1 The Gaussian Distribution

The general form of the Gaussian distribution is:

$$p_x(x) = Ne^{-\frac{1}{2}Ax^2 - Bx} \qquad (-\infty < x < \infty), \tag{1}$$

where A is a positive constant, B is any real number and N is a normalization constant given by

$$N = \left(\frac{A}{2\pi}\right)^{1/2} e^{-\frac{B^2}{2A}}. \tag{2}$$

In order to better understand the meaning of the constants it is useful to write the distribution using its mean (μ) and variance (σ^2)

$$p_x(x) = \left(2\pi\sigma^2\right)^{-1/2} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (-\infty < x < \infty). \tag{3}$$

****

The generating function (or characteristic function) is defined by:

$$\begin{aligned} G_x(k) &\equiv \int_{-\infty}^{\infty} e^{-ikx} p_x(x)\, dx = \int_{-\infty}^{\infty} \sum_{m=0}^{\infty} \frac{(-ik)^m}{m!} x^m p_x(x)\, dx \\ &= \sum_{m=0}^{\infty} \frac{(-ik)^m}{m!} \langle x^m \rangle, \end{aligned} \tag{4}$$

and the moments are given by:

$$\langle x^m \rangle = \frac{1}{(-i)^m} \frac{\partial^m G_x(k)}{\partial k^m}|_{k=0}. \tag{5}$$

****

For the Gaussian distribution of equation (1), the generating function is:

$$G_x(k) \equiv \int_{-\infty}^{\infty} e^{-ikx} \left(2\pi\sigma^2\right)^{-1/2} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = e^{-ik\mu - k^2\sigma^2/2},$$

and the first two moments are given by:

$$\langle x \rangle = \frac{1}{(-i)} \frac{\partial G_x(k)}{\partial k}|_{k=0} = \mu, \tag{6}$$

$$\langle x^2 \rangle = \frac{1}{(-i)^2} \frac{\partial^2 G_x(k)}{\partial k^2}|_{k=0} = \mu^2 + \sigma^2. \tag{7}$$

The cumulants generating function is related to the moments generating function and is defined by:

$$C_x(k) \equiv \ln(G_x(k)) = \sum_{m=0}^{\infty} \frac{(-ik)^m}{m!}\kappa_m. \tag{8}$$

The cumulants are related to the moments, for example

$$\kappa_1 = \langle x \rangle ; \kappa_2 = \left\langle x^2 \right\rangle - \langle x \rangle^2 ; \kappa_3 = \left\langle x^3 \right\rangle - 3 \langle x \rangle \left\langle x^2 \right\rangle + 2 \langle x \rangle^3 . \tag{9}$$

Note that the second cumulant is equal to the second central moment, $\left\langle (x - \langle x \rangle)^2 \right\rangle$, and $\kappa_3 = \left\langle (x - \langle x \rangle)^3 \right\rangle$ is the third central moment but higher cumulants are not equal to higher central moments. Higher cumulants may be calculated from the definition in eq.(8).

One of the important characteristics of the Gaussian distribution is easily seen from its cumulants. The cumulants generating function for the Gaussian distribution is

$$C_x(k) = \ln(G_x(k)) = -ik\mu - k^2\sigma^2/2. \tag{10}$$

Therefore,

$$\kappa_1 = \mu; \kappa_2 = \sigma^2; \tag{11}$$

and all higher cumulants vanish.

The exponential distribution is defined as:

$$f(x) = \lambda e^{-\lambda x} \qquad x \geq 0.$$

The corresponding characteristic function is,

$$\widetilde{f}(k) = \int_0^{\infty} e^{-ikx} \lambda e^{-\lambda x} dx = \frac{\lambda}{\lambda + ik};$$

First we notice that

$$\widetilde{f}(k=0) = 1,$$

implying that the probability density function (PDF) is normalized. Second, we can calculate the moments

$$\begin{aligned} \langle x \rangle &= \frac{1}{(-i)} \frac{d}{dk} \left(\frac{\lambda}{\lambda + ik} \right) |_{k=0} = -\frac{\lambda}{(-i)} \frac{i}{(\lambda + ik)^2} |_{k=0} = \frac{1}{\lambda}; \\ \left\langle x^2 \right\rangle &= \frac{1}{(-i)^2} \frac{d^2}{dk^2} \left(\frac{\lambda}{\lambda + ik} \right) |_{k=0} = \frac{-i\lambda}{(-i)^2} \frac{-2i}{(\lambda + ik)^3} |_{k=0} = \frac{2}{\lambda^2}. \end{aligned}$$

The cumulant generating function is given by,

$$C_f(k) = \ln \widetilde{f}(k) = \ln(\frac{\lambda}{\lambda + ik}) = \ln(\lambda) - \ln(\lambda + ik).$$

The two first cumulants are given by,

$$\begin{aligned}
\kappa_1 &= \frac{1}{(-i)}\frac{d}{dk}\left(\ln(\lambda)-\ln(\lambda+ik)\right)|_{k=0} = \frac{1}{(-i)}\frac{d}{dk}\left(-\ln(\lambda+ik)\right)|_{k=0} = \frac{-1}{(-i)}\frac{i}{\lambda+ik}|_{k=0} = \frac{1}{\lambda};\\
\kappa_2 &= \frac{1}{(-i)^2}\frac{d^2}{dk^2}\left(\ln(\lambda)-\ln(\lambda+ik)\right)|_{k=0} = \frac{1}{(-i)}\frac{d}{dk}\left(\frac{1}{\lambda+ik}\right)|_{k=0} = \frac{1}{(\lambda+ik)^2}|_{k=0} = \frac{1}{\lambda^2} = \left\langle x^2\right\rangle - \left\langle x\right\rangle^2;
\end{aligned}$$

2 The PDF of the sum of two Gaussian random variables

We define the variable Y as the sum of two random variables, each of which has a Gaussian PDF.

$$Y = x_1 + x_2.$$

The PDF of x_i is:

$$p_x\left(x_i\right) = \frac{1}{\sqrt{2\pi s^2}}e^{-\frac{(x_i-\mu)^2}{2s^2}}.$$

The PDF of Y is given by:

$$\begin{aligned}
p_Y\left(Y\right) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\delta\left(Y-x_1-x_2\right)p_x\left(x_1\right)p_x\left(x_2\right)dx_1dx_2\\
&= \int_{-\infty}^{\infty}p_x\left(x_1\right)p_x\left(Y-x_1\right)dx_1\\
&= \frac{1}{s\sqrt{2\pi}}\int_{-\infty}^{\infty}\frac{1}{s\sqrt{2\pi}}e^{-\frac{(x_1-\mu)^2}{2s^2}}e^{-\frac{(Y-x_1-\mu)^2}{2s^2}}dx_1\\
&= \frac{1}{\sqrt{2}s\sqrt{2\pi}}e^{-\frac{(Y-2\mu)^2}{2(\sqrt{2}s)^2}} = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(Y-m)^2}{2\sigma^2}}\\
\sigma^2 &= 2s^2; m = 2\mu;
\end{aligned}$$

We showed that the PDF of Y is again Gaussian. $\left\langle Y\right\rangle = 2\mu$ and $\left\langle\left(Y-\left\langle Y\right\rangle\right)^2\right\rangle = 2s^2$.

3 The PDF of the sum of r Gaussian random variables

When $x_1, x_2, ...x_r$ are Gaussian random variables and mutually independent, their sum $Y = \sum_{i=1}^{r} x_i$ is again a Gaussian random variable. The average is

simply the sum of the averages and the variance is the sum of the variances. To see that, consider the PDF of Y

$$\begin{aligned} p_Y(Y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}...\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\delta\left(\sum_{i=1}^{r}x_i - Y\right)\prod_{i=1}^{r}p_x(x_i)\,dx_i \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}...\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}p_x\left(Y-\sum_{i=1}^{r-1}x_i\right)\prod_{i=1}^{r-1}p_x(x_i)\,dx_i. \end{aligned} \tag{12}$$

The Fourier transform of $p_Y(Y)$ is:

$$\begin{aligned} G_Y(q) &= \int_{-\infty}^{\infty}e^{-iqY}p_Y(Y)\,dY \\ &= \int_{-\infty}^{\infty}e^{-iqY}\left(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}...\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\delta\left(\sum_{i=1}^{r}x_i - Y\right)\prod_{i=1}^{r}p_x(x_i)\,dx_i\right)dY \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}...\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}e^{-iq\sum_{i=1}^{r}x_i}\prod_{i=1}^{r}p_x(x_i)\,dx_i \\ &= \left(\int_{-\infty}^{\infty}e^{-iqx}p_x(x)\,dx\right)^r = (G_x(q))^r = \left(e^{-ik\mu-k^2\sigma^2/2}\right)^r = e^{-ikr\mu-k^2r\sigma^2/2} \end{aligned} \tag{13}$$

This implies that the PDF of Y is given by

$$p_Y(Y) = \left(2\pi r\sigma^2\right)^{-1/2}e^{-\frac{(Y-r\mu)^2}{2r\sigma^2}}.$$

For the specific case of $\mu = 0$,

$$G_Y(q) = e^{-k^2r\sigma^2/2} \rightarrow p_Y(Y) = \left(2\pi r\sigma^2\right)^{-1/2}e^{-\frac{Y^2}{2r\sigma^2}}. \tag{14}$$

It is useful to define the scaled sum

$$Z = \frac{1}{\sqrt{r}}\sum_{i=1}^{r}x_i. \tag{15}$$

Using the usual transformation rule

$$|p_Y(Y)\,dY| = |p_Z(Z)\,dZ| \tag{16}$$

we find

$$p_Z(Z) = \left(2\pi\sigma^2\right)^{-1/2}e^{-\frac{Z^2}{2\sigma^2}}. \tag{17}$$

4 The Central Limit Theorem

The central limit theorem states that even when $p_x(x)$ is not Gaussian, but some other distribution with zero mean and finite variance σ^2, equation (17) is still valid at the limit of $r \to \infty$. This fact is responsible for the dominant role of the Gaussian distribution in all fields of statistics.

To see that, we just need to write the characteristic function of an arbitrary $p_x(x)$ with zero mean and a finite variance σ^2,

$$G_x(q) = \int_{-\infty}^{\infty} e^{-iqx} p_x(x)\, dx = 1 - \frac{1}{2}\sigma^2 q^2 + ... \tag{18}$$

Hence, one finds for the characteristic function of Z

$$G_Z(q) = \left(G_x\left(q/\sqrt{r}\right)\right)^r = \left(1 - \frac{1}{2r}\sigma^2 q^2 + O\left(\frac{1}{r^{3/2}}\right)\right)^r \to e^{-q^2\sigma^2/2 + O\left(r^{-1/2}\right)}. \tag{19}$$

Therefore, in the limit $r \to \infty$ the distribution of Z is Gaussian.

Example:

Let x be a stochastic variable that takes the values 0 or 1 with probability 1/2 each.

Let Y be the sum of r such variables. Then Y takes the values $0, 1, 2...$ with probability

$$p_n = 2^{-r} \left(\begin{array}{c} r \\ n \end{array} \right) = 2^{-r} \frac{r!}{n!\,(r-n)!}. \tag{20}$$

To study the limiting behavior, we need first to define the proper new variable.

The average of Y is:

$$\langle Y \rangle = r \langle x \rangle = r/2.$$

The variance is:

$$\sigma_Y^2 = r\sigma_x^2 = \frac{1}{4} r.$$

Therefore, we define

$$\begin{aligned} Z &= \frac{Y - r/2}{\left(\frac{r}{4}\right)^{1/2}}; \\ Y &= \frac{r}{2} + \left(\frac{r}{4}\right)^{1/2} Z. \end{aligned}$$

Note that the variable Z has a zero mean and a unit variance.

The probability that Z lies between z and $z + \Delta z$ is

$$p_Z(Z)\,\Delta z = \sum_{\frac{r}{2} + \left(\frac{r}{4}\right)^{1/2} z < n < \frac{r}{2} + \left(\frac{r}{4}\right)^{1/2} (z + \Delta z)} p_n. \tag{21}$$

When r is large, the probability density is smooth. For large r, we expand p_n using the Stirling's formula

$$\lim_{m\to\infty} \ln(m!) \sim \frac{1}{2}\ln(2\pi) + \left(m+\frac{1}{2}\right)\ln m - m$$

$$\begin{aligned}
\ln p_n &\sim -r\ln 2 \\
&\quad +\frac{1}{2}\ln(2\pi) + \left(r+\frac{1}{2}\right)\ln r - r \\
&\quad -\frac{1}{2}\ln(2\pi) - \left(n+\frac{1}{2}\right)\ln n + n \\
&\quad -\frac{1}{2}\ln(2\pi) - \left(r-n+\frac{1}{2}\right)\ln(r-n) + r - n \\
&= -r\ln 2 - \frac{1}{2}\ln(2\pi) + \left(r+\frac{1}{2}\right)\ln r \\
&\quad -\left(n+\frac{1}{2}\right)\ln n - \left(r-n+\frac{1}{2}\right)\ln(r-n).
\end{aligned}$$

Expressing n in terms of Z

$$\begin{aligned}
\ln p_n &\sim \ln 2 - \frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln r \\
&\quad -\left(\frac{r}{2}+\frac{\sqrt{r}Z}{2}+\frac{1}{2}\right)\ln\left(1+\frac{Z}{\sqrt{r}}\right) \\
&\quad -\left(\frac{r}{2}-\frac{\sqrt{r}Z}{2}+\frac{1}{2}\right)\ln\left(1-\frac{Z}{\sqrt{r}}\right).
\end{aligned}$$

Expanding the logarithms

$$\begin{aligned}
\ln p_n &\sim \ln 2 - \frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln r \\
&\quad -\left(\frac{r}{2}+\frac{\sqrt{r}Z}{2}+\frac{1}{2}\right)\frac{Z}{\sqrt{r}} \\
&\quad +\frac{1}{2}\left(\frac{r}{2}+\frac{\sqrt{r}Z}{2}+\frac{1}{2}\right)\frac{Z^2}{r} \\
&\quad +\left(\frac{r}{2}-\frac{\sqrt{r}Z}{2}+\frac{1}{2}\right)\frac{Z}{\sqrt{r}} \\
&\quad +\frac{1}{2}\left(\frac{r}{2}-\frac{\sqrt{r}Z}{2}+\frac{1}{2}\right)\frac{Z^2}{r} \\
&= \ln 2 - \frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln r - Z^2/2 + O\left(1/\sqrt{r}\right).
\end{aligned}$$

Substituting that into equation (21) one obtains,

$$p_Z(Z)\,\Delta z \sim \frac{1}{2}\sqrt{r}\Delta z e^{\left(\ln 2 - \frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln r - Z^2/2\right)} = \frac{\Delta z}{\sqrt{2\pi}} e^{-Z^2/2}.$$

Therefore, the probability distribution of Y may be written as:

$$p_Y(Y) = \frac{1}{\sqrt{\frac{1}{2}\pi r}} e^{-\frac{\left(Y-\frac{r}{2}\right)^2}{\frac{1}{2}r}}.$$

Note that the same result could be obtained using the CLT (using the average ($\langle x \rangle = 1/2$) and the variance $\left(\left\langle (x - \langle x \rangle)^2 \right\rangle = 1/4\right)$ that we derived earlier).

Berry-Eseen Theorem

Consider the normalized sum of N random variables, each characterized by a PDF (probability density function) with a vanishing first moment and a finite variance σ^2.

$$\xi = \frac{1}{\sqrt{N}\sigma} \sum_{i=1}^{N} x_i \tag{A1}$$

The PDF of ξ is denoted as $p_N(\xi)$ and the cumulative distribution of ξ is given by $P_N(\xi) = \int_{-\infty}^{\xi} p_N\left(\xi'\right) d\xi'$. We assume, in addition to the previously mentioned conditions, that the third absolute moment $r_3 = \int_{-\infty}^{\infty} \left|x_i^3\right| p(x_i)\, dx_i < \infty$ is finite. The Berry-Eseen theorem states that,

$$\left| P_N(\xi) - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\xi} e^{-\frac{\xi'^2}{2}} d\xi' \right| \leq \frac{C r_3}{N^{1/2}\sigma^3}. \tag{A2}$$

This is a very strong bound since it holds for any N. This relation can be used to quantify the goodness of the Gaussian approximation to the PDF of ξ. In the limit of $N \to \infty$ this relation goes back to the CLT (central limit theorem). Over the years the value of C was reduced and the current value is $0.4748 \geq C \leq 0.40973$. The theorem holds also for the sum of non identical random variables. For random variables with zero mean and different STDs we define

$$S_N = \frac{1}{\sqrt{\sum_{i=1}^{N} \sigma_i^2}} \sum_{i=1}^{N} x_i \tag{A1}$$

Berry showed that

$$\left| P_N\left(z\right) - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{\xi'^2}{2}} d\xi' \right| \leq \frac{C_1}{\sqrt{\sum_{i=1}^{N} \sigma_i^2}} \max\left(\frac{r_{3,i}}{\sigma_i^2}\right). \tag{A2}$$

Eseen showed an even tighter bound

$$\left| P_N\left(z\right) - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{\xi'^2}{2}} d\xi' \right| \leq \frac{C_0}{\left(\sqrt{\sum_{i=1}^{N} \sigma_i^2}\right)^3} \sum_{i=1}^{N} r_{3,i}. \tag{A2}$$

when the variables ar identical both limits converge to the limit we mentioned earlier. The latest upper bound for C_0 is 0.56 and the lower bound is the same as for identical variables.