RNR / ENTO 613. Assumptions for Two-sample t-tools

The assumptions on which t-tests are based are *never* met exactly. This is not necessarily a problem because conclusions obtained from t-tests are often valid even if the assumptions are not strictly met. Alternatives exist if violation of assumptions occurs. Data transformation is one such alternative.

Statistical models for 2 *independent* samples are based on 4 assumptions, in decreasing order of importance:

1. Random sampling
2 Independence of observations within and between groups
3 Homogeneity of variance (homoscedasticity)
4 Normality

Departure from these assumptions may lead to misleading inferences.

Robustness of t-tools

> **Robustness:** A statistical procedure is *robust* to departures from a particular assumption if it is *valid* even when the assumptions are not met exactly. Valid means that the confidence interval and p-values arising from the test remain equal to the stated rate.

1. Random sampling

Statistical inferences and CI estimation can be biased if subjects are not selected randomly from populations.

Random sampling of units is done to ensure that the sample is representative of the population. Inferences about populations are not possible without random sampling of subjects (as opposed to self-selected subjects or arbitrary selection).

Example: You may underestimate how fast rabbits can run if you only study sick individuals that are easy to observe!

There are no statistical procedures that are *robust* to violations of the random sampling assumption.

2. Sample independence

Lack of independence occurs *in* a group when the value of an observation (say above the mean) allows to guess the value of another observation (also above the mean).

Lack of independence occurs *between* groups when the value of an observation in one group allows prediction of the value of an observation in the other group (e.g. Schizophrenia in twins; paired observations).

Lack of independence commonly takes 2 forms:

> **Cluster effects:** data are collected in subgroups, where individuals in the subgroups are more similar in their response to a treatment than other individuals in the group. Ex: subgroups arise within samples as a result of common environment or genetic effects (litters).
>
> **Serial effects:** measurements taken closer in time or space tend to be more similar than measurements taken farther apart in time or space. These are respectively designated as *serial* and *spatial* autocorrelation.

<< Fig. 15.4 in Sleuth>>

Lack of independence leads to incorrect estimation of the *standard error* of the difference between group means, and to incorrect estimates of the t-statistic: t-tools are *not robust* to violations of the independence assumption.

3. Homogeneity of variance

T-tools are not robust when the SD of 2 populations are unequal *and* sample sizes are different. Then the pooled estimate of standard deviation does not estimate any real population parameter, and SE ($\overline{Y}_2 - \overline{Y}_1$) is not appropriate.

When sample sizes are *approximately equal*, the t-tools are robust to violations of equal variance assumptions.

<<Fig. 3.5 in Sleuth>>

4. Normality

In theory, the *larger* the sample size, the more *robust* are t-tools to violations in the normality assumption (central limit theorem). In practice it is hard to tell how large is large…Two general rules are useful:

a) if 2 sample distributions have approximately equal variance *and* shape:

   i) if sample sizes are approximately *equal*, robustness of t-tools is affected moderately by long-tailed distributions.

      <<Fig. 3.4 in Sleuth>>

ii) If sample sizes are *unequal*, robustness is affected moderately by long-tailed distributions and *substantially* by skewness.

b) if skewness of 2 distributions differs considerably (e.g. in opposite direction), t-tools can be misleading, especially for moderate sample sizes.

Resistance of t-tools

Resistant: A statistical procedure is *resistant* if its conclusions are not sensitive to a change in a small part of the data. Specifically, a resistant statistical procedure is not affected much by *outliers.*

Practical strategies for analysis of Two-Sample Data

To meet the assumptions underlying the use of two-sample t-tool:

1. Consider Serial and Cluster effects

To detect: Plot data to detect presence of subgroups, or non-random patterns of measurements as a function of time or distance. Review the ways in which data were gathered.

<<Fig. 15.4 from Sleuth>>

To remedy:

a) Choose a good sampling design to keep track of possible lack of independence.

b) If observations are not independent, statistical models must reflect such lack of independence:

For example, cluster effects within samples could be addressed with a *nested* ANOVA or *randomized block* experiment.

Lack of independence between groups could be addressed with paired t-test or *repeated measure designs*.

Tools are available to analyze autocorrelated data (Chap 15 in Sleuth).

2. Evaluate Homogeneity of variance and Normality

To Detect: Graphically with side-by side box plots.

To Remedy: Try a data transformation and reassess; if not successful, use techniques that are more resistant and robust, such as the non-parametric rank-sum procedure (also called Wilcoxon or Mann-Whitney tests), permutation / randomization tests (programs are available to do this: e.g. Manly's program), or the Welch t-test for comparing two normal populations with unequal variance.

There are tests available for assessing homogeneity of variance (e.g. Levene's test) and normality (e.g. Shapiro-Wilk W-test). Many of these tests are sensitive to small departure from normality or homogeneity of variance that would have little consequence on robustness of t-tools. For that reason they are not recommended.

<<Fig. 3.2 in Sleuth>>

3. Assess Effect of Outliers

1. Perform statistical analysis both with and without the outliers.
2. If statistical conclusion is not qualitatively different (i.e. does not change much), leave the outliers in the data set.
3. If statistical conclusion does change, then:
   a) Investigate outliers for possible recording error.
   b) If it is found that outliers are from another population than the experimental ones, then remove them.
   c) Otherwise, use a resistant tool (e.g. non-parametric rank-sum test) or report results from both analyses.

<< Fig. 3.6 in Sleuth>>

Data transformations

A transformation of original data can improve how well the assumptions are met. Often the main goal of a transformation is to establish a scale where the 2 groups have the same spread.

Log transformation

The logarithmic (log) transformation is very useful for positive data. The common logarithm, defined as $\log_{10}(10^x) = x$ can be used. Commonly the natural log is used (e = 2.72828…): $\ln(e) = 1$; $\ln(1) = 0$; $\ln(e^x) = x$.

The log transformation spreads out small values and draws in larger ones. It is indicated *when the ratio of largest to smallest measurement in a group is > 10,* or when groups are skewed to the right and the mean is positively correlated to the variance.

<<Fig. 3.7 Sleuth>>

The logarithmic transformation expresses a relation that is *multiplicative* on the scale of measurement as an *additive* relation on a logarithmic scale. For example, the Richter scale for measuring magnitude of earthquake, and pH values that measure acidity, are logarithms.

<<Fig on Growth process>>

Interpretation after Log transformation

For randomized Experiments.

In a randomized experiment with an additive treatment effect, we assume that *every subject* exposed to a treatment show the same response denoted δ. We thus have Y* = Y + δ for subjects exposed to the treatment and Y* = Y for subjects in the control group.

If a treatment effect is *multiplicative*, a log transformation makes that treatment effect additive on a log scale:

A multiplicative treatment (e^{δ}) applied to the treatment group yields $Y^* = Y e^{\delta}$. For the control group, δ = 0 and thus we obtain: Y* = Y.

<u>On a log scale</u>, the same treatment produces log Y* = log (Y) + δ. For the control group, δ = 0 and we obtain log Y*= log (Y).

To test whether there is a treatment effect on the log scale (i.e. δ ≠ 0), one performs the usual t-test, where the null hypothesis is:

> Ho: mean [$\log(Y_2)$] – mean [$\log(Y_1)$] = 0 = δ

To describe the treatment effect <u>on the original scale</u>, we *back transform* the estimate of δ and the endpoint of its confidence interval.

> *Interpretation*: A subjects' response to treatment 2 is $\exp(\bar{Z}_2 - \bar{Z}_1)$ times as large (i.e. e^{δ}) as the subjects response for treatment 1 (where $\bar{Z}_i$ =average $\log(Y_i)$).

For the Cloud Seeding example, the difference between the treatment and control group (log-transformed observations) was 1.14 (SE = 0.449), with a 95% CI = 0.241, 2.047. Back-transformation yields $e^{1.14}$ = 3.138, and the 95% CI is $e^{0.241}$ = 1.27 to $e^{2.047}$ = 7.74.

> *Interpretation*: Volume of rainfall was 3.14 times larger in seeded than unseeded clouds (95% CI: 1.27 to 7.74)

For Observational Studies.

Inferences about population means still refer to multiplicative changes (see Sleuth p. 70-71 for logic behind this).

Interpretation of observational studies is:

The median for population 2 is estimated to be exp ($\overline{Z}_2 - \overline{Z}_1$) times as large as the median for population 1.

For the Sex Discrimination example, the average difference between the log transformed male and female salaries is 0.147. Because $e^{0.147} = 1.16$, it is estimated that the median salary for males is 1.16 times as large (16 % greater) than the median salary for females. The 95% CI on the log scale is 0.100 to 0.194. Because $e^{0.100} = 1.11$ and $e^{0.194} = 1.21$, the median salary for males is estimated with 95% confidence to be between 11% and 21% greater than the median salary for females.

Other transformations for positive measurements

Other useful transformations can be applied to improve assumptions underlying the use of t-tools. Tests based on these transformations are generally approximate on the original scale with a CI approach. This is because the back-transformed averages used to calculate a t-ratio do not exactly correspond to the original untransformed averages.

This does not represent a problem because:

a) The difference between the transformed and back-transformed averages is usually small.
b) If you were to repeat the same experiment again, it would be unlikely that you would obtain exactly the same *estimate* for the difference between the 2 population means anyway (but each CI computed from the 2 experiments would have the same probability of capturing the true difference between the population means).

So in practice, to get an idea of the difference between group means when transformations other than the log are used, we look at each back transformed means and their associated back transformed confidence limits (not at the back transformed difference between means).

1) Use transformed data to calculate an average and associated 95% CI
2) Back transform average and limits of the 95% CI

In the end you get 2 back transformed averages and their associated back transformed CI. These can be used to assess the difference between groups.

Some common problems and useful transformations.

Problem	Severity / Nature	Transformation
Positive skew (tail to the right)	Severe (waiting times, times to failure)	1/y
	Moderate	log y
	Slight (count data)	$y^{½}$, $y^{1/3}$
Negative skew (tail to the left)	Severe	y^2
	Moderate	y^3
Unequal variance	Standard deviation ∝mean	log y
	Standard deviation ∝ square of the mean	1/y
	Standard deviation ∝ root of the mean	$y^{½}$, $y^{1/3}$
Proportions	Almost all cases	log[y/1-y] arcsin(√y)
Large range of y's within groups		Log y

a) **For log transformation, use log (X + 1) if data include zeros**
b) **For SQRT transformation, use SQRT (X + 0.5) if data include zeros**

See Sokal and Rohlf (Biometry) or other textbooks for more discussions on transformations. Ramsey and Schafer suggest trial and error for choosing an appropriate transformation.