Descriptive Statistics Project - Udacity

Pratik Gandhi

March 30, 2016

Random experiment with the use of standard deck of cards for showing Descriptive Statistics

- A deck of 52 cards, divided into four suits(Spades,Hearts,Diamonds and Clubs), each containing 13 cards(Ace,numbers 2-10 and face cards Jack, Queen and King) are taken.
- Here, we would be assigning values:

a) Ace - 1
b) Numbered cards take the printed values
c) Jack, Queen and King - 10

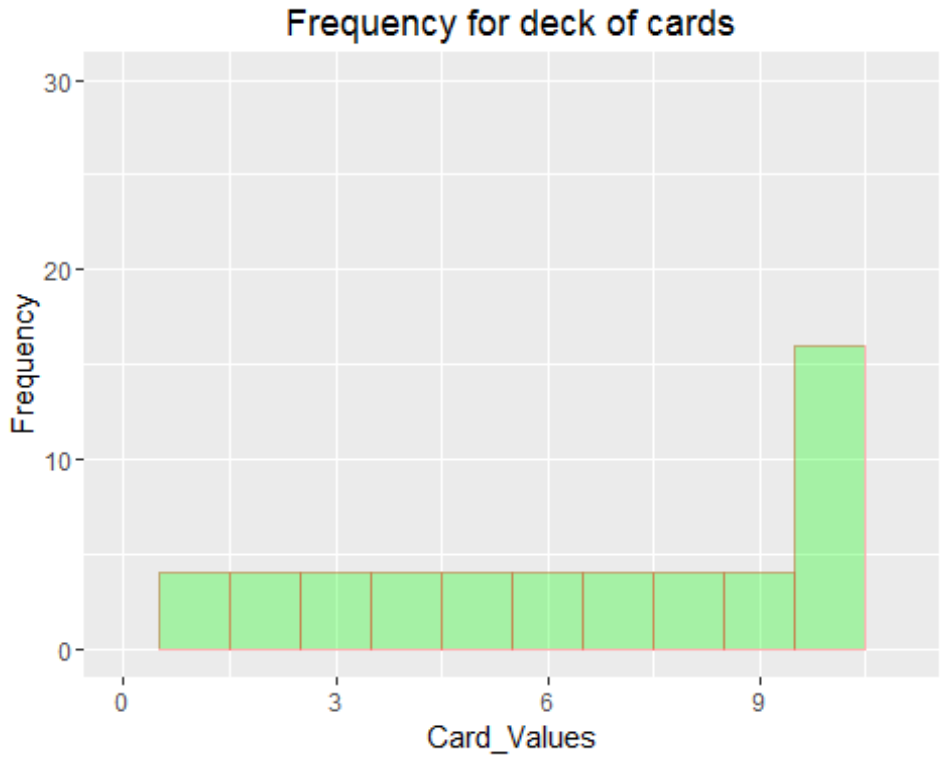
```
# Loading the libraries
library(ggplot2)
library(gridExtra)
i<-0
sum_value <- 0

# Generating the cards
suits <- c("Diamonds", "Clubs", "Hearts", "Spades")
cards <- c("Ace", "Deuce", "Three", "Four","Five", "Six", "Seven", "Eight",
"Nine", "Ten", "Jack", "Queen", "King")
values <- c(1, 2:9, rep(10, 4))
totalNumOfDecks <- 1 # The number of decks we are using here!

## Building a Deck:
deck <- expand.grid(cards=cards, suits=suits)

## Assigning values to deck.
deck$value <- values
```

1. Plotting the relative frequencies of the card values



2.Getting random samples of size = 3 from the population(entire deck) distribution.

- Sampling without replacement.
- Recording the card values and summing three of them.
- Repeating it three times.

```
# Running loop over the cards:
for (i in 1:30){
  x <- deck[sample(1:nrow(deck), 3, replace=FALSE),]
  sum_value[i] <- x[1,3] + x[2,3] + x[3,3]
}

# Storing the values in data frame form
sum_value_df <- as.data.frame(sum_value)
```

3. Distribution of this card sums. Reporting measures of central tendency (mean,median,mode) and measures of variability (range,mean absolute deviation(MD),variance,standard deviation).

```
# Calculating the mean
mean(sum_value)

## [1] 18.26667
```

```r
# Calculating the median
median(sum_value)
```

```
## [1] 20
```

```r
# Creating a function to get mode
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}
getmode(sum_value) # Putting our variable with values in the function
```

```
## [1] 15
```

```r
# Calculating the range
range_df <- range(sum_value)
diff(range_df) # Taking the difference between minimum and maximum values of
range
```

```
## [1] 20
```

```r
# Calculating the mean absolute deviation
mad(sum_value)
```

```
## [1] 7.413
```

```r
# Calculating the variance
var(sum_value)
```

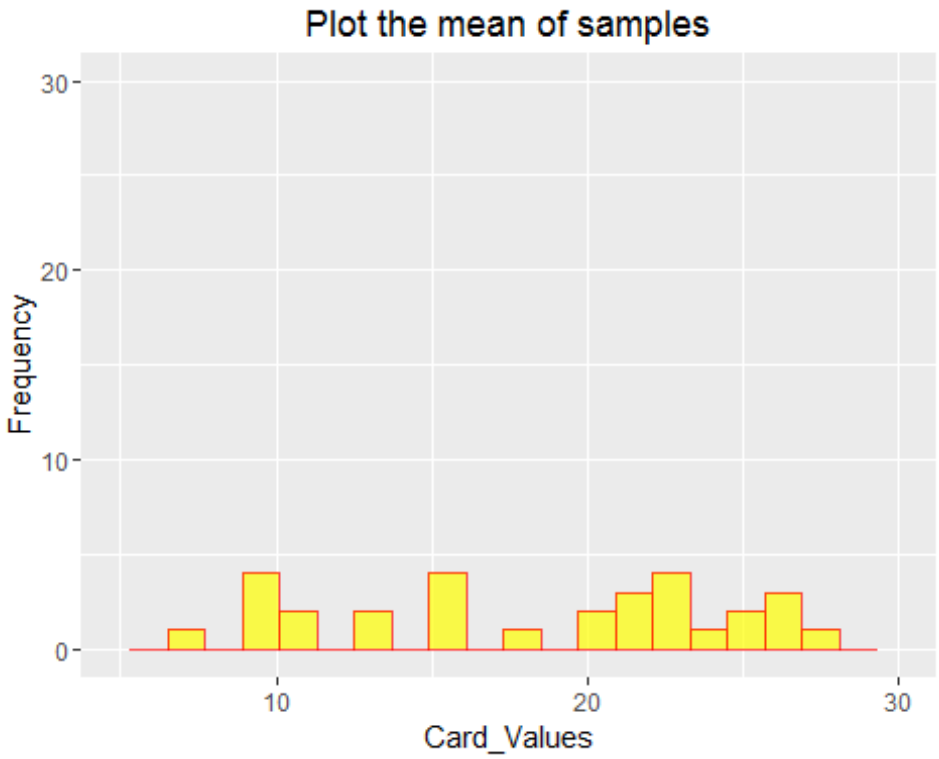
```
## [1] 39.02989
```

```r
# Calculating the standard deviation
sd(sum_value)
```

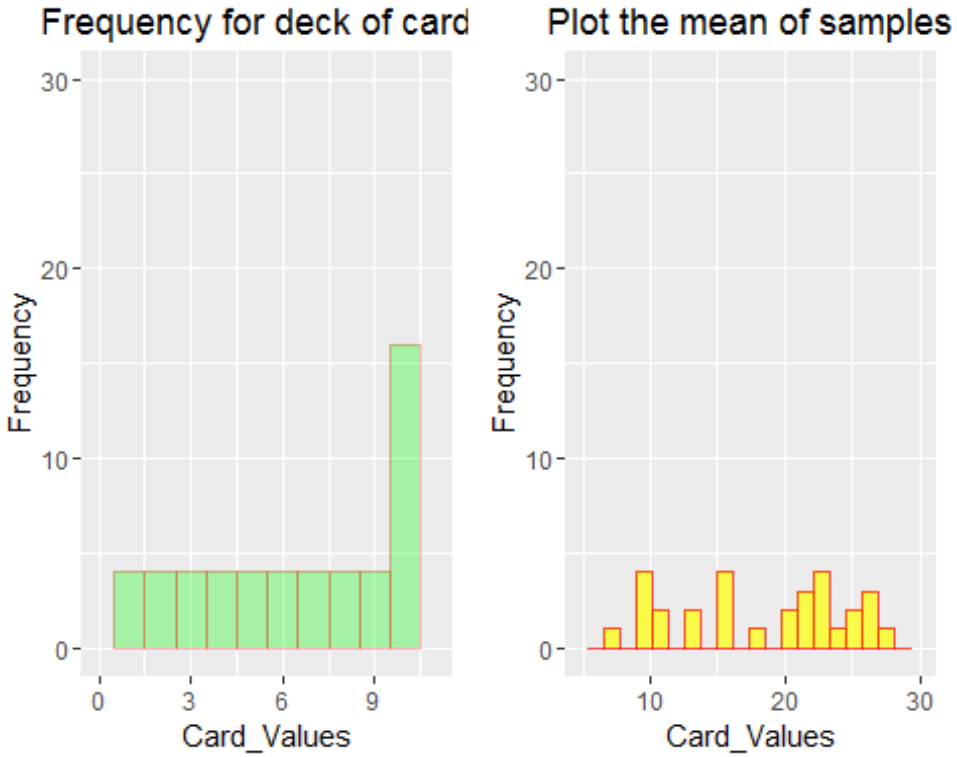
```
## [1] 6.24739
```

4. Creating histogram of the sample recorded.

```r
sample_plot <- ggplot(sum_value_df,aes(sum_value_df$sum_value)) +
geom_histogram(breaks=seq(0.5,30.5,by=1.2),col="red",fill="yellow",alpha=0.7)
+ labs(title="Plot the mean of samples")  +
labs(x="Card_Values",y="Frequency") + xlim(c(5,30)) + ylim(c(0,30))
plot(sample_plot)
```




```
# Comparing the population to the sample distribution. So, plotting in grid
grid.arrange(original_plot,sample_plot,ncol=2)
```



We can make several observations and conclusion watching both the plots:

- The original graph has a skewed distribution. Taking 30 samples and applying Central Limit Theorem would give a less uniform distribution.
- If the sampling procedure is done more times (300/3000) the distribution would have been much better normally distributed.

5. Future Predictions:

```
# The range in which we expect approximately 90% of future draws to fall
quantile(sum_value,probs=c(.05,.95))
```

```
##    5%   95%
##  9.45 26.00
```

```
# Probability of getting draw value of atleast 20:
z=1-(sqrt((20-mean(sum_value))^2/nrow(sum_value_df)))
print(z)
```

```
## [1] 0.6835381
```