Sampling Distribution

Bruce A. Desmarais

Stats I Lab

November 2, 2009

The Sampling Distribution of a Statistic

If a statistic ($T(\cdot)$), defined on a sample of data $\mathbf{x}$ has non-degenerate support, it will not be constant over repeated trials with equivalent sampling conditions.

1. In order to understand whether a statistic $T(\mathbf{x})$ is unusually high, low or mediocre, we need a measure of the *Probability* that $T(\mathbf{X}) = T(\mathbf{x})$ under certain conditions for the data generating process.
2. For instance, suppose we posit conditions on $\mathbf{X}$ that constitute a Null Hypothesis. A common method of defining hypothesis tests is to derive the distribution of $T(\mathbf{x})$ under the condition implied by the Null Hypothesis. $T(\mathbf{x})$ can then be compared to this distribution to evaluate whether it is unusual given the null condition (i.e. would the value of $T(\mathbf{x})$ persuade you to *reject* the null condition?).

Classical Parametric Sampling Distributions

There are three steps in defining a sampling distribution of $T(\mathbf{X})$ to be used in hypothesis testing.

1. Define the null distribution for $\mathbf{X}$ ($f_N(\mathbf{X})$).
2. Derive the distribution ($f_{TN}(T(\mathbf{X}))$) of $T(\mathbf{X})$ given that $\mathbf{X}$ follows the distribution $f_N(\mathbf{X})$.
3. Derive the tail probabilities of $f_{TN}(T(\mathbf{X}))$ so that $T(\mathbf{x})$ can be assessed for rarity given the null condition.

Example: The T-Statistic

Suppose we want to test the hypothesis that the *population* mean of X is different from zero using an i.i.d. sample of size n from the population. A reasonable choice of test statistic (because it involves only sufficient statistics for the parameters) is

$$T(\mathbf{x}) = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i}{S}$$

.

1. A justifiable null distribution for X is a normal (or Gaussian) distribution with 0 mean and unknown standard deviation.
2. It can be shown that if the null is true the distribution of $T(\mathbf{x})$ is Student's T with $n-1$ degrees of freedom (λ).
3. The two-tail p-value is derived as

$$p = 1 - \int_{-|T(\mathbf{x})|}^{|T(\mathbf{x})|} f_T(y,\lambda)dy$$

Exercise

Suppose a casino has the following table game. A six-sided die with the numbers 1, 2, 3 each on two sides is rolled on each trial and the player gets -1, 0, or 1 dollars if 1, 2, or 3 are rolled respectively. The casino claims that the die is balanced. In other words, the value rolled follows a discrete uniform distribution over 3 categories.

You seek to test whether the die is biased. Specifically you want to know whether the expected roll is different from 2.

Imagine you are able to passively observe the table for three rolls and you see (1, 1, 2) before security escorts you out because they know what you are up to. Do you believe the die is unbiased?