Introduction to Probability: Expectations, Bayes Theorem, Gaussians, and the Poisson Distribution.[1]

Pankaj Mehta

September 6, 2021

In this worksheet, we will review some basic concepts in probability theory. We will start with some basic ideas about probability and define various moments and cumulants. We will then discuss how moments and cumulants can be calculated using generating functions and cumulant functions. We will then focus on two widely occurring "universal distributions": the Gaussian and the Poisson distribution.

[1] Read: This will introduce some elementary ideas in probability theory that we will make use of repeatedly.

Probability distributions of one variable

Consider a probability distribution $p(x)$ of a single variable where X can take discrete values in some set $x \in \mathcal{A}$ or continuous values over the real numbers[2]. We can define the expectation value of a function of this random variable $f(X)$ as

$$\langle f(X) \rangle = \sum_{x \in \mathcal{A}} p(x) \tag{1}$$

for the discrete case and

$$\langle f(X) \rangle = \int dx f(x) p(x), \tag{2}$$

in the continuous case. For simplicity, we will write things assuming x is discrete but it is straight forward to generalize to the continuous case by replacing by integrals.

We can define the n-th *moment* of x as

$$\langle X^n \rangle = \sum_x p(x) x^n. \tag{3}$$

The first moment is often called the mean and we will denote it as $\langle x \rangle = \mu_x$. We can also define the n-th *central* of x as

$$\langle (X - \langle X \rangle)^n \rangle = \sum_x p(x)(x - \langle x \rangle)^n. \tag{4}$$

The second central moments is often called the *variance* of the distribution and in many cases "characterizes" the typical width of the distribution. We will often denote the variance of a distribution by $\sigma_x^2 = \langle (X - \langle X \rangle)^2$.

- **Exercise:** Show that the second central moment can also be written as $\langle X^2 \rangle - \langle X \rangle^2$.

[2] We will try to denote random variables by capital letters and the values they can take by the corresponding lower case letter.

Example: Binomial Distribution

Consider the Binomial distribution which describes the probability of getting n heads if one tosses a coin N times. Denote this random variable by X. Note that X can take on $N + 1$ values $n = 0, 1, \ldots, N$. If probability of getting a head on a coin toss is p, then it is clear that

$$p(n) = \binom{N}{n} p^n (1 - p)^{N-n} \tag{5}$$

Let us calculate the mean of this distribution. To do so, it is useful to define $q = (1 - p)$.

$$\begin{aligned} \langle X \rangle &= \sum_{n=0}^{N} np(n) \\ &= \sum_{n} n \binom{N}{n} p^n q^{N-n} \end{aligned}$$

Now we will use a very clever trick. We can formally treat the above expression as function of two independent variables p and q. Let us define

$$G(p,q) = \sum_{n} \binom{N}{n} p^n q^{(N-n)} = (p+q)^N \tag{6}$$

and take a partial derivative with respect to p. Formally, we know the above is the same as

$$\begin{aligned} \langle X \rangle &= \sum_{n} n \binom{N}{n} p^n q^{N-n} \\ &= p\partial_p G(p,q) \\ &= p\partial_p (p+q)^N \\ &= pN(p+q)^{N-1} = pN, \end{aligned} \tag{7}$$

where in the last line we have substituted $q = 1 - p$. This is exactly what we expect [3]. The mean number of heads is exactly equal to the number of times we toss the coin times the number probability we get a heads.

[3] When one first encounters this kind of formal trick, it seems like magic and cheating. However, as long as the probability distributions converge so that we can interchange sums/integrals with derivatives there is nothing wrong with this as strange as that seems.

Exercise: Show that the variance of the Binomial distribution is just $\sigma_n^2 = Np(1-p)$.

Probability distribution of multiple variables

In general, we will also consider probability distributions of M variables $p(X_1, \ldots, X_M)$. We can get the *marginal distributions* of a single variable by integrating (summing over) all other variables [4] We have that

$$p(X_j) = \int dx_1 \ldots dx_{j-1} dx_{j+1} \ldots dx_M p(x_1, x_2, \ldots, x_M). \tag{8}$$

[4] In this section, we will use continuous notation as it is easier and more compact. Discrete variables can be treated by replacing integrals with sums.

Similarly, we can define the marginal distribution over any subset of variables by integrating over the remaining variables.

We can also define the *conditional probability* of a variable x_i which encodes the probability that variable X_i takes on a give value given the values of all the other random variables. We denote this probability by $p(X_i|X_1, \ldots X_{i-1}, X_{i+1}, \ldots, X_M)$.

One important formula that we will make use of repeatedly is Bayes formula which relates marginals and conditionals to the full probability distribution

$$p(X_1, \ldots, X_M) = p(X_i|x_1, \ldots X_{i-1}, X_{i+1}, \ldots, X_M)p(X_1, \ldots X_{i-1}, X_{i+1}, \ldots, X_M). \tag{9}$$

For the special case of two variables, this reduces to the formula

$$p(X, Y) = p(X|Y)p(Y) = p(Y|X)p(Y). \tag{10}$$

This is one of the fundamental results in probability and we will make use of it extensively when discussing inference in our discussion of early vision.[5]

[5] Bayes formula is crucial to understanding probability and I hope you have encountered it before. If you have not, please do some exercises that use Bayes formula that are everywhere on the web.

Exercise: Understanding test results.[6]. Consider a mammogram to diagnose breast cancer. Consider the following facts (the numbers aren't exact but reasonable).

- 1% of women have breast cancer.
- Mammograms miss 20% of cancers when that are there.
- 10% of mammograms detect cancer even when it is *not* there.

Suppose you get a positive test result, what are the chances you have cancer?

[6] This is from the nice website `https://betterexplained.com/articles/an-intuitive-and-short-explanation-of-bayes-theore`

Exercise: Consider a distribution of two variables $p(x, y)$. Define a variable $Z = X + Y$ [7]

- Show that $\langle Z \rangle = \langle X \rangle + \langle Y \rangle$

- Show that $\sigma_z^2 = \langle Z^2 \rangle - \langle Z \rangle^2 = \sigma_x^2 + \sigma_y^2 + 2Cov(X, Y)$ where we have defined the covariance of X and Y as $Cov(x, y) = \langle (X - \langle X \rangle)(Y - \langle Y \rangle) \rangle$.

[7] We will make use of these results repeatedly. Please make sure you can derive them and understand them thoroughly.

- Show that if x and y are independent variables than $Cov(x, y) = 0$ and $\sigma_z^2 = \sigma_x^2 + \sigma_y^2$.

- Not let $\langle Z \rangle = aX + bZ$ with a and b constants. Calculate $\langle Z \rangle$ and σ_z^2.1

Law of Total Variance

This section is optional but this is often a result that is missing from physics education and worth knowing. An important result often used in statistics is the law of total variance. Consider two random variables X and Y. We would like to know how much the variance of Y is explained by the value of X. In other words, we would like to decompose the variance of Y into two terms: one term that captures how much of the variance of Y is "explained" by X and another term that captures the "unexplained variance".

It will be helpful to define some notation to make the expectation value clearer

$$E_X[f(X)] = \langle X \rangle_x = \int dx p(x) f(x) \tag{11}$$

Notice that we can define the *conditional expectation* of a function Y given that $X = x$ by

$$E_Y[f(Y)|X = x] = \int dy p(y|X = x) f(y). \tag{12}$$

For the special case when $f(Y) = Y^2$ we can define the *conditional variance* which measures how much does Y vary is we fix the value of X

$$Var_Y[Y|X = x] = E_Y[Y^2|X = x] - E_Y[Y|X = x]^2 \tag{13}$$

We can also define a different variance that measures the variance of the $E_Y[Y|X]$ viewed as a random function of X:

$$Var_X[E_Y[Y|X]] = E_X[E_Y[Y|X]^2] - E_X[E_Y[Y|X]]^2. \tag{14}$$

This is a well defined probability distribution since $p(y|x)$ is the marignal probability distribution over y. Notice now, that using Bayes rule we can write

$$\begin{aligned} E_Y[f(y)] &= \int dy p(y) f(y) = \int dx dy p(y|x) p(x) f(y) \\ &= \int dx p(x) \int dy p(y|x) f(y) = E_X[E_Y[f(Y)|X]] \end{aligned} \tag{15}$$

So now notice that we can write

$$\begin{aligned}\sigma_Y^2 &= E_Y[Y^2] - E_Y[Y]^2 \\ &= E_X[E_Y[Y^2|X]] - E_X[E_Y[Y|X]]^2 \\ &= E_X[E_Y[Y^2|X]] - E_X[E_Y[Y|X]^2] + E_X[E_Y[Y|X]^2] - E_X[E_Y[Y|X]]^2 \\ &= E_X[E_Y[Y^2|X] - E_Y[Y|X]^2] + Var_X[E_Y[Y|X]] \\ &= E_X[Var_Y[Y|X]] + Var_X[E_Y[Y|X]] \end{aligned} \tag{16}$$

The first term is often called the unexplained variance and the second term is the explained variance.

To see why, consider the case where the random variable Y are related by a linear function plus noise $Y = aX + \epsilon$, where ϵ is a third random variable with mean $\langle \epsilon \rangle = b$ and $\langle \epsilon^2 \rangle = \sigma_\epsilon^2 + \mu^2$. Furthermore, let $E_X[X] = \mu_x$. Then notice, that in this case all the randomness of Y is contained in ϵ. Notice that

$$E[Y] = a\mu_x + b \tag{17}$$

and

$$E_\epsilon[Y|X] = aX + E_\epsilon[\epsilon] = aX + b \tag{18}$$

We can also write

$$\begin{aligned} Var_X[E_Y[Y|X]] &= E_X[E_Y[Y|X]^2] - E_X[E_Y[Y|X]]^2 \\ &= E_X[(aX+b)^2] - (a\mu_x + b)^2 \\ &= a^2 E_X[X^2] + 2ab\mu_x + b^2 - (a\mu_x + b)^2 \\ &= a^2\sigma_X^2 \end{aligned} \tag{19}$$

This is clearly the variance in Y "explained" by the variance in X. Using the law of total variance, we see that in this case

$$E_X[Var_Y[Y|X]] = \sigma_Y^2 - a^2\sigma_X^2 \tag{20}$$

is the unexplained variance.

Generating Functions and Cumulant Generating Functions

We are now in the position to introduce some more machinery. These are the moment generating functions and cumulant generating functions. Consider a probability distribution $p(x)$. We would like an easy and efficient way to calculate the moments of this distribution. It turns out that there is an easy way to do this using ideas that will be familiar from Statistical Mechanics.

Generating Functions

Given a discrete probability distribution, $p(X)$ we can define an moment-generating function $G(t)$ for the probability distribution as

$$G(t) = \langle e^{tX} \rangle = \int dx p(x) e^{tx} \tag{21}$$

Alternatively, we can also define an (ordinary) generating function for the distributions

$$G(z) = \langle Z^X \rangle = \sum_x p(x) z^x \tag{22}$$

If the probability distribution is continuous, we define the moment-generating function as [8]

$$G(t) = \langle e^{-tX} \rangle = \int dx p(x) e^{tx}. \tag{23}$$

[8] If x is defined over the reals, this is just the Laplace transform.

Notice that $G(t = 0) = G(z = 1) = 1$ since probability distributions are normalized.

These functions have some really nice properties. Notice that the $n - th$ moment of X can be calculated by taking the n-th partial derivative of the moment generating function evaluated at $t = 0$:

$$\langle X^N \rangle = \frac{\partial^n G(t)}{\partial t^n}|_{t=0} \tag{24}$$

Alternatively, we can write in terms of the ordinary generating function

$$\langle X^N \rangle = (z\partial_z)^n G(z)|_{z=0} \tag{25}$$

where $(z\partial_z)^n$ means apply this operator n times.

Example: Generating Function of the Binomial Distribution

Let us return to the Binomial distribution above.The ordinary generating function is given by

$$G(z) = \sum_n \binom{N}{n} p^n q^{N-n} z^n = (pz + q)^N = (pz + (1 - p))^N \tag{26}$$

Let us calculate the mean using this

$$\mu_X = z\partial_z G(z) = [zNp(pz + (1 - p))^{N-1}]|_{z=1} = Np \tag{27}$$

We can also calculate the second moment

$$\begin{aligned} \langle X^2 \rangle &= z\partial_z(z\partial_z G(z)) \\ &= [zNp(pz + (1 - p))^{N-1}]|_{z=1} + \partial_z[Np(pz + (1 - p))^{N-1}]|_{z=1} \\ &= Np + N(N - 1)p^2 = Np(1 - p) + N^2p^2 \end{aligned} \tag{28}$$

and the variance

$$\sigma_X^2 = \langle X^2 \rangle - \langle X \rangle^2 = Np(1-p). \tag{29}$$

Example: Normal Distribution

We now consider a Normal Random Variable X whose probability density is given by

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{\sigma^2}}. \tag{30}$$

Let us calculate the moment generating function

$$G(t) = \int dx \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{\sigma^2}} e^{tx}. \tag{31}$$

It is useful to define a new variable $u = (x-\mu)/\sigma$ [9]. In terms of this variable we have that

[9] Notice that u is just the "z-score" in statistics.

$$\begin{aligned} G(t) &= \frac{1}{\sqrt{2\pi}} \int du e^{-u^2/2 + t(\sigma u + \mu)} \\ &= e^{\mu t + t^2\sigma^2/2} \frac{1}{\sqrt{2\pi}} \int du e^{-\frac{(u-\sigma u/2)^2}{2}} \\ &= e^{\mu t + t^2\sigma^2/2} \end{aligned} \tag{32}$$

Exercise: Show that the $n-th$ moment of a Gaussian with mean zero ($\mu = 0$) can be written as

$$\langle X^p \rangle = \begin{cases} 0 & \text{if p is odd} \\ (p-1)!\sigma^p & \text{if p is even} \end{cases} \tag{33}$$

This is one of the most important properties of the Gaussian/Normal distribution. All higher order moments can be expressed solely in terms of the mean and variance! We will make use of this again.

The Cumulant Generating Function

Another function we can define is the Cumulant generating function of a probability distribution

$$K(t) = \log \langle e^{tX} \rangle = \log G(t). \tag{34}$$

The $n-th$ derivative of $K(t)$ evaluated at $t = 0$ is called the n-th cumulant:

$$\kappa_n = K^{(n)}(0). \tag{35}$$

Let us not look the first few cumulants. Notice that

$$\kappa_1 = \frac{G'(0)}{G(0)} = \frac{\mu_X}{1} = \mu_X. \tag{36}$$

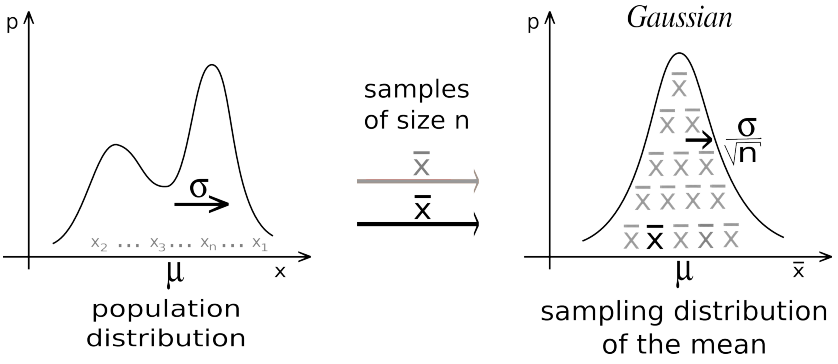

Figure 1: "Whatever the form of the population distribution, the sampling distribution tends to a Gaussian, and its dispersion is given by the Central Limit Theorem" (figure and caption from Wikipedia)

Similarly, it is easy to show that the second cumulant is just the variance:

$$\kappa_2 = \frac{G''(0)}{G(0)} - \frac{[G'(0)]^2}{G(0)^2} = \langle X^2 \rangle - \langle X \rangle^2 = \sigma_X^2 \tag{37}$$

Notice that the cumulant generating function is just the Free Energy in statistical mechanics and the moment-generating function is the partition function.

Central Limit Theorem and the Normal Distribution

One of the most ubiquitous distributions that we see in all of physics, statistics, and biology is the Normal distribution. This is because of the Central Limit theorem. Suppose that we draw some random variables $X_1, X_2, \ldots, X_N$ identically and independently from a distribution with mean μ and variance σ^2. Let us define a variable

$$S_N = \frac{X_1 + X_2 + \ldots + X_N}{N}. \tag{38}$$

Then as $N \to \infty$, the distribution of S_N is well described by a Gaussian/Normal distribution with mean μ and variance σ^2/N. We will denote such a normal distribution by $\mathcal{N}(\mu, \sigma^2/N)$. This is true regardless of the original distribution (see Figure 1). The main thing we should take away is that the variance decreases as $1/N$ with the number of samples N that we take!

Example: The Binomial Distribution We can ask how the sample mean changes with the number of samples for a Bernoulli variable with probability $p = 1/2$ for various N. This is shown in Figure 2 from Wikipedia.

The Poisson Distribution

There is a second "universal" distribution that occurs often in Biology. This is the distribution that describes the number of rare events

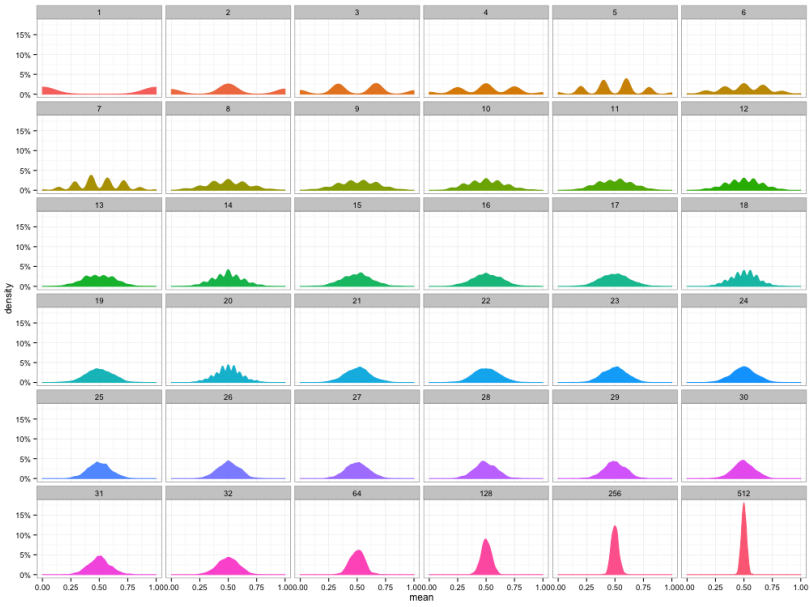

Figure 2: Means from various N samples drawn from Bernoulli distribution with $p = 1/2$. Notice it converges more and more to a Gaussian. (figure and caption from Wikipedia)

we expect in some time interval T. The Poisson distribution is applicable if the following assumptions hold:

- The number of events is discrete $k = 0, 1, 2, \ldots$.
- The events are independent, cannot occur simultaneously, occur at a constant rate per unit time r.
- Generally the number of trials is large and probability of success is small.

Examples of things that can be described by the Poisson distribution include:

- Photons arriving in a microscope (especially at low intensity)
- The number of mutations per unit length of DNA
- The number of nuclear decays in a time interval
- "The number of soldiers killed by horse-kicks each year in each corps in the Prussian cavalry. This example was made famous by a book of Ladislaus Josephovich Bortkiewicz (1868-1931) "(from Wikipedia)
- "The targeting of V-1 flying bombs on London during World War II investigated by R. D. Clarke in 1946".

Let us denote the mean number of events that occur in a time t by $\lambda = rT$. Then, the probability that k events occurs is given by

$$p(k) = e^{-\lambda} \frac{\lambda^k}{k!}. \tag{39}$$

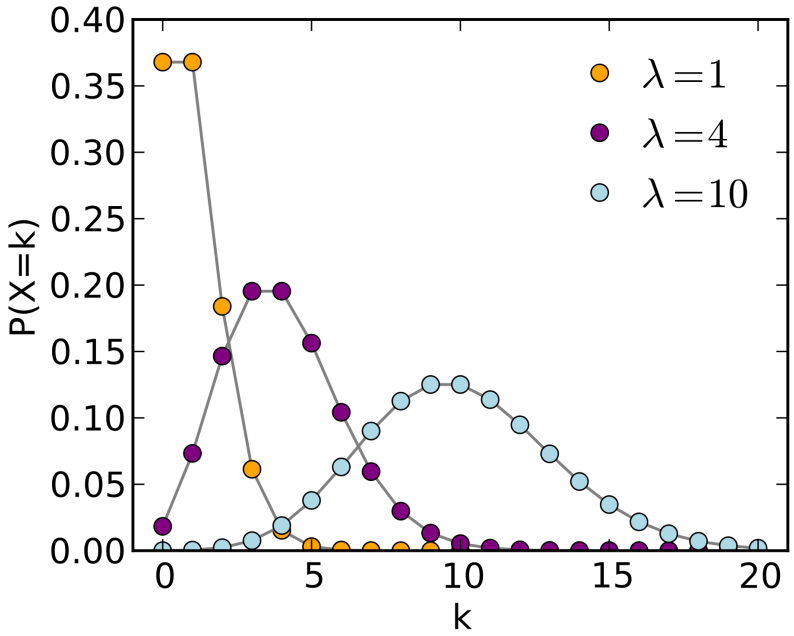

Figure 3: Examples of Poisson distribution for different means.. (figure from Wikipedia)

It is easy to check that $\sum_k p(k) = 1$ using the Taylor series of the exponential. We can also calculate the mean and variance. Notice we can define the generating function of the Poisson as

$$G(t) = \sum_z e^{-\lambda} \frac{e^{tk}(\lambda)^k}{k!} = e^{\lambda(1+e^t)}. \tag{40}$$

The cumulant-generating function is just

$$K(t) = \log G(t) = \lambda(1 + e^t) \tag{41}$$

From this we can easily get the all cumulants since this is just differentiating the expression above n times

$$\kappa_n = K^{(n)}(0) = \lambda. \tag{42}$$

In other words, all the higher order cumulants of the Poisson distribution are the same and equal to the mean.

Another defining feature of the Poisson distribution is that it has "no memory". Since things happen at a constant rate, there is no memory. We will see that in general to create non-Poisson distributions (with memory), we will have to burn energy! More on this cryptic statement later.