Lecture 18 **Math 20 Fall 2014, Dartmouth College**

1 Confidence Interval

Suppose that a polling agency wants to conduct a survey to estimate what proportion p of a population supports a reform. They can't ask eeryone in the population, so they would like to estimate the number n of people that is sufficient to estimate p to within 3% with 95% probability.

A randomly asked person answers "yes" with probability p. Let the proportion of people that answer "yes" be $\bar{p}$. What is the distribution of $\bar{p}$, if they ask n people?

The number of people that answer "yes" is a random variable S_n, the sum of n Bernoulli trials with probability p. The proportion $\bar{p} = \frac{S_n}{n}$ can be approximated by a normal random variable with mean p and variance $\frac{pq}{n}$.

What is $\bar{p}^*$, the standardized version of $\bar{p}$?

$$\bar{p}^* = \frac{\bar{p} - p}{\sqrt{pq/n}}$$

A standard normal variable is within two standard deviations of its mean with probability about 95%. What are the cutoff values of $\bar{p}$ that correspond to two standard deviations from the mean?

The standard deviation of $\bar{p}$ is $\sqrt{pq/n}$. So the cutoff values are $p - 2\sqrt{pq/n},\ p + 2\sqrt{pq/n}$. We have:

$$P(p - 2\sqrt{pq/n} \leq \bar{p} \leq p + 2\sqrt{pq/n}) \simeq 95\%.$$

What is the value n that gives the desired confidence interval of 3%, for given p? Can you find a bound that will work for any p?

$$2\sqrt{pq/n} \leq 0.03$$
$$\frac{1}{\sqrt{n}} \leq \frac{0.03}{2\sqrt{pq}}$$

But since $pq \leq 1/4$, we have:

$$\frac{1}{\sqrt{n}} \leq 0.03$$
$$n \geq\sim 1111$$

Notice anything strange about this result?

Where does the total size of the population come into this? Well... it doesn't. Most surveys ask just over 1000 people no matter how big a population they are trying to gauge.

2 DNA

Suppose that the height of an individual of the same gender is mostly determined by genetic factors, and that there are n different types of genes that can affect height. Assume that each of them can take form of k different pairs of alleles, with equal probability. Is this enough to say that the height of an individual is approximately normally distributed?

To use the Central Limit Theorem we also need them to be independent, and we need the height to depend on the sum of the alleles. So we would also assume that the genes work cumulatively.

It was observed these heights are not only normally distributed across a population, but the variance is the same from one generation to the next. Can you come up with an explanation for that? (Hint: Use a simplified model in which any two individuals in a population are equally likely to be the parents, and the population is large enough so that you can treat the different gene types as independent. Still, is this enough?)

The starting distribution of the alleles might be different, but for any nest generation it will be the same.

3 LLN

Prove the Law of Large Numbers using the Central Limit Theorem.

Let S_n be the sum of n identically distributed, numerically-valued random variables $X_1, \ldots, X_n$, each with mean μ and standard deviation σ. Then $\frac{S_n}{n}$ has mean μ and standard deviation $\sigma/\sqrt{n}$. We want to prove that for any ϵ,

$$\lim_{n\to\infty} P(|\frac{S_n}{n} - \mu| \geq \epsilon) = 0$$

By Central Limit Theorem, we know that

$$P(|\frac{S_n}{n} - \mu| \geq \epsilon) \simeq 1 - \int_a^b e^{-x^2/2} dx,$$

where $a = \frac{\mu-\epsilon}{\sigma/\sqrt{n}} = \sqrt{n}\frac{\mu-\epsilon}{\sigma}$ and $b = \frac{\mu+\epsilon}{\sigma/\sqrt{n}} = \sqrt{n}\frac{\mu+\epsilon}{\sigma}$. Notice that this is the probability that the variable is NOT between $\mu - \epsilon$ and $\mu + \epsilon$. Now, since a, b grow further apart as $n \to \infty$,

$$\lim_{n\to\infty} \int_a^b e^{-x^2/2} dx = 1,$$

and so:

$$\lim_{n\to\infty} \left(1 - \int_a^b e^{-x^2/2} dx\right) = 0 = \lim_{n\to\infty} P(|\frac{S_n}{n} - \mu| \geq \epsilon)$$

□