Assignment 13

Rahul Ramachandran

June 10, 2022

Outline

Problem Statement

Papoulis 8.10

Among 4000 newborns, 2080 are male. Find the 0.99 confidence interval of the probability $p = P(\text{male})$

Definitions

- A **Confidence Interval** is an estimate for an unknown parameter. It is governed by a number $\gamma = 1 - \delta$, which determines the accuracy of the estimation method. γ is called the **confidence coefficient**.
- A ***z*-score** is the number of standard deviations a data point is away from the mean.
- If $\hat{p}$ is a sample mean, the distribution of $\hat{p}$ is called the **sampling distribution of the sample means**.
- To find the confidence interval, we find the points on the $x-$axis between which the area under the curve equals γA

Binomial Distribution Mean

Let the random variable X be the sum of n Bernoulli random variables, i.e.

$$X = X_1 + X_2 + \ldots + X_n \tag{1}$$

For simplicity, assume $E(X_i) = p$. Then,

$$\begin{align}
E(X) &= E(X_1 + X_2 + \ldots + X_n) \tag{2}\\
&= E(X_1) + E(X_2) + \ldots + E(X_n) \tag{3}\\
&= p + p + \ldots + p \tag{4}\\
&= np \tag{5}
\end{align}$$

Binomial Distribution Variance

The variance is given by:

$$
\begin{aligned}
E((X-np)^2) &= E(X^2 - 2Xnp + n^2p^2) && (6)\\
&= E(X^2) - 2npE(X) + E(n^2p^2) && (7)\\
&= E(X^2) - n^2p^2 && (8)
\end{aligned}
$$

Now,

$$
\begin{aligned}
E(X^2) &= \sum_{k=0}^{n} k^2 \times \binom{n}{k} p^k q^{n-k} && (9)\\
&= \sum_{k=1}^{n} npk \times \binom{n-1}{k-1} p^{k-1} q^{n-k} && (10)
\end{aligned}
$$

Binomial Distribution Variance

Since $k = (k-1) + 1$, we have:

$$\begin{aligned}\sum_{k=1}^{n} npk \times \binom{n-1}{k-1} p^{k-1} q^{n-k} &= n(n-1)p^2 \times \sum_{k=2}^{n} \binom{n-2}{k-2} p^{k-2} q^{n-k} \\ &\quad + np \times \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} q^{n-k} && (11) \\ &= n(n-1)p^2 (p+q)^{n-2} + np(p+q)^{n-1} && (12) \\ &= n(n-1)p^2 + np && (13)\end{aligned}$$

Substituting this in Equation 8, we have:

$$\begin{aligned}\sigma^2 &= (n(n-1)p^2 + np) - n^2 p^2 && (14) \\ &= np - np^2 && (15)\end{aligned}$$

Binomial Distribution

Therefore:

$$\sigma^2 = np - np^2 \tag{16}$$
$$= np(1-p) \tag{17}$$
$$= npq \tag{18}$$

Let the random variable $Y = \frac{X}{n}$. Then, Y is a binomial random variable that maps to $\frac{k}{n}$ when X maps to k. Therefore,

$$\mu_Y = \frac{\mu_X}{n} \tag{19}$$
$$= p \tag{20}$$
$$\sigma_Y^2 = \frac{\sigma_X^2}{n^2} \tag{21}$$
$$= \frac{pq}{n} \tag{22}$$

De Moivre-Laplace Theorem

The **De Moivre-Laplace Theorem** states that for large *n* (and a significant value of *np*), the Normal Distribution can be used to approximate a Binomial Distribution.
Let the point *k* in a Binomial distribution equal $\mu + c\sigma$, where *c* is an arbitrary constant. Therefore, from Equation 5 and Equation 18:

$$k = np + c\sqrt{npq} \tag{23}$$

The differential equation for a Normal Distribution is:

$$\frac{f'(x)}{f(x)}\left(-\frac{\sigma^2}{x-\mu}\right) = 1 \tag{24}$$

De Moivre-Laplace Theorem

If we show that the pdf of the Binomial Distribution satisfies Equation 24, we have proved the theorem. Since the Binomial Distribution is discrete, we consider the discrete analogue of the derivative, $p(k+1) - p(k)$.

$$\frac{f'(x)}{f(x)}\left(-\frac{\sigma^2}{x-\mu}\right) = \frac{p(k+1)-p(k)}{p(k)}\left(\frac{\sqrt{npq}}{-c}\right) \tag{25}$$

$$= \frac{\binom{n}{k+1}p^{k+1}q^{n-(k+1)} + \binom{n}{k}p^k q^{n-k}}{\binom{n}{k}p^k q^{n-k}}\left(\frac{\sqrt{npq}}{-c}\right) \tag{26}$$

$$= \left(\frac{(n-k)p}{(k+1)q} - 1\right)\left(\frac{\sqrt{npq}}{-c}\right) \tag{27}$$

$$= \frac{np-k-q}{(k+1)q}\left(\frac{\sqrt{npq}}{-c}\right) \tag{28}$$

De Moivre-Laplace Theorem

Substituting for k, we get:

$$\frac{f'(x)}{f(x)}\left(-\frac{\sigma^2}{x-\mu}\right) = \frac{-c\sqrt{npq}-q}{npq+cq\sqrt{npq}+q}\left(\frac{\sqrt{npq}}{-c}\right) \tag{29}$$

As $n \to \infty$

$$\frac{-c\sqrt{npq}-q}{npq+cq\sqrt{npq}+q}\left(\frac{\sqrt{npq}}{-c}\right) \to 1 \tag{30}$$

Therefore, the De Moivre-Laplace Theorem holds.

Confidence Interval

Consider a Normal Distribution of sample means symmetric about $x = p$. Let the number z be chosen such that the area between $x = -\infty$ and $x = \mu + z\sigma = A(1 - \frac{\delta}{2})$, where A is the area under the whole curve. Then:

$$Ar(\mu - z\sigma \leq x \leq \mu + z\sigma) = A(1 - \delta) \tag{31}$$

$$\implies \Pr\left(p - z\sigma \leq \hat{p} \leq p + z\sigma\right) = 1 - \delta \tag{32}$$

Analogously, for a Normal Distribution symmetric about $\hat{p}$, with variance σ^2,

$$\Pr\left(\hat{p} - z\sigma \leq p \leq \hat{p} + z\sigma\right) = 1 - \delta \tag{33}$$

Confidence Interval

The interval $[\hat{p} - z\sigma, \hat{p} + z\sigma]$ is called the γ confidence interval of $\hat{p}$. Often, we aren't aware of the exact value of p, so we approximate it as under:

$$\sigma = \sqrt{\frac{pq}{n}} \tag{34}$$

$$\approx \sqrt{\frac{\hat{p}\hat{q}}{n}} \tag{35}$$

Since all the values are now known, we can find an estimate for the confidence interval of $\hat{p}$

Solution

Since there are 4000 babies and 2080 girls, for the sample, we have:

$$\hat{p} = 2080/4000 \tag{36}$$
$$= 0.52 \tag{37}$$

Also,

$$\sigma_{\hat{p}} = \sqrt{\frac{0.52(1-0.52)}{4000}} \tag{38}$$
$$= 0.0079 \tag{39}$$

Solution

Since we are interested in the 0.99 confidence interval, we have

$$\gamma = 0.99 \tag{40}$$
$$\implies \delta = 1 - \gamma \tag{41}$$
$$= 0.01 \tag{42}$$

Therefore, we have to find z corresponding to $\delta = 0.01$, which from the z-score table equals 2.58

Solution

Therefore, it follows that

$$p_u = \mu + z\sigma \tag{43}$$
$$= 0.52 + 2.58 \times 0.0079 \tag{44}$$
$$= 0.54 \tag{45}$$

where p_u is the upper limit of the interval. Similarly,

$$p_l = \mu - z\sigma \tag{46}$$
$$= 0.52 - 2.58 \times 0.0079 \tag{47}$$
$$= 0.49 \tag{48}$$

Therefore, the 0.99 confidence interval for p is $[0.49, 0.54]$