Class Activity#11

Sunny Lee

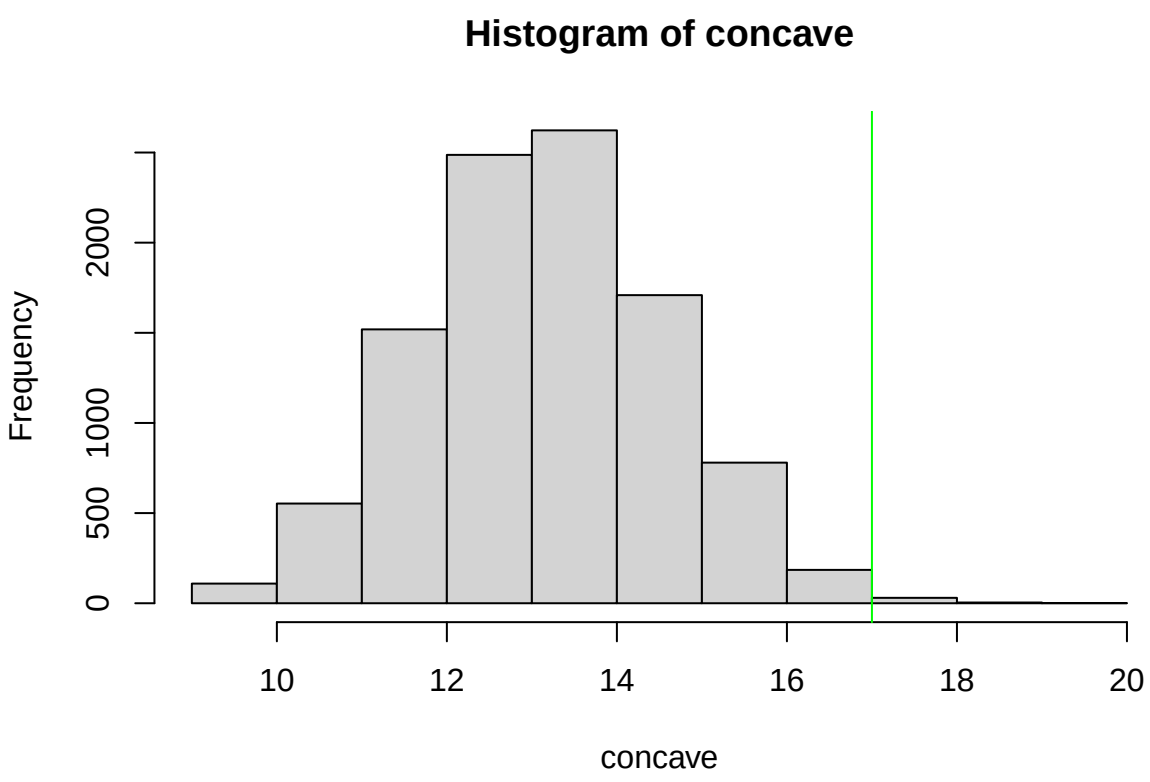
2021-03-12

Catagorical Data Analysis

```
cell<-c(rep("M",24),rep("B",13))
exp1<-sample(cell,21,replace=FALSE)
sum(exp1=="M")
```

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

```
p=10000
concave<-rep(0,p)
for(i in 1:p)
{
  temp<-sample(cell,21,replace=FALSE)
  concave[i]<-sum(temp=="M")
}
hist(concave)
abline(v=17,col="green")
```




```
p_value<-mean(concave>=17)
p_value
```

```
## [1] 0.022
```

1) Identify the observational units, the explanatory variable, and the response variable in the cancer cell data.

The observational unit are cancer cells, the explanatory variable are the concave cells, and the response is whether the concave cell is malignant or benign.

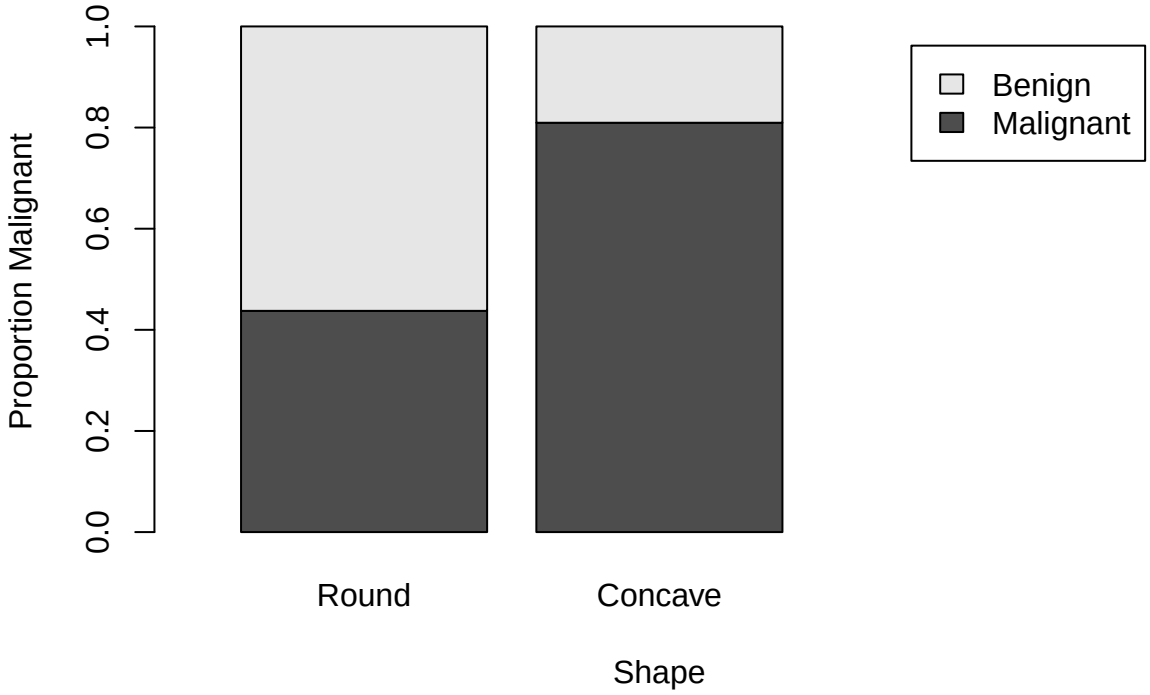
2) Calculate the proportion of round cell samples that are malignant and the proportion of concave cell samples that are malignant.

The proportion that the round cell samples are malignant is $\frac{7}{16}$ and the proportion that the concave cell samples are malignant is $\frac{17}{21}$.

3) Create a segment bar graph using the table given. The explanatory variable should be along the horizontal axis. Assuming this is a random sample from a larger population, does the graph show evidence that nucleus shape is related to the likelihood of a cell being malignant? Explain.

```
tumors <- cbind(c(7/16,9/16),c(17/21,4/21))

barplot(tumors,xlab="Shape",legend=c("Malignant","Benign"), ylab="Proportion Malignant",names.arg=c("Ro
```



Yes, we see that the percentage of Malignant cells which are concave is much higher than the percent of malignant cells which are round.

4) Use 37 index cards to represent this sample of 37 cancer cells. On 24 of the cards write "M" for malignant and on 13 of the cards write "B" for benign. Shuffle the card and randomly select 21 cards. these 21 cards can represent the concave nucleus group. How many of the 21 concave cards are also malignant?

```
cell<-c(rep("M",24),rep("B",13))
exp1<-sample(cell,21,replace=FALSE)
sum(exp1=="M")
```

```
## [1] 14
```

From the sample run above, we find that 14 of the 21 concave cards are also malignant.

5) Repeat the simulation process in Question 4 nine more times. Does it seem likely that 17 or more malignant cells would occur in the concave group by chance alone?

```
cell<-c(rep("M",24),rep("B",13))

for(i in 1:9){
exp1<-sample(cell,21,replace=FALSE)
print(sum(exp1=="M"))
}
```

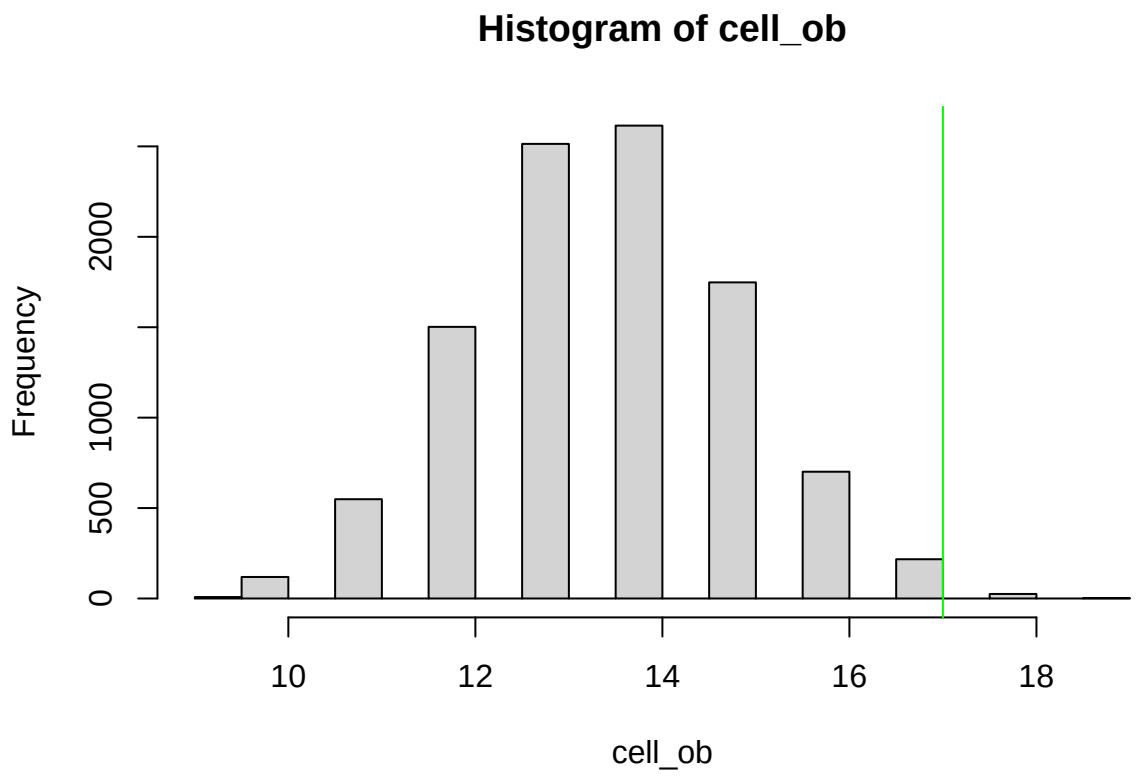
```
## [1] 12
## [1] 14
## [1] 13
## [1] 17
```

```
## [1] 14
## [1] 13
## [1] 15
## [1] 15
## [1] 14
```

From the 9 additional runs above, we see that only one of the runs had 17 malignant cells. Thus, we conclude that it is not likely taht 17 or more malignant cells would occur in the concave group.

6) Use the software instructions to repeat the computer-simulated randomization process a total of 10,000 times. Create a histogram of the 10,000 simulated counts in the concave malignant group. Estimate the p-value by dividing the number of counts greater than or equal to 17 by 10,000.

```
cell<-c(rep("M",24),rep("B",13))
p<-10000
cell_ob<-rep(0,p)
for(i in 1:p){
exp1<-sample(cell,21,replace=FALSE)
cell_ob[i]<-sum(exp1=="M")
}
hist(cell_ob)
abline(v=17,col="green")
```




```
mean(cell_ob>=17)
```

```
## [1] 0.0244
```

After running the simulation 10,000 times, most of the values on the histogram are less than 17. We can check this by calculating the p value, which, in this run, is .0237. Since this number is less than .05, we reject

the null hypothesis and conclude that the probability that a concave cell is malignant is less than $\frac{17}{21}$.

7) In this cancer study, assume $N = 37$ observations with $M = 24$ successes. If $n = 21$ observations are selected, use the technology instructions provided on the CD to calculate the exact probabilities $P(X = 17)$, $P(X = 18)$, $P(X = 19)$, $P(X = 20)$, and $P(X = 21)$.

```
(dhyper(17:21,24,13,21))
```

```
## [1] 1.921938e-02 2.989681e-03 2.574845e-04 1.072852e-05 1.571944e-07
```

Calculating the probabilities for $P(X = 17)$, $P(X = 18)$, $P(X = 19)$, $P(X = 20)$, and $P(X = 21)$., we get the array above.

8) What is the exact p-value $P(X \geq 17)$? How does this exact p-value compare to the simulated p-value?

```
sum(dhyper(17:21,24,13,21))
```

```
## [1] 0.02247743
```

```
phyper(16,24,13,21,lower.tail = FALSE)
```

```
## [1] 0.02247743
```

```
1-phyper(17,24,13,21)+dhyper(17,24,13,21)
```

```
## [1] 0.02247743
```

```
#sample(c(1,2,3,4,5,6),10,prop=c(.4,.3,.1,.075,.05,.25))
```

For the exact p value, we obtain a value of 0.02247743, which is quite close to our simulated p value of 0.0237.

9) There is nothing special about how a success or failure is defined. Assume we have a table of $N = 37$ observations with M=13 successes (here benign cell is considered a success). For a sample of size 16 (round nuclei), find $P(X \geq 9)$. How does this answer compare to your answer in question 8?

```
#qchisq(.975,df=1)
#qchisq(.025,df=1)
sum(dhyper(17:21,24,13,21))
```

```
## [1] 0.02247743
```

```
sum(dhyper(9:16,13,24,16))
```

```
## [1] 0.02247743
```

From the answer in question 8, we find that we obtain the same p value, so we could take either benign or malignant cells as our success.

10) Use the cancer table and define benign as a success and round cells to be group 1. Calculate and interpret the relative risk and the odds ratio. Relative Risk= $\frac{P(benign|round)}{P(benign|concave)} = \frac{\frac{9}{16}}{\frac{4}{21}} = \frac{9 \cdot 21}{16 \cdot 4} = 2.95$ and the odds ratio $= \frac{4 \cdot 7}{17 \cdot 9} = 0.18$. From the relative risk, we conclude that the chance of a round cell being benign is 2.95 times higher than a concave cell. From the odds ratio, we conlude that given a concave cell, it is .18 times more likely for that cell to be benign than malignant.

11) Show that the null hypothesis $H_o : p_1 = P_2$ is mathematically equivalent to the null hypothesis $H_o : \theta_1/\theta_2 = 1$, where p represents the proportion successful and θ represents the odds of success for any two groups (labeled 1 and 2).

Assume $P_1 = P_2$. Then, $\theta_1/\theta_2 = odds(S|G1)/odds(S|G2) = \frac{P(S|G1)}{P(S^c|G1)} / \frac{P(S|G2)}{P(S^c|G2)}$. Since $P(S|G1) = P(S|G2)$ and $P(S^c|G1) = P(S^c|G2)$, $\frac{P(S|G1)}{P(S^c|G1)} / \frac{P(S|G2)}{P(S^c|G2)} = 1$, and thus, $\theta_1/\theta_2 = 1$