Central Limit Theorem

by Robert D. Klauber   July 7, 2015
www.quantumfieldtheory.info

1 Background: Types of Distributions

Normal (bell curve, Gaussian) probability distributions are common. For example, height of human females,

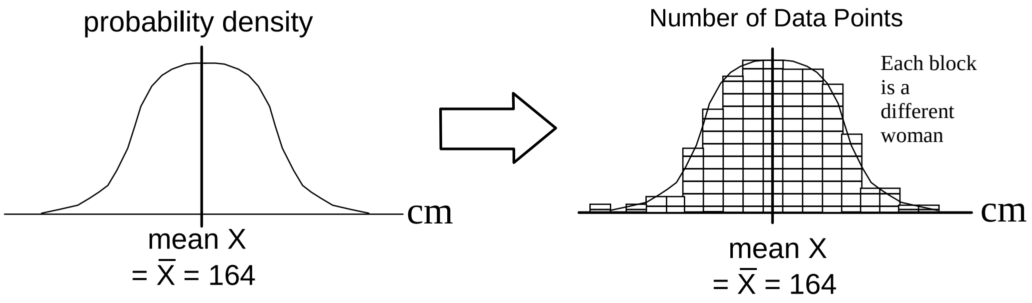

Figure 1. Gaussian Probability Distribution

There are other types of distributions. For example, consider having 10,000 balls in a bin, each ball having a number from 1 to 100, with each number being on 100 balls. The probability density and histogram look like Fig. 2, so this called a uniform, or rectangular, distribution.

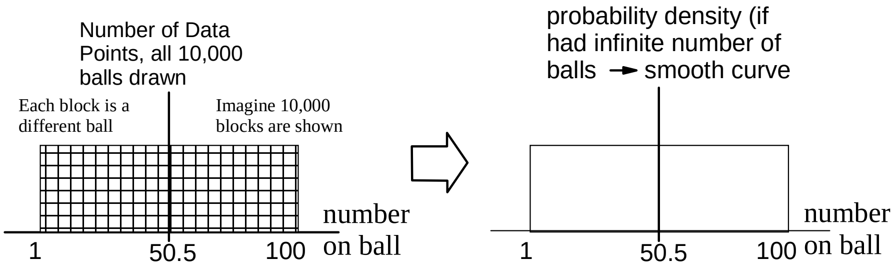

Figure 2. Uniform Probability Distribution

There are other shape distributions as well. For example, the probability density for failure of your new washing machine vs time when it fails peaks in the first few weeks, then lowers and levels out for years, and then gradually rises higher and higher as the years go on (and it gets older).

For any type of distribution we can find the mean (the average) value, denoted herein with an overbar on the variable, and the standard deviation, denoted here and virtually everywhere by σ. Note (1) and (2) below, where we differentiate between measuring every member (numbering N in all) of a population (subscript "all") and sampling some lesser number n ($< N$) of the population. As $n \to N$, then $\bar{X} \to \bar{X}_{all}$ and $\sigma \to \sigma_{all}$.

$$\text{Measuring all } N \text{ members of population} \quad \langle X \rangle_{all} = \bar{X}_{all} = \frac{\sum_{a=1}^{N} X_a}{N} \qquad \sigma_{all} = \sqrt{\frac{\sum_{a=1}^{N} \left(X_a - \bar{X}_{all} \right)^2}{N}} \qquad (1)$$

Measuring sample of *n* members of population $$\langle X \rangle = \bar{X} = \frac{\sum_{a=1}^{n} X_a}{n} \qquad \sigma = \sqrt{\frac{\sum_{a=1}^{n} \left(X_a - \bar{X} \right)^2}{n}} \tag{2}$$

Note that if we knew the true mean of the entire population of *N* members when we measured our sample, we could calculate the standard deviation of our sample of *n* a little differently. That is,

Measuring *n* members of population when we know $\bar{X}_{all}$ $$\sigma = \sqrt{\frac{\sum_{a=1}^{n} \left(X_a - \bar{X}_{all} \right)^2}{n}} \quad . \tag{3}$$

2 The Classical Central Limit Theorem

The central limit theorem has a number of variations depending on such things as whether processes are independent or not. We focus herein on the simplest variation, what is known as the classical central limit theorem.

2.1 A Simple Example

To illustrate, we will use the example of Fig. 2 above, the uniform probability distribution.

We start by sampling 10 balls at random and finding the mean of the numbers written on them. Each time we draw a ball, we put it back in the bin before we draw the next ball. We call this test sample #1 and label our resulting mean value $\bar{X}_1$ (with subscript 1). *n* in (2) equals 10 here. *N* = 10,000, but we don't use that fact in (4).

$$\langle X_1 \rangle = \bar{X}_1 = \frac{\sum_{a=1}^{10} X_{1a}}{10} \tag{4}$$

For this, we might get a histogram like Fig. 3, where $\bar{X}_1$=60.5.

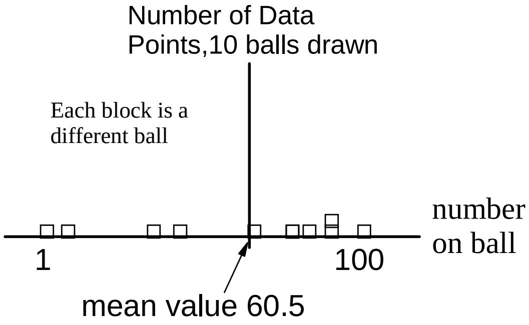

Figure 3. Sample #1 of 10 Balls from Uniform Probability Distribution of 10,000 Balls

Let's do the same thing again a second time, to get a second mean value for this second time.

$$\langle X_2 \rangle = \bar{X}_2 = \frac{\sum_{a=1}^{10} X_{2a}}{10} \tag{5}$$

Our histogram might look like Fig. 4 with a different mean value $\bar{X}_2 = 48.4$.

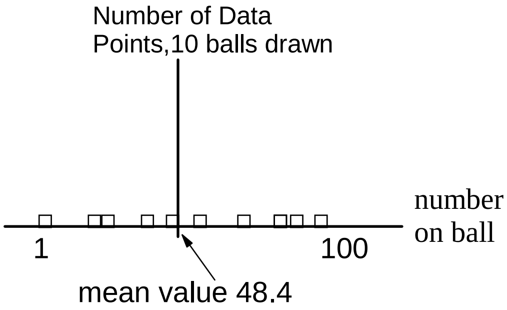

Figure 4. Sample #2 of 10 Balls from Uniform Probability Distribution of 10,000 Balls

Let’s now repeat that procedure over and over, each time (the *i*th time) getting a value

$$\langle X_i \rangle = \bar{X}_i = \frac{\sum_{a=1}^{10} X_{i\,a}}{10} \;. \tag{6}$$

Assume we do this procedure a total of N_T times (*T* subscript for “total tests”). So we now have a set $\bar{X}_i$ of N_T different mean values. Let’s plot these as a histogram in the LHS of Fig. 5.

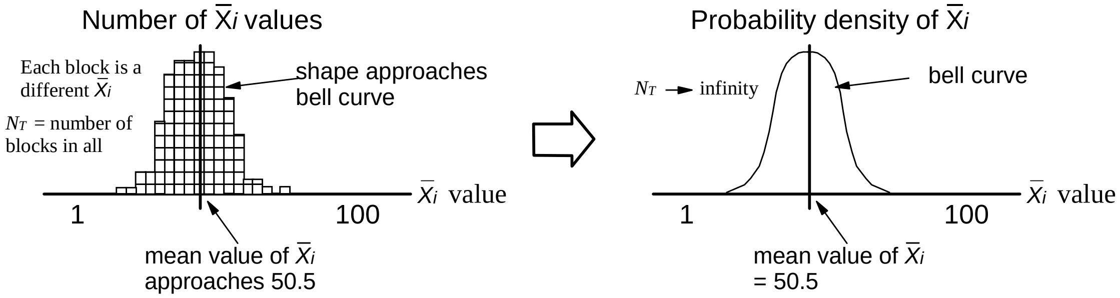

Figure 5. Histogram of Mean Values $\bar{X}_i$ for Test Samples (*n* = 10 per test) for N_T Tests

<u>Bottom line:</u> We get a histogram approaching a bell (Gaussian, normal) shape curve when we plot all our mean values from each sample test AND *this is true even for non-Gaussian underlying distributions of our whole population* (such as the uniform distribution of Fig. 2.) For total number of test samples $N_T \rightarrow \infty$, the curve becomes a perfect, smooth bell (Gaussian) shape.

<u>Proof:</u> We don’t prove it here, but a proof can be found in many statistics books. Here we only try to make the concept easy to understand.

<u>Presumption for What Follows</u>: In the following, and the way the central limit theorem is usually posed, we assume we have an infinite number of tests, i.e., $N_T \rightarrow \infty$. That is, we want to talk about a smooth distribution of test sample means $\bar{X}_i$, i.e., a probability density, as in the RHS of Fig. 5.

2.2 Quantifying the Central Limit Theorem

Note the following, again stated without proof.

2.2.1 The mean (average) of our test sample means (averages),

The mean (average) of our test sample means (averages), which we designate with $\langle \bar{X} \rangle$ (notation gets a bit unwieldy with the "mean of means") approaches the average of the underlying whole population as the number of tests N_T gets large. This is the law of large numbers and should be intuitively obvious.

$$\langle \bar{X} \rangle \to \bar{X}_{all} \quad \text{as} \quad N_T \to \infty \tag{7}$$

As noted above, we will assume the limiting case in (7) holds, so we can talk about a smooth distribution of the test sample mean (average) values $\bar{X}_i$ (RHS of Fig. 5) where the mean of that distribution equals the mean of the original underlying distribution of the whole population (Fig. 2).

2.2.2 The standard deviation of our test sample means (averages),

The standard deviation of our test sample means (averages), as displayed by the width of the bell curve in the RHS of Fig. 5, varies with *n*, the number of measurements (number of balls in Figs. 2 to 5) in each test sample. We designate this standard deviation of our test sample means (averages) with the symbol σ_n. The precise dependence is (where σ_{all} is the standard deviation of the whole underlying population as in Fig. 2 and (1))

$$\sigma_n = \frac{\sigma_{all}}{\sqrt{n}} . \tag{8}$$

Note this tells us that the greater number of measurements *n* we have in each of our test samples, the smaller the standard deviation of the distribution of the test sample means. In concrete terms, if we had taken 20 measurements per test in Figs. 2 to 5 instead of 10 measurements, the curve on the RHS of Fig. 5 would be narrower.

Intuitively this makes sense. The more measurements we have per test sample, the more likely the mean of any single test sample will be closer to the mean of the whole population. So the individual test sample means $\bar{X}_i$ should tend to cluster nearer the whole population mean $\bar{X}_{all}$.

2.2.3 Normalized distribution

A normalized (area underneath = 1) Gaussian distribution has form

$$\rho(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\bar{x})^2}{2\sigma^2}} . \tag{9}$$

For us (see RHS of Fig. 5), $x \to \bar{X}_i$, $\bar{x} \to \bar{X}_{all}$, and $\sigma \to \sigma_n$. Thus, the Gaussian distribution we get in the central limit theorem for the test sample mean values (see RHS Fig. 5) using (8), is

$$\rho(\bar{X}_i) = \frac{1}{\sigma_n\sqrt{2\pi}} e^{-\frac{(\bar{X}_i-\bar{X}_{all})^2}{2\sigma_n^2}} = \frac{\sqrt{n}}{\sigma_{all}\sqrt{2\pi}} e^{-\frac{(\bar{X}_i-\bar{X}_{all})^2}{2\frac{\sigma_{all}^2}{n}}} . \tag{10}$$

For greater measurements per sample *n*, the bell curve gets higher and narrower.

2.3 Points to Note

2.3.1 Independence of measurements

The central limit theorem, in its simplest (classical) guise as shown above, works for independent measurements. No measurement can depend on any other measurement.

By way of example, in our ball sampling case of Figs. 2 to 5, after drawing a ball and recording its number, we put the ball back into the bin. This made the next drawing of a ball independent of what had gone before (or other measurements). Had we not done that, then the drawing of the first ball, and not returning it to the bin, would have an influence on our odds for the next drawing. We would have only had 99 chances to draw the same number again, but 100 chances for every other number. The second measurement would not be independent of the first.

Similarly, each test sample (of several measurements each) must be independent of other test samples.

Bottom line: For the classical central limit theorem, measurements must be independent of one another.

2.3.2 Randomness of measurements

For the central limit theorem, measured variable (like number on a ball X_i above) must vary randomly with measurements. By way of example, if balls in the above example with numbers under 20 were hollow, like ping-pong balls, and the rest were solid, like golf balls, then the hollow, lighter balls would tend to rise to the top of the bin. And we would be more likely to pick out a lower numbered ball, so our averages would tend to be lower, below the random mean of 50.5.

Note that we can still put the selected ball back into the bin before our next measurement and have independence of measurements, but the selection process would not be random.

So, we can have independence without randomness. And we can have randomness (all balls the same weight and size), but not independence (not putting a selected ball back into the bin before the next selection). We need both for the classical central limit theorem to work.

Bottom line: For the central limit theorem, the measurement process must be random.

2.3.3 Works for any underlying distribution

Note that the distribution within any given test sample, as in Figs. 3 and 4, will tend to look like the distribution of the parent (underlying) population, as in Fig. 2. However, as noted before, the distribution of the means of the test samples will be Gaussian (Fig. 5), regardless of the form of the distribution of the underlying population.

Bottom line: Provided measurements are independent and random, the classical central limit theorem says the distribution of test sample means will be Gaussian, regardless of the distribution form (curve shape) of the underlying population.

To return to home page → www.quantumfieldtheory.info