STAT 477/STAT 577
HW 2 - Solutions

1. (15 pts) According to the company's website, the proportion of Green milk chocolate M&Ms produced is 0.16. Let the sample proportion $\hat{p}$ be the proportion of Green milk chocolate M&Ms in a large bag of 100 of the candies.

   (a) (5 pts) Determine the sampling distribution of the sample proportion $\hat{p}$.

   Since $np = 100(0.16) = 16$ and $n(1-p) = 100(0.84) = 84$ are both larger than 10, the sampling distribution of $\hat{p}$ is approximately Normal with mean $p = 0.16$ and standard deviation $\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(0.16)(0.84)}{100}} = 0.0367$.

   (b) (5 pts) Find the probability a large bag of 100 of the candies would have more than 20% green M&Ms.

   Find $P(\hat{p} > 0.2)$. Using the sampling distribution from part (a), this probability is approximately equal to:

```r
1 - pnorm(0.2, 0.16, sqrt(0.16*0.84/100))

## [1] 0.1376168
```

   (c) (5 pts) Find the probability a large bag of 100 of the candies would have less than 10% green M&Ms.

   Find $P(\hat{p} < 0.1)$. Using the sampling distribution from part (a), this probability is approximately equal to:

```r
pnorm(0.1, 0.16, sqrt(0.16*0.84/100))

## [1] 0.05085347
```

2. (39 pts) Many dog owners teach their dogs to "shake hands". Your friend has a dog and you notice over time that the dog seems to favor his right paw in doing this trick. You decide to test the accuracy of your perception. For 15 consecutive times the trick is done in the same way with the same person, you find that the dog extended his right paw 10 times.

   (a) (10 pts) Use R to give the summary table and bar graph of the sample data. You can use the data file **Dogs.csv**.

   Read in the data:

```r
dogdata <- read.csv(file.choose(), header = T)
```

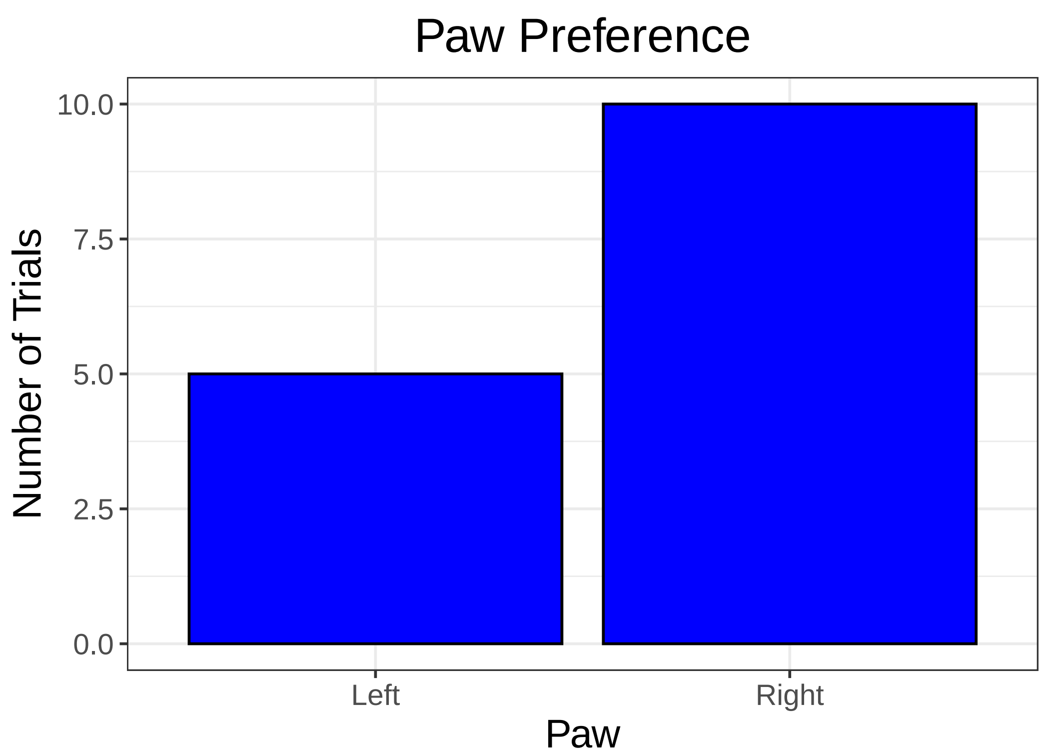
```r
paw.counts<- count(dogdata, var = 'Paw')
paw.table<- mutate(paw.counts,
                    prop = freq/sum(paw.counts[2]))
```

```r
paw.table<- rbind(paw.table, data.frame(Paw='Total',
                                t(colSums(paw.table[, -1]))))
paw.table
```

```
##     Paw freq      prop
## 1  Left    5 0.3333333
## 2 Right   10 0.6666667
## 3 Total   15 1.0000000
```

Make the bar graph for the Paw variable:

```r
ggplot(dogdata, aes(x=Paw))+
  geom_bar(fill = "blue", colour = "black")+
  labs(x = "Paw",
       y = "Number of Trials",
       title = "Paw Preference")+
    theme_bw()+
    theme(axis.title.y = element_text(size = rel(1.4)),
          axis.title.x = element_text(size = rel(1.4)),
          axis.text.x = element_text(size = rel(1.2)),
          axis.text.y = element_text(size = rel(1.2)),
          plot.title = element_text(hjust=0.5, size = rel(1.6)))
```



(b) (11 pts) Use R to conduct a binomial exact test for determining whether the dog favors his right paw when "shaking hands". Make sure to include the null and alternative hypotheses, test statistic, p-value, and conclusion.

If the dog does not favor a paw, the proportion of times the right paw is offered will be $p = 0.5$. If he does favor the right paw, the proportion of times the right paw is offered will be $p > 0.5$.

Null Hypothesis: $H_0 : p = 0.5$
Alternative Hypothesis: $H_a : p > 0.5$
Test Statistic: $Y = 10$
p-value: $P(Y \geq 10|p = 0.5) = 0.1509$
Conclusion: We have little evidence the dog prefers his right paw.

```r
binom.test(10, 15, p = 0.5, alternative = "greater")

##
##  Exact binomial test
##
## data:  10 and 15
## number of successes = 10, number of trials = 15, p-value = 0.1509
## alternative hypothesis: true probability of success is greater than 0.5
## 95 percent confidence interval:
##  0.4225563 1.0000000
## sample estimates:
## probability of success
##              0.6666667
```

(c) (5 pts) Use R to find the rejection region for this test. Use $\alpha = 0.05$.

We want to find the value of y so that $P(Y \geq y) \leq 0.05$. Using the `dbinom()` function, we have:

```r
sum(dbinom(11:15, 15, 0.5))

## [1] 0.05923462

sum(dbinom(12:15, 15, 0.5))

## [1] 0.01757812

sum(dbinom(13:15, 15, 0.5))

## [1] 0.003692627
```

This makes the rejection region $Y \geq 12$.

(d) (5 pts) Based on the rejection region you found in part (c), what is the observed Type I error rate for this test?

The observed Type I error rate will be $P(Y \geq 12|p = 0.5)$. This is:

```r
sum(dbinom(12:15, 15, 0.5))

## [1] 0.01757812
```

(e) (8 pts) Based on the rejection region you found in part (c), what is the power of your hypothesis test if the dog favors his right paw with probability 0.6, 0.75, or 0.9?

The power is $P(Y \geq 12 | p = p_a)$. For the three values of p_a above, we have:

```r
sum(dbinom(12:15, 15, 0.6))

## [1] 0.0905019

sum(dbinom(12:15, 15, 0.75))

## [1] 0.4612869

sum(dbinom(12:15, 15, 0.9))

## [1] 0.9444444
```

3. (46 pts) A company's old antacid formula provided relief from heartburn for 75% of the people who used it. The company develops a new formula in hopes of improving on the proportion of users who obtain relief. In a random sample of 400 people, 312 had relief of their heartburn.

(a) (10 pts) Use R to give the summary table and bar graph of the sample data. You can use the data file **Antacid.csv**.

Read in the data:

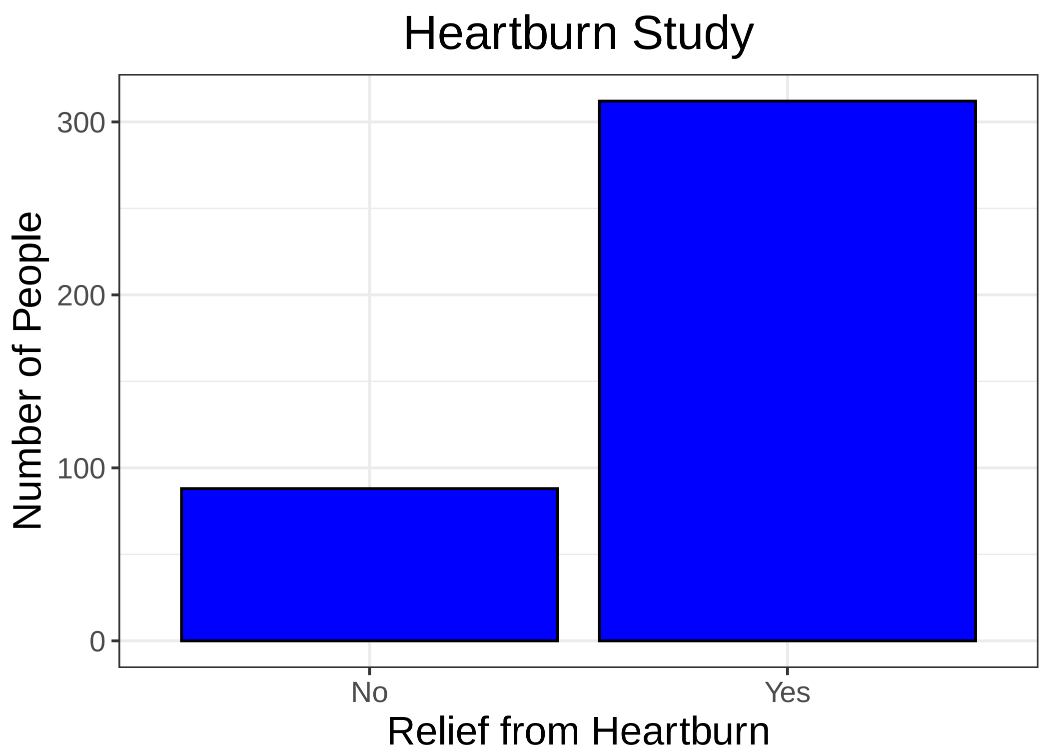
```r
reldata <- read.csv(file.choose(), header = T)
```

```r
rel.counts<- count(reldata, var = 'Relief')
rel.table<- mutate(rel.counts,
                   prop = freq/sum(rel.counts[2]))
rel.table<- rbind(rel.table, data.frame(Relief='Total',
                              t(colSums(rel.table[, -1]))))
rel.table

##   Relief freq prop
## 1     No   88 0.22
## 2    Yes  312 0.78
## 3  Total  400 1.00
```

Make the bar graph for the Relief variable:

```r
ggplot(reldata, aes(x=Relief))+
  geom_bar(fill = "blue", colour = "black")+
  labs(x = "Relief from Heartburn",
       y = "Number of People",
       title = "Heartburn Study")+
    theme_bw()+
    theme(axis.title.y = element_text(size = rel(1.4)),
          axis.title.x = element_text(size = rel(1.4)),
          axis.text.x = element_text(size = rel(1.2)),
          axis.text.y = element_text(size = rel(1.2)),
          plot.title = element_text(hjust=0.5, size = rel(1.6)))
```



(b) (5 pts) Explain why you can use the score test for this hypothesis test.

In this situation, the sample size $n = 400$ and the value of $p_0 = 0.75$ meaning both $np_0 = 400(0.75) = 300$ and $n(1 - p_0) = 400(0.25) = 100$ are greater than 10. We can use the score test since the Normal distribution is a good approximation for the sampling distribution of $\hat{p}$.

(c) (12 pts) Use R to conduct a score test for determining whether the new formula is better than the old formula. Make sure to include the null and alternative hypotheses, test statistic, p-value, and conclusion.

If the new formula is the same as the old one, the proportion of people who get relief will be $p = 0.75$. If the new formula is better than the old one, the proportion of people who get relief will be $p > 0.75$.

Null Hypothesis: $H_0 : p = 0.75$
Alternative Hypothesis: $H_a : p > 0.75$

```
prop.test(312, 400, 0.75, alternative = "greater", correct = F)

##
##  1-sample proportions test without continuity correction
##
## data:  312 out of 400, null probability 0.75
## X-squared = 1.92, df = 1, p-value = 0.08293
## alternative hypothesis: true p is greater than 0.75
## 95 percent confidence interval:
##  0.7441127 1.0000000
## sample estimates:
##    p
## 0.78
```

Test Statistic: $\sqrt{X^2} = \sqrt{1.92} = 1.3856$
p-value: $P(Z > 1.3856) = 0.0829$
Conclusion: We have weak evidence the new formula is better than the old one.

(d) (9 pts) Use R to calculate the power of this score test if the true proportion of users who obtain relief from their heartburn with the new formula is either $p = 0.8$, 0.85, or 0.9 and $\alpha = 0.01$, 0.05, and 0.1.
When $p_a = 0.8$

```
powerprop.test(400, 0.75, 0.8, alternative = "greater", 0.1)

## [1] 0.867077

powerprop.test(400, 0.75, 0.8, alternative = "greater", 0.05)

## [1] 0.7640508

powerprop.test(400, 0.75, 0.8, alternative = "greater", 0.01)

## [1] 0.4926816
```

When $p_a = 0.85$

```
powerprop.test(400, 0.75, 0.85, alternative = "greater", 0.1)

## [1] 0.9999741

powerprop.test(400, 0.75, 0.85, alternative = "greater", 0.05)

## [1] 0.9998448

powerprop.test(400, 0.75, 0.85, alternative = "greater", 0.01)
```

```
## [1] 0.9972821
```

When $p_a = 0.9$

```r
powerprop.test(400, 0.75, 0.9, alternative = "greater", 0.1)

## [1] 1

powerprop.test(400, 0.75, 0.9, alternative = "greater", 0.05)

## [1] 1

powerprop.test(400, 0.75, 0.9, alternative = "greater", 0.01)

## [1] 1
```

(e) (5 pts) Discuss the effect of the value of the population proportion p and the value of α on the power of this hypothesis test.

When p_a is further away from $p_0 = 0.75$, the power increases with the power for both $p_a = 0.85$ and $p_a = 0.9$ near 1.

When α is larger, the power is also larger.

(f) (5 pts) After the above analysis, the company decided to switch production to the new antacid formula. After several years in production, they found the new formula provided relief to 80% of the people who used it. Suppose the company would like to test another formula in the future. What sample size will they need to use to have a power of 0.9 to detect an improvement in the proportion of users who obtain relief of 0.05 if $\alpha = 0.05$.

```r
npowerprop.test(0.8, 0.85, alternative = "greater", 0.05, 0.9)

## [1] 498
```