RNR / ENTO 613. Alternatives to the t-tools

There are occasions when the assumptions of the t-tests are not met:

a) Outliers are present (especially if sample size is small). T-tools are not resistant.
b) Unequal sample sizes and long-tailed or skewed distributions
c) Censored data, i.e. for observations only known to be greater than, or lower than, some number (e.g. the cognitive load experiment).
d) Data with many ties (e.g. the O-Ring failure study).

Two alternatives exist when assumptions are not met.

1- For two-sample comparisons, permutation (randomization) tests are distribution free, not sensitive to outliers, and ties in data pose no problem. Because permutation tests involve swapping observations between groups, they can only be used to test hypothesis involving two or more groups (not applicable to conduct a test equivalent to a one-sample t-test). Commercial software is available; the programs do many things…. (e.g. t-tests, ANOVA, linear and multiple regression).

2- Other "distribution free" statistical tools exist. Many of those are based on rank transformation instead of on values of observations. Rank-sum tests are not sensitive to outliers and are appropriate to deal with censored observations. Too many tied values can cause a problem; ranks-sum tests *are not robust* to violation of the assumption of homogeneity of variance. Other methods based on different models are also available.

Confidence intervals are not computed by most statistical packages for these types of tests, but can be obtained by trial and error.

Permutation tests

Context: 2-sample or paired t-tests when assumptions are not met

A permutation test directly generates the distribution of a statistic by randomly and repeatedly drawing samples from the same data set, under the assumption that the null hypothesis is correct. When used to analyze randomized experiments, permutation tests are called randomization tests.

The logic of permutation tests is very similar to the logic underlying t-tests.

Logic of 2 sample t-test:

a) If the null hypothesis is correct ($\mu 1 = \mu 2$), the t-ratio calculated from the difference between the average of 2 samples (i.e. $\overline{Y}_2 - \overline{Y}_1$) should be centered on 0. The expected distribution of this t-ratio is described by a Student's t-distribution on $n_1 + n_2 - 2$ degrees of freedom.

b) An extreme value for t_{obs} therefore leads to rejection of the null hypothesis.

Logic of Permutation test:

a) If the null hypothesis is correct ($\mu 1 = \mu 2$), then the two samples originated from the same population. The observations are therefore our best guess about the composition of that population. Multiple randomization of observations between groups and calculation of $\overline{Y}_2 - \overline{Y}_1$ directly generates the expected distribution of differences between sample averages under the null hypothesis. That distribution of differences is always centered on 0 (with a null of no difference).

b) The *observed* difference between the sample averages is only likely to be close to zero if the null hypothesis is correct. An extreme value for the *observed* $\overline{Y}_2 - \overline{Y}_1$ leads to rejection of the null hypothesis.

<<Fig. Permutation-1>>

The choice of measuring the difference between sample averages or computing a t-statistic for than difference is arbitrary.

Example:

Compare mandible lengths for males and female golden jackals (an extinct species) in the collection of the British Museum. Since contemporary male jackals are in general bigger than females, a one-tailed comparison is used.

Data (length in millimeters):

Males: 120, 107, 110, 116, 114, 111, 113, 117, 114, 112
Females: 110, 111, 107, 108, 110, 105, 107, 106, 111, 111

One-tailed 2-sample t-test:

Assumptions of the test:

1) Random sampling from the population: questionable since museum specimens collected in unknown manner.
2) Sample independence: questionable also. We don't know whether the jackals originate from a single population; may have cluster effects in the data set.
3) Equal SD for male and female populations: appear reasonable from comparison of samples
4) Normal distribution for observations within groups: sample size too small to seriously check.

(but t-tools are robust to violation of assumption 3 and 4 if sample size is equal)

<<Fig. JMP>>

$\overline{Y}_m$ = 113.4 mm, s_m = 3.72 mm, n = 10

$\overline{Y}_f$ = 108.6 mm, s_f = 3.08 mm, n = 10

s_p = 3.08

t_{obs} = 3.484 with 18 *df* .

The one-tailed probability of such a value is 0.0013 if the null hypothesis is correct. It is concluded that males were likely larger than females, or more formally that there was an association between sex and size in this sample (no random sampling = no inference to populations).

Permutation test:

Assumptions of the test:

1) Random sampling: To make sure samples are representative of their respective populations.
2) Independence: The null / alternative hypothesis (samples from a single / two different populations) does not make sense if samples come from many populations (which could generate cluster effects).

Procedure:

1. Find the observed mean length for males and females, and the difference D_o between these.
2. Randomly reallocate 10 of the observations to a "male" group and the rest of the observations to a "female" group. Calculate the difference between the means obtained (= Di).
3. Repeat step 2 many times (e.g. 4999 times) to generate the distribution of differences expected if the null hypothesis is correct, i.e. the randomization distribution.
4. If D_o is a typical value from the randomization distribution, do not reject Ho. Formally calculate the proportions of all the observed Di values that are greater or equal to Do, incorporating the observed D_o in the numerator and denominator (since D_o is just one of the values in the randomization distribution).

Results:

Do = 113.4 – 108.6 = 4.8 mm

4999 randomizations are conducted. Adding D_o we have 5000 D values. Only 9 of these Ds are greater or equal to 4.8 mm. D_o is therefore an unlikely value (9/5000 = 0.0018), which provides strong evidence against the null hypothesis. We reject the null and conclude that that males probably had longer mandible length than females (p-value = 0.0018).

<<Fig. Permutation-2>>

Note on permutation tests:

If you perform two or more tests on the same data set, you will likely obtain a slightly different p-value. Conducting many permutations for a single test (e.g. 9999) contribute in "stabilizing" the p-value.

Nonparametric tests

Rank transformation

Many nonparametric tests assess whether the distributions of populations (or factor levels) are centered at the same location, often by using a rank transformation.

A *rank transformation* replaces each observation by its rank in a *combined* sample. Because the ranks are then separated by the same amount of units, a rank transformation eliminates many (but not all) of the concerns about the distribution of populations.

Importantly, tests based on ranked data are a *resistant* alternative to t-tests. They are more conservative when data are normal, but do much better when there are outliers or for long-tailed distributions.

Steps to rank data:
1. List all observations from both samples in increasing order.
2. Identify the sample from which each observation came.
3. Assign the rank order as a sequence of numbers from 1 to n1 + n2.
4. Tied observations (duplicated values) receive the average rank of the tied group.

   E.g. if the 13^{th} and 14^{th} observations have the same value (say Y = 55), they are both given their average rank (13 + 14)/2 = 13.5.

Plants treated with water (group 1) or a fertilizer (group 2)

Group	Y (height)	Rank
1	47.8	1
1	48.2	2
1	49.1	3
1	49.9	4
1	51.4	5
1	52	6
2	52.3	7
1	52.6	8
2	53.2	9
1	54.6	10
2	54.8	11
1	55.2	12
2	55.6	13
2	57.4	14
2	58	15
1	58.3	16

2	59.8	17
2	61.3	18

Note:

a) Transformation to ranks conserves the order of data but eliminates the importance of the population distribution.
b) Ranks are insensitive to outliers (the tallest plant above would receive rank 18 even if it measured 223 cm).

Context: 2 independent samples, equal variance but distribution assumption not met

The non-parametric equivalent to the two-sample t-test is the *Rank-Sum test* (also called Wilcoxon or Mann-Whitney U-test). The test statistic is calculated from rank scores.

The test addresses the null hypothesis:

Randomized experiment: no treatment effect

Observational study: distributions of two groups are the same

Plant example:

The *test statistic*, T, is *the sum of ranks in one group*. (often the smallest group for convenience)

The *expected value of T* under the null hypothesis of no treatment effect is:

Mean (T) = $n_1 \bar{R}$ ($\bar{R}$ is the average of the ranks).

Logic: If the null hypothesis is correct, the groups of size n1 and n2 are sampled from the same population. Thus, the expected *average* rank in group 1 (and 2) is $\bar{R}$, the average of the combined set of (n1 + n2) ranks. Thus, the expected distribution of the *sum* of the ranks (i.e. T) in group n_1 should be centered on $n_1 \times \bar{R}$.

$$SD(T) = s_R \sqrt{\frac{n_1 n_2}{(n_1 + n_2)}}$$ (s_R is the sample standard deviation of the ranks).

Note here that by estimating s_R, we assume that the 2 populations have *equal standard deviation*.

<<Fig. 4.6 in Sleuth>>

Conversion to ranks homogenizes the group distribution and generally takes care of difference in *shape of the distribution* of the populations (including presence of outliers): this is what is meant when it is said that rank-sum tests are *distribution free*.

If the spread (standard deviation) of the ranks of the groups is similar, the distribution of T under the null hypothesis can be approximated by a normal approximation, if size of samples is *greater than 5* and there are *not too many ties*.

Therefore, the p-value for the *observed* T is obtained with a Z-statistic:

$$Z - stat = \frac{T - mean(T)}{SD(T)}$$

The null hypothesis is rejected for the alternative if Zobs > Zcrit (obtained with appropriate α from a Z-table).

If the standard deviation of the groups is not similar (i.e. much more ties in one group than in the other), the assumptions of the rank-sum test *are not met* [s_R does not estimate the standard deviations of the samples, so SD(T) is not accurate].

<<Fig. 4.6 in Sleuth>>

Steps for the Rank-Sum test for the plant example:

1. n1 =10, n2 = 8, $\overline{R}$ = 9.50, s_R = 5.33, T= 67

2..Mean (T) = (10) (9.5) = 95. SD (T) = $5.33\sqrt{\frac{10x8}{10+8}}$ = 11.237

3. Z = (67 – 95) / 11.237 = - 2.49

4. For Z = -2.49, 1-tailed p-value = 0.0064, 2-tailed p-value = 0.0128.

We conclude that addition of a fertilizer likely increased plant growth (one-sided p-value =0.0064, from the Wilcoxon rank-sum test). The median height of plants in the control group was 51.7 cm, compared to 56.5 cm in the fertilizer group.

Context: 2 independent samples, equal variance assumption not met

When the equal variance assumption is not met but populations are *approximately normal*, a Welch's t-test can be conducted (see Sleuth, p. 93). The t-statistic and CI's for this test are computed exactly as for the two-sample t-test, except for calculation of the standard error of the differences between averages.

Context: Paired samples, normality assumption not met

An alternative to paired t-test when the difference between paired observations are not normal is the *Wilcoxon Signed-Rank Test.* This test uses ranked-transformed differences between pairs (see Sleuth p. 96-97).

Note on Equal-spread Models

Parametric statistics compare populations with similar *shape* (i.e. normal) and *spread* (i.e. equal standard deviation). In such case the difference in means provides a simple summary of differences between populations.

With the Welch t-test or permutations for comparing 2 normal populations with unequal variance, a difference between means may be more difficult to interpret. When the populations have *different* standard deviations, the difference between means may not always provide a sufficient summary of the data.

<<Fig. 4.11 in Sleuth>>