TWO PILLARS OF PROBABILITY THEORY

N. H. BINGHAM

Sarajevo Workshop: IECSMA-2013

August 2013

Law of Large Numbers (LLN):
"Law of Averages"

Central Limit Theorem (CLT):
"Law of Errors"

1. Compound interest

If one invests \$1 for a year at $100x$ % simple interest, after one year one has $1 + x$ (\$). One has $1 + \frac{1}{2}x$ after half a year, and investing this for the second half gives $(1 + \frac{1}{2}x)^2$. Similarly, investing with simple interest calculated n times per year gives $(1+(x/n))^n$. Clearly, we are better off the bigger n is. As n increases, our final capital increases, to the limit e^x:

$$(1 + \frac{x}{n})^n \to e^x \quad (n \to \infty).$$

This is *compound interest*, or *exponential growth* – the limit of simple interest as the interest is compounded continuously. Similarly, if $x_n \to x$,

$$(1 + \frac{x_n}{n})^n \to e^x \quad (n \to \infty).$$

This also extends to complex numbers $z_n \to z$:

$$(1 + \frac{z_n}{n})^n \to e^z \quad (n \to \infty). \qquad (*)$$

It turns out that this result is a crucial ingredient to the proofs of our two pillars of Probability Theory below.

2. Characteristic functions (CFs)

For X a random variable (rv), its *characteristic function (CF)* is

$$\phi(t) = \phi_X(t) := E[e^{itX}]$$

($t \in \mathbb{R}$ will suffice here, but $t \in \mathbb{C}$ is also possible). Then if X, Y are independent rvs,

$$\phi_{X+Y}(t) = E[e^{it(X+Y)}]$$

$$= E[e^{itX}.e^{itY}] \quad \text{(property of exponentials)}$$

$$= E[e^{itX}].E[e^{itY}] \quad \text{(independence)}$$

$$= \phi_X(t).\phi_Y(t).$$

So *adding* independent rvs corresponds to *multiplying* CFs (but to *convolution* of distributions: this involves an integration, so is harder, and much harder if we add lots of terms). So if $S_n := X_1 + \ldots + X_n$, $\phi_{S_n}(t) = \phi_X(t)^n$.
Also, from

$$e^x = 1 + x + \ldots + x^n/n! + o(x) \quad (x \to 0) :$$

if $\mu_n := E[X^n]$ exists, one would expect

$$\phi_X(t) = E[\sum_{k=0}^{n} (it)^k X^k/k! + o(t^n)]$$

$$= \sum_{0}^{n} (it)^k E[X^k]/k! + o(t^n) \quad (t \to 0).$$

This is in fact true; we assume it here. (The proof uses Measure Theory, to which we return later.)

The CF determines the distribution it comes from uniquely, and the correspondence between the distribution and its CF is suitably *continuous* (Lévy's continuity theorem for CFs, below). So the CF encodes all the information in the distribution, in a way that is often more convenient (e.g., when adding independent rvs, as here).

The CF is a special kind of *Fourier transform* (actually, a Fourier-Stieltjes transform). This is related to the *Laplace transform*. Such integral transforms are very useful, in both theory and applications.

3. The standard normal law

$$\phi(x) := \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

is a (probability) *density* (function) (non-negative, and integrates to 1).

Proof. Write I for its integral. Then

$$I^2 = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2} dx. \int_{-\infty}^{\infty} e^{-\frac{1}{2}y^2} dy.$$

Write this product of two repeated integrals as a double integral over the (x, y)-plane (integrand $e^{-\frac{1}{2}(x^2+y^2)}$). Change to plane polar coordinates (integrand $e^{-\frac{1}{2}r^2}$, $dxdy \to rdrd\theta$). Now do the r and θ integrations separately. //

The corresponding (probability) *distribution* (function) is

$$\Phi(x) := \int_{-\infty}^{x} \phi(y) dy.$$

If X is a rv with this distribution, then

$$P(X \leq x) = \Phi(x).$$

Φ is called the *standard normal* (distribution, or law), $\Phi = N(0,1)$. The general normal law $N(\mu, \sigma^2)$ has mean μ and variance $\sigma^2 > 0$; if $X \sim N(\mu, \sigma^2)$, then $(X - \mu)/\sigma \sim N(0,1)$. The moment-generating function (MGF) is

$$M_X(t) := E[e^{tX}] = \int e^{tx} . \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx$$

$$= \int \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[(x-t)^2 - t^2]} dx,$$

completing the square. Take out the $e^{\frac{1}{2}t^2}$, and write $u := x - t$. The integral is 1 ('normal density'), so

$$M_X(t) = E[e^{tX}] = e^{\frac{1}{2}t^2}.$$

Formally replace t by it ($i = \sqrt{-1}$) to get

$$\phi_X(t) = E[e^{itX}] = e^{-\frac{1}{2}t^2}.$$

All this is correct! One needs some Complex Analysis (analytic continuation gives it immediately; Cauchy's theorem gives it after a calculation).

4. The Weak Law of Large Numbers (WLLN)

For $X, X_1, X_2, \ldots$ independent and identically distributed (iid) random variables (rvs), with mean μ:

$$E[|X|] < \infty, \qquad E[X] = \mu,$$

$S_n := X_1 + \ldots + X_n$:

Theorem (Weak Law of Large Numbers (WLLN)).

$$S_n/n \to \mu \quad (n \to \infty) \text{ in probability,}$$

i.e. for all $\epsilon > 0$,

$$P(|S_n/n - \mu| > \epsilon) \to 0 \quad (n \to \infty).$$

Proof. If the X_k have CF $\phi(t)$, then as the mean μ exists $\phi(t) = 1 + i\mu t + o(t)$ as $t \to 0$. So $(X_1 + \ldots + X_n)/n$ has CF

$$E \exp\{it(X_1 + \ldots + X_n)/n\} = [\phi(t/n)]^n$$

$$= [1 + \frac{i\mu t}{n} + o(1/n)]^n,$$

for fixed t and $n \to \infty$. By $(*)$, the RHS has limit $e^{i\mu t}$ as $n \to \infty$. But $e^{i\mu t}$ is the CF of the constant μ. This suggests that

$$(X_1 + \ldots + X_n)/n \to \mu \quad (n \to \infty) \quad \text{in distribution.}$$

This is indeed true, by Lévy's continuity theorem (which we quote). As the limit μ is constant, this gives further

$$(X_1 + \ldots + X_n)/n \to \mu \quad (n \to \infty) \quad \text{in probability}$$

(one can check this easily). //

5. The Central Limit Theorem (CLT)

The *variance* of a rv X is

$$\sigma_X^2 := E[(X - E[X])^2],$$

and then

$$\sigma_X^2 = E[X^2] - (E[X])^2.$$

The proof below is just the proof of the WLLN above, but with the Taylor expansion of the CF carried one term further, because now we have one more moment.

Theorem (Central Limit Theorem (CLT)). If X_i are iid with mean μ and variance σ^2, then as $n \to \infty$

$$(S_n - n\mu)/(\sigma\sqrt{n}) \to \Phi = N(0,1) \quad \text{in distribution.}$$

Proof. When we subtract μ from each X_k, we change the mean from μ to 0 and the second moment from μ_2 to the variance σ^2. So by

the moments property of CFs, $X_k - \mu$ has CF $1 - \frac{1}{2}\sigma^2 t^2 + o(t^2)$ as $t \to 0$. So $X_1 + \ldots + X_n - n\mu$ has CF

$$E \exp\{it(X_1 + \ldots + X_n - n\mu)\} = [1 - \frac{1}{2}\sigma^2 t^2 + o(t^2)]^n$$

Replace t by $t/(\sigma\sqrt{n})$ and let $n \to \infty$:

$$E \exp\{it(X_1 + \ldots + X_n - n\mu)/(\sigma\sqrt{n})\}$$

$$= [1 - \frac{1}{2}.\frac{t^2}{n} + o(1/n)]^n \to \exp\{-t^2/2\} \quad (n \to \infty),$$

by $(*)$ again.
The left is the CF of $(S_n - n\mu)/(\sigma\sqrt{n})$;
the right is the CF of $\Phi = N(0, 1)$.
By the continuity theorem for CFs, this gives

$$(S_n - n\mu)/(\sigma\sqrt{n}) \to \Phi \text{ in distribution. //}$$

6. The Strong Law of Large Numbers (SLLN).

It turns out that the conclusion of the WLLN (convergence in probability) can be greatly strengthened, to convergence with probability one. We do not need stronger conditions, but the proof is now much harder, so is omitted.

Theorem (Strong Law of Large Numbers (SLLN)). If the X_n are iid with mean μ,

$$S_n/n \to \mu \quad (n \to \infty) \text{ with probability 1.}$$

We abbreviate 'with probability 1' to 'almost surely', or 'a.s.':

$$S_n/n \to E[X] = \mu \quad a.s.$$

7. Interpretation

The LLN (in Weak or Strong form) gives the mathematical form of the 'folklore' statement known as the *Law of Averages*. This is known to the man or woman in the street. It says, e.g., that fair coins fall heads about half the time in the long run.

The CLT gives the mathematical form of the *Law of Errors*. This is known to the physicist in the street, and says that errors are normally distributed about the mean. E.g.: to measure a physical constant (electrical conductivity of copper, specific heat of mercury, etc.): measure it n times, independently. Each reading X_n is the 'right answer', c say, plus a *measurement error*, ϵ_n say. If the experiment is *unbiased* ('right on average'), $E[\epsilon_n] = 0$. Then LLN says $\bar{\epsilon} := \sum_1^n \epsilon_k/n \sim 0$, so

$$\bar{X} \sim E[X] = c :$$

$\bar{X}$ gives us our estimate of the *answer*. Similarly, $\overline{X^2} \sim E[X^2]$, so

$$\overline{X^2} - (\bar{X})^2 \sim E[X^2] - (E[X])^2 = var(X) = \sigma^2,$$

which gives us our estimate of the *accuracy*. A conclusion of the form "$c = 7.034 \pm 0.003$" means that our estimate of the answer is 7.034, our estimate of the standard deviation σ (SD: square root of the variance σ^2) is 0.003, and that our average reading $\bar{X}$ is approximately normally distributed with this mean and this SD.

Note. The variance σ^2 has good mathematical properties. But the SD σ has the same units as the data, and so is better suited for use in Physics, etc. So we use both.

8. A little history

Jakob Bernoulli (1654-1705)
Ars Conjectandi (AC) (1713, posthumous)
WLLN for 'Bernoulli trials' (tossing a perhaps biased coin)
Abraham de Moivre (1667-1754)
The Doctrine of Chances (DC), 1718/1738/1756
Normal distribution; CLT for Bernoulli trials
Carl Friedrich Gauss (1777-1855)
Theoria motus ... (TM), 1809.
Gauss was the greatest mathematician of all time. The normal distribution is also called the *Gaussian*, after him and his work on CLT.
Paul Lévy (1886-1971)
Calcul des Probabilités (CP), 1925:
Lévy's continuity theorem for CFs; modern proof of WLLN and CLT.
Andrei Nikolaevich Kolmogorov (1903-1987)
Grundbegriffe der Wahrscheinlichkeitsrechnung, 1933 (Foundations of probability theory): SLLN.
Kolmogorov was the greatest probabilist of all time. His SLLN ended a journey Bernoulli began 220 years before!

9. Measure Theory

Henri Lebesgue (1875-1941)

Thesis, *Intégrale, longueur, aire*, 1902.

It turns out that the mathematics needed to do Probability Theory properly is Measure Theory, initiated by Lebesgue.

One also needs Complex Analysis, initiated by Augustin-Louis Cauchy (1789-1857) in 1825-29.

Moral: if you want to do Probability Theory properly, learn as much Analysis as possible!

Probability Theory is not just very good as mathematics – it is also very useful, as it provides the tools needed to do Statistics, a subject of great and growing practical importance.

NHB