Central Limit Theorem

Ryan Miller

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 probability
    - One of these experiments involved flipping a coin 10,000 times

Kerrich's Experiment and Probability

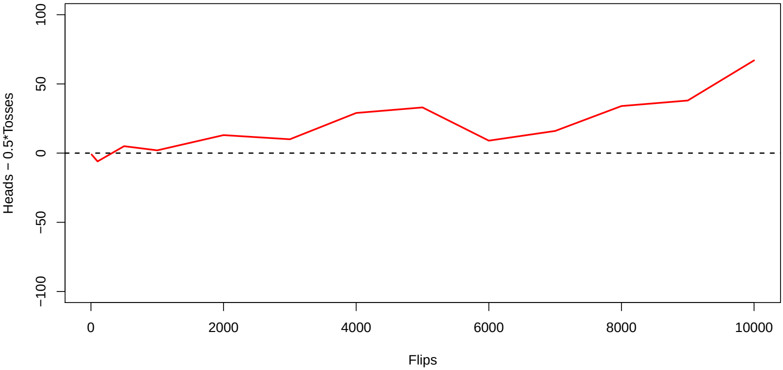
- We know that a fair coin shows "Heads" with a probability of 50%
- So, if we flip a coin many times you might expect roughly even numbers of "Heads" and "Tails"
  - We'll explore the results of Kerrich's experiment to understand more precisely what probability theory tells about flipping a coin 10,000 times

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

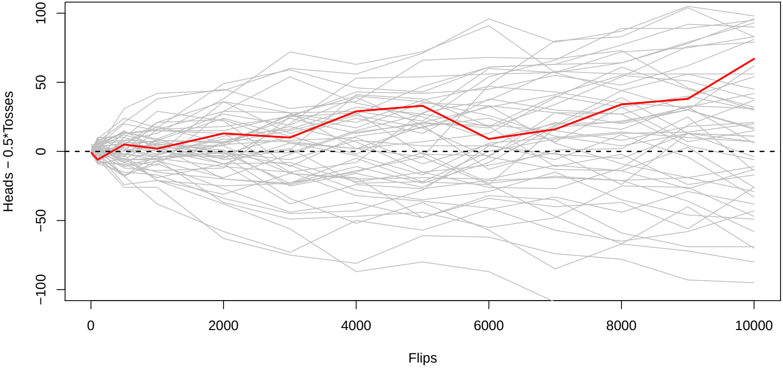
Kerrich's Results

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

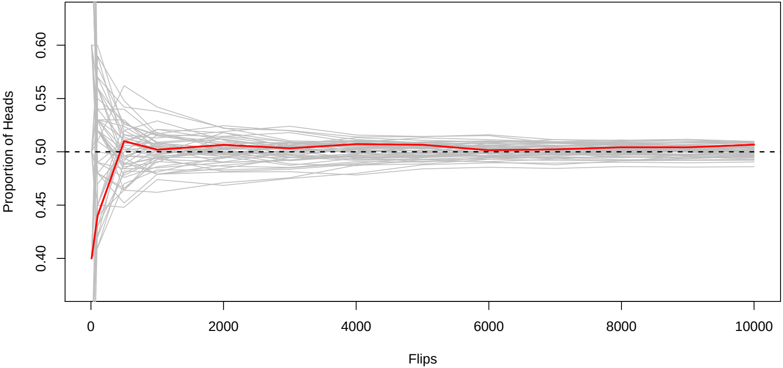


How we Summarize

- Hopefully you recognize it is a little out of the ordinary to summarize Kerrich's experiment by reporting "Heads - 0.5*Tosses"
  - The summary measure seems reasonable, but it isn't something you see very often
- Instead, it's very likely your first thought was that this experiment should be summarized using the proportion of heads
  - There is a reason for this...

Proportions

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

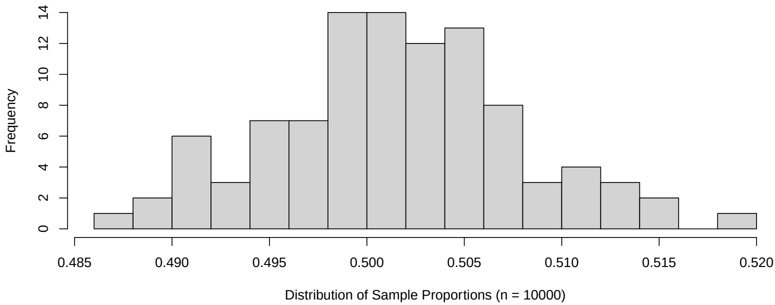


Law of Large Numbers

- Suppose $X_1, X_2, \ldots X_n$ are random variables with the same expected value $E(X) = \mu$
- The **law of large numbers** states that as a $n \to \infty$, the sample average will converge to the random variable's expected value, or $\sum_i X_i/n \to \mu$
- For binary events, the sample proportion is just the average of a sequence of Bernoulli (binary) random variables!

Distribution of the Sample Proportion

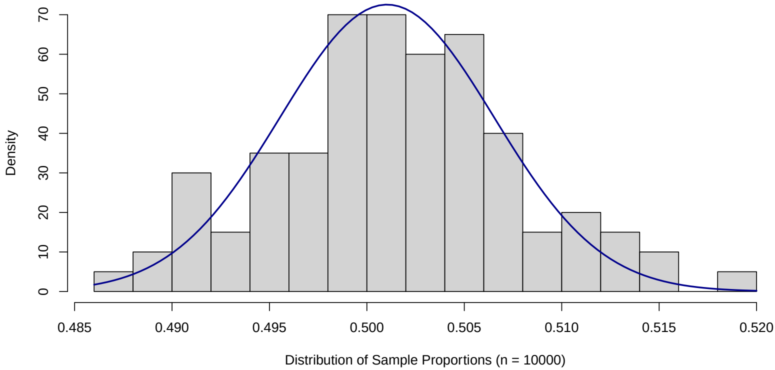
Even when conducting 10,000 coin flips, none of the sample proportions were *precisely* 0.5, below is a histogram:



However, this distribution is incredibly useful, it allows us to express the variability that can be expected in a sample proportion *by random chance alone*

Distribution of the Sample Proportion

Even more useful is that it can be proven that the *distribution of the sample proportion* follows a *normal curve*!



Central Limit Theorem

- Suppose $X_1, X_2, \ldots, X_n$ are independent random variables with expected value $E(X) = \mu$ and variance $Var(X) = \sigma^2$ (see Probability Part 2 for a definition of variance)
- Let $\bar{X}$ denote the mean of all n random variables, **Central Limit Theorem** (CLT) states:

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

Central Limit Theorem

- Suppose $X_1, X_2, \ldots, X_n$ are independent random variables with expected value $E(X) = \mu$ and variance $Var(X) = \sigma^2$ (see Probability Part 2 for a definition of variance)
- Let $\bar{X}$ denote the mean of all n random variables, **Central Limit Theorem** (CLT) states:

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

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

$$\bar{X} \sim N(\mu, \sigma/\sqrt{n})$$

The Power of CLT

- Central Limit Theorem is one of the most important theoretical results in all of statistics
- In the real word, it is nearly impossible to ever figure out the precise distribution of your data
- But if we focus on *sample averages* we don't need to worry about this, CLT tells us what the distribution of sample averages will look like

Example

- The Transport Security Administration (TSA) oversees all travel in the United States, which includes screening all persons and personal possessions traveling via airplane.
  - Each year, thousands of legal claims are filed against the TSA regarding damaged or stolen property, improper screening practices, and bodily injury

Example

- The Transport Security Administration (TSA) oversees all travel in the United States, which includes screening all persons and personal possessions traveling via airplane.
  - Each year, thousands of legal claims are filed against the TSA regarding damaged or stolen property, improper screening practices, and bodily injury
- In 2004, the average claim amount against the TSA was $820.38 with a standard deviation of $20321.43
  - Notice these data are extremely right-skewed (the median claim was only $150)

Example

- The Transport Security Administration (TSA) oversees all travel in the United States, which includes screening all persons and personal possessions traveling via airplane.
  - Each year, thousands of legal claims are filed against the TSA regarding damaged or stolen property, improper screening practices, and bodily injury
- In 2004, the average claim amount against the TSA was $820.38 with a standard deviation of $20321.43
  - Notice these data are extremely right-skewed (the median claim was only $150)
- Suppose the TSA anticipates 300 new claims in any given month (consider this a random sample from the population)
  - What is the probability the month's average claim will exceed $2000?

Example (solution)

- For a sample of size $n = 300$, Central Limit Theorem suggests:

$$\bar{X} \sim N(820.38, 20321.43/\sqrt{300})$$

- We can then calculate $P(\bar{X} \geq 1000)$ using the normal distribution:

```
pnorm(2000, mean = 820.38,
      sd = 20321.43/sqrt(300), lower.tail = FALSE)
```

```
## [1] 0.1573468
```

- So there's a 15.7% chance the sample average exceeds $2000

The "Fuzzy" Central Limit Theorem

- Look at enough datasets and you'll see that normal curve comes up remarkably often
  - Adult height, intelligence, travel times, birth weight, stock volatility, etc. all tend to follow normal distributions

The "Fuzzy" Central Limit Theorem

- Look at enough datasets and you'll see that normal curve comes up remarkably often
  - Adult height, intelligence, travel times, birth weight, stock volatility, etc. all tend to follow normal distributions
- Each of these examples depends upon thousands of genetic and/or environmental factors making small contributions
  - For an individual observation, what we see is the average of all of these numerous factors, making the population appear normally distributed

The "Fuzzy" Central Limit Theorem

- Look at enough datasets and you'll see that normal curve comes up remarkably often
  - Adult height, intelligence, travel times, birth weight, stock volatility, etc. all tend to follow normal distributions
- Each of these examples depends upon thousands of genetic and/or environmental factors making small contributions
  - For an individual observation, what we see is the average of all of these numerous factors, making the population appear normally distributed
- Put differently, CLT tells us that the distribution of averages is normal
  - So if a person's observed height reflects the average effect of thousands of genes, the distribution of heights across the population will be approximately normal

Babies Revisited (again)

- On the first day of class we looked at a study involving babies choosing between a "helper" and "hinderer" toy
  - Recall that 14 of 16 infants chose the "helper" toy
  - We used simulation to determine that this result would be very unlikely to happen by random chance alone

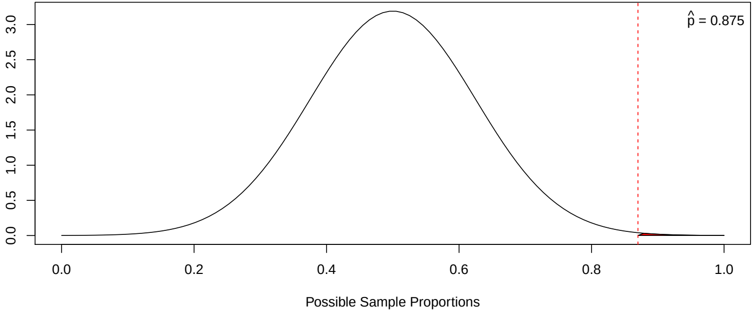
Babies Revisited (again)

- On the first day of class we looked at a study involving babies choosing between a "helper" and "hinderer" toy
  - Recall that 14 of 16 infants chose the "helper" toy
  - We used simulation to determine that this result would be very unlikely to happen by random chance alone
- Now we can use Central Limit Theorem:
  - Let X_i denote the i^{th} baby's choice, then $E(X) = p$
  - Because X is a Bernoulli random variable, $Var(X) = p * (1 - p)$
  - All together:

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

Babies Revisited (again)

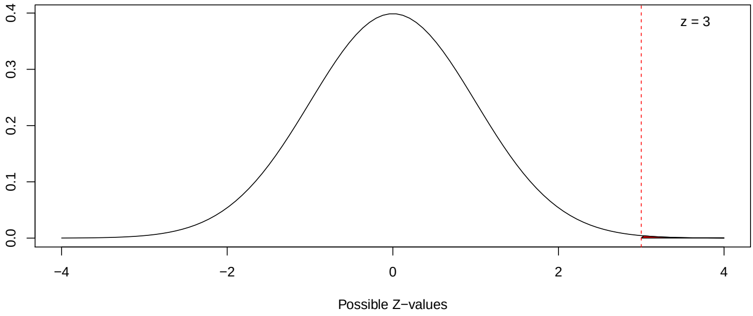
- Under the null model, $p = 0.5$ so $\hat{p} \sim N(0.5, \sqrt{\frac{.5(1-.5)}{16}})$
- The probability of observing a sample proportion at least as large as $\hat{p} = 14/16 = 0.875$ is depicted below
  - The p-value is minuscule, 0.0013



Babies Revisited (again)

Historically, we'd *standardize* our observed proportion in order to find this p-value using the standard normal distribution:

$$z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} = \frac{0.875 - 0.5}{.25/16} = 3$$



The Z-Test

The general procedure we just walked through is known as the "Z-Test", it involves the following steps:

1. Decide upon a suitable summary measure (ie: the sample proportion, $\hat{p}$)
2. Decide upon a null model for that summary measure (ie: coin flips, or $p = 0.5$)
3. Standardize the observed summary measure to obtain a z-score (ie: $z = \frac{\hat{p}-p}{\sqrt{p(1-p)/n}}$), this describes how unusual our sample is relative to other samples that we'd expect under the null model
4. Use the standard normal distribution to calculate the probability of observing a z-score as extreme as the one in our sample

The Z-Test is extremely general, but be aware that it does really on an *asymptotic result* that is only perfectly accurate in the limit.

CLT Caveats

- CLT only applies to *independent* observations
- CLT tells us about the distribution of sample averages (noting that proportions are a average of zeros and ones)
  - It doesn't tell us about other summary measures (see our original summary of Kerrich's experiment)
- CLT is an *asymptotic* result, meaning its results may not be accurate for sample sample sizes ($n = 30$ is a commonly cited threshold)

Next Steps

Broadly speaking, *statistical inference* primarily addresses two goals:

- Hypothesis testing
  - Using sample data to evaluate a null model of a population
- Estimation
  - Using sample data to accurately determine some aspect of a population (ie: the population mean, the correlation between two variables, etc.)

The next portion of the course cover these two topics in detail