The Normal Distribution from Gauss to Kalman

Divakar Viswanath

Department of Mathematics, University of Michigan

Contents

1 Introduction: random variables, mean, and variance

In this lecture, we will introduce the normal distribution, which is one of the of great ideas in mathematics. Its applicability is very wide, touching every area of the sciences. In addition, there are mathematical questions related to it that are still a subject of active research.

To gain an understanding of the normal distribution, we will begin with a simple example and rigorously deduce the central limit theorem for that example. The central limit theorem gives a partial explanation for why the normal distribution occurs in so many situations.

Our next step is to deduce a very special case of the Kalman filter. The Kalman filter is undoubtedly one of the most consequential ideas in applied mathematics. Its applications are profound. The cell phone in your pocket may have a Kalman filter. Although I have no direct knowledge, I am certain that India's Mars mission Mangalyaan makes extensive use of Kalman filters.

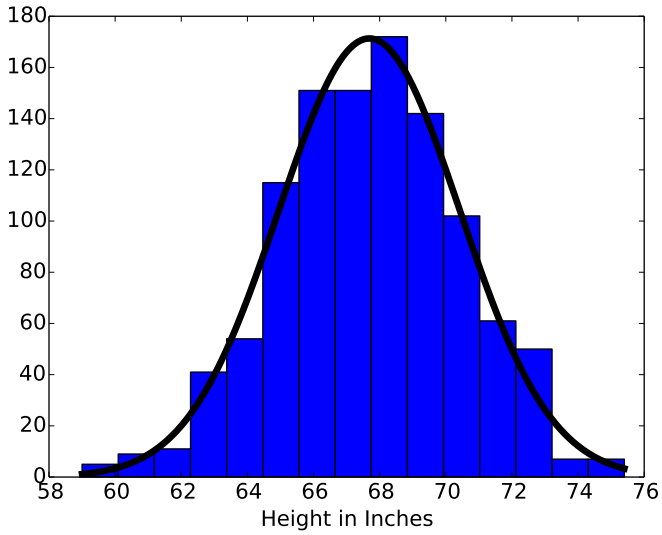

Fig. 1: Normal law fit to the height distribution of 1078 subjects (fathers) in Pearson's data set.

The Kalman filter is a method for the systematic estimation of means and variances. So we will begin by understanding random variables, means or expectations, and variances.

Figure 1 provides some motivation for our study of the normal law as well as means and variances. The figure has two items in it. The first item is a histogram. The heights of 1078 fathers (as well as their sons, but we ignore the sons) was recorded by the eminent statistician Karl Pearson. The minimum recorded height was 59 inches (which is an inch short fo 5 feet) and the maximum was 75.4 inches (which is 3.4 inches above 6 feet). To obtain this histogram, this range of heights is divided into 15 bins so that every bin is approximately an inch. The vertical axis of the figure shows the number of persons in each bin.

The second item in Figure 1 is the normal fit to the heights data. As you can see the fit is good but not very accurate. The convergence to the normal law is quite slow, typically at the rate $n^{-1/2}$ if the number of data points is n. We will understand Figure 1 better as we progress in this lecture, but as far as I am aware there are still many things in it that are yet not understood.

1.1 Random variables

The random variable is a mathematical abstraction of phenomena that we cannot calculate perfectly. The complete definition of a random variable is a little complicated. The difficulties in defining a random variable are related to the difficulties in defining a set axiomatically. We will not go into those difficulties. Instead, we assume that the notion of a random variable X is a given. That means that for any set A, we know the probability

$$\mathbb{P}\left(X \in A\right).$$

Certain random variables are discrete. That means they take values in a finite set $\{a_1, \ldots, a_M\}$ or a countable set. For example, X may take only the two values ± 1. In that case if we give the probabilities $\mathbb{P}(1)$ and $\mathbb{P}(-1)$ then we have specified the random variable completely. The probabilities must of course sum to 1.

Certain other random variables may be specified in terms of a probability density function $p(x)$, where $x \in \mathbb{R}$. In this case,

$$\mathbb{P}\left(a \leq X \leq b\right) = \int_a^b p(x)\, dx.$$

We require the normalization condition $\int_{-\infty}^{\infty} p(x)\, dx = 0$.

Notice that discrete random variables cannot be represented using densities. That problem can be dealt with fairly easily by passing to the Stieltjes integral, which is the next step from the Riemann integral. One can make up examples of random variables that cannot be represented using the Stieltjes integral. Modern definitions of random variables rely upon the Lebesgue theory and its extensions.

The probability density function $p(x, y)$ represents two random variables X, Y simultaneously. It is assumed to be continuous. We have

$$\begin{aligned}
\mathbb{P}\left(a \leq X \leq b,\, c \leq Y \leq d\right) &= \int_c^d \int_a^b p(x, y)\, dx\, dy \\
\mathbb{P}\left(a \leq X \leq b\right) &= \int_{-\infty}^{\infty} \int_a^b p(x, y)\, dx\, dy \\
\mathbb{P}\left(c \leq Y \leq c\right) &= \int_c^d \int_{-\infty}^{\infty} p(x, y)\, dx\, dy.
\end{aligned}$$

The normalization condition here is that $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x, y)\, dx\, dy = 1$. We can obtain the density functions of X and Y as

$$\begin{aligned}
p_X(x) &= \int_{-\infty}^{\infty} p(x, y)\, dy \\
p_Y(y) &= \int_{-\infty}^{\infty} p(x, y)\, dx.
\end{aligned}$$

In the same manner, we can consider joint density functions for any finite number of random variables.

The joint density function holds information about how the random variables are related to each other. If much of the mass of the density function $p(x, y)$ is concentrated near the diagonal $x = y$, then the random variables X and Y are highly correlated and take nearly the same value much of the time.

Independence is one of the key concepts. The random variables X and Y are independent if

$$\mathbb{P}\left(X \in A \,\text{and}\, Y \in B\right) = \mathbb{P}\left(x \in A\right) \mathbb{P}\left(y \in B\right),$$

where A and B are any two sets. Similarly, $X_1, \ldots, X_n$ are an independent sequence of random variables if

$$\mathbb{P}\left(X_1 \in A_1, \ldots, X_n \in A_n\right) = \prod_{i=1}^{n} \mathbb{P}\left(X_i \in A_i\right).$$

The random variables represented by the density function $p\left(x_1, \ldots, x_n\right)$ are independent if $p\left(x_1, \ldots, x_n\right) = p_1\left(x_1\right) p_2\left(x_2\right) \ldots p_n\left(x_n\right)$. They are independent and identically distributed if

$$p\left(x_1, \ldots, x_n\right) = p(x_1)p(x_2)\ldots p(x_n).$$

The definition of independence extends to discrete random variables as well.

1.2 Mean or expectation

If X is a discrete random variable that takes the values $a_1, \ldots, a_M$ with probabilities $p_1, \ldots, p_M$, respectively, then its mean is given by

$$\frac{p_1 a_1 + \cdots + p_M a_M}{M}.$$

In the mathematical literature, the mean is called expectation or expected value and denoted $\mathbb{E}X$. If the random variable is represented using the density function $p(x)$, then we have

$$\mathbb{E}X = \int_{-\infty}^{\infty} xp(x)\, dx.$$

If $f(\cdot)$ is a suitable function, then we have $\mathbb{E}f(X) = \int_{-\infty}^{\infty} f(x)\, p(x)\, dx$.

If $X_1, \ldots, X_n$ are a sequence of random variables then we may prove that

$$\mathbb{E}\left(X_1 + \cdots + X_n\right) = \mathbb{E}X_1 + \cdots + \mathbb{E}X_n.$$

The proof does not assume that the X_i are independent.

Exercise: Prove the above identity assuming the random variables $X_1, \ldots, X_n$ to be represented by the density function $p\left(x_1, \ldots, x_n\right)$.

Exercise: Let $\pi_1, \ldots, \pi_n$ be a random permutation of the numbers $1, \ldots, n$ which is equally likely to be any one of the $n!$ possible permutations. We say that the permutation has a descent at the position i if and only if $\pi_i > \pi_{i+1}$. A permutation may have d descents with $0 \leq d \leq n-1$.

- Find the permutation with 0 descents.
- Find the permutation with $n-1$ descents.
- Prove that the expected number of descents is $(n-1)/2$.

Exercise: Give an example of a random variable whose expectation is ∞. Give an example of a random variable whose expectation does not exist.

Exercise: If X and Y are independent random variables and may be represented using a density function, prove that $\mathbb{E}XY = \mathbb{E}X\,\mathbb{E}Y$.

Exercise: If X and Y are independent random variables and may be represented using a density function, If X and Y are independent random variables and may be represented using a density function, as above, and f and g are any continuous functions prove that

$$\mathbb{E}f(X)g(Y) = \mathbb{E}f(x)\,\mathbb{E}g(Y).$$

You may assume that all functions that arise in the calculation are integrable.

1.3 Variance and standard deviation X and Y are independent random variables and may be represented using a density function

If we know only the mean of a random variable, we know practically nothing about it. If we know its mean and its standard deviation, we have a great deal of information, as we will soon understand. What then is standard deviation?

The standard deviation, or the variance which is the square of the standard deviation, quantifies the typical size of fluctuations. If $\mathbb{E}X = \mu$, then the variance σ^2 of X is defined as

$$\sigma^2 = \mathbb{E}\left(X-\mu\right)^2.$$

The standard deviation is σ. It is often called the root-mean-square value since it is given by $\sqrt{\mathbb{E}\left(X-\mu\right)^2}$.

Exercise: Prove that $\sigma^2 = \mathbb{E}X^2 - \left(\mathbb{E}X\right)^2$.

Exercise: Give an example of a variable whose mean is finite and standard deviation infinite.

1.4 The normal distribution

The density function of the standard normal distribution is given by

$$\frac{1}{\sqrt{2\pi}}\exp\left(-x^2/2\right). \tag{1}$$

It is properly normalized because

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp(-x^2/2)\,dx = \frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty}\exp(-x^2)\,dx = 1.$$

This latter integral (called Gaussian integral) was first evaluated by Laplace. The popular method for evaluating it using double integrals and polar coordinates (see the wikipedia article `http://en.wikipedia.org/wiki/Gaussian_integral`) is due to Poisson and based on an earlier method of Laplace.

Given the Gaussian integral, it follows that

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 \exp(-x^2/2)\, dx = 1. \tag{2}$$

Since the mean of the standard normal distribution (1) is zero by symmetry, it follows that its variance and standard deviation are both 1.

Exercise: Prove that

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 \exp(-x^2/2)\, dx = 1.$$

Next consider the density function

$$\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\left(x-\mu\right)^2/2\sigma^2\right). \tag{3}$$

This is the density function of the normal distribution with mean μ and variance σ^2.

Exercise: Prove that the density function (3) is suitably normalized (it must integrate to 1 between $\pm\infty$. Prove that its mean is μ and variance σ^2.

Figure 1 was obtained as follows. After histogramming Pearson's height data, we considered the normal distribution (3) with mean $\mu = 67.78$ inches and standard deviation $\sigma = 2.744$ inches. The mean and variance match the Pearson data set. The graph of the density function (3) is scaled by a factor S to make the area under it equal to the area under the histogram.

Exercise: Argue that the area under a histogram with equal bins is equal to the total number of samples times the size of each bin.

2 A central limit theorem

If the mean μ and the variance σ^2 match those of the data, the normal distribution is a good fit to Pearson's heights data set (see Figure 1). Why does the normal distribution fit the height data set?

The answer to that question is not fully understood. In this section, we derive a central limit theorem. When a lot of random variables are added and then suitably normalized, the distribution of their sum tends to converge to the normal distribution. In this section, we will work through a single example to understand that phenomenon.

It used to be believed that the convergence to the normal distribution was a consequence of summing *independent* random variables. There are many, many examples where the sum of dependent random variables is known to converge to the normal distribution. Nevertheless the basic example we tackle assumes that random variables that are summed to be independent.

2.1 Statement of the theorem

Let X be a random variable with $\mathbb{P}(X = 1) = \mathbb{P}(X = -1) = 1/2$. By symmetry we have

$$\mu = \mathbb{E}X = 0. \tag{4}$$

In addition,

$$\sigma^2 = \mathbb{E}(X - \mu)^2 = \mathbb{E}X^2 = 1. \tag{5}$$

The random variable X has mean 0 and variance 1.

The distribution of the random variable X is very far from being normal. However, we will see that summing many random variables with same distribution as X gives the normal distribution.

Let $X_1, \ldots, X_n$ be a sequence of independent random variables each of which, like X, is equally likely to be $+1$ or -1. Consider the sum

$$S_n = X_1 + \cdots + X_n.$$

We begin by calculating the mean and variance of S_n.

We have

$$\mathbb{E}S_n = \mathbb{E}X_1 + \cdots + \mathbb{E}X_n = 0.$$

The variance calculation is similar.

$$\begin{aligned} \mathbb{E}S_n^2 &= \mathbb{E}\left(X_1 + \cdots + X_n\right)^2 \\ &= \sum_{i=1}^{n} \mathbb{E}X_i^2 + 2 \sum_{1 \le i < j < n} \mathbb{E}X_i X_j \\ &= \sum_{i=1}^{n} \mathbb{E}X_i^2 + 2 \sum_{1 \le i < j \le n} \mathbb{E}X_i \, \mathbb{E}X_j \\ &= n. \end{aligned}$$

The third equality above is justified because the X_i are independent. The final equality follows from (4) and (5).

So the sum S_n has mean 0, variance n, and standard deviation $\sqrt{n}$. The standard deviation is a measure of the order of fluctuations. If n independent and identical random variables are added, the standard deviation of their sum grows only like $\sqrt{n}$.

The random variable $S_n/\sqrt{n}$ has mean 0 and variance 1. It converges in distribution to the normal law.

Theorem 2.1. $\lim_{n\to\infty} \mathbb{P}\left(a \le \frac{S_n}{\sqrt{n}} \le b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b \exp(-x^2/2)\, dx.$

This central limit theorem makes precise the notion that the order of fluctuations of S_n is the same as its standard deviation.

2.2 Calculation of $\mathbb{E}\exp(tX)$

Let t be any number real or complex. We define the random variable $\exp(tX)$ using the series expansion:

$$\exp(tX) = 1 + tX + \frac{t^2X^2}{2!} + \frac{t^3X^3}{3!} + \cdots$$

Here t may be any real or complex number.

Since $\mathbb{P}(X = 1) = \mathbb{P}(X = -1) = 1/2$, we have

$$\mathbb{E}X^n = \begin{cases} 0 & \text{if } n\text{is odd} \\ 1 & \text{if } n\text{is even.} \end{cases}$$

Therefore

$$\begin{aligned} \mathbb{E}\exp(tX) &= 1 + \frac{t^2}{2!} + \frac{t^4}{4!} + \frac{t^6}{6!} + \cdots \\ &= \cos(it). \end{aligned}$$

Exercise: Use the power series of $\cos x$ to prove that

$$\cos it = 1 + \sum_{n=1}^{\infty} \frac{t^{2n}}{2n!}.$$

This calculation of the expectation$\mathbb{E}\exp(tX)$ is the basis of our proof of the central limit theorem.

2.3 Calculation of $\mathbb{E}\exp(S_n/\sqrt{n})$

Next we turn to the calculation of the expectation of the random variable

$$\frac{S_n}{\sqrt{n}} = \frac{X_1 + \cdots + X_n}{\sqrt{n}}.$$

When we see a sum like this it appears strange to divide by $\sqrt{n}$ instead of by n. Of course, division by $\sqrt{n}$ is justified by the variance calculation of S_n and normalizing the variance to 1 is key to the central limit theorem.

Since the X_i are independent, we have

$$\begin{aligned} \mathbb{E}\exp\left(\frac{S_n}{\sqrt{n}}\right) &= \mathbb{E}\exp\left(\frac{X_1}{\sqrt{n}}\right).\exp\left(\frac{X_2}{\sqrt{n}}\right)\ldots\exp\left(\frac{X_n}{\sqrt{n}}\right) \\ &= \mathbb{E}\exp\left(\frac{X_1}{\sqrt{n}}\right).\mathbb{E}\exp\left(\frac{X_2}{\sqrt{n}}\right)\ldots\mathbb{E}\exp\left(\frac{X_n}{\sqrt{n}}\right) \\ &= \left(\cos\left(\frac{it}{\sqrt{n}}\right)\right)^n \end{aligned}$$

2.4 Towards the central limit theorem

Let $\mathcal{N}$ be the normal variable with mean 0 and variance 1 (the standard normal distribution). Then

$$\begin{aligned}\mathbb{E}\exp(t\mathcal{N}) &= \frac{1}{2\sqrt{\pi}}\int_{-\infty}^{\infty} {}^{tx-x^2/2}\,dx \\ &= \frac{{}^{t^2/2}}{\sqrt{2\pi}}\int_{-\infty}^{\infty} {}^{-(x-t)^2/2}\,dx \\ &= \exp(t^2/2) \quad \text{after the substitution } u = x - t.\end{aligned}$$

Theorem 2.2. $\lim_{n\to\infty}\mathbb{E}\exp(S_n/\sqrt{n}) = \exp(t^2/2)$.

Proof. We need to show that

$$\lim_{n\to\infty}\left(\cos\left(\frac{it}{\sqrt{n}}\right)\right)^n = \exp(t^2/2).$$

This limit may be evaluated by noting that $\cos(it/\sqrt{n}) \approx 1 + t^2/n$ for large n and then using

$$\begin{aligned}\lim_{n\to\infty}\left(1+\frac{t^2}{n}\right)^n &= \lim\left(\left(1+\frac{t^2}{n}\right)^{\frac{n}{t^2}}\right)^{t^2} \\ &= \left(\lim_{x\to\infty}\left(1+\frac{1}{x}\right)^x\right)^{t^2} \\ &= {}^{t^2/2}\end{aligned}$$

The last limit with $x \to \infty$ is one of the standard ways to define the number .

Another proof would be to take the log and then use l'Hospital's rule. □

We are now almost on the cusp of completing the proof of the central limit theorem 2.1. We know that

$$\lim_{n\to\infty}\mathbb{E}\exp\left(\frac{tS_n}{\sqrt{n}}\right) = \exp(t^2/2) = \mathbb{E}\exp(t\mathcal{N}),$$

where $\mathcal{N}$ is a random variable of the standard normal distribution. This is true for any t real or complex.

To complete the proof we may let $t = i\omega$ to be purely imaginary and then use the theory of Fourier transforms. It is not right to sneak in a major idea like Fourier transforms in the middle of some other discussion. Therefore we leave out this part of the proof.

Exercise: Let X take the value $+1$ with probability p and the value -1 with probability $1-p$. Calculate the mean μ and variance σ^2 of X.

Exercise: Let $X_1, \ldots, X_n$ be a sequence of independent random variables of the same distribution as X in the previous exercise. Let $S_n = X_1 + \cdots + X_n$. Prove that

$$\lim_{n\to\infty}\mathbb{E}\exp t\left(\frac{S_n - n\mu}{\sqrt{n}\sigma}\right) = \exp(t^2/2).$$

3 The Kalman filter