Introduction to Statistics for DS
Individual Assignment 1

Zehao Qian

October 19, 2023

1 Question 1

I've highlight the keywords and calculation steps in the questions part.

1. B

2. A

3. B

4. A

5. D

2 Question 2

2.1 Find the modal tooth length across the entire data set.

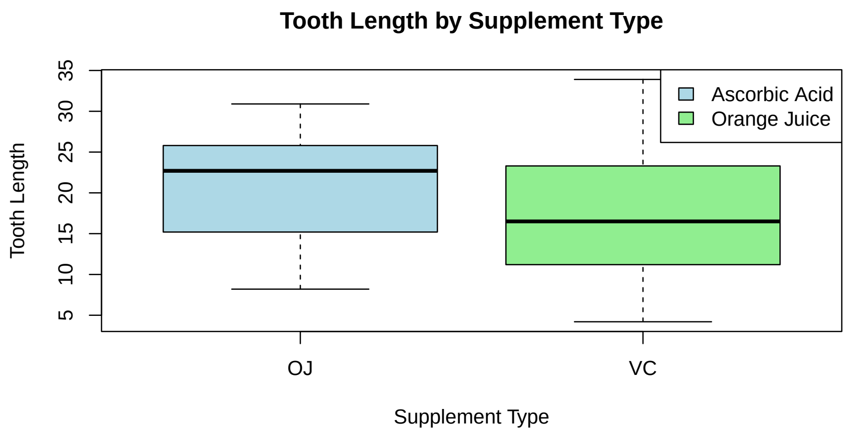
```
# clear the console area
cat("\014")

# clear the environment var area
rm(list = ls())

# Load the Tooth Growth data set
data(ToothGrowth

# Find the modal tooth length
modal_tooth_length1 = as.numeric(names(sort(
    table(ToothGrowth$len), decreasing = TRUE)[1]))
```

2.2 Mean tooth length for guinea pigs who were given their vitamins via orange juice

```
# Calculate the mean tooth length for guinea pigs
# given vitamins via orange juice
mean_tooth_length_orange = mean(
    ToothGrowth$len[ToothGrowth$supp == "OJ"])
mean_tooth_length_orange
```

2.3 Create a side-by-side box-and-whisker plot:

```
# Create a side-by-side box-and-whisker plot
# for tooth length by supplement type
boxplot(len ~ supp, data = ToothGrowth,
        col = c("lightblue", "lightgreen"),
        xlab = "Supplement␣Type", ylab = "Tooth␣Length",
        main = "Tooth␣Length␣by␣Supplement␣Type")
legend("topright",
        legend = c("Ascorbic␣Acid", "Orange␣Juice"),
        fill = c("lightblue", "lightgreen"))
```



2.4 Comment on which vitamin delivery approach is more effective:

To determine which vitamin delivery approach is more effective in promoting tooth growth, you should analyze the box-and-whisker plot created in the previous step. Look at the distribution of tooth lengths for guinea pigs given vitamins via ascorbic acid and orange juice. If the median tooth length for one group is higher than the other, it suggests that the corresponding supplement may be more effective. Additionally, look for differences in the spread (interquartile range) and the presence of outliers. A supplement with a larger median, smaller spread, and fewer outliers may be considered more effective.

3 Question 3

3.1 Expected number of party poppers

$$E(X) = Probability\ of\ failure * Total\ number\ of\ party\ poppers$$

$$E(X) = 0.6\% * 200 = 1.2$$

3.2 Distribution of fail poppers

3.2.1 Poisson distribution

To represent the number of party poppers that fail to go off in a box as a random variable, X, we can use the Poisson distribution. In this case, the Poisson distribution is appropriate because it models the number of events (party poppers failing to go off) that occur in a fixed interval (a box of party poppers) when the events are rare and random.

The parameter, λ (lambda), is the average number of events in the fixed interval. Here, λ is equal to the expected number of party poppers that will fail to go off, which was calculated in Part 1.

3.2.2 The assumptions that need to hold for justifying the use of the Poisson distribution

- Events (party poppers failing to go off) occur randomly.
- Events are rare, meaning the probability of more than one event occurring in a very short time period is negligible.
- Events are independent of each other.

3.3 Distribution of mistake box

3.3.1 Distribution for Y

For the random variable Y, representing the number of party poppers that fail to go off in a box of 2 (due to the administrative error), you can still use the Poisson distribution with the same λ value as calculated in Question 3.1, which is based on the company's claim.

3.3.2 $E(Y)$ and $Var(Y)$

According to Poisson Distribution, PMF is:

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

$$E(Y) = \mu = \lambda$$

$$Var(Y) = \sigma^2 = \lambda$$

4 Question 4

4.1 Calculate $P(X = 4)$

Poisson probability formula is: $P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$

In this case, $k = 4$ and $\lambda = 0.4$:

$$P(X = 4) = \frac{e^{-0.4} * 0.4^4}{4!}$$

4.2 Calculate $P(X < 3)$

$$P(X < 3) = P(X = 0) + P(X = 1) + P(X = 2)$$

4.3 Calculate $P(Y = 4|Y \geq 3)$

The conditional probability of an event A given event B is defined as:

$$P(A|B) = \frac{P(AB)}{B}$$

The $P(Y = 4|Y \geq 3)$ equals to $\frac{P(Y=4 \cap Y \geq 3)}{P(Y \geq 3)}$, $P(Y = 4 \cap Y \geq 3) = P(Y = 4)$.

$P(Y \geq 3) = 1 - P(Y < 3)$

Introduction To Statistics For Data Science **Individual Assignment 1**
Dr. Ric Crossman

The submission deadline for this assignment is **12pm Monday 23rd October**. You will need to submit via Gradescope, via the Ultra page. I **strongly recommend** submitting at least a few hours ahead of the deadline, in case of technical issues.

Each of the four questions carries similar but not identical weight.

Question 1

A series of multiple choice questions.

B

1. A short e-survey is released, asking the following question: "How many cups of tea do you drink a day?". The available answers were "0", "1", "2", and "3 or more".

   What type of data is the e-survey collecting?

   (a) Nominal
   (b) Ordinal
   (c) Discrete
   (d) Continuous

A

2. Data is collected relating to the continent of origin of students at Durham. The data is to be shown graphically to the University Council, to help them understand where students come from most often, and least often.

   Council will want to compare numbers between continents - they are not interested in the proportion each continent of origin makes up of the whole of the student body. Which of the following graphs would be the best method for displaying this data?

   (a) Bar chart
   (b) Pie chart
   (c) Histogram
   (d) Stem and leaf diagram

Event A

B

3. In a recent survey, 46% of the British population said they preferred dogs to cats. In another recent survey, 12% of the British population listed jazz as one of their favourite musical genres.

Event B

Assume that whether a person prefers dogs to cats is independent of whether they consider jazz one of their favourite musical genres. What probability, expressed as a percentage, should we give to the event that a British person prefers dogs to cats **and** considers jazz one of their favourite musical genres.

(a) 58%
(b) 5.52%
(c) 55.2%
(d) 5.80%

$46\% * 12\%$

mutual independant

$P(A) \cdot P(B) = P(AB)$

4. Consider an outcome space $\Omega = \{1, 8, 15, 35, 69, 732, 983\}$. Let $A = \{1, 8, 69, 983\}$ and let $B = \{1, 8, 15, 35, 69\}$.

   Assuming a uniform probability distribution on Ω, which of these is the probability that A AND B occur?

   (a) $\frac{3}{7}$
   (b) $\frac{4}{7}$
   (c) $\frac{5}{7}$
   (d) $\frac{6}{7}$

$$P(A \cup B) = P(A) + P(A) - P(AB)$$
$$= \frac{4}{7} + \frac{5}{7} - \frac{6}{7}$$

D

5. A random variable U is defined with the probability density function below:

$$f(u) = \begin{cases} \frac{u}{2} & 0 \leq u \leq 2 \\ 0 & \text{otherwise} \end{cases}.$$

   Which of the following is the value of $P(U > 0.30)$?

   (a) 0.09775
   (b) 0.03
   (c) 0.70
   (d) 0.9775

$$F(u) = \begin{cases} \frac{u^2}{4} \Big|_{low}^{high} \\ 0, \text{ otherwise} \end{cases}$$

$$P(U > 0.30) = \frac{u^2}{4} \Big|_{0.30}^{2} = 1 - \frac{0.09}{4}$$
$$= 0.9775$$

Question 2

Using the `ToothGrowth` data set in `R`:

1. Find the modal tooth length across the entire data set.
2. Find the mean tooth length for guinea pigs who were given their vitamins via orange juice.

3. Create a diagram in `R` of two box-and-whisker plots, one showing tooth length for guinea pigs given their vitamins via asorbic acid, and one showing tooth length for guinea pigs given their vitamins via orange juice. The box-and-whisker plots should be adjacent to each other, using the same axis, to enable easy comparison.

4. Using the box-and-whisker plots created in Question 1.2.3, comment on which of the two vitamin delivery approaches is more effective in promoting tooth growth. Justify your answer.

Question 3

A company that produces party poppers claims on their website that only 0.6% of their party poppers will fail to go off when the string is pulled.

1. The company sells party poppers in boxes of 200. What is the expected number of party poppers which will fail to go off in a box?

2. I decide I want to represent the number of party poppers which fail to go off in a box using a random variable, X.

   (a) Which distribution would be the most appropriate to use for X? Give both the name of the distribution, and the value of the parameters assuming the company's claim on their website is correct.

   (b) What assumptions regarding the party poppers would need to hold in order to justify that choice of distribution?

3. I order a box for myself, but due to an administrative error, the box arrives containing not 200 party poppers, but 2.

   (a) Define the distribution for Y, the random variable representing the number of party poppers which fail to go off in my box of 2, assuming the claim on the company's website is correct.

   (b) Find $E(Y)$ and $\text{Var}(Y)$, under the assumptions needed for Question 1.3.2 b), and under the assumption that the claim on the company's website is correct. NOTE: You will need to show your working.

Part D

Let $X \sim Pois(0.4)$, and let $Y \sim Pois(\lambda)$. Find:

1. $P(X = 4)$.

2. $P(X < 3)$. NOTE: You will need to show your working.

3. An algebraic expression for $P(Y = 4|Y \geq 3)$. NOTE: You will need your working.