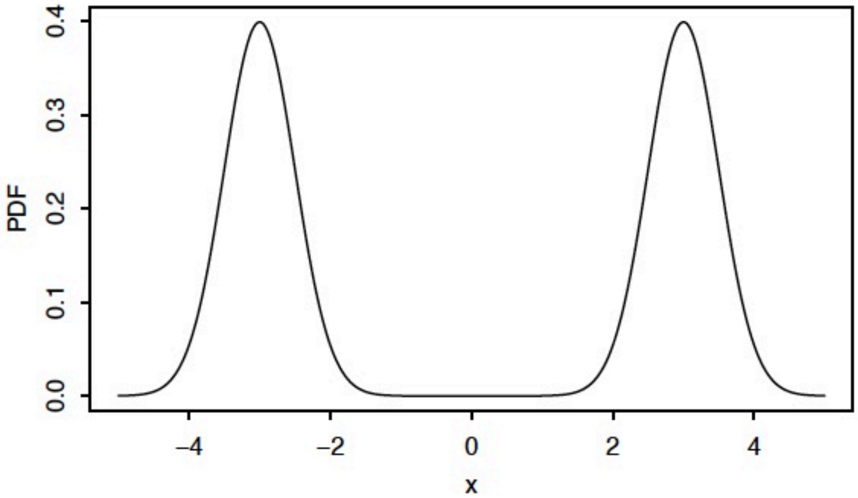

Figure 1: A distribution with two modes (−3 and 3), and infinitely many medians (all $x \in [-1, 1]$). The PDF is 0 between −1 and 1, so all values between −1 and 1 are medians of the distribution because half of the mass falls on either side.

The nth moment of a random variable X is $\mathsf{E}(X^n)$. We explain how the moments of a random variable provide relevant information about its distribution.

1 The first moment

The first moment $\mathsf{E}(X)$ or the mean is called a measure of central tendency because it tells us something about the center of a distribution, specifically its center of mass. Other measures of central tendency are the median and the mode.

Definition (Median)**.** We say that c is a *median* of a random variable X if $\mathsf{P}(X \leq c) \geq \frac{1}{2}$ and $\mathsf{P}(X \geq c) \geq \frac{1}{2}$.

Definition (Mode)**.** For a discrete random variable X, we say that c is a mode of X if it maximizes the PMF: $\mathsf{P}(X = c) \geq \mathsf{P}(X = x)$ for all x. For a continuous random variable X with PDF f, we say that c is the mode if it maximizes the PDF: $f(c) \geq f(x)$ for all x.

The mean, the median and mode of a random variable depend only on its distribution. Intuitively, the median is a value c such that half the mass of the distribution falls on either side of c, and the mode is a value that has the greatest mass or density out of all values in the support of X. Note that a distribution can have at most one mean (the mean does not necessarily exist), but it can have multiple medians and multiple modes. Medians occur side by side; modes occur all over the distribution – see Figure 1.

Suppose we are trying to guess what a not-yet-observed random variable X will be, by making a prediction c. The mean and the median are both good candidates, but which one to choose?

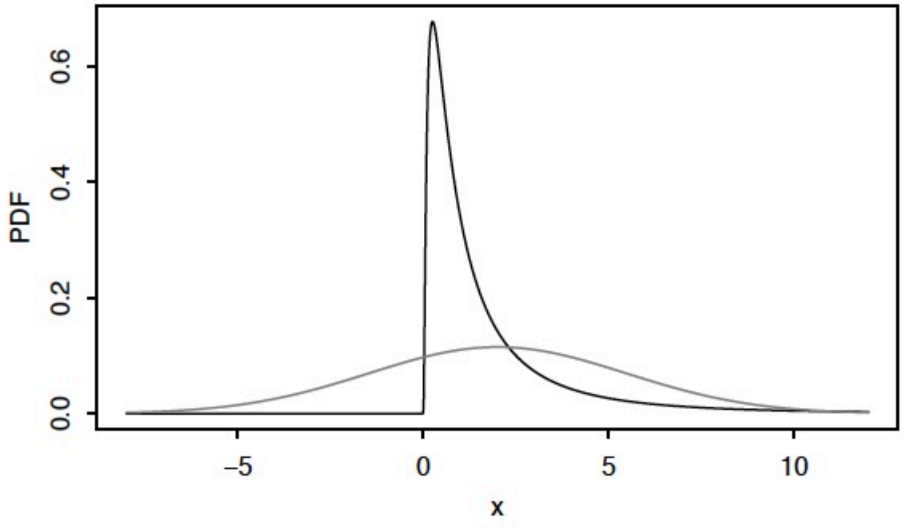

Figure 2: Two PDFs with mean 2 and variance 12. The light curve is the $\mathsf{N}(2, 12)$ PDF, and the dark curve is Log-Normal. The Normal curve is symmetric about 2, so its mean, median and mode are all 2. In contrast, the Log-Normal is heavily skewed to the right; this means its right tail is very long compared to its left tail. It has mean 2, but median 1 and mode 0.25. From the mean and variance alone, we would not be able to capture the difference between the asymmetry of the Log-Normal and the symmetry of the Normal.

Theorem. *Let X be a random variable with mean μ, and let m be the median of X.*

- *The value of c that minimizes the mean squared error $\mathsf{E}(X - c)^2$ is $c = \mu$.*
- *A value of c that minimizes the mean absolute error $\mathsf{E}|X - c|$ is $c = m$.*

Definition (Symmetry of a random variable)**.** We say that a random variable X has a *symmetric distribution about μ* if $X - \mu$ has the same distribution as $\mu - X$. The number μ must be the mean $\mathsf{E}(X)$ if it exists, and must also be a median of the distribution of X.

Proposition (Symmetry in terms of the PDF)**.** *Let X be a continuous random variable with PDF f. Then X is symmetric about μ if and only if $f(x) = f(2\mu - x)$ for all x.*

2 Higher-order moments (first, second, third,... moments)

The second moment $\mathsf{E}(X^2)$ together with the mean $\mathsf{E}(X)$ give the variance $\mathsf{Var}(X) = \mathsf{E}(X^2) - (\mathsf{E}X)^2$ – a measure of the *spread* of the distribution of X. However, there are major features of distributions that are not captured by the mean and variance – see Figure 2.

Definition (Kinds of moments)**.** Let X be a random variable with mean μ and variance σ^2. For any positive integer n, the nth moment of X is $\mathsf{E}(X^n)$, the nth central moment is $\mathsf{E}((X-\mu)^n)$, and the nth standardized moment is $\mathsf{E}(\left(\frac{X-\mu}{\sigma}\right)^n)$

The mean is the first moment and the variance is the second central moment.

Definition (Skewness)**.** The *skewness* of a random variable X with mean μ and variance σ^2 is the third standardized moment of X

$$\text{Skew}(X) = \mathsf{E}\left(\frac{X-\mu}{\sigma}\right)^3.$$

By standardizing first, we make the skewness of X independent of its location μ and scale σ. In addition, standardizing first means that the units in which X is measured will not affect the skewness.

Proposition (Odd central moments of a symmetric distribution)**.** *Let X be symmetric about its mean μ. Then for any odd number m, the mth central moment* $\mathsf{E}(X-\mu)^m$ *is* 0 *if it exists.*

This result justifies using an odd central moment as a measure of the skew of a distribution. The first standardized moment $\mathsf{E}\left(\frac{X-\mu}{\sigma}\right)$ is always 0, so the third standardized moment is taken as the definition of skewness. Positive skewness is indicative of having a long right tail relative to the left tail, and negative skewness is indicative of the reverse. The fifth or higher odd standardized moments could also be useful in characterizing skewness, but they are typically harder to be estimated from the data.

Another important descriptive feature of a distribution is how heavy (or long) its tails are. For a given variance, is the variability explained more by a few rare (extreme) events, or by a moderate number of moderate deviations from the mean? As with measuring skew, no single measure can perfectly capture the tail behavior, but there is a widely used summary based on the fourth standardized moment.

Definition (Kurtosis)**.** The *kurtosis* of a random variable X with mean μ and variance σ^2 is a shifted version of the fourth standardized moment of X:

$$\text{Kurt}(X) = \mathsf{E}\left(\frac{X-\mu}{\sigma}\right)^4 - 3.$$

We substract 3 to make any Normal distribution have kurtosis 0.

Let us call the regions within 1 standard deviation of the mean, between 1 and 2 standard deviations of the mean, and more than 2 standard deviations from the mean the *center*, *shoulders*, and *tails* of a distribution, respectively. Then a prototypical distribution with large kurtosis has a PDF with a sharp peak in the center, low shoulders, and heavy tails – see Figure 3.

3 Sample moments

In statistical inference, a central problem is how to use data to estimate unknown parameters of a distribution, or functions of unknown parameters.

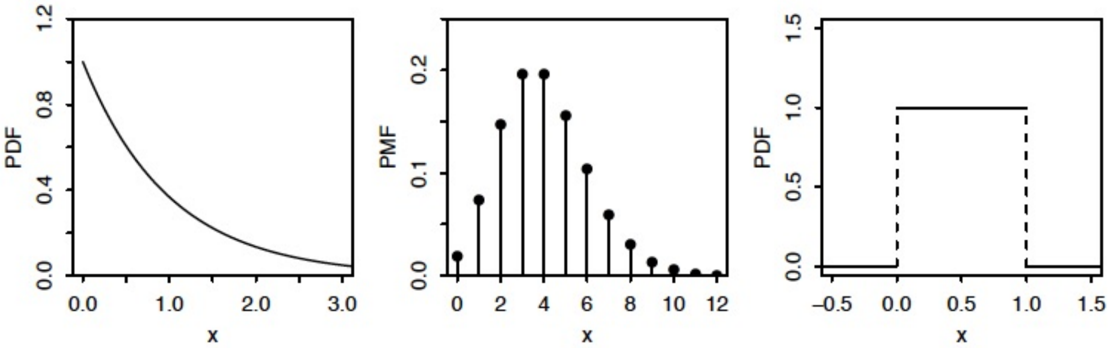

Figure 3: Skewness and kurtosis of some named distributions. Left panel: $\textsf{Expo}(1)$ PDF, skewness = 2, kurtosis = 6. Middle panel: $\textsf{Pois}(4)$ PMF, skewness = 0.5, kurtosis = 0.25. Right panel: $\textsf{Unif}(0,1)$ PDF, skewness = 0, kurtosis = −1.2. The $\textsf{Expo}(1)$ and $\textsf{Pois}(4)$ distributions (left and middle) both have positive skewness and positive kurtosis, indicating that they are right-skewed and their tails are heavier than those of a Normal distribution. The $\textsf{Unif}(0,1)$ distribution (right) has zero skewness and negative kurtosis: zero skewness because the distribution is symmetric about its mean, and negative kurtosis because it has no tails.

Definition (Sample moments)**.** Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables. The kth *sample moment* is the random variable

$$M_k = \frac{1}{n}\sum_{j=1}^{n} X_j^k.$$

The *sample mean* $\bar{X}_n$ is the first sample moment:

$$\bar{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j.$$

The *population mean* or *true mean* is $\mathsf{E}(X_j)$, the mean of the distribution from which the X_j were drawn.

The law of large numbers that will be covered later in the course, shows that the kth sample moment of i.i.d. random variables $X_1, \ldots, X_n$ converges to the kth moment $\mathsf{E}(X_1^k)$ as the sample size becomes larger and larger $n \to \infty$.

The expected value of the kth sample moment is the kth moment:

$$\mathsf{E}\left(\frac{1}{n}\sum_{j=1}^{n} X_j^k\right) = \mathsf{E}\left(X_1^k\right).$$

We say that the kth sample moment is *unbiased* for estimating the kth moment.

Theorem (Mean and variance of sample mean)**.** *Let $X_1, \ldots, X_n$ be i.i.d. random variables with mean μ and variance σ^2. Then the sample mean $\bar{X}_n$ is unbiased for estimating μ:*

$$\mathsf{E}(\bar{X}_n) = \mu.$$

The variance of $\bar{X}_n$ is

$$\mathsf{Var}(\bar{X}_n) = \frac{\sigma^2}{n}.$$

Proof. We use the fact that the variance of the sum of *independent* random variables is the sum of the variances:

$$\mathsf{Var}(\bar{X}_n) = \frac{1}{n^2}\mathsf{Var}(X_1 + X_2 + \ldots + X_n) = \frac{n}{n^2}\mathsf{Var}(X_1) = \frac{\sigma^2}{n}.$$

□

Definition (Sample variance and sample standard deviation)**.** Let $X_1, \ldots, X_n$ be i.i.d. random variables. The *sample variance* is the random variable:

$$S_n^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X}_n)^2.$$

The *sample standard deviation* S_n is the square root of the sample variance S_n^2. The motivation for the $n-1$ is that this makes S_n^2 *unbiased* for estimating σ^2, i.e., it is correct on average. However, the sample standard deviation S_n is *not* unbiased for estimating σ.

Theorem (Unbiasedness of sample variance)**.** *Let $X_1, \ldots, X_n$ be i.i.d. random variables with mean μ and variance σ^2. Then the sample variance S_n^2 is unbiased for estimating σ^2:*

$$\mathsf{E}(S_n^2) = \sigma^2.$$

Proof. The following equality holds for any $c \in \mathbb{R}$:

$$\sum_{j=1}^{n}(X_j - c)^2 = \sum_{j=1}^{n}(X_j - \bar{X}_n)^2 + n(\bar{X}_n - c)^2.$$

We choose $c = \mu$ in this equality. We take expectation on both sides:

$$n\mathsf{E}(X_1 - \mu)^2 = \mathsf{E}\left(\sum_{j=1}^{n}(X_j - \bar{X}_n)^2\right) + n\mathsf{E}(\bar{X}_n - \mu)^2.$$

We have

$$\begin{aligned}
\mathsf{E}(X_1 - \mu)^2 &= \mathsf{Var}(X_1) = \sigma^2. \\
\mathsf{E}(\bar{X}_n - \mu)^2 &= \mathsf{Var}(\bar{X}_n) = \frac{\sigma^2}{n}, \\
\sum_{j=1}^{n}(X_j - \bar{X}_n)^2 &= (n-1)S_n^2.
\end{aligned}$$

Thus

$$n\sigma^2 = (n-1)\mathsf{E}(S_n^2) + \sigma^2.$$

which proves that $\mathsf{E}(S_n^2) = \sigma^2$. □

We define the *sample skewness* to be

$$\frac{\frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X}_n)^3}{S_n^3},$$

and the *sample kurtosis* to be

$$\frac{\frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X}_n)^4}{S_n^4} - 3.$$

4 Moment generating functions

A moment generating function is a continuous function that encodes the moments of a distribution.

Definition (Moment generating function)**.** The *moment generating function* (MGF) of a random variable X is

$$M(t) = \mathsf{E}(e^{tX}),$$

as a function of $t \in \mathbb{R}$. Note that t does not have any interpretation. If $M(t)$ is finite in some open interval $(-a, a)$ it exits, otherwise the MGF does not exist. Note that $M(0) = 1$ for any valid MGF M.

Example: Bernoulli MGF
For $X \sim \mathsf{Bern}(p)$, the MGF is

$$M(t) = \mathsf{E}(e^{tX}) = e^{t\cdot 0}\mathsf{P}(X = 0) + e^{t\cdot 1}\mathsf{P}(X = 1) = pe^t + q,$$

where $q = 1 - p$. Since $M(t)$ is finite for any $t \in \mathbb{R}$, the MGF of a Bernoulli random variable is defined on the entire real line.

Example: Geometric MGF
For $X \sim \mathsf{Geom}(p)$, the MGF is

$$M(t) = \mathsf{E}(e^{tX}) = \sum_{k=0}^{\infty} e^{tk}\mathsf{P}(X = k) = \sum_{k=0}^{\infty} e^{tk}q^k p = \frac{p}{1 - qe^t}.$$

for $qe^t < 1$ or, equivalently, for $(-\infty, log(1/q))$, which is an open interval containing 0.

The next three theorems give three reasons why the MGF is important. First, the MGF encodes the moments of a random variable. Second, the MGF of a random variable determines its distribution, like the CDF and the PMF/PDF. Third, MGFs make it easy to find the distribution of a sum of independent random variables.

Theorem (Moments via derivatives of MGFs)**.** *Given the MGF $M(t)$ of a random variable X, we can get the nth moment of X by evaluating its nth derivative of the MGF at* 0*:*

$$\mathsf{E}(X^n) = M^{(n)}(0).$$

Proof. The Taylor series expansion of $M(t)$ about 0 is

$$M(t) = \sum_{n=0}^{\infty} M^{(n)}(0)\frac{t^n}{n!}.$$

On the other hand, we also have

$$M(t) = \mathsf{E}\left(e^{tX}\right) = \mathsf{E}\left(\sum_{n=0}^{\infty} X^n \frac{t^n}{n!}\right) = \sum_{n=0}^{\infty} \mathsf{E}(X^n)\frac{t^n}{n!}.$$

By matching the coefficients of these two polynomials, we obtain $\mathsf{E}(X^n) = M^{(n)}(0)$ for any $n = 1, 2, \ldots$. □

Theorem (MGF determines the distribution)**.** *The MGF of a random variable determines its distribution: if two random variables have the same MGF, they must have the same distribution.*

Theorem (MGF of a sum of independent random variables)**.** *If X and Y are independent, then the MGF of $X + Y$ is the product of the individual MGFs:*

$$M_{X+Y}(t) = M_X(t)M_Y(t).$$

Using this result, we can obtain the MGFs of the Binomial and Negative Binomial, which are sums of independent Bernoulli and Geometric i.i.d. random variables.

Example: Binomial MGF

The MGF of a $\mathsf{Bern}(p)$ random variable is $pe^t + q$, so the MGF of a $\mathsf{Bin}(n, p)$ random variable is

$$M(t) = (pe^t + q)^n.$$

Example: Negative Binomial MGF

The MGF of a $\mathsf{Geom}(p)$ random variable is $\frac{p}{1-qe^t}$ for $qe^t < 1$, thus the MGF of $X \sim \mathsf{NBin}(r, p)$ is

$$M(t) = \left(\frac{p}{1 - qe^t}\right)^n, \text{ for } qe^t < 1.$$

Location and scale transformations are a fundamental way to build a family of distributions from an initial distribution. It is easy to relate the MGFs of two random variables connected by a location and scale transformation.

Proposition (MGF of location-scale transformation)**.** *If X has MGF $M(t)$, then the MGF of $a + bX$, where $a, b \in \mathbb{R}$ is*

$$\mathsf{E}\left(e^{t(a+bX)}\right) = e^{at}\mathsf{E}\left(e^{btX}\right) = e^{at}M(bt).$$

Example: Normal MGF

The MGF of a standard Normal random variable $Z \sim \mathsf{N}(0, 1)$ is

$$M_Z(t) = \mathsf{E}(e^{tZ}) = \int_{-\infty}^{\infty} e^{tz}\frac{1}{\sqrt{2\pi}}e^{-z^2/2}dz = e^{t^2/2} = \sum_{n=0}^{\infty}\frac{(t^2/2)^n}{n!} = \sum_{n=0}^{\infty}\frac{(2n)!}{2^n n!}\frac{t^{2n}}{(2n)!}.$$

Therefore

$$\mathsf{E}\left(Z^{2n}\right) = \frac{(2n)!}{2^n n!}, \text{ and } \mathsf{E}\left(Z^{2n+1}\right) = 0.$$

The MGF of $X = \mu + \sigma Z \sim \mathsf{N}(\mu, \sigma^2)$ is

$$M_X(t) = e^{\mu t} M_Z(\sigma t) = e^{\mu t} e^{(\sigma t)^2/2} = e^{\mu t + \frac{1}{2}\sigma^2 t^2}.$$

It follows that the kurtosis of the Normal distribution is 0:

$$\text{Kurt}(X) = \mathsf{E}\left(\frac{X-\mu}{\sigma}\right)^4 - 3 = \mathsf{E}\left(Z^4\right) - 3 = 3 - 3 = 0.$$

Next, if we have $X_1 \sim \mathsf{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathsf{N}(\mu_2, \sigma_2^2)$, and X_1 and X_2 are independent, then the MGF of $X_1 + X_2$ is

$$M_{X_1+X_2}(t) = M_{X_1}(t) M_{X_2}(t) = e^{(\mu_1+\mu_2)t + \frac{1}{2}(\sigma_1^2+\sigma_2^2)t^2},$$

which is the $\mathsf{N}\left(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2\right)$ MGF. Since the MGF determines the distribution of a random variable, we have $X_1 + X_2 \sim \mathsf{N}\left(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2\right)$. The converse of this result also holds, and it is known as Cramer's theorem: if X_1 and X_2 are independent and $X_1 + X_2$ is Normal, then X_1 and X_2 are also Normal.

Example: Exponential MGF

The MGF of $X \sim \mathsf{Expo}(1)$ is

$$M(t) = \mathsf{E}(e^{tX}) = \int_0^\infty e^{tx} e^{-x} dx = \frac{1}{1-t} = \sum_{n=0}^\infty t^n = \sum_{n=0}^\infty n! \frac{t^n}{n!}, \text{ for } t < 1.$$

Thus $\mathsf{E}(X^n) = n!$ for $n = 1, 2, \ldots$.

To find the MGF of $Y \sim \mathsf{Expo}(\lambda)$, we use a scale transformation: $Y = X/\lambda$ where $X \sim \mathsf{Expo}(1)$, thus

$$M_Y(t) = M_X\left(\frac{t}{\lambda}\right) = \frac{\lambda}{\lambda - t}, \text{ for } t < \lambda.$$

Since $Y^n = X^n/\lambda^n$, we obtain

$$\mathsf{E}\left(Y^n\right) = \frac{n!}{\lambda^n}.$$

Now we can easily calculate

$$\mathsf{Var}(Y) = \mathsf{E}\left(Y^2\right) - (\mathsf{E}Y)^2 = \frac{2!}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

Example: Poisson MGF

The MGF of $X \sim \mathsf{Pois}(\lambda)$ is

$$\mathsf{E}\left(e^{tX}\right) = \sum_{k=0}^\infty e^{tk} \frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^\infty \frac{(\lambda e^t)^k}{k!} = e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t-1)}.$$

Now let $Y \sim \mathsf{Pois}(\mu)$ be independent of X. The MGF of $X + Y$ is

$$\mathsf{E}\left(e^{tX}\right)\mathsf{E}\left(e^{tY}\right) = e^{\lambda(e^t-1)} e^{\mu(e^t-1)} = e^{(\lambda+\mu)(e^t-1)},$$

which is the $\mathsf{Pois}(\lambda+\mu)$ MGF. Since the MGF determines the distribution, it follows that $X+Y \sim \mathsf{Pois}(\lambda+\mu)$. That is, a sum of independent Poisson random variables is also a Poisson random variable.