1. [13 marks] Between 10:00 am September 26 and 10:00 pm September 28 2016 the Federation of Students at the University of Waterloo conducted a referendum. The question was:

Which one of the following options do you support?

(1) Keep the mandatory, refundable \$4.75 per academic term fee for WPIRG (Waterloo Public Interest Research Group).

(2) Remove the mandatory, refundable \$4.75 per academic term fee for WPIRG (Waterloo Public Interest Research Group).

All eligible undergraduates were informed by email to cast their ballot online. Of the $31,380$ eligible voters, 8788 voted, and 7156 chose option 2.

(*a*) [3] The Federation of Students used an empirical study to determine whether or not students supported the removal of the WPIRG fee. The Plan step of the empirical study involved using an online referendum. Using complete sentences give at least one advantage and at least one disadvantage of using the online referendum in this context.

The respondents to the survey are students who heard about the online referendum and then decided to vote in the referendum. These students may not be representative of all students at the University of Waterloo. For example, it is possible that the students who took the time to vote are also the students who most want the to remove the WPIRG fee. Students who don't care about the WPIRG fee probably did not bother to vote. This is an example of sample error. Any online survey such as this **online referendum has the disadvantage that the sample of people who choose to vote are not necessarily a representative sample of the study population of interest**. The **advantage of online surveys is that they are inexpensive and easy to conduct.** To obtain a representative sample you would need to select a random sample of all students at the University of Waterloo. Unfortunately taking such a sample would be much more time consuming and costly than conducting an online referendum.

(*b*) [3] Assume the model $Y \backsim Binomial\,(8788, \theta)$ where $Y =$ number of students who chose option (2): "Remove the mandatory, refundable \$4.75 per academic term fee for WPIRG." What does the parameter θ represent in this study? Using complete sentences indicate how valid you think the Binomial model is and why?

The parameter θ represents the proportion of the $31,380$ eligible undergraduate voters (the study population) who support option (2).

Important Note: The parameters in the model are always related to attributes of interest in the **study population** not in the sample.

A Binomial model assumes independent trials (students) which might not be a valid assumption. For example, if groups of students, say within a specific faculty, all got together and voted, their responses may not be independent events.

Since a student may not vote more than once, the sample of 8788 students is actually drawn without replacement from the finite population of $31,380$ students. If the sample was drawn at random (it was not) then we could justify the Binomial model using the Binomial approximation to the Hypergeometric.

(c) [1] The maximum likelihood estimate of θ based on the observed data is

_______0.814_______. (You do not need to derive this estimate.)

$$\hat{\theta} = \frac{7156}{8788} = 0.814292$$

(d) [3] The $p - value$ for testing the hypothesis $H_0 : \theta = 0.8$ is approximately

_______0.001_______ . **Show all your steps.**

To test $H_0 : \theta = 0.8$ we use the test statistic

$$D = |Y - n\theta_0| = |Y - (8788)(0.8)| = |Y - 7030.4|$$

with observed value

$$d = |7156 - 7030.4| = 125.6$$

and

$$\begin{aligned} p - value &= P(D \geq d; H_0) \\ &= P(|Y - 7030.4| \geq 125.6) \quad \text{if} \quad Y \backsim Binomial(8788, 0.8) \\ &\approx P\left(|Z| \geq \frac{125.6}{\sqrt{8788(0.8)(1-0.8)}}\right) \quad \text{where} \quad Z \backsim G(0,1) \\ &= 2[1 - P(Z \leq 3.35)] \\ &= 2(1 - 0.9996) \\ &= 0.0008 \end{aligned}$$

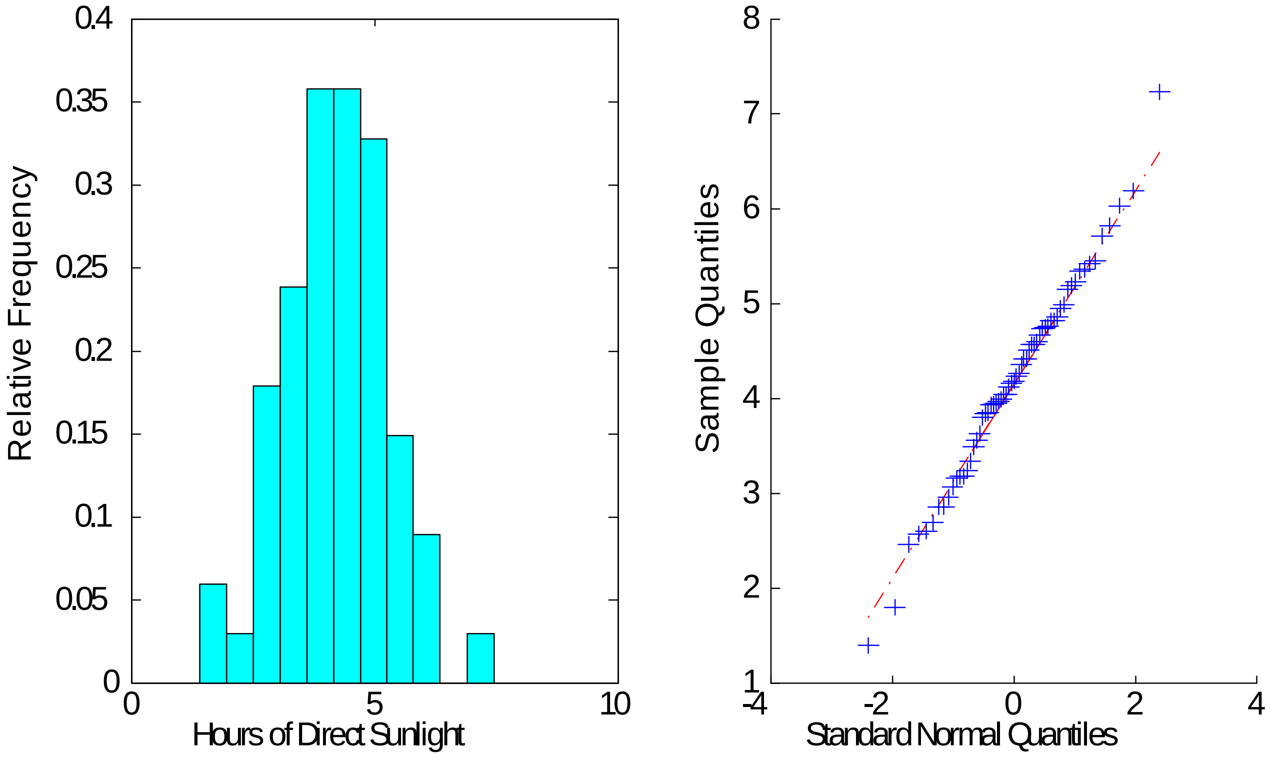
(e) [1] State your conclusion regarding the hypothesis $H_0 : \theta = 0.8$ in a sentence.

Since the approximate $p - value$ for testing $H_0 : \theta = 0.8$ is less than 0.001 we would conclude that, based on the data, there is very strong evidence against the hypothesis $H_0 : \theta = 0.8$.

(f) [2] **By reference to your answer in** (d), indicate whether the value $\theta = 0.8$ is inside an approximate 95% confidence interval or not. Justify your answer but do not construct the interval.

Since the approximate $p - value$ for testing $H_0 : \theta = 0.8$ is less than 0.001 which is less than 0.05 then we know that the value $\theta = 0.8$ is not inside an approximate 95% confidence interval.

2. [16 marks] To decide whether to install solar panels on the roof of her house a homeowner records the number of hours of full sunlight on her roof for 61 consecutive days in June and July. A relative frequency histogram and a qqplot for these data are given below:



Let y_i = number of hours of full sunlight on the i'th day, $i = 1, 2, \ldots, 61$. For these data

$$\sum_{i=1}^{61} y_i = 255.28 \quad \text{and} \quad \sum_{i=1}^{61} (y_i - \bar{y})^2 = 71.5607$$

To analyze these data the model $Y_i \backsim G(\mu, \sigma)$, $i = 1, 2, \ldots, 61$ is assumed.

(a) [2] Using complete sentences indicate how reasonable the Gaussian model is for these data.

Since the relative frequency histogram looks reasonably bell-shaped and the points in the qqplot lie reasonably along a straight line, the Gaussian model seems reasonable for these data.

(b) [2] In a complete sentence explain clearly what the parameter μ represents.

The parameter μ represents the mean number of hours of full sunlight in a day on the homeowner's roof over a year which is the study population.

Important Note: The parameters in the model are always related to attributes of interest in the **study population** not in the sample.

(c) [2] For these data the maximum likelihood estimate of μ is ____4.185____ $\frac{255.28}{61} = 4.18492$

and the maximum likelihood estimate of σ is ____1.083____. $\hat{\sigma} = \left[\frac{1}{61}(71.5607)\right]^{1/2} = 1.0831095$
(You do not need to derive these estimates.)

(e) [4] A 99% confidence interval for μ based on these data is (**show all your steps**):

[3.813, 4.557]

Note that

$$s = \left[\frac{1}{60}(71.5607)\right]^{1/2} = 1.092098$$

Since $P(T \leq 2.6603) = 0.995$ where $T \backsim t(60)$ a 99% confidence interval for μ is given by

$$\begin{aligned} & \bar{y} \pm 2.6603 s/\sqrt{61} \\ = & \frac{255.28}{61} \pm 2.6603\,(1.092098)\,/\sqrt{61} \\ = & 4.18492 \pm 0.371987 \\ = & [3.812931, 4.556905] \end{aligned}$$

(f) [3] The Solar Energy Association recommends that the average number of hours of full sunlight in a day should be at least 4 to generate enough energy to make solar panels worthwhile. Using complete sentences indicate what the homeowner should conclude about whether or not it is worthwhile placing solar panels on the roof of her house. Note any limitations of her study.

The point estimate of μ is $\hat{\mu} = \bar{y} = 4.185$ which is greater than 4. However the 99% confidence interval for μ is $[3.813, 4.557]$ which contains values less than 4. Therefore based on the data there are values of μ which are less than 4 which are reasonable in light of the observed data. Since values of μ which are less than 4 are reasonable in light of the observed data and since the Solar Energy Association recommends that the average number of hours of full sunlight in a day should be at least 4, the landowner should conclude that there is not enough evidence to suggest placing solar panels on her roof.

Note also that she only took observations in two particular months (June and July) which may be the months with most sunlight. It would be a better idea to take measurements over the different months of the year in order to make an informed decision about whether solar panels are worthwhile.

(g) [3] The $p - value$ for testing $H_0 : \sigma = 1$ is between

0.2 and 0.4. **Show all your steps.**

We use the test statistic

$$U = \frac{(n-1)\,S^2}{\sigma_0^2} = \frac{60S^2}{(1)^2} = 60S^2 \backsim \chi^2\,(60) \quad \text{if} \quad H_0 : \sigma = 1 \quad \text{is true.}$$

The observed value is $u = 60s^2 = 71.5607$.

$$p - value = 2P\,(U \geq 71.5607) \quad \text{where} \quad U \backsim \chi^2\,(60)$$

From Chi-squared table

$$P\,(U \geq 68.972) = 1 - 0.8 = 0.2 \;\; \text{and} \;\; P\,(U \geq 74.397) = 1 - 0.9 = 0.1$$

Therefore $0.2 < p - value < 0.4$.

3. [16 marks] Suppose $Y \backsim \textit{Exponential}\,(\theta)$ with probability density function

$$f(y;\theta) = \frac{1}{\theta} e^{-y/\theta} \quad \text{for } y > 0 \text{ and } \theta > 0.$$

(a) [3] Use Change of Variable to show that $W = \frac{2Y}{\theta}$ has probability density function given by

$$g(w) = \frac{1}{2} e^{-w/2}, \qquad \text{for} \quad w > 0$$

which is the probability density function of a $\chi^2(2)$ random variable.

For $w \geq 0$,

$$\begin{aligned} G(w) &= P(W \leq w) = P\left(\frac{2Y}{\theta} \leq w\right) = P\left(Y \leq \frac{\theta w}{2}\right) \\ &= F\left(\frac{\theta w}{2}\right) \quad \text{where } F(y) = P(Y \leq y) \text{ is the c.d.f. of } Y \end{aligned}$$

Therefore

$$\begin{aligned} g(w) &= G'(w) = f\left(\frac{\theta w}{2}\right) \cdot \frac{d}{dw}\left(\frac{\theta w}{2}\right) = \frac{1}{\theta} e^{-\left(\frac{\theta w}{2}\right)/\theta} \cdot \left(\frac{\theta}{2}\right) \\ &= \frac{1}{2} e^{-w/2}, \quad \text{for } w \geq 0 \end{aligned}$$

as required.

(b) [2] Suppose $Y_1, Y_2, \ldots, Y_n$ is a random sample from the *Exponential* (θ) distribution. Use your result from (a) and theorem(s) that you have learned in class to show that

$$U = \sum_{i=1}^{n} \frac{2Y_i}{\theta} \sim \chi^2(2n).$$

Since the sum of independent Chi-squared random variables has a Chi-squared distribution with degrees of freedom equal to the sum of the degrees of freedom of the Chi-squared random variables in the sum, and since $\frac{2Y_i}{\theta} \sim \chi^2(2)$, $i = 1, 2, \ldots, n$ therefore

$$U = \sum_{i=1}^{n} \frac{2Y_i}{\theta} \sim \chi^2\left(\sum_{i=1}^{n} 2\right) \quad \text{or} \quad \chi^2(2n) \quad \text{as required.}$$

(c) [4] Explain clearly how the pivotal quantity U given in (b) can be used to obtain a two-sided, equal tailed, $100p\%$ confidence interval for θ.

Using Chi-squared tables find a and b such that $P(U \leq a) = \frac{1-p}{2}$ and $P(U \leq b) = \frac{1+p}{2}$ where $U \sim \chi^2(2n)$.

Since

$$p = P\left(a \leq \sum_{i=1}^{n} \frac{2Y_i}{\theta} \leq b\right) = P\left(\frac{1}{b} \leq \frac{\theta}{2\sum_{i=1}^{n} Y_i} \leq \frac{1}{a}\right) = P\left(\frac{2\sum_{i=1}^{n} Y_i}{b} \leq \theta \leq \frac{2\sum_{i=1}^{n} Y_i}{a}\right)$$

then a $100p\%$ confidence interval for θ is given by

$$\left[\frac{2\sum_{i=1}^{n} y_i}{b}, \frac{2\sum_{i=1}^{n} y_i}{a}\right].$$

(*d*) [2] Suppose $W \backsim \chi^2(20)$ and let a and b be such that $P(W \le a) = 0.025 = P(W \ge b)$.

Then $a =$ 9.591 and $b =$ 34.170. (Use all the decimal places from the tables.)

(*e*) [5] Suppose $y_1, y_2, \ldots, y_{10}$ is an observed random sample from the *Exponential* (θ) distribution with

$$\sum_{i=1}^{10} y_i = 62.4$$

(*i*) Using your results from (*c*) and (*d*), a 95% confidence interval for θ based on the pivotal quantity U is:

[3.652, 13.012]. Show your work.

$$\left[\frac{2\sum_{i=1}^{n} y_i}{b}, \frac{2\sum_{i=1}^{n} y_i}{a}\right] = \left[\frac{2(62.4)}{34.170}, \frac{2(62.4)}{9.591}\right] = [3.65232,\ 13.01220]$$

(*ii*) An approximate 95% confidence interval for θ based on the asymptotic Normal pivotal quantity is:

[2.372, 10.108]. Show your work. Compare this interval with the interval you obtained in (*i*).

An approximate 95% confidence interval for θ is given by

$$\begin{aligned}\bar{y} \pm 1.96\frac{\bar{y}}{\sqrt{n}} &= \frac{62.4}{10} \pm 1.96\frac{(62.4/10)}{\sqrt{10}} \\ &= 6.24 \pm 3.867592 \\ &= [2.372408,\ 10.107592]\end{aligned}$$

The intervals are quite different which is what you would expect since the result in (*i*) is exact while the result in (*ii*) is based on an approximation which is poor since $n = 10$ is small.

4. [10 marks] **Circle the letter corresponding to your choice.**

(a) Suppose $Y \backsim \mathit{Binomial}\,(n, \theta)$. An experiment is to be conducted in which data y are to be collected to estimate θ. To ensure that the width of the approximate 90% confidence interval for θ is no wider that $2\,(0.03)$, the sample size n should be at least:

A: 1068

B: 2401

[✓ **C:**] 752

D: 267

E: 188

(b) For a Binomial experiment the approximate 95% confidence interval for θ based on the asymptotic Normal pivotal quantity was 0.75 ± 0.05. Which statement is **TRUE**?

A: $P\,(\theta \in [0.7, 0.8]) = 0.95$.

B: The interval $[0.7, 0.8]$ is also a 15% likelihood interval.

C: We are 95% confident that $\theta = \hat{\theta}$.

[✓ **D:**] If the Binomial experiment was repeated 100 times independently and an approximate 95% confidence interval was constructed each time then approximately 95 of these intervals would contain the true value of θ.

E: None of the above.

(c) Which statement is **FALSE**?

A: A 15% likelihood interval is an approximate 95% confidence interval.

[✓ **B:**] A 10% likelihood interval is an approximate 90% confidence interval.

C: Likelihood intervals must usually be found numerically or from a graph of the relative likelihood function.

D: Likelihood intervals are as good or better than approximate confidence intervals based on asymptotic Normal pivotal quantities.

E: The likelihood ratio statistic is an asymptotic pivotal quantity.

(d) Suppose $Y_i \backsim G\,(\mu, \sigma)\,,\ i = 1, 2, \ldots, n$ independently. Let

$$T = \frac{\bar{Y} - \mu}{S/\sqrt{n}} \quad \text{where } \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \text{ and } S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$$

The distribution of T is:

A: $G(0, 1)$

B: $G(0, \sigma)$

C: $t\,(n)$

[✓ **D:**] $t\,(n-1)$

E: $\chi^2\,(n-1)$

(e) Suppose we have n observations from a $G(\mu, \sigma)$ distribution. Which statement is **TRUE**?

A: $\tilde{\mu}$ is a point estimate of μ.

B: If σ is known then $\bar{y} \pm 1.645\sigma/\sqrt{n}$ is a 95% confidence interval for μ.

[✓ **C:**] If σ is unknown then $\bar{y} \pm as/\sqrt{n}$ is a 95% confidence interval for μ if $P\,(T \le a) = 0.975$ and $T \backsim t\,(n-1)$.

D: If σ is unknown then $\bar{y} \pm as/\sqrt{n}$ is a 95% confidence interval for μ if $P\,(T \le a) = 0.95$ and $T \backsim t\,(n-1)$.

E: S is the maximum likelihood estimator of σ.

(f) Data are collected in an experiment to test the null hypothesis H_0 using the test statistic D. The $p-value$ for testing H_0 is equal to

A: the probability that the null hypothesis H_0 is true.

B: the probability that the alternative hypothesis H_A is true.

[✓ **C:**] the probability of obtaining a value of D as unusual or more unusual than the observed value of D if H_0 is true.

D: the probability of obtaining a value of D as unusual or more unusual than the observed value of D if the alternative hypothesis H_A is true.

E: None of the above.

(g) Which statement is **FALSE**?

[✓ **A:**] For Binomial data the likelihood ratio statistic is a continuous random variable.

B: The distribution of the likelihood ratio statistic based on a random sample $Y_1, Y_2, \ldots, Y_n$ is approximately $\chi^2(1)$ for large n.

C: For Exponential data, the likelihood ratio statistic is a continuous random variable.

D: For Binomial(n, θ) data, an approximate 95% confidence interval for θ based on the asymptotic Normal pivotal quantity can contain values outside the interval $[0, 1]$.

E: For Exponential(θ) data, an approximate 95% confidence interval for θ based on a 15% likelihood interval only contains values of θ greater than zero.

For questions (h) and (i), suppose that a data set is assumed to be a random sample from a $G(\mu, \sigma)$ distribution where μ and σ are unknown. Suppose also that the data set is stored in the variable y and the following code has been run in R:

```
ybar<-mean(y)
n<-length(y)
s2<-var(y)
s<-sqrt(s2)
```

(h) Which of the following R commands gives a 95% confidence interval for the mean μ?

A: c(ybar-qnorm(0.975,0,1)*s/sqrt(n),ybar+qnorm(0.975,0,1)*s/sqrt(n))

B: c(ybar-qnorm(0.95,0,1)*s/sqrt(n),ybar+qnorm(0.95,0,1)*s/sqrt(n))

C: c(ybar+qt(0.05,n-1)*s/sqrt(n),ybar+qt(0.95,n-1)*s/sqrt(n))

D: c(ybar-qt(0.975,n)*s/sqrt(n),ybar+qt(0.975,n)*s/sqrt(n))

[✓ **E:**] c(ybar-qt(0.975,n-1)*s/sqrt(n),ybar+qt(0.975,n-1)*s/sqrt(n))

(i) Which of the following R commands gives a 95% confidence interval for the standard deviation σ?

A: c(sqrt((n-1)*s2/qchisq(0.025,n-1)),sqrt((n-1)*s2/qchisq(0.975,n-1)))

[✓ **B:**] c(sqrt((n-1)*s2/qchisq(0.975,n-1)),sqrt((n-1)*s2/qchisq(0.025,n-1)))

C: c(sqrt((n-1)*s2/qchisq(0.05,n-1)),sqrt((n-1)*s2/qchisq(0.95,n-1)))

D: c(sqrt((n-1)*s2/qchisq(0.025,n)),sqrt((n-1)*s2/qchisq(0.975,n)))

E: c(sqrt((n-1)*s2/qchisq(0.975,n)),sqrt((n-1)*s2/qchisq(0.025,n)))

(j) Suppose that a data set is assumed to be a random sample from an *Exponential* (θ) distribution. Suppose also that the data set is stored in the variable y and the following code has been run in R:

```
thetahat<-mean(y)
n<-length(y)
```

Which of the following R commands does **NOT** give the observed value of the likelihood ratio statistic evaluated at theta for these data?

A: -2*log((thetahat/theta)^n*exp(n*(1-thetahat/theta)))

[✓ **B:**] 2*log((theta/thetahat)^n*exp(n*(1-theta/thetahat)))

C: -2*n*(log(thetahat/theta)+1-thetahat/theta)

[✓ **D:**] -2*log((thetahat/theta)^n*exp(n*(thetahat/theta-1)))

E: -2*n*log((thetahat/theta)*exp(1-thetahat/theta))