# Module 4 Lab

## Loading Data

In this lab we will read in data, calculate a Z-statistic, obtain a p-value, and write a technical summary of our findings.

First we read in data. We will use umpire data from *The Statistical Sleuth*. Therefore, we load the "Sleuth3" package first.

```
library("Sleuth3")
```

The umpire data set is stored as `ex0321`, so we name it `umpire` for convenience. Use the command `?ex0321` to learn more about the data.

```
umpire <- ex0321 # pg 77 in Sleuth (2nd Ed.)
```

Alternatively, in general, you may have data stored in a .csv file. This stands for "comma separated values," and is a common data file format. If this were the case, you would read in data with the following commands.

```
ex0321 <- read.csv("/Users/ABC/Desktop/ex0321.csv") # specify file path
```

The code inside the parentheses gives the file path to tell R where the file is stored on your computer; you will need to change this portion if you use this line of code. Use `?read.csv` to learn more about the `read.csv()` command.

One final option we present, for a file stored on your computer, is to click the "Import Dataset" button toward the top right in RStudio—in the "Environment" pane. Then choose "From Text file". This will allow you to manually navigate to the file location.

## Test Statistics and p-values

Now that we have loaded our data, and named it `umpire`, let's look at it. The command `head()` returns the first few rows of our data frame, including column names. Use `?head` to learn more about this function.

```
head(umpire) # Take a look at data
```

```
##   Lifelength Censored Expected
## 1         63        0       70
## 2         69        0       71
## 3         58        0       71
## 4         61        1       70
## 5         70        0       70
## 6         68        0       69
```

This data set contains the age at which umpires died, to determine if stress is shortening umpire lifespans. It also contains the age of living umpires, but let's remove those to keep our analysis simple. We use the `subset()` command to do so.

```
umpire <- subset(umpire, Censored == 0) # Remove umpires still alive
head(umpire)
```

```
##   Lifelength Censored Expected
## 1         63        0       70
## 2         69        0       71
## 3         58        0       71
## 5         70        0       70
## 6         68        0       69
## 7         72        0       68
```

I specified "`Censored == 0`" to tell R and `subset()` to retain the rows of `umpire` where the "Censored" variable/column is equal to zero. Notice in the first few rows of the `umpire` data set that the `Censored` column is now all zeroes.

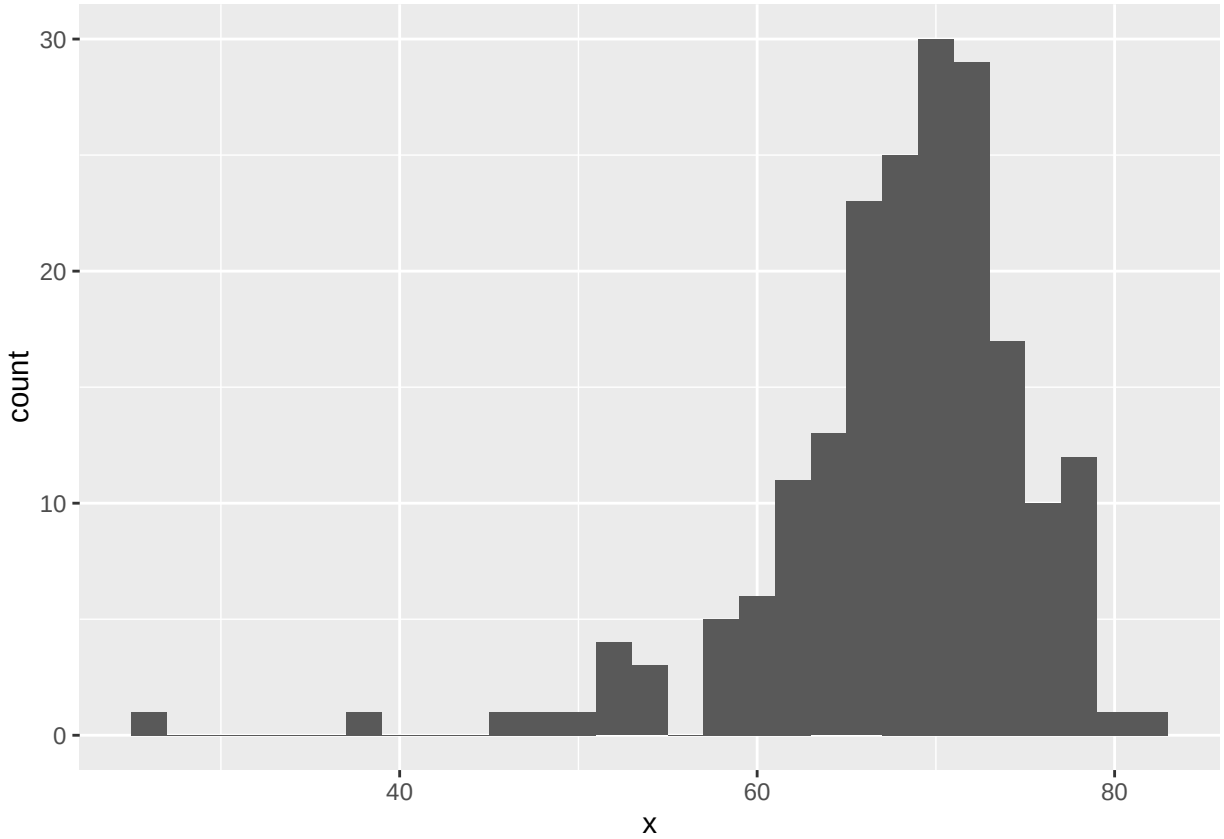
We want to analyze umpire lifespans, so let's define `x` to be the umpire `Lifelength`.

```
x <- umpire[,1] # Define x as *all rows, column 1* of "umpire"
```

This subsetting format specifies "all rows" with a comma as the first entry in the brackets, and column one with the 1 as the second entry.

Now we load the `ggplot2` package, and take a look at a histogram of our upire life spans to get a sense of our data.

```
library("ggplot2")
qplot(x, binwidth = 2)
```



The umpire lifespans are left skewed, with a mean somewhere in the 60s or low 70s. Now we will conduct a simple hypothesis test, where the null hypothesis is that the mean age of death of umpires is 69.5, and the alternative hypothesis is that mean umpire age of death is less than 69.5.

$$H_0 : \mu = 69.5$$
$$H_A : \mu < 69.5$$

With a sample size of 195, we know $\overline{x}$ is Normally distributed. Therefore, under $H_0$:

$$\frac{\overline{x} - \mu}{s/\sqrt{n}} \sim N(0, 1)$$

For more details, see Module 4 lectures, or **Section 4.3.4** of *OpenIntro Statistics*. For now, we focus on the calculating the test statistic in R, and getting a p-value. We will use the `pnorm()` fucction to get a p-value. If you provide `pnrom()` with a quantile (our test statistic) it will return a probability.

```
# pg 178, OpenIntro
Z <- (mean(x) - 69.5)/(sd(x)/sqrt(195)) # Calculate z-statistic
Z
```

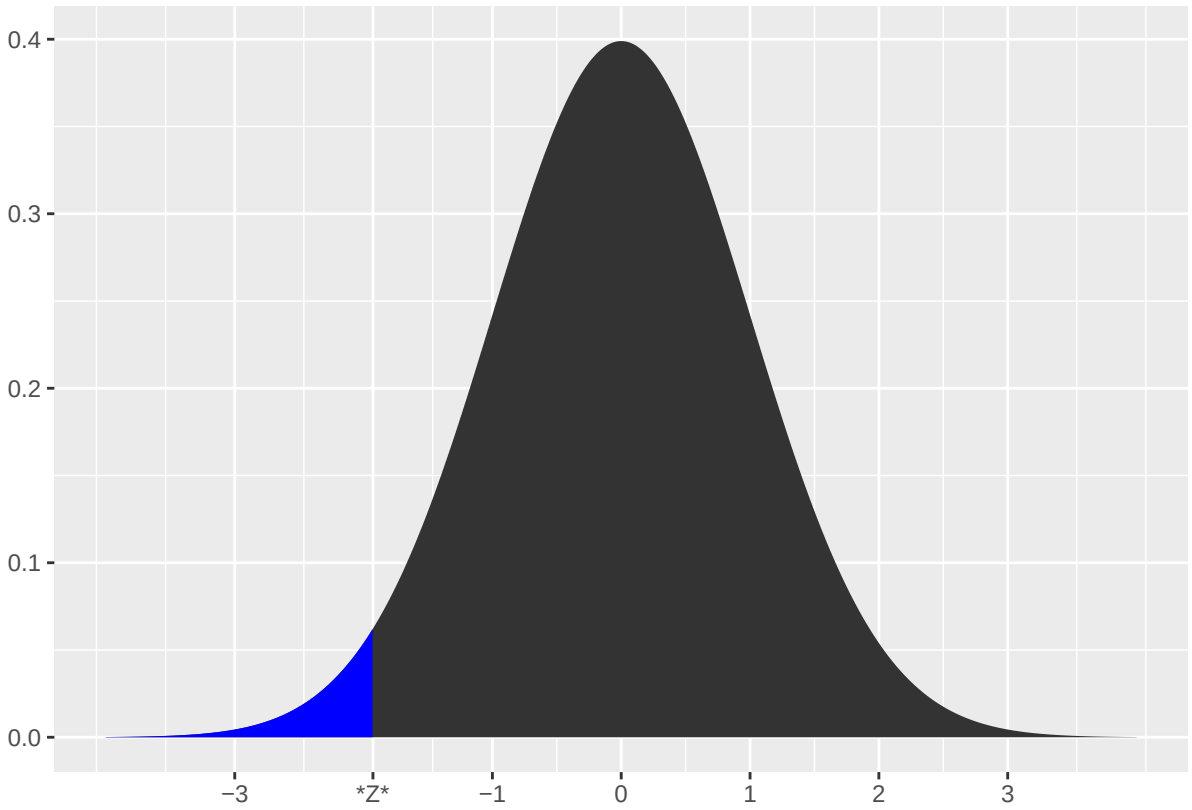
```
## [1] -1.927416
```

```
pnorm(Z, mean = 0, sd = 1, lower.tail = TRUE)
```

```
## [1] 0.02696391
```

The argument `lower.tail = TRUE` tells `pnorm()` to return the probability that a N(0,1) random variable is less than Z. Use `?pnorm()` to learn more about the function.

So what does this tell us? This tells us that if the null hypothesis ($H_0$) is true, the probability of observing a sample mean as unlikely (in this case–as small), or more unlikely (in this case–or smaller), than what we actually observed is 0.02696. That is pretty unlikely! How about a picture.




This is a N(0,1) density curve, with a blue shaded region representing the probability of observing a sample mean as unlikely as, or more unlikely than, what we actually observed.

## Technical Summary

Now we need to present our findings. As with most things, there are a few right ways, and many wrong ways to do this. Here is one example of the right way.

- There is strong evidence to suggest that the mean umpire age of death is less than 69.5 years old. The one sided p-value from a z-test, with a sample size of 195, is 0.02696. The probability of observing a sample mean less than or equal to 68.48, from a sample of size 195, is 0.02696.

Now for a few wrong ways. Here are some incorrect statements sometimes included in technical summaries.

- We conclude that the mean umpire age of death is 68.48 years old.
- The probability our conclusion (reject the null) is incorrect is 0.02696.
- We conclude the null hypothesis is false, based on a p-value of 0.02696.

Can you figure out what is wrong with each statement?