The Normal DIstirbution and the Central Limit Theorem

1 The Gaussian (Normal) Distribution

1.1 Limit Theorems in Probability

We start by noting that, just like in any other branch of Mathematics, it is sometimes useful or necessary to *approximate* formulas and quantities, due to the difficulty in handling them directly. Also, many applications involve a great number of data, and to approximate such great numbers as "infinity" works rather well.

This is true in probability as well, and checking what happens "in the limit" is also a good way to understand both theoretical concepts, and application methods.

1.1.1 An Intuitive Idea Of The Main Limit Theorems

Suppose we look at a very long sequence of independent, identically distributed RVs - in fact, let us assume that they represent successive measurements of a given quantity.

1.1.2 The Law Of Large Numbers (LLN)

If you recall, an empirical "fact" is that if you throw a fair coin many, many times, after many flips, the frequency of heads will hover very close to 50%. Of course, there is some circularity in this statement: when is a coin actually "fair"? Well, thats when, when thrown a huge number of times, the frequency of heads is very, very close to 50%!

Anyway, this "fact" is the intuitive motivating basis for constructing probability. In fact, the goal of probability is to produce a model for this behavior, so that the same can be predicted and controlled in some way. The **LLN** is a *theorem* that shows which is the model for the coin-flipping experiment (and any similar one), producing, at the mathematical level, the behavior that we recognize in nature - hence, it seems to validate the use of probability to describe "random" phenomena".

The theorem is quite easy and direct (at least in its simplest form). First of all, note that for a RV X with expectation μ, and variance σ^2, the following

elementary fact holds true (we'll work out the formula for continuous RVs, but it works exactly the same for discrete ones): for any $k > 0$

$$\sigma^2 = E\left[(X-\mu)^2\right] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx =$$

$$= \int_{-\infty}^{k+\mu} (x-\mu)^2 f(x)\, dx + \int_{|x|=k+\mu}^{\infty} (x-\mu)^2 f(x)\, dx$$

Now, observe that, in the second integral, $x \geq k+\mu$, so that $(x-\mu)^2 \geq k^2$

$$\int_{k+\mu}^{\infty} (x-\mu)^2 f(x)\, dx \geq k^2 \int_{|x|=k+\mu}^{\infty} f(x)\, dx =$$

$$= k^2 P\left[(X-\mu)^2 \geq k^2\right]$$

Also, very brutally, we can certainly say that

$$\int_{-\infty}^{|x|=k+\mu} (x-\mu)^2 f(x)\, dx \geq 0$$

so that

$$\sigma^2 \geq k^2 P\left[(X-\mu)^2 \geq k^2\right]$$

and, since

$$(X-\mu)^2 \geq k^2 \Leftrightarrow |X-\mu| \geq k$$

$$P\left[|X-\mu| \geq k\right] \leq \frac{\sigma^2}{k^2} \tag{1}$$

(1) is called "Chebyshev's Inequality", and is true for any distribution with expected value and variance (as such, it is used in some statistical contexts, when the "normal approximation", which we will discuss momentarily, just fails badly).

Now, notice the following simple fact: if $X_1, X_2, \ldots, X_n$ are *independent* (or, at least, *uncorrelated* RVs), all with the same distribution, calling $EX_i = \mu$, and $Var\left[X_i\right] = \sigma^2$, we notice that (since all covariances are zero)

$$Var\left[\sum_{i=1}^{n} X_i\right] = n\sigma^2$$

If we now take the *arithmetic mean* of these variables, we obtain

$$Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2} Var\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

Hence,

$$\lim_{n\to\infty} Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = 0$$

and, moreover, thanks to Chebyshev's Inequality,

$$P\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > k\right] \leq \frac{\sigma^2}{k^2 n}$$

or,

$$\lim_{n\to\infty} P\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > k\right] = 0$$

In other words, the RVs $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is such that $E\bar{X}_n = \mu$, $Var\left[\bar{X}_n\right] = \frac{\sigma}{n}$, and the probability that $\bar{X}_n$ will deviate from its expected value by more than any fixed number will tend to zero as $n \to \infty$.

This is the (weak) LLN, and it says that taking the average of many i.i.d. RVs will produce a RV whose distribution is more and more peaked around its expected value.

In particular, if $X = 1_A$, then $EX = P[A]$, and observing repeatedly, and independently the occurrence of a random event, taking the fraction of successes will provide an approximation to the probability of the event, getting better as we increase the number of observations.

This is one of the fundamental theorems that provide a foundation for mathematical statistics, and also provides a justification for most applications of probability theory to the real world.

1.1.3 The Central Limit Theorem

One way to look at the LLN is to think of looking at the sum of n independent variables that have each been divided by n. If, say, $n = 1000$, a quantity originally measured in, say, meters, is just like it was now being measured in kilometers. We are thus looking at the sum of many extremely small numbers, and, even though they are random, their oscillation around their average value (their expected value, that is) are going to be so small, that they are practically unobservable - thus they behave essentially like constants, and their sum is, approximately, $n \cdot E\left[\frac{\bar{X}_n}{n}\right] = EX$.

A scale of $\frac{1}{n}$ causes all fluctuations to disappear... this is another way to look at the LLN, and it suggests that if we reduced the scale in a less drastic way, we might keep track of oscillations around the mean, without being swamped by them.

By noting that the variance of the average of Bernoulli variables happens to be $\frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$ (which, again, shows that the mean becomes essentially constant for large n), we find that subtracting the mean (to avoid the sum to run off, having mean np), and dividing by $\sqrt{n}$ instead of by n, we should get a non trivial limit. This is the case, and it is the result known as the "Central Limit Theorem" (CLT), where "Central" refers to "Theorem", not to "Limit".

The precise statement is as follows

Theorem: Suppose $X_1, X_2, X_3, \ldots$ form a sequence of independent, identically distribute RVs, with $EX_i = \mu$, $Var[X_i] = \sigma^2 < \infty$. Then, as $n \to \infty$,

$$P\left[\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \leq x\right] \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$

(i.e., the cdf tends to the cdf of a standard normal RV). Note that $\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$, so that the theorem can be seen as a statement about the sum of the RVs, or about their average.

The proof of the CLT is tedious for the Bernoulli case (essentially, it is a careful application of the Stirling approximation for factorials), and is not elementary in the general case (that is, it relies on a theorem that is not elementary - if you take that theorem for granted, the rest of the proof is easy: the book carries such a proof in sec. 8.3).

2 Further Comments

2.1 The "Sample Variance" vs. the "Population Variance"

If you are going to take a course in statistics, or already took one, you have certainly met the following topic.

Suppose you take repeated, say n, independent observation of some quantity, whose behavior is modeled by a random variable. We describe this mathematically as observing n independent random variables, all with the same distribution $X_1, X_2, \ldots, X_n$. If the common distribution is assumed to be Gaussian[1] with mean μ, and variance σ^2, then, as we know, the *sample mean* $\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$ is also normal, with mean μ, and variance σ^2/n. In general, we don't know these values, and use *inferential statistics* to get estimates for them. For this purpose, we use the fact that, by the Law of Large Numbers, $\overline{X}$ should be close to μ. To find something close to σ^2, we note that $E\left[\sum_{k=1}^{n}(X_i-\mu)^2\right] = \sum_{k=1}^{n} E\left[(X_i-\mu)^2\right] = n\sigma^2$, so that, by the same theorem, if n is large enough, $\frac{1}{n}\sum(X_i-\mu)^2 \approx \sigma^2$. However, μ is generally unknown, so we substitute $\overline{X}$ for it. Now, with a little patience, it is not too hard to realize that $E\left[\sum_{k=1}^{n}\left(X_i-\overline{X}\right)^2\right] = \sum_{k=1}^{n} E\left[\left(X_i-\overline{X}\right)^2\right] = (n-1)\sigma^2$, and, since $\frac{n-1}{n} \to 1$, as $n \to \infty$, we have that

$$\frac{1}{n}\sum_{k=1}^{n}\left(X_i-\overline{X}\right)^2 \approx \frac{1}{n-1}\sum_{k=1}^{n}\left(X_i-\overline{X}\right)^2 \approx \sigma^2$$

for large enough n. The first expression is called, in statistical lingo, the *population variance*, and the second the *sample variance*. Which to use for the purpose of estimating σ^2 is clearly purely a matter of taste. First of all, if n is large enough, the difference between the two is irrelevant (the whole estimation procedure is generally much more approximate anyway). Second, if you want to find some virtue in one or the other, both have *theoretical* features that may lead you to prefer one or the other:

- The "population variance" turns out to be the *Maximum Likelihood Estimator* for σ^2, a property with some theoretical advantages, especially in Bayesian statistics.
- The "sample variance" happens to be *unbiased*, that is its expectation equals σ^2 (the expectation of the "population variance is $\frac{n-1}{n}\sigma^2$), and is always (slightly) smaller than the "sample variance". This might seem impressive, but is of very little practical significance. Being unbiased, allows some theoretical results on the optimality of unbiased estimators to apply, but the practical consequences are marginal, if any.

Nonetheless, standard practice is to use the sample variance for statistical purposes. This is mainly a historical heredity, and, in theory, we could rewrite the whole chapter in terms of "population" variances, but, obviously, the effort would not be worth it.

[1] These tools are used also in non-Gaussian situations, assuming that n is large enough for an application of the Central Limit Theorem, which tells us that $\overline{X}$ is approximately Gaussian. The size of the sample has to extra large, since the distribution of $\sum_{k=1}^{n}\left(X_i-\overline{X}\right)^2$ takes a lot longer to approach that of the Gaussian case (called a χ^2 distribution). People have shown that the procedure is justified for reasonable sample sizes when the non Gaussian distribution is symmetric, and not to dispersed.

One thing to remember, though, is that the attempt to convince students that the "sample" version is the "correct" one, some introductory statistics books come out with erroneous statements. Here are a couple of examples: you could read that

- *The population variance always underestimates the true variance* σ^2. This is so false on its face that you wonder how it got into print: both variances are random quantities wobbling around σ^2, so they can both under- or over-estimate σ^2 (and you cannot tell, since you don't know σ^2).

- *The sample variance is half the time smaller than and half the time larger than the true variance* σ^2. This author is confused between *expectation* (which is what unbiasedness is about), and *median*, which is generally different form the expectation unless the distribution is symmetric – and the distribution of $\frac{1}{n-1}\sum_{k=1}^{n}\left(X_i-\overline{X}\right)^2$ is not symmetric at all.

2.2 Multidimensional Normal Distributions

Even though there is, technically, not a great deal of extra work involved to deal with this case, we will have to leave it for future developments.

Still, if you have some familiarity with conic sections – i.e., multi-dimensional quadratic functions – the work involved here is not difficult. Basically, much like in the 1-dimensional case, you are looking at a multi-dimensional "negative definite" quadratic polynomial in the exponent (we need the density to be integrable).

Just looking at the exponent (and concentrate on the case of dimension 2 to keep some visual intuition handy), you will see that the "iso-lines" (the curves on which the exponent is constant) have to be ellipses, and, by completing the square (plus the extra tricks involved when your ellipse has axes that are not parallel to the coordinate axes), you can bring the exponents into a standard form.

Also, you will notice that a linear change of variables (in fact, a *rotation of the axes*), will change the exponent so that the mixed terms (in xy, as opposed to x^2, y^2, x, y) disappear, and the quadratic polynomial can be written as a sum of a polynomial in x and a polynomial in y. Since this is an exponent, the function can can be written as a product of a density in x with a density in y. In other words, any multidimensional Gaussian, can be transformed, via a linear transformation of the variables, into a collection of independent Gaussians. For example, starting from

$$e^{3x^2-2xy+3y^2-12x+4y+12}$$

changing to new variables $u=\frac{x+y}{2}, v=\frac{x-y}{2}$ (a rotation of the axes by $\frac{\pi}{4}$), we arrive at an expression like

$$e^{4(u-1)^2+8(v-1)^2}=e^{4(u-1)^2}e^{8(v-1)^2}$$