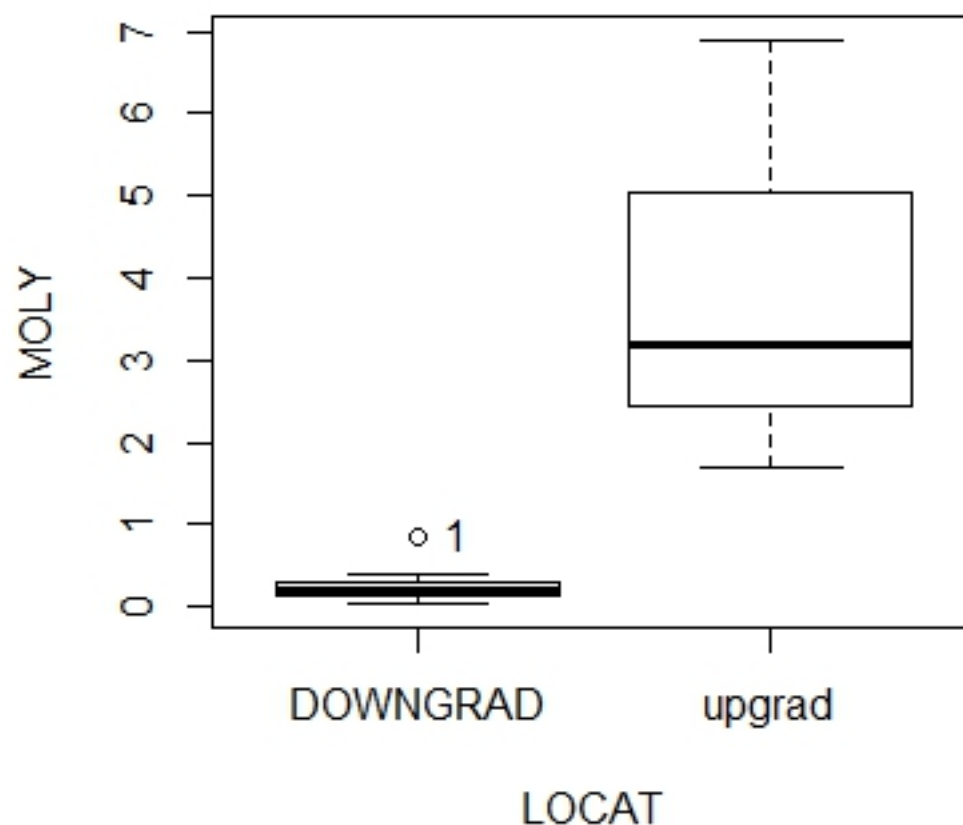


Permutation Homework Week 1

t-test versus Permutation test for
Molybdenum data

Homework: the MOLY2 Dataset



All of the upgrad data are higher than all of the DOWNGRAD data

The DOWNGRAD data are non-normal. The upgrad data may be, but there are only 3 observations so it cannot be tested with any power

1. Two sample t-test

Null Hypothesis: $\text{mean } X = \text{mean } Y$

Alternate Hyp: $\text{mean } X \neq \text{mean } Y$ (2-sided)

or

$\text{mean } X > \text{mean } Y$ (1-sided)

Assumptions: Each group's data follows a normal dist.
Each group's data have same variance

1. t-test

```
> t.test(MOLY~LOCAT)
```

Welch Two Sample t-test

data: MOLY by LOCAT

t = -2.3836, df = 2.0057, p-value = 0.1396

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

95 percent confidence interval:

-10.321151 2.949254

sample estimates:

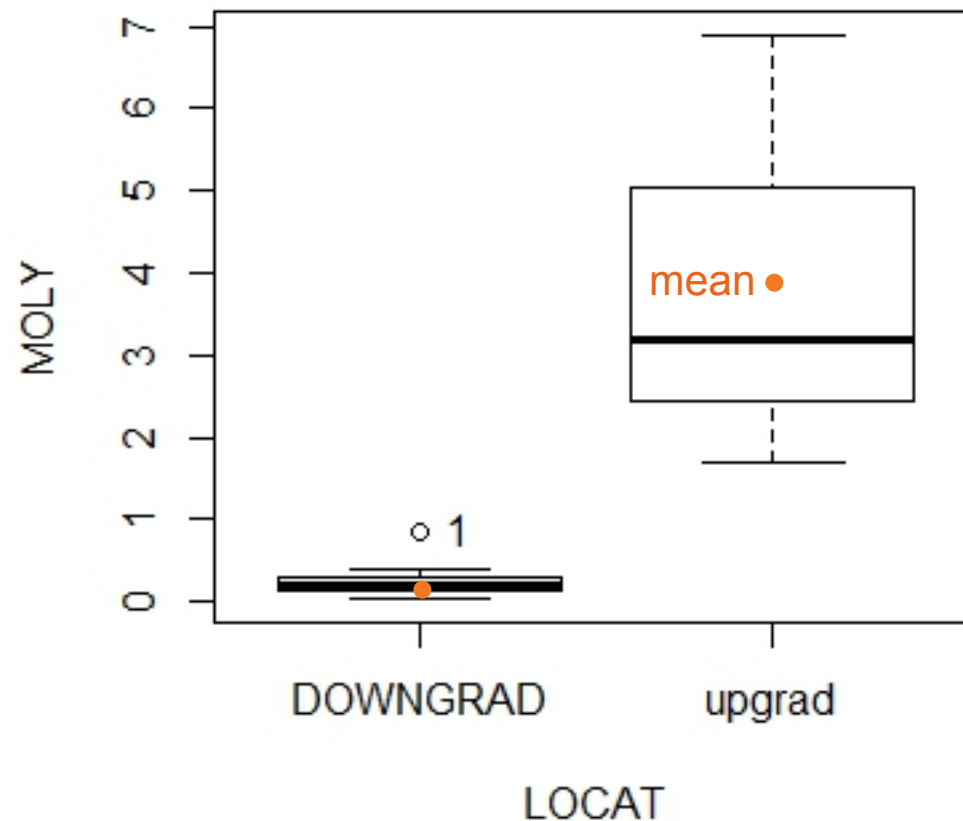
mean in group DOWNGRAD

0.2473846

mean in group upgrad

3.9333333

The means weren't found to significantly differ
using the t-test



4 vs $\frac{1}{4}$ certainly
looks different!

Did the non-normal
DOWNGRAD data
lessen the power of
the t-test?

Did the very unequal
variances of the two
groups lessen the
power of the t-test?

2. The t-test in log units

Difference in the logarithms $\overline{\ln X} - \overline{\ln Y}$

is a ratio of the geometric means
in the original units: $\frac{\text{geometric mean } X}{\text{geometric mean } Y}$

The t-test on logs no longer tests differences in means of data, but whether the ratio of medians (geo means) equals 1.

2. The t-test in log units (test for geometric means)

```
> t.test(log(MOLY)~LOCAT)
```

Welch Two Sample t-test

data: log(MOLY) by LOCAT

t = -6.1927, df = 3.5896, p-value = 0.004907

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

95 percent confidence interval:

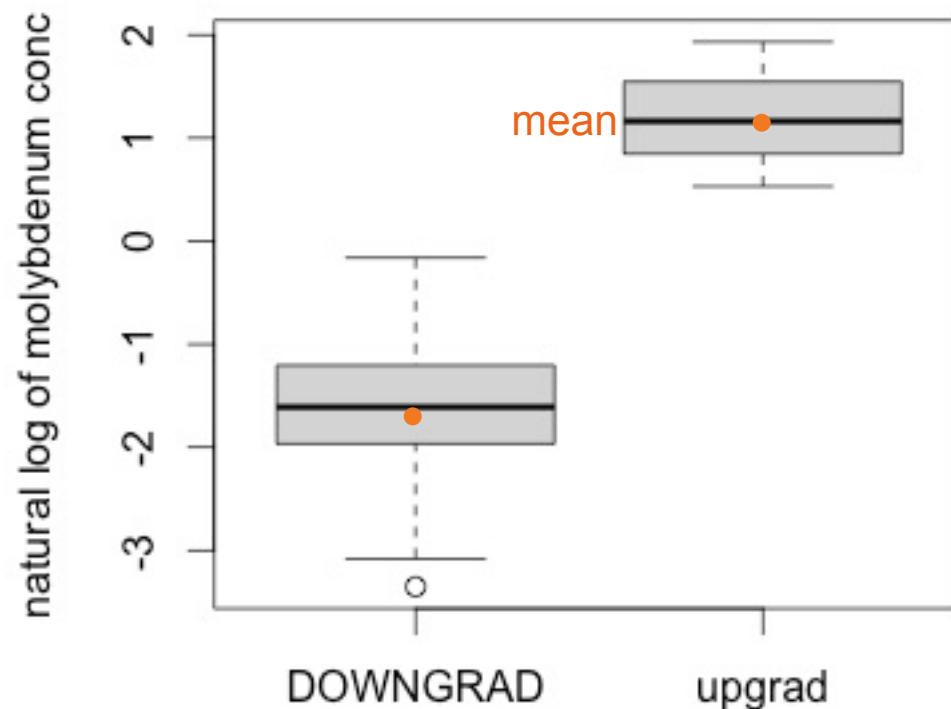
-4.287707 -1.548671

sample estimates:

mean in group DOWNGRAD	mean in group upgrad
------------------------	----------------------

-1.709755	1.208433
-----------	----------

2. The t-test in log units (test for geometric means)



The mean logs (geometric means) do significantly differ.

The means in original units did not. This shows that these are two entirely different tests.

Two-Sample Permutation Test on Means

H_0 : Mean Group 1 = Mean Group 2

If H_0 is true the data could be randomly reassigned to either group.

Take the test statistic to be the difference in the means.

“Shuffle” the SITE names.

Compute the difference in the means for each shuffle. Compute the percent of results equal to or more extreme than the one observed in your data.

4. Two-Sample Permutation Test

First approach: compute all possible results.

For 2 groups of size N and M there are $(N+M)!/(N!*M!)$ distinct rearrangements.

For $N = 13$ and $M = 3$ there are

```
> choose(16,3)
```

```
[1] 560
```

distinct combinations. Compute all 560.

4. Two-Sample Permutation Test

First approach, cont.: compute all possible results.

The perm package will compute exact results with the method="exact.ce" option, and tsmethod="abs"

```
> permTS(MOLY~LOCAT,method='exact.ce',  
control=permControl(tsmethod='abs'))
```

Exact Permutation Test (complete enumeration)

data: MOLY by LOCAT

p-value = 0.001786

alternative hypothesis: true mean LOCAT=DOWNGRAD - mean
LOCAT=upgrad is 0

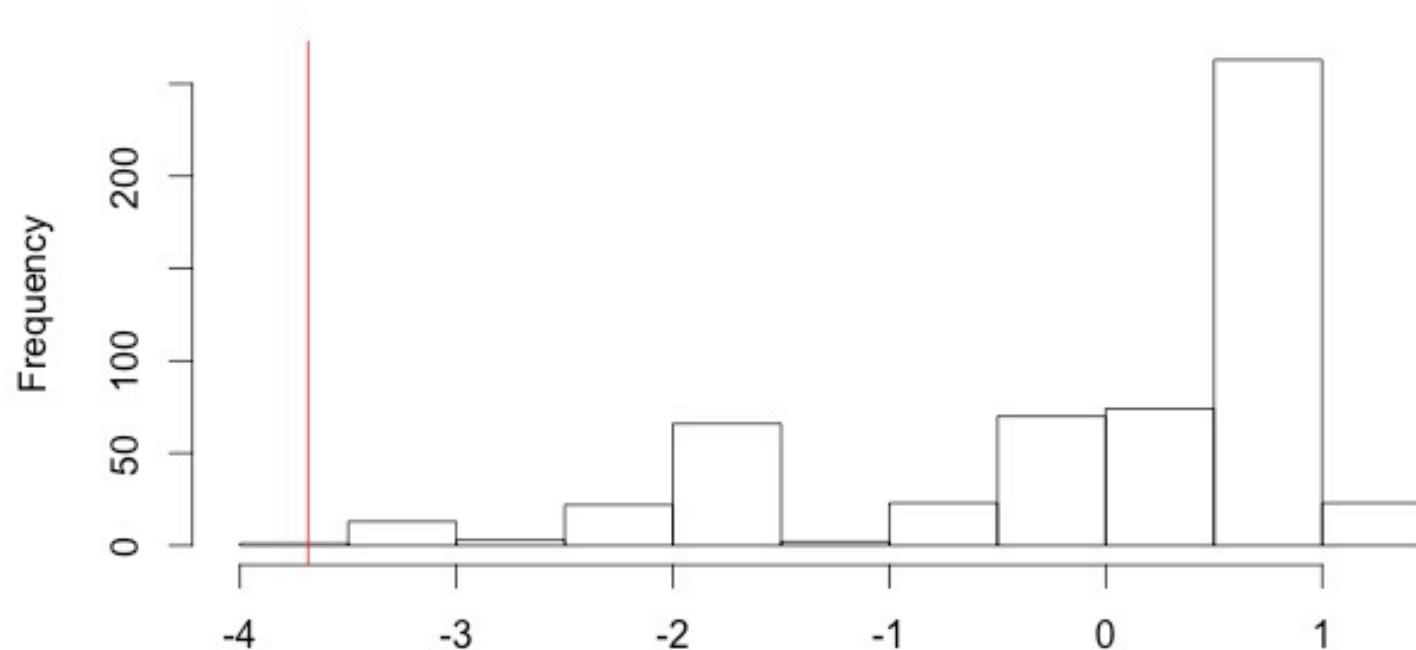
sample estimates:

mean LOCAT=DOWNGRAD - mean LOCAT=upgrad
-3.685949

4. Two-Sample Permutation Test

Histogram of differences looks very unlike the t-test's assumed normal distribution, showing why the t-test had low power to find differences.

DOWNGRAD not = upgrad , pvalue = 0.0018 , nrep = 560



Permutation estimates. Observed difference shown as red line

Comparing 3 or More Groups

Analysis of Variance

versus

Permutation Tests

Expansion of 2-gp tests to 3+ groups

Two groups

1. t-test
2. t-test on logs
3. Wilcoxon rank-sum
4. permutation
(permTS)

3+ groups

- ANOVA
- ANOVA on logs
- Kruskal-Wallis
- permutation
(permKS)

1. ANOVA

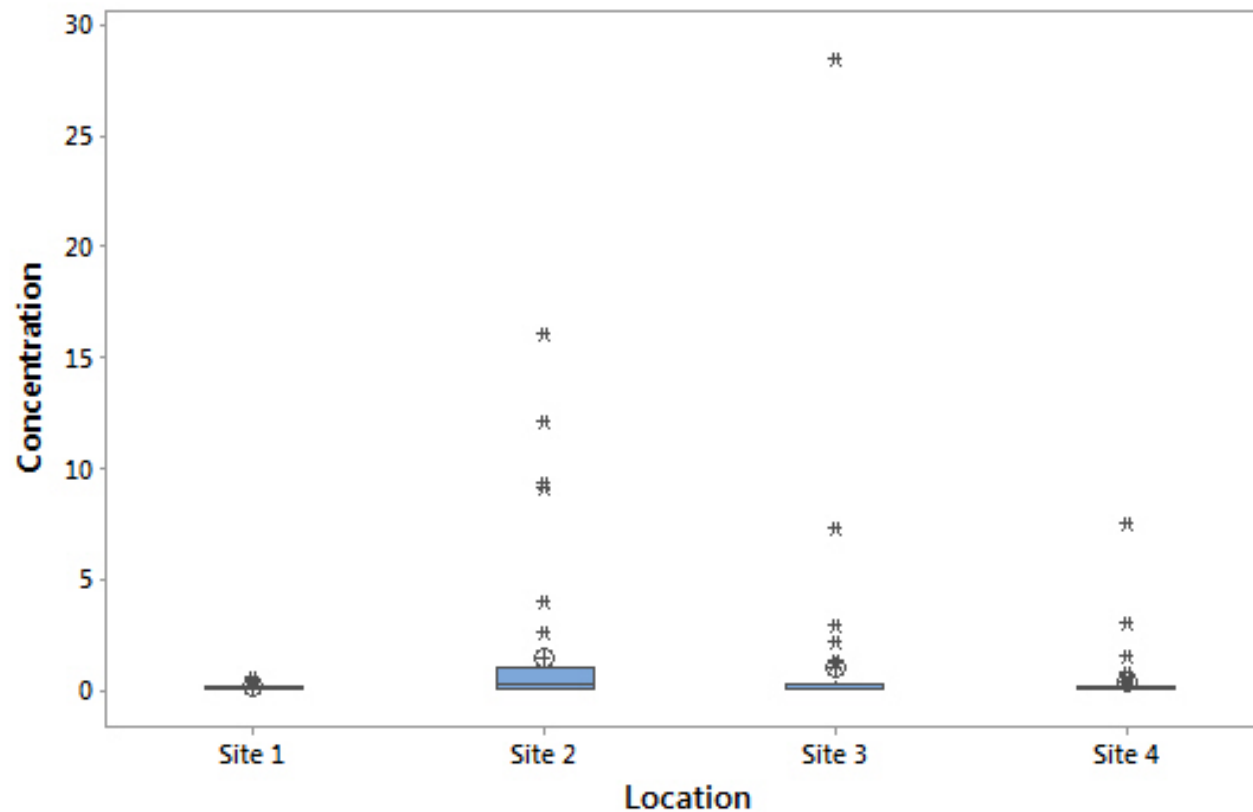
Parametric test. Assumes each group follows a normal distribution around its group mean. Best way to test is to compile residuals from each group's mean into one set of data, and test those as one gp.

Assumes each group has the same variance (“homoscedasticity”).

If these “approximations” are not correct, the p-values for the test may be too high (not significant). Instead, use a permutation test.

Example data: concentrations

From the boxplots, data at 3 of 4 sites looks non-normal

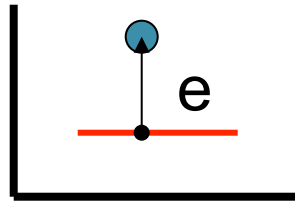


ANOVA

testing assumptions

Test residuals to see if normally distributed:

residual e = observation – its group mean



```
xres=residuals(aov(Concentration~Location))  
shapiro.test(xres)
```

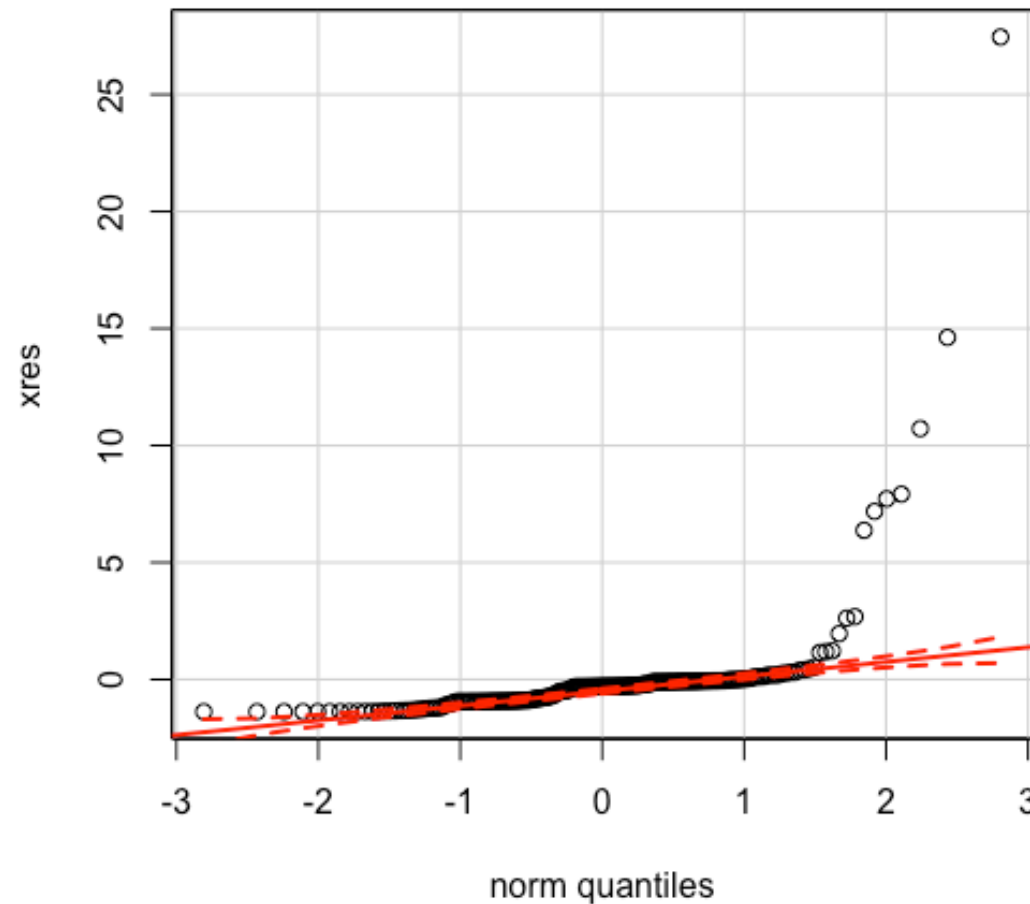
Shapiro-Wilk normality test

data: xres

$W = 0.33657$, $p\text{-value} < 2.2e-16$

Plot using a probability plot
If not normal, ANOVA will have low power

qqPlot(xres)



ANOVA

Null hypothesis: all means are equal

Alt hypothesis:

at least one group mean differs from the others (always a 2-sided test)

ANOVA does not tell which group means differ from the others.

ANOVA F-test

A signal to noise ratio, among groups is signal
within groups is noise

$$F = \frac{\text{MS treatment}}{\text{MS error}}$$

measures among groups signal
measures within group noise

If H_0 , null hypothesis, is true
F will be around 1.

ANOVA on concentration data

```
> summary(aov(Concentration~Location))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Location	3	51	16.99	2.36	0.0728
Residuals	196	1411	7.20		

Not sufficient evidence to reject H_0 and state that means differ.
Is the non-normality pushing up p-values?

ANOVA on logs of concentration

```
summary(aov(log(Concentration) ~ Location))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Location	3	53.2	17.735	4.164	0.00692 **
Residuals	196	834.8	4.259		

Small p-value: reject H_0 and state that means (of logs) differ. This tests differences in geometric means (medians) of Concentration in original units.

Permutation Tests for 3+ Groups

H_0 : All means are equal

If H_0 is true the data could be randomly reassigned to any group.

“Shuffle” the FACTOR names (or the response values) many times.

Compute an F ratio or similar statistic for each shuffle.

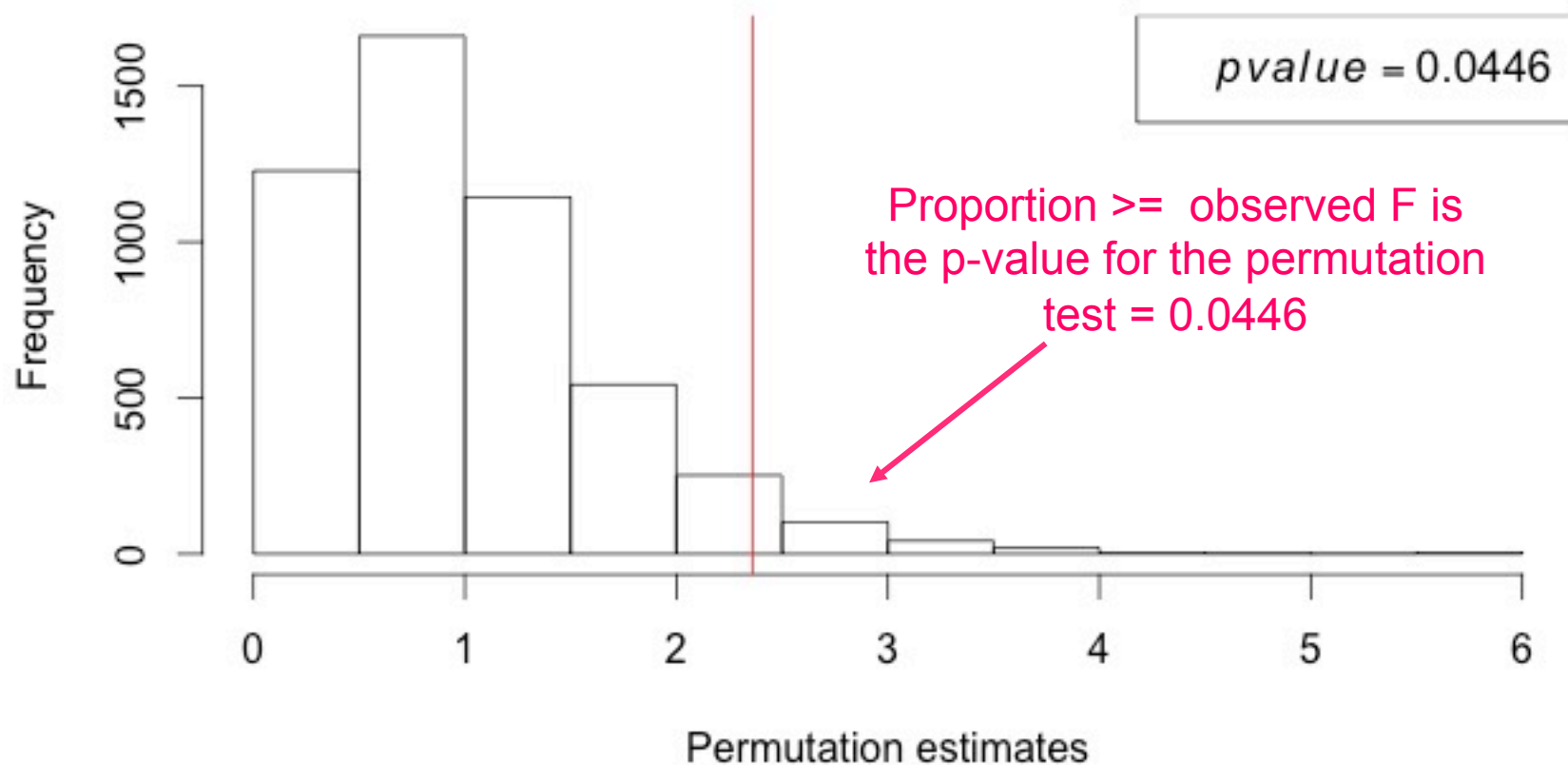
p-value equals proportion of shuffles with a test statistic that equals or exceeds the original observed statistic from the data

One of many shuffles

	Location	Concentration	Shuffle 1
1	Site 1	0.5	Site 3
2	Site 1	0.23	Site 4
3	Site 2	0.155	Site 2
4	Site 3	7.32	Site 1
5	Site 3	28.4	Site 3
....

Histogram of shuffle results: a picture of the null hypothesis

Permutation Histogram ... Observed F in RED, data = Concentration



Permutation permKS function in the perm package

K-Sample permutation test

Permutation is obtained with the
method="exact.mc" option (Monte Carlo)

Number of permutations is set using
control=permControl(nmc=10000) option

Permutation permKS function in the perm package

```
> permKS(Concentration, Location, method="exact.mc",  
control=permControl(nmc=10000, p.conf.level=0.95))
```

K-Sample Exact Permutation Test Estimated by Monte Carlo

data: Concentration and Location
p-value = 0.0481

p-value estimated from 10000 Monte Carlo replications
95 percent confidence interval on p-value:
0.04389243 0.05237330

(Remember: ANOVA p-value= 0.0728 for comparison)

Summary: Tests for the One-Way Design

ANOVA could not find a difference in mean concentrations, due to violation of the approximate test's assumptions

The ANOVA on logs found a significant difference in the geometric means (medians) of the 4 groups

The Kruskal-Wallis test found a significant difference in the medians of the 4 groups (not shown)

The permutation test found that there was a significant difference between the mean concentrations of the 4 groups

Which of the latter 3 tests you use depends entirely on the objectives of your study

Two-factor analysis of variance

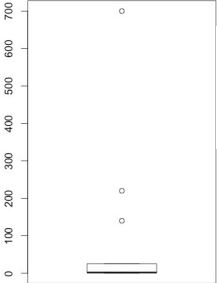
Evaluates the effects of two factors simultaneously

May have 'interactions', synergistic or antagonistic effects.

Assumes residuals follow a normal distribution.

We will use the coin package to perform an overall test.

Two-way ANOVA example: Iron concs in streams by Rock type and Mining Category

	Unmined	Mined and Reclaimed	Mined and Abandoned
Sandstone	132.4 66.7 22.7	Within each block, data assumed to follow a normal distribution	 <p>Normal Dist ?</p>
Limestone	4.5 1.8	4.7 5.9 9.6	3.9 0.4

Check whether residuals follow a normal dist

```
> a6=aov(fe~mining+rocktype+mining*rocktype)
```

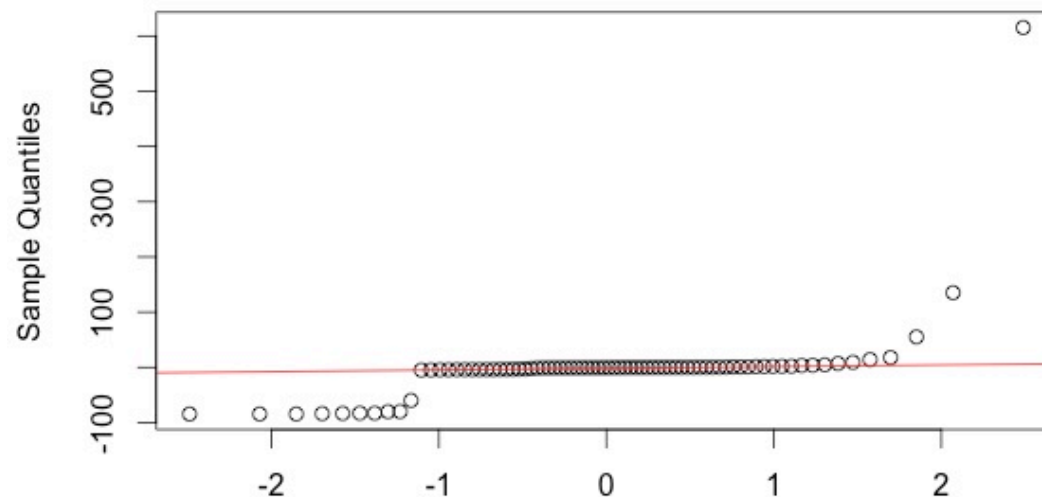
```
> shapiro.test(residuals(a6))
```

Shapiro-Wilk normality test

data: residuals(a6)

W = 0.33507, p-value < 2.2e-16

Residuals don't look like
a normal distribution



The ANOVA model with interaction

```
> summary(a6)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
mining	2	32282	16141	2.493	0.0898	.
rocktype	1	15411	15411	2.380	0.1273	
mining:rocktype	2	25869	12934	1.997	0.1431	
Residuals	72	466239	6476			

Neither mining or rocktype appear significant. Could this be due to the non-normality of the residuals?

The coin package performs an overall permutation test: is there some effect?

```
independence_test(fe~mining*rocktype,teststat="quad",  
distribution="approximate")
```

Approximative General Independence Test

data: fe by mining, rocktype

chi-squared = 6.8032, p-value = 2e-04

The significant result shows that there is some effect of the two factors. Not all groups have the same mean. But which factors are important? One? Both?

Test for an effect due to mining, blocking out the rocktype factor

```
> independence_test(fe~mining|rocktype,  
teststat="quad",distribution="approximate")
```

Approximative General Independence Test

data: fe by

mining (Abandoned, Reclaimed, Unmined)

stratified by rocktype

chi-squared = 4.6787, p-value = 7e-04

The significant result shows that the effect of mining on group means is significant when rocktype is 'blocked out'.

Test for an effect due to rocktype, blocking out the mining factor

```
> independence_test(fe~rocktype|mining,  
teststat="quad",distribution="approximate")
```

Approximative General Independence Test

data: fe by

rocktype (limestone, sandstone)

stratified by mining

chi-squared = 2.2774, p-value = 0.0263

The significant result shows that the effect of rocktype on group means is significant when mining is 'blocked out'.

This was the capability of permutation tests until the perm.fact.test (BDM test) was added to the asbio package:

The `perm.fact.test` in `asbio` performs a factor-specific permtest for two-way ANOVA

1. The original F-test is performed and saved.
2. ANOVA residuals are computed.
3. Residuals are randomly permuted across all groups (without replacement)
4. An F-statistic is computed for each effect using each replication and saved.
5. The p-value is the proportion of F values equal to or higher than the observed F for each effect.

The perm.fact.test in asbio performs a permtest for the two-way design

```
perm.fact.test(fe,mining,rocktype,perm=2000)
```

Y, X1, X2

\$Table

	Initial.F	Df	pval
X1	2.492636	2	0.0010
X2	2.379906	1	0.0105
X1:X2	1.997428	2	0.0255
Residual	NA	72	NA

There is a significant effect due to mining (X1) and rocktype (X2).

The interaction test is also significant (as plots also indicate) but there is some controversy whether permutation tests can evaluate interactions.

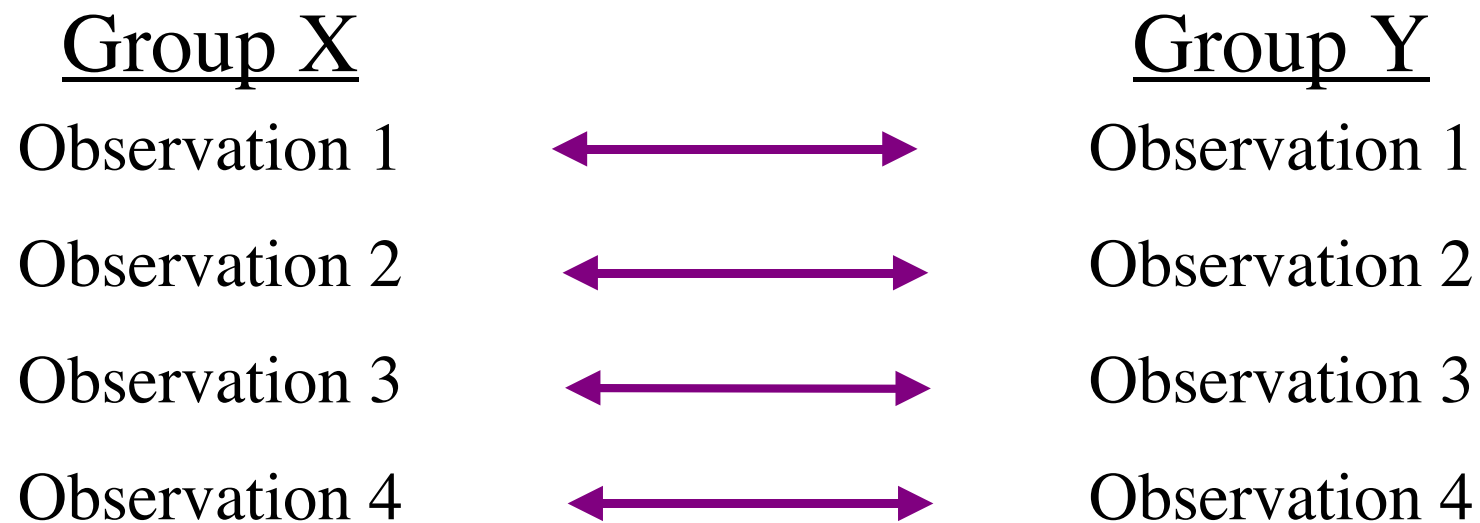
Summary for One and Two Factor Permutation Tests

- Can be used when ANOVA residuals do not follow a normal distribution.
- Can be used for data with unequal variance.
- Several varieties found in R. Presence is spotty in other packages.
- Be aware of the limitations of ANOVA, and that its p-values may be too high, missing the signal in the data.

Testing Paired Differences

Do the means of two groups of paired sets of data differ?

Matched - Pairs Tests



Is a direct relation between each observation in the first group and its equivalent in the second group.

Pairing in Environmental Studies

Most commonly by **time** or location

	Urban	Ag
→ Jan		
“Blocks” → Feb		
→ March		
→ April		

Pairing in Environmental Studies

Most commonly by time or location

	Old method	New method
→ Well 1		
“Blocks” → Well 2		
→ Well 3		
→ Well 4		

Example: Matched Pairs

Soil lead was measured at the same sites in 1996 before a major fire, and after the fire in 2001

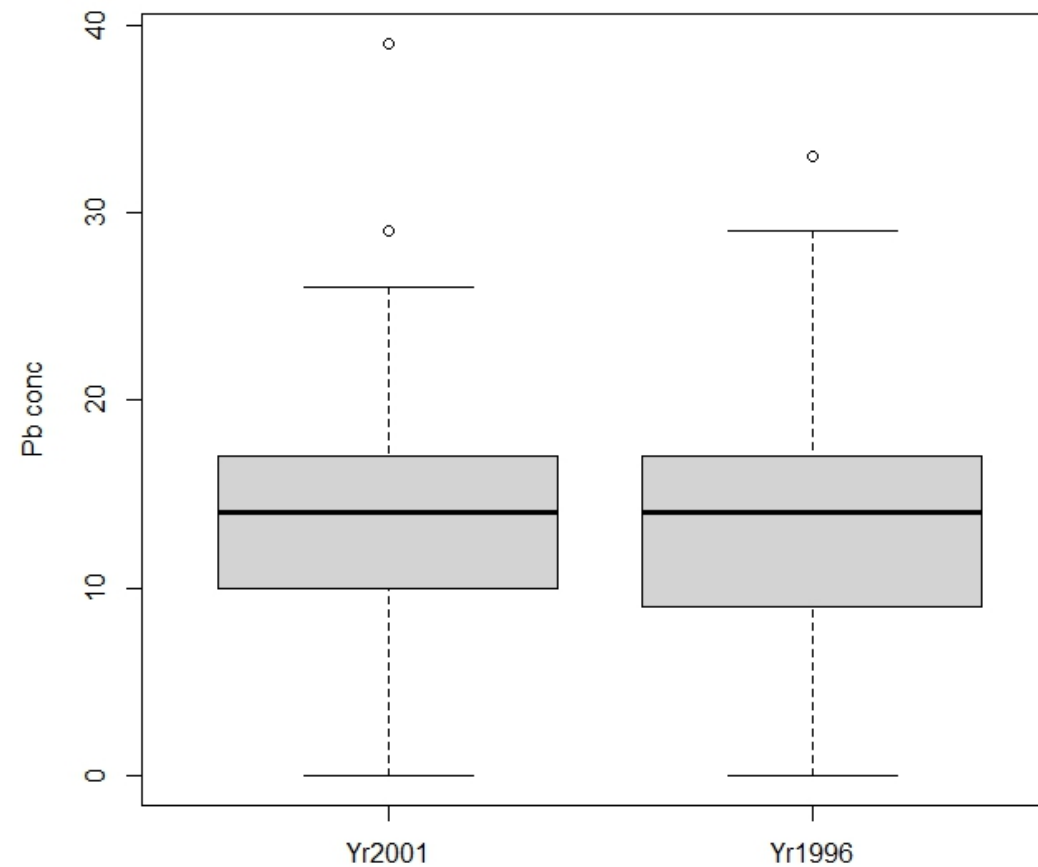
Measurements were 'blocked' by location. This minimizes causes of change other than the difference between the two years, which is attributed to the effect of the fire.

Are mean lead concentrations before the fire different than after the fire (a two-sided test)?

Compute the differences After–Before for pairs at the same site. Is the mean difference = 0?

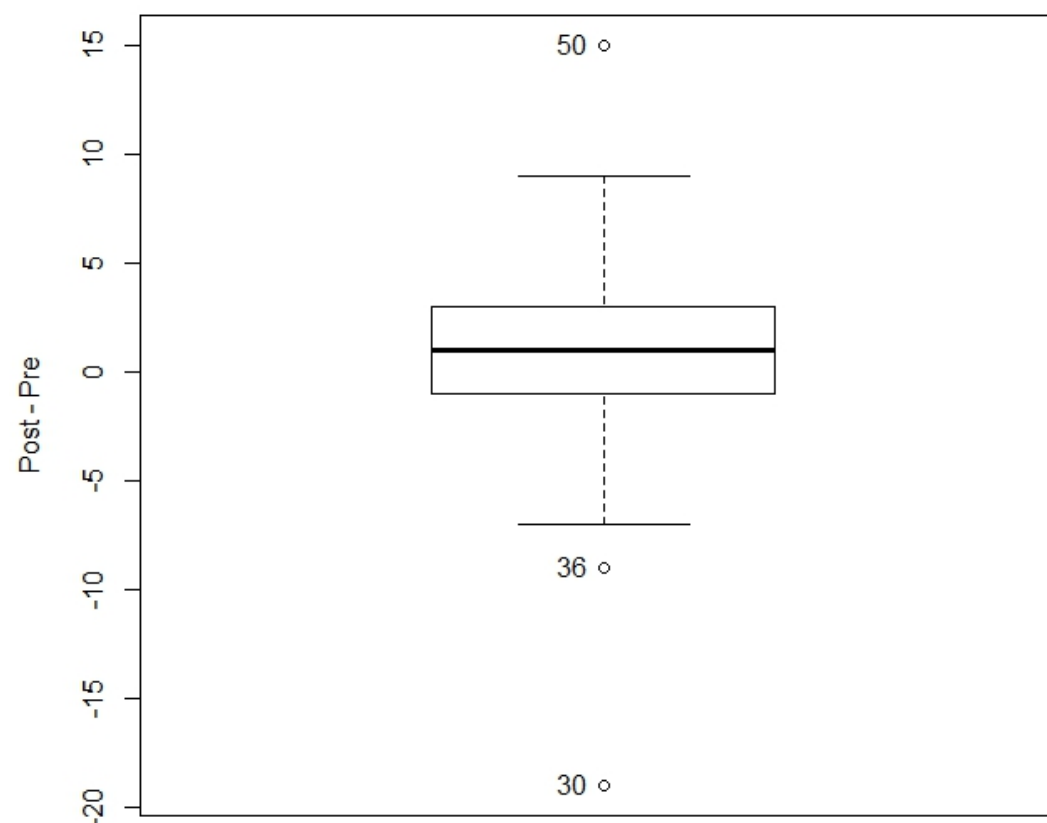
Boxplot of Data

This doesn't
show the
information
due to
pairing of
samples at
the same
site.



Boxplot of Paired Differences

Diffs are
symmetric.
Outliers,
esp. #30,
don't fit
a normal
distribution.
Is this a
problem?



We might expect low power due to non-normality

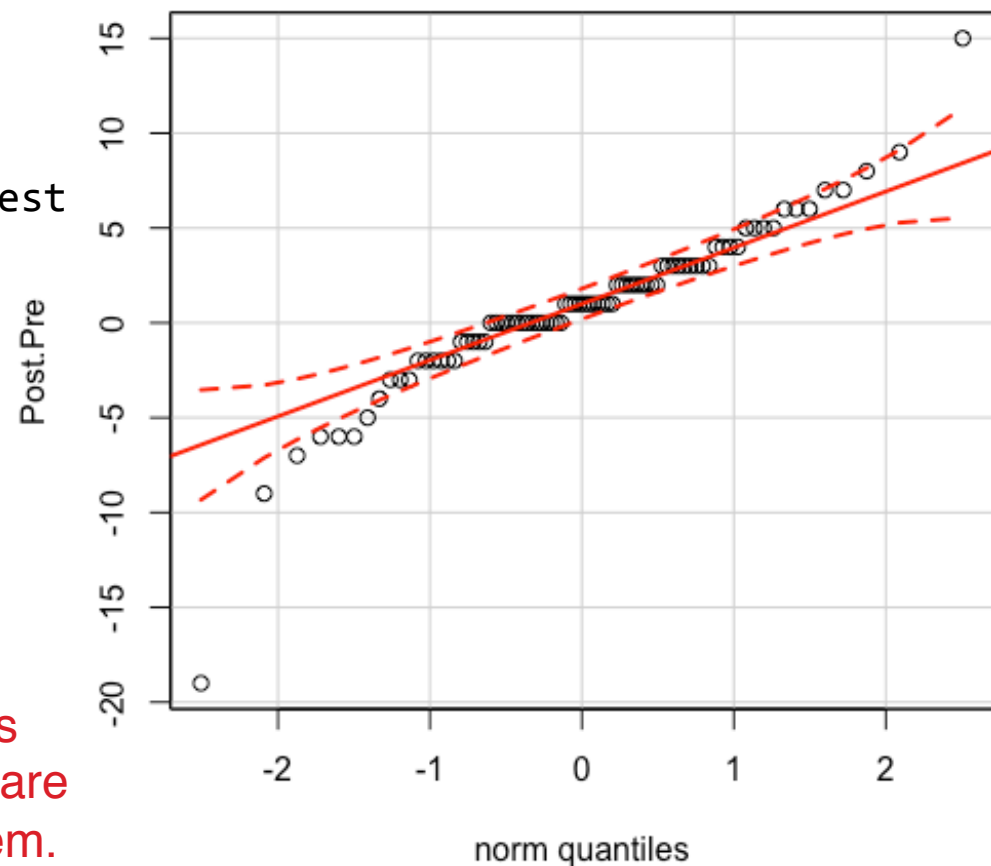
```
> shapiro.test(Post.Pre)
```

Shapiro-Wilk normality test

data: Post.Pre

W = 0.90893, p-value =
2.394e-05

Though the boxplot looks symmetric. The question is whether a few low outliers are sufficient to cause a problem.



Paired-t Test

Null Hypothesis: mean difference = 0.

Alt. Hypothesis: mean difference is NOT
= 0 (two-sided)

Alt. Hypothesis: mean difference > 0.
(one-sided)

Paired t-test

```
> t.test(Yr2001, Yr1996, alternative='two.sided', paired=TRUE)
```

Paired t-test

data: Yr2001 and Yr1996

t = 1.7406, df = 81, p-value = 0.08555

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

95 percent confidence interval:

-0.1186691 1.7772057

sample estimates:

mean of the differences

0.8292683

Permutation Test 1: analog of the t-test

Compute the differences $d_i = X_i - Y_i$ for each pair of observations $i=1$ to n .

Compute the test statistic \bar{x} , the mean of paired differences d_i

Compute the representation of H_0 : Compute a random vector of $+s$ and $-s$ of the same size as the number of pairs (n).

Multiply this sign vector times the absolute value of the differences $|d_i|$.
Compute the test statistic \bar{x}_{perm} .

Repeat 2,000–10,000 times or 2^n times, whichever is smaller.

Compute the p-value for the test. For a 2-sided test,
 $p = \text{Prob}(\bar{x}_{\text{perm}} \geq \bar{x}_{\text{obs}}) + \text{Prob}(\bar{x}_{\text{perm}} \leq -\bar{x}_{\text{obs}})$

For a 1-sided test expecting $X > Y$,
 $p = \text{Prob}(\bar{x}_{\text{perm}} \geq \bar{x}_{\text{obs}})$

Permutation matched-pair test

The two-sided test:

```
> permMatched(Yr2001,Yr1996)
```

Permutation Matched-Pair Test

Yr2001 - Yr1996 alternative = two.sided

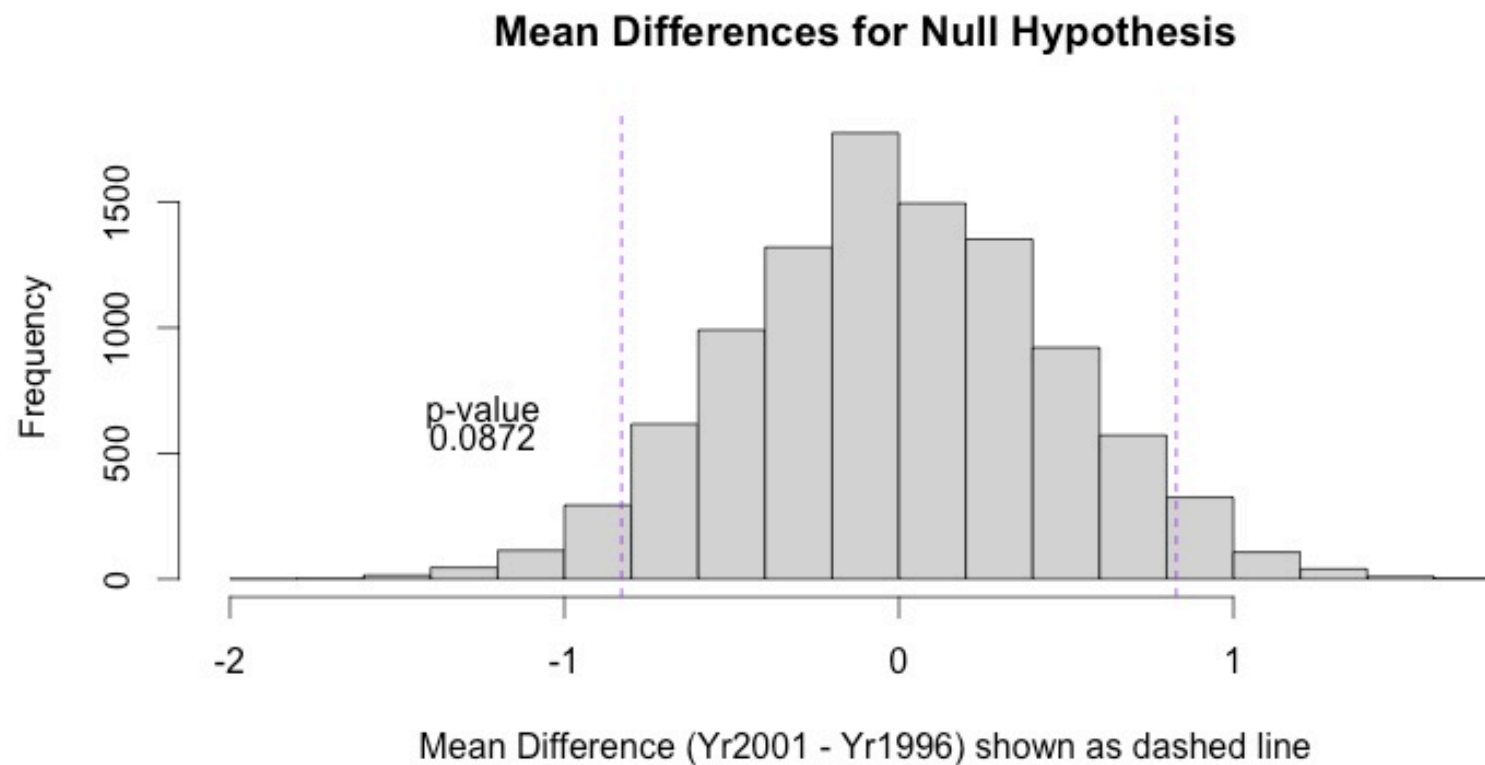
p-value = 0.0888

The normal-theory paired t test is:

t= 1.740604 p= 0.08555043

mean difference = 0.8292683 95% CI = (-0.1186691 1.777206)

Permutation matched-pair test results



Homework Exercise #1

Test for a difference in the means of two groups

Dissolved oxygen was measured in a river/estuary in Florida over several years. In 2008 a change was made upstream that was hoped to increase DO conditions somewhat. However, the uncertainty in the effect was enough to test also for a decrease in DO – the scientist did not want ignore a decrease if it was observed (a one-sided test for increased DO would ignore any decrease). Run two-sided tests for whether the mean DO has changed.

Homework Exercise #2

Test for differences in the means of four groups

Chloride concentrations were measured by Feth et al. (1964) -- *Sources of mineral constituents in water from granitic rocks, Sierra Nevada, California and Nevada*; USGS Water Supply Paper 1535-I – in shallow ephemeral springs and waters from two granitic rock types in the Sierra Nevada mountains. This was a classic paper in geochemistry, not to mention that field work must have been done in awesome scenery. Test whether the mean chloride concentration differs among the groups of springs using anova and the permutation test. Explain your findings.