Question1

Q1)

1.1) $\bar{x} = 10.467$ } by using R
$\sigma^2 = 7.559$
$\sigma = 2.749$

For CI with unknown mean and unknown variance, we use:

$$\mu \in \left(\bar{x} - t_{\alpha/2, n-1} \frac{\sigma}{\sqrt{n}}, \bar{x} + t_{\alpha/2, n-1} \frac{\sigma}{\sqrt{n}}\right)$$

$95\% \Rightarrow \frac{\alpha}{2} = 0.025$

$df = n-1 = 19$

$t_{\alpha/2, n-1} = t_{0.025, 19} = 2.09302$ (by using t table)

Then,

$$\mu \in \left(10.467 - (2.09302)\frac{2.749}{\sqrt{20}}, 10.467 + (2.09302)\frac{2.749}{\sqrt{20}}\right)$$
$$= (9.18, 11.75)$$

∴ The estimated mean fuel efficiency of vehicles that are all-wheel driver (sample size $n = 20$) is 10.467. We are 95% confident the population mean fuel efficiency for this group is between 9.18 and 11.75.

1.1
R code:

```
fueldb <- read.csv("fuel.efficiency.csv", header = TRUE)

print(head(fueldb))

a_fuel <-fueldb[fueldb$Type == "A","FA"]

a_mean <- mean(a_fuel)

a_var <- var(a_fuel) #default is the unbiased variance

a_sd <- sqrt(a_var)

a_df <- length(a_fuel)-1

#alpha/2 = 0.025, df = 20-1 = 19

# t = 2.09302

low_a_ci <- a_mean - (2.09302 * (a_sd/sqrt(a_df+1)))

high_a_ci <- a_mean+ (2.09302 * (a_sd/sqrt(a_df+1)))
```

1.2) $\hat{\mu}_A = 10.467$

$\hat{\mu}_P = 8.772$

$\hat{\sigma}^2_A = 7.559$

$\hat{\sigma}^2_P = 9.387$

$\Big\}$ by using R

$n_A = 20$

$n_B = 25$

95% => 1.96 in value

difference in fuel efficiency between A and P:

$$\hat{\mu}_A - \hat{\mu}_P = 10.467 - 8.772 = 1.695$$

interval:

$$\left(\hat{\mu}_A - \hat{\mu}_B - z_{\alpha/2}\sqrt{\frac{\hat{\sigma}^2_A}{n_A} + \frac{\hat{\sigma}^2_B}{n_B}} \; , \; \hat{\mu}_A - \hat{\mu}_B + z_{\alpha/2}\sqrt{\frac{\hat{\sigma}^2_A}{n_A} + \frac{\hat{\sigma}^2_A}{n_B}} \right)$$

hence,

$$\left(1.695 - 1.96\sqrt{\frac{7.559}{20} + \frac{9.387}{25}} \; , \; 1.695 + 1.96\sqrt{\frac{7.559}{20} + \frac{9.387}{25}} \right)$$

$$= \left(\cancel{-0.006287} \; , \; 3.3963 \right) \#$$

-0.00576

∴ The estimated difference in mean fuel efficiency between all-wheel drive vehicles (sample size n = 20) and part-time four-wheel drive vehicles (sample size n = 25) is 1.695. We are 95% confident the population mean difference in fuel efficiency is between -0.006287 up to 3.3963. As the interval includes zero, we cannot rule out the possibility of there being no difference at a population level between all-wheel drive vehicles and part-time four-wheel drive vehicles.

1.2

R code:

```
p_fuel <-fueldb[fueldb$Type == "P", "FA"]

p_mean <- mean(p_fuel)

p_var <- var(p_fuel) #default is the unbiased variance

# alpha/2 in z score = 1.96 (95%)

low_mean_diff_ci <- (a_mean - p_mean) -
1.96*(sqrt((a_var/length(a_fuel))+(p_var/length(p_fuel))))

high_mean_diff_ci <- (a_mean - p_mean) +
1.96*(sqrt((a_var/length(a_fuel))+(p_var/length(p_fuel))))
```

1.3)

$H_0 : \mu_A < \mu_p$

$H_A : \mu_A \geq \mu_p$

$\hat{\mu}_A = 10.467$ $\quad n_A = 20$

$\hat{\mu}_p = 8.772$ $\quad n_p = 25$

$\hat{\sigma}^2_A = 7.559$

$\hat{\sigma}^2_p = 9.387$

we know that $\sigma^2_A \neq \sigma^2_p$

we get test statistic :

$$z(\hat{\mu}_A - \hat{\mu}_p) = \frac{\hat{\mu}_A - \hat{\mu}_p}{\sqrt{\frac{\hat{\sigma}^2_A}{n_A} + \frac{\hat{\sigma}^2_p}{n_p}}}$$

$$\Rightarrow \quad z(\hat{\mu}_A - \hat{\mu}_p) = \frac{10.467 - 8.772}{\sqrt{\frac{7.559}{20} + \frac{9.387}{25}}}$$

$$= \frac{1.695}{0.868}$$

$$= 1.9528$$

$$p = 1 - P(Z < z(\hat{\mu}_A - \hat{\mu}_p))$$

$$= 1 - 0.9746$$

$$= 0.0254$$

∴ The p-value is smaller than the significance level 0.05 (default). There is enough evidence to reject the null hypothesis that all-wheel-drives are less efficient than part-time four-wheel-drive vehicles.

1.3

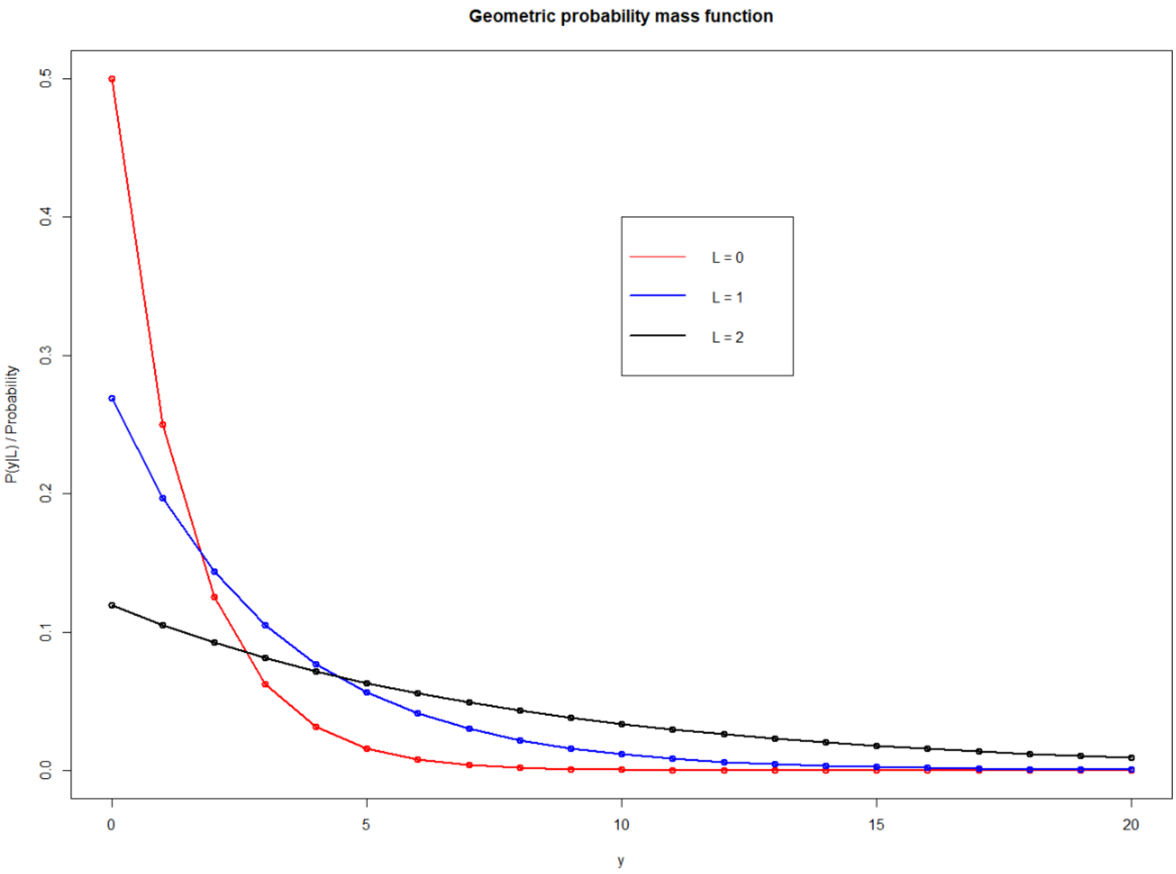
R code:

```
my_z <- (a_mean - p_mean) / (sqrt((a_var/length(a_fuel)) +
(p_var/length(p_fuel))))

z_p = 1-pnorm(my_z)

pnorm(1.9528)
```

1.4) For the estimated difference in fuel, we cannot rule out the possibility of there being no difference at population level between all-wheel drive vehicles and part-time four-wheel drive vehicles. This is because the interval ~~ine~~ includes zero. For estimated p-value, it's only an estimated p-value and the method we used is very close to true p value with the size of population being moderate only. If the size of population grows larger the closure of p-value will be lower.

Question2

2.1)



2.1

R code:

```
cal_prob <- function(L,y){
  res <- ((exp(L) + 1)^(-y-1)) * exp(y*L)
  return(res)
}
```

```
my_y <- seq(0,20,1)

prob1 <- cal_prob(0,my_y)

plot(my_y, prob1, col = "red", lwd=2.5, type ="o", xlab = "y", ylab
= "P(y|L) / Probability", main = "Geometric probability mass
function" )

prob2 <- cal_prob(1,my_y)

lines(my_y, prob2, col = "blue", lwd=2.5, type ="o")

prob3 <- cal_prob(2, my_y)

lines(my_y, prob3, col = "black",lwd=2.5, type ="o")

legend(x=10,y=0.4,c("L = 0","L = 1","L = 2"), lty=c(1,1,1),

       pch=c("","",""), col=c("red","blue","black"),
lwd=c(1,2.5,2.5))
```

* for entire Q2, $\log e = \log_e e$, which is equal to 1

Q2) Let L_s be the parameter: L, the log-odds of seeing a failure (tail) when the coin is tossed.

2.2) joint probability:

$$P(y|L_s) = \prod_{i=1}^{n} P(y_i|L_s)$$

$$= (e^{L_s}+1)^{-y_1-1} e^{y_1 L_s} \cdot (e^{L_s}+1)^{-y_2-1} e^{y_2 L_s} \cdots (e^{L_s}+1)^{-y_n-1} e^{y_n L_s}$$

$$= (e^{L_s}+1)^{-y_1-y_2-\cdots-y_n} (e^{L}+1)^{-n} e^{nL} \cdot e^{y_1+y_2+\cdots+y_n}$$

$$= \frac{e^{nL_s}}{(e^{L_s}+1)^n} \cdot \sum_{i=1}^{n} \frac{1}{(e^{L_s}+1)^{y_i}} \cdot \sum_{i=1}^{n} e^{y_i}$$

$\therefore \frac{e^{nL}}{(e^L+1)^n} \cdot \sum_{i=1}^{n} \frac{1}{(e^L+1)^{y_i}} \cdot \sum_{i=1}^{n} e^{y_i}$, where L is the parameter #

2.3)

$$L(y|L_s) = -\log P(y|L_s)$$

$$= -\log \left[\frac{e^{nL_s}}{(e^{L_s}+1)^n} \cdot \sum_{i=1}^{n} \frac{1}{(e^{L_s}+1)^{y_i}} \cdot \sum_{i=1}^{n} e^{y_i} \right]$$

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

$$= \log (e^{L_s}+1)^n - \log e^{nL_s} + \sum_{i=1}^{n} \log (e^{L_s}+1)^{y_i} - \sum_{i=1}^{n} y_i$$

$$= n \log (e^{L_s}+1) - nL_s + \sum_{i=1}^{n} y_i \log (e^{L_s}+1) - \sum_{i=1}^{n} y_i$$ #

$\therefore n \log (e^L+1) - nL + \sum_{i=1}^{n} y_i \log (e^L+1) - \sum_{i=1}^{n} y_i$, where L is the ~~ete~~ parameter #

2.4) maximum likelihood estimator $\hat{L}$ for L :

let L_{lg} be the negative log-likelihood

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

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

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

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

Set derivative to 0 :

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

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

$e^L = -1$; $-ne^L = -\sum_{i=1}^{n} y_i$

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

hence, rejected ; $L(\log e) = \log \frac{1}{n}\sum_{i=1}^{n} y_i$

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

2.5)

Assumptions : the population L

i) have a mean of $E[y_i] = \mu$

ii) have a variance of $V[y_i] = \sigma^2$

iii) and are independent

bias :

$$\cancel{b_L(\hat{L}) = E[\hat{L}(y)] - L\mu}$$

$$\cancel{Var_L(\hat{L}) = E[(\hat{L}(y) - E[\hat{L}(y)])^2] = V[\hat{L}(y)]}$$

$$\cancel{= \frac{\sigma^2}{n}}$$

$$\cancel{= \frac{e^L(e^L+1)}{n}}$$

Let $X = \log \sum_{i=1}^{n} y_i - \log \sum_{i=1}^{n} (1)$, $X \sim e^L$, $E(X) = e^L$, $V(X) = e^L(e^L+1)$

$$f(X) = \log X$$

$$\frac{df(X)}{dX} = \frac{1}{X}$$

$$\frac{d^2 f(X)}{dX^2} = -\frac{1}{X^2}$$

$$E[f(X)] = f(\mu_X) + \left[-\frac{1}{x^2}\Big|x=\mu_X\right]\frac{\sigma^2_X}{2}$$

$$= f(\mu_X) + \left[-\frac{1}{\mu_X^2}\right]\frac{\sigma^2 X}{2}$$

$$= \log e^L + \left[-\frac{1}{e^{2L}}\right]\frac{e^L(e^L+1)}{2}$$

$$= L + \left[-\frac{1}{e^{2L}}\right]\frac{e^{2L}+e^L}{2}$$

$$= L + \left[-\frac{1\cdot(e^{2L}+e^L)}{e^{2L}\cdot 2}\right]$$

$$= L - \frac{e^L+1}{2e^L}$$

Bias :

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

$$= E[\log(\tfrac{1}{n}\sum_{i=1}^{n} y_i)] - L$$

$$= E[\log(X)] - L$$

$$= L - \frac{e^L+1}{2e^L} - L$$

$$= -\frac{e^L+1}{2e^L}$$ #

$$V[f(X)] = \left[\frac{df(x)}{dx}\Big|x=\mu_X\right]^2 \sigma^2_X$$

$$= \left(\frac{1}{e^L}\right)^2 e^L(e^L+1)$$

$$= \frac{e^L+1}{e^L}$$

Variance :

$$V_L(\hat{L}) = V[\log(\tfrac{1}{n}\sum_{i=1}^{n} y_i)] - L$$

$$= V[f(X)]$$

$$= \frac{e^L+1}{e^L}$$ #

Question3

Q3

3.1) $X \sim Be(\theta)$, θ is the probability of a full moon day experiencing an above average number of dog bite admissions

$n = 26$

$\hat{\theta} = \frac{11}{26}$

$V[X] = \left(\frac{11}{26}\right)\left(1 - \frac{11}{26}\right)$

$= 0.244 \ / \ \frac{165}{676}$

95% => 1.96 in value

Confidence Interval:

$$\left(\frac{11}{26} - 1.96\sqrt{\frac{\frac{165}{676}}{26}} \ , \ \frac{11}{26} + 1.96\sqrt{\frac{\frac{165}{676}}{26}} \right)$$

$= (0.233 \ , \ 0.613)$ #

∴ The estimated probability of a full moon day experiencing an above average number of dog bite admissions (sample size $n=26$) is 0.244. We are 95% confident the population probability dog bite admissions for this group is between 0.233 and 0.613.

3.1
R code:
#3.1

```
my_var <- 11/26 * (1-11/26)

my_lowci <- 11/26 - 1.96*(sqrt(my_var/26))

my_highci <- 11/26 + 1.96*(sqrt(my_var/26))
```

3.2)

let A : number of dog bite admissions of full moon days

let B : number of dog bite admissions of non-full moon days

$H_0 : \theta_A = \theta_B$

$H_A : \theta_A \neq \theta_B$

$\theta_A = 11/26$

$\theta_B = 0.53$

$$Z_{\hat{\theta}} = \frac{11/26 - 0.53}{\sqrt{\frac{0.53(1-0.53)}{26}}}$$

$$= \frac{-139/1300}{0.09788}$$

$$= -1.0924$$

$$p = 2P(Z < -|z_{\hat{\theta}}|)$$
$$= 2P(Z < -1.0924)$$
$$= 2(0.13733) \quad \text{by using pnorm(-1.0924) in R}$$
$$= 0.274$$

∴ The p-value is greater than the significant level 0.05 (default). There is insufficient evidence to reject the null hypothelis that there's no difference in the probability of experiencing an above average number of dog bite admissions between full moon and non-full moon days.

3.2

R code:

```
the_z <- (11/26 - 0.53)/sqrt(0.53*(1-0.53)/26)

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

3.3) To find exact p-value:

binom.test (x=11, n=26, 0.53)

$p = 0.3275$ #

∴ The exact p-value is slightly greater than the approximated p-value. Yet the p-value is still greater than the significant level 0.05 (default). There's insufficient evidence to reject the null hypothesis that there's no difference in the probability of experiencing an above average number of dog bite admissions between full moon and non-full moon days.

3.4) let A = number of dogbite admissions of full moon days
let B = number of dogbite admissions of new moon days

$H_0: \theta_A = \theta_B$
$H_A: \theta_A \neq \theta_B$

we use a pooled estimate of θ:

$$\hat{\theta}_p = \frac{11+20}{26+26} = \frac{31}{52} / 0.5962$$

$$Z(\hat{\theta}_A - \hat{\theta}_B) = \frac{11/26 - \frac{20}{26}}{\sqrt{\frac{31}{52}(1-\frac{31}{52})(\frac{1}{26}+\frac{1}{26})}}$$

$$= \frac{-9/26}{\sqrt{(0.24081)\frac{2}{26}}}$$

$$= -2.5437$$

$$p = 2P(Z < -|z(\hat{\theta}_A - \hat{\theta}_B)|)$$
$$= 2P(Z < -2.5437)$$
$$= 2\,\cancel{P}\,(0.005485)$$
$$= 0.011$$ #

∴ The p-value is smaller than the significant level 0.05 (default) there is enough evidence to reject the null hypothesis that the probability of experiencing an above average number of dogbite admissions does not differ between days falling on the new moon and full moon.

3.4

R code:

```
my_Z <- (11/26 - 20/26)/(sqrt(31/52*(1-31/52)*2/26))

my_P <- 2*pnorm(-2.543629)
```