Chapter 12

Linear Algebra in Probability & Statistics

12.1 Mean, Variance, and Probability

We are starting with the three fundamental words of this chapter: *mean, variance, and probability*. Let me give a rough explanation of their meaning before I write any formulas:

The **mean** is the *average value* or expected value

The **variance σ^2** measures the average *squared distance* from the mean m

The **probabilities** of n different outcomes are positive numbers $p_1, \ldots, p_n$ adding to 1.

Certainly the mean is easy to understand. We will start there. But right away we have two different situations that you have to keep straight. On the one hand, we may have the results (*sample values*) from a completed trial. On the other hand, we may have the expected results (*expected values*) from future trials. Let me give examples:

Sample values Five random freshmen have ages $\mathbf{18}, \mathbf{17}, \mathbf{18}, \mathbf{19}, \mathbf{17}$

Sample mean $\frac{\mathbf{1}}{\mathbf{5}}(18 + 17 + 18 + 19 + 17) = \mathbf{17.8}$

Probabilities The ages in a freshmen class are $17\,(\mathbf{20\%}), 18\,(\mathbf{50\%}), 19\,(\mathbf{30\%})$

A random freshman has **expected age** $\mathbf{E}\,[\mathbf{x}] = (0.2)\,17 + (0.5)\,18 + (0.3)\,19 = \mathbf{18.1}$

Both numbers 17.8 and 18.1 are correct averages. The sample mean starts with N samples $x_1, \ldots, x_N$ from a completed trial. Their mean is the *average* of the N observed samples:

$$\textbf{Sample mean} \qquad m = \mu = \frac{\mathbf{1}}{\boldsymbol{N}}(\boldsymbol{x_1} + \boldsymbol{x_2} + \cdots + \boldsymbol{x_N}) \tag{1}$$

The **expected value of x** starts with the probabilities $p_1, \dots, p_n$ of the ages $x_1, \dots, x_n$:

$$\textbf{Expected value} \quad \boldsymbol{m = \mathrm{E}[x] = p_1 x_1 + p_2 x_2 + \cdots + p_n x_n}. \tag{2}$$

This is $\boldsymbol{p \cdot x}$. Notice that $m = \mathrm{E}[x]$ tells us what to expect, $m = \mu$ tells us what we got.

By taking many samples (large N), the sample results will come close to the probabilities. The "Law of Large Numbers" says that with probability 1, the sample mean will converge to its expected value $\mathrm{E}[x]$ as the sample size N increases. A fair coin has probability $p_0 = \frac{1}{2}$ of tails and $p_1 = \frac{1}{2}$ of heads. Then $\mathrm{E}[x] = \left(\frac{1}{2}\right) 0 + \frac{1}{2}(1)$. The fraction of heads in N flips of the coin is the sample mean, expected to approach $\mathrm{E}[x] = \frac{1}{2}$.

This does *not* mean that if we have seen more tails than heads, the next sample is likely to be heads. The odds remain 50-50. The first 100 or 1000 flips do affect the sample mean. *But* 1000 *flips will not affect its limit*—because you are dividing by $N \to \infty$.

Variance (around the mean)

The **variance $\boldsymbol{\sigma^2}$** measures expected distance (squared) from the expected mean $\mathrm{E}[x]$. The **sample variance $\boldsymbol{S^2}$** measures actual distance (squared) from the sample mean. The square root is the **standard deviation $\boldsymbol{\sigma}$ or $\boldsymbol{S}$**. After an exam, I email μ and S to the class. I don't know the expected mean and variance because I don't know the probabilities p_1 to p_{100} for each score. (After teaching for 50 years, I still have no idea what to expect.)

The deviation is always deviation *from the mean*—sample or expected. We are looking for the size of the "spread" around the mean value $x = m$. Start with N samples.

$$\textbf{Sample variance} \qquad \boldsymbol{S^2 = \frac{1}{N-1}\left[(x_1 - m)^2 + \cdots + (x_N - m)^2\right]} \tag{3}$$

The sample ages $x = 18, 17, 18, 19, 17$ have mean $m = 17.8$. That sample has variance 0.7 :

$$S^2 = \frac{1}{4}\left[(.2)^2 + (-.8)^2 + (.2)^2 + (1.2)^2 + (-.8)^2\right] = \frac{1}{4}(2.8) = \mathbf{0.7}$$

The minus signs disappear when we compute squares. Please notice ! Statisticians divide by $N - 1 = 4$ (and not $N = 5$) so that S^2 is an unbiased estimate of σ^2. One degree of freedom is already accounted for in the sample mean.

An important identity comes from splitting each $(x - m)^2$ into $x^2 - 2mx + m^2$:

$$\begin{aligned} \text{sum of } (x_i - m)^2 &= (\text{sum of } x_i^2) - 2m(\text{sum of } x_i) + (\text{sum of } m^2) \\ &= (\text{sum of } x_i^2) - 2m(Nm) + Nm^2 \\ \textbf{sum of } \boldsymbol{(x_i - m)^2} &= \textbf{(sum of } \boldsymbol{x_i^2}\textbf{)} - \boldsymbol{Nm^2}. \end{aligned} \tag{4}$$

This is an equivalent way to find $(x_1 - m)^2 + \cdots + (x_N - m^2)$ by adding $x_1^2 + \cdots + x_N^2$.

Now start with probabilities p_i (never negative !) instead of samples. We find expected values instead of sample values. The variance σ^2 is the crucial number in statistics.

$$\textbf{Variance} \quad \boldsymbol{\sigma^2 = \mathrm{E}\left[(x-m)^2\right] = p_1(x_1-m)^2 + \cdots + p_n(x_n-m)^2}. \qquad (5)$$

We are squaring the distance from the expected value $m = \mathrm{E}[x]$. We don't have samples, only expectations. We know probabilities but we don't know experimental outcomes.

Example 1 Find the variance σ^2 of the ages of college freshmen.

Solution The probabilities of ages $x_i = 17, 18, 19$ were $p_i = 0.2$ and 0.5 and 0.3. The expected value was $\boldsymbol{m = \sum p_i x_i = 18.1}$. The variance uses those same probabilities :

$$\begin{aligned}\boldsymbol{\sigma^2} &= (0.2)(\mathbf{17-18.1})^2 + (0.5)(\mathbf{18-18.1})^2 + (0.3)(\mathbf{19-18.1})^2 \\ &= (0.2)\mathbf{(1.21)} + (0.5)\mathbf{(0.01)} + (0.3)\mathbf{(0.81)} = \mathbf{0.49}.\end{aligned}$$

The **standard deviation** is the square root $\boldsymbol{\sigma = 0.7}$.
This measures the spread of $17, 18, 19$ around $\mathrm{E}[x]$, weighted by probabilities .2, .5, .3.

Continuous Probability Distributions

Up to now we have allowed for n possible outcomes $x_1, \ldots, x_n$. With ages 17, 18, 19, we only had $n = 3$. If we measure age in days instead of years, there will be a thousand possible ages (too many). Better to allow *every number between* 17 *and* 20—a continuum of possible ages. Then the probabilities p_1, p_2, p_3 for ages x_1, x_2, x_3 have to move to a **probability distribution $\boldsymbol{p(x)}$** for a whole continuous range of ages $17 \le x \le 20$.

The best way to explain probability distributions is to give you two examples. They will be the **uniform distribution** and the **normal distribution**. The first (uniform) is easy. The normal distribution is all-important.

Uniform distribution Suppose ages are uniformly distributed between 17.0 and 20.0. All ages between those numbers are "equally likely". Of course any one exact age has no chance at all. There is zero probability that you will hit the exact number $x = 17.1$ or $x = 17 + \sqrt{2}$. What you can truthfully provide (assuming our uniform distribution) is **the chance $\boldsymbol{F(x)}$ that a random freshman has age less than $\boldsymbol{x}$** :

The chance of age less than $x = 17$ is $F(17) = \mathbf{0}$	$x \le 17$ won't happen
The chance of age less than $x = 20$ is $F(20) = \mathbf{1}$	$x \le 20$ will happen
The chance of age less than x is $\boldsymbol{F(x) = \frac{1}{3}(x-17)}$	$\boldsymbol{F}$ **goes from 0 to 1**

That formula $F(x) = \frac{1}{3}(x - 17)$ gives $F = 0$ at $x = 17$; then $x \le 17$ won't happen. It gives $F(x) = 1$ at $x = 20$; then $x \le 20$ is sure. Between 17 and 20, the graph of the **cumulative distribution $\boldsymbol{F(x)}$** increases linearly for this uniform model.

Let me draw the graphs of $F(x)$ and its derivative $p(x)$ = "probability density function".

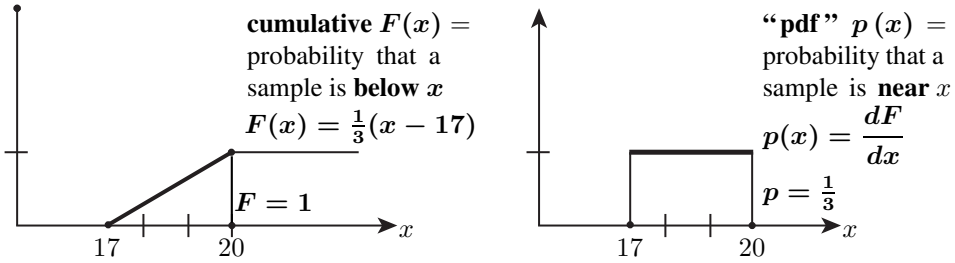

Figure 12.1: $F(x)$ is the cumulative distribution and its derivative $p(x) = dF/dx$ is the **probability density function** (**pdf**). For this uniform distribution, $p(x)$ is constant between 17 and 20. The total area under the graph of $p(x)$ is the total probability $F = 1$.

You could say that $p(x)\,dx$ is the probability of a sample falling in between x and $x + dx$. This is "infinitesimally true": $p(x)\,dx$ is $F(x + dx) - F(x)$. Here is the full truth:

$$\boldsymbol{F = \text{integral of } p} \qquad \textbf{Probability of } a \le x \le b = \int_a^b p(x)\,dx = F(b) - F(a) \quad (6)$$

$F(b)$ is the probability of $x \le b$. I subtract $F(a)$ to keep $x \ge a$. That leaves $a \le x \le b$.

Mean and Variance of $p(x)$

What are the mean m and variance σ^2 for a probability distribution? Previously we added $p_i x_i$ to get the mean (expected value). With a continuous distribution we **integrate** $xp(x)$:

$$\textbf{Mean} \qquad m = \mathbf{E}[x] = \int x\,p(x)\,dx = \int_{x=17}^{20} (x)\left(\frac{1}{3}\right) dx = 18.5$$

For this uniform distribution, the mean m is halfway between 17 and 20. Then the probability of a random value x below this halfway point $m = 18.5$ is $F(m) = \frac{1}{2}$.

In MATLAB, x = rand(1) chooses a random number uniformly between 0 and 1. Then the expected mean is $m = \frac{1}{2}$. The interval from 0 to x has probability $F(x) = x$. The interval below the mean m always has probability $F(m) = \frac{1}{2}$.

The variance is the average squared distance to the mean. With N outcomes, σ^2 is the sum of $p_i(x_i - m)^2$. For a continuous random variable x, the sum changes to an **integral**.

$$\textbf{Variance} \qquad \sigma^2 = \mathbf{E}\left[(x - m)^2\right] = \int p(x)\,(x - m)^2\,dx \quad (7)$$

When ages are uniform between $17 \le x \le 20$, the integral can shift to $0 \le x \le 3$:

$$\sigma^2 = \int_{17}^{20} \frac{1}{3}(x-18.5)^2\,dx = \int_{0}^{3} \frac{1}{3}(x-1.5)^2\,dx = \frac{1}{9}(x-1.5)^3\Big]_{x=0}^{x=3} = \frac{2}{9}(1.5)^3 = \frac{3}{4}.$$

That is a typical example, and here is the complete picture for a uniform $p(x)$, 0 to a.

Uniform distribution for $0 \le x \le a$ **Density $p(x) = \dfrac{1}{a}$** **Cumulative $F(x) = \dfrac{x}{a}$**

$$\textbf{Mean } m = \frac{a}{2} \text{ halfway} \qquad \textbf{Variance } \sigma^2 = \int_0^a \frac{1}{a}\left(x - \frac{a}{2}\right)^2 dx = \frac{a^2}{12} \tag{8}$$

The mean is a multiple of a, the variance is a multiple of a^2. For $a = 3$, $\sigma^2 = \frac{9}{12} = \frac{3}{4}$. For one random number between 0 and 1 $\left(\text{mean } \frac{1}{2}\right)$ the variance is $\sigma^2 = \frac{1}{12}$.

Normal Distribution : Bell-shaped Curve

The normal distribution is also called the "Gaussian" distribution. It is the most important of all probability density functions $p(x)$. The reason for its overwhelming importance comes from repeating an experiment and averaging the outcomes. The experiments have their own distribution (like heads and tails). *The average approaches a normal distribution.*

Central Limit Theorem (informal) The average of N samples of "any" probability distribution approaches a normal distribution as $N \to \infty$.

Start with the "standard normal distribution". It is symmetric around $x = 0$, so its mean value is $m = 0$. It is chosen to have a standard variance $\sigma^2 = 1$. It is called $\mathbf{N}(0,1)$.

$$\textbf{Standard normal distribution} \qquad p(x) = \frac{1}{\sqrt{2\pi}}\,e^{-x^2/2}. \tag{9}$$

The graph of $p(x)$ is the **bell-shaped curve** in Figure 12.2. The standard facts are

$$\textbf{Total probability} = 1 \qquad \int_{-\infty}^{\infty} p(x)\,dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2}\,dx = 1$$

$$\textbf{Mean } \mathbf{E}[x] = 0 \qquad m = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-x^2/2}\,dx = 0$$

$$\textbf{Variance } \mathbf{E}[x^2] = 1 \qquad \sigma^2 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-0)^2 e^{-x^2/2}\,dx = 1$$

The zero mean was easy because we are integrating an odd function. Changing x to $-x$ shows that "integral $=-$ integral". So that integral must be $m = 0$.

The other two integrals apply the idea in Problem 12 to reach 1. Figure 12.2 shows a graph of $p(x)$ for the normal distribution $\mathbf{N}(0, \sigma)$ and also its cumulative distribution $F(x) =$ integral of $p(x)$. From the symmetry of $p(x)$ you see *mean = zero*. From $F(x)$ you see a very important practical approximation for opinion polling:

The probability that a random sample falls between $-\sigma$ and σ is $\boldsymbol{F(\sigma) - F(-\sigma) \approx \frac{2}{3}}$.

This is because $\int_{-\sigma}^{\sigma} p(x)\,dx$ equals $\int_{-\infty}^{\sigma} p(x)\,dx - \int_{-\infty}^{\sigma} p(x)\,dx = F(\sigma) - F(-\sigma)$.

Similarly, the probability that a random x lies between -2σ and 2σ ("*less than two standard deviations from the mean*") is $F(2\sigma) - F(-2\sigma) \approx 0.95$. If you have an experimental result further than 2σ from the mean, it is fairly sure to be not accidental: chance $= 0.05$. Drug tests may look for a tighter confirmation, like probability 0.001. Searching for the Higgs boson used a hyper-strict test of 5σ deviation from pure accident.

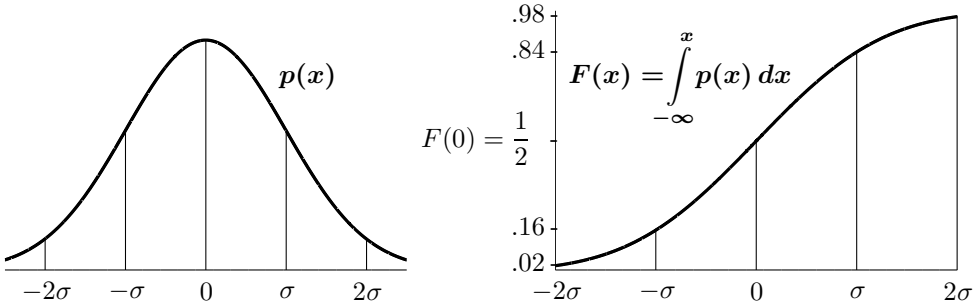

Figure 12.2: The standard normal distribution $p\,(x)$ has mean $m = 0$ and $\sigma = 1$.

The normal distribution with any mean m and standard deviation σ comes by shifting and stretching the standard $\mathbf{N}\,(0, 1)$. **Shift $\boldsymbol{x}$ to $\boldsymbol{x - m}$. Stretch $\boldsymbol{x - m}$ to $\boldsymbol{(x - m)/\sigma}$.**

Gaussian density $p(x)$
Normal distribution $\mathbf{N}(m, \sigma)$

$$\boldsymbol{p(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x-m)^2/2\sigma^2}} \tag{10}$$

The integral of $p(x)$ is $F(x)$—the probability that a random sample will fall below x. The differential $p(x)\,dx = F(x + dx) - F(x)$ is the probability that a random sample will fall between x and $x + dx$. There is no simple formula to integrate $e^{-x^2/2}$, so this cumulative distribution $F(x)$ is computed and tabulated very carefully.

N Coin Flips and $N \to \infty$

Example 2 **Suppose x is 1 or -1 with equal probabilities $p_1 = p_{-1} = \frac{1}{2}$.**

The mean value is $\boldsymbol{m} = \frac{1}{2}(1) + \frac{1}{2}(-1) = \mathbf{0}$. The variance is $\boldsymbol{\sigma^2} = \frac{1}{2}(1)^2 + \frac{1}{2}(-1)^2 = \mathbf{1}$.

The key question is the *average* $A_N = (x_1 + \cdots + x_N)/N$. The independent x_i are ± 1 and we are dividing their sum by N. The expected mean of A_N is still zero. The law of large numbers says that this sample average approaches zero with probability 1. How fast does A_N approach zero? **What is its variance $\boldsymbol{\sigma_N^2}$?**

$$\text{By linearity} \quad \boldsymbol{\sigma_N^2} = \frac{\sigma^2}{N^2} + \frac{\sigma^2}{N^2} + \cdots + \frac{\sigma^2}{N^2} = N\frac{\sigma^2}{N^2} = \boldsymbol{\frac{1}{N}} \quad \text{since } \sigma^2 = 1. \tag{11}$$

Example 3 **Change outputs from 1 or -1 to $x = 1$ or $x = 0$. Keep $p_1 = p_0 = \frac{1}{2}$.**

The new mean value $\boldsymbol{m} = \frac{1}{2}$ falls halfway between 0 and 1. The variance moves to $\boldsymbol{\sigma^2} = \frac{1}{4}$:

$$\boldsymbol{m} = \frac{1}{2}(1) + \frac{1}{2}(0) = \boldsymbol{\frac{1}{2}} \quad \text{and} \quad \boldsymbol{\sigma^2} = \frac{1}{2}\left(1 - \frac{1}{2}\right)^2 + \frac{1}{2}\left(0 - \frac{1}{2}\right)^2 = \boldsymbol{\frac{1}{4}}.$$

$$\text{The average } A_N \text{ now has mean } \frac{1}{2} \text{ and variance } \frac{1}{4N^2} + \cdots + \frac{1}{4N^2} = \boldsymbol{\frac{1}{4N}} = \boldsymbol{\sigma_N^2}. \tag{12}$$

This σ_N is half the size of σ_N in Example 2. This must be correct because the new range 0 to 1 is half as long as -1 to 1. Examples 2-3 are showing a law of linearity.

The new $\mathbf{0-1}$ variable $\boldsymbol{x_{\text{new}}}$ is $\frac{1}{2}\boldsymbol{x_{\text{old}}} + \frac{1}{2}$. So the mean m is increased to $\frac{1}{2}$ and the variance is *multiplied* by $\left(\frac{1}{2}\right)^2$. A shift changes m and the rescaling changes σ^2.

Linearity $\boldsymbol{x_{\text{new}} = ax_{\text{old}} + b}$ has $\boldsymbol{m_{\text{new}} = am_{\text{old}} + b}$ and $\boldsymbol{\sigma^2{}_{\text{new}} = a^2\sigma^2{}_{\text{old}}}$. (13)

Here are the results from three numerical tests: random 0 or 1 averaged over N trials.

[**48** 1's from $N = \mathbf{100}$] [**5035** 1's from $N = \mathbf{10000}$] [**19967** 1's from $N = \mathbf{40000}$].

The standardized $\boldsymbol{X}$ $= (x - m)/\sigma = \left(A_N - \frac{1}{2}\right) / 2\sqrt{N}$ was [$\mathbf{-.40}$] [$\mathbf{.70}$] [$\mathbf{-.33}$].

The Central Limit Theorem says that the average of many coin flips will approach a normal distribution. Let us begin to see how that happens: **binomial approaches normal**.

For each flip, the probability of heads is $\frac{1}{2}$. For $N = 3$ flips, the probability of heads all three times is $\left(\frac{1}{2}\right)^3 = \frac{1}{8}$. The probability of heads twice and tails once is $\frac{3}{8}$, from three sequences HHT and HTH and THH. These numbers $\boldsymbol{\frac{1}{8}}$ and $\boldsymbol{\frac{3}{8}}$ are pieces of $\left(\frac{1}{2} + \frac{1}{2}\right)^3 = \boldsymbol{\frac{1}{8} + \frac{3}{8} + \frac{3}{8} + \frac{1}{8}} = 1$. *The average number of heads in* 3 *flips is* 1.5.

$$\textbf{Mean} \quad m = (3 \text{ heads})\frac{1}{8} + (2 \text{ heads})\frac{3}{8} + (1 \text{ head})\frac{3}{8} + 0 = \frac{3}{8} + \frac{6}{8} + \frac{3}{8} = \textbf{1.5 heads}$$

With N flips, Example 3 (or common sense) gives a mean of $\boldsymbol{m = \Sigma\, x_i p_i = \frac{1}{2}N}$ heads.

The variance σ^2 is based on the *squared distance* from this mean $N/2$. With $N = 3$ the variance is $\sigma^2 = \frac{3}{4}$ (*which is* $N/4$). To find σ^2 we add $(x_i - m)^2\, p_i$ with $m = 1.5$:

$$\boldsymbol{\sigma^2} = (3-1.5)^2\,\frac{1}{8} + (2-1.5)^2\,\frac{3}{8} + (1-1.5)^2\,\frac{3}{8} + (0-1.5)^2\,\frac{1}{8} = \frac{9+3+3+9}{32} = \boldsymbol{\frac{3}{4}}.$$

For any N, the variance is $\boldsymbol{\sigma_N^2 = N/4}$. Then $\sigma_N = \sqrt{N}/2$.

Figure 12.3 shows how the probabilities of 0, 1, 2, 3, 4 heads in $N = 4$ flips come close to a bell-shaped Gaussian. That Gaussian is centered at the mean value $N/2 = 2$. To reach the standard Gaussian (mean 0 and variance 1) we shift and rescale that graph. If x is the number of heads in N flips—the average of N zero-one outcomes—then x is shifted by its mean $m = N/2$ and rescaled by $\sigma = \sqrt{N}/2$ to produce the standard X:

Shifted and scaled $\quad \boldsymbol{X = \dfrac{x - m}{\sigma} = \dfrac{x - \frac{1}{2}N}{\sqrt{N}/2}} \quad (N = 4 \text{ has } X = x - 2)$

Subtracting m is "centering" or "detrending". The mean of X is zero.

Dividing by σ is "normalizing" or "standardizing". The variance of X is 1.

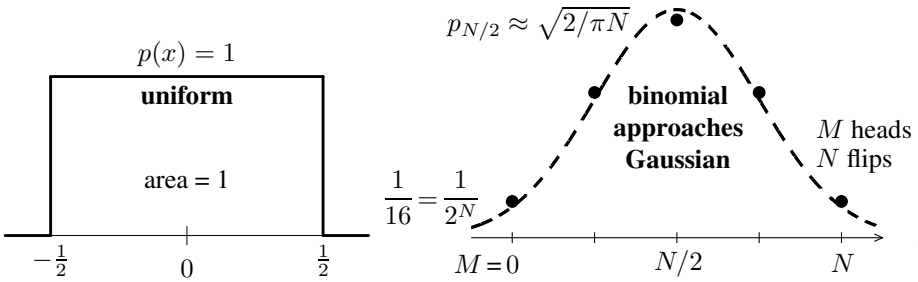

Figure 12.3: The probabilities $p = (1, 4, 6, 4, 1)/16$ for the number of heads in 4 flips. These p_i approach a Gaussian distribution with variance $\sigma^2 = N/4$ centered at $m = N/2$. For X, the Central Limit Theorem gives convergence to the normal distribution $\mathbf{N}(0, 1)$.

It is fun to see the Central Limit Theorem giving the right answer at the center point $X = 0$. At that point, the factor $e^{-X^2/2}$ equals 1. We know that the variance for N coin flips is $\sigma^2 = N/4$. The center of the bell-shaped curve has height $1/\sqrt{2\pi\sigma^2} = \boldsymbol{\sqrt{2/N\pi}}$.

What is the height at the center of the coin-flip distribution p_0 to p_N (the binomial distribution)? For $N = 4$, the probabilities for $0, 1, 2, 3, 4$ heads come from $\left(\frac{1}{2} + \frac{1}{2}\right)^4$.

Center probability $\boldsymbol{\frac{6}{16}}$ $\qquad \left(\dfrac{1}{2} + \dfrac{1}{2}\right)^4 = \dfrac{1}{16} + \dfrac{4}{16} + \boldsymbol{\dfrac{6}{16}} + \dfrac{4}{16} + \dfrac{1}{16} = 1.$

The binomial theorem in Problem 8 tells us the center probability $p_{N/2}$ for any even N:

The center probability $\left(\frac{N}{2} \text{ heads}, \frac{N}{2} \text{ tails}\right)$ is $\dfrac{1}{2^N}\dfrac{N!}{(N/2)!\,(N/2)!}$

For $N = 4$, those factorials produce $4!/2!\,2! = 24/4 = 6$. For large N, Stirling's formula $\sqrt{2\pi N}(N/e)^N$ is a close approximation to $N!$. Use Stirling for N and twice for $N/2$:

Limit of coin-flip Center probability
$$p_{N/2} \approx \frac{1}{2^N}\,\frac{\sqrt{2\pi N}(N/e)^N}{\pi N(N/2e)^N} = \frac{\sqrt{2}}{\sqrt{\pi N}} = \frac{1}{\sqrt{2\pi}\sigma}. \tag{14}$$

At that last step we used the variance $\sigma^2 = N/4$ for the coin-tossing problem. The result $1/\sqrt{2\pi}\sigma$ matches the center value (above) for the Gaussian. The Central Limit Theorem is true: The "binomial distribution" approaches the normal distribution as $N \to \infty$.

Monte Carlo Estimation Methods

Scientific computing has to work with errors in the data. Financial computing has to work with unsure numbers and uncertain predictions. All of applied mathematics has moved to **accepting uncertainty in the inputs and estimating the variance in the outputs**.

How to estimate that variance? Often probability distributions $p(x)$ are not known. What we can do is to try different inputs $\boldsymbol{b}$ and compute the outputs $\boldsymbol{x}$ and take an average. This is the simplest form of a **Monte Carlo method** (named after the gambling palace on the Riviera, where I once saw a fight about whether the bet was placed in time). Monte Carlo approximates an expected value $\mathrm{E}[x]$ by a sample average $(x_1 + \cdots + x_N)/N$.

Please understand that every x_k can be expensive to compute. We are not just flipping coins. Each sample comes from a set of data b_k. *Monte Carlo randomly chooses this data b_k, it computes the outputs x_k, and then it averages those x's.* Decent accuracy for $\mathrm{E}[x]$ often requires many samples b and huge computing cost. The error in approximating $\mathrm{E}[x]$ by $(x_1 + \cdots + x_N)/N$ is normally of order $1/\sqrt{N}$. *Slow improvement as N increases.*

That $1/\sqrt{N}$ estimate came for coin flips in equation (11). Averaging N independent samples x_k of variance σ^2 reduces the variance to σ^2/N.

"Quasi-Monte Carlo" can sometimes reduce this variance to σ^2/N^2: a big difference! The inputs b_k are selected very carefully—not just randomly. This QMC approach is surveyed in the journal *Acta Numerica* 2013. The newer idea of "Multilevel Monte Carlo" is outlined by Michael Giles in *Acta Numerica* 2015. Here is how it works.

Suppose it is much simpler to simulate another variable $y(b)$ close to $x(b)$. Then use N computations of $y(b_k)$ and only $N^* < N$ computations of $x(b_k)$ to estimate $\mathrm{E}[x]$.

2-level Monte Carlo
$$\mathbf{E}[\boldsymbol{x}] \approx \frac{\boldsymbol{1}}{\boldsymbol{N}}\sum_{\boldsymbol{1}}^{\boldsymbol{N}} \boldsymbol{y}(\boldsymbol{b_k}) + \frac{\boldsymbol{1}}{\boldsymbol{N^*}}\sum_{\boldsymbol{1}}^{\boldsymbol{N^*}} [\boldsymbol{x}(\boldsymbol{b_k}) - \boldsymbol{y}(\boldsymbol{b_k})]\,.$$

The idea is that $x - y$ has a smaller variance σ^* than the original x. Therefore N^* can be smaller than N, with the same accuracy for $\mathrm{E}[x]$. We do N cheap simulations to find the y's. Those cost C each. We only do N^* expensive simulations involving x's. Those cost C^* each. The total computing cost is $NC + N^*C^*$.

Calculus minimizes the overall variance for a fixed total cost. The optimal ratio N^*/N is $\sqrt{C/C^*}\ \sigma^*/\sigma$. Three-level Monte Carlo would simulate x, y, and z:

$$\mathrm{E}[x] \approx \frac{1}{N}\sum_1^N \boldsymbol{z(b_k)} + \frac{1}{N^*}\sum_1^{N^*} [\boldsymbol{y(b_k)} - \boldsymbol{z(b_k)}] + \frac{1}{N^{**}}\sum_1^{N^{**}} [\boldsymbol{x(b_k)} - \boldsymbol{y(b_k)}]\,.$$

Giles optimizes $N, N^*, N^{**}, \ldots$ to keep $\mathrm{E}[x] \le$ fixed E_0, and provides a MATLAB code.

Review: Three Formulas for the Mean and the Variance

The formulas for m and σ^2 are the starting point for all of probability and statistics. There are three different cases to keep straight: **sample** values X_i, **expected** values (discrete p_i), and a range of **expected** values (continuous $p(x)$). Here are the mean and the variance:

	$\boldsymbol{m}$	$\boldsymbol{S^2}$ or $\boldsymbol{\sigma^2}$
Samples X_1 to X_N	$\boldsymbol{m} = \dfrac{X_1 + \cdots + X_N}{N}$	$\boldsymbol{S^2} = \dfrac{(X_1 - m)^2 + \cdots + (X_N - m)^2}{N - 1}$
n possible outputs with probabilities p_i	$\boldsymbol{m} = \sum_1^n p_i x_i$	$\boldsymbol{\sigma^2} = \sum_1^n p_i (x_i - m)^2$
Range of outputs with probability density	$\boldsymbol{m} = \int x\, p(x) dx$	$\boldsymbol{\sigma^2} = \int (x - m)^2 p(x) dx$

A natural question: Why are there no probabilities p on the first line? How can these formulas be parallel? Answer: *We expect a fraction p_i of the samples to be $X = x_i$.* If this is exactly true, $X = x_i$ is repeated $p_i N$ times. Then lines 1 and 2 give the same m.

When we work with samples, we don't know the p_i. We just include each output X as often as it comes. We get the "empirical" mean instead of the expected mean.

Problem Set 12.1

1 Add 7 to every output x. What happens to the mean and the variance? What are the new sample mean, the new expected mean, and the new variance?

2 We know: $\frac{1}{3}$ of all integers are divisible by 3 and $\frac{1}{7}$ of integers are divisible by 7. What fraction of integers will be divisible by 3 or 7 or both?

3 Suppose you sample from the numbers 1 to 1000 with equal probabilities $1/1000$. What are the probabilities p_0 to p_9 that the last digit of your sample is $0, \ldots, 9$? What is the expected mean m of that last digit? What is its variance σ^2?

4 Sample again from 1 to 1000 but look at the last digit of the sample *squared*. That square could end with $x = 0, 1, 4, 5, 6,$ or 9. What are the probabilities $p_0, p_1, p_4, p_5, p_6, p_9$? What are the (expected) mean m and variance σ^2 of that number x?

5 (a little tricky) Sample again from 1 to 1000 with equal probabilities and let x be the *first* digit ($x = 1$ if the number is 15). What are the probabilities p_1 to p_9 (adding to 1) of $x = 1, \dots, 9$? What are the mean and variance of x?

6 Suppose you have $N = 4$ samples $157, 312, 696, 602$ in Problem 5. What are the first digits x_1 to x_4 of the squares? What is the sample mean μ ? What is the sample variance S^2 ? Remember to divide by $N - 1 = 3$ and not $N = 4$.

7 Equation (4) gave a second equivalent form for S^2 (the variance using samples) :

$$\boldsymbol{S^2} = \frac{1}{N-1} \text{ sum of } (x_i - m)^2 = \frac{1}{N-1}\left[\left(\text{sum of } x_i^2\right) - Nm^2\right].$$

Verify the matching identity for the expected variance σ^2 (using $m = \Sigma\, p_i\, x_i$) :

$$\boldsymbol{\sigma^2 = \text{ sum of } p_i\,(x_i - m)^2 = \left(\text{sum of } p_i\, x_i^2\right) - m^2}.$$

8 If all 24 samples from a population produce the same age $x = 20$, what are the sample mean μ and the sample variance S^2 ? What if $x = 20$ or 21, 12 times each ?

9 Computer experiment as on page 541 : Find the average $A_{1000000}$ of a million random 0-1 samples ! What is $X = \left(A_N - \frac{1}{2}\right)/2\sqrt{N}$?

10 The probability p_i to get i heads in N coin flips is the *binomial number* $b_i = \binom{N}{i}$ divided by 2^N. The b_i add to $(1+1)^N = 2^N$ so the probabilities p_i add to 1.

$$p_0 + \cdots + p_N = \left(\frac{1}{2} + \frac{1}{2}\right)^N = \frac{1}{2^N}(b_0 + \cdots + b_N) \text{ with } b_i = \frac{N!}{i!\,(N-i)!}$$

$N = 4$ leads to $b_0 = \dfrac{24}{24}$, $b_1 = \dfrac{24}{(1)(6)} = 4$, $b_2 = \dfrac{24}{(2)(2)} = 6$, $p_i = \dfrac{1}{16}(1, 4, 6, 4, 1)$.

Notice $b_i = b_{N-i}$. *Problem*: Confirm that the mean $m = 0p_0 + \cdots + Np_N$ equals $\frac{N}{2}$.

11 For any function $f(x)$ the expected value is $\mathrm{E}[f] = \sum p_i\, f(x_i)$ or $\int p(x)\, f(x)\, dx$ (discrete probability or continuous probability). Suppose the mean is $\mathrm{E}[x] = m$ and the variance is $\mathrm{E}[(x-m)^2] = \sigma^2$. **What is $\mathbf{E}[\boldsymbol{x^2}]$** ?

12 Show that the standard normal distribution $p(x)$ has total probability $\int p(x)\, dx = 1$ as required. A famous trick multiplies $\int p(x)\, dx$ by $\int p(y)\, dy$ and computes the integral over all x and all y ($-\infty$ to ∞). The trick is to replace $dx\, dy$ in that double integral by $r\, dr\, d\theta$ (polar coordinates with $x^2 + y^2 = r^2$). Explain each step :

$$2\pi \int_{-\infty}^{\infty} p(x)\, dx \int_{-\infty}^{\infty} p(y)\, dy = \iint_{-\infty}^{\infty} e^{-(x^2+y^2)/2} dx\, dy = \int_{\theta\,=\,0}^{2\pi} \int_{r\,=\,0}^{\infty} e^{-r^2/2}\, r\, dr\, d\theta = \boldsymbol{2\pi}.$$