STA 623 – Fall 2010 – Dr. Charnigo

Section 5.1: Basic Concepts of Random Samples

Independent and identically distributed. We say that $X_1, \ldots, X_n$ are independent and identically distributed ("iid") if their joint probability density (mass) function has the form

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = f(x_1) \times \cdots \times f(x_n),$$

where f is the marginal probability density (mass) function common to each of $X_1, \ldots, X_n$.

Example (independent and identically distributed). Let $X_1, \ldots, X_n$ be iid with common marginal probability density function $f(x) := \theta x^{\theta-1} 1_{\{0<x<1\}}$ for some $\theta \in (0, \infty)$. Fix $b \in (0, 1)$. What is the probability that all of $X_1, \ldots, X_n$ exceed b?

What is the probability that none of $X_1, \ldots, X_n$ exceed b?

What is the probability that exactly one of $X_1, \ldots, X_n$ exceeds b?

What is the probability that exactly m $(0 < m < n)$ of $X_1, \ldots, X_n$ exceed b?

Interpretation as a simple random sample. Consider an effectively infinite population whose individuals are described by some quantitative characteristic.

Suppose there exists a continuous function f such that, for any real number x, the proportion of individuals in the population with characteristic less than or equal to x is given by $\int_{-\infty}^{x} f(t)\, dt$.

Suppose also that we randomly select n individuals from the population, in such a way that any two groups of n individuals have the same probability of being selected. In particular, every individual in the population has the same probability of being selected. This is called a simple random sample.

Let the random variables $X_1, \ldots, X_n$ denote the characteristics of the n individuals in the simple random sample. Then $X_1, \ldots, X_n$ will be iid with common marginal probability density function f. In particular, X_1 will have marginal probability density function f. Therefore, statements about proportions in the whole population can be cast as statements about probabilities for a single individual randomly selected from that population.

We often employ this interpretation in methods courses. For example, whenever we say that a population is normal, what we mean is that there exist numbers $\mu \in \mathbb{R}, \sigma \in (0, \infty)$ such that $\int_{-\infty}^{x} (2\pi\sigma^2)^{-1/2} \exp[-(t-\mu)^2/(2\sigma^2)]\, dt$ gives the proportion of individuals in the population with characteristic less than or equal to x. However, we usually represent a normal population to undergraduates as one for which the random variables $X_1, \ldots, X_n$ arising from a simple random sample are normally distributed. Then we draw a picture of a bell curve to describe the probabilistic behavior of X_1.

Example (interpretation as a simple random sample). We can also draw a simple random sample from a population that is not large enough to be deemed effectively infinite. However, the above interpretation fails in that instance.

First, the probabilistic behavior of X_1 may be more appropriately described by a discrete distribution than by a continuous distribution.

Second, and perhaps more importantly, $X_1, \ldots, X_n$ will not be independent.

To appreciate these points, suppose that we have a population consisting of four individuals whose characteristics are 2, 3, 4, and 7 respectively. If we draw a simple random sample of size two, what is the probabilistic behavior of X_1?

What is $P(X_2 = 2|X_1 = 2)$? How does this compare to $P(X_2 = 2|X_1 = 3)$?

Another point that deserves mention is that simple random sampling is not as... simple... as every individual in the population having the same probability of being selected, even though that is implied by simple random sampling.

Continuing from the previous example, can you exhibit a scheme for selecting a sample of size two for which every individual in the population has the same probability of being selected but different pairs of individuals have different probabilities of being selected?