Completely Randomized Designs

Tim Hanson

Department of Statistics
University of South Carolina

January, 2017

Modified from originals by Gary W. Oehlert

3.1 Definition

Completely randomized design (CRD) recipe:

1. Fix sample sizes $n_1, n_2, \ldots, n_g$ with $n_1 + n_2 + \cdots + n_g = N$.
2. Randomly assign n_1 units to treatment 1, n_2 units to treatment 2, etc.

All possible arrangements of the N units into g groups with sizes n_1 though n_g equally likely.

Selection of treatments, experimental units, and responses also need considerable thought w/ scientists conducting experiment.

Good place to start!

This is the basic experimental design; everything else is a modification.

The CRD is

- Easiest to do.
- Easiest to analyze.
- Most resilient when things go wrong.
- Often sufficient.

Consider a CRD first when designing.

Acid rain and birch seedlings

Wood and Bormann (1974) studied the effect of acid rain on trees. "Clean" precipitation has pH in the 5.0 to 5.5 range; precipitation pH in northern New Hampshire in the 3.0 to 4.0 range. Does this harm trees? If so, does harm extent change w/ rain pH?

One experiment: $N = 240$ six-week-old yellow birch seedlings randomly divided into five groups of $n_i = 48$; seedlings in each group got acid mist treatment 6 hours a week for 17 weeks at pH's: 4.7, 4.0, 3.3, 3.0, and 2.3. Seedlings treated identically except for treatment. Total plant weight response after 17 weeks.

Much thought goes into experiment!

- Scientists suspected that damage might vary by pH level, plant developmental stage, and plant species, among other things.
- This experiment only addresses pH level (other experiments were conducted...)
- Many factors affect tree growth; experiment controlled soil type, seed source, and amounts of light, water, and fertilizer.
- Desired treatment was real acid rain; available (controllable) treatment was synthetic acid rain consisting of distilled water and sulfuric acid.
- Experiment used yellow birch seedlings; other species or more mature trees?
- Total plant weight is an important response, but other responses (possibly equally important) are also available.
- Investigators boiled broad question down to a workable experiment using artificial acid rain on seedlings of a single species under controlled conditions.
- Much nonstatistical background work and compromise goes into the planning even simple experiments.

Resin lifetimes

Nelson (1990) gave an example where the goal was to estimate the lifetime (in hours) of an encapsulating resin for gold-aluminum bonds in integrated circuits operating at 120^oC. Lifetime very long; **accelerated tests** use more extreme temperatures (higher) to induce failure quickly; then interference **extrapolates** to 120^oC.

$N = 37$ units were assigned at random to one of five different temperature stresses: 175^oC, 194^oC, 213^oC, 231^oC, and 250^oC.

Choice of units clear: integrated circuits with the resin bond of interest.

Choice of treatments, however, depended on knowing that temperature stress reduced resin bond lifetime. The actual choice of temperatures probably benefited from knowledge of the results of previous similar experiments.

Experimental design combines subject matter knowledge w/ statistical methods.

Trebuchet projectile distance

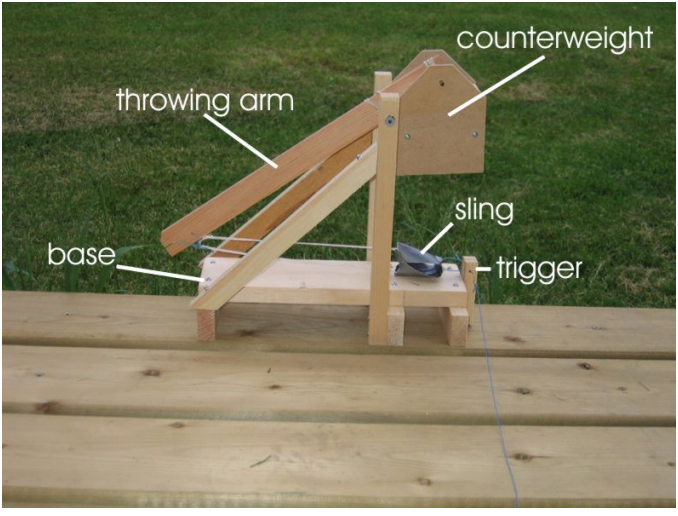
How do sling length and size of counterweight affect the throw distance of a trebuchet?

Randomly assign $N = 27$ throws to the nine combinations of three lengths and three weights, with three throws per combination.

The response is distance projectile travels; want to maximize this!

Leads to consideration of response surfaces (end of book).

Experiment small...



Build big!

3.2 Preliminary exploratory analysis

It is always a good idea to simply look at your data before modeling.

For a CRD we have data by group; therefore looking at side-by-side boxplots, and summary statistics by group is a good idea.

Gives an idea about: normality within groups, outlying observations, skew, etc. Also gives us an idea about how distributions change (and their means/medians) across groups.

Ask yourself: are data approximately normal? Are the variances roughly constant?

Resins in R

```
resin=read.table("http://users.stat.umn.edu/~gary/book/fcdae.data/exmpl3.2",header=T)
resin # data from website slightly different than in library 'oehlert'
colnames(resin)=c("temp","logTime") # column names now match
resin[,1]=c(rep(175,8),rep(194,8),rep(213,8),rep(231,7),rep(250,6))
boxplot(logTime~temp,data=resin) # side-by-side boxplots
boxplot(logTime~temp,data=resin,xlab="Temperature",ylab=expression(log[10](hours))) # bit nicer
summary(resin) # overall data set summary statistics
tapply(resin$logTime,resin$temp,summary) # by temperature
tapply(10^(resin$logTime),resin$temp,summary) # in hours
```

3.3 Models for mean response

Most of our inference is about treatment means:

- Any evidence means are not all the same?
- Which ones differ?
- Any pattern in differences?
- How can differences be described succinctly?
- Estimates/confidence intervals of means and differences.

Sometimes variances are more interesting, or quantiles such as the median, or other things like the number of modes.

Sometimes it is extremes that are of interest, e.g. minimum rainfall across a large area after seeding a cloud.

Some plausible models for resin lifetime

y_{ij} is jth lifetime in hours from group i. Different models for mean:

$$
\begin{aligned}
y_{ij} &= \mu + \epsilon_{ij}, \\
y_{ij} &= \beta_0 + \beta_1 z_i + \epsilon_{ij} \\
y_{ij} &= \beta_0 + \beta_1 z_i + \beta_2 z_i^2 + \epsilon_{ij} \\
y_{ij} &= \mu_i + \epsilon_{ij}
\end{aligned}
$$

Here, $z_1 = 175$, $z_2 = 194$, $z_3 = 213$, $z_4 = 231$, and $z_5 = 250$. Each model is special case of those that come after! Simpler models **nested** within more complicated (more parameters) models.

All errors $\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$.

Occam's razor

We seek the simplest model consistent with the data.

"All treatments have the same mean" is simpler than
"Each treatment has its own mean." If we cannot say that the complicated model is needed, we take the simple model.

Sometimes simple explanatory model is necessary. "Treatment means vary linearly with temperature" is simpler than "Each treatment has its own mean" or even "Treatment means vary quadratically with temperature." An explanatory model (especially a simple one) helps us understand the data. In particular, regression models allow us to extrapolate to temperatures not used in the experiment! The separate means model does not allow extrapolation.

Model is gross approximation to reality

All models are wrong; some models are useful. — George Box

We do not believe any model is really true, but if the data are consistent with it, we use it.

Comparing models

We gauge model fit by looking at the sum of squared residuals. We usually choose model parameters so as to minimize the sum of squared residuals.

The total sum of squares in the data SS_T is the sum of the model or explained sum of squares SS_M plus the error or residual sum of squares SS_E. For a fixed set of data, if you change the model making one SS bigger, then the other must get smaller.

$$SS_T = SS_M + SS_E$$

Always,

$$SS_T = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{\bullet\bullet})^2,$$

where $\bar{y}_{\bullet\bullet} = \frac{1}{N} \sum_{i=1}^{g} \sum_{j=1}^{n_i} y_{ij}$ is the mean of all observations. Represents total variability in responses about an overall, common mean $\bar{y}_{\bullet\bullet}$.

Separate means model

For $y_{ij} = \mu_i + \epsilon_{ij}$ estimate μ_i by $\hat{\mu}_i = \bar{y}_{i\bullet} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$ the sample mean in the ith group. Then

$$SS_M = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2$$

is variability explained by allowing means to change w/ group.

$$SS_E = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\bullet})^2$$

is the slop that is left over, i.e. the variability within groups.

Calculus can show that $\hat{\mu}_i = \bar{y}_{i\bullet}$ makes $SS_E = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij} - \hat{\mu}_i)^2$ as small as possible; $\hat{\mu}_i = \bar{y}_{i\bullet}$ are called the **least squares** estimates of μ_i because they minimize SS_E. $\hat{\mu}_1$, $\hat{\mu}_2$, $\hat{\mu}_3$, $\hat{\mu}_4$, $\hat{\mu}_5$ given by `tapply(resin$logTime,resin$temp,mean)`.

Linear regression model

For $y_{ij} = \beta_0 + \beta_1 z_i + \epsilon_{ij}$ estimate β_0 and β_1 by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(z_i - \bar{z})(y_{ij} - \bar{y}_{\bullet\bullet})}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(z_i - \bar{z})^2}, \quad \hat{\beta}_0 = \bar{y}_{\bullet\bullet} - \hat{\beta}_1 \bar{z},$$

where $\bar{z} = \frac{1}{N}\sum_{i=1}^{g}\sum_{j=1}^{n_i} z_i$. Then

$$SS_M = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\hat{\beta}_0 + \hat{\beta}_1 z_i - \bar{y}_{\bullet\bullet})^2, \quad SS_E = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij} - [\hat{\beta}_0 + \hat{\beta}_1 z_i])^2.$$

SS_M is variability explained by allowing means to change linearly with temperature;
SS_E is the slop that is left over, i.e. the variability not explained by the regression line.

Calculus shows that the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ above make $SS_E = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij} - [\hat{\beta}_0 + \hat{\beta}_1 z_i])^2$ as small as possible. $\hat{\beta}_0$ and $\hat{\beta}_1$ given by `lm(resin$logTime~resin$temp)`.

Nested models

"All treatment means are the same" is a special case of "Each treatment has its own mean." "Treatment means vary linearly with temperature" is a special case of "Treatment means vary quadratically with temperature" and, indeed, of "Each treatment has its own mean" as well.

We say that the special case model is included in the more complicated model, or perhaps that it is a restriction of (a restricted version of) the more complicated model.

We sometimes say that the special case model is nested in the more complicated model, but we will also use the descriptor "nested" in a different way later, so beware.

Nested models

When we have model A included in model B, then:

1. Model B (fit by LS) always fits at least as well as model A (fit by LS), and usually fits better. LS = "least squares", which minimizes the SS_E for whatever model you are fitting.
2. The SS_E from model B cannot be larger than the SS_E from model A, and is almost always smaller.
3. Equivalently, the SS_M for model B is always at least as large and almost always larger than the SS_M for model A.
4. The reduction in SS_E going from A to B is the same as the increase in SS_M going from A to B.

ANOVA

The partitioning of the sums of squares is called Analysis of Variance, or ANOVA.

The special case model never fits as well as the larger model, but how do we decide that it is good enough, that is, is consistent with the data?

The two basic approaches are:

- Significance testing
- Information Criteria

Significance testing

We will make an ANOVA table that has a row for the restricted model, a row for the increment from the restricted model to the larger model, and a row for all of the residual bits.

Each row in the table has a label, a sum of squares, a "degrees of freedom," and a "Mean square."

Degrees of freedom count free parameters. If there are r_1 parameters in the mean structure of the simpler, nested model, and r_2 parameters in the mean structure of the larger model, then there are $r_2 - r_1$ parameters in the improvement from the small model to the large model, and $N - r_2$ parameters for residuals (error).

An MS is SS divided by DF.

ANOVA table for comparing two models

The generic table looks like this (SS_1 is model SS for restricted model, and SS_2 is model SS for the large model):

Source	SS	DF	MS
Model 1	SS_1	r_1	SS_1/r_1
Improvement from Model 1 to Model 2	$SS_2 - SS_1$	$r_2 - r_1$	$(SS_2 - SS_1)/(r_2 - r_1)$
Error	SS_E	$N - r_2$	$SS_E/(N - r_2)$

Notation

There are simple formulae for elements of the ANOVA table for many designed experiments.

Let y_{ij} be the jth response in treatment i. $i = 1, 2, \ldots, g$ and $j = 1, 2, \ldots, n_i$.

Let

$$\bar{y}_{i\bullet} = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}$$

be the mean response in the ith treatment, and let

$$\bar{y}_{\bullet\bullet} = \frac{\sum_{i=1}^{g} \sum_{j=1}^{n_i} y_{ij}}{N}$$

be the grand mean response.

Most common ANOVA table

Suppose that the restricted model is the model that all treatments have the same mean, and the larger model is the model that each treatment has its own mean. Then:

$r_1 = 1$

$r_2 = g$

$SS_1 = N\bar{y}_{\bullet\bullet}^2$

$SS_2 = \sum_{i=1}^{g} n_i \bar{y}_{i\bullet}^2$

$SS_2 - SS_1 = \sum_{i=1}^{g} n_i (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2$

$SS_E = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\bullet})^2$

and the ANOVA table is ...

Basic ANOVA

The first four columns of the ANOVA table are:

Source	SS	DF	MS
Overall mean	$N\bar{y}_{\bullet\bullet}^2$	1	
Between Treatments	$\sum_{i=1}^{g} n_i(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2$	$g-1$	$SS_{Trt}/(g-1)$
Error	$\sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\bullet})^2$	$N-g$	$SS_E/(N-g)$

and the MS may be denoted MS_E and MS_{Trt}.

In fact, the line for the overall mean is so boring that it is usually left off. In R try something like `fit=lm(response~factor(treatment))` followed by `anova(fit)` to get the table.

ANOVA tables...

`anova(f)` tests a constant mean against whatever was fit in `f`.

`anova(f1,f2)` tests model f1 nested in f2.

Digression on Pythagorean Theorem

Note that

$$y_{ij} = \bar{y}_{\bullet\bullet} + (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}) + (y_{ij} - \bar{y}_{i\bullet})$$

Square both sides and add over all i and j and we get

$$\sum_{i=1}^{g}\sum_{j=1}^{n_i} y_{ij}^2 = N\bar{y}_{\bullet\bullet}^2 + \sum_{i=1}^{g} n_i(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2 + \sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\bullet})^2$$

plus a lot of sums of cross products. All those sums of cross products add to zero (the three components of y_{ij} are perpendicular out in N-dimensional geometry so sums of squares add up).

F statistic and test

The ANOVA is just algebra, albeit algebra with statistical intent. By assuming $y_{ij} \sim N(\mu_i, \sigma^2)$, we have

$$E(MS_E) = \sigma^2$$

and **if the restricted model is true** we also have

$$E(MS_{Trt}) = \sigma^2$$

If the restricted model is not good enough its expectation is larger than σ^2. This means that

$$F = MS_{Trt}/MS_E$$

is a test statistic for comparing the restricted model to the full model; we reject the null if F is too big.

F statistic and test

When the null is true and the normal distribution assumptions are correct, the F-test follows an F-distribution with $g - 1$ and $N - g$ df (df from numerator and denominator MS). Reject the null that the single mean model is true when the p-value for the F-test is too small.

We did the algebra for the single mean model and individual mean model, but the F test is appropriate for general restricted models versus a containing model. It's just that the computations are not always so clean.

Resin, continued...

```
attach(resin) # don't have to use resin$temp and resin$logTemp
logTime
temp
m1=lm(logTime~1) # one overall mean
m2=lm(logTime~temp) # linear in temp
m3=lm(logTime~temp+I(temp^2)) # quadratic in temp
m4=lm(logTime~factor(temp)) # separate means
anova(m1,m2) # linear better than constant mean?
anova(m2,m3) # quadratic better than linear?
anova(m4,m3) # separate means better than quadratic?
anova(m1,m2,m3,m4) # all three tests at once

summary(m4) # are alpha_i significantly different from zero?
anova(m4) # tests H0: alpha1=alpha2=alpha3=alpha4=alpha5=0
```

Information criteria

Akaike introduced the first information criterion, AIC.

Later Schwartz added a second one, BIC.

Now there are several more.

Information criteria include a measure of how well the data fit the model (smaller being better) plus a penalty for using additional parameters.

Models with smaller values of AIC or BIC are better models.

Information criteria

Let L be the maximized likelihood for the data. This is the "probability" of the data under the model, with the parameters chosen to make the probability as high as possible. This likelihood model has k parameters that we can choose. Typically these parameters are things like treatment means, or regression coefficients, or residual variances.

We'll say more later, but for now suffice it to say that big L is good.

$$AIC = -2\ln(L) + 2k$$
$$BIC = -2\ln(L) + \ln(N)k$$

Information criteria

Choose a model with smaller AIC (or BIC).

In general, AIC tends to choose models with more parameters than we get from significance testing, i.e., some things in the selected model might be "insignificant." The reverse tends to be true for BIC, especially for big data sets.

Except for very small data sets, BIC penalizes additional parameters more than AIC. BIC thus tends to choose smaller models than AIC.

AIC tends to work better when all candidate models are approximate; BIC tends to work better in large samples when one of the candidate models is really the right model.

Resin, continued...

```
AIC(m1,m2,m3,m4)
BIC(m1,m2,m3,m4)
```

Both AIC and BIC pick the quadratic model.

Parameterizations for $\mu_i = E(y_{ij})$

There are many ways to describe/parameterize the same set of means.

Some parameterizations aid in interpretation.

They can all be different yet still correct, but you need to know which ones you're working with.

Parameterizations for $\mu_i = E(y_{ij})$

Consider the resin example.

Trt (^oC)	175	194	213	231	250	All data
Average	1.933	1.629	1.378	1.194	1.057	1.465
Count	8	8	8	7	6	37

If we have a single mean model, the only parameter is the overall mean μ. Our estimate would be $\hat{\mu} = \bar{y}_{\bullet\bullet} = 1.465$.

In the separate means model, parameters are the group means, and the estimates would be $\hat{\mu}_1 = \bar{y}_{1\bullet} = 1.933$ and so on.

Parameterizations for $\mu_i = E(y_{ij})$

Sometimes we want to write

$$\mu_i = \mu^* + \alpha_i$$

Where μ^* is some kind of "central value" and α_i is a treatment effect.

We always have $\alpha_i = \mu_i - \mu^*$ and $\hat{\alpha}_i = \hat{\mu}_i - \hat{\mu}^*$, but how do we define μ^*?

There are many ways but there three are common.

Parameterizations for $\mu_i = E(y_{ij})$

Define μ	Equivalent constraint
$\mu^* = \mu_1$	$\alpha_1 = 0$
$\mu^* = \frac{\sum_i \mu_i}{g}$	$\sum_i \alpha_i = 0$
$\mu^* = \frac{\sum_i n_i \mu_i}{N}$	$\sum_i n_i \alpha_i = 0$

The first is the default in R and SAS, the second is the default in Minitab, and the third is useful in hand calculations.

Model effects for $\sum_{i=1}^{g} n_i \alpha_i = 0$ can be obtained as $\hat{\alpha}_i = \bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}$, e.g. `tapply(response,treatment,mean)-mean(response)`.

The important things ($\mu_i - \mu_j = \alpha_i - \alpha_j$) are the same in all versions.

Care about μ in the single mean model; care about μ_i and $\alpha_i - \alpha_j$ in the separate means model.

Resin, continued...

```
# default in R is to set alpha1=0; first group is "control" to compare rest to
ftemp=factor(temp) # turn temperature into a factor
m4=lm(logTime~ftemp) # need factor or fits linear regression!
anova(m4) # are treatment means significantly different at alpha=0.05?
summary(m4) # how do other temps compare to 175 degrees C?

# getting sum-to-zero treatment effects: sum_i alpha_i =0
library(cfcdae) # if you got the package to load, otherwise:
source("http://people.stat.sc.edu/hansont/stat506/cfcdae.R")
m4=lm(logTime~ftemp,contrasts=list(ftemp="contr.sum"))
# function model.effects() part of cfcdae package
model.effects(m4,ftemp) # sum-to-zero!
```

Parameterizations for $\mu_i = E(y_{ij})$

What about polynomial models? Let z_i be the temperature treatment for group i. Here are some models

$$\begin{aligned}
\mu_i &= \beta_0 \\
\mu_i &= \beta_0 + \beta_1 z_i \\
\mu_i &= \beta_0 + \beta_1 z_i + \beta_2 z_i^2 \\
\mu_i &= \beta_0 + \beta_1 z_i + \beta_2 z_i^2 + \beta_3 z_i^3 \\
\mu_i &= \beta_0 + \beta_1 z_i + \beta_2 z_i^2 + \beta_3 z_i^3 + \beta_4 z_i^4
\end{aligned}$$

The first is the same as the single mean model, the last fits the same means as the separate means model, and the others are intermediate.

Note that equivalently written parameters have different meanings (and different values) in different models.

Parameterizations for $\mu_i = E(y_{ij})$

But we don't even leave polynomials in peace. Consider

$$\begin{aligned}\mu_i &= \beta_0 + \beta_1[z_i - 210.0811] \\ &\quad + \beta_2[z_i^2 - 422.9z_i + 44043.5] \\ &\quad + \beta_3[z_i^3 - 636.4z_i^2 + 133812.3z_i - 9294576.3]\end{aligned}$$

This is equivalent to the cubic model on the last slide, but here the β_i retain values and meanings as we change linear to quadratic to cubic (and you can go higher). These are *orthogonal polynomials*.

Moral of the story...

- Parameters can be defined in many ways within a single mean structure.
- Parameters are a means to an end.
- Most parameters are arbitrary, so inference on parameters (as opposed to model comparison or comparison of means) is also somewhat arbitrary.

R will compute the estimates as well as standard errors for various parameterizations, polynomials, orthogonal polynomials, trigonometric series, and so on. They are done correctly, but they retain the arbitrariness of their definition.

Resin, finished...

```
op4=lm(logTime~poly(temp,degree=4),data=resin)
op3=lm(logTime~poly(temp,degree=3),data=resin)
op2=lm(logTime~poly(temp,degree=2),data=resin)
op1=lm(logTime~poly(temp,degree=1),data=resin)
anova(op1,op2,op3,op4) # accept that quadratic (degree=2) is adquate

library(graphics) # can use matplot() function
pred.data=data.frame(temp=seq(120,250,1))
pred.int=predict(op2,pred.data,int="predict")
matplot(pred.data$temp,10^(pred.int),lty=c(1,2,2),col=c(1,2,2),type="l",xlab="temperature",
 ylab="hours")
points(c(175,194,213,231,250),tapply(10^(resin$logTime),resin$temp,mean),pch=16)
```

`poly()` makes orthogonal polynomials in your variable. They are difficult to interpret but not subject to collinearity. Regular polynomials give the exact same predictions.

Alternatives

Kruskall-Wallace is a nonparametric generalization of Mann-Whitney-Wilcoxin to more than two groups. The Welch ANOVA adjusts the usual F test to allow non-constant variance across groups; data are still assumed normal though. Can also do a type of permutation test. Power transformations can be used to "coerce" data into a more "normal" form, possibly with constant variance.

```
kruskal.test(logTime~temp) # Kruskall-Wallace nonparametric one-way ANOVA
oneway.test(logTime~temp) # Welch adjustment for non-constant variance

library(coin) # one type of permutation test
independence_test(logTime~factor(temp))

library(rcompanion) # pairwise comparisons after perm. test (for later)
pairwisePermutationTest(logTime,factor(temp),method="fdr")

# power transformations can help stablize variances and make observations more "normal"
library(MASS)
Time=10^logTime # original time in hours; why was log10(time) used in the first place?
boxplot(Time~temp) # Yikes!!!
boxcox(Time~factor(temp)) # lambda near 0 suggests log-transformation
boxplot(log10(Time)~temp) # That's why!
```