CLARK UNIVERSITY

Nonparametric statistics
Math 218, Mathematical Statistics
D Joyce, Spring 2016

Order statistics. The minimum value among the variables $X_1, X_2, \ldots, X_n$ in a random sample $\mathbf{X}$ 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

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

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.

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.)

There's 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 $r = \frac{1}{4}n$, and the *third sample quartile* is the r^{th} order statistic $X_{(r)}$ where $r = \frac{3}{4}n$

The k^{th} sample percentile is the r^{th} order statistic $X_{(r)}$ where $r = \frac{k}{100}n$. 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.

The sign test. The sign test is used to test the hypothesis H_0 that the median $\tilde{\mu}$ of a population is a particular value $\tilde{\mu}_0$ versus the alternate hypothesis H_1 that $\tilde{\mu} > \tilde{\mu}_0$. You can imagine that if you take a sample and find a great many of the sample values lie on above μ_0, then you should reject H_0. The sign test codifies that intuition.

Let p denote the unknown probability that $X_i > \tilde{\mu}_0$. Then the number of trials S_+ in the sample $\mathbf{X}$ greater than $\tilde{\mu}_0$ follows a binomial distribution $\textsc{Binomial}(n, p)$, that is

$$P(S_+ > s_+) = \sum_{i=s_+}^{n} = \binom{n}{i} p^i (1-p)^{n-i}.$$

When H_0 holds, then $p = \frac{1}{2}$, and conversely. That allows us to compute the P-value for a sign test since it's just

$$P = \sum_{i=s_+}^{n} = \binom{n}{i} \left(\frac{1}{2}\right)^n.$$

You can use table A.1 in the text to find the P-value for $n \leq 20$. Reject H_0 at the α-level if s_+, the number of observations greater than

$$\tilde{\mu}_0 \geq b_{n,\alpha}$$

where s_+ is the number of observations greater than $\tilde{\mu}_0$.

For larger n you can use the normal approximation with a continuity correction to the binomial distribution. Reject H_0 if

$$z = \frac{x_+ - n/2 - 1/2}{\sqrt{n/4}} \geq z_\alpha,$$

or equivalently if

$$s_+ \geq \tfrac{1}{2}(n + 1 + z_\alpha \sqrt{n}).$$

The other one-sided sign test for H_1 being $\tilde{\mu} < \tilde{\mu}_0$ is similar but uses s_- instead of s_+.

The two-sided sign test for H_1 being $\tilde{\mu} \neq \tilde{\mu}_0$ uses $s_{\max}$ which is the maximum of s_+ and s_-. The criteria for rejection are the same except α is replaced by $\alpha/2$.

The confidence intervals correspond to the hypothesis tests, as usual.

Math 218 Home Page at
`http://math.clarku.edu/~djoyce/ma218/`