Class Example 1 – Single Sample Descriptive Statistics

(a) **Summary Statistics and Box-Plots**
You are working in factory producing hand held bicycle pumps and obtain a sample of 174 bicycle pump weights in grams, as in the data file *"Pumps.csv."* Carry out descriptive statistics as follows:

(i) Load the data file.

```
> library(tidyverse)
> pumps <- read.csv("Pumps.csv")
```

(ii) Output the sample size, sample mean and sample standard deviation.

```
> c(length(pumps$weight),mean(pumps$weight),sd(pumps$weight))
```

(iii) View basic summary statistics.

```
> summary(pumps$weight)
```

(iv) Create a box-plot of the data and comment on the structure of the data.

```
> ggplot(pumps,aes(y=weight))+geom_boxplot()+theme_classic()
```

(v) Present the summary statistics and the box-plot of the data in a neatly formatted RMarkdown document describing the data set.

Summary Statistics and a box-plot of 174 pump weights

A Student

05/04/2019

```
library(tidyverse)
pumps = read_csv("Pumps.csv")
```

```
## Parsed with column specification:
## cols(
##   weight = col_double()
## )
```

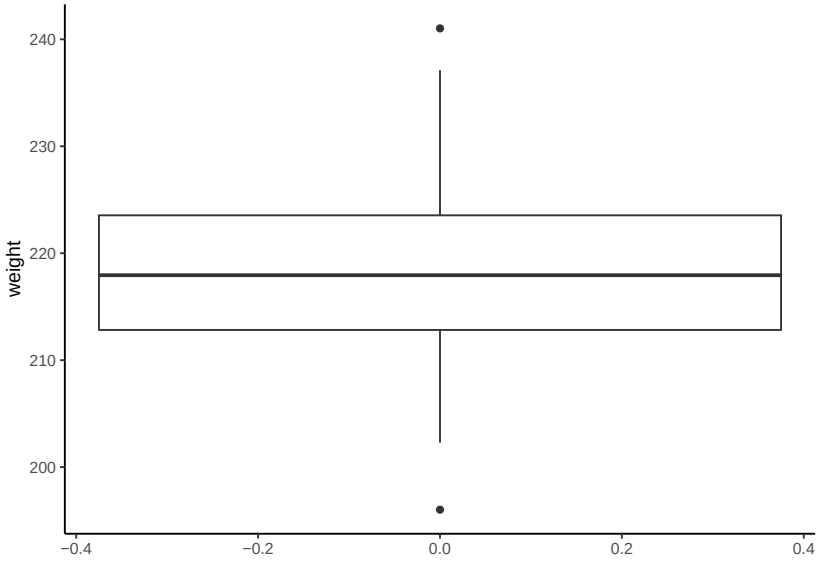
```
c(length(pumps$weight),mean(pumps$weight),sd(pumps$weight))
```

```
## [1] 174.00000 218.11448   7.71757
```

```
summary(pumps$weight)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   196.0   212.8   217.9   218.1   223.5   241.0
```

```
ggplot(pumps,aes(y=weight))+geom_boxplot()+theme_classic()
```



A brief summary of the data As can be seen the mean pump weight is at 218 grams, and the standard deviation is at 7.7 grams. The data appears to be symmetric with only two noted outliers out of 174 observations

(b) **Empirical Cumulative Distribution Function**

i Key in the code below and run it.

```
> numSamples <- 15
>
> data <- rnorm(numSamples,20,5)
> estimateCDF=ecdf(data)
>
> grid <- seq(0,40,by=0.1)
> cdf <- pnorm(grid,20,5)
```

```
>
> ggplot() + geom_line(aes(x=grid,y=cdf),colour="orange") +
+ stat_ecdf(aes(data),colour="blue")+theme_classic()
```

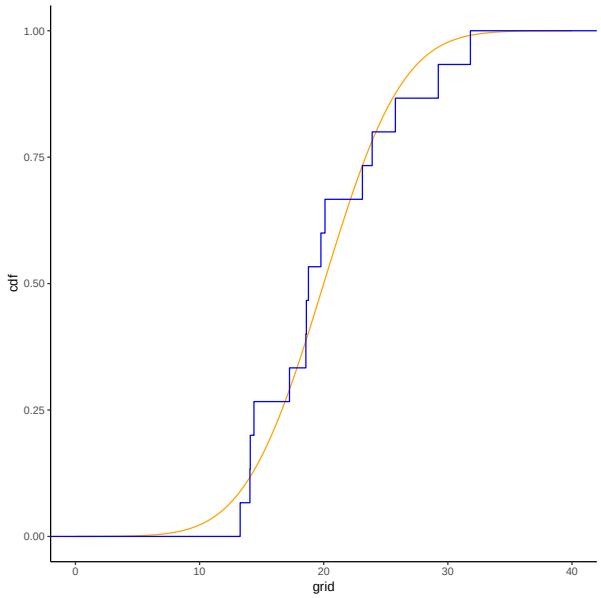


Figure 1: The resulting CDF of a theoretical Normal distribution with $\mu = 20$ and $\sigma = 5$ plotted against an ECDF of randomly generated samples from the same distribution.

ii Now modify the number of samples seeing how the ECDF converges to the true CDF as the number of samples increases.

iii Change the distribution to a distribution of your choice, e.g. Exponential or Uniform. Now run again to view the resulting graphs. Note that you may need to change the values of grid, appropriately.

(c) **Histograms and Kernel Density Estimation.**
The code below generates random variables from a *mixture of two normal distributions* with pdf represented by the blue curve below (left). It then plots a kernel density estimate of the data (red) and a histogram of the data.

Key in this code and experiment with increasing the number of samples. You may also try changing other parameters such as the means, variances and p.

```
> library(gridExtra)


Attaching package: 'gridExtra'

The following object is masked from 'package:dplyr':

    combine

> mu1 <- 10
> mu2 <- 40
> sigma1 <- 5
> sigma2 <- 12
>
```

```
> p <- 0.3
>
> grid <- seq(-20,80,by=0.1)
>
> actualPDF <- p*dnorm(grid,mean=mu1,sd=sigma1)+(1-p)*
+ dnorm(grid,mean=mu2,sd=sigma2)
>
> numSamples = 100
>
> # mixing function for mixed random variable
> mixRV <- function() {
+         ind = runif(1) <= p
+         if(ind) {
+                 return(rnorm(1,mean=mu1,sd=sigma1))
+         } else {
+                 return(rnorm(1,mean=mu2,sd=sigma2))
+         }
+ }
>
> data <- replicate(numSamples,mixRV())
> p1 <- ggplot()+geom_line(aes(x=grid,y=actualPDF),colour="blue")+
+ geom_density(aes(data),colour="red",n=512,trim=FALSE)+
+ theme_classic()
> p2 <- ggplot()+geom_histogram(aes(data),bins=15) + theme_classic()
> grid.arrange(p1,p2,ncol=2)
```

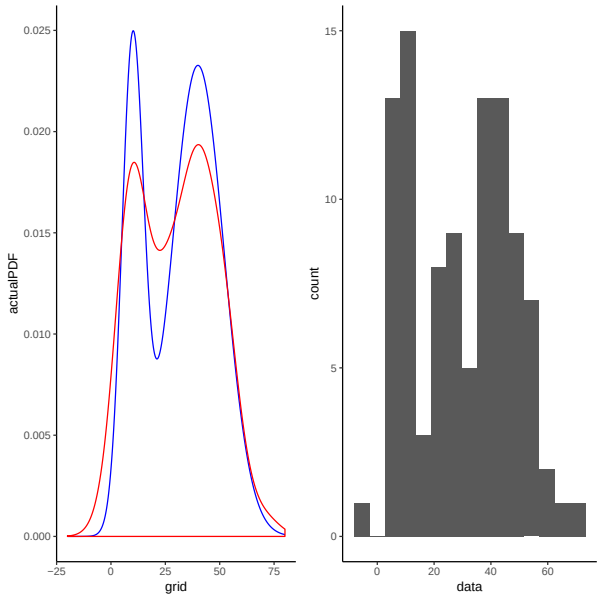


Figure 2: On left: Blue is theoretical pdf and red is a kernel density estimate (depends on data). On right: A simple histogram.

Class Example 2 – Statistical Inference Ideas.

A farmer wanted to test whether a new fertilizer was effective at increasing the yield of her tomato plants. She took 20 plants, kept 10 as controls and treated the remaining 10 with the new fertilizer. After two months, she harvested the plants and recorded the yield of each plant

(in kg), as shown in the following table:

Control	4.17	5.58	5.18	6.11	4.5	4.61	5.17	4.53	5.33	5.14
Fertilizer	6.31	5.12	5.54	5.5	5.37	5.29	4.92	6.15	5.8	5.26

From this data, the group of plants treated with the fertilizer had an average yield of 0.494 kg greater than the control group. One could argue that this difference is due to the effects of fertilizer. We will now investigate if this is a reasonable assumption.

Let us assume for a moment that the fertilizer has no effect on plant yield (it is a placebo/has no added nutrients). In such a scenario, we actually have 20 observations from the **same group** (non-nutrient enriched plants), and the average difference is purely the result of random chance.

We can investigate this, by taking all possible combinations of 10 samples from our group of 20 observations, and counting how many of these combinations results in a sample mean greater than the mean of our treatment group. Dividing this total by the total number of possible combinations, we can obtain a proportion, and hence likelihood, that the difference observed was due purely to random chance.

First calculate the number of ways of sampling $r = 10$ unique items from a total of $n = 20$. This is given by,

$$\binom{n}{r} = \frac{n!}{r!(n-r)!} = \binom{20}{10} = \frac{20!}{10!10!} = 1.84756 \times 10^5$$

This number of possible samples is computationally managable, and hence we'll use R's `combn` function to enumerate all 1.84756×10^5 combinations.

```
> tomato <- read.csv("Fertilizer.csv")
> tomatoes <- c(tomato$Control,tomato$FertilizerX)
>
> x <- combn(tomatoes,10)
>
> length(x)

[1] 1847560

> pvalue <- mean(colMeans(x) >= mean(tomato$FertilizerX))
> print(pvalue)

[1] 0.02416701
```

We can see that from this data, only 2.42% of all possible combinations have a mean greater or equal to our treated group. Therefore there is significant statistical evidence that the fertilizer increases the yield.