ST 411/511 Outline 3

Reading assignment: Chapter 3. This chapter explores the consequences of violating the assumptions of t-tests.

Chapter 3 Assumptions of the t-Tools

Three assumptions needed for t-test and t confidence interval:

1. Normality — pop'ns normally distributed
2. Equal variance — same std. dev σ (if two pop'ns)
3. Independence

(Statistical) independence: Knowing the value of one observation tells us nothing about where another observation falls in its distribution.

e.g. twins not indep. of each other
But twin pairs indep. of each other.

Case Study 3.1.1 Does cloud-seeding increase the volume of rain, and if so, by how much?

Days are a random sample. This guarantees indep. assumption is met.

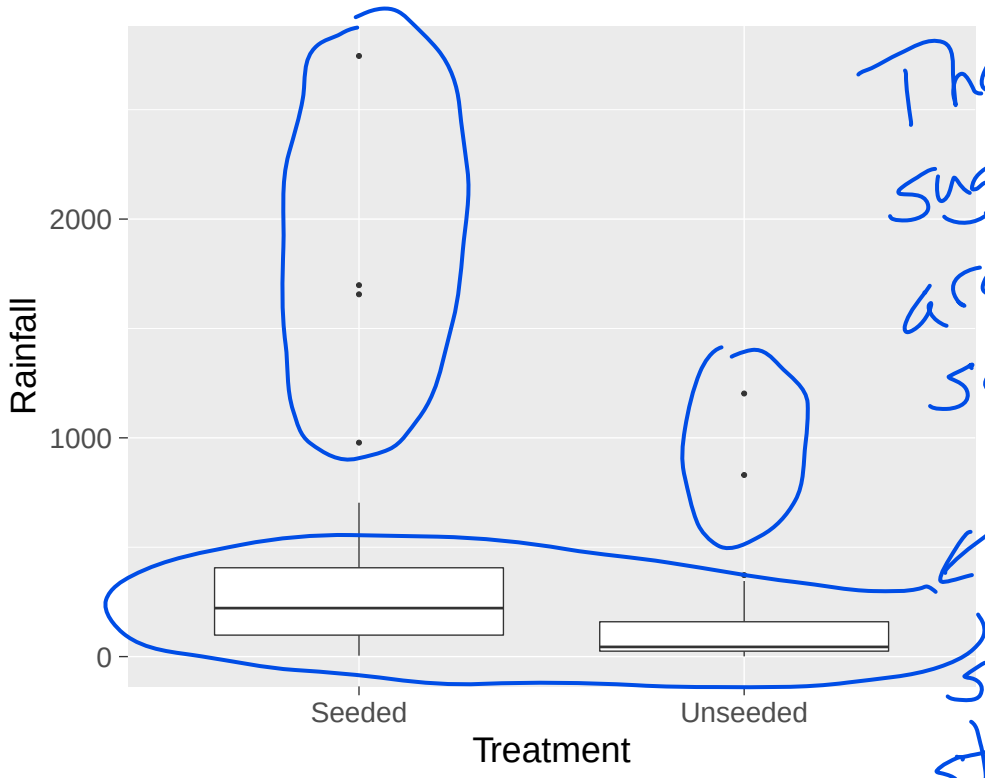
DISPLAY 3.1 Rainfall (acre-feet) for days with and without cloud seeding

Rainfall from Unseeded Days ($n=26$)								
1,202.6	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
26.0	24.4	21.4	17.3	11.5	4.9	4.9	1.0	
Rainfall from Seeded Days ($n=26$)								
2,745.6	1,697.1	1,656.4	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	

```
> library(Sleuth3, ggplot2)  # Load the Sleuth3 and ggplot2 packages.
> head(case0301) # View the first few lines of the data frame.
  Rainfall Treatment
1   1202.6  Unseeded
2    830.1  Unseeded
3    372.4  Unseeded
4    345.5  Unseeded
5    321.2  Unseeded
6    244.3  Unseeded
>
> qplot(Treatment, Rainfall, data=case0301, geom="boxplot")
```

response

grouping var.



These p/s suggest popns are skewed, so not normal

Heights different, suggesting popn std. devs are different

```
> with(case0301, summary(Rainfall[Treatment=="Seeded"]))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   4.10   98.13  221.60  442.00  406.00 2746.00
> with(case0301, summary(Rainfall[Treatment=="Unseeded"]))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00   24.82   44.20  164.60  159.20 1203.00
```

True popn quantities for

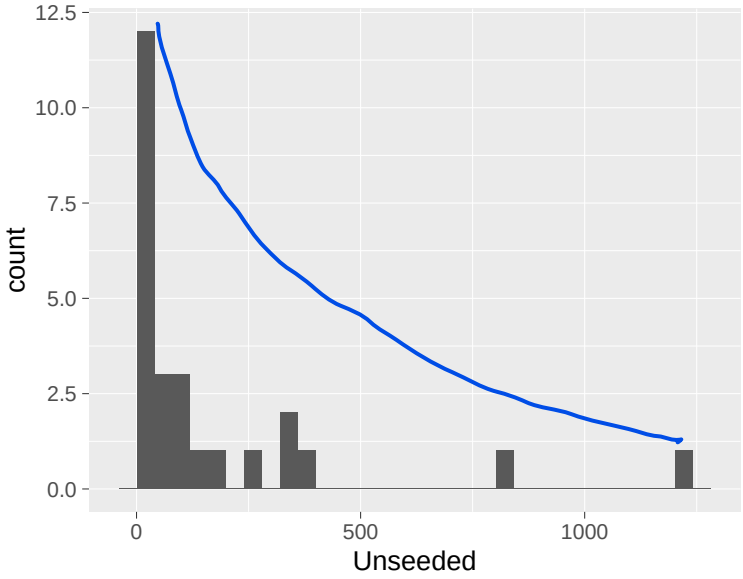
For symmetric distns, mean = median.

For the two samples, means & medians are very different, suggesting skewness, so not normal.

Log Transformations Useful for data that are multiplicative rather than additive. Such data exhibit skewness and unequal standard deviation.

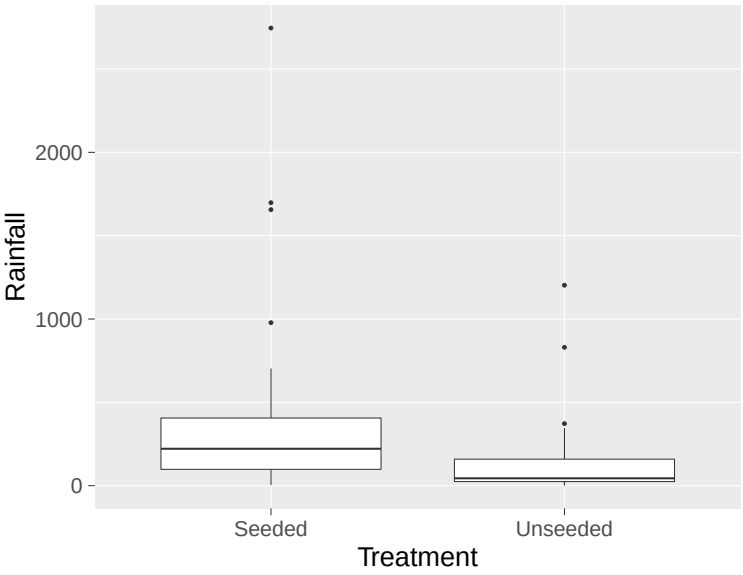
Indications:

- Right-skewed distributions



long right tail

- Larger variance (variance = standard deviation^2) associated with larger mean.



seeded data have larger rainfall and are more spread out.

Warning: Do not use a log transformation indiscriminately.

Need pop'n dist'n of logged quantities to be approx. symmetric. Also, can't take log of 0 or neg. #s.

Example of analysis with log transformation

response

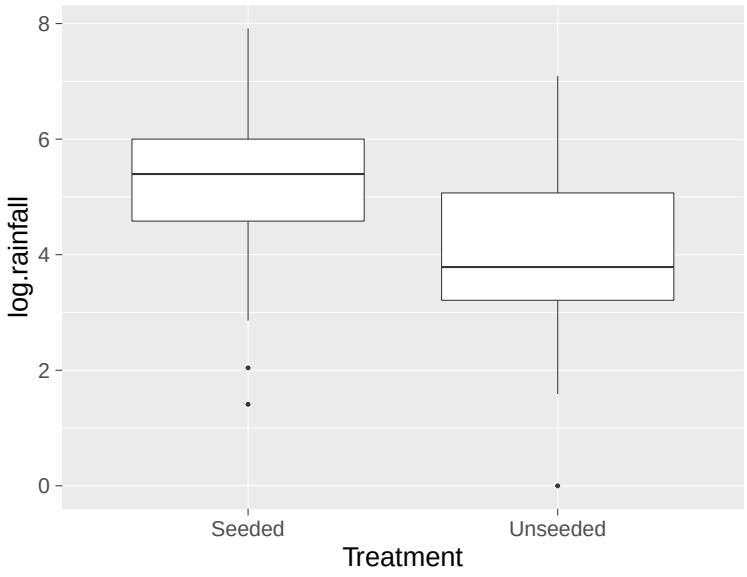
```
> # Log-transform the rainfall data.

> log_rainfall <- log(case0301$Rainfall) # log() is natural log.
```

ln

binds columns

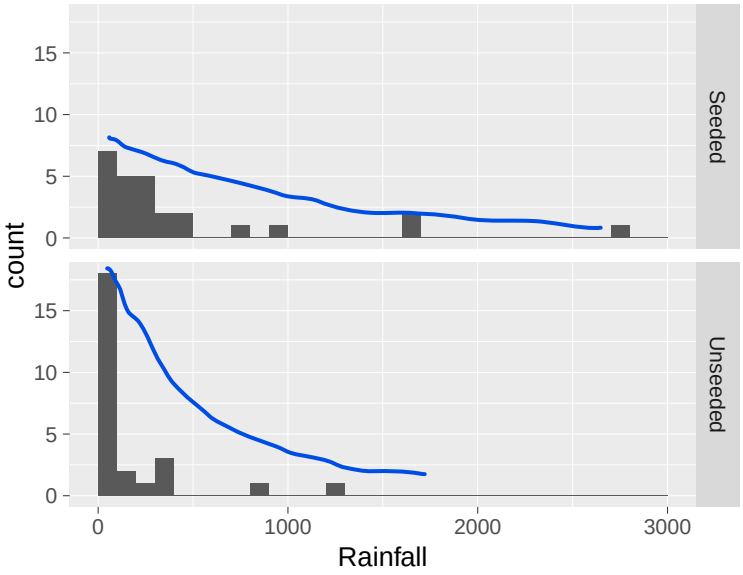
```
> head(cbind(case0301, log_rainfall))
  Rainfall Treatment log_rainfall
1   1202.6  Unseeded     7.092241
2    830.1  Unseeded     6.721546
3    372.4  Unseeded     5.919969
4    345.5  Unseeded     5.844993
5    321.2  Unseeded     5.772064
6    244.3  Unseeded     5.498397
```

```
> qplot(Treatment, log_rainfall, data=case0301, geom="boxplot")
```



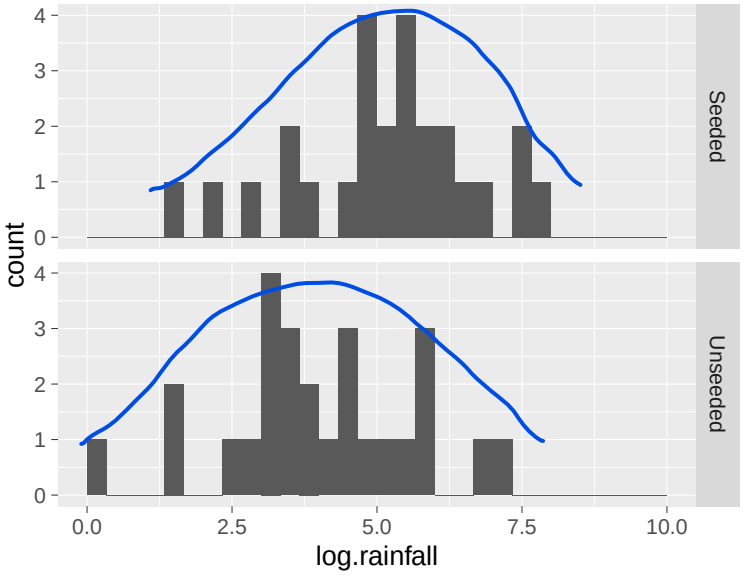
This is a big improvement.

```
> qplot(Rainfall, data=case0301, geom="histogram", xlim=c(0,3000)) +
+                 facet_grid(Treatment ~ .)
```



Long tails

```
> qplot(log_rainfall, data=case0301, geom="histogram", xlim=c(0,10)) +
+                     facet_grid(Treatment ~ .)
```



Dist'ns much more symmetric, spreads similar, so t-tool assumptions are more reasonable now.

```
> summary(case0301$Treatment)
  Seeded Unseeded
      26       26
```

Does seeding cause more rain? One-sided alternative. *

μ_1 = pop'n mean log rainfall for seeded days

μ_2 = " " unseeded "

$H_0: \mu_1 - \mu_2 = 0$ $H_A: \mu_1 - \mu_2 > 0$

```
> t.test(log_rainfall~Treatment ,data=case0301, var.equal=TRUE,
+        alternative="greater")

Two Sample t-test

data:  log_rainfall by Treatment
t = 2.5444, df = 50, p-value = 0.007041
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.3904045       Inf
sample estimates:
  mean in group Seeded mean in group Unseeded
              5.134187                3.990406
```

* since days randomly assigned to groups.

Not-quite-right statistical conclusion for hypothesis test:

Stong evidence that seeding causes * an increase in mean log rainfall ($p \approx 0.007$, one-sided t-test)

Need to change back to original units.

Getting a 2-sided CI.

```
> t.test(log_rainfall~Treatment, data=case0301, var.equal=TRUE)

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:
 0.240865 2.046697
sample estimates:
  mean in group Seeded mean in group Unseeded
              5.134187                3.990406
```

Not-quite-right statistical conclusion for confidence interval:

We estimate diff in mean log rainfall between seeded & unseeded days is 0.2409 to 2.047 log acre-feet (95% CI)

Need to transform back to orig. units.

Background for interpretating CI after an (appropriate) log transformation

Recall: Natural log and exponential function are inverses.

For example,

$$\log(1202.6 \text{ acre-feet}) = 7.0922 \log \text{ acre-feet}$$

$$e^{7.0922 \log \text{ acre-feet}} = 1202.6 \text{ acre-feet}$$

Exponentiate to transform back from log scale to original scale. (Item 3(h), Lab 3)

Recall: Addition and subtraction on the log scale correspond to multiplication and division on the original scale.

Subtraction on log scale

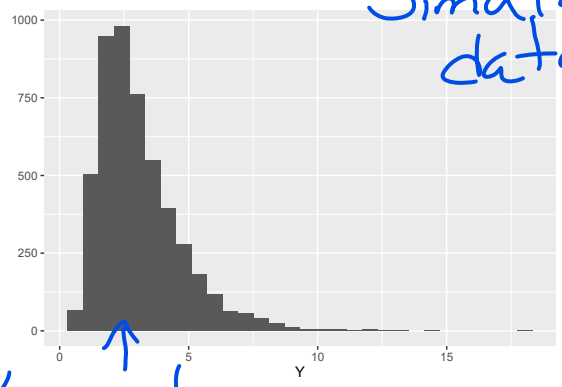
$$e^{5.134187 - 3.990406} = \frac{e^{5.134187}}{e^{3.990406}}$$

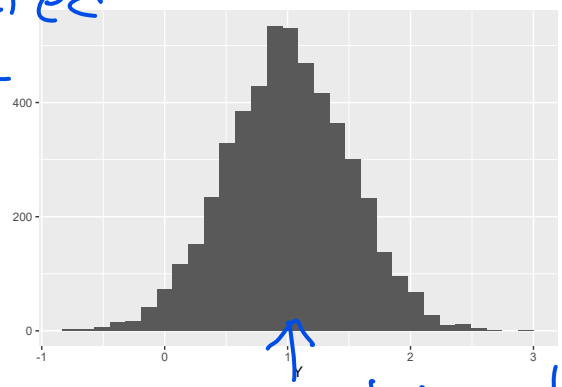
Division on orig. scale

Diff. on log scale → Ratio on orig. scale

Recall: The population *median* is the point where half the distribution is above and half is below. The population *mean* is the balancing point of the distribution.

Simulated data





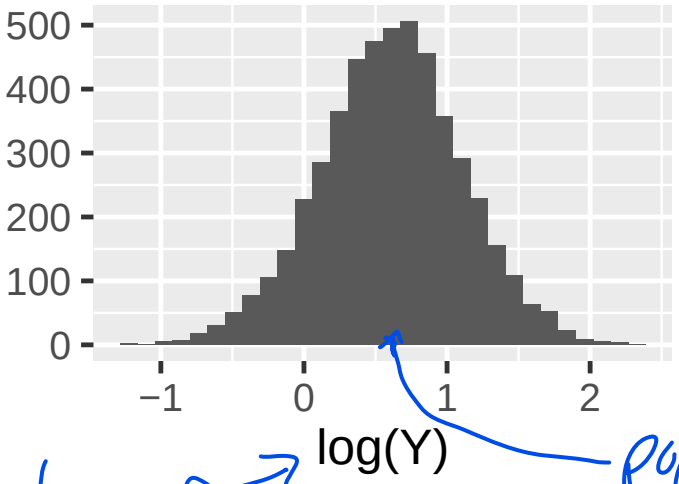
popn median ≈ 2.5
popn mean ≈ 3.0

Symmetric distn
popn mean = popn median

$Y_1 < Y_2$

Recall: Exponentiation preserves order: If $\log(Y_1) < \log(Y_2)$, then $e^{\log(Y_1)} < e^{\log(Y_2)}$.

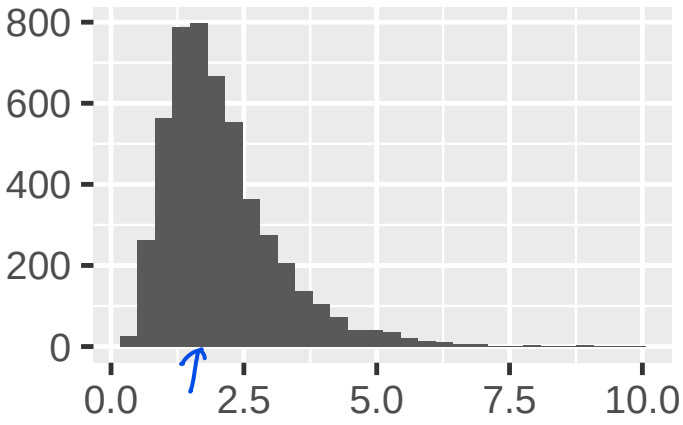
Simulated data



So median stays the same when we back-transform

log-transformed to symmetrize distn

pop'n mean = pop'n median = 0.6



pop'n median $= e^{0.6} \approx 1.8$

exponentiate logged data to recover original scale.

Inference on orig. scale is about pop'n medians

Reporting statistical conclusion of hypothesis test on page 6, Outline 3:

```
> t.test(log.rainfall~Treatment, data=case0301, var.equal=TRUE,
+        alternative="greater")

Two Sample t-test

data:  log.rainfall by Treatment
t = 2.5444, df = 50, p-value = 0.007041
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.3904045       Inf
sample estimates:
  mean in group Seeded mean in group Unseeded
              5.134187                3.990406
```

Same output as before

Statistical conclusion: for hypothesis test

Stong evidence that seeding causes* an increase in median rainfall ($p \approx 0.007$, one-sided t-test).

* because days randomized to treatments.

Reporting statistical conclusion of confidence interval on page 7, Outline 3:

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

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:
 0.240865 2.046697
sample estimates:
  mean in group Seeded mean in group Unseeded
              5.134187                3.990406

> 5.134187 - 3.990406 # Point estimate of difference in mean log rainfall
[1] 1.143781

> exp(5.134187 - 3.990406) # Point estimate of ratio of medians
[1] 3.138613

>  exp(5.134187)/exp(3.990406) # Same thing
[1] 3.138613

> exp(c(0.240865, 2.046697)) # Confidence interval for ratio of medians
[1] 1.272349 7.742286
```

Same output as on p.7

pt. est. of $\mu_1 - \mu_2$ pop'n means on log scale

e^{μ_1} & e^{μ_2} = pop'n medians on orig. scale.

pt. est. of $\frac{e^{\mu_1}}{e^{\mu_2}}$ = ratio of pop'n medians

CI for $\mu_1 - \mu_2$ on log scale.

pop'n

on orig. scale

Statistical conclusion: It is estimated that the ratio of pop'n median** rainfall on seeded days to unseeded days is 3.14* (95% CI 1.27 to 7.74).

* ratio has no units

** ratio of medians because of log

Note: We estimated that median rainfall on seeded days is 3.14 times median on unseeded days.

Steps to back-transform from log scale in two-sample t-test

1. Exponentiate difference in sample means to get a point estimate of the ratio of population medians.

$e^{\bar{Y}_1 - \bar{Y}_2}$ = pt. est. of $\frac{\text{median in popn 1}}{\text{median in popn 2}}$

2. Exponentiate endpoints of confidence interval to get them on original scale.

Conf. level and p-value stay the same

3. Confidence interval estimates the ratio of population medians

Stat conclusion reports an est. of this ratio of medians

4. Paraphrase of alternative hypothesis should be in terms of population medians, not population means.

Stat conclusion for hyp test doesn't mention popn means.

Reiterating assumptions for "t-tools" (t-tests and t-based confidence intervals):

- Normal population(s)
- Equal standard deviation
- Statistical independence

Robustness: A procedure is *robust* to departures from a certain assumption if the results are valid even when the assumption is not met.

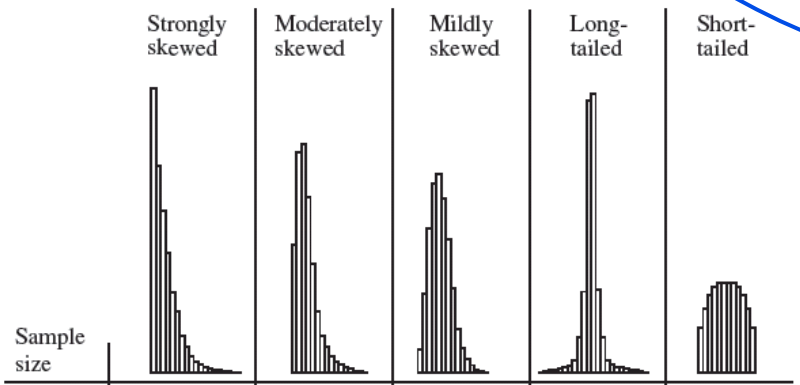
t-tests & t-CI's have good robustness properties.

Can use simulation to assess robustness

Normality assumption

Two pop'n, same std. dev. but not normal.

DISPLAY 3.4 Percentage of 95% confidence intervals that are successful when the two populations are non-normal (but with same shape and SD, and equal sample sizes) (each percentage is based on 1,000 computer simulations)



Sample size	Strongly skewed	Moderately skewed	Mildly skewed	Long-tailed	Short-tailed
5	95.5	95.4	95.2	98.3	94.5
10	95.5	95.4	95.2	98.3	94.6
25	95.3	95.3	95.1	98.2	94.9
50	95.1	95.3	95.1	98.1	95.2
100	94.8	95.3	95.0	98.0	95.6

Successful CI contains $\mu_1 - \mu_2$

95% CI should be successful about 95% of the time

Skewness not a problem as long as dist'ns are the same shape & have same std. dev.

These intervals are too wide

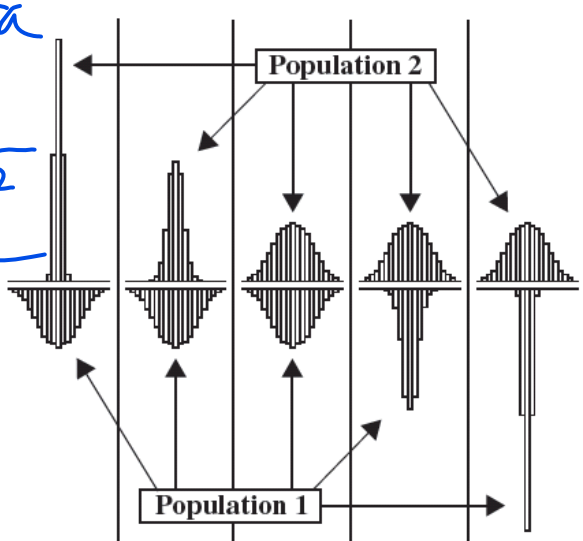
These are fine too.

Equal standard deviation assumption

DISPLAY 3.5 Percentage of successful 95% confidence intervals when the two populations have different standard deviations (but are normal) with possibly different sample sizes (each percentage is based on 1,000 computer simulations)

Simulated data

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$



n_1	n_2	$\sigma_2/\sigma_1 = 1/4$	$\sigma_2/\sigma_1 = 1/2$	$\sigma_2/\sigma_1 = 1$	$\sigma_2/\sigma_1 = 2$	$\sigma_2/\sigma_1 = 4$
10	10	95.2	94.2	94.7	95.2	94.5
10	20	83.0	89.3	94.4	98.7	99.1
10	40	71.0	82.6	95.2	99.5	99.9
100	100	94.8	96.2	95.4	95.3	95.1
100	200	86.5	88.3	94.8	98.8	99.4
100	400	71.6	81.5	95.0	99.5	99.9

Success rates for 95% intervals

Problems when samples sizes are different and pop'n std. devs. are different.

If sample sizes are close to equal, then we get good results, even when pop'n std. dev's are not equal.

Coming up in Ch. 4: Welch's t-test, an approx. of t-test that doesn't require equal std. dev assumption.

Independence assumption

Examples of **dependent** data: Twins

Clustered observations - families, households

Spatial or temporal dependence

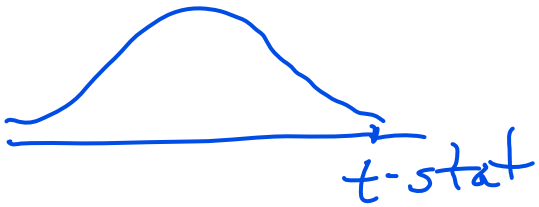
Extreme example: Measure the same subject n times (with no measurement error).

All obs equal, so sample std dev. is $s = 0$

If data are positively dependent, i.e. not identical but from similar subjects, sample standard deviation s will be too small, so standard error will be too small.

$\text{SE}(\overline{Y}) = s/\sqrt{n}$ $= 0$ in extreme example or $SE(\overline{Y})$ too small.

$t_{\text{stat}} = \dfrac{\overline{Y} - \text{value under } H_0}{\text{SE}(\overline{Y})}$ ← bigger than it should be, so p-value too small



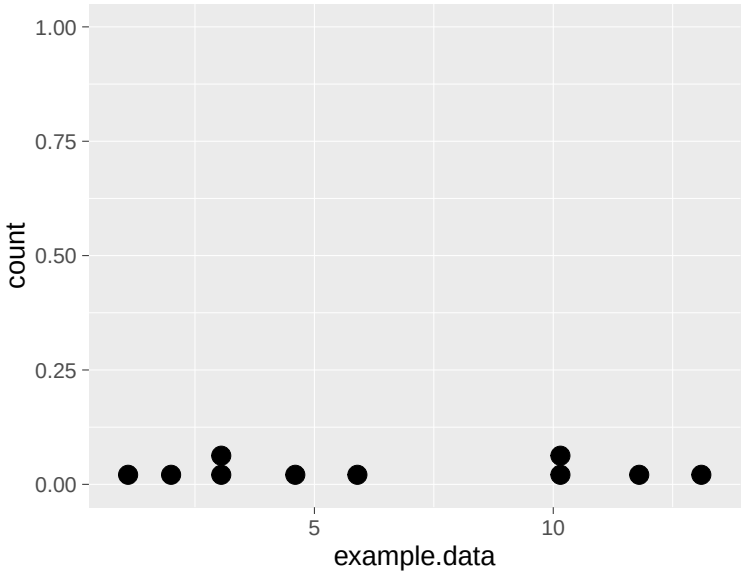
$(1-\alpha)\ 100\%\ \text{CI} = \overline{Y} \pm t_{\text{df}(1-\alpha/2)}\text{SE}(\overline{Y})$

margin of error too small so CI's too precise.

A random sample from popn satisfies indep. assumption.

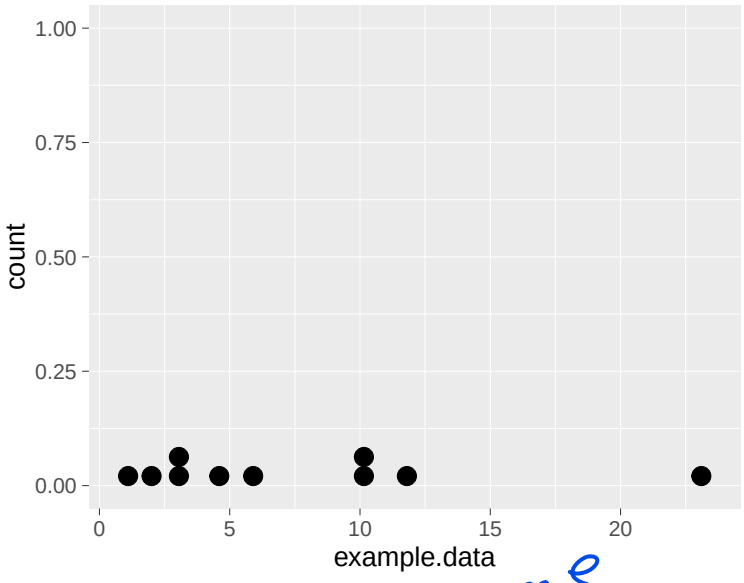
Resistance vs Robustness: A procedure is *resistant* if the results don't change much if a small part of the data changes, even if it changes by a lot.

```
> example.data <- c(1.1,2.0,3.0,3.1,4.6,5.9,10.1,10.2,11.8,13.1)
> qplot(example.data, geom="dotplot")
```



Changed one number

```
> summary(example.data)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.100   3.025   5.250   6.490  10.170  13.100
>
> example.data <- c(1.1,2.0,3.0,3.1,4.6,5.9,10.1,10.2,11.8,23.1)
> qplot(example.data, geom="dotplot")
```



Median is resistant but mean is not. t-tools base on means so they are generally not resistant.

same

```
> summary(example.data)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.100   3.025   5.250   7.490  10.170  23.100
```

Suggestions for data with a few extreme values

- Try an analysis with and without extreme points to see if conclusions are substantially different.

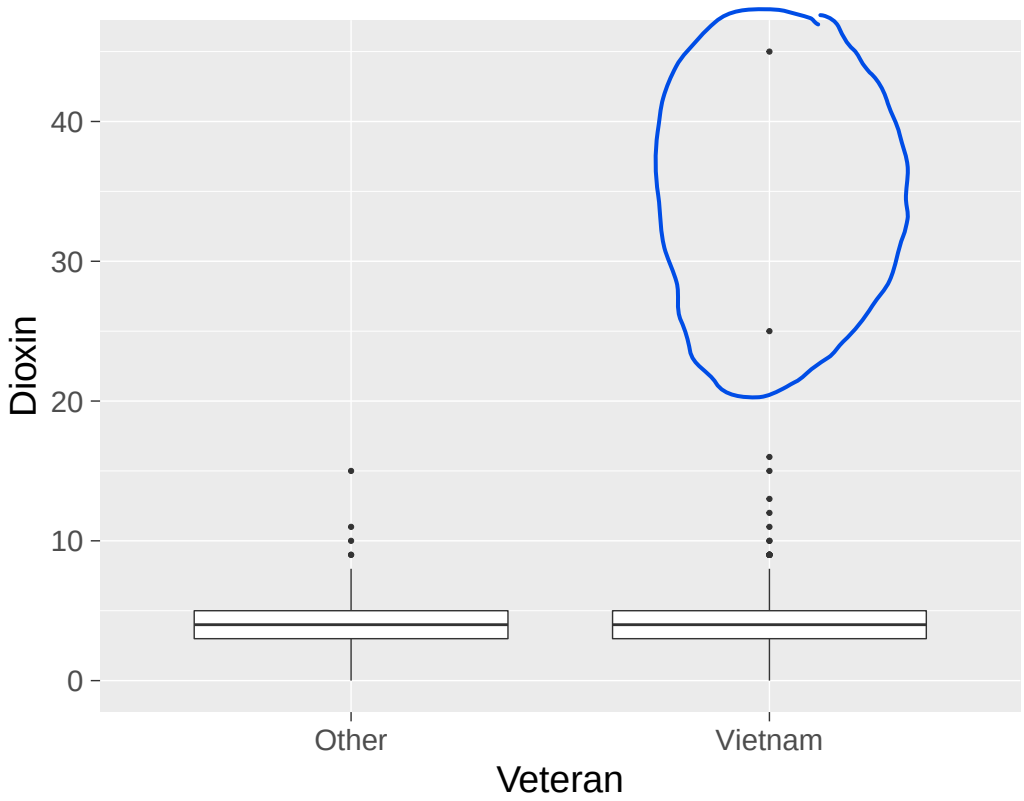
You might get a similar quantification of evidence in either case.

- If conclusions are very different, is there any justification for removing those points from the analysis?

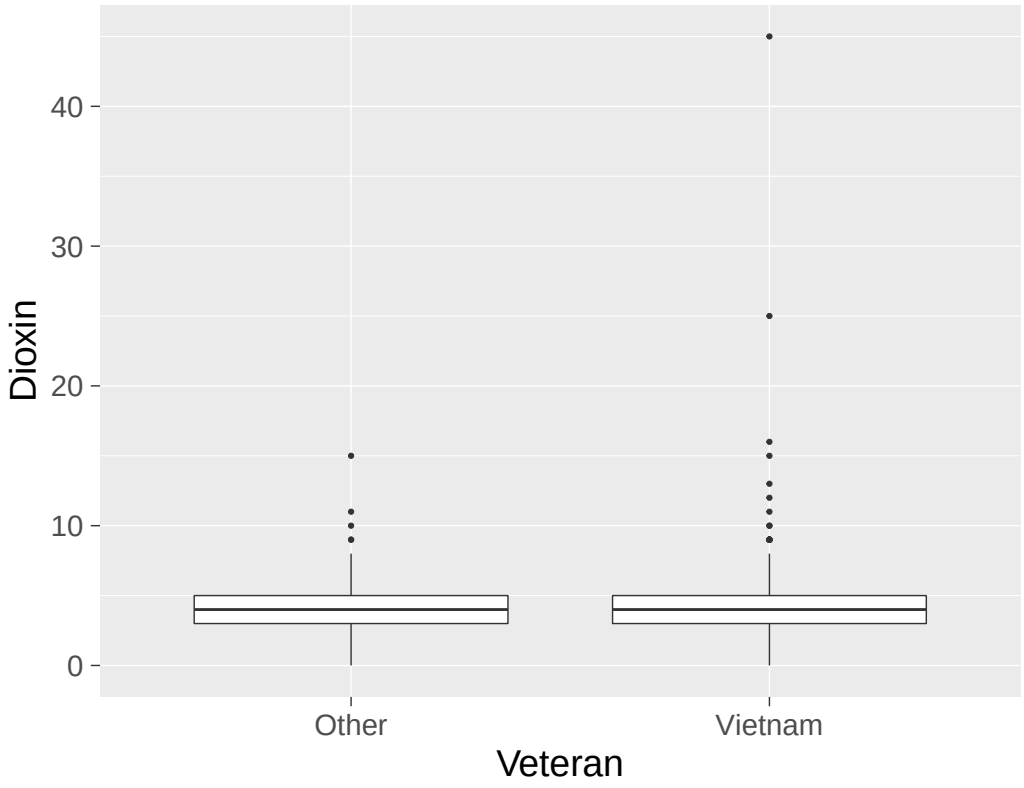
Are they errors?
Are they from a diff. pop'n?
Can remove them, but this limits scope of inference.
Must clearly state what data you removed and why.

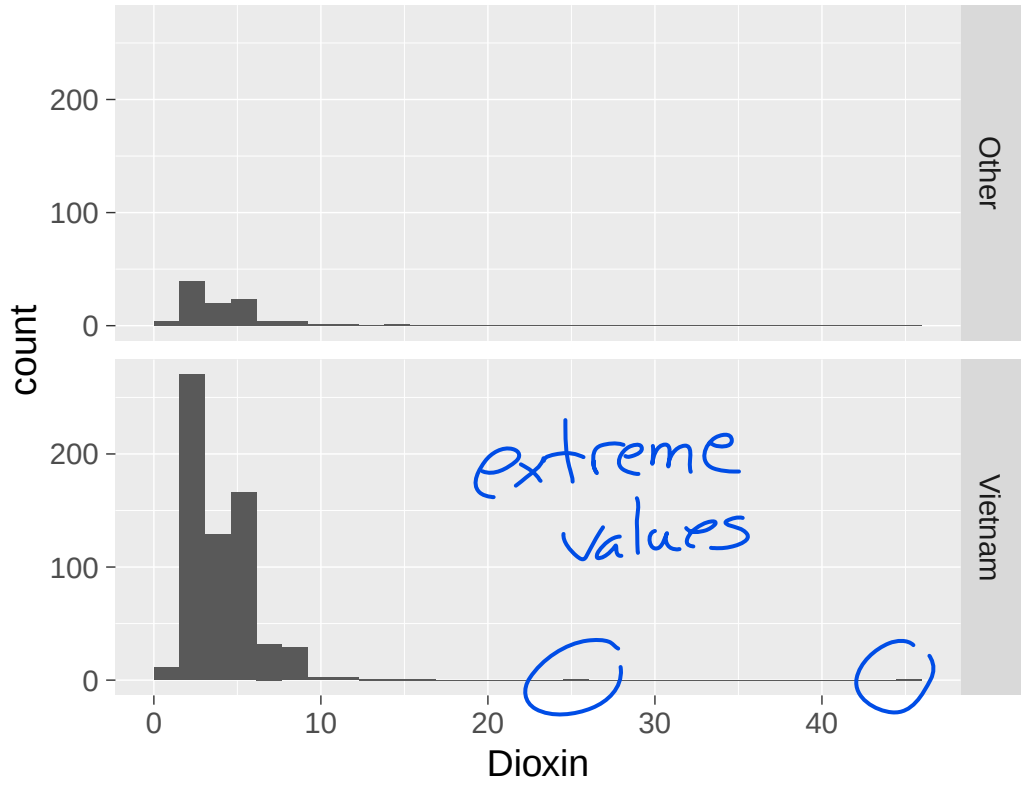
Case Study 3.1.2 Are blood dioxin levels higher in Vietnam veterans compared to veterans of other wars?

```
> head(case0302)
  Dioxin Veteran
1      0 Vietnam
2      0 Vietnam
3      0 Vietnam
4      0 Vietnam
5      0 Vietnam
6      1 Vietnam
>
> qplot(Veteran, Dioxin, data=case0302, geom="boxplot")
```



If we ignore the very large values, normality & equal std. dev assumptions look OK.


Dioxin
40
30
20
10
0
Other
Vietnam
Veteran


count
200
100
0
Other
Vietnam
extreme values
0
10
20
30
40
Dioxin

H_0 : population mean dioxin in blood is the same for Vietnam vets as for others.

```
> summary(case0302$Veteran) # Check R's ordering of the groups.
  Other Vietnam
     97     646
> t.test(Dioxin~Veteran,data=case0302,var.equal=TRUE,
+        alternative="less")

Two Sample t-test

data:  Dioxin by Veteran
t = -0.26302, df = 741, p-value = 0.3963
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 0.391951
sample estimates:
  mean in group Other mean in group Vietnam
             4.185567              4.260062

> t.test(Dioxin~Veteran,data=case0302,var.equal=TRUE,
+        alternative="less",subset=-646)

Two Sample t-test

data:  Dioxin by Veteran
t = -0.048902, df = 740, p-value = 0.4805
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
    -Inf 0.37031
sample estimates:
  mean in group Other mean in group Vietnam
             4.185567              4.196899

> t.test(Dioxin~Veteran,data=case0302,var.equal=TRUE,
+        alternative="less",subset=-c(646,645))

Two Sample t-test

data:  Dioxin by Veteran
t = 0.096911, df = 739, p-value = 0.5386
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf 0.3773514
sample estimates:
  mean in group Other mean in group Vietnam
             4.185567              4.164596
```

no evidence against H_0

all data

no evidence

take out largest pt.

no evidence

take out both pts

Summary

- The t-tools make three assumptions:
  1. Normally-distributed population(s)
  2. Equal standard deviations (if more than one population)
  3. Statistically independent observations
- The t-tools are robust under violations of some of the assumptions, and simulation studies help identify when t-tools fail.

normality & equal std. dev.

- In general, symmetric distributions, equal sample sizes, and large samples are situations where you can trust inference with t-tools.

- Log transformation can help with some populations (right-skew, larger standard deviation with larger mean).

logged distns → symmetric

- If log transformation is appropriate, inference about mean on log scale becomes inference about median on original scale.

diff in means on log scale → ratio of medians on orig. scale.

- If assumptions violated because data contain extreme points, try analysis with and without these points, but if you exclude data, you must report that.