Sampling Distributions

Montserrat Farell

Universitat Autònoma de Barcelona

March 27, 2008

Contents

General Information

Prof. M. Farell

Office: B3-154

Tel. 581-2932

Web page

There is a class web page which will serve as the distribution point for the class notes and materials. To access it, point your favorite browser at ***pareto.uab.es/mfarell/statistics***. The Facultat has a computer room in Aules 23-24-25 where you can find plenty of computers connected to the internet.

1 Introduction to Statistical Inference

Statistical inference is the set of tools used to obtain guesses on indicators that sumarize an event in the real world. Statistical inference is concerned with the generalizations about the *population* on the basis of information provided by a *sample*.

1.1 Basic Concepts of Statistical Inference

Definitions:

Population: is the totality of all possible observations on measurements or outcomes. Examples, incomes of all people in a certain country in a specific period of time, all outcomes of a given experiment, "tossing a coin". Populations can be infinite or finite. *Finite*: All the possible observations is less than infinity.

Sample: it is a set of measurements or outcomes selected from the population.

Random Sample: in finite populations, every individual in the population has the same chance of being chosen. In infinite populations the sample is random if the observations are independent.

Parameters: are numerical characteristics of the population.

Statistics: are numerical characteristics of the sample.

Random Variable: a variable that has a probability distribution.

Estimator: formula that describes a procedure of guessing the value of a given population parameter, usually using the observations in the sample.

Estimate: a realization of the formula (estimator).

Hypothesis Testing: assumption/s about values on parameter/s.

Test Statistic: is the summary of the evidence provided by the observations in the sample, used to arrive to a verdict concerning a hypothesis.

1.2 Statistical Inference: intuition

Our objective, in general, will be obtaining information about a population. Parameters are unknown numbers that give us a lot of information about the population, we are interested on giving "the best " approximation to that number (*the parameter*).

Suppose we are interested on the value of a parameter θ (is an unknown value) from a population. We will use a formula to obtain a numerical approximation to θ, let's call it $\hat{\theta}$, this formula will be a function of the sample; $\hat{\theta} = f(observations\ sample)$. As we said, we want to give "the best" approximation to θ, therefore we are interested on analyzing how "well" this formula called $\hat{\theta}$works. When using this formula, how many times do we get close to the true value θ? Answering this question is finding the characteristics, "properties", of the distribution of probability of $\hat{\theta}$, that is called its *sampling distribution*.

2 Experimental Derivation of Sampling Distributions

Sampling Distributions can be derived experimentally or theoretically. Here we will derive sampling distributions experimentally, using *Monte Carlo experiments.* To see how to derive them theroetically see Kmenta (Part One, Basic Statistical Theory, by Jan Kmenta *Elements of Econometrics,* MacMillan). Sampling distributions depend on the sample size n, i.e., an estimator, say $\hat{\theta}$,has a diferent sampling distribution for different $n's$.

2.1 Sampling Distribution of Sample Proportion of Successes

Let's consider the proportion of successes (non-smokers) in the sample, as an estimator of the proportion of successes in the population. We could derive the sampling distribution of this estimator experimentally by repeating the extraction of samples. The experiment can be thought of a container of marbles with two different colors, in this population 70% red and 30% white. A success, a non-smoker, is represented by

$n = 10$. We will generate data that represents n extractions of that uniform, then we calculate the sample mean, and this we will do m times.

We will do it using the computer software GRETL, below you have an example of instructions to generate the data we are talking about:

- *nulldata n*
- *loop m–progressive*
- *genr var1=uniform()*10 (generates a continuous uniform in the interval [0,10]*
- *genr var2=(var1>1)+(var1>2)+...(do the same for all numbers)..+(var1>9) (drops the decimals and keeps the integers)*
- *genr var3=mean(var2) (keeps in a m vector all the calculated sample means)*
- *store var4.gdt*
- *endloop*

Then we reproduce the histogram of the m values of $\bar{x}$ and that is the experimental aproximation to the sampling distribution. When m goes to ∞, we obtain experimentally the theoretical sampling distribution. Theoretical sampling distributions can be obtained using statistical theorems and propositions, for example for this case see slide of pg. 101 of Kmenta, where a theoretical sampling distribution of $\bar{x}$ dor $n = 2$ is derived. In the last slide you can see the comparetive values of the two, experimental and theoretical, probability distributions for that case, for $n = 5$ and $n = 10$. You should be able to understand and interpret all those slides. *Observe that in order to ilustrate results experimentally you always control, know, all the paremeters you are interested on, and then you study how "well" some estimators behave, if you didn't know the true population model and parameter values you would not be able to analize the behaviour of the estimators, this is very important to understand.*

2.3 Experimental illustration of the Central Limit Theorem (CLT):

General statement of the CLT: If x has any distribution of probability with mean μ_x and variance σ_x^2 then the distribution of

$$\hat{\theta} = \frac{(\bar{x} - \mu_{\bar{x}})}{\sigma_{\bar{x}}} \quad \sim \quad N(0,1)$$

i. e., approaches the standard normal, and $\bar{x}$ is the sample mean. To illustrate this theorem experimentally we should design an experiment that generates the sampling distribution of $\hat{\theta}$for n small and n big, with m the same for both $n's$ and quite big say 10.000, recall that when m goes to ∞ the experimental distribution approaches the theoretical.

GOOD LUCK
