STAT 2112: Design of Experiment

Let us start with the two parameter fixed effects linear model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \text{ where } \epsilon_i \sim N(0, \sigma^2) \tag{1}$$

This two parameter model is a simple model. We can write an even simpler model in which we we set $x_i = 0$ or $x_i = 1$.

In other words,

$$\begin{aligned} y_i &= \beta_0 + \epsilon_i \\ y_j &= \beta_0 + \beta_1 + \epsilon_j & (2) \\ & & (3) \end{aligned}$$

We write this model more elegantly by writing μ for β_0 and α_i for the second term.

$$y_i = \mu + \alpha_i + \epsilon_i \text{ where } \epsilon_i \sim N(0, \sigma^2) \tag{4}$$

We will impose a constraint of symmetry on the α_i's such that they are equidistant from μ as depicted in Figure 1.

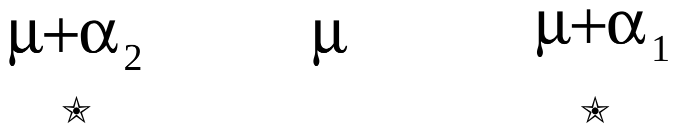

Figure 1

Mathematically, we impose the constraint that $\alpha_1 + \alpha_2 = 0$. This constraint will allow us to produce estimates and ANOVA tables otherwise the problem becomes indeterminable.

α_i is called a factor and it has two levels. We can specify more than two levels, like three level: α_1 α_2 and α_3 where the constraint of symmetry becomes $\alpha_1 + \alpha_2 + \alpha_3 = 0$. The number of levels can be as many as we want. Of course, the more levels, the more complicated the model. We will look at a two level factorial model.

In Figure 2, two factors are shown in which each factor has two levels.

The goal is to find μ. We hope that the four corners of the rectangle bracket μ. When designing an experiment, we rely on prior experience to specify the four corners.

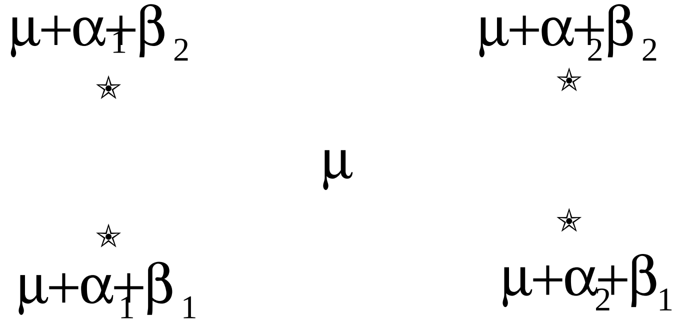

Figure 2

We will refer to the design shown in Figure 2 as 2 × 2 factorial design. In Figure 1, we will say that it depicts a 2 factorial design. We can generalize to more than one or two factor. A three factorial design would be written as 2×2×2. A schematic diagram of it would the same as the one shown in Figure 2, but instead of a rectangle, the geometric figure would be something like a cube. The mathematical expression for a 2 × 2 × 2 is given in equation 5.

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \epsilon_{ijkl} \text{ where } \epsilon_{ijkl} \sim N(0, \sigma^2) \tag{5}$$

The subscript i corresponds to factor α; the subscript j corresponds to factor β; the subscript k corresponds to factor γ. The subscript l denotes the replication. The experiment can be replicated several times; that is, the experiment can be done several times under the same experimental conditions to achieve greater precision in the estimates.

$$\begin{aligned} y_{11} &= \mu + \alpha_1 + \epsilon_{11} \text{ where } \epsilon_{11} \sim N(0, \sigma^2) \\ y_{21} &= \mu + \alpha_2 + \epsilon_{21} \text{ where } \epsilon_{21} \sim N(0, \sigma^2) \end{aligned} \tag{6}$$

For example, in equation (6), there a 2 factorial design replicated once.

$$\begin{aligned} y_{11} &= \mu + \alpha_1 + \epsilon_{11} \text{ where } \epsilon_{11} \sim N(0, \sigma^2) \\ y_{12} &= \mu + \alpha_1 + \epsilon_{12} \text{ where } \epsilon_{12} \sim N(0, \sigma^2) \\ y_{21} &= \mu + \alpha_2 + \epsilon_{21} \text{ where } \epsilon_{21} \sim N(0, \sigma^2) \\ y_{22} &= \mu + \alpha_2 + \epsilon_{22} \text{ where } \epsilon_{22} \sim N(0, \sigma^2) \end{aligned} \tag{7}$$

In equation (7), there is a 2 factorial design is replicated twice.

$$
\begin{aligned}
y_{11} &= \mu + \alpha_1 + \epsilon_{11} \text{ where } \epsilon_{11} \sim N(0,\sigma^2) \\
y_{12} &= \mu + \alpha_1 + \epsilon_{12} \text{ where } \epsilon_{12} \sim N(0,\sigma^2) \\
y_{13} &= \mu + \alpha_1 + \epsilon_{13} \text{ where } \epsilon_{13} \sim N(0,\sigma^2) \\
y_{21} &= \mu + \alpha_2 + \epsilon_{21} \text{ where } \epsilon_{21} \sim N(0,\sigma^2) \\
y_{22} &= \mu + \alpha_2 + \epsilon_{22} \text{ where } \epsilon_{21} \sim N(0,\sigma^2) \\
y_{23} &= \mu + \alpha_2 + \epsilon_{23} \text{ where } \epsilon_{23} \sim N(0,\sigma^2) \qquad (8)
\end{aligned}
$$

In equation (8), there a 2 factorial design is replicated three times.

By writing equation (7) in matrix notation, the patterns which the 2 factorial design replicated three times will be more apparent.

$$
\begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \end{bmatrix}
$$

$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$

The design matrix has an interesting structure. The first column is always filled with 1's. The second and third columns remind us a binomial random variable in which the random variable represents two outcomes: 1-0, success-failure, on-off, up-down, pass-fail, high-low. We will write design matrix again; this time we will ignore the first column, because we know that it is always filled with 1's. Instead of 1-0, let us write H for high and L for low.

$$
\begin{bmatrix} H & L \\ H & L \\ H & L \\ L & H \\ L & H \\ L & H \end{bmatrix}
$$

Actually the second column is redundant. Therefore, for the α factor, we can write the matrix shown in equation ().

$$
\begin{bmatrix} H \\ H \\ H \\ L \\ L \\ L \end{bmatrix}
$$

By looking at this matrix, we immediately recognize a 2 factorial design with two levels and replicated three times. It suggests the nature of the experiment. We will take measurements at the high condition and at the low condition. We hope that the high level and the low level will bracket μ.

Example 1. *Whenever we bake an apple pie, conditions vary especially if we do not precisely follow the recipe. In a commercial bakery, conditions needs to be highly controlled in order to produce consistently good pies. Let us pretend that we want our home kitchen to be run like a commercial bakery. The Betty Crocker recipe book was written perhaps 60 years ago when ovens then performed differently than modern ovens. We know that the temperature of the oven will determine a good pie or a bad pie. We find a judge with good discriminatory taste to evaluate our pies on a scale of 1=bad to 5=excellent. The high temperature setting will be* $400^\circ F$ *and the low temperature setting will be* $350^\circ F$. *The pies will be baked for 40 minutes. From experience, the high temperature is a little too high and the low temperature is little too low. We are confident that based on our experience, they will bracket the optimum temperature for earning a 5 from the judge.*

If were were to bake one pie at $400^\circ F$ *and another pie at* $350^\circ F$*, we will have conducted a 2 factorial design experiment replicated once. Suppose on the next day, two pies were baked at the high and low temperature settings, then the experiment will have been replicated twice. Suppose the on the third day, another two pies were baked at the high and low temperature settings, then the experiment will have been replicated three times. As the number of replicates increases, the more expensive the experiment becomes in terms of time, resources, and money.*

Following the idea of equation (), we can show the factors and levels of a 2 × 2 factorial design replicated once for the model: $y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$ where $\epsilon_{ij} \sim N(0, \sigma^2)$ by matrix shown in equation ().

$$\begin{bmatrix} H & H \\ H & L \\ L & H \\ L & L \end{bmatrix}$$

When we look at the matrix shown in equation (), we notice that there are entries for all combinations of factors and levels. It looks balances.

Suppose, on the other hand, that the matrix looked like the one shown in equation ().

$$\begin{bmatrix} H & H \\ H & L \\ L & H \end{bmatrix}$$

We see that the L-L row is missing. Not all combinations of factors and levels appear in the matrix shown in equation (). This is an example of an unbalanced design while the matrix shown in equation () refers to a balanced design. The necessary mathematics to produce ANOVA tables for an unbalanced design becomes very sophisticated. Imagine, in the case of this unbalanced design, a corner of the rectangle show in Figure 2 having disappeared. The mathematics has to deal with the missing information by employing such things as generalized inverses the nature of which conform with certain constraints of the statistician. Even though statistical software packages will produce numbers by default, the statistician must understand what mathematical constraints the software is imposing on the unbalanced design problem for producing ANOVA's. If a statistician is not careful, he may be misled by the output to make a wrong conclusion.

Example 2 (Georgetown University). *A set of data which was obtained from an experiment on examining the effects of levels of nitrogen and the structure of an habitat on the number of species of arthropods over a four month period on seven locations is based on a four parameter fixed effects linear model with the response variable, y_{ijkl}, being the number of species of arthropods:*

$$y_i = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + \epsilon_{ijkl} \textit{ where } \epsilon_{ijkl} \sim N(0, \sigma^2) \tag{9}$$

The definitions of the factors are given in Table 1.

Table 1: Definitions of Factors

Effect	Definition	Given Variable Name	Level	Meaning
y	Number of Species	Richness		
α	Fertilization Treatment	Fert	0 L H	None Low High
β	Habitat Structure	Thatch	0 Th	Thatch Removed Thatch Present
γ	Month		1 2 3 4	17 June 27 June 12 July 12 August
δ	Block	Block	1 2 3 4 5 6 7	

A model is deemed to be a good model if the underlying theory makes sense, if there appears to be a conspicuous pattern in a plot of the data, if we can reject the hypothesis that the parameters of the model are zero, and if the assumptions of the model like the assumption that the residuals are indistinguishable from white noise are valid.

The fields which are the analyst deems to be useful are: `Richness`,`Fert`,`Thatch`, *and* `Block`.

`Richness` *is the response variable. It is a measure of the number of species of arthropods which are found in an area of a certain size. According to the theory, if the habitat is fertile and healthy, arthropods should be abundant and representing many species. The scientists believe* *that applications of fertilizer* `Fert` *will improve the growth of the flora and thereby produce a better habitat for arthropods.* `Thatch` *is either removed or it is left undisturbed. Finally, there is the factor* `Block`*.*

Much of the vocabulary of experimental designs is derived from agricultural research which was conducted by Ronald Fisher and other British scientists. It seems obvious that growing conditions depend on soil, moisture, and sunlight, for example. These conditions defer by location of a plot of land. A plot might have a slope; another one might have less fertility; another might be sodden with water. The scientists called these plots, blocks. Generally, we are not interested whether one block is more productive than another, rather, we are interested in measuring the effect of fertilizer and the amount of residual thatch in affecting the richness of species of arthropods.

The term block is used in other experiments as a variable which has an effect on the response but which is not a subject of concern. It is used to eliminate a that variable as a confounding effect. For example, does attending lecture increase Final Examination scores. A blocking variable might the sex of the student; it be a significant factor, but are not interested in it. Instead, we are interested in attendance regardless of a student's sex, but block on sex as a way to eliminate it as a confounding variable.

The structure of the original set of data suggests the following four factor fixed effects linear model:

$$y_i = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + \epsilon_{ijkl} \text{ where } \epsilon_{ijkl} \sim N(0, \sigma^2) \tag{10}$$

which was formulated in the introduction as equation (9).

According to the model, we are dealing with a five dimensional set of data. As such, it is impossible to make a picture of the data at once. Instead, we will look at two dimensional slices of the data, in order to discover whether there exists any conspicuous relationships between the variables or trends in the data.

0.1 Make a Picture of the Data

To begin with, we will look at each variable individually as if there are not other factors present. The presence of other factors will muddy any patterns which might appear, but in the gross context, we might see some obvious trends.

There appears in Figure 3 an increase in Richness due to a high level of fertilizer over no fertilizer. We, therefore, should expect to reject the hypothesis that the richness is the same across levels of fertilizer in the ANOVA table.

Figure 4 suggests that removing the thatch does not affect the richness in species of the area.

Perhaps month has an affect on richness as suggested in Figure 5. We might see a corresponding significant effect in the ANOVA table.

A blocking factor is used to eliminate an effect. Presumably, `block` *accounts for seven geographic areas. Whether there is a difference in richness between locations evidently is not a*

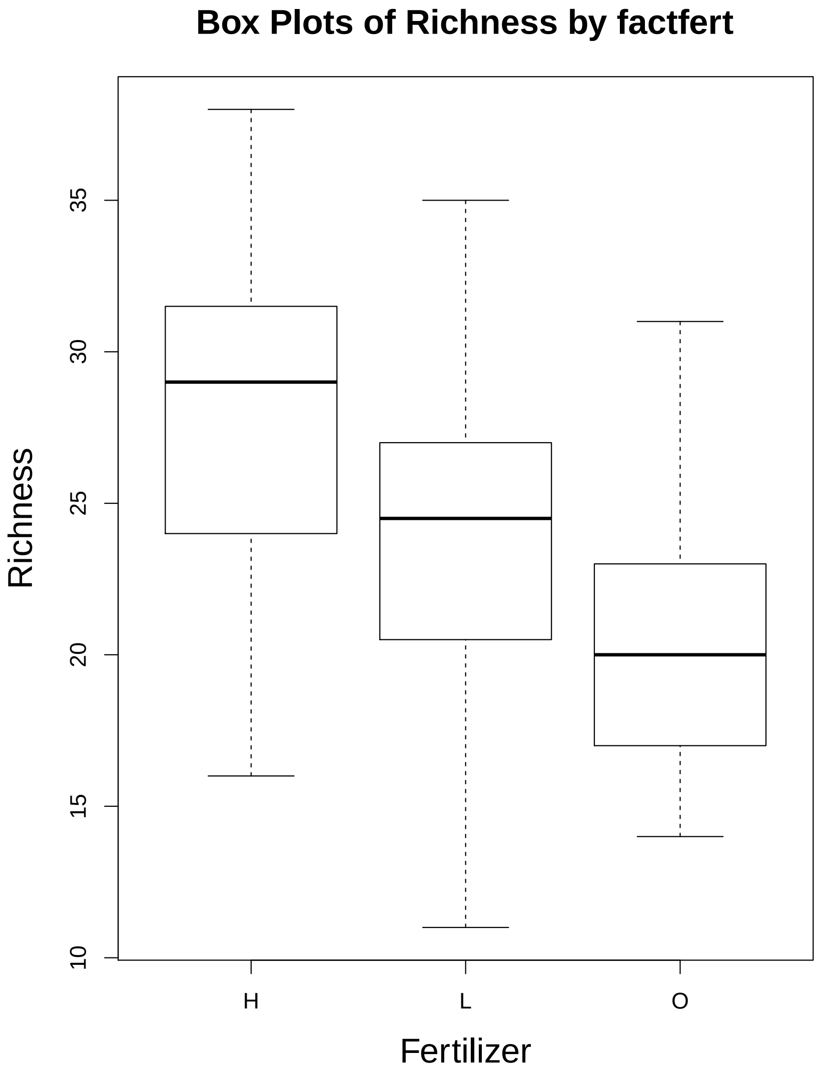

Figure 3: Box Plots of Richness by Levels of Fertilizer

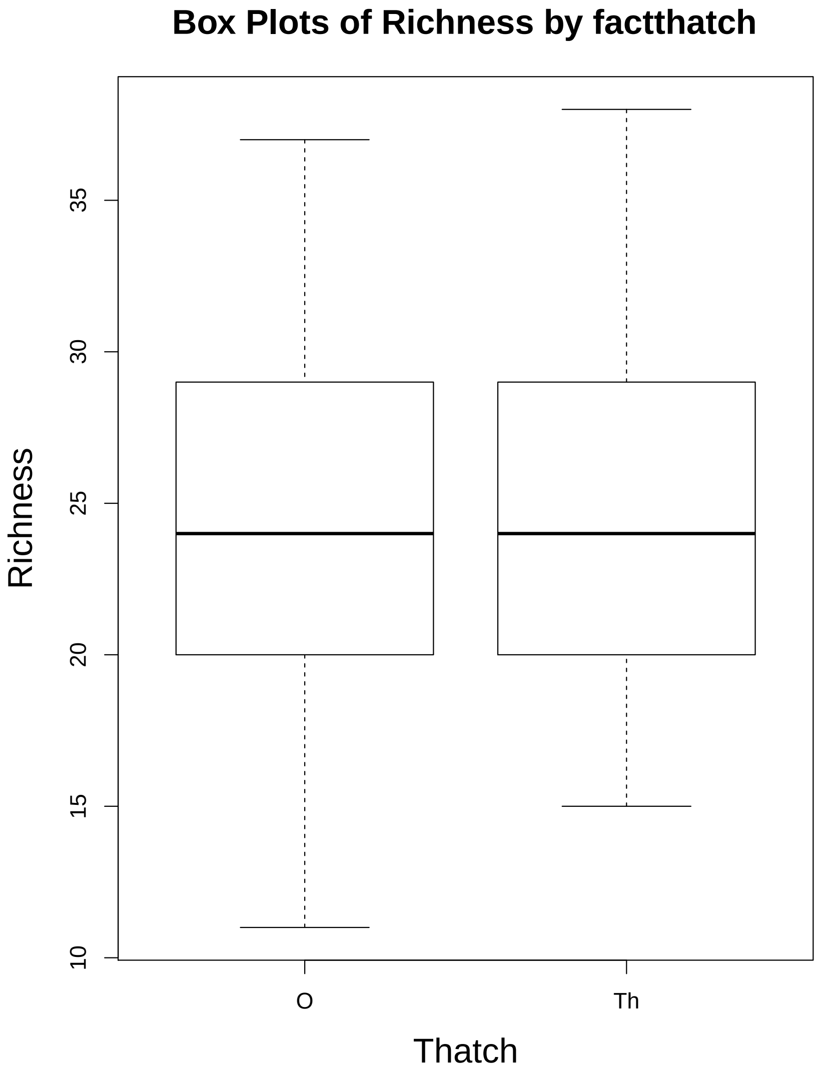

Figure 4: Box Plots of Richness by Levels of Thatching

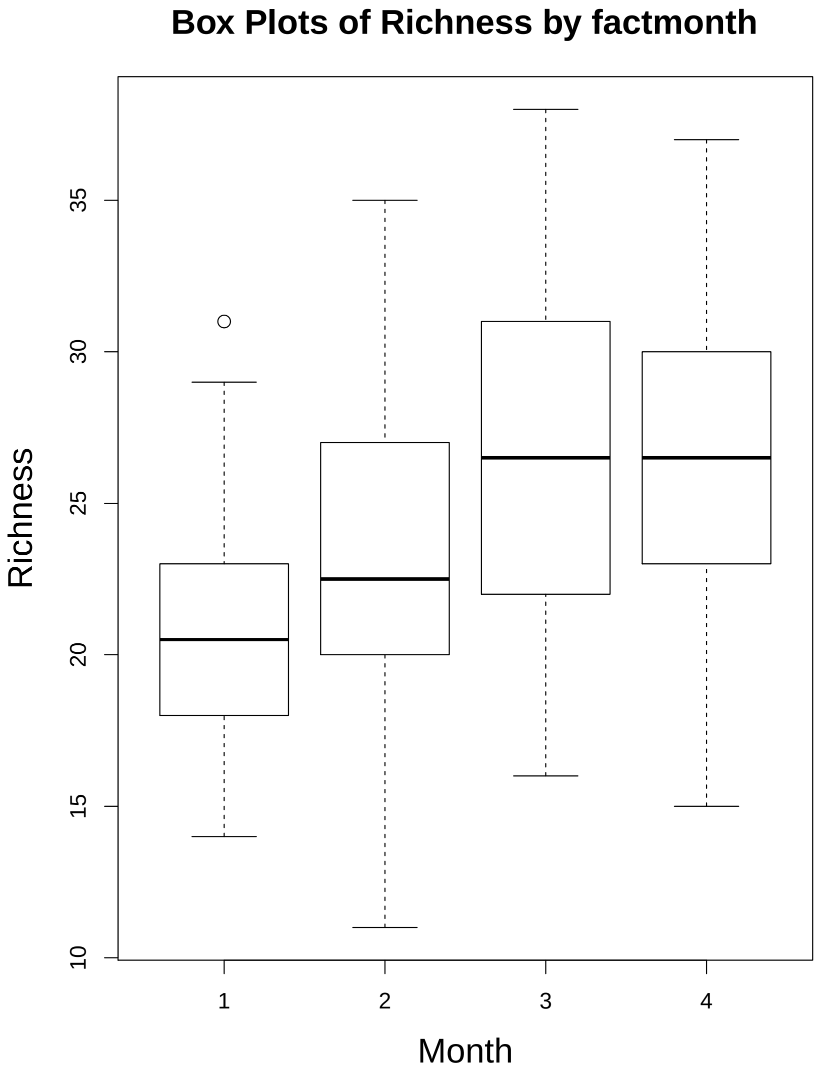

Figure 5: Box Plots of Richness by Levels of Month

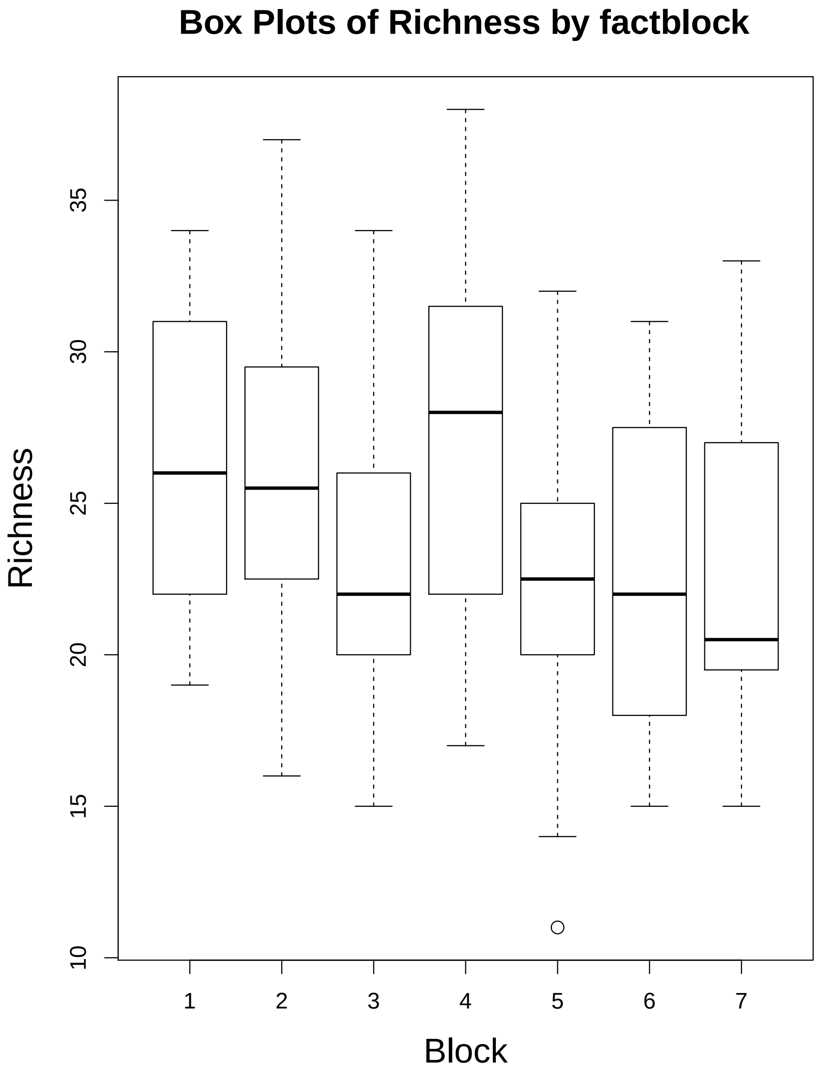

Figure 6: Box Plots of Richness by Levels of Blocks

matter of concern whereas within a block the factors of fertilizer, thatching, and month are being examined to assess whether they can explain the response.

0.2 Analysis of Variance Table

An analysis of variance table can be produced by means of some statistical software product. A model such as the one given by equation 10 and is written as:

$$y_i = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + \epsilon_{ijkl} \text{ where } \epsilon_{ijkl} \sim N(0, \sigma^2) \tag{11}$$

By means of the R software program, the following ANOVA is produced:

```
Analysis of Variance Table

Response: Richness
            Df  Sum Sq Mean Sq F value    Pr(>F)
    fert     2 1620.33  810.17 70.3243 < 2.2e-16 ***
    thatch   1   24.38   24.38  2.1163    0.1478
    block    6  685.24  114.21  9.9134 3.000e-09 ***
    month    3  917.79  305.93 26.5553 6.467e-14 ***
    Residuals  155 1785.67   11.52
    ---
    Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see that the factor for Thatch has a p-value of .1478. We may assume that it does not make a significant contribution in explaining the response variable, Richness.

As was expected from inspecting Figures 3 and 5, fertilizer and month are significant factors in explaining Richness. *Somewhat surprising is that* Block *is an important factor even though according to the box plots given in Figure 6, there does not appear to be a conspicuous difference between them. We need to keep in mind that a plot of the data like Figure 6 is only a two dimensional slice of a five dimensional set of data.*

In order to asses the validity of the assumption of the model that $\epsilon_{ijkl} \sim N(0, \sigma^2)$, that is, the assumption that the residuals resemble white noise. The plot of residuals versus predicted values shown in Figure 7 shows a random pattern; therefore, we may consider that the assumption of the ϵ_{ijkl}'s is valid.

Confidence intervals, of course, are of paramount importance in making statistical inferences. Because the current set of data lies in a five dimensional space, five dimensional confidence regions are impossible to draw. For the same reason when a series of box plots were made to examine the data in two dimensional slices, two dimensional confidence intervals are constructed for Richness *according to* month *and* Thatch. *Though the effects of* Block *are confounded in the confidence intervals, the confidence interval provide a useful portrayal of the effects of* Fert, Thatch, *and* month *on* Richness.

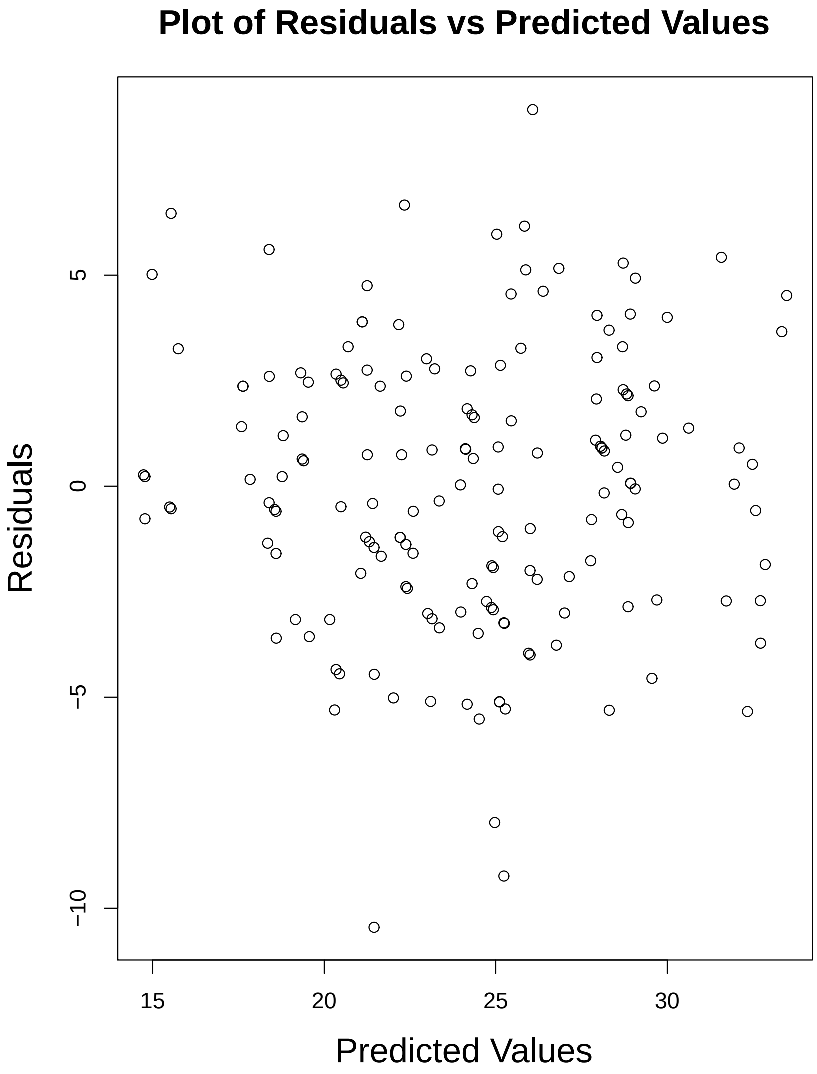

Figure 7: Diagnostic Plot for the Assumption of Normality of the Residuals

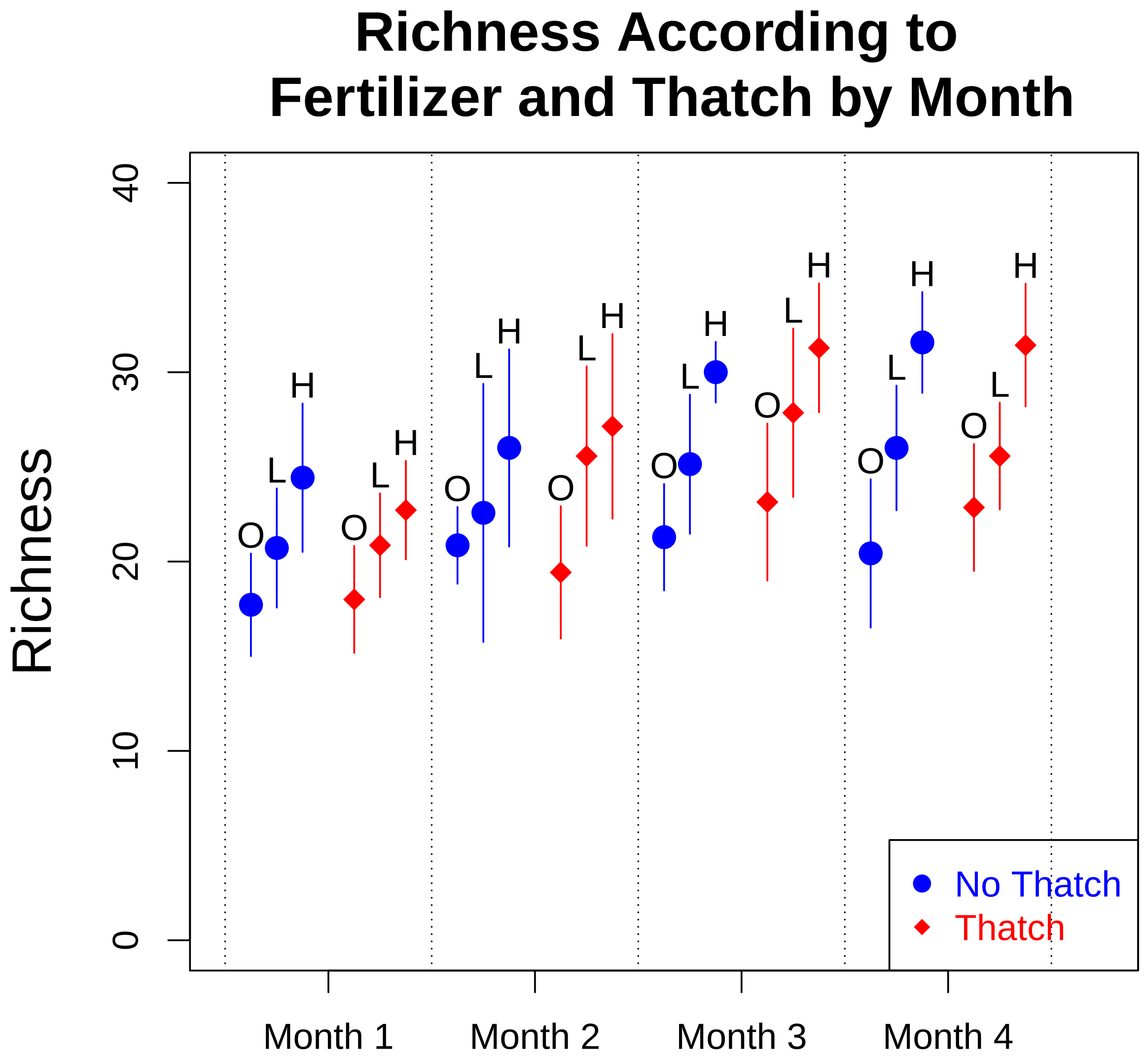

Figure 8: 95% Confidence Intervals of Richness by Fert, Thatch, and month.

The pattern of applying fertilizer improve richness is apparent in Figure 8. High fertilizer always produces a higher richness. We see in the same figure that keeping the thatch or removing will not affect richness to which the ANOVA agrees.

In conclusion, the scientists showed that the vitality of the flora which a high level of fertilizer promotes is an important factor regardless of location, month, and the presence of thatch.