STAT2203/7203: Assignment 2

Alexander White

43218307

1. Transformation of a Random Variable

(a) **Find the pdf of $Y = 10 + 2X$:**
X has pdf $f_X(x) = 1$ for $x \in (0, 1)$ and $f_X(x) = 0$ elsewhere.

Determining the range of Y:

$$Y = 10 + 2X$$

As $X \in (0, 1)$:

$$\begin{aligned} \text{For } X = 0: &\quad Y = 10 + 0 = 10, \\ \text{For } X = 1: &\quad Y = 10 + 2 = 12. \end{aligned}$$

Therefore, $Y \in (10, 12)$.

Since Y is of the form $Y = aX + b$, where $a = 2$ and $b = 10$, we can use the formula:

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y - b}{a}\right).$$

Substituting the values:

$$f_Y(y) = \frac{1}{|2|} f_X\left(\frac{y - 10}{2}\right).$$

Since $f_X(x) = 1$ for $x \in (0, 1)$, we have:

$$f_Y(y) = \frac{1}{2}, \quad \text{for } y \in (10, 12).$$

Therefore, the pdf of Y is:

$$f_Y(y) = \begin{cases} \frac{1}{2}, & 10 < y < 12, \\ 0, & \text{otherwise.} \end{cases}$$

(b) **Find the pdf of $Y = (X - \frac{1}{3})^2 - \frac{1}{3}$:**
Given that $X \in (0, 1)$, its pdf is:

$$f_X(x) = 1, \quad 0 < x < 1.$$

To find the CDF of Y, we use:

$$F_Y(y) = P\left((X - \frac{1}{3})^2 - \frac{1}{3} \leq y\right).$$

Solving for X:

$$(X - \frac{1}{3})^2 \leq y + \frac{1}{3}.$$

Taking the square root:

$$-\sqrt{y + \frac{1}{3}} \leq X - \frac{1}{3} \leq \sqrt{y + \frac{1}{3}}.$$

Rearranging for X:

$$\frac{1}{3} - \sqrt{y + \frac{1}{3}} \leq X \leq \frac{1}{3} + \sqrt{y + \frac{1}{3}}.$$

Since $X \in (0,1)$, X is within:

$$\max\left(0, \frac{1}{3} - \sqrt{y + \frac{1}{3}}\right) \leq X \leq \min\left(1, \frac{1}{3} + \sqrt{y + \frac{1}{3}}\right).$$

Thus, the CDF of Y is:

$$F_Y(y) = \min\left(1, \frac{1}{3} + \sqrt{y + \frac{1}{3}}\right) - \max\left(0, \frac{1}{3} - \sqrt{y + \frac{1}{3}}\right).$$

Differentiating to find the pdf $f_Y(y)$:

$$f_Y(y) = \frac{d}{dy}\left(\sqrt{y + \frac{1}{3}} - \left(-\sqrt{y + \frac{1}{3}}\right)\right) = \frac{d}{dy}\left(2\sqrt{y + \frac{1}{3}}\right).$$

$$f_Y(y) = \frac{2}{2\sqrt{y + \frac{1}{3}}} = \frac{1}{\sqrt{y + \frac{1}{3}}}$$

y is within the interval $-\frac{2}{9} < y < \frac{1}{9}$, outside this interval, the pdf is 0. Therefore, the pdf of Y is:

$$f_Y(y) = \begin{cases} \frac{1}{\sqrt{y+\frac{1}{3}}}, & -\frac{2}{9} < y < \frac{1}{9}, \\ 0, & \text{otherwise.} \end{cases}$$

2. Discrete Distribution

- Book has 100 pages.
- Average of 5 misprints per page.
- Occur according to a Poisson distribution.
- Occurrence of misprints on different pages is mutually independent.

(a) **Probability that a page contains no misprints:**
As the misprints on any given page follows a Poisson distribution, this can be represented with $\lambda = 5$ as:

$$f(x) = e^{-\lambda}\frac{\lambda^x}{x!}, \quad x = 0, 1, 2, ...$$

For $P(X = 0)$:

$$f(0) = e^{-5}\frac{5^0}{0!} = e^{-5} \approx 0.0067$$

Therefore, the probability that a page contains no misprints is e^{-5} or 0.67%

(b) **Probability that there will be at least 80 pages that contain more than 2 misprints:**
This can be represented as a binomial distribution, where the "success" is if the page contains more than 2 misprints. Let Y be the number of pages that contain more than 2 misprints. Then Y is a Binomial random variable:

$$Y \sim \text{Binomial}(n = 100, p = P(X > 2)).$$

Firstly, need to find $P(X > 2)$:

$$P(X > 2) = 1 - P(X \leq 2).$$

Therefore $P(X \leq 2)$ for $X \sim \text{Poisson}(\lambda = 5)$:

$$P(X = 0) = e^{-5}$$
$$P(X = 1) = 5e^{-5}$$
$$P(X = 2) = \frac{25e^{-5}}{2} = 12.5e^{-5}.$$
$$P(X \leq 2) = e^{-5} + 5e^{-5} + 12.5e^{-5} = 18.5e^{-5}.$$

Therefore,

$$P(X > 2) = 1 - 18.5e^{-5} \approx 0.875$$

To find $P(Y \geq 80)$, we can use the cumulative probability of $P(Y \leq 79)$:

$$P(Y \geq 80) = 1 - P(Y \leq 79).$$

Using R, and the pbinom function we can solve this:

```
p <- 0.8753
n <- 100
1 - pbinom(79, size = n, prob = p)
```

Result: 0.9889 or 98.89% chance of there being at least 80 pages that contain more than 2 misprints.

3. Discrete Distribution

- X_1 and X_2 are independent random variables.
- X_i has Poisson distribution with mean $\lambda_i, i = 1, 2$.

(a) **Show that $X_1 + X_2$ has a Poisson distribution with mean $\lambda_1 + \lambda_2$:**
The pmf of a Poisson distribution is:

$$P(X_i = k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

$$\begin{aligned} P(X_1 + X_2 = n) &= \sum_{k=0}^{n} P(X_1 = k)P(X_2 = n - k) \\ &= \sum_{k=0}^{n} e^{-\lambda_1}\frac{\lambda_1^k}{k!}e^{-\lambda_2}\frac{\lambda_2^{n-k}}{(n-k)!} \\ &= e^{-(\lambda_1+\lambda_2)}\sum_{k=0}^{n}\frac{\lambda_1^k\lambda_2^{n-k}}{k!(n-k)!} \end{aligned}$$

Divide by $n!$ to isolate the binomial term:

$$= \frac{e^{-(\lambda_1+\lambda_2)}}{n!}\sum_{k=0}^{n}\frac{n!}{k!(n-k)!}\lambda_1^k\lambda_2^{n-k}$$

As $\sum_{k=0}^{n}\frac{n!}{k!(n-k)!}\lambda_1^k\lambda_2^{n-k}$ is the binomial theorem:

$$= \frac{e^{-(\lambda_1+\lambda_2)}}{n!}(\lambda_1 + \lambda_2)^n$$

(b) **Calculate the conditional pmf of $X_1 + X_2$ given $X_1 = k$:**
Want to find $P(X_1 + X_2 = n \mid X_1 = k)$:

$$P(X_1 + X_2 = n \mid X_1 = k) = \frac{P(X_1 = k, X_1 + X_2 = n)}{P(X_1 = k)}.$$

Since X_1 and X_2 are independent:

$$P(X_1 = k, X_1 + X_2 = n) = P(X_1 = k)P(X_2 = n - k).$$

Therefore,

$$P(X_1 + X_2 = n \mid X_1 = k) = \frac{P(X_1 = k)P(X_2 = n - k)}{P(X_1 = k)} = P(X_2 = n - k).$$

Since $X_2 \sim \text{Poisson}(\lambda_2)$, we have:

$$P(X_2 = n - k) = \frac{\lambda_2^{n-k}e^{-\lambda_2}}{(n-k)!}, \quad n \geq k.$$

Therefore, the conditional pmf is:

$$P(X_1 + X_2 = n \mid X_1 = k) = \frac{\lambda_2^{n-k}e^{-\lambda_2}}{(n-k)!}, \quad n \geq k.$$

When $n < k$, we have $n - k < 0$. Since X_2 is a Poisson random variable, it cannot be negative. Therefore,

$$P(X_2 = n - k) = 0, \quad \text{if } n < k.$$

(c) **Determine the conditional distribution of X1 given X1 + X2 = k:**
Want to find $P(X_1 = j \mid X_1 + X_2 = k)$ for $k \in \{0, 1, 2, \ldots\}$.

- $X_1 \sim \text{Poisson}(\lambda_1)$
- $X_2 \sim \text{Poisson}(\lambda_2)$
- X_1 and X_2 are independent
- $X_1 + X_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$.

- X_1 represents the number of events attributed to the first Poisson process out of a total of k events.

- The probability of an event being attributed to X_1 is the proportion of events out of the total events: $\frac{\lambda_1}{\lambda_1+\lambda_2}$.

Therefore, the conditional distribution is:

$$X_1 \mid (X_1 + X_2 = k) \sim \text{Binomial}\left(k, \frac{\lambda_1}{\lambda_1 + \lambda_2}\right).$$

The pmf of this binomial distribution is:

$$P(X_1 = j \mid X_1 + X_2 = k) = \binom{k}{j} \left(\frac{\lambda_1}{\lambda_1 + \lambda_2}\right)^j \left(1 - \frac{\lambda_1}{\lambda_1 + \lambda_2}\right)^{k-j},$$

where $j = 0, 1, 2, \ldots, k$.

4. Continuous Distribution

- Exponential distribution with mean $35,200$

$$X \sim \text{Exponential}(\lambda),$$

- Therefore, the rate parameter of X is $\lambda = \frac{1}{35,200}$.
- The probability density function (pdf) of an exponential distribution is:

$$f(x;\lambda) = \lambda e^{-\lambda x}, \quad x \geq 0.$$

- The cumulative distribution function (CDF) is:

$$F(x;\lambda) = 1 - e^{-\lambda x}, \quad x \geq 0.$$

(a) **What is the median (0.5-quantile) income?**

Need to solve for x in the exponential distribution equation:

Using the CDF of the exponential distribution:

$$1 - e^{-\lambda x} = 0.5.$$
$$-e^{-\lambda x} = 0.5 - 1$$
$$e^{-\lambda x} = 0.5$$

Solving for x, take natural log of both sides:

$$-\lambda x = \ln(0.5).$$
$$x = -\frac{\ln(0.5)}{\lambda}.$$

Substitute $\lambda = \frac{1}{35,200}$:

$$x = -\frac{\ln(0.5)}{\frac{1}{35,200}} = -35,200 \ln(0.5) = 24398.78$$

Using R to check:

```
qexp(0.5, (1/35200))
```

output: 24398.78

Therefore, the median (0.5-quantile) income is approximately $\$24,398.78$.

(b) **What is the probability that a person's income is less than \$10,000?**

Need to find $P(X < 10,000)$.

Using the CDF of the exponential distribution:

$$P(X < 10,000) = F(10,000;\lambda) = 1 - e^{-\lambda \cdot 10,000}.$$

Substitute $\lambda = \frac{1}{35,200}$:

$$P(X < 10,000) = 1 - e^{-\frac{10,000}{35,200}} \approx 0.2473$$

Check using R:
`pexp(10000,(1/35200))`
Output: 0.2473

Thus, the probability that a person's income is less than \$10,000 is approximately **24.73**%.

(c) **What income would put a person in the top 1% of earners?**

Top 1% of earners corresponds to the 99th percentile, or 0.99-quantile.
Need to solve for x in the equation $P(X \leq x) = 0.99$.

Using the CDF to solve for x:

$$F(x;\lambda) = 1 - e^{-\lambda x} = 0.99.$$
$$e^{-\lambda x} = 0.01$$
$$-\lambda x = \ln(0.01).$$
$$x = -\frac{\ln(0.01)}{\lambda}.$$

Sub in $\lambda = \frac{1}{35,200}$:

$$x = -\frac{\ln(0.01)}{\frac{1}{35,200}} = -35,200\ln(0.01) \approx 162102$$

Therefore, the income required to be in the top 1% of earners is approximately \$162, 102.

5. Continuous Distribution

- $X \sim N(\mu, \sigma^2)$
- $Y = e^X$
- $Y \sim lnN(\mu, \sigma^2)$

(a) **Find the pdf of Y**

$$Y = e^X$$
$$ln(Y) = X$$

$$F_Y(y) = P(Y \leq y) = P(e^X \leq y) = P(X \leq \ln(y)).$$

To find the pdf of Y:

$$f_Y(y) = f_X(\ln(y)) \cdot \left| \frac{d}{dy} \ln(y) \right|, \quad y > 0.$$

The pdf of X is given by:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}.$$

Substituting $x = \ln(y)$ into the pdf:

$$f_X(\ln(y)) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln(y)-\mu)^2}{2\sigma^2}}.$$

The derivative of $\ln(y)$ with respect to y:

$$\frac{d}{dy} \ln(y) = \frac{1}{y}.$$

Therefore, the pdf of Y becomes:

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln(y)-\mu)^2}{2\sigma^2}} \cdot \frac{1}{y}.$$

$$f_Y(y) = \frac{1}{y\sigma\sqrt{2\pi}} e^{-\frac{(\ln(y)-\mu)^2}{2\sigma^2}}, \quad y > 0.$$

Therefore, the pdf of $Y \sim \ln N(\mu, \sigma^2)$ is:

$$f_Y(y) = \frac{1}{y\sigma\sqrt{2\pi}} e^{-\frac{(\ln(y)-\mu)^2}{2\sigma^2}}, \quad y > 0.$$

(b) **Find the distribution of** $1/Y$
Let $Z = 1/Y$. As $Y = 1/X$:

$$Z = \frac{1}{e^X}$$
$$= e^{-X}$$

As $X \sim N(\mu, \sigma^2)$, then we know that:

$$-X \sim N(-\mu, \sigma^2).$$

From 5.a we know $e^X \sim \ln N(\mu, \sigma^2)$, therefore e^{-X} will be log-normally distributed as well, but with parameters $-\mu$ and σ^2. Therefore,

$$Z = 1/Y = e^{-X} \sim \ln N(-\mu, \sigma^2).$$

Additionally, the pdf of Z is given by:

$$f_Z(z) = \frac{1}{z\sigma\sqrt{2\pi}} e^{-\frac{(\ln(z)+\mu)^2}{2\sigma^2}}, \quad z > 0.$$

(c) **Find the distribution of $3\sqrt{Y}$**

Let $W = 3\sqrt{Y}$. As $Y = e^X$:

$$\sqrt{Y} = \sqrt{e^X} = e^{X/2}.$$

Therefore,

$$W = 3\sqrt{Y} = 3e^{X/2}.$$

Need to find distribution of $\frac{X}{2}$.
Given that $X \sim N(\mu, \sigma^2)$, we know given the linear transformation:

$$\frac{X}{2} \sim N\left(\frac{\mu}{2}, \frac{\sigma^2}{4}\right).$$

From results of 5.a, we know $e^{X/2} \sim \ln N\left(\frac{\mu}{2}, \frac{\sigma^2}{4}\right)$.

Multiplying a log-normal random variable by a constant results in another log-normal random variable.
E.g. if $V \sim \ln N(\mu, \sigma^2)$, then for a constant $c > 0$:

$$cV \sim \ln N(\ln(c) + \mu, \sigma^2).$$

Therefore, since $\sqrt{Y} \sim \ln N\left(\frac{\mu}{2}, \frac{\sigma^2}{4}\right)$, the answer is:

$$3\sqrt{Y} \sim \ln N\left(\ln(3) + \frac{\mu}{2}, \frac{\sigma^2}{4}\right).$$

(d) **Find the distribution of the product W Y.**

- $W \sim \ln N(\mu W, \sigma^2 W)$
- $Y \sim \ln N(\mu Y, \sigma^2 Y)$

Want to find the distribution of the product:

$$Z = W \cdot Y.$$

Since $W \sim \ln N(\mu_W, \sigma_W^2)$ and $Y \sim \ln N(\mu_Y, \sigma_Y^2)$

$$\ln(W) \sim N(\mu_W, \sigma_W^2), \quad \ln(Y) \sim N(\mu_Y, \sigma_Y^2).$$

Now consider the logarithm of the product $Z = W \cdot Y$:

$$\ln(Z) = \ln(W \cdot Y) = \ln(W) + \ln(Y).$$

Since we know $\ln(W)$ and $\ln(Y)$ are independent and normally distributed, their sum is also normally distributed:

$$\ln(Z) \sim N(\mu_W + \mu_Y, \sigma_W^2 + \sigma_Y^2).$$

Therefore, the random variable $Z = W \cdot Y$ is log-normally distributed with parameters $\mu_W + \mu_Y$ and $\sigma_W^2 + \sigma_Y^2$:

$$W \cdot Y \sim \ln N(\mu_W + \mu_Y, \sigma_W^2 + \sigma_Y^2).$$

(e) **Assessing Distribution of Rainfall Data**
Can assess the distribution of the rainfall data using R.

(a) Load the data into R

(b) Transform to a logarithmic variable

(c) Use a QQ plot to check normality of the transformed variable

R Code:

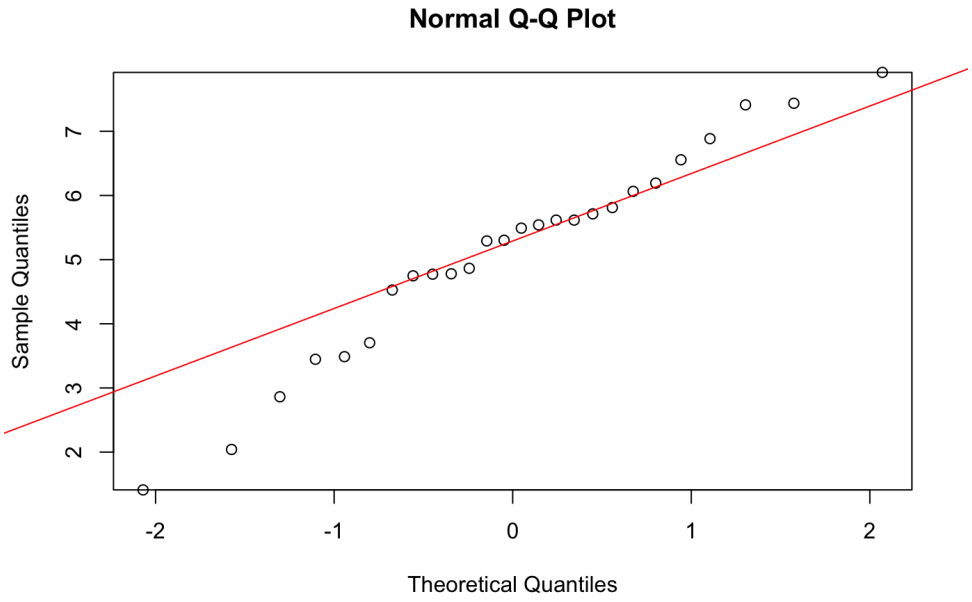
```
library(ggplot2)

rainfall <- read.csv("rainfall.csv")

# Transform the data
log_rainfall <- log(rainfall$rainfall)

# QQ plot
qqnorm(log_rainfall)
qqline(log_rainfall, col = "red")
```

As shown in the figure above, the logarithmic transformed rainfall data mostly follows the normal distribution. A normal distribution would follow the red line plotted on the chart. There is some deviance among the first 7 or so points and last 4, this indicates some skewness in the data along the tails. A normal distribution could be adequate for this data set, but as there are significant deviations among the lower and upper quantiles, some values may not be captured well by this model.



6. Multiple Random Variables

X_1, X_2, X_3 are independent Bernoulli random variables with probability p.

$$Y = (1 - X_1)(X_2 + X_3 - X_2X_3).$$

(a) **Find the expected value and variance of Y**

$$Y = (1 - X_1)(X_2 + X_3 - X_2X_3).$$

Noting that as X_1 is a Bernoulli random variable, it has a binary p value, either 1 or 0:

$$1 - X_1 = \begin{cases} 1, & \text{if } X_1 = 0, \\ 0, & \text{if } X_1 = 1. \end{cases}$$

Therefore,

$$Y = \begin{cases} X_2 + X_3 - X_2X_3, & \text{if } X_1 = 0, \\ 0, & \text{if } X_1 = 1. \end{cases}$$

Using the law of total expectation:

$$E[Y] = E[Y \mid X_1 = 0]P(X_1 = 0) + E[Y \mid X_1 = 1]P(X_1 = 1).$$

Since $P(X_1 = 1) = p$ and $P(X_1 = 0) = 1 - p$, we have:

$$E[Y \mid X_1 = 1] = 0.$$

When $X_1 = 0$, we have:

$$E[Y \mid X_1 = 0] = P(X_2 + X_3 - X_2X_3 = 1) = 1 - P(X_2 = 0, X_3 = 0).$$

Since X_2 and X_3 are independent:

$$P(X_2 = 0, X_3 = 0) = (1-p)^2.$$

Thus,

$$E[Y \mid X_1 = 0] = 1 - (1-p)^2.$$

Therefore,

$$E[Y] = (1-(1-p)^2)(1-p) + 0 \cdot p = (1-(1-p)^2)(1-p).$$

Simplifying:

$$E[Y] = (1-(1-2p+p^2))(1-p) = (2p-p^2)(1-p).$$

Therefore the expected value of Y is:

$$E[Y] = 2p(1-p) - p^2(1-p) = 2p - 3p^2 + p^3.$$

To find the variance $\mathrm{Var}(Y)$, we use:

$$\mathrm{Var}(Y) = E[Y^2] - (E[Y])^2.$$

Since Y is binary (0 or 1), $Y^2 = Y$, so $E[Y^2] = E[Y]$. Thus:

$$\mathrm{Var}(Y) = E[Y] - (E[Y])^2.$$

Substituting the value of $E[Y]$:

$$\mathrm{Var}(Y) = (2p - 3p^2 + p^3) - (2p - 3p^2 + p^3)^2.$$

(b) **Find** $E[Y \mid X_3]$

Need to find $E[Y \mid X_3]$.

$$Y = (1 - X_1)(X_2 + X_3 - X_2X_3).$$

Given X_3, there are two cases:

- $X_3 = 0$**:**

  $$Y = (1 - X_1)X_2.$$

  Therefore,

  $$E[Y \mid X_3 = 0] = E[(1 - X_1)X_2].$$

  Since X_1 and X_2 are independent:

  $$E[Y \mid X_3 = 0] = E[1 - X_1]E[X_2] = (1-p)p = p(1-p).$$

- $X_3 = 1$**:**

  $$Y = (1 - X_1)(X_2 + 1 - X_2) = (1 - X_1)(1) = 1 - X_1.$$

  Thus,

  $$E[Y \mid X_3 = 1] = E[1 - X_1] = 1 - p.$$

Combining both cases:

$$E[Y \mid X_3] = (1-p)X_3 + p(1-p)(1-X_3).$$

(c) **Find** $E[X_3 \mid Y = 1]$ Given that $Y = 1$, there are 3 possible cases:

1. $X_1 = 0, X_2 = 1, X_3 = 0$
2. $X_1 = 0, X_2 = 0, X_3 = 1$
3. $X_1 = 0, X_2 = 1, X_3 = 1$

Therefore, $E[X_3 \mid Y = 1]$:

$$E[X_3 \mid Y = 1] = P(X_3 = 1 \mid Y = 1) = \frac{P(Y = 1, X_3 = 1)}{P(Y = 1)}.$$

Need to find both $P(Y = 1, X_3 = 1)$ and $P(Y = 1)$. $P(Y = 1)$

For $Y = 1$, we must have $X_1 = 0$. Therefore, the possible cases are:

$$P(Y = 1) = P(X_1 = 0, X_2 = 1, X_3 = 0) + P(X_1 = 0, X_2 = 0, X_3 = 1) + P(X_1 = 0, X_2 = 1, X_3 = 1).$$

As X_1, X_2, X_3 are independent Bernoulli random variables:

$$P(X_1 = 0, X_2 = 1, X_3 = 0) = (1-p) \cdot p \cdot (1-p) = p(1-p)^2,$$
$$P(X_1 = 0, X_2 = 0, X_3 = 1) = (1-p) \cdot (1-p) \cdot p = p(1-p)^2,$$
$$P(X_1 = 0, X_2 = 1, X_3 = 1) = (1-p) \cdot p \cdot p = p^2(1-p).$$

Therefore, the total probability when $Y = 1$ is:

$$P(Y = 1) = p(1-p)^2 + p(1-p)^2 + p^2(1-p) = 2p(1-p)^2 + p^2(1-p).$$

Simplifying:

$$P(Y = 1) = p(1-p)(2(1-p) + p) = p(1-p)(2-p).$$

Also need to find $P(Y = 1, X_3 = 1)$

Finding the joint probability when $Y = 1$ and $X_3 = 1$:

1. $X_1 = 0, X_2 = 0, X_3 = 1$
2. $X_1 = 0, X_2 = 1, X_3 = 1$

Thus,

$$P(Y = 1, X_3 = 1) = P(X_1 = 0, X_2 = 0, X_3 = 1) + P(X_1 = 0, X_2 = 1, X_3 = 1).$$

Substituting the values:

$$P(Y = 1, X_3 = 1) = (1-p)(1-p)p + (1-p)p \cdot p.$$

$$P(Y = 1, X_3 = 1) = (1-p)^2 p + (1-p)p^2.$$

$$P(Y = 1, X_3 = 1) = p(1-p)((1-p)+p) = p(1-p).$$

Using this we can calculate $E[X_3 \mid Y = 1]$:

$$E[X_3 \mid Y = 1] = P(X_3 = 1 \mid Y = 1) = \frac{P(Y = 1, X_3 = 1)}{P(Y = 1)}.$$

Substitute the values:

$$E[X_3 \mid Y = 1] = \frac{p(1-p)}{p(1-p)(2-p)}.$$

Simplifying:

$$E[X_3 \mid Y = 1] = \frac{1}{2-p}.$$

Therefore, the conditional expectation of X_3 given that $Y = 1$ is:

$$E[X_3 \mid Y = 1] = \frac{1}{2-p}.$$

7. Confidence Interval

(a) **Sample Size for 95% Confidence Interval of a Normal Distribution**

The formula for a confidence interval for the mean μ is given by:

$$\bar{X} \pm z^* \frac{\sigma}{\sqrt{n}},$$

where:

- $\bar{X}$ is the sample mean.
- z^* is the critical value corresponding to a 95% confidence level. For 95%, $z^* \approx 1.96$.
- σ is the standard deviation of the population.
- n is the sample size.

The length of the confidence interval is:

$$2 \times z^* \frac{\sigma}{\sqrt{n}}.$$

Need this length to be less than 0.01σ. Therefore:

$$2 \times 1.96 \frac{\sigma}{\sqrt{n}} < 0.01\sigma.$$

Simplifying:

$$\frac{3.92\sigma}{\sqrt{n}} < 0.01\sigma.$$

Divide both sides by σ:

$$\frac{3.92}{\sqrt{n}} < 0.01.$$

Rearranging:

$$\sqrt{n} > \frac{3.92}{0.01} = 392$$
$$n > 392^2 = 153664$$

Therefore, the sample size n must be at least **153665** to achieve a 95% confidence interval for μ with a length less than 0.01σ.

(b) **Confidence Interval for θ with Uniform Distribution**

- X_1 and X_2 are taken at random from a uniform distribution on $[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$, where $\theta \in \mathbb{R}$ is unknown.
- Let $Y_1 = \min\{X_1, X_2\}$ and $Y_2 = \max\{X_1, X_2\}$.
- Need to show that (Y_1, Y_2) is a 50% confidence interval for θ.

As X_1 and X_2 are independent and taken from same uniform distribution $U[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$. The pdf is:

$$f(x) = \begin{cases} 1, & \text{if } x \in [\theta - \frac{1}{2}, \theta + \frac{1}{2}], \\ 0, & \text{otherwise.} \end{cases}$$

Therefore, $Y_1 = \min(X_1, X_2)$ and $Y_2 = \max(X_1, X_2)$ fall within $[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$.
For (Y_1, Y_2) to cover θ, we need $Y_1 < \theta < Y_2$. This has a 50% chance of occurring as the uniform distribution is symmetric.

Therefore, (Y_1, Y_2) is a 50% confidence interval for θ.

(c) **100% Confidence Interval for θ Given Additional Information**

- $Y_1 = y_1$ and $Y_2 = y_2$ such that $y_2 - y_1 > \frac{1}{2}$
- Show that (y_1, y_2) is a 100% confidence interval for θ.

Since X_1, X_2 are taken from a uniform distribution, $U[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$, Y_1 and Y_2 lie within this interval.

$$y_2 - y_1 > \frac{1}{2}$$

Therefore,

$$(y_1, y_2) > [\theta - \frac{1}{2}, \theta + \frac{1}{2}]$$

Thus, θ must lie within (y_1, y_2).

Consequently, (y_1, y_2) must cover θ with 100% confidence.

(d) **90% Confidence Interval for Mean Texting Speed While Sitting**

- R was used for checking normality and calculating the confidence interval.

- When graphed on a QQ plot, the sitting wpm values aligned along the constant qqline, indicating that the data followed a normal distribution.
- A confidence interval was then created by using the formula:

$$\bar{X} \pm z^* \frac{\sigma}{\sqrt{n}},$$

R Code:

```
data <- read.csv("TextSpeed.csv")

sitting_speed <- data$SitWPM

# Checking normality of sitting speed
qqnorm(sitting_speed)
qqline(sitting_speed, col = "red")

# Calculating the confidence interval
n <- length(sitting_speed)
mean_sitting <- mean(sitting_speed)
sd_sitting <- sd(sitting_speed)

# Standard error
se <- sd_sitting / sqrt(n)

# Critical value for 90% confidence (two-tailed)
t_value <- qt(0.95, df = n - 1)

# Confidence interval
lower_bound <- mean_sitting - t_value * se
upper_bound <- mean_sitting + t_value * se
cat("90% Confidence Interval for Sitting Texting Speed: (", lower_bound, ", ", uppe
```

Output:
90% Confidence Interval for Sitting Texting Speed: (40.25769 , 41.90897)