3.7 More properties of the expected value

A random variable X is a real number whose value is determined by the outcome of some random phenomenon. To write one down, we have to specify two things: its range and its distribution. To fully characterize the distribution of a random variable, we often write down objects like the cdf or the pmf/pdf, from which any probability can be computed. Having access to a complete description of the probability behavior of a random variable is nice, but we are often interested in extracting from the distribution of a random variable a few principle *features* or *summaries* of the distribution. We have already seen some:

- The expected value $E(X)$ is a single number that averages over the distribution of X to answer "for all the random variation in X, where, at the end of the day, will its value typically be concentrated?"
- The variance $\text{var}(X)$ is a single number that answers "how far should I expect X to be from its expected value?"

When they exist, these are nice ways to concisely report on the key features of the random variation of X, without having to confront every detail of what the pmf/pdf is up to. And there are more where that came from.

3.7.1 Standardization

We have seen two properties of expectation and variance:

$$\begin{aligned} E(aX+b) &= aE(X)+b \\ \text{var}(aX+b) &= a^2\text{var}(X). \end{aligned}$$

Say that a random variable X has $E(X)=\mu$, $\text{var}(X)=\sigma^2$, and $\text{sd}(X)=\sigma$. Then we can define a new random variable Y by applying a transformation to X:

$$Y = \frac{X-\mu}{\sigma} = \frac{X}{\sigma} - \frac{\mu}{\sigma}.$$

As a consequence of the properties above, you can show that $E(Y)=0$ and $\text{var}(Y)=1$:

$$\begin{aligned} E(Y) &= E\left(\frac{X}{\sigma} - \frac{\mu}{\sigma}\right) = \frac{E(X)}{\sigma} - \frac{\mu}{\sigma} = \frac{\mu}{\sigma} - \frac{\mu}{\sigma} = 0 \\ \text{var}(Y) &= \text{var}\left(\frac{X}{\sigma} - \frac{\mu}{\sigma}\right) = \frac{1}{\sigma^2}\text{var}(X) = \frac{1}{\sigma^2}\sigma^2 = 1. \end{aligned}$$

So, if you start with some random variable X, and then you subtract off its mean (center it) and divide by its standard deviation (normalize it), you **standardize** it: transform it to have mean 0 and variance 1.

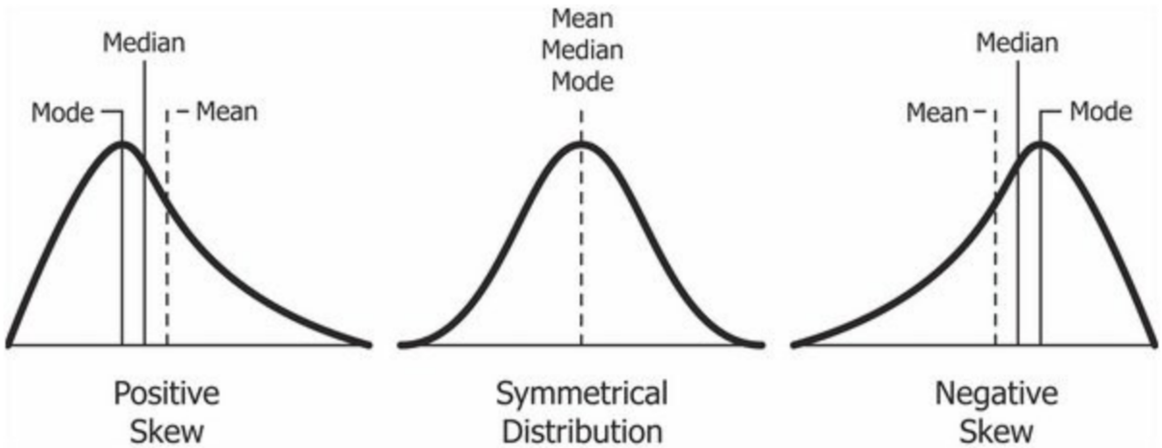

Figure 3.25

3.7.2 Moments

"Moment" is one of those vocabulary words (like "degrees of freedom," I would say) that doesn't really mean anything, but we just learn how to use it in context. Anyway, here are the definitions for the different "moments" of a random variable X that has $E(X) = \mu$ and $\text{var}(X) = \sigma^2$:

n^{th} raw moment	$E(X^n)$
n^{th} absolute moment	$E(\lvert X\rvert^n)$
n^{th} central moment	$E[(X-\mu)^n]$
n^{th} absolute, central moment	$E(\lvert X-\mu\rvert^n)$
n^{th} standardized moment	$E\left[\left(\frac{X-\mu}{\sigma}\right)^n\right]$
n^{th} absolute, standardized moment	$E\left[\left\lvert\frac{X-\mu}{\sigma}\right\rvert^n\right]$

Some of these moments get special names because they teach us something about the "shape" of the distribution of X

- The first raw moment is the **mean**, and it tells us something about the location of the distribution;
- The second central moment is the **variance**, and it tells us something about the spread of the distribution;
- The third standardized moment is called the **skewness**, and it tells us something about the asymmetry of the distribution. The Normal distribution has 0 skewness, for example, but the Gamma has positive skew;
- The fourth standardized moment is called the **kurtosis**, and it tells us something about the thickness of the tails of the distribution. Thicker tails mean extreme events are more likely.

This is summarized in Table 3.3.

	raw	central	standardized
$n = 1$	$\mu = E(X)$ "mean"	0	0
$n = 2$	$E(X^2)$	$\sigma^2 = E[(X-\mu)^2]$ "variance"	1
$n = 3$	$E(X^3)$	$E[(X-\mu)^3]$	$E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$ "skewness"
$n = 4$	$E(X^4)$	$E[(X-\mu)^4]$	$E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]$ "kurtosis"
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Table 3.3

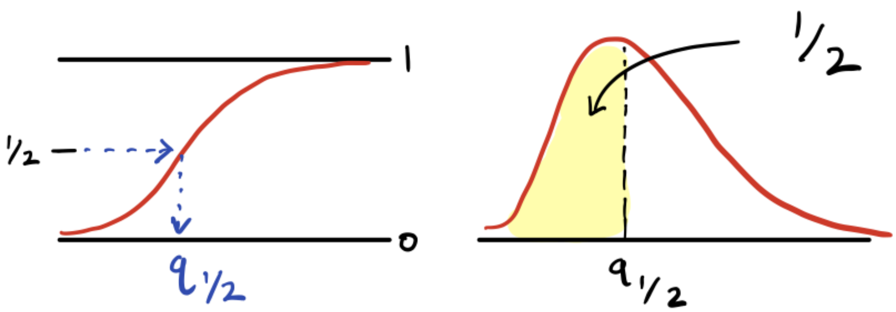

Figure 3.26

Definition 3.19. We say a moment $E(X^n)$ **exists** if the corresponding absolute moment is finite: $E(|X|^n) < \infty$. So the mean exists, for example, if $E(|X|) < \infty$.

3.7.3 Quantiles

The mean is a measure of central tendency in a distribution. But it does not always exist. What other measures can we employ it that case? The main one is the **median**. This is the point in the range that has exactly half the probability to the left and half the probability to the right, like in Figure 3.26.

When the pmf/pdf of a distribution is symmetric, then the median is obviously the center. And when the expectation exists, the median and the expectation are equal in the symmetric case. But in general they are not equal, and the median always exists, even if the mean does not. But there is nothing special about "half the probability." The median is just an example of a **quantile**. It's an

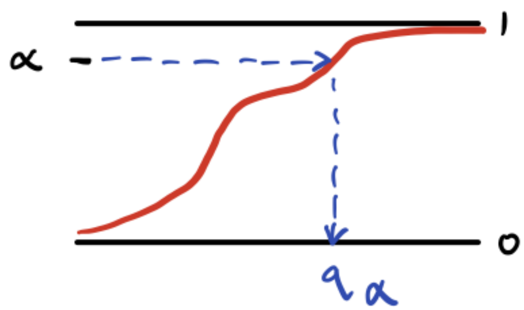

Figure 3.27

$\alpha = 0.5$ quantile to be exact. An αth quantile is any real number q_α satisfying

$$P(X \leq q_\alpha) = \alpha.$$

So q_α is a point in the range with α probability to the left of it. Note that quantiles are not unique in general. When the cdf F_X is an invertible function (no jumps or plateaus), then the quantiles are unique, and you can define them this way, displayed in Figure 3.27:

$$q_\alpha = F_X^{-1}(\alpha).$$

For this reason, the inverse cdf F_X^{-1} is called the **quantile function**.

Example 3.19. Let $X \sim \text{Exp}(\lambda)$, and recall that this means $X \sim \text{Gamma}(\alpha = 1,\ \beta = \lambda)$. So the pdf is $f_X(x) = \lambda \exp(-\lambda x)$ for $x > 0$, and the cdf is

$$F_X(x) = \int_0^x \lambda e^{-\lambda t}\,\mathrm{d}t = [-e^{\lambda t}]_0^x = 1 - e^{-\lambda x}, \quad x > 0.$$

For some $y \in (0,\ 1)$, we see that

$$\begin{aligned} y &= 1 - e^{-\lambda x} \\ e^{-y} &= 1 - y \\ -\lambda x &= \ln(1 - y) \\ x &= \frac{-\ln(1-y)}{\lambda}, \end{aligned}$$

and so the quantile function is

$$F_X^{-1}(y) = -\frac{\ln(1-y)}{\lambda},$$

and the median of the exponential distribution is

$$F_X^{-1}(0.5) = -\frac{\ln(1-0.5)}{\lambda} = -\frac{\ln(0.5)}{\lambda} = \frac{\ln 2}{\lambda}.$$

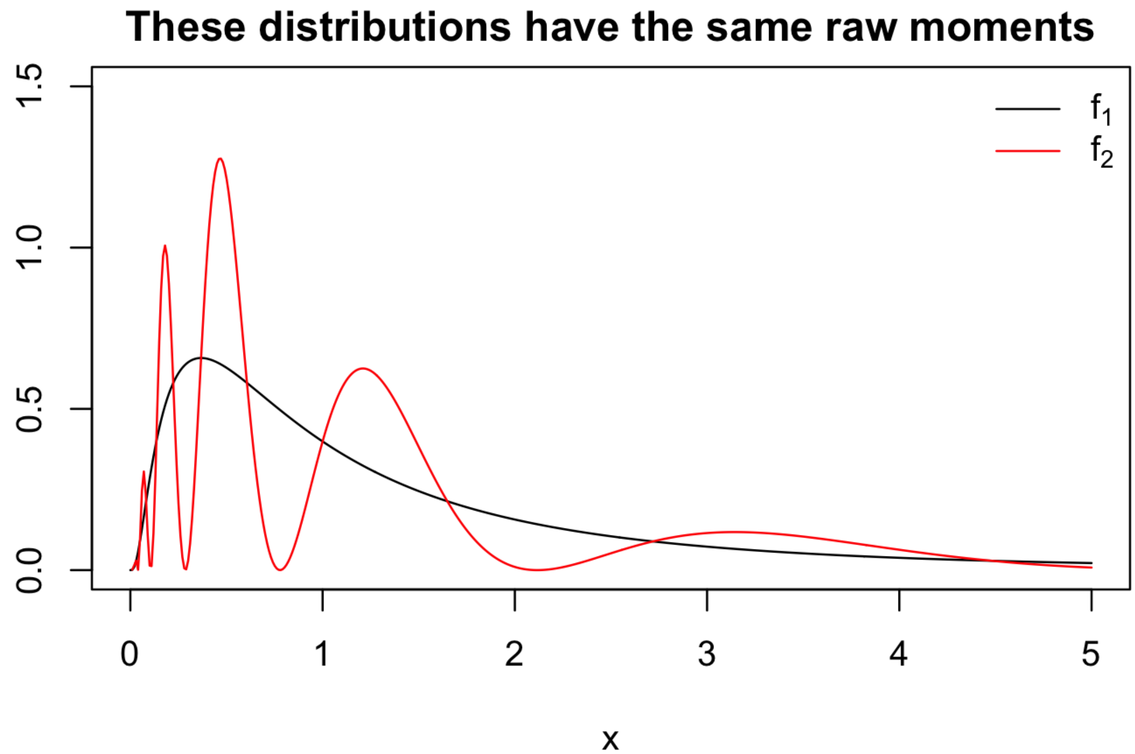

Figure 3.28

3.7.4 The moment-generating function

We have referred many times to the fact that objects like the cdf, the pmf, or the pdf fully and uniquely characterize the entire distribution of a random variable. If you have any one of these objects, you *have* the distribution, and you know everything there is to know about it. Moments are an example of an object that does *not* fully and uniquely characterize the distribution of a random variable. Two random variables can have the exact same moments, and yet be wildly different.

Example 3.20. Consider $X \sim \mathrm{N}(2,\ 2)$ and $Y \sim \text{Poisson}(2)$. So these distributions are very different (discrete versus continuous, different ranges entirely, etc), and yet $E(X) = E(Y) = 2$ and $\text{var}(X) = \text{var}(Y) = 2$.

That example shows that the first two moments are not enough to uniquely characterize the distribution, but it turns out that even if *all* of the infinitely-many moments $E(X)$, $E(X^2)$, $E(X^3)$, ... are the same, the distributions could still be very different:

Example 3.21. Consider two absolutely continuous random variables X_1 and X_2 with pdfs

$$\begin{aligned} f_1(x) &= \frac{1}{\sqrt{2\pi}}\frac{1}{x}\exp\left(-\frac{1}{2}(\ln x)^2\right), && x \geq 0 \\ f_2(x) &= f_1(x)[1 + \sin(2\pi \ln x)], && x \geq 0. \end{aligned}$$

Figure 3.28 displays these densities, and clearly they are quite different. And yet, you can show that $E(X_1^n) = E(X_2^n)$ for all $n \in \mathbb{N}$. We start by computing $E(X_2^n)$ via LOTUS:

$$\begin{aligned}
E(X_2^n) &= \int_0^\infty x^n f_2(x)\,\mathrm{d}x \\
&= \int_0^\infty x^n f_1(x)[1 + \sin(2\pi \ln x)]\,\mathrm{d}x \\
&= \int_0^\infty [x^n f_1(x) + x^n f_1(x) \sin(2\pi \ln x)]\,\mathrm{d}x \\
&= \int_0^\infty x^n f_1(x)\,\mathrm{d}x + \int_0^\infty x^n f_1(x) \sin(2\pi \ln x)]\,\mathrm{d}x \\
&= E(X_1^n) + \int_0^\infty x^n f_1(x) \sin(2\pi \ln x)]\,\mathrm{d}x \\
&= E(X_1^n) + \int_0^\infty \frac{x^n}{x\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\ln x)^2\right) \sin(2\pi \ln x)]\,\mathrm{d}x
\end{aligned}$$

So it will suffice to show that the second term is zero. We apply the change of variables

$$\begin{aligned}
u &= \ln x - n \\
x &= \exp(u + n) \\
\mathrm{d}u &= \frac{1}{x}\,\mathrm{d}x.
\end{aligned}$$

So

$$\begin{aligned}
E(X_2^n) &= E(X_1^n) + \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}} \exp(u+n)^n \exp\left(-\frac{1}{2}(u+n)^2\right) \sin(2\pi(u+n))\,\mathrm{d}u \\
&= E(X_1^n) + \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}} \exp(nu + n^2) \exp\left(-\frac{1}{2}(u+n)^2\right) \sin(2\pi u + 2\pi n)\,\mathrm{d}u \\
&= E(X_1^n) + \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}} \exp(nu + n^2) \exp\left(-\frac{1}{2}u^2 - nu - \frac{1}{2}n^2\right) \sin(2\pi u)\,\mathrm{d}u \\
&= E(X_1^n) + \int_{-\infty}^\infty \underbrace{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(u^2 - n^2)\right) \sin(2\pi u)}_{\text{odd function}}\,\mathrm{d}u \\
&= E(X_1^n) + 0 \\
&= E(X_1^n).
\end{aligned}$$

Alright alright. So moments, even infinitely many of them, are not enough to uniquely characterize a distribution. But there exists an object intimately related to the moments that does uniquely characterize its distribution:

Definition 3.20. If X is any random variable, and the expected value $E[e^{tX}]$ exists and is finite for all t in a neighborhood of zero (so for all $-b < t < b$ for some $b > 0$), then the **moment-**

generating function (**mgf**) of X is

$$M_X(t) = E[e^{tX}], \quad t \in (-b, b). \tag{3.25}$$

Remark 3.17. The moment generating function might not exist! As we have seen, not all expected values exist or are finite, so not all random variables have a moment generating function.

Theorem 3.14. If X and Y are random variables with mgfs M_X and M_Y, and $M_X = M_Y$, then X and Y have the exact same distribution.

We will not prove this result, but it means that the mgf, like the cdf or pmf or pdf, is one of those objects that fully and uniquely characterizes the entire distribution of a random variable. If you know the mgf, you know everything.

That's all well and good, but what's the point of this new object? Here's the point:

Theorem 3.15. If X is a random variable with a moment generating function M_X, then the nth derivative of the mgf evaluated at zero is equal to the nth raw moment:

$$M_X^{(n)}(0) = E(X^n) \quad \forall n \in \mathbb{N}. \tag{3.26}$$

Proof sketch. We can rewrite M_X with a Taylor series expansion about zero (ie a Maclaurin series):

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

We can also rewrite e^{tX} using the Taylor series expansion of the exponential function

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

While it is not always the case that the expected value of an infinite sum is the infinite sum of the expected values, it is true in this case (for reasons I will not go into), and so

$$E[e^{tX}] = \sum_{n=0}^{\infty} \frac{E(X^n)}{n!} t^n.$$

By definition, we know that $M_X(t) = E(e^{tX})$, and so it must be the case that

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

The only way these can be equal if is each term is equal, and so $M_X^{(n)}(0) = E(X^n)$. □

Remark 3.18. Theorem 3.15 is killer. It allows us to compute moments by taking a derivative instead of simplifying an integral or an infinite series. As a general matter, differentiation is much more straightforward than integration, so this will be very useful. Now that we have this tool, we can use it to do some moment calculations that we've been putting off.

Remark 3.19. The moment-generating function is intimately related to something called the *Laplace transform*, which you may encounter in a course on ordinary differential equations.

Example 3.22. Let $X \sim \text{Poisson}(\lambda)$, and fix any $t \in \mathbb{R}$. Then

$$\begin{aligned} M_X(t) &= E[e^{tX}] \\ &= \sum_{k=0}^{\infty} e^{tk} \frac{\lambda^k}{k!} e^{-\lambda} \\ &= \sum_{k=0}^{\infty} \left(e^t\right)^k \frac{\lambda^k}{k!} e^{-\lambda} \\ &= \sum_{k=0}^{\infty} \frac{\left(\lambda e^t\right)^k}{k!} e^{-\lambda} \\ &= e^{-\lambda} \sum_{k=0}^{\infty} \frac{\left(\lambda e^t\right)^k}{k!} \\ &= e^{-\lambda} e^{\lambda e^t} \\ &= e^{\lambda e^t - \lambda}. \end{aligned}$$

Note that we did not have to place any restrictions on t to complete the computation, so the mgf is defined for all t and we have

$$M_X(t) = \exp\left(\lambda e^t - \lambda\right), \quad t \in \mathbb{R}. \tag{3.27}$$

We can use this to compute the mean and variance of the Poisson:

$$\begin{aligned} M'_X(t) &= \lambda e^t \exp(\lambda e^t - \lambda) \\ M''_X(t) &= (\lambda e^t)^2 \exp(\lambda e^t - \lambda) + \lambda e^t \exp(\lambda e^t - \lambda) \end{aligned}$$

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

So $E(X) = \lambda$ and $\text{var}(X) = E(X^2) - E(X)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$. Once you have the mgf, getting these moment calculations done with derivatives is way easier than proceeding from LOTUS and grinding through an infinite series or integral.

Example 3.23. Let $X \sim \mathrm{N}(0, 1)$, and fix any $t \in \mathbb{R}$. Then

$$\begin{aligned}
M_X(t) &= E[e^{tX}] && \text{definition} \\
&= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}\,\mathrm{d}x && \text{LOTUS} \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2+tx}\,\mathrm{d}x && \text{combine base-}e\text{ terms} \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x^2-2tx)}\,\mathrm{d}x && \text{factor out -1/2} \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x^2-2tx+t^2-t^2)}\,\mathrm{d}x && \text{add/subtract } t^2 \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[(x-t)^2-t^2]}\,\mathrm{d}x && \text{factor first three terms} \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-t)^2+\frac{1}{2}t^2}\,\mathrm{d}x && \text{distribute -1/2} \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-t)^2} e^{\frac{1}{2}t^2}\,\mathrm{d}x && \\
&= e^{\frac{1}{2}t^2} \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-t)^2}\,\mathrm{d}x}_{\mathrm{N}(t,1)\text{ PDF}} && \text{pull constant out of integral} \\
&= e^{\frac{1}{2}t^2} \cdot 1 && \text{PDFs integrate to 1} \\
&= e^{\frac{1}{2}t^2}.
\end{aligned}$$

Note two things:

- We did not have to place any restrictions on t to complete the computation, so the mgf is defined for all t;
- This is a classic example of "massage and squint." We just pushed and pulled the integrand and massaged it with constants until it turned into something familiar. Then we applied an identity.

So the mgf of the standard normal is

$$M_X(t) = \exp\left(\frac{1}{2}t^2\right), \quad t \in \mathbb{R}. \tag{3.28}$$

We can use this to compute the mean and variance of the standard normal:

$$M'_X(t) = t\exp\left(\frac{1}{2}t^2\right)$$
$$M''_X(t) = t^2\exp\left(\frac{1}{2}t^2\right) + \exp\left(\frac{1}{2}t^2\right)$$

$$M'_X(0) = E(X) = 0$$
$$M''_X(0) = E(X^2) = 0 + 1 = 1$$

So $E(X) = 0$ and $\text{var}(X) = E(X^2) - E(X)^2 = 1 - 0 = 1$. Not exactly earth-shattering, but note that it's the first time we've actually seen a calculation of these facts.

Theorem 3.16. If the moment generating function of a random variable X exists, then all of its moments are finite: $E[|X|^n] < \infty$ for all $n \in \mathbb{N}$.

However, the converse is not true. A random variable could have all finite moments, and yet the mgf does not exist. So all moments finite is a *necessary condition* for the existence of the moment generating function, but not a sufficient one.

Summary of basic facts about moments and mgfs

- mgf exists **does imply** $E(X^n)$ finite for all $n \in \mathbb{N}$;
- If mgf exists, then $M_X^{(n)}(0) = E(X^n)$ for all $n \in \mathbb{N}$;
- $E(X^n)$ finite for all $n \in \mathbb{N}$ **does not imply** mgf exists;
- $M_X = M_Y$ **does imply** $F_X = F_Y$;
- $E(X^n) = E(Y^n)$ for all $n \in \mathbb{N}$ **does not imply** $F_X = F_Y$.