CHAPTER 3. Expectation

3.1. Expectation of a Random Variable

The *expectation* of a random variable X – denote by μ_X or $E(X)$ or EX – is the mean (or first moment) of the distribution of the variable. It is defined by $E(X) = \sum_x x f_X(x)$ in the discrete case and $E(X) = \int x f_X(x)dx$ in the continuous case, assuming the sum and integral are well defined. We say that $E(X)$ is the mean of X or sometimes we use the terminology, $E(X)$ is the mean of the distribution of X. The expectation is a one-number summary of the distribution.

Important Notation. We shall sometimes write $E(X) = \int x\, dF(x)$. You should interpret $\int x\, dF(x)$ to mean $\int x\, f(x)dx$ in the continuous case and $\sum_x x\, f(x)$ in the discrete case.

If we repeated the experiment many times and averaged the outcomes, this average will (approximately) be $E(X)$. We will make this last point more precise when we discuss the law of large numbers. The expectation is also the "balancing point" of the mass.

To ensure that $E(X)$ is well defined, we say that $E(X)$ exists if $\int_x |x| dF_X(x) < \infty$. Otherwise we say that the expectation does not exist.

EXAMPLE. 3.1.1. Flip a fair coin two times. Let X be the number of heads. Then, $E(X) = \int x dF_X(x) = \sum_x x f_X(x) = (0 \times f(0)) + (1 \times f(1)) + (2 \times f(2)) = (0 \times (1/4)) + (1 \times (1/2)) + (2 \times (1/4)) = 1$.

EXAMPLE. 3.1.2. Let $X \sim Unif(-1, 3)$. Then, $E(X) = \int x dF_X(x) = \int x f_X(x) dx = \frac{1}{4}\int_{-1}^{3} x dx = 1$.

EXAMPLE. 3.1.3. A random variable has a *Cauchy* distribution if it has density $f_X(x) = \{\pi(1 + x^2)\}^{-1}$. To see that this is indeed a density, let's do the integral:

$$\begin{aligned}\int_{-\infty}^{\infty} f(x)dx &= \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{dx}{1+x^2} \\ &= \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{d\tan^{-1}}{dx}\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\pi}\left[\tan^{-1}(\infty) - \tan^{-1}(-\infty)\right] \\
&= \frac{1}{\pi}\left[\frac{\pi}{2} - \left(-\frac{\pi}{2}\right)\right] = 1.
\end{aligned}
$$

But, using integration by parts, (set $u = x$ and $v = \tan^{-1} x$),

$$
\int |x| dF(x) = \frac{2}{\pi}\int_0^\infty \frac{x\, dx}{1+x^2} = \left[x \ \tan^{-1}(x)\right]_0^\infty - \int_0^\infty \tan^{-1} x\, dx = \infty
$$

so the mean does not exist. If you simulate a Cauchy distribution many times and take the average, you will see that the average never settles down. This is because the Cauchy has thick tails so extreme observations are common.

From now on, whenever we discuss expectations, we implicitly assume that they exist.

Let $Y = r(X)$. How do we compute $E(Y)$? One way is to do a change of variables, find $f_Y(y)$ and then compute $E(Y) = \int y f_Y(y) dy$. But there is an easier way.

THEOREM. 3.1.4. (The rule of the lazy statistician.) Let $Y = r(X)$. Then

$$
E(Y) = \int r(x) dF_X(x).
$$

This result makes intuitive sense. Think of playing a game where we draw $X \sim f_X$ at random. Then I pay you $Y = r(X)$. Your average income is $r(x)$ times the chance that $X = x$, summed (or integrated) over all values of x. This makes it easy to compute $E(Y)$; no change of variables is required. Here is an important special case. Let A be an event and let $r(x) = I_A(x)$ where $I_A(x) = 1$ if $x \in A$ and $I_A(x) = 0$ if $x \notin A$. Then $EI_A(X) = \int I_A(x) f_X(x) dx = \int_A f_X(x) dx = P(X \in A)$. So, probability is a special case of expectation.

THEOREM. 3.1.5. Let $X \sim Unif(0,1)$. Let $Y = r(X) = e^X$. Then,

$$
E(Y) = \int_0^1 e^x f(x) dx = \int_0^1 e^x dx = e - 1.
$$

Alternatively, you could find $f_Y(y)$ which turns out to be $f_Y(y) = 1/y$ for $1 < y < e$. Then, $E(Y) = \int_1^e y\, f(y) dy = e - 1$.

Functions of several variables are handled in a similar way. If $Z = r(X, Y)$ then $E(Z) = \int\int r(x, y)dxdy$.

EXAMPLE. 3.1.6. Let (X, Y) have a jointly uniform distribution on the unit square. Let $Z = r(X, Y) = X^2 + Y^2$. Then,

$$E(Z) = \int_0^1 \int_0^1 (x^2 + y^2)dxdy = \frac{2}{3}.$$

EXAMPLE. 3.1.7. Take a stick of unit length. Break it at random. Let Y be the length of the longer piece. What is the mean (expectation) of Y? Let X be the break point so that $X \sim Unif(0, 1)$. Note that $Y = r(X) = \max\{X, 1 - X\}$. Thus, $r(x) = 1 - x$ when $0 < x < 1/2$ and $r(x) = x$ when $1/2x < 1$. Hence,

$$E(Y) = \int r(x)f(x)dx = \int_0^{1/2} (1 - x)dx + \int_{1/2}^1 xdx = \frac{3}{4}.$$

3.2. Properties of Expectations

Expectations possess the following properties. (Again, we assume that all the expectations are well-defined.)

Expectation is a linear operator: if a and b are constants, then $E(aX + bY) = aE(X) + bE(Y)$. Also, if c is a constant then $E(c) = c$.

More generally, if $X_1, \ldots, X_n$ are random variables and $a_1, \ldots, a_n$ are constants, then

$$E\left(\sum_i a_i X_i\right) = \sum_i a_i E(X_i).$$

EXAMPLE. 3.2.1. Let $X \sim Binomial(n, p)$. What is the mean of X? We could try to appeal to the definition:

$$E(X) = \int xdF_X(x) = \sum_x xf_X(x) = \sum_{x=0}^n x\binom{n}{x}p^x(1 - p)^{n-x}$$

but this is not an easy sum to evaluate. Instead, use the following trick. Let $X_i = 1$ if the i^{th} toss is heads and $X_i = 0$ otherwise. Now, $E(X_i) = p(1) + (1-p)(0) = p$. Also note that $X = \sum_i X_i$. Thus, $E(X) = E(\sum_i X_i) = \sum_i E(X_i) = np$.

THEOREM. 3.2.2. Let $X_1, \ldots, X_n$ be independent random variables. Then,

$$E\left(\prod_{i=1}^{n} X_i\right) = \prod_i E(X_i).$$

Notice that the summation rule does not require independence but the multiplication rule does.

EXAMPLE. 3.2.3. Let $X_1 \sim Bin(1, p_1)$ and $X_2 \sim Bin(1, p_2)$ and suppose that $X_1 \amalg X_2$. Then

$$E(X_1(1-X_2)) = E(X_1 - X_1X_2) = E(X_1) - E(X_1X_2) = p_1 - p_1p_2 = p_1(1-p_2).$$

EXAMPLE. 3.2.4. Suppose we play a game where we start with c dollars. On each play of the game you either double or half your money, with equal probability. What is your expected fortune after n trials? Define

$$X_i = \begin{cases} 2 & \text{if you win on trial } i \\ \frac{1}{2} & \text{if you lose on trial } i. \end{cases}$$

Now,

$$E(X_i) = \left(2 \times \frac{1}{2}\right) + \left(\frac{1}{2} \times \frac{1}{2}\right) = \frac{5}{4}.$$

Let Y_n denote your fortune after n trials. Then, $Y = cX_1 \cdots X_n$. Hence, $E(Y) = cE(X_1 \cdots X_n) = c(5/4)^n$.

3.3. The Variance of a Random Variable

Let X be a random variable with mean $\mu = \mu_X = E(X)$. We define the *variance* of X – denoted by σ^2 or σ_X^2 or $Var(X)$ or $V(X)$ or VX – by $\sigma^2 = E(X - \mu)^2$ (assuming this expectation exists). Imagine drawing X at random. Then, σ^2 measures how far X is from its mean (in squared distance) on average. In a sense, σ^2 measure how "spread out" the distribution of X is. When σ is small, the distribution of X is tightly concentrated around its mean. We define the standard deviation by $\sigma = \sqrt{Var(X)}$.

THEOREM. 3.3.1. Assume that the variance of X is well-defined. Then $Var(X) = 0$ if and only if there is a constant c such that $P(X = c) = 1$. In this case we say that X has a point mass distribution.

PROOF. Suppose that $P(X = c) = 1$. Then $E(X) = c$. Now, $(X - c)^2$ is 0 with probability 1. Thus, $E(X - c)^2 = 0 \times 1 = 0$. Now suppose that $Var(X) = 0$. Let $c = E(X)$ and define $Y = (X - c)^2$. Note that $Y \geq 0$ and has mean 0. Hence, $P(Y = 0) = 1$ i.e. $P((X - c)^2 = 0) = 1$ which implies that $P(X = c) = 1$.

Here are some properties of the variance that follow easily from its definition.

Property 1: If a and b are constants then $Var(aX + b) = a^2 Var(X)$. Hence, if $Y = aX + b$ then $\sigma_Y = a\sigma_X$.

Property 2: The variance can be written as $Var(X) = E(X^2) - \mu^2$.

Property 3: If $X_1, \ldots, X_n$ are independent and $a_1, \ldots, a_n$ are constants, then $Var(\sum_i a_i X_i) = \sum_i a_i^2 Var(X_i)$.

REMARK: The additivity property holds for expectations whether or not the random variables are independent. But for variances, we need independence.

EXAMPLE. 3.3.2. Let $X \sim Bin(n, p)$. Find $Var(X)$. We can solve this as we did with the expectation. We write $X = \sum_i X_i$ where $X_i = 1$ if toss i is heads and $X_i = 0$ otherwise. Then $X = \sum_i X_i$ and the random variables are independent. Also, $P(X_i = 1) = p$ and $P(X_i = 0) = 1 - p$. Recall that $E(X_i) = (p \times 1) + (p \times 0) = p$. Now, $E(X_i^2) = (p \times 1^2) + (p \times 0^2) = p$. So, using property 2, $Var(X_i) = E(X_i^2) - p^2 = p - p^2 = p(1 - p)$. Finally, $Var(X) = Var(\sum_i X_i) = \sum_i Var(X_i) = \sum_i p(1 - p) = np(1 - p)$.

A BIT OF STATISTICS. Let us pursue the last example a bit more. Suppose we toss a possibly biased coin n times and get X heads. How can we estimate the unknown parameter p from the data? Intuitively, we might use $\hat{p} = X/n$. We will justify this estimator later in the course. For now, note that $\hat{p}$ is a random variable. Let us compute its mean and standard deviation. The mean is computed as $E(\hat{p}) = E(X/n) = n^{-1}E(X) = n^{-1}np = p$. Thus, the estimator has the right value as its mean. Now $Var(\hat{p} = Var(X/n) = n^{-2}Var(X) = n^{-2}np(1 - p) = p(1 - p)/n$. Thus, $\sigma_{\hat{p}} = \{p(1 - p)/n\}^{1/2}$. The important thing to notice is that the standard deviation is decreasing in sample size at rate $n^{-1/2}$.

EXAMPLE. 3.3.3. Let $X \sim Unif(a,b)$. Find the variance of X. Of course, $f(x) = c \equiv 1/(b-a)$ for $a < x < b$ and 0 otherwise. Then, $\mu = \int_a^b x f(x) dx = (a+b)/2$. Next,

$$E(X^2) = \int_a^b x^2 f(x) dx = \frac{1}{3}\frac{b^3 - a^3}{(b-a)}.$$

Finally,

$$Var(X) = E(X^2) - \mu^2 = \frac{1}{3}\frac{b^3 - a^3}{(b-a)} - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}.$$

3.4. Moments

The k^{th} moment of X is defined to be $E(X^k)$ assuming that $E(|X|^k) < \infty$.

THEOREM. 3.4.1. If the k^{th} moment exists and if $j < k$ then the j^{th} moment exists.

PROOF. We shall prove the continuous case. We have

$$\begin{aligned} E|X|^j &= \int_{-\infty}^{\infty} |x|^j f_X(x) dx \\ &= \int_{|x| \le 1} |x|^j f_X(x) dx + \int_{|x|>1} |x|^j f_X(x) dx \\ &\le \int_{|x| \le 1} f_X(x) dx + \int_{|x|>1} |x|^k f_X(x) dx \\ &\le 1 + E(|X|^k) < \infty. \end{aligned}$$

The k^{th} central moment is defined to be $E((X-\mu)^k)$. Thus, the first central moment is 0 and the second is just the variance.

3.5. "Equals in Distribution" and Symmetry

Two random variables X and Y are *equal in distribution* if they have the same cdf i.e. $F_X(c) = F_Y(c)$ for all c. We write $X \stackrel{d}{=} Y$. If you and I each generate a number from the same random number generator then our random numbers are equal in distribution. It does not mean that we will get the same number. Rather it means that we will make identical probability statements about our numbers. Avoid the temptation to treat $\stackrel{d}{=}$ like an ordinary equals sign.

A random variable X has a symmetric distribution if $X \stackrel{d}{=} -X$. If X has pdf f_X and X is symmetric, then $f_X(x) = f_X(-x)$. To see this, note that $F_X(x) = P(X \le x) = P(-X \ge -x) = 1 - P(-X \le -x) = 1 - F_{-X}(-x) = 1 - F_X(-x)$ since $F_X(c) = F_{-X}(c)$ for all c. Differentiating we get $f_X(x) = f_X(-x)$.

Let us compute the mean of X. We get

$$
\begin{aligned}
E(X) &= \int_{-\infty}^{\infty} x f_X(x) dx \\
&= \int_{-\infty}^{0} x f_X(x) dx + \int_0^{\infty} x f_X(x) dx \\
&= -\int_0^{\infty} u f_X(-u) du + \int_0^{\infty} x f_X(x) dx \quad u = -x \\
&= -\int_0^{\infty} u f_X(u) du + \int_0^{\infty} x f_X(x) dx \quad \text{symmetry} \\
&= 0.
\end{aligned}
$$

3.6. The Median and the Mean

The median of X is defined by $m = F_X^{-1}(1/2)$. DeGroot defines it slightly differently. One way to think about the median and mean is as solutions to predictions problems. Suppose you have to provide a guess at the random variable X. Let $\mu = E(X)$ and let $\sigma^2 = Var(X)$. What is your best guess of X? The answer depends on how we decide to measure our "loss." Suppose that d is your guess (or prediction) of X. Suppose further that when the value of X is known, you will be penalized $(X - \mu)^2$. The "risk," or average loss is defined as $r(d, \mu) = E(X - d)^2$. If we differentiate r with respect to d and set the derivative equal to 0, we conclude that the risk is minimized by choosing $d = \mu$. In summary, the mean is the best one-number prediction of X assuming squared-error loss.

Now suppose that the loss is $|X - d|$ (mean absolute deviation or MAD). The risk is now $r(d, \mu) = E|X - d|$. It can be shown that the median is the optimal prediction.

3.7. Covariance and Correlation

Let X and Y be random variables with means μ_X and μ_Y and standard deviations σ_X and σ_Y. Define the covariance between X and Y by $Cov(X, Y) = E\left[(X - \mu_X)(Y - \mu_Y)\right]$. The correlation is defined by

$$\rho = \rho_{X,Y} = \rho(X, Y) = \frac{Cov(X, Y)}{\sigma_x \sigma_Y}.$$

THEOREM. 3.7.1. Assuming the appropriate moments exist, $-1 \leq \rho(X,Y) \leq 1$.

LEMMA. 3.7.2. (The Cauchy-Schwartz inequality.) Assuming that U and V possess second moments, $[E(UV)]^2 \leq EU^2 EV^2$.

PROOF. Note that $0 \leq E(aU+bV)^2 = a^2EU^2+b^2EV^2+2abE(UV)$. Set $a = \sqrt{EU^2}$ and $b = \sqrt{EV^2}$ and rearrange the above inequality to conclude that $E(UV) \geq -\sqrt{EU^2EV^2}$. Similarly, by rearranging $0 \leq E(aU-bV)^2 = a^2EU^2 + b^2EV^2 - 2abE(UV)$ conclude that $E(UV) \leq \sqrt{EU^2EV^2}$.

PROOF OF THEOREM 3.7.1. We apply the Cauchy-Schwartz inequality to get

$$Cov(X,Y)^2 = [E(X-\mu_X)(Y-\mu_Y)]^2 \leq E(X-\mu_X)^2E(Y-\mu_Y)^2 = \sigma_X^2\sigma_Y^2.$$

Hence,

$$\rho(X,Y)^2 = \frac{Cov(X,Y)^2}{\sigma_X^2\sigma_Y^2} \leq \frac{\sigma_X^2\sigma_Y^2}{\sigma_X^2\sigma_Y^2} = 1.$$

Properties of Covariance.

(1) $Cov(X,Y) = E(XY) = E(X)E(Y)$.

(2) If $X \amalg Y$ then $Cov(X,Y) = \rho(X,Y) = 0$.

The converse of property (2) is false. There are dependent random variables for which $Cov(X,Y) = 0$. For example, if $P(X=-1, Y=1) = P(X=0, Y=0) = P(X=1, Y=1) = 1/3$, then X and Y are dependent but $Cov(X,Y) = 0$.

We have seen that, for independent random variables, $V(X+Y) = V(X) + V(Y)$. The corresponding formula for non-independent variables follows easily from the definition of covariance.

THEOREM. 3.7.3. $V(X+Y) = V(X)+V(Y)+2Cov(X,Y)$ and $V(X-Y) = V(X) + V(Y) - 2Cov(X,Y)$. More generally, for random variables $X_1, \dots, X_n$,

$$Var\left(\sum_i X_i\right) = \sum_i Var(X_i) + 2\sum\sum_{i<j} Cov(X_i, X_j).$$

3.8. Conditional Expectation

Suppose that X is a random variable with expectation $E(X)$. Now suppose we observe the value of a second random variable Y. How should we change the expectation of X to account for this information?

Before observing Y, we computed the mean of X via $E(X) = \int x f_X(x)dx$. After we observe that $Y = y$, we replace $f_X(x)$ with $f_{X|Y}(x|y)$. We compute the *conditional expectation of* X *given that* $Y = y$ using $f_{X|Y}(x|y)$:

$$E(X|Y = y) = \int x f_{X|Y}(x|y)dx.$$

We shall sometimes also denote $E(X|Y = y)$ by $\mu_X(y)$. Whereas, $E(X)$ is a number, note that $E(X|Y = y)$ is a function of y. The value of $E(X|Y = y)$ depends on the observed value of Y. Before we observe Y, we don't know the value of the conditional expectation so it is a random variable. We denote this random variable by $E(X|Y)$ or $\mu_X(Y)$.

To summarize, $E(X|Y)$ is a random variable; it depends on Y. The value of the random variable $E(X|Y)$ once $Y = y$ is observed is denoted by $E(X|Y = y)$ and is calculated by the formula $E(X|Y = y) = \int x f_{X|Y}(x|y)dx$.

Similarly, we can define $\mu_Y(X) = E(Y|X)$ and $\mu_Y(x) = E(Y|X = x)$. In statistics and machine learning, X often denotes characteristics of a person (age, blood pressure, etc.) and Y denotes some variable we wish to predict (such as life length). In this case, $\mu_Y(x)$ is called the regression of Y on X. In some cases, one assumes that $\mu_Y(x)$ has a simple form, for example linear. This is called parametric regression. In other cases, $\mu_Y(x)$ is only assumed to be some smooth function. This is called nonparametric regression. We shall discuss these issues later.

EXAMPLE. 3.8.1. Suppose we draw $X \sim Unif(0,1)$. After we observe $X = x$, we draw $Y|X = x \sim Unif(x,1)$. Intuitively, we expect that $E(Y|X = x) = (1 + x)/2$. Let's show this. First, $f_{Y|X}(y|x) = 1/(1 - x)$ for $x < y < 1$ and $f_{Y|X}(y|x) = 0$ otherwise. Hence,

$$E(Y|X = x) = \int_x^1 y f_{Y|X}(y|x)dy = \frac{1}{1-x}\int_x^1 ydy = \frac{1+x}{2}$$

as expected. Thus we can write $E(Y|X) = (1 + X)/2$.

THEOREM. 3.8.2. (The rule of iterated expectations.) For random variables X and Y, assuming the expectations exist, we have that

$$E\left[E(Y|X)\right] = E(Y) \quad \text{and} \quad E\left[E(X|Y)\right] = E(X).$$

More generally, for a function $r(x, y)$ we have

$$E\left[E(r(X,Y)|X)\right] = E(r(X,Y)) \quad \text{and} \quad E\left[E(r(X,Y)|X)\right] = E(r(X,Y))$$

where $E\left[E(r(X,Y)|X)\right]$ is the random variable taking value $E\left[E(r(X,Y)|X=x)\right] = \int r(x,y) f_{Y|X}(y|x)dy$ when $X = x$.

PROOF. $E\left[E(Y|X)\right] = \int E(Y|X=x) f_X(x)dx = \int\int y f(y|x) dy f(x) dx = \int\int y f(y|x) f(x) dx dy = \int\int y f(x,y) dx dy = E(Y)$.

What does this theorem mean? Remember that $E(Y|X = x) = \mu_Y(x)$ is a function of x. We can therefore compute its expectation by integrating $\int \mu_Y(x) f_X(x)dx$. When we do this, we get back $E(Y)$.

EXAMPLE. 3.8.3. Consider the previous example. How can we compute $E(Y)$. One method is to find the joint density $f(x,y)$ and then compute $E(Y) = \int\int y f(x,y)dxdy$. An easier way is to do this in two steps. First, we already figured out that $E(Y|X) = (1+X)/2$. Thus, $E(Y) = EE(Y|X) = E((1+X)/2) = (1+E(X))/2 = (1+(1/2))/2 = 3/4$.

The conditional variance is defined as $Var(Y|X=x) = \int (y-\mu_Y(x))^2 f(y|x)dy$.

THEOREM 3.8.4. For random variables X and Y,

$$Var(Y) = EVar(Y|X) + VarE(Y|X).$$

EXAMPLE. 3.8.5. (A two stage model). Draw a county at random from the United States. Then draw n people at random from the county. Let X be the number of those people who have a certain disease. If P denotes the proportion of of people in that county with the disease then P is also a random variable since it varies from county to county. Given $P = p$, we have that $X \sim Bin(n,p)$. Thus, $E(X|P=p) = np$ and $Var(X|P=p) = np(1-p)$. Suppose that the random variable P has a uniform (0,1) distribution. Then, $E(X) = EE(X|P) = E(nP) = nE(P) = n/2$. Let us compute the variance of X. Now, $Var(X) = EVar(X|P) + VarE(X|P)$. Let's compute these two terms. First, $EVar(X|P) = E[nP(1-P)] = nE(P(1-P)) = n\int p(1-p)f(p)dp = n\int_0^1 p(1-p)dp = n/6$. Next, $VarE(X|P) = Var(nP) = n^2 Var(P) = n^2 \int (p-(1/2))^2 dp = n^2/12$. Hence, $Var(X) = (n/6) + (n^2/12)$.

3.9. Special Distributions Revisited

We now compute the moments of some special distributions. Remember that it is important to distinguish the random variable from parameters (constants). For this reasons, we shall write these probability functions in the form $f(x;\theta)$ where θ represents parameters. For convenience, I will define the special distributions again.

Bernoulli. Suppose that X takes values 0 and 1 only. We say X has a Bernoulli distribution. Let $p = P(X = 1)$ and $q = 1 - p = P(X = 0)$. Write $X \sim Ber(p)$. This is just a coin flip. Here p is usually an unknown parameter. We have seen that $E(X) = p$ and $Var(X) = pq$. We can write the probability function as $f(x;p) = p^x q^{1-x}$ for $x = 0, 1$.

Binomial. We already know this one. Flip a coin n times and let X be the number of heads. Then $X \sim Bin(n,p)$ and

$$f(x;n,p) = \binom{n}{x} p^x q^{n-x} \quad \text{for } x = 0, \ldots n.$$

Usually, n is a known parameter and p is an unknown parameter. Recall that $E(X) = np$ and $Var(X) = npq$.

Poisson. A random variable X has a Poisson distribution, denoted by $X \sim Pois(\lambda)$, if

$$f(x;\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \quad x = 0, 1, 2, \ldots$$

Here, $\lambda > 0$ is a parameter. Often, counts have a Poisson distribution. For example, the number of atoms that experience radioactive decay in a lump of uranium, follows a Poisson distribution. If you count the number of errors in computer code, it will have (approximately) a Poisson distribution.

Let us show that $f(x;\lambda)$ sums to one. We will need the following fact.

FACT: (Series expansion for exponential). If a is a real number then

$$e^a = \sum_{x=0}^{\infty} \frac{a^x}{x!} = 1 + a + \frac{a^2}{2!} + \cdots$$

Now,

$$\sum_{x=0}^{\infty} f(x;\lambda) = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda} e^{\lambda} = 1.$$

Let us compute the mean:

$$
\begin{aligned}
E(X) &= \sum_{x=0}^{\infty} x f(x;\lambda) = \sum_{x=0}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!} \\
&= \sum_{x=1}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!} = \lambda \sum_{x=1}^{\infty} x \frac{e^{-\lambda}\lambda^{x-1}}{x!} \\
&= \lambda \sum_{x=1}^{\infty} \frac{e^{-\lambda}\lambda^{x-1}}{(x-1)!} = \lambda \sum_{y=0}^{\infty} \frac{e^{-\lambda}\lambda^{y}}{y!} \qquad y = x-1 \\
&= \lambda \sum_{y=0}^{\infty} f(y;\lambda) = \lambda.
\end{aligned}
$$

To compute the variance we do a trick. By a similar calculation as the one above, one can compute that $E[X(X-1)] = \lambda^2$. But $E[X(X-1)] = E(X^2 - X) = E(X^2) - E(X) = E(X^2) - \lambda$. So, $\lambda^2 = E(X^2) - \lambda$. Thus, $E(X^2) = \lambda^2 + \lambda$. Finally, $Var(X) = E(X^2) - \lambda^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$. Suppose that $X_1, \ldots, X_n$ are independent and that $X_i \sim \text{Pois}(\lambda_i)$. Let $Y = \sum_i X_i$. In the appendix we prove that $Y \sim \text{Pois}(\sum_i \lambda_i)$.

EXAMPLE 3.9.1. Let X be the number of errors per page of computer code. Suppose that $X \sim \text{Pois}(3)$. What is the probability that the number of errors on two pages is 10 or more? We want $P(Y \geq 10)$ where $Y = X_1 + X_2 \sim Pois(6)$. So

$$
P(Y \geq 10) = \sum_{x=10}^{\infty} \frac{e^{-6}6^x}{x!} = 1 - \sum_{x=0}^{9} \frac{e^{-6}6^x}{x!} = .0838.
$$

The last sum can be done numerically or can be obtained from the Table on page 688 of DeGroot.

The Normal (Gaussian). A random variable X has a Normal (or Gaussian) distribution, denoted by $X \sim N(\mu, \sigma^2)$, if it has pdf

$$
f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2} \right\}
$$

where x is any real number, μ is real and $\sigma > 0$. Clearly, $f(x;\mu,\sigma) > 0$ and it can be shown using some calculus tricks that $\int f(x;\mu,\sigma)dx = 1$.

Let $X \sim N(\mu, \sigma)$ and let $Z = (X - \mu)/\sigma$. Then, in the appendix we prove that $Z \sim N(0,1)$. A Normal random variable with $\mu = 0$ and $\sigma = 1$ is called a standard normal random variable. The distribution function for a standard normal is tabulated on page 689 of DeGroot. It is also available in many compute programs. From this table we can compute the following: $P(|Z| \le 1) = .68$, $P(|Z| \le 2) = .95$, $P(|Z| \le 3) = .99$.

The standard normal density is often written as $\phi(z)$ and the cdf is written as $\Phi(z)$ instead of $F(z)$. Note that ϕ is symmetric about 0. It follows that $\Phi(-z) = 1 - \Phi(z)$ which you can see by drawing a picture.

EXAMPLE 3.9.2. Suppose that $X \sim N(\mu, \sigma^2)$ with $\mu = 5$ and $\sigma = 2$. Find $P(1 < X < 8)$. To solve this, let $Z = (X - \mu)/\sigma = (X - 5)/2$ and recall that $Z \sim N(0,1)$. Then,

$$\begin{aligned}
P(1 < X < 8) &= P\left(\frac{1-5}{2} < \frac{X-5}{2} < \frac{8-5}{2}\right) \\
&= P(-2 < Z < 1.5) \\
&= P(Z < 1.5) - P(Z < -2) \\
&= \Phi(1.5) - \Phi(-2) \\
&= \Phi(1.5) - [1 - \Phi(2)] \\
&= .9332 - [1 - .9771] = .9105.
\end{aligned}$$

We wrote $\Phi(-2)$ as $1 - \Phi(2)$ since the table in DeGroot only has the cdf for positive values.

Suppose that $X_1, \ldots, X_n$ are independent and $X_i \sim N(\mu_i, \sigma_i^2)$. Let $Y = b + \sum_i a_i X_i$. Then, $Y \sim N(b + \sum_i a_i \mu_i, \sum_i a_i^2 \sigma_i^2)$. In particular, let $(X_1, \ldots, X_n)$ be iid $N(\mu, \sigma^2)$ and let $\overline{X}_n = n^{-1} \sum_i X_i$. Then, $\overline{X}_n \sim N(\mu, \sigma^2/n)$.

EXAMPLE. 3.9.3. Suppose that $X_i \sim N(\mu, \sigma^2)$, $i = 1, \ldots, n$ are iid and that $\sigma = 3$. How large should n be so that $P(|\overline{X}_n - \mu| \le 1) \ge .95$? We saw that $\overline{X}_n \sim N(\mu, \sigma^2/n) = N(\mu, 9/n)$. Hence,

$$Z = \frac{\overline{X}_n - \mu}{\frac{3}{\sqrt{n}}} \sim N(0,1).$$

So,

$$P\left(|\overline{X}_n - \mu| \leq 1\right) = P\left(\frac{\overline{X}_n - \mu}{\frac{3}{\sqrt{n}}} \leq \frac{1}{\frac{3}{\sqrt{n}}}\right)$$
$$= P\left(|Z| \leq \frac{\sqrt{n}}{3}\right)$$

and we want this to be at least .95. From the Normal table in DeGroot we see that $P(|Z| \leq 1.96) = .95$ (do you see why?) So we set $\sqrt{n}/3 = 1.96$ to conclude that $n = 34.6 \approx 35$.

The Gamma Function. For any $\alpha > 0$ define

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx.$$

The function $\Gamma(\alpha)$ is called the Gamma function. You can't do this integral in closed form but it can be evaluated numerically. Anyway, lots of properties about $\Gamma(\alpha)$ are known. Here we state some handy properties:

(1) $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$.

(2) $\Gamma(1) = 1$.

(3) If n is an integer, then $\Gamma(n) = (n-1)!$. Thus, the Gamma function is a generalization of the factorial function.

(4) If n is an integer, then

$$\Gamma\left(n + \frac{1}{2}\right) = \left(n - \frac{1}{2}\right)\left(n - \frac{3}{2}\right) \cdots \frac{1}{2}\Gamma\left(\frac{1}{2}\right).$$

(5) $\Gamma(1/2) = \sqrt{\pi}$.

All these facts are easy to prove by doing some integration though we shall not pursue the details.

The Gamma Distribution. A random variable X has a Gamma distribution with parameters α and β, denoted by $X \sim \mathit{Gamma}(\alpha, \beta)$, if $X > 0$ and

$$f(x;\ \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \quad \text{for } x > 0.$$

To see that this is a pdf,

$$\int_0^\infty f(x;\ \alpha, \beta) dx = \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-\beta x} dx$$

$$
\begin{aligned}
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty \left(\frac{u}{\beta}\right)^{\alpha-1} e^{-u}\frac{du}{\beta} \qquad [u = \beta x] \\
&= \frac{1}{\Gamma(\alpha)} \int_0^\infty u^{\alpha-1} e^{-u} du = \frac{\Gamma(\alpha)}{\Gamma(\alpha)} = 1.
\end{aligned}
$$

Now let's get the mean:

$$
\begin{aligned}
E(X) &= \int_0^\infty f(x;\ \alpha, \beta) dx \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x x^{\alpha-1} e^{-\beta x} dx \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha} e^{-\beta x} dx \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+1)}{\beta^{\alpha+1}} \int_0^\infty \frac{\beta^{\alpha+1}}{\Gamma(\alpha+1)} x^{(\alpha+1)-1} e^{-\beta x} dx \\
&= \frac{\alpha\Gamma(\alpha)}{\beta\Gamma(\alpha)} \int_0^\infty f(x; \alpha+1, \beta) dx \\
&= \frac{\alpha}{\beta}.
\end{aligned}
$$

By a similar calculation we conclude that

$$
E(X^k) = \frac{\alpha(\alpha+1)\cdots(\alpha+k-1)}{\beta^k}.
$$

Thus, $Var(X) = E(X^2) - \mu^2 = \alpha/\beta^2$. Now let $X_1, \ldots, X_n$ be independent with $X_i \sim Gamma(\alpha_i, \beta)$. Let $Y = \sum_i X_i$. Then, in the appendix we prove that $Y \sim \Gamma(\sum_i \alpha_i, \beta)$.

The Exponential Distribution. We say that X has an exponential distribution, written $X \sim exp(\beta)$ if

$$
f(x; \beta) = \beta e^{-x\beta} \quad x > 0.
$$

Note that this is just a special case of the Gamma distribution with $\alpha = 1$. Hence, we can immediately conclude that $E(X) = 1/\beta$, $Var(X) = 1/\beta^2$ and $\psi(t) = \beta/(\beta - t)$ for $t < \beta$. Hence, if $X_1, \ldots, X_n$ are iid $exp(\beta)$ then $\sum_i X_i \sim \Gamma(n, \beta)$.

Exponential and Gamma distributions are sometimes used as models for lifetimes. The exponential distribution has an intriguing property called the "memoryless property." Define the "survivor function" by $S(t) = P(X > t) = 1 - F(t)$. Suppose $X \sim exp(\beta)$. Then $S(t) = P(X > t) = \int_t^\infty \beta e^{-\beta x} dx = e^{-\beta t}$. Let t and h be positive numbers. Then,

$$\begin{aligned} P(X > t+h | X > t) &= \frac{P(X > t+h, X > t)}{P(X > t)} \\ &= \frac{P(X > t+h)}{P(X > t)} \\ &= \frac{S(t+h)}{S(t)} \\ &= e^{-\beta h} = S(h) = P(X > h). \end{aligned}$$

Thus, for an exponential distribution, the distribution of remaining lifetime does not depend on how long one has been alive. This is false for humans but approximately true for some electronic components.

The Beta Distribution. We say that X has a Beta distribution, written, $X \sim Beta(\alpha, \beta)$, if

$$f(x; \alpha, \beta) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} & 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

The parameters are both required to be positive. It can be show that $\int_0^1 f(x; \alpha, \beta) dx = 1$ so this is a pdf. We can compute the mean as follows:

$$\begin{aligned} E(X) &= \int_0^1 x f(x; \alpha, \beta) dx \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 x^\alpha (1-x)^{\beta-1} dx \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)} \int_0^1 \frac{\Gamma(\alpha+\beta+1)}{\Gamma(\alpha+1)\Gamma(\beta)} x^{(\alpha+1)-1}(1-x)^{\beta-1} dx \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)} \int_0^1 f(x; \alpha+1, \beta) dx \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\alpha\Gamma(\alpha)\Gamma(\beta)}{(\alpha+\beta)\Gamma(\alpha+\beta)} \end{aligned}$$

$$= \frac{\alpha}{\alpha+\beta}.$$

By similar reasoning one gets

$$Var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

The Multinomial Distribution. Suppose there are k possible outcomes in an experiment. For example, suppose we draw a ball from an urn and the ball can be one of four colors. Then $k = 4$. Suppose we take n independent draws and let $X = (X_1, \ldots, X_k)$ where X_1 is the number of draws that were color 1, where X_2 is the number of draws that were color 2, etc. In this case our random variable is actually a random vector. If $k = 2$ this reduces to a binomial situation (heads or tails). Let $p = (p_1, \ldots, p_k)$ where $p_i \geq 0$ and $\sum_i p_i = 1$. Here, p_i is the probability of getting a ball of color i. It can be shown that the probability mass function for X is

$$f(x; n, p) = P(X = x) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}.$$

Note that the marginal distribution of any one component of the vector is binomial, that is $X_i \sim Bin(n, p_i)$. Thus, $E(X_i) = np_i$ and $Var(X_i) = np_i(1-p_i)$. It is also interesting to compute $Cov(X_i, X_j)$. To do so, we use a trick. Note that $X_i + X_j \sim Bin(n, p_i + p_j)$. Thus, $Var(X_i + X_j) = n(p_i+p_j)(1-[p_i+p_j])$. On the other hand, using the formula for the variance of a sum, we have that $Var(X_i+X_j) = Var(X_i)+Var(X_i)+2Cov(X_i, X_j) = np_i(1-p_i) + np_j(1-p_j) + 2Cov(X_i, X_j)$. If you equate this formula with $n(p_i + p_j)(p_i + p_j)$ and solve, one gets $Cov(X_i, X_j) = -np_ip_j$.

APPENDIX 3. Moment Generating Functions

The moment generating function (mgf), or Laplace transform, of X is defined by $\psi_X(t) = Ee^{tX}$ where t varies over the real numbers. In what follows, we assume that the mgf is well defined for all t in small neighborhood of 0. It need not be defined for all t to be useful. When the mgf is well defined in this sense, it can be shown that we change interchange the operations of differentiation and "taking expectation." This leads to

$$\psi'(0) = \left[\frac{d}{dt} Ee^{tX}\right]_{t=0}$$

$$
\begin{aligned}
&= E\left[\frac{d}{dt}e^{tX}\right]_{t=0} \\
&= E\left[Xe^{tX}\right]_{t=0} = E(X).
\end{aligned}
$$

By taking for derivatives we conclude that $\psi^{(k)}(0) = E(X^k)$. This gives us a method for computing the moments of a distribution.

EXAMPLE. Let X have pdf

$$
f_X(x) = \begin{cases} e^{-x} & x > 0 \\ 0 & \text{otherwise.} \end{cases}
$$

For any $t < 1$ we have

$$
\begin{aligned}
\psi_X(t) &= Ee^{tX} = \int_0^\infty e^{tx}e^{-x}dx \\
&= \int_0^\infty e^{(t-1)x}dx = \frac{1}{1-t}.
\end{aligned}
$$

The integral is divergent if $t \geq 1$. So, $\psi_X(t) = 1/(1-t)$ for all $t < 1$. Now, $\psi'(0) = 1$ and $\psi''(0) = 2$. Hence, $E(X) = 1$ and $Var(X) = E(X^2) - \mu^2 = 2 - 1 = 1$.

Properties of the mgf.

(1) If $Y = aX + b$ then $\psi_Y(t) = e^{bt}\psi_X(at)$.

(2) If $X_1, \ldots, X_n$ are independent and $Y = \sum_i X_i$ then $\psi_Y(t) = \prod_i \psi_i(t)$ where ψ_i is the mgf of X_i.

EXAMPLE. Let $X \sim Bin(n,p)$. As before we know that $X = \sum_i X_i$ where $P(X_i = 1) = p$ and $P(X_i = 0) = 1 - p$. Now $\psi_i(t) = Ee^{X_i t} = (p \times e^t) + ((1-p)) = pe^t + q$ where $q = 1 - p$. Thus, $\psi_X(t) = \prod_i \psi_i(t) = (pe^t + q)^n$.

Here is the reason why mgf's are important.

THEOREM. Let X and Y be random variables. If $\psi_X(t) = \psi_Y(t)$ for all t in an open interval around 0, then $X \overset{d}{=} Y$.

This means that when the mgf exists, it completely characterizes the distribution.

EXAMPLE. Let $X \sim Bin(n_1, p)$ and $X \sim Bin(n_2, p)$ be independent. Let $Y = X_1 + X_2$. Now

$$\psi_Y(t) = \psi_1(t)\psi_2(t) = (pe^t + q)^{n_1}(pe^t + q)^{n_2} = (pe^t + q)^{n_1+n_2}$$

and we recognize the latter as the mgf of a $Bin(n_1+n_2, p)$ distribution. Since, by the last theorem, the mgf characterizes the distribution (i.e. there can't be another random variable which just happens to have the same mgf) we conclude that $Y \sim Bin(n_1 + n_2, p)$.

Let us also compute the mgf of some distributions. For the Poisson:

$$\begin{aligned}
\psi(t) &= Ee^{Xt} \\
&= \sum_{x=0}^{\infty} e^{tx} \frac{e^{-\lambda}\lambda^x}{x!} \\
&= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!} \\
&= e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t-1)}.
\end{aligned}$$

Suppose now that $X_1, \ldots, X_n$ are independent and that $X_i \sim Pois(\lambda_i)$. Let $Y = \sum_i X_i$. Then,

$$\psi_Y(t) = \prod_i \psi_i(t) = \prod_i e^{\lambda_i(e^t-1)} = e^{\sum_i \lambda_i(e^t-1)}.$$

The latter expression is the mgf for a Poisson with parameter $\sum_i \lambda_i$. We have thus proved that $\sum_i X_i \sim Pois(\sum_i \lambda_i)$.

Let's compute the mgf of a Normal:

$$\begin{aligned}
\psi(t) &= \int e^{xt} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\} \\
&= \exp\left\{\mu t + \frac{\sigma^2 t^2}{2}\right\} \int \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\frac{(x-\tilde{\mu})^2}{\sigma^2}\right\} dx
\end{aligned}$$

where $\tilde{\mu} = \mu + t\sigma^2$. To see that the last line equals the line before it, substitute in $\tilde{\mu} = \mu + t\sigma^2$ and combine all the terms. It is tedious but straightforward. Now, the term in the integral is a Normal density and hence it integrates to 1. Thus,

$$\psi(t) = \exp\left\{\mu t + \frac{\sigma^2 t^2}{2}\right\}.$$

Now $E(X) = \psi'(0) = \mu$, $Var(X) = E(X^2) - \mu^2 = \psi''(0) - \mu^2 = \sigma^2$. So the mean and variance are just μ and σ^2. Let $Y = aX + b$ where $a \neq 0$. Then,

$$\psi_Y(t) = e^{bt}\psi_X(at) = \exp\left\{(a\mu + b)t + \frac{1}{2}a^2\sigma^2t^2\right\}.$$

Therefore, $Y \sim N(a\mu + b, a^2\sigma^2)$.

For a Gamma, the mgf can be obtained by integration and we get

$$\psi(t) = \left(\frac{\beta}{\beta - t}\right)^{\alpha}$$

for $t < \beta$. Now let $X_1, \ldots, X_n$ be independent with $X_i \sim Gamma(\alpha_i, \beta)$. Let $Y = \sum_i X_i$. Then,

$$\psi_Y(t) = \prod_i \psi_i(t) = \left(\frac{\beta}{\beta - t}\right)^{\sum_i \alpha_i}.$$

Hence, $Y \sim \Gamma(\sum_i \alpha_i, \beta)$.