Applications of the Central Limit Theorem

October 23, 2008

Take home message. I expect you to know all the material in this note. We will get to the Maximum Liklihood Estimate material very soon!

1 Introduction

First, we state the central limit theorem

Theorem 1 *Suppose that $X_1, X_2, ...$ is an infinite sequence of independent, identically distributed random variables with common mean $\mu = \mathbb{E}(X_1)$ and finite variance $\sigma^2 = V(x_1)$. Then, if we let $S_n = X_1 + \cdots + X_n$ we have that*

$$\lim_{n\to\infty} P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \ \le \ c\right) \ = \ \Phi(c) \ = \ \int_{-\infty}^{c} \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx.$$

There are many applications of this theorem to real-world problems, and in these notes we will give two: An application to hypothesis testing, and an application to noise cancellation.

2 Hypothesis Testing

Here we will give an example of how to use the CLT to test hypotheses. We have already seen how to do this using a chi-square test to determine whether to reject a hypothesized population distribution (with finitely many classes) as being false. Here we will do this for when the population breaks down into two classes, smokers and non-smokers.

2.1 The Main Problem

Problem. You read in a newspaper that 20% of Georgians smoke, and you decide to test this hypothesis by doing a poll on $1,000$ randomly selected Georgians with replacement (if the population you are testing is very large, then you would not need to test with replacement). Suppose that 205 of the responses are "smoker", while 795 are "non-smoker". Is the claim "20% of Georgians smoke" unreasonable? (Obviously not, but let's see what the math tells us...)

Well, in order to answer this question we would need more information; we would need to know what we mean by "unreasonable". Here we will mean "unreasonable" with respect to a certain statistical test which we presently describe:

Let $X_i = 1$ if respondant i says he/she is a smoker, and let $X_i = 0$ if he/she is not a smoker. These X_i's are independent Bernoulli random variables. Let $S_n = X_1 + \cdots + X_{1,000}$. If our hypothesis that 20% of Georgians smoke were correct, then $\mu = \mathbb{E}(X_i) = 0.2$, and $V(X_i) = \mu(1-\mu) = 0.16$; and so, the Central Limit Theorem would tell us that

$$\frac{S_{1,000} - 200}{\sqrt{1,000 \cdot 0.16}} \text{ is approximately } N(0,1), \tag{1}$$

in the sense that

$$P\left(\frac{S_{1,000} - 200}{12.64911} \ \leq \ c \right) \ \approx \ \Phi(c).$$

Now, if $S^*_{1,000}$ is the observed value of $S_{1,000}$ [1], and if

$$\gamma \ = \ \frac{S^*_{1,000} - 200}{12.64911}, \tag{2}$$

then, on the basis of the central limit theorem and (1) we wouldn't expect that γ is an atypical value for $N(0,1)$. In particular, we wouldn't expect that $|\gamma|$ is too big; that is, we wouldn't expect that

$$P(|N(0,1)| \ \geq \ |\gamma|) \ < \ 0.05$$

if $\mu = 0.2$ is the true mean.

[1] That is, we observe $S^*_{1,000}$ smokers.

Thus, we have the following basic statistical test

Statistical Test. Fix an $\alpha > 0$, typically $\alpha = 0.05$ or 0.01. Compute γ as in (2). If

$$P(|N(0,1)| \ \geq\ |\gamma|)\ =\ 2\Phi(-|\gamma|)\ <\ \alpha,$$

then we reject the hypothesis that the mean value of X_1 is μ; and, if this inequality is not satisfied, we do not reject it, which is not the same as saying we accept it.

In the example given above we have that

$$\gamma\ =\ \frac{205-200}{12.64911}\ =\ 0.39528,$$

and one can readily compute that

$$2\Phi(-0.39528)\ >\ 0.05$$

Thus, we do not reject the hypothesis that 20% of Georgians smoke.

3 Noise Cancellation

Suppose that a man is driving through the desert, and runs out of gas. He grabs his cellphone to make a call for help, dialing 911, but he is just at the edge of the broadcast range for his cellphone, and so his call to the 911 dispatcher is somewhat noisy and garbled. Suppose that the 911 dispatcher has the ability to use several cellphone towers to clean up the signal. Suppose that there are about 100 towers near to the stranded driver, and suppose that the signals they each receive at a particular instant in time is given by

$$X_1, ..., X_{100},$$

where

$$X_i\ =\ S\ +\ Y_i,$$

where S is the true signal being sent to the towers, and where Y_i is the noise. Suppose that all the noises $Y_1, ..., Y_{100}$ are independent and identically distributed, and further suppose they all have mean 0 and variance σ^2. Further, it is not unreasonable to assume that the noises are all normally distributed

– i.e. they are all $N(0, \sigma^2)$ – though we will not need this assumption for what follows.

The 911 dispatcher cleans up the signal by computing the average

$$\overline{X} \ = \ \frac{X_1 + \cdots + X_{100}}{100} \ = \ S + \frac{Y_1 + \cdots + Y_{100}}{100}.$$

Now, by the Central Limit Theorem, we would expect that

$$\frac{Y_1 + \cdots + Y_{100}}{100} \text{ is approximately } N(0, \sigma^2/100). \qquad (3)$$

Of course we need to be careful here – the central limit theorem only applies for n large, and just *how* large depends on the underyling distribution of the random variables Y_i. There are more powerful versions of the central limit theorem, which give conditions on n under which (3) holds under a precise notion of "is approximately". At any rate, if we assume that the Y_is are all independent normal random variables, then we don't even need the central limit theorem, because in that case we have that $\overline{X} - S$ is *exactly* $N(0, \sigma^2/100)$.

Now, suppose that, in fact, all the noises Y_is have variance $\sigma^2 = 1$. Then, the central limit theorem in the guise (3) would be telling us that the new noise $\overline{X} - S$ is approximately normal with variance 1/100, a 100-fold improvement in the noise variance gotten just using one tower!

3.1 Just How Good is the Averaging Method for Noise Cancellation – Can we Do Better?

It turns out that not only does averaging give us a pretty good way to cancel noise, but it is, in some sense, the best thing we could possibly try. The proper language is that taking the average $\overline{X}$ gives us a maximum liklihood estimate for the signal S, which is the same as the expected value of X_i for all $i = 1, ..., 100$. Let us make this more precise:

Maximum Liklihood Estimates. Suppose X is some random variable having a distribution that depends on a list of unknown, underlying parameters $\theta_1, ..., \theta_k$.[2] Let $f(x; \theta_1, ..., \theta_k)$ denote the pdf for X, given the parameters

[2] For example, perhaps X is normal with mean μ and variance σ^2, where neither μ nor σ^2 are known, in which case $\theta_1 = \mu$ and $\theta_2 = \sigma^2$.

$\theta_1, ..., \theta_k$. Suppose we make n independent observations of our random source X, and suppose these observations are the values $x_1, ..., x_n$. Then, the liklihood value of this observation is

$$L(x_1, ..., x_n; \theta_1, ..., \theta_k) \ = \ \prod_{i=1}^{n} f(x_i; \theta_1, ..., \theta_k).$$

We say that $\hat{\theta}_1, ..., \hat{\theta}_k$ are maximum liklihood estimates for $\theta_1, ..., \theta_k$, given the observations $x_1, ..., x_n$, if these values $\hat{\theta}_i$ maximize $L(x_1, ..., x_n; \theta_1, ..., \theta_k)$. Note that the $\hat{\theta}_i$s which maximize L may not be unique (there may be more than one global max).

Question. Why the word 'liklihood', and not, say, 'probability'? To answer this, note that in the discrete setting it is easy to describe what the liklihood function computes: It is just the probability that given particular values for $\theta_1, ..., \theta_k$, the observed values for some random variable X were $x_1, ..., x_n$. So, in this case 'liklihood' and 'probability' coincide. However, as we well know, the pdf for a *continuous* random variable X does not give us probability values when we plug in varlues for x, and hence the use of the word 'liklihood'.

In our case, let us suppose that the received signals X_i are, in fact, normal, with mean S, and variance σ^2; that is to say, the noises Y_i are $N(0, \sigma^2)$. Now suppose that we have definite values for these observations (that is, our observed signals are 'instantiated'), and suppose that those values are $x_1, ..., x_{100}$. The liklihood function here is

$$L(x_1, ..., x_n; S, \sigma^2) \ = \ \frac{1}{(2\pi)^{50}\sigma^{100}} \exp\left(-[(x_1 - S)^2 + \cdots + (x_{100} - S)^2]/2\sigma^2\right).$$

If we seek S which maximizes this (for any given value for σ^2), we can ignore the factor $(2\pi)^{50}\sigma^{100}$, and we maximize the log of the remaining exponential factor; thus, we just need to maximize

$$-\frac{(x_1 - S)^2 + \cdots + (x_{100} - S)^2}{2\sigma^2}.$$

We can ignore the σ^2; so, we maximize

$$-\frac{(x_1 - S)^2 + \cdots + (x_{100} - S)^2}{2}. \tag{4}$$

Taking a derivative with respect to S and setting equal to 0 we have

$$(x_1 + \cdots + x_{100}) - 100S \; = \; 0.$$

So,

$$S \; = \; \frac{x_1 + \cdots + x_{100}}{100},$$

which is our sample mean. The fact that the expression (4) is a down-turning parabola means that, indeed, this is a maximum.

Thus, we see that by averaging we obtain a maximum liklihood estimate for S, and therefore, in some sense, this is the best we could hope to do to recover S.