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** $\boldsymbol{\sigma^2}$ 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 $\mathbf{18, 17, 18, 19, 17}$

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

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

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

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:

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

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

$$\textbf{Expected value} \quad m = \mathrm{E}[x] = p_1x_1 + p_2x_2 + \cdots + p_nx_n. \tag{2}$$

This is $\boldsymbol{p} \cdot \boldsymbol{x}$. 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 σ^2** measures expected distance (squared) from the expected mean $\mathrm{E}[x]$. The **sample variance S^2** measures actual distance (squared) from the sample mean. The square root is the **standard deviation σ or 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.

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

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.

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

We are squaring the distance from the expected value $m = E[x]$. 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 $E[x]$, 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".

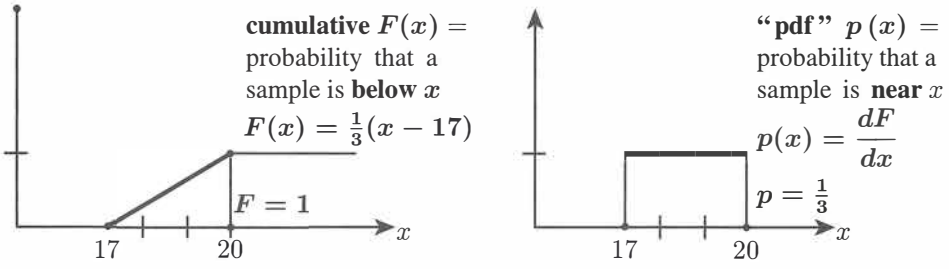

Figure 12.1: $F(x)$ is the cumulative distribution and its derivative $p(x) = dF/dx$ is the **probability density function (pdf)**. For this uniform distribution, $p(x)$ is constant between 17 and 20. The total area under the graph of $p(x)$ is the total probability $F = 1$.

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:

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

$F(b)$ 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 (mean $\frac{1}{2}$) 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.

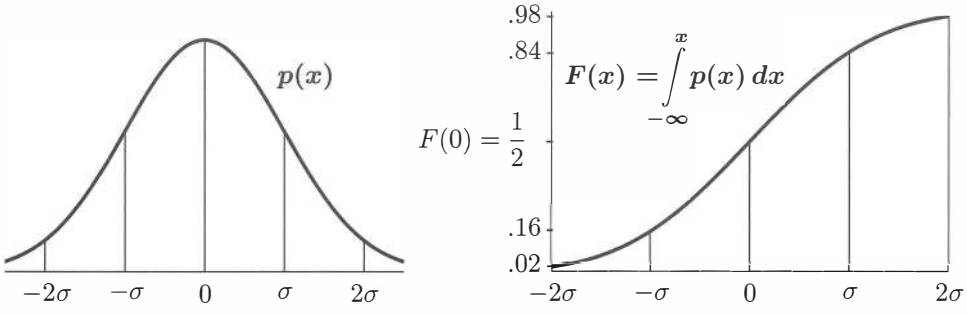

Figure 12.2: The standard normal distribution $p\,(x)$ has mean $m = 0$ and $\sigma = 1$.

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

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

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

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

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 σ_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 $\boldsymbol{\sigma^2 = \frac{1}{4}}$:

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

The average A_N now has mean $\dfrac{1}{2}$ and variance $\dfrac{1}{4N^2} + \cdots + \dfrac{1}{4N^2} = \boldsymbol{\dfrac{1}{4N}} = \boldsymbol{\sigma_N^2}$. (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 $0 - 1$ variable x_{new} is $\frac{1}{2}\,x_{\text{old}} + \frac{1}{2}$. 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 X $= (x - m)/\sigma = \left(A_N - \frac{1}{2}\right) / 2\sqrt{N}$ was [**$-.40$**] [**.70**] [**$-.33$**].

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 $\frac{1}{8}$ and $\frac{3}{8}$ are pieces of $\left(\frac{1}{2} + \frac{1}{2}\right)^3 = \frac{1}{8} + \frac{3}{8} + \frac{3}{8} + \frac{1}{8} = 1$. *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 $m = 1.5$:

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

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.

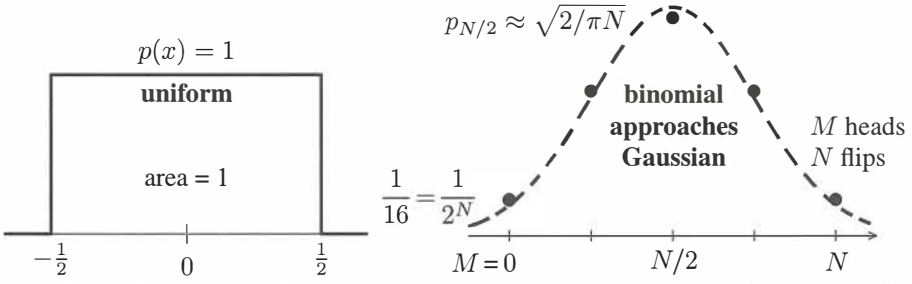

Figure 12.3: The probabilities $p = (1, 4, 6, 4, 1)/16$ for the number of heads in 4 flips. These p_i approach a Gaussian distribution with variance $\sigma^2 = N/4$ centered at $m = N/2$. For X, the Central Limit Theorem gives convergence to the normal distribution $\mathbf{N}(0, 1)$.

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

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

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

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:

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

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

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 $\mathrm{E}[\boldsymbol{x^2}]$** ?

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 :

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

12.2 Covariance Matrices and Joint Probabilities

Linear algebra enters when we run M different experiments at once. We might measure age and height and weight ($M = 3$ measurements of N people). Each experiment has its own mean value. So we have a vector $\boldsymbol{m} = (m_1, m_2, m_3)$ containing the M mean values. Those could be *sample means* of age and height and weight. Or m_1, m_2, m_3 could be *expected values* of age, height, weight based on known probabilities.

A matrix becomes involved when we look at variances. Each experiment will have a sample variance S_i^2 or an expected $\sigma_i^2 = \mathrm{E}\left[(x_i - m_i)^2\right]$ based on the squared distance from its mean. Those M numbers $\sigma_1^2, \ldots, \sigma_M^2$ will go on the main diagonal of the matrix. So far we have made no connection between the M parallel experiments. They measure M different random variables, but the experiments are not necessarily independent!

If we measure age and height and weight $(\boldsymbol{a}, \boldsymbol{h}, \boldsymbol{w})$ for children, the results will be strongly correlated. Older children are generally taller and heavier. Suppose the means m_a, m_h, m_w are known. Then $\sigma_a^2, \sigma_h^2, \sigma_w^2$ are the separate variances in age, height, weight. **The new numbers are the covariances like σ_{ah}, where age multiplies height.**

$$\textbf{Covariance} \quad \sigma_{ah} = \mathrm{E}\left[(\textbf{age} - \textbf{mean age})\ (\textbf{height} - \textbf{mean height})\right]. \tag{1}$$

This definition needs a close look. To compute σ_{ah}, it is not enough to know the probability of each age and the probability of each height. We have to know the **joint probability of each pair** (**age and height**). This is because age is connected to height.

$\boldsymbol{p_{ah}}$ = probability that a random child has age $= \boldsymbol{a}$ **and** height $= \boldsymbol{h}$: both at once

$\boldsymbol{p_{ij}}$ = **probability that experiment 1 produces x_i and experiment 2 produces y_j**

Suppose experiment 1 (age) has mean m_1. Experiment 2 (height) has mean m_2. The covariance in (1) between experiments 1 and 2 looks at **all pairs** of ages x_i, heights y_j :

$$\textbf{Covariance} \quad \sigma_{12} = \sum_{\textbf{all } i,j} \sum p_{ij}(x_i - m_1)(y_j - m_2) \tag{2}$$

To capture this idea of "joint probability p_{ij}" we begin with two small examples.

Example 1 Flip two coins separately. With 1 for heads and 0 for tails, the results can be $(1,1)$ or $(1,0)$ or $(0,1)$ or $(0,0)$. Those four outcomes all have probability $p_{11} = p_{10} = p_{01} = p_{00} = \frac{1}{4}$. **Independent experiments have Prob of (i, j) = (Prob of i) (Prob of j).**

Example 2 *Glue the coins together*, facing the same way. The only possibilities are $(1,1)$ and $(0,0)$. Those have probabilities $\frac{1}{2}$ and $\frac{1}{2}$. The probabilities p_{10} and p_{01} are zero. $(1,0)$ and $(0,1)$ won't happen because the coins stick together: both heads or both tails.

Probability matrices for Examples 1 and 2

$$P = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} \end{bmatrix} \qquad P = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}.$$

Let me stay longer with P, to show it in good matrix notation. The matrix shows the probability p_{ij} of each pair (x_i, y_j)—starting with (x_1, y_1) = (heads, heads) and (x_1, y_2) = (heads, tails). Notice the row sums p_i and column sums P_j and the total sum = 1.

$$\textbf{Probability matrix} \quad P = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} \quad \begin{matrix} p_{11}+p_{12} = \boldsymbol{p_1} \\ p_{21}+p_{22} = \boldsymbol{p_2} \end{matrix} \quad \begin{pmatrix} \text{first} \\ \text{coin} \end{pmatrix}$$

$$\text{(second coin) column sums} \quad \boldsymbol{P_1} \quad \boldsymbol{P_2} \qquad 4 \text{ entries add to } 1$$

Those numbers p_1, p_2 and P_1, P_2 are called the **marginals** of the matrix P:

$p_1 = p_{11} + p_{12}$ = chance of heads from **coin 1** (coin 2 can be heads or tails)
$P_1 = p_{11} + p_{21}$ = chance of heads from **coin 2** (coin 1 can be heads or tails)

Example 1 showed *independent* variables. Every probability p_{ij} equals p_i times p_j ($\frac{1}{2}$ times $\frac{1}{2}$ gave $p_{ij} = \frac{1}{4}$ in that example). In this case **the covariance σ_{12} will be zero**. Heads or tails from the first coin gave no information about the second coin.

Zero covariance σ_{12} for independent trials $\quad V = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$ = **diagonal covariance matrix.**

Independent experiments have $\sigma_{12} = 0$ because every $p_{ij} = (p_i)(p_j)$ in equation (2):

$$\sigma_{12} = \sum_i \sum_j (p_i)(p_j)(x_i - m_1)(y_j - m_2) = \left[\sum_i (p_i)(x_i - m_1)\right]\left[\sum_j (p_j)(y_j - m_2)\right] = [0][0].$$

The glued coins show perfect correlation. Heads on one means heads on the other. The covariance σ_{12} moves from 0 to $\sigma_1\sigma_2 = \frac{1}{4}$—this is the largest possible value of σ_{12}:

$$\textbf{Means} = \frac{1}{2} \qquad \boldsymbol{\sigma_{12}} = \frac{1}{2}\left(1-\frac{1}{2}\right)\left(1-\frac{1}{2}\right) + 0 + 0 + \frac{1}{2}\left(0-\frac{1}{2}\right)\left(0-\frac{1}{2}\right) = \frac{1}{4}$$

Heads or tails from coin 1 gives complete information about heads or tails from coin 2:

Glued coins give largest possible covariances
Singular covariance matrix: determinant = 0 $\quad V_{\text{glue}} = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2 \\ \sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$

Always $\boldsymbol{\sigma_1^2\sigma_2^2 \geq \sigma_{12}^2}$. Thus σ_{12} is *between* $-\sigma_1\sigma_2$ *and* $\sigma_1\sigma_2$. The covariance matrix V is **positive definite** (or in this singular case of glued coins, V is **positive semidefinite**). That is an important fact about M by M covariance matrices for M experiments.

Note that the **sample covariance matrix S** from N trials is certainly semidefinite. Every new sample X = (age, height, weight) contributes to the **sample mean $\overline{X}$** and to S. Each term $(X_i - \overline{X})(X_i - \overline{X})^{\mathrm{T}}$ is positive semidefinite and we just add to reach S:

$$\overline{X} = \frac{X_1 + \cdots + X_N}{N} \qquad S = \frac{(X_1 - \overline{X})(X_1 - \overline{X})^{\mathrm{T}} + \cdots + (X_N - \overline{X})(X_N - \overline{X})^{\mathrm{T}}}{N-1} \tag{3}$$

The Covariance Matrix V is Positive Semidefinite

Come back to the *expected* covariance σ_{12} between two experiments 1 and 2 (two coins):

$$\boxed{\begin{aligned}\sigma_{12} &= \text{expected value of } [(\textit{output}\,1 - \textit{mean}\,1) \text{ times } (\textit{output}\,2 - \textit{mean}\,2)] \\ &= \sum_{\text{all } i,j} \sum p_{ij}\,(x_i - m_1)\,(y_j - m_2).\end{aligned}} \tag{4}$$

$p_{ij} \geq 0$ is the probability of seeing output x_i in experiment 1 **and** y_j in experiment 2. Some pair of outputs must appear. Therefore the N^2 probabilities p_{ij} add to 1.

Total probability (all pairs) is 1 $$\sum_{\text{all } i,j} \sum p_{ij} = 1. \tag{5}$$

Here is another fact we need. *Fix on one particular output* x_i in experiment 1. Allow *all outputs* y_j in experiment 2. Add the probabilities of $(x_i, y_1), (x_i, y_2), \ldots, (x_i, y_n)$:

$$\textbf{Row sum } p_i \textbf{ of } P \quad \sum_{j=1}^{n} p_{ij} = \textbf{probability } p_i \textbf{ of } x_i \textbf{ in experiment 1.} \tag{6}$$

Some y_j must happen in experiment 2! Whether the two coins are completely separate or glued together, we still get $\frac{1}{2}$ for the probability $p_H = p_{HH} + p_{HT}$ that coin 1 is heads:

$$\text{(separate) } P_{HH} + P_{HT} = \frac{1}{4} + \frac{1}{4} = \mathbf{\frac{1}{2}} \qquad \text{(glued) } P_{HH} + P_{HT} = \frac{1}{2} + 0 = \mathbf{\frac{1}{2}}.$$

That basic reasoning allows us to write one matrix formula that includes the covariance σ_{12} along with the separate variances σ_1^2 and σ_2^2 for experiment 1 and experiment 2. We get the whole covariance matrix V by adding the matrices V_{ij} for each pair (i, j):

$$\boxed{\begin{aligned}&\textbf{Covariance matrix} \\ &\boldsymbol{V = \Sigma\Sigma\, V_{ij}}\end{aligned} \qquad V = \sum_{\text{all } i,j} \sum p_{ij} \begin{bmatrix} (x_i - m_1)^2 & (x_i - m_1)(y_j - m_2) \\ (x_i - m_1)(y_j - m_2) & (y_j - m_2)^2 \end{bmatrix}} \tag{7}$$

Off the diagonal, this is equation (2) for the covariance σ_{12}. On the diagonal, we are getting the ordinary variances σ_1^2 and σ_2^2. I will show in detail how we get $V_{11} = \sigma_1^2$ by using equation (6). Allowing all j just leaves the probability p_i of x_i in experiment 1:

$$\boldsymbol{V_{11}} = \sum_{\text{all } i,j} \sum p_{ij}(x_i - m_1)^2 = \sum_{\text{all } i} (\text{probability of } x_i)\,(x_i - m_1)^2 = \boldsymbol{\sigma_1^2}. \tag{8}$$

Please look at that twice. It is the key to producing the whole covariance matrix by one formula (7). The beauty of that formula is that it combines 2 by 2 matrices V_{ij}. And the matrix V_{ij} in (7) for each pair of outcomes i, j is **positive semidefinite**:

V_{ij} has diagonal entries $p_{ij}(x_i - m_1)^2 \geq 0$ and $p_{ij}(y_j - m_2)^2 \geq 0$ and $\det(V_{ij}) = 0$.

That matrix V_{ij} has rank 1. Equation (7) multiplies p_{ij} *times column* $\boldsymbol{U}$ *times row* $\boldsymbol{U}^{\mathrm{T}}$:

$$\begin{bmatrix} (x_i - m_1)^2 & (x_i - m_1)(y_j - m_2) \\ (x_i - m_1)(y_j - m_2) & (y_j - m_2)^2 \end{bmatrix} = \begin{bmatrix} x_i - m_1 \\ y_j - m_2 \end{bmatrix} \begin{bmatrix} x_i - m_1 & y_j - m_2 \end{bmatrix} \tag{9}$$

Every matrix $\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}$ *is positive semidefinite.* So the whole matrix V (combining these matrices $\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}$ with weights $p_{ij} \geq 0$) is **at least semidefinite**—and probably V is definite.

The covariance matrix V is positive definite unless the experiments are dependent.

Now we move from two variables x and y to M variables like age-height-weight. The output from each trial is a vector X with M components. (Each child has an age-height-weight vector with 3 components.) The covariance matrix V is now M by M. V is created from the output vectors $\boldsymbol{X}$ and their average $\overline{\boldsymbol{X}} = \mathbf{E}[\boldsymbol{X}]$:

$$\textbf{Covariance matrix} \qquad \boldsymbol{V} = \mathbf{E}\left[(\boldsymbol{X} - \overline{\boldsymbol{X}})(\boldsymbol{X} - \overline{\boldsymbol{X}})^{\mathrm{T}}\right] \tag{10}$$

Remember that $\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ and $\overline{\boldsymbol{X}}\,\overline{\boldsymbol{X}}^{\mathrm{T}}$ = (column)(row) are M by M matrices.

For $M = 1$ (one variable) you see that $\overline{X}$ is the mean m and V is σ^2 (Section 12.1). For $M = 2$ (two coins) you see that $\overline{X}$ is (m_1, m_2) and V matches equation (10). The expectation E always adds up outputs times their probabilities. For age-height-weight the output could be X = (5 years, 31 inches, 48 pounds) and its probability is $\boldsymbol{p_{5,31,48}}$.

Now comes a new idea. *Take any linear combination* $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{X} = c_1X_1 + \cdots + c_MX_M$. With $c = (6, 2, 5)$ this would be $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{X} = 6\,(\text{age}) + 2\,(\text{height}) + 5\,(\text{weight})$. By linearity we know that its expected value $\mathrm{E}[\boldsymbol{c}^{\mathrm{T}}\boldsymbol{X}]$ is $\boldsymbol{c}^{\mathrm{T}}\mathrm{E}[\boldsymbol{X}] = \boldsymbol{c}^{\mathrm{T}}\overline{\boldsymbol{X}}$:

$$\mathbf{E}[\boldsymbol{c}^{\mathrm{T}}\boldsymbol{X}] = \boldsymbol{c}^{\mathrm{T}}\mathbf{E}[\boldsymbol{X}] = 6\,(\text{expected age}) + 2\,(\text{expected height}) + 5\,(\text{expected weight}).$$

More than that, we also know the *variance* σ^2 of that number $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{X}$:

$$\begin{aligned} \text{Variance of } \boldsymbol{c}^{\mathrm{T}}\boldsymbol{X} &= \mathrm{E}\left[(\boldsymbol{c}^{\mathrm{T}}\boldsymbol{X} - \boldsymbol{c}^{\mathrm{T}}\overline{\boldsymbol{X}})(\boldsymbol{c}^{\mathrm{T}}\boldsymbol{X} - \boldsymbol{c}^{\mathrm{T}}\overline{\boldsymbol{X}})^{\mathrm{T}}\right] \\ &= \boldsymbol{c}^{\mathrm{T}}\mathrm{E}\left[(\boldsymbol{X} - \overline{\boldsymbol{X}})(\boldsymbol{X} - \overline{\boldsymbol{X}})^{\mathrm{T}}\right]\boldsymbol{c} = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{c}! \end{aligned} \tag{11}$$

Now the key point: *The variance of* $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{X}$ *can never be negative.* So $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{c} \geq \boldsymbol{0}$. *The covariance matrix* V *is therefore positive semidefinite by the energy test* $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{c} \geq \boldsymbol{0}$.

Covariance matrices V open up the link between probability and linear algebra: V equals $Q\Lambda Q^{\mathrm{T}}$ with eigenvalues $\lambda_i \geq 0$ and orthonormal eigenvectors $\boldsymbol{q}_1$ to $\boldsymbol{q}_M$.

Diagonalizing the covariance matrix means finding M *independent* experiments as combinations of the original M experiments.

Confession I am not entirely happy with that proof based on $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{c} \geq \mathbf{0}$. The expectation symbol **E** is burying the key idea of **joint probability**. Allow me to show directly that V is positive semidefinite (at least for the age-height-weight example). The proof is simply that **V is the sum of the joint probability p_{ahw} of each combination (age, height, weight) times the positive semidefinite matrix $\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}$.** Here U is $X - \overline{X}$:

$$\boldsymbol{V} = \sum_{\text{all } a,h,w} p_{ahw}\, \boldsymbol{U}\,\boldsymbol{U}^{\mathrm{T}} \quad \text{with} \quad \boldsymbol{U} = \begin{bmatrix} \text{age} \\ \text{height} \\ \text{weight} \end{bmatrix} - \begin{bmatrix} \text{mean age} \\ \text{mean height} \\ \text{mean weight} \end{bmatrix}. \tag{12}$$

This is exactly like the 2 by 2 coin flip matrix V in equation (7). Now $M = 3$.

The value of the expectation symbol E is that it also allows *pdf*'s (probability density functions like $p(x, y, z)$ for continuous random variables x and y and z). If we allow all numbers as ages and heights and weights, instead of age $i = 0, 1, 2, 3 \ldots$, then we need $p(x, y, z)$ instead of p_{ijk}. The sums in this section of the book would all change to integrals. But we still have $V = \mathrm{E}\,[\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}]$:

$$\textbf{Covariance matrix} \quad V = \iiint p(x, y, z)\, \boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\, dx\, dy\, dz \quad \text{with} \quad \boldsymbol{U} = \begin{bmatrix} x - \overline{x} \\ y - \overline{y} \\ z - \overline{z} \end{bmatrix}. \tag{13}$$

Always $\iiint p = 1$. Examples 1–2 emphasized how p can give diagonal V or singular V:

Independent variables x, y, z $\quad p(x, y, z) = \boldsymbol{p_1(x)\, p_2(y)\, p_3(z)}$.

Dependent variables x, y, z $\quad p(x, y, z) = \mathbf{0}$ except when $cx + dy + ez = 0$.

The Mean and Variance of $z = x + y$

Start with the sample mean. We have N samples of x. Their mean (= average) is m_x. We also have N samples of y and their mean is m_y. **The sample mean of $z = x + y$ is clearly $m_z = m_x + m_y$:**

$$\textbf{Mean of sum = Sum of means} \quad \frac{1}{N}\sum_{1}^{N}(x_i + y_i) = \frac{1}{N}\sum_{1}^{N} x_i + \frac{1}{N}\sum_{1}^{N} y_i. \tag{14}$$

Nice to see something that simple. The *expected* mean of $z = x + y$ doesn't look so simple, but it must come out as $\mathbf{E}[z] = \mathbf{E}[x] + \mathbf{E}[y]$. Here is one way to see this.

The joint probability of the pair (x_i, y_j) is p_{ij}. Its value depends on whether the experiments are independent, which we don't know. But for the mean of the sum $z = x + y$,

dependence or independence of x and y doesn't matter. *Expected values still add*:

$$\mathbf{E}[\boldsymbol{x}+\boldsymbol{y}] = \sum_i \sum_j p_{ij}(x_i + y_j) = \sum_i \sum_j p_{ij}x_i + \sum_i \sum_j p_{ij}y_j. \tag{15}$$

All the sums go from 1 to N. We can add in any order. For the first term on the right side, add the p_{ij} along row i of the probability matrix P to get p_i. That double sum gives $\mathrm{E}[x]$:

$$\sum_i \sum_j p_{ij}x_i = \sum_i (p_{i1} + \cdots + p_{iN})x_i = \sum_i p_i x_i = \mathrm{E}[x].$$

For the last term, add p_{ij} down column j of the matrix to get the probability P_j of y_j. Those pairs (x_1, y_j) and (x_2, y_j) and ... and (x_N, y_j) are all the ways to produce y_j:

$$\sum_i \sum_j p_{ij}y_j = \sum_j (p_{1j} + \cdots + p_{Nj})y_j = \sum_j P_j y_j = \mathrm{E}[y].$$

Now equation (15) says that $\mathbf{E}[\boldsymbol{x}+\boldsymbol{y}] = \mathbf{E}[\boldsymbol{x}] + \mathbf{E}[\boldsymbol{y}]$.

What about the variance of $z = x + y$? The joint probabilities p_{ij} and the covariance σ_{xy} will be involved. Let me separate the variance of $x + y$ into three simple pieces:

$$\begin{aligned}\boldsymbol{\sigma_z^2} &= \sum\sum p_{ij}(x_i + y_j - m_x - m_y)^2 \\ &= \sum\sum p_{ij}(x_i - m_x)^2 + \sum\sum p_{ij}(y_j - m_y)^2 + 2\sum\sum p_{ij}(x_i - m_x)(y_j - m_y)\end{aligned}$$

The first piece is $\boldsymbol{\sigma_x^2}$. The second piece is $\boldsymbol{\sigma_y^2}$. The last piece is $\boldsymbol{2\sigma_{xy}}$.

$$\textbf{The variance of } \boldsymbol{z = x + y} \quad \textbf{is} \quad \boldsymbol{\sigma_z^2 = \sigma_x^2 + \sigma_y^2 + 2\sigma_{xy}}. \tag{16}$$

The Covariance Matrix for $Z = AX$

Here is a good way to see σ_z^2 when $z = x + y$. Think of (x, y) as a column vector $\boldsymbol{X}$. Think of the 1 by 2 matrix $A = \begin{bmatrix} 1 & 1 \end{bmatrix}$ multiplying that vector $\boldsymbol{X}$. Then $A\boldsymbol{X}$ is the sum $z = x + y$. The variance $\boldsymbol{\sigma_z^2}$ in equation (16) goes into matrix notation as

$$\boldsymbol{\sigma_z^2} = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \text{which is} \quad \boldsymbol{\sigma_z^2 = AVA^{\mathrm{T}}}. \tag{17}$$

You can see that $\sigma_z^2 = AVA^{\mathrm{T}}$ in (17) agrees with $\sigma_x^2 + \sigma_y^2 + 2\sigma_{xy}$ in (16).

Now for the main point. The vector $\boldsymbol{X}$ could have M components coming from M experiments (instead of only 2). Those experiments will have an M by M covariance matrix $\boldsymbol{V_X}$. The matrix A could be K by M. Then $A\boldsymbol{X}$ is a vector with K combinations of the M outputs (instead of 1 combination $x + y$ of 2 outputs).

That vector $\boldsymbol{Z} = A\boldsymbol{X}$ of length K has a K by K covariance matrix $\boldsymbol{V_Z}$. Then the great rule for covariance matrices—of which equation (17) was only a 1 by 2 example—is this beautiful formula: Covariance matrix of $A\boldsymbol{X}$ is A (covariance matrix of $\boldsymbol{X}$) A^{T}:

$$\boxed{\textbf{The covariance matrix of } \boldsymbol{Z = AX} \textbf{ is } \boldsymbol{V_Z = AV_X A^{\mathrm{T}}}} \tag{18}$$

To me, this neat formula shows the beauty of matrix multiplication. I won't prove this formula, just admire it. It is constantly used in applications—coming in Section 12.3.

The Correlation ρ

Correlation ρ_{xy} is closely related to covariance σ_{xy}. They both measure dependence or independence. Start by rescaling or "standardizing" the random variables x and y **The new $X = x/\sigma_x$ and $Y = y/\sigma_y$ have variance $\sigma_X^2 = \sigma_Y^2 = 1$.** This is just like dividing a vector v by its length to produce a unit vector $v/||v||$ of length 1.

The correlation of x and y is the covariance of X and Y. If the original covariance of x and y was σ_{xy}, then rescaling to X and Y will divide by σ_x and σ_y :

$$\textbf{Correlation } \rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \textbf{covariance of } \frac{x}{\sigma_x} \textbf{ and } \frac{y}{\sigma_y} \qquad \textbf{Always } -1 \le \rho_{xy} \le 1$$

Zero covariance gives zero correlation. *Independent random variables* produce $\rho_{xy} = 0$.

We know that always $\sigma_{xy}^2 \le \sigma_x^2 \sigma_y^2$ (the covariance matrix V is at least positive semidefinite). Then $\rho_{xy}^2 \le 1$. Correlation near $\rho = +1$ means strong dependence in the same direction: often voting the same. Negative correlation means that y tends to be below its mean when x is above its mean: Voting in opposite directions.

Example 3 *Suppose that y is just $-x$.* A coin flip has outputs $x = 0$ or 1. The same flip has outputs $y = 0$ or -1. The mean m_x is $\frac{1}{2}$ for a fair coin, and m_y is $-\frac{1}{2}$. The covariance is $\sigma_{xy} = -\sigma_x \sigma_y$. The correlation divides by $\sigma_x \sigma_y$ to get $\rho_{xy} = -1$. In this case the correlation matrix R has determinant zero (singular and only semidefinite):

$$\textbf{Correlation matrix } R = \begin{bmatrix} 1 & \rho_{xy} \\ \rho_{xy} & 1 \end{bmatrix} \qquad R = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \text{ when } y = -x$$

R always has 1's on the diagonal because we normalized to $\sigma_X = \sigma_Y = 1$. R is the correlation matrix for x and y, and the covariance matrix for $X = x/\sigma_x$ and $Y = y/\sigma_y$.

That number ρ_{xy} is also called the Pearson coefficient.

Example 4 Suppose the random variables x, y, z are *independent. What matrix is R?*

Answer *R is the identity matrix.* All three correlations $\rho_{xx}, \rho_{yy}, \rho_{zz}$ are 1 by definition. All three cross-correlations $\rho_{xy}, \rho_{xz}, \rho_{yz}$ are zero by independence.

The correlation matrix R comes from the covariance matrix V, when we rescale every row and every column. Divide each row i and column i by the ith standard deviation σ_i.

(a) $R = DVD$ for the diagonal matrix $D = \text{diag}\,[1/\sigma_1, \ldots, 1/\sigma_M]$.

(b) If covariance V is positive definite, correlation $R = DVD$ is also positive definite.

■ WORKED EXAMPLES ■

12.2 A Suppose x and y are independent random variables with mean 0 and variance 1. Then the covariance matrix $\boldsymbol{V_X}$ for $\boldsymbol{X} = (x, y)$ is the 2 by 2 identity matrix. What are the mean $\boldsymbol{m_Z}$ and the covariance matrix $\boldsymbol{V_Z}$ for the 3-component vector $\boldsymbol{Z} = (x, y, ax + by)$?

Solution

$$\boldsymbol{Z}\textbf{ is connected to }\boldsymbol{X}\textbf{ by }\boldsymbol{A} \quad Z = \begin{bmatrix} x \\ y \\ ax+by \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ a & b \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = AX.$$

The vector $\boldsymbol{m_X}$ contains the means of the M components of $\boldsymbol{X}$. The vector $\boldsymbol{m_Z}$ contains the means of the K components of $\boldsymbol{Z} = A\boldsymbol{X}$. The matrix connection between the means of $\boldsymbol{X}$ and $\boldsymbol{Z}$ has to be linear: $\boldsymbol{m_Z} = A\boldsymbol{m_X}$. The mean of $ax + by$ is $am_x + bm_y$.

The covariance matrix for $\boldsymbol{Z}$ is $\boldsymbol{V_Z} = AA^{\mathrm{T}}$, when $\boldsymbol{V_X}$ is the 2 by 2 identity matrix:

$$\boldsymbol{V_Z} = \begin{matrix}\textbf{covariance matrix for} \\ \boldsymbol{Z} = (\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{ax} + \boldsymbol{by})\end{matrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ a & b \end{bmatrix} \begin{bmatrix} 1 & 0 & a \\ 0 & 1 & b \end{bmatrix} = \begin{bmatrix} \boldsymbol{1} & \boldsymbol{0} & \boldsymbol{a} \\ \boldsymbol{0} & \boldsymbol{1} & \boldsymbol{b} \\ \boldsymbol{a} & \boldsymbol{b} & \boldsymbol{a^2 + b^2} \end{bmatrix}.$$

Interpretation: x and y are independent so $\sigma_{xy} = 0$. Then the covariance of x with $ax + by$ is a and the covariance of y with $ax + by$ is b. Those just come from the two independent parts of $ax + by$. Finally, equation (18) gives the variance of $ax + by$:

$$\textbf{Use } \boldsymbol{V_Z} = \boldsymbol{AV_XA^{\mathrm{T}}} \qquad \sigma^2_{ax+by} = \sigma^2_{ax} + \sigma^2_{by} + 2\sigma_{ax,by} = a^2 + b^2 + 0.$$

The 3 by 3 matrix $\boldsymbol{V_Z}$ is *singular*. Its determinant is $a^2 + b^2 - a^2 - b^2 = 0$. The third component $z = ax + by$ is completely dependent on x and y. The rank of $\boldsymbol{V_Z}$ is only 2.

GPS Example The signal from a GPS satellite includes its departure time. The receiver clock gives the arrival time. The receiver multiplies the travel time by the speed of light. Then it knows the distance from that satellite. Distances from four or more satellites pinpoint the receiver position (using least squares !).

One problem: The speed of light changes in the ionosphere. But the correction will be almost the same for all nearby receivers. If one receiver stays in a known position, we can take differences from that position. **Differential GPS** reduces the error variance:

$$\textbf{Difference matrix } \boldsymbol{A} = \begin{bmatrix} 1 & -1 \end{bmatrix} \quad \textbf{Covariance matrix } \boldsymbol{V_Z} = \boldsymbol{AV_XA^{\mathrm{T}}}$$

$$V_Z = \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \boldsymbol{\sigma_1^2 - 2\sigma_{12} + \sigma_2^2}$$

Errors in the speed of light are gone. Then centimeter positioning accuracy is achievable. (The key ideas are on page 320 of *Algorithms for Global Positioning* by Borre and Strang.) The GPS world is all about time and space and amazing accuracy.

Problem Set 12.2

1 (a) Compute the variance σ^2 when the coin flip probabilities are p and $1-p$ (tails $= 0$, heads $= 1$).

(b) The sum of N independent flips (0 or 1) is the count of heads after N tries. The rule (16-17-18) for the variance of a sum gives $\sigma^2 =$ _____.

2 What is the covariance σ_{kl} between the results $x_1, \ldots, x_n$ of Experiment 3 and the results $y_1, \ldots, y_n$ of Experiment 5 ? Your formula will look like σ_{12} in equation (2). Then the $(3,5)$ and $(5,3)$ entries of the covariance matrix V are $\sigma_{35} = \sigma_{53}$.

3 For $M = 3$ experiments, the variance-covariance matrix V will be 3 by 3. There will be a probability p_{ijk} that the three outputs are x_i and y_j and z_k. Write down a formula like equation (7) for the matrix V.

4 What is the covariance matrix V for $M = 3$ independent experiments with means m_1, m_2, m_3 and variances $\sigma_1^2, \sigma_2^2, \sigma_3^2$?

Problems 5–9 are about the conditional probability that $Y = y_j$ when we know $X = x_i$.
Notation: **Prob $(Y = y_j | X = x_i)$** = probability of the outcome y_j given that $X = x_i$.

Example 1 *Coin* 1 *is glued to coin* 2. Then Prob $(Y =$ heads when $X =$ heads$)$ is **1**.

Example 2 *Independent coin flips*: X gives no information about Y. Useless to know X. Then Prob $(Y =$ heads $| X =$ heads$)$ is the same as Prob $(Y =$ heads$)$.

5 Explain the **sum rule** of conditional probability :

$$\text{Prob}\,(Y = y_j) = \text{ sum over all outputs } x_i \text{ of Prob}\,(Y = y_j | X = x_i).$$

6 The n by n matrix P contains **joint probabilities p_{ij}** = Prob $(X = x_i$ **and** $Y = y_j)$.

Explain why the conditional Prob $(Y = y_j | X = x_i)$ equals $\dfrac{p_{ij}}{p_{i1} + \cdots + p_{in}} = \dfrac{p_{ij}}{p_i}$.

7 For this joint probability matrix with Prob $(x_1, y_2) = 0.3$, find Prob $(y_2 | x_1)$ and Prob (x_1).

$$P = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} = \begin{bmatrix} 0.1 & 0.3 \\ 0.2 & 0.4 \end{bmatrix}$$

The entries p_{ij} add to 1.
Some i, j must happen.

8 Explain the **product rule** of conditional probability:

p_{ij} = Prob $(X = x_i$ **and** $Y = y_j)$ equals Prob $(Y = y_j | X = x_i)$ times Prob $(X = x_i)$.

9 Derive this **Bayes Theorem** for p_{ij} from the product rule in Problem 8:

$$\text{Prob}\,(Y = y_j \text{ and } X = x_i) = \frac{\text{Prob}\,(X = x_i | Y = y_j)\ \text{Prob}\,(Y = y_j)}{\text{Prob}\,(X = x_i)}$$

"Bayesians" use prior information. "Frequentists" only use sampling information.

12.3 Multivariate Gaussian and Weighted Least Squares

The normal probability density $p(x)$ (the Gaussian) depends on only two numbers:

Mean m and variance σ^2 $$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x-m)^2/2\sigma^2}. \quad (1)$$

The graph of $p(x)$ is a bell-shaped curve centered at $x = m$. The continuous variable x can be anywhere between $-\infty$ and ∞. With probability close to $\frac{2}{3}$, that random x will lie between $m - \sigma$ and $m + \sigma$ (less than one standard deviation σ from its mean value m).

$$\int_{-\infty}^{\infty} p(x)\,dx = 1 \quad \text{and} \quad \int_{m-\sigma}^{m+\sigma} p(x)\,dx = \frac{1}{\sqrt{2\pi}} \int_{-1}^{1} e^{-X^2/2}\,dX \approx \frac{2}{3}. \quad (2)$$

That integral has a change of variables from x to $X = (x - m)/\sigma$. This simplifies the exponent to $-X^2/2$ and it simplifies the limits of integration to -1 and 1. Even the $1/\sigma$ from p disappears outside the integral because dX equals dx/σ. Every Gaussian turns into a **standard Gaussian** $p(X)$ with mean $m = 0$ and variance $\sigma^2 = 1$. Just call it $p(x)$:

The standard normal distribution $N(0,1)$ has $$p(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}. \quad (3)$$

Integrating $p(x)$ from $-\infty$ to x gives the cumulative distribution $F(x)$: the probability that a random sample is below x. That probability will be $F = \frac{1}{2}$ at $x = 0$ (the mean).

Two-dimensional Gaussians

Now we have $M = 2$ Gaussian random variables x and y. They have means m_1 and m_2. They have variances σ_1^2 and σ_2^2. If they are *independent*, then their probability density $p(x, y)$ is just $\boldsymbol{p_1(x)}$ **times** $\boldsymbol{p_2(y)}$. Multiply probabilities when variables are independent:

Independent x and y $$p(x,y) = \frac{1}{2\pi\sigma_1\sigma_2} e^{-(x-m_1)^2/2\sigma_1^2} e^{-(y-m_2)^2/2\sigma_2^2} \quad (4)$$

The covariance of x and y will be $\boldsymbol{\sigma_{12} = 0}$. The covariance matrix V will be *diagonal*. The variances σ_1^2 and σ_2^2 are always on the main diagonal of V. The exponent in $p(x, y)$ is just the sum of the x-exponent and the y-exponent. Good to notice that the two exponents can be combined into $-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{m})^{\mathrm{T}} V^{-1} (\boldsymbol{x} - \boldsymbol{m})$ with V^{-1} in the middle:

$$-\frac{(x-m_1)^2}{2\sigma_1^2} - \frac{(y-m_2)^2}{2\sigma_2^2} = -\frac{1}{2} \begin{bmatrix} x-m_1 & y-m_2 \end{bmatrix} \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}^{-1} \begin{bmatrix} x-m_1 \\ y-m_2 \end{bmatrix} \quad (5)$$

Non-independent x and y

We are ready to give up independence. The exponent (5) with V^{-1} is still correct when V is no longer a diagonal matrix. **Now the Gaussian depends on a vector m and a matrix V.**

When $M = 2$, the first variable x may give partial information about the second variable y (and vice versa). Maybe part of y is decided by x and part is truly independent. It is the M by M covariance matrix V that accounts for dependencies between the M variables $\boldsymbol{x} = x_1, \ldots, x_M$. Its inverse V^{-1} goes into $p(\boldsymbol{x})$:

Multivariate Gaussian probability distribution

$$\boxed{p(\boldsymbol{x}) = \frac{1}{(\sqrt{2\pi})^M \sqrt{\det V}}\, e^{-(\boldsymbol{x}-\boldsymbol{m})^{\mathrm{T}} V^{-1}(\boldsymbol{x}-\boldsymbol{m})/2}} \quad (6)$$

The vectors $\boldsymbol{x} = (x_1, \ldots, x_M)$ and $\boldsymbol{m} = (m_1, \ldots, m_M)$ contain the random variables and their means. The M square roots of 2π and the determinant of V are included to make the total probability equal to 1. Let me check that by linear algebra. I use the eigenvalues λ and orthonormal eigenvectors $\boldsymbol{q}$ of the symmetric matrix $V = Q\Lambda Q^{\mathrm{T}}$. So $\boldsymbol{V^{-1} = Q\Lambda^{-1}Q^{\mathrm{T}}}$:

$$\boldsymbol{X = x - m} \qquad \boldsymbol{(x-m)^{\mathrm{T}} V^{-1}(x-m) = X^{\mathrm{T}} Q \Lambda^{-1} Q^{\mathrm{T}} X = Y^{\mathrm{T}} \Lambda^{-1} Y}$$

Notice! The combinations $\boldsymbol{Y = Q^{\mathrm{T}} X = Q^{\mathrm{T}}(x - m)}$ are statistically independent. *Their covariance matrix Λ is diagonal.*

This step of diagonalizing V by its eigenvector matrix Q is the same as "uncorrelating" the random variables. Covariances are zero for the new variables $X_1, \ldots X_m$. This is the point where linear algebra helps calculus to compute multidimensional integrals.

The integral of $p(\boldsymbol{x})$ is not changed when we center the variable $\boldsymbol{x}$ by subtracting $\boldsymbol{m}$ to reach $\boldsymbol{X}$, and rotate that variable to reach $\boldsymbol{Y = Q^{\mathrm{T}} X}$. The matrix $\boldsymbol{\Lambda}$ is diagonal! So the integral we want splits into M separate one-dimensional integrals that we know:

$$\int \cdots \int e^{-\boldsymbol{Y^{\mathrm{T}}\Lambda^{-1}Y/2}}\, \boldsymbol{dY} = \left(\int_{-\infty}^{\infty} e^{-y_1^2/2\lambda_1}\, dy_1 \right) \cdots \left(\int_{-\infty}^{\infty} e^{-y_M^2/2\lambda_M}\, dy_M \right)$$

$$= \left(\sqrt{2\pi\lambda_1}\right) \cdots \left(\sqrt{2\pi\lambda_M}\right) = \left(\sqrt{2\pi}\right)^M \sqrt{\det V}. \quad (7)$$

The determinant of V (also the determinant of Λ) is the product $(\lambda_1) \ldots (\lambda_M)$ of the eigenvalues. Then (7) gives the correct number to divide by so that $p(x_1, \ldots, x_M)$ in equation (6) has integral $= 1$ as desired.

The mean and variance of $p(\boldsymbol{x})$ are also M-dimensional integrals. The same idea of diagonalizing V by its eigenvectors and introducing $\boldsymbol{Y = Q^{\mathrm{T}} X}$ will find those integrals:

Vector m of means
$$\int \cdots \int \boldsymbol{x}\, p(\boldsymbol{x})\, \boldsymbol{dx} = (m_1, m_2, \ldots) = \boldsymbol{m} \quad (8)$$

Covariance matrix V
$$\int \cdots \int (\boldsymbol{x} - \boldsymbol{m})\, p(\boldsymbol{x}) (\boldsymbol{x} - \boldsymbol{m})^{\mathrm{T}}\, \boldsymbol{dx} = \boldsymbol{V}. \quad (9)$$

Conclusion: Formula (6) for the probability density $p(\boldsymbol{x})$ has all the properties we want.

Weighted Least Squares

In Chapter 4, least squares started from an unsolvable system $A\boldsymbol{x} = \boldsymbol{b}$. We chose $\widehat{\boldsymbol{x}}$ to minimize the error $||\boldsymbol{b} - A\boldsymbol{x}||^2$. That led us to the least squares equation $A^{\mathrm{T}}A\widehat{\boldsymbol{x}} = A^{\mathrm{T}}\boldsymbol{b}$. The best $A\widehat{\boldsymbol{x}}$ is the projection of $\boldsymbol{b}$ onto the column space of A. But is this squared distance $E = ||\boldsymbol{b} - A\boldsymbol{x}||^2$ the right error measure to minimize?

If the measurement errors in $\boldsymbol{b}$ are independent random variables, with mean $m = 0$ and variance $\sigma^2 = 1$ and a normal distribution, Gauss would say **yes**: *Use least squares.* If the errors are not independent or their variances are not equal. Gauss would say **no**: *Use* ***weighted*** *least squares.* This section will show that the good measure of error is $\boldsymbol{E} = (\boldsymbol{b} - A\boldsymbol{x})^{\mathrm{T}}\boldsymbol{V^{-1}}(\boldsymbol{b} - A\boldsymbol{x})$. The equation for the best $\widehat{\boldsymbol{x}}$ uses the covariance matrix V:

Weighted least squares $$A^{\mathrm{T}}V^{-1}A\widehat{\boldsymbol{x}} = A^{\mathrm{T}}V^{-1}\boldsymbol{b}. \tag{10}$$

The most important examples have m *independent* errors in $\boldsymbol{b}$. Those errors have variances $\sigma_1^2, \ldots, \sigma_m^2$. By independence, V is a diagonal matrix. The good weights $1/\sigma_1^2, \ldots, 1/\sigma_m^2$ come from V^{-1}. *We are weighting the errors in* $\boldsymbol{b}$ *to have* **variance** $= \mathbf{1}$:

Weighted least squares
Independent errors in b

$$\text{Minimize} \quad E = \sum_{i=1}^{m} \frac{(\boldsymbol{b} - A\boldsymbol{x})_i^2}{\sigma_i^2} \tag{11}$$

By weighting the errors, we are "whitening" the noise. **White noise** is a quick description of independent errors based on the standard Gaussian $\mathbf{N}(0, 1)$ with mean zero and $\sigma^2 = 1$.

Let me write down the steps to equations (10) and (11) for the best $\widehat{\boldsymbol{x}}$:

Start with $A\boldsymbol{x} = \boldsymbol{b}$ (m equations, n unknowns, $m > n$, no solution)

Each right side b_i has mean zero and variance σ_i^2. The b_i are independent.

Divide the ith equation by σ_i to have variance $= 1$ for every b_i/σ_i

That division turns $A\boldsymbol{x} = \boldsymbol{b}$ into $V^{-1/2}A\boldsymbol{x} = V^{-1/2}\boldsymbol{b}$ with $V^{-1/2} = \text{diag}\,(1/\sigma_1, \ldots, 1/\sigma_m)$

Ordinary least squares on those weighted equations has $A \to V^{-1/2}A$ and $\boldsymbol{b} \to V^{-1/2}\boldsymbol{b}$

$$(V^{-1/2}A)^{\mathrm{T}}(V^{-1/2}A)\widehat{\boldsymbol{x}} = (V^{-1/2}A)^{\mathrm{T}}V^{-1/2}\boldsymbol{b} \;\text{ is }\; A^{\mathrm{T}}V^{-1}A\widehat{\boldsymbol{x}} = A^{\mathrm{T}}V^{-1}\boldsymbol{b}. \tag{12}$$

Because of $1/\sigma^2$ in V^{-1}, more reliable equations (*smaller* σ) get heavier weights. This is the main point of weighted least squares.

Those diagonal weightings (uncoupled equations) are the most frequent and the simplest. They apply to *independent errors in the* b_i. When these measurement errors are not independent, V is no longer diagonal—but (12) is still the correct weighted equation.

In practice, finding all the covariances can be serious work. Diagonal V is simpler.

The Variance in the Estimated $\widehat{x}$

One more point: Often the important question is not the best $\widehat{x}$ for one particular set of measurements $\boldsymbol{b}$. This is only one sample! The real goal is to know the reliability of the whole experiment. That is measured (as reliability always is) by the **variance in the estimate** $\widehat{x}$. First, zero mean in $\boldsymbol{b}$ gives zero mean in $\widehat{x}$. Then the formula connecting variance $\boldsymbol{V}$ in the inputs $\boldsymbol{b}$ to variance $\boldsymbol{W}$ in the outputs $\widehat{x}$ turns out to be beautiful:

Variance-covariance matrix W for $\widehat{x}$ $\quad \mathrm{E}[(\widehat{x}-x)(\widehat{x}-x)^{\mathrm{T}}] = (\boldsymbol{A^{\mathrm{T}}V^{-1}A})^{-1}$. (13)

That smallest possible variance comes from the best possible weighting, which is V^{-1}.

This key formula is a perfect application of Section 12.2. **If $\boldsymbol{b}$ has covariance matrix V, then $\widehat{x} = \boldsymbol{Lb}$ has covariance matrix $\boldsymbol{LVL^{\mathrm{T}}}$.** Equation (12) above tells us that L is $(A^{\mathrm{T}}V^{-1}A)^{-1}A^{\mathrm{T}}V^{-1}$. Now substitute this into LVL^{T} and watch equation (13) appear:

$$\boldsymbol{LVL^{\mathrm{T}}} = (A^{\mathrm{T}}V^{-1}A)^{-1}A^{\mathrm{T}}V^{-1} \quad V \quad V^{-1}A(A^{\mathrm{T}}V^{-1}A)^{-1} = \boldsymbol{(A^{\mathrm{T}}V^{-1}A)^{-1}}.$$

This is the covariance W of the output, our best estimate $\widehat{x}$. It is time for examples.

Example 1 Suppose a doctor measures your heart rate x three times $(m = 3, n = 1)$:

$$\begin{matrix} x = b_{\cdot} \\ x = b_2 \\ x = b_3 \end{matrix} \quad \text{is} \quad A\boldsymbol{x} = \boldsymbol{b} \quad \text{with} \quad A = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \quad \text{and} \quad V = \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{bmatrix}$$

The variances could be $\sigma_1^2 \cdot \quad /9$ and $\sigma_2^2 = 1/4$ and $\sigma_3^2 = \mathbf{1}$. You are getting more nervous as measurements are taken: ; is less reliable than b_2 and b_1. All three measurements contain some information, so they all go into the best (weighted) estimate $\widehat{x}$:

$$V^{-1/2}A\widehat{x} = V^{-1/2}\boldsymbol{b} \quad \text{is} \quad \begin{matrix} 3x = 3b_1 \\ 2x = 2b_2 \\ 1x = 1b_3 \end{matrix} \quad \text{leading to} \quad \boldsymbol{A^{\mathrm{T}}V^{-1}A\widehat{x} = A^{\mathrm{T}}V^{-1}b}$$

$$\begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 9 & & \\ & 4 & \\ & & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \widehat{x} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 9 & & \\ & 4 & \\ & & 1 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

$$\boxed{\widehat{x} = \frac{9b_1 + 4b_2 + b_3}{14} \quad \textbf{is a weighted average of } b_1, b_2, b_3}$$

Most weight is on b_1 since its variance σ_1 is smallest. The variance of $\widehat{\boldsymbol{x}}$ has the beautiful formula $W = (A^{\mathrm{T}}V^{-1}A)^{-1} = 1/14$:

$$\textbf{Variance of } \widehat{x} \quad \left(\begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 9 & & \\ & 4 & \\ & & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right)^{-1} = \frac{1}{14} \quad \text{is smaller than} \quad \frac{1}{9}$$

The BLUE theorem of Gauss (proved on the website) says that our $\widehat{\boldsymbol{x}} = L\boldsymbol{b}$ is the best linear unbiased estimate of the solution to $A\boldsymbol{x} = \boldsymbol{b}$. Any other unbiased choice $\boldsymbol{x}^* = L^*\boldsymbol{b}$ has greater variance than $\widehat{\boldsymbol{x}}$. All unbiased choices have $L^*A = I$ so that an exact $A\boldsymbol{x} = \boldsymbol{b}$ will produce the right answer $\boldsymbol{x} = L^*\boldsymbol{b} = L^*A\boldsymbol{x}$.

Note. I must add that there are reasons not to minimize squared errors in the first place. One reason: This $\widehat{\boldsymbol{x}}$ often has many small components. The squares of small numbers are very small, and they appear when we minimize. It is easier to make sense of *sparse* vectors—only a few nonzeros. Statisticians often prefer to minimize **unsquared errors**: **the sum of** $|(\boldsymbol{b} - A\boldsymbol{x})_i|$. *This error measure is* L^1 *instead of* L^2. Because of the absolute values, the equation for $\widehat{\boldsymbol{x}}$ becomes nonlinear (it is actually piecewise linear).

Fast new algorithms are computing a sparse $\widehat{\boldsymbol{x}}$ quickly and the future may belong to L^1.

The Kalman Filter

The "Kalman filter" is the great algorithm in dynamic least squares. That word *dynamic* means that new measurements $\boldsymbol{b}_k$ keep coming. So the best estimate $\widehat{\boldsymbol{x}}_k$ keeps changing (based on all of $\boldsymbol{b}_0, \ldots, \boldsymbol{b}_k$). More than that, the matrix A is also changing. So $\widehat{\boldsymbol{x}}_2$ will be our best least squares estimate of the latest solution $\boldsymbol{x}_k$ to the **whole history of observation equations and update equations (state equations) up to time 2**:

$$A_0\boldsymbol{x}_0 = \boldsymbol{b}_0 \qquad \boldsymbol{x}_1 = F_0\boldsymbol{x}_0 \qquad A_1\boldsymbol{x}_1 = \boldsymbol{b}_1 \qquad \boldsymbol{x}_2 = F_1\boldsymbol{x}_1 \qquad A_2\boldsymbol{x}_2 = \boldsymbol{b}_2 \tag{14}$$

The Kalman idea is to introduce one equation at a time. There will be errors in each equation. With every new equation, we update the best estimate $\widehat{\boldsymbol{x}}_k$ for the current $\boldsymbol{x}_k$. But history is not forgotten! This new estimate $\widehat{\boldsymbol{x}}_k$ uses all the past observations $\boldsymbol{b}_0$ to $\boldsymbol{b}_{k-1}$ and all the state equations $\boldsymbol{x}_{\text{new}} = F_{\text{old}}\,\boldsymbol{x}_{\text{old}}$. A large and growing least squares problem.

One more important point. Each least squares equation is **weighted** using the covariance matrix V_k for the error in b_k. There is even a covariance matrix C_k for errors in the update equations $\boldsymbol{x}_{k+1} = F_k x_k$. The best $\widehat{\boldsymbol{x}}_2$ then depends on $\boldsymbol{b}_0, \boldsymbol{b}_1, \boldsymbol{b}_2$ and V_0, V_1, V_2 and C_1, C_2. The good way to write $\widehat{\boldsymbol{x}}_k$ is as an update to the previous $\widehat{\boldsymbol{x}}_{k-1}$.

Let me concentrate on a simplified problem, without the matrices F_k and the covariances C_k. We are estimating the same true $\boldsymbol{x}$ at every step. How do we get $\widehat{\boldsymbol{x}}_1$ from $\widehat{\boldsymbol{x}}_0$?

OLD $A_0\,\boldsymbol{x}_0 = \boldsymbol{b}_0$ leads to the weighted equation $A_0^{\mathrm{T}}\,V_0^{-1}\,A_0\,\widehat{\boldsymbol{x}}_0 = A_0^{\mathrm{T}}\,V_0^{-1}\,\boldsymbol{b}_0$. (15)

NEW $\begin{bmatrix} A_0 \\ A_1 \end{bmatrix} \widehat{\boldsymbol{x}}_1 = \begin{bmatrix} \boldsymbol{b}_0 \\ \boldsymbol{b}_1 \end{bmatrix}$ leads to the following weighted equation for $\widehat{\boldsymbol{x}}_1$:

$$\begin{bmatrix} A_0^{\mathrm{T}} & A_1^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} V_0^{-1} & \\ & V_1^{-1} \end{bmatrix} \begin{bmatrix} A_0 \\ A_1 \end{bmatrix} \widehat{\boldsymbol{x}}_1 = \begin{bmatrix} A_0^{\mathrm{T}} & A_1^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} V_0^{-1} & \\ & V_1^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{b}_0 \\ \boldsymbol{b}_1 \end{bmatrix}. \quad (16)$$

Yes, we could just solve that new problem and forget the old one. But the old solution $\widehat{\boldsymbol{x}}_0$ needed work that we hope to reuse in $\widehat{\boldsymbol{x}}_1$. What we look for is **an update to $\widehat{\boldsymbol{x}}_0$**:

Kalman update gives $\widehat{\boldsymbol{x}}_1$ from $\widehat{\boldsymbol{x}}_0$ $\quad \widehat{\boldsymbol{x}}_1 = \widehat{\boldsymbol{x}}_0 + K_1(\boldsymbol{b}_1 - A_1\widehat{\boldsymbol{x}}_0).$ (17)

The update correction is the mismatch $\boldsymbol{b}_1 - A_1\widehat{\boldsymbol{x}}_0$ between the old state $\widehat{\boldsymbol{x}}_0$ and the new measurements $\boldsymbol{b}_1$—multiplied by the *Kalman gain matrix* K_1. The formula for K_1 comes from comparing the solutions $\widehat{\boldsymbol{x}}_1$ and $\widehat{\boldsymbol{x}}_0$ to (15) and (16). And when we update $\widehat{\boldsymbol{x}}_0$ to $\widehat{\boldsymbol{x}}_1$ based on new data $\boldsymbol{b}_1$, **we also update the covariance matrix $\boldsymbol{W_0}$ to $\boldsymbol{W_1}$**. Remember $W_0 = (A_0^{\mathrm{T}} V_0^{-1} A_0)^{-1}$ from equation (13). Update its inverse to W_1^{-1}:

Covariance W_1 of errors in $\widehat{\boldsymbol{x}}_1$ $\quad \boldsymbol{W_1^{-1} = W_0^{-1} + A_1^{\mathrm{T}} V_1^{-1} A_1}$ (18)

Kalman gain matrix K_1 $\quad \boldsymbol{K_1 = W_1 A_1^{\mathrm{T}} V_1^{-1}}$ (19)

This is the heart of the Kalman filter. Notice the importance of the W_k. Those matrices measure the reliability of the whole process, where the vector $\widehat{\boldsymbol{x}}_k$ estimates the current state based on the particular measurements $\boldsymbol{b}_0$ to $\boldsymbol{b}_k$.

Whole chapters and whole books are written to explain the dynamic Kalman filter, when the states $\boldsymbol{x}_k$ are also changing (based on the matrices F_k). There is a *prediction* of $\boldsymbol{x}_k$ using F, followed by a *correction* using the new data $\boldsymbol{b}$. Perhaps best to stop here.

This page was about **recursive least squares**: adding new data b_k and updating both $\widehat{\boldsymbol{x}}$ and W: the best current estimate based on all the data, and its covariance matrix.

Problem Set 12.3

1 Two measurements of the same variable x give two equations $x = b_1$ and $x = b_2$. Suppose the means are zero and the variances are σ_1^2 and σ_2^2, with independent errors: V is diagonal with entries σ_1^2 and σ_2^2. Write the two equations as $A\boldsymbol{x} = \boldsymbol{b}$ (A is 2 by 1). As in the text Example 1, find this best estimate $\widehat{\boldsymbol{x}}$ based on b_1 and b_2:

$$\widehat{\boldsymbol{x}} = \frac{b_1/\sigma_1^2 + b_2/\sigma_2^2}{1/\sigma_1^2 + 1/\sigma_2^2} \qquad \mathrm{E}\left[\widehat{\boldsymbol{x}}\,\widehat{\boldsymbol{x}}^{\mathrm{T}}\right] = \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)^{-1}.$$

2 (a) In Problem 1, suppose the second measurement b_2 becomes super-exact and its variance $\sigma_2 \to 0$. What is the best estimate $\widehat{\boldsymbol{x}}$ when σ_2 reaches zero?

(b) The opposite case has $\sigma_2 \to \infty$ and no information in b_2. What is now the best estimate $\widehat{\boldsymbol{x}}$ based on b_1 and b_2?

3 If x and y are independent with probabilities $p_1(x)$ and $p_2(y)$, then $p(x,y) = p_1(x)\,p_2(y)$. By separating double integrals into products of single integrals ($-\infty$ to ∞) show that

$$\iint p(x,y)\,dx\,dy = \mathbf{1} \qquad \text{and} \qquad \iint (x+y)\,p(x,y)\,dx\,dy = \boldsymbol{m_1 + m_2}.$$

4 Continue Problem 3 for independent x, y to show that $p(x,y) = p_1(x)\,p_2(y)$ has

$$\iint (x-m_1)^2\,p(x,y)\,dx\,dy = \boldsymbol{\sigma_1^2} \qquad \iint (x-m_1)(y-m_2)\,p(x,y)\,dx\,dy = \mathbf{0}.$$

So the 2 by 2 covariance matrix V is diagonal and its entries are _____.

5 Show that the inverse of a 2 by 2 covariance matrix V is

$$V^{-1} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}^{-1} = \frac{1}{1-\rho^2} \begin{bmatrix} 1/\sigma_1^2 & -\rho/\sigma_1\sigma_2 \\ -\rho/\sigma_1\sigma_2 & 1/\sigma_2^2 \end{bmatrix} \quad \begin{matrix} \text{with correlation} \\ \rho = \sigma_{12}/\sigma_1\sigma_2. \end{matrix}$$

This produces the exponent $-(\boldsymbol{x}-\boldsymbol{m})^{\mathrm{T}}\,\boldsymbol{V}^{-1}(\boldsymbol{x}-\boldsymbol{m})$ in a 2-variable Gaussian.

6 Suppose $\widehat{x}_k$ is the average of $b_1, \ldots, b_k$. A new measurement b_{k+1} arrives and we want the new average $\widehat{x}_{k+1}$. The Kalman update equation (17) is

$$\textbf{New average} \qquad \widehat{x}_{k+1} = \widehat{x}_k + \frac{1}{k+1}\,(b_{k+1} - \widehat{x}_k)\,.$$

Verify that $\widehat{x}_{k+1}$ is the correct average of $b_1 \ldots, b_{k+1}$.

7 Also check the update equation (18) for the variance $W_{k+1} = \sigma^2/(k+1)$ of this average $\widehat{x}$ assuming that $W_k = \sigma^2/k$ and b_{k+1} has variance $V = \sigma^2$.

8 (**Steady model**) Problems 6–7 were *static* least squares. All the sample averages $\widehat{x}_k$ were estimates of the same x. To make the Kalman filter *dynamic*, include also a *state equation $x_{k+1} = Fx_k$ with its own error variance s^2*. The dynamic least squares problem allows x to "drift" as k increases:

$$\begin{bmatrix} 1 & \\ -F & 1 \\ & 1 \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} b_0 \\ 0 \\ b_1 \end{bmatrix} \text{ with variances } \begin{bmatrix} \sigma^2 \\ s^2 \\ \sigma^2 \end{bmatrix}.$$

With $F = 1$, divide both sides of those three equations by σ, s, and σ. Find $\widehat{x_0}$ and $\widehat{x_1}$ by least squares, which gives more weight to the recent b_1. The Kalman filter is developed in *Algorithms for Global Positioning* (Borre and Strang, Wellesley-Cambridge Press).

Change in A^{-1} from a Change in A

This final page connects the beginning of the book (inverses and rank one matrices) with the end of the book (dynamic least squares and filters). Begin with this basic formula:

The inverse of $M = I - uv^{\mathrm{T}}$ is $M^{-1} = I + \dfrac{uv^{\mathrm{T}}}{1 - v^{\mathrm{T}}u}$

The quickest proof is $MM^{-1} = I - uv^{\mathrm{T}} + (1 - uv^{\mathrm{T}})\dfrac{uv^{\mathrm{T}}}{1 - v^{\mathrm{T}}u} = I - uv^{\mathrm{T}} + uv^{\mathrm{T}} = I$.

M is not invertible if $v^{\mathrm{T}}u = 1$ (then $Mu = 0$). Here $v^{\mathrm{T}} = u^{\mathrm{T}} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}$:

Example The inverse of $M = I - \begin{bmatrix} 1&1&1 \\ 1&1&1 \\ 1&1&1 \end{bmatrix}$ is $M^{-1} = I + \dfrac{1}{1-3}\begin{bmatrix} 1&1&1 \\ 1&1&1 \\ 1&1&1 \end{bmatrix}$

But we don't always start from the identity matrix. Many applications need to invert $M = A - uv^{\mathrm{T}}$. After we solve $Ax = b$ we expect a rank one change to give $My = b$. The division by $1 - v^{\mathrm{T}}u$ above will become a division by $c = 1 - v^{\mathrm{T}}A^{-1}u = 1 - v^{\mathrm{T}}z$.

Step 1 Solve $Az = u$ and compute $c = 1 - v^{\mathrm{T}}z$.

Step 2 If $c \neq 0$ then $M^{-1}b$ is $y = x + \dfrac{v^{\mathrm{T}}x}{c}z$.

Suppose A is easy to work with. A might already be factored into LU by elimination. Then this Sherman-Woodbury-Morrison formula is the fast way to solve $My = b$. Here are three problems to end the book!

9 Take Steps 1– 2to find y when $A = I$ and $u^{\mathrm{T}} = v^{\mathrm{T}} = [1\ \ 2\ \ 3]$ and $b^{\mathrm{T}} = [2\ \ 1\ \ 4]$.

10 Step 2 in this "update formula" claims that $My = \left(A - uv^{\mathrm{T}}\right)\left(x + \dfrac{v^{\mathrm{T}}x}{c}z\right) = b$.

Simplify this to $\dfrac{uv^{\mathrm{T}}x}{c}[1 - c - v^{\mathrm{T}}z] = 0$. This is true since $c = 1 - v^{\mathrm{T}}z$.

11 When A has a new row v^{T}, $A^{\mathrm{T}}A$ in the least squares equation changes to M:

$$M = \begin{bmatrix} A^{\mathrm{T}} & v \end{bmatrix}\begin{bmatrix} A \\ v^{\mathrm{T}} \end{bmatrix} = A^{\mathrm{T}}A + vv^{\mathrm{T}} = \text{ rank one change in } A^{\mathrm{T}}A.$$

Why is that multiplication correct? The updated $\widehat{x}_{\text{new}}$ comes from Steps 1 and 2.

For reference here are four formulas for M^{-1}. The first two were given above, when the change was uv^{T}. Formulas 3 and 4 go beyond rank one to allow matrices U, V, W.

1	$M = I - uv^{\mathrm{T}}$	and	$M^{-1} = I + uv^{\mathrm{T}}/(1 - v^{\mathrm{T}}u)$ (*rank*1 *change*)
2	$M = A - uv^{\mathrm{T}}$	and	$M^{-1} = A^{-1} + A^{-1}uv^{\mathrm{T}}A^{-1}/(1 - v^{\mathrm{T}}A^{-1}u)$
3	$M = I - UV$	and	$M^{-1} = I_n + U(I_m - VU)^{-1}V$
4	$M = A - UW^{-1}V$	and	$M^{-1} = A^{-1} + A^{-1}U(W - VA^{-1}U)^{-1}VA^{-1}$

Formula **4** is the "matrix inversion lemma" in engineering. Not seen until now! The Kalman filter for solving block tridiagonal systems uses formula **4** at each step.