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

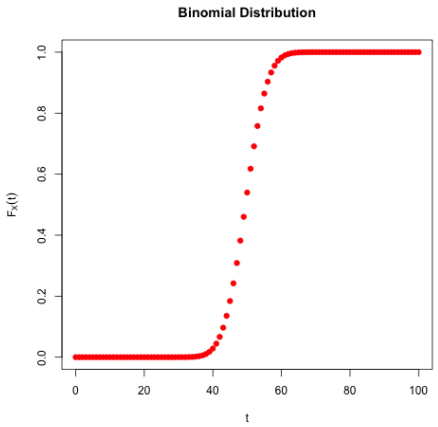
Once we have a random variable defined for an experiment that maps outcomes to real numbers, the next question is to ask how the values of the random variable are likely to be distributed along the real axis after performing the experiment many times. A central concept to answering this question is the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

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

And the variance.

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

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

$$\text{Var}[X] = \sum_{i=-\infty}^{\infty} (i - E[X])^2 \cdot f_X(i)$$

$$
\begin{aligned}
&= \sum_{i=-\infty}^{\infty} (i^2 - 2iE[X] + E[X]^2) \cdot f_X(i) \\
&= \sum_{i=-\infty}^{\infty} i^2 \cdot f_X(i) - 2E[X] \sum_{i=-\infty}^{\infty} i \cdot f_X(i) - E[X]^2 \sum_{i=-\infty}^{\infty} f_X(i) \\
&= E[X^2] - 2E[X]E[X] - E[X]^2 \\
&= E[X^2] - E[X]^2 \qquad (3)
\end{aligned}
$$

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

$$m_X(t) = E(e^{it}) = \sum_{i=-\infty}^{\infty} e^{it} f_X(i) \qquad (4)$$

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 (**??**), the first and second moments of X are, respectively,

$$E(X) = m'_X(0) \tag{5}$$
$$E(X^2) = m''_X(0) \tag{6}$$

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

$$f(i) = \binom{n}{i} p^i (1-p)^{n-i}, i \in 0, \ldots, n \tag{7}$$

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 (**??**) 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.

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

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 (**??**).

$$\begin{aligned} \mathrm{Var}(X) &= E(X^2) - E(X)^2 \\ &= (np)^2 + npq - (np)^2 \\ &= npq \end{aligned} \tag{8}$$

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.

$$f_X(k; \lambda) = e^{-\lambda}\frac{\lambda^k}{k!} \quad k = 0, 1, 2, \ldots \tag{9}$$

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.

$$\begin{aligned} E(X) &= m'_X(0) \\ &= \lambda e^{(0-\lambda)+\lambda e^0} \\ &= \lambda e^0 \\ &= \lambda \\ E(X^2) &= m''_X(0) \\ &= \lambda e^{(0-\lambda)+\lambda e^0}(1 + \lambda e^0) \\ &= \lambda e^0(1 + \lambda) \\ &= \lambda + \lambda^2 \end{aligned}$$

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

$$\begin{aligned} \mathrm{Var}(X) &= E(X^2) - E(X)^2 \\ &= \lambda + \lambda^2 - \lambda^2 \\ &= \lambda \end{aligned} \tag{10}$$

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 (**??**)?

It turns out that this particular "humped curve" does indeed have some special properties. In fact, (**??**) 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. $P[N(t) \geq 1] = o(t)$

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. $o(t) = a_2 t^2 + a_3 t^3 \ldots$).

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.

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx \tag{11}$$

$$\text{Var}[X] = \int_{-\infty}^{\infty} (E[X] - x)^2 f_X(x) dx \tag{12}$$

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

$$m_X(t) = E[e^{xt}] = \int_{-\infty}^{\infty} e^{xt} f_X(x) dx \tag{13}$$

This is the continuous version of (**??**).

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.

$$f_X(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{14}$$

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

$$f_X(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} \tag{15}$$

The moment generation function is defined in the usual way.

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

$$= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2+tx\right]dx$$

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

$$\begin{aligned}
m_X(t) &= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}\exp\left[\frac{1}{2}\left(\frac{x-(\mu+t\sigma^2)}{\sigma}\right)^2+\frac{1}{2}(2\mu t+t^2\sigma^2)\right]dx\\
&= \exp\left[\frac{1}{2}(2\mu t+t^2\sigma^2)\right]\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}\exp\left[\frac{1}{2}\left(\frac{x-(\mu+t\sigma^2)}{\sigma}\right)^2\right]dx\\
&= \exp\left[\mu t+\frac{1}{2}t^2\sigma^2\right] \qquad (16)
\end{aligned}$$

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.

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

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 random variable has a cumulative distribution function of the form

$$F_X(x;\lambda) = \begin{cases} 1 - e^{-\lambda x} & : \quad x > 0 \\ 0 & : \quad x \leq 0 \end{cases} \tag{17}$$

where $1/\lambda$ is the expected wait-time of a Poisson processes. As usual, the density function is the derivative of the CDF.

$$f_X(x;\lambda) = \lambda e^{-\lambda x} \tag{18}$$

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

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

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.

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

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

$$\begin{aligned} \mathrm{Var}(X) &= E(X^2) - E(X)^2 \\ &= \frac{2}{\lambda^2} - \frac{1}{\lambda^2} \\ &= \frac{1}{\lambda^2} \end{aligned} \tag{19}$$

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.