The Central Limit Theorem
And the Limits of the Central Limit Theorem

Elena Sellentin

Université de Genève
Département de Physique Théorique

9. Mai 2017

Gaussian and Non-Gaussian

Two true statements:

Averaging over non-Gaussian random variables is an effective Gaussianization.

Non-linear functions of Gaussian random variables follow non-Gaussian distributions.

Which is your case?

Towards Gaussianity: The Central Limit Theorem

Descriptive statistics

- Noncentral moments: $m_n = \langle x^n \rangle$
- Cumulants: $\kappa_n = m_n - \sum_{k=1}^{n-1} \binom{n-1}{k-1} \kappa_k m_{n-k}$ (combinatorical superposition of moments)
- Gaussian: has infinitely many even moments, but only the first and second cumulant

Moment-generating function:

$$m_x(t) = \langle e^{tx} \rangle \tag{1}$$

Cumulant-generating function:

$$c_x(t) = \log\left(m_x(t)\right) \tag{2}$$

Factors i can appear when a solution in the complex plane exists, but not on the real axis.

The Central Limit Theorem

Take n iid random variables x_i, meaning $x_i \sim p(x)\ \forall i$.
Now take the mean

$$\bar{x} = \frac{1}{n} \sum_{i}^{n} x_i. \tag{3}$$

$p(x)$ must have a finite mean μ and variance σ^2... ... but is not further specified and can be non-Gaussian.
Then, the sample average $\bar{x}$ for $n \to \infty$ will be

$$\bar{x} \sim \mathcal{G}(\mu, \sigma^2/n), \tag{4}$$

where $\mathcal{G}(\mu, \sigma^2/n)$ is the Gaussian of mean μ and variance σ^2/n.
Equivalently, the whitened variable Y follows

$$Y = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1). \tag{5}$$

The Central Limit Theorem

Proof: we show that the higher order cumulants die out.

We whiten the data points

$$z_i = \frac{x_i - \mu}{\sigma} \;\forall i, \tag{6}$$

such that by definition

$$\langle z_i \rangle = 0, \quad \langle z_i^2 \rangle = 1. \tag{7}$$

In terms of the z_i, the variable Y is then

$$Y = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} z_i, \tag{8}$$

and its moment-generating function is

$$m_Y(t) = \langle e^{tY} \rangle = \left\langle \exp\left(\frac{t z_i}{\sqrt{n}} \right) \right\rangle^n. \tag{9}$$

The Central Limit Theorem

Now expand the average into a power series and use linearity of the average:

$$\left\langle \exp\left(\frac{tz_i}{\sqrt{n}}\right)\right\rangle = \left\langle 1 + \frac{tz_i}{\sqrt{n}} + \frac{t^2 z_i^2}{2n} + \frac{t^3 z_i^3}{3! n^{3/2}} + ... \right\rangle, \tag{10}$$

the second term is zero due to $\langle z_i \rangle = 0$, and in the third term $\langle z_i^2 \rangle = 1$. Plug back into the power-n:

$$\begin{aligned} m_Y(t) &= \left[1 + \frac{t^2}{2n} + \frac{t^3 \langle z_i^3 \rangle}{3! n^{3/2}} + ...\right]^n \\ &= \left[1 + \frac{1}{n}\left(\frac{t^2}{2} + \frac{t^3 \langle z_i^3 \rangle}{3! n^{1/2}} + ...\right)\right]^n . \end{aligned} \tag{11}$$

The Central Limit Theorem

Now introduce a shorthand for the power series

$$u = \frac{t^2}{2} + \frac{t^3 \langle z_i^3 \rangle}{3! n^{1/2}} + ..., \tag{12}$$

such that

$$m_Y(t) = \left[1 + \frac{u}{n} \right]^n. \tag{13}$$

For averaging over ever more samples, we have $n \to \infty$, and in this limit have

$$u \to \frac{t^2}{2}, \tag{14}$$

because the other fractions will be supressed by the powers of n in their denominator. But each of these fractions contains a moment $\langle z_i^n \rangle$ and hence the higher moments die out, if we average over increasingly more samples n.

The Central Limit Theorem

At the same time, for $n \to \infty$, we have

$$\lim_{n\to\infty} \left[1 + \frac{u}{n}\right]^n = e^u. \tag{15}$$

Consequently, in the limit of $n \to \infty$, the moment-generating function of Y is

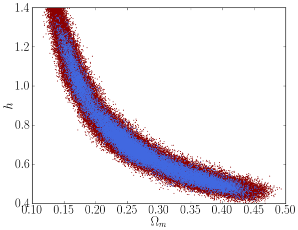
$$m_Y(t) = e^{\frac{t}{2}}, \tag{16}$$

which is the Laplace transform of the standard normal distribution $\mathcal{N}$. Hence, the central limit theorem is proven. Most importantly, we also see the limit in which it arises: if n is large but finite, higher moments of the initial distribution $p(x)$ can survive – if they exist.

Away from Gaussianity: Non-linear functions of Gaussian random variables

Non-Gaussianity

Posteriors of Gaussian data will still remain forever non-Gaussian, if you have perfect parameter degeneracies no matter how many new data you get.



→ Uncertainty on parameters can be driven by the model, instead of by the data. $\Rightarrow$ Discussion about degeneracy breaking in cosmology.

Non-linear functions

Non-linear functions of Gaussian random variables follow non-Gaussian distributions, and the CLT agrees with this.

- the ratio of two Gaussian rv follows a Cauchy distribution
- the exponential of a Gaussian rv is log-normally distributed
- the absolute value of Fourier modes from a Gaussian random field follow a Rayleigh distribution
- many more $\Rightarrow$ Exercises: Are astronomical magnitudes Gaussianly distributed? And what is the distribution of $|x|$ if x is Gaussianly distributed?

→ Homepage: http://theory.physics.unige.ch/sellentin/