Unit 1 DDS Live Session Assignment

Dawson

May 5, 2020

```
sessionInfo()
```

```
## R version 3.5.2 (2018-12-20)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 18363)
## 
## Matrix products: default
## 
## locale:
## [1] LC_COLLATE=English_United States.1252 
## [2] LC_CTYPE=English_United States.1252   
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C                          
## [5] LC_TIME=English_United States.1252    
## 
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base     
## 
## loaded via a namespace (and not attached):
##  [1] compiler_3.5.2  magrittr_1.5    tools_3.5.2     htmltools_0.4.0
##  [5] yaml_2.2.0      Rcpp_1.0.1      stringi_1.4.3   rmarkdown_1.12 
##  [9] knitr_1.28      stringr_1.4.0   xfun_0.5        digest_0.6.18  
## [13] rlang_0.4.5     evaluate_0.13
```

```
#1. Make a bar plot for your data science profile : computer programming', 'math', 'statistic
s', 'machine learning', 'domain    expertise','communication and presentation skills', 'data
visualization'
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.5.3
```

```
categories = c('Coding', 'Math', 'Stats', 'ML', 'Expertise','Comm', 'Vis')
num_categories = c(1,2,3,4,5,6,7)
ranking = c(1, 4, 4, 2, 5, 3,5)
categories
```

```
## [1] "Coding"    "Math"      "Stats"     "ML"        "Expertise" "Comm"     
## [7] "Vis"
```

```
ranking
```

```
## [1] 1 4 4 2 5 3 5
```

```
class(categories)
```

```
## [1] "character"
```

```
class(ranking)
```

```
## [1] "numeric"
```

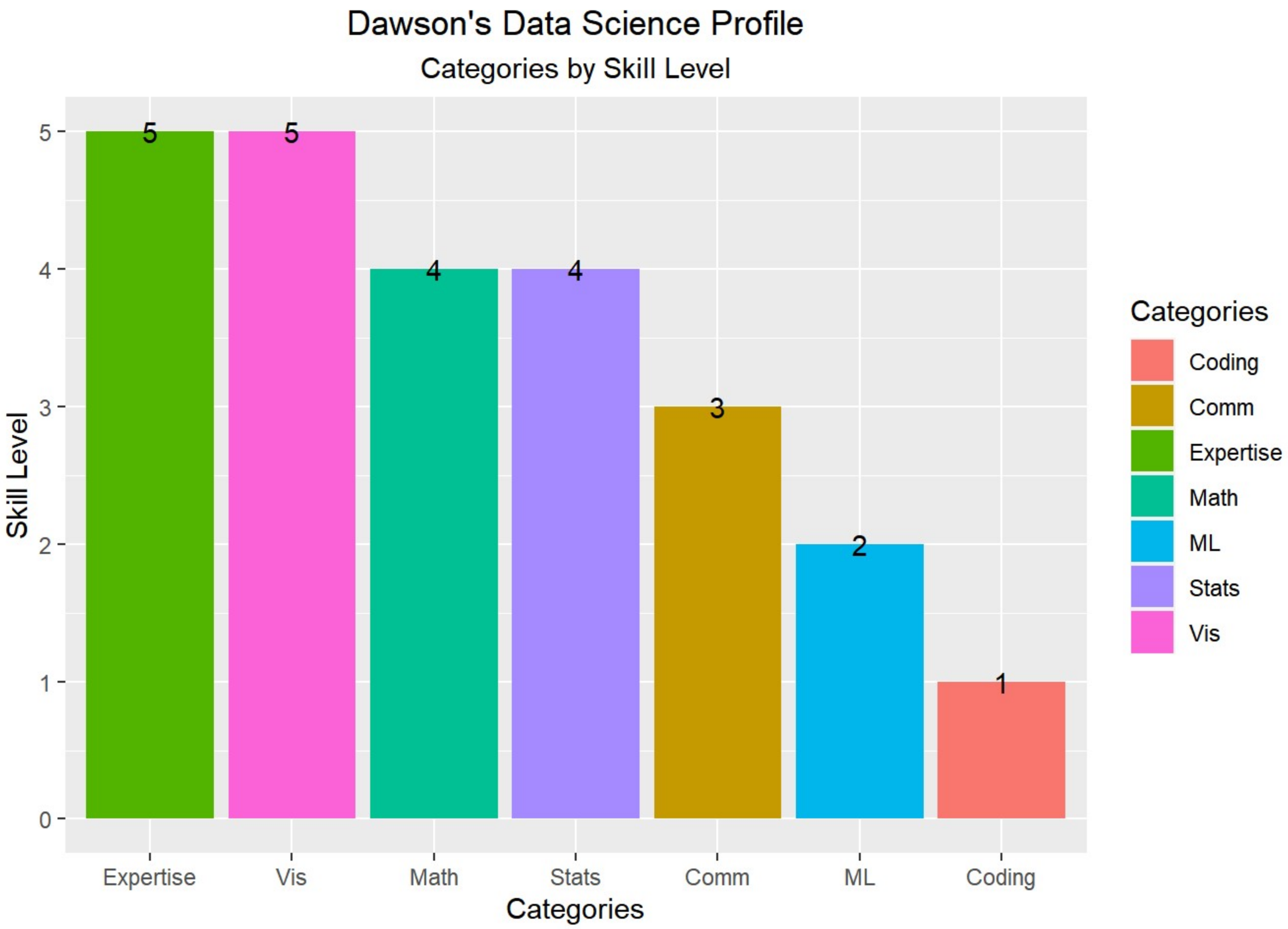
```
Dawson = data.frame(Categories = categories, Ranking = as.numeric(ranking), Num_Categories =
num_categories)
Dawson
```

```
##   Categories Ranking Num_Categories
## 1     Coding       1              1
## 2       Math       4              2
## 3      Stats       4              3
## 4         ML       2              4
## 5  Expertise       5              5
## 6       Comm       3              6
## 7        Vis       5              7
```

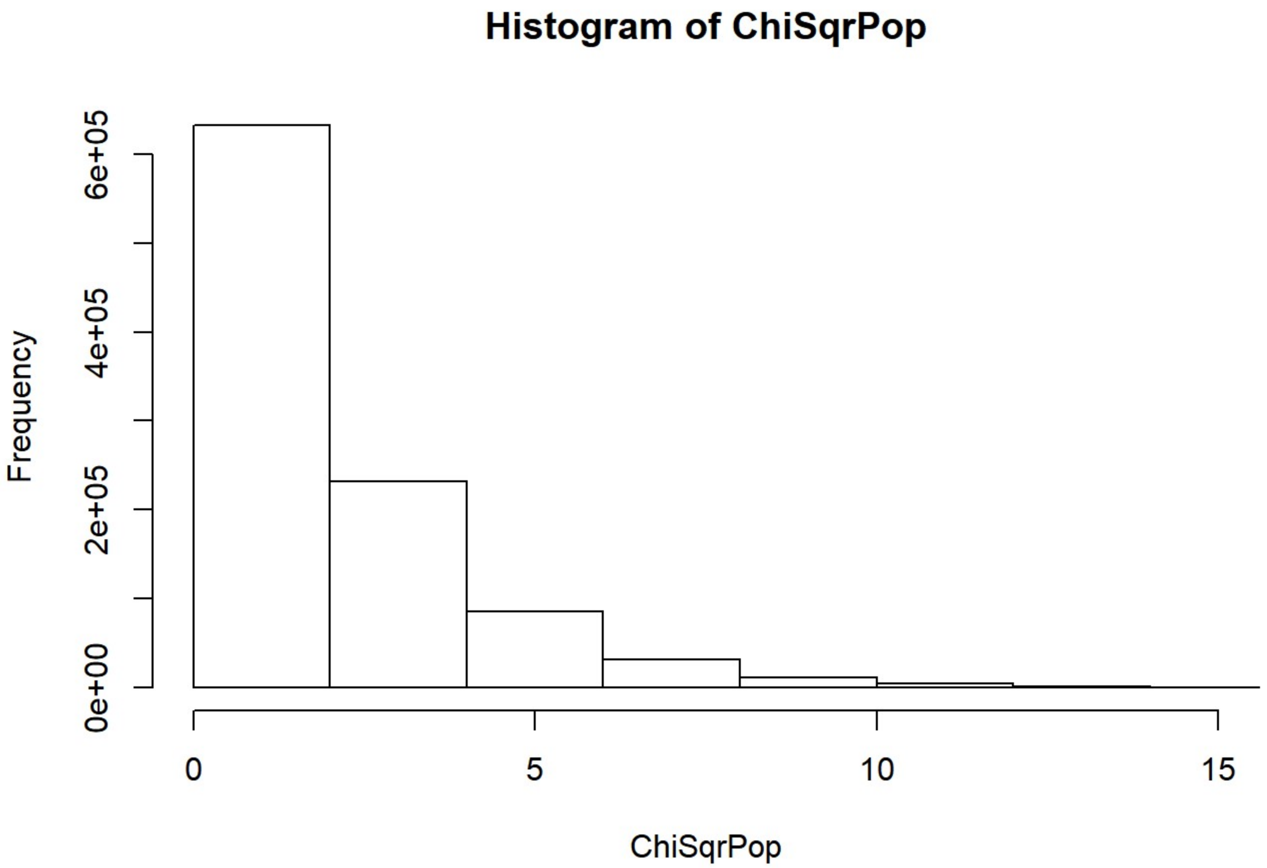
```
Dawson_Sorted = Dawson[order(-Dawson$Ranking),]
Dawson_Sorted
```

```
##   Categories Ranking Num_Categories
## 5  Expertise       5              5
## 7        Vis       5              7
## 2       Math       4              2
## 3      Stats       4              3
## 6       Comm       3              6
## 4         ML       2              4
## 1     Coding       1              1
```

```
#help("barplot")
ggplot(Dawson_Sorted, aes(x =reorder(Categories, -Ranking), y = Ranking)) + geom_col(aes(fill
= Categories)) + theme( plot.title = element_text(hjust = 0.5),
  plot.subtitle = element_text(hjust = 0.5)) + ggtitle("Dawson's Data Science Profile") + xla
b("Categories") + ylab("Skill Level") + geom_text(aes(label =Ranking)) + labs(subtitle = "Cat
egories by Skill Level")
```



```
#2. a) Adapt population 10M from chi-sqr distribution w/ 2 DoF (rchisq())
n = 1000000
ChiSqrPop = rchisq(n,2)
##b) Provide Hist
hist(ChiSqrPop, xlim =c(0,15))
```



```
##c) std & mean of this population
sd(ChiSqrPop)
```

```
## [1] 2.000106
```

```
mean(ChiSqrPop)
```

```
## [1] 2.000698
```

```
summary(ChiSqrPop)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.5754  1.3863  2.0007  2.7765 37.0018
```

```
##d) According to CLT, what should be the Approx sample means of sample size 50 from righ ske
w + What should be the mean & std error of the mean
### The mean & standard deviation should be both the same for the population and the sample
sd(ChiSqrPop)
```

```
## [1] 2.000106
```

```
mean(ChiSqrPop)
```

```
## [1] 2.000698
```

```
summary(ChiSqrPop)
```

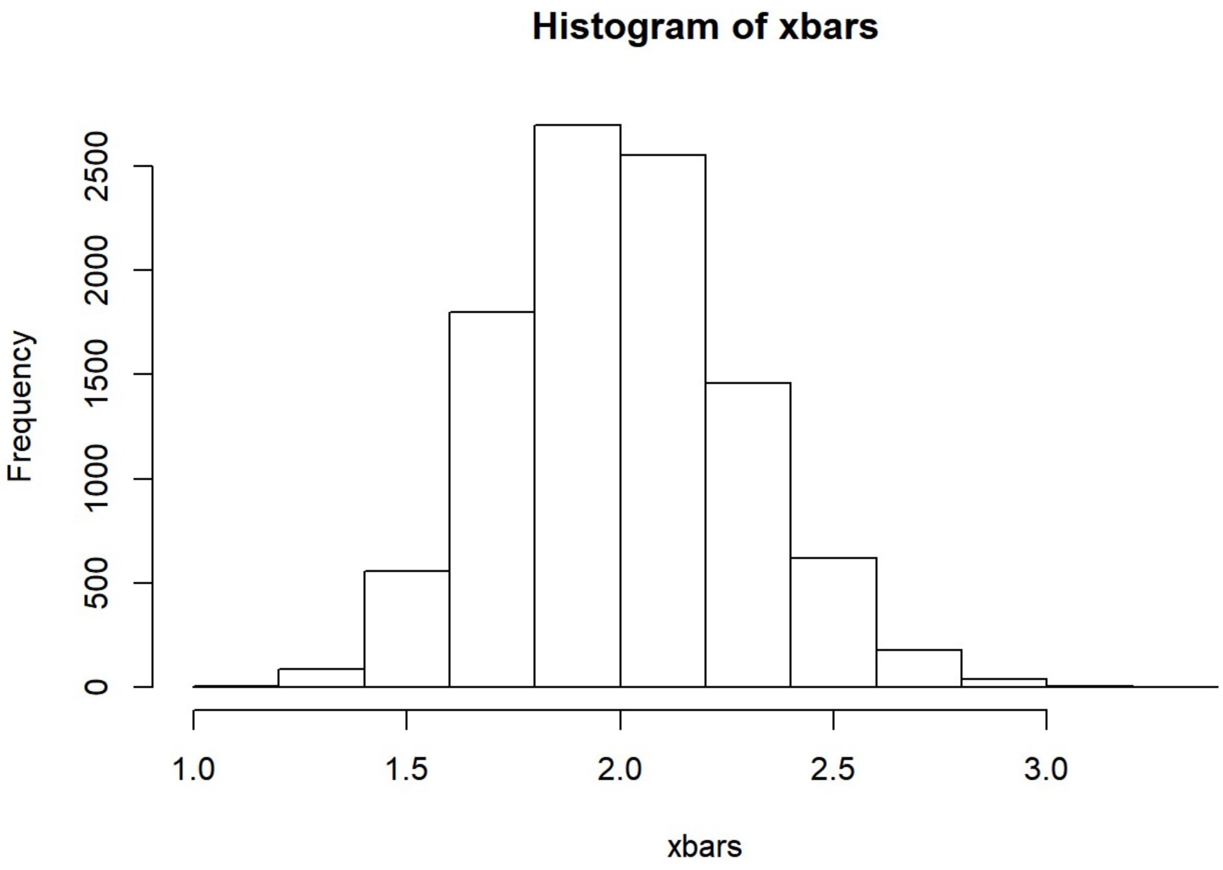
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.5754  1.3863  2.0007  2.7765 37.0018
```

```
##e/f) What is the mean & standard deviation of sample size = 50 from 10k # of samples?
xbarGenerator = function(sampleSize = 50,number_of_samples = 10000) {
    xBarVec = c()
    for(i in 1:number_of_samples) {
        theSample = sample(ChiSqrPop,sampleSize)
        xbar = mean(theSample)
        xBarVec = c(xBarVec, xbar)
    }
    return(xBarVec)
}

xbars = xbarGenerator(50,10000)
length(xbars)
```

```
## [1] 10000
```

```
#> [1] 1000
hist(xbars)
```



```
sd(xbars)
```

```
## [1] 0.2793164
```

```
mean(xbars)
```

```
## [1] 2.001019
```

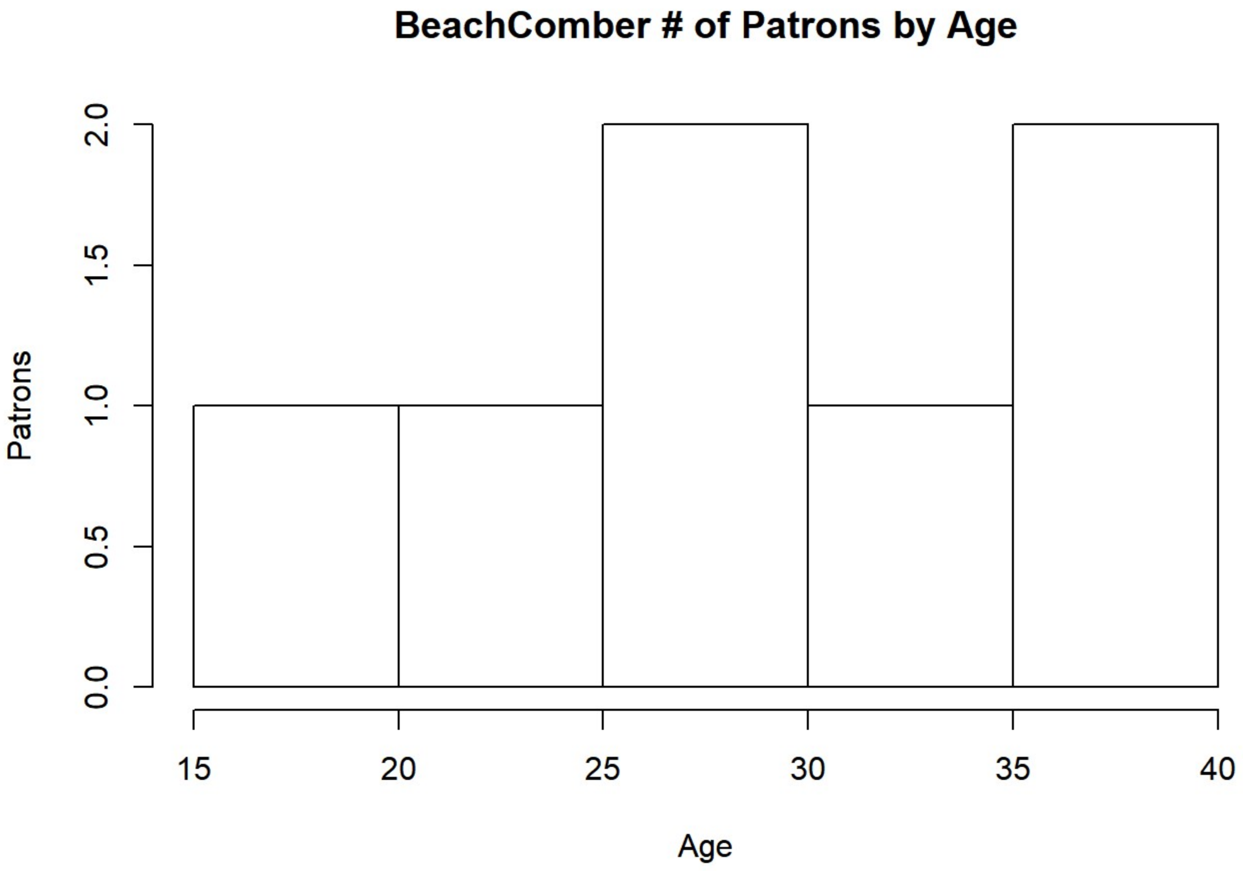
```
summary(xbars)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   1.067   1.804   1.990   2.001   2.182   3.215
```

```
#3. T-Test 6-step hypothesis test
## Test Mean Age is different from populaiton21. Assume Normal Distribution.
### Ho sample age (mu) = 21
### Ha sample age (mu) NE 21
PatronID = as.factor(c(1,2,3,4,5,6,7))
age = c(25, 19, 37, 29, 40, 28, 31)
BeachComber = data.frame(Patron = PatronID, Age = age)
BeachComber
```

```
##   Patron Age
## 1      1  25
## 2      2  19
## 3      3  37
## 4      4  29
## 5      5  40
## 6      6  28
## 7      7  31
```

```
hist(BeachComber$Age, main = "BeachComber # of Patrons by Age", xlab = "Age", ylab = "Patrons")
```



```
summary(BeachComber)
```

```
##  Patron      Age
##  1:1    Min.   :19.00
##  2:1    1st Qu.:26.50
##  3:1    Median :29.00
##  4:1    Mean   :29.86
##  5:1    3rd Qu.:34.00
##  6:1    Max.   :40.00
##  7:1
```

```
sd(BeachComber$Age)
```

```
## [1] 7.081162
```

```
t.test(BeachComber$Age, mu=21)
```

```
##
##  One Sample t-test
##
## data:  BeachComber$Age
## t = 3.3093, df = 6, p-value = 0.01622
## alternative hypothesis: true mean is not equal to 21
## 95 percent confidence interval:
##  23.30816 36.40613
## sample estimates:
## mean of x
##  29.85714
```

```
# Ho: mu=21
#Reject the null hypothesis. Beach Comber patrons mean age is different (p-value<0.01).
```

```
#4.
##Key Take-aways
#1) CLT population and sample mimic eachother statistically.
#2) Reproducibility: All code, figures, and dependencies outlined for variation of results, i
ncluding Session(); while all could be wrong
#3) Data is not objective. A model is an attempt to understand and represent the nature of re
ality through a lense(). Starting with EDA & known assumptions are key. Starting with basic s
ummary(), hist(), and box-plot-whisker() goes a long way fundamentally.
#4) Scientific Method: Question > background research > hypothesis > test with experiment, an
alysis & conclusion > Communication of Results.
#5 Data Science: Creating order from chaos. Ask questions.

##Questions
#1) Explain the relationship between Standard Deviation & Standard Error.
#2) How would you organize into your personal Github.
#3) Assume Normal distribution? Real data isn't this way, or is it?
```