BMI 713: Computational Statistics for Biomedical Sciences

# Assignment 5 – Solutions

## 1. Sampling Distribution of Sample Proportions

(a). Suppose we have a random sample of size 100 from a Binomial distribution with the population proportion of 0.3. What are the expected mean $E(\hat{p})$ and variance $Var(\hat{p})$ of the sample proportion $\hat{p}$? What is the sampling distribution of the sample proportion $\hat{p}$?

**Sol'n.** *The expected mean is* $E(\hat{p}) = p = 0.3$, *and the variance is* $Var(\hat{p}) = \frac{pq}{n} = \frac{0.3 \times (1-0.3)}{100} = 0.0021$.

*Check the adequacy of normal approximation,* $npq = 100 \times 0.3 \times (1-0.3) = 21 > 5$ *(n is large enough).*

*The sampling distribution of* $\hat{p}$ *is approximately normal with mean 0.3 and variance 0.0021.*

*The exact distribution of* $100 \cdot \hat{p}$ *(the number of "successes") is Binomial (100, 0.3).*

(b). Generate 1000 samples of size 100 from a Binomial distribution with population proportion of 0.3, calculate the sample proportion for each sample, and save them in the vector `sample.p`.

**Sol'n.** *We can generate the 1000 sample proportions in R:*

```
sample.p <- rbinom(1000, 100, 0.3) / 100
```

(c). Calculate the mean and variance of the 1000 sample proportions, and compare to the results in (a).

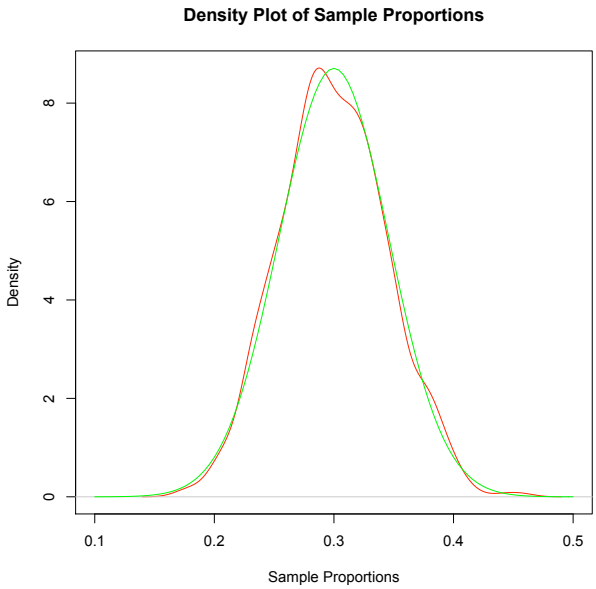
**Sol'n.** *Calculate the mean and variance of the 1000 sample proportions in R:*

```
mean(sample.p)
var(sample.p)
```

*You may get different answers given different simulations, but the mean and variance of the 1000 sample proportions should be very close to the expected mean and variance in (a), respectively.*

(d). Plot the density of the sample proportions, and compare it to the curve of Normal distribution with mean `E.p` and variance `V.p`, where `E.p` is the expected mean $E(\hat{p})$ and `V.p` is the variance $Var(\hat{p})$ from (a), using the following commands:

```
plot(density(sample.p), main="Density Plot of Sample Proportions", xlab="Sample
Proportions", xlim=c(0.1, 0.5), col="red")
curve(dnorm(x, mean=E.p, sd=sqrt(V.p)), from=0.1, to=0.5, add=TRUE, col="green")
```




**Sol'n.** *From the plot, we can see that the shape of the density of simulated sample proportions matches very well with the shape of the normal distribution with mean p and variance pq/n. The sampling distribution of the sample proportion is approximately normal.*

**2. Hypothesis Testing**

Alcohol and breast cancer: The following are partial results from a *case-control* study involving a sample of *cases* (women with breast cancer) and a sample of *controls* (demographically similar women without breast cancer). (Data from L. Rosenberg, *et. al.*, A Case-Control Study of Alcoholic Beverage Consumption and Breast Cancer, *American Journal of Epidemiology* 131 (1990): 6-14). Are the occurrences of women breast cancer related to drinking habits?

| | Breast Cancer | |
|---|---|---|
| | Cases | Controls |
| Fewer than 4 drinks per week | 330 | 658 |
| 4 or more drinks per week | 204 | 386 |

(a). State the null hypothesis and the alternative hypothesis.

**Sol'n.** *If $p_1$ denotes the proportion of breast cancer among women who have fewer than 4 drinks per week, and $p_2$ denotes the proportion of breast cancer among women who have 4 or more drinks per week, then the null hypothesis is $H_0: p_1 = p_2$, and the alternative hypothesis is $H_1: p_1 \neq p_2$.*

(b). Compute the sample proportions of women with breast cancer among women who have fewer than 4 drinks per week and who have 4 or more drinks per week, $\hat{p}_1$ and $\hat{p}_2$ respectively.

**Sol'n.** *The sample proportion of breast cancer among women who have fewer than 4 drinks per week is*

$$\hat{p}_1 = \frac{330}{330+658} \approx 0.3340,$$

*and the sample proportion of breast cancer among women who have 4 or more drinks per week is*

$$\hat{p}_1 = \frac{204}{204+386} \approx 0.3458.$$

(c). How large should the sample sizes be for the adequacy of the normal approximation for the sampling distribution of the sample proportions?

**Sol'n.** *Denote the sample size of women who have fewer than 4 drinks per week by $n_1$, and the sample size of women who have 4 or more drinks per week by $n_2$.*

*For normal approximation, it is required that $n_1\hat{p}_1(1-\hat{p}_1) \geq 5$ and $n_2\hat{p}_2(1-\hat{p}_2) \geq 5$,*

*so, $n_1$ should be equal to or greater than $\frac{5}{\hat{p}_1(1-\hat{p}_1)} = \frac{5}{0.3340 \times (1-0.3340)} \approx 23$,*

*and $n_2$ should be equal to or greater than $\frac{5}{\hat{p}_2(1-\hat{p}_2)} = \frac{5}{0.3458 \times (1-0.3458)} \approx 23$.*

*Here $n_1 = 330+658 = 988$, and $n_2 = 204+386 = 590$, which are large enough for the adequacy of normal approximation.*

(d). What test do you need to perform for the hypothesis testing? What is the value of the test statistic? What is the *p-value*? What conclusion can you make?

**Sol'n.** *We can perform Z-test to test if the two population proportions are equal.*

*The combined sample proportion is*

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1+n_2} = \frac{988 \times 0.3340 + 590 \times 0.3458}{988+590} \approx 0.3384.$$

*The z-statistic is*

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{0.3340 - 0.3458}{\sqrt{0.3384 \times (1-0.3384)\left(\frac{1}{988}+\frac{1}{590}\right)}} \approx -0.48$$

*Under the null hypothesis $H_0$, the distribution of test statistic Z is approximately standard normal. The two-sided p-value is then*

$$P(|Z| > |z|) = 2 \times P(Z < -0.48) \approx 0.63.$$

*We can also use the function* `prop.test` *in R to perform the test:*

```
> prop.test(c(330, 204), c(988, 590), correct=FALSE)
       2-sample test for equality of proportions without continuity correction
data:  c(330, 204) out of c(988, 590)
X-squared = 0.228, df = 1, p-value = 0.633
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.06010485  0.03659562
sample estimates:
   prop 1    prop 2
0.3340081 0.3457627
```

*There is no significant evidence that the two proportions are different, so we fail to reject the null hypothesis.*

**3. Confidence Interval**

(a). Given a sample of size $n$ from a population with proportion $p$ (unknown), and the estimated sample proportion $\hat{p}$ (known), write a function `CI(n, p.hat, a)` in R to calculate the confidence interval (CI) for $p$, at the significance level of $\alpha$. The function should return a vector of two elements: the first is the lower bound of the CI, and the second is the upper bound of the CI. Note: for this problem, please do not use the function `prop.test` or `binom.test`.

**Sol'n.** *Write a function to calculate the confidence interval in R:*

```
CI <- function(n, p.hat, a) {
       CI.lower <- p.hat - qnorm(1-a/2) * sqrt(p.hat*(1-p.hat)/n)
       CI.upper <- p.hat + qnorm(1-a/2) * sqrt(p.hat*(1-p.hat)/n)
       return(c(CI.lower, CI.upper))
       }
```

(b). Generate 1000 samples of size 50 from a Binomial distribution with proportion of 0.75, and use the `CI` function in (a) to derive the 95% confidence interval of the population proportion $p$ for each sample, based on the estimated sample proportion $\hat{p}$ (pretending that $p$ is unknown). How many times out of the 1000 simulations do the confidence intervals contain the true population proportion $p$ (i.e., 0.75)? How would you interpret confidence interval based on this problem?

**Sol'n.** *Simulation in R:*

```
p <- 0.75
count <- 0
for (i in 1:1000) {
       x <- rbinom(1, 50, p)
       p.hat <- x/50
       conf.int <- CI(50, p.hat, 0.05)
       if (conf.int[1] < p & p < conf.int[2]) count <- count + 1
       }
> count
[1] 937
```

*The number of confidence intervals that contain the true population proportion should be close to 950.*

*A (1-$\alpha$)% confidence interval for a population parameter, $\theta$, is a <u>random interval</u>, calculated from the sample, that contains $\theta$ with probability (1-$\alpha$). For example, here, we draw 1000 random samples and form a 95% confidence interval of population proportion p from each sample, and about 95% of these intervals contain p.*