Chapter 6: The Distribution of a Sample Statistic

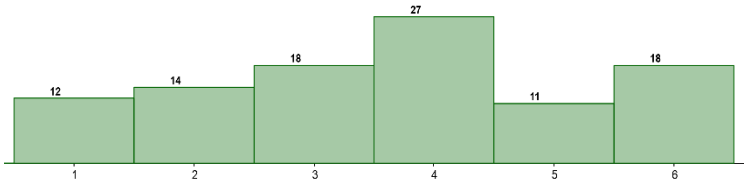
Convergence in Distribution

- **A Population Distribution** represents the entire distribution of a particular random variable.
  - Some populations are theoretically known and has an infinite size (e.g., the distribution of possible dice roll outcomes).
  - Some populations are finite, but might be difficult to catalog (e.g., the distribution of heights from every adult in the world).
- **A Sample Distribution** is the distribution of measurements collected from our ***sample***.
  - A sample is an _____ of the population, but we hope that our sample is representative of the population,
  - We also know that our sample distribution will approach the population distribution in shape as we collect more data!

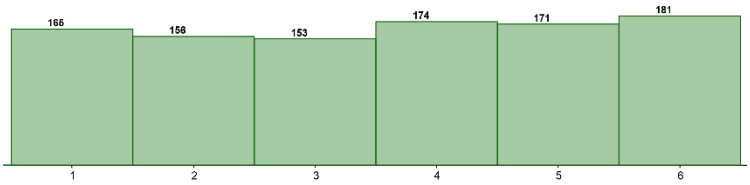
Consider what happens when we roll one die 100 times and record the results on a histogram. Let's use this GeoGebra Simulation (https://www.geogebra.org/m/UsoH4eNl) to draw some different possibilities.

What should happen if we took more dice rolls? Like 1,000? Or 10,000?

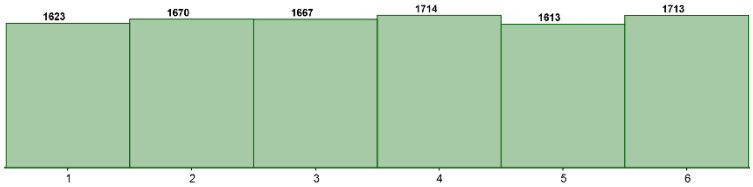
N = 100



N = 1,000



N = 10,000



Convergence in Distribution

The larger your sample, the more your _____ Distribution will converge toward the _____ Distribution in shape!

Convergence of an Estimator

Example: Consider this research question: "*What is the true average salary for U.S. workers?"* Two different research groups try to answer this question by taking a random sample from the population

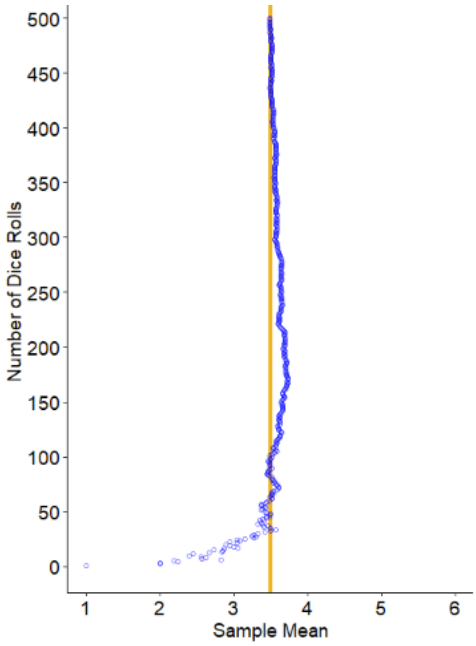
- Group A: Samples 100 people randomly
- Group B: Samples 200 people randomly

Which group would you *expect* to have a sample mean closer to the true mean salary?

The Law of Large Numbers: The sample mean from results obtained from a large number of trials should be _____ to the expected value and _____ to the expected value as _____ trials are performed

You can explore this with rolling a die here:
https://digitalfirst.bfwpub.com/stats_applet/stats_applet_11_largenums.html

In addition to noticing that the sample mean tends closer to the expected value as sample size increases, you may also notice that the _____ in the sample mean _____ as sample size increases.

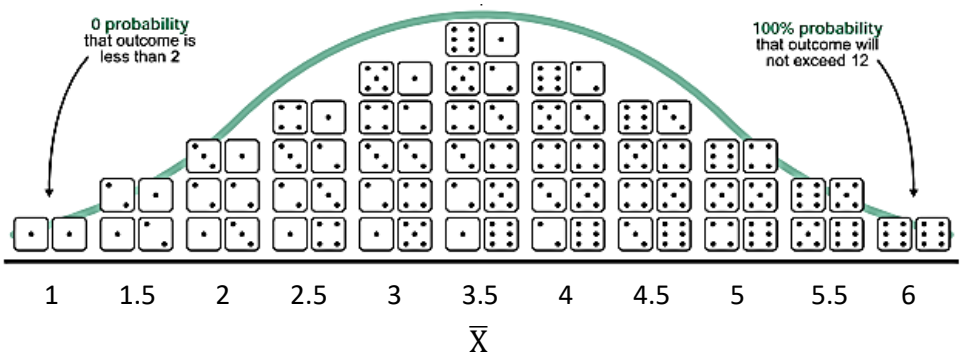


Convergence of an Estimator

The larger your sample, the more your Sample Mean will converge toward the Population Mean in value.

- **The Distribution of a Sample Mean**
  - Mathematically, we can think about a sample mean ($\overline{X}$) as a random variable.
    - Rather than taking a singular observation, we take a sample that holistically produces a singular value in our sample mean
    - $\overline{X}$ varies with each sample we take
    - There is a distribution of possible $\overline{X}$'s I could get every time I sample n units from the population.

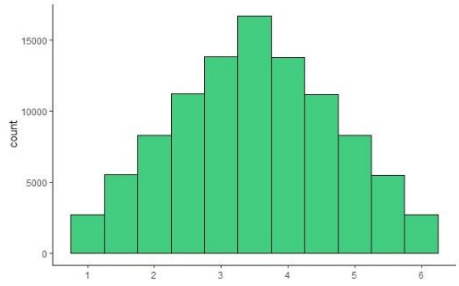
Example: Consider the distribution of $\overline{X}$ when $\overline{X}$ represents the mean of two dice rolls.



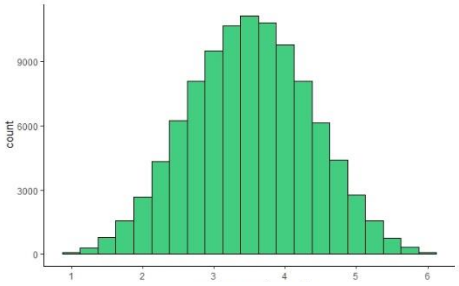
What is the lowest and highest sample mean you could get?

Are all sample means equally likely, or are some more likely than others?

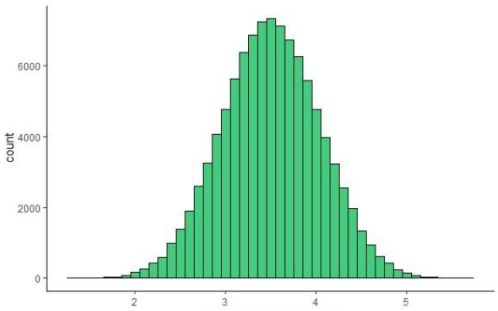
How do you think the distribution of $\overline{X}$ would change if it represented the average of more than 2 dice rolls?



Means from 2 dice rolls



Means from 4 dice rolls



Means from 10 dice rolls

What value is the sampling distribution converging to? How does this relate to the Law of Large Numbers?

Search "OnlineStatBook sampling distributions" (link) → http://onlinestatbook.com/stat_sim/sampling_dist/
Click begin on the upper left and play with the simulation. You'll notice that you can adjust the population distribution up top by clicking directly and reshaping it how you wish!

Practice: Consider this question: What is the earliest age at which a U.S. resident takes their first ride on an airplane?

- Create a distribution that you think would represent that population. *What age you think is most common for a first time? Do you think this distribution is symmetric or skewed?*
- Next, set n = 10 for the box directly below. This will now simulate what would happen if we took 10 random people from this population and reported their sample mean…do this a few thousand times.
- For the bottom box, set to Mean and n to 25 and repeat this process again

Sketch these 3 distributions below: 1) population, 2) distribution of $\overline{X}$ for n = 10, and 3) distribution of $\overline{X}$ for n = 25).

What is the actual mean age according to your **population distribution**? (It should be reported on the top left in blue). Compare this to the means of each of your sampling distributions.

How would you compare the distribution of sample means we get when n = 10 to the distribution we get when n = 25?

If you continued to increase the sample size, what do you think will happen to the distribution of possible sample means?

- A _____ **Distribution** is the general name for the distribution of a sample statistic. But don't confuse it with "SAMPLE distribution."
  - ***Sample*** distributions converge to the population distribution in shape.
  - ***Sampling*** distributions converge to a singular point—the parameter your statistic is estimating!
- We have talked extensively about the sampling distribution for of $\bar{X}$, but we could create a sampling distribution for the median, the standard deviation, or any other sample statistic!

Properties of the Sampling Distribution for <u>Sample Means</u>

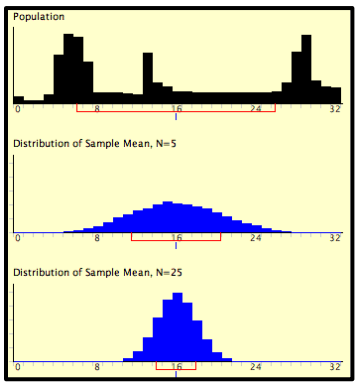
1) The mean of a distribution of means is equal to the population mean. In other words, of $\bar{X}$ is an "unbiased" estimator of μ

2) The standard deviation of $\bar{X}$ will be smaller than the standard deviation of X. It is equal to the population standard deviation divided by the square root of the sample size. We call this the "Standard Error of $\bar{X}$"

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

3) (a) If the distribution of a random variable is normally distributed, then the distribution of $\bar{X}$ is also normally distributed, regardless of the sample size n

3 (b) **Central Limit Theorem**: Even if a variable is **not** normally distributed, the distribution of $\bar{X}$ will become normally distributed if the sample size generating those sample means is large enough.



If the population is not normal, but *doesn't* have a large skew (long tail), then the distribution of $\bar{X}$ will typically be normal if **n > 30**.

If the population is heavily skewed and has a long tail, then n will need to be bigger for $\bar{X}$ to be normally distributed. **n > 120** is sufficient for most common cases!

- The Standard Error of a Sample Mean
  - The "**Standard Error** of the Sample Mean" (abbreviated as $\mathbf{SE_{\bar{x}}}$**)** can be described as _____.

Practice: Consider the population of UIUC students. Let's say that UIUC reports the average ACT score of the student body as 30.35 with a standard deviation of 2.8. If we took a random sample of 50 students and calculated *their* average ACT score, what would be the standard error of our sample mean?

If taking a sample of size 50 from this population, then the expected amount of error in our sample mean as an estimate for the population mean is around _____.

Practice: How would our sample mean change if we sampled **100 students instead of 50?**

A. It would definitely be closer to the true mean of 30.35
B. We would expect it to be closer to 30.35, but it's possible it may not be
C. It would definitely be farther from the true mean of 30.35
D. We would expect it to be farther from 30.35, but it's possible it may not be

- **Estimating Standard Error with s instead of σ**
  - When we do not know σ, we estimate with the sample standard deviation, abbreviated "s." So often, we're actually just *estimating* the standard error.

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad SE_{\bar{x}} \approx \frac{s}{\sqrt{n}}$$

  - This adjustment is important when doing "inference" with only the sample information. It will come up later when we talk about hypothesis testing and confidence intervals!

- **Quick Recap:** Standard Deviation (SD) vs Standard Error (SE)
  - Standard Deviation represents the expected deviation of a data point from the mean
    - σ representing the true standard deviation of a population
    - S representing the standard deviation of a sample of data
  - Standard Error is the standard deviation of a sample statistic. It represents the *expected error* of a sample statistic as an estimate of a parameter.

The Distribution of a Sample Proportion

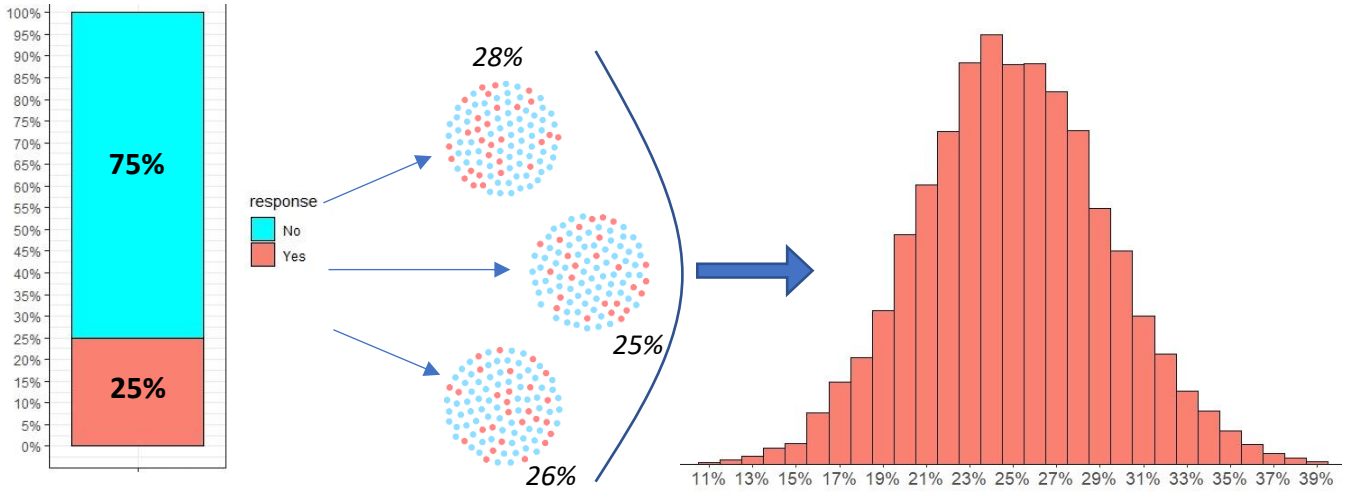
- Proportions Reminders!
  - A Proportion would be the "mean" equivalent when our responses are binary (e.g., "yes" and "no" data).
    - Have you taken a college statistics course before?
    - Did this individual test positive for skin cancer?
    - Does this individual live inside or outside the city limits?
  - For sample data, we can calculate a **sample proportion** (notated $\hat{p}$ [p-hat]) as an estimate of a **population proportion** (p)

Practice: COVID-19 produces new variants over time, and epidemiologists track the spread of new variants to see if it will become the dominant strain. An epidemiologist does extensive testing with 100 randomly selected people with COVID-19 in her community. Of the 100 tested, 21 are positive for the new variant.

What is n?

What is $\hat{p}$?

- Visualizing a Distribution of Sample Proportions
  - Let's consider what *could* be true of the population. Let's assume the true proportion of COVID cases in this community with the new variant is 0.25. Then we can simulate possible values for $\hat{p}$ when taking samples of size 100 from this population.



Population (Actual Proportion) | **Take Samples (of Size 100)** | **Distribution of Possible Sample Proportions**

  - As with sample means, the distribution of $\hat{p}$ will converge and **distribute normally around p**.

- Is the distribution of $\hat{p}$ always normally distributed?
  - Almost always! The only time this wouldn't be true is if p is too close to 0 or 1, and the sample size is not particularly large.
  - If we have **at least 10 "yes" and 10 "no" responses** in our sample, then the distribution of $\hat{p}$ is likely approximately normally distributed.

- The Standard Error and Distribution of a sample proportion
  - A nice shortcut to calculating the **standard deviation** for binary responses is $\sqrt{p(1-p)}$
    - When we don't know p, just substitute $\hat{p}$ into that formula! $\sqrt{\hat{p}\,(1-\hat{p}\,)}$
    - *The "long way" would be converting our binary data to 0's and 1's and using the standard deviation formula.*
  - Standard Error of a Proportion **($SE_{\hat{p}}$)**
    - Once we calculate the standard deviation of our binary data, the formula for the standard error of a sample proportion will still be $\frac{\sigma}{\sqrt{n}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}} \approx \frac{\sqrt{\hat{p}\,(1-\hat{p}\,)}}{\sqrt{n}}$

Practice: Using the COVID-19 variant example, calculate the (approximate) standard error for our sample proportion ($SE_{\hat{p}}$) if our sample size is 100 *using only our sample information.*

Practice: Let's look at the birth ward of two hospitals in one day:
Hospital A records 10 births, with 7 of the babies being female and 3 being male.

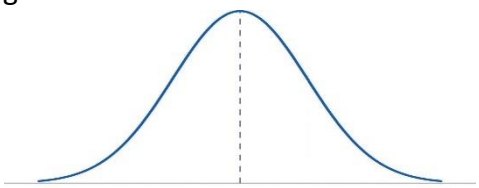
Hospital B records 100 births for the same day, with 70 births being female and 30 being male.

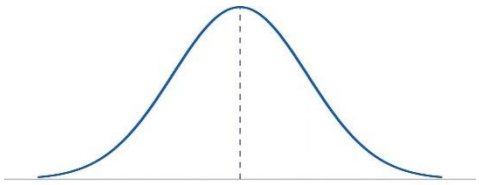
a) Let's assume the true population proportion in both cases to be approximately 50% (about half of babies born will be male and about half born will be female). Use your intuition—is one of these situations more unusual than the other?

b) If p = 50%, then how much error would we typically expect to have in our sample proportions for a sample of size 10? For a sample of size 100?

- **Z-Scores for Statistics: Assigning Positions from the Center**
    - Since sample means and sample proportions distribute in a normal shape around the true parameter, we can convert these statistics to z-scores to identify their positions in their respective sampling distributions.
    - A z-score in the context of sample statistics is measuring **how many standard errors off your sample statistic is away from the parameter.**
    - We can translate this into an equation as well: $z = \frac{\hat{p} - p}{SE_{\hat{p}}}$ (or for means: $z = \frac{\bar{x} - \mu}{SE_{\bar{x}}}$)

Practice: Calculate the z scores for the Hospital A sample proportion and the Hospital B sample proportion. Then identify where each sample proportion is on their respective sampling distributions





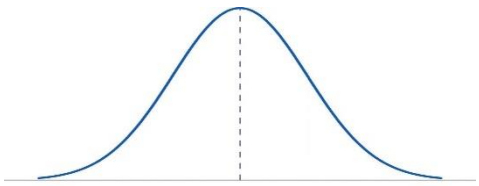
Now interpret each: The sample proportion for Hospital A is _____ standard errors away from the population proportion of 50%. The sample proportion for Hospital B is _____ standard errors away from the population proportion of 50%.

Practice: We are investigating the average amount of arrival delay for U.S. domestic flights. We take a sample of 125 flights and record when they arrived with respect to their scheduled arrival time.

- In our sample, the mean discrepancy was 14 minutes delayed.
- The standard deviation of arrival time discrepancies came out to be 18 minutes for our sample.

What is the (estimated) standard error of $\overline{X}$?

Find the z-score for our specific sample mean, and then visualize its position on the distribution.



Practice: Consider two different distributions. One is the distribution of typical haircut prices that *Male* UIUC students report and the other distribution is the typical haircut prices that *Female* UIUC students report. We collect a random sample of 40 male students and ask them how much they spend on haircuts, and then did the same with 40 females. We notice that female haircut prices tend to have higher variability.

Which sample mean will have the higher standard error? The Male mean or Female mean?

Let's Review Some Important Insights about Sampling Variability!

- ✓ Whenever we want to approximate a parameter, a sample statistic is *usually* our best guess.
- ✓ But samples are incomplete, and our sample statistic will almost surely have some error.
- ✓ In fact, we can create a distribution to represent the possible sample statistics we would reasonably get with our sample size.
- ✓ The Central Limit Theorem shows that for sample means and sample proportions, the distribution of sample statistics will be normally distributed when the sample sizes those statistics come from are sufficiently large. The fact that sampling distributions are normally distributed is nice because we can use normal distribution properties to quantify how likely we are to be a certain distance from the parameter
- ✓ The Standard Error is a way to quantify how much error we likely have in our sample statistic.
- ✓ The Standard Error will get smaller when our sample is larger—that's because a larger sample should produce a more reliable sample statistic.