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?

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

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?

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?

Notice anything strange about this result?

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?

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?)

3 LLN

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