# Stat 411/511

## Log Transform

Oct 27 2014

Charlotte Wickham

stat511.cwick.co.nz

# Log transform

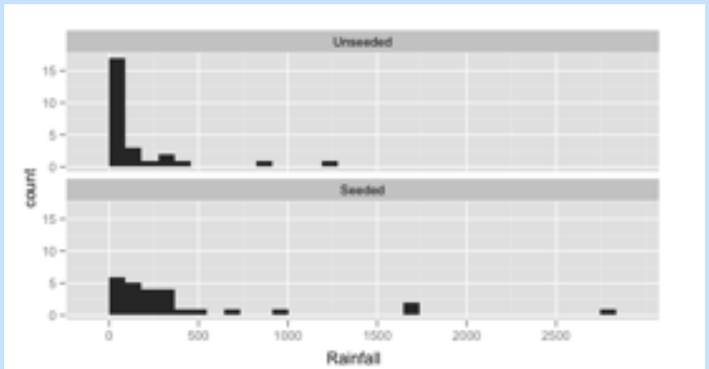
Sometimes assumptions can be met by transforming the data.

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.

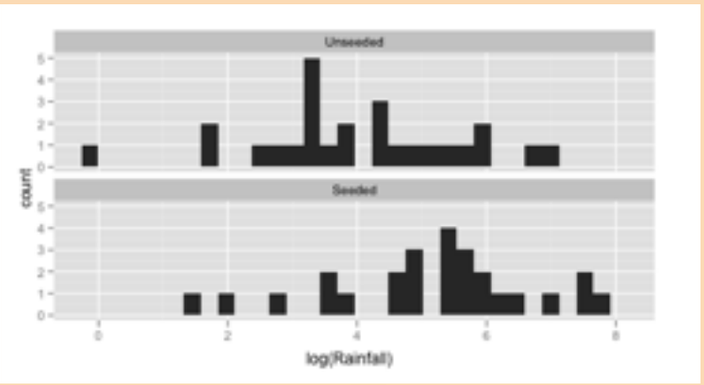
# Using the log transform

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
```

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

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**

# Cloud seeding

**Display 3.1** **p. 57**

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

Rainfall from unseeded days (n = 26)

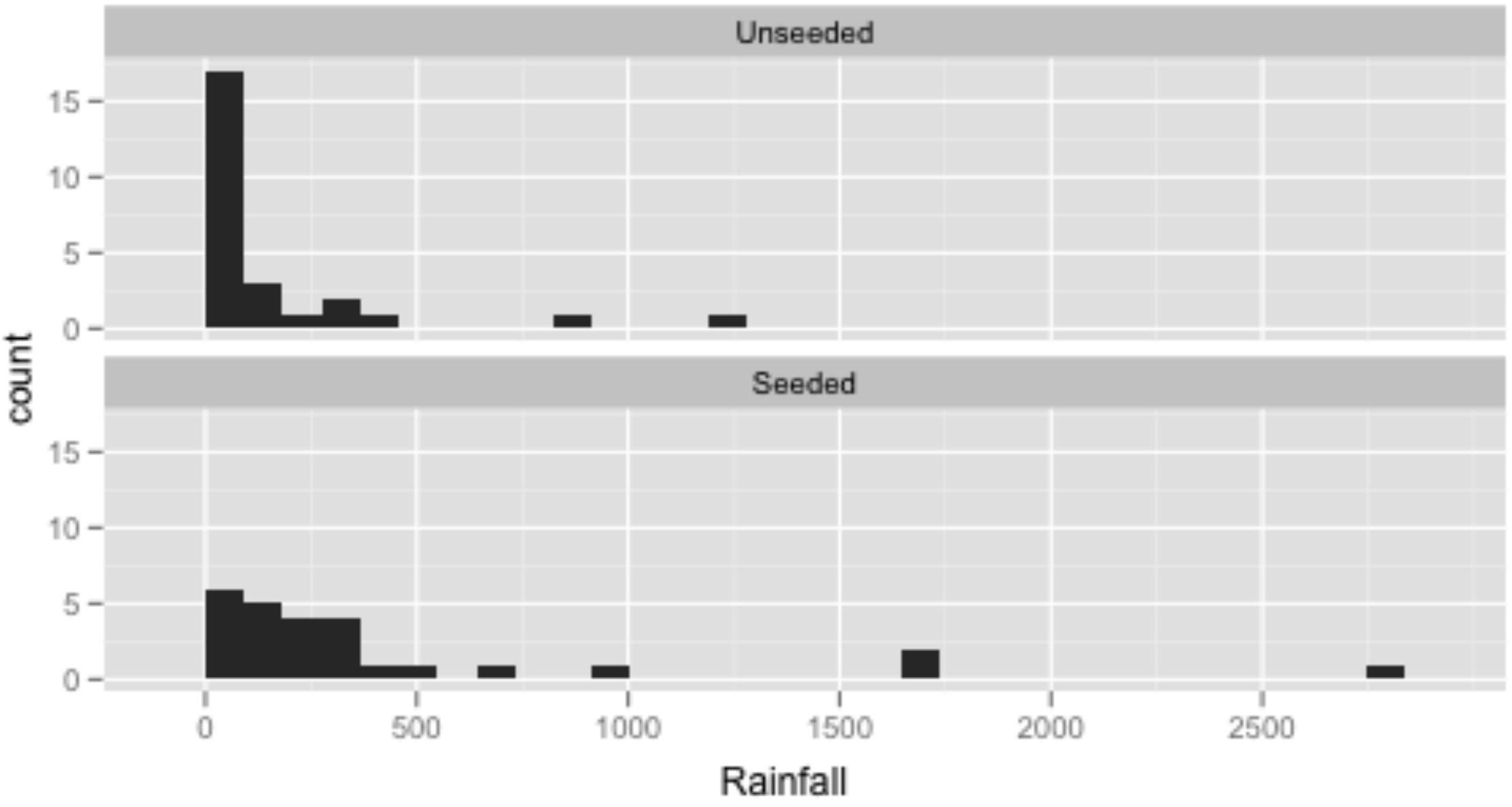
| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1202.6 | 830.1 | 372.4 | 345.5 | 321.2 | 244.3 | 163.0 | 147.8 | 95.0 |
| 87.0 | 81.2 | 68.5 | 47.3 | 41.1 | 36.6 | 29.0 | 28.6 | 26.3 |
| 26.1 | 24.4 | 21.7 | 17.3 | 11.5 | 4.9 | 4.9 | 1.0 | |

Rainfall from seeded days (n = 26)

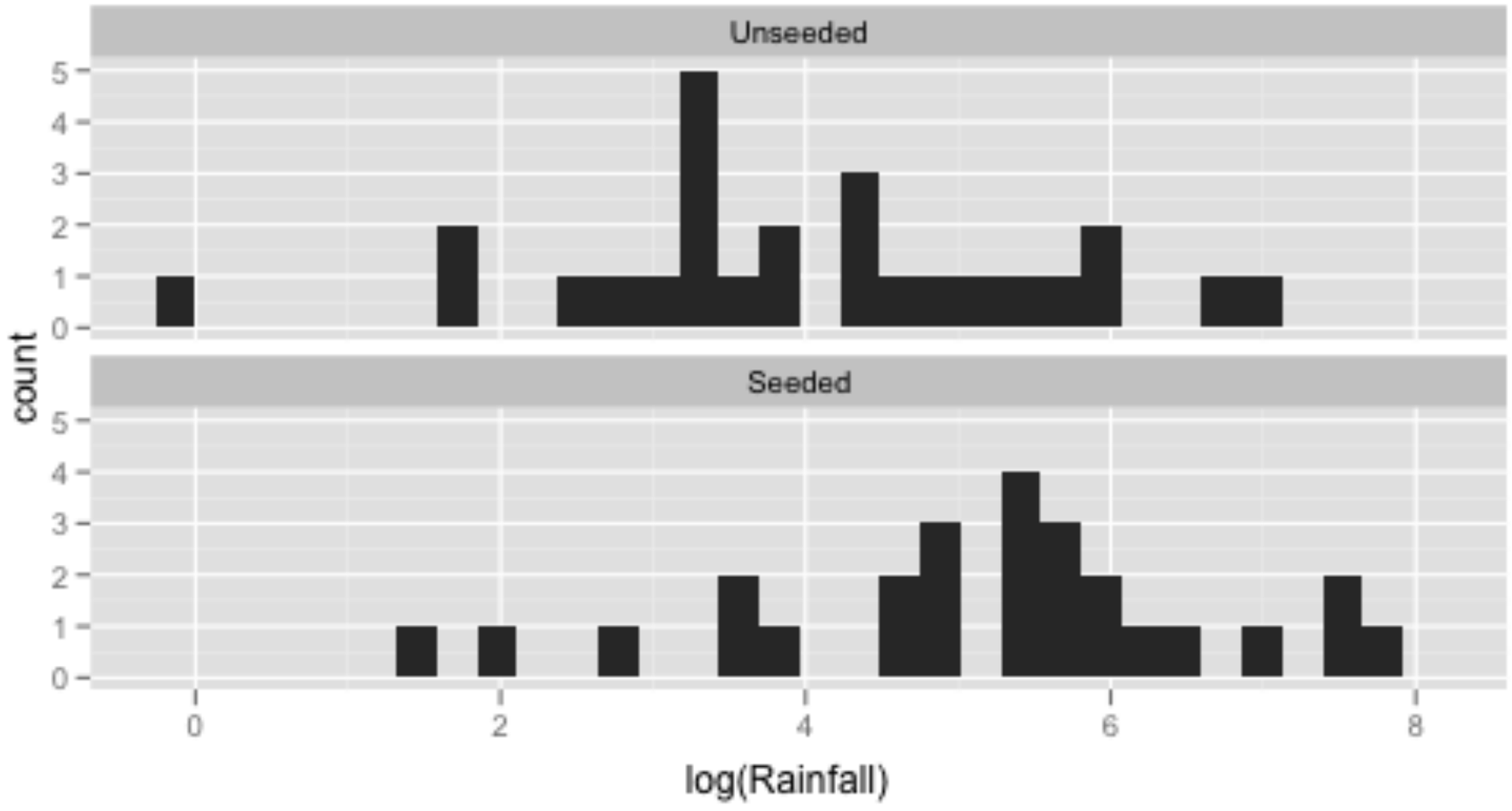
| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 2745.6 | 1697.8 | 1656.0 | 978.0 | 703.4 | 489.1 | 430.0 | 334.1 | 302.8 |
| 274.7 | 274.7 | 255.0 | 242.5 | 200.7 | 198.6 | 129.6 | 119.0 | 118.3 |
| 115.3 | 92.4 | 40.6 | 32.7 | 31.4 | 17.5 | 7.7 | 4.1 | |

# Randomized experiment

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



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



# Facts about log

We will only use the natural logarithm

(log to the base e, ln)

$\exp(\log(x)) = x$

$\log(AB) = \log(A) + \log(B)$

$\exp(A + B) = \exp(A)\exp(B)$

# Cloud seeding

After a log transform, we are satisfied a two sample t-test is appropriate.

(the two sample t p-value and CI, will be a good approximation to the p-value and CI from a randomization test on log rainfall)

for randomized experiments

# Procedure

1. Take the logarithm of the data,

   $Z_1 = \log(Y_1)$, $Z_2 = \log(Y_2)$

2. Perform t-test using $Z_1$ and $Z_2$. If the p-value is small we have evidence the **<u>additive</u> treatment effect** on the **log outcome** is not zero

3. We estimate the **<u>multiplicative</u> treatment effect** on the **untransformed outcome** is **$\exp(\bar{Z}_2 - \bar{Z}_1)$**

CI's need to be "back"-transformed too.

# Why multiplicative?

δ = the additive treatment effect on (outcome) of
being assigned to (Group 2) compared to (Group 1)

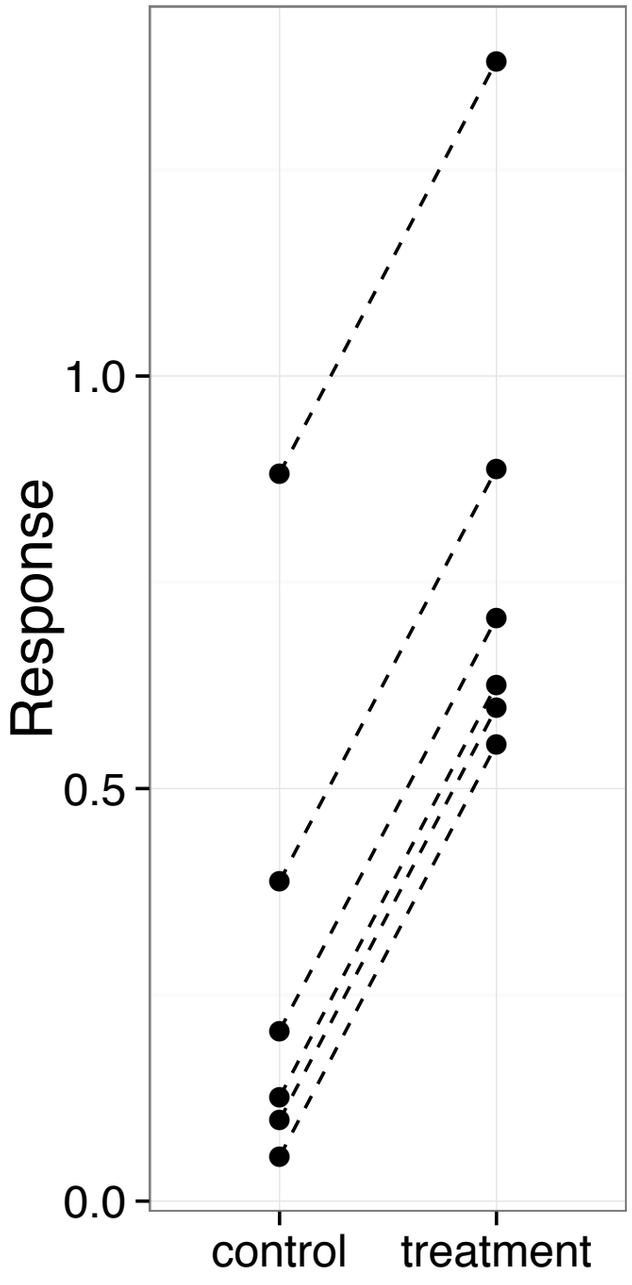
$$Z_2 = Z_1 + \delta$$ (definition of treatment effect)

$$\exp(Z_2) = \exp(Z_1 + \delta)$$ (back transform to original scale)

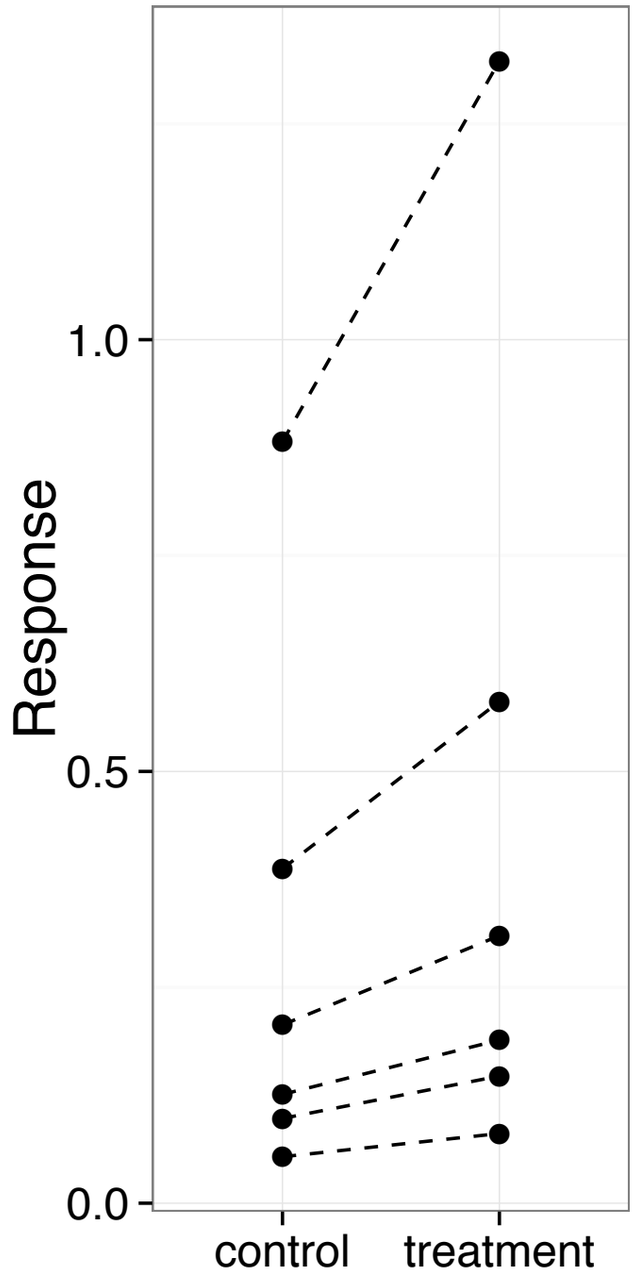
$$\exp(Z_2) = \exp(Z_1)\exp(\delta)$$ (property of exp)

$$Y_2 = Y_1 \exp(\delta)$$ (definition of $Z_1$ & $Z_2$)

## Additive treatment




## Multiplicative treatment



```
t.test(log(Rainfall) ~ Treatment, data = case0301,
       var.equal = TRUE)
```

```
Two Sample t-test
```

There is moderate evidence against the hypothesis that treatment effect on log rainfall for Seeding is zero compared to Not Seeding (two sample t-test, pvalue = 0.014).

```
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
```

exp(5.13 - 3.99) = 3.13

We estimate the treatment effect of seeding a cloud is to increase rainfall by 3.13 times the rainfall from an unseeded cloud.

```
95 percent confidence interval:
 -2.0466973 -0.2408651
```

exp(0.2408) = 1.27

exp(2.0466) = 7.74

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

this is the wording for a multiplicative treatment effect

# Statistical Summary

There is moderate evidence the seeding treatment is not the same as the control treatment (two sample t-test of log rainfall, two sided p-value = 0.014).

Alternative: There is moderate evidence against the hypothesis that treatment effect on log rainfall for Seeding is zero compared to Not Seeding (two sample t-test, pvalue = 0.014).

We estimate the treatment effect of seeding a cloud is to increase rainfall by 3.13 **times** that of unseeded clouds.

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

# What about a randomization test?

**We could do one on the Rainfall scale:**

```
> oneway_test(Rainfall ~ Treatment, data = case0301,
              distribution = approximate(B = 9999))

    Approximative 2-Sample Permutation Test

data:  Rainfall by Treatment (Seeded, Unseeded)
Z = 1.9421, p-value = 0.0439
alternative hypothesis: true mu is not equal to 0
```

**Or the log Rainfall scale:**

```
> oneway_test(log_rainfall ~ Treatment, data = case0301,
              distribution = approximate(B = 9999))

    Approximative 2-Sample Permutation Test

data:  log_rainfall by Treatment (Seeded, Unseeded)
Z = 2.4179, p-value = 0.0134
alternative hypothesis: true mu is not equal to 0
```

different p-value, but similar conclusion. There is moderate evidence that the seeding treatment does something.

We wouldn't be happy talking about an additive treatment model on this scale.

This test is legitimate here (we satisfied the assumptions), but we are being more vague about what the treatment does.

p-value is close to the two-sample t. On this scale the two sample t-test is a good approximation to the randomization test.

We would be happy talking about an additive treatment model on this scale.

This test is also legitimate here (we satisfied the assumptions), and we could even use the corresponding confidence intervals.

for randomized experiments

# Procedure

1. Take the logarithm of the data,

   $Z_1 = \log(Y_1)$, $Z_2 = \log(Y_2)$

2. Perform t-test using $Z_1$ and $Z_2$. If the p-value is small we have evidence the **treatment effect** on the **log outcome** is not zero

3. We estimate the **treatment effect** is to **multiply** the outcome by **$\exp(\bar{Z}_2 - \bar{Z}_1)$** CI's need to be "back"-transformed too.

for observational studies

i.e. two samples from two populations

# Procedure

1. Take the logarithm of the data,

   $Z_1 = \log(Y_1)$, $Z_2 = \log(Y_2)$

2. Perform t-test using $Z_1$ and $Z_2$. If the p-value is small we have evidence the **population mean** of $\log(Y_1)$ differs to $\log(Y_2)$

3. We estimate the **median** value of population 2 is **$\exp(\bar{Z}_2 - \bar{Z}_1)$ times** the **median** value of population 1. CI's need to be "back"-transformed too.

# Why ratio of medians?

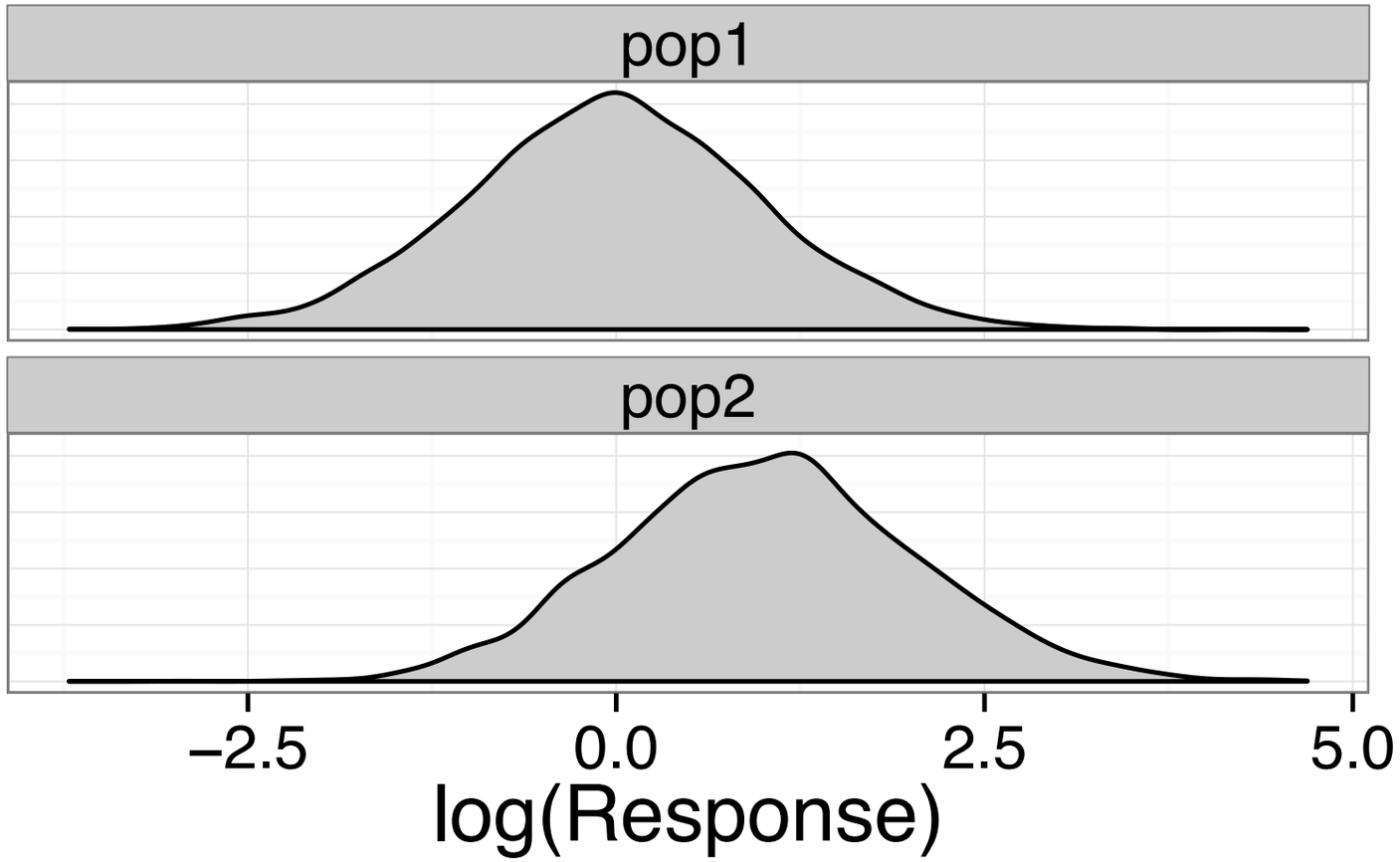
$$\exp(\text{mean of } \log(Y_2) - \text{mean of } \log(Y_1)) \neq \text{mean}(Y_2) - \text{mean}(Y_1)$$

$$\exp(\text{median of } \log(Y_2) - \text{median of } \log(Y_1)) = \text{median of } Y_2 \,/\, \text{median of } Y_1$$

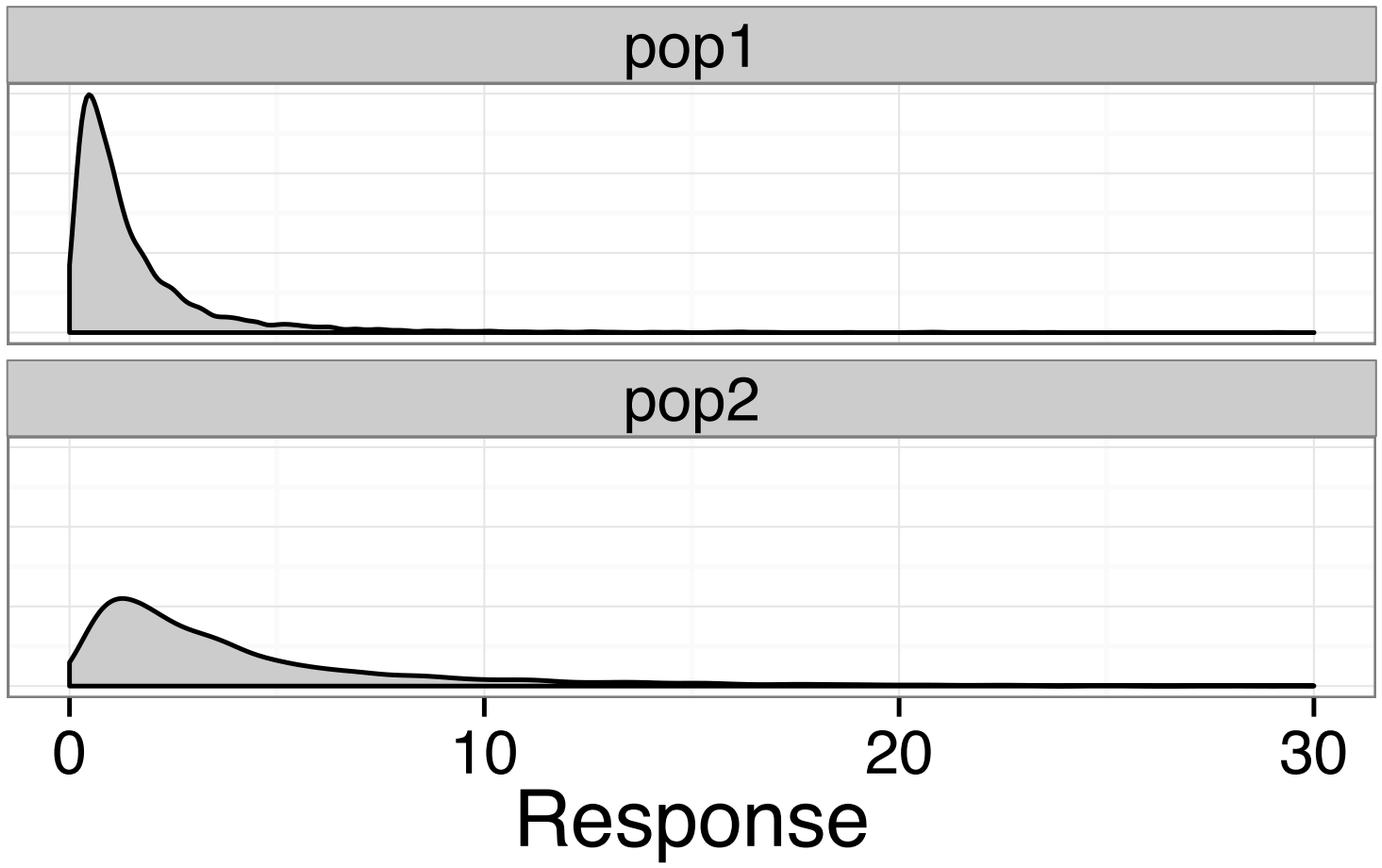
**Hidden assumption:** for the t-test to tell us about the median of the transformed data, the population mean of the log outcome must be the same as the population median of the log outcome.

I.e. we assume the population is symmetric on the log scale

On the log scale, the populations are symmetric with the same spread.




On the original scale, the populations are skewed with the different spreads.



# Statistical Summary

## for observational study

There is **moderate** evidence that the median ***response*** of ***population 1*** is not the same as the median ***response*** of ***population 2*** *(*two sample t-test on log transformed response, two sided p-value = **0.014**).

(Alternative) There is **moderate** evidence that the mean log ***response*** of ***population 1*** is not the same as the mean log ***response*** of ***population 2*** *(*two sample t-test, two sided p-value = **0.014**).

We estimate that the median ***response*** of ***population 1*** is **XX** times the median ***response*** of ***population 2***.

With 95% confidence, the median ***response*** of ***population 1*** is between **XX** and **YY** times that of ***population 2***.

replace **bold** terms with correct context and values