

Stat 411/511

# OUTLIERS & TRANSFORMATIONS

Feb 3 2012

Charlotte Wickham

[stat511.cwick.co.nz](http://stat511.cwick.co.nz)

# 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?

# Today

Outliers

The log transformation

# Resistance

# Resistance

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

# Resistance

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

The average **is not** resistant.

# 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.

# 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 **is not** resistant.

# 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 **is not** resistant.

It can be sensitive to a few outlying observations



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

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

If there are outliers retry the analysis without them.

**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.

**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.

**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

**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

# Log transform

Sometimes assumptions can be met by transforming the data.

# Log transform

Sometimes assumptions can be met by transforming the data.

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

# 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.

# 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.

# 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	

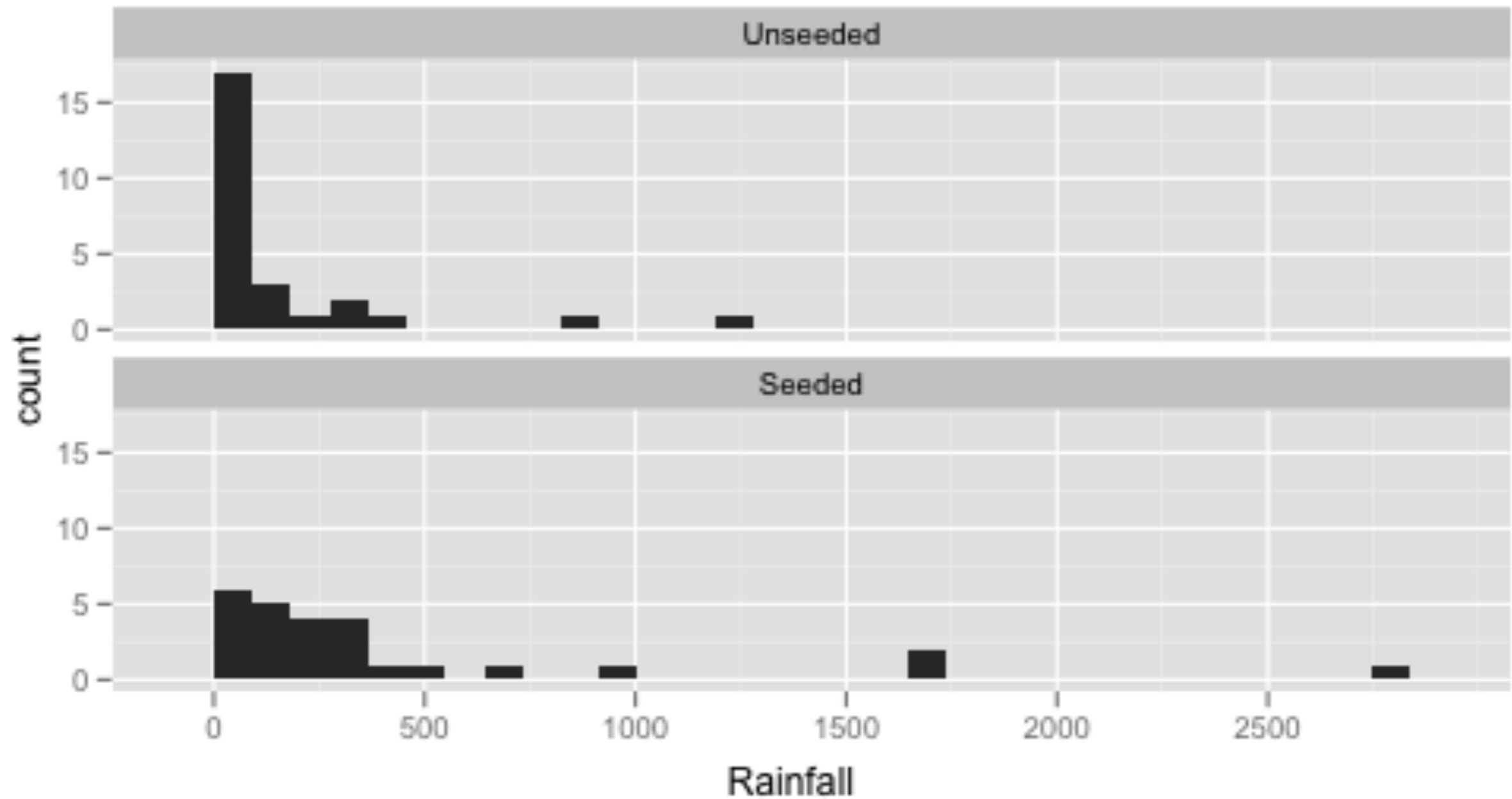
### Rainfall from seeded days (n = 26)

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	

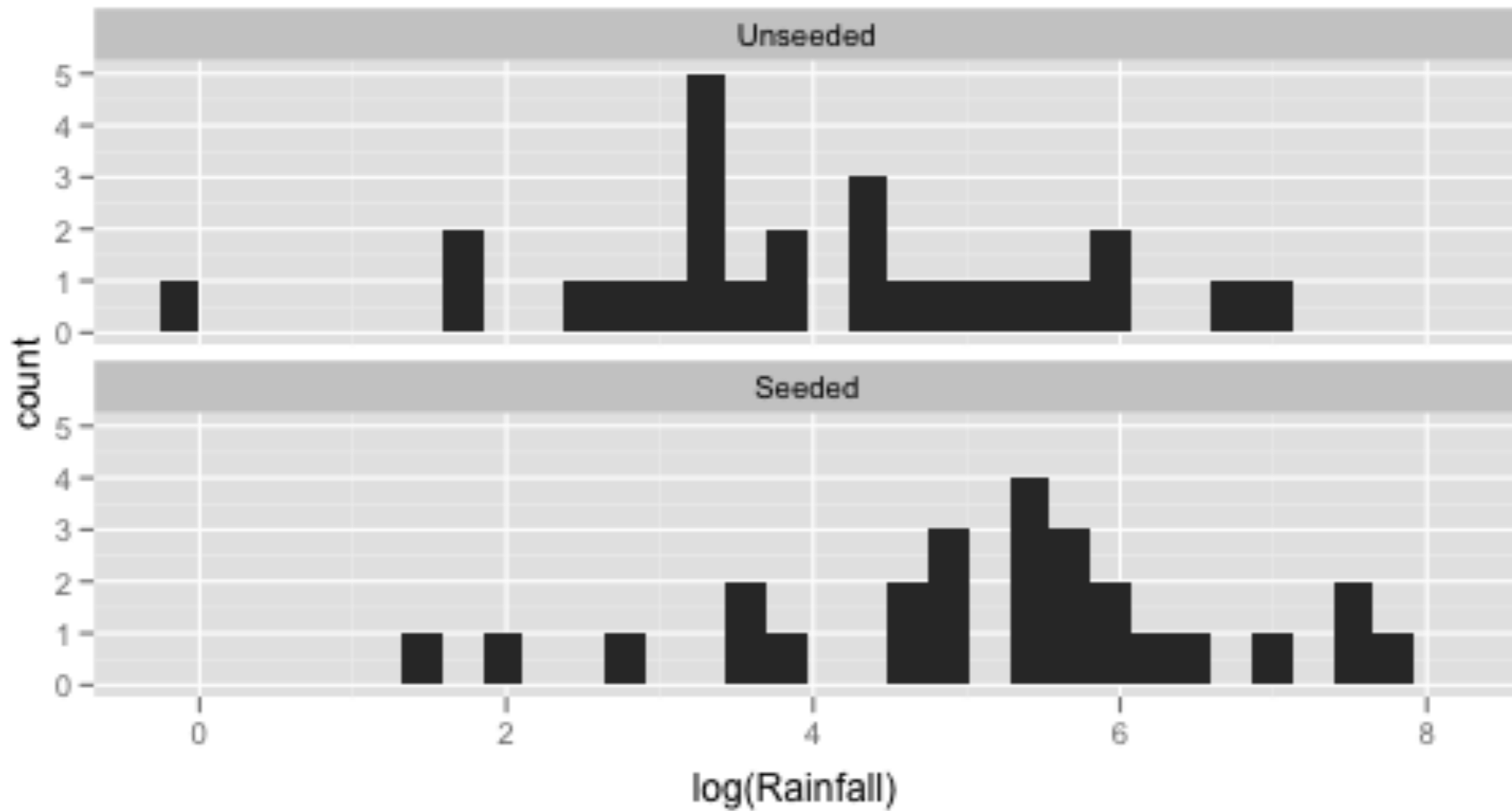
---

**Controlled 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)

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

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

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

# 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.

# Why multiplicative?

$\delta$  = the treatment effect on (outcome) of being assigned to (Group 2) compared to (Group 1)

# Why multiplicative?

$\delta$  = the treatment effect on (outcome) of being assigned to (Group 2) compared to (Group 1)

$$Z_2 = Z_1 + \delta \quad (\text{definition of treatment effect})$$

# Why multiplicative?

$\delta$  = the treatment effect on (outcome) of being assigned to (Group 2) compared to (Group 1)

$$Z_2 = Z_1 + \delta \quad \text{(definition of treatment effect)}$$

$$\exp(Z_2) = \exp(Z_1 + \delta) \quad \text{(back transform to original scale)}$$

# Why multiplicative?

$\delta$  = the treatment effect on (outcome) of being assigned to (Group 2) compared to (Group 1)

$$Z_2 = Z_1 + \delta \quad \text{(definition of treatment effect)}$$

$$\exp(Z_2) = \exp(Z_1 + \delta) \quad \text{(back transform to original scale)}$$

$$\exp(Z_2) = \exp(Z_1) \exp(\delta) \quad \text{(property of exp)}$$

# Why multiplicative?

$\delta$  = the treatment effect on (outcome) of being assigned to (Group 2) compared to (Group 1)

$$Z_2 = Z_1 + \delta \quad \text{(definition of treatment effect)}$$

$$\exp(Z_2) = \exp(Z_1 + \delta) \quad \text{(back transform to original scale)}$$

$$\exp(Z_2) = \exp(Z_1) \exp(\delta) \quad \text{(property of exp)}$$

$$Y_2 = Y_1 \exp(\delta) \quad \text{(definition of } Z_1 \text{ \& } Z_2)$$

```
seeded <- subset(case0301, Treatment == "Seeded")$Rainfall
unseeded <- subset(case0301, Treatment == "Unseeded")$Rainfall
t.test(log(seeded), log(unseeded), var.equal = TRUE)
```

## Two Sample t-test

data: log(seeded) and log(unseeded)

t = 2.5444, df = 50, p-value = 0.01408

alternative hypothesis: true difference in means is  
not equal to 0

95 percent confidence interval:

0.2408651 2.0466973

sample estimates:

mean of x mean of y

5.134187 3.990406

```
seeded <- subset(case0301, Treatment == "Seeded")$Rainfall
unseeded <- subset(case0301, Treatment == "Unseeded")$Rainfall
t.test(log(seeded), log(unseeded), var.equal = TRUE)
```

## Two Sample t-test

data: log(seeded) and log(unseeded)

t = 2.5444, df = 50, p-value = 0.01408

alternative hypothesis: true difference in means is  
not equal to 0

95 percent confidence interval:

0.2408651 2.0466973

sample estimates:

mean of x mean of y

5.134187 3.990406

$$\exp(5.13 - 3.99) = 3.13$$

```
seeded <- subset(case0301, Treatment == "Seeded")$Rainfall
unseeded <- subset(case0301, Treatment == "Unseeded")$Rainfall
t.test(log(seeded), log(unseeded), var.equal = TRUE)
```

## Two Sample t-test

data: log(seeded) and log(unseeded)

t = 2.5444, df = 50, p-value = 0.01408

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

0.2408651 2.0466973

sample estimates:

mean of x mean of y

5.134187 3.990406

$$\exp(5.13 - 3.99) = 3.13$$

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

95 percent confidence interval:

0.2408651 2.0466973

$$\exp(0.2408) = 1.27$$

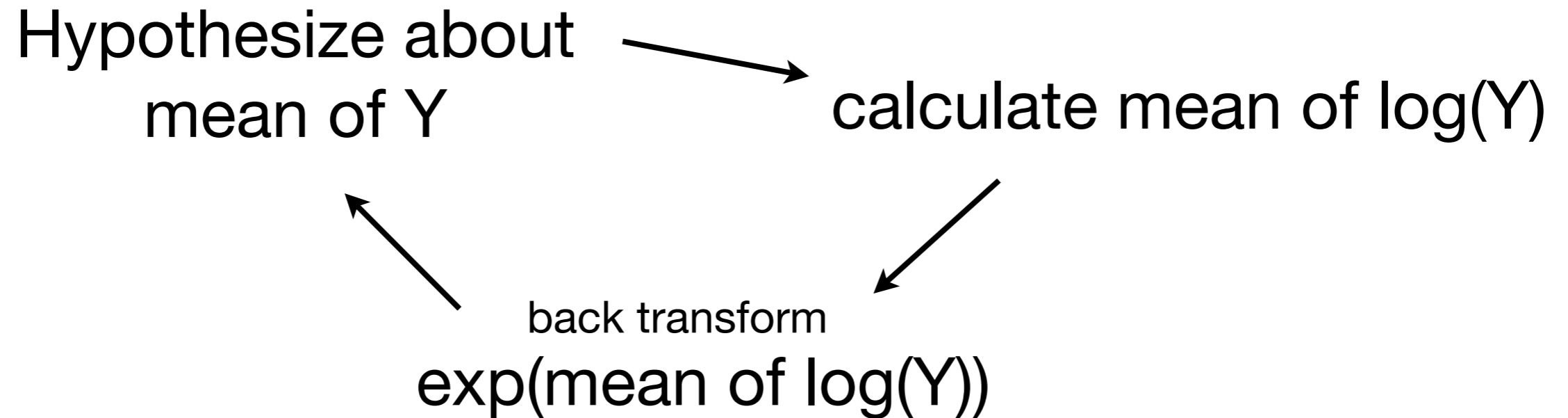
$$\exp(2.0466) = 7.74$$

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

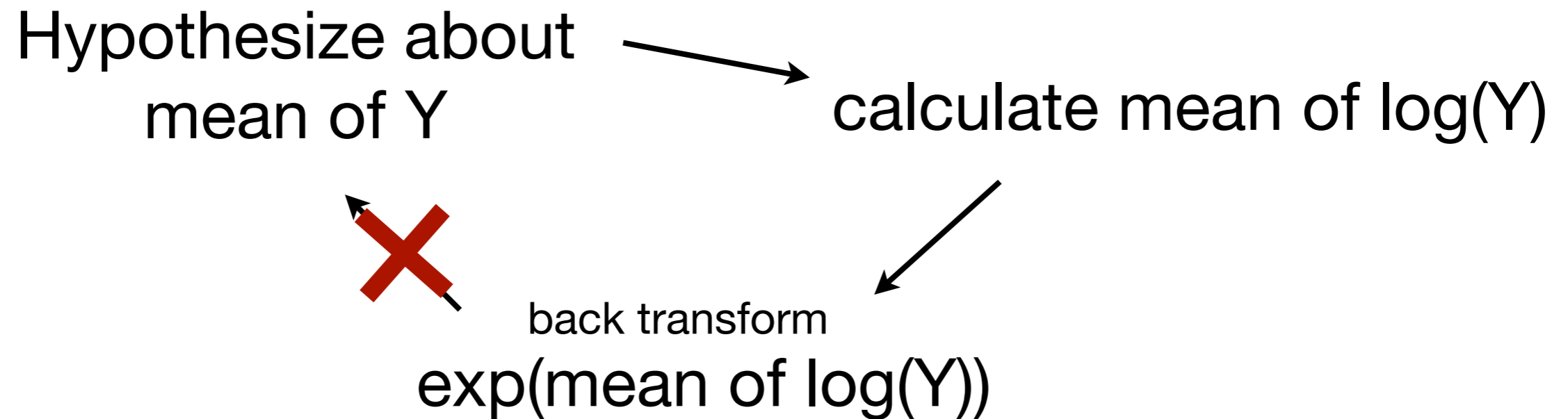
# Procedure

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.

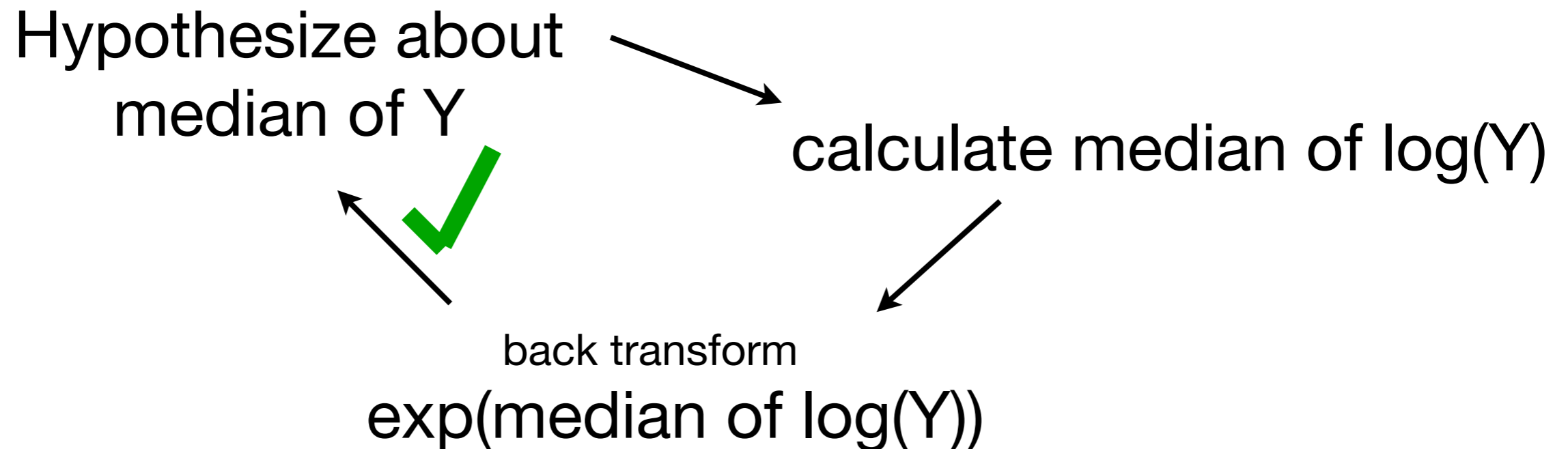
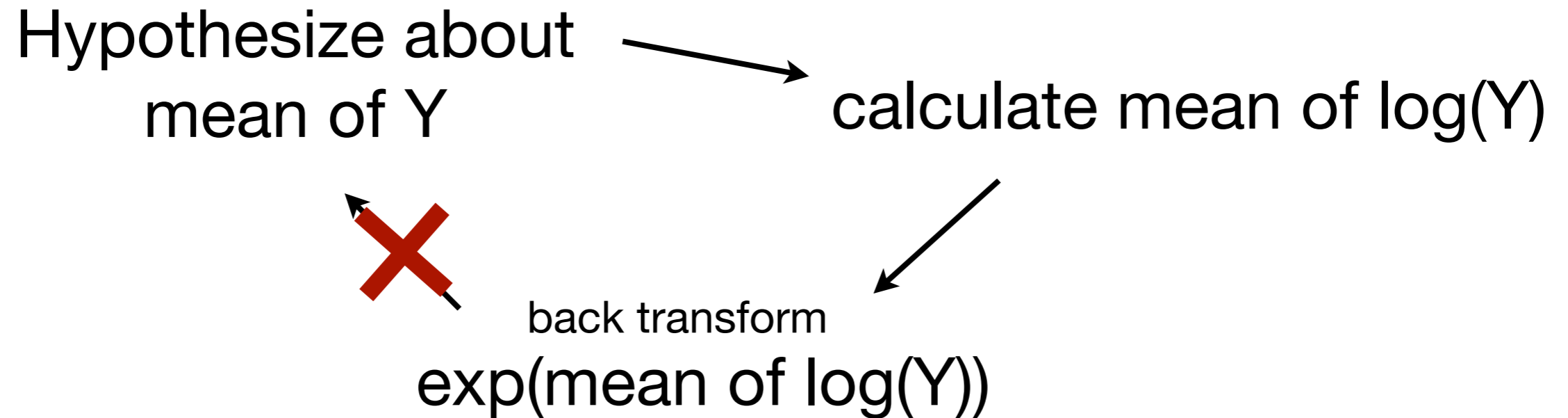
# Why the median?



# Why the median?



# Why the median?



```
x <- 1:21  
mean(x)  
median(x)
```

```
exp(mean(log(x)))
```

```
exp(median(log(x)))
```