Rohit Wason
Spring 2021

Math 560 Homework (#9, Inference of Two-Way Tables)

Problem 1. A table of two-variables is given:

Coffee	Male	Female	Total
Always	18	15	33
Sometimes	36	36	72
Never	36	9	45
Total	90	60	150

Solution. The expected cell counts are as follows:

Coffee	Male	Female
Always	$\frac{33\times 90}{150} = 19.8$	$\frac{33\times 60}{150} = 13.2$
Sometimes	$\frac{72\times 90}{150} = 43.2$	$\frac{72\times 60}{150} = 28.8$
Never	$\frac{45\times 90}{150} = 27$	$\frac{45\times 60}{150} = 18$

Hypothesis: H_0: There is no association between the row & the column variables vs. H_a: There is an association between them.

The test statistic, χ^2

$$= \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$
$$= \frac{(18-19.8)^2}{19.8} + \frac{(15-13.2)^2}{13.2} + \frac{(36-43.2)^2}{43.2} + \frac{(36-28.8)^2}{28.8} + \frac{(36-27)^2}{27} + \frac{(9-18)^2}{18}$$
$$= 0.1636364 + 0.2454545 + 1.2 + 1.8 + 3 + 4.5$$
$$= 10.90909$$

Looking up the χ^2 *distribution critical values* table, we see that the **Critical value** at $\alpha = 0.01$ and with **Degrees of freedom**, $df = (r-1)(c-1) = 2$ comes out to be, $\chi^{2*} = 4.605$.

Conclusion: Since $\chi^2 > \chi^{2*}$ we **reject** the hypothesis H_0, that there is no association between the coffee consumption & gender of the students. □

Problem 2. Given

Outcome	1	2	3	4	5	6
# of occurances	153	184	160	175	162	166

Solution. We are to perform a Goodness of Fit test at $\alpha = 0.05$ that the die is fair.

Hypothesis: H_0: The outcome from this die follows a fair distribution (where each outcome is $\frac{1}{6}$ likely) vs. H_a: The die is not fair.

The **expected counts** for a fair die, when rolled 1000 times would be ≈ 166.67

The test statistic, χ^2

$$= \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$
$$= \frac{1}{166.67}[(153 - 166.67)^2 + (184 - 166.67)^2 + (160 - 166.67)^2$$
$$+ (175 - 166.67)^2 + (162 - 166.67)^2 + (166 - 166.67)^2]$$
$$\approx 3.739926$$

Looking up the χ^2 *distribution critical values* table, we see that the **Critical value** at $\alpha = 0.05$ and with **Degrees of freedom**, $df = (k - 1) = 5$ comes out to be, $\chi^{2*} = 11.070$.

Conclusion: Since $\chi^2 < \chi^{2*}$ we **fail to reject** the hypothesis H_0, that this is a fair die. □

Problem 3. Given:
$n = 500$
Mean height of the fathers, $\bar{x} = 67.9$
Standard deviation of the height of fathers, $s_x = 2.75$
Mean height of the sons, $\bar{y} = 68.7$
Standard deviation of the height of sons, $s_y = 2.83$

Solution. (a) The correlation between the heights of the sons and their fathers in the sample was $r = 0.5$. The equation of the least-squares regression line for predicting the height of a son (y) from this population based on his father's height (x) is given by

$$\hat{y} = b_0 + b_1 x$$

where $b_1 = r(s_y/s_x) = 0.5(\frac{2.83}{2.75}) \approx 0.5145$
and $b_0 = \bar{y} - b_1\bar{x} = 68.7 - 0.5145(67.9) \approx 33.7654$

Therefore the equation of the least squares regression line is

$$\hat{y} = 33.7654 + 0.5145x \tag{3.1}$$

(b) Suppose instead that the correlation was 1. In this case the equation of the least-squares regression line for predicting the height of a son from this population based on his father's height will use $b_1 = s_y/s_x = \frac{2.83}{2.75} \approx 1.0291$
and $b_0 = \bar{y} - b_1\bar{x} = 68.7 - 1.0291(67.9) \approx -1.1759$

Therefore the equation of the least squares regression line will be

$$\hat{y} = -1.1759 + 1.0291x \tag{3.2}$$

(c) If, instead, the correlation was 0. $b_1 = 0(s_y/s_x) = 0$
and $b_0 = \bar{y} - b_1\bar{x} = 68.7 - 0 = 68.7$

This time, the equation of the least squares regression line will be

$$\hat{y} = 68.7 \tag{3.3}$$

(d) Predicting the height of a son from this population whose father is 6 feet 6 inches tall ($x = 78$) based on the regression line obtained in (3.1) will be

$$\hat{y} = 33.7654 + 0.5145(78) = 73.8964$$

Or approximately 6 feet 2 inches.

(e) If r was equal to 1, using (3.2),

$$\hat{y}(x = 78) = -1.1759 + 1.0291(78) \approx 79.0939$$

Or approximately 6 feet 7 inches.

(f) And finally, if r was equal to 0, using (3.3),

$$\hat{y}(x = 78) = 68.7$$

Or approximately 5 feet 9 inches. □

Project details

Introduction

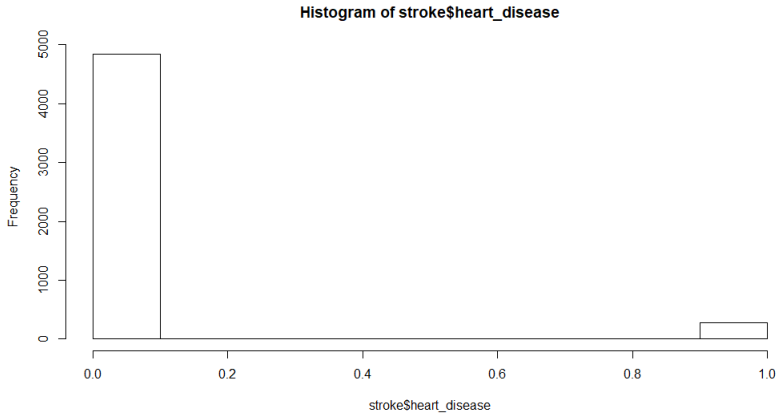
According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. This dataset is used to predict whether a patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking status. We can use this dataset to state hypotheses of correlation between various explanatory variables (like, age, bmi, marital status, residence type, etc.) and the result that they have had a stroke.
The dataset used here can be found on Kaggle, a public competetion forum where data scientists and novices collaborate to find answers in complex datasets (kaggle.com/fedesoriano 2021). This dataset contains 5110 data points with 12 variables:

1. **id:** unique identifier
2. **gender:** "Male", "Female" or "Other"
3. **age:** age of the patient
4. **hypertension:** 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
5. **heart_disease:** 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
6. **ever_married:** "No" or "Yes"
7. **work_type:** "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
8. **Residence_type:** "Rural" or "Urban"
9. **avg_glucose_level:** average glucose level in blood
10. **bmi:** body mass index
11. **smoking_status:** "formerly smoked", "never smoked", "smokes" or "Unknown"*
12. **stroke:** 1 if the patient had a stroke or 0 if not

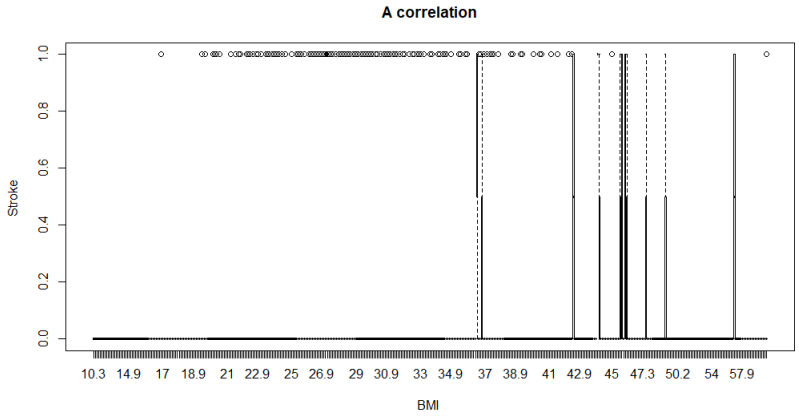
Methods

According to stroke.org, a non-profit organization dedicated to public awareness about, and prevention of common causes of stroke, smoking, lack of physical activity, diabetes and obesity are leading factors that could lead to stroke (stroke.org 2021).
In particular the following questions are of interest:

1. **Is there a strong correlation between stroke and other conditions, like heart disease, or hypertension?**
   A histogram of whether heart-disease is present in a data point:



2. **Does BMI indicate the occurence of stroke?**



Works Cited

kaggle.com/fedesoriano. "Stroke Prediction Dataset". 2021. Web. 31/03/2021. <https://www.kaggle.com/fedesoriano/stroke-prediction-dataset?select=healthcare-dataset-stroke-data.csv>.

stroke.org. "Stroke Risk Factors You Can Control". 2021. Web. 31/03/2021. <https://www.stroke.org/en/about-stroke/stroke-risk-factors/stroke-risk-factors-you-can-control-treat-and-improve>.