DDS For Live Session

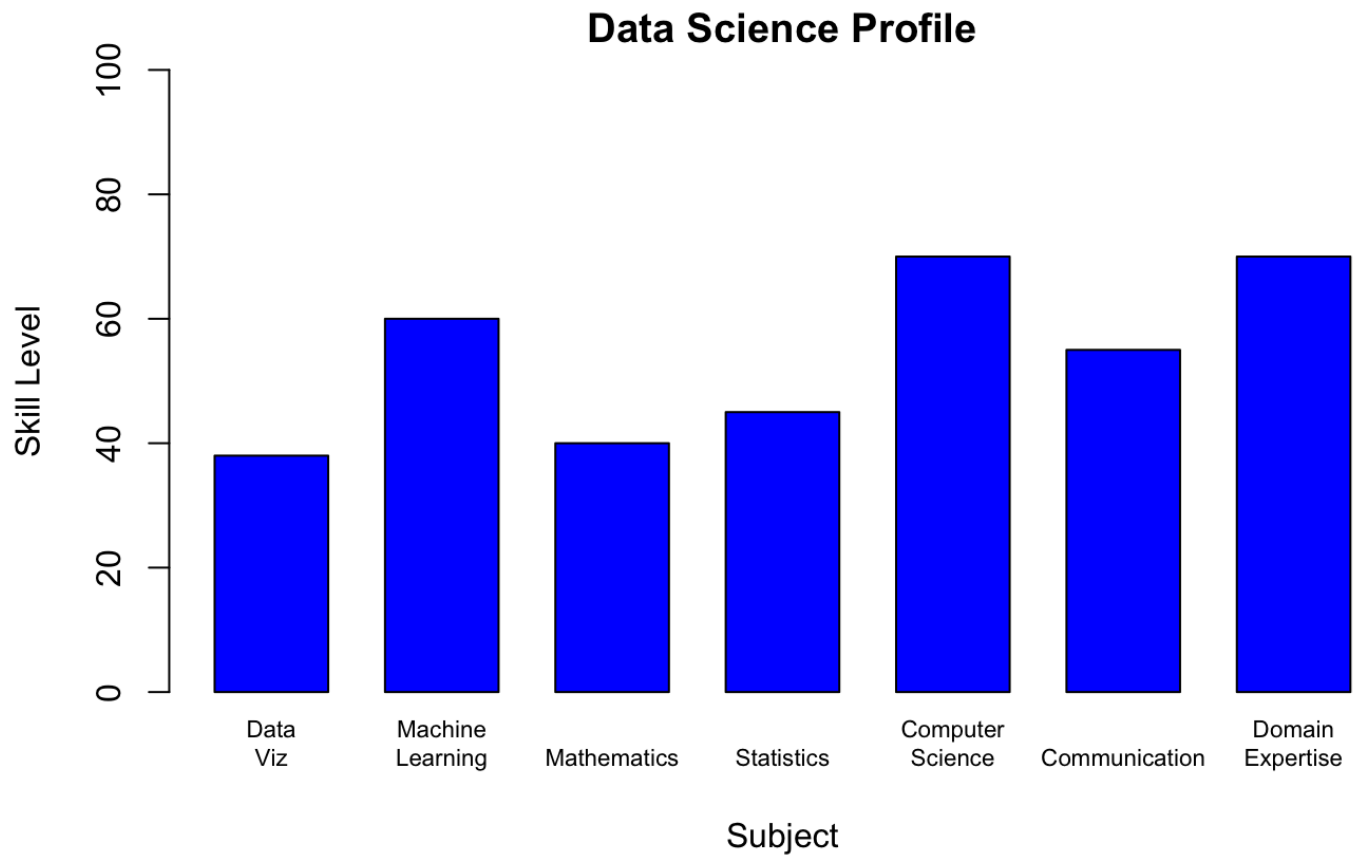
UNIT 1

David Grijalva

Data Science Profile

Code:

```
# Create arrays with subjects and skill levels
values = c(38, 60, 40, 45, 70, 55, 70)
area = c("Data\nViz", "Machine\nLearning", "Mathematics",
"Statistics",
        "Computer\nScience", "Communication",
"Domain\nExpertise")


dfp = data.frame(Area = area, Values = values)


# Build Barchart
barplot(dfp$Values, names.arg = dfp$Area, col="blue",
ylab="Skill Level",
       xlab ="Subject", main="Data Science Profile", space=0.5,
cex.names=0.70,
       ylim=c(0,100))

# Sample standard deviation
sd(ages)
```

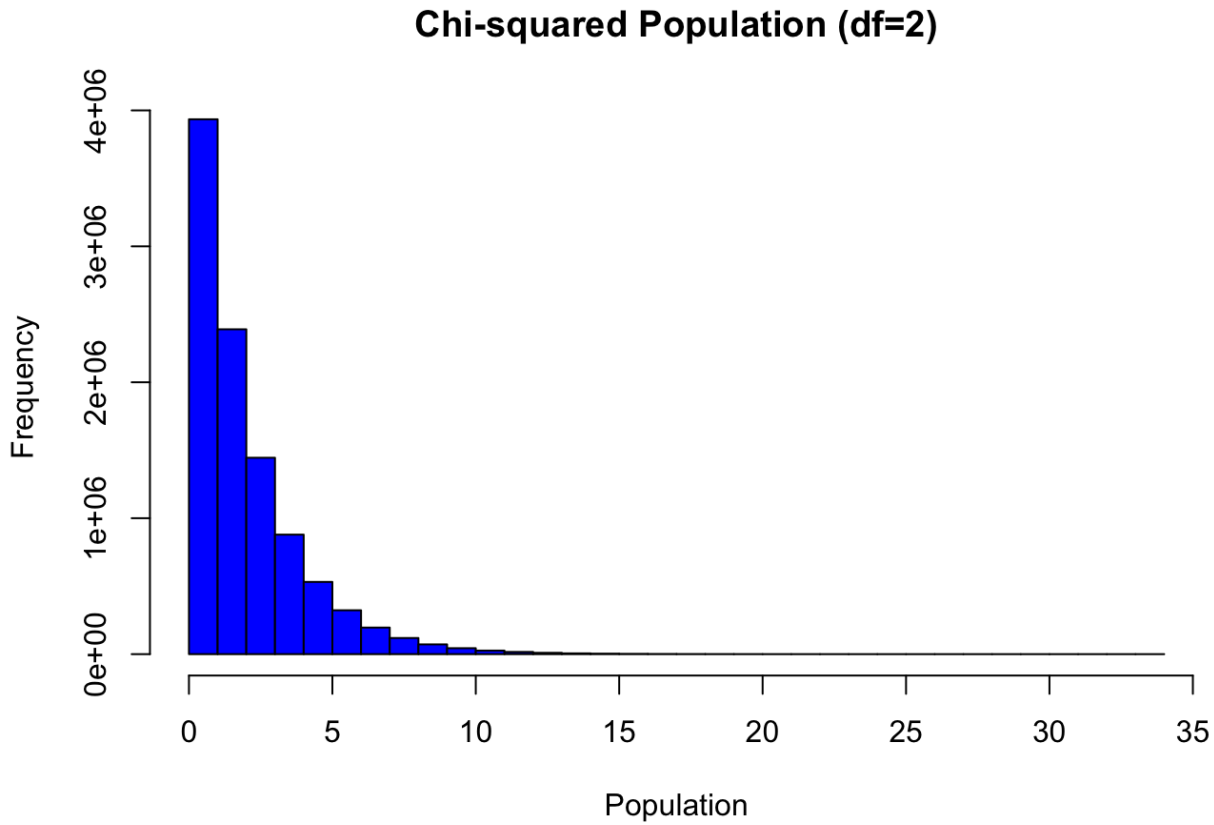


CLT

1. Adapt the code to generate a population of 10,000,000 from a chi-square distribution with 2 degrees of freedom. This is a heavily right skewed distribution.
2. Provide a histogram of this population… display the right skewness.

Code:

```
population = rchisq(10000000, 2)
 hist(population, main = "Chi-squared Population (df=2)", xlab="Population", col="blue")
```



CLT

3. Record the mean and standard deviation of this population.

Population Mean: 1.9998
Population Standard Deviation: 2.000214

Code:

```
# Population Mean
mean(population)

# Population Standard Deviation
sd(population)
```

CLT

4. According to the central limit theorem, what should be the approximate distribution of sample means of size 50 from this right skewed population? What should be the mean and standard error of the mean (standard deviation of the distribution of sample means)?

According to the CLT we expect that the sample means approximates to the one of the population and that the standard deviation, given a large enough sample will be smaller than the one from the original distribution.
The expected mean is 2.
The expected standard deviation is less than 2. I believe the number of samples should be large enough.

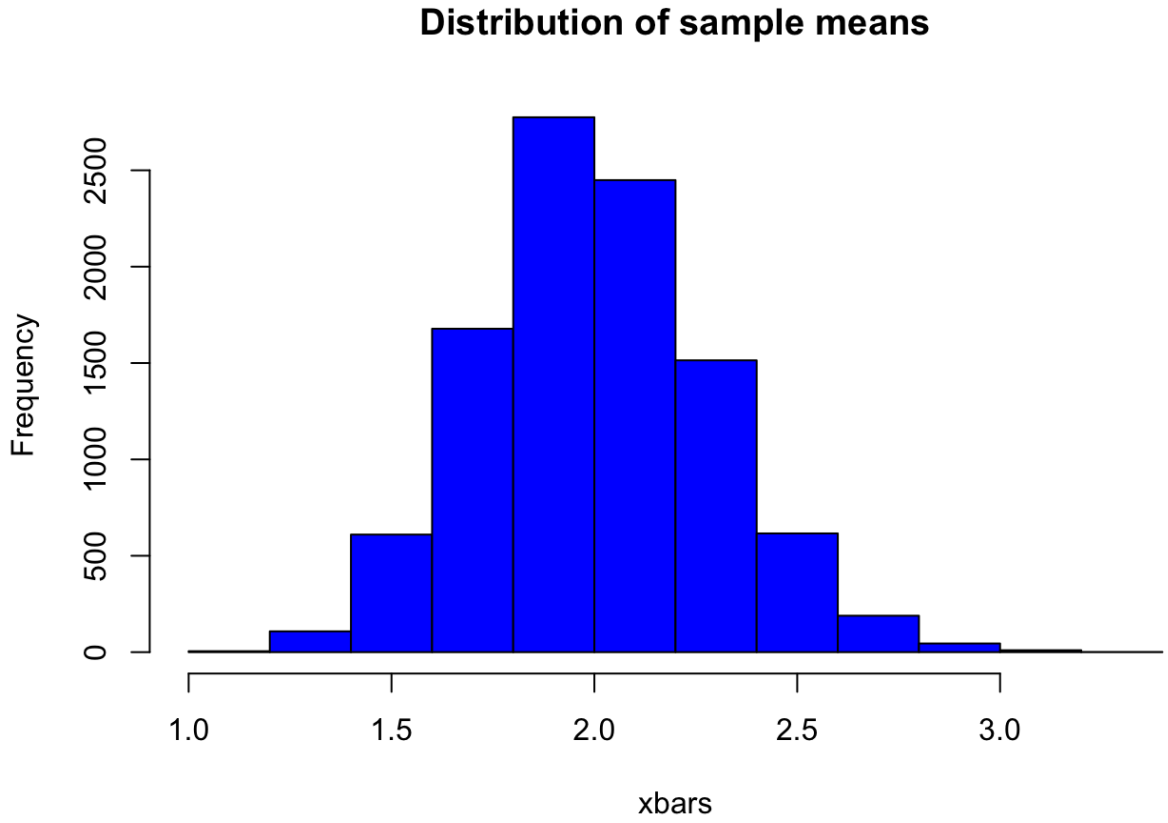
Since population distribution is not normal, we must use n>30, since we are using sample size of 50 we do meet this requirement.

CLT

5. Now let's check this: Adapt the CLT code to draw 10,000 means each of size 50 from this population and provide the sampling distribution of this sample mean. Provide a histogram of these 10,000 sample means.

Code:

```
xbarGenerator = function(sampleSize = 50,number_of_samples = 10000)
{
 for(i in 1:number_of_samples)
 {
  theSample = sample(population,sampleSize)
  xbar = mean(theSample)
  xBarVec = c(xBarVec, xbar)
 }
 return(xBarVec)
}

# Run function
xBarVec = c()
xbars = xbarGenerator(50,10000)
length(xbars)
hist(xbars, col="Blue", main="Distribution of sample means")
```



CLT

6. What is the mean and standard deviation of these 10,000 sample means?

Xbar Mean: 2.002255
Xbar Standard Deviation: 0.2847198

Code:

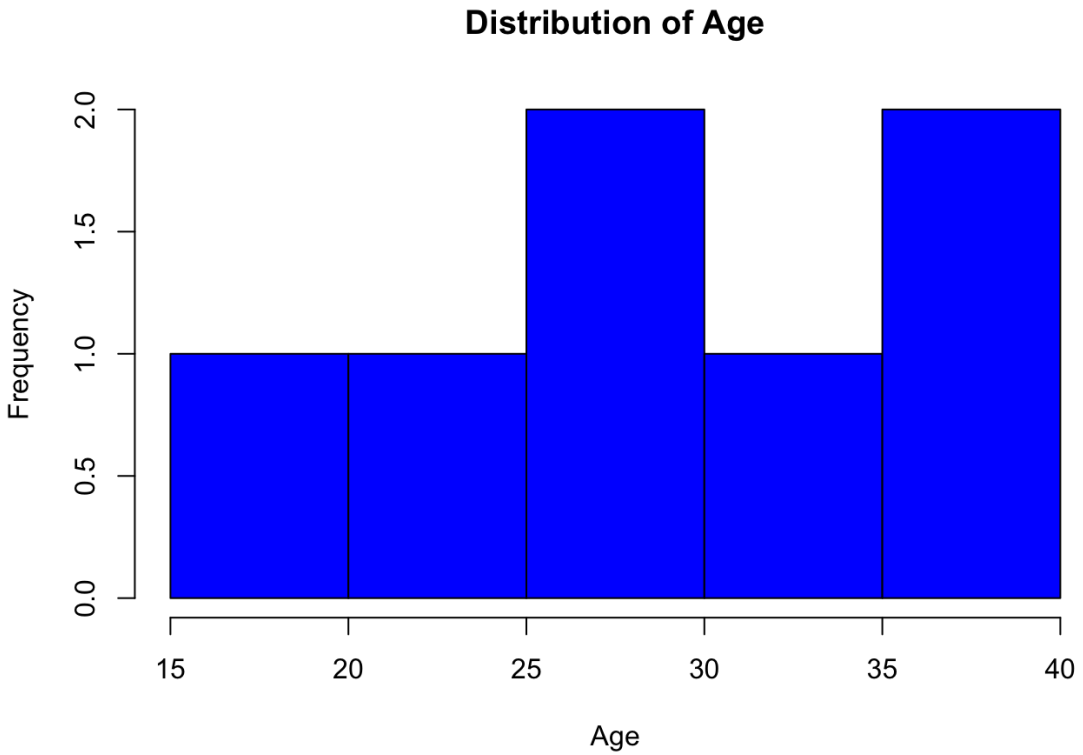
```
# xbars Mean
mean(xbars)

# xbars Standard Deviation
sd(xbars)
```

T-test & Hypothesis Testing

Conduct a 6 step hypothesis test to test this claim.

The following are ages of 7 randomly chosen patrons seen leaving the Beach Comber in South Mission Beach at 7pm! We assume that the data come from a normal distribution and would like to test the claim that the mean age of the distribution of Comber patrons is different than 21.

Dataset: 25, 19, 37, 29, 40, 28, 31



Code:

```
# Data
ages = c(25,19,37,29,40,28,31)
hist(ages, main="Distribution of Age", col="Blue", xlab="Age")
# Sample mean
mean(ages)
# Sample standard deviation
sd(ages)
```

Mean: 29.85714

Standard Deviation: 7.081162

T-test & Hypothesis Testing

Step 1

This is a one sample two sides t-test.

h0 = 21

h1 != 21

Step 2

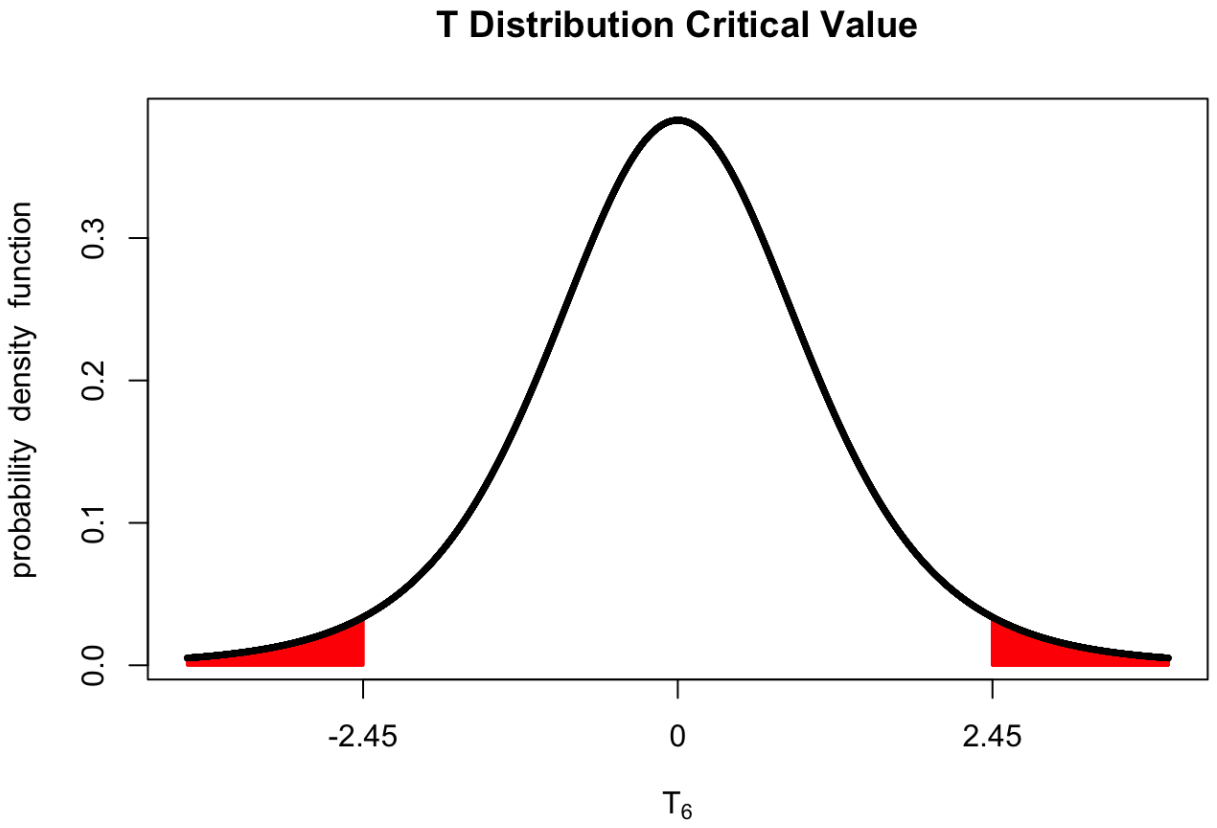
df = 6

Critical value = 2.447*

* Found using t-table

Code:

```
library(fastGraph)

shadeDist( c(- 2.447, 2.447), "dt", 6, main = "T Distribution Critical Value" )
```



T-test & Hypothesis Testing

Step 3, 4

One Sample t-test

data: ages

t = 3.3093, df = 6, p-value = 0.01622

alternative hypothesis: true mean is not equal to 21

5 percent confidence interval:

29.68217 30.03211

sample estimates:

mean of x 29.85714

Code:

```
t.test(ages, mu=21, conf.level=0.05)
```

T-test & Hypothesis Testing

Step 5

The P-value is less than 0.05, hence we reject the null hypothesis

Step 6

Conclusion: There is enough evidence that suggests that the mean age of patrons is different from 21 years old (P-value: 0.01622).

Takeaways

- R coding for charts it's much easier than in python
- The interactivity of using r-markdown files is nice. It remands me to jupyter notebooks.
- t.test is easy to use. You just need to provided data, type of test and confidence level

Questions

- How do we add the t-value and p-value information to the T-Distribution chart used in CLT?
- In the hypothesis test section we determined that we have enough evidence to reject the null, in the reality a sample size of 7 seems like a very small sample to have a valid conclusion even if we have statistical significance *(P<CV)*. How do we know what is the correct sample size required to ensure that the test has enough power?

Appendix

- Link to code: https://github.com/DavidG16/SMU-MSDS-6036-Doing-Data-Science/blob/master/Units/Unit1/David-Grijalva-For-Live-Session-Unit1.Rmd