

A sampling theorem.

Suppose X_1, X_2, \dots, X_N are independent Bernoulli random variables with parameter p , which is **unknown**. N here is large and, possibly, unknown as well. As an example, you could think of each X_i as being one if an individual i in a population of N individuals will vote for Candidate A and zero otherwise. If it's a congressional election N would be around a half a million. Thus $100p\%$ of the population of voters will vote for Candidate A, and we want to know what p is. So why not ask them? Well, that's too difficult; we'll try to get away with asking n of them where n is determined by how much we want to pay the pollster. So set $S_n = \sum_{i=1}^n X_i$ and use

$$\frac{S_n}{n},$$

the **sample mean**, as an approximation to p . How good an approximation is it? Common sense tells you that the larger n is the better an approximation you get, but it does not tell us any more than that. We shall we will show that "you can be 95% confident that $\frac{S_n}{n}$ is within $.98/\sqrt{n}$ of p ".

Let Z be standard normal. We begin by solving

$$P(|Z| \leq c) = 2\Phi(c) - 1 = .95$$

for c and find that

$$c \approx 1.96;$$

we further note that

$$\frac{c}{2} \approx .98.$$

Since $\sqrt{pq} \leq 1/2$ we find that

$$\left\{ \left| \frac{S_n - np}{\sqrt{npq}} \right| \leq c \right\} = \left\{ \left| \frac{S_n}{n} - p \right| \leq \frac{c\sqrt{pq}}{\sqrt{n}} \right\} \subset \left\{ \left| \frac{S_n}{n} - p \right| \leq \frac{c}{2\sqrt{n}} \right\} \approx \subset \left\{ \left| \frac{S_n}{n} - p \right| \leq \frac{.98}{\sqrt{n}} \right\}.$$

Thus, by the central limit theorem,

$$P\left(\left| \frac{S_n}{n} - p \right| \leq \frac{.98}{\sqrt{n}} \right) \approx \geq .95$$

this last \approx doesn't mean much if n is not large enough for the central limit theorem to provide a good approximation. That's not a big deal. What *is* a big deal is whether you have sampled *independently* or not. That is a another very interesting subject.