Test 1

STAT 292 Applied Statistics 2A

08/05/2020

1. **c.**
2. **d.**
3. **e.**
4. **a.** unless $n(1-p)$ is a typo and it's supposed to be the variance $np(1-p)$ then it's **b.**
5. **b.**
6. **b.**
7. **a.**
8. **d.**
9.

- a.

$\hat{\theta}_{XY(M)} = \frac{647\times27}{622\times2} = \mathbf{14.0426}$

$\hat{\theta}_{XY(F)} = \frac{41\times32}{28\times19} = \mathbf{2.4662}$

- b. The odds of male smokers having lung cancer are 14 times the odds of male non-smokers having lung cancer and the odds of female smokers are 2.5 times the odds of female non-smokers that have lung cancer. $\frac{14.0426}{2.4662} = 5.6940$, the odds of males who smoke having lung cancer is 5.6940 times the odds of females who smoke having lung cancer. Since there's a big difference between the odds ratio for males and the odds ratio for females there's evidence that the sex of the individual and the smoker status do interact.

- c.

$O\hat{d}ds_Y = \frac{\frac{688}{709}}{\frac{21}{709}} \approx 32.7619047619$

$O\hat{d}ds_N = \frac{\frac{650}{709}}{\frac{59}{709}} \approx 11.0169491525$

$\hat{\theta}_{XY} = \frac{O\hat{d}ds_Y}{O\hat{d}ds_N} \approx \frac{32.7619047619}{11.0169491525} \approx \mathbf{2.97377289378}$

The odds of a smoker having lung cancer is **2.9738** times more than the odds of a non-smoker having lung cancer.

$log\hat{\theta}_{XY} \pm z_{1-\frac{\alpha}{2}} \times \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$

$log\ 2.97377289378 \pm 1.960 \times \sqrt{\frac{1}{688} + \frac{1}{21} + \frac{1}{650} + \frac{1}{59}}$

$1.08983148123 \pm 1.960 \times 0.25992335422$

$1.08983148123 + 0.50944977427 \approx 1.5992812555$

$1.08983148123 - 0.50944977427 \approx 0.58038170696$

$(exp(0.58038170696),\ exp(1.5992812555)) \approx (1.7867,\ 4.9495)$

With 95% confidence, the true odds ratio is between 1.7867 and 4.9495

- d.

– i. $\hat{\mu}_{11} = \frac{n_{1+}n_{+1}}{n} = \frac{709 \times 1338}{1418} = 669$, $\hat{\mu}_{12} = \frac{n_{1+}n_{+2}}{n} = \frac{709 \times 80}{1418} = 40$, $\hat{\mu}_{21} = \frac{n_{2+}n_{+1}}{n} = \frac{709 \times 1338}{1418} = 669$, $\hat{\mu}_{22} = \frac{n_{2+}n_{+2}}{n} = \frac{709 \times 80}{1418} = 40$. All $\hat{\mu}_{ij}$ for all i, j are ≥ 5 so we can assume a chi-square test of independence would be appropriate for the data.

– ii.

$$\mathcal{H}_0 \; : \; \textit{Lung cancer and smoker status are independent}$$

$$\mathcal{H}_1 \; : \; \textit{Lung cancer and smoker status are not independent}$$

– iii.

Pearson χ^2 statistic:

– $X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} \approx \frac{(688-669)^2}{669} + \frac{(21-40)^2}{40} + \frac{(650-669)^2}{669} + \frac{(59-40)^2}{40} \approx \mathbf{19.1292}$

– Under $\mathcal{H}_0$, $X^2 \sim X^2_{(I-1)(J-1)} \implies X^2_{(2-1)(2-1)} \implies X^2_1$

Likelihood ratio χ^2 statistic:

– $G^2 = 2 \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} log(\frac{n_{ij}}{\hat{\mu}_{ij}}) \approx 2 \times 688 \; log(\frac{688}{669}) + 2 \times 21 \; log(\frac{21}{40}) + 2 \times 650 \; log(\frac{650}{669}) + 2 \times 59 \; log(\frac{59}{40}) \approx$ $\mathbf{19.8780}$

– Under $\mathcal{H}_0$, $G^2 \sim X^2_{(I-1)(J-1)} \implies X^2_{(2-1)(2-1)} \implies X^2_1$

– iv. The SAS output shows us that the p-value for both the Pearson test (denoted as Chi-Square) and the Likelihood ratio test is $< .0001$, incredibly small.

– v. The p-value is significantly smaller than the significance level $\alpha = 0.05$, thus we reject the null hypothesis $\mathcal{H}_0$ meaning there is sufficient evidence that lung cancer and smoke status are **not independent**.

- e.

– i.

$$\mathcal{H}_0 \; : \; \theta = 1$$

$$\mathcal{H}_1 \; : \; \theta > 1$$

– ii. We're doing a one-sided hypothesis test so we're looking for the `Right-sided Pr >= F` row of the results table which gives us $< .0001$ a significantly small p-value.

– iii. A mid-p-value is used to adjust for discreteness in small sample sizes. Since our estimated expected frequencies are all above 5 our sample size is sufficiently large so we don't need to calaculate the mid-p-value.

– iv. Our p-value is significantly smaller than the significance level $\alpha = 0.05$ so we reject the null hypothesis meaning the someone having lung cancer is not independent of smoke status.

10.

- a.

$P(Y \geq 3) = 1 - (P(Y=2) + P(Y=1) + P(Y=0)) = 1 - (0.2019 + 0.32303 + 0.25843) = \mathbf{0.21664}$

$\hat{f}_0 = P(Y=0) \times n = 0.2019 \times 30 = \mathbf{6.057}$

- b.

$\chi^2 = \sum_{r=1}^{k} \frac{(f_r - \hat{f}_r)^2}{\hat{f}_r} = \frac{(6-6.057)^2}{6.057} + \frac{(6-9.6909)^2}{9.6909} + \frac{(12-7.7529)^2}{7.7529} + \frac{(6-6.4992)^2}{6.4992} =$

$0.00053640416 + 1.40572524843 + 2.32659500445 + 0.03834327917 = \mathbf{3.77119993621}$

- c.

Under $\mathcal{H}_0,\ \chi^2 \sim \chi^2_{k-m-1} \implies \chi^2 \sim \chi^2_{4-1-1} \implies \chi^2 \sim \chi^2_2.$

The degree of freedom is 2 because k represents the number of categories which is 4, and m represents the number of parameters we had to estimate which is 1, λ the population mean. Thus $k - m - 1 = 4 - 1 - 1 = 2$

- d.

Since our p-value is larger than our significance level $\alpha = 0.05$, we have insufficient evidence to reject $\mathcal{H}_0$. In other words there's not sufficient evidence to conclude that our model does not fit a Poisson distribution.