

ENV 710: Lecture 9

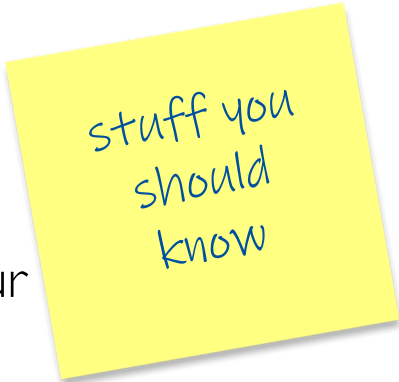
data transformation

data transformation

“normalization”

learning goals

- what to do if your data do not fit the assumptions of your statistical test?
- data transformations
- non-parametric tests

A yellow sticky note with a slight shadow, tilted at an angle, containing handwritten text in blue ink.

stuff you
should
know

data transformation

one of the assumptions of the t-test and other parametric tests is that data are normally distributed

but ... what if they aren't?

example

cloud seeding

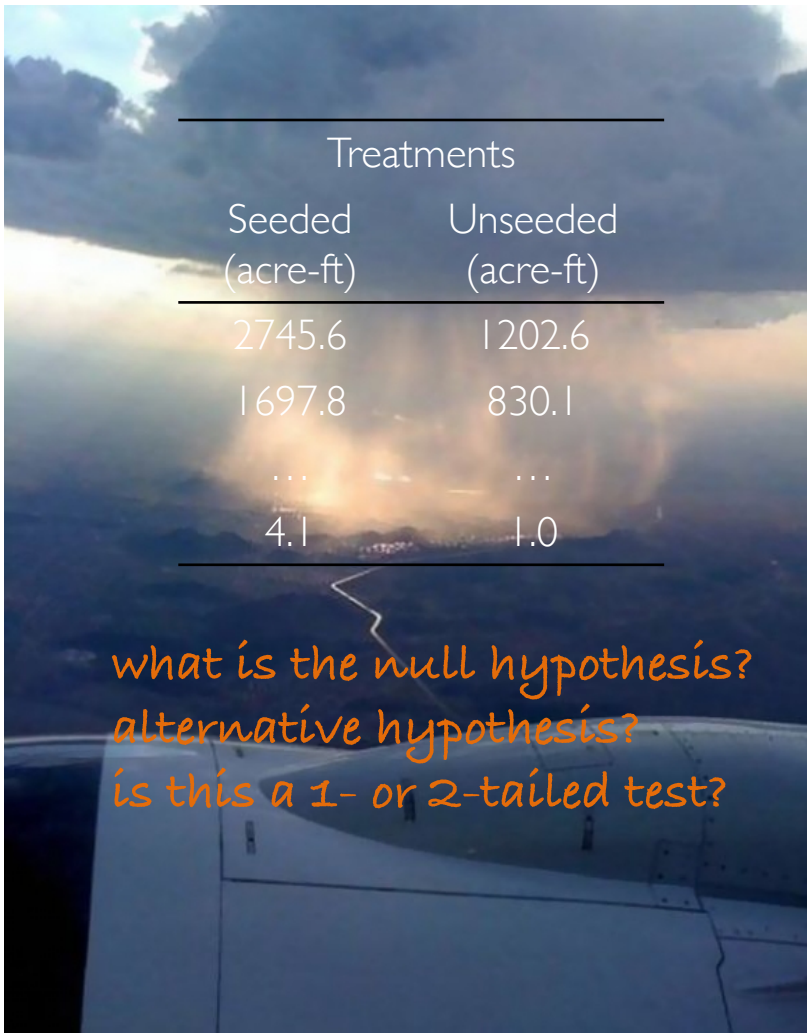
- test hypothesis that injection of silver iodide into cumulus clouds increases rainfall
- treatments randomly chosen over 52 days
 - seed clouds or don't seed clouds with AgI
- measured resulting precipitation as total rain volume



example

cloud seeding

- test hypothesis that injection of silver iodide into cumulus clouds increases rainfall
- treatments randomly chosen over 52 days
 - seed clouds or don't seed clouds with AgI
- measured resulting precipitation as total rain volume

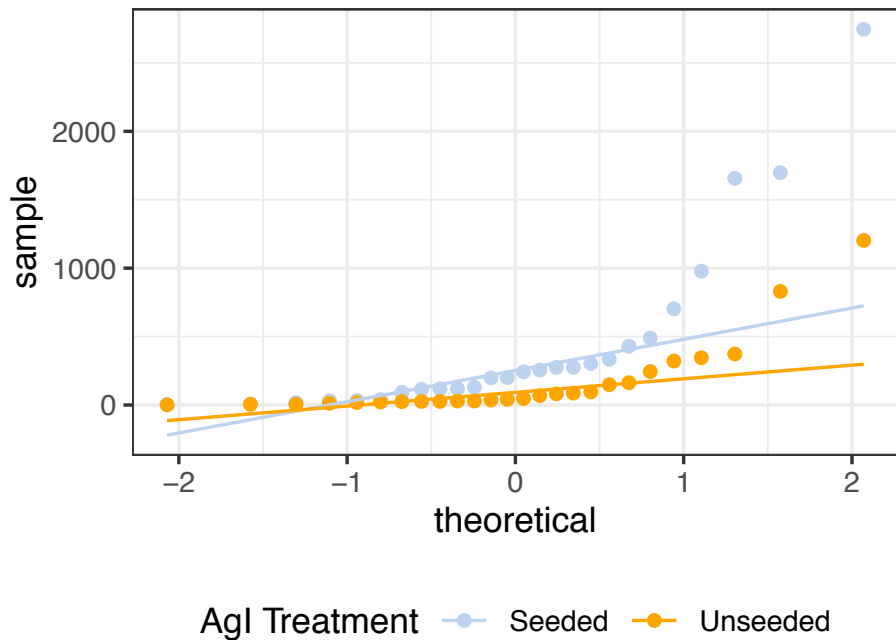


Treatments	
Seeded (acre-ft)	Unseeded (acre-ft)
2745.6	1202.6
1697.8	830.1
...	...
4.1	1.0

what is the null hypothesis?
alternative hypothesis?
is this a 1- or 2-tailed test?

example

cloud seeding



```
var.test(rf1$precip~rf1$treat)
```

F test to compare two variances

```
data: rf1$precip by rf1$treat
F = 5.4633, num df = 25, denom df = 25,
p-value = 6.695e-05
alternative hypothesis: true ratio of
Variances is not equal to 1
95 percent confidence interval:
 2.44959 12.18487
sample estimates:
ratio of variances
 5.463326
```




data transformation

a data transformation is a mathematical function applied to all the observations of a given variable

why transform your data?

- to understand and communicate patterns in data
- meet the assumptions of analyses

if we have a skewed distribution of a sample, a simple transformation often renders the distribution symmetric and may more closely approximate a normal distribution



most common transformation is the logarithm (natural log)

pulls the right tail
(extreme values) in
closer to the mean

possible remedy
for right skewness
or when variation
increases with the
mean

values must be
positive to be log-
transformed

transformation

apply the same mathematical function to all the observations of a sample

Treatments			
Seeded (acre-ft)	Unseeded (acre-ft)	log(Seeded)	log(Unseeded)
2745.6	1202.6	7.92	7.09
1697.8	830.1	7.44	6.72
...
4.1	1.0	1.41	0

```
l.seeded <- log(rfl$precip[rfl$treat == "seeded"])
l.unseeded <- log(rfl$precip[rfl$treat == "unseeded"])
```

review of log rules

logarithm rules

$$\log(\bar{a}_i) \neq \overline{\log(a_i)}$$

$$\log(a/b) = \log(a) - \log(b)$$

$$\log(a \cdot b) = \log(a) + \log(b)$$

$$\log(a^b) = b \cdot \log(a)$$

$$e^{\log(a)} = a$$

back transform by exponentiating

$$\ln(X) \dots e^X$$

$$\log_{10} X \dots 10^X$$

R functions

natural logarithm, ln: `log()`

`log(25)=3.218876; exp(3.218876)=25`

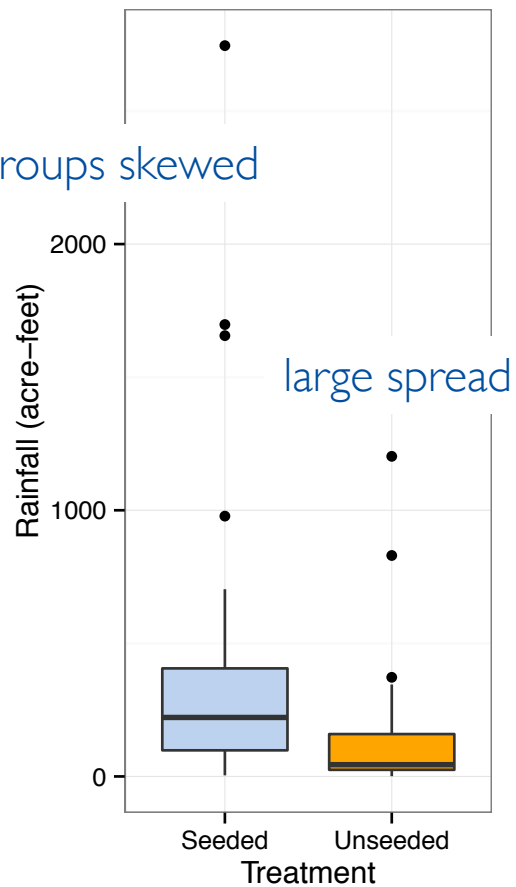
log base 10: `log10()`

`log10(1000)=3; 10^3=1000`

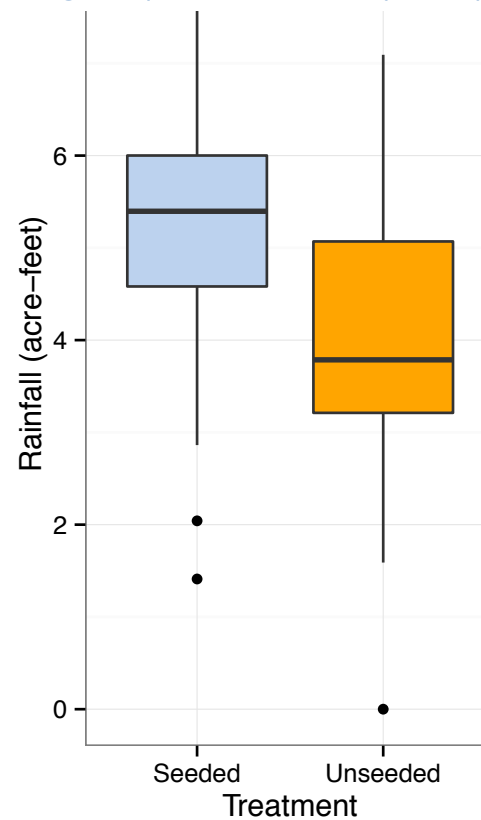
any other base: `log(x, base)`

`log(9, base=3)=2; 3^2=9`

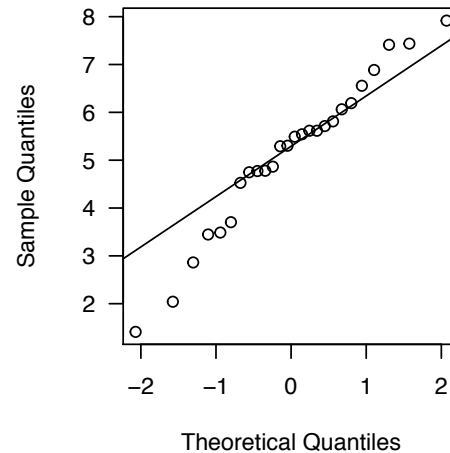
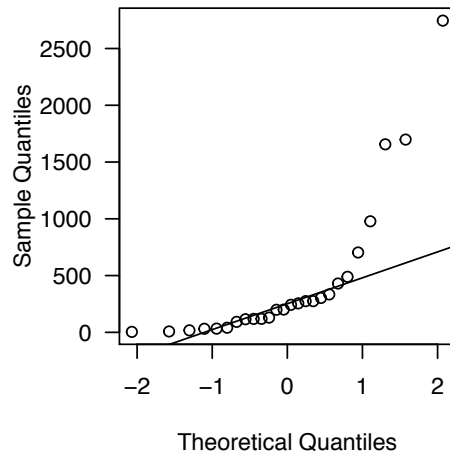
both groups skewed



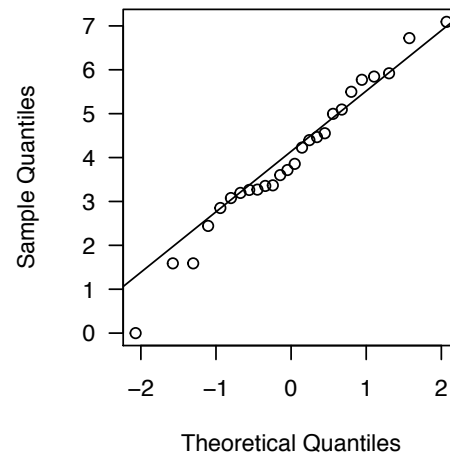
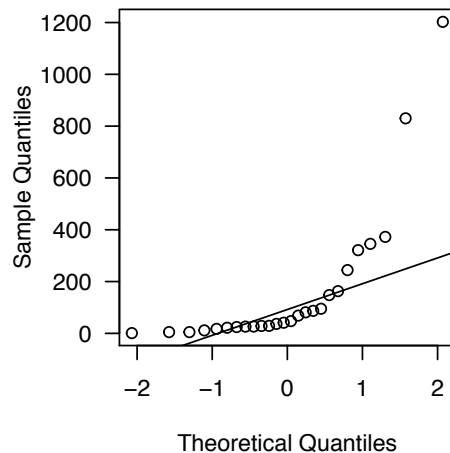
no skewness
groups have ~ equal spread



QQ plot:
seeded and log(seeded)

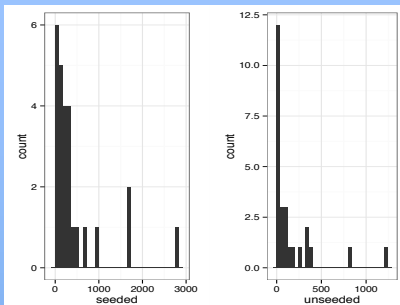


QQ plot
unseeded and log(unseeded)



using the log transform

original scale
acre feet of rainfall



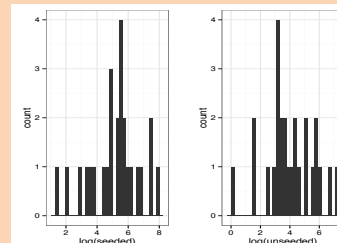
interpret

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

log transform
data



log scale
 $\log(\text{acre feet of rainfall})$



recheck assumptions



conduct analysis

Welch Two Sample t-test

```
data: log(seeded) and log(unseeded)
t=2.5444, df=49.966, p-value=0.007042
95 percent confidence interval:
 0.3903948 Inf
sample estimates:
mean of x mean of y
 5.134187  3.990406
```

back-transform
estimates & CI's



using the log transform

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

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

back-transform



Welch Two Sample t-test

```
data: log(seeded) and log(unseeded)
t=2.5444, df=49.966, p-value=0.007042
95 percent confidence interval:
 0.3903948 Inf
sample estimates:
mean of x mean of y
 5.134187  3.990406
```

t-test results tell us about differences in means of the populations of log response

example

cloud seeding

“There is convincing evidence that seeding increased rainfall (one-sided p -value = 0.007). The mean volume of rainfall produced by a seeded cloud is estimated to be 3.14 times as large as the volume that would have been produced in the absence of seeding (95% confidence: 1.27 to 7.74 times)”

Or, ... “We estimate that seeding a cloud will increase rainfall by 3.14 times that of an unseeded cloud.”

Cloud Seeding: natural log

with natural log-transformed data: $t = 2.54$, $df = 50$, $p = 0.014$

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:
 0.2408650 2.0466972
sample estimates:
mean of x mean of y
 5.134187  3.990406
```

95% CI - log scale
(0.241, 2.047)

95% CI original scale

$$e^{0.241} = 1.27$$

$$e^{2.047} = 7.74$$

note: this is a 2-sided test

$$\frac{1}{n} \sum (\log(y_{\text{seed}}) - \log(y_{\text{unseed}}))$$

$$e^{1.143} = 3.14$$

$$\exp(5.13 - 3.99) = 3.14$$

Cloud Seeding: log10

with log10-transformed data: $t = 2.54$, $df = 50$, $p = 0.014$

Two Sample t-test

```
data: log10(seeded) and  
log10(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:  
 0.1046064 0.8888693  
sample estimates:  
mean of x mean of y  
 2.229749  1.733011
```

95% CI - log scale
(0.104, 0.889)

95% CI original scale

$$10^{0.104} = 1.27$$

$$10^{0.889} = 7.74$$

$$\frac{1}{n} \sum (\log_{10} y_{\text{seed}} - \log_{10} y_{\text{unseed}}) = 0.497$$

$$10^{0.497} = 3.14$$

note: this is a 2-sided test

interpretation of transformed data

- for ease of comprehension, back-transform results of t -test
- transform data, conduct test, back-transform estimate (e.g., differences in averages) by taking `exp(estimate)`
- back transform ends of confidence interval
- report conclusions in original scale

interpretation of transformed data

- back-transformed mean is the geometric mean (median on the original scale)
- interpretation of back-transformed difference of the means of the logs is the **ratio of the geometric** means of the two samples

$$\log(a/b) = \log(a) - \log(b)$$

*back-transformed mean is
the geometric mean and
CI's will not be
symmetrical*

think... before transformation

For count data, our results suggest that transformations perform poorly. An additional problem with regression of transformed variables is that it can lead to impossible predictions, such as negative numbers of individuals. Instead statistical procedures designed to deal with counts should be used, i.e. methods for fitting Poisson or negative binomial models to data. The development of statistical and computational methods over the last 40 years has made it easier to fit these sorts of models...

O'Hara and Kotze

“The first principle for understanding data is that no data have meaning apart from their context... there must always be a link between what you do with the data and the original context for the data. Any transformation of the data risks breaking this linkage. If a transformation makes sense both in terms of the original data and the objectives of the analysis, then it will be okay to use that transformation.”

Dan Wheeler

other common transformations

- square root

- often used in count data
- can't include negative numbers
- moderate effect on distribution shape
- if 0's present in data, add constant so minimum is 1

$$\sqrt{x}$$

- reciprocal

- waiting times and rates (offspring/female)
- interpreted as rate or speed

$$\frac{1}{x}$$

- logit transformation

- used for proportions
(e.g., proportion of trees infected by an insect)

$$\log\left(\frac{x}{1-x}\right)$$

other common transformations

- arcsine
 - proportions (and percentages) $\arcsin \sqrt{x}$
 - data must be proportions
 - do not turn count data into proportions
 - to use this test! proportions from smaller samples will have a higher variance than proportions from larger samples, information that is disregarded by this transformation – use logistic regression.

data transformation

non-parametric alternatives

non-parametric tests

when assumptions of normality are badly violated:

one sample t-test: Sign Test

- compare a sample to a hypothesized value
- allocates a sign (+ or -) to each observation if it is greater or less than the hypothesized value, then evaluates if the number of +'s or -'s are different from what would occur by chance alone

```
library(BSDA)
```

```
SIGN.test(x = data, md = median)
```

non-parametric tests

one sample t-test: Wilcoxon test

```
wilcox.test(seeded-unseeded, mu=0, alternative="greater", conf.int=T)
```

Wilcoxon signed rank test

data: seeded - unseeded

V = 351, p-value = 1.49e-08

alternative hypothesis: true location is greater than 0

95 percent confidence interval:

110.2 Inf

sample estimates:

(pseudo)median

172.95

alternatives

two sample t-test: Mann-Whitney-Wilcoxon Test

- also used when data are ranks, rather than direct measurements
- `wilcox.test(y, x)`

```
wilcox.test(x = seeded, y = unseeded, alternative="greater")
```

paired sample t-test: Wilcoxon Signed-Rank Test

```
wilcox.test(x = seeded, y = unseeded, paired=T)
```

pros and cons of non-parametric tests

- limited assumptions made about the format of the data
- useful for dealing with unexpected, outlying observations
- simple to carry out by hand, for small samples at least
- useful in the analysis of ordered categorical data
- may lack power as compared with more traditional approaches
- nonparametric methods are geared toward hypothesis testing rather than estimation of effects; CI's are not straightforward
- tied values can be problematic when these are common, and adjustments to the test statistic may be necessary

Post your questions to be
answered during lecture