# Assignment_1

Ted Ladas - s2124289

1/30/2021

## Question 1

### a)

**Test for independence** in a 2-way contingency table. This is because we want to test for independence between the two variables $X :=$ Smoker, $Y :=$ Lung Cancer, where $X, Y$ are boolean variables.

## b)

Formulation of the problem. If smoking and lung cancer **are** independent, then $P_{rc} = P((X,Y) = (r,c)) = P(X = r)P(Y = c) \quad r = 1,2$ and $c = 1,2$, representing the rows and the columns of the given table respectively. Or in words, the joint probability is the product of the marginals. So our hypothesis test is the following

$$\begin{aligned} H_0 : P_{rc} = P_r P_c \;\; & r = 1,2 \text{ and } c = 1,2 \\ \text{s.t. } \sum_{r=1}^{2}\sum_{c=1}^{2} P_{rc} = 1 & \\ H_0 : P_{rc} \neq P_r P_c \;\; & r = 1,2 \text{ and } c = 1,2 \\ \text{s.t. } \sum_{r=1}^{2}\sum_{c=1}^{2} P_{rc} = 1 & \end{aligned}$$

$\hat{P_{rc}} = \frac{O_{r.}O_{.c}}{n^2}$

where $n$ is the total number of observations and

$E_{rc} = n\hat{P_{rc}}$

so we now need to calculate $E_{11}, E_{12}, E_{21}, E_{22}$

$E_{11} = \frac{O_{1.}O_{.1}}{n} = \frac{1338 \times 709}{1418} = 669$

$E_{12} = \frac{O_{1.}O_{.2}}{n} = \frac{1338 \times 709}{1418} = 669$

$E_{21} = \frac{O_{2.}O_{.1}}{n} = \frac{80 \times 709}{1418} = 40$

$E_{22} = \frac{O_{2.}O_{.2}}{n} = \frac{80 \times 709}{1418} = 40$

so finally:

$$E = \begin{pmatrix} 669 & 669 \\ 40 & 40 \end{pmatrix}$$

## c)

We know that the distribution we need to use is the chi-squared distribution. Firstly, we need to calculate the the degrees of freedom. df $= (R-1)(C-1) = (2-1)(2-1) = 1$. Therefore, In order to correctly calculate the test statistic, we need to apply Yates' correction to the chi-squared statistic.

$X_1^2 = \sum_{r=1}^{2}\sum_{c=1}^{2} \frac{(|O_{rc}-E_{rc}|-0.5)^2}{E_{rc}}$

We know the $E$ matrix from question b) and the $O$ matrix from the exercise question, which is

$$O = \begin{pmatrix} 688 & 650 \\ 21 & 59 \end{pmatrix}$$

Therefore,

$$\begin{aligned} X_1^2 = & \frac{(|O_{11}-E_{11}|-0.5)^2}{E_{11}} + \\ & \frac{(|O_{12}-E_{12}|-0.5)^2}{E_{12}} + \\ & \frac{(|O_{21}-E_{21}|-0.5)^2}{E_{21}} + \\ & \frac{(|O_{22}-E_{22}|-0.5)^2}{E_{22}} = 18.1357 \end{aligned}$$

From tables we know that the critical value for $a = 0.05$ is: $X^2_{critical} = 3.841$, which is lower than our $X_1^2 = 18.1357$. Therefore, we conclude that we reject the Null hypothesis $H_0$, on the 5% significance level.

### d)

```r
rm(list = ls())
report_results <- function(diff_check = diff, p_value = p_auto,
    p_value_manual = p_manual, sig_lvl = alpha) {
    if (diff_check == 0) {
        cat("Reject the null hypothesis under the 5% significance level" %s*%
            (p_value < sig_lvl))
        cat("\nNo evidence to reject the null hypothesis under the 5% significance level" %s*%
            (p_value > sig_lvl))
        cat("\n     with p-value:", p_value)
    } else {
        warning("\nWARNING: manual and automatic p-value are not the same")
        cat("\ndiff between p-value calculations:", diff_check)
        cat("\np-value manual calc:", p_value_manual)
        cat("\np-value auto   calc:", p_value)
    }
}


# setting up the table and helper variables.
O <- as.table(rbind(c(688, 650), c(21, 59)))
dimnames(O) <- list(smokers = c("Smokers", "Nonsmokers"), lung_cancer = c("Cancer",
    "Control"))
alpha <- 0.05

# Manual Calculation of p-value.
n <- sum(O)
row_n <- rowSums(O)
col_n <- colSums(O)
E <- outer(row_n, col_n)/n  # expected cell counts.
X2 <- sum((abs(O - E) - 0.5)^2/E)  # Yates' continuity correction.
df <- (nrow(O) - 1) * (ncol(O) - 1)  # df=1 because we have 2x2 matrix.
p_manual <- 1 - pchisq(X2, df = df)


# using the chisq.test method.
p_auto <- chisq.test(O)$p.value
diff = round(p_manual - p_auto, 8)

# report final results.
report_results(diff_check = diff, p_value = p_auto, p_value_manual = p_manual,
    sig_lvl = alpha)
```

```
## Reject the null hypothesis under the 5% significance level
##      with p-value: 2.057117e-05
```

### e)

We have seen that from our test, we end up rejecting $H_0$ under the 5% significance level. That means that our variables of interest, smoking and Lung Cancer are **not** independent to one another. Therefore, we can claim that there is a relationship between $X$ and $Y$. However, from the test alone, we **cannot determine the directionality of the causation**, whether $X$ causes $Y$ or $Y$ causes $X$. We cannot safely conclude that smoking causes cancer from the data alone. If we take into consideration, all the past research on the topic, on the chemicals inside the cigarettes, and the fact that for most patients, smoking come before the fact that they develop cancerous cells, and other factors, we can conclude that smoking causes cancer.