STAT 240 Homework 3

Rebecca Barter, Andrew Do, and Kellie Ottoboni

March 13, 2015

1) Show that a permutation test based on $\bar{X}$ and a permutation test based on t are equivalent when $m = n$

Note that the t statistic is defined by

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{Var(X)}{n} + \frac{Var(Y)}{m}}}$$

and that a permutation test based on t involves

1. Fill a box with the observed data.
2. Draw a simple random sample of size n and call if X, call the remaining elements in the box Y.
3. Compute t as you would if X and Y were your original data. Call this $t^{*(1)}$.
4. Repeat steps 1-3 L times to get $t^{*(2)}, t^{*(3)}, ..., t^{*(L)}$.
5. The distribution of the $t^{*(\ell)}$ approximates the true probability distribution of t under the strong null. In particular, a (left-tail) p-value can be computed as

$$\frac{1}{L}\#\{t^{*(\ell)} \leq t\}$$

We thus need to show that we can write

$$t^* = \frac{\bar{X}^* - \bar{Y}^*}{\sqrt{\frac{Var(X^*)}{n} + \frac{Var(Y^*)}{n}}}$$

in terms of $\bar{X}^*$ only.

We note, however, that if we simply write $A = \sum_{i=1}^n X_i + \sum_{i=1}^n Y_i = \sum_{i=1}^n X_i^* + \sum_{i=1}^n Y_i^*$ to the the sum of all observations (which can be considered our new population from which we are drawing), we have that

$$\bar{X}^* - \bar{Y}^* = \left(1 + \frac{n}{n}\right)\bar{X}^* - \frac{1}{n}A = 2\bar{X}^* - \frac{1}{n}A$$

Thus the RHS depends only on $\bar{X}^*$. Next, if we write $B = \sum_{i=1}^n X_i^2 + \sum_{i=1}^n Y_i^2 = \sum_{i=1}^n X_i^{*2} + \sum_{i=1}^n Y_i^{*2}$, then

$$\begin{aligned}
\frac{Var(X^*)}{n} + \frac{Var(Y^*)}{n} &= \frac{1}{n}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^{*2} - \frac{1}{n^2}\left(\sum_{i=1}^{n} X_i^*\right)^2\right] + \frac{1}{n}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i^{*2} - \frac{1}{n^2}\left(\sum_{i=1}^{n} Y_i^*\right)^2\right] \\
&= \left[\frac{1}{n^2}\sum_{i=1}^{n} X_i^{*2} + \frac{1}{n^2}\sum_{i=1}^{n} Y_i^{*2}\right] - \frac{1}{n}\bar{X}^{*2} - \frac{1}{n^3}\left[A - n\bar{X}^*\right]^2 \\
&= \frac{1}{n^4}\left[n^2\sum_{i=1}^{n} X_i^{*2} + n^2\left(B - \sum_{i=1}^{n} X_i^{*2}\right)\right] - \frac{1}{n}\bar{X}^{*2} - \frac{1}{n^3}\left[A - n\bar{X}^*\right]^2 \\
&= \frac{B}{n^2} - \frac{1}{n}\bar{X}^{*2} - \frac{1}{n^3}\left[A - n\bar{X}^*\right]^2
\end{aligned}$$

which depends only on $\bar{X}^*$. Thus we have shown that we can write t^* in terms of $\bar{X}^*$ only, implying that t and $\bar{X}$ are equivalent test statistics for a permutation test when $m = n$.

2) Construct a hypothetical dataset (with at least 3 data points in treatment and at least 3 in control) for which the p-value of a permutation test based on $\bar{X}$ is smaller than the p-value of a permutation test based on t. Try to make the difference substantial.

Suppose that the treatment group X contains only 5 observations, drawn from a Gaussian distribution with mean 0 and standard deviation 20. In this example, the generated sample is

$$X = \{10.9934, -16.8321, 0.6600, 10.4830, -34.5521\}$$

Let the control group Y contains 100 standard normal observations. A situation like this might occur if a treatment is very expensive to administer and has variable effects.

Suppose we want to test the strong null hypothesis using a two-sided test. The observed difference in means is $\bar{X} - \bar{Y} = -5.7513$ and the observed t-statistic is -0.6560. Using 10000 simulations, the permutation test p-values for the two-sided test are

$$\begin{aligned}
P(|\bar{X} - \bar{Y}| \geq 5.7513) &= 0.05 \\
P(|t| \geq 0.6560) &= 0.559
\end{aligned}$$

The p-value of the permutation test based on $\bar{X}$ is smaller than the p-value of the permutation test based on t because of the extreme noise and small sample size in the treatment group.

3) Construct a hypothetical dataset (with at least 3 data points in treatment and at least 3 in control) for which the p-value of a permutation test based on t is smaller than the p-value of a permutation test based on $\bar{X}$. Try to make the difference substantial.

The treatment group consists of 10 i.i.d. observations drawn from a normal distribution with mean 0 and unit standard deviation. The control group are 1000 i.i.d observations drawn from a normal distribution with mean 1 and standard deviation 100. The exact values we used can be found in the `hw3.Rdata` file

turned in with this assignment. The treatment and control group data are stored in the objects `tr_3` and `ctrl_3`, respectively.

The observed different in means is $\bar{X}-\bar{Y} = 6.4$, and the observed t-statistic is 2.00. Using 10000 simulations, the permutation test p-values for the two-sided test are:

$$P(|\bar{X} - \bar{Y}| \geq 6.4) = 0.84$$
$$P(|t| \geq 2.0) = 0.07$$

The p-value of the permutation test based on $\bar{X}$ is larger than its t-statistic counterpart due to the large amount of noise coupled with a large sample size in the control group.

4) Construct a hypothetical dataset (with at least 3 data points in treatment and at least 3 in control) for which the p-value of a permutation test based on $\bar{X}$ is smaller than the p-value of a standard t test. Try to make the difference substantial.

Let the treatment group consist of the observations

$$X = \{-10, -9, \ldots, 9, 10, 500, 1000, 2000, 5000\}$$

and the control group consist of the numbers 0 through 5, each appearing 5 times. Suppose that in the treatment group, the large observations are highly unusual. If we were to exclude these, the two groups would have the same means. We would like our test to be robust to these outliers and tell us that the difference in means between the treatment and control groups is not significantly different from 0 under the strong null. We will compare the strong null to the alternative hypothesis that the difference in means is greater than 0.

The observed difference in means is $\bar{X} - \bar{Y} = 337.5$ and the observed t-statistic is 1.5805 ($df = 53$). Using 10000 simulations, the permutation test p-value for the mean is

$$P(|\bar{X} - \bar{Y}| \geq 337.5) = 0.0394$$

The t-test p-value is

$$P(|t_{53}| \geq 1.5805) = 0.1200$$

The p-value of the permutation test based on $\bar{X}$ is smaller than the p-value of the usual two-sample t-test based because of the extreme outliers in the treatment group. When the data is highly skewed, the normality assumptions needed for the t-test are violated.