FIT2086 Assignment 2

Due Date: 11:55PM, Friday, 17/9/2021

1 Introduction

There are total of three questions worth $10+10+8=28$ marks in this assignment. This assignment is worth a total of 20% of your final mark, subject to hurdles and any other matters (e.g., late penalties, special consideration, etc.) as specified in the FIT2086 Unit Guide or elsewhere in the FIT2086 Moodle site (including Faculty of I.T. and Monash University policies).

Students are reminded of the Academic Integrity Awareness Training Tutorial Activity and, in particular, of Monash University's policies on academic integrity. In submitting this assignment, you acknowledge your awareness of Monash University's policies on academic integrity and that work is done and submitted in accordance with these policies.

Submission Instructions: Please follow these submission instructions:

1. No files are to be submitted via e-mail. Submissions are to be made via Moodle.
2. Please provide a single file containing your report, i.e., your answers to these questions. Provide code/code fragments as required in your report, and make sure the code is written in a fixed width font such as `Courier New` (or a screen shot is taken and inserted – please make sure this is neat and readable), or similar, and is grouped with the question the code is answering. You can submit hand-written answers, but if you do, please make sure they are clear and legible. Do not submit multiple files – all your files should be combined into a single PDF file as required. Please ensure that your assignment answers the questions in the order specified in the assignment. Multiple files and questions out of order make the life of the tutors marking your assignment much more difficult than it needs to be, and may attract penalties, so please ensure you assignment follows these requirements.

Question 1 (10 marks)

In this question we will revisit our analysis of the COVID-19 recovery data that we began in Assignment 1. The file `covid.19.ass2.csv` contains a subset of the New South Wales days-to-recovery data we examined previously; this time, patients with recovery times over four weeks (28 days) were removed as these recovery times are unusual and likely represent a sub-population of people more susceptible to the virus. We know from Assignment 1 that the Poisson distribution is not a good fit to the recovery data: instead, for this question we will use a normal distribution as it provides an improved fit to the data due to its increased flexibility, while accepting this assumption is also not necessarily correct; to quote the famous statistician G.E.P.Box: "*all models are wrong – but some are more useful than others*".

Important: you may use R to determine the means and variances of the data, as required, and the R functions `qt()` and `pnorm()` but you must perform all the remaining steps by hand. Please provide appropriate R code fragments and all working out.

1. Calculate an estimate of the average number of days to recovery using the provided data. Calculate a 95% confidence interval for this estimate using the t-distribution, and summarise/describe your results appropriately. Show working as required. **[4 marks]**

2. Similar data was collected in 2020 by the Israeli Ministry of Health. While the specific data was not available, the summary statistics were provided, and from these I have simulated a dataset of $n = 494$ individuals from the Israeli study. The days to recovery in this group are provided in the file `israeli.covid.19.ass2.csv`. Using the provided data and the approximate method for difference in means with (different) unknown variances presented in Lecture 4, calculate the estimated mean difference in recovery times between the Israeli patients and the patients from NSW, and provide an approximate 95% confidence interval. Summarise/describe your results appropriately. Show working as required. **[3 marks]**

3. It is of interest to determine if there are any differences, at a population level, in recovery times for patients in different countries. Test the hypothesis that the population average time taken to recover for the Israeli cohort is the same as in the NSW cohort. Write down explicitly the hypothesis you are testing, and then calculate a p-value using the approximate hypothesis test for differences in means with (different) unknown variances presented in Lecture 5. What does this p-value suggest about the difference in mean recovery time between the two cohorts of patients? **[3 marks]**

Question 2 (10 marks)

The Erlang distribution is a probability distribution for non-negative real numbers. It models the time taken for k events to occur, under the condition that the events occur at the same rate over time. It finds extensive use in telecommunications and medical research. The version that we will look at has a probability density function of the form

$$p(y \mid v, k) = \left(\frac{y^{k-1}}{(k-1)!} \right) \exp\left(-e^{-v} y - kv\right) \tag{1}$$

where $e^x \equiv \exp(x)$ and $y \in \mathbb{R}_+$, i.e., y can take on the values of non-negative real numbers. In this form it has two parameters: the "shape" $k > 0$, which is the number of events we are interested in waiting for, and $v \in \mathbb{R}$, which is the log-inverse-rate (also called a log-scale) of the distribution, which controls the rate at which the events we are modelling occur. Often k is not treated as a learnable parameter, but is rather set by the user depending on the context. If a random variable follows an Erlang distribution with log-inverse-rate v and shape k we say that $Y \sim \mathrm{Er}(v, k)$. If $Y \sim \mathrm{Er}(v, k)$, then $\mathbb{E}[Y] = k\, e^v$ and $\mathbb{V}[Y] = k\, e^{2v}$.

1. Produce a plot of the Erlang probability density function (1) for the values $y \in (0, 15)$, for $(v = 0,\, k = 1)$, $(v = 1,\, k = 1)$ and $(v = -1/2,\, k = 2)$. Ensure that the graph is readable, the axis are labelled appropriately and a legend is included. **[2 marks]**

2. Imagine we are given a sample of n observations $\mathbf{y} = (y_1, \ldots, y_n)$. Write down the joint probability of this sample of data, under the assumption that it came from an Erlang distribution with log-inverse-rate v and shape k (i.e., write down the likelihood of this data). Make sure to simplify your expression, and provide working. *(hint: remember that these samples are independent and identically distributed.)* **[2 marks]**

3. Take the negative logarithm of your likelihood expression and write down the negative log-likelihood of the data $\mathbf{y}$ under the Erlang model with log-inverse-rate v and shape k. Simplify this expression. **[1 mark]**

4. Derive the maximum likelihood estimator $\hat{v}$ for v, under the assumption that k is fixed; that is, find the value of v that minimises the negative log-likelihood, treating k as a fixed quantity. You must provide working. **[2 marks]**

5. Determine expressions for the approximate bias and variance of the maximum likelihood estimator $\hat{v}$ of v for the Erlang distribution, under the assumption that k is fixed. *(hints: utilise techniques from Lecture 2, Slide 22 and the mean/variance of the sample mean)* **[3 marks]**

Question 3 (8 marks)

In this question we will analyse some binary data; in particular, this question examines the compressive strength of concrete as a function of its setting time. Obviously this is an extremely important problem as concrete is the single most important material in civil engineering and construction. According to the International Building Code (IBC) (Section 1905.1.1), a mixture of concrete is said to be "construction grade" if its compressive strength is $17MPa$ (mega Pascals) or greater. Imagine we are interested in testing the quality control of a company supplying concrete to our construction firm; the company claims to have high quality concrete and guarantees that after 14 days setting time, 90% of the concrete it supplies will have a compressive strength of $17MPa$ or greater. To test this claim we conduct an experiment on 62 pours of concrete, and after setting for 14 days, we perform a strength test and find that 53 of these pours resulted in concrete with strength of $17MPa$ or greater.

1. Using this data, calculate an estimate of the probability that a pour of concrete made by this company will, after setting for 14 days, will have a compressive strength of $17MPa$ or greater, and provide an approximate 95% confidence interval for this estimate. Summarise/describe your results appropriately. **[3 marks]**

2. Test the null hypothesis that at least 90% of the pours made by this company, after setting for 14 days, have a compressive strength of $17MPa$ or greater. Write down explicitly the hypothesis you are testing, and then calculate a p-value using the approximate approach for testing a Bernoulli population discussed in Lecture 5. What does this p-value suggest? **[2 marks]**

3. Using R, calculate an exact p-value to test the above hypothesis. What does this p-value suggest? Please provide the appropriate R command that you used to calculate your p-value. **[1 mark]**

4. Someone at your construction company suggests that waiting for 28 days instead of 14 days should substantially improve the probability that the resulting concrete will have a compressive strength of $17MPa$ or greater. To test this, you conduct an experiment on 425 new pours, and after letting them set for 28 days, you test them and find 399 of these have compressive strength of $17MPa$ or greater. Using the approximate hypothesis testing procedure for testing two Bernoulli populations from Lecture 5, test the hypothesis that the probability of the concrete having compressive strength of $17MPa$ or greater does not differ between pours that set for 14 days and pours that set for 28 days. Summarise your findings. What does the p-value suggest? **[2 marks]**

Question 1 (10 marks)

In this question we will revisit our analysis of the COVID-19 recovery data that we began in Assignment 1. The file `covid.19.ass2.csv` contains a subset of the New South Wales days-to-recovery data we examined previously; this time, patients with recovery times over four weeks (28 days) were removed as these recovery times are unusual and likely represent a sub-population of people more susceptible to the virus. We know from Assignment 1 that the Poisson distribution is not a good fit to the recovery data: instead, for this question we will use a normal distribution as it provides an improved fit to the data due to its increased flexibility, while accepting this assumption is also not necessarily correct; to quote the famous statistician G.E.P.Box: "*all models are wrong – but some are more useful than others*".

Important: you may use R to determine the means and variances of the data, as required, and the R functions `qt()` and `pnorm()` but you must perform all the remaining steps by hand. Please provide appropriate R code fragments and all working out.

1. Calculate an estimate of the average number of days to recovery using the provided data. Calculate a 95% confidence interval for this estimate using the t-distribution, and summarise/describe your results appropriately. Show working as required. **[4 marks]**

$\hat{\mu}_{ML}$ = sample mean = $\frac{1}{n}\sum_{i=1}^{n} y_i$; [bias is 0]

= 14.2580

$\hat{\mu}_{ML} \sim N(\mu, \sigma^2/n)$

```
> mean(df$Recovery.Time)
[1] 14.25797
```

CI : $P\left(\hat{\mu}_{ML} - t_{\alpha/2,n-1}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \hat{\mu}_{ML} + t_{\alpha/2,n-1}\left(\frac{\sigma}{\sqrt{n}}\right)\right) = 0.95$

$\Rightarrow \left(\hat{\mu}_{ML} - t_{\alpha/2,n-1}\left(\frac{\sigma}{\sqrt{n}}\right), \hat{\mu}_{ML} + t_{\alpha/2,n-1}\left(\frac{\sigma}{\sqrt{n}}\right)\right)$

$100(1-\alpha) = 95$; $t_{\alpha/2} = t_{0.025,2353-1} = 1.9610$

$\alpha = 5\%$

$= 0.05$

```
> qt(p = 1 - 0.05 / 2, df = length(df$Recovery.Time) - 1)
[1] 1.960973
```

σ^2 = 44.15324 [unbiased estimator]

```
> var(df$Recovery.Time)
[1] 44.15324
```

$\sigma = \sqrt{\sigma^2}$

= 6.64479

$\Rightarrow$ CI : $\left(14.2580 - 1.9610\left(\frac{6.64479}{\sqrt{2353}}\right), 14.2580 + 1.9610\left(\frac{6.64479}{\sqrt{2353}}\right)\right)$

$\therefore$ confidence interval at 5% significance level

$\mu \in (13.9894, 14.5266)$

2. Similar data was collected in 2020 by the Israeli Ministry of Health. While the specific data was not available, the summary statistics were provided, and from these I have simulated a dataset of $n = 494$ individuals from the Israeli study. The days to recovery in this group are provided in the file `israeli.covid.19.ass2.csv`. Using the provided data and the approximate method for difference in means with (different) unknown variances presented in Lecture 4, calculate the estimated mean difference in recovery times between the Israeli patients and the patients from NSW, and provide an approximate 95% confidence interval. Summarise/describe your results appropriately. Show working as required. **[3 marks]**

95 % confidence interval :

Let A : population from NSW

B : population from Israeli

$\mu_A = 14.2580$ [earlier calculation]

$\mu_B = 14.6498$

```
> mean(df2$Recovery.Time)
[1] 14.6498
```

$\sigma_A^2 = 44.15324$ [earlier calculation]

$\sigma_B^2 = 30.47549$

```
> var(df2$Recovery.Time)
[1] 30.47549
```

$n_A = 2353$, $n_B = 494$

```
> length(df2$Recovery.Time)
[1] 494
```

$100(1-\alpha) = 95$; $z_{0.025} = 1.96$

$\alpha = 5\% = 0.05$

$$\mu_A - \mu_B \in \left((\bar{x} - \bar{y}) - z_{\alpha/2} \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}} \, , \, (\bar{x} - \bar{y}) + z_{\alpha/2} \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}} \right)$$

$$\Rightarrow \mu_A - \mu_B \in \left((14.2580 - 14.6498) - 1.96 \sqrt{\frac{44.15324}{2353} + \frac{30.47549}{494}} \, , \, (14.2580 - 14.6498) + 1.96 \sqrt{\frac{44.15324}{2353} + \frac{30.47549}{494}} \right)$$

$$\mu_A - \mu_B \in (-0.9477 \, , \, 0.1641)$$

3. It is of interest to determine if there are any differences, at a population level, in recovery times for patients in different countries. Test the hypothesis that the population average time taken to recover for the Israeli cohort is the same as in the NSW cohort. Write down explicitly the hypothesis you are testing, and then calculate a p-value using the approximate hypothesis test for differences in means with (different) unknown variances presented in Lecture 5. What does this p-value suggest about the difference in mean recovery time between the two cohorts of patients? **[3 marks]**

$H_0 : \mu_A = \mu_B$ } $\mu_A - \mu_B = 0$ [two side test]
$H_A : \mu_A \neq \mu_B$ $\mu_A - \mu_B \neq 0$

where μ_A is the population mean of time taken to recover in NSW
μ_B is the population mean of time taken to recover in Israeli

Test statistic : (estimate)

$$z = (\hat{\mu}_A - \mu_B) / \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}$$
$$= (14.2580 - 14.6498) / \sqrt{\frac{44.15324}{2353} + \frac{30.47549}{494}}$$
$$= -1.3813$$

$$p \approx \begin{cases} 2\mathbb{P}(Z < -|z_{(\hat{\mu}_x - \hat{\mu}_y)}|) & \text{if } H_0 : \mu_x = \mu_y \text{ vs } H_A : \mu_x \neq \mu_y \\ 1 - \mathbb{P}(Z < z_{(\hat{\mu}_x - \hat{\mu}_y)}) & \text{if } H_0 : \mu_x \leq \mu_y \text{ vs } H_A : \mu_x > \mu_y \\ \mathbb{P}(Z < z_{(\hat{\mu}_x - \hat{\mu}_y)}) & \text{if } H_0 : \mu_x \geq \mu_y \text{ vs } H_A : \mu_x < \mu_y \end{cases}$$

p-value $= 2 \times P(z < -|-1.3813|)$
$= 2 \times P(Z < -1.3813)$
$= 0.16719$

```
> 2 * pnorm(-1.3813)
[1] 0.1671867
```

∴ The p-value is larger than the significance level 0.05.
The p-value says that there is a difference at the population level in the Israeli and NSW cohort.
We would expect to see a difference in recovery times 16.7% of the time if we drew samples from these populations.
The p-value suggests that there is very weak evidence against the null, that there are differences in recovery times of the two cohorts

Question 2 (10 marks)

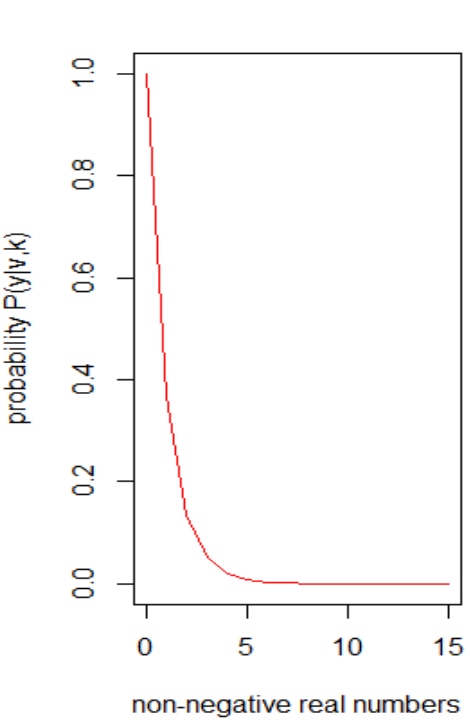
The Erlang distribution is a probability distribution for non-negative real numbers. It models the time taken for k events to occur, under the condition that the events occur at the same rate over time. It finds extensive use in telecommunications and medical research. The version that we will look at has a probability density function of the form

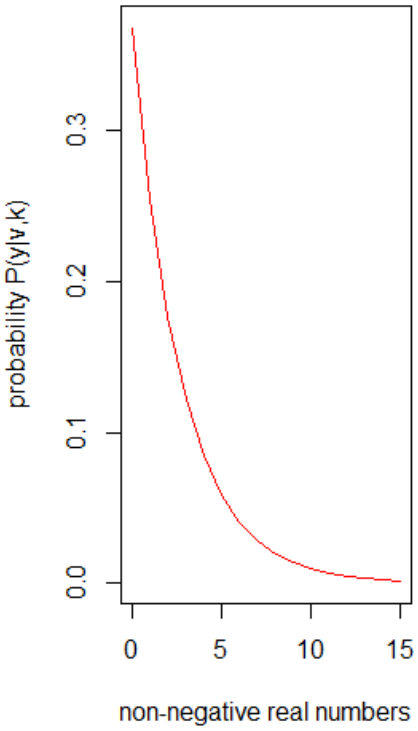
$$p(y \mid v, k) = \left(\frac{y^{k-1}}{(k-1)!} \right) \exp\left(-e^{-v} y - kv\right) \tag{1}$$

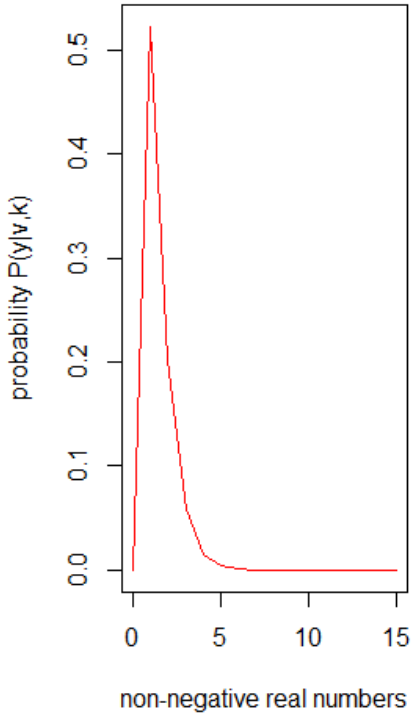
where $e^x \equiv \exp(x)$ and $y \in \mathbb{R}_+$, i.e., y can take on the values of non-negative real numbers. In this form it has two parameters: the "shape" $k > 0$, which is the number of events we are interested in waiting for, and $v \in \mathbb{R}$, which is the log-inverse-rate (also called a log-scale) of the distribution, which controls the rate at which the events we are modelling occur. Often k is not treated as a learnable parameter, but is rather set by the user depending on the context. If a random variable follows an Erlang distribution with log-inverse-rate v and shape k we say that $Y \sim \text{Er}(v, k)$. If $Y \sim \text{Er}(v, k)$, then $\mathbb{E}[Y] = k\, e^v$ and $\mathbb{V}[Y] = k\, e^{2v}$.

1. Produce a plot of the Erlang probability density function (1) for the values $y \in (0, 15)$, for $(v = 0, k = 1)$, $(v = 1, k = 1)$ and $(v = -1/2, k = 2)$. Ensure that the graph is readable, the axis are labelled appropriately and a legend is included. **[2 marks]**

$(v = 0, k = 1)$ $(v = 1, k = 1)$ $(v = -0.5, k = 2)$







2. Imagine we are given a sample of n observations $\mathbf{y} = (y_1, \ldots, y_n)$. Write down the joint probability of this sample of data, under the assumption that it came from an Erlang distribution with log-inverse-rate v and shape k (i.e., write down the likelihood of this data). Make sure to simplify your expression, and provide working. *(hint: remember that these samples are independent and identically distributed.)* **[2 marks]**

Joint Probability / Likelihood:

$$p(y \mid v, k) = \left(\frac{y^{k-1}}{(k-1)!}\right) e^{(-e^{-v}y - kv)}$$

$$p(\mathbf{y} \mid v, k) = \prod_{i=1}^{n} p(y_i \mid v, k)$$

$$= \frac{(y_1)^{k-1}}{(k-1)!} e^{(-e^{-v}(1) - kv)} \times \frac{(y_2)^{k-1}}{(k-1)!} e^{(-e^{-v}(2) - kv)} \times \ldots \times \frac{(y_n)^{k-1}}{(k-1)!} e^{(-e^{-v}(n) - kv)}$$

$$= \frac{e^{-kv}}{(k-1)!}\left[(y_1)^{k-1} e^{-e^{-v}}\right] \times \frac{e^{-kv}}{(k-1)!}\left[(y_2)^{k-1} e^{-2e^{-v}}\right] \times \ldots \times \frac{e^{-kv}}{(k-1)!}\left[(y_n)^{k-1} e^{-ne^{-v}}\right]$$

$$= \left(\frac{e^{-kv}}{(k-1)!}\right)^n \left[(y_1)^{k-1} e^{-e^{-v}} \times (y_2)^{k-1} e^{-2e^{-v}} \times \ldots \times (y_n)^{k-1} e^{-ne^{-v}}\right]$$

$$= \left(\frac{e^{-kv}}{(k-1)!}\right)^n \left[\prod_{i=1}^{n} (y_i)^{k-1} e^{-e^{-v}\sum_{i=1}^{n} y_i}\right]$$

3. Take the negative logarithm of your likelihood expression and write down the negative log-likelihood of the data **y** under the Erlang model with log-inverse-rate v and shape k. Simplify this expression. **[1 mark]**

negative - log likelihood

$$
\begin{aligned}
L(p \mid v,k) &= -\log\left[P(y \mid v,k)\right] \\
&= -\log\left[\left(\frac{e^{-kv}}{(k-1)!}\right)^n \left(\prod_{i=1}^n (y_i)^{k-1} e^{-e^{-v}\sum_{i=1}^n y_i}\right)\right] \\
&= -\left[\log\left(\left(\frac{e^{-kv}}{(k-1)!}\right)^n\right) + \log\prod_{i=1}^n (y_i)^{k-1} + \log e^{-e^{-v}\sum_{i=1}^n y_i}\right] \\
&= -\left[\log e^{-nkv} - \log((k-1)!)^n + \log\prod_{i=1}^n (y_i)^{k-1} + \log e^{-e^{-v}\sum_{i=1}^n y_i}\right] \\
&= -\left[-nkv - n\log((k-1)!) + \log\left(1^{k-1}\times 2^{k-1}\ldots\times n^{k-1}\right) + \left(-e^{-v}\sum_{i=1}^n y_i\right)\right] \\
&= nkv + n\log((k-1)!) - \left(\log 1^{k-1} + \log 2^{k-1} + \ldots + n\log^{k-1}\right) + e^{-v}\sum_{i=1}^n y_i \\
&= nkv + n\log[(k-1)!] - \sum_{i=1}^n \log(y_i)^{k-1} + e^{-v}\sum_{i=1}^n y_i \\
&= nkv + n\log[(k-1)!] - (k-1)\sum_{i=1}^n \log(y_i) + e^{-v}\sum_{i=1}^n y_i
\end{aligned}
$$

4. Derive the maximum likelihood estimator $\hat{v}$ for v, under the assumption that k is fixed; that is, find the value of v that minimises the negative log-likelihood, treating k as a fixed quantity. You must provide working. **[2 marks]**

ML estimator of $\hat{v}$: solve partial derivatives for v

$$L(P|v,k) = nkv + n\log[(k-1)!] - (k-1)\sum_{i=1}^{n}\log(y_i) + e^{-v}\sum_{i=1}^{n} y_i$$

$$\frac{dL(P|v,k)}{dv} = nk - e^{-v}\sum_{i=1}^{n} y_i = 0$$

$$nk = e^{-v}\sum_{i=1}^{n} y_i$$

$$e^{-v} = \frac{nk}{\sum_{i=1}^{n} y_i}$$

$$v = -\log\left(\frac{nk}{\sum_{i=1}^{n} y_i}\right)$$

$$\hat{v}_{ML} = \log\left(\frac{\sum_{i=1}^{n} y_i}{nk}\right)$$

5. Determine expressions for the approximate bias and variance of the maximum likelihood estimator $\hat{v}$ of v for the Erlang distribution, under the assumption that k is fixed. *(hints: utilise techniques from Lecture 2, Slide 22 and the mean/variance of the sample mean)* **[3 marks]**

$-\frac{1}{2}k - \log n$

Approximate bias of estimator:

$b_\theta(\hat{\theta}) = E[\hat{\theta}(Y)] - \theta$

using taylor series to approximate $E[\hat{\theta}(Y)]$ around μ,

$\Rightarrow E[\hat{v}_{ML}(Y)] = E[f(\bar{Y})]$ ($\bar{Y}$: sample mean)

$\approx E[f(\mu) + f'(\mu)(\bar{Y}-\mu) + \frac{f''(\mu)}{2}(\bar{Y}-\mu)^2]$

$= f(\mu) + f'(\mu)E[(\bar{Y}-\mu)] + \frac{f''(\mu)}{2}E[(\bar{Y}-\mu)^2]$

$= f(\mu) + \frac{f''(\mu)}{2}V[\bar{Y}]$

$= f(\mu) + \frac{f''(\mu)}{2}\left(\frac{\sigma^2}{n}\right)$

; $\hat{v}_{ML} = \ln\left(\frac{\sum_{i=1}^n y_i}{nk}\right)$

$\Rightarrow$ we know that $\frac{1}{n}\sum_{i=1}^n y_i = \bar{Y}$,

so $\hat{v}_{ML} = \log\left(\frac{\bar{Y}}{k}\right)$

$\mu = E[\bar{Y}]$, the expected value of the sample mean

$= E\left[\frac{Y_1 + Y_2 + Y_3 + \ldots + Y_n}{n}\right]$

$= \frac{E[Y_1]}{n} + \frac{E[Y_2]}{n} + \frac{E[Y_3]}{n} + \ldots + \frac{E[Y_n]}{n}$

AND $\frac{\sigma^2}{n} = V[\bar{Y}]$

$= V\left[\frac{Y_1 + Y_2 + \ldots + Y_n}{n}\right]$

$= \frac{V[Y_1]}{n^2} + \frac{V[Y_2]}{n^2} + \ldots + \frac{V[Y_n]}{n^2}$

From the question,

parameter, but is rather set by the user depending on the context. If a random variable follows an Erlang distribution with log-inverse-rate v and shape k we say that $Y \sim \mathrm{Er}(v, k)$. If $Y \sim \mathrm{Er}(v, k)$, then $\mathbb{E}[Y] = k\,e^v$ and $\mathbb{V}[Y] = k\,e^{2v}$.

$\Rightarrow \mu = \frac{ke^v}{n} + \frac{ke^v}{n} + \ldots \frac{ke^v}{n}$

$= n\left(\frac{ke^v}{n}\right)$

$= ke^v$

; $\frac{\sigma^2}{n} = \frac{ke^{2v}}{n^2} + \frac{ke^{2v}}{n^2} + \ldots + \frac{ke^{2v}}{n^2}$

$= n\left(\frac{ke^{2v}}{n^2}\right)$

$= \frac{ke^{2v}}{n}$

$\Rightarrow E[f(\bar{Y})] = f(\mu) + \frac{f''(\mu)}{2}\left(\frac{\sigma^2}{n}\right)$

$= \log\left(\frac{ke^v}{k}\right) + \frac{1}{2}\left(-\frac{1}{\mu^2}\right)\left(\frac{ke^{2v}}{n}\right)$

$= \log(e^v) + \frac{1}{2}\left(-\frac{1}{(ke^v)^2}\right)\left(\frac{ke^{2v}}{n}\right)$

$= v + \frac{1}{2}\left(\frac{ke^{2v}}{k^2e^{2v}n}\right)$

$= v + \frac{1}{2kn}$

; $f(\mu) = \log\left(\frac{\mu}{k}\right) = \log(\mu) - \log(k)$

$f'(\mu) = \frac{1}{\mu}$

$f''(\mu) = -\frac{1}{\mu^2}$

$\therefore$ bias $= E[f(\bar{Y})] - \theta$

$= v + \frac{1}{2kn} - v$

$= \frac{1}{2kn}$

Approximate variance of estimator:

$$\text{Var}_\theta(\hat{\theta}) = V[\hat{\theta}(Y)] \quad ; \quad \hat{\theta}(Y) = f(\bar{Y}) = \log\left(\frac{\bar{Y}}{k}\right)$$

$$= V[f(\bar{Y})]$$

$$= V[f(\mu) + f'(\mu)(\bar{Y} - \mu)]$$

$$= (f'(\mu))^2 V[(\bar{Y} - \mu)]$$

$$= (f'(\mu))^2 V[\bar{Y}]$$

$$= \left(\frac{1}{\mu}\right)^2 \left(\frac{\sigma^2}{n}\right) \quad ; \quad \mu = ke^v \quad , \quad \sigma^2 = ke^{2v}$$

$$= \frac{1}{k^2 e^{2v}} \left(\frac{ke^{2v}}{n}\right)$$

$$= \frac{1}{kn}$$

Question 3 (8 marks)

In this question we will analyse some binary data; in particular, this question examines the compressive strength of concrete as a function of its setting time. Obviously this is an extremely important problem as concrete is the single most important material in civil engineering and construction. According to the International Building Code (IBC) (Section 1905.1.1), a mixture of concrete is said to be "construction grade" if its compressive strength is $17MPa$ (mega Pascals) or greater. Imagine we are interested in testing the quality control of a company supplying concrete to our construction firm; the company claims to have high quality concrete and guarantees that after 14 days setting time, 90% of the concrete it supplies will have a compressive strength of $17MPa$ or greater. To test this claim we conduct an experiment on 62 pours of concrete, and after setting for 14 days, we perform a strength test and find that 53 of these pours resulted in concrete with strength of $17MPa$ or greater.

1. Using this data, calculate an estimate of the probability that a pour of concrete made by this company will, after setting for 14 days, will have a compressive strength of $17MPa$ or greater, and provide an approximate 95% confidence interval for this estimate. Summarise/describe your results appropriately. **[3 marks]**

This is a bernoulli distribution:

$X \sim e(\theta)$, where X is the RV of a pour of concrete with compressive strength $\geq 17MPa$

$$\hat{\theta}_{ML} = \frac{1}{n}\sum_{i=1}^{n} y_i$$
$$= \frac{53}{62}$$ ← successes we've observed

95% confidence interval, 5% significance level:

Using CLT, $\hat{\theta}_{ML}$ = sample mean:

$$\hat{\theta}_{ML} \xrightarrow{d} N\left(E[Y_i], \frac{V[Y_i]}{n}\right) = N\left(\theta, \frac{\theta(1-\theta)}{n}\right)$$

$$CI = \left(\hat{\theta}_{ML} - 1.96\sqrt{\frac{\hat{\theta}_{ML}(1-\hat{\theta}_{ML})}{n}}, \hat{\theta}_{ML} + 1.96\sqrt{\frac{\hat{\theta}_{ML}(1-\hat{\theta}_{ML})}{n}}\right)$$
$$= \left(\frac{53}{62} - 1.96\sqrt{\frac{53/62\,(1-53/62)}{62}}, \frac{53}{62} + 1.96\sqrt{\frac{53/62\,(1-53/62)}{62}}\right)$$
$$= (0.76715, 0.94252)$$

It means that, we can be 95% sure that the compressed strength of one pour of concrete $\geq 17MPa$, is between the interval of $(0.76715, 0.94252)$.

2. Test the null hypothesis that at least 90% of the pours made by this company, after setting for 14 days, have a compressive strength of $17MPa$ or greater. Write down explicitly the hypothesis you are testing, and then calculate a p-value using the approximate approach for testing a Bernoulli population discussed in Lecture 5. What does this p-value suggest? **[2 marks]**

$H_0: \theta \geq 0.9$

$H_A: \theta < 0.9$ [one-sided testing]

where θ is the probability that a single concrete pour has $\geq$ 17MPa after setting for 14 days.

By CLT: $\hat{\theta} - \theta_0 \xrightarrow{d} N\left(0, \frac{\theta_0(1-\theta_0)}{n}\right) = \hat{\theta} - \theta_0 \xrightarrow{d} \left(0, \frac{0.9(1-0.9)}{62}\right)$

$= \hat{\theta} - \theta_0 \xrightarrow{d} \left(0, \frac{0.09}{62}\right)$

Test statistic:

$$z_{\hat{\theta}} = \frac{\hat{\theta} - \theta_0}{\sqrt{\theta_0(1-\theta_0)/n}}$$

$$= \frac{53/62 - 0.9}{\sqrt{0.9(0.1)/62}}$$

$$= -1.1853$$

p-value:

```
> pnorm(-1.1853)
[1] 0.1179494
```

$P(Z < -1.1853) = 0.1179$

∴ For a given threshold α, if the p-value is less than α, we don't have enough evidence to reject the null hypothesis.
In contrast, if the p-value $\geq \alpha$, then we would have sufficient evidence to reject the null hypothesis.

This means that almost 11.7% of the time, a sample of 62 pours of concrete will result in: 53 pours or more with $\geq$ 17MPa strength

By convention, p-value > 0.1, we have very weak evidence against the null.

3. Using R, calculate an exact p-value to test the above hypothesis. What does this p-value suggest? Please provide the appropriate R command that you used to calculate your p-value. **[1 mark]**

```
> binom.test(53,62, p=0.9, alternative=c("less"))

        Exact binomial test

data:  53 and 62
number of successes = 53, number of trials = 62, p-value = 0.1634
alternative hypothesis: true probability of success is less than 0.9
95 percent confidence interval:
 0.0000000 0.9221288
sample estimates:
probability of success
             0.8548387
```

p-value = 0.1634

This p-value is greater than the p-value in the approximate procedure. If the sample size is greater, we would expect both p-values to be closer as the normal approximation gets better with increasing sample size.
Same as the conclusion earlier in Question 3.2 this means that:
Almost 16.3% of the time, the 62 pours of concrete will result in 53 pours or more with $\geq$ 17MPa strength

4. Someone at your construction company suggests that waiting for 28 days instead of 14 days should substantially improve the probability that the resulting concrete will have a compressive strength of $17MPa$ or greater. To test this, you conduct an experiment on 425 new pours, and after letting them set for 28 days, you test them and find 399 of these have compressive strength of $17MPa$ or greater. Using the approximate hypothesis testing procedure for testing two Bernoulli populations from Lecture 5, test the hypothesis that the probability of the concrete having compressive strength of $17MPa$ or greater does not differ between pours that set for 14 days and pours that set for 28 days. Summarise your findings. What does the p-value suggest? **[2 marks]**

$H_0 : \theta_x = \theta_y$

$H_A : \theta_x \neq \theta_y$ [two - side testing]

where: θ_x is the probability that a single concrete pour has $\geq$ 17MPa after setting for 14 days.

θ_y is the probability that a single concrete pour has $\geq$ 17MPa after setting for 28 days.

$$\hat{\theta}_p = \frac{m_x + m_y}{n_x + n_y} = \frac{53 + 399}{62 + 425} = \frac{452}{487} ; \hat{\theta}_x = \frac{53}{62} ; \hat{\theta}_y = \frac{399}{425}$$

Test statistic :

$$Z_{(\hat{\theta}_x - \hat{\theta}_y)} = \frac{\hat{\theta}_x - \hat{\theta}_y}{\sqrt{\hat{\theta}_p(1-\hat{\theta}_p)(1/n_x + 1/n_y)}}$$

$$= \frac{53/62 - 399/425}{\sqrt{(452/487)(35/487)(1/62 + 1/425)}}$$

$$= -2.3920$$

p-value $= 2 \times P(Z < -2.3920)$

```
> 2*pnorm(-2.3920)
[1] 0.01675684
```

$= 0.01676$

Summary :

Informally, for p-values : $0.01 < p < 0.05$, we have moderate evidence against the null.

This means that almost 1.7% of the time, there is a difference between setting the concrete pour for 14 days and 28 days.