Chapter 14 Nonparametric Statistics

A.K.A. "distribution-free" statistics! Does not depend on the population fitting any particular type of distribution (e.g, normal). Since these methods make fewer assumptions, they apply more broadly... at the expense of a less powerful test (needing more observations to draw a conclusion with the same certainty).

Let's think about the median $\tilde{\mu}$. Given a sample $x_1, \ldots, x_n$ drawn randomly from an unknown *continuous* distribution, say we want to test:

$$\begin{aligned} H_0 &: \tilde{\mu} = \tilde{\mu}_0 \\ H_1 &: \tilde{\mu} > \tilde{\mu}_0 \end{aligned}$$

For example, test whether the median household income exceeds 25K.

Sign Test

Step 1 Count the number of x_i's that exceed $\tilde{\mu}_0$. Call this s_+. Let $s_- = n = s_+$.

Step 2 Reject H_0 if s_+ is too large (or if s_- is too small).

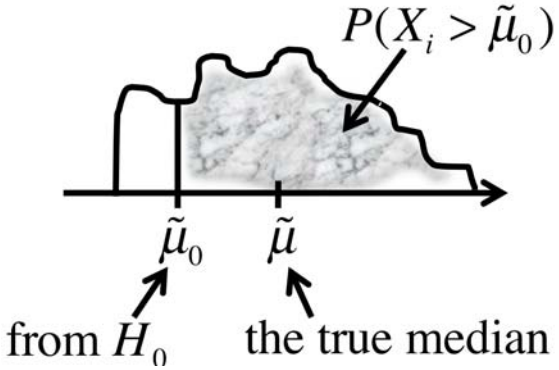
Why does this make sense? What if the true median $\tilde{\mu}$ is 1000 and $\tilde{\mu}_0$ is 1?

How large should s_+ be in order to reject? To find out, we need to know the distribution of the r.v. for s_+. Call that r.v. S_+.

Let

$$p = P(X_i > \tilde{\mu}_0) \text{ and } 1 - p = P(X_i < \tilde{\mu}_0).$$

Here's a helpful picture. Note that the distribution of the population isn't normal!



If you think of:

$$Y_i = \begin{cases} 1 \text{ if } X_i > \tilde{\mu}_0 \\ 0 \text{ otherwise} \end{cases}$$

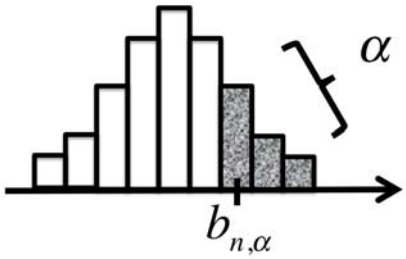
as a Bernoulli r.v. with parameter p, then S_+ is a sum of the Y_i's. So S_+ is a sum of Bernoulli's. So it's binomial!

$$S_+ \sim Bin(n, p) \text{ and } S_- \sim Bin(n, 1-p). \tag{1}$$

Now, if H_0 is true, $\tilde{\mu}_0$ is the true median and $p = 1/2$, so:

$$S_+ \sim Bin(n, 1/2) \text{ and } S_- \sim Bin(n, 1/2). \tag{2}$$

So reject when $s_+ \geq b_{n,\alpha}$, where $b_{n,\alpha}$ is the upper α critical point for $Bin(n, 1/2)$.



That is, $\alpha = \sum_{i=b_{n,\alpha}}^{n} \binom{n}{i} \left(\frac{1}{2}\right)^n$.

(Or reject when $s_- \leq b_{n,1-\alpha}$.)

Let's calculate the pvalue using the binomial distribution:

$$\begin{aligned} \text{pvalue} &= P(S_+ \geq s_+) = \sum_{i=s_+}^{n} \binom{n}{i} \left(\frac{1}{2}\right)^n \\ &\overset{(*)}{=} P(S_- \leq s_-) = \sum_{i=0}^{s_-} \binom{n}{i} \left(\frac{1}{2}\right)^n. \end{aligned}$$

The step with the (*) is from symmetry of $Bin(n, 1/2)$.

As usual, reject if pvalue $< \alpha$.

(Also if n is large, the binomial distribution can be replaced with the normal distribution and we could use a z-test.)

Example

Can you see now why we needed the assumption of a *continuous* r.v.?

(Think about p under the null hypothesis.)

Also we could rewrite the hypotheses:

$$\begin{aligned} H_0 &: p = 1/2 \\ H_1 &: p > 1/2. \end{aligned}$$

Summary of Sign Test:

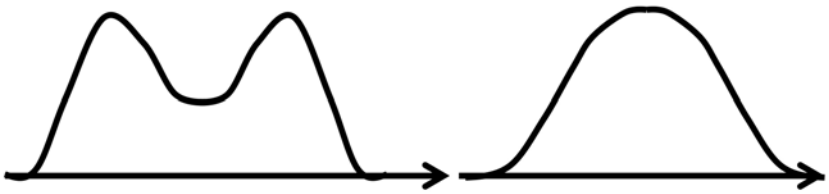
Data & Assumptions: $X_1, \ldots, X_n$ ∼unknown continuous distribution, no other assumptions!

Test Statistic: $S_+ =$ number of observations X_i that exceed $\tilde{\mu}_0$ (or $s_- = n - s_+$).

Hypotheses	Reject when	pvalue
$H_0 : \tilde{\mu} \leq \tilde{\mu}_0$ $H_1 : \tilde{\mu} > \tilde{\mu}_0$	$s_+ \geq b_{n,\alpha}$	$P(S_+ \geq s_+) = \sum_{i=s_+}^{n} \binom{n}{i} \left(\frac{1}{2}\right)^n$
$H_0 : \tilde{\mu} \geq \tilde{\mu}_0$ $H_1 : \tilde{\mu} < \tilde{\mu}_0$	$s_- \geq b_{n,\alpha}$	$P(S_- \geq s_-) = \sum_{i=s_-}^{n} \binom{n}{i} \left(\frac{1}{2}\right)^n$
$H_0 : \tilde{\mu} = \tilde{\mu}_0$ $H_1 : \tilde{\mu} \neq \tilde{\mu}_0$	$s_{\max} \geq b_{n,\alpha}$ where $s_{\max} := \max(s_+, s_-)$	$2 \sum_{i=s_{\max}}^{n} \binom{n}{i} \left(\frac{1}{2}\right)^n$

Wilcoxon Signed Rank Test

Let us add an assumption in order to gain more power from the test. Namely, the *assumption that the distribution is symmetric.*



Symmetric means that reflection around the median yields the same thing. (The sign test did not require this... remember, generally more assumptions means more conclusions.)

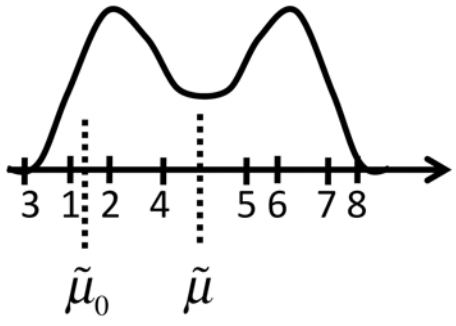
The Wilcoxon Signed Rank Test looks at magnitudes

$$d_i = X_i - \tilde{\mu}_0.$$

Also assume no ties: $d_i = 0$ for any i, and no absolute ties $|d_j| = |d_j|$ for any i, j.

$$\begin{aligned} H_0 &: \tilde{\mu} = \tilde{\mu}_0 \\ H_1 &: \tilde{\mu} > \tilde{\mu}_0. \end{aligned}$$

Step 1 Rank the $|d_i|$'s. Let r_i be the rank of $|d_i|$. Here, $r_i = 1$ for the smallest $|d_i|$.



Step 2 Let

$$\begin{aligned} w_+ &= \text{sum of ranks of the positive } d_i\text{'s} \\ w_- &= \text{sum of ranks of the negative } d_i\text{'s.} \end{aligned}$$

$$\left(\text{So, } w_+ + w_- = r_1 + r_2 + \cdots + r_n = 1 + 2 + \cdots + n = \frac{n(n+1)}{2}.\right)$$

Step 3 Reject H_0 if w_+ is too large (or if w_- is to small.)

Example

How large to reject? Our r.v. is W_+ which is a sum of ranks. We've never seen W_+'s distribution before, but tail probabilities for it are in Appendix A10 on page 683.

As an aside: To make the distribution of W_+, take all 2^n possible assignments of signs to the ranks of $|d_i|$'s:

$$\begin{array}{rcccccccc} i & = & 1 & 2 & 3 & 4 & \cdots & n & \\ \text{possible assignments} & = & 2 \times 2 \times 2 \times 2 & & & & \cdots & 2 & = & 2^n \end{array}$$

(Each assignment gets a + or − so there are 2 possibilities of signs for each rank.) For each assignment, calculate w_+. Since assignments are equally likely, we get a distribution over w_+ values.

It can be shown that W_+ and W_- have the same distribution. So call $W = W_+ = W_-$. Then we can use the table to get the pvalues:

$$\text{pvalue } = P(W \geq w_+) = P(W \leq w_-).$$

Reject H_0 if pvalue $\leq \alpha$ or if $w_+ \geq w_{n,\alpha}$.

(For large n, can approximate null distribution of W by a normal distribution.)

Summary of Wilcoxon Signed Rank Test:

Data & Assumptions: $X_1, \ldots, X_n \sim$unknown symmetric distribution

Test Statistic: $w_+ =$ sum of ranks of positive d_i's where $d_i = x_i - \tilde{\mu}_0$.

Hypotheses	Reject when	pvalue
$H_0 : \tilde{\mu} \leq \tilde{\mu}_0$ $H_1 : \tilde{\mu} > \tilde{\mu}_0$	$w_+ \geq w_{n,\alpha}$	$P(W \geq w_+)$
$H_0 : \tilde{\mu} \geq \tilde{\mu}_0$ $H_1 : \tilde{\mu} < \tilde{\mu}_0$	$w_- \geq w_{n,\alpha}$	$P(W \geq w_-)$
$H_0 : \tilde{\mu} = \tilde{\mu}_0$ $H_1 : \tilde{\mu} = \tilde{\mu}_0$	$w_{\max} = \max(w_+, w_-) \geq w_{n,\alpha}$	$2P(W \geq w_{\max})$

Example continued

Why do we need the assumption of a symmetric distribution?

Important**** There are many cases in which H_0 is rejected by the Wilcoxon Signed Rank Test but not the Sign Test

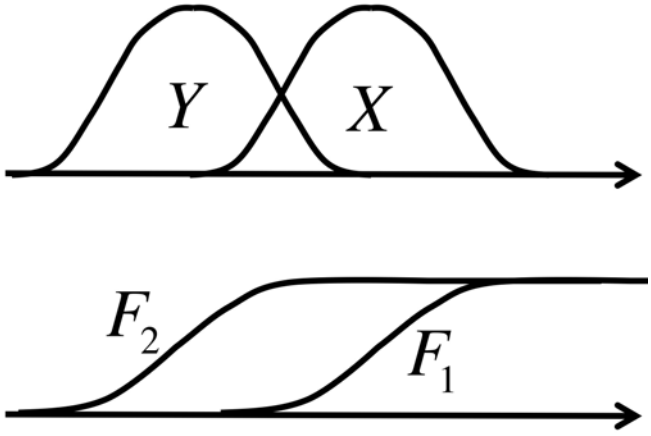
Inferences for Two Independent Samples (Rank Sum Test and Mann-Whitney U Test

We want to know whether observations from one population (given sample $x_1, \ldots, x_{n_1}$) tend to be larger than those from another population (given $y_1, \ldots, y_{n_2}$).

Mouse Data Example

Let's make precise X "larger than" Y.

Given r.v.'s X and Y with cdf's F_1 and F_2,



X is *stochastically larger* than Y (denoted $X \succ Y$) if for all real numbers u,

$$\begin{aligned} F_1(u) &\leq F_2(u), \\ \text{in other words } P(X \leq u) &\leq P(Y \leq u), \end{aligned}$$

with strict inequality for at least one u. Denote $F_1 < F_2$ to mean $X \succ Y$.

Let us test:

$$\begin{aligned} H_0 &: F_1 = F_2 \\ H_1 &: F_1 < F_2 \end{aligned}$$

Wilcoxon-Mann-Whitney U Test and Wilcoxon Rank Sum Test (2 equivalent tests)

Wilcoxon Rank Sum

Step 1 Rank all $N = n_1 + n_2$ observations in ascending order (assume no ties)

Step 2 Sum the ranks of the $x's$ and $y's$ separately. Denote sums by w_1 and w_2.

Step 3 Reject H_0 if w_1 is large (or equivalently if w_2 is small).

Example

To do testing, we need the distribution of W_1 (the random variable for w_1) or W_2 under H_0 (soon).

Mann-Whitney U

Step 1 Compare each x_i with each y_j

Step 2 Let u_1 be the number of pairs in which $x_i > y_j$. Let u_2 be the number of pairs in which $x_i < y_j$.

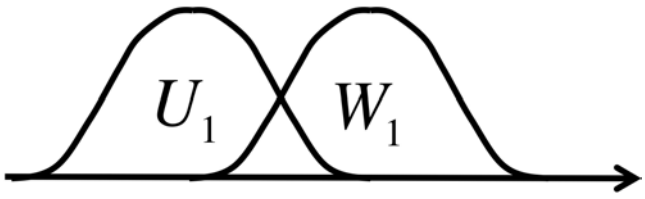
Step 3 Reject H_0 if u_1 is large (or equivalently if u_2 is small).

It is true that

$$u_1 = w_1 - \frac{n_1(n_1+1)}{2} \text{ and } u_2 = w_2 - \frac{n_2(n_2+1)}{2}.$$

Demo of this fact

Since u_1 and w_1 are just a constant apart, the distributions of u_1 (r.v. U_1) and w_1 (r.v. W_1) have the same shape:



The distribution of U_1 turns out to be symmetric about $(n_1 n_2)/2$ and in fact, U_2 has the same distribution as U_1. Tail probabilities for this distribution are in Table A.11. So we define $U := U_1 = U_2$.

So given $x_1, \ldots, x_{n_1}$, $y_1, \ldots, y_{n_2}$, to test:

$$\begin{aligned} H_0 &: F_1 = F_2 \\ H_1 &: F_1 < F_2 \end{aligned}$$

Steps 1 and 2 Compute u_1= number of pairs in which $x_i > y_i$.

$$\text{or } u_1 = w_1 - \frac{n_1(n_1+1)}{2}$$

where remember that w_1 is the sum of ranks of the x_i's.

Step 3 Reject H_0 when $u_1 \geq u_{n_1,n_2,\alpha}$ (using the table) or compute:

$$\text{pvalue } = P(U \geq u_1) = P(U \leq u_2), \text{ reject if it's less than } \alpha.$$

(If n_1 and n_2 are large, we can approximate the distribution of U under H_0 by a normal distribution.)

Example

15.075J / ESD.07J Statistical Thinking and Data Analysis

Fall 2011