

## Discussion of *t*-Tool Assumptions

**One-Sample Case:** In the one-sample case, the formal assumptions for the use of *t*-tools require a simple random sample from a population with a normal distribution. The *t*-tools provide useful results in a much broader range of situations. Below are some of the main situations that can prevent the valid use of *t*-tools in the one-sample problem.

1. The sample size is small AND the population from which the sample is drawn is far from normal. What is meant by “small” and “far from normal” is intentionally vague. The more different the population distribution is from normal, the larger the sample size needs to be before the *t*-tools will produce reliable results.
2. The sample size is small AND there are one or more extreme outlying observations. What is meant by “small” and “extreme outlying observations” is intentionally vague. The more extreme the outlying values, the larger the sample size needs to be before the *t*-tools will produce reliable results.

One strategy for dealing with outliers is to analyze the data both with and without the outliers. Report the results including outlying values if the basic conclusions of the analysis are the same both with and without the outliers. If the conclusions depend on the outlying data points, it is important to make sure the outliers are not recording errors (this should always be checked) or observations from outside the population of interest. A data point should never be excluded unless it can be determined (using information other than the unusual numerical value of the data point) that the data point is an error. When data points are excluded from an analysis, the reason for their exclusion should be stated. Rather than removing values, it is usually better to report results both with and without outlying values or to use analysis methods that are resistant to outlying values.

3. The data is not a simple random sample from a population of interest. We often apply *t*-tools when the sample is not a simple random sample from a population of interest, but these results must always be interpreted with caution.
4. The sampled values are not independent of one another. Loosely speaking, two observations are independent if knowing that one observation is above (or below) average indicates nothing about whether the other observation will be above (or below average). Treating dependent observations as independent observations often leads to underestimated standard errors.

**Two-Sample Case:** In the two-sample case, the formal assumptions for the use of *t*-tools require independent simple random samples from each of two populations with normal distributions and common standard deviation. The *t*-tools provide useful results in a much broader range of situations. Below are some of the main situations that can prevent the valid use of *t*-tools in the two-sample problem.

1. One or more of the problems mentioned for the one-sample problem apply to one or both of the samples in the two-sample problem. Number 1 above can be particularly problematic if the shapes of the two population distributions are considerably different from one another.
2. The standard deviation for population 1 ( $\sigma_1$ ) is quite different from the standard deviation for population 2 ( $\sigma_2$ ) AND the size of the sample taken from population 1 ( $n_1$ ) is quite different than the size of the sample taken from population 2 ( $n_2$ ). The most problematic situation occurs when  $\sigma_1 > \sigma_2$  and  $n_1 < n_2$  or, equivalently, when  $\sigma_1 < \sigma_2$  and  $n_1 > n_2$ . This will lead to *p*-values that are too small and confidence intervals that have lower coverage probability than claimed, even if both  $n_1$  and  $n_2$  are large. There are methods for comparing the means of groups with different standard deviations, but it is usually better to seek a transformation of the data that makes  $\sigma_1$  and  $\sigma_2$  more similar.

Generally, if....	$s_p$ (pooled spread estimator)	confidence interval	p-value
$\sigma_1 = \sigma_2$ , any $n_1, n_2$	OK	OK	OK
$\sigma_1 \neq \sigma_2, n_1 = n_2$	OK	OK	OK
$\sigma_1 > \sigma_2, n_1 > n_2$ or $\sigma_2 > \sigma_1, n_2 > n_1$	typically too large	typically too large	typically too large
$\sigma_1 < \sigma_2, n_1 > n_2$ or $\sigma_2 < \sigma_1, n_2 > n_1$	typically too small	typically too small	typically too small

The previous points made are a summary of a detailed discussion of the validity of  $t$ -tools under various conditions presented in Chapter 3 of *The Statistical Sleuth* by Ramsey and Schafer. Here are some questions to help you better understand.

1. An experiment is conducted to determine if a pain-killing drug affects the length of time (in seconds) a rat can support its own weight. The amount of time each of three rats is able to support itself both with and without the drug is presented below.

- (a) Conduct a  $t$ -test to determine if the drug has an effect. Comment on the validity of the  $t$ -tools.

Rat	Drug	Placebo
1	51	48
2	15	11
3	16	14

$$\bar{Y} = \frac{3 + 4 + 2}{3} = 3 \quad s = \sqrt{\frac{0^2 + 1^2 + (-1)^2}{3 - 1}} = 1$$

- (b) Suppose that one additional rat is included in the experiment. The rat supports himself for 78 seconds while on the drug compared to 27 seconds while on the placebo. Repeat part (a) with this new rat added to the data set.

$$\bar{Y} = 15 \quad s = \sqrt{\frac{(-12)^2 + (-11)^2 + (-13)^2 + 36^2}{4 - 1}} = 24.014$$

- (c) Should the data from the new rat be excluded from the analysis?

2. Suppose a researcher wishes to estimate the mean weight of approximately 1000 hogs. The hogs are housed 150 pens with around 6 to 8 hogs per pen. The hogs in any given pen are all from a single litter (family). The researcher randomly picks one pen that contains 8 hogs. The average and standard deviation of the weights of the 8 hogs are 185 and 7 pounds, respectively. Use this data to compute a 95% confidence interval for the mean weight of all 1000 hogs. Comment on the validity of the interval.

3. Suppose you want to find a 95% confidence interval for the difference between the means of two roughly normal populations. State whether the interval will tend to be too narrow, about right, or too wide for each of the following situations:

- (a)  $\sigma_1 = 10, \sigma_2 = 20, n_1 = 50, n_2 = 100$
- (b)  $\sigma_1 = 10, \sigma_2 = 40, n_1 = 10, n_2 = 10$
- (c)  $\sigma_1 = 20, \sigma_2 = 20, n_1 = 10, n_2 = 100$