Theorem 2.1 *Let X be a random variable and let a, b, and c be constants. Then for any functions $g_1(x)$ and $g_2(x)$ whose expectations exist,*

a. $E(ag_1(X) + bg_2(X) + c) = aEg_1(X) + bEg_2(X) + c.$

b. If $g_1(x) \geq 0$ *for all* x*, then* $Eg_1(X) \geq 0.$

c. If $g_1(x) \geq g_2(x)$ *for all* x*, then* $Eg_1(X) \geq Eg_2(X).$

d. If $a \leq g_1(x) \leq b$ *for all* x*, then* $a \leq Eg_1(X) \leq b.$

Example 2.4 *(Minimizing distance) Find the value of b which minimizes the distance $E(X-b)^2$.*

$$\begin{aligned} E(X-b)^2 &= E(X - EX + EX - b)^2 \\ &= E(X-EX)^2 + (EX-b)^2 + 2E((X-EX)(EX-b)) \\ &= E(X-EX)^2 + (EX-b)^2. \end{aligned}$$

Hence $E(X-b)^2$ is minimized by choosing $b = EX$.

When evaluating expectations of nonlinear functions of X, we can proceed in one of two ways. From the definition of $Eg(X)$, we could directly calculate

$$Eg(X) = \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

But we could also find the pdf $f_Y(y)$ of $Y = g(X)$ and we would have

$$Eg(X) = EY = \int_{-\infty}^{\infty} y f_Y(y) dy.$$

3 Moments and moment generating functions

Definition 3.1 *For each integer n, the n^{th} moment of X (or $F_X(x)$), μ'_n, is*

$$\mu'_n = EX^n.$$

The n^{th} central moment of X, μ_n, is

$$\mu_n = E(X-\mu)^n,$$

where $\mu = \mu'_1 = EX$.

Theorem 3.1 *The variance of a random variable X is its second central moment, $VarX = E(X - EX)^2$. The positive square root of $VarX$ is the standard deviation of X.*

Example 3.1 *(Exponential variance) Let $X \sim$ exponential(λ). We have calculated $EX = \lambda$, then*

$$VarX = E(X-\lambda)^2 = \int_0^\infty (x-\lambda)^2 \frac{1}{\lambda} e^{-x/\lambda} dx = \lambda^2.$$

Theorem 3.2 *If X is a random variable with finite variance, then for any constants a and b,*

$$Var(aX+b) = a^2\, VarX.$$

The variance can be calculated using an alternative formula:

$$\mathrm{Var}X = EX^2 - (EX)^2.$$

Example 3.2 *(Binomial variance) Let $X \sim$ binomial(n,p), that is,*

$$P(X=x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n.$$

We have known $EX = np$. Now we calculate

$$\begin{aligned} EX^2 &= \sum_{x=0}^{n} x^2 \binom{n}{x} p^x (1-p)^{n-x} \\ &= n \sum_{x=1}^{n} x \binom{n-1}{x-1} p^x (1-p)^{n-x} \\ &= n \sum_{y=0}^{n-1} (y+1) \binom{n-1}{y} p^{y+1} (1-p)^{n-1-y} \\ &= np \sum_{y=0}^{n-1} y \binom{n-1}{y} p^{y+1} (1-p)^{n-1-y} + np \sum_{y=0}^{n-1} \binom{n-1}{y} p^{y+1} (1-p)^{n-1-y} \\ &= np[(n-1)p] + np. \end{aligned}$$

Hence, the variance is:

$$VarX = n(n-1)p^2 + np - (np)^2 = np(1-p).$$

The moment generating function (mgf), as its name suggests, can be used to generate moments. In practice, it is easier in many cases to calculate moments directly than to use the mgf. However, the main use of the mdf is not to generate moments, but to help in characterizing a distribution. This property can lead to some extremely powerful results when used properly.

Definition 3.2 *Let X be a random variable with cdf F_X. The moment generating function (mgf) of X (or F_X), denoted by $M_X(t)$, is*

$$M_X(t) = Ee^{tX},$$

provided that the expectation exists for t in some neighborhood of 0. That is, there is an h such that, for all t in $-h < t < h$, Ee^{tX} exists. If the expectation does not exist in a neighbor of 0, we say that the moment generating function does not exist.

More explicitly, we can write the mgf of X as

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x)dx \quad \text{if } X \text{ is continuous,}$$

or

$$M_X(t) = \sum_x e^{tx} P(X = x) \quad \text{if } X \text{ is discrete.}$$

Theorem 3.3 *If X has mgf $M_X(t)$, then*

$$EX^n = M_X^{(n)}(0),$$

where we define

$$M_X^{(n)}(0) = \frac{d^n}{dt^n} M_X(t)|_{t=0}.$$

That is, the n^{th} moment is equal to the n^{th} derivative of $M_X(t)$ evaluated at $t = 0$.

Proof: Assuming that we can differentiate under the integral sign, we have

$$\begin{aligned}\frac{d}{dt} M_X(t) &= \frac{d}{dt}\int_{-\infty}^{\infty} e^{tx} f_X(x)dx \\ &= \int_{-\infty}^{\infty} (\frac{d}{dt} e^{tx}) f_X(x)dx \\ &= \int_{-\infty}^{\infty} (xe^{tx}) f_X(x)dx \\ &= EXe^{tX}.\end{aligned}$$

Thus, $\frac{d}{dt} M_X(t)|_{t=0} = EXe^{tX}|_{t=0} = EX$. Proceeding in an analogous manner, we can establish that

$$\frac{d^n}{dt^n} M_X(t)|_{t=0} = EX^n e^{tX}|_{t=0} = EX^n.$$

□

Example 3.3 *(Gamma mgf) The gamma pdf is*

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta}, \quad 0 < x < \infty, \quad \alpha > 0, \quad \beta > 0,$$

where $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt$ *denotes the gamma function. The mgf is given by*

$$\begin{aligned}
M_X(t) &= \frac{1}{\Gamma(\alpha)\beta^\alpha}\int_0^\infty e^{tx}x^{\alpha-1}e^{-x/beta}dx \\
&= \frac{1}{\Gamma(\alpha)\beta^\alpha}\int_0^\infty x^{\alpha-1}e^{-x/(\frac{\beta}{1-\beta t})}dx \\
&= \frac{1}{\Gamma(\alpha)\beta^\alpha}\Gamma(\alpha)(\frac{\beta}{1-\beta t})^\alpha \\
&= (\frac{1}{1-\beta t})^\alpha \quad \text{if } t < \tfrac{1}{\beta}
\end{aligned}$$

If $t \geq 1/\beta$*, then the quantity* $1-\beta t$ *is nonpositive and the integral is infinite. Thus, the mgf of the gamma distribution exists only if* $t < 1/\beta$*.*

The mean of the gamma distribution is given by

$$EX = \frac{d}{dt}M_X(t)|_{t=0} = \frac{\alpha\beta}{(1-\beta t)^{\alpha+1}}|_{t=0} = \alpha\beta.$$

Example 3.4 *(Binomial mgf) The binomial mgf is*

$$\begin{aligned}
M_X(t) &= \sum_{x=0}^{n} e^{tx}\binom{n}{x}p^x(1-p)^{n-x} \\
&= \sum_{x=0}^{n}(pe^t)^x(1-p)^{n-x}
\end{aligned}$$

The binomial formula gives

$$\sum_{x=0}^{n}\binom{n}{x}u^x v^{n-x} = (u+v)^n.$$

Hence, letting $u = pe^t$ *and* $v = 1-p$*, we have*

$$M_X(t) = [pe^t + (1-p)]^n.$$

The following theorem shows how a distribution can be characterized.

Theorem 3.4 *Let* $F_X(x)$ *and* $F_Y(y)$ *be two cdfs all of whose moments exist.*

a. If X *and* Y *have bounded supports, then* $F_X(u) = F_Y(u)$ *for all* u *if and only if* $EX^r = EY^r$ *for all integers* $r = 0, 1, 2, \ldots$*.*

b. If the moment generating functions exist and $M_X(t) = M_Y(t)$ *for all* t *in some neighborhood of 0, then* $F_X(u) = F_Y(u)$ *for all* u*.*

Theorem 3.5 *(Convergence of mgfs) Suppose* $\{X_i, i = 1, 2, \ldots\}$ *is a sequence of random variables, each with mgf* $M_{X_i}(t)$*. Furthermore, suppose that*

$$\lim_{i\to\infty} M_{X_i}(t) = M_X(t), \quad \text{for all } t \text{ in a neighborhood of 0,}$$

and $M_X(t)$ is an mgf. Then there is a unique cdf F_X whose moments are determined by $M_X(t)$ and, for all x where $F_X(x)$ is continuous, we have

$$\lim_{i\to\infty} F_{X_i}(x) = F_X(x).$$

That is, convergence, for $|t| < h$, of mgfs to an mgf implies convergence of cdfs.

Before going to the next example, we first mention an important limit result, one that has wide applicability in statistics. The proof of this lemma may be found in many standard calculus texts.

Lemma 3.1 *Let $a_1, a_2, \ldots$ be a sequence of numbers converging to a, that is, $\lim_{n\to\infty} a_n = a$. Then*

$$\lim_{n\to\infty} (1 + \frac{a_n}{n})^n = e^a.$$

Example 3.5 *(Poisson approximation) The binomial distribution is characterized by two quantities, denoted by n and p. It is taught that the Poisson approximation is valid "when n is large and np is small," and rules of thumb are sometimes given.*

The Poisson(λ) pmf is given by

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

where λ is a positive constant. The approximation states that if $X \in$ binomial(n, p) and $Y \sim$ Poisson(λ), with $\lambda = np$, then

$$P(X = x) \approx P(Y = x)$$

for large n and small np. We now show the mgfs converge, lending credence to this approximation. Recall that

$$M_X(t) = [pe^t + (1 - p)]^n.$$

For the Poisson(λ) distribution, we can calculate

$$M_Y(t) = e^{\lambda(e^t - 1)},$$

and if we define $p = \lambda/n$, then

$$M_X(t) = [1 + \frac{1}{n}(e^t - 1)(np)]^n = [1 + \frac{1}{n}(e^t - 1)\lambda]^n.$$

Now set $a_n = a = (e^t - 1)\lambda$, and apply the above lemma to get

$$\lim_{n\to\infty} M_X(t) = e^{\lambda(e^t - 1)} = M_Y(t).$$

The Poisson approximation can be quite good even for moderate p and n.

Theorem 3.6 *For any constants a and b, the mgf of the random variable $aX+b$ is given by*

$$M_{aX+b}(t) = e^{bt} M_X(at).$$

PROOF: By definition

$$\begin{aligned} M_{aX+b}(t) &= Ee^{(aX+b)t} = e^{bt} Ee^{(at)X} \\ &= e^{bt} M_X(at). \end{aligned}$$

□

4 Differentiating under an integral sign

The purpose of this section is to characterize conditions under which this operation is legitimate. We will also discuss interchanging the order of differentiation and summation. Many of these conditions can be established using standard theorems from calculus and detailed proofs can be found in most calculus textbooks. Thus, detailed proofs will not be presented here.

Theorem 4.1 *(Leibnitz's Rule) If $f(x,\theta)$, $a(\theta)$, and $b(\theta)$ are differentiable with respect to θ, then*

$$\frac{d}{d\theta}\int_{a(\theta)}^{b(\theta)} f(x,\theta)dx = f(b(\theta),\theta)\frac{d}{d\theta}b(\theta) - f(a(\theta),\theta)\frac{d}{d\theta}a(\theta) + \int_{a(\theta)}^{b(\theta})\frac{\partial}{\partial\theta}f(x,\theta)dx.$$

Notice that if $a(\theta)$ and $b(\theta)$ are constants, we have a special case of Leibnitz's rule:

$$\frac{d}{d\theta}\int_a^b f(x,\theta)dx = \int_a^b \frac{\partial}{\partial\theta}f(x,\theta)dx.$$

Thus, in general, if we have the integral of a differentiable function over a finite range, differentiation of the integral poses no problem. If the range of integration is infinite, however, problems can arise.

The question of whether interchanging the order of differentiation and integration is justified is really a question of whether limits and integration can be interchanged, since a derivative is a special kind of limit. Recall that if $f(x,\theta)$ is differentiable, then

$$\frac{\partial}{\partial\theta}f(x,\theta) = \lim_{\delta\to 0}\frac{f(x,\theta+\delta)-f(x,\theta)}{\delta},$$

so we have

$$\int_{-\infty}^{\infty}\frac{\partial}{\partial\theta}f(x,\theta)dx = \int_{-\infty}^{\infty}\lim_{\delta\to 0}\left(\frac{f(x,\theta+\delta)-f(x,\theta)}{\delta}\right)dx,$$

while

$$\frac{d}{d\theta}\int_{-\infty}^{\infty} f(x,\theta)dx = \lim_{\delta\to 0}\int_{-\infty}^{\infty}\left(\frac{f(x,\theta+\delta)-f(x,\theta)}{\delta}\right)dx.$$

The following theorems are all corollaries of Lebesgue's Dominated Convergence Theorem.

Theorem 4.2 *Suppose the function $h(x, y)$ is continuous at y_0 for each x, and there exists a function $g(x)$ satisfying*

i. $|h(x, y)| \leq g(x)$ *for all x and y.*

ii. $\int_{-\infty}^{\infty} g(x)dx < \infty$.

Then

$$\lim_{y \to y_0} \int_{-\infty}^{\infty} h(x, y)dx = \int_{-\infty}^{\infty} \lim_{y \to y_0} h(x, y)dx.$$

The key condition in this theorem is the existence of a dominating function $g(x)$, with a finite integral, which ensures that the integrals cannot be too badly behaved.

Theorem 4.3 *Suppose $f(x, \theta)$ is differentiable at $\theta = \theta_0$, that is,*

$$\lim_{\delta \to 0} \frac{f(x, \theta_0 + \delta) - f(x, \theta_0)}{\delta} = \frac{\partial}{\partial \theta} f(x, \theta)|_{\theta = \theta_0}$$

exists for every x, and there exists a function $g(x, \theta_0)$ and a constant $\delta_0 > 0$ such that

i.

$$\left|\frac{f(x, \theta_0 + \delta) - f(x, \theta_0)}{\delta}\right| \leq g(x, \theta_0), \quad \textit{for all } x \textit{ and } |\delta| \leq \delta_0,$$

ii.

$$\int_{-\infty}^{\infty} g(x, \theta_0)dx < \infty.$$

Then

$$\frac{d}{d\theta} \int_{-\infty}^{\infty} f(x, \theta)dx\Big|_{\theta = \theta_0} = \int_{-\infty}^{\infty} \Big[\frac{\partial}{\partial \theta} f(x, \theta)|_{\theta = \theta_0}\Big]dx. \tag{4}$$

It is important to realize that although we seem to be treating θ as a variable, the statement of the theorem is for one value of θ. That is, for each value θ_0 for which $f(x, \theta)$ is differentiable at θ_0 and satisfies conditions (i) and (ii), the order of integration and differentiation can be interchanged. Often the distinction between θ and θ_0 is not stressed and (4) is written

$$\frac{d}{d\theta} \int_{-\infty}^{\infty} f(x, \theta)dx = \int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} f(x, \theta)dx. \tag{5}$$

Example 4.1 *(Interchanging integration and differentiation-I) Let X have the exponential(λ) pdf given by $f(x) = \frac{1}{\lambda} e^{-x/\lambda}$, $0 < x < \infty$, and suppose we want to calculate*

$$\frac{d}{d\lambda} EX^n = \frac{d}{d\lambda} \int_0^{\infty} x^n \frac{1}{\lambda} e^{-x/\lambda} dx$$

for integer $n > 0$. If we could move the differentiation inside the integral, we would have

$$\begin{aligned}\frac{d}{d\lambda}EX^n &= \int_0^\infty \frac{\partial}{\partial\lambda} x^n \frac{1}{\lambda} e^{-x/\lambda} dx \\ &= \int_0^\infty \frac{x^n}{\lambda^2}\left(\frac{x}{\lambda} - 1\right) e^{-x/\lambda} dx \\ &= \frac{1}{\lambda^2} EX^{n+1} - \frac{1}{\lambda} EX^n.\end{aligned}$$

To justify the interchange of integration and differentiation, we bound the derivative as follows.

$$\left|\frac{\partial}{\partial\lambda}\left(x^n \frac{1}{\lambda} e^{-x/\lambda}\right)\right| = \frac{x^n e^{-x/\lambda}}{\lambda^2}\left|\frac{x}{\lambda} - 1\right| \leq \frac{x^n e^{-x/\lambda}}{\lambda^2}\left(\frac{x}{\lambda} + 1\right),$$

since $x/\lambda > 0$. For some constant δ_0 satisfying $0 < \delta_0 < \lambda$, take

$$g(x, \lambda) = \frac{x^n e^{-x/(\lambda+\delta_0)}}{(\lambda - \delta_0)^2}\left(\frac{x}{\lambda - \delta_0} + 1\right).$$

We then have

$$\left|\frac{\partial}{\partial\lambda}\left(x^n \frac{1}{\lambda} e^{-x/\lambda}\right)\Big|_{\lambda=\lambda'}\right| \leq g(x, \lambda) \quad \textit{for all } \lambda' \textit{ such that } |\lambda' - \lambda| \leq \delta_0.$$

Since the exponential distribution has all of its moments, $\int_{-\infty}^\infty g(x, \lambda) dx < \infty$ as long as $\lambda - \delta_0 > 0$, so the interchange of integration and differentiation is justified.

Note that this example gives us a recursion relation for the moments of the exponential distribution,

$$EX^{n+1} = \lambda EX^n + \lambda^2 \frac{d}{d\lambda} EX^n.$$

Example 4.2 *(Interchanging summation and differentiation) Let X be a discrete random variable with the geometric distribution*

$$P(X = x) = \theta(1-\theta)^x, \quad x = 0, 1, \ldots, \quad 0 < \theta < 1.$$

We have that $\sum_{x=0}^\infty \theta(1-\theta)^x = 1$ and, provided that the operations are justified,

$$\begin{aligned}\frac{d}{d\theta}\sum_{x=0}^\infty \theta(1-\theta)^x &= \sum_{x=0}^\infty \frac{d}{d\theta}\theta(1-\theta)^x \\ &= \sum_{x=0}^\infty [(1-\theta)^x - \theta x(1-\theta)^{x-1}] \\ &= \frac{1}{\theta}\sum_{x=0}^\infty \theta(1-\theta)^x - \frac{1}{1-\theta}\sum_{x=0}^\infty x\theta(1-\theta)^x.\end{aligned}$$

Since $\sum_{x=0}^\infty \theta(1-\theta)^x = 1$ for all $0 < \theta < 1$, its derivative is 0. So we have

$$\frac{1}{\theta}\sum_{x=0}^\infty \theta(1-\theta)^x - \frac{1}{1-\theta}\sum_{x=0}^\infty x\theta(1-\theta)^x = 0,$$

that is,

$$\frac{1}{\theta} - \frac{1}{1-\theta} EX = 0,$$

or

$$EX = \frac{1}{\theta} - 1.$$

Theorem 4.4 *Suppose that the series $\sum_{x=0}^{\infty} h(\theta, x)$ converges for all θ in an interval (a, b) of real numbers and*

i. $\frac{\partial}{\partial\theta} h(\theta, x)$ *is continuous in θ for each x,*

ii. $\sum_{x=0}^{\infty} \frac{\partial}{\partial\theta} h(\theta, x)$ *converges uniformly on every closed bounded subinterval of (a, b).*

Then

$$\frac{d}{d\theta} \sum_{x=0}^{\infty} h(\theta, x) = \sum_{x=0}^{\infty} \frac{\partial}{\partial\theta} h(x, \theta).$$

The condition of uniform convergence is the key one to the theorem. Recall that a series converges uniformly if its sequence of partial sums converges uniformly.

Example 4.3 *(Continuation of Example 4.2) Since $h(\theta, x) = \theta(1-\theta)^x$ and*

$$\frac{\partial}{\partial\theta} h(\theta, x) = (1-\theta)^x - \theta x(1-\theta)^{x-1},$$

the uniform convergence of $\sum_{x=0}^{\infty} \frac{\partial}{\partial\theta} h(\theta, x)$ can be verified as follows. Define

$$S_n(\theta) = \sum_{x=0}^{n} [(1-\theta)^x - \theta x(1-\theta)^{x-1}].$$

The convergence will be uniform on $[c, d] \subset (0, 1)$ if, given $\epsilon > 0$, we can find an N such that

$$n > N \Rightarrow |S_n(\theta) - S_\infty(\theta)| < \epsilon \quad \textit{for all } \theta \in [c, d].$$

Since

$$\sum_{x=0}^{n} (1-\theta)^x = \frac{1 - (1-\theta)^{n+1}}{\theta},$$

and

$$\begin{aligned} \sum_{x=0}^{n} \theta x(1-\theta)^{x-1} &= \theta \sum_{x=0}^{n} -\frac{\partial}{\partial\theta}(1-\theta)^x \\ &= -\theta \frac{d}{d\theta} \sum_{x=0}^{n} (1-\theta)^x = -\theta \frac{d}{d\theta} [\frac{1-(1-\theta)^{n+1}}{\theta}] \\ &= \frac{[1-(1-\theta)^{n+1}] - (n+1)\theta(1-\theta)^n}{\theta}. \end{aligned}$$

Hence,

$$S_n(\theta) = \frac{1-(1-\theta)^{n+1}}{\theta} - \frac{[1-(1-\theta)^{n+1}] - (n+1)\theta(1-\theta)^n}{\theta}$$
$$= (n+1)(1-\theta)^n.$$

It is clear that, for $0 < \theta < 1$, $S_\infty = \lim_{n\to\infty} S_n(\theta) = 0$. *Since* $S_n(\theta)$ *is continuous, the convergence is uniform on any closed bounded interval. Therefore, the series of derivatives converges uniformly and the interchange of differentiation and summation is justified.*

Theorem 4.5 *Suppose the series* $\sum_{x=0}^{\infty} h(\theta, x)$ *converges uniformly on* $[a, b]$ *and that, for each* x, $h(\theta, x)$ *is a continuous function of* θ. *Then*

$$\int_a^b \sum_{x=0}^{\infty} h(\theta, x) d\theta = \sum_{x=0}^{\infty} \int_a^b h(\theta, x) d\theta.$$