# STAT 457 Homework 05

*Martha Eichlersmith*

*2019-11-16*

## Problem 1

Consider two urns each containing an unknown mixture of blue and white marbles. A random sample of size 18 (with replacement) is drawn from urn #1 and a random sample of size 6 (with replacement) is drawn from urn #2. Of the 18 selected marbles from urn #1, 14 are blue. The corresponding number of blue marbles from urn #2 is 2.

$$L(\pi|Y) \propto \pi^y(1-\pi)^{n-y} = \pi^{14}(1-\pi)^4 \sim \text{Binomial}, n_\pi = 18, \ y_\pi = 14$$
$$\implies p(\pi \mid Y) \sim \text{Beta}(y + \alpha_0, n - y + \beta_0) \quad \text{where} \quad p(\pi) \sim \text{Beta}(\alpha_0, \beta_0)$$

$$L(\psi|Y) \propto \psi^y(1-\psi)^{n-y} = \psi^2(1-\psi)^4 \sim \text{Binomial}, n_\pi = 6, \ y_\pi = 2$$
$$\implies p(\psi \mid Y) \sim \text{Beta}(y + \alpha_0, n - y + \beta_0) \quad \text{where} \quad p(\psi) \sim \text{Beta}(\alpha_0, \beta_0)$$

|       | Blue | White |
|-------|------|-------|
| Urn 1 | 14   | 4     |
| Urn 2 | 2    | 4     |

### Problem 1a

Let $\pi$ denote the proportion of blue marbles in urn #1 and let $\psi$ denote the corresponding proportion in urn #2. Under the (i) Haldane, (ii) flat and (iii) non-informative priors, compute $p\left(\ln\left[\frac{\pi}{1-\pi}\right] > \ln\left[\frac{\psi}{1-\psi}\right] \mid \text{data}\right)$ using the normal approximation.

$$p\left(\ln\left[\frac{\pi}{1-\pi}\right] > \ln\left[\frac{\psi}{1-\psi}\right] \mid \text{data}\right) = p\left(\ln\left[\frac{\pi}{1-\pi}\right] - \ln\left[\frac{\psi}{1-\psi}\right] > 0 \mid \text{data}\right)$$
$$p(\pi) \sim \text{Beta}(\alpha_0, \beta_0)$$
$$p(\psi) \sim \text{Beta}(\alpha_0, \beta_0)$$
$$p(\pi \mid Y) \sim \text{Beta}(y_\pi + \alpha_0, \ n_\pi - y_\pi + \beta_0) \overset{\text{def}}{=} \text{Beta}(\alpha, \beta)$$
$$p(\pi \mid Y) \sim \text{Beta}(y_\pi + \alpha_0, \ n_\pi - y_\pi + \beta_0) \overset{\text{def}}{=} \text{Beta}(\gamma, \delta)$$
$$\text{Normal Approx Mean} = \ln\left(\frac{\alpha \cdot \delta}{\beta \cdot \gamma}\right)$$
$$\text{Normal Approx Variance} = \frac{1}{\alpha} + \frac{1}{\beta} + \frac{1}{\gamma} + \frac{1}{\delta}$$
$$\text{Normal Approx} \sim \mathcal{N}\left(\ln\left(\frac{\alpha \cdot \delta}{\beta \cdot \gamma}\right), \ \sqrt{\frac{1}{\alpha} + \frac{1}{\beta} + \frac{1}{\gamma} + \frac{1}{\delta}}\right)$$

Normal Approx: Probability difference in logodds is greater than 0

|         | Haldane      | Flat         | Non-informative |
|---------|--------------|--------------|-----------------|
| Prior   | Beta(0, 0)   | Beta(1, 1)   | Beta(.5, .5)    |
| p-value | 0.96994      | 0.96402      | 0.96706         |

## Problem 1b

Repeat (1a) by drawing deviates from the appropriate beta distributions. Quantify the Monte Carlo error in your value.

Probability that the log differences are greater than 0

|  | Haldane Beta(0, 0) | | | Flat Beta(1, 1) | | | Non-informative Beta(.5, .5) | | |
|---|---|---|---|---|---|---|---|---|---|
| Iterations | 10000 | 1e+05 | 1e+06 | 10000 | 1e+05 | 1e+06 | 10000 | 1e+05 | 1e+06 |
| p-value | 0.97293 | 0.97254 | 0.97256 | 0.97303 | 0.97258 | 0.97265 | 0.97105 | 0.97251 | 0.97281 |
| Standard Error | 0.01137 | 0.00360 | 0.00114 | 0.01125 | 0.00359 | 0.00113 | 0.01136 | 0.00359 | 0.00113 |

## Problem 1c

Compare your results in (1a) and (1b) to the p-value obtained via Fisher's exact test.

$$\text{Odds Ratio} > 1 \implies \frac{\pi}{1-\pi} > \frac{\psi}{1-\psi}$$

```
##
##  Fisher's Exact Test for Count Data
##
## data:  blue
## p-value = 0.9931
## alternative hypothesis: true odds ratio is less than 1
## 95 percent confidence interval:
##    0.00000 63.37908
## sample estimates:
## odds ratio
##    6.334078

## [1] "Probability that log differences are greater than 0: 0.00686"
```

The Fisher Exact test produces a a probability that is lower than the methods in (1a) and (1b).

**Problem 1d**

Is delinquency related to birth order?

Row Proportion

|  | Oldest | In-between | Youngest | Only Child | Total |
|---|---|---|---|---|---|
| Most Delinquent | 0.35278 | 0.34167 | 0.25833 | 0.04722 | 1 |
| Least Delinquent | 0.44402 | 0.26898 | 0.20335 | 0.08366 | 1 |
| (1/2)Difference (Most - Least) | -0.04562 | 0.03634 | 0.02749 | -0.01822 | 0 |

Difference in Proportion for Oldest : Most Delinquent - Least Delinquent

| Iterations | 10000 | 100000 | 1000000 |
|---|---|---|---|
| Mean | -0.04753 | -0.04558 | -0.04565 |
| 95% CI | ( -0.79796, 0.71730 ) | ( -0.80043, 0.73221 ) | ( -0.79722, 0.73221 ) |
| P(diff>0) | 0.4313 | 0.43721 | 0.436071 |

Difference in Proportion for In-Between : Most Delinquent - Least Delinquent

| Iterations | 10000 | 100000 | 1000000 |
|---|---|---|---|
| Mean | 0.04349 | 0.03782 | 0.03614 |
| 95% CI | ( -0.66624, 0.74943 ) | ( -0.67684, 0.73751 ) | ( -0.67381, 0.74165 ) |
| P(diff>0) | 0.5723 | 0.56754 | 0.564099 |

Difference in Proportion for Youngest : Most Delinquent - Least Delinquent

| Iterations | 10000 | 100000 | 1000000 |
|---|---|---|---|
| Mean | 0.02591 | 0.02642 | 0.02761 |
| 95% CI | ( -0.64223, 0.68651 ) | ( -0.61413, 0.67724 ) | ( -0.61398, 0.68130 ) |
| P(diff>0) | 0.558 | 0.56111 | 0.563085 |

Difference in Proportion for Only : Most Delinquent - Least Delinquent

| Iterations | 10000 | 100000 | 1000000 |
|---|---|---|---|
| Mean | -0.01886 | -0.01806 | -0.01811 |
| 95% CI | ( -0.44356, 0.29082 ) | ( -0.43012, 0.28790 ) | ( -0.42690, 0.28280 ) |
| P(diff>0) | 0.3603 | 0.36121 | 0.359778 |

Combined p-values: $X = -2\sum_{i=1}^{4}(p_i) \sim \chi^2_{df=8}$

| Iterations | 10000 | 1e+05 | 1e+06 |
|---|---|---|---|
| Combined p-value | 0.02310 | 0.02168 | 0.02265 |

There is evidence that birth order has an effect on delinquency rates. There may be an argument that younger brothers are more delinquent, but my younger brother is not a delinquent (anecdotal evidence!).

# Problem 2

Suppose a sample of size $n$ is drawn at random and with replacement from some population. For large $n$ the sample proportion $(\hat{p})$ is normally distributed with mean $p$ and variance $\frac{p(1-p)}{n}$. Find the asymptotic distribution of $2\sin^{-1}\sqrt{\hat{p}}$ using the delta method.

$$\hat{p} \sim \mathcal{N}\left(p, \ \frac{p(1-p)}{n}\right)$$

$$\text{Let } g(\hat{p}) = 2\sin^{-1}\sqrt{\hat{p}}$$

$$(g(\hat{p}) - g(p)) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \ \sigma^2[g'(p)]^2\right)$$

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

$$g'(p) = \frac{1}{\sqrt{1-p}\sqrt{p}}$$

$$[g'(p)]^2 = \frac{1}{(1-p)p}$$

$$\sigma^2[g'(p)]^2 = \frac{p(1-p)}{n} \cdot \frac{1}{(1-p)p} = \frac{1}{n}$$

$$(g(\hat{p}) - g(p)) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \ \tfrac{1}{n}\right)$$

$$g(\hat{p}) \xrightarrow{\mathcal{D}} \mathcal{N}\left(g(\hat{p}), \ \frac{1}{n}\right)$$

$$2\sin^{-1}\sqrt{\hat{p}} \xrightarrow{\mathcal{D}} \mathcal{N}\left(2\sin^{-1}\sqrt{p}, \ \frac{1}{n}\right)$$

# Problem 3

Let $x_1, \cdots, x_n$ be an iid sample from $\mathcal{N}(\theta, 1)$ and let $y_1, \cdots, y_n$ be an independent iid sample from $\mathcal{N}(\phi, 1)$. Derive the distribution of $\overline{x}/\overline{y}$ (where $\overline{y} \neq 0$) via the delta method.

$$\overline{x} \sim \mathcal{N}(\theta, 1/n)$$

$$\overline{y} \sim \mathcal{N}(\phi, 1/n)$$

$$\text{Let } h(x,y) = x/y \quad \text{so} \quad h(B) = \overline{x}/\overline{y} \quad \text{and} \quad h(\beta) = \theta/\phi$$

$$\sqrt{n}\left(h(B) - h(\beta)\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \nabla h(\beta)^T \cdot \Sigma \cdot \nabla h(\beta)\right)$$

$$\Sigma = \begin{bmatrix} 1/n & 0 \\ 0 & 1/n \end{bmatrix}$$

$$\nabla h(\beta)^T = \begin{bmatrix} \frac{\partial h}{\partial x} & \frac{\partial h}{\partial y} \end{bmatrix}_{\theta, \phi} = \begin{bmatrix} \frac{1}{y} & -\frac{x}{y^2} \end{bmatrix}_{\theta, \phi} = \begin{bmatrix} \frac{1}{\phi} & -\frac{\theta}{\phi^2} \end{bmatrix}$$

$$\nabla h(\beta)^T \cdot \Sigma \cdot \nabla h(\beta) = \begin{bmatrix} \frac{1}{\phi} & -\frac{\theta}{\phi^2} \end{bmatrix} \begin{bmatrix} 1/n & 0 \\ 0 & 1/n \end{bmatrix} \begin{bmatrix} \frac{1}{\phi} \\ -\frac{\theta}{\phi^2} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{n\phi} & -\frac{\theta}{n\phi^2} \end{bmatrix} \begin{bmatrix} \frac{1}{\phi} \\ -\frac{\theta}{\phi^2} \end{bmatrix}$$

$$= \frac{1}{n}\left(\frac{1}{\phi^2} - \frac{\theta^2}{\phi^4}\right)$$

$$\sqrt{n}\left(\overline{x}/\overline{y} - \theta/\phi\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{1}{n}\left(\frac{1}{\phi^2} - \frac{\theta^2}{\phi^4}\right)\right)$$

$$\frac{\overline{x}}{\overline{y}} \xrightarrow{\mathcal{D}} \mathcal{N}\left(\frac{\theta}{\phi}, \ \frac{1}{n^2}\left(\frac{1}{\phi^2} - \frac{\theta^2}{\phi^4}\right)\right)$$

# Problem 4

197 animals are distributed into four categories: $Y = (y_1, y_2, y_3, y_4)$ according to the genetic linkage model $\left(\frac{2+\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right)$. In HW#4 you derived the likelihood for the data $Y = (125, 18, 20, 34)$ and you derived the likelihood for the data $Y = (14, 0, 1, 15)$. In that homework, you also used Newton-Raphson algorithm to obtain the MLE $(\hat{\theta})$ of $\theta$ and the standard error of $\hat{\theta}$.

$$L(\theta \mid \boldsymbol{Y}) = \frac{(y_1 + y_2 + y_3 + y_4)!}{y_1! y_2! y_2! y_4!} \left(\frac{2+\theta}{4}\right)^{y_1} \left(\frac{1-\theta}{4}\right)^{y_2} \left(\frac{1-\theta}{4}\right)^{y_3} \left(\frac{\theta}{4}\right)^{y_4}$$

$$\propto (2+\theta)^{y_1} \cdot (1-\theta)^{y_2+y_3} \cdot (\theta)^{y_4}$$

$$\ell(\theta \mid \boldsymbol{Y}) \propto y_1 \log(2+\theta) + (y_2 + y_3) \log(1-\theta) + y_4 \log(\theta)$$

$$\frac{\partial \ell}{\partial \theta} = \frac{y_1}{2+\theta} - \frac{y_2 + y_3}{1-\theta} + \frac{y_4}{\theta}$$

$$\frac{\partial^2 \ell}{\partial \theta^2} = -\frac{y_1}{(2+\theta)^2} - \frac{y_2 + y_3}{(1-\theta)^2} - \frac{y_4}{\theta^2}$$
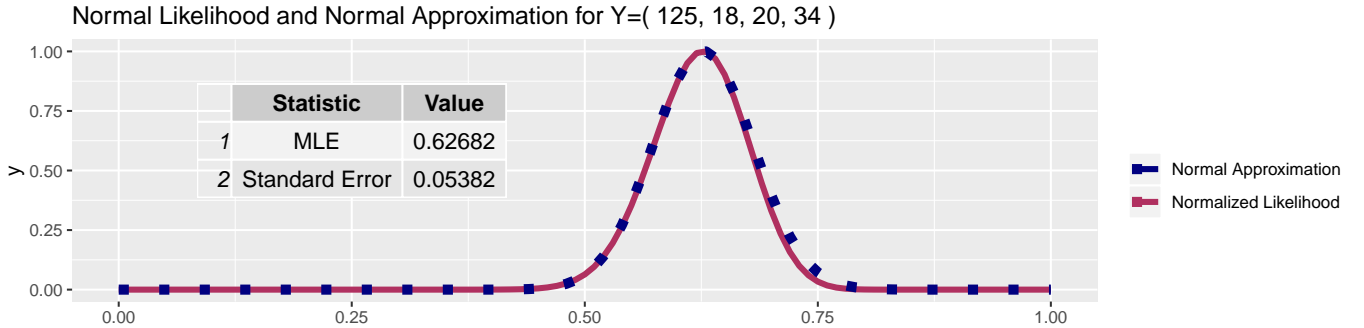
$$\theta^{(i+1)} = \theta^{(i)} - \frac{\frac{y_1}{2+\theta^{(i)}} - \frac{y_2+y_3}{1-\theta^{(i)}} + \frac{y_4}{\theta^{(i)}}}{-\frac{y_1}{(2+\theta^{(i)})^2} - \frac{y_2+y_3}{(1-\theta^{(i)})^2} - \frac{y_4}{(\theta^{(i)})^2}} \overset{\text{Newton-Raphson}}{\implies} \hat{\theta}$$

$$s.e.(\hat{\theta}) = \sqrt{1/\mathcal{I}(\theta)}$$

$$\mathcal{I}(\theta) = \left[\frac{\partial^2 \ell}{\partial \theta^2}\right]_{\hat{\theta}} = -\frac{y_1}{(2+\hat{\theta})^2} - \frac{y_2 + y_3}{(1-\hat{\theta})^2} - \frac{y_4}{\hat{\theta}^2}$$
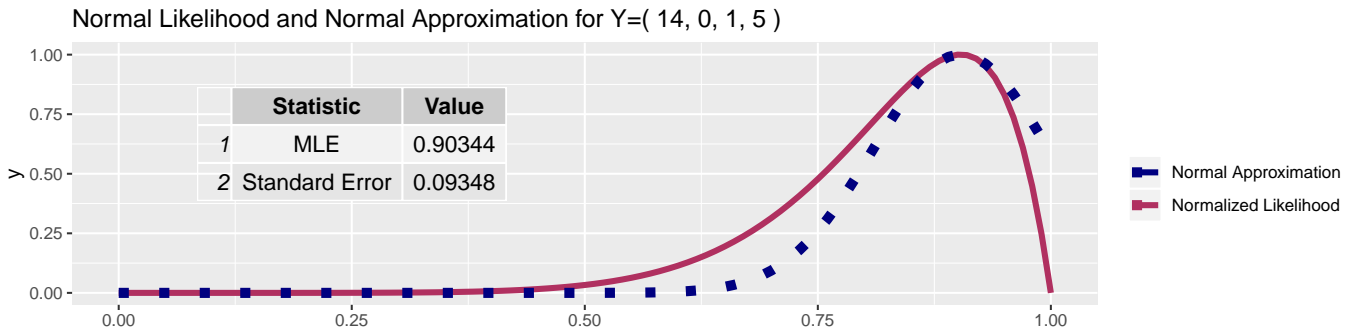
## Problem 4a

Plot the normalized likelihood and the associated normal approximation in the same figure for the data $Y = (125, 18, 20, 34)$. Discuss the adequacy of the normal approximation.



Normal Likelihood and Normal Approximation for Y=( 125, 18, 20, 34 )

| | Statistic | Value |
|---|---|---|
| 1 | MLE | 0.62682 |
| 2 | Standard Error | 0.05382 |

## Problem 4b

Repeat (4a) for $Y = (14, 0, 1, 5)$



Normal Likelihood and Normal Approximation for Y=( 14, 0, 1, 5 )

| | Statistic | Value |
|---|---|---|
| 1 | MLE | 0.90344 |
| 2 | Standard Error | 0.09348 |

## Problem 5

Use Laplace's method (second order) to compute the posterior mean (under a flat prior) for the genetic linkage model for both data sets.

$$L(\theta \mid Y) \propto (2 + \theta)^{y_1} (1 - \theta)^{y_2 + y_3} (\theta)^{y_4}$$

$$\ell(\theta \mid Y) \propto y_1 \ln(2 + \theta) + (y_2 + y_3) \ln(1 - \theta) + y_4 \ln(\theta)$$

$$p(\theta) = \theta^{\alpha - 1}(1 - \theta)^{\beta - 1}$$

$$\text{Flat Prior } p(\theta) = \theta^{1-1}(1 - \theta)^{1-1} = 1$$

$$-nh(\theta) = \ell(\theta \mid Y) + \ln(p(\theta))$$

$$= y_1 \ln(2 + \theta) + (y_2 + y_3) \ln(1 - \theta) + y_4 \ln(\theta) + ln(1)$$

$$= y_1 \ln(2 + \theta) + (y_2 + y_3) \ln(1 - \theta) + y_4 \ln(\theta)$$

$$-nh^*(\theta) = \ell(\theta \mid Y) + \ln(p(\theta)) + \underbrace{\ln(g(\theta))}_{\ln(\theta)}$$

$$= y_1 \ln(2 + \theta) + (y_2 + y_3) \ln(1 - \theta) + (y_4 + 1) \ln(\theta)$$

$$\hat{\theta} = \text{Newton-Raphson Result}$$

$$\theta^* = \text{Newton-Raphson Result}$$

$$\hat{\sigma} = [h''(\theta)]_{\hat{\theta}}^{-1/2} \qquad h(\theta) = \frac{1}{-n} \cdot -nh(\theta)$$

$$\sigma^* = [(h^*)''(\theta)]_{\theta^*}^{-1/2} \quad h^*(\theta) = \frac{1}{-n} \cdot -nh^*(\theta)$$

$$\mathbb{E}_\theta [\theta] = \frac{\sigma^*}{\hat{\sigma}} \cdot \frac{\exp\{-nh^*(\theta^*)\}}{\exp\{-nh(\hat{\theta})\}}$$

Laplace's Method (Second Order) of Posterior Mean for Y=( 125, 18, 20, 34 )

| Statistic | Value |
|---|---|
| theta.hat | 0.62682 |
| theta.star | 0.63099 |
| sigma.hat | 0.72238 |
| sigma.star | 0.71529 |
| Posterior Mean | 0.62275 |

Laplace's Method (Second Order) of Posterior Mean for Y=( 14, 0, 1, 5 )

| Statistic | Value |
|---|---|
| theta.hat | 0.90344 |
| theta.star | 0.91217 |
| sigma.hat | 0.41696 |
| sigma.star | 0.38003 |
| Posterior Mean | 0.82752 |