Homework 2

Daniel Hartig

September 27, 2017

Problem 3.3

a., b., c.
For a 95% confidence interval, $\alpha/2 = 0.025$ and $z_{\alpha/2} = -1.960$.

$$\log\hat{\theta} = \log\left(\frac{60(61)}{2(44)}\right) = 3.728$$
$$\hat{\sigma}\left(\log\hat{\theta}\right) = \sqrt{\frac{1}{60} + \frac{1}{61} + \frac{1}{2} + \frac{1}{44}} = .746$$
$$\log\hat{\theta} \pm z_{\alpha/2}\hat{\sigma}\left(\log\hat{\theta}\right) = 3.728 \pm 1.960(0.746) = (2.267, 5.189)$$
$$\text{Confidence Interval}[\hat{\theta}] = (e^{2.267}, e^{5.189}) = (9.648, 179.299)$$

The main factor causing the confidence interval to be so large is the sample of two in the category of "Strong Republican, Strong Agree." In the calculation of standard error, it can be seen that the term $1/2$ dominates within the square root.

Problem 3.4

a. Odds Ratio

$$\log\hat{\theta} = \log\left(\frac{1085(441239)}{703(55623)}\right) = 2.505$$
$$\hat{\sigma}\left(\log\hat{\theta}\right) = \sqrt{\frac{1}{1085} + \frac{1}{55623} + \frac{1}{703} + \frac{1}{441239}} = 0.0486$$
$$\log\hat{\theta} \pm z_{\alpha/2}\hat{\sigma}\left(\log\hat{\theta}\right) = 2.505 \pm 1.960(0.0486) = (2.464, 2.546)$$
$$\text{Confidence Interval}[\hat{\theta}] = (e^{2.464}, e^{2.546}) = (11.756, 12.751)$$

A driver or passenger involved in an auto accident is between 11.8 and 12.8 times more likely to become a fatality if that person is not wearing a seat-belt compared to if that person is wearing a seat belt.

b. Difference of Proportions
Let $\hat{\pi}_{no}$ be the probability of a fatality for a driver or passenger not using a seat-belt and $\hat{\pi}_{yes}$ be the probability if using a seat-belt.

$$\hat{\pi}_{no} = \frac{y_{no}}{n_{no}} = \frac{1085}{1085 + 55623} = 0.0191$$
$$\hat{\pi}_{yes} = \frac{y_{yes}}{n_{yes}} = \frac{703}{703 + 441239} = 0.00159$$
$$\hat{\pi}_{no} - \hat{\pi}_{yes} = 0.0191 - 0.00159 = 0.0175$$
$$\hat{\sigma}\left(\hat{\pi}_{no} - \hat{\pi}_{yes}\right) = \sqrt{\frac{\hat{\pi}_{no}(1-\hat{\pi}_{no})}{n_{no}} + \frac{\hat{\pi}_{no}(1-\hat{\pi}_{no})}{n_{no}}} = \sqrt{\frac{0.0191(1-0.0191)}{55623} + \frac{0.00159(1-.00159)}{441239}} = 0.000584$$
$$\text{Confidence Interval}[\hat{\pi}_{no} - \hat{\pi}_{yes}] = \hat{\pi}_{no} - \hat{\pi}_{yes} \pm z_{\alpha/2}\hat{\sigma}\left(\hat{\pi}_{no} - \hat{\pi}_{yes}\right) = 0.0175 \pm 1.960(0.000584) = (0.0164, 0.0187)$$

A driver or passenger involved in an auto accident is between 1.6 and 1.9 percentage points more likely to become a fatality if that person is not wearing a seat-belt compared to if that person is wearing a seat belt.

c. Relative Risk

$$\log r = \log\left(\frac{\hat{\pi}_{no}}{\hat{\pi}_{yes}}\right) = \log\left(\frac{0.0191}{0.00159}\right) = 2.487$$

$$\hat{\sigma}(\log r) = \sqrt{\frac{1-\hat{\pi}_{no}}{y_{no}} + \frac{1-\hat{\pi}_{yes}}{y_{yes}}} = \sqrt{\frac{1-.0191}{1085} + \frac{1-.00159}{703}} = 0.0482$$

$$\log r \pm z_{\alpha/2}\hat{\sigma}(\log r) = 2.505 \pm 1.960(0.0482) = (2.410, 2.599)$$

$$\text{Confidence Interval}[r] = (e^{2.410}, e^{2.599}) = (10.943, 13.220)$$

A driver of passenger involved in an auto accident is between 10.9 and 13.2 times more likely to become a fatality if that person is not wearing a seat belt compared to all people involved in car accidents. The reason this number is so close to the values obtained in the Odds Ratio confidence interval is that the probability of becoming a fatality is low, so the number of non-fatal accidents and the number of total accidents are close.

Problem 3.9

The 2014 edition of the GSS yielded the following table for the variables "EDUCATION" and "FUND"

	Fundamentalist	Moderate	Liberal	Row Total
K-12	307	461	270	1038
13-18	262	508	472	1242
19 or greater	16	56	69	141
Column Total	585	1025	811	2421

Solved using python 3.5 code attached at end of assignment.

a.
The X^2 statistic is 63.822 with p-value 4.556×10^{-13}. The G^2 statistic is 65.443 with p-value 2.075×10^{-13}. There are 4 degrees of freedom. Because of the very high p-values, we have very high confidence that education and fundamentalism are not independent variables.

b.
There is a clear relationship between the fundamentalist and liberal positions and having or not having a college education. The number of fundamentalists with no college education is 5.39 standard deviations above expected, while the number of fundamentalists with a college education is 3.62 stdev below expected and graduate school is 3.66 stdev below expected. Conversely, the number of liberals with no college education is 6.76 stdev lower than expected while the number of liberals with a college education is 4.81 stdev above expected and grad school is 4.00 stdev above expected.

The moderate position has a similar relationship to education as the fundamentalist, but this relationship is weaker. There is limited difference in the ideologies of those people who have been to college, and those who have been to graduate school. Overall, the significant finding is that fundamentalism is more likely in those without college education and liberalism is more likely in those with college education.

Problem 3.31

a.
For the given table, the marginal distributions in both directions for the given table are

$$\pi_{1+} = \pi_{+1} = \theta^2 + \theta(1-\theta) = \theta$$
$$\pi_{2+} = \pi_{+2} = (1-\theta)^2 + \theta(1-\theta) = 1-\theta$$

so that the resulting table with marginals is

	Success in j	Failure in j	Marginal in j
Success in i	θ^2	$\theta(1-\theta)$	θ
Failure in i	$\theta(1-\theta)$	$(1-\theta)^2$	$(1-\theta)$
Marginal in i	θ	$(1-\theta)$	1

To prove independence, we calculate π_{ij} from the marginals for all entries of the table and verify that we see the expected values in the table.

$$\begin{aligned}\pi_{11} &= \theta^2 = \theta(\theta) = \pi_{1+}\pi_{+1} \\ \pi_{12} = \pi_{21} &= \theta(1-\theta) = \pi 1 + \pi + 2 = \pi + 1 \pi 2 + \\ \pi_{22} &= (1-\theta)^2 = (1-\theta)(1-\theta) = \pi_{2+}\pi_{+2}\end{aligned}$$

b.
Since $\hat{\theta}$ represents the overall proportion of expected successes of either trial i or trial j, it is represented by

$$\begin{aligned}\hat{\theta} &= \frac{\hat{n}_1 + \hat{n}_2}{\hat{n} + \hat{n}} = \frac{1}{2}(\pi_1 + \pi_2) \\ &= \frac{\pi_{1+} + \pi_{+1}}{2}\end{aligned}$$

c.
To test a null hypothesis of independence, we use the Chi Squared test. The test statistic is

$$\chi^2 = \sum_i \sum_j \frac{\left(n_{ij} - E(n_{ij})\right)^2}{E(n_{ij})}$$

If we divide through by n, the unknown sample size, we obtain an expression in terms of sample and expected proportions

$$\chi^2 = \sum_i \sum_j \frac{\left(\pi_{ij} - \hat{\pi}_{ij}\right)^2}{\hat{\pi}_{ij}}.$$

The degrees of freedom of a test of H_0 are $(I-1) + (J-1) = (2-1) + (2-1) = 2$. We expect the two degrees of freedom because the table is symmetrical in addition to being independent. We then find a p-value for the obtained test statistic with the given degrees of freedom and use our desired confidence level to determine whether or not to reject H_0.

d.
The contingency table of sampled shots in its proportional form is

	Made 2nd	Missed 2nd	
Made 1st	0.661	0.143	0.804
Missed 1st	0.161	0.035	0.196
	0.822	0.178	1

The overall proportion of shots made is 0.813, which is $\hat{\theta}$. Therefore, the expected shots table in proportional form is

	Made 2nd	Missed 2nd	
Made 1st	0.661	0.152	0.813
Missed 1st	0.152	0.035	0.187
	0.813	0.187	1

The chi squared test statistic is

$$\begin{aligned}\chi^2 &= \frac{(0.661-0.661)^2}{0.661} + \frac{(0.143-0.152)^2}{0.152} + \frac{(0.161-0.152)^2}{0.152} + \frac{(0.035-0.035)^2}{0.035} \\ &= 4.379 \times 10^{-8} + 4.782 \times 10^{-4} + 5.171 \times 10^{-4} + 8.281 \times 10^{-7} \\ &= 0.000996\end{aligned}$$

Since the p-value is so low we cannot reject the null hypothesis. It is very plausible that Kobe Bryant's free throws are both independent and identically distributed.

Stat 665 HW2 Prob 3.9

September 26, 2017

```
In [28]: import scipy.stats as stats, numpy as np

In [29]: ctable = np.array([[307,461,270],[262,508,472],[16,56,69]])

In [30]: chi2, p, df, ex = stats.chi2_contingency(ctable)

In [31]: chi2, p, df

Out[31]: (63.821970741954601, 4.5559193424393558e-13, 4)

In [32]: chi2_g, p_g, df_g, ex_g = stats.chi2_contingency(ctable, lambda_ = 'log-likelihood')

In [33]: chi2_g, p_g, df_g

Out[33]: (65.442909051143943, 2.0756740637919442e-13, 4)

In [34]: (ctable - ex)/np.sqrt(ex)

Out[34]: array([[ 3.54747372,  1.02715989, -4.16766792],
                [-2.199961  , -0.77782631,  2.74290262],
                [-3.09587378, -0.47841592,  3.16720923]])

In [35]: n = np.sum(ctable)

In [36]: ed_sum, fund_sum = stats.contingency.margins(ctable)

In [37]: (ctable - ex)/np.sqrt(ed_sum*(n-ed_sum) * fund_sum*(n-fund_sum) / n**3)

Out[37]: array([[ 5.38972427,  1.78969507, -6.7618241 ],
                [-3.62006634, -1.46783691,  4.81987015],
                [-3.66331508, -0.64921759,  4.00212992]])
```