Chapter 5–6: Limit Theorems & Normal-Based Distributions

Donghyun Ko

May 27, 2026

Chapter 5 (Why averages behave so well). Imagine repeating the same measurement many times. Individual outcomes bounce around, but the *average*

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

starts to settle down. The **Law of Large Numbers (LLN)** explains *why* $\bar{X}_n$ gets close to the true mean μ as n grows. Then the **Central Limit Theorem (CLT)** explains something even more useful: it tells you the *shape and size* of the remaining randomness. Roughly, the error $\bar{X}_n - \mu$ is typically about $\sigma/\sqrt{n}$, and for large n it is well-approximated by a normal curve. This is the reason normal approximations show up everywhere in statistics.

Chapter 6 (What appears when you work with normal data). Once normal distributions enter the picture, three new distributions appear almost automatically because of the operations statisticians use all the time: we *square* deviations to measure variability, we *add* squared deviations across many data points, and we often take *ratios* when we standardize or compare variances. These natural steps produce the **chi-square** (χ^2), t, and F distributions. They are not just extra topics—they are the probability laws behind the most common tools for confidence intervals, hypothesis tests, and ANOVA.

Contents

1 How to read these notes

At first, probability statements in Chapters 5 and 6 can feel confusing because many results look similar but mean different things. The main challenge here is *not* complicated algebra. Instead, it is understanding **what is random**, **what is fixed**, and **what kind of statement is being made**.

Whenever you see a limit such as $Z_n \to Z$, stop and ask a simple question: *Are the random values themselves getting close most of the time (convergence in probability), or is only the overall distribution getting close in shape (convergence in distribution)?* This distinction is essential for understanding the Law of Large Numbers and the Central Limit Theorem.

When you encounter the χ^2, t, or F distributions, try not to treat them as new formulas to memorize. Instead, ask: *What operation produced this distribution?* In these chapters, the answer is almost always one of the following: squaring a normal variable, adding squared deviations, or taking a ratio involving a sample variance. Thinking this way makes the roles of χ^2, t, and F much easier to remember. Throughout these notes, we write Φ for the standard normal cumulative distribution function. We also use $\bar{X}_n$ to denote the sample mean based on n observations and S^2 for the sample variance. Keeping this notation in mind will help you focus on ideas rather than symbols.

2 Limit Theorems

Suppose we repeatedly observe an outcome X from the same stable random mechanism. In practice, the distribution of X is unknown, but we are often interested in learning a simple and interpretable feature of it, most notably its mean $\mu = \mathbb{E}[X]$. The most natural estimator of μ is the sample mean

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

where $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.) observations. Averages appear throughout statistics because they are simple, they use all available data, and they exhibit remarkably regular mathematical behavior. To begin understanding this behavior, we first examine two basic properties. If $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$, then by linearity of expectation,

$$\mathbb{E}[\bar{X}_n] = \mu,$$

so the sample mean is correctly centered at the population mean. Next, we consider variability: how much does $\bar{X}_n$ fluctuate around μ? Since the observations are independent and variances add,

$$\mathrm{Var}(\bar{X}_n) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{\sigma^2}{n}.$$

This single formula already reveals the key idea of this chapter: as the sample size n increases, the variability of the average shrinks at rate $1/n$. Consequently, large deviations of $\bar{X}_n$ from μ become increasingly unlikely.

Theorem

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variables with

$$\mathbb{E}[X_i] = \mu, \qquad \mathrm{Var}(X_i) = \sigma^2 < \infty.$$

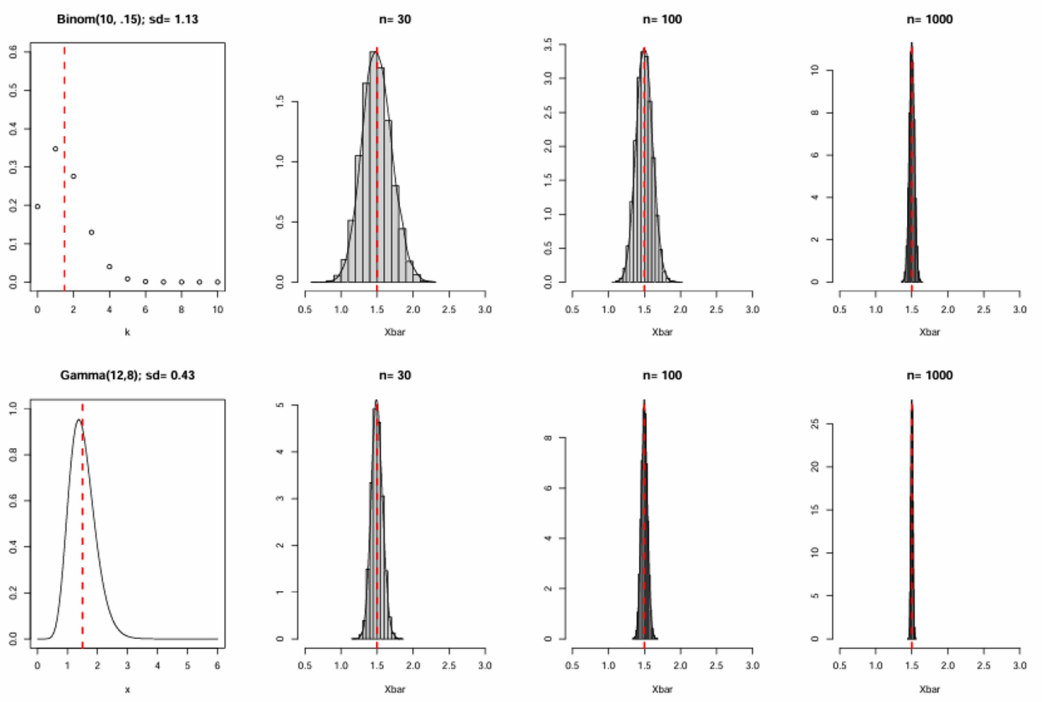
Define the sample mean

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then the following properties hold:

1. **Unbiasedness:**
$$\mathbb{E}[\bar{X}_n] = \mu.$$
2. **Variance reduction:**
$$\mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n}.$$
3. **Exact distribution under normality:** If $X_i \sim \mathcal{N}(\mu, \sigma^2)$, then $\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$.
4. **General case:** If the X_i are not normally distributed, the finite-sample distribution of $\bar{X}_n$ need not be normal. Nevertheless, as $n \to \infty$, the sample mean $\bar{X}_n$ converges to the population mean μ, and its fluctuations around μ decrease at rate $1/\sqrt{n}$.

Figure below illustrates this behavior graphically. For both symmetric (Binomial) and highly skewed (Gamma) underlying distributions, the sampling distribution of $\bar{X}_n$ becomes increasingly concentrated around the true mean μ as n grows. Despite the differing shapes of the original distributions, the sample mean converges to the same limit μ, motivating the formal study of the Law of Large Numbers and the Central Limit Theorem.



2.1 What does it mean for random variables to "converge"?

In calculus, a sequence of real numbers $\{a_n\}$ is said to converge to a limit a if the distance $|a_n - a|$ becomes arbitrarily small as n increases. When dealing with random variables, however, this notion

must be refined. Since random variables are themselves random, there are several meaningful but non-equivalent ways to formalize the idea that a sequence "becomes close" to a limit. The sample mean $\bar{X}_n$, indexed by the sample size n, forms a sequence of random variables. Just as with sequences of numbers, it is natural to ask whether this sequence converges, and if so, in what sense. Unlike the deterministic setting, there is no single universal definition of convergence for random variables. Instead, several distinct notions are used, each capturing a different probabilistic meaning of convergence and each useful in different contexts. In this chapter, we will encounter three fundamental types of convergence for sequences of random variables:

- **Almost sure convergence** (or convergence with probability one), which describes pointwise convergence on almost every outcome.
- **Convergence in probability**, which formalizes the idea that large deviations become increasingly unlikely.
- **Convergence in distribution**, which concerns the convergence of the distributions themselves.

Understanding the distinctions between these modes of convergence is essential for interpreting limit theorems such as the Law of Large Numbers and the Central Limit Theorem, and for understanding precisely in what sense the sample mean $\bar{X}_n$ converges to the population mean μ.

1) Almost sure convergence: Almost sure convergence is the strongest and most literal notion of convergence for random variables. It asserts that, except on a set of probability zero, the sequence behaves like an ordinary convergent numerical sequence.

Theorem

Almost sure convergence. We say that Z_n converges almost surely to Z if

$$\mathbb{P}\big(\{\omega : Z_n(\omega) \to Z(\omega)\}\big) = 1.$$

We write $Z_n \xrightarrow{\text{a.s.}} Z$.

In words, almost sure convergence means that for almost every outcome ω, the numerical sequence $Z_n(\omega)$ converges to $Z(\omega)$. Once ω is fixed, no probability remains: convergence happens deterministically along that sample path. This notion is particularly important in the Strong Law of Large Numbers, which states that the sample mean $\bar{X}_n$ converges almost surely to μ.

Example

Strong Law intuition. Let $X_1, X_2, \ldots$ be i.i.d. with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) < \infty$. The Strong Law of Large Numbers asserts that

$$\bar{X}_n \xrightarrow{\text{a.s.}} \mu.$$

Thus, with probability one, the sequence of averages eventually stays arbitrarily close to μ and never leaves.

2) Convergence in probability: Convergence in probability weakens almost sure convergence by allowing occasional deviations, as long as they become increasingly unlikely.

Theorem

Convergence in probability. We say that Z_n converges in probability to Z if for every $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\big(|Z_n - Z| > \varepsilon\big) = 0.$$

We write $Z_n \xrightarrow{p} Z$.

The phrase "for every $\varepsilon > 0$" is crucial. It means that no matter how strict the accuracy requirement is, the probability of violating it can be made arbitrarily small by taking n large enough. Convergence in probability does not require convergence along individual sample paths. Instead, it ensures that *most realizations* are close to the limit when n is large. This is the mode of convergence appearing in the Weak Law of Large Numbers.

Example

A simple but instructive example. Let

$$\mathbb{P}(Z_n = 0) = 1 - \frac{1}{n}, \qquad \mathbb{P}(Z_n = 1) = \frac{1}{n}.$$

Fix any $\varepsilon > 0$. If $\varepsilon \leq 1$, then the event $\{|Z_n| > \varepsilon\}$ occurs exactly when $Z_n = 1$, so

$$\mathbb{P}(|Z_n| > \varepsilon) = \frac{1}{n} \to 0.$$

If $\varepsilon > 1$, the probability is identically zero. Hence $Z_n \xrightarrow{p} 0$. However, Z_n does not converge almost surely, since the value 1 occurs infinitely often with positive probability.

3) Convergence in distribution: Convergence in distribution is the weakest of the three notions. It does not require Z_n to be close to Z on the same probability space. Instead, it only concerns the shapes of their distributions.

Theorem

Let F_n be the cumulative distribution function (CDF) of Z_n and F the CDF of Z. We say that

$$Z_n \xrightarrow{d} Z$$

if for every real number x at which F is continuous,

$$\lim_{n\to\infty} F_n(x) = F(x).$$

convergence in probability is about values, whereas convergence in distribution is about curves. Two random variables can have nearly identical distributions even if they are never close with high probability. Convergence in distribution is central to the Central Limit Theorem, which states that a suitably normalized version of $\bar{X}_n$ converges in distribution to a normal random variable.

4) A minimal map of implications The three notions of convergence are related but not equivalent. The following implications always hold:

$$Z_n \xrightarrow{\text{a.s.}} Z \;\Rightarrow\; Z_n \xrightarrow{p} Z \;\Rightarrow\; Z_n \xrightarrow{d} Z.$$

In general, none of the reverse implications is true. However, when the limit Z is a constant c,

$$Z_n \xrightarrow{d} c \;\Rightarrow\; Z_n \xrightarrow{p} c.$$

Thus, for convergence to constants, convergence in probability and convergence in distribution essentially coincide.

2.2 Chebyshev's inequality: a bridge from variance to probability

The Law of Large Numbers can be proved in several ways. In an introductory course, the most transparent approach relies on *Chebyshev's inequality*. Its importance is conceptual rather than technical: it provides a direct bridge between variance, which measures average squared deviation, and probability, which measures the likelihood of large deviations. The key idea is simple. If a random variable has small variance, then its values must cluster tightly around their mean. Chebyshev's inequality makes this intuition precise by giving a universal bound on tail probabilities that holds for *any* distribution with finite variance.

Theorem

Chebyshev's inequality. Let Y be a random variable with

$$\mathbb{E}[Y] = m \qquad \text{and} \qquad \mathrm{Var}(Y) = v < \infty.$$

Then for any $\varepsilon > 0$,

$$\mathbb{P}(|Y - m| > \varepsilon) \le \frac{v}{\varepsilon^2}.$$

Proof. Consider the nonnegative random variable

$$W = (Y - m)^2 \ge 0.$$

By Markov's inequality,

$$\mathbb{P}(W \ge \varepsilon^2) \le \frac{\mathbb{E}[W]}{\varepsilon^2}.$$

Since

$$\mathbb{E}[W] = \mathbb{E}[(Y - m)^2] = \mathrm{Var}(Y) = v,$$

and the event $\{W \ge \varepsilon^2\}$ is exactly the same as $\{|Y - m| \ge \varepsilon\}$, we obtain

$$\mathbb{P}(|Y - m| \ge \varepsilon) \le \frac{v}{\varepsilon^2}.$$

□

The proof is short, but its message is powerful: *variance controls tails*. Regardless of the shape of the distribution of Y, the probability of a deviation larger than ε from the mean cannot exceed v/ε^2. Chebyshev's inequality becomes particularly meaningful when applied to the sample mean. If $X_1, \ldots, X_n$ are i.i.d. with mean μ and variance σ^2, then

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

Applying Chebyshev's inequality to $\bar{X}_n$ yields

$$\mathbb{P}(|\bar{X}_n - \mu| > \varepsilon) \le \frac{\sigma^2}{n\varepsilon^2}.$$

For any fixed $\varepsilon > 0$, the right-hand side converges to 0 as $n \to \infty$. This directly implies that

$$\bar{X}_n \xrightarrow{p} \mu,$$

which is precisely the statement of the Weak Law of Large Numbers. Thus, Chebyshev's inequality serves as the key analytical link between the shrinking variance of the sample mean and its convergence in probability to the population mean.

2.3 Weak and Strong Laws of Large Numbers

We now formalize the intuitive statement that *the sample mean gets close to the true mean as the sample size increases*. There are two closely related results that make this idea precise: the Weak Law of Large Numbers (WLLN) and the Strong Law of Large Numbers (SLLN). They differ not in the limit itself, but in the *mode of convergence*.

Theorem

Weak Law of Large Numbers (WLLN). Let $X_1, X_2, \ldots$ be i.i.d. random variables with

$$\mathbb{E}[X_i] = \mu \qquad \text{and} \qquad \mathrm{Var}(X_i) = \sigma^2 < \infty.$$

Then

$$\bar{X}_n \xrightarrow{p} \mu.$$

Equivalently, for every $\varepsilon > 0$,

$$\mathbb{P}\big(|\bar{X}_n - \mu| > \varepsilon\big) \to 0 \quad \text{as } n \to \infty.$$

Proof (Chebyshev-based). We have already shown that

$$\mathbb{E}[\bar{X}_n] = \mu, \qquad \mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n}.$$

Applying Chebyshev's inequality to $Y = \bar{X}_n$ yields

$$\mathbb{P}(|\bar{X}_n - \mu| > \varepsilon) \le \frac{\mathrm{Var}(\bar{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \xrightarrow[n\to\infty]{} 0.$$

This is precisely the statement $\bar{X}_n \xrightarrow{p} \mu$ as $n \to \infty$. □

The Weak Law of Large Numbers states that for large n, *most realizations* of the sample mean are close to μ. However, it does not assert that the sequence $\bar{X}_n(\omega)$ converges for individual outcomes ω, nor does it describe the exact distribution of $\bar{X}_n$. Its conclusion is purely probabilistic: large deviations become unlikely. A stronger statement is given by the Strong Law of Large Numbers.

Theorem

Strong Law of Large Numbers (SLLN). Let $X_1, X_2, \ldots$ be i.i.d. random variables with

$$\mathbb{E}[X_i] = \mu \qquad \text{and} \qquad \mathrm{Var}(X_i) < \infty.$$

Then,

$$\bar{X}_n \xrightarrow{\text{a.s.}} \mu. \qquad \text{equivalently,} \qquad \mathbb{P}\Big(\lim_{n\to\infty} \bar{X}_n = \mu\Big) = 1.$$

The SLLN asserts that, with probability one, the sequence of sample means converges to μ *along almost every sample path.* Once n is large enough, $\bar{X}_n(\omega)$ stays arbitrarily close to μ and never drifts away again, except on a set of outcomes of probability zero. Although the conclusion of the SLLN is stronger than that of the WLLN, both laws describe the same limiting value. The difference lies entirely in the sense of convergence:

- The **WLLN** guarantees closeness with high probability.
- The **SLLN** guarantees pathwise convergence almost surely.

An important practical consequence of the WLLN is that convergence in probability is preserved under continuous transformations. Since the function $g(x) = x^2$ is continuous, the WLLN implication

$$\bar{X}_n \xrightarrow{p} \mu$$

immediately yields

$$\bar{X}_n^2 \xrightarrow{p} \mu^2.$$

More generally, for any continuous function g, we have

$$g(\bar{X}_n) \xrightarrow{p} g(\mu).$$

This type of reasoning is frequently used to derive convergence results for functions of sample averages without repeating full proofs. Thus, while the SLLN provides a strong, pathwise guarantee of stabilization, the WLLN already supplies many useful asymptotic implications for statistical estimators through convergence in probability. In particular, the SLLN implies the WLLN, since almost sure convergence always implies convergence in probability. In practice, the WLLN explains why averaging reduces noise in large samples, while the SLLN justifies treating long-run averages as stable, deterministic quantities. Neither law describes the detailed shape of the distribution of $\bar{X}_n$. That finer description is provided by the Central Limit Theorem, which explains how $\bar{X}_n$ fluctuates around μ at the $1/\sqrt{n}$ scale.

2.4 From convergence to convergence of functions (continuous mapping idea)

In statistics, we often plug $\bar{X}_n$ into functions: squares, logs, exponentials, and so on. The good news is that if $\bar{X}_n$ converges to μ, then any reasonable continuous function of it converges to the corresponding function of μ.

Theorem

Continuous mapping principle (practical form). If $Z_n \xrightarrow{p} Z$ and g is continuous, then

$$g(Z_n) \xrightarrow{p} g(Z).$$

If $Z_n \xrightarrow{d} Z$ and g is continuous, then

$$g(Z_n) \xrightarrow{d} g(Z).$$

This principle is one of the reasons convergence in probability is such a powerful concept: it behaves like ordinary limits when you apply continuous transformations.

2.5 Central Limit Theorem: why normal approximations appear

The WLLN tells us that $\bar{X}_n$ is close to μ for large n. But practitioners need more than that. We want to quantify uncertainty: how big are typical errors $\bar{X}_n - \mu$? For inference, we want approximate probabilities of the form $\mathbb{P}(\bar{X}_n \leq a)$. The CLT answers these questions by describing the *shape* of the fluctuations.

Theorem

Central Limit Theorem. Let $X_1, X_2, \ldots, X_n$ be iid random variables with

$$\mathbb{E}[X_i] = \mu \qquad \text{and} \qquad \text{Var}(X_i) = \sigma^2 < \infty.$$

Then, as $n \to \infty$,

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

Equivalently, if $\Phi(x)$ denotes the cumulative distribution function of the standard normal distribution $N(0,1)$, then for every $x \in \mathbb{R}$,

$$\mathbb{P}\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq x\right) \longrightarrow \Phi(x).$$

In particular, for large n, the sample mean $\bar{X}_n$ behaves approximately like a normal random variable with mean μ and variance σ^2/n:

$$\bar{X}_n \approx N\left(\mu, \frac{\sigma^2}{n}\right).$$

A helpful way to read the Central Limit Theorem is to focus on *scaling*. The difference $\bar{X}_n - \mu$ typically has magnitude on the order of $\sigma/\sqrt{n}$. This explains why the standard error of the mean is $\sigma/\sqrt{n}$ and why statistical errors shrink at rate $1/\sqrt{n}$ rather than the faster $1/n$ rate suggested by variance alone. One practical implication of this scaling is that the CLT allows us to approximate tail probabilities for the sample mean. For large n, we may write

$$\mathbb{P}\big(|\bar{X}_n - \mu| > k\sigma/\sqrt{n}\big) \;\approx\; \mathbb{P}(|Z| > k), \qquad Z \sim N(0,1),$$

which can be read directly from standard normal tables. In contrast, Chebyshev's inequality yields the much cruder bound

$$\mathbb{P}\big(|\bar{X}_n - \mu| > k\sigma/\sqrt{n}\big) \leq \frac{1}{k^2},$$

highlighting how the CLT provides far sharper and more informative approximations for large deviations when the sample size is sufficiently large. Another important implication is that many commonly used distributions can be viewed as sums of independent random variables. For example, Binomial(n,p), Gamma(n,λ), and Negative Binomial(n,p) random variables can all be expressed as sums of i.i.d. components. The CLT therefore allows these distributions to be approximated by a normal distribution when the number of components is large, even when the individual summands are not normal. This perspective explains the widespread appearance of normal approximations throughout statistics and probability.

Example

Normal approximation to a binomial probability (with continuity correction). Let $S \sim \text{Binomial}(n, p)$. We can write $S = \sum_{i=1}^{n} X_i$ where $X_i \sim \text{Bernoulli}(p)$ i.i.d. Then $\mathbb{E}[S] = np$ and $\text{Var}(S) = np(1-p)$. The CLT suggests that for large n,

$$\frac{S - np}{\sqrt{np(1-p)}} \approx N(0,1), \qquad \text{or} \qquad S \approx N\big(np,\ np(1-p)\big).$$

Suppose we want $\mathbb{P}(S \le k)$. Because S is discrete but the normal is continuous, we improve the approximation by matching the area under the curve to the probability mass. The standard continuity correction replaces k by $k + 0.5$:

$$\mathbb{P}(S \le k) \approx \Phi\left(\frac{k + 0.5 - np}{\sqrt{np(1-p)}}\right).$$

Similarly, for an upper tail,

$$\mathbb{P}(S \ge k) \approx 1 - \Phi\left(\frac{k - 0.5 - np}{\sqrt{np(1-p)}}\right).$$

This example matters because it shows the CLT in action: even though S is not normal, its distribution becomes close to a normal curve when n is large, especially when p is not too close to 0 or 1.

3 Distributions Derived from the Normal

This section can feel like a list of new distributions, but there is a simple storyline behind them. When we estimate means and variances from normal data, two fundamental objects appear:

- **Sums of squared standardized deviations**
- **Ratios of a mean deviation to an estimated standard deviation**

Because squares and ratios appear naturally in these statistics, the distributions that govern them are not optional; they are forced upon us by the algebra. We begin with the simplest case: a single standard normal variable.

3.1 From $N(0,1)$ to χ^2

If $Z \sim N(0,1)$, then Z can be positive or negative, symmetric around 0. Squaring removes the sign and produces a nonnegative, right-skewed variable. That new variable is the building block of the chi-square family.

Theorem

Chi-square distribution: definitions and core properties.

- If $Z \sim N(0,1)$, then $U = Z^2$ is said to have a chi-square distribution with 1 degree of freedom:

$$U \sim \chi^2_1.$$

- If $U_1, \ldots, U_m$ are independent χ^2_1 variables, then their sum is chi-square with m degrees of freedom:

$$V = \sum_{i=1}^{m} U_i \sim \chi^2_m.$$

The mean and variance are

$$\mathbb{E}[\chi^2_m] = m, \qquad \mathrm{Var}(\chi^2_m) = 2m.$$

While the definition above already tells you how χ^2 is built (sum of squared standard normals), it is also helpful to remember what it represents: it is a standardized measure of total squared fluctuation.

Example

Why "standardizing then squaring" gives χ^2_1. Let $X \sim N(\mu, \sigma^2)$. Standardize:

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1).$$

Now square:

$$Z^2 = \left(\frac{X - \mu}{\sigma}\right)^2 \sim \chi^2_1.$$

This is the single-observation prototype of what happens for a full sample: sums of squared standardized deviations produce chi-square distributions.

3.2 Normal samples: $\bar{X}$ is normal, and S^2 leads to χ^2

Now assume a normal sampling model:

$$X_1, \ldots, X_n \text{ i.i.d. } N(\mu, \sigma^2).$$

Define

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

The normal family has a remarkable property: linear combinations of normal variables are again normal. Since $\bar{X}$ is a linear combination of the X_i, we obtain an *exact* distribution (not an approximation):

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

The sample variance S^2 is not normal; it is built from squares. This is where chi-square enters.

Theorem

Scaled sample variance has a chi-square distribution. If $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$, then

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

Proof (textbook-style, with the key decomposition). Let $Z_i = (X_i - \mu)/\sigma$. Then $Z_1, \ldots, Z_n$ are i.i.d. $N(0,1)$, so by definition

$$\sum_{i=1}^{n} Z_i^2 \sim \chi_n^2.$$

Rewrite the left-hand side in terms of X_i:

$$\sum_{i=1}^{n} Z_i^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2.$$

Now use the fundamental identity that separates "total variation around μ" into "variation around $\bar{X}$" plus "variation of $\bar{X}$ around μ":

$$\sum_{i=1}^{n} (X_i - \mu)^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2.$$

Divide by σ^2:

$$\sum_{i=1}^{n} Z_i^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \bar{X})^2 + \frac{n(\bar{X} - \mu)^2}{\sigma^2}.$$

The first term is exactly $(n-1)S^2/\sigma^2$, because $\sum (X_i - \bar{X})^2 = (n-1)S^2$:

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \bar{X})^2 = \frac{(n-1)S^2}{\sigma^2}.$$

The second term is the square of a standard normal variable: since $\bar{X} \sim N(\mu, \sigma^2/n)$,

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim N(0,1) \quad \Rightarrow \quad \frac{n(\bar{X} - \mu)^2}{\sigma^2} \sim \chi_1^2.$$

Finally, a special property of normal samples is that $\bar{X}$ and S^2 are independent, which implies that the two terms in the decomposition are independent. We therefore have a representation of a χ_n^2 variable as a sum of two independent chi-square variables:

$$\chi_n^2 = \frac{(n-1)S^2}{\sigma^2} + \chi_1^2.$$

The only way this can happen is if the first term is χ_{n-1}^2. Thus

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2.$$

□

Example

Turning the theorem into an actual probability. Assume $X_1, \dots, X_{10} \sim N(\mu, \sigma^2)$ and compute

$$\mathbb{P}(S^2 \le 0.5\,\sigma^2).$$

Use the chi-square fact with $n = 10$:

$$\frac{(10-1)S^2}{\sigma^2} = \frac{9S^2}{\sigma^2} \sim \chi^2_9.$$

Then

$$\mathbb{P}(S^2 \le 0.5\,\sigma^2) = \mathbb{P}\left(\frac{9S^2}{\sigma^2} \le \frac{9(0.5\sigma^2)}{\sigma^2}\right) = \mathbb{P}(\chi^2_9 \le 4.5).$$

The final number is obtained from a χ^2 table or software (e.g., calculator, R, Python). The important skill is setting up the probability correctly; the theorem tells you exactly how to convert a statement about S^2 into a statement about χ^2.

3.3 From normal and chi-square to the t distribution

In practice, σ is usually unknown. When we standardize $\bar{X}$ using the true σ, we obtain a standard normal:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

But because σ is unknown, we replace it by S. This produces the statistic

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}},$$

which is no longer normal because S is random. The t distribution is exactly the distribution of this ratio under normal sampling.

Theorem

Definition of the t distribution. Let $Z \sim N(0,1)$ and $U \sim \chi^2_\nu$ be independent. Then

$$T = \frac{Z}{\sqrt{U/\nu}}$$

has a t distribution with ν degrees of freedom, written $T \sim t_\nu$.

Theorem

The classical t statistic from a normal sample. If $X_1, \dots, X_n$ are i.i.d. $N(\mu, \sigma^2)$, then

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

Proof (by matching to the definition). Define

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

Because $\bar{X} \sim N(\mu, \sigma^2/n)$, we have $Z \sim N(0,1)$. Next define

$$U = \frac{(n-1)S^2}{\sigma^2}.$$

From the previous chi-square theorem, $U \sim \chi^2_{n-1}$. For a normal sample, $\bar{X}$ and S^2 are independent, so Z and U are independent.
Now rewrite the statistic:

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \cdot \frac{\sigma}{S} = Z \cdot \frac{1}{\sqrt{S^2/\sigma^2}} = \frac{Z}{\sqrt{U/(n-1)}}.$$

This is exactly the t definition with $\nu = n-1$. Therefore the statistic has the t_{n-1} distribution. □

Example

Why the t curve looks like a wider normal curve. The difference between Z and T is the denominator. In Z, the denominator is fixed ($\sigma/\sqrt{n}$). In T, we divide by $S/\sqrt{n}$. When S happens to be small, T becomes larger in magnitude; when S is large, T becomes smaller. This extra randomness creates heavier tails than the normal distribution. As n increases, S becomes a more stable estimate of σ, so the t distribution approaches the standard normal distribution.

3.4 From chi-square ratios to the F distribution

The final normal-based distribution arises when we compare two independent variance-like quantities. If we have two independent normal samples, each sample variance can be converted into a chi-square variable by scaling. Taking a ratio of these scaled variances produces an F distribution.

Theorem

Definition of the F distribution. Let $U \sim \chi^2_m$ and $V \sim \chi^2_n$ be independent. Then

$$W = \frac{U/m}{V/n}$$

has an F distribution with (m,n) degrees of freedom, written $W \sim F_{m,n}$.

Example

Ratio of sample variances from two independent normal samples. Let $X_1, \ldots, X_m$ be i.i.d. $N(\mu_X, \sigma_X^2)$ and $Y_1, \ldots, Y_n$ be i.i.d. $N(\mu_Y, \sigma_Y^2)$, and assume the two samples are independent. Let S_X^2 and S_Y^2 be the sample variances.
From the chi-square theorem applied to each sample,

$$\frac{(m-1)S_X^2}{\sigma_X^2} \sim \chi^2_{m-1}, \qquad \frac{(n-1)S_Y^2}{\sigma_Y^2} \sim \chi^2_{n-1},$$

and these two chi-square variables are independent because the samples are independent.
Form the ratio

$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} = \frac{\big((m-1)S_X^2/\sigma_X^2\big)/(m-1)}{\big((n-1)S_Y^2/\sigma_Y^2\big)/(n-1)} \sim F_{m-1,n-1}.$$

In the special case $\sigma_X^2 = \sigma_Y^2$, this simplifies to

$$\frac{S_X^2}{S_Y^2} \sim F_{m-1,n-1}.$$

This fact is the core probabilistic reason why the F distribution appears in classical tests for comparing variances and in ANOVA.