Las Vegas hotel reviews

Michael Barrera

12/12/2018

```
path <- file.path("~", "desktop", "lasvegas_tripadvisor.csv")
mydata <- read.table(path, header = TRUE, sep = ",",
                     stringsAsFactors = FALSE) #
```

```
library(plotly)
```

Dataset Details

This is a dataset of 492 guest reviews from the travel site tripadvisor.com. The reviews contain both numerical and categorical variables about 21 Las Vegas hotels in addition to data about the individual reviewers. Reviewers are asked to rate their overall satisfaction level on a scale of 1 to 5 (1 being the lowest, 5 being the highest). Reviewers also provided personal information about themselves such as where they are from and who they traveled with. The dataset also contains information about property amenities like pool, spa, casino, free internet, etc. The amenities information was not provided by the reviewers but rather was automatically populated by tripadvisor.com in order to maintain the consistency and integrity of the data.

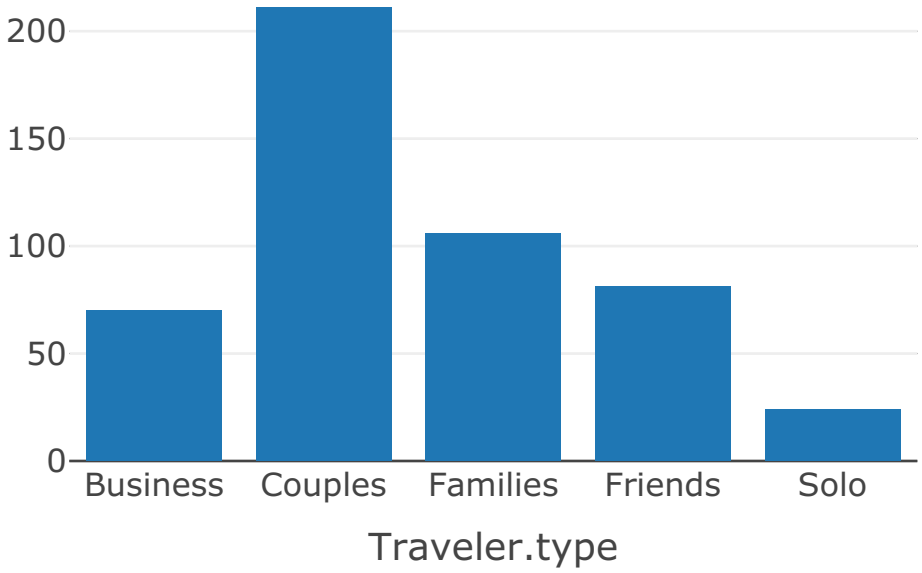
Objective

This project takes a hypothetical business approach. The objective of this project is to gain insight into the demographics of Las Vegas tourists, the overall satisfaction of Las Vegas visitors, and the types of amenities that are most important to them. Such information could be very valuable for organizations (e.g.: Hotels) with a vested interest in improving guest satisfaction.

Categorical variable #1

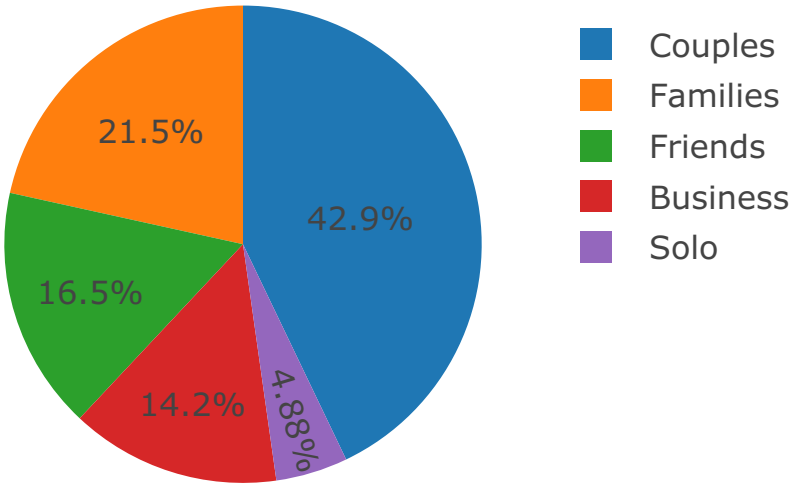
Which groups most commonly visit Las Vegas?

```
plot_ly(mydata, x = ~Traveler.type, type = "histogram")
```



Categorical variable #1 (cont.)

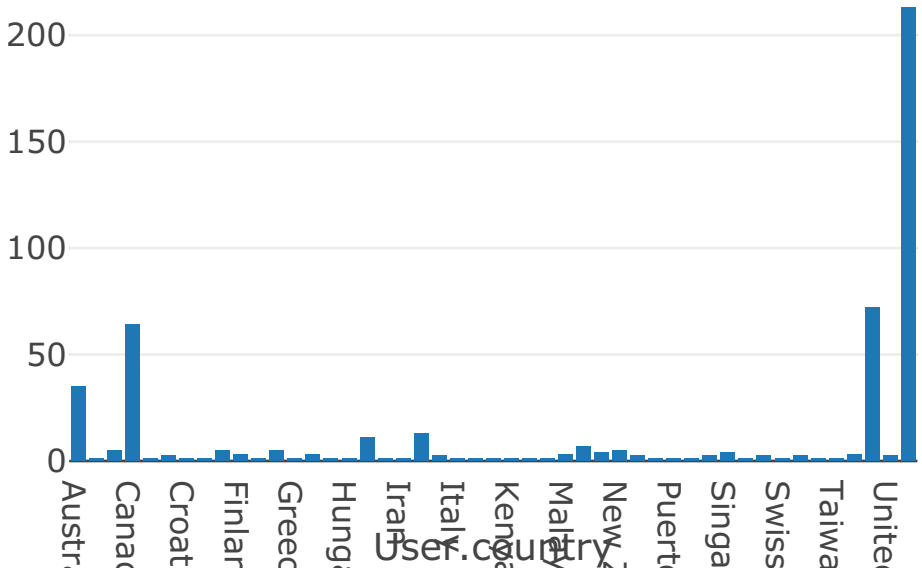
```
plot_ly(mydata, labels = ~Traveler.type, type = "pie")
```



Categorical variable #2

What countries are the reviewers from?

```
plot_ly(mydata, x = ~User.country, type = "histogram")
```



Numeric variable #1

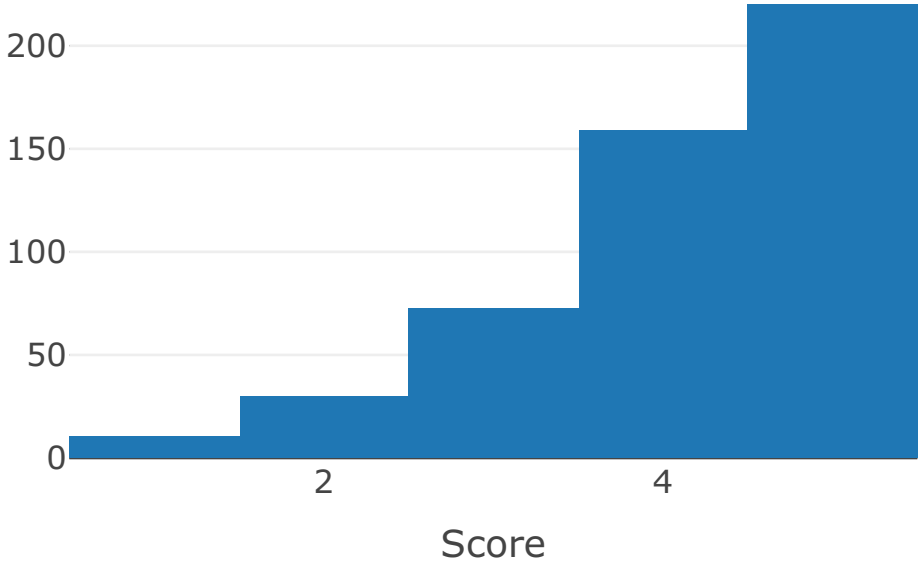
Let's explore the most important piece of data... the overall Score!

```
scores <- mydata$Score
range(scores)
## [1] 1 5
mean(scores)
## [1] 4.111789
sd(scores)
## [1] 1.014007
median(scores)
## [1] 4
table(scores)
## scores
##   1   2   3   4   5
##  11  30  72 159 220
```

Numeric variable #1 (cont.)

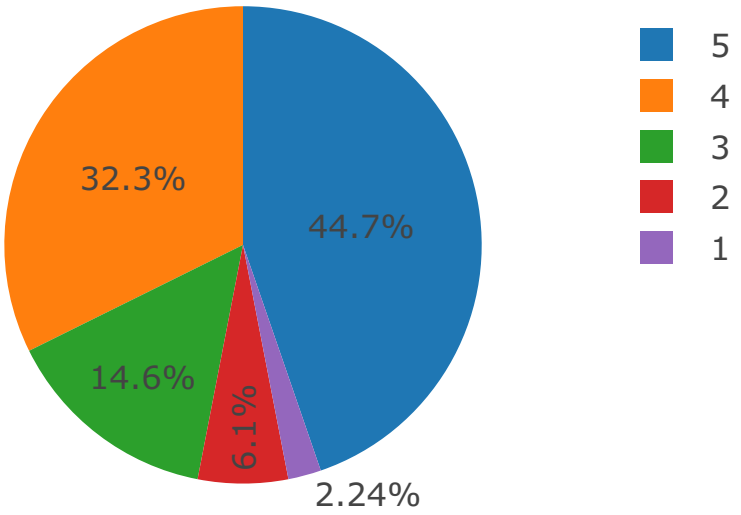
Histogram of Scores

```
plot_ly(mydata, x = ~Score, type = "histogram")
```



Numeric variable #1 (cont.)

```
plot_ly(mydata, labels = ~Score, type = "pie")
```



Numeric variable #2

Another variable that may be of interest (depending on the audience) is the number of hotel reviews each reviewer has submitted. tripadvisor.com tracks this data and makes it publicly available on each reviewer's profile. This helps determine the reviewer's experience & credibility. There are a number of different types of analyses that could be explored with this data. For example, this data may give a glimpse into the frequecy with which someone travels and how their expectations compare to others who travel more/less frequently.

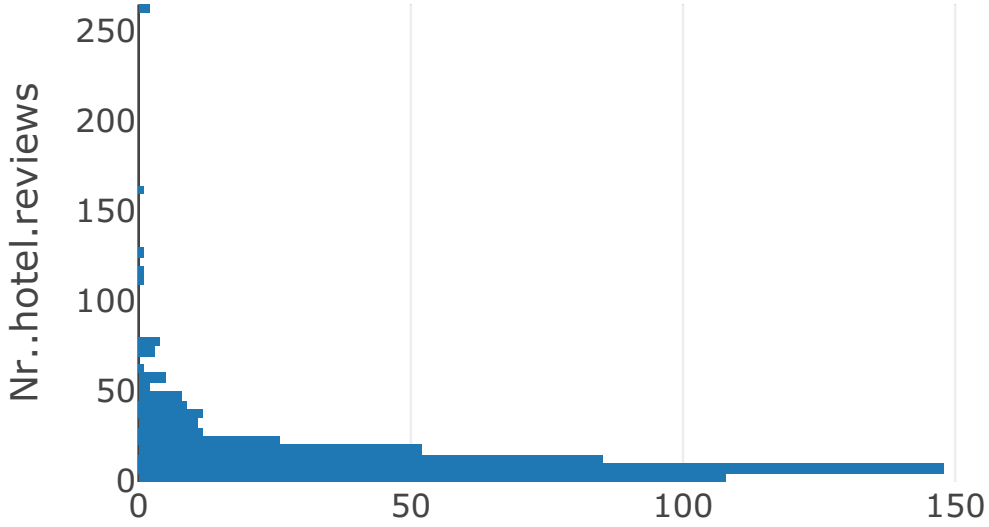
Numeric variable #2 (cont.)

```
range(mydata$Nr..hotel.reviews)
## [1]   0 263
mean(mydata$Nr..hotel.reviews)
## [1] 15.50407
median(mydata$Nr..hotel.reviews)
## [1] 9
var(mydata$Nr..hotel.reviews)
## [1] 550.2057
sd(mydata$Nr..hotel.reviews)
## [1] 23.45646
```

Numeric variable #2 (cont.)

Histogram of the # of hotel reviews

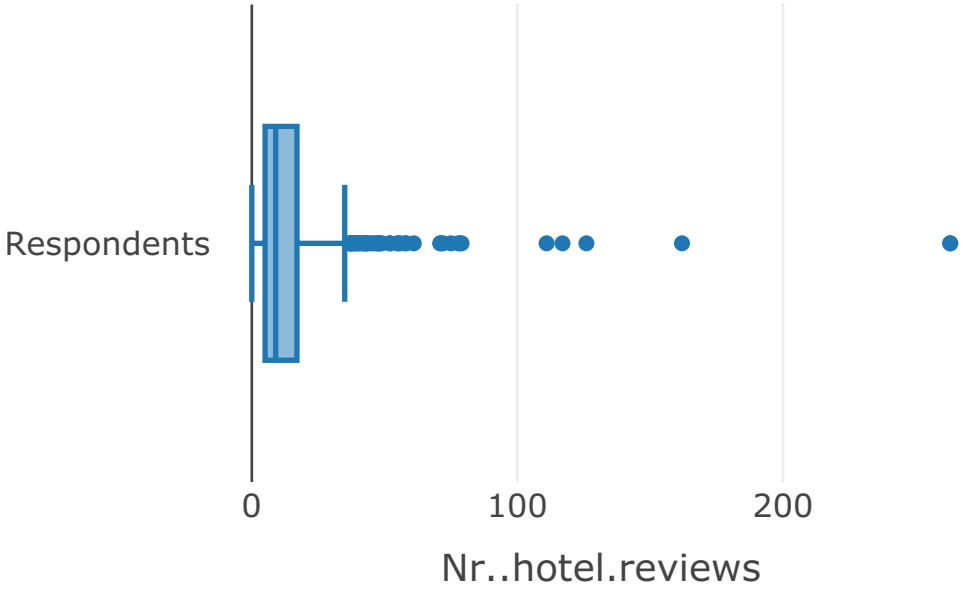
```
plot_ly(mydata, y = ~Nr..hotel.reviews, type = "histogram")
```



Numeric variable #2 (cont.)

Boxplot of the # of hotel reviews

```
plot_ly(mydata, x = ~Nr..hotel.reviews, type = "box", name = "Respondents")
```




```
attach(mydata)
```

Bivariate analysis

An important factor for most travelers when rating a hotel are the property's amenities. There are 6 different variables for amenities in the dataset, but I chose to focus on 2 of them: Pool & free internet. I wanted to see if there was any relation between the availability of these amenities and the scores. This type of analysis may be useful for hotel management when trying to determine which amenities to invest in or how to market them.

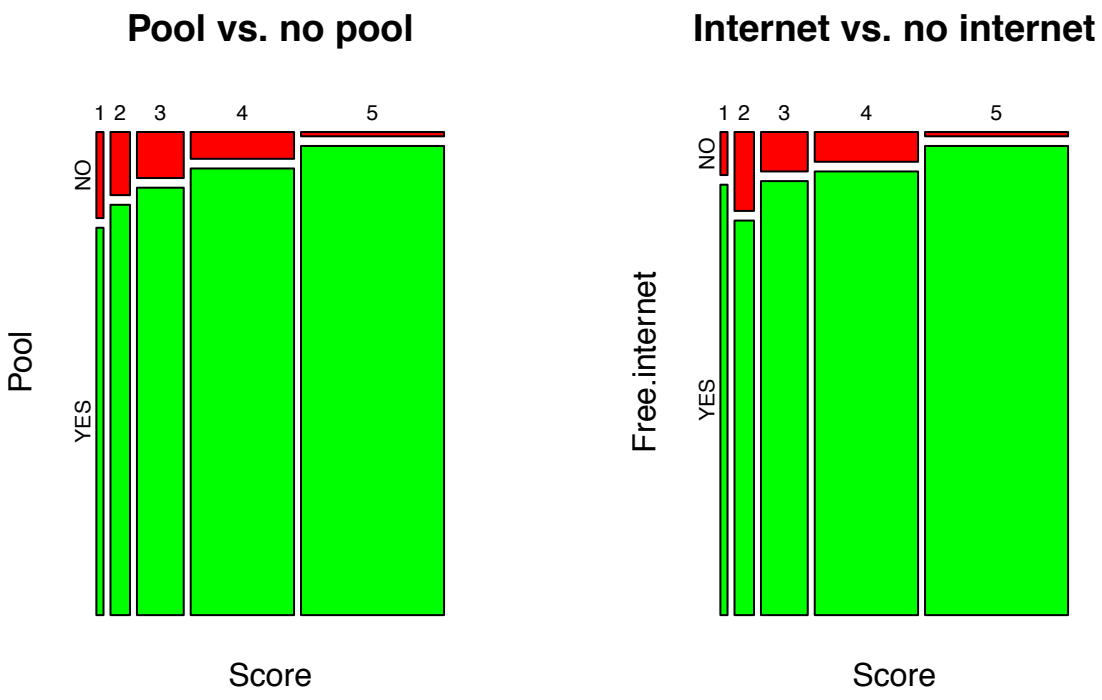
Contingency table

```
ftable(table(Score, Pool, Free.internet),
       col.vars = c("Free.internet", "Pool"))
```

```
##       Free.internet  NO     YES
##       Pool           NO YES  NO YES
## Score
## 1                     0   1   2   8
## 2                     0   5   4  21
## 3                     0   6   7  59
## 4                     0  10   9 140
## 5                     0   2   2 216
```

How do ammenities impact scores?

```
par(mfrow = c(1,2))
mosaicplot(table(Score,Pool), color = c("red","green"),
           main = "Pool vs. no pool")
mosaicplot(table(Score,Free.internet), color = c("red","green"),
           main = "Internet vs. no internet")
```

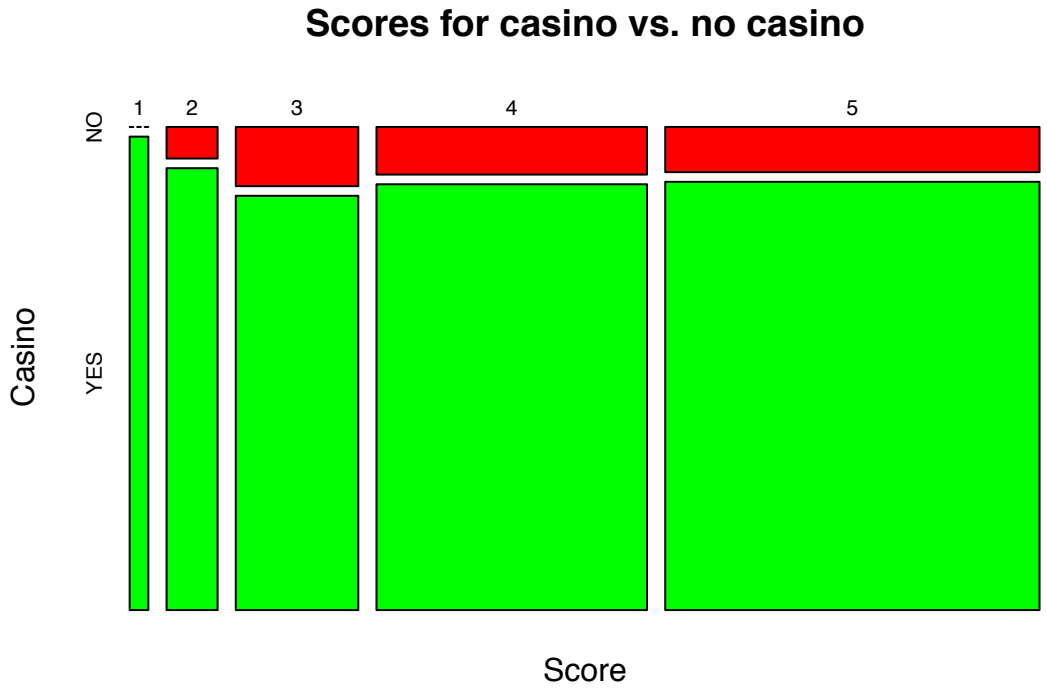


An interesting observation...

Whether a property had a casino or not seemed to have no influence on guest satisfaction. This makes sense because these guests likely deliberately booked a property without this amenity.

```
par(mfrow = c(1,1))
```

```
mosaicplot(table(Score,Casino), color = c("red","green"),
           main = "Scores for casino vs. no casino")
```



One more bivariate analysis. . .

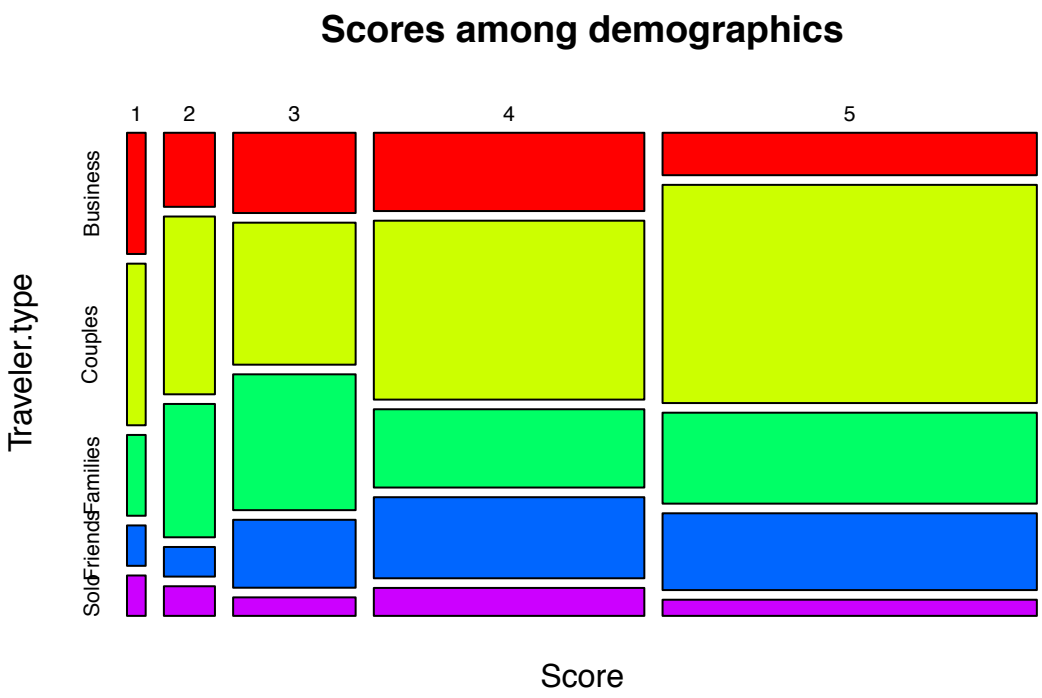
Another interesting observation that a hotel management or marketing team may be interested in is the distibution of scores among the different traveler types. The following plot seems to suggest that couples are more likely to give favorable scores than business travelers. . .

```
table(Score, Traveler.type)
```

```
##       Traveler.type
## Score Business Couples Families Friends Solo
##     1        3       4        2       1    1
##     2        5      12        9       2    2
##     3       13      23       22      11    3
##     4       28      64       28      29   10
##     5       21     108       45      38    8
```

Mosaic plot of scores among the demographics

```
mosaicplot(prop.table(table(Score, Traveler.type)),color = rainbow(5),
           main = "Scores among demographics")
```



Central Limit Theorem

Various samples of the data were drawn from the Score variable to demonstrate the applicability of the Central Limit theorem. As mentioned earlier, here is a summary of the data from the entire dataset:

```
nrow(mydata)
## [1] 492
median(mydata$Score)
## [1] 4
mean(mydata$Score)
```

```
## [1] 4.111789
sd(mydata$Score)
## [1] 1.014007
table(mydata$Score)
##
##   1   2   3   4   5
##  11  30  72 159 220
```

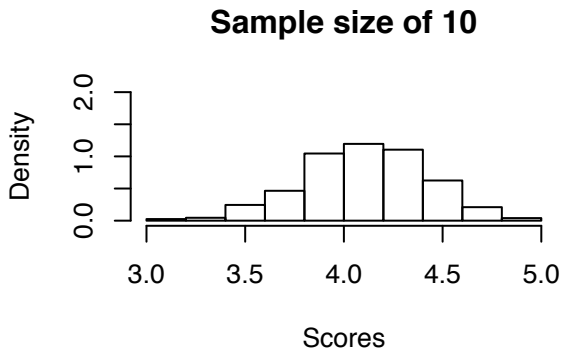
Central Limit Theorem (cont.)

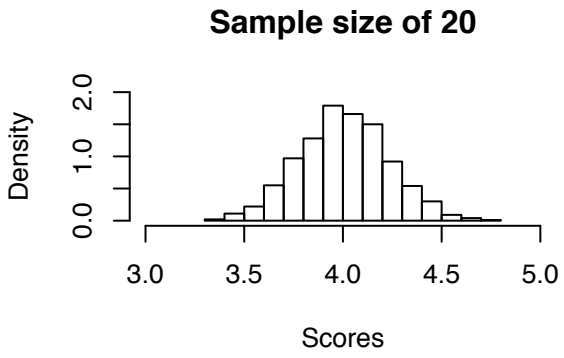
The central limit theorem states that the distribution of sample means, taken from independent random sample sizes, follows a normal distribution even if the original population is not normally distributed. This is important because there are times when statistical procedures require normality in the data set. As we saw earlier, the distribution of the Scores was heavily left-skewed. To demonstrate the applicability of the Central Limit Theorem to this variable, the next slide contains sample means of 1000 random samples of sample sizes 10, 20, 30, and 40...

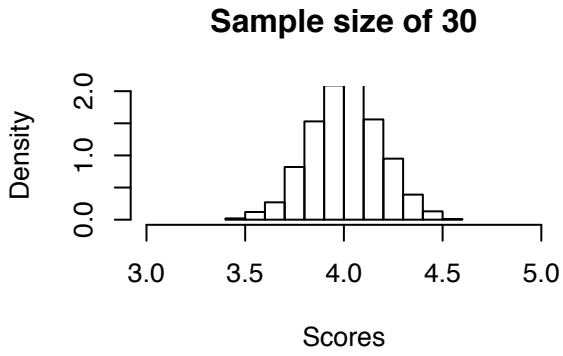
Central Limit Theorem (cont.)

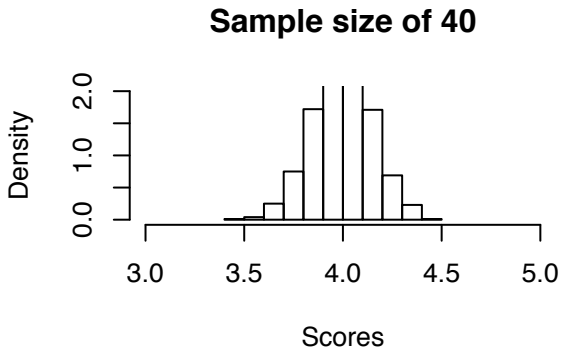
```
library(sampling)
par(mfrow = c(2,2))
set.seed(223)
samples <- 1000
sample.size <- 10

xbar <- numeric(samples)

for (i in 1:samples) {
  xbar[i] <- mean(rnorm(sample.size,
                        mean = 4.111789, sd = 1.014007))
}

hist(xbar, prob = TRUE,xlim = c(3,5),
     ylim = c(0,2), main = "Sample size of 10",
     xlab = "Scores")

set.seed(234)
samples <- 1000
sample.size <- 20

xbar <- numeric(samples)

for (i in 1:samples) {
  xbar[i] <- mean(rnorm(sample.size,
                        mean = 4, sd = 1))
}

hist(xbar, prob = TRUE,xlim = c(3,5),
     ylim = c(0,2), main = "Sample size of 20",
     xlab = "Scores")
```

```r
set.seed(456)
samples <- 1000
sample.size <- 30

xbar <- numeric(samples)

for (i in 1:samples) {
  xbar[i] <- mean(rnorm(sample.size,
                        mean = 4, sd = 1))
}

hist(xbar, prob = TRUE,xlim = c(3,5),
     ylim = c(0,2), main = "Sample size of 30",
     xlab = "Scores")

set.seed(678)
samples <- 1000
sample.size <- 40

xbar <- numeric(samples)

for (i in 1:samples) {
  xbar[i] <- mean(rnorm(sample.size,
                        mean = 4, sd = 1))
}

hist(xbar, prob = TRUE,xlim = c(3,5),
     ylim = c(0,2), main = "Sample size of 40",
     xlab = "Scores")
```










```
set.seed(321)
s1 <- srswor(20, nrow(mydata))
sample.2 <- mydata[s1 != 0, ]
sample.2$Score
```

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

```
table(sample.2$Score)
```

```
##
##  1  2  4  5
##  1  1 10  8
```

```
mean(sample.2$Score)
```

```
## [1] 4.15
```

```
sd(sample.2$Score)
```

```
## [1] 1.03999
```

```
set.seed(765)
N <- nrow(mydata)
n <- 20
k <- ceiling(N / n)
k
```

```
## [1] 25
```

```
r <- sample(k, 1)
r
```

```
## [1] 12
```

```
  #select every kth item
s2 <- seq(r, by = k, length = n)
sample.3 <- mydata[s2, ]
sample.3$Score
```

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

```
table(sample.3$Score)
```

```
##
##  2  3  4  5
##  1  3  4 12
```

```
mean(sample.3$Score)
```

```
## [1] 4.35
```

```
sd(sample.3$Score)
```

```
## [1] 0.933302
```

```
set.seed(212)
order.index <- order(mydata$Score)
ord.data <- mydata[order.index, ]
freq <- table(mydata$Score)
st.sizes <-ceiling(20 * freq / sum(freq))
st.1 <- strata(ord.data, stratanames = c("Score"),
               size = st.sizes, method = "srswor",
               description = TRUE)
```

```
## Stratum 1
##
## Population total and number of selected units: 11 1
## Stratum 2
##
## Population total and number of selected units: 30 2
## Stratum 3
##
## Population total and number of selected units: 72 3
## Stratum 4
##
## Population total and number of selected units: 159 7
## Stratum 5
##
## Population total and number of selected units: 220 9
## Number of strata  5
## Total number of selected units 22
```

```
sample.5 <- getdata(ord.data, st.1)
mean(sample.5$Score)
```

```
## [1] 3.954545
```

```
sd(sample.5$Score)
```

```
## [1] 1.174218
```

Various Samling methods

Various samples of the data were drawn using simple random sampling without replacement, systematic sampling, and stratified sampling. In all three cases, the mean and standard deviation were close to those of the raw data.

```
mean(sample.2$Score) #SWSROR
## [1] 4.15
sd(sample.2$Score)
## [1] 1.03999

mean(sample.3$Score) #Systematic Sampling
## [1] 4.35
sd(sample.3$Score)
## [1] 0.933302

mean(sample.5$Score) # Stratified Sampling
## [1] 3.954545
sd(sample.5$Score)
## [1] 1.174218
```

Conclusion

The dataset shows that overall, most visitors to Las Vegas are very satisfied with their hotels. If the management at any of these hotels wanted to improve their overall ratings, they can use a few key points from this analysis to develop the best strategy: 1) The most common traveler types to Las Vegas are couples 2) The vast majority of visitors are from the USA, UK, Canada, & Austrailia 3) Business travelers tend to be harsher reviewers compared to other groups 4) Properties with more amenities tend to get better ratings One thing to bear in mind is this was a relatively small data set captured over a small period of time with reviews from only 1 website. More robust analysis may be needed depending on one's goal, but this data provides a great starting point.