Introduction to Statistics for DS
Individual Assignment 1

Zehao Qian

October 22, 2023

1 Question 1

1. B

Analytics: The answer options "0," "1," "2," and "3 or more" represent an ordered range, but the gaps between them are not necessarily equal. In this case, the data is ordered, but cannot perform exact mathematical operations, and is therefore classified as ordinal data.

2. A

Analytics: Not interested in the proportion each continent of origin make up.

3. B

Analytics: Let Event A = British population said they preferred dogs to cats,
Event B = British population listed jazz as one of their favourite musical genres.
A and B are mutual independent $\Rightarrow P(AB) = P(A)P(B) = 46\% * 12\% = 5.52\%$

4. A

Analytics: $P(A \cup B) = P(A) + P(B) - P(AB) = \frac{4}{7} + \frac{5}{7} - \frac{6}{7} = \frac{3}{7}$

5. D

Analytics:

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

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

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_lengt = as.numeric(names(sort(
    table(ToothGrowth$len), decreasing = TRUE)[1]))
```

Env Output: $modal\_tooth\_length = 26.4$

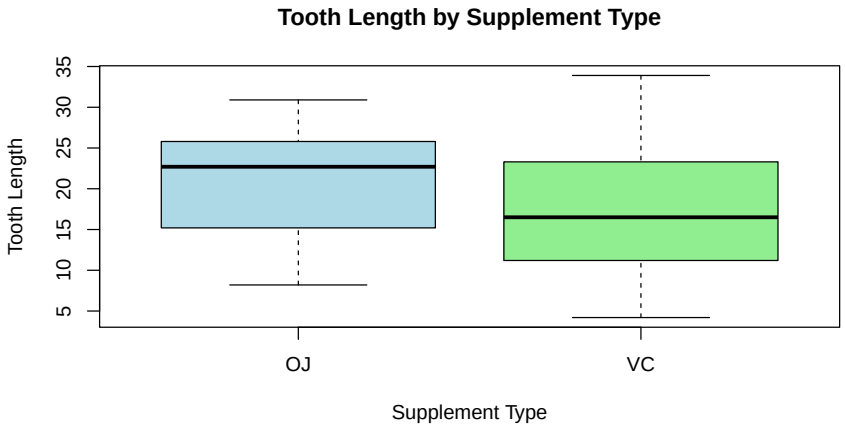
2.2 Mean tooth length for guinea pigs who were given their vitamins via orange juice

```
# 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
```

Env Output: $mean\_tooth\_length\_orange = 20.66333$

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")
```



2.4 Comment on which vitamin delivery approach is more effective:

Analytics

- OJ has a higher **Median Value** than VC.
- OJ data is basically normally distributed, while VC is skewed.
- The upper and lower **quartiles** of VC's span longer distances, so the data is more dispersed.
- **The length of the box:** I don't think there's much difference between the two, and the degree of centralization in the data set is basically the same.

Conclusion I think OJ supplement is 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 and Binomial Distribution

I don't know if the condition provided in the first question (200 in a box) holds true in the second question, so I think both the binomial distribution and the Poisson distribution can describe the probability of this problem.

Poisson Distribution: $X \sim Poisson(\lambda)$, $\lambda = 0.6\% * n$

Binomial Distribution: $X \sim B(n, p)$, $p = 0.006$

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

- Events (party poppers failing to go off) occur randomly (only for Poison).
- Events are rare, meaning the probability of more than one event occurring in a very short time period is negligible (only for Poison).
- 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), we can still use the Poisson or Binomial distribution with the same λ (Poison) $= n * p$ (Binomial) $= 0.012$ 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:

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

$$E(Y) = \mu = \lambda = 0.012$$

$$Var(Y) = \sigma^2 = \lambda = 0.012$$

According to Binomial Distribution, PMF is:

$$P(Y = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$E(Y) = n * p = 0.012$$

$$Var(Y) = n * p * (1-p) = 0.01193$$

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

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