CLARK UNIVERSITY 1887

Order statistics
Math 217 Probability and Statistics
Prof. D. Joyce, Fall 2014

A random sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ of size n consists of n independent and identically distributed random variables. We'll call the distribution that the random variables come from the *population distribution* and denote the density for the population distribution f_X.

Any random variable which is a function of these is called a *sample statistic* of the random sample. We've looked at the most important one, the sample mean $\overline{X} = \frac{X_1+X_2+\cdots+X_n}{n}$. We'll also look at the sample variance. Today we'll look at the sample median and the order statistics.

Order statistics. The minimum value among the variables $X_1, X_2, \ldots, X_n$ is also called the *first order statistic* and denoted $X_{(1)}$, and the maximum value among them is also called the n^{th} *order statistic*, denoted $X_{(n)}$. There are also intermediate order statistics which result from ordering the values $X_1, X_2, \ldots, X_n$ from smallest to largest

$$X_{(1)} \le X_{(2)} \le \cdots \le X_{(n-1)} \le X_{(n)}$$

These order statistics form for basis for nonparametric statistical inference since some of their properties do not depend on the family of distributions that the random comes from. For parametric statistical inference, there is a assumption that the distribution comes from a certain family. In many cases, such an assumption can be justified since there's an underlying Bernoulli process, Poisson process, or the sample is large and the Central Limit Theorem can usually be invoked. But in other cases, it's not clear what form the distribution might take, and in that case nonparametric methods may be the best way to go.

As described in the text, you can determine the distribution of the r^{th} order statistic $X_{(r)}$ when the population distribution is continuous and its formula is

$$\begin{aligned} f_{(r)}(x) &= \frac{n!}{(r-1)!(n-r)!}\left(\int_{-\infty}^{x} f_X(t)\,dt\right)^{r-1} f_X(x)\left(\int_{x}^{\infty} f_X(t)\,dt\right)^{n-r} \\ &= \frac{n!}{(r-1)!(n-r)!}(F_X(x))^{r-1} f_X(x)(1-F_X(x))^{n-r} \end{aligned}$$

The sample median. If n is an odd number, $n = 2r+1$, then there is a middle order statistic $X_{(r)}$. It's index is $(n+1)/2$, and that order statistic is called the *sample median*. (When n is even, there won't be a single middle order statistic, but it's not uncommon to average the two middle order statistics and call that the median.)

When the population distribution is continuous, then its c.d.f. F_X takes on the value $\frac{1}{2}$, and where it takes on that value, that is, where $F_X(\tilde{\mu}) = \frac{1}{2}$, is called the *median* of the population distribution.

We'll prove later on the Central Limit Theorem that says that as n approaches ∞, the distribution of the sample mean $\overline{X}$ for a population distribution with a finite mean and variance approaches a normal distribution whose mean is μ, the mean of population distribution. That is, the sample mean is asymptotically normal with mean μ.

There's also a limit theorem for the sample median, but we won't prove it. It says that for a continuous population distribution, the sample median is asymptotically normal with mean $\tilde{\mu}$, the median of the population distribution.

Quartiles and percentiles. The *first sample quartile* is the r^{th} order statistic $X_{(r)}$ where $r = n/4$, and the *third sample quartile* is the r^{th} order statistic $X_{(r)}$ where $r = 3n/4$

The k^{th} sample percentile is the r^{th} order statistic $X_{(r)}$ where $r = kn/100$. Thus, the 50^{th} sample percentile is the sample median, the 25^{th} sample percentile is the first sample quartile, and the 75^{th} sample percentile is the third sample quartile.

Quartile and percentile statistics enjoy limit theorems analogous to that for medians.

Math 217 Home Page at `http://math.clarku.edu/~djoyce/ma217/`