

# N-Way Analysis of Variance

## 1 Introduction

A good example when to use a n-way ANOVA is for a factorial design. A factorial design is an efficient way to conduct an experiment. Each observation has data on all factors, and we are able to look at one factor while observing different levels of another factor. Table 1 shows an analysis of variance table for a three-factor design. This can obviously be extended to more factors or be used for two factors.

Table 1: Two-Way ANOVA

Source	DF	SS	Mean Square	F
Total	$rab - 1$	SS Total		
Factor A	$a - 1$	SS(A)	MS(A)	$MS(A)/MSE$
Factor B	$b - 1$	SS(B)	MS(B)	$MS(B)/MSE$
Factor C	$c - 1$	SS(C)	MS(C)	$MS(C)/MSE$
A*B Interaction	$(a-1)(b-1)$	SS(AB)	MS(AB)	$MS(AB)/MSE$
B*C Interaction	$(b-1)(c-1)$	SS(BC)	MS(BC)	$MS(BC)/MSE$
A*B*C Interaction	$(a-1)(b-1)(c-1)$	SS(ABC)	MS(ABC)	$MS(ABC)/MSE$
Error	$abc(r-1)$	SSE	MSE	

## 2 Analysis of Variance Preparation

This example will use the dataset *npk* (Nitrogen - N, Phosphate - P, Potassium - K) that is available in R with the MASS library (Modern Applied Statistics with S). The first thing to do with analysis of variance is to inspect the data to ensure it is formatted and designed properly. Inspecting the balance of the design can be done with the following R code. This shows that the design is balanced and the analysis not require additional considerations.

```
>replications(yield ~ N*P*K, data=npk);  
  
  N     P     K  N:P  N:K  P:K N:P:K  
12  12   12    6    6    6     3
```

There are several ways to graphically review the data. A good graphical approach is to create side-by-side boxplots of each of the experimental combinations. Figure 1 show the boxplot for the *npk* data.

Because the *npk* data has three factors each with two factor levels the interactions should be reviewed. A good way to accomplish this is graphically using an interaction plot. Figure 2 shows the output from the following R code. This code is a concise way to display two interaction plots for the nitrogen and potassium and the nitrogen and phosphate.

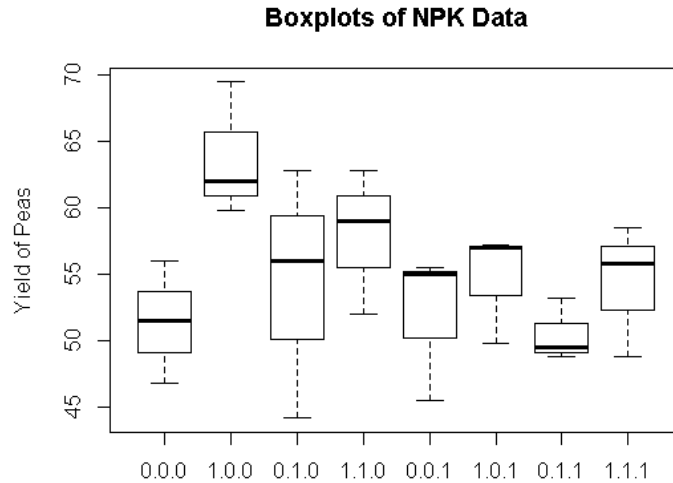


Figure 1: Box Plot of Experimental Combinations

```
>with(npk, {
  interaction.plot(N, P, yield, type="b", legend="T",
    ylab="Yield of Peas", main="Interaction Plot", pch=c(1,19) )
  interaction.plot(N, K, yield, type="b", legend="T",
    ylab="Yield of Peas", main="Interaction Plot", pch=c(1,19))
})
```

It is generally a good idea to look at the basic summary statistics for the data. This will give a researcher a good idea what to expect from the data and to identify any observation that may be outliers or may have suffered from data input errors. The following code provides table summaries of each of the factors compared to the other factors.

```
with(npk, tapply(yield, list(N,P), mean));
with(npk, tapply(yield, list(N,K), mean));
with(npk, tapply(yield, list(P,K), mean));

> with(npk, tapply(yield, list(N,P), mean));
  0      1
0 51.71667 52.41667
1 59.21667 56.15000
> with(npk, tapply(yield, list(N,K), mean));
  0      1
0 52.88333 51.25000
1 60.85000 54.51667
> with(npk, tapply(yield, list(P,K), mean));
  0      1
```

```
0 57.60000 53.33333
1 56.13333 52.43333
```

### 3 Analysis of Variance

Once the data has been reviewed and a researcher has a complete understanding of the design and scope of the data then the ANOVA table can be produced. Analysis of variance in R is quite simple and can generally be done using only a couple of lines of code. The following code produces an analysis of variance table and shows the difference in means using the Tukey HSD (Honestly Significant Difference) test. Additionally, the difference and the family-wise confidence levels can be easily graphed as seen in Figures 3 and 4. The following code produces quite a bit of output showing the differences and the confidence intervals for all comparisons of means.

```
>npk.aov <- aov(yield ~ N*P*K, data=npk);
>TukeyHSD(npk.aov, conf.level=.99);
>plot(TukeyHSD(npk.aov, conf.level=.99));
```

At this point the analysis of variance table can be produced. We can look at the output a few different ways. First we can look at the traditional ANOVA table to determine the significance of the main effects and the interactions (interactions are identified by a colon ':'). A second way to look at the ANOVA output is by looking at treatment effects. The output from the below code shows that the mean when all factors are set to zero is 51.4333. From there we can see the effect of the other variables.

```
>summary(npk.aov);
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
N	1	189.3	189.28	6.161	0.0245 *
P	1	8.4	8.40	0.273	0.6082
K	1	95.2	95.20	3.099	0.0975 .
N:P	1	21.3	21.28	0.693	0.4175
N:K	1	33.1	33.13	1.078	0.3145
P:K	1	0.5	0.48	0.016	0.9019
N:P:K	1	37.0	37.00	1.204	0.2887
Residuals	16	491.6	30.72		

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

```
>options("contrasts");
>summary.lm(npk.aov);
```

```
Call:
aov(formula = yield ~ N * P * K, data = npk)
```

```
Residuals:
      Min       1Q   Median       3Q      Max
-10.133  -4.133   1.250   3.125   8.467
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  51.4333     3.2002  16.072  2.7e-11 ***
N1           12.3333     4.5258   2.725   0.015 *
P1            2.9000     4.5258   0.641   0.531
K1            0.5667     4.5258   0.125   0.902
N1:P1       -8.7333     6.4004  -1.365   0.191
N1:K1       -9.6667     6.4004  -1.510   0.150
P1:K1       -4.4000     6.4004  -0.687   0.502
N1:P1:K1    9.9333     9.0515   1.097   0.289
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 5.543 on 16 degrees of freedom
Multiple R-squared: 0.4391, Adjusted R-squared: 0.1937
F-statistic: 1.789 on 7 and 16 DF, p-value: 0.1586
```

## 4 Final Diagnostics

Finally, diagnostics should be conducted to ensure that ANOVA is a proper fit and that all of the basic assumptions are met. These primarily include normality of data and homogeneity of variance. A good way to visually determine if the data are normally distributed is by using a Q-Q plot as seen in Figure 5. If the point follow Q-Q line then data follow a normal distribution if there is a lot of deviation from the line then the data should be inspected further. Additionally, the Shapiro-Wilk test for normality is another good way to identify (using a null hypothesis of normal data) whether or not the data is from a normal distribution. The homogeneity of variance can be determined using the Bartlett Test of Homogeneity (seen in the output below) of Variances or the Fligner-Killeen Test of Homogeneity of Variances.

```
plot(npk.aov);
plot.design(yield~N*P*K, data=npk);

qqnorm(npk$yield); qqline(npk$yield, col=4);

##Shapiro-Wilk Normality Test
by(npk$yield, npk$N, shapiro.test);
by(npk$yield, npk$P, shapiro.test);
by(npk$yield, npk$K, shapiro.test);

> by(npk$yield, npk$N, shapiro.test);
npk$N: 0

Shapiro-Wilk normality test
```

```
data: dd[x, ]
W = 0.9578, p-value = 0.7522
```

---

```
npk$N: 1
```

```
Shapiro-Wilk normality test
```

```
data: dd[x, ]
W = 0.9642, p-value = 0.8414
```

```
> by(npk$yield, npk$P, shapiro.test);
npk$P: 0
```

```
Shapiro-Wilk normality test
```

```
data: dd[x, ]
W = 0.9629, p-value = 0.824
```

---

```
npk$P: 1
```

```
Shapiro-Wilk normality test
```

```
data: dd[x, ]
W = 0.9574, p-value = 0.7456
```

```
> by(npk$yield, npk$K, shapiro.test);
npk$K: 0
```

```
Shapiro-Wilk normality test
```

```
data: dd[x, ]
W = 0.9721, p-value = 0.9313
```

---

```
npk$K: 1
```

```
Shapiro-Wilk normality test
```

```
data: dd[x, ]
W = 0.9189, p-value = 0.2766
```

```
> bartlett.test(npk$yield ~ npk$N);
```

```
Bartlett test of homogeneity of variances
```

```
data: npk$yield by npk$N
Bartlett's K-squared = 0.0577, df = 1, p-value = 0.8102
```

```
> bartlett.test(npk$yield ~ npk$P);
```

Bartlett test of homogeneity of variances

data: npk\$yield by npk\$P

Bartlett's K-squared = 0.1555, df = 1, p-value = 0.6933

```
> bartlett.test(npk$yield ~ npk$K);
```

Bartlett test of homogeneity of variances

data: npk\$yield by npk\$K

Bartlett's K-squared = 3.0059, df = 1, p-value = 0.08296

```
fligner.test(npk$yield ~ npk$N);
```

```
fligner.test(npk$yield ~ npk$P);
```

```
fligner.test(npk$yield ~ npk$K);
```

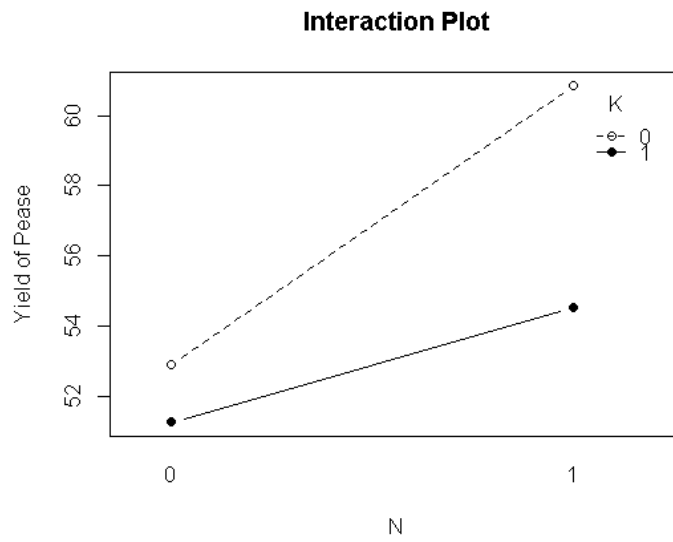
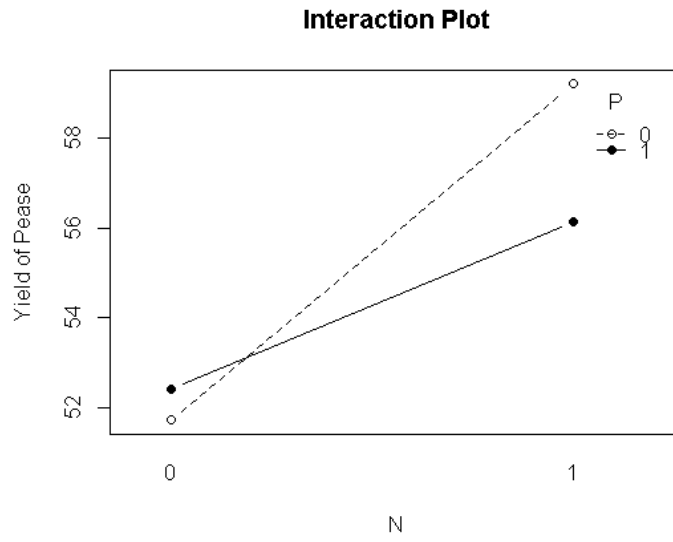


Figure 2: Interaction Plot of Factor Levels

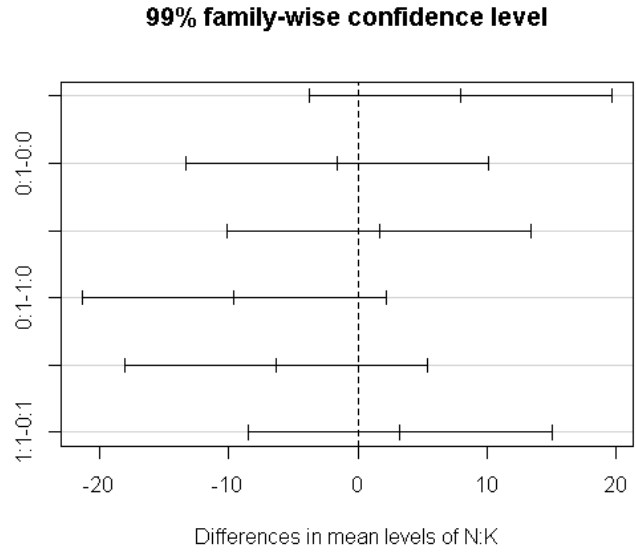
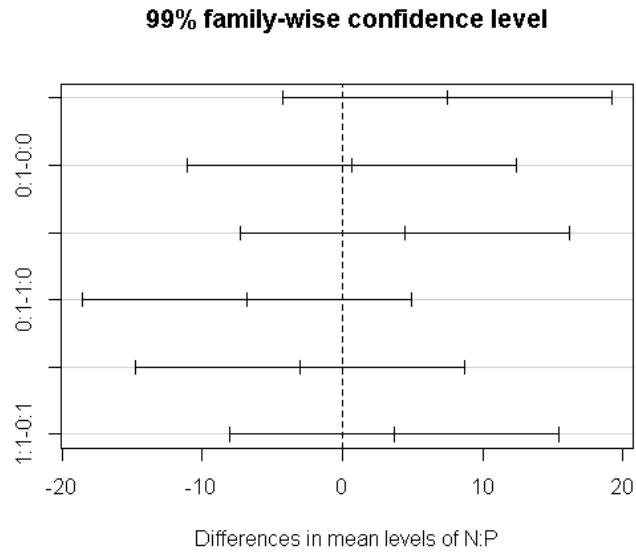


Figure 3: Interaction Plot of Factor Levels



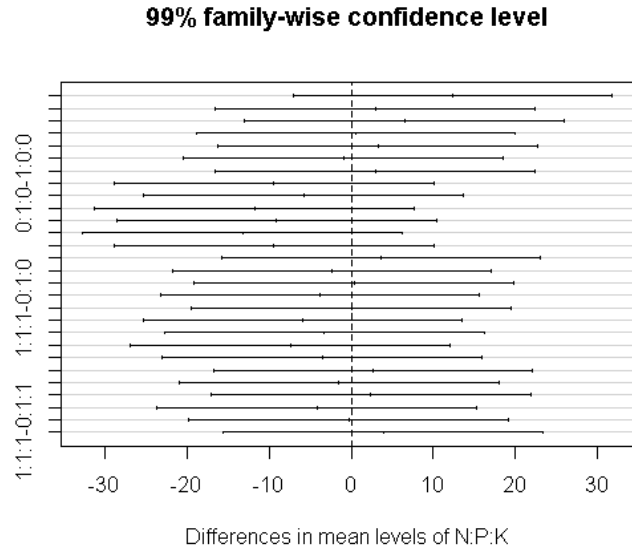
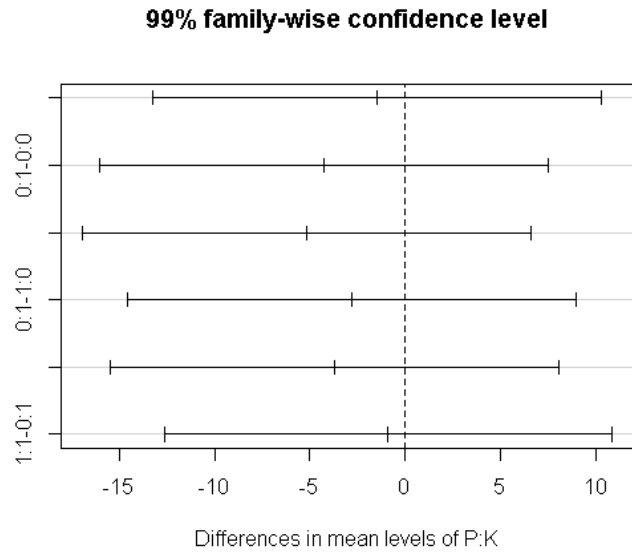


Figure 4: Interaction Plot of Factor Levels

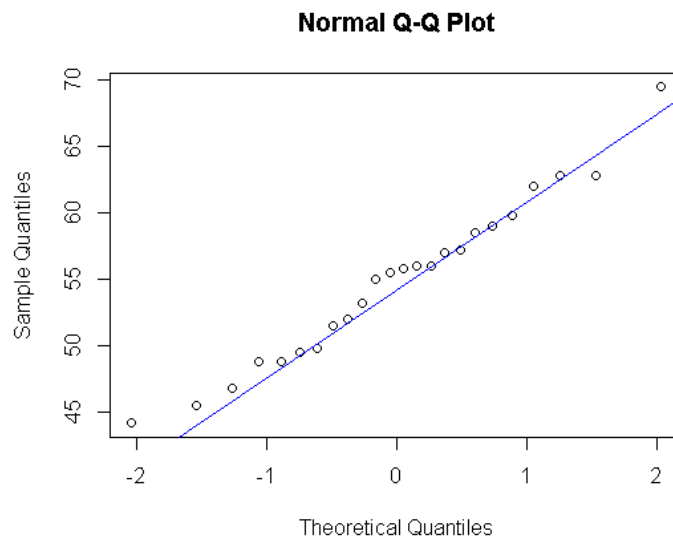


Figure 5: Normal Q-Q Plot