Lab 12

Categorical Data Analysis

Kristen Fitzgerald

2024-07-15

Question 1

Data were collected in an effort to relate the safety of certain vehicles to different aspects of those vehicles. This dataset has the following variables:

- `Unsafe`: binary safety designation (1 = below average (unsafe), 0 = average or above average (safe))
- `Type`: type of car (Large, Medium, Small, Sport/Utility, Sports)
- `Region`: manufacturing region (Asia, N America)
- `Weight`: integer value for car weight ranging from 1 to 6
- `Size`: size of car corresponding to Type (1 = Small/Sports, 2 = Medium, 3 = Large or Sport/Utility)

Part (a)

Which variables are continuous, nominal, ordinal?

Solution:

- `Unsafe`: Ordinal
- `Type`: Nominal
- `Region`: Nominal
- `Weight`: Ordinal
- `Size`: Ordinal

Part (b)

Examine the association between Region and Unsafe.

a. What percentage of cars manufactured in Asia were classified as unsafe?

Solution: 42.9% of cars manufactured in Asia were classified as unsafe.

```
CrossTable(safety$Region, safety$Unsafe)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  96
##
##
##                | safety$Unsafe
## safety$Region |         0 |         1 | Row Total |
## --------------|-----------|-----------|-----------|
##          Asia |        20 |        15 |        35 |
##               |     0.686 |     1.509 |           |
##               |     0.571 |     0.429 |     0.365 |
##               |     0.303 |     0.500 |           |
##               |     0.208 |     0.156 |           |
## --------------|-----------|-----------|-----------|
##     N America |        46 |        15 |        61 |
##               |     0.394 |     0.866 |           |
##               |     0.754 |     0.246 |     0.635 |
##               |     0.697 |     0.500 |           |
##               |     0.479 |     0.156 |           |
## --------------|-----------|-----------|-----------|
##  Column Total |        66 |        30 |        96 |
##               |     0.688 |     0.312 |           |
## --------------|-----------|-----------|-----------|
##
##
```

b. What percentage of cars classified as safe were manufactured in North America?

Solution: 75.4% of cars manufactured in North America were classified as unsafe.

c. What is the appropriate test to use?

- State the null and alternative hypothesis for the test.
- At an alpha of 0.05, what is your decision?

Solution: We will use Fisher's Exact test since we are dealing with one ordinal variable and one nominal variable and we fail the assumptions necessary for Pearson's and the Likelihood Ratio Chi-Square Tests. The null hypothesis is that there is no association. The distribution of one variable does not change across levels of another. The alternative hypothesis is that there is an association. The distribution of one variable changes across levels of another. We fail to reject the null hypothesis, there is not statistically significant evidence to conclude that the distribution of `Region` changes across the levels of `Unsafe`.

```
fisher.test(table(safety$Region, safety$Unsafe))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(safety$Region, safety$Unsafe)
## p-value = 0.07175
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.1634104 1.1635862
## sample estimates:
## odds ratio
##  0.4387958
```

d. Regardless of significance, interpret the odds ratio in the context of the problem.

Solution: Cars made in Asia have 0.43 times the likelihood (odds) of being classified as safe as compared to cars made in North America.

```
OddsRatio(table(safety$Region, safety$Unsafe))
```

```
## [1] 0.4347826
```

Part (c)

Examine the association between `Size` and `Unsafe`.

a. What is the appropriate test to use for association?

- At an alpha of 0.05, what is your decision?

Solution: We will use the Mantel-Haenszel Chi-Square Test because both variables are ordinal. We reject the null hypothesis in favor of the alternative. There is statistically significant evidence to conclude that the distribution of `Unsafe` changes across different levels of `Size`.

```
CMHtest(table(safety$Size, safety$Unsafe))$table[1,]
```

```
##        Chisq           Df         Prob
## 2.770978e+01 1.000000e+00 1.409484e-07
```

```
CrossTable(safety$Size, safety$Unsafe)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  96
##
##
##              | safety$Unsafe
##  safety$Size |         0 |         1 | Row Total |
## -------------|-----------|-----------|-----------|
##            1 |        12 |        23 |        35 |
##              |     6.047 |    13.303 |           |
##              |     0.343 |     0.657 |     0.365 |
##              |     0.182 |     0.767 |           |
##              |     0.125 |     0.240 |           |
## -------------|-----------|-----------|-----------|
##            2 |        24 |         5 |        29 |
##              |     0.828 |     1.821 |           |
##              |     0.828 |     0.172 |     0.302 |
##              |     0.364 |     0.167 |           |
##              |     0.250 |     0.052 |           |
## -------------|-----------|-----------|-----------|
##            3 |        30 |         2 |        32 |
##              |     2.909 |     6.400 |           |
##              |     0.938 |     0.062 |     0.333 |
##              |     0.455 |     0.067 |           |
```

```
##              |     0.312 |     0.021 |           |
## -------------|-----------|-----------|-----------|
## Column Total |        66 |        30 |        96 |
##              |     0.688 |     0.312 |           |
## -------------|-----------|-----------|-----------|
##
##
```

b. How strong is the association between these variables?

Solution: Using Spearman's rank test we find a measure of association of $\rho = -0.54$.

```
cor.test(x = as.numeric(ordered(safety$Size)),
         y = as.numeric(ordered(safety$Unsafe)),
         method = "spearman")
```

```
## Warning in cor.test.default(x = as.numeric(ordered(safety$Size)), y =
## as.numeric(ordered(safety$Unsafe)), : Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  as.numeric(ordered(safety$Size)) and as.numeric(ordered(safety$Unsafe))
## S = 227423, p-value = 1.136e-08
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.5424769
```