Physics 115/242
Derivation of the central limit theorem.

Peter Young

(Dated: November 29, 2007)

Consider a random variable x with distribution $P(x)$. It has mean μ and standard deviation σ. According to the **central limit theorem**, if μ and σ are finite, the distribution of the sum of N independent such variables,

$$X = \sum_{i=1}^{N} x_i \,, \tag{1}$$

is, for $N \to \infty$, a **Gaussian with mean $N\mu$ and standard deviation $\sqrt{N}\sigma$**, i.e.

$$\lim_{N \to \infty} P_N(X) = \frac{1}{\sqrt{2\pi N}\sigma} \, \exp\left(-\frac{(X - N\mu)^2}{2N\sigma^2}\right) \,. \tag{2}$$

The purpose of this handout is to **derive this result**.

The distribution of X is given by integrating over all possible values for the x_i subject to the constraint that $\sum_i x_i = X$, i.e.

$$P_N(X) = \int_{-\infty}^{\infty} P(x_1)\, dx_1 \int_{-\infty}^{\infty} P(x_2)\, dx_2 \cdots \int_{-\infty}^{\infty} P(x_N)\, dx_N \ \delta(x_1 + x_2 + \cdots + x_N - X) \,. \tag{3}$$

Using the integral representation of the delta function,

$$\delta(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{ikx} \, dk, \tag{4}$$

we have

$$P_N(X) = \frac{1}{2\pi} \int_{-\infty}^{\infty} dk \, e^{-ikX} \int_{-\infty}^{\infty} P(x_1) e^{ikx_1} \, dx_1 \int_{-\infty}^{\infty} P(x_2) e^{ikx_2} \, dx_2 \cdots \int_{-\infty}^{\infty} P(x_N) e^{ikx_N} \, dx_N \,, \tag{5}$$

or

$$P_N(X) = \frac{1}{2\pi} \int_{-\infty}^{\infty} dk \, e^{-ikX} \, \widetilde{P}_N(k) \,, \tag{6}$$

where

$$\widetilde{P}_N(k) = \left[\widetilde{P}(k)\right]^N \,, \tag{7}$$

in which

$$\widetilde{P}(k) = \int_{-\infty}^{\infty} P(x)\, e^{ikx} \, dx \tag{8}$$

is the Fourier transform of $P(x)$. Equation (7) shows that $\widetilde{P}_N(k)$, the Fourier transform of the distribution of the sum of N variables $P_N(X)$, is the N-th power of $\widetilde{P}(k)$.

We see that Fourier transform of the distribution is important since the connection between the distribution of the sum and the distribution of a single variable is much simpler when Fourier transformed. We therefore consider the Fourier transform of a distribution in more detail. Expanding the exponential gives

$$\widetilde{P}(k) = \int_{-\infty}^{\infty} e^{ikx} P(x)\, dx = \sum_{n=0}^{\infty} \frac{(ik)^n}{n!} \int_{-\infty}^{\infty} x^n \, P(x)\, dx = \sum_{n=0}^{\infty} \frac{(ik)^n}{n!} \langle x^n \rangle \,, \tag{9}$$

where $\langle \cdots \rangle$ indicates an average over $P(x)$. The coefficient of $(ik)^n$ is therefore the n-th moment of $P(x)$ divided by $n!$. We say that $\widetilde{P}(k)$ is the *generating function* of the moments of the distribution.

In view of Eq. (7) it is useful to re-exponentiate the last expression in Eq. (9), i.e.

$$\widetilde{P}(k) \equiv \int_{-\infty}^{\infty} P(x)\, e^{ikx}\, dx = \sum_{n=0}^{\infty} \frac{(ik)^n}{n!} \langle x^n \rangle = \exp\left[\sum_{n=1}^{\infty} \frac{(ik)^n}{n!} \langle x^n \rangle_c \right] , \tag{10}$$

which defines the **"cumulant"** averages $\langle \cdots \rangle_c$. Expanding out the exponential in the last expression and comparing powers of k one finds that the first few cumulants are

$$\langle x \rangle_c = \langle x \rangle \equiv \mu , \tag{11a}$$

$$\langle x^2 \rangle_c = \langle (x - \langle x \rangle)^2 \rangle = \langle x^2 \rangle - \langle x \rangle^2 \equiv \sigma^2 , \tag{11b}$$

$$\langle x^3 \rangle_c = \langle (x - \langle x \rangle)^3 \rangle , \tag{11c}$$

$$\langle x^4 \rangle_c = \langle (x - \langle x \rangle)^4 \rangle - 3 \langle (x - \langle x \rangle)^2 \rangle^2 . \tag{11d}$$

Note that $\langle x \rangle_c$ is the mean, μ, $\langle x^2 \rangle_c$ is the variance, σ^2, $\langle x^3 \rangle_c$ is proportional to the skewness, and $\langle x^4 \rangle_c$ is related to the kurtosis. Roughly speaking the n-th order cumulant average represents the n-th moment with the effects of lower moments subtracted off.

In particular, if we add a constant C to x, thereby adding C to μ, all the higher order cumulants are unchanged. This can be seen explicitly for $n = 2, 3$ and 4 in Eq. (11). To prove this in general consider $P'(x) = P(x - C)$. $P'(x)$ and $P(x)$ have the same shape but one is shifted relative to the other such that if the mean of $P(x)$ is μ then the mean of $P'(x)$ is $\mu + C$. Fourier transforming $P'(x)$ gives

$$\widetilde{P'}(k) = \int_{-\infty}^{\infty} P'(x)\, e^{ikx}\, dx = \int_{-\infty}^{\infty} P(x - C)\, e^{ikx}\, dx = \int_{-\infty}^{\infty} P(x)\, e^{ikx} e^{ikC}\, dx = \exp\left[(ik)(\mu + C) + \sum_{n=2}^{\infty} \frac{(ik)^n}{n!} \langle x^n \rangle_c \right] . \tag{12}$$

where we used Eq. (10) in the last equality. We see that all cumulants except the first are unchanged.

It is particularly interesting to determine the cumulant averages of the the Gaussian distribution

$$P^G(x) = \frac{1}{\sqrt{2\pi}\, \sigma} \exp\left[(x - \mu)^2 / 2\sigma^2 \right] . \tag{13}$$

We have already discussed that the Fourier transform of a Gaussian is a Gaussian and that this result is obtained by "completing the square". We therefore simply quote the result for the Fourier transform

$$\widetilde{P}^G(k) = \frac{\sigma}{\sqrt{2\pi}} \exp\left[ik\mu - \sigma^2 k^2 / 2 \right] . \tag{14}$$

Comparing with Eq. (10) we see that $\langle x \rangle_c = \mu$ and $\langle x^2 \rangle_c = \sigma^2$ (which we have already seen are general results) and **all cumulants higher than the second vanish**. This last result. which is special to the Gaussian, is the key to the derivation of the central limit theorem.

We now return to the derivation of this theorem. From Eqs. (7) and (10) we see that each of the cumulants of $P_N(X)$ is N times as big as the corresponding cumulant of $P(x)$, so for example

$$\langle X \rangle_{N,c} = N \langle x \rangle_c = N\mu , \tag{15a}$$

$$\langle X^2 \rangle_{N,c} = N \langle x^2 \rangle_c = N\sigma^2 , \tag{15b}$$

$$\langle X^3 \rangle_{N,c} = N \langle x^3 \rangle_c . \tag{15c}$$

If we define a variable $Y = (X - N\mu)/\sqrt{N}$ then we see from Eqs. (15a) and (15b) that Y has zero mean and standard deviation unity. What about the higher cumulants like $\langle Y^3 \rangle_c$? We showed above in Eq. (12) that subtracting a constant (such as the mean) doesn't affect any of the cumulants beyond the first. Hence

$$\langle Y^3 \rangle_{N,c} = N^{-3/2} \langle X^3 \rangle_{N,c} = N^{-1/2} \langle x^3 \rangle_c , \tag{16}$$

where the last expression used Eq. (15c). Hence $\langle Y^3 \rangle_{N,c}$ **tends to zero for $N \to \infty$** provided the first three moments of x exist. Similarly $\langle Y^n \rangle_{N,c} \propto N^{-(n-2)/2}$, and hence vanishes for $N \to \infty$ for all n greater than 2.

Since all cumulants of Y beyond the second vanish for $N \to \infty$, the distribution of Y must be a Gaussian in this limit. Now Y has zero mean and variance 1, so $X = \sqrt{N}(Y + N\mu)$ must also be Gaussian but with mean $N\mu$ and variance $N\sigma^2$, i.e.

$$P_N(X) = \frac{1}{\sqrt{2\pi N}\,\sigma} \exp\left[\frac{(X - N\mu)^2}{2N\sigma^2}\right]. \tag{17}$$

for $N \to \infty$. This is the famous **central limit theorem**. We emphasize that the distribution of the sum becomes Gaussian for $N \to \infty$ even though the distribution of the individual variables x_i is *not* in general Gaussian.

Our derivation assumed that all moments of x exist. A more careful treatment shows that the central limit theorem holds provided only that the first two moments exist (i.e. μ and σ^2 exist). However, *convergence* to a Gaussian distribution as N increases is faster if the higher moments also exist.

In a separate handout I show numerically and example of the convergence to the Gaussian distribution.

An important aspect of the central limit theorem is that the standard deviation (a measure of the width of the distribution) only grows like $N^{1/2}$, whereas the mean increases like N. Hence, **at large N the distribution becomes sharp**; a large deviation from the mean becomes increasingly unlikely.

As an example, consider coin tossing. If we denote the result "heads" by $x = 1$ and "tails" by $x = 0$ the mean for a single toss is $\mu \equiv \langle x \rangle = 1/2$, and the standard deviation is $\sqrt{1/2 - (1/2)^2} = 1/2$. For N tosses, the central limit theorem tells us that the average number of heads is $N/2$ and the standard deviation of this result (obtained if one repeated the N tosses very many times) is $\sqrt{N}/2$. Hence the *fraction* of heads obtained in N tosses is

$$\frac{1}{2} \pm \frac{1}{2\sqrt{N}}, \tag{18}$$

where the "error bar" indicates one standard deviation.

The fact that the fraction of heads should be closer and closer to $1/2$ for larger N is consistent with our intuition. For example, for 16 tosses, we realize that we could very well get 10 heads. Equation (18) agrees with this since $1/(2\sqrt{N}) = 0.125$, so expect a fraction 0.5 ± 0.125 of heads. The actual result of $10/16 = 0.625$ is therefore just 1σ away from the mean, which is quite probable.

However, if we have 10^4 times as many tosses, $N = 1.6 \times 10^5$, then our intuition tells us that the same fraction of heads, 0.625, (i.e. 10^5 heads) would be most unlikely. The central limit theorem is in agreement with this because a fraction 0.625 of heads is now 100 standard deviations away from the mean (because the standard deviation is $\sqrt{10^4} = 100$ times smaller). The probability of getting a deviation of a least this big for a Gaussian distribution is $\text{erfc}(100/\sqrt{2}) \sim 10^{-2174}$ (where $\text{erfc}(x) = 1 - \text{erf}(x)$ is the complementary error function).

This is an **incredibly** small number. Such and event would **never** happen even if one had an army of monkeys tossing coins for the age of the universe. To see this, lets imagine how many events (i.e. tossings of 1.6×10^6 coins) would be possible; the age of the inverse is about 10^{10} years, there are about 3×10^7 seconds in a year, and suppose we have 10^9 monkeys each able to toss the required 1.6×10^5 coins in one second (fast!). Multiplying the numbers together gives about 5×10^{26} events. To get the overall probability of the occurence of 10^5 heads, we multiply the probability that this happens in one event, $\sim 10^{-2174}$, by the number of events, $\sim 10^{26}$. However, the power of 10 in the number of events (26) hardly makes any impression on the power in the probability (-2174). Hence, by any reasonable measure, such a rare event will **NEVER** happen.