Lab 2 BIOS 4120

Sean DeVries

January 24, 2017

Lab 2

For today's lab we will first begin with a short review over what we learned last week by working our way through an example using IQ's near or far from a lead smelter. This data can be found in the data sets on the course page under lead-IQ.

R Review Excercise

First open RStudio and work through the problems below.

1. Set a variable named IQdata as the data using a read.delim function.
2. Using the summary function what is the minimum, maximum, and average IQ's in the data?
3. How many people were there in the study? How many were near the smelter? How many were far from the smelter?

Obtaining variables from a data set

In R if you are looking at a dataset that has numerous variables on one subject you can pull the one variable out using the dollar sign, $. An example of this is if you had the above IQdata saved you could pull out just the IQ scores with the line:

```
IQdata$IQ
```

```
##   [1]  70  85  86  76  96  94 115  97 128  99 118 141  80 101 125  96  99
##  [18]  96  50  99  88 120  93  78 100 105  87  94  85  94 107  72  97 101
##  [35]  89 104  72  90  92  86  79  83 100  93  91  98  46  85 107 104  86
##  [52]  89  76  96 101 108 102  92  76  80  79  75  82  97  92  77 111  84
##  [69]  56  77  80  86  88  96  96 107  86 107  91  99 115 106 105  85  87
##  [86]  98  89  80 111 104  75  73  76  88  89  96  76  82  93  85  75  85
## [103]  80  80  94  88 104  88  88 112  83 101  92  71 114  91  85  76  95
## [120]  77  74  96  91  78
```

With this information you can use these new functions of min(), max(), mean(), and sd() to find the minimum, maximum, average, and standard deviation respectively.

With this new information can you find solutions for problems 2 and 3 from above without using the summary function?

Estimate Vs Parameter

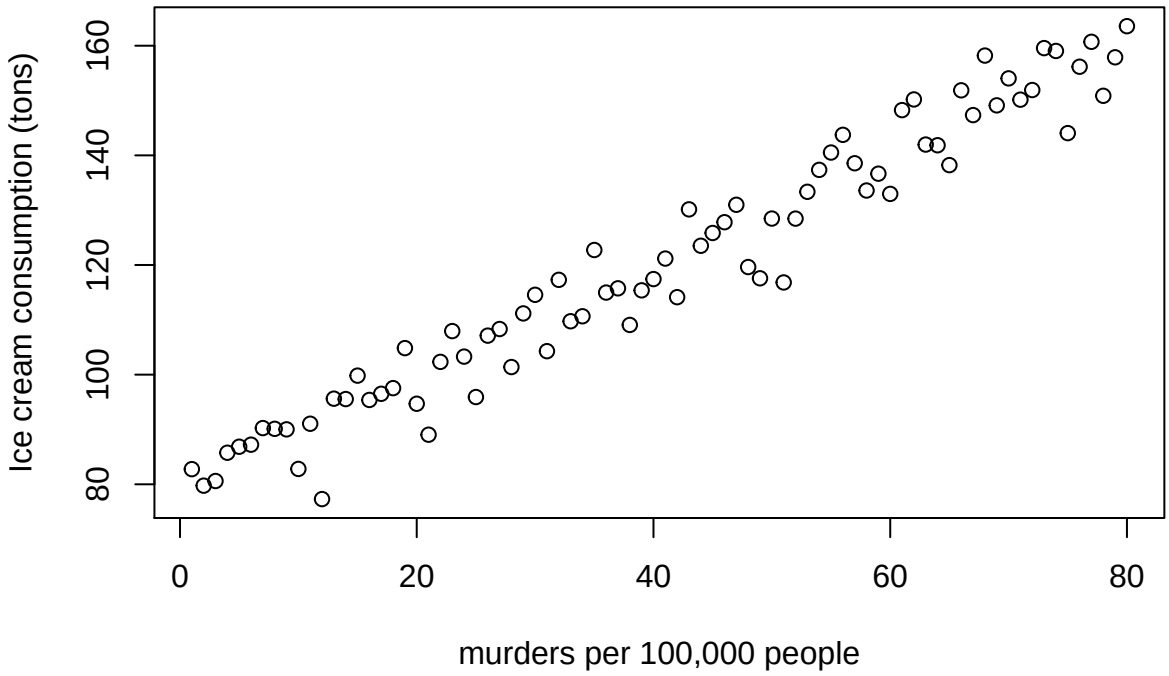
To reiterate what was explained in class a PARAMETER is a population quantity and is conventionally denoted with a greek letter such as π which represents the population proportion, μ which represents the mean of the population, and σ which represents the standard deviation of the population.

ESTIMATES are quantities derived from the sample (actual data) that we use to ESTIMATE (whoa, creative naming) the population quantities. Symbols that you will see that represent ESTIMATES are $\widehat{\pi}$ which represents the sample proportion, $\widehat{\mu}$ which is the sample mean, and $\widehat{\sigma}$ which is the sample standard deviation. Others you will see throughout the semester include p, $\bar{x}$, and s.

Confounding Factors and Bias

You may be familiar with the saying "Correlation does not imply causation". This is the case many times because of confounding factors. The best way to dismiss confounding factors is by a using a randomization process.

Example 1: Ice cream and murders It has been shown that ice cream sales are highly positively correlated with how many murdersoccur in a given month, that is, as ice cream sales increase, murders increase as well. Consider the plot below. *this data is not real



1) Why might this be?

2) Can you think of a way we may be able to solve whether or not ice cream consumption causes more homicides using the randomization process? (There are many possible answers here)

Hypothesis Testing - "Null Until Proven Alternative"

In class, you learned that there are a lot of wrong ways to think about p-values. The courtroom is a helpful example that illustrates the correct usage of p-values and hypothesis tests. Look at it in terms of "innocent until proven guilty": As the person analyzing data, you are the judge. The hypothesis test is the trial, and the null hypothesis is the defendant. The alternative hypothesis is like the prosecution, which needs to make its case beyond a reasonable doubt (say, with 95% certainty).

If the evidence presented doesn't prove the defendant is guilty beyond a reasonable doubt, you still have not proved that the defendant is innocent. But based on the evidence, you can't reject that possibility.

So how would that verdict be announced? It enters the court record as "Not guilty." That phrase is perfect: "Not guilty" doesn't mean the defendant is innocent, because that has not been proven. It just means the prosecution couldn't prove its case to the necessary, "beyond a reasonable doubt" standard. It failed to convince the judge to abandon the assumption of innocence.

If you follow that rationale, then you can see that "failure to reject the null" is just the statistical equivalent of "not guilty." In a trial, the burden of proof falls to the prosecution. When analyzing data, the entire burden of proof falls to the sample data you've collected. Just as "not guilty" is not the same thing as "innocent," neither is "failing to reject" the same as "accepting" the null hypothesis.

This method of thinking about hypothesis tests will come in handy when we start formally testing our own hypotheses.

Source: http://blog.minitab.com/blog/understanding-statistics/things-statisticians-say-failure-to-reject-the-null-hypothesis