MTL108
Transformation of Random Variables

Rahul Singh

Let X be a random variable with a known probability distribution (either a Probability Mass Function (PMF) for discrete variables or a Probability Density Function (PDF) for continuous variables), and let $Y = g(X)$ be a new random variable.

- In practice, transformations arise in almost every branch of study. A popular example is Body-Mass-Index (BMI), which is defined as

$$\text{BMI} = \frac{\text{weight in KG}}{(\text{height in meter})^2}$$

- Simple example: If X is the result of rolling a die, $Y = (X-3)^2$ is another random variable defined from X.

How do we find the probability distribution of $Y = g(X)$?

Transformation of Discrete Random Variables

Let X be a discrete random variable with a PMF $p_X(x)$. If we define a new random variable $Y = g(X)$, its PMF $p_Y(y)$ is found by summing the probabilities for all values of X that map to a specific value of Y.

$$p_Y(y) = \mathbb{P}(Y = y) = \sum_{\{x:g(x)=y\}} p_X(x)$$

Example 1 (Discrete transformation)**.** Let X be the outcome of a fair die roll, so $p_X(x) = 1/6$ for $x \in \{1, 2, 3, 4, 5, 6\}$. Let $Y = (X-3)^2$. We find the PMF of Y.

- $Y = 0$: Occurs when $X = 3$; $p_Y(0) = p_X(3) = 1/6$.
- $Y = 1$: Occurs when $X = 2$ or $X = 4$; $p_Y(1) = p_X(2) + p_X(4) = 1/6 + 1/6 = 1/3$.
- $Y = 4$: Occurs when $X = 1$ or $X = 5$; $p_Y(4) = p_X(1) + p_X(5) = 1/6 + 1/6 = 1/3$.
- $Y = 9$: Occurs when $X = 0$ or $X = 6$; $p_Y(9) = p_X(0) + p_X(6) = 0 + 1/6 = 1/6$.

Thus the PMF of Y is

$$p_Y(y) = \begin{cases} 1/6 & y = 0, \\ 1/3 & y = 1, \\ 1/3 & y = 4, \\ 1/6 & y = 9, \\ 0 & \text{otherwise.} \end{cases}$$

CDF Method (General Approach)

The CDF method works for both discrete and continuous random variables, especially when g is not one-to-one. Steps of this method are as follows:

1. Define $F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y)$.
2. Express the event $\{g(X) \leq y\}$ in terms of X.
3. Evaluate using F_X (the distribution function of X), by solving the inequality for X in terms of y. This might require considering different cases if g is not one-to-one.
4. If Y is discrete, identify mass points using jump points in the CDF and compute them.
5. If Y is continuous, differentiate to obtain $f_Y(y)$.

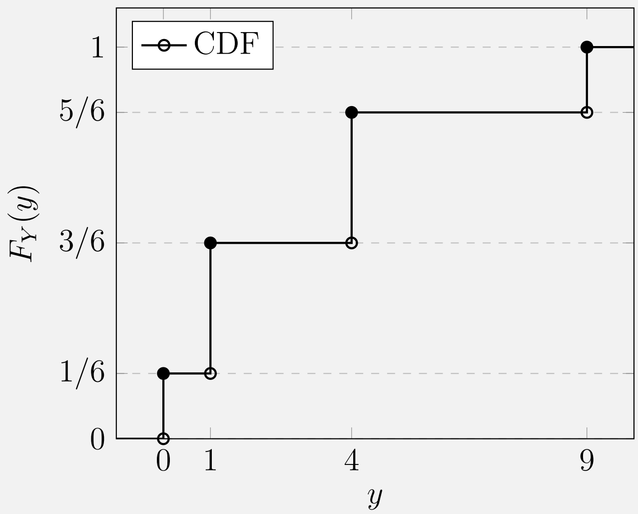
Example 2 (Transformation of a die roll)**.** Let X be a fair die outcome, $p_X(x) = 1/6$ for $x = 1, 2, 3, 4, 5, 6$. Define $Y = (X - 3)^2$. Then,

$$\begin{aligned} Y = 0 &\Leftrightarrow X = 3, \\ Y = 1 &\Leftrightarrow X = 2 \text{ or } 4, \\ Y = 4 &\Leftrightarrow X = 1 \text{ or } 5, \\ Y = 9 &\Leftrightarrow X = 6. \end{aligned}$$

Thus the CDF of Y is

$$F_Y(y) = \mathbb{P}(Y \leq y) = \begin{cases} \mathbb{P}(\emptyset) = 0, & y < 0, \\ \mathbb{P}(X = 3) = 1/6, & 0 \leq y < 1, \\ \mathbb{P}(X \in \{2, 3, 4\}) = 3/6, & 1 \leq y < 4, \\ \mathbb{P}(X \in \{1, 2, 3, 4, 5\}) = 5/6, & 4 \leq y < 9, \\ \mathbb{P}(X \in \{1, 2, 3, 4, 5, 6\}) = 1, & y \geq 9. \end{cases}$$

The CDF plot is



Clearly, the jump points for CDF are $\{0, 1, 4, 9\}$. Thus the PMF of Y is given by

$$p_Y(y) = \begin{cases} F_Y(0) - \lim F_Y(0-) = 1/6 - 0 = 1/6 & y = 0, \\ F_Y(1) - \lim F_Y(1-) = 3/6 - 1/6 = 1/3 & y = 1, \\ F_Y(4) - \lim F_Y(4-) = 5/6 - 3/6 = 1/3 & y = 4, \\ F_Y(9) - \lim F_Y(9-) = 1 - 5/6 = 1/6 & y = 9, \\ 0 & \text{otherwise.} \end{cases}$$

Example 3 (Squaring a continuous random variable)**.** Let X be a continuous random variable with PDF $f_X(x) = \frac{1}{2}x$ for $0 < x < 2$. We want to find the PDF of $Y = X^2$.

- CDF of X: $F_X(x) = \int_0^x \frac{1}{2}t\,dt = \left[\frac{t^2}{4}\right]_0^x = \frac{x^2}{4}$ for $0 < x < 2$.
- CDF of Y:
$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(X^2 \leq y).$$
Since $X > 0$, we have $X \leq \sqrt{y}$. The support of Y is $y \in (0, 4)$.
$$F_Y(y) = \mathbb{P}(X \leq \sqrt{y}) = F_X(\sqrt{y}) = \frac{(\sqrt{y})^2}{4} = \frac{y}{4} \quad \text{for} \quad 0 < y < 4.$$
- Differentiating CDF to find the PDF of Y:
$$f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{d}{dy}\left(\frac{y}{4}\right) = \frac{1}{4} \quad \text{for} \quad 0 < y < 4.$$
This shows that Y is a uniform random variable on the interval $(0, 4)$, i.e.,
$$Y \sim Uniform(0, 4).$$

One-to-One Continuous Transformation

When the transformation $Y = g(X)$ is a one-to-one monotonic function, we can use the following formula directly to find the PDF of Y.

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|$$

where $g^{-1}(y)$ is the inverse function of $g(x)$.

Example 4 (Transformation of an Exponential distribution)**.** Let $X \sim \text{Exp}(\lambda)$, with PDF $f_X(x) = \lambda e^{-\lambda x}$ for $x > 0$. Let $Y = 3X$. Find the PDF of Y.

- The transformation is
$$g(x) = 3x.$$
- The inverse is
$$g^{-1}(y) = y/3.$$
- The derivative of the inverse is
$$\frac{d}{dy} g^{-1}(y) = \frac{1}{3}.$$
- The PDF of Y is:
$$f_Y(y) = f_X(y/3) \cdot \left| \frac{1}{3} \right| = \lambda e^{-\lambda(y/3)} \cdot \frac{1}{3} = \frac{\lambda}{3} e^{-(\lambda/3)y}.$$

  This shows that Y is also exponentially distributed, with rate parameter $\lambda/3$, i.e., $Y \sim \text{Exp}(\lambda/3)$.

Method of Moment Generating Functions (MGF)

For a sum of independent random variables, the MGF of the sum is the product of the individual MGFs, that is, if $X_1, \ldots, X_n$ are independent RVs and $Y = X_1 + \cdots + X_n$ then

$$M_Y(t) = M_{X_1+\cdots+X_n}(t) = \prod_{i=1}^{n} M_{X_i}(t).$$

So, the steps involved are

1. Identify the MGF of each independent random variable.
2. Multiply the MGFs to find the MGF of the sum.
3. Identify the distribution corresponding to the resulting MGF.

Example 5 (Sum of Poisson random variables)**.** Let $X_1 \sim \text{Poisson}(\lambda_1)$ and $X_2 \sim \text{Poisson}(\lambda_2)$ be independent random variables. Find the distribution of $Y = X_1 + X_2$.

1. The MGF of a Poisson variable is $M_X(t) = e^{\lambda(e^t-1)}$.
2. The MGF of Y is the product of the MGFs of X_1 and X_2: $M_Y(t) = M_{X_1}(t) \cdot M_{X_2}(t) = e^{\lambda_1(e^t-1)} \cdot e^{\lambda_2(e^t-1)} = e^{(\lambda_1+\lambda_2)(e^t-1)}$.
3. This is the MGF of a Poisson distribution with parameter $\lambda = \lambda_1 + \lambda_2$. Thus, $Y \sim \text{Poisson}(\lambda_1 + \lambda_2)$.

Multivariate Transformations (optional)

For a random vector $\mathbf{X} = (X_1, \ldots, X_n)$ with joint PDF $f_{\mathbf{X}}(\mathbf{x})$, and a one-to-one transformation $\mathbf{Y} = g(\mathbf{X})$, the joint PDF of $\mathbf{Y}$ is given by:

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(g^{-1}(\mathbf{y})) \cdot |J|$$

where J is the Jacobian determinant of the inverse transformation $\mathbf{X} = g^{-1}(\mathbf{Y})$.

$$J = \det \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial y_1} & \cdots & \frac{\partial x_n}{\partial y_n} \end{pmatrix}.$$

Example 6 (Sum and Difference of Normal variables)**.** Let $X_1, X_2 \sim N(0,1)$ be independent. Find the joint PDF of $Y_1 = X_1 + X_2$ and $Y_2 = X_1 - X_2$.

1. The joint PDF of X_1, X_2 is
$$f_{X_1,X_2}(x_1, x_2) = \frac{1}{2\pi} e^{-\frac{1}{2}(x_1^2+x_2^2)}.$$
2. The inverse transformation is
$$x_1 = \frac{1}{2}(y_1 + y_2) \quad \textit{and} \quad x_2 = \frac{1}{2}(y_1 - y_2).$$
3. The Jacobian determinant is
$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix} = \begin{vmatrix} 1/2 & 1/2 \\ 1/2 & -1/2 \end{vmatrix} = (1/2)(-1/2) - (1/2)(1/2) = -1/2.$$
4. The joint PDF of Y_1, Y_2 is:
$$f_{Y_1,Y_2}(y_1, y_2) = f_{X_1,X_2}(\frac{y_1+y_2}{2}, \frac{y_1-y_2}{2})|J| = \frac{1}{2\pi} e^{-\frac{1}{2}\left[\left(\frac{y_1+y_2}{2}\right)^2 + \left(\frac{y_1-y_2}{2}\right)^2\right]} \cdot \frac{1}{2}.$$
So,
$$f_{Y_1,Y_2}(y_1, y_2) = \frac{1}{4\pi} e^{-\frac{1}{2}\left[\frac{y_1^2+2y_1y_2+y_2^2}{4} + \frac{y_1^2-2y_1y_2+y_2^2}{4}\right]} = \frac{1}{4\pi} e^{-\frac{1}{4}(y_1^2+y_2^2)}.$$
This can be factored into two independent normal distributions, $N(0,2)$.

Example 7. Let X_1, X_2 be IID standard normal variables, and consider the polar coordinate transformation:

$$Y_1 = \sqrt{X_1^2 + X_2^2} \quad \text{and} \quad Y_2 = \arctan\left(\frac{X_2}{X_1}\right)$$

The inverse transformation is

$$X_1 = Y_1 \cos(Y_2) \quad \textit{and} \quad X_2 = Y_1 \sin(Y_2).$$

The Jacobian determinant is

$$J = \det\begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{pmatrix} = \det\begin{pmatrix} \cos(y_2) & -y_1 \sin(y_2) \\ \sin(y_2) & y_1 \cos(y_2) \end{pmatrix} = y_1 \cos^2(y_2) + y_1 \sin^2(y_2) = y_1$$

Using the transformation formula and the fact that

$$f_{X_1,X_2}(x_1, x_2) = \frac{1}{2\pi} e^{-\frac{1}{2}(x_1^2+x_2^2)},$$

we get the joint PDF of Y_1 and Y_2:

$$f_{Y_1,Y_2}(y_1, y_2) = \frac{1}{2\pi} e^{-\frac{1}{2}((y_1\cos(y_2))^2+(y_1\sin(y_2))^2)} \cdot |y_1| = \frac{1}{2\pi} e^{-\frac{1}{2}y_1^2} \cdot y_1$$

This factorization shows that Y_1 and Y_2 are independent, and reveals their respective distributions.

Order Statistics (Optional)

In statistical analysis, it is often useful to arrange a sample of random variables in ascending order. The resulting ordered variables are known as *order statistics*. Order statistics are fundamental in non-parametric statistics and provide a basis for robust estimation and outlier detection. They allow us to analyze quantities like the minimum, maximum, median, and percentiles of a sample, which are often of greater interest than the average value.

Definition 1 (Order Statistics)**.** Let $X_1, X_2, \ldots, X_n$ be a random sample of n independent and identically distributed (IID) random variables from a continuous distribution with cumulative distribution function (CDF) $F_X(x)$ and probability density function (PDF) $f_X(x)$. Let these random variables be sorted in non-decreasing order:

$$X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$$

The variable $X_{(r)}$ is called the r**-th order statistic**.

Remark 1. For a continuous distribution, the probability that any two observations are equal is zero, so with probability 1 we have $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$. The discussion of order statistics focuses mainly developed for continuous random variables, so henceforth we assume RVs are continuous.

Distributions of Extremes (Minimum and Maximum)

The most straightforward order statistics are the minimum, $X_{(1)}$, and the maximum, $X_{(n)}$. Their distributions are simple to derive using the CDF method.

CDF and PDF of the Maximum $X_{(n)}$

The event $\{X_{(n)} \leq x\}$ occurs if and only if *all* of the random variables $X_1, \ldots, X_n$ are less than or equal to x. Due to the independence and identical distribution of the random variables, the CDF of $X_{(n)}$ is:

$$F_{X_{(n)}}(x) = \mathbb{P}(X_{(n)} \leq x) = \mathbb{P}(X_1 \leq x, \ldots, X_n \leq x) = \prod_{i=1}^{n} \mathbb{P}(X_i \leq x) = [F_X(x)]^n$$

The PDF is found by differentiating the CDF using the chain rule:

$$f_{X_{(n)}}(x) = \frac{d}{dx} F_{X_{(n)}}(x) = n[F_X(x)]^{n-1} f_X(x)$$

CDF and PDF of the Minimum $X_{(1)}$

The event $\{X_{(1)} > x\}$ occurs if and only if *all* of the random variables $X_1, \ldots, X_n$ are greater than x. Therefore, we can express the CDF of $X_{(1)}$ as:

$$F_{X_{(1)}}(x) = 1 - \mathbb{P}(X_{(1)} > x) = 1 - \mathbb{P}(X_1 > x, \ldots, X_n > x) = 1 - \prod_{i=1}^{n} \mathbb{P}(X_i > x) = 1 - [1 - F_X(x)]^n$$

The PDF is found by differentiating:

$$f_{X_{(1)}}(x) = \frac{d}{dx} F_{X_{(1)}}(x) = -n[1 - F_X(x)]^{n-1}(-f_X(x)) = n[1 - F_X(x)]^{n-1} f_X(x)$$

PDF of the r-th Order Statistic $X_{(r)}$

The PDF of the r-th order statistic $X_{(r)}$ is given by:

$$f_{X_{(r)}}(x) = \frac{n!}{(r-1)!(n-r)!} [F_X(x)]^{r-1} [1 - F_X(x)]^{n-r} f_X(x)$$

This formula can be understood intuitively:

- $f_X(x)dx$: The probability that one of the random variables falls in a small interval around x.
- $[F_X(x)]^{r-1}$: The probability that $r-1$ variables are less than x.
- $[1 - F_X(x)]^{n-r}$: The probability that $n-r$ variables are greater than x.
- $\frac{n!}{(r-1)!(n-r)!}$: The number of ways to choose which variables are less than x, which one is at x, and which are greater than x.

Example 8 (Median of a Uniform Distribution)**.** Let X_1, X_2, X_3 be IID $U(0,1)$. The PDF and CDF are $f_X(x) = 1$ and $F_X(x) = x$ for $0 < x < 1$. We want the distribution of the median, $X_{(2)}$. Using the formula for $r = 2, n = 3$:

$$f_{X_{(2)}}(x) = \frac{3!}{(2-1)!(3-2)!}[x]^{2-1}[1-x]^{3-2}(1) = 6x(1-x), \quad 0 < x < 1$$

This is the PDF of a Beta(2,2) distribution.

Joint PDF of Two Order Statistics $X_{(r)}$ and $X_{(s)}$

For $1 \leq r < s \leq n$, the joint PDF of $X_{(r)}$ and $X_{(s)}$ is:

$$f_{X_{(r)},X_{(s)}}(x,y) = \frac{n!}{(r-1)!(s-r-1)!(n-s)!}[F_X(x)]^{r-1}[F_X(y)-F_X(x)]^{s-r-1}[1-F_X(y)]^{n-s}f_X(x)f_X(y)$$

This formula helps derive the distributions of other useful statistics like the sample range, $R = X_{(n)} - X_{(1)}$.

Applications of Order Statistics

Order statistics have broad applications across various fields due to their robustness and direct connection to data ranking.

1. In reliability engineering, products are tested to determine their lifetime, often following an exponential or Weibull distribution. Order statistics are used to analyze these failure times.

Example 9 (System Lifetime in Life Testing)**.** A system is composed of $n = 5$ identical components. The lifetime of each component is exponentially distributed with a mean of 1000 hours.

(a) **Series System (Weakest Link):** The system fails if any one component fails. The lifetime of the system is the minimum of the component lifetimes, $X_{(1)}$. For an exponential distribution with rate $\lambda = 1/1000$, we have $F_X(x) = 1 - e^{-\lambda x}$.

$$f_{X_{(1)}}(x) = 5[1-(1-e^{-\lambda x})]^4 \lambda e^{-\lambda x} = 5\lambda e^{-5\lambda x}$$

The system lifetime is exponentially distributed with rate 5λ. The expected lifetime is $1/(5\lambda) = 1000/5 = 200$ hours. This illustrates how a complex system can be significantly less reliable than its individual components.

(b) **Parallel System (Redundancy):** The system fails only when all components have failed. The lifetime of the system is the maximum of the component lifetimes, $X_{(5)}$.

$$F_{X_{(5)}}(x) = [F_X(x)]^5 = (1-e^{-\lambda x})^5$$

$$f_{X_{(5)}}(x) = 5(1-e^{-\lambda x})^4 \lambda e^{-\lambda x}$$

The expected lifetime of the parallel system is $E[X_{(5)}] = \int_0^\infty 5x(1-e^{-\lambda x})^4 \lambda e^{-\lambda x} dx$, which is significantly greater than the single component lifetime.

2. Order statistics are a cornerstone of statistical process control, which monitors manufacturing processes to ensure quality standards.

> **Example 10** (Detecting Outliers in Manufacturing)**.** A company produces parts with a target weight of 5 grams and a standard deviation of 0.01 grams, assumed to be normally distributed. To check for a quality problem, they take a sample of $n = 25$ candies. If the range of the sample, $X_{(25)} - X_{(1)}$, is too large, it may indicate an issue with the process. Using the joint PDF of the extremes, one can calculate the probability of observing a certain range under normal conditions and set a control limit. If a sample's range exceeds this limit, a quality investigation is triggered.

3. Extreme order statistics are essential for studying extreme weather events, such as floods, droughts, and heatwaves.

> **Example 11** (Modeling Environmental Extremes)**.** Consider the annual maximum river levels over a period of n years. The maximum river level each year can be modeled as an IID random variable. The annual maximum over n years is then the maximum order statistic $X_{(n)}$ of these measurements. Extreme value theory, a branch of statistics heavily reliant on order statistics, is used to model and predict the probability of rare, catastrophic events like floods. These models inform the design of infrastructure like dams and levees.

4. In finance, order statistics are used to analyze extreme events, such as maximum losses or minimum asset values, which is crucial for risk management.

References

[1] Blitzstein, J. K., & Hwang, J. (2019). *Introduction to probability.* Chapman and Hall/CRC.

Disclaimer