

# ST 516: Foundations of Data Analytics

## Robustness

## Robustness

Assumptions

Evaluating Robustness

# Robustness

A statistical procedure is **robust** to departures from a particular assumption if it is valid even when that assumption is not met.

- Recall that by “valid” we mean that we can interpret a procedure the way it is intended—for example, a 95% confidence interval is valid if it really does have a 95% confidence level associated with it.
- In this lecture, we’ll talk about the underlying assumptions of the t-based procedures that you have been learning about, and how robust those procedures are.

# Assumptions of Inferential Procedures

First and foremost, the estimation and hypothesis testing procedures that you have learned out in this course rely on the assumption that the observations within samples are independent, and, in the case of two samples, that two samples themselves are independent.

- It's generally true that if independence is violated in anyway, the Normal-based and t-based inferential procedures that you've learned about are not valid—more on this in a subsequent lecture.

## Assumptions of Inferential Procedures

For independent samples, we have used the Central Limit Theorem to justify the use of Normal- and t-based inferential procedures.

- But the CLT relies on large sample sizes, and it's reasonable to ask how large is large enough and/or whether there are certain situations that may require especially large sample sizes.

A result from statistical theory ensures that *if* the populations from which the samples are drawn are Normal, then the one- and two-sample procedures based on the t-distribution are exact, regardless of sample size.

- This leads us to ask about the robustness of these t-based procedures against departures from Normality for small sample sizes.

# Evaluating Robustness

In the computer lab activity and in the homework for this week, you will use R to perform simulations to examine the robustness of

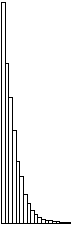


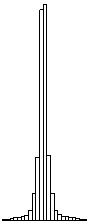

1. the t-test to small sample sizes and non-Normal population distributions
2. the t-based confidence intervals to small sample sizes and non-Normal population distributions
3. the t-based procedures to population distributions with extreme skewness

# Evaluating Robustness

Displays 3.4 and 3.5 in *The Statistical Sleuth* show results of simulations to evaluate the robustness of the t-based procedures.

- The simulations were based upon evaluating the *equal variance* t-based procedures, and we don't in general recommend using the equal variance procedure. *You can disregard Display 3.5.*
- In Display 3.4, you'll see that even for very small sample sizes, the t-based procedures are robust to departures from Normality (the "heavy-tailed" example appears to be not robust, but it's a fairly pathological example).
- Bottom line: the t-based procedures are quite robust—they are valid even for small sample sizes when the population distributions are not highly skewed, and even problems with skewed population distributions can be overcome by increasing the sample sizes.

**Percentage of 95% confidence intervals that are successful when the two populations are non-normal (but same shape and SD, and equal sample sizes); each percentage is based on 1,000 computer simulations**

	strongly skewed	moderately skewed	mildly skewed	long-tailed	short-tailed
$n_1, n_2$					
5	95.5	95.4	95.2	98.3	94.5
10	95.5	95.4	95.2	98.3	94.6
25	95.3	95.3	95.1	98.2	94.9
50	95.1	95.3	95.1	98.1	95.2
100	94.8	95.3	95.0	98.0	95.6