TU Dortmund
Department of Statistics

Case Studies I

Project 3: Contingency tables

Aspirin, the miracle drug?

Lecturers:
Prof. Dr. Paul Bürkner
Dr. Daniel Habermann
Lars Kühmichel

Author: Rahul Ramesh Vishwkarma
Matriculation Number.: 236862

Group Number.: 17

Group members: Md Jubirul Alam, Rafay Maqsood,
Syed Hassan Saqlain Tayyab

May 29, 2024

Contents

1 Introduction

Aspirin, also known as acetylsalicylic acid, is a common medication known for relieving pain, reducing fever, lowering inflammation, and preventing blood clots. It is used for headaches, muscle pain, and arthritis, and apart from this, it is also used to prevent heart attacks and strokes in those at risk. Because of these benefits, aspirin is widely used and trusted around the world. However, aspirin should be used carefully because it can cause side effects like stomach bleeding and can interact with other medications.

First, the distribution of the variable that corresponds to a heart attack is compared among the different groups formed by categorical variables using descriptive statistical tools. Then, a 2 x 2 contingency table is formed to analyse the joint distribution of two binary variables, `HeartAttack` and `Group`. Using the contingency table, the two measures of association, namely Risk Ratio and Odds Ratio are calculated to determine the strength of the association among these binary variables. Secondly, the Chi-squared test is conducted to test the independence of the variables `HeartAttack` and `Group`.

The second section describes the dataset and explains the meaning of variables. In the third section, important statistical methods for the analysis are discussed. In the fourth section, descriptive and inferential statistical tools are used to determine the strength of the dependency of two variables, and also to test the independency between the variables `HeartAttack` and `Group`.

2 Problem Statement

2.1 Dataset and its quality

The given dataset is from an unpublished experimental study held in 1993 and compiled as `Aspirin.csv` in a `.csv` file by the instructors of ICS at TU Dortmund. The dataset contains 20021 observations in total and has no missing values.

In this experiment, the effect of aspirin is investigated on heart attacks. All the participants were male and given randomly one of the two tablets, i.e., aspirin or placebo, for every two days over a period of five years. To make observations less likely to be

biased, the study was held in a double-blind manner, in which neither the participants nor the experimenters knew which tablet was being taken by the participants until the experiment was over.

2.2 Explanation of variables

There are four categorical variables (or factors) `Group`, `HeartAttack`, `Smoking` and `Age` with different categories (or levels). The description of these variables with their levels, are as follows,

- `Group` (with 2 levels) - which of the tablets was taken by the participant? : "Aspirin", "Placebo"
- `HeartAttack` (with 2 levels) - did the participant have a heart attack? : "Yes", "No"
- `Smoking` (with 2 levels) - is the participant a smoker? : "Smoker", "Non-Smoker"
- `Age` (with 6 levels) - the age of the participants in years : "61", "62", "63", "64", "65"

2.3 Goal of this project

The project has three goals. The first is to analyse if the distribution of the number of heart attacks is different in different groups. The second goal requires a two-way contingency table to analyse a bivariate distribution and to calculate two measures of association from it to know the strength of association between them. Finally, Chi-squared is used to justify the dependency between two categorical variables.

3 Statistical Methods

3.1 Two-way contingency table

Suppose we have two categorical variables A and B, where A and B have $A_1, A_2, .., A_r$ and $B_1, B_2, .., B_s$ categories, respectively. Then, the joint distribution of A and B follows a multinomial distribution with probabilities p_{ij} for $i = 1, 2, ..., r$ and $j = 1, 2, ..., s$. The probability density for a single observation $x = (x_{11}, x_{12}, ..., x_{1s}, x_{2,s}, ..., x_{rs})$ is (Mood (2012)),

$$f(x) = \prod_{i=1}^{r} \prod_{j=1}^{s} p_{ij}^{x_{ij}}, \tag{1}$$

where, $x_{ij} = 0$ or 1, with $\sum_{i=1}^{r} \sum_{j=1}^{s} x_{ij} = 1$.

Suppose, we have a sample of size n from a multinomial distributed population having a bivariate variable (A, B), and n_{ij} is the entry of (i, j)th cells (i.e., the table entry in the i^{th} row and j^{th} column) corresponds to the number of observations belonging to both categories, A_i and B_j, then $\sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij} = n$, and we have a r x s contingency table as follows,

Table 1: Two-way contingency table representing joint distribution of A and B

	B_1	B_2	$\cdots$	B_s
A_1	n_{11}	n_{12}	$\cdots$	n_{1s}
A_2	n_{21}	n_{22}	$\cdots$	n_{2s}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
A_r	n_{r1}	n_{r2}	$\cdots$	n_{rs}

Now, let's denote the row sums and column sums by $n_{i\cdot}$ and $n_{\cdot j}$, and are calculated as $n_{i\cdot} = \sum_{j=1}^{s} n_{ij}$ and $n_{\cdot j} = \sum_{i=1}^{r} n_{ij}$, respectively, and it also holds that, $\sum_{j=1}^{s} n_{\cdot j} = \sum_{i=1}^{r} n_{i\cdot} = n$.

Relative Frequency

The relative frequency f_{ij} is the proportion of the observations observed in a sample of size n of categories A_i and B_j, and it is defined as (Mood (2012)),

$$f_{ij} = \frac{n_{ij}}{n} \quad ; \text{where } i = 1, 2, ..., r \text{ and } j = 1, 2, ..., s. \tag{2}$$

The relative frequency f_{ij} is the estimator of the probability p_{ij} (i.e., $f_{ij} = \hat{p_{ij}}$), and if the sample size is large, this f_{ij} approximates p_{ij}.

Note that, when the cells $(i, j)^{'s}$ of the table contain the probabilities $p_{ij}^{'s}$, it gives the joint distribution of the variables A and B.

3.2 Measure of association

For simplicity, we consider a 2 x 2 contingency table (i.e., $r = 2$, $s = 2$), then consider A as a response with outcomes, A_1 as an event and A_2 as a non-event, and B as an exposure variable with B_1 as an exposed group and B_2 as an unexposed group. The risk ratio and odds ratio summarise the association of an event in the exposed group with the unexposed group with a single number.

3.2.1 Risk Ratio

The risk R is the probability of an event in a group and is given by a conditional probability as follows (Mood (2012)),

$$R = P(A|B) = \frac{P(A, B)}{P(B)} \quad ...\text{(by Conditonal probability)} \tag{3}$$

$$= \frac{P(A, B)}{P(A, B) + P(A, B)} \quad ...\text{(by Law of total probablity).} \tag{4}$$

Then, the risk ratio RR is defined as the ratio of the risk in the exposed group to the risk in the unexposed group and given by (Agresti (2012)),

$$RR = \frac{P(A_1, B_1)/P(B_1)}{P(A_1, B_2)/P(B_2)} = \frac{\frac{P(A_1,B_1)}{P(A_1,B_1)+P(A_2,B_1)}}{\frac{P(A_1,B_2)}{P(A_1,B_2)+P(A_2,B_2)}} \quad \text{...(by Law of total probablity)} \tag{5}$$

$$= \frac{p_{11}/(p_{11} + p_{21})}{p_{12}/(p_{12} + p_{22})} \tag{6}$$

Note that an estimator of RR can be derived from relative frequencies as follows,

$$\widehat{RR} = \frac{f_{11}/(f_{11} + f_{21})}{f_{12}/(f_{12} + f_{22})} = \frac{n_{11}/(n_{11} + n_{21})}{n_{12}/(n_{12} + n_{22})}. \quad \text{...(since, } n \cdot f_{ij} = n_{ij}) \tag{7}$$

The sampling distribution of $\ln(\widehat{RR})$ follows a normal distribution with a large size, i.e.,

$$\ln(\widehat{RR}) \sim \mathcal{N}(\ln(RR), \sigma^2_{RR}), \tag{8}$$

and estimator for standard error of $\ln(\widehat{RR})$ is calculated as,

$$\hat{\sigma}_{RR} = \sqrt{\frac{n_{21}}{n_{11}(n_{11} + n_{21})} + \frac{n_{22}}{n_{12}(n_{12} + n_{22})}}. \tag{9}$$

The large-sample normality of $\ln(\widehat{RR})$ implies,

$$\tilde{L}_{RR} = \ln(\widehat{RR}) - Z_{\alpha/2} \cdot \hat{\sigma}_{RR} \quad \text{and} \quad \tilde{U}_{RR} = \ln(\widehat{RR}) + Z_{\alpha/2} \cdot \hat{\sigma}_{RR}, \tag{10}$$

where $Z_{\alpha/2}$ is the $\frac{\alpha}{2}^{th}$ quantile of the standard normal distribution. The C.I. for RR is obtained by exponentiating $\tilde{L}_{RR}$ and $\tilde{U}_{RR}$, and hence, (1-α)% C.I.$(OR) = (e^{\tilde{L}_{RR}}, e^{\tilde{U}_{RR}})$.

3.2.2 Odds Ratio

The odds of an event are the ratio of the probability of an event to the probability of a complement of the event in a group, given by the equation below (Mood (2012)),

$$O = \frac{P(A|B)}{P(A^c|B)} = \frac{\frac{P(A,B)}{P(B)}}{\frac{P(A^c,B)}{P(B)}} \quad \text{...(by coniditional probablity)} \tag{11}$$

$$= \frac{P(A,B)}{P(A^c,B)}. \tag{12}$$

Now, the odds ratio OR is the ratio of the odds of the event in the exposed group to the odds in the unexposed group, as follows (Agresti (2012)),

$$OR = \frac{P(A_1,B_1)/P(A_2,B_1)}{P(A_1,B_2)/P(A_2,B_2)} \quad \text{...(since, } A_1^c = A_2) \tag{13}$$

$$= \frac{p_{11}/p_{21}}{p_{12}/p_{22}} \tag{14}$$

Also, the estimator of OR can be expressed in terms of frequencies, and it is given as,

$$\widehat{OR} = \frac{f_{11}/f_{21}}{f_{12}/f_{22}} = \frac{n_{11}/n_{21}}{n_{12}/n_{22}} \quad \text{...(since, } n \cdot f_{ij} = n_{ij}) \tag{15}$$

The sampling distribution of the log of $\widehat{OR}$ follows a normal distribution, i.e.,

$$\ln(\widehat{OR}) \sim \mathcal{N}(\ln(OR), \sigma^2_{OR}), \tag{16}$$

where estimator for standard error of $\ln(\widehat{OR})$ is,

$$\hat{\sigma}_{OR} = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}. \tag{17}$$

The normality of $\ln(\widehat{OR})$ holds due to the large sample size, and then the (1-α)% confi-

dence interval (C.I.) can be constructed for $\ln(OR)$ as follows,

$$\tilde{L}_{OR} = \ln(\widehat{OR}) - Z_{\alpha/2} \cdot \hat{\sigma}_{OR} \quad \text{and} \quad \tilde{U}_{OR} = \ln(\widehat{OR}) + Z_{\alpha/2} \cdot \hat{\sigma}_{OR} \tag{18}$$

Where $Z_{\alpha/2}$ is the $\frac{\alpha}{2}^{th}$ quantile of the standard normal distribution. The C.I. for OR is obtained by exponentiating $\tilde{L}_{OR}$ and $\tilde{U}_{OR}$, and hence, (1-α)% C.I.$(OR) = (e^{\tilde{L}_{OR}}, e^{\tilde{U}_{OR}})$.

3.3 Chi-squared test

First, note that two discrete variables are independent if and only if their joint pmf (probability mass function) is equal to the product of their marginal pmf's (Ross (2022)). Now suppose we have a multinomial distribution of A and B with pdf f in an r x s contingency table with probabilities p_{ij} as defined in the earlier section 3.1. The statistical hypothesis testing problem for independence of A and B is defined as (Agresti (2012)),

$$H_0 : p_{ij} = p_{i\cdot} \cdot p_{\cdot j} \quad \text{and} \quad H_1 : p_{ij} \neq p_{i\cdot} \cdot p_{\cdot j} \quad ; \text{for } i = 1, 2, .., r \text{ and } j = 1, 2, .., s \tag{19}$$

Suppose μ_{ij} is the mean of variables A and B for the categories A_i and B_j in n trials, then $\mu_{ij} = n{\cdot}p_{ij}$. Then, under H_0, $\mu_{ij} = n{\cdot}p_{i\cdot}{\cdot}p_{\cdot j}$ and estimated μ_{ij}, i.e., $\hat{\mu}_{ij} = (n_{i\cdot}{\cdot}n_{\cdot j})/n$ (since $n \cdot \hat{p}_{i\cdot} = n_{i\cdot}$ and $n \cdot \hat{p}_{\cdot j} = n_{\cdot j}$). Now the Pearson statistic X^2 follows the χ^2-distribution with $(r-1)(s-1)$ degrees of freedom under H_0 and is defined as,

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} \sim \chi^2(r-1)(s-1) \tag{20}$$

The decision rule at α (significance level) can be defined with respect to the value of the test statistic X^2 as,

$$\text{Reject } H_0 : \text{if } X^2 > \chi^2_{(1-\alpha)}, \tag{21}$$

where $\chi^2_{(1-\alpha)}$ is the $(1-\alpha)^{th}$-quantile of χ^2-distribution with $(r-1)(s-1)$ degrees of freedom. Similarly, the decision rule can be defined in terms of p-value $= P(X^2 > \chi^2_{(1-\alpha)})$ as, Reject H_0 : if p-value $< \alpha$

3.4 Software

For our analysis, we used the statistical programming language R (2023-03-15 ucrt)(R Development Core Team (2020)). Apart from the base R package, several other packages are also used, such as *epitools*(Tomas J. Aragon (2020)) for handling contingency tables, *dplyr*(Wickham *et al.* (2021)) for data manipulation, and *ggplot*(Hadley Wickham ORCID (2020)) for visualization.

4 Statistical analysis

4.1 Distribution of heart attack in different groups

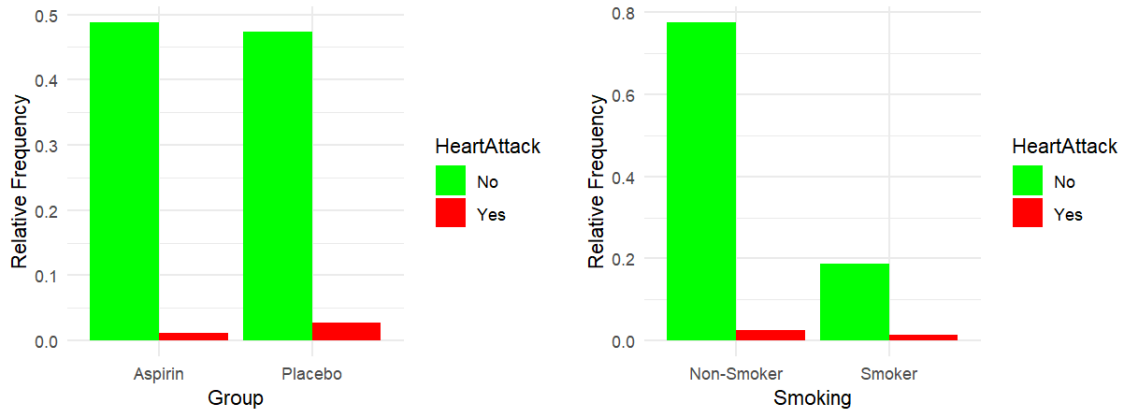
Distribution of heart attacks due to aspirin intake

From the Figure 1 (a), the relative frequency distribution of the variable `HearAttack` is changed with respect to the groups (or levels), "Aspirin" and "Placebo" of the variable `Group`. Moreover, about only 2.32 % participants have heart attack in the group of "Aspirin" whereas, about 5.57% cases of heart attack is observed among the participants in "Placebo" group. This implies that the rate of heart attack is almost half within the "Aspirin" group compared to "Placebo" group.

more than twice in placebo compared to "Aspirin".

Distribution of heart attacks for smokers and non-smokers

Similarly, as can be seen from Figure 1 (b), the distribution of `HeartAttack` in the variable `Smoking` is different, and around 7.87% of the participants in the "Smoker"



(a) `HeartAttacks` in `Group` (b) `HeartAttacks` in `Smoking`

Figure 1: Comparison of heart attacks in different groups

group have a heart attack, while the number of heart attacks in the group of "Non-Smoker" is just 3.29%. It suggests that the participants who smokes are more likely to have a heart attack than the ones who do not.

Distribution of heart attacks in different age groups

In contrast to the above two scenarios, the distribution of `HearAttack` seems to be quite similar among the different age groups, i.e., with different values of `Age` (see Figure 2). Hence, it can be said that the chance of having a heart attack is not influenced by participants' age.

4.2 Risk of a heart attack with aspirin

According to Table 2, the association between the variables `HeartAttack` and `Group` in terms of risk ratio suggests that the risk of having a heart attack is approximately 41.69% of the risk in the placebo group.Also, we have $[0.345, 0.471]$ being an 95% interval estimate for the RR which means the true RR lies in this interval with a probability of 95%.Hence, it can be concluded that aspirin reduces the risk of having a heart attack by around 58%. And from the 95% C.I., aspirin reduces the risk of having a heart attack by between around 51.5% and 64.% with a probability of 95%.

Table 2: Risk ratio with 95% C.I.

`Group`	estimate	lower	upper
Placebo	1.0000000	NA	NA
Aspirin	0.41698	0.358704	0.4847237

4.3 Odds of a heart attack with aspirin

The odds ratio of 0.40311 implies that the odds of the occurrence of a heart attack are less, and almost 40% of the odds of a heart attack in the placebo group (see Table 3). Further, the 95% confidence interval for the OR (Odds Ratio) is, $[0.345, 0.471]$ which says that the true OR value is lies in this interval with probability of 95%. Hence, this concludes that the odds of heart attack is reduced by almost 60% due to intake of aspirin compared to the participants who just had a placebo treatment. Also, with 95% probability, aspirin reduces the odds by nearly from 52.9% to 65%.

Table 3: Odds ratio with 95% C.I.

`Group`	estimate	lower	upper
Placebo	1.0000000	NA	NA
Aspirin	0.4031142	0.3450062	0.4710091

4.4 Does aspirin prevent heart attacks significantly?

To test the independence between the two categorical variables, we defined the testing problem as below,

$$H_0 : p(\texttt{Group}, \texttt{HeartAttack}) = p(\texttt{Group}) \cdot p(\texttt{HeartAttack})$$
$$\text{vs}$$
$$H_1 : p(\texttt{Group}, \texttt{HeartAttack}) \neq p(\texttt{Group}) \cdot p(\texttt{HeartAttack}), \tag{22}$$

where the null hypothesis H_0 states that `Group` and `HeartAttack` are independent, whereas H_1 states that they are not. The p-value, from Table 5, is less than $2.2e-16$.

Table 4: result of χ^2-test for independence

Pearson's Chi-squared test		
X-squared = 139.15,	df = 1,	p-value < 2.2e-16

It implies rejecting H_0 and suggests that there is dependency between the two variables `Group` and `HeartAttack`.

Hence, it infers that there is a strong dependency exists between regular consumption of aspirin and having a heart attack. Moreover, Risk Ratio and Odds Ratio explain that aspirin helps to prevent heart attack significantly.

5 Summary

This project aims to understand the effect of aspirin on heart attack prevention by analyzing a 2 x 2 contingency table using association measures and a statistical test for independence.

The analysis begins by comparing the frequency distribution of heart attacks across different groups categorized by the variables `Group`, `Smoking`, and `Age`. The data shows that the incidence of heart attacks differs between the "Aspirin" and "Placebo" groups, with the "Aspirin" group having nearly half the rate of heart attacks compared to the "Placebo" group. Furthermore, smokers have almost double the rate of heart attacks compared to non-smokers. In contrast, the rate of heart attacks remains fairly consistent across different age groups, indicating that age does not significantly influence heart attack rates.

Next, two association measures, the Risk Ratio (RR) and Odds Ratio (OR), are calculated to quantify the strength of the association between aspirin intake and heart attacks. The RR of approximately 41.69% suggests that regular aspirin use reduces the risk of heart attacks by nearly 60% compared to non-users. The OR of 0.403 indicates that the odds of a heart attack for those taking aspirin are about 40% of the odds for those in the placebo group.

Finally, a chi-squared test is performed to assess the independence between aspirin consumption and heart attacks. The null hypothesis (H_0) is states that there is no significant dependency between the two variables. The test results in a p-value of less than 2.2e-16, providing strong evidence to reject the null hypothesis. Consequently, it is concluded that regular aspirin intake significantly reduces the chance of heart attacks.

Bibliography

Agresti, Alan. 2012. *Categorical Data Analysis*. John Wiley Sons Inc;.

Hadley Wickham ORCID, Winston Chang. 2020. *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. R package version 3.5.1.

Mood, Alexander McFarlane. 2012. *Introduction to the Theory of Statistics*. McGraw-Hill Education Ltd.

R Development Core Team. 2020. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ross, Kevin. 2022. *An Introduction to Probability and Simulation*.

Tomas J. Aragon, Michael P. Fay, Daniel Wollschlaeger Adam Omidpanah. 2020. *epitools: Epidemiology Tools*. R package version 0.5-10.1.

Wickham, Hadley, François, Romain, Henry, Lionel, & Müller, Kirill. 2021. *dplyr: A Grammar of Data Manipulation*. R package version 1.0.6.

Appendix

A Additional Tables and Figures

Table 5: 2 x 2 Contingency table with row sums and column sums

	No (HeartAttack)	Yes (HeartAttack)	Sum
Non-Smoker	15495	497	15992
Smoker	3735	294	4029
Sum	19230	791	20021

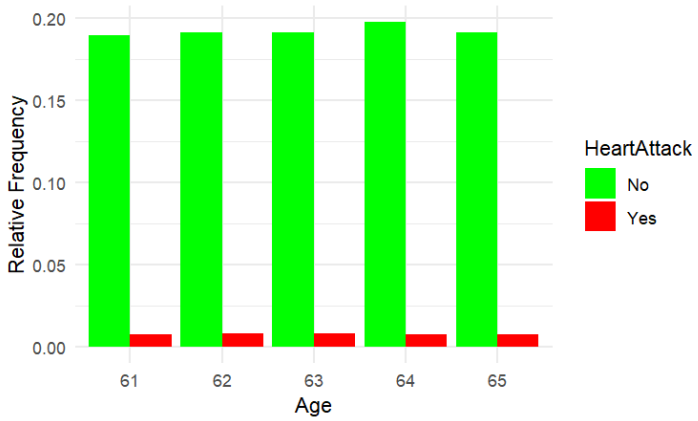


Figure 2: Comparing heart attacks in different age groups