LECTURE 2: TRANSFORMATION AND EXPECTATION OF RANDOM VARIABLES

MECO 7312.
INSTRUCTOR: DR. KHAI CHIONG
SEPTEMBER 2, 2020

Let X be a random variable distributed with the cdf F_X. Suppose $g(\cdot)$ is some function, what is the distribution of $Y = g(X)$?

In many settings, we want to know the behavior of *functions* of random variables. Any function of random variable is also a random variable. Transformation and change of random variables is very important. For example, if $X \sim N(0,1)$ is the standard Gaussian random variable, then $Y = X^2$ is a Chi-squared distribution, which is an important class of distributions used in hypothesis testing. Further if $Y = e^X$, then Y has a log-normal distribution, which is used to model variables that take positive real values, such as income, asset prices, etc.

1. Transformation of Continuous Random Variables

Let $Y = g(X)$, F_Y denotes the cdf of Y and F_X denotes the cdf of X. From the definition of the cdf of Y:

$$\begin{aligned} F_Y(y) &= P_Y(Y \leq y) \\ &= P_Y(g(X) \leq y) \\ &= P_X(X \leq g^{-1}(y)) \text{ assuming } g \text{ is a strictly increasing, continuous function} \\ &= F_X(g^{-1}(y)) \end{aligned}$$

Therefore, we have expressed the cdf of Y in terms of the pdf of the original random variable X.

Examples:

1.) $X \sim U[-1,1]$ and $Y = \exp(X)$.

That is:

$$f_X(x) = \begin{cases} \frac{1}{2}, \text{ if } x \in [-1,1] \\ 0, \text{ otherwise} \end{cases}$$

$$F_X(x) = \begin{cases} 0, \text{ if } x < -1 \\ \frac{1}{2} + \frac{1}{2}x, \text{ if } x \in [-1,1] \\ 1, \text{ if } x \geq 1 \end{cases}$$

Therefore

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(\exp(X) \leq y) \\ &= P(X \leq \log y) \\ &= F_X(\log y) \end{aligned}$$

$$F_X(\log y) = \begin{cases} 0, \text{ if } \log y < -1 \\ \frac{1}{2} + \frac{1}{2}\log y, \text{ if } \log y \in [-1,1] \\ 1, \text{ if } \log y \geq 1 \end{cases}$$

As such, the cdf of Y is,

$$F_Y(y) = \begin{cases} 0, \text{ if } y < \frac{1}{e} \\ \frac{1}{2} + \frac{1}{2}\log y, \text{ if } y \in [\frac{1}{e}, e] \\ 1, \text{ if } y \geq e \end{cases}$$

The pdf of Y is $f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{1}{2y}$ for $y \in [\frac{1}{e}, e]$, and $f_Y(y) = 0$ for $y \notin [\frac{1}{e}, e]$.

2.) $X \sim U[-1,1]$ and $Y = X^2$.

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(X^2 \leq y) \\ &= P(-\sqrt{y} \leq X \leq \sqrt{y}) \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}) \\ &= 2F_X(\sqrt{y}) - 1 \\ &= \sqrt{y} \quad \text{for } y \in (0,1] \end{aligned}$$

The second last equality follows from $F_X(-\sqrt{y}) = 1 - F_X(\sqrt{y})$.

2. A general formula for the transformation of random variables

It is easy to derive the cdf and the pdf of Y when the transformation function g is a continuous monotonic function. A function is monotone (strictly) increasing if $u > v \implies g(u) > g(v)$, and a function is monotone (strictly) decreasing if $u > v \implies g(u) < g(v)$. We now derive a general formula for the transformation of a continuous random variable X.

Setup: let X be the random variable with the support $\mathcal{X}$. The support[1] of X is the region where the pdf of X is positive; outside of the support, the pdf is zero. Now let $Y = g(X)$, where g is monotone over $\mathcal{X}$. The support of Y is then $\mathcal{Y} = \{y \in \mathbb{R} : y = g(x) \text{ for some } x \in \mathcal{X}\}$.

If the transformation is monotone, then there is a bijection (one-to-one and onto)[2] between $\mathcal{X}$ and $\mathcal{Y}$. As such, $g^{-1}(y) = \{x \in \mathcal{X} : g(x) = y\}$ exists, and it is a single-valued monotone function, that is increasing if g is increasing, and it is decreasing if g is decreasing.

Hence, if $g(x)$ is a (strictly) increasing function, then:

$$\begin{aligned}
F_Y(y) &= P_Y(Y \le y) \\
&= P_X(g(X) \le y) \\
&= \int_{\{x \in \mathcal{X}: g(x) \le y\}} f_X(x)dx \\
&= \int_{\{x \in \mathcal{X}: x \le g^{-1}(y)\}} f_X(x)dx = P_X(X \le g^{-1}(y)) \\
&= \int_{-\infty}^{g^{-1}(y)} f_X(x)dx \\
&= F_X(g^{-1}(y))
\end{aligned}$$

The pdf is:

[1] The support is also the sample space

[2] One-to-one (injective): for all $x, x' \in \mathcal{X}$, $g(x) = g(x') \implies x = x'$. Onto (surjective): for each $y \in \mathcal{Y}$, there is an $x \in \mathcal{X}$ such that $g(x) = y$.

$$\begin{aligned} f_Y(y) &= \frac{dF_Y(y)}{dy} \\ &= \frac{dg^{-1}(y)}{dy}\frac{dF_X(g^{-1}(y))}{dx} \text{ by the chain rule} \\ &= \frac{dg^{-1}(y)}{dy} f_X(g^{-1}(y)) \end{aligned}$$

If $g(x)$ is a (strictly) decreasing function, then

$$\begin{aligned} F_Y(y) &= P_Y(Y \leq y) \\ &= P_X(g(X) \leq y) \\ &= P_X(X \geq g^{-1}(y)) \\ &= \int_{\{x \in \mathcal{X}: x \geq g^{-1}(y)\}} f_X(x)dx \\ &= \int_{g^{-1}(y)}^{\infty} f_X(x)dx \\ &= 1 - F_X(g^{-1}(y)) \end{aligned}$$

$$\begin{aligned} f_Y(y) &= \frac{dF_Y(y)}{dy} \\ &= -\frac{dg^{-1}(y)}{dy}\frac{dF_X(g^{-1}(y))}{dx} \end{aligned}$$

Moreover, $\frac{dg^{-1}(y)}{dy}$ has a negative sign when g is decreasing, and $\frac{dg^{-1}(y)}{dy}$ has a positive sign when g is increasing. Therefore we can succinctly rewrite the pdf of $Y = g(X)$ as:

$$f_Y(y) = \left|\frac{dg^{-1}(y)}{dy}\right| \frac{dF_X(g^{-1}(y))}{dx}, \text{ for } y \in \mathcal{Y}$$

Example:

Suppose $X \sim U[0,1]$, then $F_X(x) = x$ for $0 < x < 1$, and $f_X(x) = 1$ for $0 < x < 1$ (remember to specify the pdf and cdf completely, which is not done here). Further suppose that $Y = g(X) = -\log(X)$. Check that $g(x)$ is a monotone decreasing function over $0 < x < 1$ (whose derivative is $-\frac{1}{x} < 0$). As such, when the domain is restricted to $(0,1)$, the inverse of g exists and it is given by $g^{-1}(y) = e^{-y}$. Check that $g^{-1}(g(x)) = e^{\log x} = x$.

But what is the support (sample space) of Y? The function g maps $(0,1)$ bijectively to $(0,\infty)$. Therefore, the pdf of Y is:

$$f_Y(y) = \begin{cases} 0 & \text{if } y \leq 0 \\ e^{-y} & \text{if } y > 0 \end{cases}$$

The cdf of Y is:

$$F_Y(y) = \begin{cases} 0, & \text{if } y \leq 0 \\ 1 - F_X(g^{-1}(y)) = 1 - e^{-y}, & \text{if } y > 0 \end{cases}$$

2.1. Piecewise monotonic transformation

What if the function g is not monotone over the sample space $\mathcal{X}$? By Theorem 2.1.8 in Casella-Berger, we can partition $\mathcal{X}$ into $A_0, A_1, \ldots, A_k$ such that the function g is monotone over each $A_1, \ldots, A_k$. Then we can just apply the previous transformation formula separately over these sets, and then summing up the individual pdfs to obtain the overall pdf.

Let X has the pdf $f_X(x)$. Let the transformation be $Y = g(X)$. Let $A_0, A_1, \ldots, A_k$ be a partition of the support of X. Further, let $g_1, \ldots, g_k$ be monotone functions such that $g(x) = g_i(x)$ for $x \in A_i$. That is, g_i is the function g whose domain is restricted to the set A_i.

A_0 is an "exception" set $P_X(X \in A_0) = 0$, which can be ignored. We also assume that the pdf $f_X(x)$ is a continuous function on each A_i. Further, the functions g_i have identical range, in the sense that $\mathcal{Y} = \{y : y = g_i(x), \exists x \in A_i\}$ is the same for each i. In another words, each g_i is a one-to-one transformation from A_i onto $\mathcal{Y}$. Finally, $g_i^{-1}(y)$ has continuous derivative on $\mathcal{Y}$.

The pdf of $Y = g(X)$ is:

$$f_Y(y) = \begin{cases} \sum_{i=1}^k f_X(g_i^{-1}(y)) \left| \frac{dg_i^{-1}(y)}{dy} \right|, \text{ for } y \in \mathcal{Y} \\ 0, \text{ otherwise} \end{cases}$$

Example:

Let X have the standard Normal distribution. $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$ for $x \in (-\infty, \infty)$. Consider $Y = X^2$. The function $g(x) = x^2$ is monotone on $(-\infty, 0)$ and on $(0, \infty)$.

Let $A_0 = \{0\}$, $A_1 = (-\infty, 0)$, $A_2 = (0, \infty)$. Let $g_1(x) = x^2$ for $x < 0$, and $g_2(x) = x^2$ for $x > 0$. The respective inverses are: $g_1^{-1}(y) = -\sqrt{y}$ for $y > 0$, and $g_2^{-1}(y) = \sqrt{y}$ for $y > 0$. Thus the pdf of Y is:

$$\begin{aligned} f_Y(y) &= \frac{1}{\sqrt{2\pi}} e^{\frac{-(-\sqrt{y})^2}{2}} \left| -\frac{1}{2\sqrt{y}} \right| + \frac{1}{\sqrt{2\pi}} e^{\frac{-(\sqrt{y})^2}{2}} \left| \frac{1}{2\sqrt{y}} \right| \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{y}} e^{\frac{-y}{2}} \text{ for } y \in (0, \infty) \end{aligned}$$

Y is a chi-squared random variable with 1 degree of freedom. We check that all the technical conditions are satisfied. $P(X = 0) = 0$. $\frac{dg_1^{-1}(y)}{dy} = -\frac{1}{2\sqrt{y}}$ is continuous for $y > 0$. Finally, each g_i is a one-to-one function from A_i onto $\mathcal{Y} = \{y \in \mathbb{R} : y > 0\}$.

The inverse function theorem can be helpful in deriving $\frac{dg^{-1}(y)}{dy}$. It says that if $g(x)$ is a continuously differentiable function with nonzero derivative at the point $x = g^{-1}(y)$, then g is invertible in a neighborhood of $g^{-1}(y)$, the inverse is continuously differentiable, and the derivative of the inverse function at y is the reciprocal of the derivative of g at $g^{-1}(y)$:

$$\frac{dg^{-1}(y)}{dy} = \frac{1}{g'(x)} \bigg|_{x = g^{-1}(y)}$$

2.2. Probability integral transformation

Let X have continuous cdf $F_X(x)$ and define $Y = F_X(X)$. Then Y is uniformly distributed on $(0, 1)$, that is, $P(Y \leq y) = y$ for $0 < y < 1$.

$$\begin{aligned} P(Y \leq y) &= P(F_X(X) \leq y) \\ &= P(X \leq F_X^{-1}(y)) \text{ since } F_X \text{ is increasing} \\ &= F_X(F_X^{-1}(y)) \\ &= y \end{aligned}$$

Similarly, let Y be uniformly distributed on $(0, 1)$, and let $Z = F_X^{-1}(Y)$. Then Z has the cdf:

$$\begin{aligned} P(Z \leq z) &= P(F_X^{-1}(Y) \leq z) \\ &= P_Y(Y \leq F_X(z)) \text{ since } F_X^{-1} \text{ is increasing} \\ &= F_X(z) \end{aligned}$$

Z and X are identically distributed and has the same cdf. This result is important, as it allows us to generate random samples from any probability distribution. Suppose we want to draw a random sample x from a population with cdf F_X. First, we draw a uniform random number u between 0 and 1, then apply the transformation $F_X^{-1}(u)$.

Example:

Suppose we want to draw random samples $(x_1, \ldots, x_n)$ from the exponential distribution $F_X(x) = 1 - \exp(-x)$. First we draw $(u_1, \ldots, u_n)$ from $U[0, 1]$. Then let $x_i = F_X^{-1}(u_i) = \log(\frac{1}{1-u_i})$.

An example using R: see web appendix.

3. Transformation of Discrete Random Variables*

Let X be a discrete random variable, then $\mathcal{X}$, the sample space of X, is countable. Let the pmf of X be f_X, the sample space (or support) is $\mathcal{X} = \{x \in \mathbb{R} : f_X(x) > 0\}$.

The sample space for $Y = g(X)$ is $\mathcal{Y} = \{y \in \mathbb{R} : y = g(x), x \in \mathcal{X}\}$, which is also a countable set. Thus, Y is also a discrete random variable. The pmf for Y is:

$$\begin{aligned} f_Y(y) &= P_Y(Y = y) \\ &= P_X(g(X) = y) \\ &= P_X(\{x \in \mathcal{X} : g(x) = y\}) \\ &= \sum_{x \in \mathcal{X}: g(x) = y} P_X(X = x) \\ &= \sum_{x \in \mathcal{X}: g(x) = y} f_X(x) \end{aligned}$$

4. Expectation

The expected value, or mean, of a random variable $g(X)$ is:

$$\mathbb{E}[g(X)] = \begin{cases} \int_{-\infty}^{\infty} g(x) f_X(X) dx & \text{if } X \text{ is continuous} \\ \sum_{x \in \mathcal{X}} g(x) P(X = x) & \text{if } X \text{ is discrete} \end{cases}$$

In the formula above, expectation is the average of the values of the random variable, weighted by the probability distribution. Expected value is a commonly used measure of "central tendency" of a random variable X. Note: expectation is a population average, not sample average.

Properties of the expectation operator:

1.) Expectation is a linear operator: $\mathbb{E}[ag_1(X) + bg_2(X) + c] = a\,\mathbb{E}[g_1(X)] + b\,\mathbb{E}[g_2(X)] + c$.

2.) If $g_1(x) \geq 0$ for all $x \in \mathcal{X}$, then $\mathbb{E}[g_1(X)] \geq 0$.

3.) If $g_1(x) \geq g_2(x)$ for all $x \in \mathcal{X}$, then $\mathbb{E}[g_1(X)] \geq \mathbb{E}[g_2(X)]$.

Example:

If X has a binomial distribution $Bin(n, p)$ where n and p are parameters, its pmf is given by

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

1.) The binomial distribution is the discrete probability distribution of the number of successes in a sequence of n independent trials with binary outcomes, and where the probability of success in each trial is p.

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{x=1}^{n} x\binom{n}{x}p^x(1-p)^{n-x} \\
&= \sum_{x=1}^{n} n\binom{n-1}{x-1}p^x(1-p)^{n-x} \\
&= \sum_{y=0}^{n-1} n\binom{n-1}{y}p^{y+1}(1-p)^{n-(y+1)}\,, \text{ substitute } y = x-1 \\
&= np\sum_{y=0}^{n-1}\binom{n-1}{y}p^y(1-p)^{n-1-y} \\
&= np
\end{aligned}$$

Since $\sum_{y=0}^{n-1}\binom{n-1}{y}p^y(1-p)^{n-1-y}$ is the sum over all possible values of a binomial pmf with parameters $(n-1)$ and p.

Example:

Suppose X is Exponentially distributed with the parameter λ and has the pdf $f_X(x) = \lambda e^{-\lambda x}$. What is $\mathbb{E}[X]$?

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^\infty x\lambda e^{-\lambda x}dx \\
&= \left[-xe^{-\lambda x}\right]_0^\infty - \int_0^\infty -e^{-\lambda x}dx \\
&= 0 + \int_0^\infty e^{-\lambda x}dx \\
&= \left[-\frac{1}{\lambda}e^{-\lambda x}\right]_0^\infty \\
&= \frac{1}{\lambda}
\end{aligned}$$

5. Other central-tendency measures

The expected value of a random variable may not exist. A well-known example is the Cauchy distribution. However other central-tendency measures such as the median and the mode are well-defined in the case of the Cauchy distribution.

The Cauchy distribution has the pdf $f(x) = \frac{1}{\pi(1+x^2)}$.

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} xf(x)dx \\
&= \lim_{u\to\infty} \lim_{l\to-\infty} \int_{l}^{u} xf(x)dx \\
&= \lim_{l\to-\infty} \lim_{u\to\infty} \int_{l}^{u} xf(x)dx \qquad \text{for a well-defined intergral}
\end{aligned}$$

For an integral to be well-defined[3], we require that $\lim_{u\to\infty} \lim_{l\to-\infty} \int_l^u \frac{x}{\pi(1+x^2)} = \lim_{l\to-\infty} \lim_{u\to\infty} \int_l^u \frac{x}{\pi(1+x^2)}$.

For the Cauchy distribution, it can be shown that:

$$\lim_{u\to\infty} \lim_{l\to-\infty} \int_{l}^{u} \frac{x}{\pi(1+x^2)} = \lim_{u\to\infty} \lim_{l\to-\infty} \frac{\log(1+u^2)}{\pi} - \frac{\log(1+l^2)}{\pi} = -\infty$$

$$\lim_{l\to-\infty} \lim_{u\to\infty} \int_{l}^{u} \frac{x}{\pi(1+x^2)} = \lim_{l\to-\infty} \lim_{u\to\infty} \frac{\log(1+u^2)}{\pi} - \frac{\log(1+l^2)}{\pi} = \infty$$

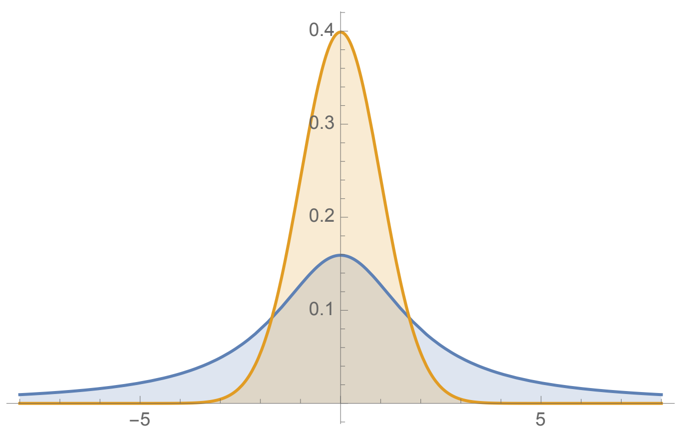

Figure 1. Cauchy distribution vs Normal distribution

The Cauchy distribution has much fatter tail than the Normal distribution. Consider another intuition for why the mean of the Cauchy distribution is undefined. For X that is distributed as Cauchy, $\mathbb{E}[X] = \int_{-\infty}^{0} xf(x)dx + \int_{0}^{\infty} xf(x)dx = -\infty + \infty$, which is undefined.

[3]Note that $\int_{-\infty}^{\infty} g(x)dx \neq \lim_{t\to\infty} \int_{-t}^{t} g(x)dx$. Indeed for an odd function $g(x)$, we always have $\lim_{t\to\infty} \int_{-t}^{t} g(x)dx = 0$. Therefore, $\lim_{t\to\infty} \int_{-t}^{t} \frac{x}{\pi(1+x^2)} dx = 0$. The Cauchy principal value is defined as $\lim_{t\to\infty} \int_{-t}^{t} g(x)dx$, for a function g even when $\int_{-\infty}^{\infty} g(x)dx$ is undefined.

5.1. Median

The median of the random variable X is $med(X) = m$ such that $F_X(m) = 0.5$. That is, the median is the value such that $\int_{-\infty}^{m} f_X(x)dx = 0.5$. It is robust to outliers, and has a nice invariance property: for $Y = g(X)$ and g monotonic increasing, then $med(Y) = g(med(X))$.

Example 1:

Suppose $X \sim \text{Exp}(\lambda)$ and has the pdf $f_X(x) = \lambda e^{-\lambda x}$ for $x > 0$. What is the median of X?

$$\begin{aligned} 0.5 &= \int_0^m \lambda e^{-\lambda x} dx \\ 0.5 &= \left[-e^{-\lambda x}\right]_0^m \\ 0.5 &= 1 - e^{-\lambda m} \\ m &= \frac{1}{\lambda}\log(2) \end{aligned}$$

Example 2:

What about the mean and median of $Y = \log(X)$, where X has the pdf e^{-x}?

$$\begin{aligned} \mathbb{E}[Y] &= \int_0^\infty \log(x)e^{-x} dx \\ &= -\text{Euler's constant} \\ &\approx -0.577216 \end{aligned}$$

In general, when X has the pdf $f_X(x) = \lambda e^{-\lambda x}$, we have $\mathbb{E}[\log(X)] = -\gamma - \log(\lambda)$.

Show that the median of Y is $\log(\log(2)) - \log(\lambda)$.

5.2. Mode

The mode of X is $Mode(X) = \text{argmax}_x f_X(x)$. That is, the mode is the peak of the pdf of X.

Suppose X has the pdf $f_X(x) = \lambda e^{-\lambda x}$. The mode of X is $\text{argmax}_{x \geq 0} \lambda e^{-\lambda x} = 0$.

Suppose $Y = \log(X)$. The mode of Y is $\operatorname{argmax}_y e^y e^{-e^y} = \operatorname{argmax}_{y\in\mathbb{R}} y - e^y = 0$.

The solution can be found by calculating the first-order condition and then checking the second-order condition. For more complicated functions, we can find the solution numerically using just one line of code in MATLAB:

`fminunc(@(y)exp(y)-y,-2)`.

In R, it can be implemented as:

```
optimize(function(x) exp(x) - x,c(-10,10))
```

6. Higher moments

For each integer n, the n-th moment of X is defined as $\mathbb{E}[X^n]$.

The n-th centered moment of X is $\mathbb{E}[(X - \mathbb{E}[X])^n]$.

The mean $\mathbb{E}[X]$ is the first moment of X, and the variance is the second centered moment of X.

The variance of the random variable X is defined as $\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$. The positive square root of $\text{Var}(X)$ is the standard deviation of X.

Properties of the variance:

1.) $\text{Var}(aX + b) = a^2\text{Var}(X)$. Variance is not a linear operation. Moreover, variance measures the spread of a distribution around its mean, and so it is unaffected when a constant is added to the X.

2.) $\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$ (alternative formula for the variance)

Example 1:

Suppose X has the pdf $f_X(x) = \lambda e^{-\lambda x}$. What is the variance of X?

$$\begin{aligned}
\mathbb{E}[X^2] &= \int_0^\infty x^2 \lambda e^{-\lambda x} dx \\
&= \left[-x^2 e^{-\lambda x}\right]_0^\infty - \int_0^\infty -2xe^{-\lambda x} dx \\
&= 0 + \int_0^\infty 2xe^{-\lambda x} dx \\
&= \frac{2}{\lambda^2}
\end{aligned}$$

Therefore $\text{Var}(x) = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$

Example 2:

Suppose X has the pdf $f_X(x) = \lambda e^{-\lambda x}$. What is the variance of $Y = \log(X)$?

$$\begin{aligned}
\mathbb{E}[Y^2] &= \mathbb{E}[\log(X)^2] \\
&= \int_0^\infty \log(x)^2 \lambda e^{-\lambda x} dx \\
&= \gamma^2 + \frac{\pi^2}{6} + 2\gamma \log(\lambda) + \log(\lambda)^2 \\
&= \frac{\pi^2}{6} + (\gamma + \log(\lambda))^2
\end{aligned}$$

Therefore $\text{Var}(Y) = \frac{\pi^2}{6} \approx 1.64493$, which does not depend on λ.

What information does the third moment convey? Consider the third-centered moment of a random variable, $\mathbb{E}[(X - \mathbb{E}[X]^3)]$.

Going back to our example. Let $X \sim \text{Exp}(\lambda)$.

$$\begin{aligned}
\mathbb{E}[(X - \mathbb{E}[X])^3] &= \int_0^\infty \left(x - \frac{1}{\lambda}\right)^3 \lambda e^{-\lambda x} dx \\
&= \frac{2}{\lambda^3} \\
&> 0
\end{aligned}$$

Now let $Y = \log(X)$, and consider the third-centered moment of Y.

$$\begin{aligned}
\mathbb{E}[(Y - \mathbb{E}[Y])^3] &= \int_{-\infty}^\infty (y + \gamma + \log(\lambda))^3 \lambda e^y e^{-\lambda e^y} dy \\
&= -2\zeta(3) \\
&\approx -2.40411 < 0
\end{aligned}$$

Where $\zeta(s)$ is the Riemann-Zeta function. In particular, $\zeta(3) = \sum_{n=1}^\infty \frac{1}{n^3}$.

The third-centered moment conveys information about the **skewness** of a random variable. A negative skewness value means the tail is on the left side of the distribution, and positive skewness indicates that the tail is on the right. Verify this visually, using the fact that $Y = \log X$ has the pdf $f_Y(y) = \lambda e^{y - \lambda e^y}$.

In order for the third-centered moment to be comparable across different scales of random variables, **skewness** is defined as the third *standardized* moment. Going back to our examples:

$$\begin{aligned}\mathbb{E}\left[\left(\frac{X-\mathbb{E}[X]}{\sqrt{\text{Var}(X)}}\right)^3\right] &= \frac{1}{\text{Var}(X)^{3/2}}\mathbb{E}[(X-\mathbb{E}[X])^3] \\ &= \lambda^3\frac{2}{\lambda^3} \\ &= 2\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\left(\frac{Y-\mathbb{E}[Y]}{\sqrt{\text{Var}(Y)}}\right)^3\right] &= \frac{-2\zeta(3)}{(\frac{\pi^2}{6})^{3/2}} \\ &\approx -1.13955\end{aligned}$$

Which does not depend on λ. Again, we emphasize that these are the *population* moments (population variance, population skewness, etc). These are theoretical values – true values associated with a random variable. Later, we will talk about *sample* moments: given a vector of numbers $x_1, \dots, x_n$, how do we calculate the sample mean, sample variance, sample skewness, etc.

7. Moments Generating Function

The moments of a random variable are summarized in the moment generating function (mgf). Definition: the moment-generating function of X is $M_X(t) \equiv \mathbb{E}[\exp(tX)]$, provided that the expectation exists in some neighborhood $t \in [-h, h]$ of zero.

Specifically,

$$M_X(t) = \begin{cases} \int_{-\infty}^{\infty} e^{tx} f_X(x)dx, & \text{for } X \text{ continuous} \\ \sum_{x\in\mathcal{X}} e^{tx} P(X=x), & \text{for } X \text{ discrete} \end{cases}$$

The mgf has the property that

$$\mathbb{E}[X^n] = \frac{d^n}{dt^n} M_X(t)\bigg|_{t=0}$$

That is, the n-th derivative of the MGF evaluated at $t = 0$ gives the n-th moment of the corresponding random variable. Another notation for $\frac{d^n}{dt^n} M_X(t)\Big|_{t=0}$ is $M_X^{(n)}(0)$.

When it exists, then mgf provides alternative description of a probability distribution. Mathematically, it is a Laplace transform, which can be convenient for certain mathematical calculations.

Example:

Let X be the standard Normal distribution. As such $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$.

$$\begin{aligned}
M_X(t) &= \mathbb{E}[e^{tX}] \\
&= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2+2tx}{2}} dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-(x-t)^2+t^2}{2}} dx \\
&= e^{\frac{t^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-t)^2}{2}} dx \\
&= e^{\frac{t^2}{2}}
\end{aligned}$$

First moment of X is $M^{(1)}(0) = e^{\frac{t^2}{2}}\big|_{t=0} = 0$.

The second moment of X is $M^{(2)}(0) = \frac{d}{dt} te^{\frac{t^2}{2}}\Big|_{t=0} = \left(e^{\frac{t^2}{2}} + t(te^{\frac{t^2}{2}})\right)\Big|_{t=0} = 1$.

The third moment of X is $M^{(3)}(0) = \left(te^{\frac{t^2}{2}} + 2te^{\frac{t^2}{2}} + t^3 e^{\frac{t^2}{2}}\right)\Big|_{t=0} = 0$.

Moment generating function will be useful later when we talk about the central limit theorem. Moreover, mgf has the nice property that:

Let $S = \sum_{i=1}^{n} a_i X_i$, where X_i are independent random variables. The mgf for S is given by $M_S(t) = M_{X_1}(a_1 t) \times M_{X_2}(a_2 t) \times \cdots \times M_{X_n}(a_n t)$.

To see the intuition behind mgf, consider the Taylor series expansion of e^{tx} around $t = 0$

$$
\begin{aligned}
g(t) &= g(0) + \frac{1}{1!} t g^{(1)}(0) + \frac{1}{2!} t^2 g^{(2)}(0) + \dots \\
e^{tx} &= 1 + \frac{1}{1!} tx + \frac{1}{2!} t^2 x^2 + \frac{1}{3!} t^3 x^3 + \dots \\
\mathbb{E}[e^{tX}] &= 1 + t\,\mathbb{E}[X] + \frac{1}{2!} t^2\,\mathbb{E}[X^2] + \frac{1}{3!} t^3\,\mathbb{E}[X^3] + \dots
\end{aligned}
$$

Hence the first-derivative of $\mathbb{E}[e^{tX}]$ with respect to t evaluated at $t = 0$ is $\mathbb{E}[X]$, the second-derivative of $\mathbb{E}[e^{tX}]$ w.r.t t evaluated at $t = 0$ is $\mathbb{E}[X^2]$, and so on.