The Normal Model for Sample Averages

Ryan Miller

Introduction

- Last presentation introduced the *sample average* as a random variable of interest to statisticians
  - As you may have anticipated, this is because the *distribution of sample averages* has been proven to follow a Normal distribution under certain conditions

Introduction

- Last presentation introduced the *sample average* as a random variable of interest to statisticians
  - As you may have anticipated, this is because the *distribution of sample averages* has been proven to follow a Normal distribution under certain conditions
- At first this might not seem very special, but it turns out the relatively few things have sampling distributions that are this well-understood!

John Kerrich

- John Kerrich, a South African mathematician, was visiting Copenhagen in 1940
- When Germany invaded Denmark he was sent to an internment camp, where he spend the next five years
- To pass time, Kerrich conducted experiments exploring sampling and probability theory
  - One of these experiments involved flipping a coin 10,000 times

Kerrich's Experiment and Probability

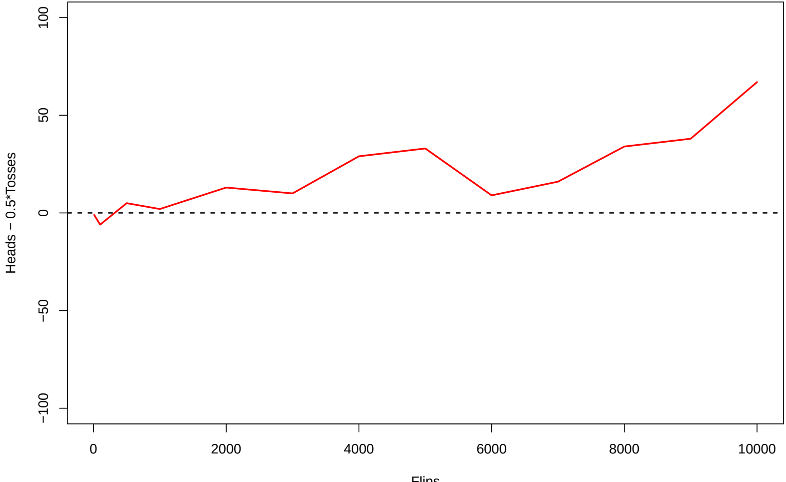
- We know that a fair coin shows "Heads" with a probability of 50%
- So, in a random sample of n coin flips, we'd expect roughly even numbers of "Heads" and "Tails"
  - We'll explore the results of Kerrich's experiment to see why the *sample average* is so special

Kerrich's Results

Number of Tosses (n)	Number of Heads	Heads - 0.5*Tosses
10	4	-1
100	44	-6
500	255	5
1,000	502	2
2,000	1,013	13
3,000	1,510	10
4,000	2,029	29
5,000	2,533	33
6,000	3,009	9
7,000	3,516	16
8,000	4,034	34
9,000	4,538	38
10,000	5,067	67

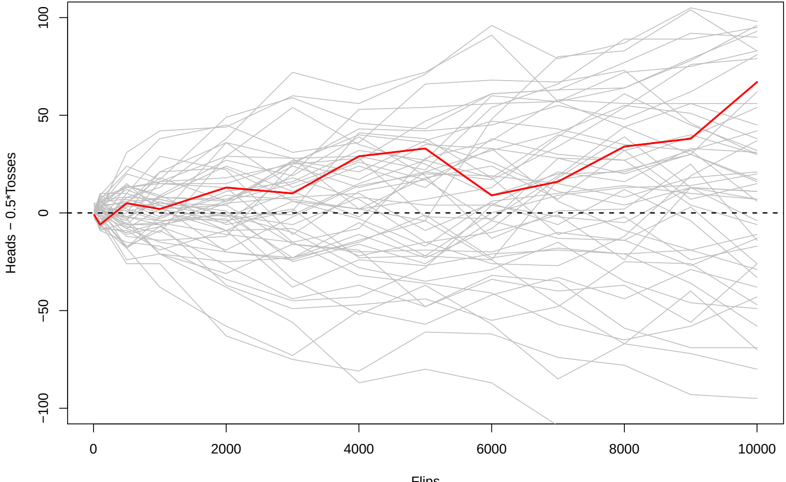
Kerrich's Results

It seems like the number of heads and tails are actually getting further apart... could this be a fluke?



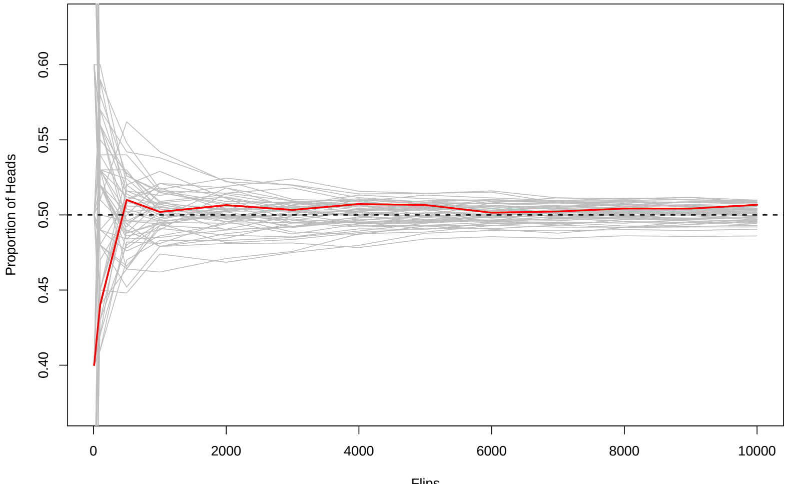
Kerrich's Experiment (repeated 50 times)

No, the phenomenon occurs systematically when repeating Kerrich's experiment



Kerrich's Experiment (sample proportions)

The *sample proportion* of heads behaves exactly as we'd expect, but why?



Central Limit Theorem

- Suppose $X_1, X_2, \dots, X_n$ are independent random variables with a common expected value $E(X)$ and variance $Var(X)$ (see previous notes for definitions of these two terms)
- Let $\bar{X}$ denote the average of all n random variables, **Central Limit Theorem** (CLT) states:

$$\sqrt{n}\left(\frac{\bar{X} - E(X)}{\sqrt{Var(X)}}\right) \to N(0, 1)$$

Central Limit Theorem

- Suppose $X_1, X_2, \ldots, X_n$ are independent random variables with a common expected value $E(X)$ and variance $Var(X)$ (see previous notes for definitions of these two terms)
- Let $\bar{X}$ denote the average of all n random variables, **Central Limit Theorem** (CLT) states:

$$\sqrt{n}\left(\frac{\bar{X} - E(X)}{\sqrt{Var(X)}}\right) \rightarrow N(0,1)$$

- Often it is more useful to think of CLT in the following way (which abuses notation):

$$\bar{X} \sim N\left(E(X), \frac{SD(X)}{\sqrt{n}}\right)$$

Central Limit Theorem and Sample Proportions

- The sample proportion is comprised of n different binary variables (taking on values of 1 and 0)
  - Each one of these binary variables has the same expected value and variance

Central Limit Theorem and Sample Proportions

- The sample proportion is comprised of n different binary variables (taking on values of 1 and 0)
  - Each one of these binary variables has the same expected value and variance
  - $E(X) = p * 1 + 0 * (1 - p) = p$
  - $Var(X) = p * (1 - p)^2 + (1 - p) * (0 - p)^2 = p * (1 - p)$

Central Limit Theorem and Sample Proportions

- The sample proportion is comprised of n different binary variables (taking on values of 1 and 0)
  - Each one of these binary variables has the same expected value and variance
  - $E(X) = p * 1 + 0 * (1 - p) = p$
  - $Var(X) = p * (1 - p)^2 + (1 - p) * (0 - p)^2 = p * (1 - p)$
- Thus, the *sampling distribution* of sample proportions is:

$$\hat{p} \sim N(p, \sqrt{p(1 - p)/n})$$

The Power of CLT

- Central Limit Theorem is one of the most important theoretical results in all of statistics
- In real-world applications, it is nearly impossible to know the probability distribution of something that is only observed once (remember that real researchers can only afford to collect a single sample)

The Power of CLT

- Central Limit Theorem is one of the most important theoretical results in all of statistics
- In real-world applications, it is nearly impossible to know the probability distribution of something that is only observed once (remember that real researchers can only afford to collect a single sample)
- But by focusing on the *sample average* this isn't an issue, as CLT provides us the distribution of sample averages
  - That is, we are able to use CLT to understand the *sampling variability* of our study, despite only getting to see a single sample!

Example

- Let's consider a random sample of $n = 100$ coin flips
  - What proportion of heads might we expect? It'll likely be close to 50%, but we know there's sampling variability, the question is how much...

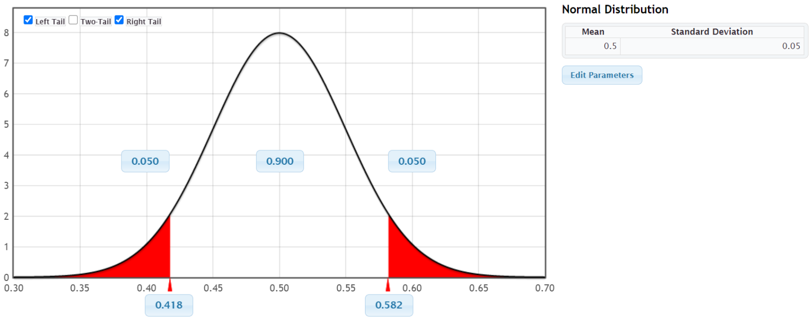
Example

- Let's consider a random sample of $n = 100$ coin flips
  - What proportion of heads might we expect? It'll likely be close to 50%, but we know there's sampling variability, the question is how much...
- Each coin flip is a random variable an expected value of 0.5, so Central Limit Theorem tells us that proportion of heads in random samples of $n = 100$ coin flips follows a Normal distribution:

$$\hat{p} \sim N(0.5, \sqrt{0.5(1-0.5)/100})$$

- To understand the sampling variability of $n = 100$ coin flips, we might look at the *interval* that defines what we'd expect to see 90% of the time

Example



- We'd expect 90% of different random samples to result in sample proportions between 0.418 and 0.582

Assumptions

Using the Central Limit theorem to determine the distribution of sample averages is only appropriate when the following conditions are met:

1) Independence - the cases in the sample (ie: the individual contributions to the sample average) are not related to each other
2) Large population - less that 10% of the population is being sampled (otherwise removing the already sampled individuals has too much of an impact on the probability of selection)
3) Large sample - $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$

In most applications, only the third condition is problematic

Conclusion

- Central Limit theorem provides a theoretical basis for focusing on sample averages when attempting to describe the characteristics of a population
  - Put differently, CLT allows us to understand the sampling variability of the sample average

Conclusion

- Central Limit theorem provides a theoretical basis for focusing on sample averages when attempting to describe the characteristics of a population
  - Put differently, CLT allows us to understand the sampling variability of the sample average
- In the next video, we'll approach the task of *estimation* in much greater detail