STAT430 Homework #6: Due Friday, March 29, 2019 (after break).

Name: Oliver Shanklin

---

0. We will continue with **Chapter 8** on estimation. You should have read **Sections 8.1-8.4** on point estimation; please continue with **Sections 8.5-8.10** on interval estimation and related topics.
1. Complete **Exercise 8.56** of the text.

Answer:

$$\theta = p, \hat{\theta} = \hat{p} = Y/n \sim N(p, \frac{p(1-p)}{n})$$

$$\hat{\sigma}^2_{\hat{\theta}} = \frac{\hat{p}(1-\hat{p})}{n}$$

Where $Z_{\alpha/2} = 2.33$.

A 98% CI for $\theta = p$ is

$$[0.409, 0.4909]$$

The interval does not include 0.50 so a majority of adults do not think that movies are getting better.

---

2. Complete **Exercise 8.60** of the text. In part (b), "What conclusions can you draw?", think about what you can conclude if your temperature is above 98.6.

Answer:

For a 99% CI, $Z_{\alpha/2} = 2.57$

$\theta = \mu$

$$\hat{\theta} = \bar{Y} \sim N(\mu, \sigma^2/n)$$

$$\hat{\sigma}^2_{\hat{\theta}} = S^2/n$$

So a 99% CI for $\theta = \mu$ is

$$[98.0855, 98.4145]$$

This interval does not include 98.6 so everyone is cold all the time.

---

3. In class, we considered the Physician's Health Study, in which a large number of physicians were randomized to receive a treatment of an aspirin every other day or a placebo every other day. They were followed for five years, and the number of heart attacks in each group was recorded. Among the $n_1 = 11034$ physicians taking placebo, there were 189 heart attacks, and among the $n_2 = 11037$ taking aspirin, there were 104 heart attacks. Give a point estimate for the log(relative risk) and an

approximate 99% confidence interval. Then exponentiate your point estimate and the endpoints of your CI to give an approximate 99% confidence interval for relative risk. Interpret your point estimate and confidence interval.

Answer:

$\theta = ln(p_1/p_2) = ln(p_1) - ln(p_2)$

$\hat{\theta} = ln(\hat{p}_1) - ln(\hat{p}_2)$

Where, $ln(p_1) \sim N(p_1, (1-p_1)/(n_1 p_1)$

So,

$$\hat{\theta} \sim N(ln(p_1) - ln(p_2), \frac{1-p_1}{n_1 p_1} + \frac{1-p_2}{n_2 p_2})$$

A 99% CI for the log-relative-risk is,

$$[0.28577, 0.90949]$$

And when we exponentiate, we get the relative risk,

$$[1.33, 2.48]$$

and this shows that p_1 has a higher success rate than p_2.

---

4. In a previous homework, you used the delta method to determine the approximate distribution of the estimated log-odds for large sample size, where the odds of success are defined as

$$\frac{\text{probability of success}}{\text{probability of failure}}$$

Now suppose that $Y_1 \sim \text{Binomial}(n_1, p_1)$, independent of $Y_2 \sim \text{Binomial}(n_2, p_2)$, where both n_1 and n_2 are large. The **odds ratio** is defined as

$$\frac{\text{odds in group 1}}{\text{odds in group 2}} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}.$$

4(a). Consider the log-odds ratio,

$$\theta = \ln\left(\frac{p_1/(1-p_1)}{p_2/(1-p_2)}\right) = \ln\left(\frac{p_1}{1-p_1}\right) - \ln\left(\frac{p_2}{1-p_2}\right).$$

Construct an approximate $(1-\alpha)$ 100% confidence interval for the log-odds ratio.

Answer:

$$\theta = ln(p_1/(1-p_1)) - ln(p_2/(1-p_2))$$

$$\hat{\theta} = ln(\hat{p}_1/(1-\hat{p}_1)) - ln(\hat{p}_2/(1-\hat{p}_2))$$

$$\hat{\theta} \sim N(ln(\hat{p}_1/(1-\hat{p}_1)) - ln(\hat{p}_2/(1-\hat{p}_2)), \frac{1}{n_1 p_1} + \frac{1}{n_1(1-p_1)} + \frac{1}{n_2 p_2} + \frac{1}{n_2(1-p_2)})$$

So the $(1-\alpha)100\%$ CI for the log odds ratio is,

$$[ln(\hat{p}_1/(1-\hat{p}_1)) - ln(\hat{p}_2/(1-\hat{p}_2)) \pm Z_{\alpha/2}\sqrt{\frac{1}{n_1 p_1} + \frac{1}{n_1(1-p_1)} + \frac{1}{n_2 p_2} + \frac{1}{n_2(1-p_2)}}]$$

4(b). The *British Medical Journal* reported the following data for 114 patients with spondyloarthropathies (a kind of joint disease) and 334 patients without spondyloarthropathies. In the "with" group, 54 had the ABO secretor state (a genetic feature), and in the "without" group, 89 had the ABO secretor state. The journal reported an odds ratio of 2.48 and a 95% confidence interval of $[1.59, 3.85]$. Use your computed confidence interval from 4(a) to reproduce the journal's reported confidence interval.

Answer:

$\hat{p}_1 = 54/114$

$\hat{p}_2 = 89/334$

$Z_{\alpha/2} = 1.96$

$$[ln(54/114) - ln(89/334) \pm 1.96 * \sqrt{\frac{1}{114 * 54/114} + \frac{1}{114(1-54/114)} + \frac{1}{334 * 89/334} + \frac{1}{334(1-89/334)}}]$$

$$\implies [0.72125, 1.09326]$$

Which is the log-odds-ratio, and even if I exponentiate, the interval is way off, so my variance is probably incorrect.

---

5. In baseball, a *complete game* is one in which a single pitcher does all of the pitching for one team, without the benefit of a relief pitcher. A *complete game shutout* is one in which that pitcher allows no runs by the opposing team. Such games are rare. For the 2016 Major League Baseball (MLB) season, there were 36 complete game shutouts out of 2,427 games played. The dataset `2016_MLB_Pitching.csv` contains pitching statistics for each of the 30 MLB teams for the 2016 season, including the number of complete game shutouts (`cSho`) for each team. If you put the dataset in a subfolder called `Data` of the homework folder where this markdown file is stored, you can read it as follows:

```r
MLB <- read.csv("2016_MLB_Pitching.csv", header = TRUE) # read in the dataset and call it MLB
head(MLB)        # look at the first few lines of the dataset
```

```
##    Tm NumP PAge RA.G   W  L  W.L.  ERA   G  GS  GF CG tSho cSho SV     IP
## 1 ARI   29 26.4 5.49  69 93 0.426 5.09 162 162 160  2    7    2 31 1451.1
## 2 ATL   35 26.4 4.84  68 93 0.422 4.51 161 161 160  1    9    1 39 1447.2
## 3 BAL   27 27.9 4.41  89 73 0.549 4.22 162 162 161  1    9    0 54 1432.0
## 4 BOS   25 29.0 4.28  93 69 0.574 4.00 162 162 153  9    5    1 43 1439.2
## 5 CHC   26 29.9 3.43 103 58 0.640 3.15 162 162 157  5   15    2 38 1459.2
## 6 CHW   28 28.6 4.41  78 84 0.481 4.10 162 162 155  7   10    1 43 1446.2
##      H   R  ER  HR  BB IBB   SO HBP BK WP   BF ERA.  FIP  WHIP  H9 HR9 BB9
## 1 1563 890 821 202 603  57 1318  57  7 69 6437   88 4.50 1.492 9.7 1.3 3.7
## 2 1414 779 725 177 547  55 1227  70  4 83 6250   92 4.32 1.355 8.8 1.1 3.4
## 3 1408 715 671 183 545  23 1248  51  3 59 6122  101 4.31 1.364 8.8 1.2 3.4
## 4 1342 694 640 176 490  16 1362  65  0 52 6073  111 4.00 1.273 8.4 1.1 3.1
## 5 1125 556 511 163 495  24 1441  63  0 80 5933  132 3.77 1.110 6.9 1.0 3.1
## 6 1422 715 659 185 521  30 1270  68  2 72 6196   99 4.27 1.343 8.8 1.2 3.2
##   SO9 SO.W  LOB
## 1 8.2 2.19 1193
## 2 7.6 2.24 1128
## 3 7.8 2.29 1111
## 4 8.5 2.78 1060
## 5 8.9 2.91  998
## 6 7.9 2.44 1141
```

```
table(MLB$cSho)  # table the values of the variable cSho in the MLB dataset
```

```
## 
##  0  1  2  3 
##  9 11  5  5
```

This dataset and additional details can be found at
https://www.baseball-reference.com/leagues/MLB/2016-standard-pitching.shtml

We might want to model the number of complete shutouts for each team as Binomial with number of trials equal to number of games played (161 or 162) and some small success probability. Since the binomial model would involve a large number of trials and a small probability of success, another alternative would be to model the number of complete shutouts per team as Y_i iid Poisson(λ) ($i = 1, 2, \ldots, 30$). Treat the observations as independent, even though this cannot be exactly true (if one team has a complete shutout in a game, then the opposing team certainly did not.) Treat the observations as identically distributed, though this cannot be exactly true either.

(a). Estimate the target λ, the mean of the Poisson distribution, with $\hat{\lambda} = \bar{Y}$ = sample mean of the $n = 30$ observations.

Answer:

(b). Assume that $n = 30$ is a large sample, and give an approximate 95% confidence interval for the target, λ.

Answer:

(c). Under the Poisson model, the probability that a team has exactly zero complete game shutouts is

$$\theta = P(Y_i = 0) = \frac{\lambda^0 e^{-\lambda}}{0!} = e^{-\lambda}.$$

Give a point estimate of this probability by plugging in $\hat{\lambda}$. Then use the approximate 95% CI from (b) and the fact that

$$0.95 \simeq P\left(\hat{\lambda}_L \leq \lambda \leq \hat{\lambda}_U\right)$$

to construct an approximate 95% confidence interval for the new target, θ.

Answer:

(d). Use the delta method to derive a 95% confidence interval for the probability that a team has exactly one complete game shutout under the Poisson model,

$$\theta = P(Y_i = 1) = \frac{\lambda^1 e^{-\lambda}}{1!} = \lambda e^{-\lambda}.$$

Answer:

Optional problems: Good optional review problems are 8.39, 8.43, 8.48 for confidence intervals with a general pivot; and any of 8.57 thru 8.67 for large-sample normal confidence intervals on simple targets like means, proportions, differences of means, and differences of proportions.