Experiment Planning and Design

Lecture 6: Techniques for Hypothesis Testing, Part II

Claus Aranha
caranha@cs.tsukuba.ac.jp

Department of Computer Science

2015-06-09

The story so far...

Outline for this class

A few more techniques for dealing with special experiments

- Paired Testing;
- ANOVA;
- Model Selection (Parameter Selection);

Paired Design

Imagine this situation

A scientist develop an optimization algorithm, *A*, for a family of problems, and wants to compare its convergence speed against the state-of-art algorithm, *B*.

The Researcher implements both methods, and wants to determine whether the proposed one has a better average performance for problems of that particular family.

The measurements are made under homogeneous conditions (same computer, same operational conditions, etc). And the running time is measured in a way that is not sensitive to other processes running in the system.

Dependent Populations

This problem has some important questions worth considering:

- What is the actual question of interest?
- What is the population for which this question is interesting?
- What are the independent observations for that population?
- What is the relevant sample size for the experiment?

Please think a little and propose answers to these questions.

Paired Design

The variability due to the different test problems is a strong source of spurious variation that can and must be controlled.

An elegant solution to eliminate the influence of this nuisiance parameter is the pairing of the measurements by problem:

- Observations are considered in pairs (A,B) for each problem.
- Hypothesis testing is done on the sample of *differences*

Another example (1)

How would you test the differences between two types of shoes (A and B)?

Can we use this information (same kid used both shoes) to improve our statistical testing?

Image from http://leagueathletics.com/

Another example (1)

How would you test the differences between two types of shoes (A and B)?

1. Give some kids the shoe A, and some kids the shoe B
   The personality of the kid becomes a factor – more agressive kids may wear the shoe faster?
2. Give each kid one of each shoe: Left shoe A, right shoe B (or random)
   In this way each kid will use both shoes equally.

Can we use this information (same kid used both shoes) to improve our statistical testing?

Image from http://leagueathletics.com/

Paired Tests (2)

Interface Testing

Imagine you are comparing two HCI technologies using a survey. You ask each surveyed person to compare two systems using a score.

- Scores given by the same person should show the same biases;
- Thus, you want to compare the scores described by the same voters;
- Important! Don't forget to change the order of voting!

Paired T-Test

- The Paired T-test is based on the idea of calculating the difference between the samples.
- This difference is then treated as a one-sample test.

```
> attach(intake)
> pre; post; post-pre
 [1] 5260 5470 5640 6180 6390 6515 6805 7515 7515 8230 8770
 [1] 3910 4220 3885 5160 5645 4680 5265 5975 6790 6900 7335
 [1] -1350 -1250 -1755 -1020  -745 -1835 -1540 -1540  -725 -1330 -
```

Paired T-Test

- The Paired T-test is based on the idea of calculating the difference between the samples.
- This difference is then treated as a one-sample test.

```
> t.test(pre,post,paired=T)

        Paired t-test

data:  pre and post
t = 11.9414, df = 10, p-value = 3.059e-07
alternative hypothesis: true difference in means
                              is not equal to 0
95 percent confidence interval:
 1074.072 1566.838
sample estimates:
mean of the differences
                1320.455
```

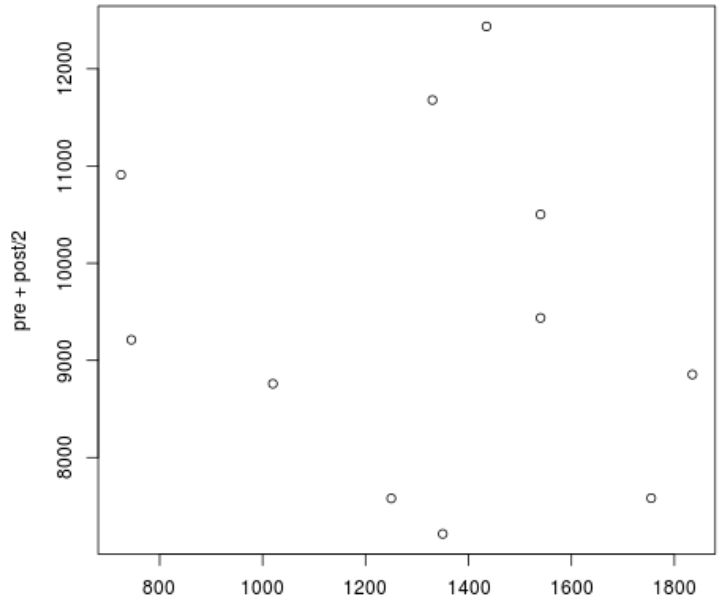
Paired T-Test – Important points 1

- If your data is paired (there is a correspondence between the two samples), you MUST use a paired test for correct results;
- Conversely, if your data is NOT paired, you should NOT use a paired test;

Paired T-Test – Important points 2

- The paired t-test assumes that the difference between the samples is independent of the level;
- This means that higher values do not have a bigger difference than lower values;
- You can examine this by plotting the difference of the pair against the average of the pair;
- Even if the differences scales with the level, a transformation of the data can be used to fix this problem.

```
plot(pre-post,pre+post/2)
```



Paired Wilcoxon Test

- The basic idea of the Paired Wilcoxon test is similar to the paired T-test;
- The “rank of the differences” is calculated; (so the size of the differences has a smaller effect)

```
> wilcox.test(pre,post,paired=T)

        Wilcoxon signed rank test with continuity correction

data:  pre and post
V = 66, p-value = 0.00384
alternative hypothesis: true location shift is
                        not equal to 0

Warning message:
In wilcox.test.default(pre, post, paired = T) :
  cannot compute exact p-value with ties
```

Paired Testing: Examples

Can anyone give me examples from your research of paired testing?

Part II: Analysis of Variance (ANOVA)

Analysis of Variance

Introduction

The Analysis of Variance (ANOVA) is a statistical method for testing many samples at the same time.

- We want to find the best values for 4 parameters in our method. Each parameter can be valued between 0.0 and 1.0;
- Can we experiment the parameter values two at a time? Why? Why not?

Analysis of Variance (Model)

Let each observation x_{ij} be the j-th observation in group i. We decompose this observation as:

$$x_{ij} = \bar{x} + (\bar{x}_i - \bar{x}) + (x_{ij} - \bar{x}_i)$$

In other words, eath observation is composed by the grand average of all samples, plus the deviation of the group, plus the (independent) error of the sample.

Analysis of Variance (Model)

- The Analysis of Variance starts from the assumption that the variance of all groups is the same;
- If this assumption holds, then we can compare the variance difference between groups to test the null hypothesis that “All groups have the same mean”;

Although it is called Analysis of Variance, and the main calculations are based on the variance of the groups, the final goal is a hypothesis centered on the means of the groups.

Analysis of Variance (Example – 1)

```
> attach(red.cell.folate)
> summary(red.cell.folate)
folate          ventilation
Min.   :206.0   N2O+O2,24h:8
1st Qu.:249.5   N2O+O2,op  :9
Median :274.0   O2,24h     :5
Mean   :283.2
3rd Qu.:305.5
Max.   :392.0
```

The red cell folate data set has three different categories: N20+02 24h, op, and O2 24h;

Analysis of Variance (Example – 2)

```
> anova(lm(folate~ventilation))
Analysis of Variance Table

Response: folate
            Df Sum Sq Mean Sq F value  Pr(>F)
ventilation  2  15516  7757.9  3.7113 0.04359 *
Residuals   19  39716  2090.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Sum Sq and Mean Sq show the variances calculated for the groups;
- The F statistic shows that the null hypothesis that all groups are the same;

Comparing multiple groups with ANOVA

The F statistic in the Anova test tells us that not all groups are the same. How do we calculate the actuall differences betwee these groups?

```
> summary(lm(folate~ventilation))
Call:
lm(formula = folate ~ ventilation)

Residuals:
    Min      1Q  Median      3Q     Max
-73.625 -35.361  -4.444  35.625  75.375

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)            316.62      16.16  19.588 4.65e-14 ***
ventilationN2O+O2,op   -60.18      22.22  -2.709   0.0139 *
ventilationO2,24h      -38.62      26.06  -1.482   0.1548
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 45.72 on 19 degrees of freedom
Multiple R-squared: 0.2809,     Adjusted R-squared: 0.2052
F-statistic: 3.711 on 2 and 19 DF,  p-value: 0.04359
```

Comparing multiple groups with ANOVA (2)

```
    Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)             316.62      16.16  19.588 4.65e-14 ***
ventilationN2O+O2,op    -60.18      22.22  -2.709   0.0139 *
ventilationO2,24h       -38.62      26.06  -1.482   0.1548
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The Intercept is the mean of the first group (N2O+02,24h);
- The Estimates are the differences between the corresponding groups and the first;
- The Test Statistic (last column), shows the p-value for the alternate hypothesis “The mean of this group is different from the mean of the first group”;
- This version of the test compares only the first group (baseline) against all;

Comparing multiple groups with ANOVA(3)

Comparing All against All

As we increase the number of comparisons, we increase the chance of fininding a “significant” result. Therefore, our p-values must be adjusted to compensate for this.

```
> pairwise.t.test(folate,ventilation)

    Pairwise comparisons using t tests with pooled SD

data:  folate and ventilation

      N2O+O2,24h N2O+O2,op
N2O+O2,op 0.042      -
O2,24h    0.310      0.408

P value adjustment method: holm
```

Non-parametric version

The Kruskal-Wallis test is the non-parametric counterpart of the one-way ANOVA:

```
> kruskal.test(folate~ventilation)

        Kruskal-Wallis rank sum test

data:  folate by ventilation
Kruskal-Wallis chi-squared = 4.1852, df = 2, p-value = 0.1234
```

Experimental Models

When executing an experiment, we have to make many choices:

- What parameter values?
- How many repetitions?
- How many samples?
- Which data combinations?

The answers for these questions will be your Experimental Model

First choice on Experimental Design

Before we can make any choices in the experimental design, you must decide: What is the goal of your experiment?

- Discover the influence of parameters (Sensibility Analysis);
- Find out best parameter values;
- Analyze one particular factor;
- Compare multiple methods;
- Test performance in data;

The Completely Randomized Design (1)

- The most basic (and yet most important) experiment design pattern is the "Completely Randomized Design" (CRD).
- When we are not interested in any of the factors (Pilot Experiments)
- Attention: If we are not interested in the factors, we are not interested in the results for those factors!
- We want to reduce the effects of factors in our results

The Completely Randomized Design (2)

Selecting Values

- We distribute the repetitions (samples) randomly between all factors.
- However a purely random distribution will tend to overrepresent some factor combinations over others;
- The Latin Hypercube can be used to evenly distribute experiments among factors;

Selecting Levels and Parameters

The Latin HyperCube Sampling (LHS)

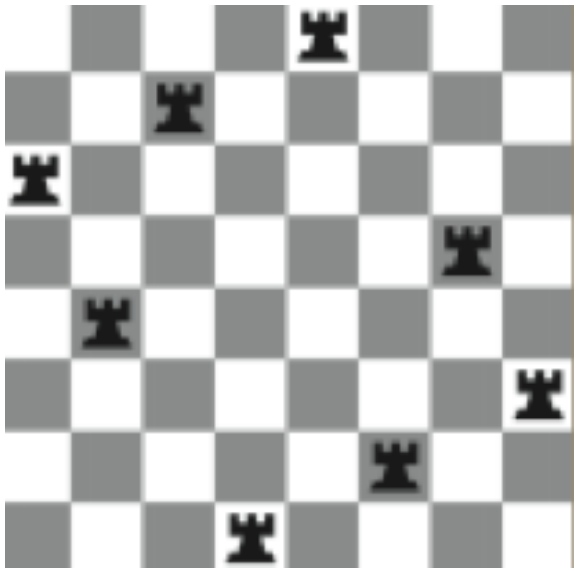
The LHS is a strategy for generating a random, robust parameter set from a large number of dimensions.

- Define the desired number of experimental runs k, and the number of parameter (dimensions) n;
- For each parameter x_i, divide the range of that parameter into k parts of equal size;
- The range of all n parameters are arranged in a hypercube;
- Choose k sets of parameters so that for each “row” and “column” in the hypercube there is only one sample;

Selecting Levels and Parameters (2)

When to use LHS

- As a initial exploration of the parameter space;
- You are interested in seeing if there is a change in performance, but you don't know where to look;
- You want to show that your model/method is resistant to change in parameters;
- You want to concentrate on one parameter (not included in LHS), but don't care about the others;



When NOT to use LHS

- You want to study the sensibility of one particular parameter;

Fixed Effect Model

Based on domain knowledge, we may decide to fix the values for some factors we know to be of interest.

- Fixed Effects: Parameter values fixed by domain knowledge – not relevant, or a single relevant level.
- Random Effects: Parameter values which we don't want to interfere in the result;
- Observed Effects: Calculated changes in levels to observe desired results;

Analysis of Variance (ANOVA) can be used on the observed levels to find significant trends in the results;

Randomized Complete Block Design (RCBD)

In the analysis of Completely Random Design (CRD), observations are assigned to different Experimental Units

Experimental Unit

A set of parameter values for an experiment

Sometimes an effect is not relevant – so we want to remove its influence from the experiment. But its value influences the result, so we can't just randomize it away.

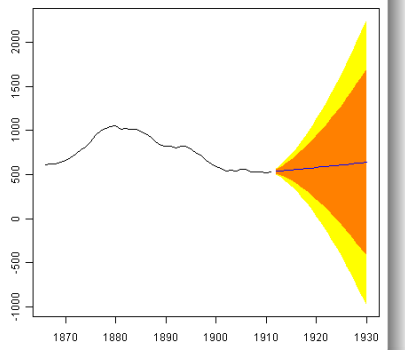
Blocking Model

When these factors are known and controllable, an elegant way to reduce their effect in the experiment is by blocking.
Blocking consists of separating the experiments into blocks, according with the levels of the experimental factor that we want to isolate.

Blocking Example (1)

Optimization Problems

A researcher develops a new algorithm for the optimization of time-series prediction models. He wants to test the performance of his model against recent algorithms in a variety of problems: Weather data, Financial data, Sun Radiation data, etc.

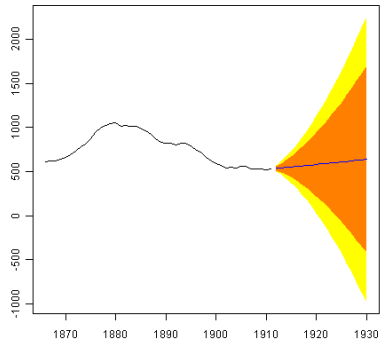


- The performance on each problem is not of interest (as opposed to the comparison with the other method);
- On the other hand, the performance is influenced by the problem type;
- We block the problem type, by separating the experiment into subgroups based on the data set, and analyzing the results as paired samples;

Blocking Example (2)

Optimization Problems

A researcher develops a new algorithm for the optimization of time-series prediction models. He wants to test the performance of his model against recent algorithms in a variety of problems: Weather data, Financial data, Sun Radiation data, etc.



- Remember that $\mu_{\text{global}} = \mu_0 + \mu_{\text{factor}}$
- If we employ Complete Randomization (CRD) in this problem, the experimental error will also contain the variability from the different data sets;
- Since we can control the allocation of the data sets, we can group the experiments according to the problem in a sistematic way, to isolate the error due to the different data sets;

Blocking x Randomization

When we randomize the run order of our observations, we are guarding against unknown factors that may affect our results.
Blocking comes into play whenever we know, from the beginning, that certain factors can influence our response variable. And we know that we are not interested in this influence.
Examples:

- The effects of different benchmark in algorithm related research;
- The effects of different batches of materials in the experiment;

Blocking Analysis (1)

In the general case, we have *a* levels of the experimental factor, and *b* levels of the blocking variable.

Our statistical model would be:

$$y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij} (i = 1 \ldots a, j = 1 \ldots b)$$

As seen in the last class for ANOVA, τ and β are the group errors for the factors, and ϵ are the individual Errors.

Blocking Analysis (2)

Null Hypothesis

We are interested in the experimental variable, *a*. Therefore, our null hypothesis is that:

$$H_0 : \sum \tau_i = 0 (\text{for} i = 0 \ldots a)$$

Statistical Relevance of Blocking

It might be interesting to also test the statistical significance of the difference between blocks. The null hypothesis would be:

$$H_0^b : \sum \beta_j = 0$$

Even if we are not interested in the effects of blocking – testing for the significance of the difference tells us if blocking is necessary or not.

Blocking Analysis (3)

Merging Blocks

Suppose that there are more than one factor that we are interesting in blocking. As long as we are not interested in the blocked effects (if we were, we wouldn't be blocking), we can merge the multiple factors into a single block factor.

Example

In the previous experiment, we want to block our algorithms by different *problem types* and different *Data Set Lengths*. Although these are different factors, we can block them together into mixed factors.

Blocking Analysis (4)

Blocking Efficiency

We can calculate the Blocking Efficiency, which shows how much larger a Complete Random Design would have to be to reach the same power as a Random Complete Block Design.

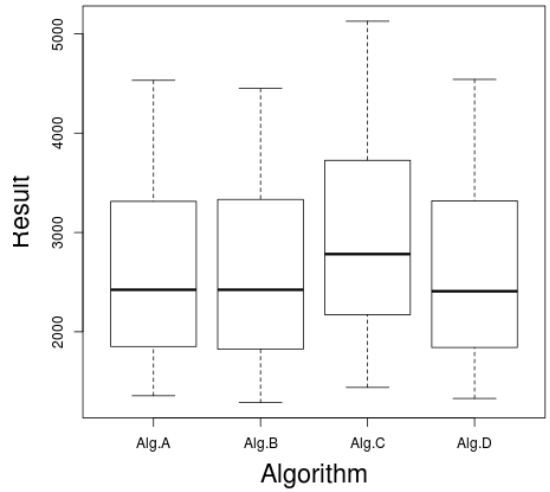
$$E = \frac{(b-1), MS_{blocks} + b(a-1)MS_E}{(ab-1)MS_E}$$

A value of 1.3, for example, would indicate that an CRD would need 30% more observations to achieve the same power.

Blocking Example (1)

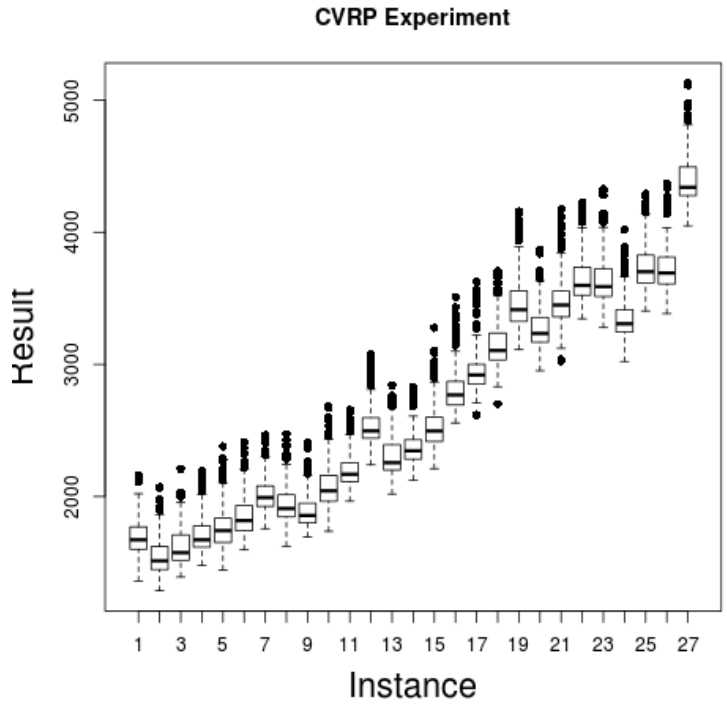
An student tries four different algorithms to the Capacitated Vehicle Routing Problem (CVRP1). These algorithms are compared by their application to 26 different benchmark problem instances. We want to compare their performances while blocking the effect of the problem instances.

```
> data = read.table("cvrp.txt", header=T)
> data$Instance <- as.factor(data$Instance)
> with(data,boxplot(Result~Algorithm,
      xlab="Algorithm",ylab="Result",cex.lab=1.8))
```



Blocking Example (2)

Graphical Analysis of the Data

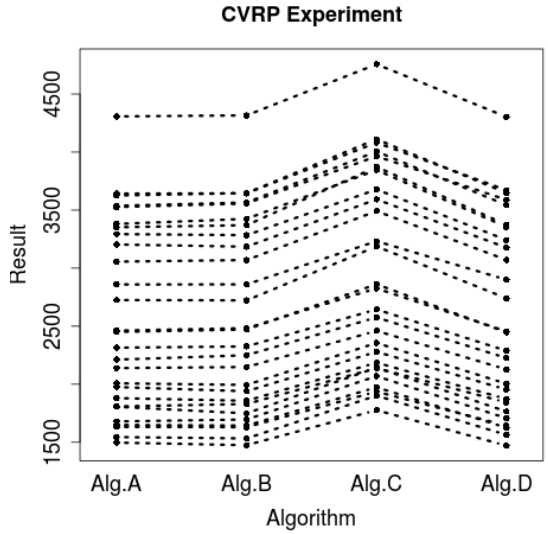



```
with(data,boxplot(Result~Instance, xlab="Instance", ylab="Result",
     main="CVRP Experiment", cex.lab=1.8,pch=16))
```

We can see clearly the influence of the different instances

Blocking Example (2)

Graphical Analysis of the Data




```
> with(data,interaction.plot(Algorithm,Instance,Result,
  xlab="Algorithm",ylab="Result", main="CVRP Experiment",
  lwd=3,cex.lab=1.5,cex.axis=1.5,cex.main=1.5,legend=F,
  type="b",pch=16,lty=3))
```

The influence of the instances seems consistent with the Algorithm Factor

Blocking Example (3)

Statistical model fit:

```
> model.blocks<-aov(Result~Algorithm+Instance,data=data)
> summary(model.blocks)
Df Sum Sq Mean Sq F value Pr(>F)
Algorithm 3 2967410 989137 916.5 <2e-16 ***
Instance 26 71386444 2745632 2543.9 <2e-16 ***
Residuals 78 84186 1079
---
```

Significant effects for both the experimental and blocking factors. The blocking efficiency is given by:

```
> b<-27; a<-4; MSb<-2745632; MSe<-1079
> ((b-1)*MSb + b*(a-1)*MSe)/((a*b-1)*MSe)
[1] 618.8988
```

suggesting that it was a very good decision to include blocking in this experimental design.

Blocking Example (3)

Now that we have our data of interest, we can apply sequential t.tests(ANOVA), to understand the relationship between the attibutes of interest.

Factorial Designs

In many experiments, we are interested in multiple factors:

- Multiple Parameters;
- Multiple Algorithms;
- Multiple Data Sets;

Using a Factorial Design we can try to measure all effect combinations.

Important Concepts

- Main Effect: Change of the response variable based on changes in one factor;
- Interaction Effect: Change of the response variable based on simultaneous changes of two or more factors;

Factorial Design Example (1)

Experiment Outline

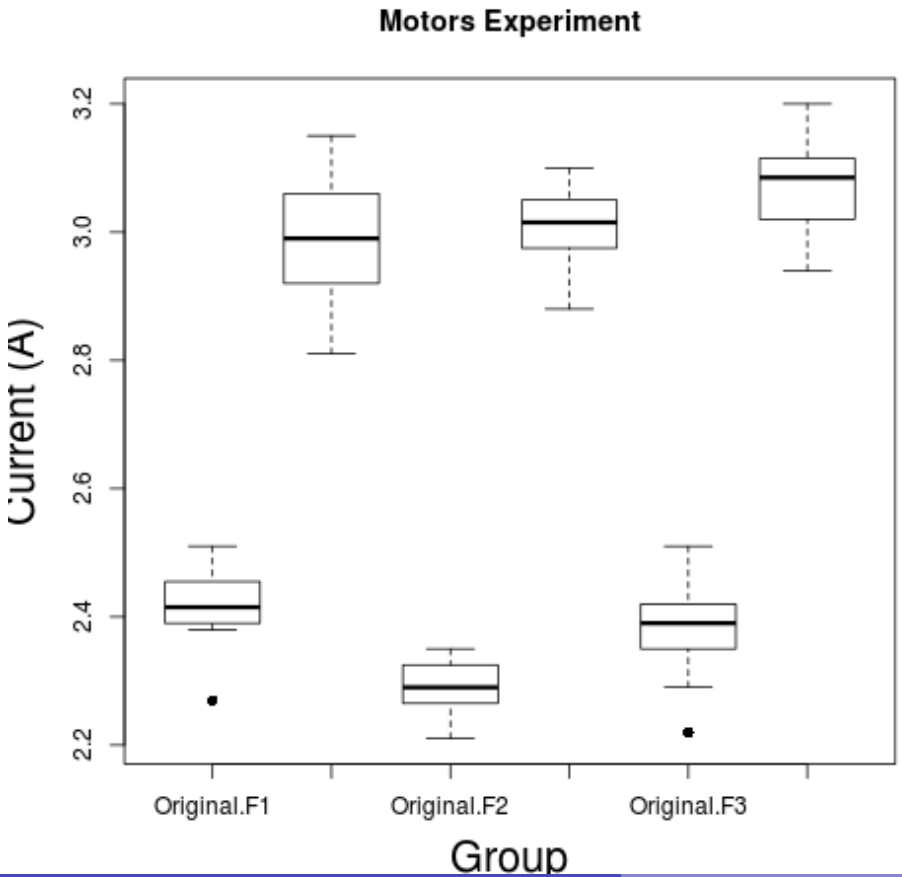
Two students wish to investigate factors that affect the electricity demand of an industrial fan.

From previous experiments, they determined that there are two factors which are likely candidates to explain the variability: Manufacturer (A or B), and State (normal or rewinded) of the motor.

Factorial Design Example (2)

Initial Data Visualization

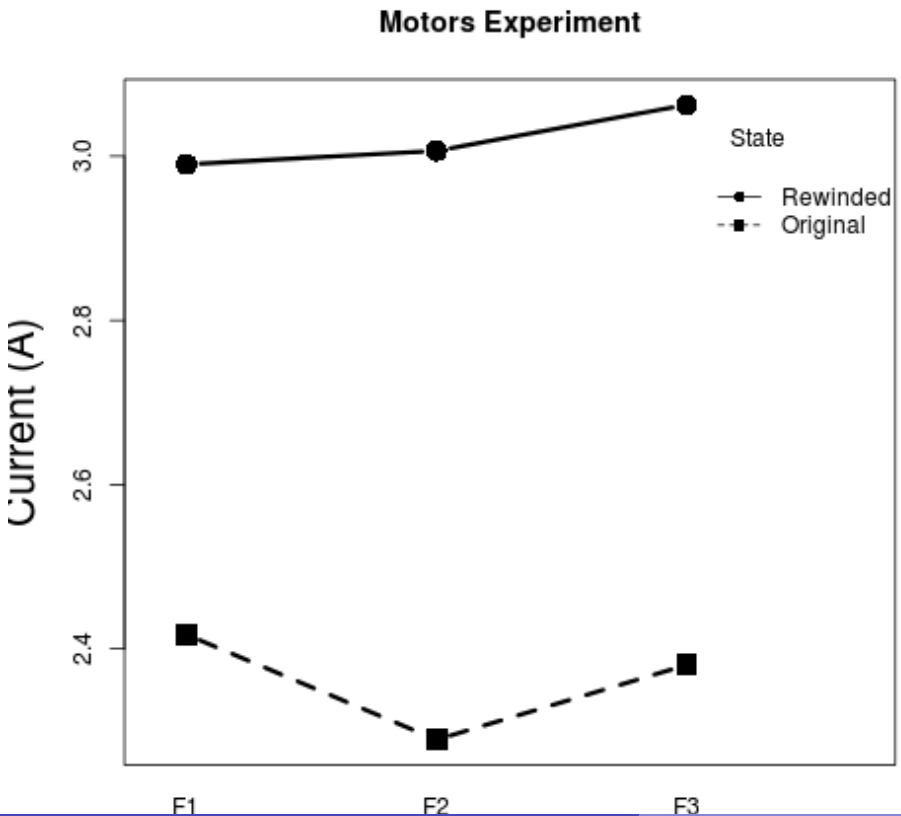
```
> with(data,boxplot(Current~State*Manufacturer, xlab="Group",
      ylab="Current (A)", main="Motors Experiment",
      cex.lab=1.8,pch=16))
```



Factorial Design Example (3)

Initial Data Visualization

```
> with(data,interaction.plot(Manufacturer,State,Current,
      xlab="Manufacturer", ylab="Current (A)",
      main="Motors Experiment", cex.lab=1.8,lwd=3,
      type="b",pch=c(15,16),cex=2))
```



- Can we see any sort of iteraction or influence between the factors?
- We don't know! Need to investigate via experiment.

General Test Statistic for Factorial Design

- a levels for factor A;
- b levels for factor B;
- n samples for each combination of levels;
- Completely randomized combination of other factors;

$$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \epsilon_{ijk}$$

Where $i = 1 \ldots a$, $j = 1 \ldots b$, $k = 1 \ldots n$. As usual, Null hypothesis is that the group errors are equal to 0.

General Test Statistic for Factorial Design

If the usual assumptions are met, MS_A/MS_E , MS_B/MS_E , and MS_{AB}/MS_E are distributed under their null hypotheses as an F variable with their respective degrees of freedom, and the hypotheses can be tested in the usual manner (i.e., comparing the obtained value of F_0 against the critical value of $F_{0;df_1;df_2}$).

```
> model<-aov(Current~State*Manufacturer,data=data)
> summary(model)
                    Df Sum Sq Mean Sq F value   Pr(>F)
State                1 12.956  12.956 2798.41  < 2e-16 ***
Manufacturer         2  0.118   0.059   12.71 1.04e-05 ***
State:Manufacturer   2  0.114   0.057   12.27 1.49e-05 ***
Residuals          114  0.528   0.005
```

If ANOVA indicates the existence of significant effects, these effects must be investigated by further statistical testing. The standard R functions discussed in last class can be used (one against all, all agains all, etc).

Paper: Ant Colony Optimization with Immigrant Schemes for the DTSP

Summary

Modifications to an ACO algorithm applied to a Dynamic Optimization Problem.

- Section 2–6: Introductions to the algorithms;
- Section 7.1: Experimental Setup;
- Section 7.2: Parameter Settings;
- Section 7.3: Variation on the Data Sets;
- Section 7.4: Variation on the Data Set Parameters;
- Section 7.5: Analysis on solution diversity;
- Section 7.6: Analysis of the “traffic factor” parameter;
- Section 7.7: Comparison with other algorithms;

Paper: Ant Colony Optimization with Immigrante Schemes for the DTSP

Ant Colony Optimization With Immigrants Schemes For The Dynamic Travelling Salesman Problem with Traffic Factors
Michalis MAvrovouniotis, Shengxiang Yang
Applied Soft Computing 13(2013)4023–4037

Take a read at section 7 of this paper – it is very instructive of Experimental design. It is not perfect! Note the positive and the negative points of the experiment Design in this paper.