

## Stat 4473 - Data Analysis

### Robustness of t-methods

The t-methods for  $\mu$  and  $\mu_d$  are exactly correct when the population from which we sample is normally distributed. Practically speaking, there's no way to be certain this is true. (In fact, it's almost certainly not true, since the normal distribution is an idealized population model.) The usefulness of the t-methods is dependent on how strongly they are affected by lack of normality. Luckily, the t-methods for  $\mu$  and  $\mu_d$  are quite robust against non-normality of the population except when outliers or strong skewness are present.

Defn: A confidence interval or hypothesis test is called robust if the confidence level or p-value does not change much when the conditions for use of the procedure are violated.

We might say that before using the t-methods, we should check the "nearly normal assumption" that the data appear close to normal — symmetric, single peak, no outliers. We will use box plots, histograms and/or stem plots, and formal tests of normality to check the nearly normal assumption. Modified boxplots are often useful in spotting outliers.

#### Ways to not be normal

- More than one peak: The nearly normal assumption is violated if a histogram or stem plot of the data shows two or more peaks. When you see this, look for the possibility that your data come from two groups. If so, try to separate the data into its separate groups, then analyze each group separately.
- Skewness: If the box plot, histogram, or stem plot indicates the data are strongly skewed, you can't use the t-methods unless your sample is large.
- Outliers: The nearly normal assumption is also violated if the data have outliers. Outliers should be investigated. Sometimes, it's obvious that a data value is wrong and the justification for removing or correcting it is clear. When there's no clear justification for removing outliers, you might want to run the analysis both with and without the outliers and note any differences in your conclusions. Any time data values are set aside, you must report on them individually. An analysis of the non-outlying points, along with a separate discussion of the outliers, is often very informative and can reveal important aspects of the data. See more about handling outliers on the next page.

P.S. As tempting as it is to get rid of annoying values, you can't just throw away outliers and not discuss them. It isn't appropriate to lop off the highest or lowest values just to improve your results.

The importance of the normality assumption for the t-methods for  $\mu$  and  $\mu_d$  depends on the sample size. Unfortunately, it matters most when it's hardest to check. Here are some practical guidelines for using the t-methods to perform hypothesis tests and calculate confidence intervals for  $\mu$  and  $\mu_d$ .

Sample size less than 15: The data should appear close to normal (symmetric, single peak, no outliers). With so little data, it's rather hard to tell. But if you do find outliers or skewness, you should not use t.

Sample size at least 15: The t-methods can be used except in the presence of outliers or strong skewness. The t methods will work well as long as the data are single peaked and reasonably symmetric.

Large samples: The t-methods can be used even for clearly skewed distributions when the sample size is large, roughly  $n \geq 40$ . If you find outliers in the data, it's a good idea to perform the analysis twice, reporting the results with and without the outliers, even for large samples. See below.

Outliers: Note that the t-methods are not resistant to the effects of outliers (since the means and standard deviations on which they are based are not resistant to the effects of outliers). The nonparametric methods that serve as alternatives to the t-methods are based on ranks and ARE resistant to the effects of outliers.

One approach to handling outliers is to perform the analysis with and without the outlier(s), if the sample size is reasonable. Sometimes the outliers don't affect the analysis anyway. If the presence or absence of the outlier doesn't make any significant difference in the results of the analysis, then its presence is of no concern. If, on the other hand, the presence of the outlier does make a significant difference in the results of the analysis, then a statistical method that is resistant to outliers should be used.

The nonparametric alternatives to the t-methods are based on ranks and have dual usage: (1) as distribution-free alternatives to the t-methods when the normality assumption is violated, and (2) as resistant (i.e., resistant to the effect of outliers) alternatives to the t-methods when outliers affect the analysis (even if the normality assumption checks out).