The Central Limit Theorem

Patrick Breheny

October 1

Kerrich's experiment

- A South African mathematician named John Kerrich was visiting Copenhagen in 1940 when Germany invaded Denmark
- Kerrich spent the next five years in an interment camp
- To pass the time, he carried out a series of experiments in probability theory
- One of them involved flipping a coin 10,000 times

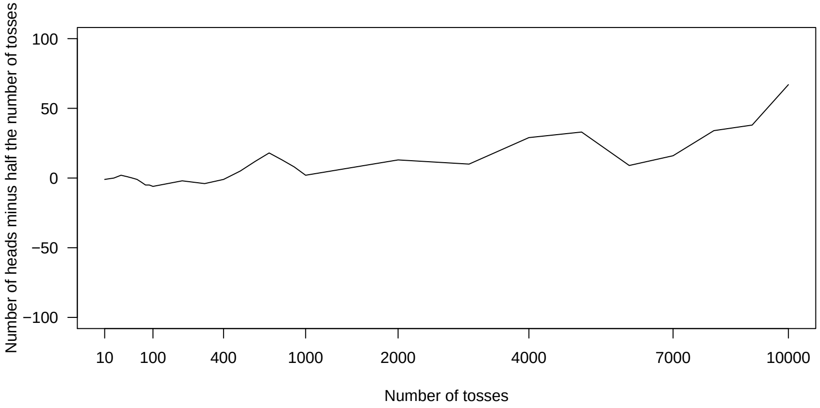
The law of averages?

- We know that a coin lands heads with probability 50%
- So, loosely speaking, if we flip the coin a lot, we should have about the same number of heads and tails
- The subject of today's lecture, though is to be much more specific about exactly what happens and precisely what probability theory tells us about what will happen in those 10,000 coin flips

Kerrich's results

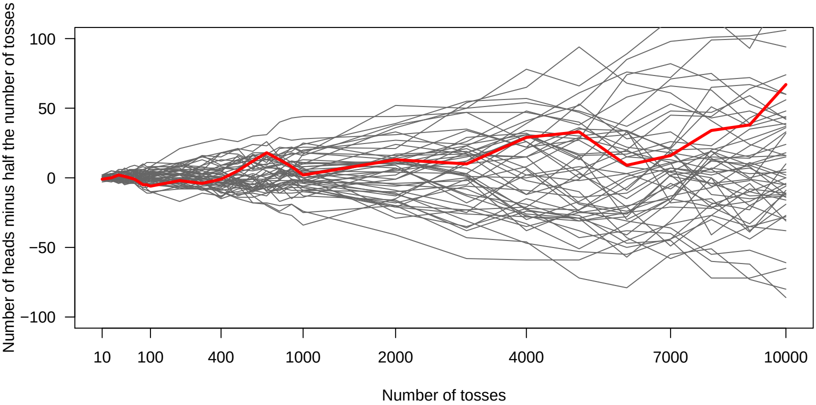
Number of tosses (n)	Number of heads	Heads - 0.5·Tosses
10	4	-1
100	44	-6
500	255	5
1,000	502	2
2,000	1,013	13
3,000	1,510	10
4,000	2,029	29
5,000	2,533	33
6,000	3,009	9
7,000	3,516	16
8,000	4,034	34
9,000	4,538	38
10,000	5,067	67

Kerrich's results plotted



Instead of getting closer, the numbers of heads and tails are getting farther apart

Repeating the experiment 50 times

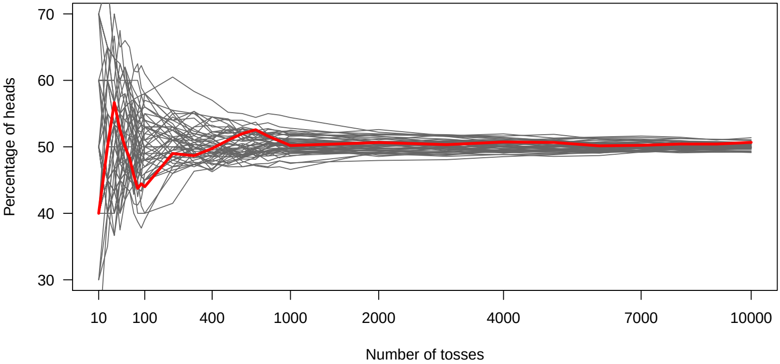


This is not a fluke – instead, it occurs systematically and consistently in repeated simulated experiments

Where's the law of averages?

- As the figure indicates, simplistic notions like "we should have about the same number of heads and tails" are inadequate to describe what happens with long-run probabilities
- We must be more precise about what is happening – in particular, what is getting *more predictable* as the number of tosses goes up, and what is getting *less predictable*?
- Consider instead looking at the percentage of flips that are heads

Repeating the experiment 50 times, Part II



What happens as n gets bigger?

- We now turn our attention to obtaining a precise mathematical description of what is happening to the mean (i.e, the proportion of heads) with respect to three trends:
  - Its expected value
  - Its variance
  - Its distribution
- First, however, we need to define joint distributions and prove a few theorems about the expectation and variance of sums

Joint distributions

- We can extend the notion of a distribution to include the consideration of multiple variables simultaneously
- Suppose we have two discrete random variables X and Y. Then the *joint probability mass function* is a function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f(x, y) = P(X = x, Y = y)$$

- Likewise, suppose we have two discrete random variables X and Y. Then the *joint probability density function* is a function $f : \mathbb{R}^2 \to \mathbb{R}$ satisfying

$$P((X, Y) \in A) = \int \int_A f(x, y)\, dx\, dy$$

for all rectangles $A \in \mathbb{R}^2$

Marginal distributions

- Suppose we have two discrete random variables X and Y. Then the pmf

$$f(x) = \sum_y f(x, y)$$

is called the *marginal* pmf of X

- Likewise, if X and Y are continuous, we integrate out one variable to obtain the marginal pdf of the other:

$$f(x) = \int f(x, y)\, dy$$

Expectation of a sum

- **Theorem:** Let X and Y be random variables. Then

$$\mathrm{E}(X + Y) = \mathrm{E}(X) + \mathrm{E}(Y)$$

  provided that $\mathrm{E}(X)$ and $\mathrm{E}(Y)$ exist
- Note in particular that X and Y do not have to be independent for this to work

Variance of a sum

- **Theorem:** Let X and Y be independent random variables. Then

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$

  provided that $\mathrm{Var}(X)$ and $\mathrm{Var}(Y)$ exist
- Note that X and Y must be independent for this to work

The expected value of the mean

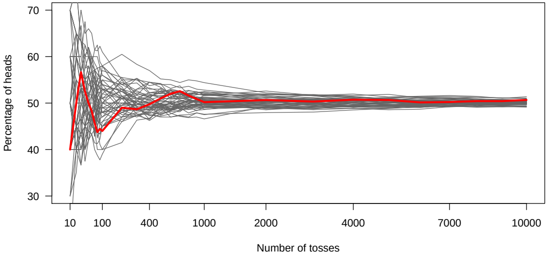
- **Theorem:** Suppose $X_1, X_2, \ldots X_n$ are random variables with the same expected value μ, and let $\bar{X}$ denote the mean of all n random variables. Then

$$\mathrm{E}(\bar{X}) = \mu$$

- In other words, for any value of n, the expected value of the sample mean is expected value of the underlying distribution
- When an estimator $\hat{\theta}$ has the property that $\mathrm{E}(\hat{\theta}) = \theta$, the estimator is said to be *unbiased*
- Thus, the above theorem shows that the sample mean is an unbiased estimator of the population mean

Applying this result to the coin flips example

- **Theorem:** For the binomial distribution, $\mathrm{E}(X) = n\pi$
- Thus, letting $\hat{\pi} = X/n$, $\mathrm{E}(\hat{\pi}) = \pi$, which is exactly what we saw in the earlier picture:



The variance of the mean

- Our previous result showed that the sample mean is always "close" to the expected value, at least in the sense of being centered around it
- Of course, how close it is also depends on the variance, which is what we now consider
- **Theorem:** Suppose $X_1, X_2, \ldots X_n$ are independent random variables with expected value μ and variance σ^2. Letting $\bar{X}$ denote the mean of all n random variables,

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

- **Corollary:** $\mathrm{SD}(\bar{X}) = \sigma/\sqrt{n}$

The square root law

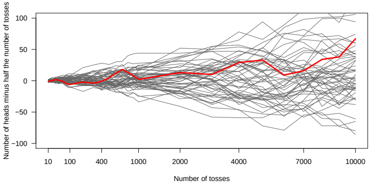
- To distinguish between the standard deviation of the data and the standard deviation of an estimator (e.g., the mean), estimator standard deviations are typically referred to as *standard errors*
- As the previous slide makes clear, these are not the same, and are related to each other by a very important way, sometimes called the *square root law*:

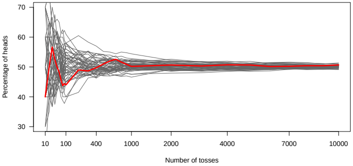
$$SE = \frac{SD}{\sqrt{n}}$$

- This is true for all averages, although as we will see later in the course, must be modified for other types of estimators

Standard errors

Note that $\mathrm{Var}(\sum X)$ goes up with n, while $\mathrm{Var}(\bar{X})$ goes down with n, exactly as we saw in our picture from earlier:

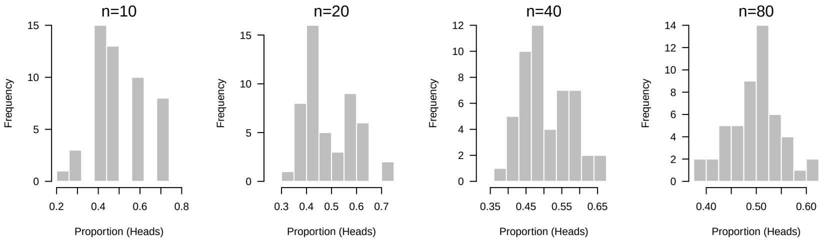




Indeed, $\mathrm{Var}(\bar{X})$ actually goes down all the way to zero as $n \to \infty$

The distribution of the mean

Finally, let's look at the distribution of the mean by creating histograms of the mean from our 50 simulations



The central limit theorem (informal)

- In summary, there are three very important phenomena going on here concerning the sample average:
  - **#1** The expected value is always equal to the population average
  - **#2** The standard error is always equal to the population standard deviation divided by the square root of n
  - **#3** As n gets larger, its distribution looks more and more like the normal distribution
- Furthermore, these three properties of the sample average hold for **any distribution**

The central limit theorem (formal)

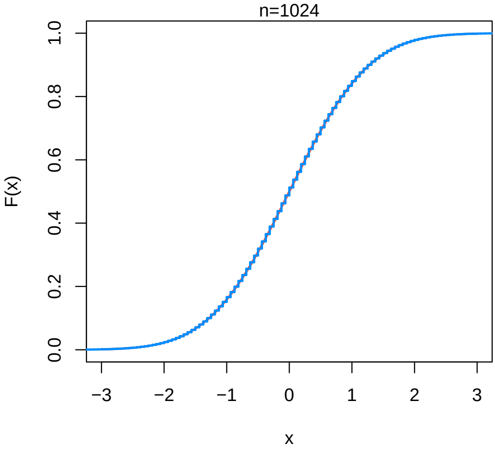
- **Central limit theorem:** Suppose $X_1, X_2, \ldots X_n$ are independent random variables with expected value μ and variance σ^2. Letting $\bar{X}$ denote the mean of all n random variables,

$$\sqrt{n}\frac{\bar{X}-\mu}{\sigma} \xrightarrow{\text{d}} \mathrm{N}(0,1)$$

- The notation $\xrightarrow{\text{d}}$ is read "converges in distribution to", and means that the limit as $n \to \infty$ of the CDF of the quantity on the left is equal to the CDF on the right (at all points x where the CDF on the right is continuous)

Graphical idea of convergence in distribution

For the binomial distribution (red=normal, blue=$\sqrt{n}(\bar{X}-\mu)/\sigma$):



Power of the central limit theorem

- This result is one of the most important, remarkable, and powerful results in all of statistics
- In the real world, we rarely know the distribution of our data
- But the central limit theorem says: we don't have to

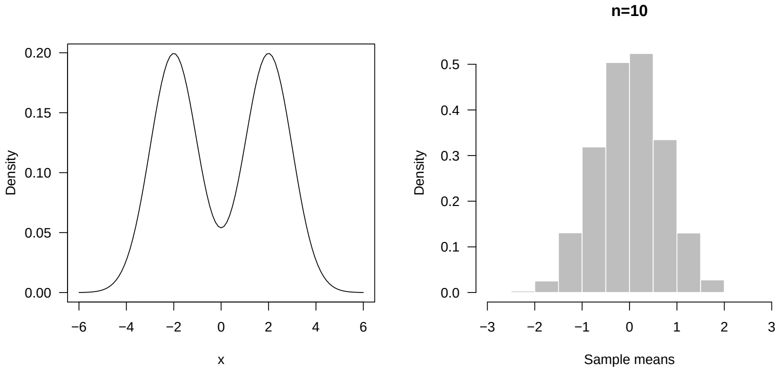
Power of the central limit theorem

- Furthermore, as we have seen, knowing the mean and standard deviation of a distribution that is approximately normal allows us to calculate anything we wish to know with tremendous accuracy – and the distribution of the mean is always approximately normal
- The caveat, however, is that for any finite sample size, the CLT only holds approximately
- How good is this approximation? It depends...

How large does n have to be?

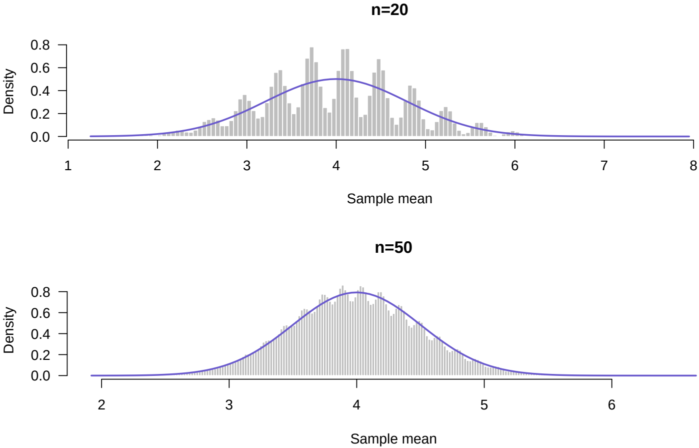
- Rules of thumb are frequently recommended that $n = 20$ or $n = 30$ is "large enough" to be sure that the central limit theorem is working
- There is some truth to such rules, but in reality, whether n is large enough for the central limit theorem to provide an accurate approximation to the true distribution depends on how close to normal the population distribution is, and thus must be checked on a case-by-case basis
- If the original distribution is close to normal, $n = 2$ might be enough
- If the underlying distribution is highly skewed or strange in some other way, $n = 50$ might not be enough

Example #1


n=10
0.20
0.15
0.10
0.05
0.00
Density
−6
−4
−2
0
2
4
6
x
0.5
0.4
0.3
0.2
0.1
0.0
Density
−3
−2
−1
0
1
2
3
Sample means

Example #2

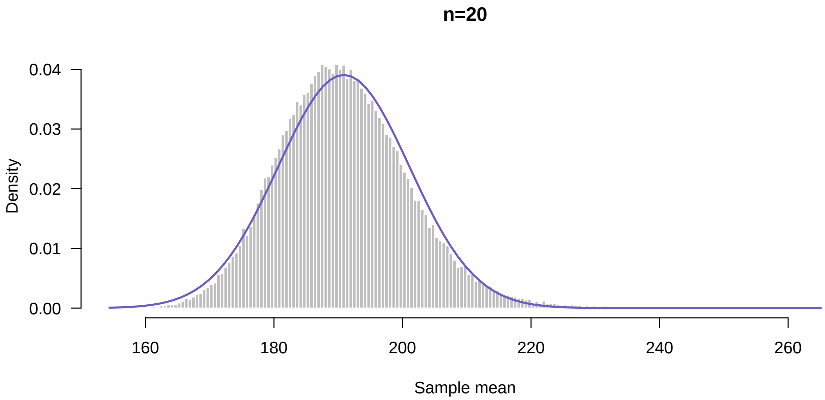
Now imagine an urn containing the numbers 1, 2, and 9:



Example #3

- Weight tends to be skewed to the right (more people are overweight than underweight)
- Let's perform an experiment in which the NHANES sample of adult men is the population
- I am going to randomly draw twenty-person samples from this population (*i.e.* I am re-sampling the original sample)

Example #3 (cont'd)



Why do so many things follow normal distributions?

- We can see now why the normal distribution comes up so often in the real world: any time a phenomenon has many contributing factors, and what we see is the average effect of all those factors, the quantity will follow a normal distribution
- For example, there is no one cause of height – thousands of genetic and environmental factors make small contributions to a person's adult height, and as a result, height is normally distributed
- On the other hand, things like eye color, cystic fibrosis, broken bones, and polio have a small number of (or a single) contributing factors, and do not follow a normal distribution

Summary

- $\mathrm{E}(X+Y)=\mathrm{E}(X)+\mathrm{E}(Y)$
- $\mathrm{Var}(X+Y)=\mathrm{Var}(X)+\mathrm{Var}(Y)$ if X and Y are independent
- Central limit theorem:
    - The expected value of the average is always equal to the population average
    - $\mathrm{SE}=\mathrm{SD}/\sqrt{n}$
    - As n gets larger, the distribution of the sample average looks more and more like the normal distribution
- Generally speaking, the sampling distribution looks pretty normal by about $n=20$, but this could happen faster or slower depending on the underlying distribution, in particular by how skewed it is