Opinionated Lessons in Statistics

by Bill Press

#10 The Central Limit Theorem

The Central Limit Theorem is the reason that the Normal (Gaussian) distribution is uniquely important. We need to understand where it does and doesn't apply.

Let $S = \frac{1}{N}\sum X_i = \sum \frac{X_\mathrm{i}}{N}$ with $\langle X_i \rangle \equiv 0$

Can always subtract off the means, then add back later.

Then

$$\phi_S(t) = \prod_i \phi_{X_\mathrm{i}/N}(t) = \prod_i \phi_{X_\mathrm{i}}\left(\frac{t}{N}\right)$$

$$= \prod_i \left(1 - \tfrac{1}{2}\sigma_i^2 \frac{t^2}{N^2} + \cdots \right)$$

Whoa! It better have a convergent Taylor series around zero! (Cauchy doesn't, e.g.)

$$= \exp\left[\sum_i \ln\left(1 - \tfrac{1}{2}\sigma_i^2 \frac{t^2}{N^2} + \cdots \right)\right]$$

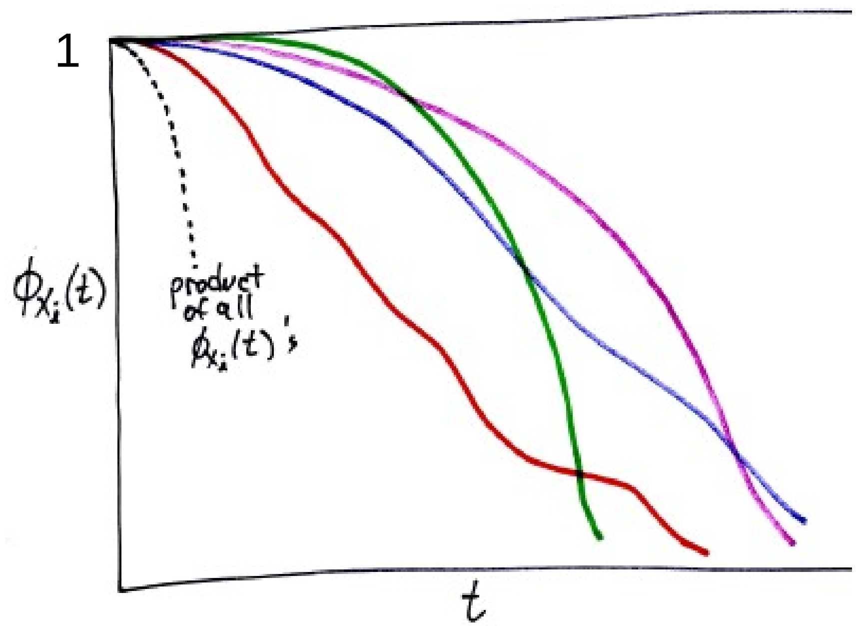
These terms decrease with N, but how fast?

$$\approx \exp\left[-\frac{1}{2}\left(\frac{1}{N^2}\sum_i \sigma_i^2\right) t^2 + \cdots\right]$$

So, S is normally distributed

$$p_S(\cdot) \sim \mathrm{Normal}(0, \tfrac{1}{N^2}\sum \sigma_i^2)$$

Intuitively, the product of a lot of arbitrary functions that all start at 1 and have zero derivative looks like this:



Because the product falls off so fast, it loses all memory of the details of its factors except the starting value 1 and fact of zero derivative. In characteristic function space that's basically the CLT.

CLT is usually stated about the sum of RVs, not the average, so

$$p_S(\cdot) \sim \mathrm{Normal}(0, \tfrac{1}{N^2} \sum \sigma_i^2)$$

Now, since

$NS = \sum X_i$ and $\mathrm{Var}(NS) = N^2 \mathrm{Var}(S)$

it follows that the simple sum of a large number of r.v.'s is normally distributed, with variance equal to the sum of the variances:

$$p_{\sum X_\mathrm{i}}(\cdot) \sim \mathrm{Normal}(0, \sum \sigma_i^2)$$

<u>If</u> N is large enough, and <u>if</u> the higher moments are well-enough behaved, and <u>if</u> the Taylor series expansion exists!

Also beware of borderline cases where the assumptions technically hold, but convergence to Normal is slow and/or highly nonuniform. (This can affect p-values for tail tests, as we will soon see.)

Since Gaussians are so universal, let's learn estimate the parameters μ and σ of a Gaussian from a set of points drawn from it:

For now, we'll just find the maximum of the posterior distribution of (μ,σ), given some data, for a uniform prior. This is called "maximum a posteriori (MAP)" by Bayesians, and "maximum likelihood (MLE)" by frequentists.

The data is: $x_i, \; i = 1, \ldots, N$

The statistical model is: $$P(\mathbf{x}|\mu,\sigma) = \prod_i \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}}$$

The posterior estimate is: $$P(\mu,\sigma|\mathbf{x}) \propto \frac{1}{\sqrt{2\pi}\sigma^N} e^{-\frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2} \times \cancel{P(\mu,\sigma)}$$ uniform

Now find the MAP (MLE):

$$0 = \frac{\partial P}{\partial \mu} = \frac{P}{\sigma^2}\left(\sum_i x_i - N\mu\right) \;\Rightarrow\; \mu = \frac{1}{N}\sum_i x_i$$

Ha! The MAP mean is the sample mean, the MAP variance is the sample variance!

$$0 = \frac{\partial P}{\partial \sigma} = \frac{P}{\sigma^3}\left[-N\sigma^2 + \sum_i (x_i-\mu)^2\right] \;\Rightarrow\; \sigma^2 = \frac{1}{N}\sum_i (x_i-\mu)^2$$

Bessel's Correction: $N-1$

It won't surprise you that I did the algebra by computer, in Mathematica:

```
p =
 (1 / s ^N)
  Exp[-(1 / (2 s ^2)) Sum[(x[i] - mu) ^2, {i, 1, N}]]
```

$$e^{-\frac{\sum_{i=1}^{N} (-mu+x[i])^2}{2 s^2}} s^{-N}$$

```
Simplify[D[p, mu]]
```

$$-\frac{1}{2} e^{-\frac{\sum_{i=1}^{N} (-mu+x[i])^2}{2 s^2}} s^{-2-N} \sum_{i=1}^{N} -2 (-mu + x[i])$$

```
Simplify[D[p, s]]
```

$$e^{-\frac{\sum_{i=1}^{N} (-mu+x[i])^2}{2 s^2}} s^{-3-N} \left(-N s^2 + \sum_{i=1}^{N} (-mu + x[i])^2\right)$$