A picture containing calendar

Description automatically generated

|  |
| --- |
| Statistics Assignment  2018 |
|  |
| 2021-22  Module: Statistics and Probability  Lecturer: Frank Leonard  Authored by: Jack Duggan |



Pet Food

An animal foods producer blends a pet food product and a data set is produced by taking samples from batches blended over a number of different runs. A measure of calorie content per 100g will have been taken for each sample as will a second measure from the same batch where this second sample will have been frozen for a period of 48 hours first. The primary ingredients in a blend include grain for which there are three suppliers A, B and C and meat for which there are four suppliers W, X, Y and Z. Additionally, a vitamin supplement might (or might not) be added to each blend. Some process parameters are recorded for each production run, in particular the temperature (oC) at which the blend is cooked and the total cooking time (minutes)

Contents

[Q1 3](#_Toc102745477)

[Q2 4](#_Toc102745478)

[Q3 5](#_Toc102745479)

[Q4 6](#_Toc102745480)

[Q5 7](#_Toc102745481)

[Q6 8](#_Toc102745482)

[Q7 10](#_Toc102745483)

[Q8 11](#_Toc102745484)

[R-Script 12](#_Toc102745485)

# Q1

Does cooking temperature tend to impact on calorie content?

Which plot is most suitable?

The scatter plot shown below graphs cooking temperature (horizontal axis) against calorie content (vertical axis). The plot clearly indicates a correlation between the two as it follows a straight-line trend. Cooking temperature is directly proportional to calorie content.

Chart, scatter chart

Description automatically generated

A Pearson’s correlation test is the most suitable test to measure the correlation between cooking temperature and calorie content. The test produced the output below, revealing a correlation value of 0.979793 which is statistically different from zero. In fact, it’s close to 1, meaning there’s a considerable correlation between the two. This is consistent with the plot.

Text

Description automatically generated

Thus, on the analysis of this data, we have significant evidence that that the cooking temperature is related to the calorie content.

# Q2

With respect to calorie content, does it matter which meat supplier is used?

The boxplot shown below portrays calorie content (vertical axis) across the 4 different meat suppliers (horizontal axis): W, X, Y, and Z. An initial glance at the plot suggests a minimal difference in calorie content between the different suppliers.

Chart, box and whisker chart

Description automatically generated

The table below shows the difference in a more readable manner. As you can see, there’s a slight increase in the mean calorie content for Supplier X in comparison to the other 3 suppliers. Supplier Z has the lowest standard deviation whereas supplier Y has the highest.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Mean | Standard Deviation | Sample Size |
| Supplier W | 101.1379 | 14.60608 | 29 |
| Supplier X | 107.2222 | 14.35359 | 28 |
| Supplier Y | 101.2727 | 16.73558 | 34 |
| Supplier Z | 100.5000 | 13.17999 | 34 |

For this instance, the ANOVA test is the most suitable.

The output from this test yielded a p value of 0.502. As the p value > 0.05, we can report that there is no statistically significant difference between the suppliers, and therefore it doesn’t make much difference which supplier is used.

# Q3

How, if all, does adding a vitamin supplement impact on calorie content?

The graph seen below is known as a conditional density plot. Such a plot provides a smoothed overview of how a categorical, binary variable (the presence of vitamin, in our case) changes across various levels of continuous, numerical variable (calorie content).

The presence of vitamin was recorded in the data set as either a 1 (vitamin present) or 2 (no vitamin present). If there was no impact on calorie content, the curved line would hover in and around the 0.5 mean mark (see the axis on the right). From the graph below you can see that the mean is slightly below the 0.5 mark, suggesting that there may be a slight impact on calorie content when the vitamin supplement is present. Further analysis of the graph substantiates this theory as the area of the light gray (presence of vitamin) is slightly larger than the area of the dark gray (no presence of vitamin).

Chart, histogram

Description automatically generated

While the graph suggested a possible impact on calorie content when the vitamin supplement was added that evidence alone is inconclusive. Thus, we’ll be performing a t-test on the data to validate the theory.

The t-test produced a p-value of 0.5493. P > 0.05 so we can report that is no statistically significant difference between the calorie content of samples with or without the vitamin supplement.

# Q4

Is cooking time and calorie content independent?

Below is a scatterplot with a lowess curve, demonstrating the possibility of any correlation between the cooking time (horizontal axis) and the calorie content (vertical axis).

From an initial inspection of the graph, I struggle to see any correlation between the two data sets. The data looks normally distributed around the 39-minute mark.

Chart

Description automatically generated

I performed a Pearson’s correlation test on the data to check if there was any relationship between the two data sets. Pearson’s test measures correlation from -1 to +1 with 0 meaning no correlation. The test gave a correlation value of -0.0237 which is very close to zero, meaning there is little to no correlation between the data. Therefore, we can consider the cooking time and calorie content to be independent of each other.

# Q5

Is there a greater tendency to add a vitamin supplement when grain from one supplier is used compared to others?

The clustered bar chart below shows decisively the different grain suppliers’ (A, B, C) tendencies to add vitamin or to not. The vertical axis shows the three different suppliers, and the horizontal axis shows whether the vitamin supplement was added.

We can clearly see from the graph below that suppliers A & C have a greater tendency to add the vitamin supplement, whereas supplier B tends not to.

A picture containing chart

Description automatically generated

I performed a chi-squared test on this data, resulting in a chi-squared value of 13.607 and a p-value of 0.00111. With a chi-squared test, if the p-value is less than 0.05, we reject the null hypothesis, which it is in our case. From this we can conclude that the two variables are dependent.

Upon analyzing the graph and data it’s quite safe to say that when suppliers A or C are used, they have a greater tendency to add the vitamin supplement than supplier B.

# Q6

Does the grain used in the blending process give rise to different calorie contents?

The boxplot shown below portrays calorie content (vertical axis) across the 3 different grain suppliers (horizontal axis): A, B, and C. An initial glance at the plot suggests a slight difference in calorie content between the different suppliers.

Chart, box and whisker chart

Description automatically generated

The table below shows the difference in a more readable manner. As you can see, supplier A has the lowest mean calorie content and supplier C has the highest.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Mean | Standard Deviation | Sample Size |
| Supplier A | 98.4600 | 15.35460 | 50 |
| Supplier B | 103.0976 | 13.10306 | 42 |
| Supplier C | 106.6333 | 14.57793 | 30 |

For this instance, the ANOVA test is the most suitable.

The output from this test yielded a p value of 0.0131. As the p value < 0.05, we can report that there is a statistically significant difference between the suppliers, and that the grain used in the blending process does in fact give rise to different calorie contents.

# Q7

Is the percentage of meat suppliers used to make up the blends about the same for all grain suppliers?

The bar chart shown below effectively visualizes the distribution of meat suppliers used to make up the blends for each grain supplier. As you can see from the graph, the percentage of meat suppliers is not similar for all grain suppliers. Grain supplier A is somewhat balanced across the four meat suppliers. Supplier B prefers meat suppliers W and Z over X and Y, and grain supplier C sparsely uses the products of meat supplier W.

Chart, bar chart

Description automatically generated

I performed a chi-squared test on this data, resulting in a chi-squared value of 7.905 and a p-value of 0.2451. With a chi-squared test, if the p-value is less than 0.05, we reject the null hypothesis. In our case however, the p-value is above this predetermined value.

To answer the question, the percentage of meat suppliers used to make up the blends is not similar for all grain suppliers. In fact, there’s a significant variation between the three.

# Q8

Is there any evidence that freezing the product effects the calorie content?

I used a boxplot (shown below) to visualize the difference in calorie contents (vertical axis) between frozen and unfrozen samples (horizontal axis). The box on the left represents the calorie contents of the batch that was frozen for a period of 48 hours beforehand, whereas the box the right represents the calorie contents of the batch that was left unfrozen.

It’s clear from the plot that the freezing period may well influence the calorie content, as the left box (frozen) sits lower on the calorie axis than the right box (unfrozen).

To verify this, we’ll perform a paired t-test on the data.

Chart, box and whisker chart

Description automatically generated

From the test, the test statistic is t = -4.9167 and the p-value is very small (p < 0..001) so the null hypothesis is rejected, since p < 0.05 and evidence of a statistically significant difference is found. Upon analyzing the data, we can conclude that freezing the product does in fact effect the calorie content.

# R-Script

Note: I did most of my workings from the RStudio console but compiled a script of the relevant commands as I went.

###########

# Statistics Assignment 2021/22

# Jack Duggan | 20094010

# Frank Leonard | Stats & Probability

###########

# Load the data from file

pet\_food = read.table("C:\\Users\\jackd\\OneDrive - Waterford Institute of Technology\\Statistics & Probability\\StatsAssign\\JackDuggan.csv", sep=",", header=TRUE) #reading from file

#Left a commented line so you can copy data from clipboard if required.

#pet\_food = read.table("clipboard", header=TRUE) #Reading from clipboard

# Attach the data

attach(pet\_food)

######

# Q1

######

plot(temperature, calorie) # Scatter Plot

cor.test(temperature, calorie, method="pearson") # Pearson's

######

# Q2

######

boxplot(calorie ~ meat)

tapply(calorie, meat, mean, na.rm=T)

tapply(calorie, meat, sd, na.rm=T)

tapply(calorie, meat, length)

q2\_anova = aov(calorie ~ meat)

summary(q2\_anova)

#####

# Q3

#####

vitamin\_presence <- factor(vitamin)

cdplot(vitamin\_presence ~ calorie)

t.test(calorie ~ vitamin)

#####

# Q4

#####

scatter.smooth(time, calorie)

cor.test(time, calorie, method="pearson") # Pearson's

#####

# Q5

#####

q5\_barchart\_table <- table(vitamin, grain)

row.names(q5\_barchart\_table) = c("Vitamin Added", "No Vitamin")

colnames(q5\_barchart\_table) = c("Supplier A", "Supplier B", "Supplier C")

plot(q5\_barchart\_table)

chisq.test(vitamin, grain, correct = F)

#####

# Q6

#####

boxplot(calorie ~ grain)

tapply(calorie, grain, mean, na.rm=T)

tapply(calorie, grain, sd, na.rm=T)

tapply(calorie, grain, length)

q6\_anova = aov(calorie ~ grain)

summary(q6\_anova)

#####

# Q7

#####

barplot(table(meat, grain), beside = TRUE, xlab = "1: Supplier A | 2: Supplier B | 3: Supplier C", main = "Meat Suppliers Distribution Per Grain Supplier", legend.text = c("Supplier W", "Supplier X", "Supplier Y", "Supplier Z"), col = c("red", "blue", "green", "yellow"))

chisq.test(meat, grain, correct = F)

######

# Q8

######

boxplot(frozen, calorie, main="Frozen/Calorie", paired=T)

t.test(frozen,calorie,paired = T)