Unit 5: Sampling Distributions

Merrick Fanning

July 25, 2025

Unit 5 Outline: Sampling Distributions

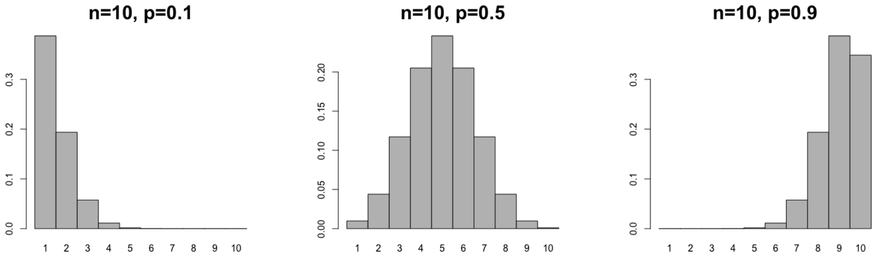
1. Distributions that converge to normal
2. Sampling distributions
3. Point estimates and bias
4. Sampling distribution for $\hat{p}$
5. Sampling distribution for a difference of proportions $(\hat{p}_1 - \hat{p}_2)$
6. Sampling distribution for μ
7. Sampling distribution for a difference of means $(\mu_1 - \mu_2)$

Normal Approximation to the Binomial Distribution

- A discrete binomial variable can be approximated by a continuous normal variable.
- This is useful when the binomial formula becomes computationally intensive for large n.
- This concept will be very important later in statistical inference.

Binomial Shape at Small $n = 10$

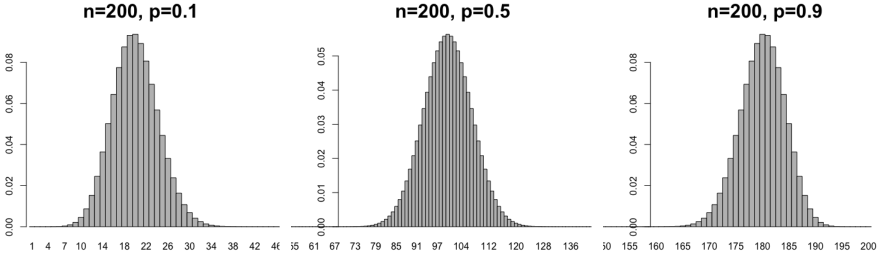
Let's consider several values of p with $n = 10$.



- $p = 0.5$: symmetric, bell-shaped
- $p = 0.1$: right-skewed
- $p = 0.9$: left-skewed

Effect of Increasing n (to 200)

Let's keep the same values of p, but increase n to 200:



As n increases, the binomial distribution looks more normal - even for skewed p values.

When is the Normal Approximation Valid?

Rule of Thumb

The normal approximation is appropriate if:

$$np \geq 10 \quad \text{and} \quad n(1-p) \geq 10$$

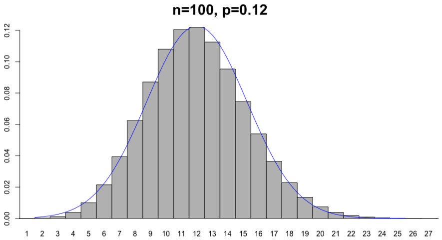
- Interpreted as having at least 10 expected successes and 10 expected failures.
- Use:

$$\mu = np, \quad \sigma = \sqrt{np(1-p)}$$

Example: No Continuity Correction

A basketball player has a 12% chance of making a free throw. Estimate the probability they make 18 or more in 100 shots.
Check approximation validity:

$$np = 12, \quad n(1-p) = 88 \Rightarrow \text{valid}$$



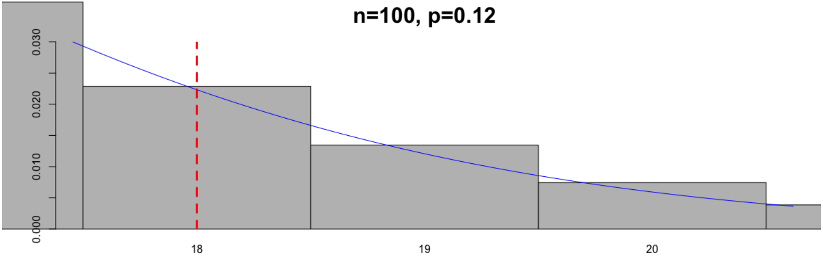
$$\mu = 12, \quad \sigma = \sqrt{100(0.12)(0.88)} \approx 3.25$$

$$z = \frac{18-12}{3.25} \approx 1.85 \Rightarrow P(Z > 1.85) \approx 0.0324$$

About 3.2% chance of making at least 18 shots.

Why Use a Continuity Correction?

We're approximating a discrete variable with a continuous one. So we may also apply a continuity correction.



- Without correction, we ignore part of the probability mass.
- For better accuracy, use $x = 17.5$ instead of 18.

Continuity Correction Summary

Binomial	Normal Approximation
$P(X = x)$	$P(x - 0.5 \leq X \leq x + 0.5)$
$P(X \leq x)$	$P(X \leq x + 0.5)$
$P(X < x)$	$P(X \leq x - 0.5)$
$P(X > x)$	$P(X \geq x + 0.5)$
$P(X \geq x)$	$P(X \geq x - 0.5)$

Tip: Don't memorize - just sketch the histogram and think logically!

Distributions That Converge to Normal

- Many important probability distributions become approximately normal under the right conditions (typically as sample size or degrees of freedom increase).

Examples:

- **Binomial:** Normal approximation valid as $n \to \infty$.
- **Hypergeometric:** As $n \to \infty$ for small n relative to N.
- **Poisson:** For large λ, the distribution becomes approximately normal.
- **Chi-Square (χ^2):** Becomes more symmetric and bell-shaped as degrees of freedom increase.
- **t-distribution:** Approaches standard normal as degrees of freedom increase.

Takeaway: The normal distribution plays a central role in inference because many statistics follow a normal distribution in large samples.

Point Estimation

- Statistics are used to estimate **population parameters**.
- A **point estimate** is a single value used to estimate a target parameter.

$$\overline{x} \text{ is a point estimate for } \mu, \quad \hat{p} \text{ is a point estimate for } p$$

Bias of a Point Estimator

We say $\hat{\theta}$ is an **unbiased estimator** of the parameter θ if:

$$E(\hat{\theta}) = \theta$$

The bias of an estimator is defined as:

$$B(\hat{\theta}) = E(\hat{\theta}) - \theta$$

- We also care about the **variance** and **distribution** of estimators.

What is a Sampling Distribution?

- A **sampling distribution** is the distribution of a statistic over all possible samples.
- Imagine repeating a random sample process infinitely many times and recording a statistic each time.
- The distribution of all these sample statistics forms the sampling distribution.

Why It Matters

Sampling distributions are **essential** for statistical inference. They allow us to:

- Understand variability in estimates
- Construct confidence intervals (Unit 6-9)
- Perform hypothesis testing (Unit 6-9)

Sampling Distribution for $\hat{p}$

- Take a random sample of size n from a population of size N.
- Let X be the number of sample elements with a certain characteristic.

$$\hat{p} = \frac{X}{n}$$

- The population has r total successes, so:

$$p = \frac{r}{N}$$

- $X \sim \text{Hypergeometric}(r, n, N)$

Expected Value of $\hat{p}$ (Hypergeometric)

Prove $\hat{p}$ is an unbiased estimator for p:

Expected Value of $\hat{p}$ (Hypergeometric)

Prove $\hat{p}$ is an unbiased estimator for p:

$$\begin{aligned} E(\hat{p}) &= E\left(\frac{X}{n}\right) \\ &= \frac{1}{n}E(X) \\ &= \frac{1}{n} \cdot n\left(\frac{r}{N}\right) \\ &= \frac{r}{N} = p \end{aligned}$$

- $\hat{p}$ is an **unbiased estimator** of p.
- The sampling distribution of $\hat{p}$ is centered at the true population proportion.

Variance of $\hat{p}$ (Hypergeometric)

Determine the variance of the statistics $\hat{p}$:

Variance of $\hat{p}$ (Hypergeometric)

Determine the variance of the statistics $\hat{p}$:

$$\begin{aligned} \text{Var}(\hat{p}) &= \text{Var}\left(\frac{X}{n}\right) \\ &= \frac{1}{n^2} \cdot n \cdot \frac{r}{N}\left(1 - \frac{r}{N}\right)\left(\frac{N-n}{N-1}\right) \\ &= \frac{p(1-p)}{n} \cdot \left(\frac{N-n}{N-1}\right) \end{aligned}$$

$$SD(\hat{p}) = \sqrt{\frac{p(1-p)}{n} \cdot \left(\frac{N-n}{N-1}\right)}$$

- This is most appropriate for small, finite populations.
- We **DON'T** use this for AP statistics (why not)?

When Can We Use the Binomial Approximation?

- We approximate $X \sim \text{Binomial}(n, p)$, which is valid when:

Independence Condition

Sample size n is less than 10% of the population: $n < 0.1N$

Assuming $X \sim \text{Binomial}(n, p)$, we get:

$$E(\hat{p}) = p, \quad Var(\hat{p}) = \frac{p(1-p)}{n}, \quad SD(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

1. Why is $X \sim \mathrm{Binomial}(n, p)$ when $n < 10\%N$?
2. Prove the expected value for $\hat{p}$ and variance for $\hat{p}$ using the binomial approximation.

Normal Approximation to the Sampling Distribution of $\hat{p}$

- Even with the binomial model, exact computations can be complex.
- So we use a normal approximation for $\hat{p}$, if the following condition is met:

Normality Condition

$$np > 10 \quad \text{and} \quad n(1-p) > 10$$

(At least 10 expected successes and failures)

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0, 1)$$

Conditions for Using the Normal Sampling Distribution of $\hat{p}$

To use the normal model for $\hat{p}$, the following must be true:

- **Random Sampling:** Sample is collected randomly.
- **Independence:** Population is at least 10 times larger than the sample ($n < 0.1N$).
- **Normality:** $np > 10$ and $n(1 - p) > 10$

These are assumptions - they are not always verifiable but are necessary to use this model.

Example: Sampling Distribution for $\hat{p}$

It is known that across North America, 65% of university students take longer than four years to complete their undergraduate degree. You survey 100 University of Calgary graduates.

(a) **Distribution for X:**
Since $n = 100 < 0.1N$, we approximate using a binomial model:

$$X \sim \text{Binomial}(n = 100, p = 0.65)$$

Example: Sampling Distribution for $\hat{p}$

It is known that across North America, 65% of university students take longer than four years to complete their undergraduate degree. You survey 100 University of Calgary graduates.

(a) **Distribution for X:**
Since $n = 100 < 0.1N$, we approximate using a binomial model:

$$X \sim \text{Binomial}(n = 100, p = 0.65)$$

(b) **Sampling distribution for $\hat{p}$:**

Conditions:

- Independence: $n = 100 < 0.1N$ ✓
- Normality: $np = 65 > 10$, $n(1 - p) = 35 > 10$ ✓

$$\hat{p} \sim \text{Normal}\left(0.65, \frac{0.65(0.35)}{100}\right)$$

Example: Sampling Distribution for $\hat{p}$

It is known that across North America, 65% of university students take longer than four years to complete their undergraduate degree. You survey 100 University of Calgary graduates.

(a) **Distribution for X:**
Since $n = 100 < 0.1N$, we approximate using a binomial model:

$$X \sim \text{Binomial}(n = 100, p = 0.65)$$

(b) **Sampling distribution for $\hat{p}$:**

Conditions:

- Independence: $n = 100 < 0.1N$ ✓
- Normality: $np = 65 > 10$, $n(1 - p) = 35 > 10$ ✓

$$\hat{p} \sim \text{Normal}\left(0.65, \frac{0.65(0.35)}{100}\right)$$

(c) **Probability that $\hat{p} > 0.70$:**

$$z = \frac{0.70 - 0.65}{\sqrt{\frac{0.65 \cdot 0.35}{100}}} = 1.048$$

$$P(\hat{p} > 0.70) = P(Z > 1.048) \approx 0.147$$

There is about a 15% chance that more than 70% of your sample took over four years to graduate.

The Sampling Distribution of $\overline{x}$

Suppose $X_1, X_2, \ldots, X_n$ are independent and identically distributed random variables.

$$\overline{x} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Assume each $X_i \sim \text{Normal}(\mu, \sigma^2)$. Then:

$$E(\overline{x}) = \mu \quad \text{(Unbiased)}$$

$$Var(\overline{x}) = \frac{\sigma^2}{n}, \quad SD(\overline{x}) = \frac{\sigma}{\sqrt{n}}$$

So:

$$\overline{x} \sim \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$$

Independence condition: $n < 10\%$ of the population

Standardizing the Sampling Distribution

We often standardize $\overline{x}$ using:

$$Z = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Where:

- $Z \sim \text{Normal}(0, 1)$
- Requires known σ
- Assumes random sampling and independence

If σ is unknown, we will require a new distribution - this will be **covered later.**

Example: Pale-Throated Sloths (Setup)

The weights of pale-throated sloths are normally distributed:

$$\mu = 4.5 \text{ kg}, \quad \sigma = 1.1 \text{ kg}$$

You randomly sample 20 sloths.

(a) **Describe the sampling distribution of** $\overline{x}$:

Example: Pale-Throated Sloths (Setup)

The weights of pale-throated sloths are normally distributed:

$$\mu = 4.5 \text{ kg}, \quad \sigma = 1.1 \text{ kg}$$

You randomly sample 20 sloths.

(a) **Describe the sampling distribution of** $\overline{x}$:

Since $n = 20 < 0.1N$, and the parent distribution is normal:

$$\overline{x} \sim \text{Normal}\left(4.5, \frac{(1.1)^2}{20}\right)$$

(b) What is the probability the sample mean is between 2.3 kg and 4.3 kg?

Example: Pale-Throated Sloths (Setup)

The weights of pale-throated sloths are normally distributed:

$$\mu = 4.5 \text{ kg}, \quad \sigma = 1.1 \text{ kg}$$

You randomly sample 20 sloths.

(a) **Describe the sampling distribution of** $\overline{x}$:

Since $n = 20 < 0.1N$, and the parent distribution is normal:

$$\overline{x} \sim \text{Normal}\left(4.5, \frac{(1.1)^2}{20}\right)$$

(b) What is the probability the sample mean is between 2.3 kg and 4.3 kg?

$$z_{\text{low}} = \frac{2.3 - 4.5}{\frac{1.1}{\sqrt{20}}} = -8.94, \quad z_{\text{high}} = \frac{4.3 - 4.5}{\frac{1.1}{\sqrt{20}}} = -0.81$$

$$P(2.3 \leq \overline{x} \leq 4.3) = P(-8.9 \leq Z \leq -0.8) \approx 0.2119$$

There is approximately a 21.2% chance the sample mean falls in this range.

The Central Limit Theorem (CLT)

Question: What happens when the parent distribution is not normal?

The Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with

$$E(X_i) = \mu, \quad Var(X_i) = \sigma^2$$

Then:

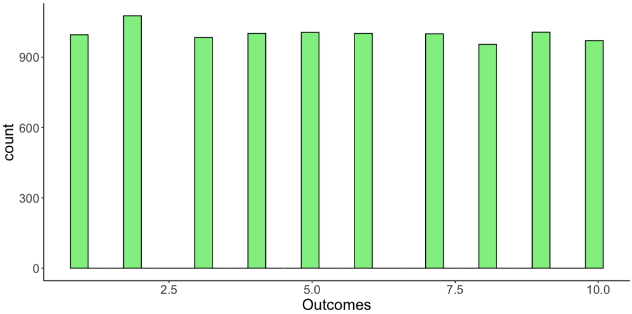
$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} \xrightarrow{n \to \infty} \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$$

That is, the sampling distribution of $\overline{X}$ becomes normal as n increases - regardless of the parent distribution.

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim Z \quad \text{for } n \geq 30$$

Example: 10-Sided Die

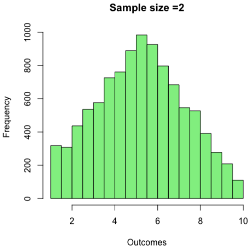
Let X represent the outcome of a 10-sided die roll. The parent distribution is uniform.



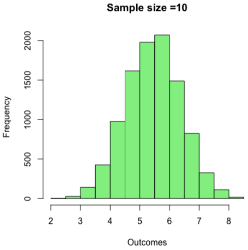
Even though this parent distribution is not normal, the CLT applies as n increases.

Sampling Distributions of $\overline{X}$

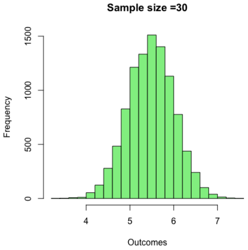
Below are sampling distributions from 10,000 samples for sample sizes of 2, 10, and 30:



Sample size = 2



Sample size = 10



Sample size = 30

As the sample size increases, the sampling distribution of $\overline{X}$ becomes more normal **regardless of the parent population**.

Example: Carnival Game - Profit Distribution

A carnival game has the following profit distribution:

Profit ($)	-1	1	5	20
Probability	0.95	0.03	0.02	0.01

Let X be your profit from a single play.

(a) **Determine expected value for X:**

Example: Carnival Game - Profit Distribution

A carnival game has the following profit distribution:

Profit ($)	-1	1	5	20
Probability	0.95	0.03	0.02	0.01

Let X be your profit from a single play.

a. **Determine expected value for X:**

$$E(X) = -1(0.95) + 1(0.03) + 5(0.02) + 20(0.01) = -0.62$$

b. **Determine variance for X:**

Example: Carnival Game - Profit Distribution

A carnival game has the following profit distribution:

Profit ($)	-1	1	5	20
Probability	0.95	0.03	0.02	0.01

Let X be your profit from a single play.

(a) **Determine expected value for X:**

$$E(X) = -1(0.95) + 1(0.03) + 5(0.02) + 20(0.01) = -0.62$$

(b) **Determine variance for X:**

$$E(X^2) = 1(0.95) + 1(0.03) + 25(0.02) + 400(0.01) = 5.48$$

$$Var(X) = E(X^2) - (E(X))^2 = 5.48 - (-0.62)^2 = 5.4556$$

Carnival Game: CLT Approximation

- Suppose you play the game 30 times ($n = 30$).
- CLT applies: large sample size.
- Then:

$$\mu_{\overline{x}} = -0.62, \quad \sigma_{\overline{x}} = \sqrt{\frac{5.4556}{30}} = 0.4264$$

$$\overline{x} \sim \text{Normal}(-0.62, 0.4264)$$

1. What is the probability that your profit is positive after playing the 30 games?

Carnival Game: CLT Approximation

- Suppose you play the game 30 times ($n = 30$).
- CLT applies: large sample size.
- Then:

$$\mu_{\overline{x}} = -0.62, \quad \sigma_{\overline{x}} = \sqrt{\frac{5.4556}{30}} = 0.4264$$

$$\overline{x} \sim \text{Normal}(-0.62, 0.4264)$$

1. What is the probability that your profit is positive after playing the 30 games?

$$\begin{aligned} P(\overline{x} > 0) &= P\left(\frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} > \frac{0 - (-0.62)}{\frac{0.4264426}{\sqrt{30}}}\right) \\ &= P(Z > 7.963275) \\ &= 1 - P(Z \leq 7.963275) \\ &\approx 0 \end{aligned}$$

Assumptions for Using a Normal Model

To use the normal model for $\overline{x}$, we must assume:

- **Normality:** Either the parent population is normal or $n \geq 30$
- **Independence:** Sample size $n < 10\%$ of population size N
- **Random Sampling:** Sample is collected using a random method

$$\overline{x} \sim \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$$

Similar to assumptions for the sampling distribution of $\hat{p}$

A Sampling Distribution Involving s^2

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a normal population with mean μ and variance σ^2.
Then the following distribution holds:

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

Example: Pale-Throated Sloths

The weights of sloths are normally distributed with $\mu = 4.5$ kg, $\sigma = 1.1$ kg. A random sample of $n = 20$ sloths is taken. What is the probability that the sample standard deviation is at least 0.9?

$$P(s^2 > 0.9^2) = P\left(\chi^2_{19} > \frac{(0.9)^2 \cdot 19}{(1.1)^2}\right)$$

$$= P\left(\chi^2_{19} > 12.72\right) = 0.8526$$

Conclusion: There's an 85% chance of observing a sample standard deviation of 0.9 or greater.

Standard Deviation vs. Standard Error

Problem: Many sampling distributions involve unknown population parameters.

- For the sampling distribution of the sample mean:

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$

- But the population standard deviation σ is usually unknown.
- We estimate it using the sample standard deviation s.

Standard Error

The **standard error** is the estimated standard deviation of a statistic:

$$SE_{\overline{x}} = \frac{s}{\sqrt{n}}$$

What happens to the distribution?

$$\frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim Z \quad \text{(when } \sigma \text{ is known)} \qquad \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}} \sim ??? \quad \text{(new distribution)}$$

Using the t-Distribution

Consider a standard normal random variable Z, and a chi-square random variable with k degrees of freedom. The t-distribution is defined as:

$$t = \frac{Z}{\sqrt{\frac{\chi_k^2}{k}}}$$

Recall the following known distributions (when assumptions are met):

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim Z, \quad \frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$$

Using these, we construct the t-statistic:

Using the t-Distribution

Consider a standard normal random variable Z, and a chi-square random variable with k degrees of freedom. The t-distribution is defined as:

$$t = \frac{Z}{\sqrt{\frac{\chi_k^2}{k}}}$$

Recall the following known distributions (when assumptions are met):

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim Z, \quad \frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$$

Using these, we construct the t-statistic:

$$\frac{\left(\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}\right)}{\sqrt{\frac{\left(\frac{(n-1)s^2}{\sigma^2}\right)}{n-1}}} = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

Assumptions for Using the t-Distribution

To use the t-distribution, the following assumptions must hold:

- **Simple Random Sampling**
- **Independence:** $n < 0.1N$
- **Normality:**
  Ideally, the population is normal with mean μ, variance σ^2. Then:

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim Z, \quad \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

  If the parent population is unknown:

  - If $n \geq 30$, the CLT allows:

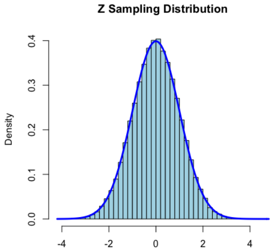
$$\frac{\overline{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

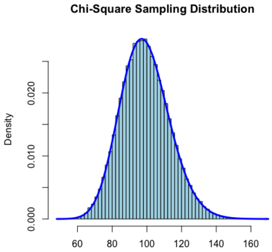
  - If $n < 30$, we require the population to be approximately normal (unimodal, symmetric, no outliers).

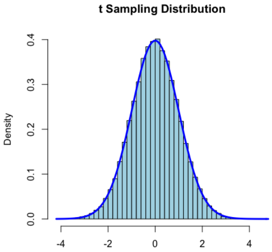
Caution: Small, skewed, or heavy-tailed samples may make the t-distribution inappropriate.

Case I: Normal Parent Distribution, Large Sample ($n = 100$)

Assume $X_1, X_2, \ldots, X_{100} \sim \text{Normal}(10, 2)$ Histograms below show the sampling distributions (100,000 simulations), with theoretical curves superimposed.





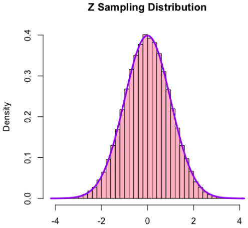


$$\frac{\overline{x} - \mu}{\sigma/\sqrt{n}} \sim Z \qquad \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{99} \qquad \frac{\overline{x} - \mu}{s/\sqrt{n}} \sim t_{99}$$

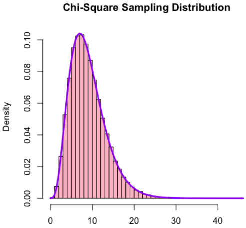
Conclusion: With a large sample size and normal parent population, the theoretical distributions are a very good fit.

Case II: Normal Parent Distribution, Small Sample ($n = 10$)

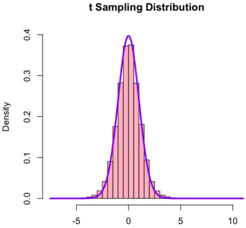
Assume $X_1, X_2, \ldots, X_{10} \sim \text{Normal}(10, 2)$. Again, histograms show empirical sampling distributions with theoretical curves.



$$\frac{\overline{x} - \mu}{\sigma/\sqrt{n}} \sim Z$$



$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_9$$

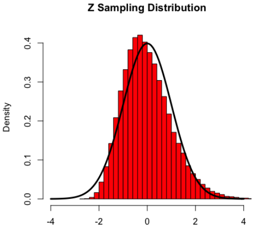


$$\frac{\overline{x} - \mu}{s/\sqrt{n}} \sim t_9$$

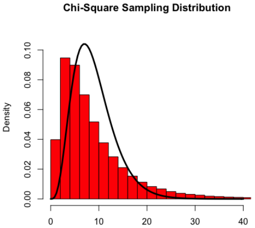
Conclusion: Even with a small sample, normality in the parent distribution ensures that the *t*-distribution is appropriate.

Case III: Skewed Parent Distribution, Small Sample ($n = 10$)

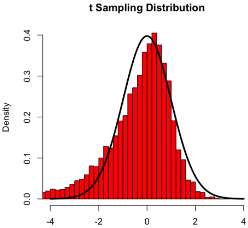
Assume $X_1, X_2, \ldots, X_{10} \sim \text{Exponential}(3)$, a **highly right skew distribution**.



$$\frac{\overline{x} - \mu}{\sigma/\sqrt{n}} \sim ?$$



$$\frac{(n-1)s^2}{\sigma^2} \sim ?$$



$$\frac{\overline{x} - \mu}{s/\sqrt{n}} \sim ?$$

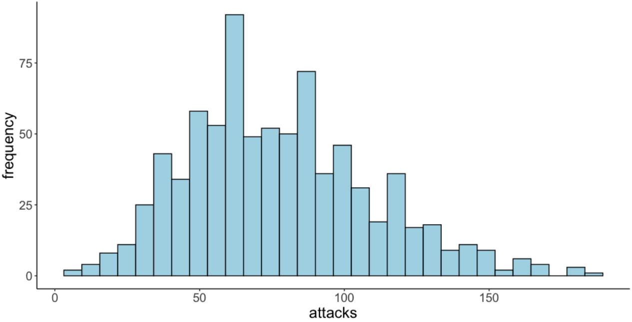
Conclusion: With a skewed parent distribution and small n, the theoretical distributions do not fit. Use caution when applying the t-distribution in this scenario.

Example: Pokémon Attack Scores

A random sample of $n = 801$ Pokémon has:

$$\overline{x} = 78, \quad s = 32$$

Suppose the true population mean is $\mu = 70$. The sample distribution is shown below:



(a) What is the probability that a future sample has a mean attack score less than 70?

Example: Pokémon Attack Scores Solution

Solution: Large n and approximately normal data $\rightarrow$ use the t-distribution.

$$P(\overline{x} < 70) = P\left(\frac{\overline{x} - \mu}{s/\sqrt{n}} < \frac{70 - 78}{32/\sqrt{801}}\right)$$

$$= P(t_{800} < -2.6533) = 0.0041$$

Conclusion: There's about a 0.41% chance that a random sample of 801 Pokémon would have a mean attack below 70.

Summary of Sampling Distributions

Let's summarize the sampling distributions we've developed so far:

Distribution	**Assumptions**
$\hat{p} \sim \text{Normal}\left(p, \frac{p(1-p)}{n}\right)$	Random sampling, independence, and normality condition: $np > 10$, $n(1-p) > 10$
$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim Z$	Same as above (standardized version)
$\overline{x} \sim \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$	Random sampling, independence ($n < 0.1N$), and normal population or large $n \geq 30$
$\frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim Z$	When σ is known, with same assumptions as above
$\frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$	When σ is unknown. Requires normality or large sample, plus random sampling and independence

Sampling Distribution of a Difference in Proportions

We often compare two sample proportions:

- $\hat{p}_1$: the sample proportion from a group of size n_1
- $\hat{p}_2$: the sample proportion from a second group of size n_2

Our goal: Understand the behavior of the statistic $\hat{p}_1 - \hat{p}_2$. **Assumptions:**

- **Random Sampling:** Each sample is drawn using a random method.
- **Independence:** Observations are independent within and between samples. Assume this if:

$$n_1 < 0.1N_1 \quad \text{and} \quad n_2 < 0.1N_2$$

- **Normality:** Each sample must have at least 10 successes and 10 failures:

$$n_1 p_1 > 10, \quad n_1(1 - p_1) > 10, \quad n_2 p_2 > 10, \quad n_2(1 - p_2) > 10$$

- What is $E(\hat{p}_1 - \hat{p}_2)$?
- What is $\text{Var}(\hat{p}_1 - \hat{p}_2)$?
- What distribution does $\hat{p}_1 - \hat{p}_2$ follow?

Example: Difference in Proportions - Two Towns

In one town, 51% of voters are conservative; in another, 44% are conservative. A random sample of 100 voters is taken from each town.

(a) **Is a normal model appropriate for** $\hat{p}_1 - \hat{p}_2$?

Example: Difference in Proportions - Two Towns

In one town, 51% of voters are conservative; in another, 44% are conservative. A random sample of 100 voters is taken from each town.

(a) **Is a normal model appropriate for** $\hat{p}_1 - \hat{p}_2$?
   - **Simple Random Sample:** Assumed for both towns.
   - **Independence:** $n_1 = n_2 = 100 < 0.1N$ so we assume independence.
   - **Normality:**

$$n_1 p_1 = 51, \quad n_1(1 - p_1) = 49$$
$$n_2 p_2 = 44, \quad n_2(1 - p_2) = 56$$

(b) **What is the probability that** $\hat{p}_1 < \hat{p}_2$?

Example: Difference in Proportions - Two Towns

In one town, 51% of voters are conservative; in another, 44% are conservative. A random sample of 100 voters is taken from each town.

(a) **Is a normal model appropriate for** $\hat{p}_1 - \hat{p}_2$?

- **Simple Random Sample:** Assumed for both towns.
- **Independence:** $n_1 = n_2 = 100 < 0.1N$ so we assume independence.
- **Normality:**

$$n_1 p_1 = 51, \quad n_1(1 - p_1) = 49$$
$$n_2 p_2 = 44, \quad n_2(1 - p_2) = 56$$

(b) **What is the probability that** $\hat{p}_1 < \hat{p}_2$?

$$P(\hat{p}_1 - \hat{p}_2 < 0) = P\left(Z < \frac{0 - (0.51 - 0.44)}{\sqrt{\frac{0.51(0.49)}{100} + \frac{0.44(0.56)}{100}}}\right)$$
$$= P(Z < -0.994) = 0.1602$$

Conclusion: There is about a 16% chance the first sample yields a lower proportion than the second.

Sampling Distribution for a Difference in Sample Means

Suppose we take two independent random samples:

- $\overline{x}_1$ is the mean of a sample of size n_1, from a population with mean μ_1 and standard deviation σ_1
- $\overline{x}_2$ is the mean of a sample of size n_2, from a population with mean μ_2 and standard deviation σ_2

We are interested in the statistic $\overline{x}_1 - \overline{x}_2$

Assumptions:

- **Random Sampling:** Each sample is randomly drawn
- **Independence:** Each sample satisfies $n_1 < 0.1N_1$, $n_2 < 0.1N_2$
- **Normality:** Either:
  - Both populations are approximately normal
  - OR sample sizes are large: $n_1 \geq 30$ and $n_2 \geq 30$

- What is $E(\overline{x}_1 - \overline{x}_2)$?
- What is $\text{Var}(\overline{x}_1 - \overline{x}_2)$?
- What is the sampling distribution of $\overline{x}_1 - \overline{x}_2$?

Difference in Sample Means (Unknown Variances)

When population standard deviations σ_1 and σ_2 are unknown, we use the sample standard deviations s_1 and s_2 to estimate them.
Sampling Distribution:

$$\frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_{\text{df}}$$

Degrees of Freedom (df):

$$\min(n_1 - 1, n_2 - 1) \leq \text{df} \leq n_1 + n_2 - 2$$

Which degree of freedom would be the most conservative?
Conditions:

- **Random Sampling:** Both samples are independently and randomly drawn.
- **Independence:** $n_1 < 10\%$ of N_1, $n_2 < 10\%$ of N_2
- **Normality:** Each sample is from a normal population or both $n_1, n_2 \geq 30$

Welch-Satterthwaite Approximation

When population variances are unknown and unequal, we estimate the degrees of freedom using the Welch-Satterthwaite formula:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Use in:

$$\frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_{\text{df}}$$

Note: This formula often gives a non-integer df; statistical software typically handles this automatically.

Sampling Distribution with Pooled Variance (Enrichment)

Suppose we take two independent random samples from two populations, and we assume that the population variances are equal:

$$\sigma_1^2 = \sigma_2^2 = \sigma^2$$

We estimate the common variance using the **pooled sample variance**:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

If all assumptions are satisfied, then the sampling distribution of the difference in sample means is:

$$\frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

Assumptions:

- Random sampling
- Independence: $n_1 < 10\% N_1$, $n_2 < 10\% N_2$
- Normal populations or large sample sizes
- **Equal population variances**