Project 2: Simulation and p value

STAT35000, Fall 2016

Due: November 15 (Monday), 2016

Abstract

This is the 2nd R project. This problem is about sampling distribution by using `R` and calculation of the so called p value.

For the Bernoulli distribution $X \sim \text{Ber}(p) = \text{Bin}(1, p)$, figure out the following problems:

1. Sample 50 points from $X \sim \text{Ber}(p_0 = 0.2)$.
2. Calculate the sample proportion $\hat{p}^*$ of "1"s shown in the above sample you obtained.
3. Do you think it's reasonably good enough? Explain why.
4. Pretend that you forgot the probability of success p_0 you used to generate the above sample of size 50. Your guess now is 0.4. And you want to figure out whether it is 0.4. Here are two ways for your choice, and which one do you think is more reasonable?
   (a) Compare the sample proportion you got in above part 2 with your guess 0.4. If they are reasonably close, you probably will adopt your guess 0.4. Think about how could you judge the closeness.
   (b) Here is the other procedure. **The logic behind this is that if $p_0 = 0.4$, what will happen? If something strange happened, then you should doubt your guess; if everything happend is reasonable under $p_0 = 0.4$, then it seems no reason for you to reject your guess.** Here is the implementation. Use $p_0 = 0.4$ to generate $N = 10000$ samples from $X \sim \text{Ber}(p_0 = 0.4)$ with sample size $n = 50$. And calculate the sample proportion for each of these $N = 10000$ samples, denoted by $\{\hat{p}_{50}^k, k = 1, 2 \cdots, N = 10000\}$. And then plot the histogram of $\{\hat{p}_{50}^k, k = 1, 2 \cdots, N = 10000\}$ to see the distribution of the random variable sample proportion $\hat{p}_{50}$ under the

assumption that the true $p_0 = 0.4$. And if the observed $\hat{p}^*$ is not in the extreme region of the distribution, you probably will adopt your guess 0.4. In particular, calculate the probability that $\hat{p}_{50} < \hat{p}^*$ through the frequency of $\{\hat{p}_{50}^k < \hat{p}^*, k = 1, 2 \cdots, N = 10000\}$. This probability is actually related with the important concept in Statistics, **p value**!

5. For the above part (b), we are actually using simulation to approximate the probability of $\hat{p}_{50} = \frac{X_1+X_2+\cdots+X_{50}}{50} < \hat{p}^*$ given that $\{X_i, i = 1, 2, \cdots, 50\}$ are independent and identically distributed as $X \sim \text{Ber}(p_0 = 0.4)$. Could you figure the probability out exactly without any approximation? What is the exact probability?

6. For the above probability, we are actually also be able to approximate it without simulation. Remember by the central limit theorem for $n = 50 > 30$,
$$\hat{p}_{50} = \frac{X_1 + X_2 + \cdots + X_{50}}{50} \overset{approximate}{\sim} N(\mu, \sigma^2/50)$$
with $\mu = E(X)$ and $\sigma^2 = Var(X)$ for $X \sim \text{Ber}(p_0 = 0.4)$. Use this approximation fact, please calculate the probability.

7. Compare the p values you obtained by the above three ways (simulation approximation, exact, CLT approximation), you should expect to see that the CLT approximation is as good as the simulation approximation. There is some empirical continuity correction about this CLT approximation. Please check out the online material
https://people.richland.edu/james/lecture/m170/ch07-bin.html
to figure out how to conduct the correction to make the approximation better. Calculate the corrected probability.