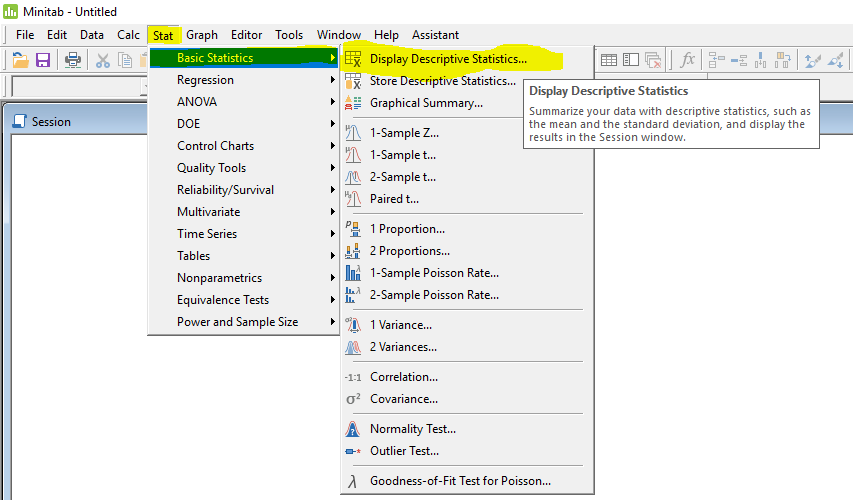
### Question 11

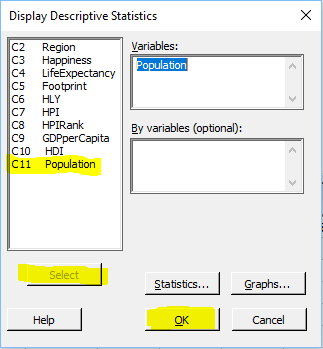
Is the distribution of random number choices symmetric or skewed? Do you believe these number choices are actually random? Explain why or not you believe they are random in 2-3 sentences. Feel to create a table to further investigate this question.

## Summary Statistics for One Quantitative Variable

Histograms are helpful for visualizing a distribution, but we can also summarize the important characteristics of the distribution numerically.

* The **mean** is the numerical average of the data:
* The **median** is the middle entry if data values are ordered from largest to smallest (or vice versa)
  + If the data contain an even number of cases, the median is the average of the middle two values
* The **standard deviation** is a measure of how much *variability* there is in a distribution, or how *spread out* the data are.
* The Pth **percentile** is the value of a quantitative variable that is greater than P percent of the data (note: sometimes percentiles are defined using greater than or equal to P percent of the data)
  + It is common to report the **First Quartile** and the **Third Quartile:**
    - First Quartile = 25th percentile
    - Third Quartile = 75th percentile
    - Interquartile Range (IQR) =
* The **minimum**, or smallest value, and the **maximum**, or largest value, are also commonly reported when summarizing the distribution of a quantitative variable.

In Minitab we can obtain these summary statistics using the “Stats” menu: 



Statistics

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Variable | N | N\* | Mean | SE Mean | StDev | Minimum | Q1 | Median | Q3 | Maximum |
| Population | 143 | 0 | 44.1 | 12.2 | 145.5 | 0.3 | 4.4 | 10.5 | 32.3 | 1304.5 |

The set of numbers: (Min, Q1, Median, Q3, Max) is known as the **five number summary**.

### Question 12

Report the mean, median, standard deviation and IQR for the responses to Q2 “Randomly choose a number between 1 and 20 and write it in the space below”.

### Question 13

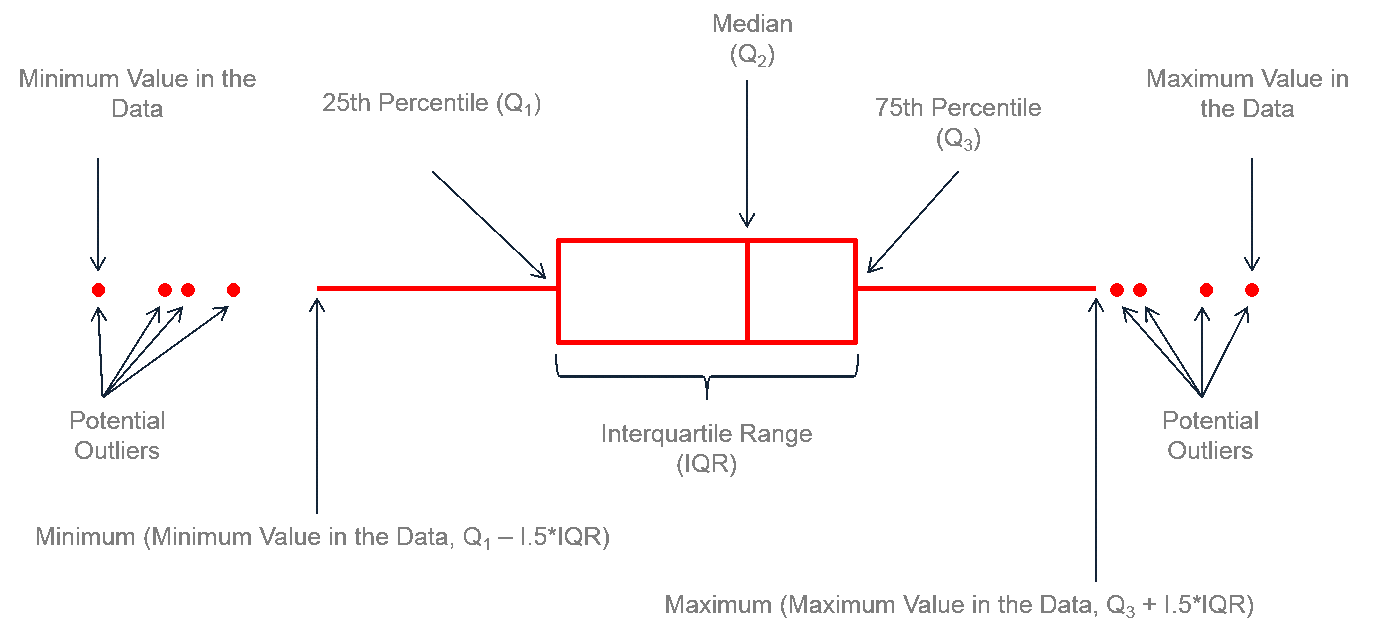
Suppose a student misread Q2 and chose the number 88, how would this impact your answer to Question 12? For the mean, median, standard deviation and IQR state whether you’d expect that statistic to increase, decrease, or remain the same.

### Question 14

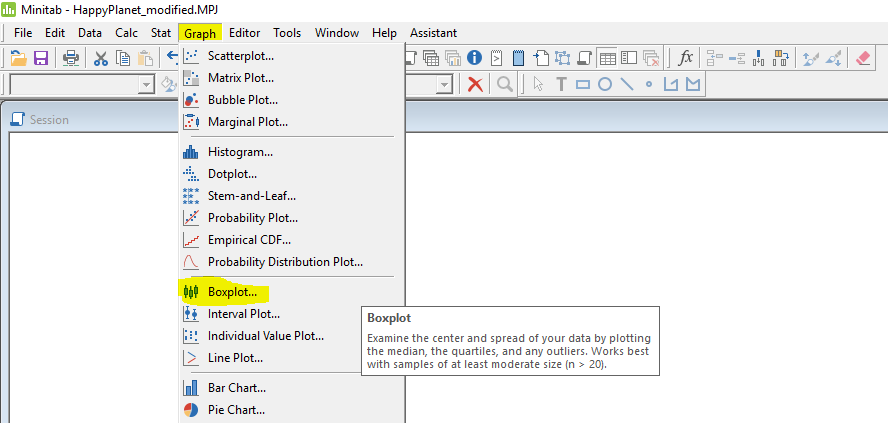
In 2-3 sentences explain how you could use summary statistics such as the mean, median, Q1, and Q3 to determine whether a not a distribution is skewed and the direction of the skew. Feel free to include an example using real or made up numbers in your justification.

## Boxplots for Quantitative/Categorical Relationships

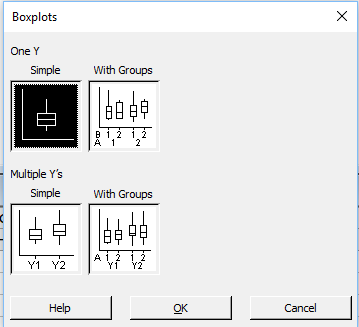
Histograms provide a detailed picture of the distribution of a single quantitative variable, but often we are interested in more than just summarizing, we want to see how a quantitative variable relates to a categorical variable. We could separate the data by category and overlay the histograms, but that quickly gets messy. **Boxplots** are a better way of assessing the relationship between quantitative and categorical variables. If you’re rusty on boxplots see the diagram below for a refresher:



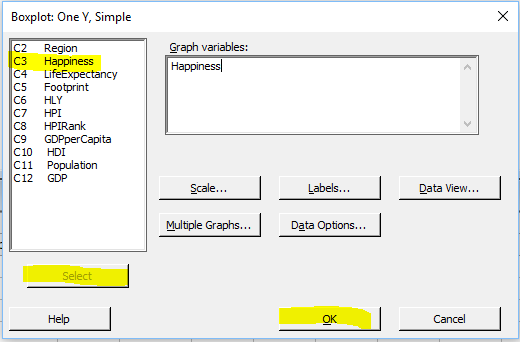
In Minitab you can create a box plot using the “Graph” menu and selecting boxplot:



The “One Y”, “Simple” option allows you to create a boxplot for a single variable, the “One Y”, “With Groups” option allows you to split that single variable by categories.

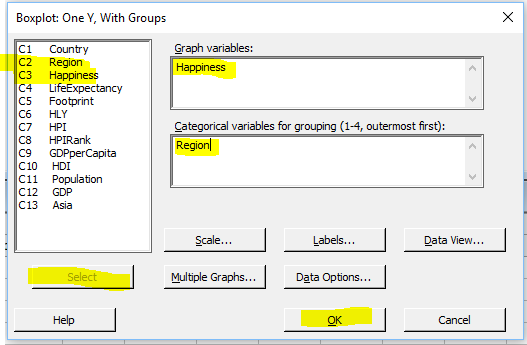


Here is the boxplot for “Happiness” in the Happy Planet Data:





Here are boxplots for Happiness by Region:





### Question 15

Use boxplots to answer the question “Are students who’ve taken a statistics course better at generating random numbers between 1 and 20?” Include your boxplots along with 2-3 sentences explaining your reasoning. Your explanation should mention quantiles, the IQR, and the median.

## PART 2:

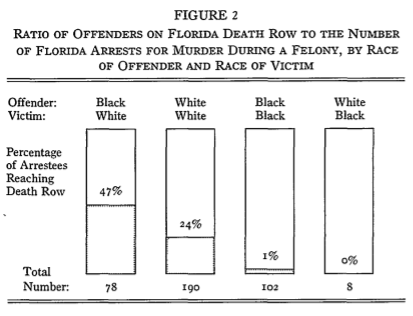
### Question 16

Think of an interesting research question (or multiple questions!) that you can answer using the class survey data. Record your question, the approach you used to answer the question, and a summary of your results. Be sure to include all relevant tables or graphs. This should be at least 3-4 sentences accompanied by 1-2 tables or figures. Time permitting, we will share the findings of your analyses with the class!

## Enhancements and Formatting (Optional – only if your group has time)

### Question 17 (Optional)

Recreate this graph that was used in the journal publication of the Florida Courts Racial Bias study:



It might be helpful to look at Question 18 below for where to find help.

### Question 18 (Optional)

Minitab has extensive documentation on how to format output. I have no problems with the default format for assignments in this course, but knowing how to make things look prettier is a worthwhile skill. Visit the link below and follow the menus:

<https://support.minitab.com/en-us/minitab/18/>

Help-and-How-To -> Graphs -> How-To -> ?

Explore these pages and use the newly gained knowledge to change the formatting of your answers (in particular your answer Question 16) to make them more visually appealing.