Probability Distribution Workshop

Los Angeles County
ISAB

1 Introduction

This workshop is a review of common probability distributions. It has two principal goals.

1. Serve as a review for people that have not studied this material for a while.
2. Provide exercise for people to familiarize themselves with their Python and R distributions.

The exercises themselves are detailed in *Jupyter notebooks* for Python and *R markdown* for R. This mathematical supplement serves to provide some theoretical motivation for functions encountered in the Python and R libraries.

1.1 Availability

The source code for this document along with the Python and R exercises are available from the ISAB GitHub repository.

https://github.com/lacounty-isab/workshops/tree/master/distributions

1.2 Who are We?

ISAB (Information Systems Advisory Body) is a subcommittee of the CCJCC (Countywide Criminal Justice Coordination Committee) of Los Angeles County, California. ISAB is a multi-jurisdictional organization serving the justice communities within the county. The ISAB Data Science Committee (IDSC) addresses data science issues faced by the community. This article addresses skill development. It is not intended as an endorsement of any one product or technology.

More details on ISAB and CCJCC can be found on their websites.

ISAB http://ccjcc.lacounty.gov/Subcommittees-Task-Forces/Information-Systems-Advisory-Board-ISAB

CCJCC http://ccjcc.lacounty.gov/

Random Variables

Since this article is not intended to be a foundational exposition on probability theory, we'll jump right in with a review of random variables. Recall that an event space Ω is a set of possible outcomes for an experiment. In the case of flipping a single coin, we would have

$$\Omega = \{\text{heads}, \text{tails}\}$$

A random variable X is an assignment from the event space to the real numbers.

$$X : \Omega \to \mathbb{R}$$

An example for the case of a coin flip might be

$$\begin{aligned} X(\text{tails}) &= 0 \\ X(\text{heads}) &= 1 \end{aligned}$$

A less trivial example would be flipping a coin 100 times where the outcome is the number of heads. There are many random varialbes you could define on this event space.

- the number of heads,
- the number of tails,
- 0 if even number of heads, 1 otherwise,
- greatest number consecutive heads,
- number of heads minus number of tails.

There are many more examples. Usually the mapping is straight forward. In fact, many times the mapping is so straight forward that we forget that the event and the random variable are not the same thing!

The goal of the random variable is to convert the experiment outcomes from an abstract "set of things" to a set of numbers. This isn't hard when the set of things naturally maps to a set of numbers in a useful way. When it's not so natural, we sometimes have to go back to the definition the right mapping. Ultimately the goal is to get a set of number with which we can analyze the experiment quantitatively.

2 Distributions

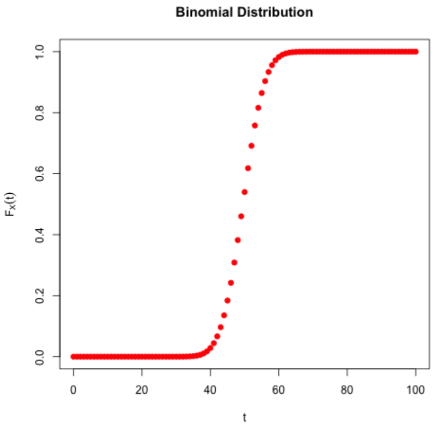
Once we have a random variable defined for an experiment that maps outcomes to real numbers, the next question is to ask how the values of the random variable are likely to be distributed along the real axis after performing the experiment many times. A central concept to answering this question is the *distribution* or *cumulative distribution.*

$$F_X(t) = P(X \leq t)$$

If we think back to the example in the previous section of flipping a coin 100 times where the random variable X represented the number of heads, then $F_X(t)$ represents the probability that the number of heads is less than or equal t. Some obvious values are

$$\begin{aligned} F_X(-1) &= 0 \\ F_X(100) &= 1 \end{aligned}$$

The first equality follows from the fact we can't have a negative count of the number of heads. The second equality follows from the fact that we can't encounter more than 100 heads if we only flip the coin 100 times. A graph for all values of t is shown below.



The function F_X is often called the *cumulative distribution* since it represents the accumulation of probability as t covers more of the real axis. In the figure above for a binomial cumulative distribution for 100 trials of a fair coin, we can see that probability of observing less than 40 heads is close to zero. The probability of observing less than 60 is close to one. That tells us that the most likely scenarios are between 40 and 60.

2.1 Discrete Densities

Discrete distributions describe experiments where the random variable takes on discrete values. These are often associated with counts. Their notation will often be written as $P(X \leq n)$ to emphasize the discrete character of n. (Note that n doesn't have to be an integer.)

A *discrete density* f_X corresponding to a discrete distribution F_X assigns a probability to each discrete value of the random variable. If the random variable only takes integer values, then the density and the distribution are related in the following ways.

$$\begin{aligned} f(n) &= F(n) - F(n-1) \\ F(n) &= \sum_{i \leq n} f(i) \\ \sum_{i=-\infty}^{\infty} f(i) &= 1 \end{aligned}$$

The X subscripts were dropped for brevity; but in general one writes F_X or f_X to associate the function with the associated random variable X. This becomes more important when multiple random variables are considered.

Many useful quantities can be expressed in terms of a density. First, there is the *mean*.

$$E[X] = \sum_{i=-\infty}^{\infty} i \cdot f_X(i) \tag{1}$$

The variance is defined in terms of the mean as

$$\mathrm{Var}[X] = E\left[(X - E[X])^2\right] \tag{2}$$

In terms of a density function the variance is

$$\mathrm{Var}[X] = E\left[(X - E[X])^2\right] = \sum_{i=-\infty}^{\infty} (i - E(X))^2 \cdot f_X(i) \tag{3}$$

A popular alternative expression for the variance can be obtained by expanding the square.

$$\begin{aligned}
\mathrm{Var}[X] &= \sum_{i=-\infty}^{\infty} (i - E[X])^2 \cdot f_X(i) \\
&= \sum_{i=-\infty}^{\infty} (i^2 - 2iE[X] + E[X]^2) \cdot f_X(i) \\
&= \sum_{i=-\infty}^{\infty} i^2 \cdot f_X(i) - 2E[X] \sum_{i=-\infty}^{\infty} i \cdot f_X(i) - E[X]^2 \sum_{i=-\infty}^{\infty} f_X(i) \\
&= E[X^2] - 2E[X]E[X] - E[X]^2 \\
&= E[X^2] - E[X]^2 \qquad (4)
\end{aligned}$$

This is often expressed in words as “the mean squared minus the square of the mean.” In the special case where the mean is zero, the variance is equal the mean squared.

The mean is a special cases of a *moments*. The n^{th} moment is defined as

$$E[X^n] = \sum_{i=-\infty}^{\infty} i^n \cdot f_X(i)$$

We’ve already seen that the first moment is the mean. The second moment can be used to determine the variance. Higher moments are not discussed as often, but they still come in handy. The third moment is related to the *skew*, which describes the degree to which a distribution is lopsided with respect to its mean. The fourth moment is related to the *kurtosis*. It describes the extent to which a distribution avoids its mean.

An important tool used in working with distributions is the *moment generating function*. For a discrete random variable, it’s defined as

$$m_X(t) = E(e^{it}) = \sum_{i=-\infty}^{\infty} e^{it} f_X(i) \qquad (5)$$

At first glance it looks like way more trouble than it could possibly be worth. But moment generating functions turn out to be important both practically and theoretically.

The theoretical importance derives from analytic function theory which describes a class of functions that are completely determined in some neighborhood of a point by the derivatives of all orders at the point. Moment generating functions are used to show how, in most cases, a distribution is uniquely determined by the value of all its moments. This result is encountered in many proofs to show how a particular distribution is, in fact, equal to a known distribution.

The practical use is that moment generating functions often simplify the symbolic calculation of the mean and variance for many distributions. The results we need for the mean and variance are the following.

Given $m_X(t)$ as described in (5), the first and second moments of X are, respectively,

$$\begin{aligned} E(X) &= m'_X(0) & (6) \\ E(X^2) &= m''_X(0) & (7) \end{aligned}$$

It's not immediately obvious how this makes things any easier. The examples below will bear it out.

2.1.1 Binomial Distribution

We've been using the binomial distribution as an example for much of the introduction. Recall that the experiment is that we perform a sequence of n independent Bernoulli trials, each of which has probability p of being successful. The outcome is the number of successful trials. Let's consider what the binomial density looks like.

Let i be the number of successes. Since i is a count between zero and n, i cannot be less than zero or greater than n. So

$$p(i) = 0, i \notin 0, 1, 2, \ldots n$$

For $i \in 0 \ldots n$ there are i successes and $n - i$ failures. The probability of such a sequence is $i^p \cdot (n-i)^{(1-p)}$ since each outcome is independent of the one before it. We then have to account for the number of positions the i occurrences could have occurred among the n trials. There are n ways to pick the first one. For each one of these, there are $n - 1$ ways to pick the next one. This continues until we have picked i times down to $n - i + 1$. This leads to the following expression with i factors for the i chosen positions.

$$n \cdot n - 1 \cdots n - i + 1$$

But this product distinguishes between the order of the successes. For example, if the first two are successes, it counts first then second separately from second and then first. The order doesn't matter. So we divide by the number of permutations of i elements, which is $i!$: i ways to choose the first one, $i - 1$ ways to choose the second one, down to 1. So the number of ways we can choose i elements from n elements without replacement divided by the number of ways we could have ordered i elements, we get

$$\begin{aligned}
\frac{n(n-1)\cdots(n-i+1)}{i!} &= \frac{n(n-1)\cdots(n-i+1)}{i!}\cdot\frac{n-i}{n-i}\cdot\frac{n-i-1}{n-i-1}\cdots\frac{2}{2}\cdot\frac{1}{1} \\
&= \frac{n!}{i!(n-i)!} \\
&= \binom{n}{i}
\end{aligned}$$

The notation in the last expression is read "n choose i".

Whew!

So a particular occurrence of i successes and $n-i$ failures occurs with probability $p^i\cdot(1-p)^{n-i}$. Since there are $\binom{n}{i}$ ways this can happen, the probability this can happen in *some way* (whichever order, we don't care) is

$$f(i)=\binom{n}{i}p^i(1-p)^{n-i}, i\in 0,\ldots,n \tag{8}$$

This is a probability, the sum of all possible values should add to one. Using the binomial formula, we have

$$\begin{aligned}
\sum_{i=-\infty}^{\infty} f(i) &= \sum_{i=0}^{n} f(i) \\
&= \sum_{i=0}^{n}\binom{n}{i}p^i(1-p)^{n-i} \\
&= [p+(1-p)]^2 \\
&= 1^2
\end{aligned}$$

Great! The probabilities sum to one. Now let's evaluate the mean and variance of this distribution. From (1) we have to evaluate

$$\sum_{i=0}^{n} i\binom{n}{i}p^i(1-p)^{n-i}$$

This is not easy to solve in a closed form. It's one of those times where we appeal to the moment generating function.

$$m_X(t) = E(e^{it})$$

$$
\begin{aligned}
&= \sum_{i=0}^{n} e^{it} \binom{n}{i} p^i (1-p)^{(n-i)} \\
&= \sum_{i=0}^{n} \binom{n}{i} (pe^t)^i (1-p)^{(n-i)} \\
&= (pe^t + 1 - p)^n \\
&= (pe^t + q)^n
\end{aligned}
$$

In the last step, $1 - p$ is replaced with q. We just remember that $p + q = 1$. Now differentiate the moment generating function twice with respect to t.

$$
\begin{aligned}
m'_X(t) &= n(pe^t + q)^{(n-1)} pe^t \\
m''_X(t) &= n(n-1)(pe^t + q)^{(n-2)} pe^t pe^t + n(pe^t + q)^{(n-1)} pe^t \\
&= npe^t (pe^t + q)^{(n-2)} ((n-1)pe^t + pe^t + q) \\
&= npe^t (pe^t + q)^{(n-2)} (npe^t + q)
\end{aligned}
$$

And evaluate each of the derivatives at $t = 0$ for the required moments.

$$
\begin{aligned}
E(X) &= m'_X(0) \\
&= n(pe^0 + q)^{n-1} pe^0 \\
&= n(p + q)^{n-1} p \\
&= n1^{n-1} p \\
&= np \\
E(X^2) &= m''_X(0) \\
&= npe^0 (pe^0 + q)^{(n-2)} (npe^0 + q) \\
&= np(p + q)^{n-2} (np + q) \\
&= np(np + q) \\
&= (np)^2 + npq
\end{aligned}
$$

The mean is np like we would expect. For the variance, substitute these values into (4).

$$
\begin{aligned}
\mathrm{Var}(X) &= E(X^2) - E(X)^2 \\
&= (np)^2 + npq - (np)^2 \\
&= npq \qquad (9)
\end{aligned}
$$

2.1.2 Poisson Process

A Poisson process is a special type of counting process. A counting process is a function $N(t), t \geq 0$ that only takes non-negative integer values, is non-decreasing, and $N(0) = 0$; i.e. it represents a count of occurrences over time. The Python and R workshops provide graphs of the Poisson probability mass function.

$$f_X(k; \lambda) = e^{-\lambda} \frac{\lambda^k}{k!} \quad k = 0, 1, 2, \ldots \tag{10}$$

The workshops experiment with $\lambda = 5$. You get a PMF curve with a hump near λ.

Mean and Variance

For the mean and variance we appeal once again to the moment generation function technique.

$$\begin{aligned} m_X(t) &= E(e^{it}) \\ &= \sum_{i=0}^{n} e^{it} e^{-\lambda} \frac{\lambda^k}{k!} \\ &= e^{-\lambda} \sum_{i=0}^{n} \frac{(e^t \lambda)^i}{i!} \\ &= e^{-\lambda} e^{e^t \lambda} \end{aligned}$$

Now differentiate the moment generating function twice with respect to t.

$$\begin{aligned} m'_X(t) &= e^{-\lambda} e^{e^t \lambda} e^t \lambda \\ &= \lambda e^{(t-\lambda)+\lambda e^t} \\ m''_X(t) &= \lambda e^{(t-\lambda)+\lambda e^t} (1 + \lambda e^t) \end{aligned}$$

Now evaluate each of the derivatives at $t = 0$ for the required moments.

$$\begin{aligned} E(X) &= m'_X(0) \\ &= \lambda e^{(0-\lambda)+\lambda e^0} \\ &= \lambda e^0 \\ &= \lambda \\ E(X^2) &= m''_X(0) \end{aligned}$$

$$\begin{aligned} &= \lambda e^{(0-\lambda)+\lambda e^0}(1+\lambda e^0) \\ &= \lambda e^0(1+\lambda) \\ &= \lambda + \lambda^2 \end{aligned}$$

The mean is λ like we expect. For the variance we substitute these values into (4).

$$\begin{aligned} \mathrm{Var}(X) &= E(X^2) - E(X)^2 \\ &= \lambda + \lambda^2 - \lambda^2 \\ &= \lambda \end{aligned} \tag{11}$$

Motivation

The curve looks plausible enough. But why this function? Is it just the most convenient way to express a curve with a hump in a certain place? Or is there something special about the expression in (10)?

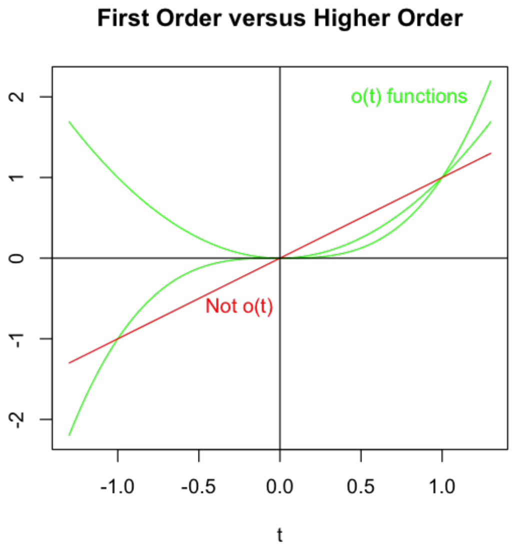
It turns out that this particular "humped curve" does indeed have some special properties. In fact, (10) can be derived from four properties.

1. Non-overlapping intervals are independent.
2. The processes is *stationary*. That is, the probability of an occurrence within time h from zero is the same as the probability of an occurrence within time t to $t+h$, it doesn't depend on t. In symbols: $P[N(t+h) - N(t)]$ depends only on h, not t.
3. $P[N(t) = 1] = \lambda t + o(t)$
4. $P[N(t) > 1] = o(t)$

The symbol $o(t)$ represents any quantity second order or higher. Namely,

$$\lim_{t\to 0} \frac{o(t)}{t} = 0$$

If $o(t)$ was analytic (and we're not saying it is), then its power series representation would only have second order and above terms (i.e. $o(t) = a_2t^2 + a_3t^3 \ldots$).



Conditions 3 and 4 are saying that $P[N(t) = 1]$ is similar to the red line for small t while $P[N(t) > 1]$ looks like a linear combination of green lines.

2.2 Continuous Distributions

Continuous distributions describe experiments where the random variable can take continuous values. Examples are time intervals and averages of counts.

A *continuous density* f_X corresponding to a continuous distribution is related through its integral in a way similar to how a discrete mass function is related through a sum.

$$\begin{aligned} F_X(x) &= \int_{-\infty}^{x} f(t)dt \\ f_X(x) &= \frac{dF(x)}{dx} \end{aligned}$$

The analogous formulas for expectation and variance also hold.

$$E[X] = \int_{-\infty}^{\infty} x f_X(x)dx \tag{12}$$

$$\text{Var}[X] = \int_{-\infty}^{\infty} (E[X] - x)^2 f_X(x)dx \tag{13}$$

Our friend, the moment generating function, continues to be useful with continuous random variables.

$$m_X(t) = E[e^{xt}] = \int_{-\infty}^{\infty} e^{xt} f_X(x) dx \tag{14}$$

This is the continuous version of (5).

2.2.1 Normal Distribution

The granddaddy of all distributions is the *normal distribution*. It has two parameters: the mean μ and standard deviation σ. The density function for the normal distribution in one dimension is given by the following formula.

$$f_X(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{15}$$

It has the familiar bell-shaped curve centered about the mean. A common practice is to normalize the random variable through the transformation

$$z = \frac{x - \mu}{\sigma}$$

Then it takes the form

$$f_X(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} \tag{16}$$

The moment generation function is defined in the usual way.

$$\begin{aligned} m_X(t) &= E\left[e^{tx}\right] \\ &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 + tx\right] dx \end{aligned}$$

From this point it becomes an exercise in completing the square within the exponent.

$$\begin{aligned}
-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}+tx &= \frac{1}{2\sigma^2}\left[(x-\mu)^2-2tx\sigma^2\right]\\
&= \frac{1}{2\sigma^2}\left[x^2-2x\mu+\mu^2-2tx\sigma^2\right]\\
&= \frac{1}{2\sigma^2}\left[x^2-2x(\mu+t\sigma^2)+\mu^2\right]\\
&= \frac{1}{2\sigma^2}\left[[x^2-2x(\mu+t\sigma^2)+(\mu+t\sigma^2)^2]-(\mu+t\sigma^2)^2+\mu^2\right]\\
&= \frac{1}{2\sigma^2}\left[[x-(\mu+t\sigma^2)]^2-\mu^2+2\mu t\sigma^2-t^2\sigma^4+\mu^2\right]\\
&= \frac{1}{2}\left[\frac{x-(\mu+t\sigma^2)}{\sigma}\right]^2+\frac{1}{2}\left[2\mu t+t^2\sigma^2\right]
\end{aligned}$$

Now if we substitute this exponent expression back into our moment generating function, we get

$$\begin{aligned}
m_X(t) &= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}\exp\left[\frac{1}{2}\left(\frac{x-(\mu+t\sigma^2)}{\sigma}\right)^2+\frac{1}{2}(2\mu t+t^2\sigma^2)\right]dx\\
&= \exp\left[\frac{1}{2}(2\mu t+t^2\sigma^2)\right]\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}\exp\left[\frac{1}{2}\left(\frac{x-(\mu+t\sigma^2)}{\sigma}\right)^2\right]dx\\
&= \exp\left[\mu t+\frac{1}{2}t^2\sigma^2\right] \qquad (17)
\end{aligned}$$

Now differentiate $m_X(t)$ twice.

$$\begin{aligned}
m'_X(t) &= (\mu+t\sigma^2)\exp\left[\mu t+\frac{1}{2}t^2\sigma^2\right]\\
m''_X(t) &= (\mu+t\sigma^2)(\mu+t\sigma^2)\exp\left[\mu t+\frac{1}{2}t^2\sigma^2\right]+\sigma^2\exp\left[\mu t+\frac{1}{2}t^2\sigma^2\right]
\end{aligned}$$

Evaluate at zero to obtain the moments.

$$\begin{aligned}
E[X] &= m'_X(0)\\
&= (\mu+0)\exp[0]\\
&= \mu\\
E[X^2] &= m''_X(0)
\end{aligned}$$

$$\begin{aligned} &= (\mu + 0)(\mu + 0)\exp[0] + \sigma^2 \exp[0] \\ &= \mu^2 + \sigma^2 \end{aligned}$$

The variance comes out as expected.

$$\begin{aligned} \text{var}[X] &= E[X^2] - E[X]^2 \\ &= \mu^2 + \sigma^2 - \mu^2 \\ &= \sigma^2 \end{aligned}$$

A linear combination of normal distributions $X_1, X_2, ..., X_n$ are *jointly normal* if the density of linear combination is given by

$$\frac{1}{(\sqrt{2\pi})^n} \frac{1}{\det|G|} \exp\left[-\frac{1}{2}(x-\alpha)^t G^{-1}(x-\alpha)\right]$$

This seems menacing at first. But we can gain an intuition for it if we consider some simple cases. Perhaps the most menacing introduction is the correlation matrix G. This matrix contains the covariance of each pair of X_i.

$$G_{ij} = \text{cov}(X_i, X_j)$$

Let's consider the case where the X_i are independent of each other. Then $\text{cov}(X_i, X_j) = 0$ for $i \neq j$ and G is diagonal where each diagonal element $G_{ii} = var(X_i) = \sigma_i^2$.

Let's assume the X_i are centered so that $\alpha = 0$. Then the expression in the exponent simplifies to

$$\begin{aligned} -\frac{1}{2}(x-\alpha)^t G^{-1}(x-\alpha) &= -\frac{1}{2}x^t G^{-1} x \\ &= -\frac{1}{2}\sum_{i=1}^{n} x_i \frac{1}{G_{ii}} x_i \\ &= -\frac{1}{2}\sum_{i=1}^{n} \frac{x_i^2}{G_{ii}} \end{aligned}$$

With these simplifications, the exponent becomes a weighted some of squares among the x components. If we further stipulate that the variances are equal, then $G = \sigma^2 I$ is a constant and the expression simplifies to

$$-\frac{1}{2}\sum_{i=1}^{n}\frac{x_i^2}{\sigma^2} = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}x_i^2 = -\frac{1}{2}\frac{|x|^2}{\sigma^2}$$

So we can see how the general form simplifies to our familiar one if we make assumptions about covariances.

2.2.2 Exponential Distribution

The Exponential distribution is, in a sense, the complement of the Poisson distribution introduced in Subsection (2.1.2). Whereas the Poisson distribution applies to the question of "how many occurrences of an event happen per unit time for Poisson process," the Exponential distribution addresses the question "how long until the next occurrence."

Let's get a feel for how the characterization of a Poisson process from Page (10) leads to an Exponential distribution. Let $P_0(t)$ denote the probability of no occurrences between times 0 and t. We can rewrite this in terms of characterics (iii) and (iv) of a Poisson process.

$$\begin{aligned}
P_0(t) &= \text{nothing happens from zero to time } t \\
&= 1 - [\text{exactly one thing happens}] - [\text{more than one thing happens}] \\
&= 1 - [\lambda t) + o(t)] - [o(t)] \\
&= 1 - \lambda t + o(t)
\end{aligned}$$

The $o(t)$ arithmetic may look funny. But remember that $o(t)$ refers to a class of functions which are closed under linear combinations.

Another property that $P_0(t)$ has from Poisson characteristic (i) of Page 10 is that non-overlapping intervals are independent. That means the probability of two things happening in separate (non-overlapping) intervals is the product of the separate probabilities.

$$P_0(t+h) = P_0(t)P_0(h)$$

We can the above two equations to derive a differential equation for $P_0(t)$.

$$\begin{aligned}
\frac{dP_0(t)}{dt} &= \lim_{h\to 0}\frac{P_0(t+h) - P_(t)}{h} \\
&= \lim_{h\to 0}\frac{P_0(t)P_0(h) - P_0(t)}{h} \\
&= \lim_{h\to 0}\frac{P_0(t)(P_0(h) - 1)}{h}
\end{aligned}$$

$$
\begin{aligned}
&= P_0(t) \lim_{h\to 0} \frac{(1-\lambda h + o(h) - 1)}{h} \\
&= P_0(t) \left[\lim_{h\to 0} \frac{-\lambda h}{h} + \lim_{h\to 0} \frac{o(h)}{h} \right] \\
&= P_0(t) \left[\lim_{h\to 0} (-\lambda) + 0 \right] \\
&= -\lambda P_0(t) \qquad (18)
\end{aligned}
$$

This first order differential equation has a well known solution of $P_0(t) = Ke^{\lambda t}$ for arbitrary constant K. This applies at a particular point t. But since the Poisson process is *stationary* (characteristic (ii) on Page 10) it translates to all $t > 0$. Since we are constraining $P_0(t)$ to be a probability density function, we require that

$$1 = \int_0^\infty Ke^{-\lambda t} dt$$

which constrains K to be λ. So our expression for the exponential density of a Poisson process with rate $1/\lambda$ is

$$f_X(x;\lambda) = \lambda e^{-\lambda x} \qquad (19)$$

The exponential random variable has a cumulative distribution function of the form

$$F_X(x;\lambda) = \begin{cases} 1 - e^{-\lambda x} & : \quad x > 0 \\ 0 & : \quad x \leq 0 \end{cases} \qquad (20)$$

To calculate the mean and variance, we resort once again to the moment generating function.

$$
\begin{aligned}
m_X(t) &= E[e^{xt}] \\
&= \int_0^\infty e^{xt} \lambda e^{-x\lambda} dx \\
&= \lambda \int_0^\infty e^{-x(\lambda - t)} dx \\
&= \lambda \cdot \frac{-1}{\lambda - t} e^{-x(\lambda - t)} \Big|_{x=0}^{x=\infty} \\
&= \frac{-\lambda}{\lambda - t}(0-1) \\
&= \frac{\lambda}{\lambda - t} \qquad (21)
\end{aligned}
$$

Note the definite integral only converges for $\lambda - t > 0$. This is fine for our needs since we intend to evaluate the moment generating function close to zero for its derivatives at zero.

Now differentiate the moment generating function twice with respect to t.

$$\begin{aligned} m'_X(t) &= \lambda(\lambda - t)^{-2} \\ m''_X(t) &= \lambda(-2)(\lambda - t)^{-3}(-1) \\ &= \frac{2\lambda}{(\lambda - t)^3} \end{aligned}$$

Now evaluate each of the derivatives at $t = 0$ for the required moments.

$$\begin{aligned} E(X) &= m'_X(0) = \frac{\lambda}{\lambda^2} = \frac{1}{\lambda} \\ E(X^2) &= m''_X(0) = \frac{2\lambda}{\lambda^3} = \frac{2}{\lambda^2} \end{aligned} \tag{22}$$

The mean is $1/\lambda$. For the variance we substitute these values into (4).

$$\mathrm{Var}(X) = E(X^2) - E(X)^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

It turns out the first and second moments are not difficult to derive directly from their integral formulas.

$$\begin{aligned} E[X] &= \int_0^\infty x\lambda e^{-\lambda x} dx \\ E[X^2] &= \int_0^\infty x^2\lambda e^{-\lambda x} dx \end{aligned}$$

It's a straight forward pair of calculations using integration by parts. But moment generating functions still save a little bit of work.

2.2.3 Gamma Distribution

The *Gamma function* is defined by

$$\Gamma(r) = \int_0^\infty x^{r-1} e^{-x} dx \tag{23}$$

for $r \in [1, \infty)$. One application of integration by parts reveals its special property.

$$\begin{aligned}
\Gamma(r) &= \int_0^\infty x^{r-1} e^{-x} dx \\
&= -x^{r-1} e^{-x} \Big|_0^\infty + \int_0^\infty (r-1) x^{r-2} e^{-x} dx \\
&= -(0-0) + (r-1) \int_0^\infty x^{(r-1)-1} e^{-x} dx \\
&= (r-1)\Gamma(r-1)
\end{aligned}$$

It's easy to show $\Gamma(1) = 1$. So for positive integers $\Gamma(n) = (n-1)\Gamma(n-1) = (n-1)!$. But the Gamma function is defined for all real $r > 1$. So the special property of the Gamma function is that it "fills in the gaps" of the factorial function.

The Gamma function lends its name to the *Gamma density function.*

$$\text{Gamma}(x; r, \lambda) = \frac{\lambda}{\Gamma(r)} (\lambda x)^{r-1} e^{-\lambda x} \tag{24}$$

We'll see in the next section how the Gamma random variable arises from summing r Exponential random variables with parameter λ. For now let's content ourselves with determining its moment generating function. About halfway through you see the change of variable $u = (\lambda - t)x$.

$$\begin{aligned}
m_X(t) &= E[e^{xt}] \qquad (25) \\
&= \int_0^\infty e^{xt} \frac{\lambda}{\Gamma(r)} (\lambda x)^{r-1} e^{-\lambda x} dx \\
&= \frac{\lambda}{\Gamma(r)} \int_0^\infty (\lambda x)^{r-1} e^{-(\lambda - t)x} dx \\
&= \frac{\lambda}{\Gamma(r)} \int_0^\infty \left(\frac{\lambda u}{\lambda - t} \right)^{r-1} e^{-u} \frac{du}{\lambda - t} \\
&= \frac{\lambda^r}{(\lambda - t)^r \Gamma(r)} \int_0^\infty u^{r-1} e^{-u} du \\
&= \frac{\lambda^r \Gamma(r)}{(\lambda - t)^r \Gamma(r)} \\
&= \left(\frac{\lambda}{\lambda - t} \right)^r \qquad (26)
\end{aligned}$$

This reduces to (21) for the Exponential distribution when $r = 1$.

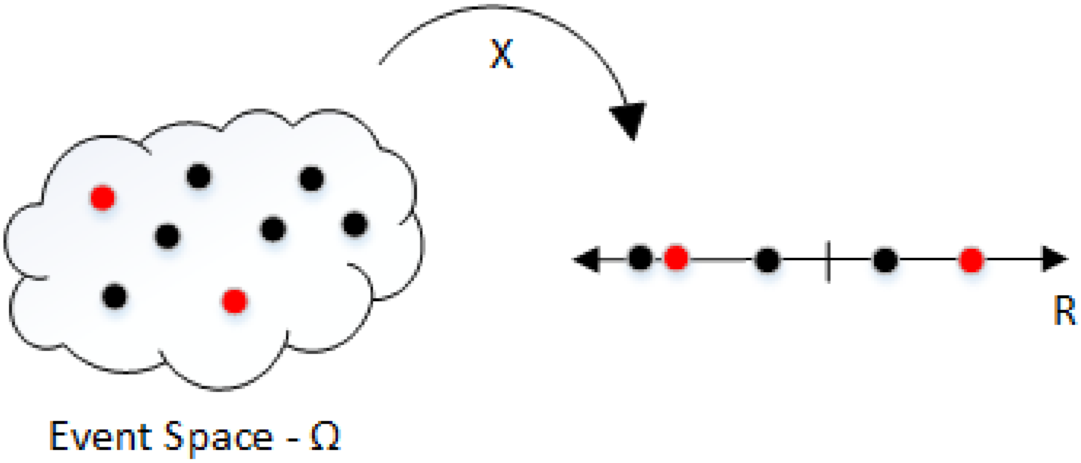
3 Random Variable Functions

Most of the random variables we've examined so far directly directly model some conceptually simple experiment like flipping a coin, counting the successes of a number of trials, or counting the number of occurrences over a fixed time interval. But we'll find there is value in constructing new random variables that are functions of random variables we have already encountered. The most common cases are squares and linear combinations of variables. This section will outline two techniques useful for determining the distribution function that arises from a function of a random variable.

3.1 CDF Method

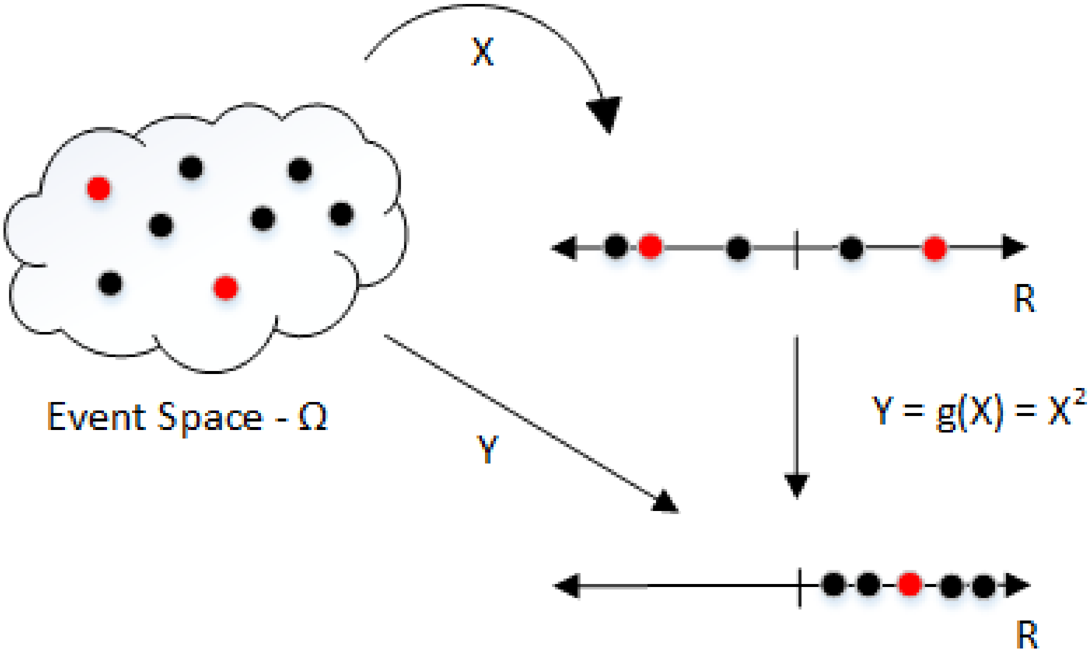
Consider random variable $Y = X^2$ where we already know the density function $f_X(x)$. What can we say about $f_Y(y)$. The technique we'll demonstrate in this section starts with the CDF. Since Y can only take on positive values, we know that the CDF is zero for all $Y \leq 0$.

Let's recall what a random variable is doing. A random variable maps from an abstract event space to points on the real number line. Often the event space won't be so abstract and the mapping will be quite natural; so natural that we often just write X rather than $X(\omega) \in \mathcal{R}$ for $\omega \in \Omega$.



By mapping from an "event space" to the concrete real number line, random variables allow us to directly employ analytic techniques that have long existed for the real number line. Our cumulative distribution functions and probability density functions are defined on the real number line by virtue of the random variable that maps events to the real number line.

A function of a random variable simply maps the points from one real number line to another real number line. In this way, a function of a random variable creates a new random variable on the same event space, i.e. a new way to map events to real numbers.



In the figure above, $g(x) = x^2$ is such a function. If we define Y as the result of mapping $\Omega \rightarrow \mathcal{R} \rightarrow \mathcal{R}$, then Y is also a random variable.

Assuming that such a mapping is useful (and we'll find out later that this $g(X) = X^2$ is), we'll want to know the CDF and PDF of the new random variable. Unfortunately, it's not as simple as plugging x^2 everywhere you see an x.

Using the figure above as an example, let's determine $F_Y(y)$ from basic principles. The basic principle is

$$F_Y(y) = P[Y \leq y]$$

From this principle we plug in the function to derive the expression. Since $Y = g(X) = X^2$, we restrict our attention to $y > 0$ since $F_Y(y) = 0$ for $y \leq 0$.

$$\begin{aligned} F_Y(y) &= P[Y \leq y] \\ &= P[X^2 \leq y] \\ &= P\left[-\sqrt{y} \leq X \leq \sqrt{y}\right] \\ &= P\left[X \leq \sqrt{y}\right] - P\left[X \leq -\sqrt{y}\right] \end{aligned}$$

This gives us the CDF for Y. It's close to simply substituting $\sqrt{y}$ for x into the expression for F_X. Now that we have the CDF, we can derive the PDF.

$$\begin{aligned}
f_Y(y) &= \frac{d}{dy}F_Y(y) \\
&= \frac{d}{dy}\left[F_x(\sqrt{y}) - F_X(-\sqrt{y})\right] \\
&= \left.\frac{dF_X}{dx}\right|_{x=\sqrt{y}} \frac{d(\sqrt{y})}{dy} - \left.\frac{dF_X}{dx}\right|_{x=-\sqrt{y}} \frac{d(-\sqrt{y})}{dy} \\
&= \frac{1}{2\sqrt{y}}\left[f_x(\sqrt{y}) + f_x(-\sqrt{y})\right]
\end{aligned}$$

where

$$\frac{dx}{dy} = \frac{dy^{\frac{1}{2}}}{dy} = \frac{1}{2}y^{-\frac{1}{2}}$$

The expression for f_Y in terms of f_X in this specific case of $Y = X^2$ demonstrates two important complications.

1. If g is not 1-to-1, then we must account for all the points in $g^{-1}(y)$. In the case of a quadratic, there are usually two.

2. The f_Y expression introduced an additional multiplicative factor, known in vector calculus circles as *the Jacobian*. The Jacobian incorporates the stretch factor introduced by the transformation.

It's reasonable to ask why the CDF didn't need a stretch factor. It's because we don't integrate the CDF. We just evaluate it at various points. **The value of the CDF at any point represents a probability.**

This is not the case with a continuous density function. The value of f_X at a single point does not really mean anything **by itself. It certainly does not represent a probability.** We only get probabilities from f_X by integrating it over some interval or combination of intervals. In other words, it's only the area under f_X that has a probability interpretation. You can't say anything about an area with just the height. You need to also specify the width. So in the case of our transformation g, it's not enough to determine the height of f_Y at a certain point through corresponding values of f_X. We need to know how g is stretching the differential widths to get the full picture of how areas under f_X transform to *areas* under f_Y. The Jacobian factor provides this "width stretching" information.

We're still messing around in the realm of probability. In the next workshop, we'll consider some special functions of a random variable (like the average of a sample) and cross into the realm of *statistics* proper.

3.2 Moment Generating Functions

Another way to determine the density function $f_Y(y)$ for $Y = g(X)$ is to compute the moment generating function of $g(x)$ and determine whether it's something we recognize.

3.2.1 Linear Combinations

Let's consider the case where we have n independent samples, each one from its own distribution. Let Y be a linear combination of the independent samples. That is, $Y = g(X) = a_1X_1 + \cdots + a_nX_n$ where X_i represents the distribution i from which we are sampling. The a_i are constants. The expression for the moment generating function is

$$\begin{aligned} m_Y(t) &= E\left[e^{tY}\right] = E\left[e^{tg(x)}\right] = E\left[e^{t\sum_{i=1}^n a_iX_i}\right] = E\left[e^{\sum_{i=1}^n ta_iX_i}\right] = E\left[\prod_{i=1}^n e^{ta_iX_i}\right] \\ &= \prod_{i=1}^n E\left[e^{ta_iX_i}\right] = \prod_{i=1}^n m_{X_i}(ta_i) \end{aligned} \tag{27}$$

We used the independence between the samples in going from the first line to the second. So the moment generating function for the sum of independent random variables can be expressed as product of the individual moment generating function.

Let's consider an example where X represents an exponential random variable with parameter λ. Recall that the value of the exponential random variable is the expected wait time for a Poisson process with rate $1/\lambda$. Let $Y = X_1 + \cdots + X_n$, the time for the nth occurrence. Then the product rule in (27) where each m_{X_i} is the exponential moment generating function with parameter λ. We established this in Equation (21) on Page 16.

$$m_X(t) = \frac{\lambda}{\lambda - t}$$

Plugging this into (27) yields

$$m_Y(t) = \prod_{i=1}^n m_{X_i}(t) = \prod_{i=1}^n \frac{\lambda}{\lambda - t} = \left(\frac{\lambda}{\lambda - t}\right)^n$$

And this, as we saw in Equation (26) on Page 18, is the moment generating function for the Gamma distribution.

Another popular application of (27) yields the moment generating function for an average of independent samples.

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

In this case, we have $X_i = X$ and $a_i = 1/n$.

$$m_{\bar{X}}(t) = \left[m_X\left(\frac{t}{n}\right)\right]^n \tag{28}$$

3.2.2 Normal Squares

The above are nice general results. Let's focus our attention on the normal distribution. The formulas above apply to the normal distribution; but here we'll investigate random variables that arise from the squares of a normal random variable.

Why would we care about the square of a normal random variable? This is actually two questions.

1. Why do we care about squares?
2. Why do we care about the normal distribution?

We care about the square of the distance from the mean because it provides us information about the spread around its mean. Combinations of positions arounds its mean can cancel out. But squares are always positive and they accumulate. Moreover, they accumulate in expected patterns which, if not followed, lead us to suspect the underlying assumptions. One popular application of this is the so-called Chi-Squared goodness of fit test.

Consider the case of n independent samples from a normal distribution with mean μ and variance σ^2; then standardize the usual way.

$$Z_i = \frac{X_i - \mu}{\sigma}$$

Our new random variable will be

$$U = \sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2$$

A single value of U represents the outcome of the experiment of independently sampling the X random variable n times, squaring each value, and summing these squares. If we do this many times, the values of U are distributed in a certain way. We don't yet know how to expect these values to be distributed. A density function for U would certainly help. To this end, let's see if we can determine the moment generating function for U.

$$m_U(t) = E\left[e^{tU}\right] = E\left[e^{t\sum_{i=1}^{n} Z_i^2}\right] = E\left[\prod_{i=1}^{n} e^{tZ_i^2}\right] = \prod_{i=1}^{n} E\left[e^{tZ_i^2}\right] \tag{29}$$

This is just the linear combination stuff we did in the last section. Now let's dig into this square.

$$\begin{aligned} E\left[e^{tZ_i^2}\right] &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{tz_i^2}e^{-\frac{1}{2}z_i^2}dz_i \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(1-2t)z_i^2}dz_i \end{aligned} \tag{30}$$

At this point we introduce a standard change of variable for this kind of thing.

$$\begin{aligned} u^2 &= (1-2t)z_i^2 \qquad & 2udu &= (1-2t)2z_idz_i \\ \frac{u}{z} &= \sqrt{1-2t} \qquad & dz_i &= \frac{2udu}{(1-2t)2z} = \frac{\sqrt{1-2t}}{1-2t}du = \frac{du}{\sqrt{1-2t}} \end{aligned}$$

Let's dump all this into (30) to get something we can work with.

$$\begin{aligned} E\left[e^{tZ_i^2}\right] &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}u^2}\frac{du}{\sqrt{1-2t}} \\ &= \frac{1}{\sqrt{1-2t}}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}u^2}du \\ &= \frac{1}{\sqrt{1-2t}} \\ &= \left(\frac{1/2}{1/2-t}\right)^{1/2} \end{aligned} \tag{31}$$

Substitute (31) back into (29).

$$\begin{aligned} m_U(t) &= \prod_{i=1}^{n}\left(\frac{1/2}{1/2-t}\right)^{1/2} \\ &= \left(\frac{1/2}{1/2-t}\right)^{n/2} \end{aligned} \tag{32}$$

Now let's step back and think: have we seen a moment generating function like (32) anywhere? It should look familiar. If not, review the previous section where we showed how the Gamma random variable arises from successive exponential trials. The Gamma moment generating function is Equation (26) on Page 18. The expression in (32) is the Gamma moment generating function for $r = n/2$ and $\lambda = 1/2$.

So our sum of squares of n standard normals has a distribution that we have already encountered before: Gamma$(n/2, 1/2)$. This distribution is special enough to get its own name: the *Chi-Squared* distribution.

$$\chi_n^2(x) = \text{Gamma}(x; n/2, 1/2) = \frac{1/2}{\Gamma(n/2)} \left(\frac{x}{2}\right)^{\frac{n}{2}-1} e^{-x/2} \tag{33}$$

The n is a parameter of our distribution called the *degrees of freedom*. So the above expression is said to be a "Chi-Squared distribution with n degrees of freedom."

3.3 Samples

3.3.1 Sample Average

Let's define a random variable to represent the average of n samples from a distribution: $\{X_1, \cdots, X_n\}$.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{34}$$

$\bar{X}$ is a random variable just like X. When we write $\{\bar{X}_1, \bar{X}_2, \cdots\}$, we mean a sequence of averages. It's not a big surprise that if we calculate the expectation of the sample averages, we get the mean.

$$\begin{aligned} E[\bar{X}] &= E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} E\left[X_i\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mu \\ &= \mu \end{aligned}$$

This says that the expected value of the sample average is the mean. This sounds so natural that we when we first encounter it, we often confuse $\bar{X}$ and μ. They are **not** the same. $E[\bar{X}] = \mu$; but $\bar{X} \neq \mu$. This will become more apparent as we director our attention to the sample variance.

We know that if we have a whole bunch of samples, the sample average will converge to the population mean. But how close to the mean can we expect these averages to be? A fundamental indicator of how the samples are likely to be spread is the *standard deviation*, which is the positive square root of the variance. We can use (53) on Page 34 of the Appendix to find an expression for the variance of an average. Since the sampling is done independently,

$$\begin{aligned} \mathrm{Var}[\bar{X}] &= \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\ &= \sum_{i=1}^{n} \frac{1}{n^2}\mathrm{Var}[X_i] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 \\ &= \frac{\sigma^2}{n} \end{aligned} \tag{35}$$

where σ^2 is the population variance. The standard deviation of the averages is the square root of this quantity.

$$\frac{\sigma}{\sqrt{n}}$$

This is how much we expect the average of samples to dance around the mean for a given sample size n. For $n = 1$, we get the original standard deviation of the population, which is what we should expect. As we bump up n, the standard deviation decreases. In other words, the distribution of the average clusters more closely to the mean.

3.3.2 Central Limit Theorem

We've determined that for a sample size n, if we perform the sampling many times, and accumulate many average ($\bar{X}$) values, they'll

- tend to have an average around the population mean μ
- tend to have a standard deviation near $\sigma/\sqrt{n}$ where σ is the population standard deviation.

But what else can we say about the distribution of $\bar{X}$? Is it symmetric? Does it look anything like the original X? The answer is amazing and is provided by the Central Limit Theorem (CLT). The CLT says that the distribution $\bar{X}_n$ (of averages from n samples) from a population with mean μ and standard deviation σ will asymptotically approach a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$.

$$\bar{X}_n \rightarrow \mathcal{N}(\mu, \sigma)$$

It's amazing that the shape of the population distribution (from which the samples were taken) doesn't matter. The shape of the sample average values will always approach a normal distribution. Now you may be wondering about the menacing qualifiers like *asymptotically* and *approach*. They signify that $\bar{X}_n$ isn't exactly normal. And that could be a deal-breaker if "not exactly" meant "not even close". Fortunately, in most cases, "not exactly" means "good enough". Some special cases are worth noting.

- If the population is already a normal distributon, then $\bar{X}_n$ will be exactly normal for all n.
- If the sample size is less than 30 (just a rule of thumb), there is more concern that n is too small for $\bar{X}_n$ to approach close enough to a normal. In this case we resort to a *t distributon* which is designed to compensate for the smaller sample size.

The CLT is the basis for most work on statistical inference and hypothesis testing.

3.3.3 Sample Variance

How close to μ can we reasonably expect our estimate $\bar{x}$ to be? We need to examine variance to answer that. The expression we'll ultimately use is not obvious upon inspection. For this reason we'll embark on an expository approach to arrive at a suitable expression for the variance.

We use the symbol S^2 to distinguish sample variance from the variance of the distribution, σ^2. Let's start with a candidate we'll call S_0^2.

$$S_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \tag{36}$$

This seems like a reasonable candidate for the variance. Let's check its expected value. This will tell us what to expect if we take more and more samples, resulting in more and more sample variances, to what can we expect their average to converge?

$$E\left[S_0^2\right] \quad = \quad E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2\right]$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n} E\left[(X_i-\mu)^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \sigma^2 \\
&= \sigma^2
\end{aligned}
$$

Great! We got the expected value we wanted. The sample variances will eventually converge to the population variance, the thing we're trying to estimate. The statistical term for this quality is that it is *unbiased.* Conversely, an estimator whose expected value is *not* equal to its counterpart for the population is called *biased.*

But there is a hitch. S_0^2 is *not a statistic.* That is, we can't evaluate it solely from data in our sample. We have all the $\{x_i\}$. But we don't know μ. (If we know this about the population, we probably wouldn't need to sample it; though there are cases where we know the mean of a population and we sample it to find other things.)

The best we can do without the actual mean is estimate it with $\bar{X}$. So let's try that with our new candidate S_1^2.

$$
S_1^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 \tag{37}
$$

$\bar{X}$ qualifies as a statistic because everything we need to determine its value is contained in a sample. You're probably on pins and needles wondering whether this estimator is biased. It *is biased.* However, don't despair; it happens to be a bias that is easy to correct. The correction is simple to apply; but not as simple to understand.

We could derive the expectation of (37) directly and readily see the needed correction. That would show us *what* correction is needed; but it would not provide us with intuition on *why* it is needed. To address the "why" as well as the "what," we'll take a somewhat circuitous route in our analysis of (37).

The first stop is Appendix A (page 31). There you'll derive a least squares formula that forms the crux of arguments that follow. Equations (47) and (48) from Appendix A are combined below for convenience.

$$
\frac{1}{n}\sum_{i=1}^{n}(x_i-\theta)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 + (\bar{x}-\theta)^2 \tag{38}
$$

In (38) the $\{x_i\}$ are a sample set, $\bar{x}$ is the average of the sample set, and θ is any real number. An important quality of (38) is that it divides the lefthand side into a sum of two parts: one that is independent of θ and one that contains the θ contribution. Note that the first part, independent of θ, is none other than our candidate for sample variance, S_1^2 from (37).

Let μ be the population mean. We don't know what μ is; we can only estimate it using $\bar{x}$. Set $\theta = \mu$ in (38). For every sample $\{x_i\}$, there will be a different $\bar{x}$ value that marks the minimal value of the least squares expression. μ, on the other hand, is always the same.

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 = &\quad \tfrac{1}{n}\textstyle\sum_{i=1}^{n}(x_i-\bar{x})^2 &+(\bar{x}-\mu)^2 \\ s_0^2 = &\quad s_1^2 &+(\bar{x}-\mu)^2 \end{aligned} \tag{39}$$

Here we let the lower case s_0^2 and s_1^2 denote the respective values of sampling the random variables S_0^2 and S_1^2. We know that $s_1^2 < s_0^2$ from the least square discussion. So it shouldn't be a surprise that the expectation of s_1^2 should also be less. This is a bit of a concern since s_0^2 is unbiased. That means s_1^2 probably **is** biased, unless the expectation of $(\bar{X}-\mu)^2$ is zero, which is unlikely since it is never negative. Let's calculate this expectation.

$$\begin{aligned}
E\left[(\bar{X}-\mu)^2\right] &= E\left[\bar{X}^2 - 2\bar{X}\mu + \mu^2\right] \\
&= E\left[\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)\left(\frac{1}{n}\sum_{j=1}^{n}X_j\right)\right] - 2\mu E[\bar{X}] + \mu^2 \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}E\left[X_iX_j\right] - 2\mu\mu + \mu^2 \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\left[E[X_i^2] + \sum_{j\neq i}E\left[X_iX_j\right]\right] - \mu^2 && (40) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\left[E[X^2] + \sum_{j\neq i}E\left[X_i\right]E\left[X_j\right]\right] - \mu^2 && (41) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\left[E[X^2] + \sum_{j\neq i}E\left[X\right]E\left[X\right]\right] - \mu^2 && (42) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\left[E[X^2] + \mu^2\sum_{j\neq i}1\right] - \mu^2 \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\left[E[X^2] + \mu^2(n-1)\right] - \mu^2 \\
&= \frac{1}{n}\left[E[X^2] + \mu^2(n-1)\right] - \mu^2 \\
&= \frac{1}{n}\left[E[X^2] - \mu^2\right]
\end{aligned}$$

$$= \frac{\sigma^2}{n} \tag{43}$$

In (40) we separate the expectation of a product into two cases: when $i = j$ and when $i \neq j$. When they are equal, it's just the expectation of a product of a random variable with itself, i.e. the second moment. When they aren't equal, it's the expectation of two independent random variables. In this case, $0 = \mathrm{cor}(A, B) = E[AB] - E[A]E[B]$. This is used in (41). In (42) we use again the fact that $E[X_i] = E[X]$ for all i.

Many will have guessed this result the well-known variance of the average of n samples from a distribution with variance σ^2. Now we also know this is how biased our s_1^2 estimator is. From (39) we have

$$\begin{aligned} E\left[S_1^2\right] &= E\left[S_0^2\right] - E\left[(\bar{X} - \mu)^2\right] \\ &= \sigma^2 - \frac{\sigma^2}{n} \\ &= \frac{n-1}{n}\sigma^2 \end{aligned} \tag{44}$$

S_1^2 is a true statistic because it consists only of items we know from the sample. As an estimator it's biased by the amount in (39) resulting in an expectation revealed in (44). By refining our statistic via the multiplicative constant revealed in (44), we get a statistic that is unbiased.

$$\begin{aligned} S^2 &= \frac{n}{n-1} S_1^2 \\ &= \frac{n}{n-1} \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 \\ &= \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 \end{aligned} \tag{45}$$

so that

$$E\left[S^2\right] = \sigma^2$$

This why we settle on the expression in (45) for our *sample variance*.

A Least Squares Formula

We assume n independent samples from a population with mean μ and variance σ^2. We don't know these two parameters. We can only estimate them through sampling.

Our first estimator is $\bar{X}$, defined by

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{46}$$

For a given sampling, $\{x_i\}, i = 1 \ldots n$, we have the value of the estimator (the estimate)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

First we show that for any sample, $\bar{x}$ minimizes

$$f(\theta) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \theta)^2 \tag{47}$$

It's easy to show this with Calculus. But the algebraic route with a few tricks will shed more light elsewhere in this article.

$$\begin{aligned}
f(\theta) &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \theta)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[(x_i - \bar{x}) + (\bar{x} - \theta)\right]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[(x_i - \bar{x})^2 + 2(x_i - \bar{x})(\bar{x} - \theta) + (\bar{x} - \theta)^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{2(\bar{x} - \theta)}{n}\sum_{i=1}^{n}(x_i - \bar{x}) + \frac{1}{n}(\bar{x} - \theta)^2\sum_{i=1}^{n}1 \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{2(\bar{x} - \theta)}{n}\left[\sum_{i=1}^{n}x_i - \sum_{i=1}^{n}\bar{x}\right] + \frac{n}{n}(\bar{x} - \theta)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{2(\bar{x} - \theta)}{n}\left[n\bar{x} - n\bar{x}\right] + (\bar{x} - \theta)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 + (\bar{x} - \theta)^2
\end{aligned} \tag{48}$$

Equation (48) makes it clear that the more θ differs from $\bar{x}$, the larger $f(\theta)$. Therefore the quantity reaches its least value when $\theta = \bar{x}$.

B Variance Formulas

In this section we'll derive some formulas for the variance and covariance of random variables. We first encountered the variance in (2) on Page 4. We used the discrete density function to derive an alternative expression (4) that would have derived more generally.

$$\begin{aligned}
\text{Var}[X] &= E\left[(X - E[X])^2\right] \\
&= E\left[X^2 - 2XE[X] + E[X]E[X]\right] \\
&= E[X^2] - E\left[2XE[X]\right] + E[X]E[X] \\
&= E[X^2] - 2E[X]E[X] + E[X]E[X] \\
&= E[X^2] - E[X]^2
\end{aligned}$$

B.1 Covarance

The *covariance* of two random variables X and Y is

$$\text{cov}(X, Y) = E\left[(X - E[X])(Y - E[Y])\right] \tag{49}$$

Much like with the "alternative formula" for variance, we can derive one for the covariance.

$$\begin{aligned}
\text{cov}(X, Y) &= E\left[(X - E[X])(Y - E[Y])\right] \\
&= E\left[XY - XE[Y] - E[X]Y - E[X]E[Y]\right] \\
&= E[XY] - E[X]E[Y] - E[X]E[Y] - E[X]E[Y] \\
&= E[XY] - E[X]E[Y] \qquad (50)
\end{aligned}$$

Note that the covariance of a random variable with itself is just the variance: $cov(X, X) = \text{Var}(X)$. An import property of covariance is that the covariance of two *independent* random variables is always zero. (The converse is not always true; but one has to try hard to find examples.)

When two random variables are independent, their covariance is zero. Equation (50) then implies

$$E[XY] = E[X]E[Y] \tag{51}$$

This is an important identity. We use it in many places; but it's important to understand we can only do so when X and Y are independent.

B.2 Linear Combinations

Expectations are linear.

$$E\left[\sum a_i X_i\right] = \sum a_i E[X_i]$$

Variances are not. In this section we'll determine formulas for a linear combination of random variables. In the present case, $X = \sum_{i=1}^n a_i X_i$.

$$\begin{aligned}
\mathrm{Var}\left[\sum_{i=1}^n a_i X_i\right] &= E\left[\left(\sum_{i=1}^n a_i X_i - E\left[\sum_{i=1}^n a_i X_i\right]\right)^2\right] \\
&= E\left[\left(\sum_{i=1}^n a_i X_i\right)\left(\sum_{i=1}^n a_i X_i\right) - 2\left(\sum_{i=1}^n a_i X_i\right) E\left[\sum_{i=1}^n a_i X_i\right] + \left(E\left[\sum_{i=1}^n a_i X_i\right]\right)^2\right] \\
&= E\left[\sum_{i=1}^n \sum_{j=1}^n a_i a_j X_i X_j - 2\left(\sum_{i=1}^n a_i X_i\right) \sum_{i=1}^n a_i E[X_i] + \left(\sum_{i=1}^n a_i E[X_i]\right)^2\right] \\
&= \sum_{i=1}^n \sum_{j=1}^n a_i a_j E[X_i X_j] - 2\left(\sum_{i=1}^n a_i E[X_i]\right) \sum_{i=1}^n a_i E[X_i] + E\left[\sum_{i=1}^n a_i E[X_i]\right]^2 \\
&= \sum_{i=1}^n \sum_{j=1}^n a_i a_j E[X_i X_j] - \sum_{i=1}^n a_i E[X_i] \sum_{i=1}^n a_i E[X_i] \\
&= \sum_{i=1}^n \sum_{j=1}^n a_i a_j E[X_i X_j] - \sum_{i=1}^n \sum_{j=1}^n a_i a_j E[X_i] E[X_j] \\
&= \sum_{i=1}^n \sum_{j=1}^n a_i a_j (E[X_i X_j] - E[X_i] E[X_j]) \\
&= \sum_{i=1}^n \sum_{j=1}^n a_i a_j \mathrm{cov}(X_i, X_j)
\end{aligned} \tag{52}$$

We've appealed to several linearity of averages sited above. Remember that $E[X]$ is a constant. So when it appears inside another average expression, it can be factored out. $E[aE[X]] = aE[E[X]] = aE[X]$. We've also used the expression for covariance in (50) in the last step.

One special cases is when all the X_i are independent of each other. Then $cov(X_i, X_j) = 0$ when $i \neq j$. Then (52) simplifies to

$$\begin{aligned}
\mathrm{Var}\left[\sum_{i=1}^{n} a_i X_i\right] &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \mathrm{cov}(X_i, X_j) \\
&= \sum_{i=1}^{n} a_i a_i \mathrm{cov}(X_i, X_i) \\
&= \sum_{i=1}^{n} a_i^2 \mathrm{Var}[X_i,] \qquad (53)
\end{aligned}$$