Basic Statistical Concepts: Definitions and Notation

A *population* is the group of all objects (or subjects) of interest. In a statistical study, the population is defined by the researcher. For example, a population might be defined as all women between 18 and 34 in Michigan, if a researcher was interested in studying that group of subjects. A *sample* is a subset of the population. For example, a sample from this population could be 100 randomly selected women between 18 and 34 in Michigan. While it is populations that we are ultimately interested in knowing about, we usually do not observe them directly. Instead, samples are what we actually observe and measure in statistical studies. *Statistical inference* is the process of drawing conclusions about a population, based on a sample taken from that population.

Why is randomization important? Only *random* samples can provide the basis for statistical inference, because only random samples are representative of the population as a whole.

It should be clear from the foregoing discussion that we will be dealing with characteristics of populations and characteristics of samples from those populations in this class. Therefore, it will be important to differentiate carefully between them. We do that by using different terms (when possible) and different notation (always) when referring to populations and samples.

A *parameter* is a characteristic of a population. A *sample statistic* is a characteristic of a sample. For example, the mean of a population is a parameter, while the mean of a sample is a sample statistic. Both are averages, but they are measured on two different groups, and so they are two different things. Characteristics measured on samples *estimate* the corresponding characteristics of the populations those samples came from. *Sample statistics*, in other words, are *estimates* of their corresponding *population parameters*.

The values of sample statistics almost always differ slightly from the values of the corresponding parameters in the underlying population due to *random variation*, also called *random error*. So, even a representative, random sample is likely to differ slightly from the underlying population. Suppose you had a population with a mean of 10. Further suppose that you took 5 different random samples from it. In the first sample, just randomly, a few more low values might be chosen. So the mean of the sample would come out a little lower than the population mean, like 9.4. But then in the second sample, a few higher values might be chosen instead. So then the second sample would come out with a mean on the high side, like 10.2. And so it would go with the other three samples – each one chosen randomly, each slightly different from the other samples, and from the underlying population they all came from. In the process of statistical inference, it will be our task to distinguish between this type of random variation, due to random chance, and true variation that can give us information about the value of a population parameter. We will do this by using a technique called hypothesis testing.

Notation is a part of life when learning statistics, and you should approach it like learning a language. The symbols are shorthand ways to refer to important concepts. You will need to become familiar with the following notation:

|  |  |  |  |
| --- | --- | --- | --- |
| Population Parameters | | Sample Statistics | |
|  | The mean of a population  The Greek letter “mu”, pronounced “mew” |  | The mean of a sample  Pronounced “x bar” |
|  | The standard deviation of a population  The lowercase Greek letter “sigma” |  | The standard deviation of a sample |
|  | The variance of a population  “Sigma squared” |  | The variance of a sample |
|  | A proportion of a population |  | A proportion of a sample  Pronounced “p bar” |
|  | The slope coefficient for a population  The Greek letter “beta” |  | The slope coefficient for a sample |

|  |  |
| --- | --- |
| More Notation | |
|  | The null hypothesis  Pronounced “H nought” or “H O” |
|  | The alternative hypothesis  Pronounced “H A” for short |
|  | A significance level  The Greek letter “alpha” |
|  | A sample size |
| p P | a probability |
|  | The summation operator  The uppercase Greek letter “sigma”  Read this as “The sum of all….” |
|  | Read this as “the change in…”  The Greek letter “delta” |
|  | A variable with a subscript  A subscript denotes one of several variables  The , pronounced “ sub ” or “" for short  For example: is the 2nd in a set of variables and could be called “ sub 2” or “ 2” |
|  | A variable with a hat  This example is pronounced “y hat” |
|  | Random error  The Greek letter “epsilon” |
|  | greater than |
|  | less than |
|  | greater than or equal to |
|  | less than or equal to |
|  | not equal to |