**Q1: Correlation (One-tailed test)**

You are given the dateset AgeWeight.csv which maps age to weight of a sample. Run this code.

**AgeWeight <- read.csv("AgeWeight.csv")**

**View(AgeWeight)**

**summary(AgeWeight)**

**plot(AgeWeight)**

**cor(AgeWeight)**

**cor.test(AgeWeight$Wt, AgeWeight$Age, alternative = ‘greater’)**

**(This code helps you test for correlation between age and weight)**

**You are required to modify the code and run it to answer this question:**

**Using the sample data, determine if weight increases with age for the population.**

**State your null and alternate hypotheses:**H0: there is no correlation between age and weight (weight is not increasing with age)

H1: there is a positive correlation between age and weight (correlation >0)

**Include the results of the correlation test  
  
Pearson's product-moment correlation**

data: AgeWeight$Wt and AgeWeight$Age

t = 5.4979, df = 15, p-value = 3.065e-05

alternative hypothesis: true correlation is greater than 0

95 percent confidence interval:

0.6104785 1.0000000

sample estimates:

cor 0.8175185

**State your conclusions based on the test results**

We can state with the results of the test that the p value is very low and we can reject the null hypothesis that correlation is less than or equal to zero.

The correlation value of 0.8175 is high and positive and thus statistically significant.

**Q2: ChiSquare Test with Single Categorical Variable:**

(Here is the code to run)

**AfterExercise <- c(255,125,90)  
BeforeProbability=c(.60,.25,.15)  
chisq.test(AfterExercise , p= BeforeProbability)**  
**Explain the result here:**

Chi-squared test for given probabilities

data: AfterExercise

X-squared = 8.4574, df = 2, p-value = 0.01457

H0=there is no difference between expected F(expected) and observed F(observed) distributions

H1=Ha: there is a difference between observed and expected frequencies

To test out whether there was a statistically significant change in the number of smokers after a talk by Amy, we conducted a chi-sqr test. The p value is low enough to say that there is a statistically significant change between observed and expected frequencies.

**Q3: ChiSquare Test for Independence with Two Categorical Variables:**

**#Create a matrix in R as follows:**

**cancer <- matrix(data = c(33, 250, 196, 136, 32, 55, 293, 190, 71, 13), nrow = 2, ncol = 5, byrow = T)**

**# Give names to the rows and columns in the matrix**

**dimnames(cancer) <- list(c("Lung cancer", "Control"), c("1 to 4", "5 to 14", "15 to 24", "25 to 49", "50+"))**

**#Display the matrix ‘cancer’**

**cancer**

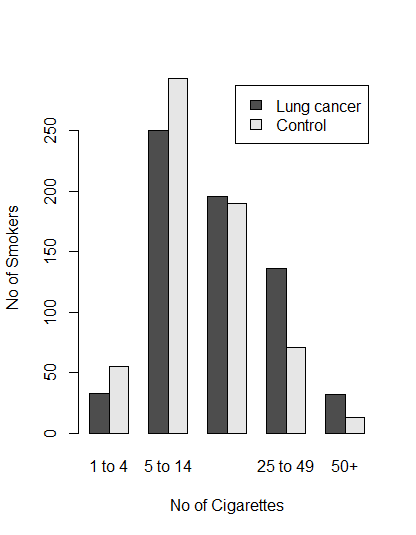
**barplot(cancer, beside=T, legend=TRUE, xlab="No of Cigarettes", ylab="No of Smokers") #(beside=T to plot the two rows beside each other)**

**#What does this show? Can you make any conclusions from observing these two distributions?**

**chisq.test(cancer)**

**Interpret the result (See class slides)**

1. **State your null and alternate hypotheses:**

H0: There is no association between smoking and cancer.

H1: There is a correlation between smoking and cancer.

1. **Include the results of the chisq test here**

Pearson's Chi-squared test

data: cancer

X-squared = 36.9531, df = 4, p-value = 1.842e-07

1. **State your conclusions based on the test results**

Since the P-value is very low, we reject the null hypothesis that there is no association between smoking and cancer in smokers.

**Problems for Practice (May appear in a quiz some time after Spring Break):**1. Given this dataset about the frequency distribution of sick and healthy people of different age groups in a small city, test the hypothesis that some age groups are particularly prone to disease (or in other words disease is independent of the age group)?

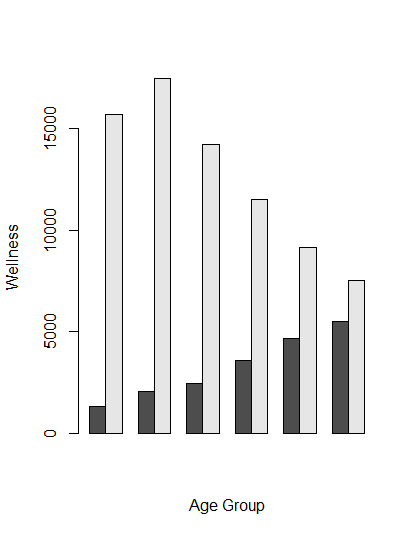
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 34-39 | 40-44 | 45-49 | 50-54 | 55-59 | 60-64 |
| Sick | 1327 | 2072 | 2456 | 3611 | 4688 | 5490 |
| Healthy | 15702 | 17454 | 14237 | 11519 | 9174 | 7526 |

**wellness <- matrix(data = c(1327, 2072, 2456, 3611, 4688, 5490, 15702, 17454, 14237, 11519, 9174, 7526), nrow = 2, ncol = 6, byrow = T)**

**dimnames(wellness) <- list(c("Sick", "Healthy"), c("34 to 39", "40 to 44", "45 to 49", "50 to 54", "55 to 59", “60 to 64”))**

wellness

**barplot(wellness, beside=T, legend=TRUE, xlab="Age Group", ylab="Wellness")**



**chisq.test(wellness)**

H0: There is no correlation between age groups and illness

H1: There is a correlation between age groups and illness

Results: Pearson's Chi-squared test

data: wellness

X-squared = 8530.795, df = 5, p-value < 2.2e-16

Conclusions:

Since P value is very low (<0.05), we reject the null hypothesis that there is no correlation between age group and illness.

2.The National Center for Health Statistics (NCHS) provides data on the distribution of weight among Americans every year.

Americans in 2002 were distributed as follows: 2% Underweight, 39% Normal Weight, 36% Overweight, and 23% Obese. Suppose we want to assess whether the distribution of BMI is different in the following sample of the population (Courtesy: Framingham Study and Boston University Medical Center for this data).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Underweight**  **BMI<18.5** | **Normal Weight**  **BMI 18.5-24.9** | **Overweight**  **BMI 25.0-29.9** | **Obese**  **BMI > 30** | **Total** |
| **# of Participants** | 20 | 932 | 1374 | 1000 | 3326 |

3. **Are lung cancer patients more likely to be smokers?**

|  |  |  |
| --- | --- | --- |
|  | **Non Smokers** | **Smokers** |
| **Cancer Patients** | 2 | 647 |
| **Other Patients** | 27 | 622 |

**Chi-square Test is not good for this since the tables have low values in one or more cells**

**>fisher.test(smokers)**

#Use Fisher's Exact Test for Count Data