FIT2086 Assignment 2

Name: Haw Xiao Ying

Student ID: 29797918

Tutorial: Group 04 (Friday 10a.m. - 12p.m.)

Question 1

1) First, let's find the estimate of the average fuel efficiency of vehicles that are all-wheel drive using the code below:

```
a <- fuel[fuel$Type=="A",]
mean(a$FA)
```

$$\hat{\mu}_a = 10.46713$$

Then, we need to obtain the variance, sample size and $t_{0.025,19}$ (95% confidence interval and sample size is n_a) for the later calculation. Use the R command below to find them:

```
var(a$FA)
nrow(a)
```

$$\hat{\sigma}^2{}_a = 7.559223$$
$$n_a = 20$$

To use the interval $\left(\hat{\mu} - t_{\frac{\propto}{2},n-1}\frac{\hat{\sigma}}{\sqrt{n}}, \hat{\mu} + t_{\frac{\propto}{2},n-1}\frac{\hat{\sigma}}{\sqrt{n}},\right)$, we also need to determine $t_{\frac{\propto}{2},n-1}$. Let us construct a 95% confidence interval by taking $\propto = 0.05$. Our sample size is $n = 20$, so using R we can find our multiplier as $t_{0.025,19} = qt(1 - 0.05/2, 20 - 1) \approx 2.093024$.

```
qt(1-0.05/2, 20-1)
```

Using this, we can find our interval to be

$$\left(10.46713 - 2.093024\sqrt{\frac{7.559223}{20}}, 10.46713 + 2.093024\sqrt{\frac{7.559223}{20}}\right)$$

which is equal to $(9.180373, 11.75389) \approx (9.1804, 11.7539)$. The estimated average of fuel efficiency of all-wheel-drive vehicles is 10.4671 km/L. This means that we are 95% confident the population mean fuel efficiency for this group is between 9.1804 km/L and 11.7539 km/L.

2) To find the population mean, variances and sample size of part-time four-wheel-drive vehicles respectively, use the R commands below:

```
p <- fuel[fuel$Type=="P",]
mean(p$FA)
var(p$FA)
nrow(p)
```

The estimated means and the unbiased estimates of variance for all-wheel drive vehicles (a) and part-time four-wheel-drive vehicles (p) are shown as below:

$$\hat{\mu}_a = 10.46713,\ \hat{\mu}_p = 8.771594$$
$$\hat{\sigma}^2{}_a = 7.559223,\ \hat{\sigma}^2{}_p = 9.386937$$

and the observed difference in fuel efficiency between the two groups is

$$\hat{\mu}_a - \hat{\mu}_p = 10.46713 - 8.771594 = 1.695536$$

Using these estimates in $\left(\hat{\mu}_a - \hat{\mu}_p - z_{\propto/2}\sqrt{\frac{\hat{\sigma}^2{}_a}{n_a} + \frac{\hat{\sigma}^2{}_p}{n_p}}, \hat{\mu}_a - \hat{\mu}_p + z_{\propto/2}\sqrt{\frac{\hat{\sigma}^2{}_a}{n_a} + \frac{\hat{\sigma}^2{}_p}{n_p}}\right)$ yields an approximate 95% confidence interval by taking $\propto = 1 - 0.95 = 0.05$

$$\left(1.695536 - 1.96\sqrt{\frac{7.559223}{20} + \frac{9.386937}{25}}, 1.695536 + 1.96\sqrt{\frac{7.559223}{20} + \frac{9.386937}{25}}\right)$$

which is $(-0.005760535, 3.396833) \approx (-0.0058, 3.3968)$. We could summarise these results using a statement such as:

"The estimated difference in fuel efficiency between all-wheel-drive vehicles (sample size n = 20) and part-time four-wheel-drive vehicles (sample size n = 25) is 1.6955 km/L. We are 95% confident the population mean difference in fuel efficiency of all-wheel-drive vehicles and part-time four-wheel-drive vehicles is between -0.0058 km/L (fuel efficiency is lower in part-time four-wheel-drive vehicles) up to 3.3968 km/L (fuel efficiency is greater in all-wheel-drive vehicles). As the interval includes zero, we cannot rule out the possibility of there being no difference at a population level between fuel efficiency of all-wheel-drive vehicles and part-time four-wheel-drive vehicles."

3) Using the provided data, we believe that all-wheel-drives are less efficient than part-time four-wheel-drive vehicles. i.e., To test the above hypothesis, we set the null hypothesis as: all-wheel-drive vehicles is more efficient than or equally efficient to part-time four-wheel-drive vehicles. If the null hypothesis is rejected, we can conclude that all-wheel-drives are less efficient than part-time four-wheel-drive vehicles.

$$H_0: \mu_a \leq \mu_p$$
$$vs$$
$$H_A: \mu_a > \mu_p$$

The two sample means are $\hat{\mu}_a = 10.46713\ and\ \hat{\mu}_p = 8.771594$ respectively and the unbiased estimates of variance are $\hat{\sigma}^{2}{}_{a} = 7.559223$ and $\hat{\sigma}^{2}{}_{p} = 9.386937$. The z-score for difference in $\hat{\mu}_p$ and $\hat{\mu}_a$ is:

$$z_{\hat{\mu}_p - \hat{\mu}_a} = \frac{8.771594 - 10.46713}{\sqrt{\frac{7.559223}{20} + \frac{9.386937}{25}}} = -1.953364 \approx -1.9534$$

Then, by using the R command below to find the p-value of $P(Z < -1.9534)$:

```
pnorm(-1.9534)
```

It gives us an answer of

$$P(Z < -1.9534) = 0.02538611 \approx 0.0254$$

The p-value, 0.0254 means that only some sample we could observe (2.54% of possible sample), the fuel efficiency of all-wheel-drive vehicles is less than or equal to the fuel efficiency of part-time four-wheel-drive vehicles. The p-value indicates that we have moderate evidence against the null hypothesis. This p-value suggests that if the null was true, only some of the sample (2.54% of the sample) we could observe would lead to a z-score as large or larger than the one we have observed; this also says that the data we have observed is at most at odds with the null distribution, and therefore offers moderate evidence against the null. Hence, the null is rejected (all-wheel-drives are less efficient than part-time four-wheel-drive vehicles) and we can conclude the alternative hypothesis (part-time four-wheel-drive vehicles are more efficient than all-wheel-drive vehicles).

4) Conclusion 1

In Question 1 (2), we have found the 95% confidence interval for difference in fuel efficiency between all-wheel-drive vehicles and part-time four-wheel-drive vehicles and we have concluded that it is possibly no difference in population level between the fuel efficiency of all-wheel-drive vehicles and part-time four-wheel-drive vehicles because the interval included zero.

Conclusion 2

For Question 1 (3), we are asked to test the hypothesis that all-wheel-drives are less efficient than part-time four-wheel-drive vehicles and our conclusion in Question 1 (3) is that we have evidence to reject the null hypothesis (fuel efficiency of all-wheel-drive vehicles is less than or equal to the fuel efficiency of part-time four-wheel-drive vehicles).

Now, as we can see, conclusion 1 is conflicted with conclusion 2. One claimed that it is possibly no difference in population level of the fuel efficiency between all-wheel-drive vehicles and part-time four-wheel-drive vehicles whereas another one rejected that fuel efficiency of all-wheel-drive vehicles is less than or equal to the part-time four-wheel-drive vehicles. If we want to make it clearer, by using the p-value in Question 1 (3): 0.0254, we can find out the p-value of the hypothesis: fuel efficiency of all-wheel-drives and part-time four-wheel-drive vehicles are the same. The p-value of the hypothesis above is 2*(0.0254) = 0.0508 so the hypothesis above (fuel efficiency of all-wheel-drives and part-time four-wheel-drive vehicles are the same) is rejected.

Question 2

1) To plot a graph of the geometric probability mass function given for the values $y \in \{0, 1, \dots, 20\}$, for L = 0, L = 1 and L = 2, use the R commands below:

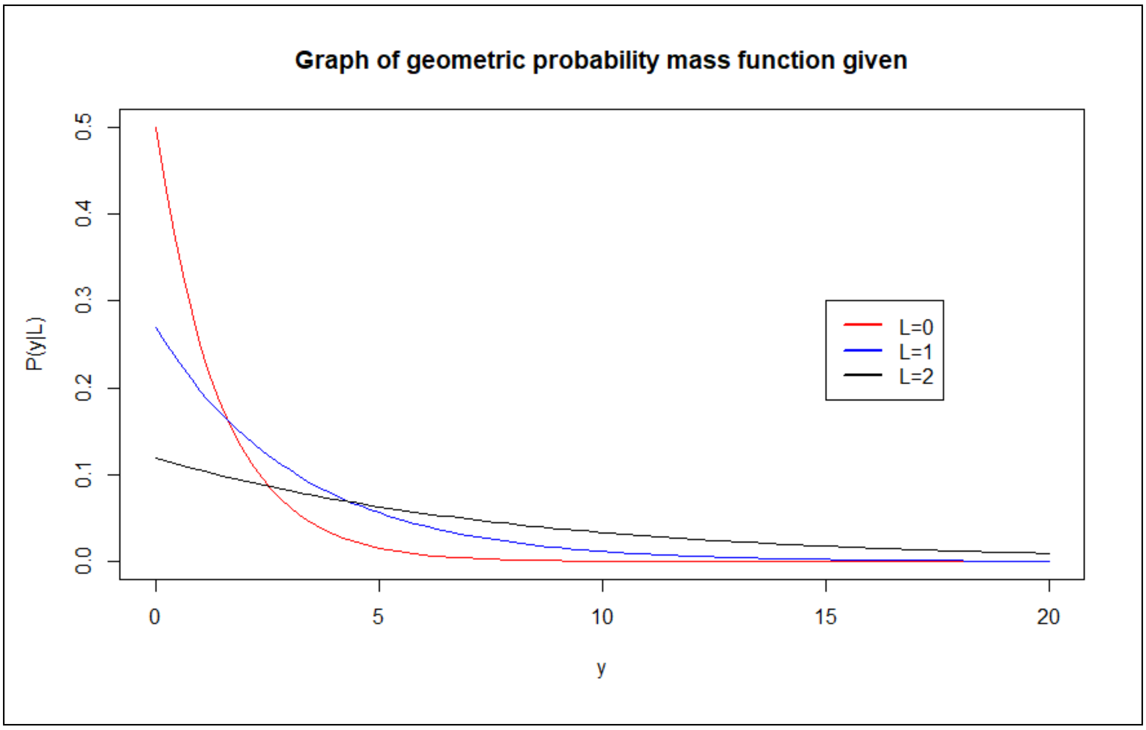
```
gpmf0 <- function(y){
    ((exp(0)+1)^(-y-1)) * (exp(y*0)) }

gpmf1 <- function(y){
    ((exp(1)+1)^(-y-1)) * (exp(y*1)) }

gpmf2 <- function(y){
    ((exp(2)+1)^(-y-1)) * (exp(y*2)) }

curve(gpmf0, from=0, to=20, xlab="y", ylab="P(y|L)", main="Geometric
probability mass function", col="red")
curve(gpmf1, from=0, to=20, add=TRUE, col="blue")
curve(gpmf2, from=0, to=20, add=TRUE, col="black")
legend(x=15, y=0.3, legend=c("L=0","L=1","L=2"),
col=c("red","blue","black"), lwd=2, lty=c(1, 1))
```

The graph of the geometric probability mass function given for the values $y \in \{0, 1, \dots, 20\}$, for L = 0, L = 1 and L = 2 is shown as below:



2) $p(y|L) = \prod_{i=1}^{n} p(y_i|L)$

$p(y|L) = ((e^L + 1)^{-y_1-1}\, e^{y_1 L}) \times ((e^L + 1)^{-y_2-1}\, e^{y_2 L}) \ldots ((e^L + 1)^{-y_n-1}\, e^{y_n L})$

$p(y|L) = (e^L + 1)^{(-y_1-1)+(-y_2-1)+\cdots+(-y_n-1)} \times e^{(y_1 L)+(y_2 L)+\cdots+(y_n L)}$

$p(y|L) = (e^L + 1)^{\sum_{i=1}^{n} -y_i-1} \times e^{L\sum_{i=1}^{n} y_i}$

The joint probability of this sample of data is $(e^L + 1)^{\sum_{i=1}^{n}(-y_i-1)} \times e^{L\sum_{i=1}^{n} y_i}$

3) $L(y|L) = -\sum_{i=1}^{n} \log p(y_i|L)$

$L(y|L) = -\ln\, ((e^L + 1)^{\sum_{i=1}^{n} -y_i-1} \times e^{L\sum_{i=1}^{n} y_i}\ )$

$L(y|L) = -(\ln\big((e^L + 1)^{\sum_{i=1}^{n} -y_i-1}\big) + \ln\,(e^{L\sum_{i=1}^{n} y_i}))$

$L(y|L) = -(\sum_{i=1}^{n} -y_i - 1)\ln(e^L + 1) - L\sum_{i=1}^{n} y_i \ln e$

$L(y|L) = -(\sum_{i=1}^{n} -y_i - 1)\ln(e^L + 1) - L\sum_{i=1}^{n} y_i$

The negative loglikelihood of the data y is $-(\sum_{i=1}^{n}(-y_i - 1))\ln(e^L + 1) - L\sum_{i=1}^{n} y_i$

4) Now, to derive the maximum likelihood estimator $\hat{L}$ for L, we need to differentiate the negative log-likelihood with respect to L:

$$\frac{d\big(L(y|L)\big)}{dL} = -\frac{(\sum_{i=1}^{n}(-y_i - 1))(e^L)}{e^L + 1} - \sum_{i=1}^{n} y_i$$

Now we set this derivative to zero, and solve for L, $\frac{d\big(L(y|L)\big)}{dL} = 0$:

$$-\frac{(\sum_{i=1}^{n}(-y_i - 1))(e^L)}{e^L + 1} - \sum_{i=1}^{n} y_i = 0$$

$$-\frac{(\sum_{i=1}^{n}(-y_i - 1))(e^L)}{e^L + 1} = \sum_{i=1}^{n} y_i$$

$$-\left(\sum_{i=1}^{n}(-y_i - 1)\right)(e^L) = (e^L + 1)\sum_{i=1}^{n} y_i$$

$$\left(\sum_{i=1}^{n} y_i\right) + \left(\sum_{i=1}^{n} 1\right) = \sum_{i=1}^{n} y_i + \frac{\sum_{i=1}^{n} y_i}{e^L}$$

$$\left(\sum_{i=1}^{n} y_i\right) + \sum_{i=1}^{n} 1 - \sum_{i=1}^{n} y_i = \frac{\sum_{i=1}^{n} y_i}{e^L}$$

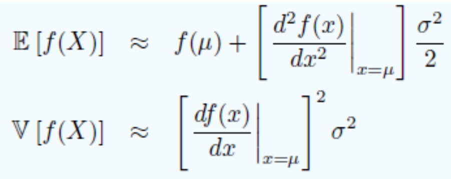
$$e^L = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} 1}$$

$$\hat{L} = \ln \sum_{i=1}^{n} y_i - \ln \sum_{i=1}^{n} 1$$

The maximum likelihood estimator $\hat{L}$ is $\ln \sum_{i=1}^{n} y_i - \ln\ \sum_{i=1}^{n} 1$

5) Since it is said that the random variable follows a geometric distribution with log-odds L, therefore $\hat{L}(Y) = f(Y) = \ln(Y)$. Given that $\mu = E[Y] = e^L$ and $\sigma^2 = V[Y] = e^L(e^L + 1)$, to find the approximate bias and variance, we first need to find $E[f(Y)]$ and $V[f(Y)]$ using the Taylor series approach.

Formula we are going to use:



$$\begin{aligned} \mathbb{E}\left[f(X)\right] &\approx f(\mu) + \left[\left. \frac{d^2 f(x)}{dx^2} \right|_{x=\mu} \right] \frac{\sigma^2}{2} \\ \mathbb{V}\left[f(X)\right] &\approx \left[\left. \frac{df(x)}{dx} \right|_{x=\mu} \right]^2 \sigma^2 \end{aligned}$$

By substituting our data into the formula above, we can get $E[f(Y)]$ and $V[f(Y)]$:

$$E[f(Y)] \approx \ln(e^L) + \left(-\frac{1}{(e^L)^2}\right)\left(\frac{e^L(e^L+1)}{2}\right)$$

$$E[f(Y)] \approx L - \left(\frac{e^L+1}{2e^L}\right)$$

$$V[f(Y)] \approx \left(\frac{1}{e^L}\right)^2 (e^L)(e^L + 1)$$

$$V[f(Y)] \approx \frac{e^L+1}{e^L}$$

The bias of an estimator is given by $b_\theta(\hat{\theta}) = E[\hat{\theta}(Y)] - \theta$.

$$b_L(\hat{L}) = E[f(Y)] - L$$

$$b_L(\hat{L}) \approx L - \left(\frac{e^L + 1}{2e^L}\right) - L$$

$$b_L(\hat{L}) \approx -\frac{e^L + 1}{2e^L}$$

The variance of an estimator is given by $Var_\theta(\hat{\theta}) = E\left[\left(\hat{\theta}(Y) - E[\hat{\theta}(Y)]\right)^2\right] = V[\hat{\theta}(Y)]$

$$Var_L(\hat{L}) = V[f(Y)] \approx \frac{e^L + 1}{e^L}$$

Question 3

1) From the given data, we know that there are 26 days (n_f) which fells on full moon and among that, 11 days (m_f) are above the average number of dog bite admissions. So, the estimate of the probability of a full moon day experiencing an above average number of dog bite admissions is

$$\hat{\theta}_f = \frac{m_f}{n_f} = \frac{11}{26} \approx \ 0.4230769$$

Then, we can find an approximate 95% confidence interval for this estimate by using the formula $\left(\hat{\theta} - z_{\alpha/2}\sqrt{v(\hat{\theta})/n}\,,\ \hat{\theta} + z_{\alpha/2}\sqrt{v(\hat{\theta})/n}\right)$ by taking $\propto = 1 - 0.95 = 0.05$:

$$\left(\frac{11}{26} - 1.96\sqrt{\frac{\frac{11}{26}\left(1 - \frac{11}{26}\right)}{26}}, \frac{11}{26} + 1.96\sqrt{\frac{\frac{11}{26}\left(1 - \frac{11}{26}\right)}{26}}\right)$$

which gives us a result of $(0.2331712, 0.6129826) \approx (0.2332, 0.6130)$. Therefore, we are 95% confident that the true population probability of a full moon day experiencing an above average number of dog bite admissions lies between 0.2332 and 0.6130.

2) To test the hypothesis that there is no difference in the probability of experiencing an above average number of dog bite admissions between full moon and non-full moon days, i.e., we set the null hypothesis to be: probability of experiencing an above average number of dog bite admissions between full moon and non-full moon days are the same. It is clarified that the proportion of non-full moon days that experienced an above average number of dog bite admissions is 0.53 and can be treated as a constant, known without error. Therefore, we want to test:

$$H_0 \colon \theta_f = \theta_0$$

$$vs$$

$$H_A \colon \theta_f \neq \theta_0$$

where θ_f is the probability of experiencing an above average number of dog bite admissions during full moon and θ_0 is the proportion of non-full moon days that experienced an above average number of dog bite admissions, which is 0.53. We know that $\hat{\theta}_f = \frac{11}{26}$ as calculated in Question 3 (1). Using all of the data collected, we can obtain an approximate z-score to use as our test statistic: $z_{\hat{\theta}_f} = \frac{\hat{\theta}_f - \theta_0}{\sqrt{\theta_0(1-\theta_0)/n}}$,

$$z_{\hat{\theta}_f} = \frac{\frac{11}{26} - 0.53}{\sqrt{\frac{0.53(1-0.53)}{26}}} = -1.092374 \approx -1.0924$$

Then, we can use the R command below to find the p-value:

```
2*pnorm(-1.0924)
```

which gives us an answer of

$$2P\left(Z < -\left|z_{\hat{\theta}_f}\right|\right) = 2P(Z < -1.4498) = 0.2746573 \approx 0.2747$$

A p-value of 0.2747 means that if the probability of experiencing an above average number of dog bite admissions between full moon and non-full moon days are the same, and we observe it for 26 days (sample size $n_f = 26$), almost 27.47% of the time these 26 days will result in a number of four or more (above average) dog bite admissions. We have no evidence against the null, which means that we are not able to disprove the null (probability of experiencing an above average number of dog bite admissions between full moon and non-full moon days are the same). But this does not prove that the null hypothesis is true ($\theta_f = \theta_0$) because we can only collect evidence against the null, never in favour of the null. In conclusion, this p-value says that the data we have observed is not at all at odds with the null distribution, and therefore offers no evidence against the null.

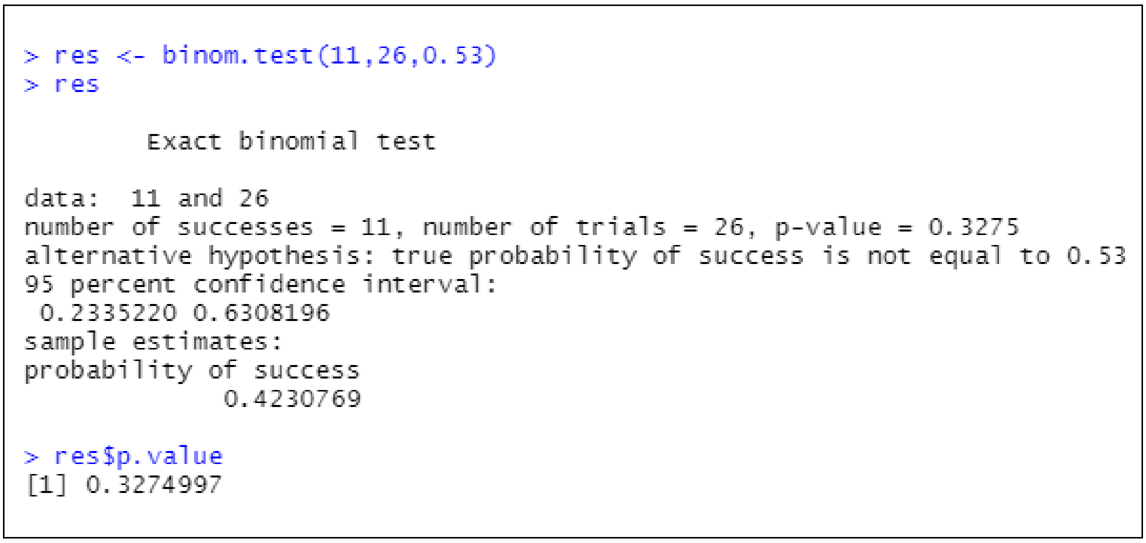
3) To calculate the exact p-value for the above hypothesis, the R command below is used:

```
res <- binom.test(11,26,0.53)

res

res$p.value
```

And this gives us the result as shown below:



```
> res <- binom.test(11,26,0.53)
> res

        Exact binomial test

data:  11 and 26
number of successes = 11, number of trials = 26, p-value = 0.3275
alternative hypothesis: true probability of success is not equal to 0.53
95 percent confidence interval:
 0.2335220 0.6308196
sample estimates:
probability of success
             0.4230769

> res$p.value
[1] 0.3274997
```

We can see the exact p-value is 0.3275, which is a bit larger than our approximate procedure, but gives the same overall conclusion. If the sample size was larger, we would expect the two p-values to be closer, as the normal approximation on which our approximate method is based would be better. This p-value also suggests that we have no evidence to disprove the null hypothesis: there is no difference in the probability of experiencing an above average number of dog bite admissions between full moon and non-full moon days. But this does not prove that the null hypothesis is true $(\theta_f = \theta_0)$ because we can only collect evidence against the null, never in favour of the null. In conclusion, this p-value says that the data we have observed is not at all at odds with the null distribution, and therefore offers no evidence against the null.

4) To test the hypothesis that the probability of experiencing an above average number of dog bite admissions does not differ between days falling on the new moon and the full moon, we set the null hypothesis to be: probability of experiencing an above average number of dog bite admissions between full moon and new moon days are the same.

$$H_0: \theta_f = \theta_m$$
$$vs$$
$$H_A: \theta_f \neq \theta_m$$

where θ_f and θ_m are the probability of experiencing an above average number of dog bite admissions during full moon and new moon respectively. Now, we know that $\hat{\theta}_f = \frac{11}{26}$ and we need to find $\hat{\theta}_m$.

It is given that there were 26 days (sample size = $n_m = 26$) that fell on a new moon, and of these 20 experienced an above average number of dog bite admissions (m_m).

Now, substitute m_m and n_m into the formula below to find $\hat{\theta}_m$.

$$\hat{\theta}_m = \frac{m_m}{n_m}$$
$$\hat{\theta}_m = \frac{20}{26} \approx 0.7692308$$

Under the null hypothesis, we assumed that $\theta_f = \theta_m$, so we can use a pooled estimate of θ now:

$$\hat{\theta}_p = \frac{m_f + m_m}{n_f + n_m}$$
$$\hat{\theta}_p = \frac{11 + 20}{26 + 26}$$
$$\hat{\theta}_p = \frac{31}{52}$$

Now, let's define the test statistic of the difference in the probability of experiencing an above average number of dog bite admissions between full moon and new moon days using this formula: $z_{\hat{\theta}_f - \hat{\theta}_m} = \frac{\hat{\theta}_f - \hat{\theta}_m}{\sqrt{\hat{\theta}_p(1-\hat{\theta}_p)(1/n_f + 1/n_m)}}$,

$$z_{\hat{\theta}_f-\hat{\theta}_m} = \frac{\frac{11}{26}-\frac{20}{26}}{\sqrt{\frac{31}{52}\left(1-\frac{31}{52}\right)\left(\frac{1}{26}+\frac{1}{26}\right)}} = -2.543629 \approx -2.5436$$

Then, we can use the R command below to find the p-value of $2P\left(Z < -\left|z_{\hat{\theta}_f-\hat{\theta}_m}\right|\right) = 2P(Z < -2.5436)$:

```
2*pnorm(-2.5436)
```

which gives us an answer of

$$2P\left(Z < -\left|z_{\hat{\theta}_f-\hat{\theta}_m}\right|\right) = 2P(Z < -2.5436) = 0.01097166 \approx 0.0110$$

A p-value of 0.0110means that if the null was true, the probability of experiencing an above average number of dog bite admissions between full moon and new moon days are the same, and we observe it for 26 days (sample size = 26), almost 1.1% of the time these 26 days will result in a number of four or more (above average) dog bite admissions. If the null was true, almost none of the samples (1.1% of the samples) would lead to a z-score as large or larger than the one we have observed. The p-value suggests that we have moderate to strong evidence against the null, which means that we are able to disprove that there is no difference in the probability of experiencing an above average number of dog bite admissions between full moon and new moon days. In conclusion, the data we have observed is at all at odds with the null distribution, and therefore offers evidence against the null and therefore, the alternative hypothesis will be concluded: there is difference in the probability of experiencing an above average number of dog bite admissions between full moon and new moon days.