
CS/MATH 3414 — Numerical Methods

Lecture 8 Probability and Statistics

Virginia Tech Department of Computer Science, Fall 2026

Instructor: Sumeet Khatri

Date: October 13, 2026

1	Introduction	2
2	Basic notions of probability and discrete random variables	3
2.1	Outcomes, events, and probability	4
2.2	Conditional probability and independence	5
2.3	Random variables and their distributions	6
2.4	Four useful distribution models	7
2.5	Histograms: from observations to a distribution	9
2.6	Joint distributions and marginal distributions	11
3	Expectation and variance	13
3.1	Covariance and independence	15
3.2	Expectation of the time to a first event	16
4	Sampling, estimation, and the law of large numbers	17
4.1	From a probability model to a sequence of observations	17
4.2	Estimating the mean	18
4.3	Estimating the variance	19
4.4	Why averaging works: the law of large numbers	19
4.5	Estimating probabilities	20
4.6	Exam bell curves and the normal distribution	21
4.7	The central limit theorem and sampling error bars	31

Learning Objectives

- Describe discrete probability models, interpret histograms, and calculate expectation and variance.
- Sample a distribution and estimate its mean, variance, and event probabilities.
- State and interpret laws of large numbers and finite-sample concentration bounds.
- Generate reproducible pseudorandom samples using NumPy.

Primary References

- [1] W. J. Stewart, *Probability, Markov Chains, Queues, and Simulation: The Mathematical Basis of Performance Modeling* (Princeton University Press, 2009).
- [2] C. M. Grinstead and J. L. Snell, *Introduction to probability*, 2nd ed., Chapters 1, 4, 6, 8, and 9: discrete probability, conditioning, moments, laws of large numbers, and the central limit theorem (American Mathematical Society, 1997), https://chance.dartmouth.edu/teaching_aids/books_articles/probability_book/pdf.html (visited on 10/10/2026).

Primary References (continued)

- [3] J. Duchi, *Hoeffding's inequality*, CS229 supplemental lecture notes: Markov, Chebyshev, and bounded-sum inequalities, Stanford University, https://cs229.stanford.edu/notes_archive/cs229-notes-all/hoeffding.pdf (visited on 10/10/2026).
- [4] National Institutes of Health, *Understanding clinical studies*, <https://www.nih.gov/about-nih/science-health-public-trust/tools/understanding-clinical-studies> (visited on 10/10/2026).
- [5] National Cancer Institute, *Cancer screening overview (pdq): patient version*, Section on absolute risk, relative risk, and odds ratios, <https://www.cancer.gov/about-cancer/screening/patient-screening-overview-pdq> (visited on 10/10/2026).
- [6] National Institute of Standards and Technology, *What are confidence intervals?*, NIST/SEMATECH e-Handbook of Statistical Methods, Section 7.1.4, <https://www.itl.nist.gov/div898/handbook/prc/section1/prc14.htm> (visited on 10/11/2026).
- [7] National Institute of Standards and Technology, *Normal distribution*, <https://www.itl.nist.gov/div898/handbook/eda/section3/eda3661.htm> (visited on 10/11/2026).

1 Introduction

A news report says that a treatment “reduces the risk by 40%.” Another says that people with a certain habit are “twice as likely” to develop a condition. A survey reports that 58% of respondents support a proposal. These statements connect observations on a finite group to claims about a larger population. To interpret them, we need a language for uncertainty and methods for deciding how much information a sample provides.

Another familiar setting is exam grading. An instructor may announce that grades will be “curved,” sometimes using a bell-shaped distribution as a reference. What does a score of 80 mean in a class with an average of 65? The answer depends on how widely the scores are spread, as well as on the grading rule. This example motivates the mean, standard deviation, and *normal (Gaussian) distribution*: a symmetric bell-shaped model that can relate a score’s distance from the average to its percentile. Curving does not by itself make exam scores normally distributed. In Section 4.6, we will use this grading example to introduce the normal distribution, which also describes the approximate behavior of sample averages through the central limit theorem.

Probability starts with a model and predicts the behavior of observations. Statistics starts with observations and estimates features of the model. Numerical computation links the two: we can generate samples, calculate estimates, repeat entire studies on a computer, and investigate how the answers change with the sample size.

In a *randomized controlled trial*, eligible participants are randomly assigned to treatment and comparison groups. The comparison may use another treatment, usual care, or a placebo when appropriate. Random assignment helps balance other relevant characteristics in the groups; it does not guarantee exact balance in a particular trial. Blinding, consistent outcome measurement, and accounting for missing outcomes help prevent other sources of bias. A well-conducted randomized trial can support a causal conclusion about the assigned treatment [4].

An *observational study* records existing exposures or behaviors. People who exercise frequently, for example, may differ from other people in age, health, or access to care. A difference in outcomes may reflect these other variables. Such a variable is a *confounder* when it influences the exposure comparison and the outcome. Establishing causation from observational data requires additional design and assumptions; the observed association alone is insufficient.

Example 1. Suppose researchers want to determine whether a new preventive treatment reduces the chance of developing a particular infection during the following year. They conduct a hypothetical randomized trial involving 2000 participants, assigning each person at random to one of two groups:

- The **treatment group** contains 1000 participants who receive the new treatment.
- The **control group** contains 1000 participants who receive a placebo: an inactive preparation resembling the treatment. This group provides a comparison for what happens without the new treatment.

Both groups otherwise receive the same usual care. Researchers follow everyone for the entire year and use the same criteria to determine whether each participant develops the infection.

Here, **risk** means the probability that a person develops the infection during that year. The proportion of participants who develop the infection provides an estimate of this risk in each group.

Suppose 100 people in the control group and 60 people in the treatment group develop the infection. The estimated one-year risks are therefore

$$\hat{p}_{\text{control}} = \frac{100}{1000} = 0.10, \quad \hat{p}_{\text{treatment}} = \frac{60}{1000} = 0.06. \quad (1.1)$$

Thus, the observed risk is 10% in the control group and 6% in the treatment group.

The **absolute risk reduction** is the difference between these risks:

$$\hat{p}_{\text{control}} - \hat{p}_{\text{treatment}} = 0.10 - 0.06 = 0.04. \quad (1.2)$$

This is a reduction of *four percentage points*. In this trial, there were 40 fewer people with the infection per 1000 participants in the treatment group.

The **relative risk reduction** expresses that difference as a fraction of the control group's risk:

$$\frac{\hat{p}_{\text{control}} - \hat{p}_{\text{treatment}}}{\hat{p}_{\text{control}}} = \frac{0.04}{0.10} = 0.40 = 40\%. \quad (1.3)$$

Equivalently, the **risk ratio** is

$$\frac{\hat{p}_{\text{treatment}}}{\hat{p}_{\text{control}}} = \frac{0.06}{0.10} = 0.60, \quad (1.4)$$

so the treatment group's observed risk is 60% of the control group's observed risk.

A headline saying “the treatment reduced risk by 40%” is reporting the *relative* reduction. It does not mean that 40 fewer people out of every 100 developed the infection. The observed change was from 10 to 6 per 100, a difference of 4 per 100.

These percentages summarize one finite trial. They estimate population risks; another trial would generally produce different numbers. Assessing how much uncertainty accompanies such estimates is one of the main purposes of statistics.

2 Basic notions of probability and discrete random variables

Before we can interpret a reported risk or simulate an experiment, we need a mathematical description of the possible outcomes and their probabilities. A probability model does not predict exactly what will happen on the next trial. It specifies how uncertainty is distributed among the possible outcomes. We begin with

discrete models, for which the relevant probabilities can be described using sums rather than integrals. See [2] for additional examples and exercises.

We will model a binary outcome by a variable X that equals 1 if the outcome occurs and 0 otherwise. The population probability p is unknown in a real study. A sample gives an estimate $\hat{p}$, and repeating the study would usually give a different estimate. Our mathematical questions are:

1. How do we describe the distribution of X ?
2. What do the mean and variance tell us about its behavior?
3. Why should the sample proportion approximate p ?
4. How many observations are enough to make a specified error unlikely?
5. How can a computer simulate this process reproducibly?

The law of large numbers answers the third question. Concentration inequalities give explicit answers to the fourth under stated assumptions. Neither corrects a biased measurement process or an inappropriate model.

2.1 Outcomes, events, and probability

Definition 1 (Discrete probability space). A **sample space** Ω is the set of possible outcomes of an experiment. In a discrete probability space, Ω is finite or countably infinite^a. An **event** A is a subset of Ω . A probability assignment $\Pr$ satisfies the following axioms:

- (i) $\Pr[A] \geq 0$ for every event A .
- (ii) $\Pr[\Omega] = 1$.
- (iii) If $A_1, A_2, \dots$ are pairwise disjoint events, then

$$\Pr\left[\bigcup_{j=1}^{\infty} A_j\right] = \sum_{j=1}^{\infty} \Pr[A_j]. \quad (2.1)$$

Equivalently, we may assign a nonnegative weight q_ω to each $\omega \in \Omega$, with $\sum_{\omega \in \Omega} q_\omega = 1$, and set

$$\Pr[A] = \sum_{\omega \in A} q_\omega. \quad (2.2)$$

^aThe word “countably” includes a set such as $\{0, 1, 2, \dots\}$.

For a finite sample space with m equally likely outcomes, the familiar counting formula is valid:

$$\Pr[A] = \frac{|A|}{|\Omega|} = \frac{|A|}{m}. \quad (2.3)$$

Here $|A|$ denotes the number of elements of A . Equation (2.3) requires the equal-likelihood assumption; counting outcomes alone is not sufficient.

Example 2 (A fair six-sided die). Let $\Omega = \{1, 2, 3, 4, 5, 6\}$, with probability $1/6$ assigned to each outcome. For $A = \{2, 4, 6\}$ and $B = \{4, 5, 6\}$, the events “even” and “at least four” have probabilities $\Pr[A] = \Pr[B] = 1/2$. Their intersection is $A \cap B = \{4, 6\}$, so $\Pr[A \cap B] = 1/3$.

The complement A^c means that A does not occur; the union $A \cup B$ means that at least one of A and B occurs. The axioms imply

$$\Pr[\emptyset] = 0, \quad (2.4)$$

$$\Pr[A^c] = 1 - \Pr[A], \quad (2.5)$$

$$\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B]. \quad (2.6)$$

The subtraction in the union formula removes the outcomes counted twice. More generally, the **union bound** gives

$$\Pr\left[\bigcup_{j=1}^m A_j\right] \leq \sum_{j=1}^m \Pr[A_j]. \quad (2.7)$$

This is useful when we can bound individual failure probabilities but do not know exactly how the failures overlap.

2.2 Conditional probability and independence

Definition 2 (Conditional probability and independent events). If $\Pr[B] > 0$, the probability of A **conditional on** B is

$$\Pr[A \mid B] = \frac{\Pr[A \cap B]}{\Pr[B]}. \quad (2.8)$$

Events A and B are **independent** if

$$\Pr[A \cap B] = \Pr[A] \Pr[B]. \quad (2.9)$$

When $\Pr[B] > 0$, independence is equivalent to $\Pr[A \mid B] = \Pr[A]$.

Conditioning restricts attention to the outcomes in B and rescales their probabilities to sum to one. In the die example, knowing that the result is at least four leaves the outcomes $\{4, 5, 6\}$, so the conditional probability of an even result is $2/3$, rather than $1/2$.

Key Idea

Disjoint does not mean independent. Disjoint events cannot occur together. Independent events do not change one another’s probabilities. If two disjoint events both have positive probability, then $\Pr[A \cap B] = 0$ whereas $\Pr[A] \Pr[B] > 0$, so they cannot be independent.

Proposition 3 (Total probability and Bayes' rule). Suppose $B_1, \dots, B_m$ partition the sample space into disjoint events of positive probability. Then

$$\Pr[A] = \sum_{j=1}^m \Pr[A | B_j] \Pr[B_j]. \quad (2.10)$$

If $\Pr[A] > 0$, then

$$\Pr[B_k | A] = \frac{\Pr[A | B_k] \Pr[B_k]}{\sum_{j=1}^m \Pr[A | B_j] \Pr[B_j]}. \quad (2.11)$$

Proof. The sets $A \cap B_j$ are disjoint and their union is A . Apply additivity and then the definition of conditional probability. For Bayes' rule, substitute the resulting expression for $\Pr[A]$ into $\Pr[B_k | A] = \Pr[A \cap B_k] / \Pr[A]$. ■

2.3 Random variables and their distributions

Definition 4 (Discrete random variable and probability mass function). A **random variable** is a function $X : \Omega \rightarrow \mathbb{R}$ that assigns a numerical value to an outcome. It is **discrete** if its set of possible values is finite or countably infinite. Its **probability mass function** (pmf) is

$$p_X(x) = \Pr[X = x] = \Pr[\{\omega \in \Omega : X(\omega) = x\}]. \quad (2.12)$$

Its support is the set $\mathcal{S}_X = \{x : p_X(x) > 0\}$. The pmf satisfies

$$p_X(x) \geq 0, \quad \sum_{x \in \mathcal{S}_X} p_X(x) = 1. \quad (2.13)$$

Distinguish three objects: ω is an outcome, X is the rule that converts outcomes into numbers, and $x = X(\omega)$ is a realized value. If an experiment tosses two coins, an outcome might be $\omega = HT$, while the random variable “number of heads” has realized value $x = 1$. Multiple outcomes may give the same value of X .

Example 3 (A running three-point distribution). Throughout the note, we will use a random variable taking the values 0, 1, and 2 with the following probabilities:

x	0	1	2
$p_X(x)$	1/5	1/2	3/10

For instance, X might represent the number of failed components in a small system, according to an assumed model. The probability of at least one failure is $\Pr[X \geq 1] = 1/2 + 3/10 = 4/5$. These probabilities describe the distribution; a sample from the distribution is one value such as 0, 1, or 2.

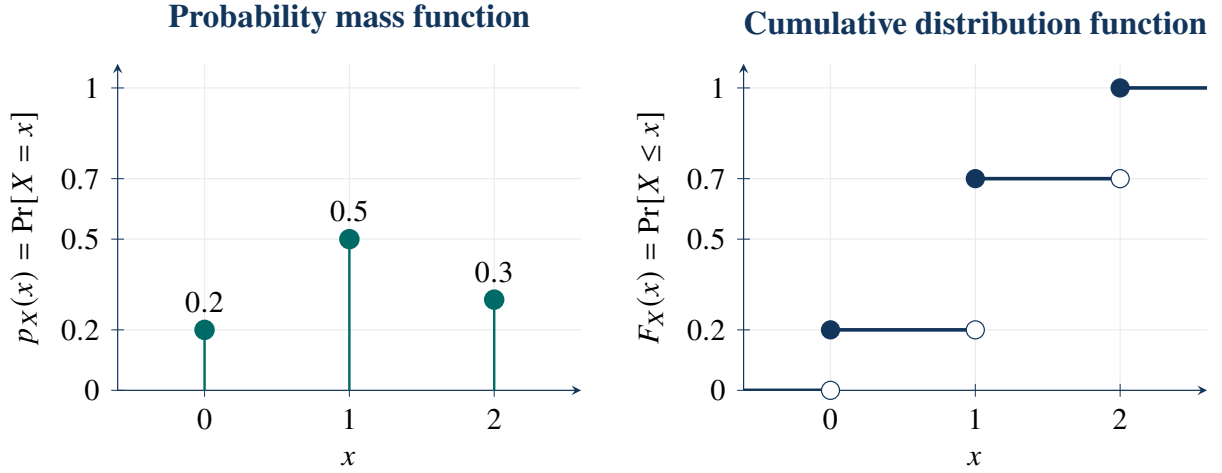


FIGURE 1: Two representations of the same discrete distribution. The pmf assigns masses to 0, 1, and 2. The CDF accumulates those masses. Filled circles include an endpoint and open circles exclude it, so the value at each jump belongs to the higher step.

Definition 5 (Cumulative distribution function). The **cumulative distribution function** (CDF) of X is

$$F_X(t) = \Pr[X \leq t] = \sum_{\substack{x \in S_X \\ x \leq t}} p_X(x), \quad t \in \mathbb{R}. \quad (2.14)$$

The CDF is nondecreasing and right-continuous. For a discrete variable, it has a jump of size $p_X(x)$ at each support point x . Our running example has

$$F_X(t) = \begin{cases} 0, & t < 0, \\ 1/5, & 0 \leq t < 1, \\ 7/10, & 1 \leq t < 2, \\ 1, & t \geq 2. \end{cases} \quad (2.15)$$

The pmf records the probability *at* each value, whereas the CDF accumulates the probability *up to and including* a value. This distinction will become useful when we turn uniform random numbers into samples from another distribution.

2.4 Four useful distribution models

Example 4 (Lost packets in a communication network). An application sends data over a network in small units called *packets*. A packet may arrive successfully or be lost, in which case it must be sent again. We would like to answer two questions: Will a particular packet be lost? How many packets in a batch will need to be sent again?

One packet: a Bernoulli random variable. Suppose a packet is lost with probability p . Define

$$X = \begin{cases} 1, & \text{if the packet is lost,} \\ 0, & \text{if the packet arrives successfully.} \end{cases} \quad (2.16)$$

Then

$$\Pr[X = 1] = p, \quad \Pr[X = 0] = 1 - p. \quad (2.17)$$

A random variable with these two possible values and probabilities has a **Bernoulli distribution**, written $X \sim \text{Bernoulli}(p)$.

The outcome labeled 1 is conventionally called a “success” in probability terminology. This simply means that the event being counted occurred; here, that event is a lost packet.

A batch of packets: a binomial random variable. Now consider the first transmission of each of n packets. Assume that each packet is lost with the same probability p , independently of the other packets. Let X_i indicate whether packet i is lost. The total number of lost packets is

$$Y = X_1 + \cdots + X_n. \quad (2.18)$$

Thus Y counts how many of the original packets need to be sent again. It can take the values $0, 1, \dots, n$. For exactly k packets to be lost, we must choose which k packets are lost. There are $\binom{n}{k}$ possible choices. By independence, each particular choice has probability $p^k(1-p)^{n-k}$. Therefore,

$$\Pr[Y = k] = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \dots, n. \quad (2.19)$$

This is the **binomial distribution**, written $Y \sim \text{Binomial}(n, p)$.

For example, suppose $n = 10$ and $p = 0.1$. The probability that exactly two of the ten packets are lost on their first transmission is

$$\Pr[Y = 2] = \binom{10}{2} (0.1)^2 (0.9)^8 \approx 0.1937. \quad (2.20)$$

There is therefore approximately a 19.4% chance that exactly two packets in the batch will need to be sent again.

The Bernoulli variable records whether loss occurs for one packet; the binomial variable counts losses across a fixed number of independent packets with the same loss probability.

Definition 6 (Bernoulli and binomial distributions). A **Bernoulli** random variable X with parameter $p \in [0, 1]$ records whether an event occurs:

$$\Pr[X = 1] = p, \quad \Pr[X = 0] = 1 - p. \quad (2.21)$$

We write $X \sim \text{Bernoulli}(p)$. A **binomial** random variable X with parameters m and p counts successes in m independent Bernoulli trials with the same parameter. For $0 < p < 1$,

$$\Pr[X = k] = \binom{m}{k} p^k (1-p)^{m-k}, \quad k = 0, 1, \dots, m. \quad (2.22)$$

We write $X \sim \text{Binomial}(m, p)$. When $p = 0$, $X = 0$ with probability one; when $p = 1$, $X = m$ with probability one.

The binomial coefficient counts the arrangements containing exactly k successes. Independence gives probability $p^k(1-p)^{m-k}$ for each such arrangement. For example, the number of heads in ten independent fair coin tosses has a binomial distribution with $m = 10$ and $p = 1/2$. A clinical count can be modeled this way only when the equal-probability and independence assumptions are suitable; sharing a label such as “patient” does not guarantee those assumptions.

A fair die is a third useful model: it is **discrete uniform** on $\{1, \dots, 6\}$, with pmf $1/6$ at each of its six values. Uniformity is a property of the distribution, not a general property of random variables.

Definition 7 (Geometric distribution). Repeat independent Bernoulli trials, each with the same success probability $p \in (0, 1]$. Let T count the number of trials *up to and including the first success*. Then T has a **geometric distribution**, written $T \sim \text{Geometric}(p)$. For $0 < p < 1$, its support is $\{1, 2, 3, \dots\}$ and its pmf is

$$\Pr[T = k] = (1-p)^{k-1}p, \quad k = 1, 2, 3, \dots \quad (2.23)$$

When $p = 1$, the first trial succeeds with certainty, so $\Pr[T = 1] = 1$.

The event $T = k$ requires $k - 1$ failures followed by one success. The event $T > k$ requires failures on all of the first k trials. Therefore, for $0 < p < 1$,

$$\Pr[T > k] = (1-p)^k, \quad k = 0, 1, 2, \dots \quad (2.24)$$

This tail probability tends to zero, so a success eventually occurs with probability one. For $p = 1$, the tail probability is 1 at $k = 0$ and 0 for every $k \geq 1$. If $p = 0$, a first success never occurs, which is why that case is excluded from this waiting-time model.

Some references define a geometric variable to count *failures before* the first success. Under that convention the variable is $T - 1$ and takes values $0, 1, 2, \dots$. Always check the convention. Our variable counts attempts, including the successful one; Section 3.2 turns this count into elapsed time in a practical example.

2.5 Histograms: from observations to a distribution

A probability mass function describes a model. A **histogram** summarizes a collection of observed numerical values: exam scores, delivery times, or the outcomes of a simulation. It helps us see where observations are concentrated, how widely they are spread, and whether the data look roughly symmetric or have a long tail. Constructing this picture requires choosing both the bins and the meaning of the vertical axis.

Definition 8 (Histogram and its normalizations). Let $x_1, \dots, x_n$ be $n \geq 1$ observations, and choose strictly increasing bin edges $b_0 < b_1 < \dots < b_m$ such that every observation lies in $[b_0, b_m]$. Use disjoint bins $B_j = [b_{j-1}, b_j)$, except that the final bin includes its right endpoint: $B_m = [b_{m-1}, b_m]$. The width and count of bin j are

$$w_j = b_j - b_{j-1} > 0, \quad c_j = \#\{i \in \{1, \dots, n\} : x_i \in B_j\}. \quad (2.25)$$

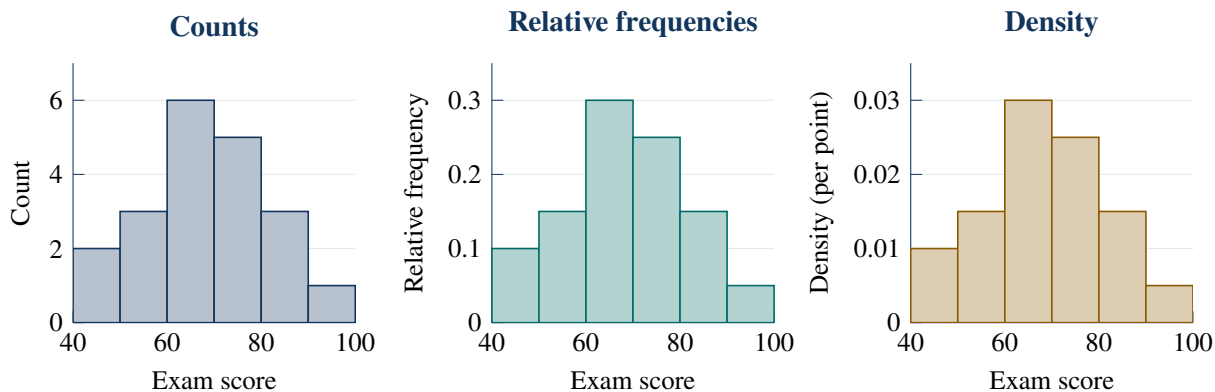


FIGURE 2: Three histograms of the same 20 exam scores using the same width-10 bins. Left: rectangle heights are counts. Middle: heights are observed fractions. Right: heights are fractions divided by bin width, so rectangle areas are observed fractions and the total area is 1. The shapes agree because all bin widths are equal; the vertical scales and their interpretations differ.

A histogram draws a rectangle over each bin. Three common choices of rectangle height are

$$\begin{aligned}
 &\text{count: } c_j, \\
 &\text{relative frequency: } r_j = \frac{c_j}{n}, \\
 &\text{density: } h_j = \frac{c_j}{nw_j}.
 \end{aligned} \tag{2.26}$$

Each observation belongs to exactly one bin, so

$$\sum_{j=1}^m c_j = n, \quad \sum_{j=1}^m r_j = 1, \quad \sum_{j=1}^m h_j w_j = 1. \tag{2.27}$$

The fraction r_j is exactly the fraction of the observed sample in B_j . Under an appropriate sampling model, it estimates the population probability $\Pr[X \in B_j]$. The density height h_j has units inverse to those of x : for exam scores, its units are “per point.” In a density histogram, *area*, rather than height, gives the fraction of observations in a whole bin, because $h_j w_j = r_j$. Density heights need not sum to 1 and can exceed 1 when the bins are sufficiently narrow.

Example 5 (Twenty exam scores). Suppose the following hypothetical exam scores are observed, out of 100:

$$\begin{aligned}
 &42, 48, 51, 55, 58, 60, 62, 64, 65, 66, \\
 &68, 70, 72, 74, 76, 78, 81, 84, 88, 93.
 \end{aligned} \tag{2.28}$$

Use bins of width 10 covering 40 through 100. A score on an interior edge goes into the bin to its right; for example, 60 belongs to $[60, 70)$. The three histogram normalizations give the following heights:

Bin B_j	Count c_j	Fraction r_j	Density h_j
[40, 50)	2	0.10	0.010
[50, 60)	3	0.15	0.015
[60, 70)	6	0.30	0.030
[70, 80)	5	0.25	0.025
[80, 90)	3	0.15	0.015
[90, 100]	1	0.05	0.005

For the bin [60, 70), the count is 6, the relative frequency is $6/20 = 0.30$, and the density is $0.30/10 = 0.030$ per point. Its rectangle has area $10(0.030) = 0.30$. These are three descriptions of the same six observations, not three different data sets.

2.6 Joint distributions and marginal distributions

An experiment can produce several quantities that we want to study together. For example, when tossing two coins, we might record both whether the first coin shows heads and the total number of heads. These quantities are random variables defined on the same sample space. Their individual distributions describe each quantity separately; their **joint distribution** also describes how their values occur together.

Definition 9 (Joint probability mass function). Let X and Y be discrete random variables with supports $\mathcal{S}_X$ and $\mathcal{S}_Y$. Their **joint probability mass function**, or joint pmf, is

$$p_{X,Y}(x, y) = \Pr[X = x, Y = y]. \quad (2.29)$$

The comma inside the probability means that both conditions hold. The joint pmf satisfies

$$p_{X,Y}(x, y) \geq 0, \quad \sum_{x \in \mathcal{S}_X} \sum_{y \in \mathcal{S}_Y} p_{X,Y}(x, y) = 1. \quad (2.30)$$

Some pairs may have probability zero, even when each coordinate is individually a possible value.

Definition 10 (Marginal distributions). The individual distributions of X and Y are called the **marginal distributions** of their joint distribution. Their pmfs are obtained by summing over the other variable:

$$\begin{aligned} p_X(x) &= \Pr[X = x] = \sum_{y \in \mathcal{S}_Y} p_{X,Y}(x, y), \\ p_Y(y) &= \Pr[Y = y] = \sum_{x \in \mathcal{S}_X} p_{X,Y}(x, y). \end{aligned} \quad (2.31)$$

This operation is called **marginalization**.

To understand the first formula, fix x . The event $X = x$ occurs together with exactly one of the possible values of Y . The events $\{X = x, Y = y\}$ are disjoint as y varies, and their union is $\{X = x\}$. Adding their probabilities therefore gives $\Pr[X = x]$. Marginalization does not require independence.

Example 6 (Two quantities from the same coin experiment). Toss two independent fair coins. The four equally likely outcomes are TT, TH, HT, and HH, where the first letter records the first coin. Let $X = 1$ if the first coin shows heads and $X = 0$ otherwise. Let Y be the total number of heads. The corresponding pairs are

$$\begin{aligned} \text{TT} &\mapsto (0, 0), & \text{TH} &\mapsto (0, 1), \\ \text{HT} &\mapsto (1, 1), & \text{HH} &\mapsto (1, 2). \end{aligned} \tag{2.32}$$

Each of these four pairs has probability $1/4$. The joint pmf and its marginals can be displayed in one table:

$p_{X,Y}(x, y)$	$y = 0$	$y = 1$	$y = 2$	$p_X(x)$
$x = 0$	1/4	1/4	0	1/2
$x = 1$	0	1/4	1/4	1/2
$p_Y(y)$	1/4	1/2	1/4	1

Adding across a row gives the marginal probability for X . Adding down a column gives the marginal probability for Y . For example,

$$p_X(0) = \frac{1}{4} + \frac{1}{4} + 0 = \frac{1}{2}, \quad p_Y(1) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}. \tag{2.33}$$

Thus X is Bernoulli with parameter $1/2$, while Y is binomial with parameters 2 and $1/2$.

The table contains more information than these marginals alone. For instance, $X = 0$ and $Y = 2$ are each individually possible, but they cannot occur together: if both coins show heads, the first coin must show heads.

Joint probabilities also determine **conditional distributions**. Whenever $p_Y(y) > 0$,

$$p_{X|Y}(x | y) = \Pr[X = x | Y = y] = \frac{p_{X,Y}(x, y)}{p_Y(y)}. \tag{2.34}$$

In the coin example,

$$\Pr[X = 1 | Y = 2] = \frac{1/4}{1/4} = 1, \quad \Pr[X = 1 | Y = 1] = \frac{1/4}{1/2} = \frac{1}{2}. \tag{2.35}$$

Knowing the total number of heads can therefore change the distribution of the first coin's indicator.

Definition 11 (Independent discrete random variables). Discrete random variables X and Y are **independent** if

$$p_{X,Y}(x, y) = p_X(x)p_Y(y) \quad \text{for every } x \in \mathcal{S}_X, y \in \mathcal{S}_Y. \tag{2.36}$$

Otherwise, they are dependent.

The variables X and Y in the preceding example are dependent, since

$$p_{X,Y}(0,2) = 0 \neq p_X(0)p_Y(2) = \frac{1}{2} \cdot \frac{1}{4} = \frac{1}{8}. \quad (2.37)$$

Although the two coin tosses are independent, quantities constructed from them need not be independent: X and Y both use the first toss.

More than two random variables. A collection $(X_1, \dots, X_m)$ is called a **random vector**. Its joint pmf is

$$p_{X_1, \dots, X_m}(x_1, \dots, x_m) = \Pr[X_1 = x_1, \dots, X_m = x_m]. \quad (2.38)$$

To obtain the marginal distribution of any selected variables, sum over all possible values of the variables that are not retained. For example, for three discrete random variables,

$$\begin{aligned} p_{X,Y}(x,y) &= \sum_{z \in \mathcal{S}_Z} p_{X,Y,Z}(x,y,z), \\ p_X(x) &= \sum_{y \in \mathcal{S}_Y} \sum_{z \in \mathcal{S}_Z} p_{X,Y,Z}(x,y,z). \end{aligned} \quad (2.39)$$

The variables are **mutually independent** if

$$p_{X_1, \dots, X_m}(x_1, \dots, x_m) = \prod_{j=1}^m p_{X_j}(x_j) \quad (2.40)$$

for every choice of their values.

3 Expectation and variance

A distribution contains more information than one number. Nevertheless, its center and spread often answer the first questions in an application: what is the average cost, and how much can individual outcomes vary? Expectation and variance quantify these two features.

Definition 12 (Expectation). For a discrete random variable X with $\sum_x |x| p_X(x) < \infty$, its **expectation** or **mean** is

$$\mu = \mathbb{E}[X] = \sum_{x \in \mathcal{S}_X} x p_X(x). \quad (3.1)$$

More generally, if $\sum_x |g(x)| p_X(x) < \infty$, then

$$\mathbb{E}[g(X)] = \sum_{x \in \mathcal{S}_X} g(x) p_X(x). \quad (3.2)$$

The absolute-convergence condition matters for distributions with infinitely many values. It is automatic when the support is finite. Equation (3.2) lets us average a cost, loss, or other function of an outcome without first deriving the distribution of $g(X)$.

Example 7 (The mean need not be a possible outcome). For the running distribution,

$$\begin{aligned}\mathbb{E}[X] &= 0 \left(\frac{1}{5}\right) + 1 \left(\frac{1}{2}\right) + 2 \left(\frac{3}{10}\right) = \frac{11}{10}, \\ \mathbb{E}[X^2] &= 0^2 \left(\frac{1}{5}\right) + 1^2 \left(\frac{1}{2}\right) + 2^2 \left(\frac{3}{10}\right) = \frac{17}{10}.\end{aligned}\tag{3.3}$$

Although the mean is 1.1, every realization is an integer. An expectation is a weighted average, not a prediction that a single trial will equal the average. Similarly, a fair die has mean $(1 + 2 + 3 + 4 + 5 + 6)/6 = 3.5$.

Proposition 13 (Linearity of expectation). If $X_1, \dots, X_m$ have finite expectations and $a_1, \dots, a_m, b$ are constants, then

$$\mathbb{E}\left[b + \sum_{j=1}^m a_j X_j\right] = b + \sum_{j=1}^m a_j \mathbb{E}[X_j].\tag{3.4}$$

Independence is not required.

Proof. For two discrete random variables, sum over their joint outcomes:

$$\begin{aligned}\mathbb{E}[aX + bY] &= \sum_{x,y} (ax + by) \Pr[X = x, Y = y] \\ &= a \sum_x x p_X(x) + b \sum_y y p_Y(y) = a\mathbb{E}[X] + b\mathbb{E}[Y].\end{aligned}\tag{3.5}$$

Absolute integrability justifies regrouping the sums. Extend the argument to finitely many terms and use that the expectation of a constant is the same constant. ■

For any event A , define its **indicator variable** I_A to equal 1 when A occurs and 0 otherwise. Then

$$\mathbb{E}[I_A] = \Pr[A].\tag{3.6}$$

Thus estimating a probability is a special case of estimating an expectation. If K counts successes across m trials with success probabilities $p_1, \dots, p_m$, linearity gives $\mathbb{E}[K] = \sum_{j=1}^m p_j$, even when the trial outcomes are dependent. Their dependence will matter for the variance.

Definition 14 (Variance and standard deviation). If $\mathbb{E}[X^2] < \infty$ and $\mu = \mathbb{E}[X]$, the **variance** of X is

$$\sigma^2 = \text{Var}[X] = \mathbb{E}[(X - \mu)^2] = \sum_{x \in \mathcal{S}_X} (x - \mu)^2 p_X(x).\tag{3.7}$$

Its **standard deviation** is $\sigma = \sqrt{\text{Var}[X]}$.

Variance averages *squared* deviations from the mean, preventing positive and negative deviations from canceling. It is nonnegative, and it equals zero precisely when $X = \mu$ with probability one. If X is measured in seconds, variance has units of seconds squared, while standard deviation has units of seconds. Standard deviation describes the spread of individual observations; later we will distinguish this from uncertainty in an estimated mean.

Proposition 15 (Useful variance identities). For a random variable with a finite second moment and constants a, b ,

$$\begin{aligned}\text{Var}[X] &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2, \\ \text{Var}[aX + b] &= a^2 \text{Var}[X].\end{aligned}\tag{3.8}$$

Proof. Expand $(X - \mu)^2 = X^2 - 2\mu X + \mu^2$ and apply linearity. The first result follows because $\mathbb{E}[X] = \mu$. For the second, subtract $\mathbb{E}[aX + b] = a\mu + b$ from $aX + b$: the centered quantity is $a(X - \mu)$, whose square is $a^2(X - \mu)^2$. ■

For our running example, Equation (3.3) gives

$$\text{Var}[X] = \frac{17}{10} - \left(\frac{11}{10}\right)^2 = \frac{49}{100}, \quad \sigma = \frac{7}{10}.\tag{3.9}$$

The variance identity is convenient for algebra, but directly subtracting two nearly equal, large floating-point numbers can lose accuracy. For numerical data, a centered or stable incremental computation is preferable to blindly evaluating “mean of squares minus square of mean.” This is the same cancellation issue encountered earlier in numerical computation.

3.1 Covariance and independence

Discrete random variables X and Y are independent when their joint pmf factors as

$$\Pr[X = x, Y = y] = p_X(x)p_Y(y) \quad \text{for every } x, y.\tag{3.10}$$

For a collection $X_1, \dots, X_m$, mutual independence requires the corresponding factorization for every joint assignment of all the variables. Having the same distribution is a separate property: two random variables can share a pmf and still be dependent. For example, setting $Y = X$ produces identical distributions but perfect dependence when X is nonconstant.

Definition 16 (Covariance). For random variables with finite second moments, their **covariance** is

$$\text{Cov}[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].\tag{3.11}$$

Expanding the square of a centered sum gives

$$\text{Var}\left[\sum_{j=1}^m a_j X_j\right] = \sum_{j=1}^m a_j^2 \text{Var}[X_j] + 2 \sum_{i < j} a_i a_j \text{Cov}[X_i, X_j].\tag{3.12}$$

If X and Y are independent, factorization of their joint pmf gives $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$, so their covariance is zero. Consequently, for independent variables,

$$\text{Var}\left[\sum_{j=1}^m a_j X_j\right] = \sum_{j=1}^m a_j^2 \text{Var}[X_j].\tag{3.13}$$

Zero covariance is sufficient for this variance formula, but it does not generally imply independence. Unlike expectation, variance is not additive without an appropriate condition on the dependence.

Example 8 (Bernoulli and binomial moments). If $X \sim \text{Bernoulli}(p)$, then $X^2 = X$, and hence

$$\mathbb{E}[X] = p, \quad \text{Var}[X] = p - p^2 = p(1 - p). \quad (3.14)$$

If $Y = X_1 + \cdots + X_m$ is binomial, the indicators are independent with a common success probability p . Therefore,

$$\mathbb{E}[Y] = mp, \quad \text{Var}[Y] = mp(1 - p). \quad (3.15)$$

For the observed proportion Y/m , the corresponding model has

$$\mathbb{E}\left[\frac{Y}{m}\right] = p, \quad \text{Var}\left[\frac{Y}{m}\right] = \frac{p(1 - p)}{m}. \quad (3.16)$$

This is our first quantitative explanation of why averaging more independent observations can produce a more precise estimate.

3.2 Expectation of the time to a first event

How many attempts should a program expect to make before an operation succeeds? How long should a device expect to wait before a message gets through? A geometric random variable models the *number of discrete attempts* when each attempt independently succeeds with the same probability. To model elapsed time, we must also specify when attempts occur and how long they take.

Proposition 17 (Geometric moments). If $T \sim \text{Geometric}(p)$ counts attempts including the first success, then

$$\mathbb{E}[T] = \frac{1}{p}, \quad \text{Var}[T] = \frac{1 - p}{p^2}, \quad 0 < p \leq 1. \quad (3.17)$$

Proof. The case $p = 1$ is immediate because $T = 1$ with probability one. For $0 < p < 1$, the geometric decay of the pmf makes both the first and second moments finite. Write $\mu = \mathbb{E}[T]$ and $m_2 = \mathbb{E}[T^2]$. If the first attempt succeeds, $T = 1$. If it fails, we have spent one attempt and start over: the number T' of additional attempts has the same distribution as T , because the future trials are independent with the same success probability. Averaging over these two cases gives

$$\mu = p \cdot 1 + (1 - p)\mathbb{E}[1 + T'] = 1 + (1 - p)\mu, \quad (3.18)$$

and therefore $p\mu = 1$. The same argument for the square gives

$$\begin{aligned} m_2 &= p \cdot 1 + (1 - p)\mathbb{E}[(1 + T')^2] \\ &= 1 + 2(1 - p)\mu + (1 - p)m_2. \end{aligned} \quad (3.19)$$

Substitute $\mu = 1/p$ and rearrange to obtain $m_2 = (2 - p)/p^2$. Hence $\text{Var}[T] = m_2 - \mu^2 = (1 - p)/p^2$. ■

Example 9 (Waiting for a sensor packet to get through). Suppose a sensor sends a packet in successive two-second attempts, retrying immediately after a failure. Start the clock at time zero; the first attempt completes at two seconds, the second at four seconds, and so on. Assume each attempt independently succeeds with probability $p = 0.2$, and the device continues until success.

The attempt count is $T \sim \text{Geometric}(0.2)$, and the elapsed time, measured in seconds, is $D = 2T$. Thus

$$\begin{aligned}\mathbb{E}[T] &= \frac{1}{0.2} = 5, & \text{Var}[T] &= \frac{0.8}{(0.2)^2} = 20, \\ \mathbb{E}[D] &= 2\mathbb{E}[T] = 10 \text{ seconds}, \\ \text{Var}[D] &= 4 \text{Var}[T] = 80 \text{ seconds}^2.\end{aligned}\tag{3.20}$$

The standard deviation of the duration is $\sqrt{80} \approx 8.944$ seconds. Although the mean is ten seconds, a delivery can take considerably longer. In particular, the probability that the packet has still not arrived after ten completed attempts, or twenty seconds, is

$$\Pr[D > 20] = \Pr[T > 10] = 0.8^{10} \approx 0.107374.\tag{3.21}$$

That is about a 10.7% chance of waiting longer than twenty seconds under this model.

For at least a 95% chance of delivery by the end of attempt k , we need

$$\begin{aligned}\Pr[T \leq k] = 1 - 0.8^k \geq 0.95 &\iff k \geq \frac{\log(0.05)}{\log(0.8)} \approx 13.425, \\ k_{\min} &= 14.\end{aligned}\tag{3.22}$$

The inequality reverses when dividing by $\log(0.8) < 0$. Fourteen attempts correspond to twenty-eight seconds, with delivery probability $1 - 0.8^{14} \approx 0.9560$. This is a probability target, not a guaranteed deadline.

4 Sampling, estimation, and the law of large numbers

4.1 From a probability model to a sequence of observations

A probability distribution describes possible outcomes and their probabilities. *Sampling* means producing an outcome according to that distribution. The outcome of one draw is uncertain before the draw, even when the distribution itself is known exactly. A simulation typically repeats this operation many times.

Definition 18 (Independent and identically distributed samples). Random variables $X_1, \dots, X_n$ are *independent and identically distributed*, abbreviated *iid*, if they are independent and each has the same distribution. For discrete variables with common probability mass function p , this means that

$$\Pr[X_1 = x_1, \dots, X_n = x_n] = \prod_{i=1}^n p(x_i).\tag{4.1}$$

Sampling independently from the same distribution corresponds to repeating an experiment under the same probabilistic conditions.

Suppose a distribution has unknown mean $\mu = \mathbb{E}[X]$ and finite variance $\sigma^2 = \text{Var}[X]$. If both μ and σ are unknown, and we wish to figure them out, then we have to do so via sampling.

4.2 Estimating the mean

The *sample mean* is

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i. \quad (4.2)$$

Before observing the data, $\bar{X}_n$ is a random variable called an *estimator*. After observing $x_1, \dots, x_n$, its numerical value $\bar{x}_n = n^{-1} \sum_{i=1}^n x_i$ is an *estimate*. Repeating the whole experiment generally gives a different estimate. The probability distribution of those possible estimates is the *sampling distribution*.

Proposition 19 (Mean and variance of the sample mean). If $X_1, \dots, X_n$ are iid with mean μ and finite variance σ^2 , then

$$\mathbb{E}[\bar{X}_n] = \mu, \quad \text{Var}[\bar{X}_n] = \frac{\sigma^2}{n}. \quad (4.3)$$

Thus the sample mean is *unbiased*: its expectation equals the parameter being estimated.

Proof. Linearity of expectation gives

$$\mathbb{E}[\bar{X}_n] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i] = \frac{n\mu}{n} = \mu. \quad (4.4)$$

Independence implies $\text{Cov}[X_i, X_j] = 0$ for $i \neq j$, so

$$\begin{aligned} \text{Var}[\bar{X}_n] &= \frac{1}{n^2} \text{Var}\left[\sum_{i=1}^n X_i\right] \\ &= \frac{1}{n^2} \left(\sum_{i=1}^n \text{Var}[X_i] + 2 \sum_{1 \leq i < j \leq n} \text{Cov}[X_i, X_j] \right) = \frac{n\sigma^2}{n^2}. \end{aligned} \quad (4.5)$$

The expectation calculation does not require independence; the variance calculation uses it to eliminate the covariance terms. ■

Key Idea

The standard deviation σ describes the spread of *individual observations*. The *standard error* of the sample mean describes the spread of *sample means across repeated experiments*:

$$\text{SE}(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}. \quad (4.6)$$

Collecting more observations makes the mean more precise; it does not make the individual observations less variable. Reducing this standard error by a factor of 10 requires 100 times as many independent observations.

For the three-point distribution above, direct calculation gives

$$\mu = 0(0.2) + 1(0.5) + 2(0.3) = 1.1, \quad \sigma^2 = (0^2(0.2) + 1^2(0.5) + 2^2(0.3)) - 1.1^2 = 0.49. \quad (4.7)$$

Consequently, an individual draw has standard deviation 0.7, whereas the average of 100 independent draws has standard error 0.07.

4.3 Estimating the variance

The *sample variance* is

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2, \quad n \geq 2. \quad (4.8)$$

The denominator $n - 1$ corrects for estimating the center from the same data. The deviations sum to zero, so knowing $n - 1$ of them determines the last one. The following calculation makes the correction precise.

Proposition 20 (Unbiased sample variance). For iid observations with finite variance σ^2 and $n \geq 2$, $\mathbb{E}[S_n^2] = \sigma^2$.

Proof. Write $X_i - \bar{X}_n = (X_i - \mu) - (\bar{X}_n - \mu)$, expand the squares, and use $\sum_i (X_i - \mu) = n(\bar{X}_n - \mu)$. This gives the identity

$$\sum_{i=1}^n (X_i - \bar{X}_n)^2 = \sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X}_n - \mu)^2. \quad (4.9)$$

Since $\bar{X}_n$ is unbiased, taking expectations yields

$$\mathbb{E} \left[\sum_{i=1}^n (X_i - \bar{X}_n)^2 \right] = n\sigma^2 - n \frac{\sigma^2}{n} = (n-1)\sigma^2. \quad (4.10)$$

Divide by $n - 1$ to obtain the claim. ■

Example 10 (A small observed sample). For the observations (0, 1, 1, 2, 1), the mean and unbiased variance estimates are

$$\bar{x}_5 = \frac{0 + 1 + 1 + 2 + 1}{5} = 1, \quad s_5^2 = \frac{(0-1)^2 + 0 + 0 + (2-1)^2 + 0}{4} = \frac{1}{2}. \quad (4.11)$$

The estimated standard error is $s_5/\sqrt{5} = \sqrt{0.1} \approx 0.316$. Using denominator 5 would instead give 0.4, the variance of the empirical distribution that puts mass 1/5 on each observation. These answer different questions: 0.4 describes the observed data using their empirical distribution; 0.5 is the unbiased estimator of the underlying population variance.

4.4 Why averaging works: the law of large numbers

Definition 21 (Convergence in probability). A sequence of random variables Y_n converges *in probability* to a constant c , written $Y_n \xrightarrow{\text{Pr}} c$, if, for every $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} \Pr[|Y_n - c| > \varepsilon] = 0. \quad (4.12)$$

Theorem 1 (Weak law of large numbers; finite-variance version). Let $X_1, X_2, \dots$ be iid random variables with mean μ and finite variance σ^2 . Then

$$\bar{X}_n \xrightarrow{\text{Pr}} \mu. \quad (4.13)$$

More quantitatively, for every $\varepsilon > 0$ and every $n \geq 1$,

$$\Pr[|\bar{X}_n - \mu| \geq \varepsilon] \leq \frac{\sigma^2}{n\varepsilon^2}. \quad (4.14)$$

The theorem says that, for any fixed error tolerance, the probability of exceeding that tolerance tends to zero. An estimator with this convergence property is called *consistent*.

Theorem 2 (Strong law of large numbers). Let $X_1, X_2, \dots$ be iid and suppose $\mathbb{E}[|X_1|] < \infty$. Then

$$\Pr\left[\lim_{n \rightarrow \infty} \bar{X}_n = \mathbb{E}[X_1]\right] = 1. \quad (4.15)$$

Equivalently, the sample mean converges *almost surely* to the expectation.

The strong law concerns an entire infinite sequence: outside an exceptional set of sequences of probability zero, its running averages converge to the mean. It does not assert convergence for every imaginable sequence. Almost-sure convergence implies convergence in probability. Notice also the weaker moment assumption: the strong law stated here needs a finite absolute first moment, not a finite variance. We state it without proof; see [2] for background on laws of large numbers.

The running-average error need not decrease at each step. An observation far from μ can temporarily move the average away from μ . The law also gives no universal sample size that guarantees a specified accuracy with certainty.

The strong law also justifies estimating variance. For iid samples with $\mathbb{E}[X^2] < \infty$, apply it separately to X_i and X_i^2 . The identity

$$S_n^2 = \frac{n}{n-1} \left(\frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}_n^2 \right) \xrightarrow{\text{almost surely}} \mathbb{E}[X^2] - \mu^2 = \sigma^2 \quad (4.16)$$

shows that the unbiased sample variance is consistent. A finite fourth moment is not needed for this consistency statement.

4.5 Estimating probabilities

For an event A , let I_i be 1 when A occurs in trial i and 0 otherwise. If the trials are independent with the same event probability p , then

$$\hat{p}_n = \frac{1}{n} \sum_{i=1}^n I_i = \frac{\text{number of occurrences of } A}{n}, \quad \mathbb{E}[\hat{p}_n] = p, \quad \text{Var}[\hat{p}_n] = \frac{p(1-p)}{n}. \quad (4.17)$$

Thus the observed relative frequency is a sample mean, and the law of large numbers explains its convergence to the event probability. This is the link between probability models and percentages reported in studies. For

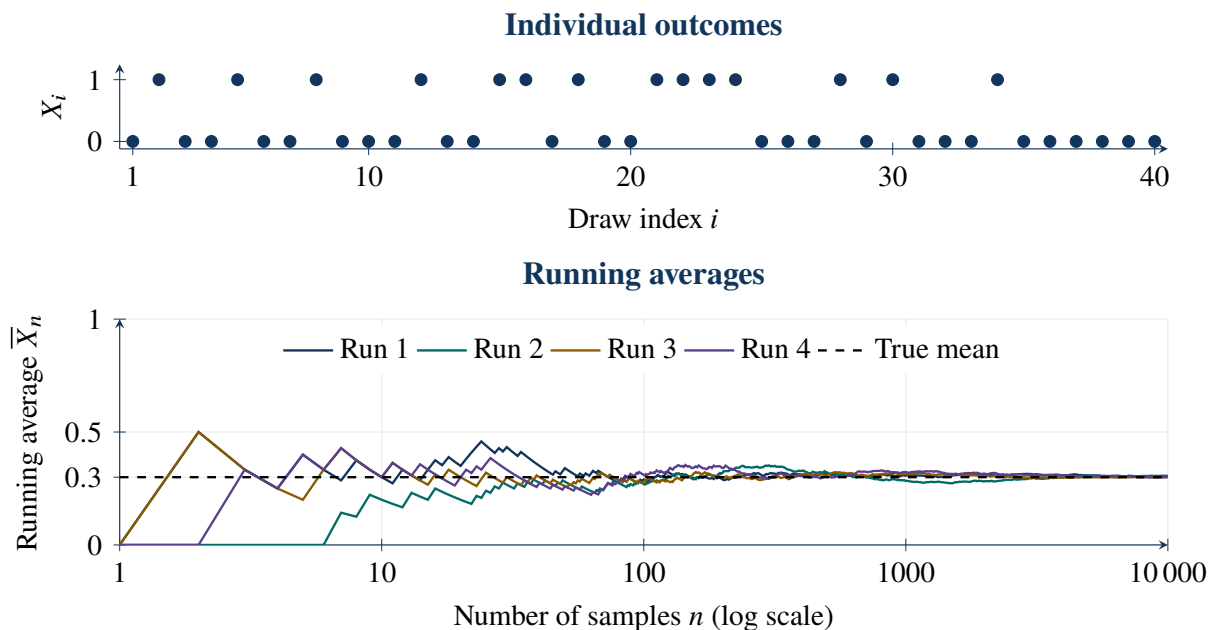


FIGURE 3: Running averages of independent Bernoulli observations with success probability $p = 0.3$. Different runs fluctuate differently, while their averages settle near 0.3. Individual observations remain 0 or 1; they do not approach 0.3. A finite plot illustrates the law of large numbers but is not a proof of it.

a clinical study, the Bernoulli model here assumes a common outcome definition, a fixed follow-up period, complete outcome data, and independent participants drawn under the specified model. Violating these assumptions can change the analysis.

4.6 Exam bell curves and the normal distribution

An instructor may decide that an exam's numerical scores should have a different average or a different spread. For example, the instructor might want the class average to be 75 and the standard deviation to be 8. A **curving procedure** specifies how to transform each student's original score into a new score.

The phrase "bell curving" does not identify a unique procedure. Here we study one explicit rule: **standardize every score relative to the class, then rescale to a chosen mean and standard deviation**. We first explain the algorithm and work through its consequences. We then introduce the normal distribution, which supplies the bell-shaped reference model associated with this procedure.

The grading algorithm. Suppose the original scores are $x_1, \dots, x_n$, measured out of 100. Choose a target mean m_* and a target standard deviation $s_* > 0$. These are grading-policy choices, not quantities determined by a probability theorem.

Apply the following steps to the entire class:

1. **Compute the original mean and sample standard deviation.** For $n \geq 2$, calculate

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i, \quad s_n = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2}. \quad (4.18)$$

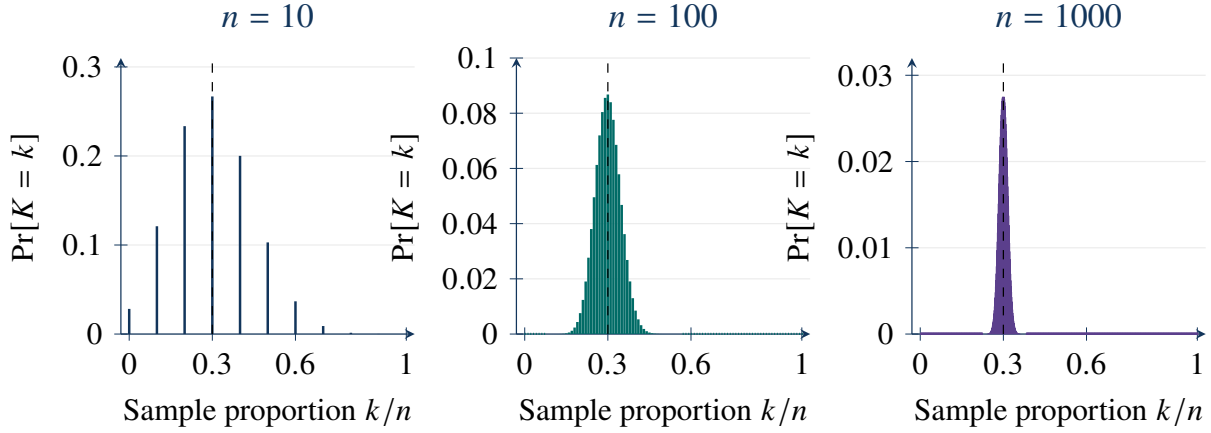


FIGURE 4: Exact sampling distributions of the Bernoulli sample proportion K/n , where $K \sim \text{Binomial}(n, 0.3)$. Each panel describes repeating the entire experiment at that fixed size. Horizontal scales agree; vertical scales differ. The distributions narrow around 0.3 as the standard error $\sqrt{0.3(0.7)/n}$ decreases.

The procedure below requires $s_n > 0$. If every student has the same score, then $s_n = 0$ and a separate grading rule is needed.

2. **Express each score in standard-deviation units.** For each student, calculate the **z-score**

$$z_i = \frac{x_i - \bar{x}_n}{s_n}. \quad (4.19)$$

A student with $z_i = 0$ scored at the class mean. A student with $z_i = 1$ scored one standard deviation above it, while a student with $z_i = -1$ scored one standard deviation below it.

3. **Transform each standardized score to the target scale.** Calculate

$$g_i = m_* + s_* z_i = m_* + \frac{s_*}{s_n} (x_i - \bar{x}_n). \quad (4.20)$$

Thus a score at the original mean becomes m_* , and a score one original standard deviation above the mean becomes $m_* + s_*$.

4. **Apply any separately announced reporting rules.** Rounding, restricting scores to $[0, 100]$, and converting numerical scores to letter grades are additional policy decisions. In the examples below, all transformed scores already lie in $[0, 100]$, and we use them without rounding.

Before any rounding or restriction to a score range, the transformed scores have exactly the chosen sample mean and sample standard deviation:

$$\frac{1}{n} \sum_{i=1}^n g_i = m_*, \quad \sqrt{\frac{1}{n-1} \sum_{i=1}^n (g_i - m_*)^2} = s_*. \quad (4.21)$$

This follows because the standardized scores have sample mean 0 and sample standard deviation 1.

Because $s_*/s_n > 0$, the transformation preserves students' ordering: a higher original score produces a higher transformed score, and tied scores remain tied. However, preserving the ordering does not mean that every student's score increases.

Example 11 (Applying the algorithm to an entire class). Suppose six students receive the original scores

$$50, \quad 60, \quad 65, \quad 65, \quad 70, \quad 80. \quad (4.22)$$

Their mean and sample standard deviation are

$$\begin{aligned} \bar{x}_6 &= \frac{50 + 60 + 65 + 65 + 70 + 80}{6} = 65, \\ s_6 &= \sqrt{\frac{(-15)^2 + (-5)^2 + 0^2 + 0^2 + 5^2 + 15^2}{5}} = 10. \end{aligned} \quad (4.23)$$

Choose target mean $m_* = 75$ and target standard deviation $s_* = 8$. The transformation is therefore

$$g_i = 75 + 8 \left(\frac{x_i - 65}{10} \right). \quad (4.24)$$

Applying it to every student gives

Student	Original score x_i	z_i	Curved score g_i
1	50	-1.5	63
2	60	-0.5	71
3	65	0	75
4	65	0	75
5	70	0.5	79
6	80	1.5	87

For the student who originally scored 80,

$$z = \frac{80 - 65}{10} = 1.5, \quad g = 75 + 8(1.5) = 87. \quad (4.25)$$

The student remains 1.5 standard deviations above the class mean, but the numerical score changes because the grading scale has changed.

The transformed scores have mean 75 and sample standard deviation 8. Every score increases in this particular example.

Example 12 (Curving can lower a student's grade). Now suppose a different class receives

$$60, \quad 70, \quad 75, \quad 75, \quad 80, \quad 90. \quad (4.26)$$

The class mean is 75 and the sample standard deviation is 10. Use the same target mean $m_* = 75$ and target standard deviation $s_* = 8$. The algorithm becomes

$$g_i = 75 + 8 \left(\frac{x_i - 75}{10} \right) = 75 + 0.8(x_i - 75). \quad (4.27)$$

The results are

Original score x_i	Curved score g_i	Change $g_i - x_i$
60	63	+3
70	71	+1
75	75	0
75	75	0
80	79	-1
90	87	-3

The student who scored 90 now receives 87: the curve *lowers* this student's numerical grade by three points. If the same illustrative letter-grade cutoffs are used before and after curving, with A for scores at least 90 and B for scores from 80 up to but not including 90, this changes an A to a B.

The reason is that the target spread is smaller than the original spread. Distances from the mean are multiplied by 0.8. Scores above the mean move down toward it, and scores below the mean move up toward it. The average remains 75.

Even raising the target mean does not guarantee that every score rises. For this same class, choosing $m_* = 77$ and $s_* = 8$ would send the score 90 to

$$g = 77 + 8 \left(\frac{90 - 75}{10} \right) = 89. \quad (4.28)$$

The class mean would increase from 75 to 77, while this student's score would still decrease.

More generally, subtracting the original score from the transformed score gives

$$g_i - x_i = (m_* - \bar{x}_n) + \left(\frac{s_*}{s_n} - 1 \right) (x_i - \bar{x}_n). \quad (4.29)$$

The first term shifts the class average. The second changes the student's distance from that average. Their combined effect determines whether the student's score increases, decreases, or stays the same.

A rule that never lowers a score would be an additional policy, such as replacing g_i by $\max\{x_i, g_i\}$. That modified rule would generally no longer produce exactly the target mean and standard deviation.

Example 13. Consider a hypothetical class of 300 students taking an exam scored out of 100. The integer scores range from 31 to 95; the table summarizes those scores in ten-point intervals. These are invented data, with uneven frequencies and a longer tail toward low scores, rather than an exactly symmetric grade distribution.

Scores	30–39	40–49	50–59	60–69	70–79	80–89	90–99
Students	4	12	29	65	97	76	17

The accompanying file `class-grades-300.csv` contains every original score and both curved scores. We compute the statistics from those individual scores, not from the interval midpoints:

$$\begin{aligned}\bar{x}_{300} &= \frac{21768}{300} = 72.56, \\ s_{300} &= \sqrt{\frac{1628800 - 300(72.56)^2}{299}} \approx 12.8425.\end{aligned}\tag{4.30}$$

Here 1628800 is the sum of the squared scores, and the denominator 299 corresponds to the sample standard deviation.

Method 1: shift and rescale. Apply the algorithm from the preceding examples with target mean 75 and target sample standard deviation 8:

$$g_i = 75 + \frac{8}{s_{300}}(x_i - 72.56) \approx 75 + 0.6229(x_i - 72.56).\tag{4.31}$$

For example,

$$\begin{aligned}x_i = 50 &\implies g_i \approx 60.95, \\ x_i = 80 &\implies g_i \approx 79.63, \\ x_i = 90 &\implies g_i \approx 85.86.\end{aligned}\tag{4.32}$$

The mean increases from 72.56 to 75, yet some students' grades decrease. The new sample standard deviation is exactly 8. Because this rule shifts and compresses every score in the same way, it preserves the distribution's asymmetry. It does not turn this class's original grades into a bell-shaped distribution.

Method 2: assign grades by rank using a bell-shaped target. If the instructor instead wants the resulting grades to follow a bell shape, a different procedure is needed. First choose a fixed, increasing list of 300 target grades, denoted by $t_1, \dots, t_{300}$. Choose this list to have many grades near 75 and progressively fewer toward the two extremes, forming an approximately symmetric bell-shaped histogram. The list is an input to the grading rule, just as the target mean and standard deviation were inputs to Method 1.

For this example, the target grades have mean 75 and sample standard deviation 8; selected entries, rounded here to two decimal places, are

Rank j	1	16	76	150	151	225	285	300
Target t_j	51.51	61.96	69.64	74.97	75.03	80.36	88.04	98.49

The next discussion introduces the smooth bell-shaped probability model used to construct this target list. For now, the complete table makes the grading algorithm explicit:

1. Sort the original grades from lowest to highest.
2. Assign positions $1, \dots, 300$ in this ordered list. If a score occurs only once, at position j , give that student grade t_j .

3. If several students have the same original score and occupy positions $a, \dots, b$, give all of them the same grade: the average of the target grades at those positions,

$$h_i = \frac{1}{b - a + 1} \sum_{j=a}^b t_j. \quad (4.33)$$

4. Return these assigned grades to the students in their original order.

This procedure preserves the ordering of students and gives equal curved grades to students with equal original scores.

For a concrete calculation, five students scored 70. Exactly 110 students scored below 70, so those five students occupy positions 111 through 115. Each receives

$$h_i = \frac{t_{111} + t_{112} + t_{113} + t_{114} + t_{115}}{5} \approx 72.45. \quad (4.34)$$

The three students who scored 90 occupy positions 284 through 286; each receives approximately 88.04, so this method can also lower a grade. By contrast, the three students who scored 95 occupy positions 298 through 300 and each receive approximately 96.09.

Without ties, the assigned grades would be exactly the prescribed target list. Averaging within tied groups changes it slightly: the resulting sample mean is still exactly 75, while the sample standard deviation is approximately 7.9939. Thus the resulting histogram is approximately bell-shaped, with some irregularity caused by ties and the finite class size. Neither the target values nor the transformed grades are rounded or capped in these calculations; all grades already lie between 0 and 100.

Figure 5 compares both procedures using the same students. Every histogram uses five-point bins centered at 30, 35, $\dots$, 100 and the same axis scales. A bar for n_k students has height

$$\frac{n_k}{300 \cdot 5}, \quad (4.35)$$

so its area is the fraction $n_k/300$ of the class in that interval. The first method changes the location and spread; the second also changes the shape by assigning grades according to rank. The bell-shaped pattern in the second method is a consequence of the chosen grading rule, not evidence that the original scores were symmetric. This motivates introducing a precise mathematical model of a bell-shaped distribution.

Where does the bell curve enter? The preceding 300-student example illustrates two different meanings of “curving” grades. Method 1 chooses a target mean and standard deviation, then shifts and rescales the original scores. It preserves their asymmetry. Method 2 chooses an entire bell-shaped target distribution and assigns grades according to students’ ranks. Its approximately bell-shaped output is therefore built into the grading rule.

The smooth black curve in the lower panels of Figure 5 represents the target shape used in Method 2. Its center is 75, and its standard deviation is 8. The mathematical model for this curve is the **normal**, or **Gaussian**, distribution. We first define this distribution, then explain exactly how it produced the rank and target table.

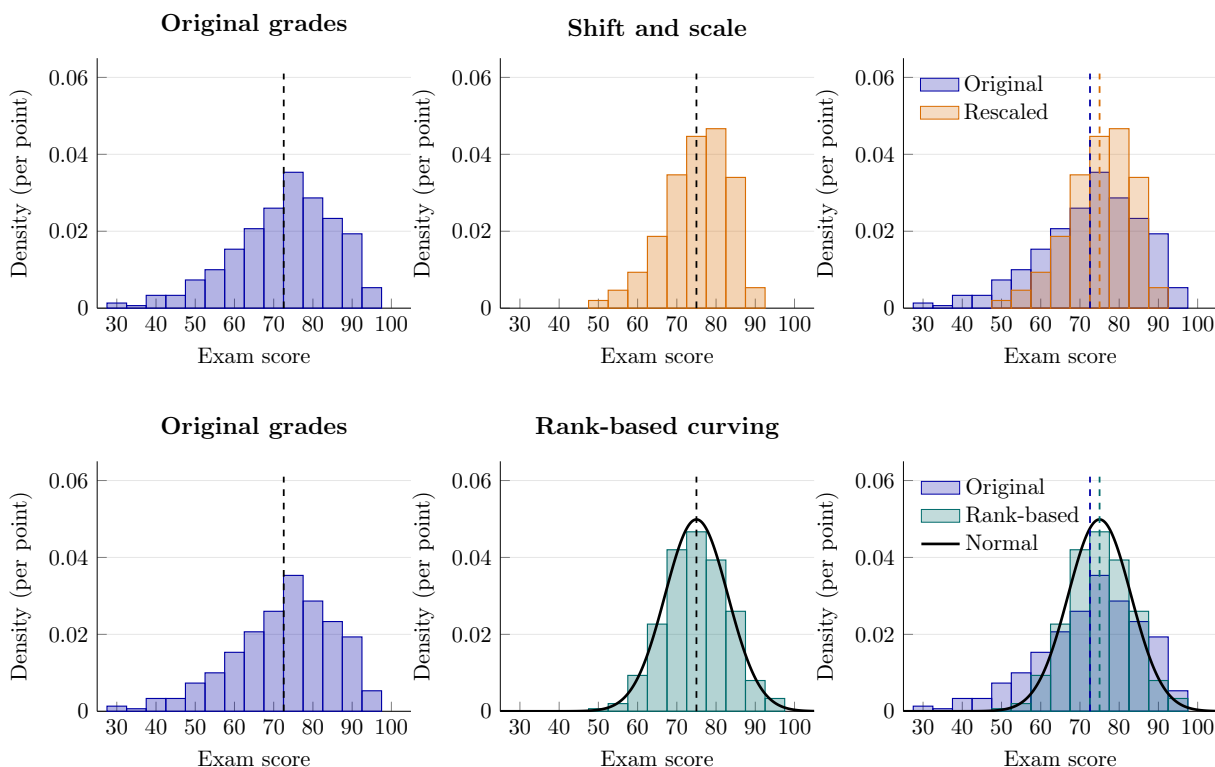


FIGURE 5: Two grading procedures applied to the same 300 hypothetical students. In each row, the left panel shows the original grades, the middle shows the curved grades, and the right shows both. The top row uses a shift and rescaling, which preserves the original asymmetry; the bottom row uses the bell-shaped target list and the rank-based assignment rule, averaging target grades for tied scores. The original mean and sample standard deviation are 72.56 and approximately 12.8425. Both curved means are 75; the respective sample standard deviations are 8 and approximately 7.9939. Dashed lines mark the means. The solid black curve in the lower middle and lower right panels is the normal density with mean 75 and standard deviation 8, the bell-shaped reference for the rank-based procedure. Its mathematical definition follows this example. All panels use the same five-point bins and axis scales; each individual histogram has total area one.

Definition 22 (Normal or Gaussian distribution). A continuous random variable X has distribution $\mathcal{N}(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma > 0$, if its probability density is

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad -\infty < x < \infty. \quad (4.36)$$

Its mean and variance are

$$\mathbb{E}[X] = \mu, \quad \text{Var}[X] = \sigma^2. \quad (4.37)$$

Thus the second parameter in $\mathcal{N}(\mu, \sigma^2)$ is the variance; σ is the standard deviation.

The curve is symmetric about μ . Changing μ moves its center; increasing σ makes it wider and lowers its peak. Its total area remains one. For a continuous distribution, probabilities are areas:

$$\int_{-\infty}^{\infty} f_X(x) dx = 1, \quad \Pr[a \leq X \leq b] = \int_a^b f_X(x) dx. \quad (4.38)$$

This is why the example uses density histograms: a bar's area is the fraction of students in its score interval, and the area under the smooth curve over that interval is the corresponding model probability. The height of the curve at a single score is not a probability. Indeed, under this continuous model, $\Pr[X = 80] = 0$.

For our grading example, the target model is $\mathcal{N}(75, 8^2)$, with density

$$f(x) = \frac{1}{8\sqrt{2\pi}} \exp\left(-\frac{(x-75)^2}{2 \cdot 8^2}\right). \quad (4.39)$$

This is the black curve in the figure. It specifies a desired pattern for the curved grades; it is not a fitted description of the original, asymmetric scores. A finite list of grades only approximates this continuous curve. Also, actual grades are bounded, whereas a normal distribution extends over the entire real line.

Standardization and Method 1. If X is exactly normal, standardizing it gives a **standard normal** random variable:

$$X \sim \mathcal{N}(\mu, \sigma^2) \implies Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1). \quad (4.40)$$

Moving to a target mean and standard deviation then gives

$$G = m_* + s_* Z \sim \mathcal{N}(m_*, s_*^2). \quad (4.41)$$

This is the normal-model interpretation of Method 1, the shift-and-rescale procedure. For observed grades, using the sample mean and sample standard deviation achieves the desired sample mean and spread, before rounding or capping. It produces a normal shape only if that shape was already present. In particular, it does not remove the asymmetry of the 300-student class.

Normal percentiles and quantiles. For $Z \sim \mathcal{N}(0, 1)$, define the density and cumulative distribution function by

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad \Phi(z) = \Pr[Z \leq z] = \int_{-\infty}^z \phi(u) du. \quad (4.42)$$

The value $\Phi(z)$ is the fraction of the model distribution at or below z . Conversely, given a fraction p with $0 < p < 1$, the **quantile** $\Phi^{-1}(p)$ is the unique value q satisfying $\Phi(q) = p$. For example, $\Phi^{-1}(0.5) = 0$: half of the standard normal distribution lies below its center.

How the rank and target table was obtained. The ranks come from the original grades: sort the 300 scores from lowest to highest and number their positions $j = 1, \dots, 300$. The target table is constructed independently of the numerical values of those scores. It assigns a target grade to each of these positions.

Step 1: assign a probability to each position. Divide the interval from 0 to 1 into 300 equal pieces and use their midpoints:

$$p_j = \frac{j - \frac{1}{2}}{300}, \quad j = 1, \dots, 300. \quad (4.43)$$

Position j therefore corresponds to the $100p_j$ th percentile of the target model. The midpoint convention places one target in each equal-probability piece. It also avoids probabilities 0 and 1, whose normal quantiles would be infinite.

Step 2: convert probabilities to normal scores. For each position, solve

$$\Phi(q_j) = p_j, \quad \text{equivalently} \quad q_j = \Phi^{-1}(p_j). \quad (4.44)$$

These values can be found using an inverse-normal table or numerical software. They are deterministic values, not random samples. Successive probabilities differ by $1/300$, so the area under the standard normal curve between successive quantiles is $1/300$. Near the center, where the density is high, a narrow interval contains that area; in the tails, a wider interval is needed. Thus the target scores cluster near the center and become more widely spaced toward the extremes.

Step 3: put the finite target list on the desired grading scale. The values $75 + 8q_j$ already approximately follow the target model $\mathcal{N}(75, 8^2)$. However, these are 300 selected quantiles, so their sample standard deviation need not be exactly 8. To obtain the exact target moments used in the example, compute

$$\bar{q} = \frac{1}{300} \sum_{j=1}^{300} q_j = 0, \quad s_q = \sqrt{\frac{1}{299} \sum_{j=1}^{300} (q_j - \bar{q})^2} \approx 0.9995206, \quad (4.45)$$

and set

$$t_j = 75 + 8 \frac{q_j - \bar{q}}{s_q}, \quad j = 1, \dots, 300. \quad (4.46)$$

The equality $\bar{q} = 0$ follows from the symmetry $q_{301-j} = -q_j$. This small adjustment gives the target list sample mean exactly 75 and sample standard deviation exactly 8. It produces the table supplied with the preceding example; selected entries are shown below, rounded only for display.

Rank j	Probability p_j	Normal score q_j	Target grade t_j
1	0.001667	-2.935199	51.51
16	0.051667	-1.628904	61.96
76	0.251667	-0.669254	69.64
150	0.498333	-0.004178	74.97
151	0.501667	0.004178	75.03
225	0.748333	0.669254	80.36
285	0.948333	1.628904	88.04
300	0.998333	2.935199	98.49

Step 4: assign the targets, keeping tied grades equal. A student with a unique score at position j receives t_j . If students with the same original score occupy positions a through b , each receives

$$h = \frac{1}{b - a + 1} \sum_{j=a}^b t_j. \quad (4.47)$$

For example, the five students who scored 70 occupy positions 111 through 115. Their target grades, before averaging, are approximately 72.3085, 72.3792, 72.4497, 72.5199, and 72.5900. Each student therefore receives approximately 72.45, as in the example. The averaging is performed on the target grades after the quantile conversion.

Without ties, this procedure assigns exactly the chosen target list. Averaging within tied groups preserves its mean and slightly reduces its spread: the 300 assigned grades have mean 75 and sample standard deviation approximately 7.9939. Their histogram follows the target bell shape approximately; it is not an exactly continuous normal distribution. This explains both the bell shape in the lower panels and the small differences between the bars and the smooth black curve.

Model percentiles and observed class ranks. A percentile inferred from a normal model is different from a rank computed from the observed scores. In the earlier six-student example, the score 80 has

$$z = \frac{80 - 65}{10} = 1.5, \quad \Phi(1.5) \approx 0.9332. \quad (4.48)$$

This is approximately the 93.3rd percentile under a normal model; it is not a claim that 93.3% of those classmates scored below 80. Method 1 preserves this standardized score when 80 becomes 87. Method 2 instead starts from the actual ordered positions and uses normal quantiles to assign new scores. Its bell shape comes from the chosen target distribution, regardless of whether the original scores were bell-shaped.

Normal probabilities and the connection to error bars. For a normal distribution, integration gives

$$\begin{aligned} \Pr[|X - \mu| \leq \sigma] &\approx 0.6827, \\ \Pr[|X - \mu| \leq 2\sigma] &\approx 0.9545, \\ \Pr[|X - \mu| \leq 3\sigma] &\approx 0.9973. \end{aligned} \quad (4.49)$$

This is the 68–95–99.7 rule. For the model $\mathcal{N}(65, 10^2)$, approximately 68.27% of its probability lies between 55 and 75. These percentages describe a normal model; they are not guaranteed percentages for an arbitrary class.

A closely related calculation is

$$\Pr[-1.96 \leq Z \leq 1.96] = \Phi(1.96) - \Phi(-1.96) \approx 0.95. \quad (4.50)$$

This explains where the number 1.96 comes from in many approximate confidence intervals.

The next subsection uses the normal distribution for a different purpose. The central limit theorem describes the distribution of a standardized sample mean across repeated samples. It does not claim that individual exam scores, or arbitrary observations, are normally distributed.

4.7 The central limit theorem and sampling error bars

A plot of a sample mean often includes a line segment called an *error bar*. The point marks the estimate; the endpoints summarize a specified kind of variability or uncertainty. The phrase “error bar” alone does not tell us what was calculated. The caption must say whether the bars represent standard deviations, standard errors, or a confidence interval, and must state the sample size and the confidence level when applicable.

For observed iid data $x_1, \dots, x_n$, let $\bar{x}_n$ be the sample mean and s_n the sample standard deviation computed with denominator $n - 1$. Three common choices are:

Label	Endpoints	Interpretation
One SD	$\bar{x}_n \pm s_n$	Describes the spread of individual observations.
One SE	$\bar{x}_n \pm s_n/\sqrt{n}$	Describes the estimated sampling variability of the mean.
Approximate 95% CI	$\bar{x}_n \pm 1.96s_n/\sqrt{n}$	Uses a large-sample normal approximation to give an interval for the population mean.

The SD bars do not generally shrink as more data are collected: they estimate the population spread σ . The SE and confidence-interval bars for the mean shrink at the typical rate $n^{-1/2}$. A one-SD interval need not contain any particular fixed percentage of observations for an arbitrary distribution, and a one-SE interval is not a 95% confidence interval.

Theorem 3 (Central limit theorem; iid finite-variance version). If $X_1, X_2, \dots$ are iid with mean μ and variance $0 < \sigma^2 < \infty$, then

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{\text{distribution}} Z, \quad Z \sim \mathcal{N}(0, 1). \quad (4.51)$$

Here $\mathcal{N}(0, 1)$ is the standard normal distribution introduced through exam bell curves in Section 4.6. Convergence in distribution means that the CDF of the standardized error approaches Φ at every real argument.

We state this theorem without proof; see [2, Chapter 9]. It describes the distribution of the scaled estimation error, whereas the LLN establishes convergence of the estimate itself.

Seeing the central limit theorem. For each sample size n , imagine repeatedly drawing a fresh independent sample of n observations, computing its mean, and standardizing it using the population mean and standard deviation:

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}, \quad \mathbb{E}[Z_n] = 0, \quad \text{Var}[Z_n] = 1. \quad (4.52)$$

The figures below show the resulting *sampling distributions* for $n = 1, 5, 30, 100$. Each panel describes the outcomes of repeating the whole size- n experiment, not the sequence of running means from one experiment. The standard normal reference curve is $\phi(z) = e^{-z^2/2}/\sqrt{2\pi}$.

Because Bernoulli and geometric observations are integer-valued, Z_n remains discrete at every finite n . To compare its shape with a density, we draw a rectangle centered at each possible value z , with

$$\text{bar width } \Delta_n = \frac{1}{\sigma\sqrt{n}}, \quad \text{bar height } \frac{\Pr[Z_n = z]}{\Delta_n}. \quad (4.53)$$

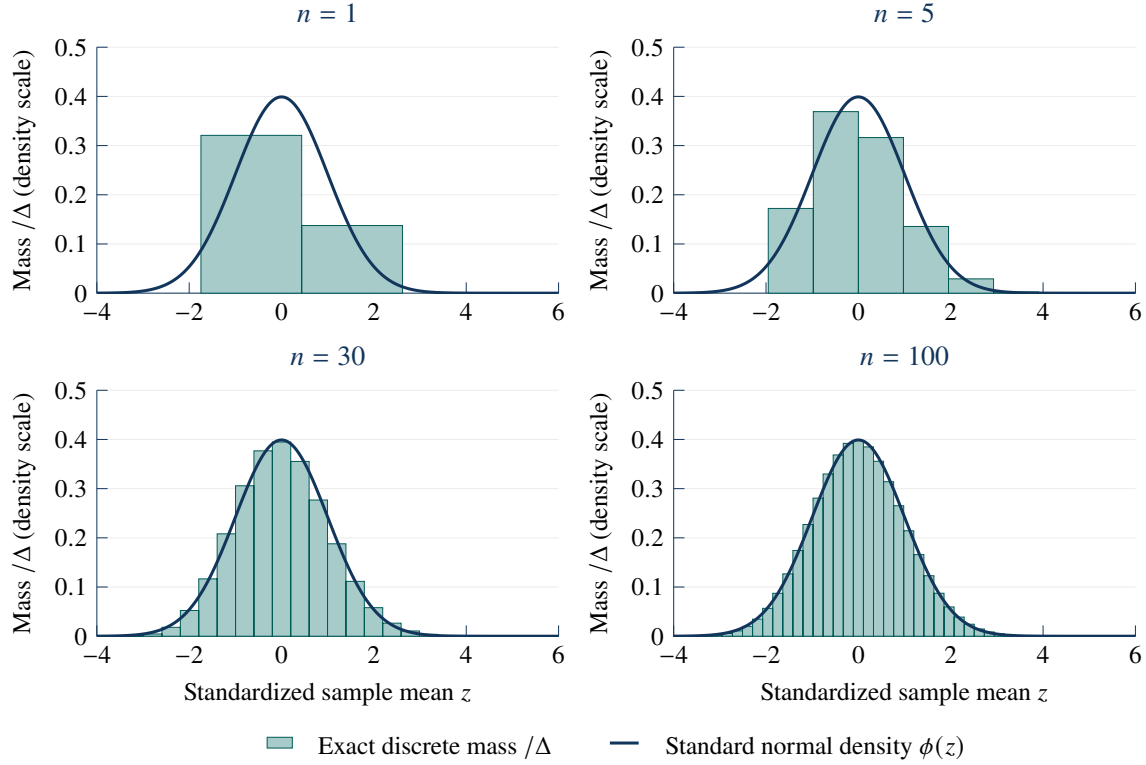


FIGURE 6: Central limit theorem for iid Bernoulli(0.3) samples. Bars represent the exact distribution of $Z_n = (K_n - 0.3n) / \sqrt{0.21n}$, which is discrete at every finite n . Each complete bar has width $\Delta_n = 1/\sqrt{0.21n}$ and area $\Pr[Z_n = z]$; the overlaid curve is the standard normal density. Tails outside the displayed range are not renormalized. Increasing n reveals convergence of the standardized distribution's shape, while its mean remains 0 and its variance remains 1.

Each complete bar therefore has area $\Pr[Z_n = z]$. The rectangles are a visualization of discrete probability masses, not a continuous density of Z_n . Their probabilities are evaluated from exact formulas; the plots contain no Monte Carlo sampling noise. All panels use the same horizontal scale, and the vertical scale is fixed within each figure. Tails outside the displayed range are omitted without renormalizing.

Example 14 (Bernoulli sample means). Let $X_i \sim \text{Bernoulli}(0.3)$ independently. Then $\mu = 0.3$, $\sigma^2 = 0.21$, and the success count $K_n = \sum_{i=1}^n X_i$ has a binomial distribution. Thus

$$Z_n = \frac{K_n - 0.3n}{\sqrt{0.21n}}, \quad \Pr[K_n = k] = \binom{n}{k} (0.3)^k (0.7)^{n-k}, \quad k = 0, \dots, n. \quad (4.54)$$

At $n = 1$ there are only two possible standardized outcomes. As n grows, the possible values become more closely spaced and their probability pattern increasingly follows the Gaussian curve. The individual X_i still take only the values 0 and 1.

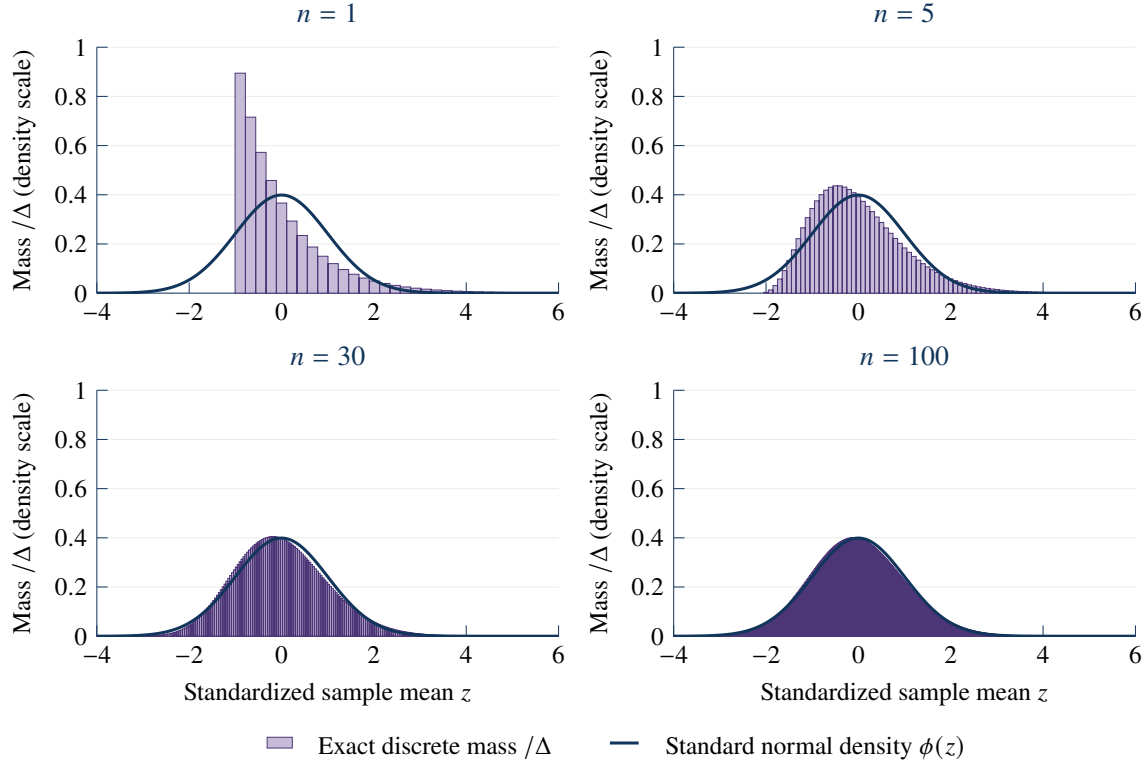


FIGURE 7: Central limit theorem for iid geometric(0.2) waiting times, counting the first success. The exact sampling probabilities of $Z_n = (W_n - 5n)/\sqrt{20n}$ form a discrete distribution at every finite n . Each complete bar has width $\Delta_n = 1/\sqrt{20n}$ and area $\Pr[Z_n = z]$; the curve is the standard normal density. The right skew is pronounced for small n and still visible at moderate n . Truncated tails are not renormalized, so the visible bar area can be less than one.

Example 15 (Geometric sample means). Let $T_i \sim \text{Geometric}(0.2)$ independently, counting attempts through the first success. In the sensor example, one observation is the waiting time for one completed delivery; a size- n sample records n independent deliveries. Here $\mu = 5$ and $\sigma^2 = 20$. For $W_n = \sum_{i=1}^n T_i$, the standardized mean is

$$Z_n = \frac{\bar{T}_n - 5}{\sqrt{20/n}} = \frac{W_n - 5n}{\sqrt{20n}}. \quad (4.55)$$

The total W_n has the negative-binomial probability mass function

$$\Pr[W_n = s] = \binom{s-1}{n-1} (0.2)^n (0.8)^{s-n}, \quad s = n, n+1, \dots \quad (4.56)$$

Here W_n counts attempts through the n th success. At $n = 1$ the distribution is strongly right-skewed; as n increases, its standardized shape approaches the same Gaussian curve as in the Bernoulli example.