Analysis of variance

Science, Statistics

Introduction to Biostatistics November 14, Statistics Homework 6 Analysis of Variance a) The null hypothesis is , H0: The mean tensile strength is equal at different levels of weight percent of cotton.

(b) The assumptions of the ANOVA method include;

Normality of the populations from which the samples has been drawn

Equality of variance for all the samples i. e. Homoscedasticity

(c) Boxplots of the observations

#Let the cotton weights assume the alphabets as follows, 15= A, 20= B, 25= C, 30= D and 35= E

> Strength Weights Strength

[1] 7 7 15 11 9 12 17 12 18 18 14 18 18 19 19 19 25 22 19 23 7 10 11 15 11

> Weights

[1] A A A A A B B B B B C C C C C D D D D D E E E E E

Levels: A B C D E

> Tablets boxplot(Strength~Weights, xlab=" Weights", ylab=" Tensile Strength", main=" Boxplot of Tensile Strength by Weight of Cotton")

> (d) By observing the box plots above, we observe that the normality assumptions have been violated. The medians (represented by the thick black line) are expected to be in the middle while the whiskers should have equal lengths, but this is not the case. Further, there are outliers in the weights C and E. However, this violation of normality could be due to the relatively small size of the data sets.

(e) A linear model on the data is as follows;

> MB summary(MB)

Call:

lm(formula = Strength ~ Weights)

Residuals:

Min 1Q Median 3Q Max

-3. 8 -2. 6 0. 4 1. 4 5. 2

Coefficients:

Estimate Std. Error t value Pr(>| t|)

(Intercept) 9. 800 1. 270 7. 719 2. 02e-07 ***

WeightsB 5. 600 1. 796 3. 119 0. 005409 **

WeightsC 7. 800 1. 796 4. 344 0. 000315 ***

WeightsD 11. 800 1. 796 6. 572 2. 11e-06 ***

WeightsE 1. 000 1. 796 0. 557 0. 583753

Signif. codes: 0 ‘***’ 0. 001 ‘**’ 0. 01 ‘*’ 0. 05 ‘.’ 0. 1 ‘ ’ 1

Residual standard error: 2. 839 on 20 degrees of freedom

Multiple R-squared: 0. 7469, Adjusted R-squared: 0. 6963

F-statistic: 14. 76 on 4 and 20 DF, p-value: 9. 128e-06

Interpretation of the Output

The linear model is of the form;

Strength = 5. 6*WeightsB + 7. 8*WeightsC + 11. 8*WeightsD + 1.

0*WeightsE

The strength of the model, R-squared = 74. 69%

(f) The ANOVA procedure is as follows;

> AoV AoV

Analysis of Variance Table

Response: Strength

Df Sum Sq Mean Sq F value Pr(> F)

Weights 4 475. 76 118. 94 14. 757 9. 128e-06 ***

Residuals 20 161. 20 8. 06

Signif. codes: 0 ‘***’ 0. 001 ‘**’ 0. 01 ‘*’ 0. 05 ‘.’ 0. 1 ‘ ’ 1

> Interpretation of the Output

Criteria: We reject H0 if p-value is very small

In this case, p = 9. 128e-06 ≈ 0, which is very small.

Conclusion: The results do not sustain our earlier assertion that the mean tensile strength is equal at different levels of weight percent of cotton. Therefore, H0 is very, very strongly rejected. This implies significant differences in the mean tensile strength of the fiber at different weight percent of cotton.

(g) Residual plots

#We produce the 4 plots on a single frame

> par(mfrow= c(2, 2))

> plot(aov(Strength~Weights))

> Interpretation of the plots

The residual vs. fitted plot checks for the homoscedacity assumption. In this case, the residuals seem to be evenly distributed on both sides of the fitted red line, hence we validate the assumption.

The normal Q-Q plot checks for the normality assumption. In this context, the normality assumption is not validated. It is not tenable to assume normality of the samples.

Proportions

2. (a) The null hypothesis is, H0: The difference between the number of deaths in those wearing seat belts and those not wearing seat belts is not significant i. e. (p1 - p2)= 0.

(b) Let p1= Probability that the child died in the accident while wearing a seat belt and, p2= Probability that the child died in the accident while not wearing a seat belt.

The point estimate of p1= 3/123 while that of p2= 13/290

The difference p1-p2 = -0. 020437342

We wish to test whether the difference between the proportions is significant. We proceed as follows;

> prop. test(c(3, 13), c(123, 290), alternative=" two. sided")

2-sample test for equality of proportions with continuity correction

data: c(3, 13) out of c(123, 290)

X-squared = 0. 4976, df = 1, p-value = 0. 4805

alternative hypothesis: two. sided

95 percent confidence interval:

-0. 06242524 0. 02155056

sample estimates:

prop 1 prop 2

0. 02439024 0. 04482759

Warning message:

In prop. test(c(3, 13), c(123, 290), alternative = " two. sided") :

Chi-squared approximation may be incorrect

> Conclusion: Since the 95% confidence interval contains 0, the null hypothesis is not rejected. Moreover, the p-value = 0. 4805, which is greater

than 0. 05. Therefore, the difference between the proportions of children who die while wearing seatbelts and those who die while not wearing seatbelts is not significant.

(c) A 95% confidence interval for the difference in proportions involved in (b) is obtained as follows;

The confidence interval is given as;

, where and

Using R;

> n1 n2 p1 p2 z s1 s2 lci uci lci

[1] -0. 05663673

> uci

[1] 0. 01576205

> The 95% confidence interval is (-0. 05663673, 0. 01576205)

Similarly (as in (b)), the 95% confidence interval contains 0; hence the null hypothesis is not rejected.

Correlation

3. (a) Scatter plot of the data

#First, we input the data into R

> Nation PAttended MaternalMortality cordata cordata

Nation PAttended MaternalMortality

1 Bangladesh 5 600

2 Belgium 100 3

3 Chile 98 67

4 Ecuador 84 170

5 Hong Kong 100 6

6 Hungary 99 15

7 Iran 70 120

8 Kenya 50 170

9 Morocco 26 300

10 Nepal 6 830

11 Netherlands 100 10

12 Nigeria 37 800

13 Pakistan 35 500

14 Panama 96 60

15 Philippines 55 100

16 Portugal 90 10

17 Spain 96 5

18 Switzerland 99 5

19 United States 99 8

20 Vietnam 95 120

> plot(cordata$PAttended, cordata$MaternalMortality, pch= 16, col=" red", xlab=" Percent Attended", ylab=" Maternal Mortality (per 100, 000 live births)", main=" Scatter Plot")

From the scatter plot, it is evident that there is a linear relationship between maternal mortality and the percentage number of patients attended. It can be observed that as maternal mortality decreases as the percentage of patients attended increase.

(b) The sample correlation coefficient is obtained as follows;

> cor(cordata$PAttended, cordata$MaternalMortality)

[1] -0. 8768071

The correlation coefficient is highly negative indicating a strong negative linear relationship between the percentage of patients attended and maternal mortality.

(c) We wish to test that H0: ρ = 0. We proceed as follows using R.

> cor. test(cordata$PAttended, cordata$MaternalMortality)

Pearsons product-moment correlation

data: cordata$PAttended and cordata$MaternalMortality

t = -7. 7364, df = 18, p-value = 3. 936e-07

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0. 9505207 -0. 7096243

sample estimates:

cor

-0. 8768071

> Conclusion: The p-value= 3. 936e-07 which is very negligible, hence we reject the null hypothesis. Moreover, the 95% confidence interval is -0. 9505207≤ ρ ≤ -0. 7096243, which does not contain 0. Therefore, the population correlation coefficient is not equal to zero.