

Lecture 8

Probability and Statistics

Virginia Tech Department of Computer Science
Fall 2026 • Sumeet Khatri

Why probability and statistics?

A treatment “reduces risk by 40 %.” How much can we infer from a finite sample?

Grades are “curved.” What does that mean?

Probability predicts observations from a model. Statistics estimates features of a model from observations.

A hypothetical clinical trial

2000 people are assigned at random and followed for one year using the same outcome criteria.

Treatment group: 1000 people receive the new preventive treatment.

Control group: 1000 receive a placebo, an inactive lookalike.

Risk is the probability of infection during the year. There are 100 infections in the control group and 60 in the treatment group:

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

Three ways to describe the result

$$\text{Absolute reduction} = 0.10 - 0.06 = 0.04 \quad (\text{four percentage points})$$

$$\text{Relative reduction} = \frac{0.10 - 0.06}{0.10} = 0.40 = 40 \%$$

$$\text{Risk ratio} = \frac{0.06}{0.10} = 0.60$$

The change is from 10 to 6 infections per 100 people: 4 fewer, not 40 fewer.

These are estimates from one finite trial. Repeating the trial would generally give different numbers.

Outcomes, events, and probability

Definition: The *sample space* Ω is finite or countably infinite, e.g., $\{1, 2, \dots\}$. An *event* is a subset $A \subseteq \Omega$.

$$\Pr[A] \geq 0, \quad \Pr[\Omega] = 1$$

$$\Pr\left[\bigcup_{j=1}^{\infty} A_j\right] = \sum_{j=1}^{\infty} \Pr[A_j] \quad \text{for pairwise disjoint events}$$

$$\Pr[A] = \sum_{\omega \in A} q_{\omega}, \quad q_{\omega} \geq 0, \quad \sum_{\omega \in \Omega} q_{\omega} = 1$$

For equally likely outcomes only, $\Pr[A] = |A| / |\Omega|$.

A fair six-sided die

$$\Omega = \{1,2,3,4,5,6\}, \quad q_\omega = \frac{1}{6}$$

$$A = \{2,4,6\} \quad (\text{even}), \quad B = \{4,5,6\} \quad (\text{at least four})$$

$$\Pr[A] = \frac{3}{6} = \frac{1}{2}, \quad \Pr[B] = \frac{3}{6} = \frac{1}{2}$$

$$A \cap B = \{4,6\}, \quad \Pr[A \cap B] = \frac{2}{6} = \frac{1}{3}$$

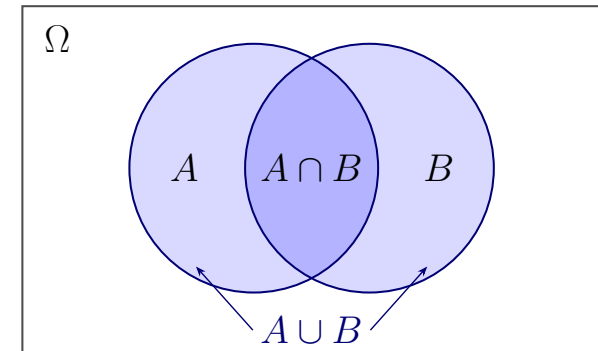
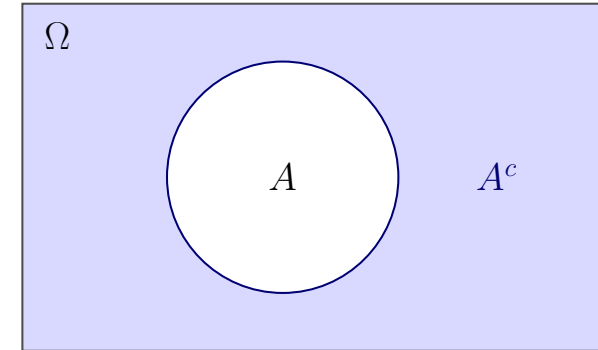
Combining events

$$\Pr[\emptyset] = 0, \quad \Pr[A^c] = 1 - \Pr[A]$$

$$\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B]$$

The intersection term removes outcomes counted twice.

$$\Pr\left[\bigcup_{j=1}^m A_j\right] \leq \sum_{j=1}^m \Pr[A_j]$$



The union bound controls the chance of at least one failure using the individual failure probabilities.

Conditioning and independence

Definition: For $\Pr[B] > 0$, conditional probability is

$$\Pr[A \mid B] = \frac{\Pr[A \cap B]}{\Pr[B]}$$

Die example: $\Pr[A \mid B] = \frac{1/3}{1/2} = \frac{2}{3}$

$$A = \{2,4,6\} \quad (\text{even}), \quad B = \{4,5,6\} \quad (\text{at least four})$$

Independence: $\Pr[A \cap B] = \Pr[A] \Pr[B]$

A random variable assigns a number to an outcome

Definition: A *random variable* $X : \Omega \rightarrow \mathbb{R}$ assigns a number to each outcome. It is discrete if its possible values are finite or countable.

Probability mass function (pmf): $p_X(x) = \Pr[X = x] = \Pr[\{\omega \in \Omega : X(\omega) = x\}]$

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

For two coin tosses: $\omega = \text{H, T}$ is an outcome, X counts heads, and $X(\text{H, T}) = 1$ is a realized value.

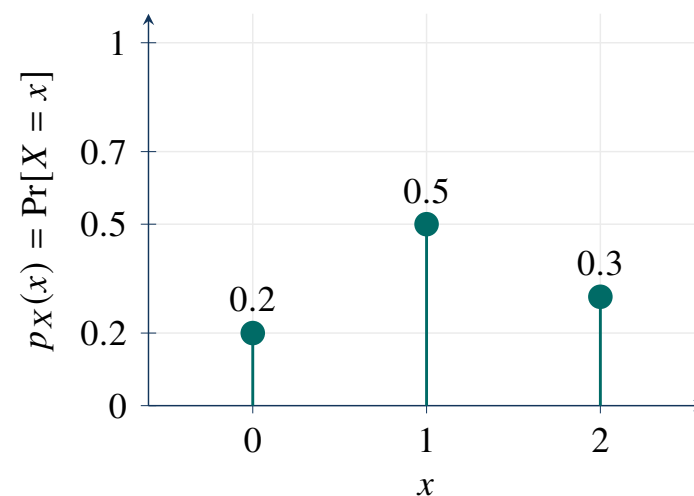
For three die rolls: $\omega = \{1, 3, 6\}$ is an outcome; if X counts the number of evens, then $X(\{1, 3, 6\}) = 1$.

A three-point distribution

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

$$\Pr[X \geq 1] = \frac{1}{2} + \frac{3}{10} = \frac{4}{5}$$

Probability mass function



Will one network packet be lost?

Bernoulli distribution

A network packet may arrive or be lost. Suppose its loss probability is $p \in [0,1]$.

$$X = \begin{cases} 1, & \text{the packet is lost,} \\ 0, & \text{the packet arrives.} \end{cases}$$

$$\Pr[X = 1] = p, \quad \Pr[X = 0] = 1 - p, \quad X \sim \text{Bernoulli}(p)$$

The label “success” means the event counted as 1 occurred. Here, that event is an undesirable packet loss.

How many packets are lost?

Binomial distribution

Transmit n packets once each. Assume losses are independent, each with the same probability p . Let X_i indicate loss of packet i .

$$Y = X_1 + \cdots + X_n \sim \text{Binomial}(n, p)$$

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

There are $\binom{n}{k}$ choices for the lost packets. At $p = 0$, $Y = 0$ with certainty; at $p = 1$, $Y = n$ with certainty.

$$n = 10, p = 0.1 : \quad \Pr[Y = 2] = \binom{10}{2} (0.1)^2 (0.9)^8 \approx 0.1937$$

Uniform outcomes and a geometric waiting time

A fair die is **uniform**: $p_X(x) = 1/6$ for $x = 1, \dots, 6$. Uniformity is a property of the model.

Geometric distribution: For independent trials with a common success probability p , let T count attempts until the first success.

$$\Pr[T = k] = (1 - p)^{k-1}p, \quad k = 1, 2, \dots \quad (0 < p < 1)$$

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

At $p = 1$, $T = 1$ with certainty; $p = 0$ is excluded. Our convention counts the successful attempt; counting failures instead gives $T - 1$.

Twenty exam scores

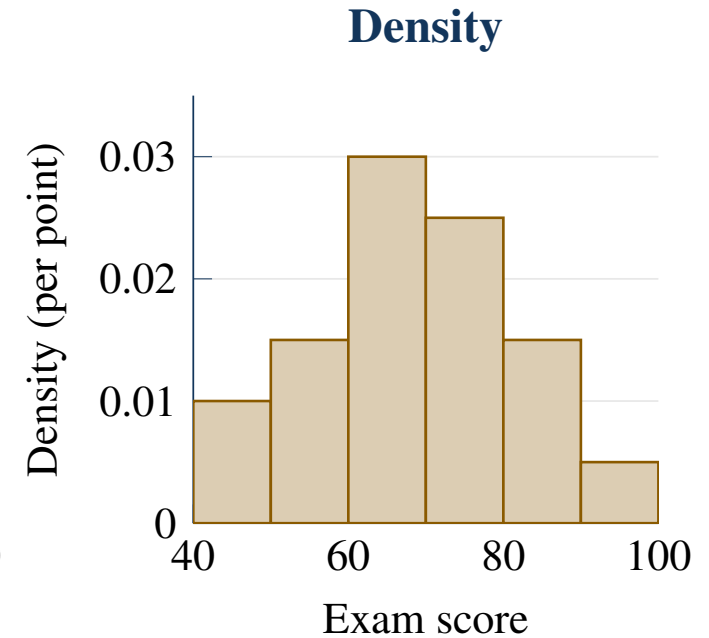
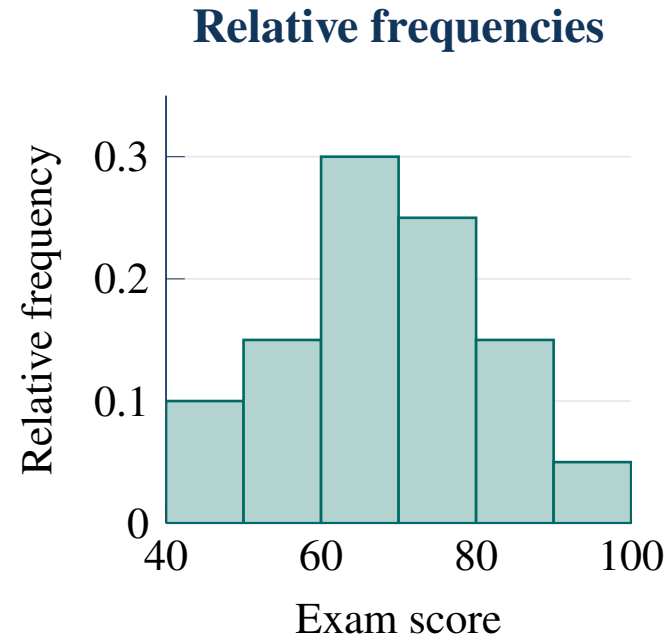
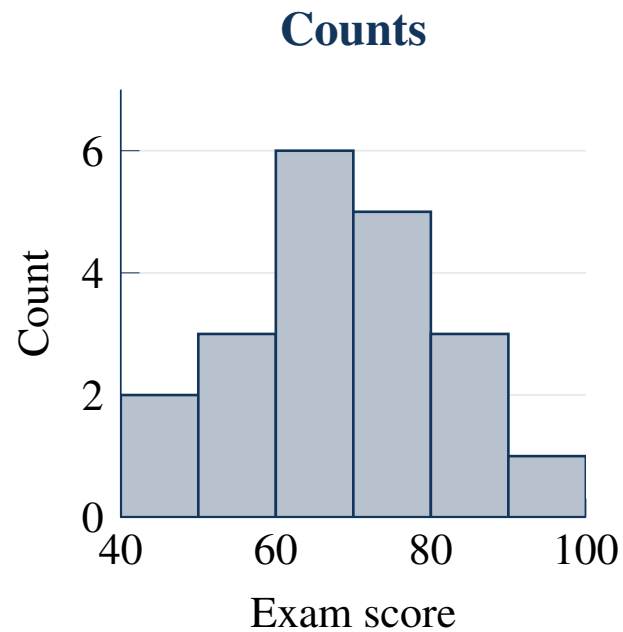
42,48,51,55,58,60,62,64,65,66,
68,70,72,74,76,78,81,84,88,93

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

A score of 60 enters [60,70). This bin has 6 scores: fraction $6/20 = 0.30$, density 0.030, and area $10(0.030) = 0.30$.

count: c_j , relative frequency: $r_j = \frac{c_j}{n}$, density: $h_j = \frac{c_j}{nw_j}$

Visualizing the data



count: c_j , relative frequency: $r_j = \frac{c_j}{n}$, density: $h_j = \frac{c_j}{nw_j}$

A histogram summarizes observed data

Choose bin edges $b_0 < \dots < b_m$ covering n observations. Use $B_j = [b_{j-1}, b_j)$, with the final right endpoint included.

$$w_j = b_j - b_{j-1}, \quad c_j = \#\{i : x_i \in B_j\}$$

$$\text{count: } c_j, \quad \text{relative frequency: } r_j = \frac{c_j}{n}, \quad \text{density: } h_j = \frac{c_j}{nw_j}$$

$$\sum_j c_j = n, \quad \sum_j r_j = 1, \quad \sum_j h_j w_j = 1$$

Recording several random variables together

Definition: The *joint pmf* records two quantities from the same experiment:

$$p_{X,Y}(x, y) = \Pr[X = x, Y = y], \quad p_{X,Y}(x, y) \geq 0, \quad \sum_x \sum_y p_{X,Y}(x, y) = 1$$

Definition: Obtain a *marginal distribution* by summing over the other variable:

$$p_X(x) = \sum_y p_{X,Y}(x, y), \quad p_Y(y) = \sum_x p_{X,Y}(x, y)$$

Two quantities from two coin tosses

Toss two independent fair coins. Let X indicate heads on the first coin; let Y count the total number of heads.

$$TT \mapsto (0,0), \quad TH \mapsto (0,1), \quad HT \mapsto (1,1), \quad HH \mapsto (1,2)$$

	$y = 0$	$y = 1$	$y = 2$	$p_X(x)$
$x = 0$	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

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

More than two random variables

A random vector $(X_1, \dots, X_m)$ has joint pmf

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

$$p_{X,Y}(x, y) = \sum_z p_{X,Y,Z}(x, y, z), \quad p_X(x) = \sum_y \sum_z p_{X,Y,Z}(x, y, z)$$

To retain selected variables, sum over all the others. *Mutual independence* means

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

Expectation as a weighted average

If I do an experiment multiple times, what is the outcome I will get “on average”?

For a discrete random variable, its **mean (expectation)** is

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

To average a cost or another function $g(X)$, use

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

The mean need not be a possible outcome

$$p_X(0) = \frac{1}{5}, \quad p_X(1) = \frac{1}{2}, \quad p_X(2) = \frac{3}{10}$$

$$\mathbb{E}[X] = 0 \left(\frac{1}{5} \right) + 1 \left(\frac{1}{2} \right) + 2 \left(\frac{3}{10} \right) = \frac{11}{10}$$

$$\mathbb{E}[X^2] = 0^2 \left(\frac{1}{5} \right) + 1^2 \left(\frac{1}{2} \right) + 2^2 \left(\frac{3}{10} \right) = \frac{17}{10}$$

The mean is 1.1, but every draw is 0, 1, or 2.

A fair die similarly has mean $(1 + 2 + 3 + 4 + 5 + 6)/6 = 3.5$.

Linearity of expectation

Let $X_1, \dots, X_m$ have finite expectations. For constants $a_1, \dots, a_m, b$,

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

Independence is not required.

The joint distribution is enough to calculate the expectation of a sum.

Variance and standard deviation

How “spread out” is the data from the experiments?

Assume $\mathbb{E}[X^2] < \infty$ and let $\mu = \mathbb{E}[X]$.

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

$$\text{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

$$\text{Var}[aX + b] = a^2 \text{Var}[X]$$

If X is in seconds, variance is in seconds squared and standard deviation is in seconds.

Variance in the running example

$$p_X(0) = \frac{1}{5}, \quad p_X(1) = \frac{1}{2}, \quad p_X(2) = \frac{3}{10}$$

$$\mathbb{E}[X] = 0 \left(\frac{1}{5} \right) + 1 \left(\frac{1}{2} \right) + 2 \left(\frac{3}{10} \right) = \frac{11}{10}$$

$$\mathbb{E}[X^2] = 0^2 \left(\frac{1}{5} \right) + 1^2 \left(\frac{1}{2} \right) + 2^2 \left(\frac{3}{10} \right) = \frac{17}{10}$$

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

Bernoulli and binomial moments

For $X \sim \text{Bernoulli}(p)$, the identity $X^2 = X$ gives

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

For a binomial $Y = X_1 + \dots + X_m$,

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

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

Time to the first event

Let $T \sim \text{Geometric}(p)$ count attempts including the first success, with $0 < p \leq 1$.

$$\mathbb{E}[T] = \frac{1}{p}, \quad \text{Var}[T] = \frac{1-p}{p^2}$$

The number of attempts and elapsed time are different quantities.

To turn attempts into seconds, specify when attempts occur and how long each takes.

A sensor retries until a packet gets through

A sensor makes successive two-second attempts. Each independently succeeds with probability $p = 0.2$.

A failed attempt is followed immediately by the next attempt.

$$T \sim \text{Geometric}(0.2)$$

$$\mathbb{E}[T] = 5, \quad \text{Var}[T] = 20$$

Repeated draws from the same distribution

In practice, we do not know the underlying distribution of the data — so we have to take *samples*.

Sampling means producing an outcome according to a probability distribution.

Samples $X_1, \dots, X_n$ are iid when they are independent and all have the same distribution.

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

We use the observations to estimate an unknown mean μ and variance σ^2 .

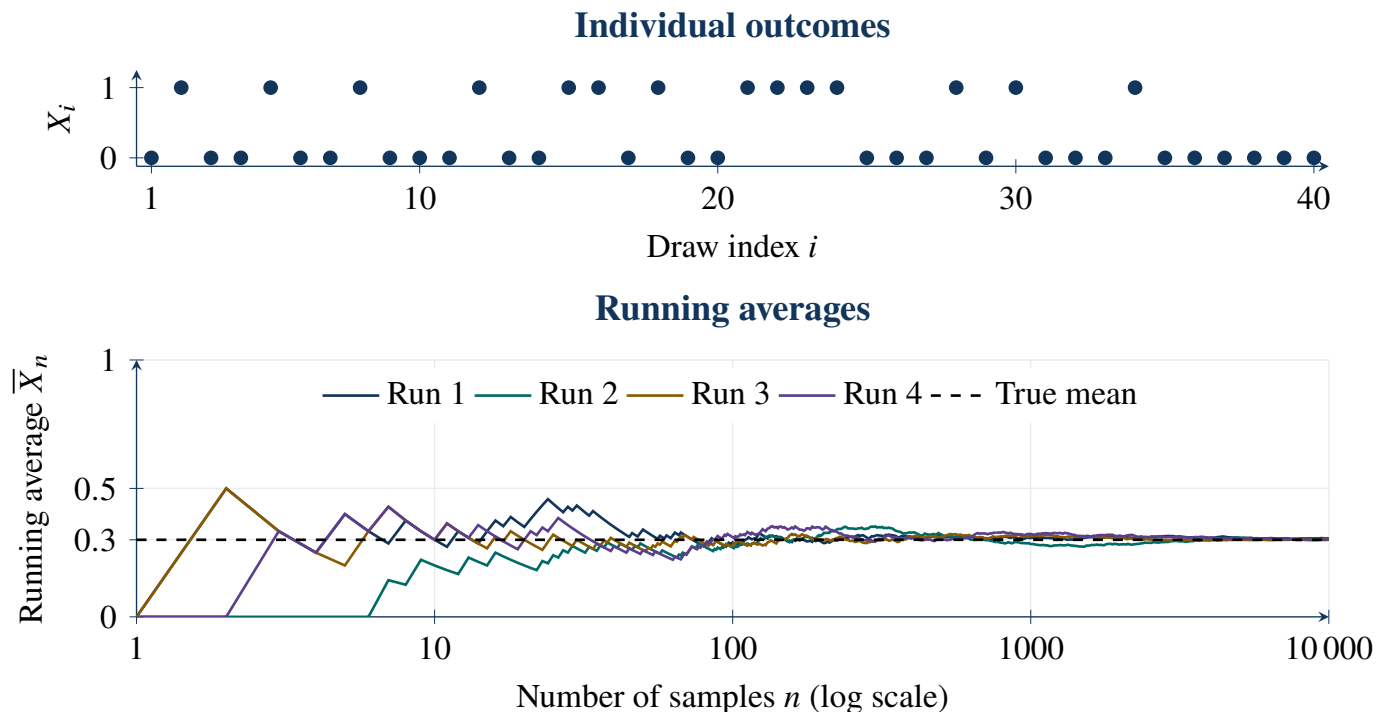
Examples: tossing a coin many times; rolling a die many times; collecting data from a clinical trial.

Estimating the probability of success

Consider an experiment that succeeds with probability $p = 0.3$ and fails with probability $1 - p$. We don't know p , so we must estimate it from data.

$$\Pr[X = 1] = p, \quad \Pr[X = 0] = 1 - p$$

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



Estimating the mean and variance

In practice, we do not know the underlying distribution of the data — so we have to take *samples*.

More generally: the underlying random variable X has mean μ and variance σ^2 — how do we estimate these from the data?

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

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

$$(x_1, x_2, x_3, x_4, x_5) = (0, 1, 1, 2, 1), \quad \bar{x}_5 = 1 \quad s_5^2 = \frac{(0-1)^2 + 0 + 0 + (2-1)^2 + 0}{4} = \frac{1}{2}$$

Why does sampling work? The law of large numbers

Let $X_1, X_2, \dots$ be iid with mean μ and finite variance σ^2 . The sample mean approaches the true mean as the number of samples increases:

$$\bar{X}_n \xrightarrow{\text{Pr}} \mu \iff \lim_{n \rightarrow \infty} \Pr[|\bar{X}_n - \mu| > \varepsilon] = 0 \quad (\varepsilon > 0)$$

$$\Pr[|\bar{X}_n - \mu| \geq \varepsilon] \leq \frac{\sigma^2}{n\varepsilon^2}, \quad n \geq 1$$

Adjusting/“curving” grades

Original scores: 50, 60, 65, 65, 70, 80.

$$\bar{x}_6 = 390/6 = 65, \quad s_6 = \sqrt{(225 + 25 + 0 + 0 + 25 + 225)/5} = 10$$

Target mean 75, target standard deviation 8.

Define the new grades as follows:

$$g_i = 75 + 8 \underbrace{\left(\frac{x_i - 65}{10} \right)}_{z_i}$$

Apply the rule to every student

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

For $x = 80$: $z = 1.5$ and $g = 87$. The new mean is 75 and the new sample SD is 8.

Curving scores to a chosen mean and spread

For $n \geq 2$, choose target mean m_* and sample standard deviation $s_* > 0$. Compute the class statistics first.

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

$$z_i = \frac{x_i - \bar{x}_n}{s_n}, \quad g_i = m_* + s_* z_i$$

Requires $s_n > 0$. Before rounding or capping, the curved mean is m_* and the curved sample SD is s_* . Ordering and ties are preserved.

Curving can lower a grade!

The class 60, 70, 75, 75, 80, 90 has mean 75 and sample standard deviation 10. Choose $m_* = 75$, $s_* = 8$.

$$g_i = 75 + 0.8(x_i - 75)$$

x_i	Students	g_i	$g_i - x_i$
60	1	63	+3
70	1	71	+1
75	2	75	0
80	1	79	-1
90	1	87	-3

A class of 300 students

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

$$\bar{x}_{300} = \frac{21768}{300} = 72.56$$

$$s_{300} = \sqrt{\frac{1628800 - 300(72.56)^2}{299}} \approx 12.8425$$

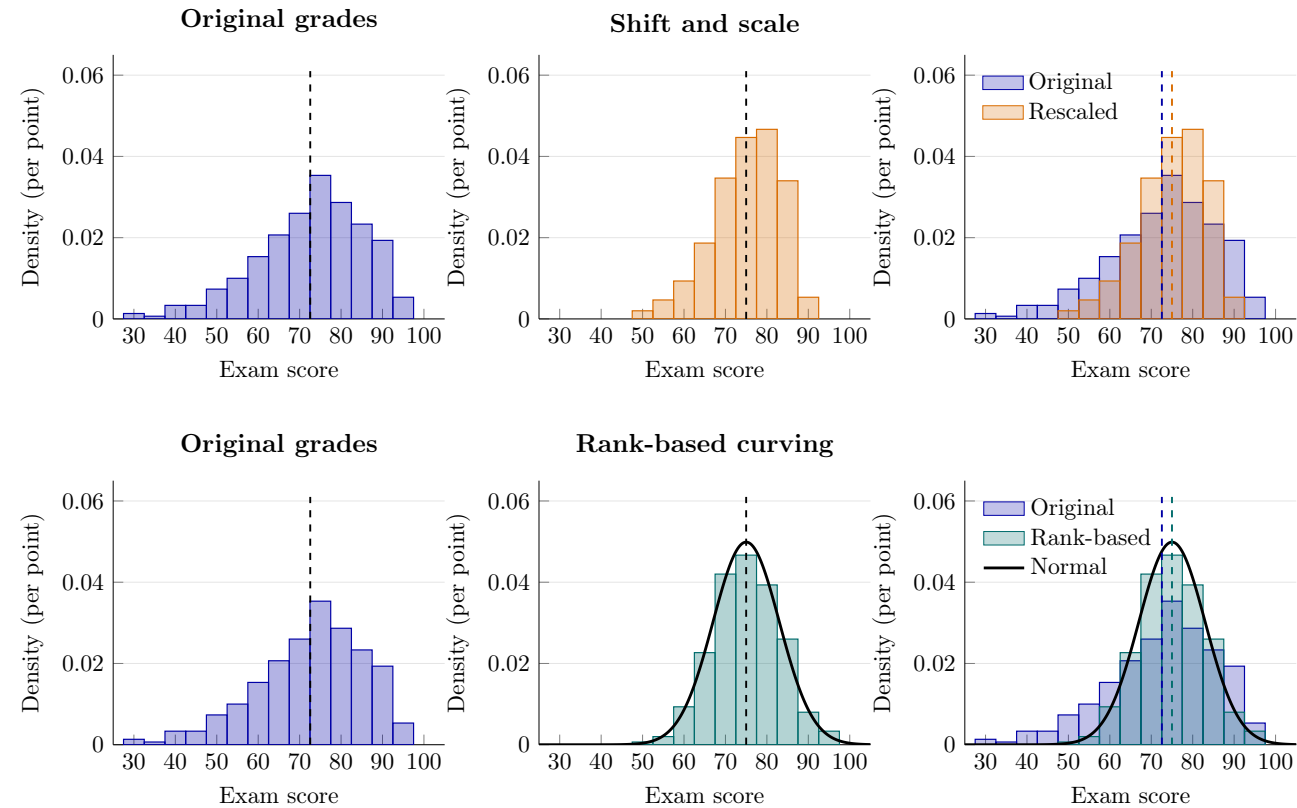
Shift and rescale the large class

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

x_i	g_i (approximately)
50	60.95
80	79.63
90	85.86

The new mean is exactly 75 and the new sample standard deviation is exactly 8.

The two grading procedures compared



Top: shift and rescale. **Bottom:** assign a bell-shaped target.

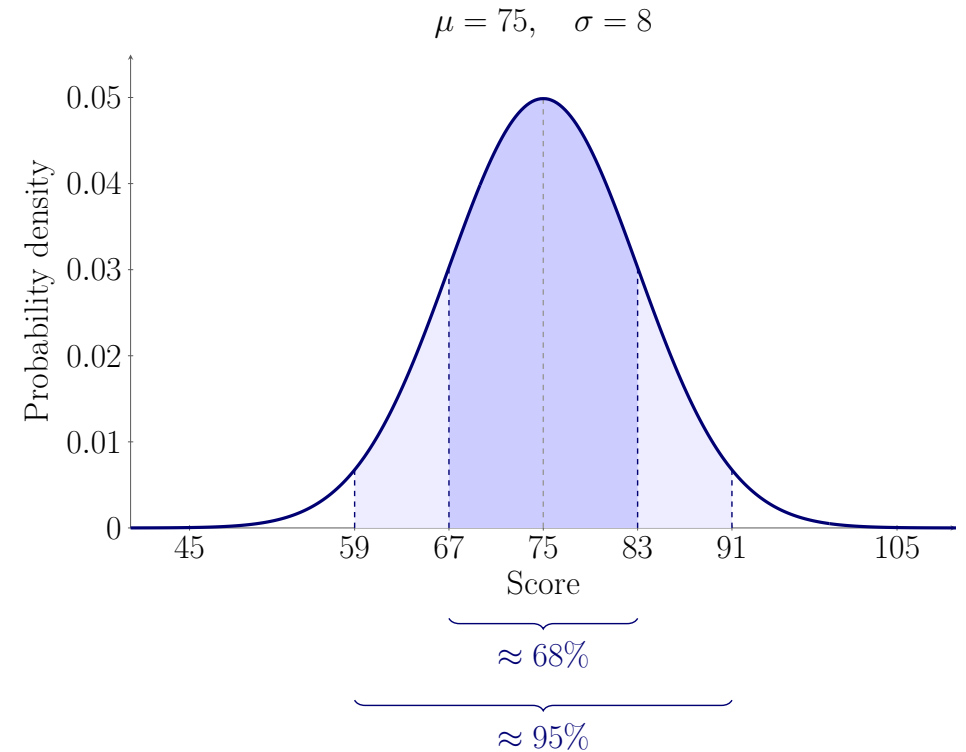
The normal distribution models a bell-shaped density

For $\mu \in \mathbb{R}$ and $\sigma > 0$, $X \sim \mathcal{N}(\mu, \sigma^2)$ has mean μ and variance σ^2 .

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}$$

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

Grade target: $f(x) = \frac{1}{8\sqrt{2\pi}} e^{-(x-75)^2/(2 \cdot 8^2)}$



The standardized mean approaches a normal distribution

Let $X_1, X_2, \dots$ be iid with mean μ and variance $0 < \sigma^2 < \infty$.

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

$$\Pr[Z_n \leq z] \longrightarrow \Phi(z) \quad \text{for every real } z$$

The LLN controls the estimate itself. The CLT describes its standardized error across repeated samples.

Reading exact sampling distributions on a density scale

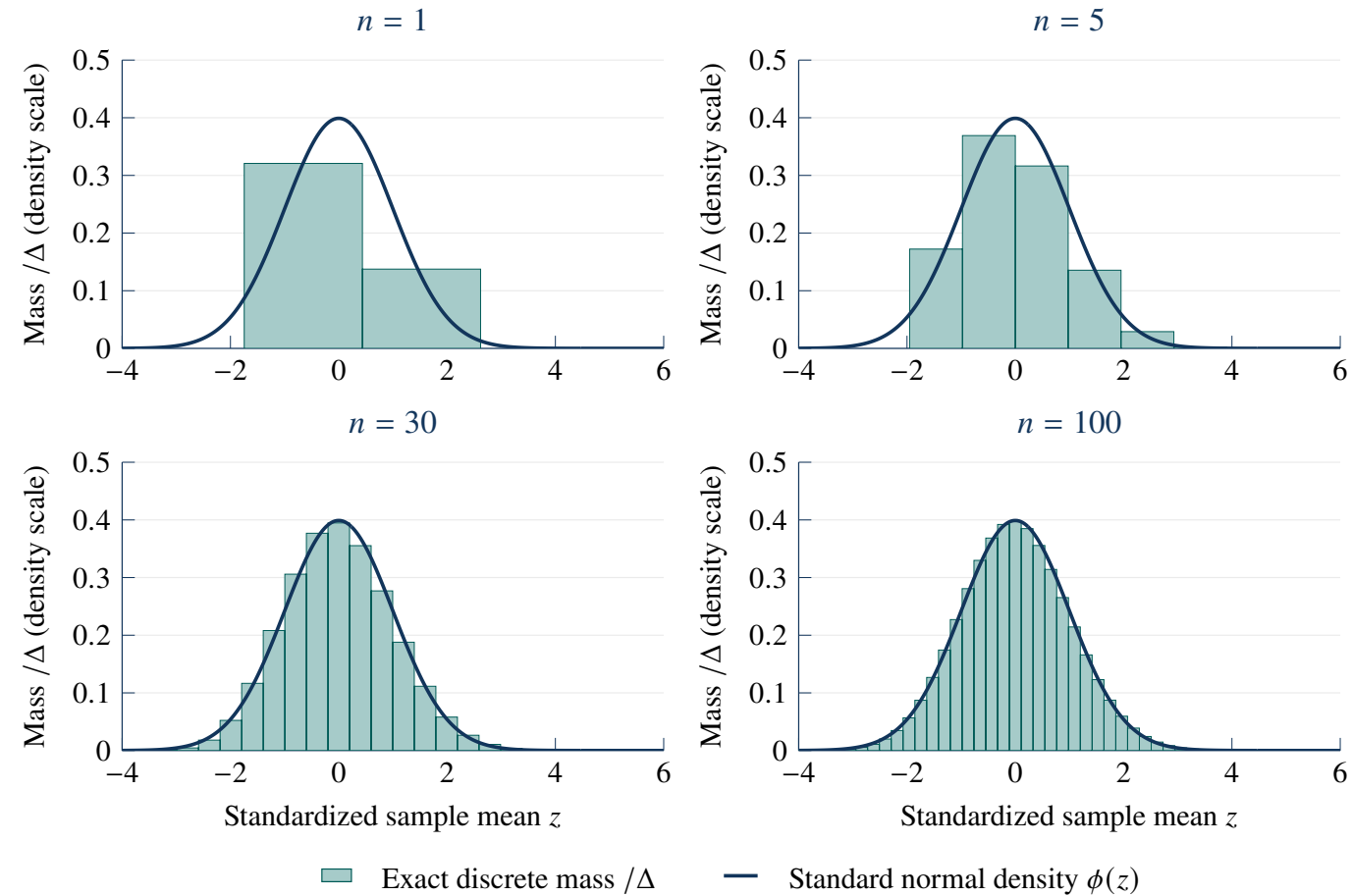
For each n , repeatedly draw a fresh size- n sample, compute its mean, and standardize it.

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

$$\Delta_n = \frac{1}{\sigma\sqrt{n}}, \quad \text{height} = \frac{\Pr[Z_n = z]}{\Delta_n}, \quad \text{full bar area} = \Pr[Z_n = z]$$

For Bernoulli and geometric observations, exact formulas determine these discrete masses. Omitted tails are not renormalized.

Binomial distributions approach the bell curve



Geometric distribution also approaches the curve

