This assessment is from: Tooba Zahid 23721219

Question # 1

In Lake Serenity, there are two types of fish: Bass and Trout. The probability of catching a **Bass is P(B) = 0.4**, and the probability of catching a **Trout is P(T) = 0.6**. Additionally, it is known that on sunny days, the probability of catching a Bass increase **to P(B|S) = 0.8,** The probability of **a sunny day is P(S) = 0.4** and the probability of **a cloudy day is P(C).**

a) Complete the contingency table. **Hint P(B | S) = P(B∩S) P(S)** [4]
b) Find the probability of the weather being Sunny OR catching a Bass. [2]
c) Find the probability of catching a Trout on a Cloudy day i.e. P(T | C). [2]
d) Are the events a Sunny day AND catching a Trout independent [4]

Solution:

we have given values,

- probability of catching a **Bass is P(B) = 0.4**
- probability of catching a **Trout is P(T) = 0.6**
- probability of **a sunny day is P(S) = 0.4**
- probability of **a cloudy day is P(C) = ?**
- and also
  on sunny days, the probability of catching a Bass is **P(B|S) = 0.8**

(a)

Complete the contingency table. **Hint P(B | S) = P(B∩S) / P(S)**

	P(S)	**P(C)**	**Total**
P(B)			
P(T)			
Total			**1**

So, by using hint **P(B | S) = P(B∩S) / P(S),** we calculate P(B∩S)

By rearranging the formula and then putting the values we are given,

We get,

$P(B \cap S) = P(B|S) * P(S)$ as $P(B|S)= 0.8$, $P(S)= 0.4$

$= 0.8 * 0.4$

$=0.32$

Hence $P(B \cap S)=0.32$

Now we calculate P(C) which is probability of Cloudy Day

$P(C) = 1- P(S)$

$= 1 – 0.4 => 0.6$

$P(C) = 0.6$

As, $P(B) = P(B \cap S) + P(B \cap C)$

So, from here we calculate $P(B \cap C)$ by rearranging and putting the values, we get,

$P(B \cap C) = P(B) - P(B \cap S)$ as $P(B) = 0.4$ and $P(B \cap S) = 0.32$

$= 0.4 – 0.32$

$= 0.08$

$P(B \cap C) = 0.08$

$P(T \cap S) = P(S) - P(B \cap S)$ as $P(S) = 0.4$ and $P(B \cap S) = 0.32$

$= 0.4 – 0.32$

$= 0.08$

$P(T \cap S)=0.08$

assuming that catching a Trout and a Bass are mutually exclusive events.

$P(T \cap C) = P(C) - P(T \cap S)$ as $P(C) = 0.6$ and $P(T \cap S) = 0.08$

$= 0.6 – 0.08$

$= 0.52$

$P(T \cap C) = 0.52$

By completing the contingency table, we get

	P(S)	**P(C)**	**Total**
P(B)	0.32	0.08	0.4
P(T)	0.08	0.52	0.6
Total	0.4	0.6	**1**

(b)

Find the probability of the weather being Sunny OR catching a Bass.

Solution:

We have to calculate

$P(S\cup B) = ?$

As $P(S\cup B) = P(S) + P(B) - P(B\cap S)$

By putting the values in it, we get

$= 0.4 + 0.4 – 0.32$

$P(S\cup B) = 0.48$

(c)

Find the probability of catching a Trout on a Cloudy day i.e. $P(T \mid C)$.

Solution:

We have to calculate

$P(T \mid C) = ?$

By using the formula,

$P(T \mid C) = P(T\cap C) / P(C)$

By putting the values $P(T\cap C) = 0.52$ and $P(C) = 0.6$

$P(T \mid C) = 0.52/0.6$

$P(T \mid C) = 0.8667$

(d)

Are the events a Sunny day AND catching a Trout independent?

Solution:

As we know that Two events A and B are independent if

$P(A\cap B) = P(A) * P(B)$

So, in this case

$P(T \cap S) = 0.08$ i.e. $P(T \cap S) = P(S \cap T)$

$P(T) * P(S) = 0.6 * 0.4$

$= 0.24$

Conclusion:

As $P(T \cap S) \neq P(T) * P(S)$ i.e. $0.08 \neq 0.24$

Hence , events are not independent.

Question # 2

A Poisson distribution is often used in Football to model the probability of a team's scoring. Manchester United's **average number of goals per match is 2.1**.

a) What is the probability of them scoring exactly 2 goals? [4]

b) What is the probability of them scoring between 1 and 4 inclusive?

c) Scoring no goals? [4]

d) Score 7 goals over 2 games? [4

Solution:

Let G be the number of goals per match , then $G \sim Pois(2.1)$.

(a)

What is the probability of them scoring exactly 2 goals?

We have $\mu = 2.1$

As $P(X=x) = P(X \leq x) - P(X \leq x - 1)$, So,

$P(G=2) = P(G \leq 2) - P(G \leq 1)$

By using table,

= 0.6496-0.3796

=0.27

(b)

What is the probability of them scoring between 1 and 4 inclusive?

We have **$\mu = 2.1$**

As, $P(a \le X \le b) = P(X \le b) - P(X \le a - 1)$ So,

$P(1 \le G \le 4) = P(G \le 4) - P(G \le 0)$

By using table,

= 0.9379 – 0.1225

=0.8154

(c)

Scoring no goals?

We have **$\mu = 2.1$**

Scoring n goals means G = 0

$P(G= 0) = P(G \le 0)$

By using table,

= 0.1225

(d)

Score 7 goals over 2 games?

The average number of goals over 2 games would be

⇨ $2 \times 2.1 = 4.2$

Now we have **$\mu = 4.2$**

For $G = 7$

As $P(X=x) = P(X \le x) - P(X \le x - 1)$, So

$P(G=7) = P(G \le 7) - P(G \le 6)$

By using table,

$= 0.9361 - 0.8675$

$= 0.0686$

Question # 3:

Researchers are investigating the distribution of resting heart rates in a population. The resting heart rates in the population follow a normal distribution with a **mean (μ) of 71 beats per minute** and a standard **deviation (σ) of 7 beats per minute**.

a) Calculate the probability that a randomly selected individual from this population has a resting heart rate of **less than 80 beats per minute**. [6]

b) If the top **10%** of individuals with the highest resting heart rates are considered at risk for heart disease, what is the threshold resting heart rate that would put people in the at-risk group? [5]

Solution:

We have mean and standard deviation.

$\mu = 71$ (beats per minute)

$\sigma = 7$ (beats per minute)

(a)

Calculate the probability that a randomly selected individual from this population has a resting heart rate of **less than 80 beats per minute**

Let B is the heart rate of less than 80 beats per minute so, For P(B < 80)

Where Z= (x – σ / μ)

By putting values we get,

Z = 80 – 71 / 7

⇨ 9/7 = 1.29

From table,

$P\ (Z < +z) = 1 - P(\ Z > +z)$

So P(Z<1.29) = 1 – P(Z > 1.29)

= 1 - 0.0985

⇨ 0.9015

So, the probability that a randomly selected individual has a resting heart rate of less than 80 BPM is about 0.9015.

(b)

If the top **10%** of individuals with the highest resting heart rates are considered at risk for heart disease, what is the threshold resting heart rate that would put people in the at-risk group?

We have to find the 90^{th} percentile in a normal distribution

As the threshold of the top is 10%

So, 100% – 10% = 90% => 0.9 so the value for Z_{90} = 1.2816 (using table for inverse normal distribution)

And the inverse transformation is, $\boldsymbol{X} = \boldsymbol{\sigma z} + \boldsymbol{\mu}$.

Where

$\boldsymbol{\mu} = 71$

$\boldsymbol{\sigma} = 7$

And Z = 1.2816

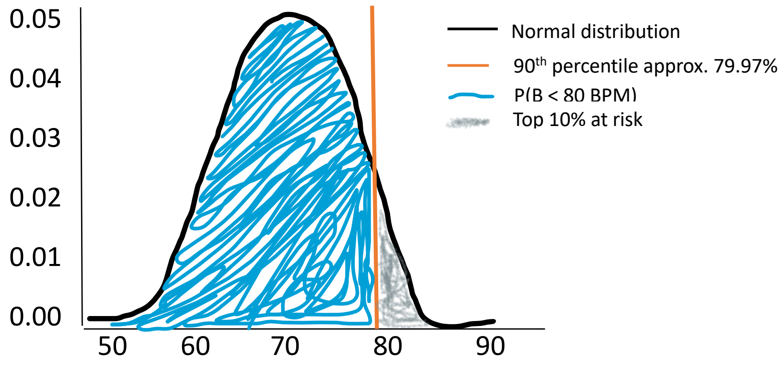
By putting values in formula

.

We get

X = 7*1.2816 + 71

X = 79.97 which is nearly equal to 80



Therefore, the threshold resting heart rate for the top 10% at-risk group is approximately 80 BPM.

Question # 4:

A Jonny Roadhouse music sells guitars. The guitars come in three different brands: **Fender, Gibson, and Ibanez**. Fender constitutes **44%** of the store's inventory, Gibson constitutes **33%,** and Ibanez constitutes **23%.** Jonny carries out a satisfaction survey. The results show that **86%** of customers 1 who purchase Fender guitars are satisfied **(P(S | F)), 85% for Gibson (P(S | G)), and 72% for Ibanez (P(S | I)).**

a) Using total probability, calculate the probability that a random customer is satisfied with their guitar. Use a diagram to fully explain your workings. [12]

b) If a customer is satisfied with their guitar, what is the probability they purchased a Fender (P(F | S)). Hint: use Bayes' Theorem. [3]

Solution:

We have,

Probability of purchasing a Fender => **P(F) =0.44**

Probability of being satisfied given that the guitar is a Fender => **P(S|F)= 0.86**

Probability of purchasing a Gibson => **P(G) =0.33**

Probability of being satisfied given that the guitar is a Gibson => **P(S|G)= 0.85**

Probability of purchasing an Ibanez => **P(I) =0.23**

Probability of being satisfied given that the guitar is an Ibanez => **P(S|I)= 0.72**

(a)

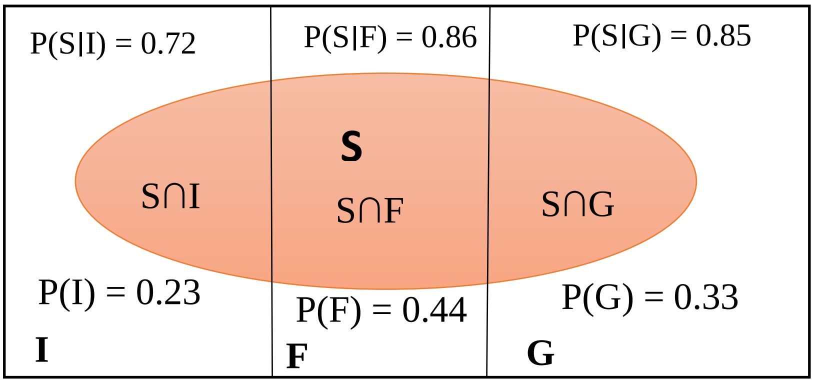
Using total probability, calculate the probability that a random customer is satisfied with their guitar. Use a diagram to fully explain your workings

We have to calculate the probability of a customer being satisfied with their guitar. So by using total probability

- ⇨ $P(S)=P(S|F)P(F)+P(S|G)P(G)+P(S|I)P(I)$
  By putting values in it, we get

  $P(S) = (0.86\times0.44) + (0.85\times0.33) + (0.72\times0.23)$
  $= 0.8245$

- ⇨ So, the probability that a random customer is satisfied with their guitar is 82.45%



(b)

If a customer is satisfied with their guitar, what is the probability they purchased a Fender (P(F | S)). Hint: use Bayes' Theorem.

We have to find P(F|S), the probability of a customer being satisfied, purchased a Fender by using Bayes' Theorem:

So, according to Bayes' Theorem

$$P(F|S)= P(S|F)*P(F) / P(S)$$

By putting values, we get

$$P(F|S)= (0.86)(0.44) / 0.8245$$

$$= 0.3784 / 0.8245$$

$$= 0.4589$$

So, probability that a customer purchased a Fender given they are satisfied is approximately 45.89%.

Question # 5:

A glass manufacturer wanted to evaluate the performance of protective coating for lenses. A random selection of lenses was given one of four different coatings, and then subjected to the same degree of simulated abrasion. After the abrasion impairment to light was tested. The higher the number, the worse more the lens was scratched. The results the results can be found in **lenses.csv**

Q. Based on the data, which coating would you recommend?

a) Describe your initial thoughts on the data.

1. Produce a table of descriptive statistics (include the mean and the standard deviation). [3]
2. Produce a well-presented boxplot.
3. Comment on your findings. Do not simply provide the raw values for the statistics, these are presented in parts 1 and 2 [10]

b) Confirm or contradict your initial findings.

1. Using an appropriate statistical test, explore the customer's question. Remember to declare all appropriate hypotheses for all the tests you use. Use a significance level of 0.05 throughout.[15]
2. Comment on the results of the statistical tests in the context of the customer's question. [6

(a)

Describe your initial thoughts on the data.

Produce a table of descriptive statistics (include the mean and the standard deviation)

1st of all I import the dataset as shown below,

```
# import dataset
lenses <- read_csv("lenses.csv")
# Rows: 28 Columns: 2
# ── Column specification ────────────────
# Delimiter: ","
# chr (1): Coating
# dbl (1): impairment

lenses$Coating = factor(lenses$Coating)
```

Calculate the descriptive statistics including mean and SD:

```
#descriptive statistics
library(mosaic)
favstats(impairment~ Coating,data=lenses)

#Result
# Coating min   Q1 median   Q3 max     mean        sd n missing
# 1       A 3.6 3.80    4.0 4.10 4.2 3.942857 0.2225395 7       0
# 2       B 3.9 4.05    4.2 4.25 4.4 4.157143 0.1718249 7       0
# 3       C 4.8 5.00    5.1 5.25 5.6 5.142857 0.2636737 7       0
# 4       D 3.2 3.30    3.4 3.45 3.5 3.371429 0.1112697 7       0


# Coatings A and B show moderate abrasion resistance,
# while Coating C is less effective.
# Coating D offers superior performance with minimal variability,
# making it recommended for its effectiveness.
```

Descriptive Statistics:

Average coating performance is shown by mean scores and standard deviation, with lower means indicating better abrasion resistance.

The descriptive statistics for each coating type are as follows:

Coating A: Mean impairment = 3.94, SD = 0.22

Coating B: 4.16 impairment, 0.17 SD

Coating C: 5.14 impairment, 0.26 SD

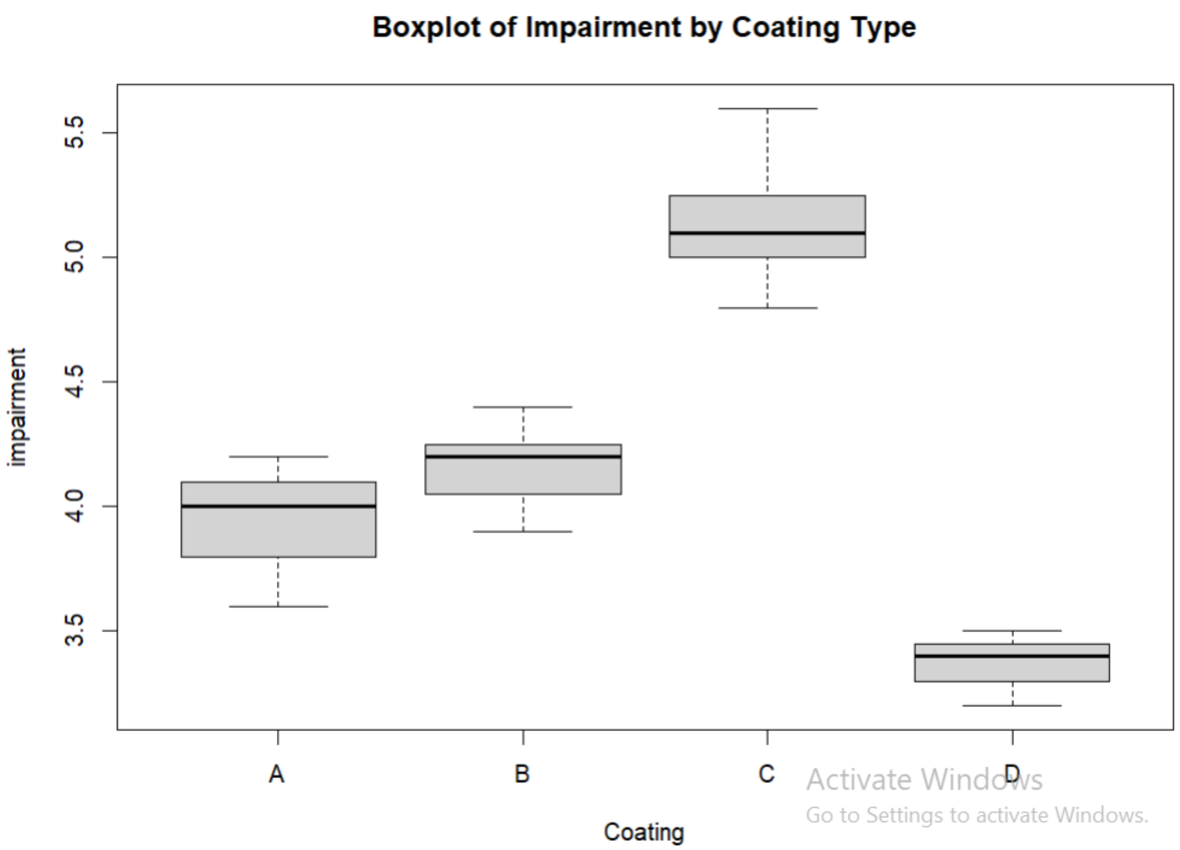
Coating D: 3.37 impairment, 0.11 SD

visualize the distribution of impairment scores for each coating!

```r
boxplot(impairment ~ Coating, data = lenses, main = "Boxplot of Impairment by Coating Type")

# In the boxplot, coating C has the highest median impairment
# and broadest interquartile range,
# coatings A and B have equal medians,
# and coating D performs best.

# Initial thoughts:
#   With the lowest mean impairment and narrow distribution,
# Coating D performs best.
# Coating C has the highest mean impairment and is least effective.
```



This boxplot shows each coating's impairment score dispersion. Lower median values and box sizes imply higher performance (less impairment).

Initial Thoughts:

- ⇨ The mean and median are used to assess a coating's abrasion resistance. Coating A and B show moderate abrasion resistance, while Coating C is less effective.
- ⇨ Coating D's lowest mean impairment score, standard deviation, and compact IQR make it the best pick.

(b)

Confirm or contradict your initial findings.

Using an appropriate statistical test, explore the customer’s question. Remember to declare all appropriate hypotheses for all the tests you use. Use a significance level of 0.05 throughout.

- ⇨ **Levene's Test** for Homogeneity of Variance:

```
library(car)
leveneTest(impairment ~ Coating, data = lenses)


# Levene's Test for Homogeneity of Variance (center = median)
#       Df F value Pr(>F)
# group  3  0.8451 0.4827
#       24
```

Result

p-value = 0.483

This high p-value indicates that the variances are homogeneous across the groups, which is an important assumption for ANOVA.

I assume a significance level of 5% (0.05)

- ⇨ **Anova test**

We do it to test if there are statistically significant differences in the mean impairment across different coatings. This is appropriate for comparing means across more than two groups.

```
#H0: There is no difference in the mean impairment among the different coatings.
# H1: There is a difference in the mean impairment among at least one pair of different coatings.
anova_results <- aov(impairment ~ Coating, data = lenses)
summary(anova_results)

# Df Sum Sq Mean Sq F value   Pr(>F)
# Coating      3 11.444   3.815    94.8 1.91e-13 ***
#   Residuals   24  0.966   0.040
# ---
#   Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
# >
```

Result:

⇨ p-value < 0.0001

This highly significant p-value suggests there are significant differences in mean impairment across the different coatings.the conclusion is to reject the null hypothesis.
At least one coating performs differently, as the p-value is much lower.

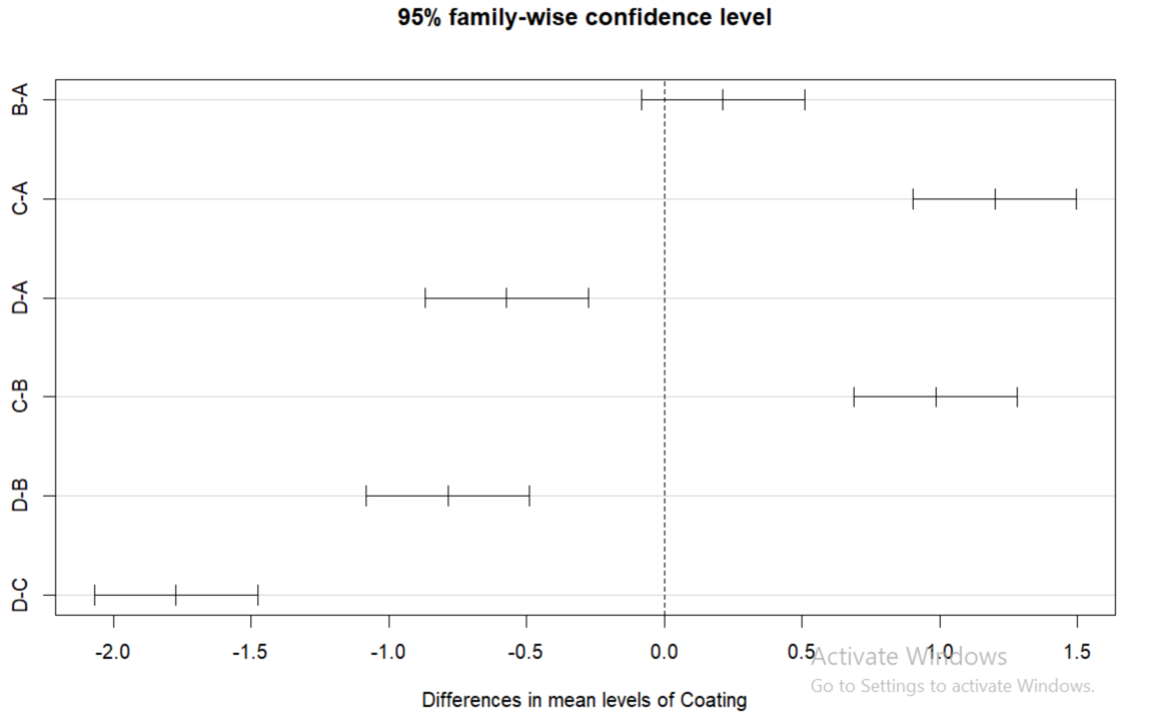
⇨ To determine which pairs are significantly different I will carry out a **Tukey HSD test**, assuming a 5% significance level

```
#to determine which coating are significantly different I will carry out
# a Tukey HSD test, assuming a 5% significance level.

TukeyHSD(model)

# Tukey multiple comparisons of means
# 95% family-wise confidence level
#
# Fit: aov(formula = impairment ~ Coating, data = lenses)
#
# $Coating
# diff          lwr         upr      p adj
# B-A  0.2142857 -0.08149837  0.5100698 0.2164491
# C-A  1.2000000  0.90421592  1.4957841 0.0000000
# D-A -0.5714286 -0.86721265 -0.2756445 0.0001007
# C-B  0.9857143  0.68993020  1.2814984 0.0000000
# D-B -0.7857143 -1.08149837 -0.4899302 0.0000008
# D-C -1.7714286 -2.06721265 -1.4756445 0.0000000

plot(TukeyHSD(model))
```



Result:

- ⇨ Coating C performs poorly compared to the others.
- ⇨ Performance is similar for coatings A and B.
- ⇨ Coating D outperforms other coatings in terms of lower impairment score, indicating its superior effectiveness in resisting abrasion

Comment on the results of the statistical tests in the context of the customer's question.

⇨ ANOVA shows substantial coating differences.
⇨ The Levene's test validates the ANOVA results.
⇨ A close examination at Tukey's HSD results would disclose which coatings differ significantly.

Based on the data, which coating would you recommend?

Conclusion:

As we expected,

    ⇨ Coatings A and B difference in mean impairment is not statistically significant ($p > 0.05$). This suggests that their performance is quite similar.
    ⇨ Coating C vs. Others (A, B, and D) has a significant difference in mean impairment when comparing Coating C with A, B, and D ($p < 0.05$). Coating C performs worse than the others.

⇨ Coating D significantly outperforms Coatings A and B ($p < 0.05$), indicating its superior effectiveness in resisting abrasion.
⇨ This analysis provides a clear recommendation based on statistical evidence, favouring Coating D for the best performance in terms of minimizing impairment due to abrasion.