Stat 411/511

Assumptions, Outliers & Log

Oct 21 2015

Charlotte Wickham

stat511.cwick.co.nz

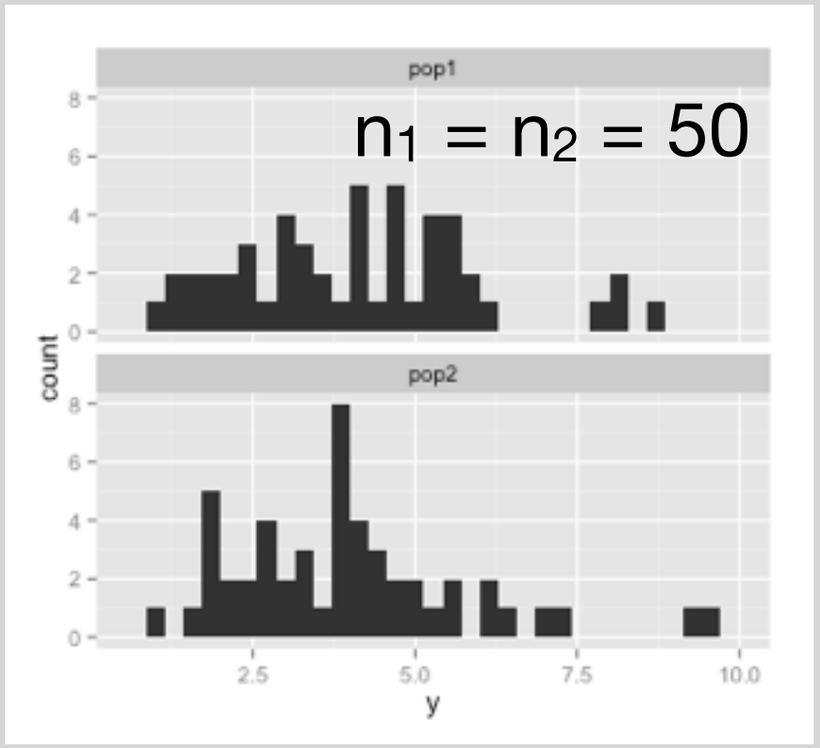
Your turn

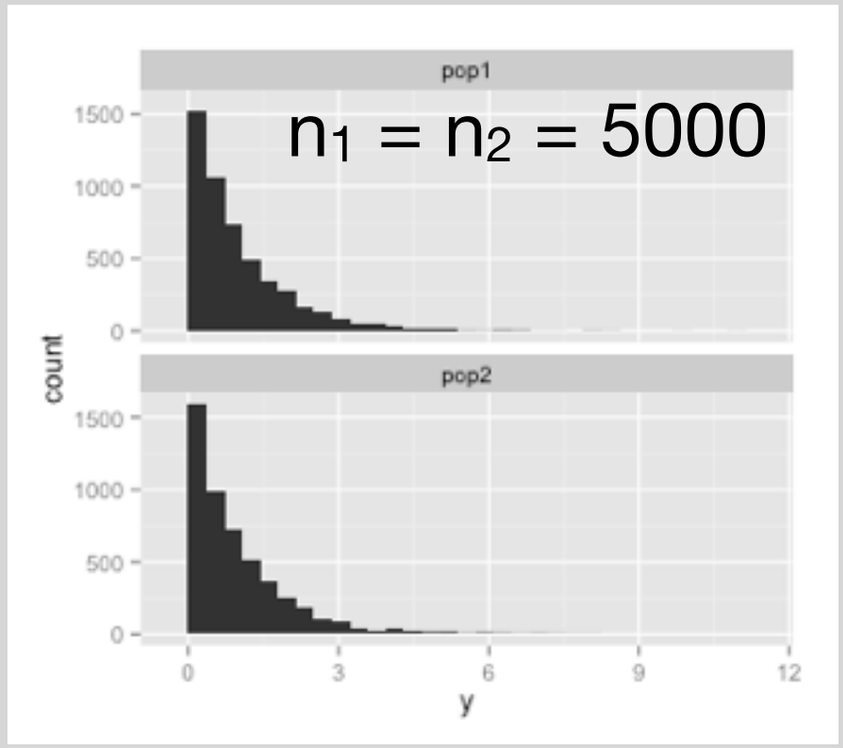
We are going to do a two sample t-test.

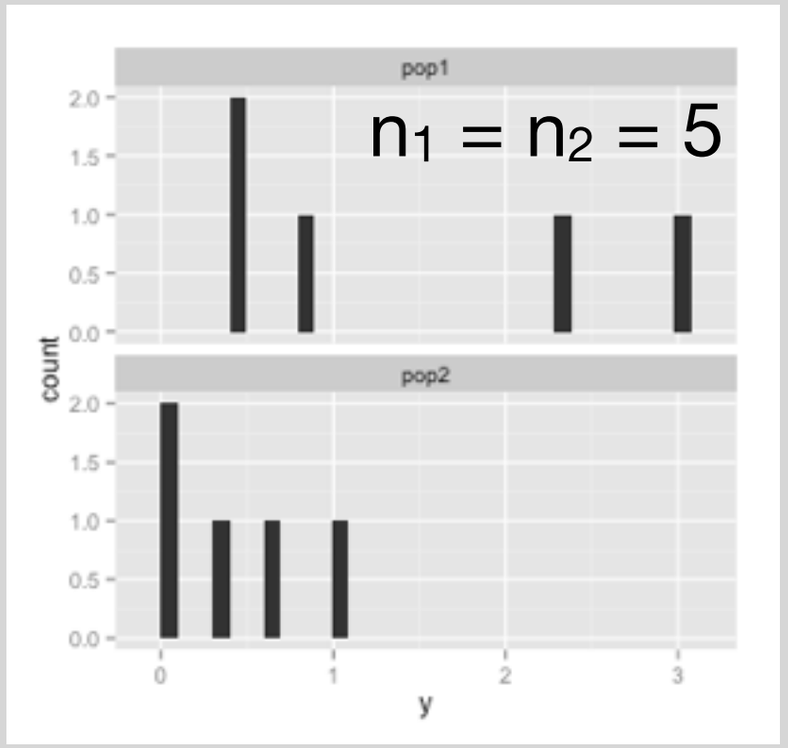
Put these datasets in order from:

“I would be **very worried** about the Normality assumption” to

“I would **not be worried** about the Normality assumption at all”







Today

The two other assumptions

Outliers & Resistance

The log transform

Quiz: you will need to calculate a pooled standard deviation, you will need a calculator/R.

Equal standard deviations

Equal population standard deviations

We assumed that $\sigma_1 = \sigma_2$ and estimated the value with the pooled standard deviation.

If your **sample sizes** are roughly **equal**, the *t*-test will be fairly robust to **unequal population standard deviations**.

If your **sample sizes are not equal**, the *t*-test **will not** be valid with **unequal population standard deviations**.

Paired t-test: taking differences is probably not appropriate if the two groups have different spreads.

Equal population SDs

Check: by looking at histograms of samples.

Remedy:

use a transformation **or**

or always use a Welch *t*-test

Independence

Independence

Observations must be independent of one another for *t*-tools to be valid.

If knowledge about one observation allows us to make a better guess about another observation, there is a lack of independence.

If there is dependence, t-tools give misleading results.

Two common types of dependence

Cluster effects

There is some kind of subgroup within groups, and subjects in the same subgroup are more similar.

Serial effects

Measurements are made in time (or space) and observations made close in time (or space) are more similar.

Independence

Check: read study design carefully, and look for sources of dependence.

Remedy:

use more complicated methods (ST512)

The one solution we already know about is the paired t-test

Assumptions for using the randomization test

You have a randomized experiment

(i.e. you have assigned subjects to treatment groups completely at random)

Assumptions for using the two sample t-**test** instead of the randomization **test**

i.e. you want a p-value using the
null hypothesis: there is no difference between the treatments

You have a randomized experiment

(i.e. you have assigned subjects to treatment groups completely at random) and

the samples aren't too non-Normal

(same deal as two sample t, you can get away with more non-Normality when you have more data)

Assumptions for using the two sample **CI** for difference in means instead of the randomization **CI** for an additive treatment effect

or you want a p-value using the
null hypothesis: the additive treatment effect is zero.

You have a randomized experiment

(i.e. subjects were assigned to treatment groups completely at random) and

the samples aren't too non-Normal

(same deal as two sample t, you can get away with more non-Normality when you have more data)

you assume the additive treatment model

(implies equal standard deviations of treatment outcomes)

In practice, you look at all the same diagnostic plots as the two sample t-test

Your turn

Consider the numbers:

1, 2, 3, 5, 9, 10

What is their average?

What is their median?

Imagine there was a mistake in recording the numbers and you were actually given:

1, 2, 3, 5, 9, 100

What is their average?

What is their median?

Resistance

A procedure is **resistant** if it doesn't change much when a few subjects change.

The average **is not** resistant.

The median **is** resistant.

The t-statistic i**s not** resistant.

It can be sensitive to a few outlying observations

Outliers should not be deleted unless you **know** they are mistakes

If there are outliers retry the analysis without them.

If the conclusions don't change, leave them in and say so.

If the conclusions do change, investigate further, report both analyses.

OR

Use a **resistant** method (Chap 4)

Log transform

Sometimes assumptions can be met by transforming the data.

A particularly useful transformation is the **logarithmic** transformation.

- Useful, when variation increases with mean, or right skewed data.
- Values must be positive to take logarithm.
- Always use the same transformation on both groups.

Cloud seeding

Display 3.1 **p. 57**

Rainfall (acre-feet) for days with and without cloud seeding

Rainfall from unseeded days (n = 26)

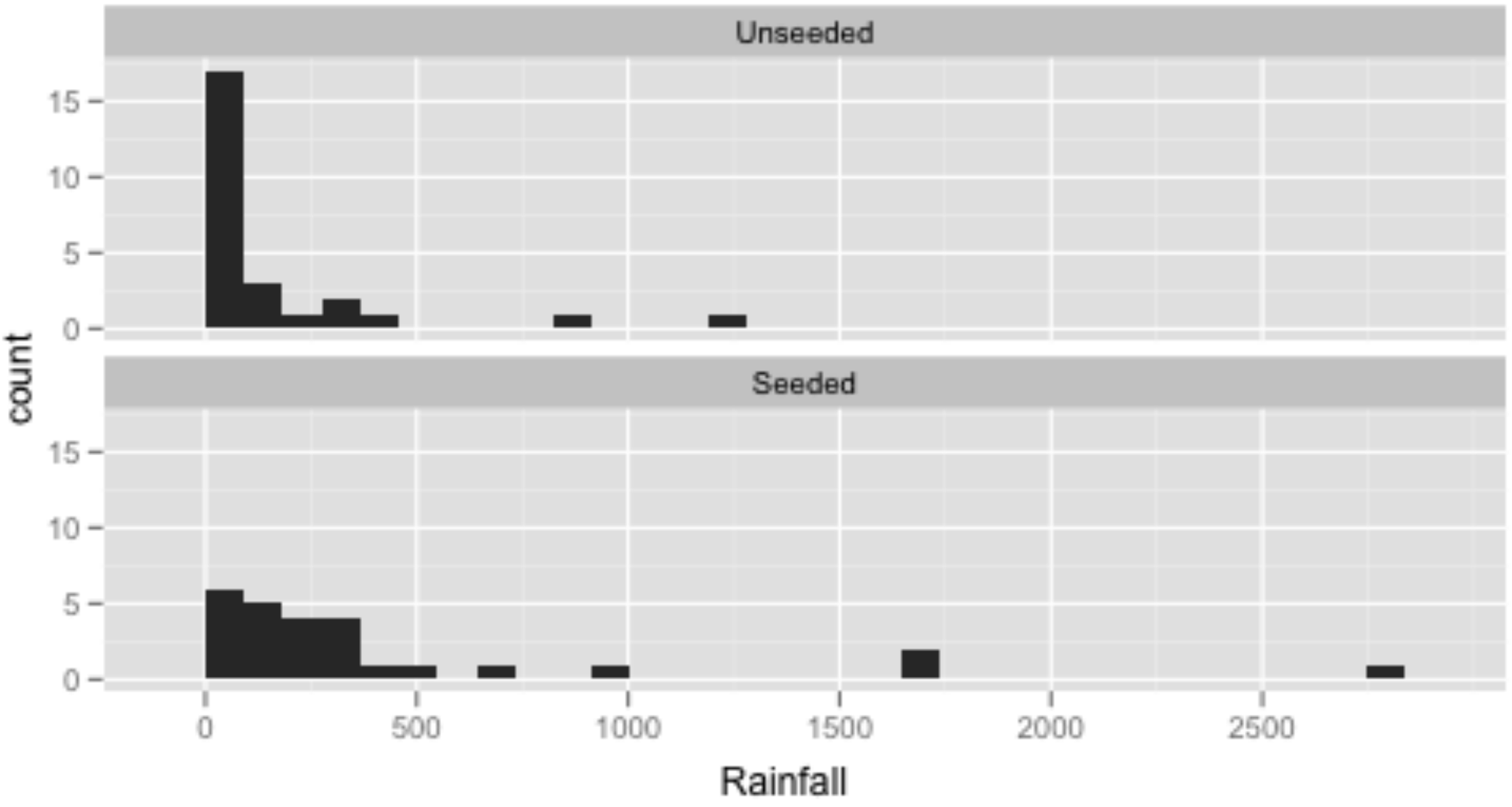
1202.6	830.1	372.4	345.5	321.2	244.3	163.0	147.8	95.0
87.0	81.2	68.5	47.3	41.1	36.6	29.0	28.6	26.3
26.1	24.4	21.7	17.3	11.5	4.9	4.9	1.0	

Rainfall from seeded days (n = 26)

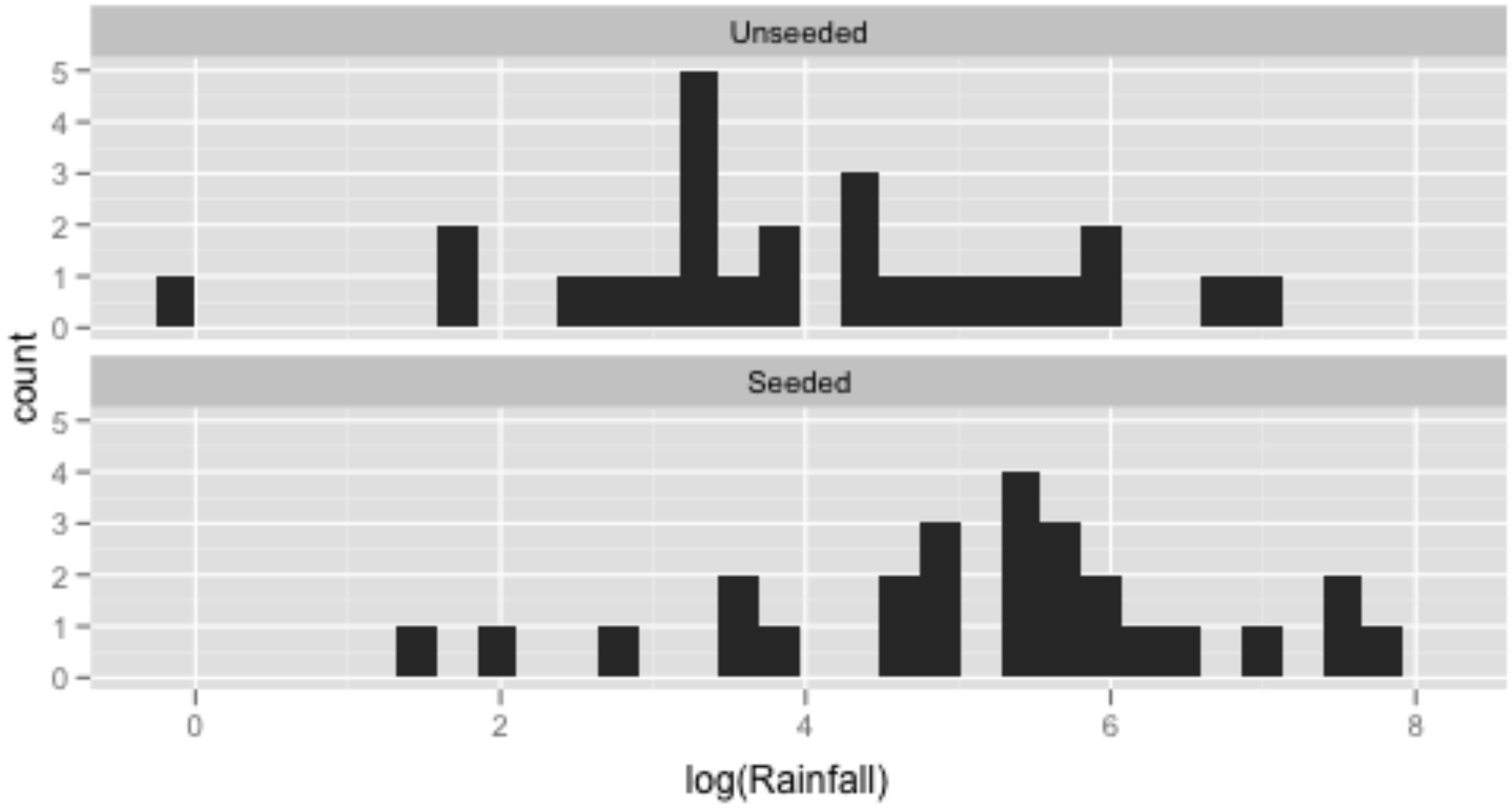
2745.6	1697.8	1656.0	978.0	703.4	489.1	430.0	334.1	302.8
274.7	274.7	255.0	242.5	200.7	198.6	129.6	119.0	118.3
115.3	92.4	40.6	32.7	31.4	17.5	7.7	4.1	

Randomized experiment

```
qplot(Rainfall, data = case0301) +
  facet_wrap(~ Treatment, ncol = 1)
```



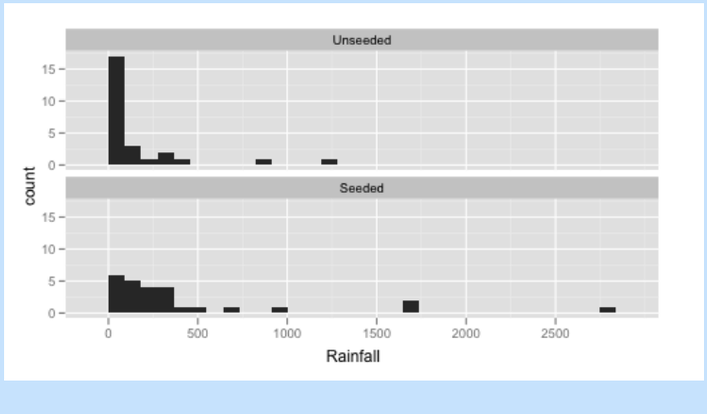
```
qplot(log(Rainfall), data = case0301)
  + facet_wrap(~ Treatment, ncol = 1)
```



details on Mon

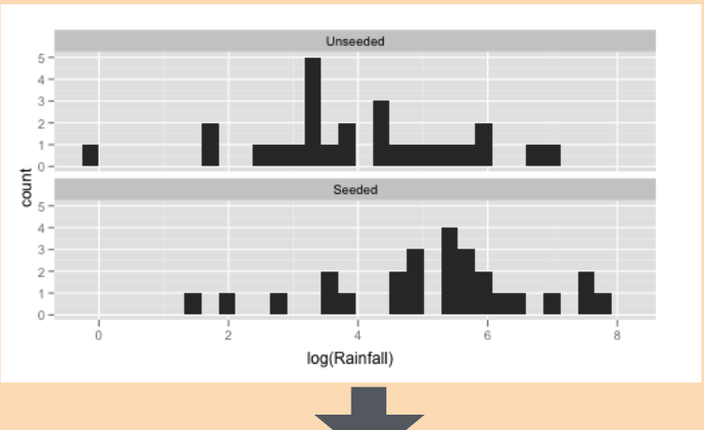
Using the log transform

original scale
acre feet of rainfall



log transform
data

log scale
log(acre feet of rainfall)



Recheck assumptions

if things aren't improved, don't proceed

Conduct analysis (t-tests etc.)

```
Two Sample t-test

data:  log(Rainfall) by Treatment
t = -2.5444, df = 50, p-value = 0.01408
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.0466973 -0.2408651
sample estimates:
mean in group Unseeded   mean in group Seeded
              3.990406               5.134187
```

back transform
estimates and
confidence intervals
but **not** p-values

Make interpretations

With 95% confidence seeding clouds increases rainfall between 1.27 and 7.74 times that of unseeded clouds.

once back transformed those t-tests tell us
about **ratios of medians** of the populations of **response**

t-tests over here tell us about **differences**
in means of the populations of **log response**