Transformations

1 Transformations of Random Variables

Let's suppose that we have got some random variables. We can make new random variables out of those, using functions. How do those new random variables behave? We will take a look at that now. This chapter of the summary is rather difficult, while it is not incredibly important. So do not stare yourself blind on this part.

1.1 Finding the CDF

Suppose we have a random variable $\underline{x}$. We can define a new variable $\underline{y}$ to be a function of $\underline{x}$, so $\underline{y} = g(\underline{x})$. Now we would like to know how we can find the CDF $F_{\underline{y}}(y)$. It can be found using

$$F_{\underline{y}}(y) = P(\underline{y} \leq y) = P(g(\underline{x}) \leq y) = P(\underline{x} \in I_y), \tag{1.1}$$

where the set I_y consists of all x such that $g(x) \leq y$. So, to find the CDF $F_{\underline{y}}(y)$, we first need to find I_y: We need to know for what x we have $g(x) \leq y$. The intervals that are found can then be used to express $F_{\underline{y}}$ in $F_{\underline{x}}$.

Let's look at a special case. When $g(x)$ is strictly increasing, or strictly decreasing, we have

$$F_{\underline{y}}(y) = F_{\underline{x}}(g^{-1}(y)) \quad \text{for increasing } g(x) \text{ and} \qquad F_{\underline{y}}(y) = 1 - F_{\underline{x}}(g^{-1}(y)) \quad \text{for decreasing } g(x). \tag{1.2}$$

Here the function $g^{-1}(y)$ is the inverse of $g(x)$. It is defined such that if $y = g(x)$, then $x = g^{-1}(y)$.

1.2 Finding the PDF

Now that we've got the CDF $F_{\underline{y}}(y)$, it's time to find the PDF $f_{\underline{y}}(y)$. You probably remember that the PDF is simply the derivative of the CDF. That rule can be used to find the PDF.

Let's consider the special case that $g(x)$ is either strictly increasing or strictly decreasing. Now we have

$$f_{\underline{y}}(y) = \frac{dF_{\underline{y}}(y)}{dy} = \begin{cases} f_{\underline{x}}(g^{-1}(y))\frac{dg^{-1}(y)}{dy} & \text{for increasing } g(x) \\ -f_{\underline{x}}(g^{-1}(y))\frac{dg^{-1}(y)}{dy} & \text{for decreasing } g(x) \end{cases} = f_{\underline{x}}(g^{-1}(y))\left|\frac{dg^{-1}(y)}{dy}\right|. \tag{1.3}$$

Note that we have simply taken the derivative of equation (1.2), using the chain rule. Also note that if $g(x)$ is decreasing, also $g^{-1}(y)$ is decreasing, and thus $dg^{-1}(y)/dy$ is negative. This explains the last step in the above equation, where the absolute stripes $|\ldots|$ suddenly appear.

Now what should we do if $g(x)$ is not increasing or decreasing? In this case no inverse function $g^{-1}(y)$ exists. Let's suppose that for a given y the equation $y = g(x)$ has n solutions $x_1, x_2, \ldots, x_n$. Now we can say that

$$f_{\underline{y}}(y) = \sum_{i=1}^{n} \frac{f_{\underline{x}}(x_i)}{\left|\frac{dg(x_i)}{dx}\right|}. \tag{1.4}$$

If only one solution x_i is present, then this equation reduces back to equation (1.3).

1.3 Functions of two random variables

Let's suppose we now have two random variables $\underline{x}_1$ and $\underline{x}_2$. Also, let's define $\underline{y} = g(\underline{x}_1, \underline{x}_2)$. In this case, we can find the CDF $F_{\underline{y}}(y)$ using

$$F_{\underline{y}}(y) = P(\underline{y} \leq y) = P(g(\underline{x}_1, \underline{x}_2) \leq y) = P((\underline{x}_1, \underline{x}_2) \in D_y), \tag{1.5}$$

where the set D_y consists of all the pairs (x_1, x_2) such that $g(x_1, x_2) \leq y$. To find the PDF, we can use

$$f_{\underline{y}}(y) = \int_{-\infty}^{\infty} f_{\underline{x}_1,\underline{x}_2}(x_1, g^{-1}(x_1, y)) \left| \frac{dg^{-1}(x_1, y)}{dy} \right| dx_1 = \int_{-\infty}^{\infty} f_{\underline{x}_1,\underline{x}_2}(g^{-1}(y, x_2), x_2) \left| \frac{dg^{-1}(y, x_2)}{dy} \right| dx_2. \tag{1.6}$$

1.4 Transformations of two random variables

Now let's not only define $\underline{y}_1 = g_1(\underline{x}_1, \underline{x}_2)$, but also $\underline{y}_2 = g_2(\underline{x}_1, \underline{x}_2)$. Now we can find the joint CDF using

$$F_{\underline{y}_1,\underline{y}_2}(y_1, y_2) = P(\underline{y}_1 \leq y_1, \underline{y}_2 \leq y_2) = P(g_1(\underline{x}_1, \underline{x}_2) \leq y_1, g_2(\underline{x}_1, \underline{x}_2) \leq y_2) = P((\underline{x}_1, \underline{x}_2) \in D_{y_1,y_2}), \tag{1.7}$$

where the region D_{y_1,y_2} is the intersection of the regions D_{y_1} and D_{y_2}. We can now find the joint PDF by differentiating the CDF. We then get

$$f_{y_1,y_2}(y_1, y_2) = \frac{\partial^2 F_{\underline{y}_1,\underline{y}_2}(y_1, y_2)}{\partial y_1 \, \partial y_2} = \frac{\partial^2}{\partial y_1 \, \partial y_2} \int \int_{D_{y_1,y_2}} f_{\underline{x}_1,\underline{x}_2}(x_1, x_2) \, dx_1 \, dx_2. \tag{1.8}$$

There is, however, another way to find the joint PDF. For that, let's examine the matrix

$$\mathbf{g}(x_1, x_2)\partial_{\mathbf{x}}{}^T = \begin{bmatrix} g_1(x_1, x_2) \\ g_2(x_1, x_2) \end{bmatrix} \begin{bmatrix} \frac{\partial}{\partial x_1} & \frac{\partial}{\partial x_2} \end{bmatrix} = \begin{bmatrix} \frac{\partial g_1(x_1,x_2)}{\partial x_1} & \frac{\partial g_1(x_1,x_2)}{\partial x_2} \\ \frac{\partial g_2(x_1,x_2)}{\partial x_1} & \frac{\partial g_2(x_1,x_2)}{\partial x_2} \end{bmatrix}. \tag{1.9}$$

The determinant of this matrix is called the **Jacobian** of $\mathbf{g}$. The joint PDF can now be found using

$$f_{\underline{y}_1,\underline{y}_2}(y_1, y_2) = \frac{f_{\underline{x}_1,\underline{x}_2}(x_1, x_2)}{\left| \det\left(\mathbf{g}(x_1, x_2)\partial_{\mathbf{x}}{}^T \right) \right|}. \tag{1.10}$$

The above equation also works for dimension higher than 2. In fact, it works for any pair of n-dimensional vectors $\mathbf{y}$ and $\mathbf{x}$ for which $\mathbf{y} = \mathbf{g}(\mathbf{x})$.

1.5 The multi-dimensional mean

Let's suppose we have an n-dimensional random vector $\underline{\mathbf{x}}$, an m-dimensional random vector $\underline{\mathbf{y}}$ and a function $G(\mathbf{x})$ such that $\underline{\mathbf{y}} = G(\underline{\mathbf{x}})$. It would be interesting to know the **expectation vector** $E(\underline{\mathbf{y}})$. It can be found using

$$E(\underline{\mathbf{y}}) = \int_{\mathbb{R}^m} \mathbf{y} f_{\underline{\mathbf{y}}}(\mathbf{y}) \, dy \qquad \Leftrightarrow \qquad E(\underline{y}_i) = \int_{\mathbb{R}^m} y_i f_{\underline{\mathbf{y}}}(\mathbf{y}) \, dy. \tag{1.11}$$

Using the right part of the above equation, you can find one component of $E(\underline{\mathbf{y}})$. The left part is the general (vector) equation. Note that in both cases you need to integrate m times. Once for every component of $\underline{\mathbf{y}}$.

Generally, we don't know $f_{\underline{\mathbf{y}}}(\mathbf{y})$ though. But we do know $f_{\underline{\mathbf{x}}}(\mathbf{x})$. So to find $E(\underline{\mathbf{y}})$, we can first find $f_{\underline{\mathbf{y}}}(\mathbf{y})$. This is, however, not always necessary. There is a way to find $E(\underline{\mathbf{y}})$ without finding $f_{\underline{\mathbf{y}}}(\mathbf{y})$. You then have to use

$$E(\underline{\mathbf{y}}) = E(G(\underline{\mathbf{x}})) = \int_{\mathbb{R}^n} G(\mathbf{x}) f_{\underline{\mathbf{x}}}(\mathbf{x}) \, dx. \tag{1.12}$$

The above equation is called the **expectation law**. If the function $G(\mathbf{x})$ is linear (so you can write it as $G(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$ for a constant matrix A), then the above equation simplifies greatly. In that case we have $\overline{\mathbf{y}} = A\overline{\mathbf{x}} + \mathbf{b}$, where $\overline{\mathbf{y}} = E(\underline{\mathbf{y}})$ and $\overline{\mathbf{x}} = E(\underline{\mathbf{x}})$.

1.6 Multi-dimensional variance and covariance

Calculating the variance $D(\underline{\mathbf{y}})$ of $\underline{\mathbf{y}}$ goes more or less similar to calculating the mean. There is a slight difference though. While $E(\underline{\mathbf{y}})$ was an $m \times 1$ vector, $D(\underline{\mathbf{y}})$ is an $m \times m$ matrix. To find this matrix, we can use either of the following two equations

$$D(\underline{\mathbf{y}}) = E\left((\mathbf{y} - \overline{\mathbf{y}})(\mathbf{y} - \overline{\mathbf{y}})^T\right) = \int_{\mathbb{R}^m} (\mathbf{y} - \overline{\mathbf{y}})(\mathbf{y} - \overline{\mathbf{y}})^T f_{\underline{\mathbf{y}}}(\mathbf{y})\, dy, \tag{1.13}$$

$$D(\underline{\mathbf{y}}) = D(G(\underline{\mathbf{x}})) = E\left((G(\mathbf{x}) - \overline{\mathbf{y}})(G(\mathbf{x}) - \overline{\mathbf{y}})^T\right) = \int_{\mathbb{R}^n} (G(\mathbf{x}) - \overline{\mathbf{y}})(G(\mathbf{x}) - \overline{\mathbf{y}})^T f_{\underline{\mathbf{x}}}(x)\, dx. \tag{1.14}$$

If $G(\mathbf{x})$ is, once more, linear, we can simplify the above equation. In this case we have

$$D(\underline{\mathbf{y}}) = AD(\underline{\mathbf{x}})A^T \qquad \Leftrightarrow \qquad Q_{yy} = AQ_{xx}A^T, \tag{1.15}$$

where $Q_{yy} = D(\underline{\mathbf{y}})$ and $Q_{xx} = D(\underline{\mathbf{x}})$. From these two matrices, we can also find the **covariance matrices** Q_{yx} and Q_{xy}, according to

$$C(\underline{\mathbf{y}}, \underline{\mathbf{x}}) = Q_{yx} = AQ_{xx} \qquad \text{and} \qquad C(\underline{\mathbf{x}}, \underline{\mathbf{y}}) = Q_{xy} = Q_{xx}A^T. \tag{1.16}$$

Here Q_{yx} is an $m \times n$ matrix, while Q_{xy} is an $n \times m$ matrix. So in the multi-dimensional situation we do not have $C(\underline{\mathbf{y}}, \underline{\mathbf{x}}) = C(\underline{\mathbf{x}}, \underline{\mathbf{y}})$. However, since Q_{xx} is symmetric, we do have $Q_{yx} = Q_{xy}^T$.

2 The Central Limit Theorem

If we put together multiple random variables, interesting things start happening. And it has something to do with the normal distribution. If you want to know more about it, then quickly read the chapter below.

2.1 The central limit theorem

Let's suppose we have a number of (possibly different) independent random variables $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$. Now let's define a new random variable $\underline{y}$ as the sum of all these variables, so $\underline{y} = \underline{x}_1 + \underline{x}_2 + \ldots + \underline{x}_n$. Let's suppose we know the mean $\overline{y}$ and the standard deviation and σ_y. The **central limit theorem** states that as n increases, we have

$$F_{\underline{y}}(y) \approx \Phi\left(\frac{y - \overline{y}}{\sigma_y}\right). \tag{2.1}$$

In words, we see that as n increases, $\underline{y}$ behaves like a normal distribution with average $\overline{y}$ and standard deviation σ_y. The corresponding PDF then is

$$f_{\underline{y}}(y) \approx \frac{1}{\sigma_y\sqrt{2\pi}} e^{-\frac{(y-\overline{y})^2}{2\sigma_y^2}}. \tag{2.2}$$

Let's now look at a special case. Suppose $\underline{x}_1 = \underline{x}_2 = \ldots = \underline{x}_n = \underline{x}$. Also, all these distributions have mean $\overline{x}$ and standard deviation σ_x. In this case we can find $\overline{y}$ and σ_y. We have $\underline{y} = n\underline{x}$. The average of $\underline{y}$ evidently becomes $\overline{y} = n\overline{x}$. To find the standard deviation of $\underline{y}$, we first look at the variance of $\underline{y}$. Since the random variables $\underline{x}$ are independent, we find that $\sigma_y^2 = n\sigma_x^2$. From this follows that $\sigma_y = \sqrt{n}\sigma_x$. The random variable $\underline{y}$ thus behaves like a normal distribution with the just found mean $\overline{y}$ and standard deviation σ_y.

2.2 The De Moivre-Laplace theorem

Let's suppose the random variable $\underline{x}$ is binomially distributed. So it is a discrete variable with as mean $\overline{x} = np$ and as variance $\sigma_x^2 = np(1-p)$. The **De Moivre-Laplace theorem** now states that for certain conditions the (discrete) binomial distribution of $\underline{x}$ also starts behaving like a (continuous) normal distribution. So,

$$P_{\underline{x}}(k) = \binom{n}{k} p^k (1-p)^{n-k} \approx \frac{1}{\sigma_x\sqrt{2\pi}} e^{-\frac{(k-\overline{x})^2}{2\sigma_x^2}} = \frac{1}{\sqrt{2\pi np(1-p)}} e^{-\frac{(k-np)^2}{2np(1-p)}}. \tag{2.3}$$

The condition for which the above equation is accurate is that k must be in the interval $(\overline{x}-3\sigma_x, \overline{x}+3\sigma_x)$. Of course k can't be smaller than 0 or bigger than n.

Suppose we do n experiments, with $\underline{x}$ denoting the amount of successes. We would like to know the chance that we have exactly k successes. We now know how to calculate that (approximately). We simply insert $\frac{k-\overline{x}}{\sigma_x}$ in the PDF of the standard normal distribution. But what should we do if we want to know the chance that we have at least

$$P(k_1 \leq \underline{x} \leq k_2) = \Phi\left(\frac{k_2 + 1/2 - \overline{x}}{\sigma_x}\right) - \Phi\left(\frac{k_1 - 1/2 - \overline{x}}{\sigma_x}\right). \tag{2.4}$$

Note the halves in the above equation. They are present because the binomial distribution is discrete, while the normal distribution is continuous. If we, for example, want to have at least 42 successes, and at most 54, then for the normal distribution we should take as boundaries 41.5 and 54.5.

3 Composed Distributions

There are some distributions we haven't treated yet. That was because they were a bit too difficult to start with right away. Often this was because they are composed of multiple other distributions. But now the time has come to take a look at them.

3.1 The multivariate normal distribution

Suppose we have an n-dimensional random vector $\underline{\mathbf{x}}$ with mean $\overline{\mathbf{x}}$ and variance matrix Q_{xx}. We say that $\underline{\mathbf{x}}$ has a **multivariate normal distribution** ($\underline{\mathbf{x}} \sim N_n(\overline{\mathbf{x}}, Q_{xx})$) if its PDF has the form

$$f_{\underline{\mathbf{x}}}(\mathbf{x}) = \frac{1}{\sqrt{\det(2\pi Q_{xx})}} e^{\left(-\frac{1}{2}(\mathbf{x}-\overline{\mathbf{x}})^T Q_{xx}^{-1}(\mathbf{x}-\overline{\mathbf{x}})\right)}, \tag{3.1}$$

where the variance matrix Q_{xx} has only positive entries.

Let's suppose $\underline{\mathbf{x}}$ is 2-dimensional. If we plot $f_{\underline{\mathbf{x}}}(\mathbf{x})$, we get a 3-dimensional graph. For this graph, we can draw **contour lines** (lines for which $f_{\underline{\mathbf{x}}}(\mathbf{x})$ is constant). This implies that

$$(\mathbf{x}-\overline{\mathbf{x}})^T Q_{xx}^{-1} (\mathbf{x}-\overline{\mathbf{x}}) = r^2, \tag{3.2}$$

for some constant r. The shapes we then get are ellipses. We can do the same if $\underline{\mathbf{x}}$ is 3-dimensional. However, we then draw contour areas, which take the shape of ellipsoids. In situations with even more dimensions, we get hyper-ellipsoids. All these shapes are called the **ellipsoids of concentration**.

The shape of these ellipsoids depends on the variance matrix Q_{xx}. If Q_{xx} is the identity matrix I_n, or a multiple of it, then the ellipsoids will be circles/spheres/hyperspheres. If Q_{xx} is just a diagonal matrix, then the principal axes of the ellipsoids will be the axes $x_1, x_2, \ldots, x_n$ itself. In other cases, the axes of the ellipsoid will have shifted.

Many things can be derived from the PDF, for which we just gave the equation. Examples are the marginal distributions and the conditional distributions. An interesting thing is that those distributions are, in turn, also normal distributions. And if that wasn't interesting enough, also all linear transformations of a multivariate normal distribution are (multivariate) normal distributions.

3.2 The χ^2 distribution

Let's suppose $\underline{x}_1, \underline{x}_2, \ldots \underline{x}_n$ are all normally distributed random variables with mean $\overline{x}_i$ and variance 1, so $\underline{x}_i \sim N(\overline{x}_i, 1)$. The **Chi-square** distribution with n degrees of freedom, denoted as $\chi^2(n, \lambda)$, is now defined as

$$\underline{\chi}^2 = \sum_{i=1}^{n} \underline{x}_i^2. \tag{3.3}$$

The **non-centrality parameter** λ is defined as

$$\lambda = \sum_{i=1}^{n} \overline{x}_i^2. \tag{3.4}$$

If $\lambda = 0$, we are dealing with the **central Chi-square distribution** $\chi^2(n, 0)$.

The Chi-square distribution has mean $E(\underline{\chi}^2) = n + \lambda$ and variance $D(\underline{\chi}^2) = 2n + 4\lambda$. If two (independent) Chi-square distributions $\underline{\chi}_1^2$ and $\underline{\chi}_2^2$ are added up, we once more get a Chi-square distribution, but now with $(n_1 + n_2)$ degrees of freedom and non-centrality parameter $(\lambda_1 + \lambda_2)$.

3.3 The t distribution

Suppose that $\underline{x} \sim N(\nabla, 1)$ and $\underline{\chi}^2 \sim \chi^2(n, 0)$ are independent random variables. The **(Student's)** t **distribution** with n degrees of freedom, denoted as $t(n, \nabla)$, is now defined as

$$\underline{t} = \frac{\underline{x}}{\sqrt{\underline{\chi}^2 / n}}. \tag{3.5}$$

Here ∇ is the non-centrality parameter. If $\nabla = 0$, we are dealing with the **central** t **distribution**.

3.4 The F distribution

Suppose that $\underline{\chi}_1^2 \sim \chi^2(n_1, \lambda)$ and $\underline{\chi}_2^2 \sim \chi^2(n_2, 0)$ are two independent Chi-square distributions. The F **distribution**, denoted as $F(n_1, n_2, \lambda)$, is then defined as

$$\underline{F} = \frac{\underline{\chi}_1^2 / n_1}{\underline{\chi}_2^2 / n_2}. \tag{3.6}$$

It is said to have n_1 and n_2 degrees of freedom. Also, λ is the non-centrality parameter. When $\lambda = 0$, we are dealing with a **central** F **distribution**.