Supplementary material for Lesson 3

1 Indicator Functions

The concept of an indicator function is a really useful one. This is a function that takes the value one if its argument is true, and the value zero if its argument is false. Sometimes these functions are called Heaviside functions or unit step functions. I write an indicator function as $I_{\{A\}}(x)$, although sometimes they are written $1_{\{A\}}(x)$. If the context is obvious, we can also simply write $I_{\{A\}}$.

Example:

$$I_{\{x>3\}}(x) = \begin{cases} 0 & x \leq 3 \\ 1 & x > 3 \end{cases}$$

Indicator functions are useful for making sure that you don't take the log of a negative number, and things like that. Indicator functions are always first in the order of operations—if the indicator function is zero, you don't try to evaluate the rest of the expression. When taking derivatives they just go along for the ride. When taking integrals, they may affect the range over which the integral is evaluated.

Example: The density function for the exponential distribution (see Section 4.1) can be written as $f(x) = \lambda \exp(-\lambda x) I_{\{x \geq 0\}}(x)$. We can integrate the density and show that it is indeed a proper density and integrates to one:

$$\int_{-\infty}^{\infty} \lambda \exp(-\lambda x) I_{\{x \geq 0\}}(x) dx = \int_{0}^{\infty} \lambda \exp(-\lambda x) dx = -\exp(-\lambda x)|_{0}^{\infty} = -(0 - 1) = 1.$$

The derivative of the density function would be:

$$\frac{d}{dx} \lambda \exp(-\lambda x) I_{\{x \geq 0\}}(x) = -\lambda^2 \exp(-\lambda x) I_{\{x \geq 0\}}(x).$$

2 Expected Values

The expected value, also known as the expectation or mean, of a random variable X is denoted $E(X)$. It is the weighed average of all values X could take, with weights given by

the probabilities of those values. If X is discrete-valued, then

$$E(X) = \sum_x x \cdot P(X = x) = \sum_x x \cdot f(x)\,.$$

If X is a continuous random variable with probability density function (PDF) $f(x)$, we replace the summation with an integral

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

The expectation of a random variable can be thought of as the average value. If, for example, we observed many realizations of a random variable X and took their average, it would be close to $E(X)$.

One nice property of expected values is that they are easy to compute for linear functions of random variables. To see this, let X and Y be random variables with $E(X) = \mu_X$ and $E(Y) = \mu_Y$. Suppose we are interested in a new random variable $Z = aX + bY + c$ where a, b, and c are any real constants. The mean of Z is easy to compute: $E(Z) = E(aX + bY + c) = aE(X) + bE(Y) + c = a\mu_X + b\mu_Y + c$.

We can also compute expectations of functions of X. For example, suppose $g(X) = 2/X$. Then we have $E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx = \int_{-\infty}^{\infty} \frac{2}{x} f(x)dx$. Note however that in general, $E(g(X)) \neq g(E(X))$.

Example: Let's say continuous random variable X has PDF $f(x) = 3x^2 I_{\{0 \leq x \leq 1\}}(x)$. We want to find $E(X)$ and $E(X^2)$. First,

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x \cdot 3x^2 I_{\{0 \leq x \leq 1\}}(x)dx \\ &= \int_0^1 x \cdot 3x^2 dx = \int_0^1 3x^3 dx = \frac{3}{4}x^4|_{x=0}^{x=1} = \frac{3}{4}(1-0) \\ &= \frac{3}{4}\,, \end{aligned} \tag{1}$$

and second,

$$\begin{aligned} E(X^2) &= \int_{-\infty}^{\infty} x^2 \cdot 3x^2 I_{\{0 \leq x \leq 1\}}(x)dx \\ &= \int_0^1 x^2 \cdot 3x^2 dx = \int_0^1 3x^4 dx = \frac{3}{5}x^5|_{x=0}^{x=1} = \frac{3}{5}(1-0) \\ &= \frac{3}{5}\,. \end{aligned} \tag{2}$$

2.1 Variance

The variance of random variable measures how spread out its values are. If X is a random variable with mean $E(X) = \mu$, then the variance is $E[(X - \mu)^2]$. In words, the variance is the expected value of the squared deviation of X from its mean. If X is discrete, this is calculated as

$$Var(X) = \sum_x (x - \mu)^2 \cdot P(X = x)$$

and if X is continuous, it is

$$Var(X) = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x) dx \,.$$

For both discrete and continuous X, a convenient formula for the variance is $Var(X) = E[X^2] - (E[X])^2$. The square root of variance is called the standard deviation.

Variance has a linear property similar to expectation. Again, let X and Y be random variables with $Var(X) = \sigma_X^2$ and $Var(Y) = \sigma_Y^2$. It is also necessary to assume that X and Y are independent. Suppose we are interested in a new random variable $Z = aX + bY + c$ where a, b, and c are any real constants. The variance of Z is then $Var(Z) = Var(aX + bY + c) = a^2 Var(X) + b^2 Var(Y) + 0 = a^2\sigma_X^2 + b^2\sigma_Y^2$. Because c is constant, it has variance 0.

Example: Continuing the previous example, let's say continuous random variable X has PDF $f(x) = 3x^2 I_{\{0 \le x \le 1\}}(x)$. We found in Equations 1 and 2 that $E(X) = 3/4$ and $E(X^2) = 3/5$. Then, $Var(X) = E[X^2] - (E[X])^2 = 3/5 - (3/4)^2 = 3/80$.

3 Additional Discrete Distributions

3.1 Geometric

The geometric distribution is the number of trials needed to get the first success, i.e., the number of Bernoulli events until a success is observed, such as the first head when flipping a coin. It takes values on the positive integers starting with one (since at least one trial is

needed to observe a success).

$$X \sim \text{Geo}(p)$$
$$P(X = x|p) = p(1-p)^{x-1} \text{ for } x = 1, 2, \ldots$$
$$E[X] = \frac{1}{p}$$

If the probability of getting a success is p, then the expected number of trials until the first success is $1/p$.

Example: What is the probability that we flip a fair coin four times and don't see any heads? This is the same as asking what is $P(X > 4)$ where $X \sim Geo(1/2)$. $P(X > 4) = 1 - P(X = 1) - P(X = 2) - P(X = 3) - P(X = 4) = 1 - (1/2) - (1/2)(1/2) - (1/2)(1/2)^2 - (1/2)(1/2)^3 = 1/16$. Of course, we could also have just computed it directly, but here we see an example of using the geometric distribution and we can also see that we got the right answer.

3.2 Multinomial

Another generalization of the Bernoulli and the binomial is the multinomial distribution, which is like a binomial when there are more than two possible outcomes. Suppose we have n trials and there are k different possible outcomes which occur with probabilities $p_1, \ldots, p_k$. For example, we are rolling a six-sided die that might be loaded so that the sides are not equally likely, then n is the total number of rolls, $k = 6$, p_1 is the probability of rolling a one, and we denote by $x_1, \ldots, x_6$ a possible outcome for the number of times we observe rolls of each of one through six, where $\sum_{i=1}^{6} x_i = n$ and $\sum_{i=1}^{6} p_i = 1$.

$$f(x_1, \ldots, x_k|p_1, \ldots, p_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}.$$

Recall that $n!$ stands for n factorial, which is the product of n times $n-1$ times $\ldots 1$, e.g., $4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$. The expected number of observations in category i is np_i.

3.3 Poisson

The Poisson distribution is used for counts, and arises in a variety of situations. The parameter $\lambda > 0$ is the rate at which we expect to observe the thing we are counting.

$$\begin{aligned} X &\sim \text{Pois}(\lambda) \\ P(X = x|\lambda) &= \frac{\lambda^x \exp(-\lambda)}{x!} \text{ for } x = 0, 1, 2, \ldots \\ E[X] &= \lambda \\ Var[X] &= \lambda \end{aligned}$$

A Poisson process is a process wherein events occur on average at rate λ, events occur one at a time, and events occur independently of each other.

Example: Significant earthquakes occur in the Western United States approximately following a Poisson process with rate of two earthquakes per week. What is the probability there will be at least 3 earthquakes in the next two weeks? Answer: the rate per two weeks is $2 \times 2 = 4$, so let $X \sim \text{Pois}(4)$ and we want to know $P(X \geq 3) = 1 - P(X \leq 2) = 1 - P(X = 0) - P(X = 1) - P(X = 2) = 1 - e^{-4} - 4e^{-4} - \frac{4^2 e^{-4}}{2} = 1 - 13e^{-4} = 0.762$. Note that $0! = 1$ by definition.

4 Continuous Distributions

4.1 Exponential

The exponential distribution is often used to model the waiting time between random events. Indeed, if the waiting times between successive events are independent from an $\text{Exp}(\lambda)$ distribution, then for any fixed time window of length t, the number of events occurring in that window will follow a Poisson distribution with mean $t\lambda$.

$$\begin{aligned} X &\sim \text{Exp}(\lambda) \\ f(x|\lambda) &= \lambda e^{-\lambda x} I_{\{x \geq 0\}}(x) \\ E[X] &= \frac{1}{\lambda} \\ Var[X] &= \frac{1}{\lambda^2} \end{aligned}$$

Similar to the Poisson distribution, the parameter λ is interpreted as the rate at which the events occur.

4.2 Gamma

If $X_1, X_2, \ldots, X_n$ are independent (and identically distributed $\text{Exp}(\lambda)$) waiting times between successive events, then the total waiting time for all n events to occur $Y = \sum_{i=1}^{n} X_i$ will follow a gamma distribution with shape parameter $\alpha = n$ and rate parameter $\beta = \lambda$.

$$\begin{aligned} Y &\sim \text{Gamma}(\alpha, \beta) \\ f(y|\alpha, \beta) &= \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y} I_{\{y \geq 0\}}(y) \\ E[Y] &= \frac{\alpha}{\beta} \\ Var[Y] &= \frac{\alpha}{\beta^2} \end{aligned}$$

where $\Gamma(\cdot)$ is the gamma function, a generalization of the factorial function which can accept non-integer arguments. If n is a positive integer, then $\Gamma(n) = (n-1)!$. Note also that $\alpha > 0$ and $\beta > 0$.

The exponential distribution is a special case of the gamma distribution with $\alpha = 1$. The gamma distribution commonly appears in statistical problems, as we will see in this course. It is used to model positive-valued, continuous quantities whose distribution is right-skewed. As α increases, the gamma distribution more closely resembles the normal distribution.

4.3 Uniform

The uniform distribution is used for random variables whose possible values are equally likely over an interval. If the interval is (a, b), then the uniform probability density function (PDF) $f(x)$ is flat for all values in that interval and 0 everywhere else.

$$\begin{aligned} X &\sim \text{Uniform}(a, b) \\ f(x|a, b) &= \frac{1}{b-a} I_{\{a \leq x \leq b\}}(x) \\ E[X] &= \frac{a+b}{2} \\ Var[X] &= \frac{(b-a)^2}{12} \end{aligned}$$

The standard uniform distribution is obtained when $a = 0$ and $b = 1$.

4.4 Beta

The beta distribution is used for random variables which take on values between 0 and 1. For this reason (and other reasons we will see later in the course), the beta distribution is commonly used to model probabilities.

$$\begin{aligned} X &\sim \text{Beta}(\alpha, \beta) \\ f(x|\alpha, \beta) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} I_{\{0<x<1\}}(x) \\ E[X] &= \frac{\alpha}{\alpha+\beta} \\ Var[X] &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \end{aligned}$$

where $\Gamma(\cdot)$ is the gamma function introduced with the gamma distribution. Note also that $\alpha > 0$ and $\beta > 0$. The standard Uniform$(0, 1)$ distribution is a special case of the beta distribution with $\alpha = \beta = 1$.

4.5 Normal

The normal, or Gaussian distribution is one of the most important distributions in statistics. It arises as the limiting distribution of sums (and averages) of random variables. This is due to the Central Limit Theorem, introduced in Section 5. Because of this property, the normal distribution is often used to model the "errors," or unexplained variation of individual observations in regression models.

The standard normal distribution is given by

$$\begin{aligned} Z &\sim \text{N}(0, 1) \\ f(z) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \\ E[Z] &= 0 \\ Var[Z] &= 1 \end{aligned}$$

Now consider $X = \sigma Z + \mu$ where $\sigma > 0$ and μ is any real constant. Then $E(X) = E(\sigma Z + \mu) = \sigma E(Z) + \mu = \sigma \cdot 0 + \mu = \mu$ and $Var(X) = Var(\sigma Z + \mu) = \sigma^2 Var(Z) + 0 = \sigma^2 \cdot 1 = \sigma^2$. Then, X follows a normal distribution with mean μ and variance σ^2 (standard deviation σ)

denoted as

$$X \sim \mathrm{N}(\mu, \sigma^2)$$
$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

The normal distribution is symmetric about the mean μ, and is often described as a "bell-shaped" curve. Although X can take on any real value (positive or negative), more than 99% of the probability mass is concentrated within three standard deviations of the mean.

The normal distribution has several desirable properties. One is that if $X_1 \sim \mathrm{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathrm{N}(\mu_2, \sigma_2^2)$ are independent, then $X_1 + X_2 \sim \mathrm{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. Consequently, if we take the average of n independent and identically distributed (iid) normal random variables,

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

where $X_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$ for $i = 1, 2, \ldots, n$, then

$$\bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right) . \tag{3}$$

4.6 t

If we have normal data, we can use Equation 3 to help us estimate the mean μ. Reversing the transformation from the previous section, we get

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0, 1) \,. \tag{4}$$

However, we may not know the value of σ. If we estimate it from data, we can replace it with $S = \sqrt{\sum_i (X_i - \bar{X})^2/(n-1)}$, the sample standard deviation. This causes the expression (4) to no longer be distributed as standard normal, but as a standard t distribution with $\nu = n-1$ degrees of freedom.

$$Y \sim \mathrm{t}_\nu$$
$$f(y) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi}} \left(1 + \frac{y^2}{\nu}\right)^{-(\frac{\nu+1}{2})}$$
$$E[Y] = 0 \text{ if } \nu > 1$$
$$Var[Y] = \frac{\nu}{\nu - 2} \text{ if } \nu > 2$$

The t distribution is symmetric and resembles the normal distribution, but with thicker tails. As the degrees of freedom increase, the t distribution looks more and more like the standard normal distribution.

5 Central Limit Theorem

The Central Limit Theorem is one of the most important results in statistics, basically saying that with sufficiently large sample sizes, the sample average approximately follows a normal distribution. This underscores the importance of the normal distribution, as well as most of the methods commonly used which make assumptions about the data being normally distributed.

Let's first stop and think about what it means for the sample average to have a distribution. Imagine going to the store and buying a bag of your favorite brand of chocolate chip cookies. Suppose the bag has 24 cookies in it. Will each cookie have the exact same number of chocolate chips in it? It turns out that if you make a batch of cookies by adding chips to dough and mixing it really well, then putting the same amount of dough onto a baking sheet, the number of chips per cookie closely follows a Poisson distribution. (In the limiting case of chips having zero volume, this is exactly a Poisson process.) Thus we expect there to be a lot of variability in the number of chips per cookie. We can model the number of chips per cookie with a Poisson distribution. We can also compute the average number of chips per cookie in the bag. For the bag we have, that will be a particular number. But there may be more bags of cookies in the store. Will each of those bags have the same average number of chips? If all of the cookies in the store are from the same industrial-sized batch, each cookie will individually have a Poisson number of chips. So the average number of chips in one bag may be different from the average number of chips in another bag. Thus we could hypothetically find out the average number of chips for each bag in the store. And we could think about what the distribution of these averages is, across the bags in the store, or all the bags of cookies in the world. It is this distribution of averages that the central limit theorem says is approximately a normal distribution, with the same mean as the distribution for the individual cookies, but with a standard deviation that is divided by the square root of the number of samples in each average (i.e., the number of cookies per bag).

In formal mathematical notation, the Central Limit Theorem says: Let $X_1, \ldots, X_n$ be independent and identically distributed with $E[X_i] = \mu$ and $Var(X_i) = \sigma^2, 0 < \sigma^2 < \infty$.

Then

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \Rightarrow N(0, 1)\,.$$

That is, $\bar{X}_n$ is approximately normally distributed with mean μ and variance σ^2/n or standard deviation $\sigma/\sqrt{n}$.

6 Bayes Theorem for continuous distributions

When dealing with a continuous random variable θ, we can write the conditional density for θ given y as:

$$f(\theta|y) = \frac{f(y|\theta)f(\theta)}{\int f(y|\theta)f(\theta)d\theta}\,.$$

This expression does the same thing that the versions of Bayes' theorem from Lesson 2 do. Because θ is continuous, we integrate over all possible values of θ in the denominator rather than take the sum over these values. The continuous version of Bayes' theorem will play a central role from Lesson 5 on.