FIT2086 Assignment 2

Ian Wong 30612616

15/09/2021

Loading Libraries

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Question 1

1.

Using **read_csv**, we are going to read in the New South Wales days-to-recovery data:

```
covid_data <- read_csv('covid.19.ass2.csv')
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   Recovery.Time = col_double()
## )
```

The confidence interval for a t-distribution is given by:

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

where $t_{\alpha/2,n-1}$ is the 100(1-α/2)-th percentile of the standard Student-t distribution with (n-1) degrees of freedom.

```r
X = covid_data$Recovery.Time
n = length(X)                   # length of Recovery Time
mu_ml = sum(X)/n                # sample mean
var_ml = sum((X - mu_ml)^2)/n   # sample variance
sd_ml = sqrt(var_ml)            # sample standard deviation

t = qt(p = 1-0.05/2, df=n-1)    # multiplier
```

The average number of days to recovery is 14.2579686.
$t_{\alpha/2,n-1}$ is equal to 1.9609731.
The 95% confidence interval for the estimated mean is then

$$(14.26 - 1.96\frac{6.64}{\sqrt{2353}}, 14.26 + 1.96\frac{6.64}{\sqrt{2353}})$$

which is equal to

$$(13.99, 14.53)$$

Therefore, we can say that the estimated average number of days to recovery (sample size n = 2353) is 14.26 days. We are 95% confident that the population mean for this group is between 13.99 and 14.53 days.

2.

Using **read_csv**, we are going to read in the Israeli days-to-recovery data:

```r
israeli_data <- read_csv('israeli.covid.19.ass2.csv')
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   Recovery.Time = col_double()
## )
```

The confidence interval for the mean difference in recovery times for NSW and Israel is given by:

$$\left(\hat{\mu}_{NSW} - \hat{\mu}_{IL} - z_{\alpha/2}\sqrt{\frac{\hat{\sigma}^2_{NSW}}{n_{NSW}} + \frac{\hat{\sigma}^2_{IL}}{n_{IL}}}, \hat{\mu}_{NSW} - \hat{\mu}_{IL} + z_{\alpha/2}\sqrt{\frac{\hat{\sigma}^2_{NSW}}{n_{NSW}} + \frac{\hat{\sigma}^2_{IL}}{n_{IL}}}\right)$$

```r
Y = israeli_data$Recovery.Time
Y_n = length(Y)                     # length of Recovery Time
Y_mu_ml = sum(Y)/n                  # sample mean
Y_var_ml = sum((Y - Y_mu_ml)^2)/n   # sample variance
Y_sd_ml = sqrt(Y_var_ml)            # sample standard deviation


diff = mu_ml - Y_mu_ml              # mean difference
lower = diff - 1.96*sqrt((var_ml/n)+(Y_var_ml/Y_n))
upper = diff + 1.96*sqrt((var_ml/n)+(Y_var_ml/Y_n))
```

The estimated mean difference in recovery times between the Israeli patients and the patients from NSW is 11.1823204.

$z_{\alpha/2}$ is approximately equal to 1.96.
The 95% confidence interval for the estimated mean difference is then

$$\left(14.26 - 3.08 - 1.96\sqrt{\frac{44.13}{2353} + \frac{34.51}{494}}, 14.26 - 3.08 + 1.96\sqrt{\frac{44.13}{2353} + \frac{34.51}{494}}\right)$$

which is equal to

$$(10.60, 11.77)$$

Therefore, we can say that the estimated mean difference in recovery times between the Israeli patients (sample size n = 494) and the patients from NSW (sample size n = 2353) is 11.18 days. We are 95% confident the population mean difference in recovery times is between 10.60 and 11.76 days. As the interval is entirely positive, this suggests that there is a positive difference in the recovery times of Israeli patients and NSW patients at population level.

3.

The null hypothesis and alternative hypothesis of testing whether the population average time taken to recover is the same in NSW and IL are:

$$H_0 : \mu_{NSW} = \mu_{IL}$$
$$H_1 : \mu_{NSW} \neq \mu_{IL}$$

where H_0 is that there is no difference between average time taken to recover in NSW and IL, and H_1 is there is a difference.

The z-test statistic for testing difference of means is given by:

$$z_{(\hat{\mu}_{NSW} - \hat{\mu}_{IL})} = \frac{\hat{\mu}_{NSW} - \hat{\mu}_{IL}}{\sqrt{\frac{\hat{\sigma}^2_{NSW}}{n_{NSW}} + \frac{\hat{\sigma}^2_{IL}}{n_{IL}}}}$$

```
z = (mu_ml - Y_mu_ml)/sqrt((var_ml/n)+(Y_var_ml/Y_n))
z
```

```
## [1] 37.56473
```

The z-calc is then equal to:

$$z = \frac{14.26 - 3.08}{\sqrt{\frac{44.13}{2353} + \frac{34.51}{494}}}$$
$$= 37.56$$

We can then approximate p-values using:

$$p \approx 2\mathbb{P}(Z < -|z_{(\hat{\mu}_{NSW} - \hat{\mu}_{IL}}|)$$

```
pval = 2*pnorm(-abs(z))
pval
```

```
## [1] 0
```

The p-value is equal to 0.
As the p-value is < 0.01 we have strong evidence against the null.
Therefore, we reject the null hypothesis and conclude that at the 5% level of significance, there is a difference in average time taken to recover in Israeli and NSW patients.

Question 2.

1.

First, we create a function that calculates the pdf of the Erlang Distribution:

```
erlang <- function(y, v, k) {

  retval = 0

  factorial = 1                     # factorial = 1 initially
  if (k != 1) {                     # if k = 1, 0! = 1; else go into for loop
    for (i in 1:(k-1)) {            # loops until reaches k-1
      factorial = factorial * i
    }
  }

  retval = (y^(k-1)/factorial)*exp((-exp(-v)*y)-(k*v))

  return(retval)
}
```

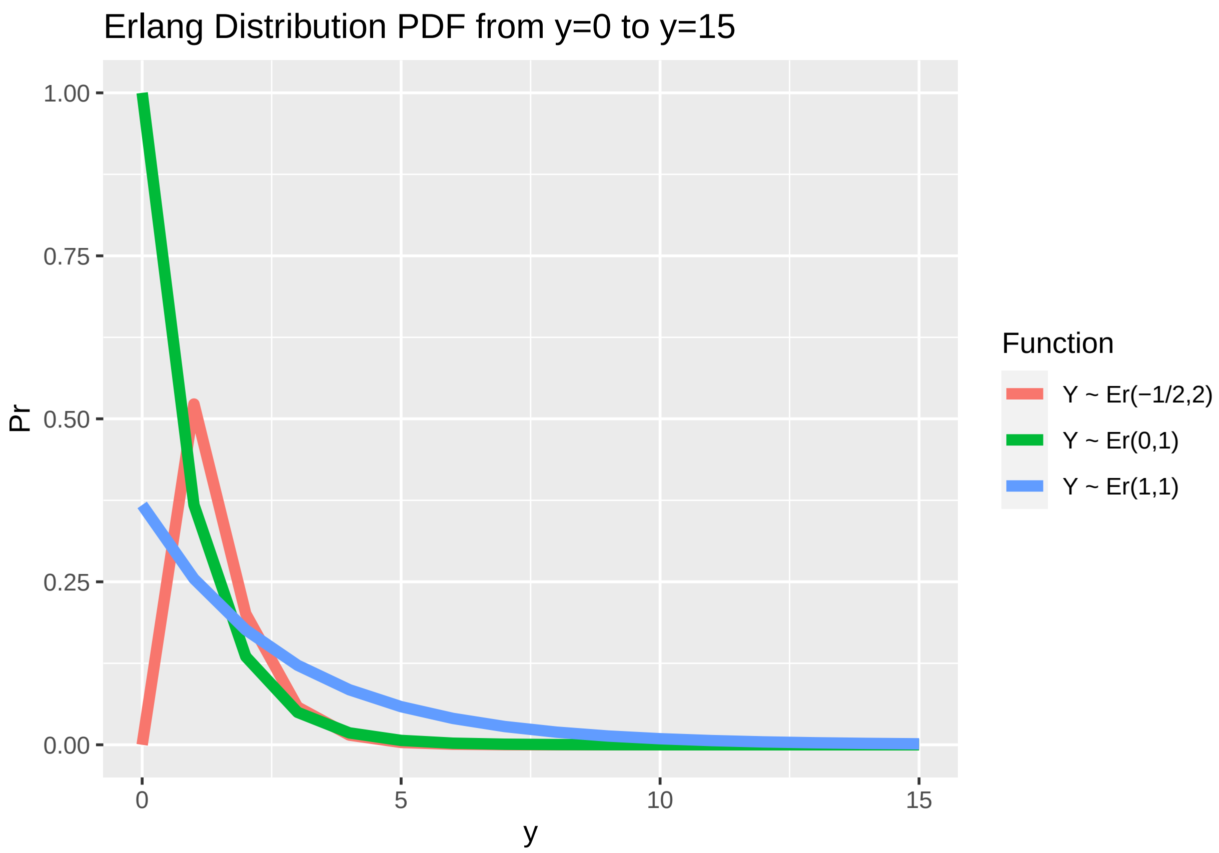
Using **data.frame**, we are creating a data frame that contains data that we can then pass through to ggplot:

```
erlang_dist <- data.frame(
  y = 0:15,                                      # y values from 0 to 15
  Pr = c(erlang(0:15,0,1),                       # get 16 vals for each func
         erlang(0:15,1,1),
         erlang(0:15,-1/2,2)),
  Function = rep(c("Y ~ Er(0,1)",                # adding on labels for func
                   "Y ~ Er(1,1)",
                   "Y ~ Er(-1/2,2)"), each = 16)
                          )
head(erlang_dist)                                # checking the data frame
```

```
##   y          Pr    Function
## 1 0 1.000000000 Y ~ Er(0,1)
## 2 1 0.367879441 Y ~ Er(0,1)
## 3 2 0.135335283 Y ~ Er(0,1)
## 4 3 0.049787068 Y ~ Er(0,1)
## 5 4 0.018315639 Y ~ Er(0,1)
## 6 5 0.006737947 Y ~ Er(0,1)
```

Creating the graph:

```
erlang_dist %>% ggplot(aes(x = y, Pr, col = Function)) +
  geom_line(size = 2) +
  ggtitle("Erlang Distribution PDF from y=0 to y=15")
```



2.

$$\mathbf{y} = y_1, y_2, ..., y_n \overset{iid}{\sim} Er(v, k)$$

As these samples are independently and identically distributed, the joint probability of this sample of data is given by:

$$\begin{aligned} p(y_1, y_2, ..., y_n|v, k) &= p(y_1|v, k) * p(y_2|v, k) * ... * p(y_n|v, k) \\ &= \prod_{i=1}^{n} p(y_i|v, k) \\ &= \prod_{i=1}^{n} \left[\left(\frac{y_i^{k-1}}{(k-1)!} \right) \exp(-e^{-v} y_i - kv) \right] \\ &= \left(\frac{(y_1 y_2 ... y_n)^{k-1}}{(k-1)^n!} \right) \exp(-e^{-v}(y_1 + y_2 + ... + y_n) - nkv) \\ \therefore p(\mathbf{y}|v, k) &= \left(\frac{(\prod_{i=1}^{n} y_i)^{k-1}}{(k-1)^n!} \right) \exp(-(e^{-v}(\sum_{i=1}^{n} y_i) + nkv)) \end{aligned}$$

3.

The negative log-likelihood function is given by:

$$\begin{aligned} L(\mathbf{y}|v,k) &= -\log p(\mathbf{y}|v,k) \\ &= -\log\left(\left(\frac{(\prod_{i=1}^n y_i)^{k-1}}{(k-1)^n!}\right)\exp(-(e^{-v}(\sum_{i=1}^n y_i)+nkv))\right) \\ &= n\log(k-1)! - (k-1)\log\prod_{i=1}^n y_i + (e^{-v}(\sum_{i=1}^n y_i)+nkv) \end{aligned}$$

4.

$$\begin{aligned} \frac{\partial L(\mathbf{y}|v,k)}{\partial v} &= 0 - 0 - e^{-v}(\sum_{i=1}^n y_i) + nk \\ 0 &= -e^{-v}(\sum_{i=1}^n y_i) + nk \\ e^{-v} &= \frac{nk}{\sum_{i=1}^n y_i} \\ -v &= \log nk - \log\sum_{i=1}^n y_i \\ \hat{v} &= \log\sum_{i=1}^n y_i - \log nk \end{aligned}$$

5.

$$E[Y] = \mu_Y = ke^v \text{ and } V[Y] = \sigma_Y^2 = ke^{2v}$$

The bias of the ML estimator $\hat{v}$ of v is given by:

$$b_v(\hat{v}) = E[\hat{v}(Y)] - v$$

The variance of the ML estimator $\hat{v}$ of v is given by:

$$Var_v(\hat{v}) = V[\hat{v}(Y)]$$

Let $\hat{v} = f(Y)$.

Variance of Estimator

The approximate variance of a function of a RV is given by:

$$\mathbb{V}[f(Y)] = \left[\frac{df(Y)}{dy}|_{y=\mu_Y}\right]^2 \sigma_Y^2$$

$$\begin{aligned}
\frac{df(Y)}{dy} &= \frac{d}{dy}(\log \sum_{i=1}^{n} y_i - \log nk) \\
&= \frac{n}{\sum y_i} \text{ (as } (\sum_{i=1}^{n} y_i)' = (1 + ... + 1) = n)
\end{aligned}$$

Substitute in $y = \mu_Y$:

$$\begin{aligned}
\frac{df(Y)}{dy}|_{y=\mu_Y} &= \frac{n}{\sum ke^v} \\
&= \frac{n}{nke^v} \\
&= \frac{1}{ke^v} \\
\therefore V[\hat{v}] &= (\frac{1}{ke^v})^2 ke^{2v} \\
&= \frac{ke^{2v}}{k^2 e^{2v}} \\
&= \frac{1}{k}
\end{aligned}$$

As $Var_v(\hat{v}) = V[\hat{v}(Y)]$, the variance of ML estimator $\hat{v}$ is equal to $\frac{1}{k}$.

Mean of Estimator

The approximate mean of a function of a RV is given by:

$$\mathbb{E}[f(Y)] \approx f(\mu_Y) + \left[\frac{d^2 f(Y)}{dy^2}|_{y=\mu_Y} \right] \frac{\sigma_Y^2}{2}$$

$$
\begin{aligned}
f(\mu_Y) &= \log \sum_{i=1}^{n} y_i - \log nk \\
&= \log nke^v - \log nk \\
&= \log nk + \log e^v - \log nk \\
&= \log e^v \\
&= v \\
\frac{d^2 f(Y)}{dy^2} &= \frac{d}{dy}(\frac{n}{\sum y_i}) \\
&= \frac{d}{dy}(n(\sum y_i)^{-1}) \\
&= -n(\sum y_i)^{-2} n \\
&= -n^2(\sum y_i)^{-2} \\
\text{Substitute in } y = \mu_Y :& \\
\frac{d^2 f(Y)}{dy^2}|_{y=\mu_Y} &= -\frac{n^2}{n^2 k^2 e^{2v}} \\
&= -\frac{1}{k^2 e^{2v}} \\
\therefore \mathbb{E}[\hat{v}] &= v - \frac{1}{k^2 e^{2v}}(\frac{ke^{2v}}{2}) \\
&= v - \frac{1}{2k} \\
\Rightarrow bias_v(\hat{v}) &= (v - \frac{1}{2k}) - v \\
&= -\frac{1}{2k}
\end{aligned}
$$

Therefore, the bias of ML estimator $\hat{v}$ is equal to $-\frac{1}{2k}$.

Question 3.

1.

The estimate of a success in this Bernoulli distribution is given by:

$$
\begin{aligned}
\hat{\theta} &= \frac{m}{n} \\
&= \frac{53}{62} \\
&= 0.854839
\end{aligned}
$$

Therefore, the probability that a pour of concrete made by this company will have a compressive strength of 17MPa or greater after 14 days is 85.48%.

The approximate 95% CI for a Bernoulli distribution is given by:

$$
\left(\hat{\theta} - 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}, \hat{\theta} + 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}\right)
$$

The 95% confidence interval then:

$$\left(0.8548 - 1.96\sqrt{\frac{0.8548(1-0.8548)}{62}}, 0.8548 + 1.96\sqrt{\frac{0.8548(1-0.8548)}{62}}\right)$$

which is equal to:

$$(0.7671, 0.9425)$$

.

Therefore, we are 95% confident that the population estimate that a pour of concrete having a strength equal to or greater than 17MPa, is contained within the interval (0.7671, 0.9425).

2.

The null and alternative hypothesis that at least 90% of the pours made by this company have a compressive strength of 17MPa or greater is given by:

$$H_0 : \theta \geq 0.90$$
$$H_1 : \theta < 0.90$$

Under the null hypothesis, then by the CLT:

$$\hat{\theta} - \theta_0 \xrightarrow{d} N(0, \frac{\theta_0(1-\theta_0)}{n})$$

Our test statistic is then the approximate z-score:

$$\begin{aligned} z_{\hat{\theta}} &= \frac{\hat{\theta} - \theta_0}{\sqrt{\theta_0(1-\theta_0)/n}} \\ &= \frac{0.8548 - 0.90}{\sqrt{0.90 * 0.10/62}} \\ &= -1.1864 \end{aligned}$$

```
pval = pnorm(-1.1864)
pval
```

```
## [1] 0.1177322
```

The p-value is given by (by looking at the z-score table):

$$\text{p-value} = P(Z < -1.1864) \approx 0.1177$$

3.

```
pbinom(53, 62, 0.9)
```

```
## [1] 0.1633807
```

As the p-value > 0.05, we have weak/no evidence against the null hypothesis. Therefore, we do not reject the null hypothesis and conclude at the 5% level of significance, that at least 90% of the pours made by this company have a strength of 17MPa or greater, after 14 days of setting.

4.

The null and alternative hypothesis that the probability of the concrete having compressive strength of 17MPa or greater does not differ between pours that set for 14 days vs 28 days is given by:

$$\begin{aligned} H_0 &: \theta_{14} = \theta_{28} \\ H_1 &: \theta_{14} \neq \theta_{28} \end{aligned}$$

Under the null hypothesis, $\theta_{14} = \theta_{28} = \theta$.
We use a pooled estimate of θ:

$$\begin{aligned} \theta_{14} &= \frac{m_{14}}{n_{14}} \\ &= \frac{53}{62} \\ &= 0.8548 \\ \theta_{28} &= \frac{m_{28}}{n_{28}} \\ &= \frac{399}{425} \\ &= 0.9388 \\ \hat{\theta}_p &= \frac{m_{14} + m_{28}}{n_{14} + n_{28}} \\ &= \frac{53 + 399}{62 + 425} \\ &= 0.9281 \end{aligned}$$

where m_x, m_y are the number of successes in the two samples, and n_x, n_y is the total number of trials.

In this case, our test statistic is given by:

$$\begin{aligned} z_{(\theta_{14} - \theta_{28})} &= \frac{\theta_{14} - \theta_{28}}{\sqrt{\theta_p(1 - \theta_p)(1/n_{14} + 1/n_{28})}} \\ &= \frac{0.8548 - 0.9388}{\sqrt{0.9281(1 - 0.9281)(1/62 + 1/425)}} \\ &= -2.3919 \end{aligned}$$

```
pval = 2*pnorm(-abs(-2.3919))
pval
```

```
## [1] 0.01676141
```

Therefore, the p-value is given by:

$$\text{p-value } = 2\mathbb{P}(Z < -|z_{(\theta_{14} - \theta_{28})}|) = 0.01676$$

As the p-value is 0.01 < p < 0.05, we have moderate evidence against the null hypothesis. Therefore, we reject the null hypothesis and conclude at the 5% level of significance, that there is a difference in the probability of the concrete having compressive strength of 17MPa or greater between pours that have set for 14 days vs 28 days.