Problems

Problem 1

A researcher wants to determine if there is an association between smoking status and the presence of lung disease. The following table summarizes the data collected from a sample of 200 participants:

Smoking Status	Lung Disease (Yes)	Lung Disease (No)	Total
Smoker	50	30	80
Non Smoker	20	100	120
Total	70	130	200

a. If we want to determine if there is a significant association between smoking status and the presence of lung disease, what test should be used?

b. Carry out the test at $\alpha = 0.05$. Report the test-statistic and interpret the results.

c. Let X represent the number of smokers who develop lung disease in the sample above. Now suppose we randomly select 5 individuals. Find the expected number and variance of smokers with lung disease from the 5 selected.

d. Suppose the following summary statistics is obtained of smokers and non-smokers from the sample

Smoking Status	Sample Mean Age	Standard Deviation of Age
Smoker	40	10
Non Smoker	45	12

Suppose the standard deviation obtained above is representative of their respective populations. Determine if there is a significant difference in mean ages between smokers and non-smokers.

(*Hint: the test statistics for a two-sample z-test is* $Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$)

e. Calculate the relative risk (RR) and odds ratio (OR) for smokers developing lung disease compared to non-smokers.

Problem 2

A study seeks to examine whether gender is associated with being a smoker. The researchers survey a population in the current month. Out of 500 total individuals, they find that there are 200 smokers. Out of all the male participants, 150 were smokers, while 75 were not.

Find a point estimate and a 95% confidence interval for this point estimate comparing the difference in risk of smoking between male and females. What can we conclude from these results?

Solution

Problem 1

A researcher wants to determine if there is an association between smoking status and the presence of lung disease. The following table summarizes the data collected from a sample of 200 participants:

Smoking Status	Lung Disease (Yes)	Lung Disease (No)	Total
Smoker	50	30	80
Non Smoker	20	100	120
Total	70	130	200

a. If we want to determine if there is a significant association between smoking status and the presence of lung disease, what test should be used?

Solution: Chi-Square Test of Independence.

b. Carry out the test at $\alpha = 0.05$. Report the test-statistic and interpret the results.

Solution:
H_0: There is not an association between smoking status and the presence of lung disease.
H_1: There is an association between smoking status and the presence of lung disease.

First, we calculate the expected table:

Smoking Status	Lung Disease (Yes)	Lung Disease (No)
Smoker	$E_{11} = \frac{80 \cdot 70}{200} = 28$	$E_{12} = \frac{80 \cdot 130}{200} = 52$
Non Smoker	$E_{21} = \frac{70 \cdot 120}{200} = 42$	$E_{22} = \frac{120 \cdot 130}{200} = 78$

Then the test statistic:

$$X^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 44.3223$$

with degrees of freedom $(r-1)(c-1) = 1$.

Finally, the p-value becomes:

$$p = P(X^2_{df=1} > 44.3223) \approx 0$$

So we reject the H_0 and confidently conclude that their is an association between those with lung cancer and their smoking status.

c. Let X represent the number of smokers who develop lung disease in the sample above. Now suppose we randomly select 5 individuals. Find the expected number and variance of smokers with lung disease from the 5 selected.

Solution:
Note that $X \sim (n = 5, p = \frac{50}{80} = 0.625)$.
Then $E(X) = 5 \cdot 0.625 = 3.125$ and $Var(X) = 5 \cdot 0.625(1 - 0.625) = 1.1718$.

d. Suppose the following summary statistics is obtained of smokers and non-smokers from the sample

Smoking Status	Sample Mean Age	Standard Deviation of Age
Smoker	40	10
Non Smoker	45	12

Suppose the standard deviation obtained above is representative of their respective populations. Determine if there is a significant difference in mean ages between smokers and non-smokers.

(*Hint: the test statistics for a two-sample z-test is* $Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$)

Solution:
Let 1 = smokers and 2 = non smokers.
We perform a *two sample z-test.*
H_0: $\mu_1 = \mu_2$
H_1: $\mu_1 \neq \mu_2$
Then the test statistic:

$$Z = \frac{(40 - 45)}{\sqrt{\frac{10^2}{80} + \frac{12^2}{120}}} = -3.19$$

Finally, the p-value

$$p = 2 \cdot P(Z \leq -3.19) = 2 \cdot 0.00071 \approx 0.$$

We reject the H_0 and conclude that their is difference between age of smokers and non-smokers.

e. Calculate the relative risk (RR) and odds ratio (OR) for smokers developing lung disease compared to non-smokers.

Solution:

$$RR = \frac{50/80}{20/120} = 3.75$$

$$OR = \frac{50 \cdot 100}{30 \cdot 20} = 8.333$$

Problem 2

A study seeks to examine whether gender is associated with being a smoker. The researchers survey a population in the current month. Out of 500 total individuals, they find that there are 200 smokers. Out of all the male participants, 150 were smokers, while 75 were not.

Find a point estimate and a 95% confidence interval for this point estimate comparing the difference in risk of smoking between male and females. What can we conclude from these results?

Solution:

	Male	Female	
Non-Smokers	75	225	300
Smokers	150	50	200
	225	275	500

$$\widehat{RD} = \frac{150}{225} - \frac{50}{275} = 0.4848$$

Note that our $\alpha = 0.05$.

$$\widehat{RD} \pm z_{1-\alpha/2} * \sqrt{\frac{ac}{n_1^3} + \frac{bd}{n_2^3}} = 0.4848 \pm 1.96 * \sqrt{\frac{75 * 150}{225^3} + \frac{225 * 50}{275^3}} = (0.408, 0.561)$$

```
# calculating in R
lwr <- (150/225 - 50/275) - (1.96 * sqrt((75*150)/(225^3)+(225*50)/(275^3)))
upr <- (150/225 - 50/275) + (1.96 * sqrt((75*150)/(225^3)+(225*50)/(275^3)))
paste("(", round(lwr, 3), ",", round(upr, 3),  ")", sep = "")
```

```
[1] "(0.408,0.561)"
```

We conclude that there is a significant difference in the risk of smoking between males and females.