Lsn 20 - MA206Y

Clark

Admin

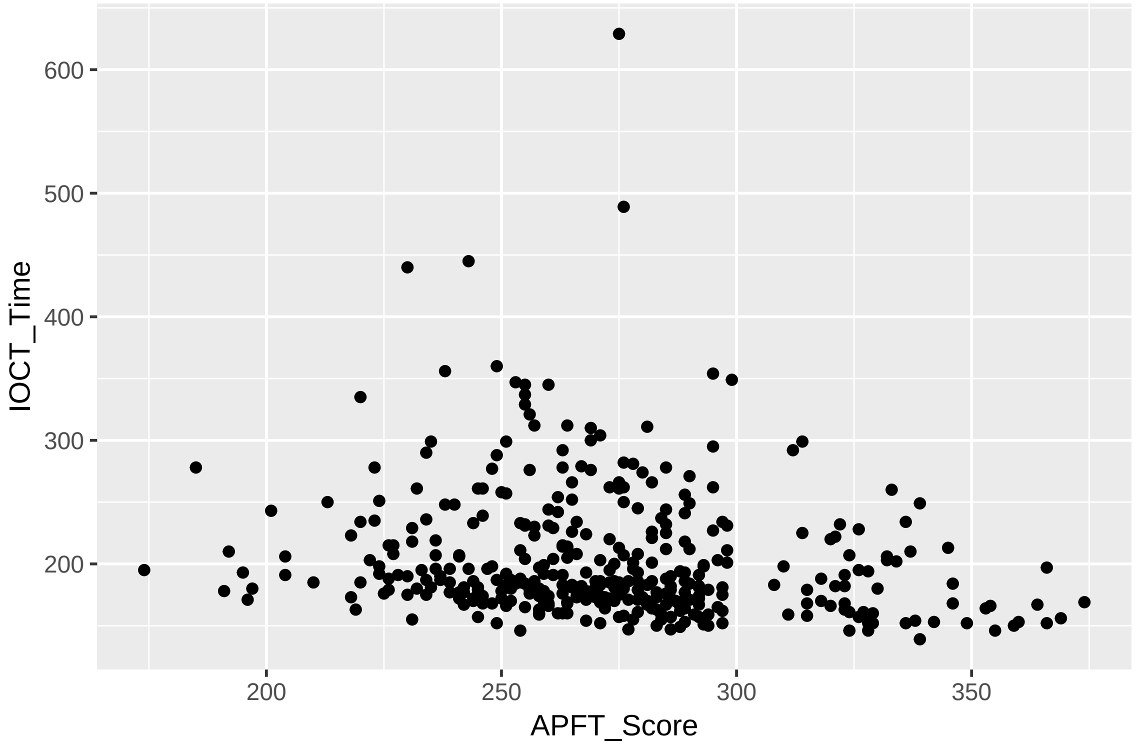
Over the next several lessons we are going to build up regression models. The basic idea is we want to determine if there is a relationship between two quantitative variables. Note here we're not talking about causality, but rather just asking the question, as one quantitative variable changes, do we expect, on average, to observe a change in another quantitative variable.

For example, we might have Cadets APFT and IOCT data. Prior to looking at the data, what would we expect the relationship between APFT score and IOCT time to be?

```r
dat=read.csv("APFT.csv")
```

One way to explore the data is to examine a **scatter plot**. Here we can show:

```r
dat %>% ggplot(aes(x=APFT_Score,y=IOCT_Time))+geom_point()
```



Is there anything unexpected in the data? Are there any unusual observations?

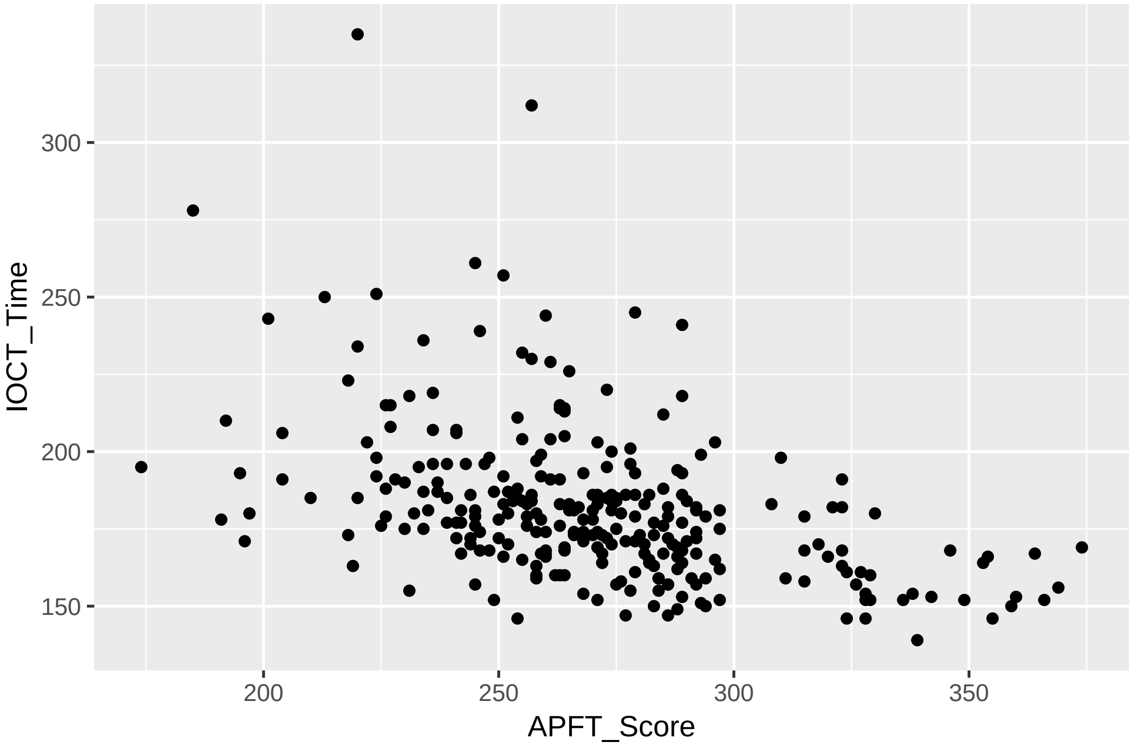
What is the direction of the relationshp?

Does the data appear to be linear?

How strong is the relationship? Are you surprised?

What about if we only look at Males?

```
males= dat %>% filter(sex=="M")
males %>%ggplot(aes(x=APFT_Score,y=IOCT_Time))+geom_point()
```



Does the relationship change? What about the strength of relationship?

One way to quantify the relationship is to calculate the **correlation coefficient** between the two variables. In R this can be done using `cor`

```
cor(males$APFT_Score,males$IOCT_Time)
```

```
## [1] -0.4608226
```

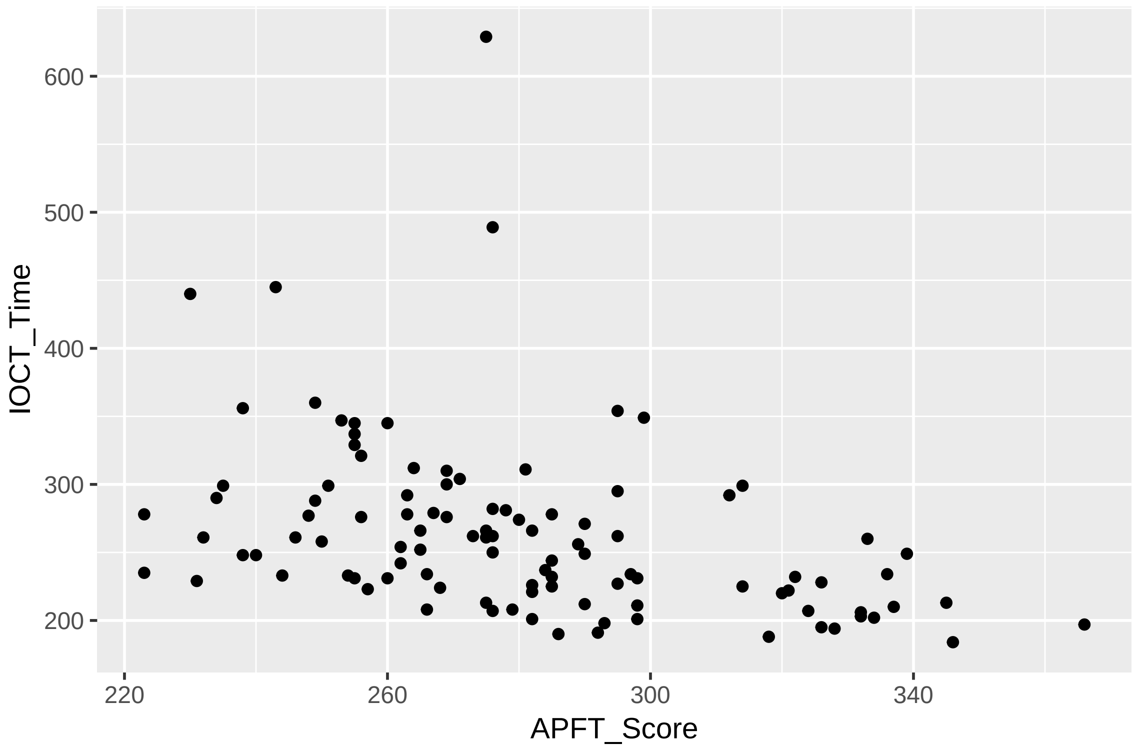
We can do the same thing with just females

```
females= dat %>% filter(sex=="F")
cor(females$APFT_Score,females$IOCT_Time)
```

```
## [1] -0.3953225
```

Does there appear to be a significnat difference in the correlation coefficient? Let's look at the female data

```
females %>%ggplot(aes(x=APFT_Score,y=IOCT_Time))+geom_point()
```



Let's see what happens to the correlation coefficient if we remove the unusual observations (maybe everything over 400 seconds?)

```
mod.females<-females %>% filter(IOCT_Time<400)
cor(mod.females$APFT_Score,mod.females$IOCT_Time)
```

```
## [1] -0.4720926
```

Does the relationshp strengthen or weaken?

While this seems like there's a relationship between IOCT Time and APFT Score for our females, does this prove that there's a relationship?

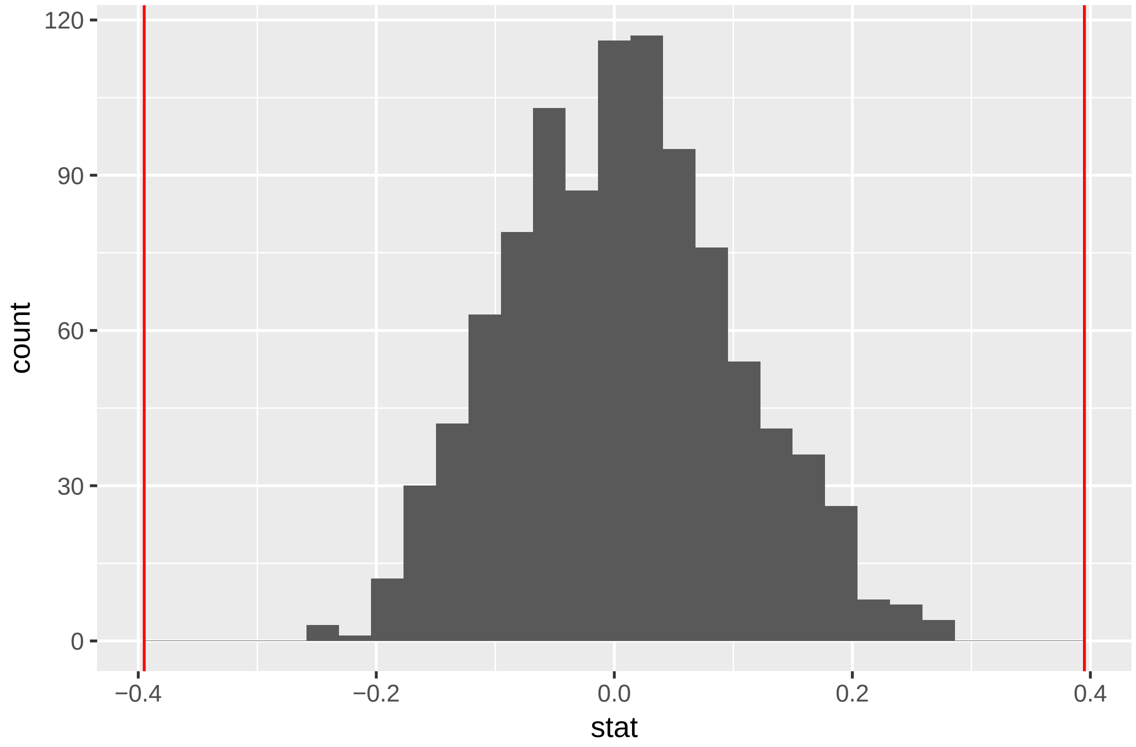
In order to do this, we need to define a statistic to test the null and alternative hypothesis:

One statistic we could use is the correlation coefficient, r. Keeping our unusual observations in our study, our realization of r is -0.395. Under H_0 what are we saying? In order to test our null and alternative hypothesis we need to get a feel for the distribution of r under H_0. If we don't know the name of the distribution we can simulate it:

```
M=1000
stat.df=data.frame(trial=seq(1,M),stat=NA)
for(i in 1:M){
  females.shuff<-females %>% mutate(shuff.IOCT=sample(females$IOCT_Time))
  stat.df[i,]$stat=cor(females.shuff$APFT_Score,females.shuff$shuff.IOCT)
}

stat.df %>% ggplot(aes(x=stat))+geom_histogram()+
  geom_vline(xintercept=0.395,color="red")+
  geom_vline(xintercept=-0.395,color="red")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Which, as previously done, we want to see how rare our statistic is, so we want the area to the right and left of the red lines, which is obviously pretty rare…

Why does it make sense that this is centered at 0?

We can do the same thing in the applet.

Let's work through the Draft Lottery