Homework 3

Akeem Ajede

10/9/2019

Question (1)

Suppose X and Y are random variables with nonzero means μ_X and μ_Y, respectively. Let $g(\mu_X, \mu_Y) = \mu_X/\mu_Y$. Show that: $E\left(\frac{X}{Y}\right) \approx \frac{\mu_X}{\mu_Y}$ and
$Var\left(\frac{X}{Y}\right) \approx \left(\frac{\mu_X}{\mu_Y}\right)^2 \left(\frac{VarX}{\mu_X^2} + \frac{VarY}{\mu_Y^2} - 2\frac{Cov(X,Y)}{\mu_X \mu_Y}\right)$

$$T = \begin{pmatrix} X \\ Y \end{pmatrix} \qquad \theta = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} \qquad g(T) = \frac{X}{Y}$$

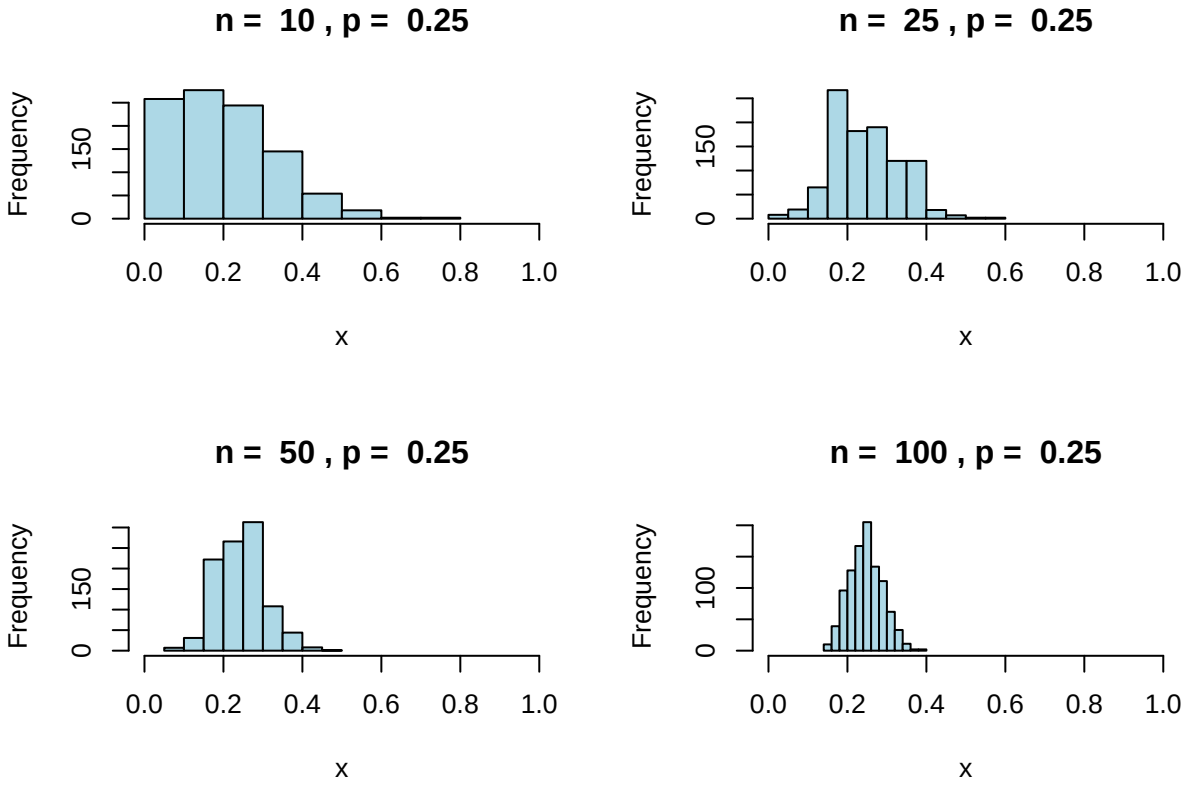
$$\begin{aligned} E\left(\frac{X}{Y}\right) &\approx g(\theta) = g(\mu_X, \mu_Y) \\ &\approx \frac{\mu_X}{\mu_Y} \end{aligned}$$

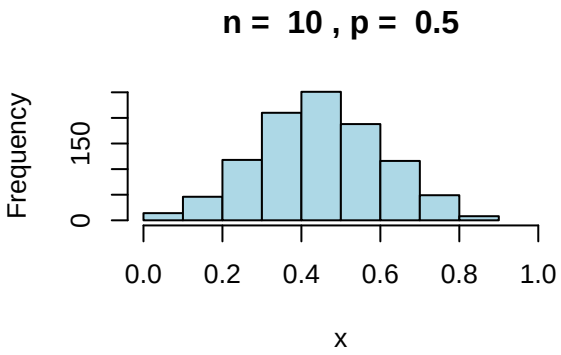
$$\begin{aligned} Var(X/Y) &\approx \left[\frac{\partial g(\theta)}{\partial X}\right]^2 VarX + \left[\frac{\partial g(\theta)}{\partial Y}\right]^2 VarY + 2\left(\frac{\partial g(\theta)}{\partial X}\frac{\partial g(\theta)}{\partial Y} Cov(X,Y)\right) \\ &\approx \left[\frac{1}{\mu_Y}\right]^2 VarX + \left[-\frac{\mu_X}{\mu_Y^2}\right]^2 VarY - 2\left(\frac{1}{\mu_Y}\frac{\mu_X}{\mu_Y^2} Cov(X,Y)\right) \\ &\approx \frac{1}{\mu_Y^2} VarX + \frac{\mu_X^2}{\mu_Y^4} VarY - 2\left(\frac{\mu_X}{\mu_Y^3} Cov(X,Y)\right) \\ &\approx \left(\frac{\mu_X}{\mu_Y}\right)^2 \left(\frac{VarX}{\mu_X^2} + \frac{VarY}{\mu_Y^2} - 2\frac{Cov(X,Y)}{\mu_X \mu_Y}\right) \end{aligned}$$

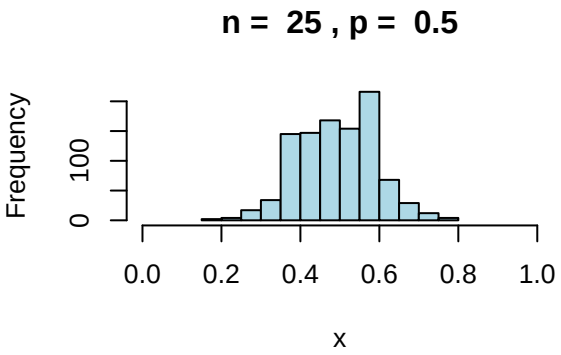
Question (2)

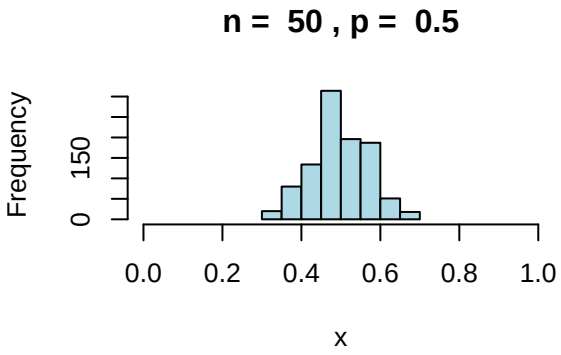
(a)

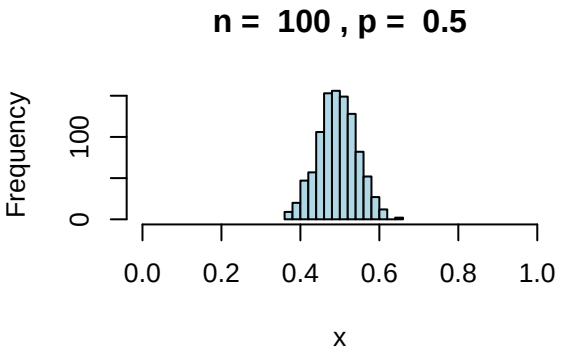
Use R to generate 1000 samples of sizes $n = 10, 25, 50, 100$ from Bernoulli(p) for each of the following values of p; $p = 0.25, 0.5$ and 0.75. For each of the specified values of p, plot the histograms of the sample means based on each of the 4 sample sizes. Plot each of the four on the same page using the $par(mfrow = c(2, 2))$ statement in R.

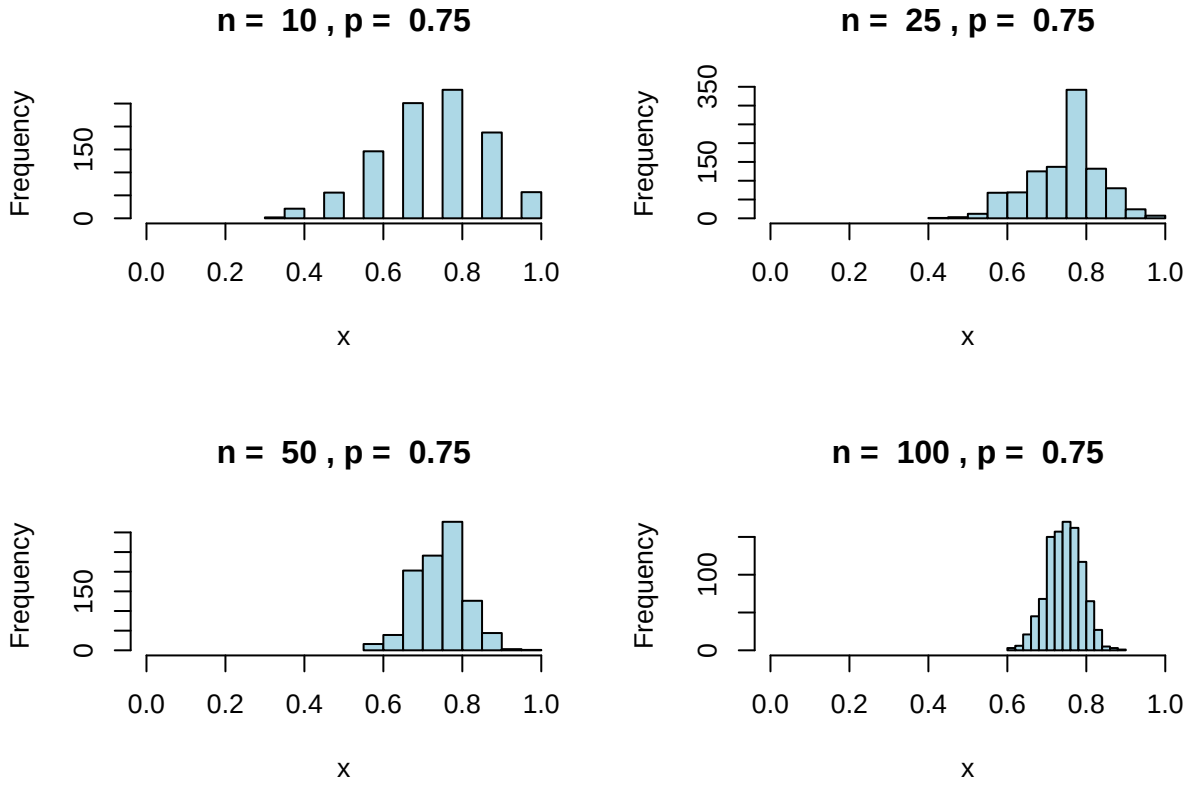



n = 10 , p = 0.5
Frequency
0
150
0.0
0.2
0.4
0.6
0.8
1.0
x


n = 25 , p = 0.5
Frequency
0
100
0.0
0.2
0.4
0.6
0.8
1.0
x


n = 50 , p = 0.5
Frequency
0
150
0.0
0.2
0.4
0.6
0.8
1.0
x


n = 100 , p = 0.5
Frequency
0
100
0.0
0.2
0.4
0.6
0.8
1.0
x



(b)

A data scientist was testing the performance of a classification model using an independent data set containing 100 data points (random sample of size $n = 100$). For each observation in the testing dataset, the predicted value was computed and each point was labeled as a correct classficiation or an incorrrect classification (binary). The results were that 77 out of the 100 were correctly classified. Find a point estimate of the true correct classfication proportion of the predictive model and its associated 95% confidence interval.

Point estimate of the correct classification is given as:

$$\hat{p} = \bar{X} = \frac{77}{100} = 0.77$$

95% C.I:

$$\hat{p} \pm Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \hat{p} \pm 1.96. \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.77 \pm 1.96. \sqrt{\frac{0.77(0.23)}{100}} = (0.688,\ 0.853)$$

(c)

From part (b), compute an estimate of the odds of correct classification and its associated 95% confidence interval.

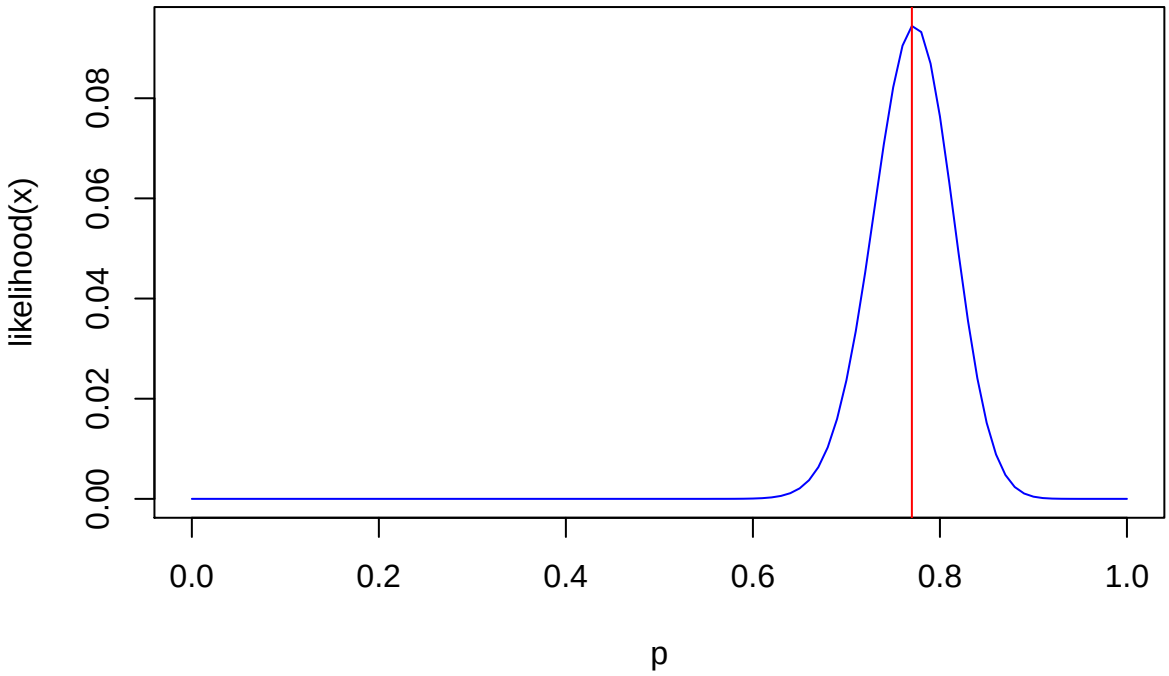
$$Odds\ estimate = \hat{\theta} = \frac{\bar{X}}{1-\bar{X}} = \frac{0.77}{1-0.77} = 3.348$$

95% C.I:

$$\frac{\hat{p}}{1-\hat{p}} \pm Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}}{n(1-\hat{p})^3}} = \frac{\hat{p}}{1-\hat{p}} \pm 1.96\sqrt{\frac{\hat{p}}{n(1-\hat{p})^3}} = 3.348 \pm 1.96.\sqrt{\frac{0.77}{100(0.23)^3}} = (1.789,\ 4.907)$$

(d)

For this sample, n = 100 and x = 77, using R plot the likelihood function, i.e., the likelihood function on the vertical axis and p on the horizontol axis. Draw a vertical line at the value of p where the likelihood function is maximized.



Question (3)

Let $X_{11}, ..., X_{1n_1} \overset{\text{iid}}{\sim}$ Bernoulli (θ_1) and $X_{21}, ..., X_{2n_2} \overset{\text{iid}}{\sim}$ Bernoulli (θ_2), and assume that the two samples are independent.

(a)

Find the maximum likelihood estimator for θ.

$$\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}$$

$$L(\theta_1 |\ X_{11}, ..., X_{1n_1}) = \prod_{i=1}^{n} (P^{x_i}(1-P)^{1-x_i})$$

$$= P^{\sum x_i}(1-P)^{\sum 1-x_i}$$

$$LL(\theta_1 |\ X_{11}, ..., X_{1n_1}) = \sum x_i LogP + \sum (1-x_i) Log(1-P)$$

$$Scorefunction : S(\theta_1 \mid \vec{x}) = \frac{\sum x_i}{P} - \frac{\sum (1-x_i)}{1-P} \overset{\text{set}}{=} 0$$

$$(1-P)\sum x_i - P(\sum (1-x_i) = 0$$

$$P = \frac{1}{n}\sum x_i = \bar{X_1}$$

If the assumption holds,

$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} \bar{X_1} \\ \bar{X_2} \end{pmatrix}$$

(b)

Find the MLE of the odds ratio between population 1 and 2, $\tau = g(\theta) = \frac{\frac{\theta_1}{1-\theta_1}}{\frac{\theta_2}{1-\theta_2}}$

By using the invariance property of MLE; $\hat{\tau} = g(\hat{\theta}) = \frac{\frac{\bar{X_1}}{1-\bar{X_1}}}{\frac{\bar{X_2}}{1-\bar{X_2}}}$

(c)

Let $\hat{\tau}$ be the MLE for the odds ratio and find expressions for approximate mean $\mathrm{E}(\hat{\tau})$ and its approximate variance $Var(\hat{\tau})$.

$$E(\hat{\tau}) = g(\theta) = \frac{\frac{E(\bar{X_1})}{1-E(\bar{X_1})}}{\frac{E(\bar{X_2})}{1-E(\bar{X_2})}}$$

$$Var(\hat{\tau}) \approx [g'_{\theta_1}(\theta)]^2.Var(\bar{X}_1) + [g'_{\theta_2}(\theta)]^2.Var(\bar{X}_2)]$$
$$(cov\ =\ 0,\ for\, independent\, samples)$$
$$\approx \frac{\theta_1(1-\theta_2)}{\theta_2^2(1-\theta_1)}.\left(\frac{1-\theta_2}{n_1(1-\theta_1)} + \frac{\theta_1}{n_2\theta_2}\right)$$

(d)

Specify the approximate distribution for $\hat{\tau}$.

$$\hat{\tau} \overset{\text{d}}{\sim} N\left(E(\hat{\tau}), Var(\hat{\tau})\right)$$

(e)

To determine the approval rating of the mayor of a very large large city, random sample of size 150 adult males was selected and an independent random sample of 150 adult females was selected. From these samples 100 of the females approved of the mayor and 120 of the males approved. Compute a point estimate of the odds ratio, its standard error and a 95% confidence interval (interpret this interval).

$$Odds\ ratio = g(\theta) = \frac{\frac{\theta_1}{1-\theta_1}}{\frac{\theta_2}{1-\theta_2}} = \frac{\frac{4/5}{1/5}}{\frac{2/3}{1/3}} = 2$$

$$Standard\ error = \sqrt{Var(\hat{\tau})} = \left[\frac{4/5(1-(2/3))}{(2/3)^2(1-4/5))}.\left(\frac{1-(2/3)}{150(1-(4/5))} + \frac{4/5}{150(2/3)}\right)\right]^{0.5} = (0.057\bar{3})^{0.5} = 0.239$$

$$\frac{\frac{\theta_1}{1-\theta_1}}{\frac{\theta_2}{1-\theta_2}} \pm Z_{\frac{\alpha}{2}}.\sqrt{\frac{\theta_1(1-\theta_2)}{\theta_2^2(1-\theta_1)}\left(\frac{1-\theta_2}{n_1(1-\theta_1)} + \frac{\theta_1}{n_2\theta_2}\right)} = 2 \pm (1.96 \times 0.239) = (1.532,\ 2.468)$$

The 95% C.I of the point estimate of the odds ratio suggests that the mayor has a positive approval rating (i.e., C.I > 0) across male and female adults in the city.