Assignment 1

MAST20005: Statistics

Kevin Yu

Student Number: 1462539

August 21, 2024

Exercise 1

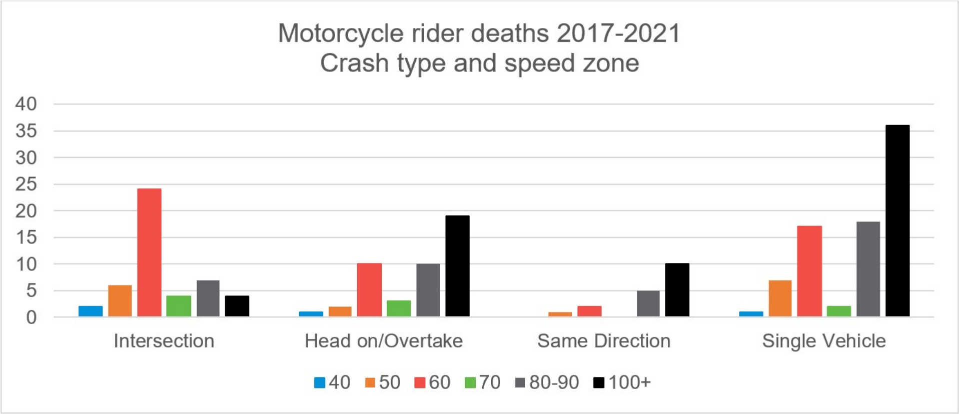


Figure 1: Motorcycle Rider Deaths 2017-2021 from TAC

Part 1

The above graph represents the number of motorcycle rider deaths in Victoria from 2017 to 2021, retrieved from the Transport Accident Commission (TAC) website [1]. Its purpose is to provide a visual representation to inform riders on the danger of riding motorcycles and to encourage them to take necessary precautions to reduce the number of deaths. Precautions can include placing yourself where you can see and be seen in intersections, appropriate Crash Avoidance Space, and maintaining a safe speed.

Part 2

- The data is shown clearly, with distinct colour-coded bars and its respective legend representing different speed zones across various crash types. The title is concise and informative, and there are no unnecessary distractions in the graph.

- The graph uses a common scale (y-axis) for comparing number of deaths acorss different crash type, which makes us easier to compare. This directly contrasts the still birthrate graph presented in the lecture, where it used different scales for domiciliary and overall deliveries, making it difficult to compare the two. Furthermore, the alingment is uniform and consistent, making it easier to read and interpret the data. Lastly, they have used a bar graph, which is a great choice for comparing data across different categories.

- The graph is flat, with no 3D effects or shadows that may distort the data; while also containing all the necessary information to successfully convey the message to the audience. The graph may be improved by encorporating patterns to the bar, which would make it more accessible to individuals with colour blindness.

- The graph has good visual encoding with each color representing a specific speed zone, and the height of the bars representing the number of deaths. The legend is placed at the bottom of the graph, which is a good location as it does not obstruct the data.

- This graph uses one of the standard form of visualisation, a bar chart, which is a good choice for comparing data across different categories (in this case, crash type). ■

[1] https://www.tac.vic.gov.au/road-safety/statistics/summaries/motorcycle-crash-data

Exercise 2

Part 1

Please note that I have abbreviated $\sum_{i=1}^{m}$ as $\sum$ at some places to avoid overcrowding

$$\begin{aligned}
L(\sqrt{p}) &= \prod_{i=1}^{m} \binom{n}{x_i} (\sqrt{p})^{x_i} (1-\sqrt{p})^{n-x_i} \\
&= \prod_{i=1}^{m} \left[\binom{n}{x_i} \right] (\sqrt{p})^{\sum x_i} (1-\sqrt{p})^{\sum (n-x_i)} \\
\implies \ln L(\sqrt{p}) &= \ln \prod_{i=1}^{m} \left[\binom{n}{x_i} \right] + \sum_{i=1}^{m} (x_i) \ln \sqrt{p} + \sum_{i=1}^{m} (n-x_i) \ln (1-\sqrt{p}) \\
\implies \frac{d}{d\sqrt{p}}[\ln L(\sqrt{p}))] &= \frac{\sum x_i \cdot \frac{1}{2} p^{1/2}}{\sqrt{p}} + \frac{(mn - \sum x_i) \cdot -\frac{1}{2} p^{1/2}}{1-\sqrt{p}} \\
&= \frac{(1-\sqrt{p}) p^{-1/2} \sum x_i - mn + \sum x_i}{2\sqrt{p}(1-\sqrt{p})}
\end{aligned}$$

Setting $\frac{d}{d\sqrt{p}}[\ln L(\sqrt{p}))] = 0$ yields,

$$\begin{aligned}
& (1-\sqrt{p}) p^{-1/2} \sum_{i=1}^{m} x_i - mn + \sum_{i=1}^{m} x_i = 0 \\
\implies & (p^{-1/2} - 1) \sum_{i=1}^{m} x_i - mn + \sum_{i=1}^{m} x_i = 0 \\
\implies & p^{-1/2} \sum_{i=1}^{m} x_i - mn = 0 \\
\implies & p^{-1/2} = \frac{mn}{\sum x_i} = \frac{n}{\bar{X}_m} \\
\implies & p^{1/2} = \frac{\bar{X}_m}{n} \\
& \therefore p = \frac{\bar{X}_m^2}{n^2} \qquad \blacksquare
\end{aligned}$$

Part 2

$$\begin{aligned}
\bar{x}_3 &= \frac{1}{3}(1+3+3) = \frac{7}{3} \\
\therefore p &= \frac{(7/3)^2}{5^2} = \frac{49}{9 \cdot 25} = \frac{49}{225} = 0.21\dot{7} \qquad \blacksquare
\end{aligned}$$

Exercise 3

Part 1

$$\begin{aligned}\mu(\theta) &= \int_1^\infty x\theta \left(\frac{1}{\theta}\right)^{\theta+1} dx \\ &= \theta \int_1^\infty x^{-\theta}\, dx \\ &= \theta \cdot \frac{1}{\theta - 1} = \frac{\theta}{\theta - 1}\end{aligned}$$

now let $\mu(\theta) = \bar{X}$,

$$\begin{aligned}&\implies \frac{\theta}{\theta - 1} = \bar{X} \\ &\implies \theta = \bar{X}\theta - \bar{X} \\ &\implies \theta(1 - \bar{X}) = -\bar{X} \\ \therefore \hat{\Theta}_{MM} &= \frac{\bar{X}}{\bar{X} - 1} \qquad \blacksquare\end{aligned}$$

Part 2

In addition to the variance, we can calculate

- the range (max - min) to measure the spread of the data,
- the interquartile range (IQR), the difference between the third and first quartiles, which is less sensitive to outliers,
- the median as a measure of central tendency that is less sensitive to outliers.

Now, let $y_{(1)}, y_{(2)}, \ldots, y_{(9)}$ and $z_{(1)}, z_{(2)}, \ldots, z_{(9)}$ be the ordered samples from Sample 1, Sample 2 respectively.

For Sample 1, $y_{(1)} = 1.333$ and $y_{(9)} = 1.684$ so the range is $1.684 - 1.333 = 0.351$.

For Sample 2, $z_{(1)} = 1.333$ and $z_{(9)} = 1.523$ so the range is $1.523 - 1.333 = 0.190$.

Now to calculate the interquantile range, we recall from the lecture that having $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ as the ordered observations; let the p-th quantile of the observations be denoted by $\hat{c}_p$ where $0 < p < 1$. Then, letting $k = 1 + (n-1)p$ and t and w be the whole and fractional part of k respectively, (i.e. $t = \lfloor k \rfloor$ and $w = k - t$),

$$\hat{c}_p = x_{(t)} + w(x_{(t+1)} - x_{(t)}).$$

Therefore, for Sample 1, the first quartile "position" is at $1 + (9-1)(0.25) = 3$ so the first quartile is $\hat{q}_1 = \hat{c}_{0.25} = y_{(3)} = 1.447$. The third quartile "position" is at $1 + (9-1)(0.75) = 7$ so the third quartile is $y_{(7)} = 1.577$. Hence, the IQR is $1.577 - 1.447 = 0.130$.

For Sample 2, the first and third quartile "position" is the same as Sample 1. Therefore, the first quartile is $z_{(3)} = 1.333$ and the third quartile is $z_{(7)} = 1.333$. Hence, the IQR is $1.333 - 1.333 = 0$. We also note that since IQR of Sample 2 is 0, so $z_{(9)} = 1.523$ is an extreme outlier.

Finally, the median of Sample 1 is $y_{(5)} = 1.529$ and the median of Sample 2 is $z_{(5)} = 1.333$.

Ultimately, the range and IQR of Sample 1 are greater than those of Sample 2. This suggests that Sample 1 has a greater variability, while Sample 2 has the same value for the first, second (medium) and third quartiles, indicating that the data is more concentrated around a single value. ■

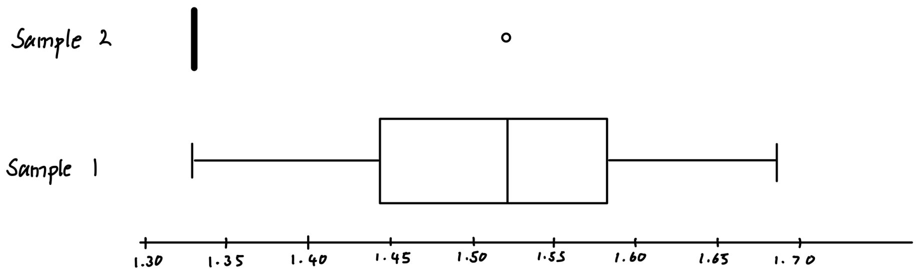


Figure 2: Boxplot of Sample 1 and Sample 2

Exercise 4

Part 1

Unbiasedness: An estimator T is unbiased for μ if $\mathbb{E}(T) = \mu$.

- **For T_1:**
$$\mathbb{E}(T_1) = \mathbb{E}\left(\frac{1}{6}\sum_{i=1}^{6} X_i\right) = \frac{1}{6}\sum_{i=1}^{6}\mathbb{E}(X_i) = \frac{1}{6} \times 6\mu = \mu$$
So, T_1 is an unbiased estimator of μ.
- **For T_2:** Following the same steps as above, T_2 is also an unbiased estimator of μ (since $X_i,\ Y_i \sim_{i.i.d.} N(\mu, \sigma^2 = 25\,\mathrm{km/h}^2)$).
- **For T_3:**
$$\mathbb{E}(T_3) = \mathbb{E}\left(\sum_{i=1}^{6} Y_i\right) = \sum_{i=1}^{6}\mathbb{E}(Y_i) = 6\mu \neq \mu$$
Therefore, T_3 is **not** an unbiased estimator of μ.

Part 2

Given $T_4 = aT_1 + (1-a)T_2$, the MSE of T_4 is defined as $\mathrm{MSE}(T_4) = \mathbb{E}[(T_4-\mu)^2] = \mathrm{var}(T_4) - [\mathrm{bias}(T_4)]^2$.

We find the bias of T_4 first. Since T_1 and T_2 are unbiased estimators of μ, we have

$$\mathbb{E}(T_4) = \mathbb{E}[aT_1 + (1-a)T_2] = a\mathbb{E}(T_1) + (1-a)\mathbb{E}(T_2) = a\mu + (1-a)\mu = \mu$$

so

$$\mathrm{bias}(T_4) = \mu - \mu = 0.$$

.

Now for the variance of T_4, we first note that $\mathrm{var}(T_1) = \mathrm{var}(\bar{X}) = \frac{\sigma^2}{6}$ and similarly, $\mathrm{var}(T_2) = \frac{\sigma^2}{6}$, where $\sigma^2 = 25\,\mathrm{km/h}^2$. Then

$$\begin{aligned}\mathrm{var}(T_4) &= \mathrm{var}[aT_1 + (1-a)T_2] \\ &= a^2\,\mathrm{var}(T_1) + (1-a)^2\,\mathrm{var}(T_2) \quad (\text{since } T_1 \text{ and } T_2 \text{ are independent}) \\ &= a^2\frac{\sigma^2}{6} + (1-a)^2\frac{\sigma^2}{6} \\ &= \frac{\sigma^2}{6}(2a^2 - 2a + 1) \\ &= \frac{25}{6}(2a^2 - 2a + 1).\end{aligned}$$

Therefore, the MSE of T_4 is just

$$\mathrm{MSE}(T_4) = \mathrm{var}(T_4) = \frac{25}{6}(2a^2 - 2a + 1).$$

Part 3

We note that T_1 and T_2 are both sample means of normal distributions, so they follow normal distributions:

$$T_1 \sim N\left(\mu, \frac{\sigma^2}{6}\right), \quad T_2 \sim N\left(\mu, \frac{\sigma^2}{6}\right)$$

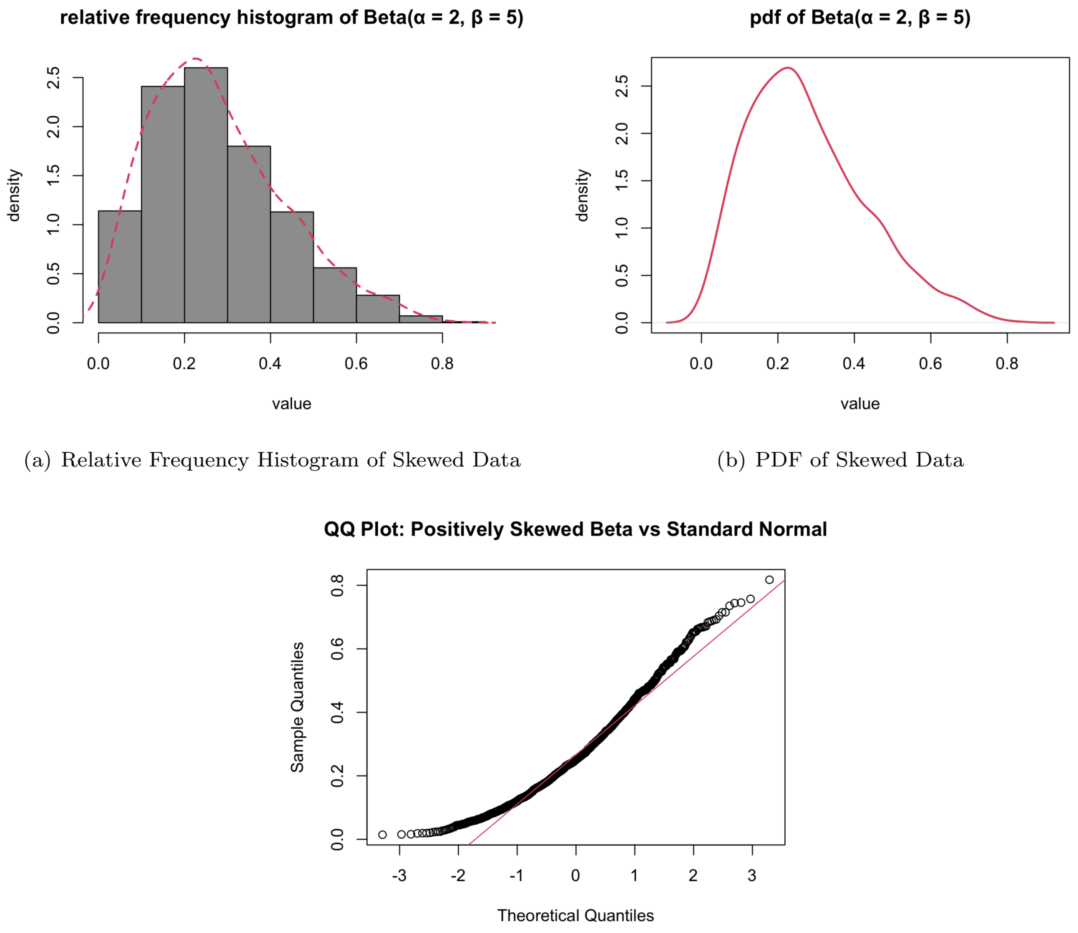
Since T_4 is a linear combination of two independent normal variables T_1 and T_2, it also follows a normal distribution given by

$$T_4 \sim N\left(\mu, \frac{25}{6}(2a^2 - 2a + 1)\right) \quad \blacksquare.$$

Exercise 5

```
set.seed(1125)

beta_skewed_positive <- rbeta(1000, shape1 = 2, shape2 = 5)

hist(beta_skewed_positive, freq = FALSE,
     main = "relative␣frequency␣histogram␣of␣Beta(a␣=␣2,␣b␣=␣5)",
     xlab = "value", ylab = "density", col = 8)
smooth.density = density(beta_skewed_positive)
lines(smooth.density, lty = 2, lwd = 2, col = 2)

plot(density(beta_skewed_positive), main = "pdf␣of␣Beta(a␣=␣2,␣b␣=␣5)",
     xlab = "value", ylab = "density", col = 2, lwd = 2)

qqnorm(beta_skewed_positive,
       main = "QQ␣Plot:␣Positively␣Skewed␣Beta␣vs␣Standard␣Normal")
qqline(beta_skewed_positive, col = 2)
```

Listing 1: R code for generating and plotting a positively skewed Beta distribution



(a) Relative Frequency Histogram of Skewed Data

(b) PDF of Skewed Data

(c) Q-Q Plot of Skewed Data

Figure 3: Q-Q Plot of Skewed Data

The QQ plot in Figure 3(c) shows an upward curvature where the points deviate above the line, especially in the tails of the plot. This makes intuitive sense as the lower quantiles are larger in the positive skewed data than in the normal distribution. Furthermore, the right tail also deviates above the line, since in a positively skewed distribution, the right tail is longer and thicker. This leads to higher quantiles at the upper end of the distribution compared to the normal distribution.

The following results suggests that if the data is negatively skewed, for example, $Beta(\alpha = 5, \beta = 2)$, we will see both tails deviate below the line in the QQ plot. ■