ST 516: Foundations of Data Analytics

Assessment and Remedies

How do we decide if assumptions are satisfied?

Assumptions are very rarely provably true. Our goal is to argue that they are at least reasonable.

There are four general approaches:

1. *Carefully thinking about the study design* We saw this in the previous lecture, suspicion of dependence often comes from scrutinizing the way the study was conducted.

2. *Using prior or external knowledge* Sometimes you'll be in a situation that is very similar to a previous study, or there is theory to predict what should happen. You can use this knowledge to argue for the reasonableness of the assumptions.

How do we decide if assumptions are satisfied?

3. *Relying on robustness* Some assumptions are less important in certain situations. For example, we generally don't even worry about the Normality assumption for large sample sizes due to the robustness of the t-based procedures to violations of this assumption.

4. *Examining the sampled data* for evidence of gross violations of the assumptions. But beware, deciding on the type of analysis to do based on examining the data can affect the validity of your analysis.

The problems with looking at data

We can never prove an assumption is satisfied by examining the data.

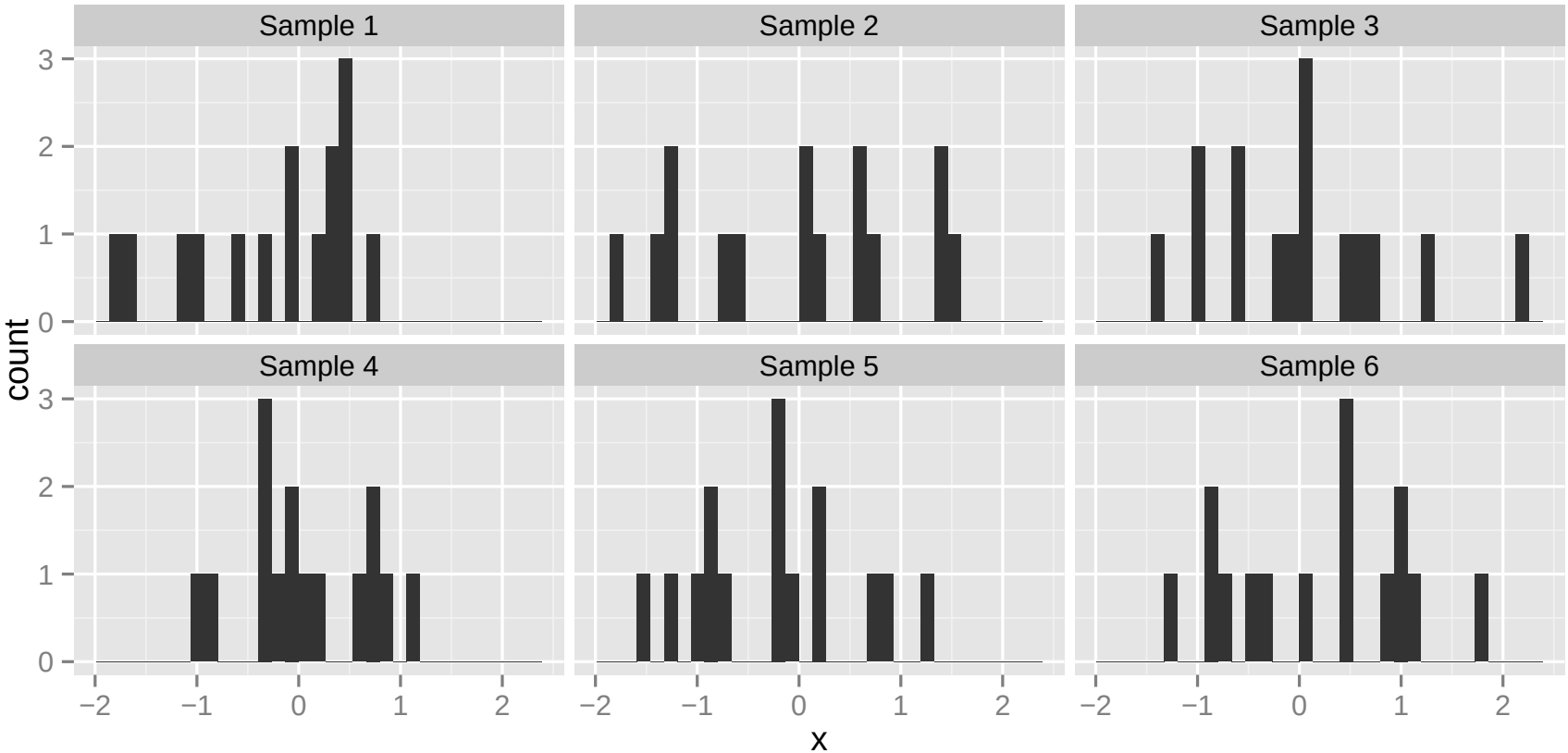
We can only check if the data seems consistent with the assumption.

Consequently, if people present a figure as support for the assumptions they will usually use language like:

- "*There appears to be no indication* of extreme skew"
- "*There is no evidence of a gross violation* of the . . . assumption"

Because we are only looking at a sample, we are easily fooled by sampling variation.

Which of these 6 samples came from a non-Normal population?



Which of these 6 samples came from a non-Normal population?

Trick question! None, they were all sampled from a Normal distribution.

Due to sampling variability, a sample can look skewed (like Sample 1), flat (like Sample 2) or clumpy even when sampled exactly from a Normal distribution.

The larger the sample size the easier it is to tell the difference between a sample from a Normal distribution and a sample from a non-Normal distribution, but the less it matters! (CLT)

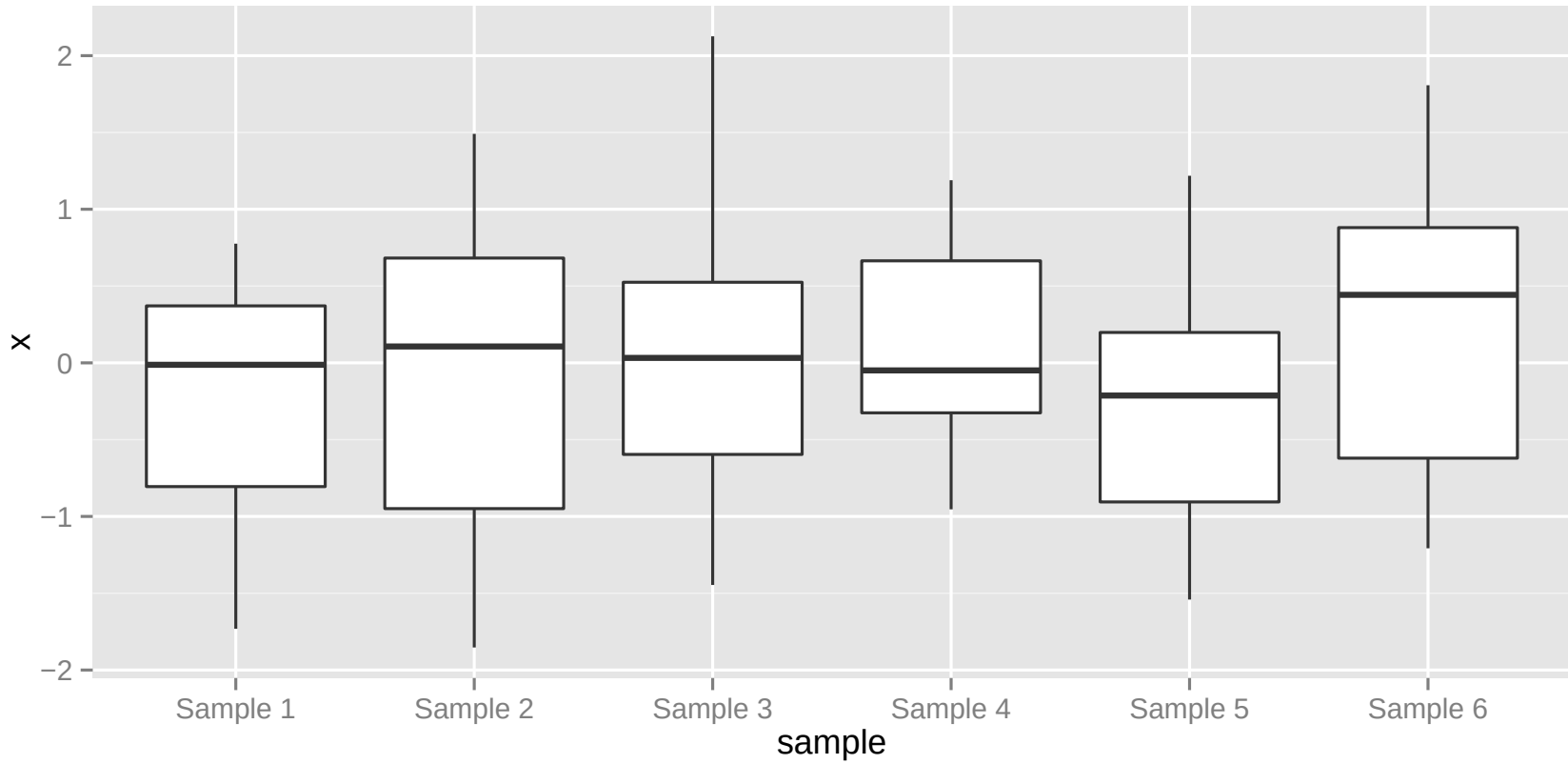
For the one and two sample t-based procedures, you should look at the histograms for each sample:

- Look for the unexpected: gross outliers, values that don't make sense etc.
- Don't expect the shape of a small sample to tell you much about the shape of the population.

Boxplots

Boxplots are quite a common way to summarize a sample.

Here are side by side boxplots for the 6 samples from earlier.



Boxplots

Boxplots can be useful as a more compact alternative to histograms, and they directly display a measure of center and spread.

However, much information is lost through the summary, and a boxplot may not reveal important features of the sample (e.g a bimodal distribution).

Be aware that different programs might have different definitions for how to draw the boxplot.

What do we do if the assumptions are not satisfied?

- *Get more data!* If the only concern is small samples from non-Normal populations, taking a larger sample provides more assurance the central limit theorem applies and the t-based methods will be valid.
- *Use more complicated methods* There are methods beyond those we have seen, that can appropriately deal with non-independence, and special types of response data.
- *A transformation may help* The assumptions may look problematic because you are working with the response variable on the wrong scale. One example, which is used quite commonly is the logarithmic transform.

The log transform

General Idea: transform the raw responses, then use a standard procedure on the transformed responses.

Often variables that vary over several orders of magnitude are more naturally measured on a log scale. You already know some: magnitude of an earthquake (Richter), loudness of a sound (decibel) and musical pitch (octave).

The cloud seeding study

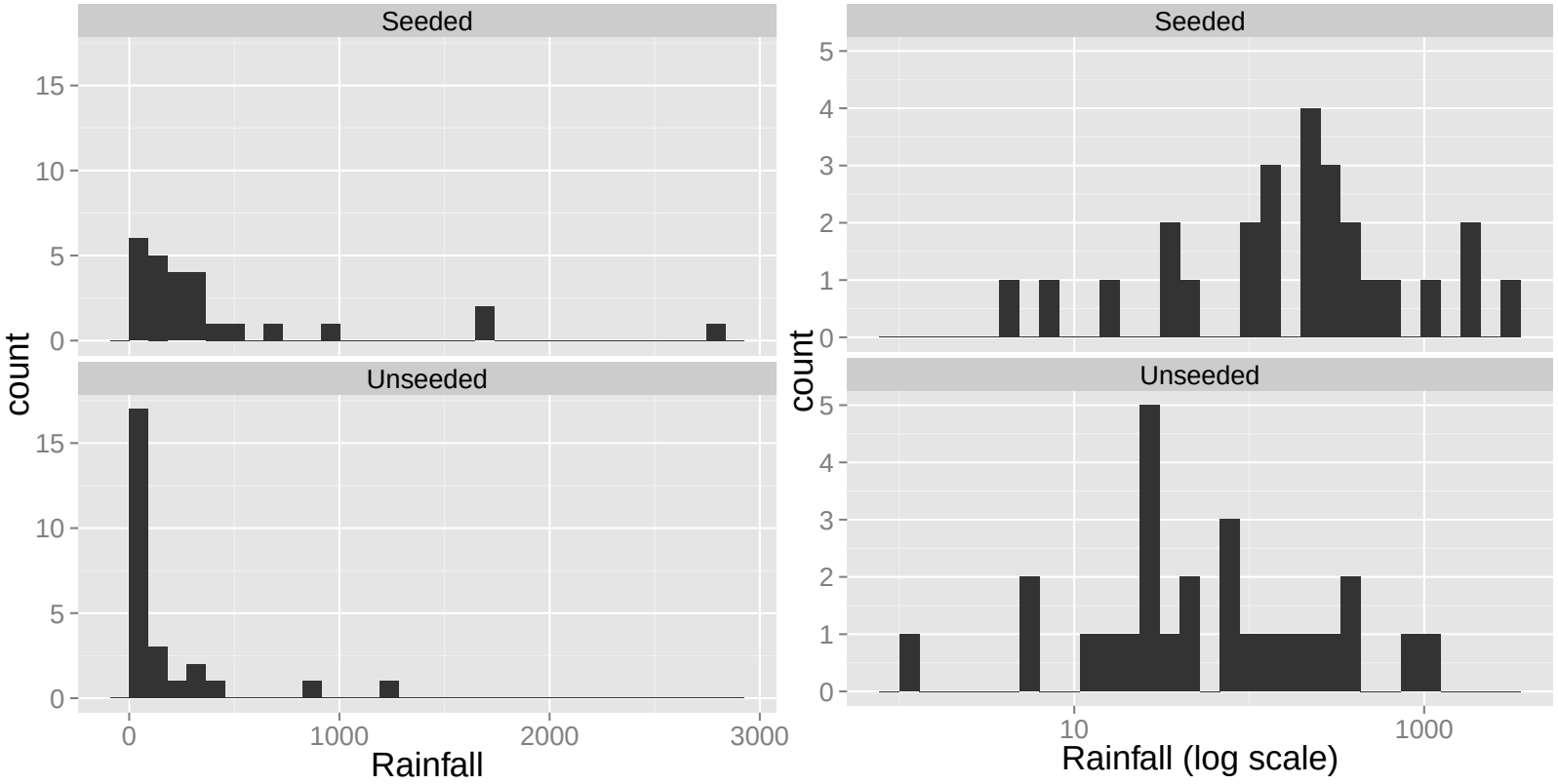
An experiment was conducted to examine if seeding a cloud with silver iodide would increase rainfall.

On 52 days, a plane flew through a target cloud and injected its cargo. On 26 days, chosen randomly, the cargo was the chemical of interest, on the other 26 the cargo was empty.

Rainfall from the target cloud was measured by radar.

For each day, the response is the rainfall, the explanatory variable, is the treatment, whether the cloud was *Seeded* or *Unseeded*.

The cloud seeding study



The log transform

Another situation in which a log transform makes sense is when the effects of a treatment are multiplicative not additive.

An **additive treatment effect**, describes the change in the response as a change by a fixed amount for all units. For example, if seeding a cloud increased the rainfall 100 acre-feet for any cloud, the effect of seeding is additive.

A **multiplicative treatment effect**, describes the change in the response as a change by an amount relative to the untreated response. For example, if seeding a cloud doubled the rainfall (or increased the rainfall by a factor of 2) for any cloud, the effect of seeding is multiplicative.

Example

In the cloud seeding study, a multiplicative treatment effect seems reasonable.

The two sample t-test is used on the log Rainfall.

A small p-value, suggests evidence for different population mean log Rainfall between Seeded and Unseeded clouds.

Caveats about using t-based tools on log transformed data

The interpretation of back-transformed confidence intervals as inference about medians (Section 3.5.2 in the Statistical Sleuth), relies crucially on the assumption that the populations are symmetric on the log scale.

Since this assumption is rarely true in practice, and very hard to verify, we don't recommend this interpretation.