

Project 2: Simulation and Statistical Inference

STAT35000, Instructor: Honglang Wang

Abstract

There's an ongoing debate in the academic community about whether Calculus is a necessary pre-requisite for Statistics. But in age of ubiquitous computing resources (not to mention open source programming languages like R), there's a fair argument to be made that all you really need is simulation, which is a numerical technique for conducting experiments on the computer involving random sampling from probability distributions. However complex the statistical proposition, you can always find useful information about its properties simply by generating a bunch of random numbers and seeing what happens. In this project, we make use of simulation to learn statistical inference procedures.

Instruction: Type your solutions to each question in the project. For the problems involving R coding, please insert the code as well as the output (including numerical output and graphs) in your solution document. And try to explain the R output to answer the questions.

John and Jack found a coin on the sidewalk. They argued about the fairness of the coin. John claimed 40% to have Heads according to his careful observation of the coin. Jack doubted and in order to infer the fairness of the coin, he tossed the coin for 50 times and got the results as shown below with 1 representing as heads and 0 as tails:

```
## [1] 0 0 0 1 1 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0
## [36] 0 1 0 0 0 0 0 1 1 0 0 0 0 0 0
```

Let p^* as the true probability to have heads for the coin. Note that p^* is a characteristic of the coin, and we want to make some inference about this unknown parameter. And denote X as the random variable which takes 0 if tails show up or 1 if heads show up for tossing the coin.

1. (10pts) What is the distribution of the random variable X ?
2. (5pts) Calculate the sample proportion $\hat{p}^*$ which is the proportion of “1”s shown in the above given sample. (*Note that this hat version is your “best” guess for the unknown p^* based on the sample. These two notations, p^* and $\hat{p}^*$, are indeed different.*)

3. How could we help Jack make use of the data given above to show some evidence for/against John's guess? One proposal here is to use proof by contradiction: **what if John's claim is true?** If John's claim is true, then we have a virtual coin with known fairness in your computer. And you can make use of simulation to figure out something hopefully in contradiction with John's guess. Here is how:

- (i) (10pts) Use simulation to flip this virtual coin with 40% to get Heads for 50 times again and again for $N = 1000$ repetition:

```
virtual = replicate(1000, rbinom(50, 1, 0.4))
```

Why this: If $p^* = 0.4$ according to John's claim, then all of these 1000 samples should be comparable to the given sample Jack obtained since we assumed they are all from the same distribution $\text{Ber}(0.4)$. Next we will use one way to check whether they are indeed comparable.

- (ii) (15pts) Calculate the sample proportion for each of these $N = 1000$ samples, denoted by $\{\hat{p}_{50}^k, k = 1, 2, \dots, N = 1000\}$:

```
phat50s = colSums(virtual)/50
```

Why this: If $p^* = 0.4$ according to John's claim, then $\hat{p}^*$ should be comparable to most of the $\{\hat{p}_{50}^k, k = 1, 2, \dots, N = 1000\}$. (*Do not get confused about the notations: p^* is the unknown true fairness of the coin, $\hat{p}^*$ is the sample proportion for Jack's particular sample. $\{\hat{p}_{50}^k, k = 1, 2, \dots, N = 1000\}$ are the sample proportions from the simulation for the virtual coin.*)

- (iii) (15pts) Plot the histogram of $\{\hat{p}_{50}^k, k = 1, 2, \dots, N = 1000\}$ and label where the observed $\hat{p}^*$ in the histogram:

```
hist(phat50s, prob = TRUE)
points(phatstar, 0, col = "red", pch = 20)
```

Why this: Visually check out whether $\hat{p}^*$ is comparable to most of the $\{\hat{p}_{50}^k, k = 1, 2, \dots, N = 1000\}$. If it is in the extreme region of the histogram, you probably will conclude that it's not comparable so that Jack could use this graph to argue against John's claim.

- (iv) (15pts) To be precise, calculate the relative frequency that $\hat{p}^*$ is larger than those 1000 $\hat{p}_{50}^k$'s:

```
mean(phat50s < phatstar)
```

Why this: This relative frequency is a measurement for how extreme of $\hat{p}^*$ located among the 1000 $\hat{p}_{50}^k$'s. This relative frequency is actually related with the important concept in Statistics, **p-value!** (*Do not get confused with the p in the notation p -value and other p 's above.*)

- (v) (10pts) John will accept Jack's objection if the p-value is smaller than 5%, otherwise Jack has to accept John's guess. Based on this rule, what is the conclusion based on your above analysis?
4. For the calculation of the above relative frequency, we are actually using simulation to approximate the probability of $\hat{p}_{50} = \frac{X_1 + X_2 + \dots + X_{50}}{50} < \hat{p}^*$ given that $\{X_i, i = 1, 2, \dots, 50\}$ are independent and identically distributed as $X \sim \text{Ber}(0.4)$. We could use the following theoretical ways to re-calculate it. (*Note that $\hat{p}_{50}$ is the random variable notation from which we got the above 1000 observations $\hat{p}_{50}^k$. Again $\hat{p}^*$ is just a particular value as the sample proportion for Jack's particular sample.*)
- (a) (10pts) Use $\Pr(\hat{p}_{50} < \hat{p}^*) = \Pr(X_1 + X_2 + \dots + X_{50} < 50\hat{p}^*)$ to calculate the probability. *Hint: $X_1 + X_2 + \dots + X_{50}$ follows Binomial distribution.*
- (b) (10pts) Use central limit theorem for $\hat{p}_{50} = \frac{X_1 + X_2 + \dots + X_{50}}{50}$ to approximate the probability.
5. (**bonus:** 10pts) If John kept guessing with different proportions, we could repeat the above simulation procedure to reject or accept the guesses: reject the guess if the corresponding p-value is less than 5%. Please try out the sequence of guesses: $\{0, 0.01, 0.02, \dots, 0.99, 1\}$ to get the maximum and minimum guesses in this sequence which are acceptable.
6. (**bonus:** 10pts) Intuitively, if the guess, say p_0 , differs a lot with Jack's sample proportion $\hat{p}^*$, then the guess is probably not acceptable. Based on this intuition, let's propose the following rejection rule: reject the guess p_0 if $|\hat{p}^* - p_0| > 2\sqrt{p_0(1-p_0)/n}$. Please try out the sequence: $p_0 \in \{0, 0.01, 0.02, \dots, 0.99, 1\}$ to get the maximum and minimum p_0 's which are acceptable by using this rule.

Note that we used a lot different notations related with p . As long as they are different, they represent different things. Do not get confused about these proportion notations.