Session 4 - Type I Error Inflation and Analysis of Variance

Brian Leroux

Wednesday, February 6, 2019

Outline

1. Multiple Testing and Type I Error Inflation
2. Analysis of Variance (ANOVA)
3. Checking Assumptions of ANOVA - Residual Diagnostics
4. Other Experimental Designs

1. Multiple Testing and Type I Error Inflation

- The problem of multiple testing
- Pairwise comparisons of 3 or more population means
- Bonferroni correction

The Problem of Multiple Testing

The logic of hypothesis testing applies to a **single** hypothesis test:

If *this* H_0 is true then my chance of *falsely* rejecting **it** is α.

But many studies or data analyses involve multiple testing, i.e., conducting multiple hypothesis tests. The above statement applies to *each test separately* but does not say anything about the overall ("experiment-wise") chance of making a type I error or **how many** type I errors I might have made.

If several hypothesis tests are performed each at significance level 0.05, then then the chance of *at least one* type I error *if all null hypotheses are true*, may be much larger than 0.05 (this is called "inflation of the type I error rate".)

"The Religion of Statistics as Practiced in Medical Journals"

- Satirizes the use and abuse of p-values in medical journals ("p-hacking")
- Highly recommended reading (article posted in canvas).
- Might seem dated in some respects but ideas just as pertinent today.
- Note: Most scientific fields have the same problem!

> *The gods are a bit stupid. Even if you run various modifications of the data through the same computer program again and again, the gods never catch on and keep presenting you with new p values.*

The Religion of Statistics as Practiced in Medical Journals by D.S. Salsburg, *The American Statistician* 39(3):220-223, 1985.

P-hacking and its synonyms

- Data dredging
- Data snooping
- Fishing expedition
- Significance questing
- etc.

A crisis of reproducibility in scientific research

"Why most published research findings are false," Ioannidis, J.P.A. PLOS MEDICINE 2(8):696-701, 2005.

> *There is increasing concern that most current published research findings are false.*

"Trouble at the lab." *The Economist* Oct 18, 2013.

> *Scientists like to think of science as self-correcting. To an alarming degree, it is not.*

Examples of Multiple Testing

1. Multiple outcome variables: e.g., testing effects of experimental conditions on output of a process when there are several measures of "output" (e.g., total output, average output per batch, defect-free output, etc.)
2. Multiple looks at the data: a particular problem in clinical trials, in which it is necessary to **monitor** the data to ensure patients are not being harmed - but we must guard against stopping the trial as soon as we see a significant difference ("sequential analysis" methods are designed to control type I error rates in this context).
3. Sub-group analyses: testing for treatment effects within sub-groups of experimental units (e.g., types of patients in clinical trials).
4. Comparing populations using categorical analysis of multiple cut-points of a continuous variable.
5. Pairwise comparisons of 3 or more means.
6. Step-wise regression: with enough candidate variables, you will always find a model with "significant" predictors.

Pairwise comparisons of 3 or more means

Review: The Two-Sample Equal-Variance T-Test

The Setting: We have two independent random samples from two populations. The means of the populations are denoted μ_1 and μ_2, and the variances are σ_1^2 and σ_2^2, respectively. The sample means are $\bar{X}_1$ and $\bar{X}_2$. The sample variances are s_1^2 and s_2^2 and the pooled sample variance is
$s^2 = [(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]/(n_1 + n_2 - 2)$.

For now, we will focus on the equal-variance t-test. The reason is that methods such as Analysis of Variance and linear regression which we will be studying next, typically assume equal variances.

The equal-variance t-statistic is calculated as

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s^2(1/n_1 + 1/n_2)}}.$$

Decision rule: Reject $H_0 : \mu_A = \mu_B$ if $|Z| > t_{\alpha, n_A+n_B-2}$, where t_{α, n_A+n_B-2} is the critical value from the $t_{n_A+n_B-2}$ distribution with 2-sided tail probability α.

Note: if the sample sizes are sufficiently large, we may use the critical value from the N(0,1) distribution because it is almost the same as the value from the t-distribution.

Example: Comparing Species in the Iris Data

Test the null hypothesis that mean sepal width is different for *setosa* versus *versicolor*. We are assuming that we have two independent random samples of plants from the two species and that the variances for the two populations are equal.

```
options(width=64)
t.test(iris$Sepal.Width[iris$Species=="setosa"],
       iris$Sepal.Width[iris$Species=="versicolor"],var.equal=T)
```

```
    Two Sample t-test

data:  iris$Sepal.Width[iris$Species == "setosa"] and iris$Sepal.Width[iris$Spe
t = 9.455, df = 98, p-value = 1.845e-15
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.519895 0.796105
sample estimates:
mean of x mean of y
    3.428     2.770
```

Verify using formulas:

```
m=with(iris,tapply(Sepal.Width,Species,mean))
s=with(iris,tapply(Sepal.Width,Species,sd))
n=with(iris,tapply(Sepal.Width,Species,length))
z=(m[1]-m[2])/sqrt(s[1]^2/n[1]+s[2]^2/n[2])
z
```

```
  setosa
9.454976
```

```
2*(1-pt(z,df=n[1]+n[2]-2))
```

```
      setosa
1.776357e-15
```

Comparing the Means for 3 Species

Let the mean Sepal Width for the 3 species by μ_1, μ_2, μ_3, for *setosa*, *versicolor*, and *virginica*, respectively. Suppose that we want to test the null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3$ against the alternative hypothesis that the 3 means are *not* all equal.

Intuitively, we might try to do this by testing all the pairwise hypotheses, e.g., $H_0 : \mu_1 = \mu_2$, $H_0 : \mu_1 = \mu_3$, and $H_0 : \mu_2 = \mu_3$. The problem with this is that we have 3 answers to our one question. For the iris data, all of these 3 hypothesis tests reject the null hypotheses (see next slide) and so there is no great problem with interpretation. We conclude that all 3 species differ.

Example: pairwise comparisons of species

```
setosa.vs.versicolor=t.test(iris$Sepal.Width[iris$Species=="setosa"],
       iris$Sepal.Width[iris$Species=="versicolor"],var.equal=T)
setosa.vs.versicolor$p.value
```

```
[1] 1.84526e-15
```

```
setosa.vs.virginica=t.test(iris$Sepal.Width[iris$Species=="setosa"],
       iris$Sepal.Width[iris$Species=="virginica"],var.equal=T)
setosa.vs.virginica$p.value
```

```
[1] 4.246355e-09
```

```
versicolor.vs.virginica=t.test(iris$Sepal.Width[iris$Species=="versicolor"],
       iris$Sepal.Width[iris$Species=="virginica"],var.equal=T)
versicolor.vs.virginica$p.value
```

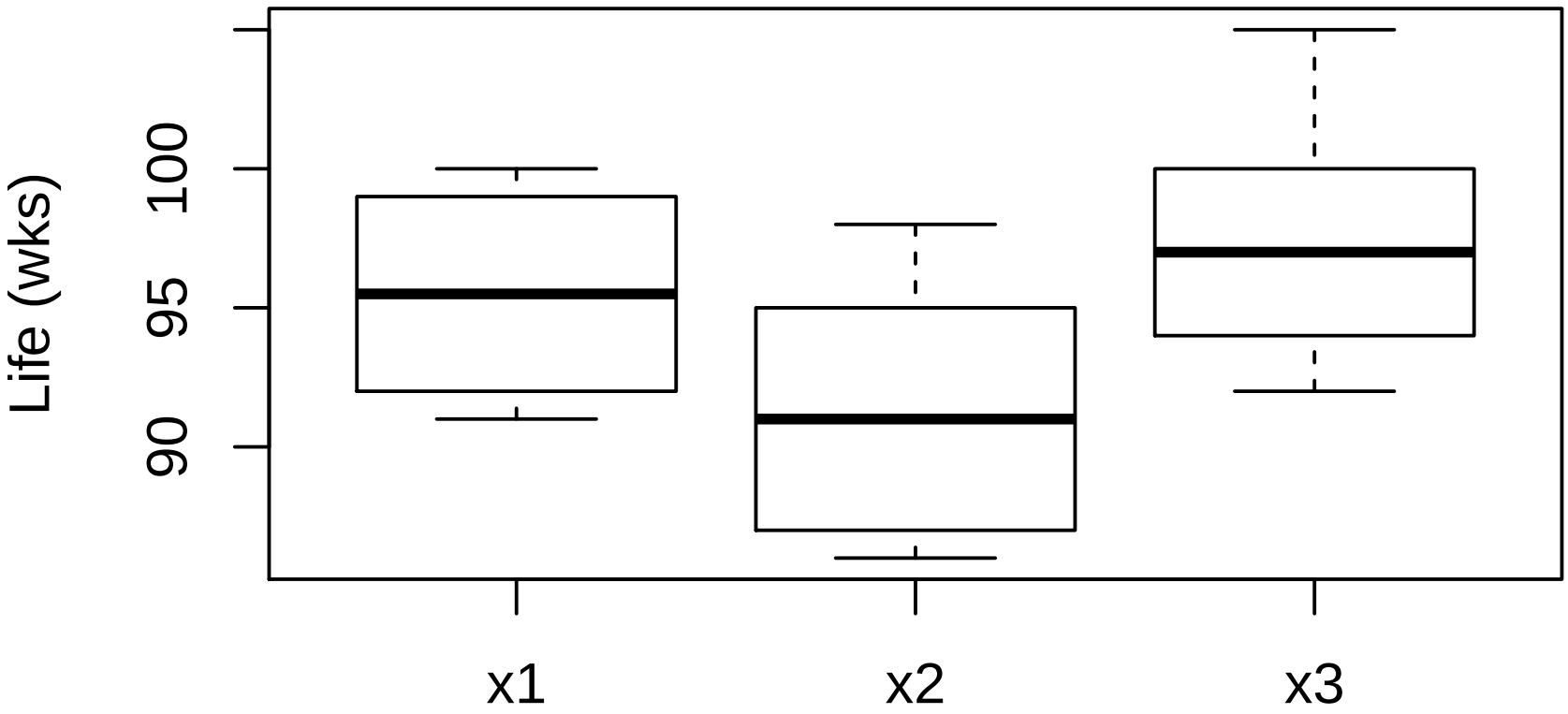
```
[1] 0.0018191
```

When all 3 comparisons result in rejection, the interpretation is clear. However, it can be difficult to interpret results in other situations.

Example 2: comparison of life of 3 brands of batteries

Random samples of 5 batteries from each of 3 different brands are tested and the following life lengths (in weeks) are obtained. Do the brands have different mean life lengths?

```
x1=c(100,96,92,99,91,95)
x2=c(86,90,95,87,92,98)
x3=c(105,94,96,98,100,92)
boxplot(cbind(x1,x2,x3),ylab="Life (wks)")
```



Pairwise comparisons

```
p12=t.test(x1,x2,var.equal=T)$p.value
p13=t.test(x1,x3,var.equal=T)$p.value
p23=t.test(x2,x3,var.equal=T)$p.value
data.frame(p12,p13,p23)
```

```
        p12       p13       p23
1 0.1132238 0.4243729 0.0439232
```

The test comparing brands 2 and 3 rejects the null hypothesis at significance level 0.05. But the other tests do not reject the null hypothesis. What conclusions should we draw?

The problem of multiplicity

The difficulty in interpretation of multiple pairwise t-tests is due to the fact that the *overall* type I error of the experiment is inflated.

Each of the 3 tests has a 5% chance of rejecting the null hypothesis. Therefore, the chance that at least one of the tests rejects even if all 3 brands are equal (i.e., $H_0 : \mu_1 = \mu_2 = \mu_3$ is true) is **greater than** 0.05.

Inflation of Type I Error due to Multiple Testing

Let's do a simulation study to illustrate the problem. Generate 3 samples of data from the same distribution, perform all pairwise t-tests, and see how often we get at least one of the tests resulting in a type I error.

```
set.seed(5)
n=5
m=90
sigma=2
reps=2000
pvalues=data.frame(p12=rep(NA,reps),p13=rep(NA,reps),p23=rep(NA,reps))
for(i in 1:reps){
  x1=rnorm(n,m,sigma)
  x2=rnorm(n,m,sigma)
  x3=rnorm(n,m,sigma)
  pvalues$p12[i]=t.test(x1,x2,var.equal=T)$p.value
  pvalues$p13[i]=t.test(x1,x3,var.equal=T)$p.value
  pvalues$p23[i]=t.test(x2,x3,var.equal=T)$p.value
}
reject = data.frame(pvalues < 0.05, any.rejection=apply(pvalues<0.05, 1, any))
apply(reject,2,mean)
```

```
          p12           p13           p23 any.rejection
       0.0520        0.0510        0.0570        0.1335
```

Each of the three tests has an appropriate type I error probability of close to 0.05 but the overall type I error probability is greater than 0.13.

A simple (but not adequate) fix

If we perform 3 tests, each of which has probability 0.05 of rejecting, then the probability of at least one of the tests rejecting cannot be greater than 0.15.[1]

Note that in our simulation the probability of rejecting was 0.1335, less than 0.15.

The idea of the *Bonferroni* correction is to adjust the significance level of each test so that the overall probability of a type I error is controlled at the desired level. For example, if we perform 3 tests and we want an overall type I error of 0.05, we would perform each test with significance level 0.05/3.

This is a very general way to correct for multiple testing in any setting.

[1] The proof uses the result from probability theory that the probability of a union of events is less than or equal to the sum of their probabilities.

Illustrating the Bonferroni Correction

```
reject.Bonf=data.frame(pvalues < 0.05/3, any.rejection=apply(pvalues<0.05/3, 1,
apply(reject.Bonf,2,mean)
```

```
      p12       p13       p23 any.rejection
   0.0175    0.0165    0.0190        0.0465
```

Now the overall type I error probability is approximately 0.05.

This means that the Bonferroni correction procedure is a valid test of the null hypothesis of equal group means.

A disadvantage of the Bonferroni procedure is that it is conservative, i.e., it leads to a test that will have less than the desired type I error probability. Note that in the example the type I error probability was estimated to be 0.0465.

The Bonferroni procedure is a useful procedure in a wide variety of settings and we will return to it. However, for comparison of means there is a preferred procedure, the Analysis of Variance.

2. Analysis of Variance (ANOVA)

- The ANOVA F-test
- The equal-variance t-test as a special case of ANOVA

Analysis of Variance (ANOVA)

Analysis of Variance is designed to provide a test of a null hypothesis of equal group means with a desired significance level. It is a generalization of the equal-variance t-test to the case where the number of means to be compared is greater than 2.

The intuitive idea behind ANOVA is quite simple. In the case of two population, the equal-variance t-test statistic has the form

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2(1/n_1 + 1/n_2)}}$$

The statistic is large when the *difference between groups* $(\bar{x}_1 - \bar{x}_2)$ is large relative to the variability within groups (represented by the pooled variance s^2).

Similarly, the ANOVA test statistic has a similar form based on a comparison of the *variability* between group means with the variability within groups.

The ANOVA decomposition

ANOVA is based on a decomposition of variability into two sources: variability between groups and variability within groups. Variability **within** groups is represented by the sum of squares of deviations of the observations from their respective group means, i.e.,

$$\mathrm{SS}_W = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

Variability **between** groups is represented by the sum of squares of the group means from the overall mean, weighted by the group sample sizes (n_i), i.e.,

$$\mathrm{SS}_B = \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2$$

The ANOVA decomposition says that the sum of these two sources of variation add up to the total sum of squares of deviations of the observations from the overall mean, i.e.,

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

Illustration of the ANOVA Decomposition

For the battery life data the sums of squares are calculated as follows:

```
x1=c(100,96,92,99,91,95)
x2=c(86,90,95,87,92,98)
x3=c(105,94,96,98,100,92)
n1=length(x1)
n2=length(x2)
n3=length(x3)
xbar1=mean(x1)
xbar2=mean(x2)
xbar3=mean(x3)
xbar=mean(c(x1,x2,x3))
ss.total = sum((x1-xbar)^2) + sum((x2-xbar)^2) + sum((x3-xbar)^2)
ss.within = sum((x1-xbar1)^2) + sum((x2-xbar2)^2) + sum((x3-xbar3)^2)
ss.between = n1*(xbar1-xbar)^2+n2*(xbar2-xbar)^2+n3*(xbar3-xbar)^2
data.frame(ss.between,ss.within,ss.total)
```

```
##   ss.between ss.within ss.total
## 1   118.7778  280.3333 399.1111
```

Note that ss.between+ss.within=ss.total.

The F-Statistic

The ANOVA F-statistic is based on the relative sizes of the between and within sums of squares. First the sums of squares are converted to "mean squares" as follows:

$$\text{MS}_B = \frac{\text{SS}_B}{k-1}$$

where k is the number of groups, and

$$\text{MS}_W = \frac{\text{SS}_W}{N-k}$$

where $N = \sum_{i=1}^{k} n_i$ is the total sample size. Then the F statistic is the ratio of the mean-squares:

$$F = \frac{\text{MS}_B}{\text{MS}_W}$$

A large value of F provides evidence against the null hypothesis.

Calculations using the aov R function

When we perform ANOVA we usually have our data arranged in a data.frame with a group variable indicating the group for each observation. (Note the use of 'factor' here to force R to treat the group variable as a categorical rather than quantitative variable.)

```
d=data.frame(x=c(x1,x2,x3),group=factor(c(rep(1,n1),rep(2,n2),rep(3,n3))))
summary(aov(x ~ group, data=d))
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
group        2  118.8   59.39   3.178 0.0707 .
Residuals   15  280.3   18.69
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The 'aov' function calculates the sums of squares, mean squares, and the F-statistic. It also gives a p-value for the F-statistic. In this case, $p = 0.49$, which indicates that there is not evidence against the null hypothesis of equal group means.

The F-Distribution

The p-value for the F test is calculated by referring the value of F to the $F_{k-1,N-k}$ distribution, which is the F-distribution with $k-1$ numerator degrees of freedom and $N-k$ denominator degrees of freedom.

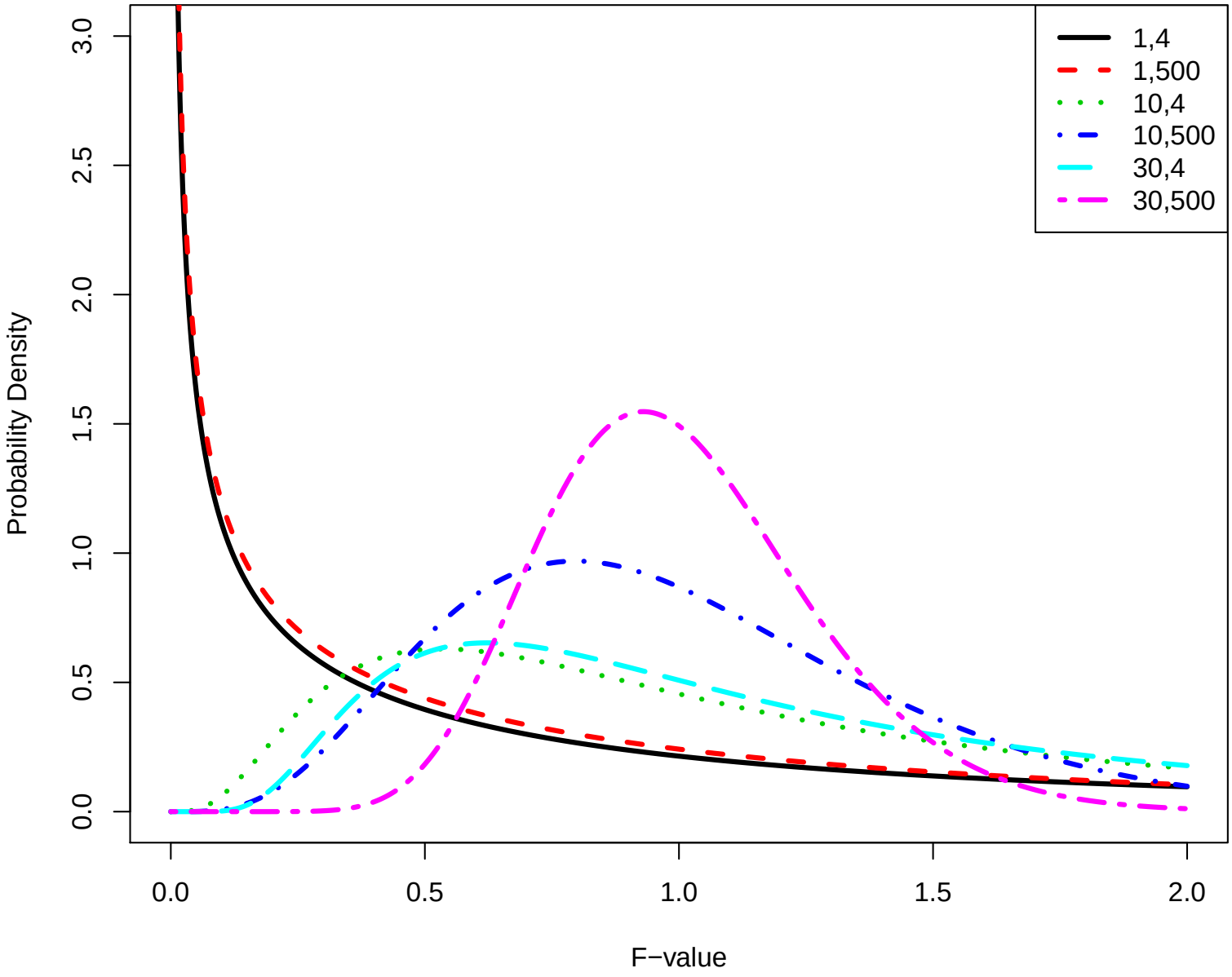
The probability density function of the F-distribution has a very complicated mathematical form but it's properties can be explored using the usual R functions (rf,pf,qf, etc).

The pdf of the F distribution

Examples with various combinations of numerator degrees of freedom and denominator degrees of freedom.

With numerator df $= 1$, the distributions are extremely skewed and very similar for different values of the denominator df. With numerator df $= 10$, the distributions are still right-skewed although not as much as for numerator df=1. If the numerator and denominator df are large, the distribution is almost symmetric (and in fact, approximates a Normal distribution).

Density functions of F-distributions



Application of ANOVA to the Iris Data

```
with(iris,tapply(Sepal.Width,Species,mean))
```

```
    setosa versicolor  virginica
     3.428      2.770      2.974
```

```
with(iris,tapply(Sepal.Width,Species,sd))
```

```
    setosa versicolor  virginica
 0.3790644  0.3137983  0.3224966
```

```
summary(aov(Sepal.Width ~ Species, data=iris))
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
Species       2  11.35   5.672   49.16 <2e-16 ***
Residuals   147  16.96   0.115
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-statistic is highly significant, this is intuitively reasonable because the differences between the group means are quite large relative to the within-group SDs.

MS_W is equal to the pooled sample variance

Note that the MS_W (the "Residuals" MS in the R output) is equal to the pooled sample variance, which is just the average of the sample variances for the 3 groups here because the group sample sizes are equal.

```
v=with(iris,tapply(Sepal.Width,Species,var))
v
```

```
    setosa versicolor  virginica
0.14368980 0.09846939 0.10400408
```

```
mean(v)
```

```
[1] 0.1153878
```

Demonstration that MS_W is the Pooled Sample Variance

The pooled sample variance from the t-test has the following form:

$$\begin{aligned} s^2 &= \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \\ &= \frac{\sum_{i=1}^{n_1}(x_{1j}-\bar{x}_1)^2 + \sum_{i=1}^{n_2}(x_{2j}-\bar{x}_2)^2}{n_1+n_2-2} \\ &= \frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2}{\sum_{i=1}^{2} n_i - 2} \end{aligned}$$

Here we have used the following formula for sample variance:
$s_1^2 = \sum_{j=1}^{n_1}(x_{1j}-\bar{x}_1)^2/(n_1-1)$, and similarly for s_2^2.

This means that for two groups, the pooled sample variance is equal to the mean-squared within, $s^2 = \mathrm{MS}_W$. In other words the MS_W is an estimate of the population variance. (Recall that the variance is assumed to be the same for all populations being compared.)

The term "mean-square error" and notation MS_E is also used for MS_W or mean-square for "Residuals" in the output from the 'aov' function in R.

Relationship between the F-test and the T-test

When there are just 2 groups, the F-test is equivalent to the equal-variance t-test. So if you use ANOVA with 2 groups you will get exactly the same conclusions as you would from the t-test.

```
summary(aov(x ~ group, data=subset(d,!(group==3))))
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
group        1  52.08   52.08   3.014  0.113
Residuals   10 172.83   17.28
```

```
t.test(d$x[d$group==1],d$x[d$group==2],var.equal=T)
```

```
    Two Sample t-test

data:  d$x[d$group == 1] and d$x[d$group == 2]
t = 1.7359, df = 10, p-value = 0.1132
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.181382  9.514716
sample estimates:
mean of x mean of y
 95.50000  91.33333
```

Note that the p-values from the two procedures are the same. Also, the F-statistic (3.014) is equal to the square of the t-statistic (at least up to round-off error). This property holds for the theoretical distributions as well: the square of random variable with a t_d distribution has a $F_{1,d}$ distribution.

Demonstration that the Equal-Variance T-test is Equivalent to the F-test

We showed previously that

$$s^2 = \frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2}{\sum_{i=1}^{2} n_i - 2} = \frac{\mathrm{SS}_W}{N-k} = \mathrm{MS}_W,$$

because $k = 2$. Also, because $k = 2$, so $k - 1 = 1$, and

$$\mathrm{MS}_B = \sum_{i=1}^{2} n_i(\bar{x}_i - \bar{x})^2 = n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 = (\bar{x}_1 - \bar{x}_2)^2 n_1 n_2/(n_1 + n_2)$$

Therefore,

$$F = \frac{\mathrm{MS}_B}{\mathrm{MS}_W} = \frac{(\bar{x}_1 - \bar{x}_2)^2}{s^2(1/n_1 + 1/n_2)} = Z^2,$$

where Z is the equal-variance t-statistic $= (\bar{x}_1 - \bar{x}_2)/\sqrt{s^2(1/n_1 + 1/n_2)}$.

3. Checking Assumptions of ANOVA - Residual Diagnostics

- Equality of variances
- Normality
- Independence

The 3 key assumptions of ANOVA

1. Independence (of samples and of observations within each sample).
2. Equal variances
3. Large sample sizes *or* normal distributions.

This is roughly in order of importance. Non-independence (either of the samples or of observations within samples) can have a huge effect on the performance of the F-test. The effect can be to make the test either overly conservative (type I error probability too small) or anti-conservative (type I error probability too large).

Similarly, non-equal variances can case the F-test to be either conservative or anti-conservative.

> *A key point about the first two assumptions is that departures from independence or equal variances are a problem* ***no matter how large the sample size****.*

In contrast with the assumptions of independence and equal variances, normality of the populations becomes unnecessary if the sample sizes are large.

The ANOVA Model

For thinking about how to check model assumptions, it is helpful to have a mathematical equation to represent the underlying statitsical model.

For ANOVA, the statistical model is quite simple: it simply says that each group has a unique population mean value for the outcome variable. We can define this as follows:

$$x_{ij} = \mu_i + \epsilon_{ij}$$

where μ_i is the population mean for the ith group and the random variable ϵ_{ij} represents the "error" for the given observation, i.e., how far the observation is away from its predicted value μ_i. Note that we only get to observe x_{ij}; we do not know the values of μ_i and ϵ_{ij}.

For above equation for the ANOVA model is useful for thinking about checking the assumptions of equality of variances and normality. Similar model equations will become even more important in the context of regression analysis.

The ANOVA Assumptions Revisited

Recall the 3 assumptions: 1) independence, 2) equal variances, and 3) normality or large samples. In terms of the ANOVA model above these assumptions can be expressed as follows:

1. The errors ϵ_{ij} are independent within and across groups.
2. The errors ϵ_{ij} have equal variances in all groups.
3. Either the sample sizes are large or the errors have normal distributions.

Residuals

Methods for checking both equal variances and normality rely on the concept of "residuals". A general definition of a residual (which applies to linear regression and other methods as well as ANOVA) is the following.

> *A **residual** is the difference between an observation and its predicted value based on a statistical model.*

For ANOVA the underlying model is $x_{ij} = \mu_i + \epsilon_{ij}$ as described above. The predicted value of x_{ij} based on this model is our estimate of the population mean μ_i for the ith group, namely, the sample mean for this group. We can write this as

$$\text{Predicted value of } x_{ij} = \hat{\mu}_i = \bar{x}_i = \sum_{i=1}^{n_i} x_{ij}/n_i.$$

Therefore, the residual for this observation is

$$e_{ij} = x_{ij} - \bar{x}_i.$$

(Sometimes the notation r_{ij} is used instead of e_{ij}.)

Relationship between Errors and Residuals

Note that we have the following equation:

$$x_{ij} = \hat{\mu}_i + e_{ij},$$

which corresponds to the model equation $x_{ij} = \mu + \epsilon_{ij}$.

Therefore, the residuals e_{ij} can be thought of as a form of estimates of the errors ϵ_{ij}.

What this means is that we should use the residuals to check assumptions about the errors.

Calculation of Residuals

To calculate the residuals, we can use the tapply function or extract them from the aov output.

```
options(width=64)
d=data.frame(x=c(x1,x2,x3),group=factor(c(rep(1,n1),rep(2,n2),rep(3,n3))))
# The following creates a vector containing the respective group mean
# for each observation.
m=tapply(d$x,d$group,mean)[tapply(d$x,d$group)]
e=d$x-m
summary(e)
```

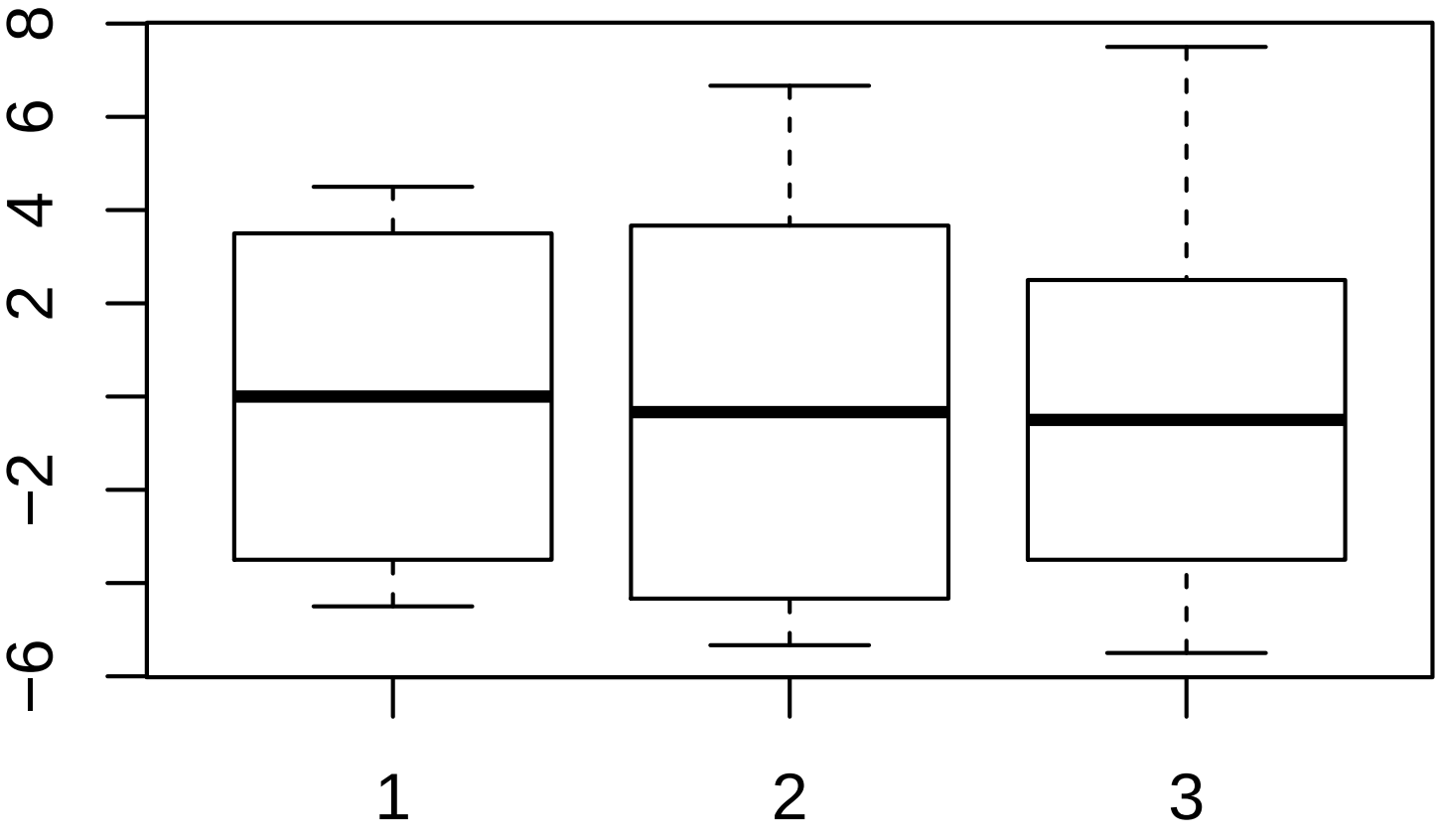
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  -5.50   -3.50    0.00    0.00    3.25    7.50
```

```
out=aov(x ~ group, data=d)
summary(out$residuals)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  -5.50   -3.50    0.00    0.00    3.25    7.50
```

Using Residuals to Check Equality of Variances

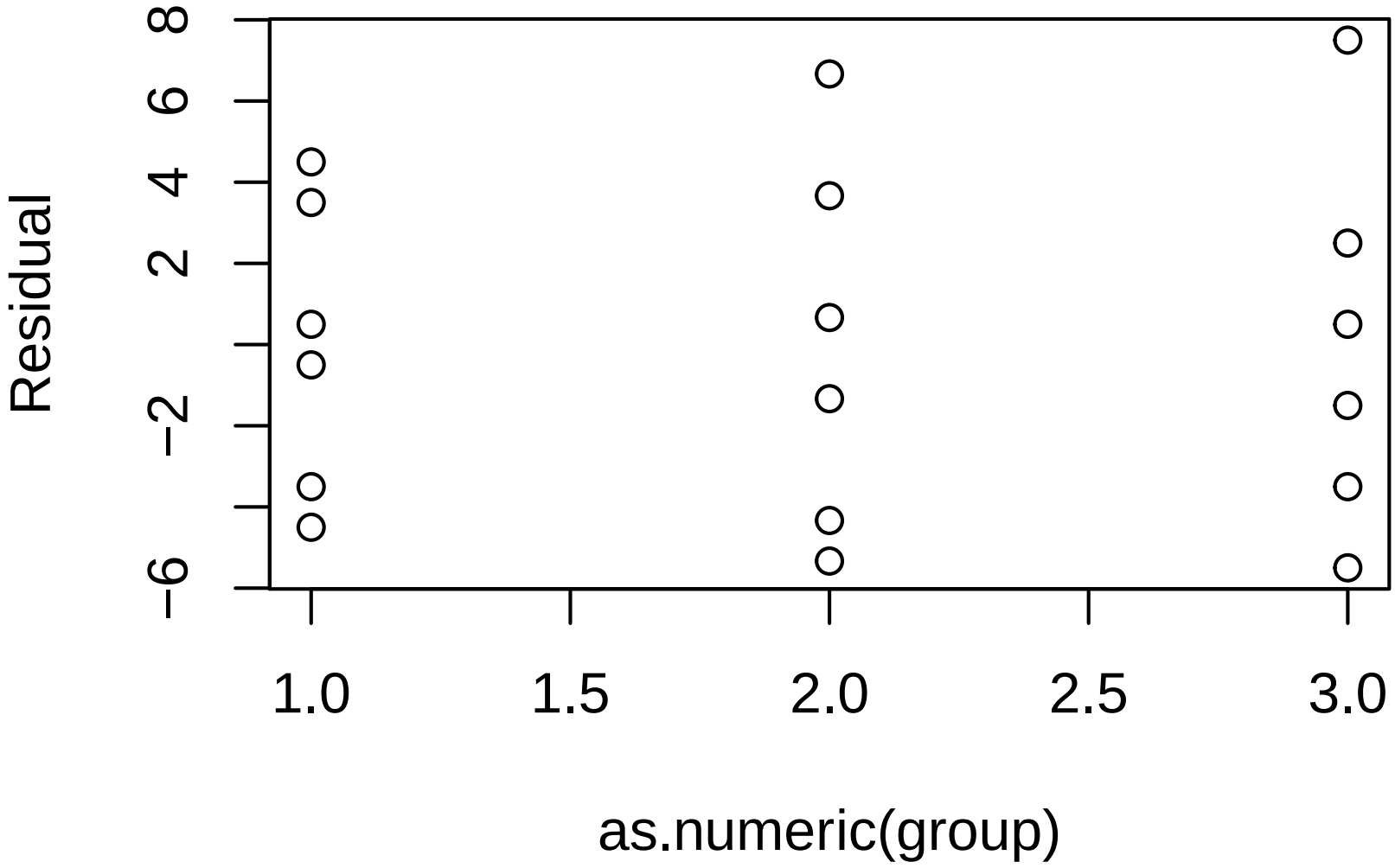
The equality of variances can be assessed graphically using a boxplot or dotplot. The boxplot shows that the variances of the residuals for the three groups are similar. This plot would be interpreted as indicating no evidence against equality of variances.



Dotplot of Residuals

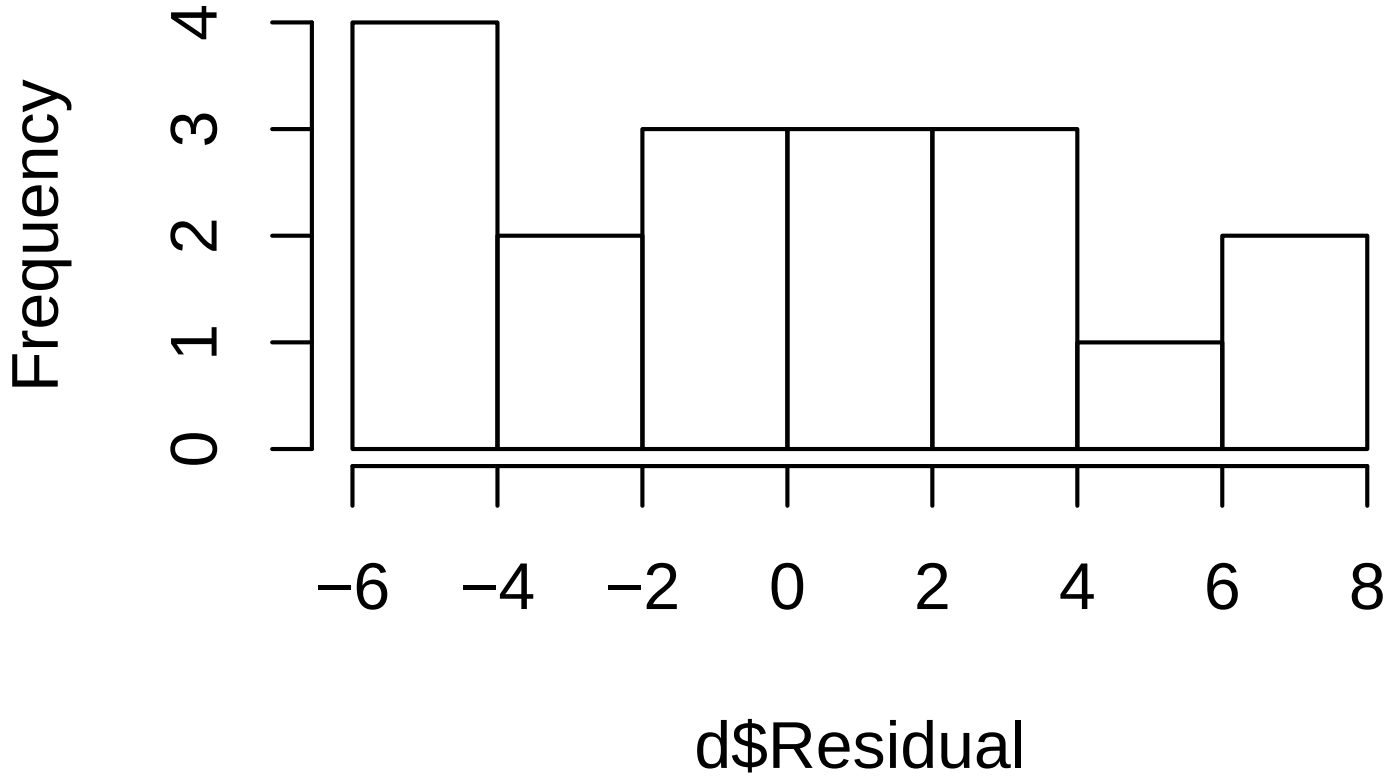
The dotplot similarly shows no evidence against equality of variances.

```
plot(Residual ~ as.numeric(group), data=d)
```



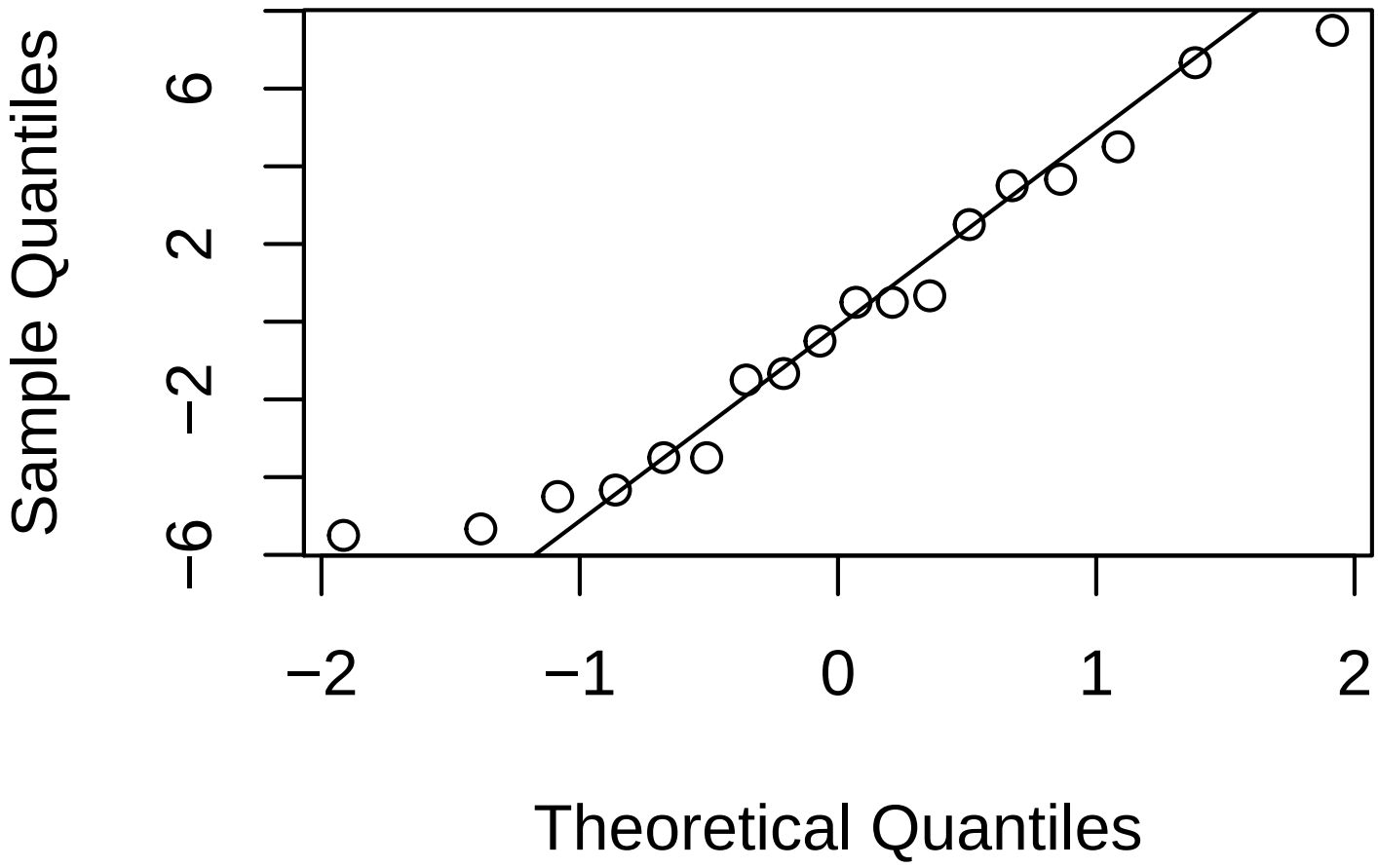
Using Residuals to Check Normality

We can also use the residuals to check normality by applying the usual graphical methods (histograms and q-q plots) to the residuals.



The histogram looks quite non-normal (more like a uniform distribution). However, the sample size is quite small, so it is difficult to say anything conclusive. This example illustrates how hard it is to assess normality from small data sets.

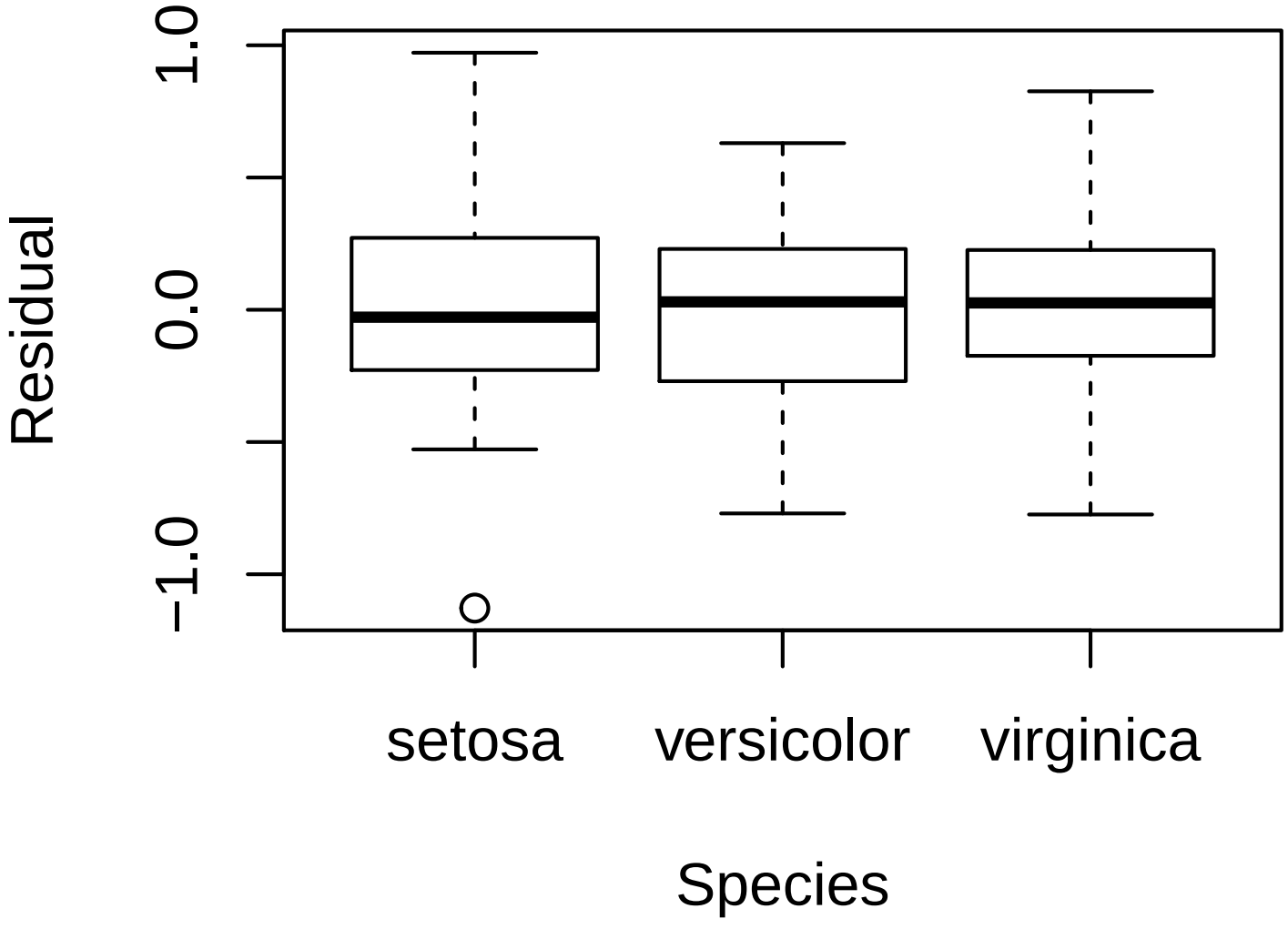
The Q-Q Plot of the Residuals



The q-q plot does not clearly indicate non-normality. The unusual shape of the distribution does show up in the 3 points above the line in the left tail and the one point below the line in the right tail, but the small number of points makes a clear decision impossible.

Checking the Residuals for the Iris Data

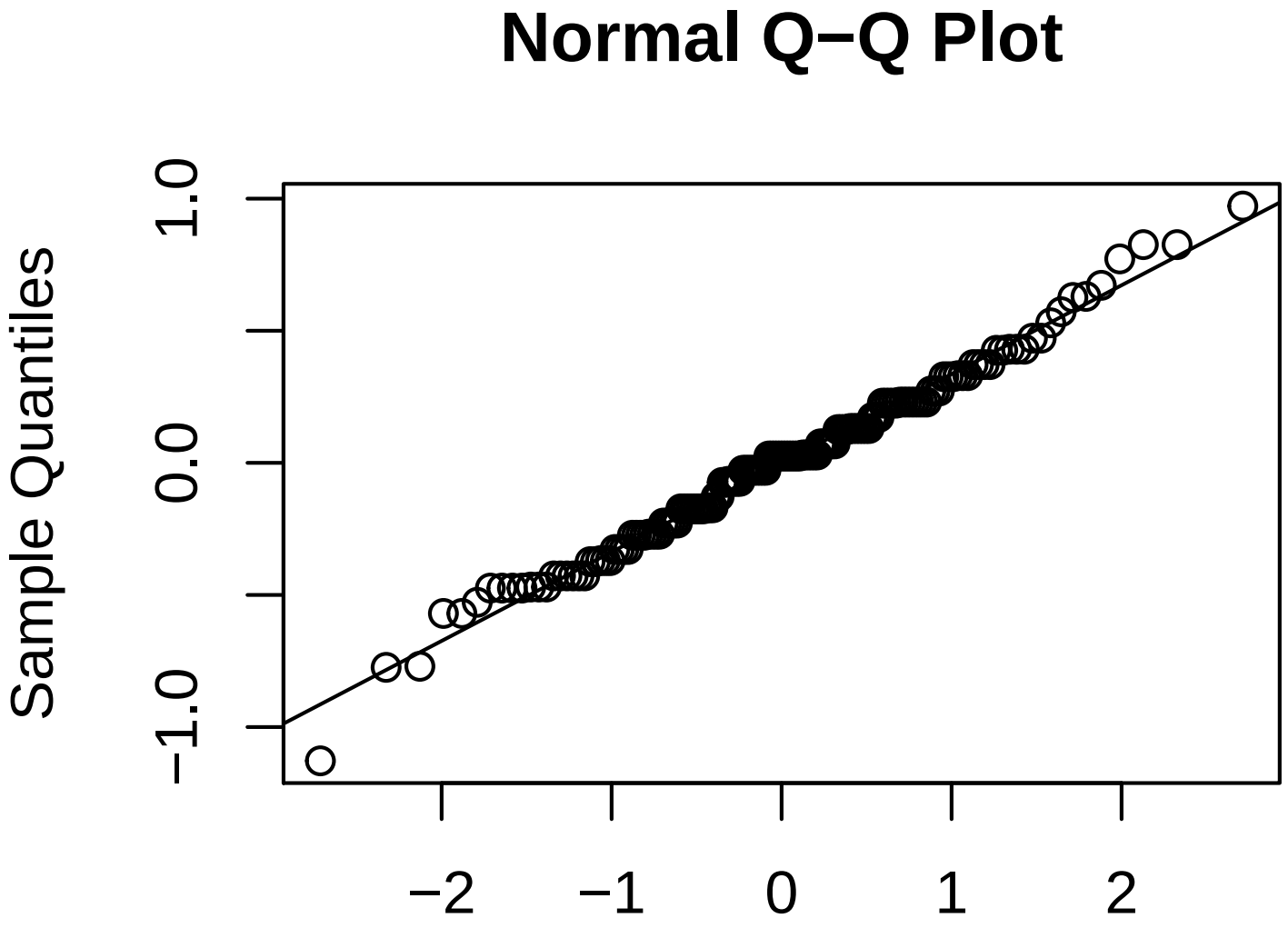
The plot indicates no departure from equality of variances.



Checking Normality for the Iris Data

No evidence of non-normality in the residuals.

```
qqnorm(b$Residual)
qqline(b$Residual)
```



Verifying the Independence Assumptions

Independence is an assumption that is typically verified through our knowledge of the study design and data collection mechanisms. We typically do not use the data to check on the assumption of independence.

For example, in a "completely-randomized" experimental design, we randomly assign the experimental units to the different conditions completely randomly. This design is often used in randomized clinical trials, in which each patient is randomly assigned to one of the treatments under study using a random number generator in such a way that their treatment assignment is not influenced by the assignments of any other patients in the trial. For a completely randomized design, the method of randomization justifies the independence assumptions.

Similarly, in an experiment comparing quality of manufactured items using three different processes, we would take random samples of items from each process in such a way that the sampling was done independently for each process.

4. Other Experimental Designs

- Factorial Designs
- Randomized Block Designs
- Cluster-Randomized Designs

Factorial Designs

A factorial design is used to test the effects of 2 or more treatment variables.

Example: The NPK Experiment

The R dataset 'npk' gives results of a factorial experiment to assess the effects of nitrogen (N), phosphorous (P) and potassium (K), on crop yield of peas. A randomized block design was used in which there were 6 blocks each consisting of 4 plots of land. Within each block the plots were randomized to receive various combinations of N, P and K. Each was either applied or not applied on a given plot. The yield of peas (pounds per plot) was recorded as the response variable.

This is a *factorial* design in which different combinations of treatments are used.

Full factorial design

There were 24 plots of land available and there are 8 different combinations of N, P, and K. Thus, the researchers assigned 3 plots to each of the 8 combinations. (We will assume for now that a completely-randomized design was used, i.e., the 8 conditions were assigned to the 24 plots completely randomly.)

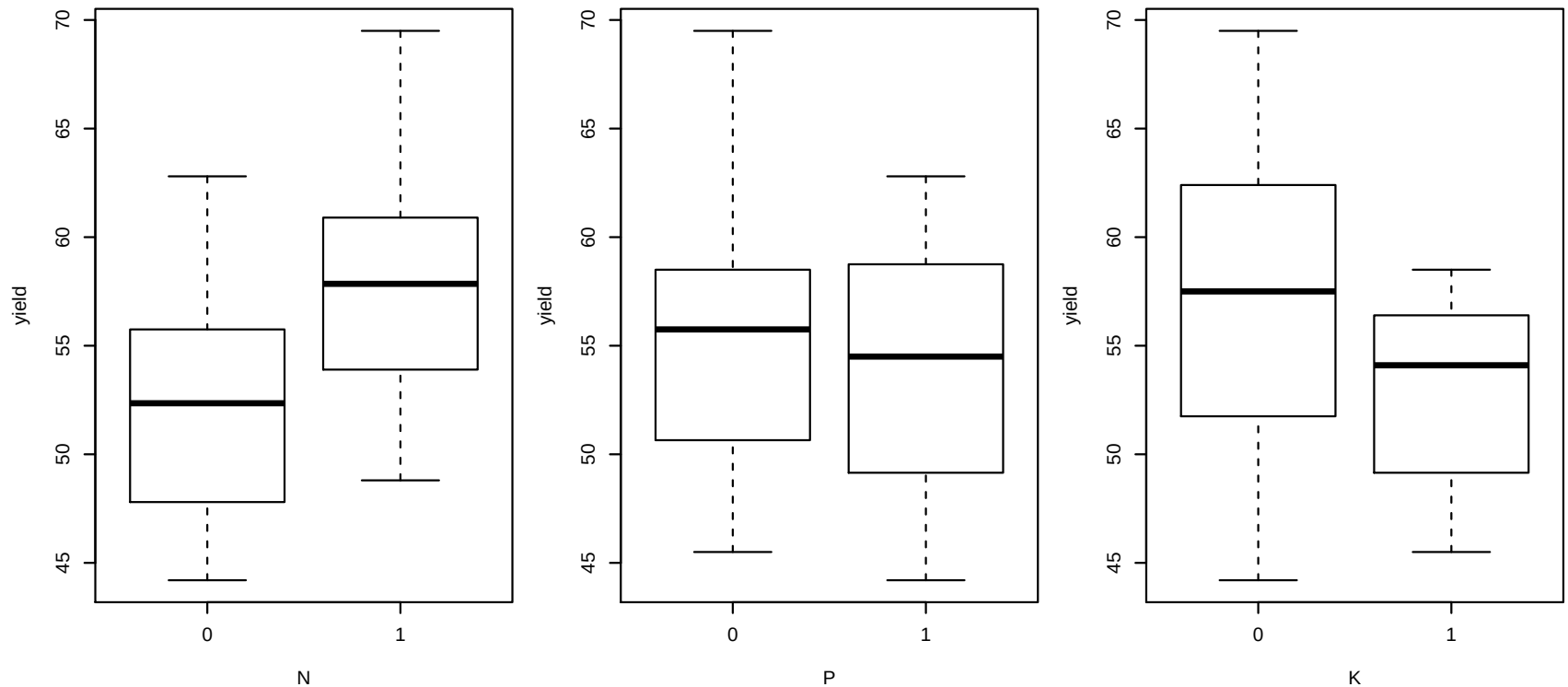
When all combinations of factors are used we call it a "full factorial" design.

```
npk$combination =paste(npk$N,npk$P,npk$K,sep="")
table(npk$combination)
```

```
000 001 010 011 100 101 110 111
  3   3   3   3   3   3   3   3
```

Plots of yield versus N, P, and K separately

```
par(mfrow=c(1,3),mar=c(5,4,4,1))
plot(yield ~ N, data=npk)
plot(yield ~ P, data=npk)
plot(yield ~ K, data=npk)
```



Note that for each factor (N, P, and K), we have a comparison of 12 plots for each treatment condition. Thus, we have 3 experiments with sample sizes of 12 per group contained within one experiment with a total of 24 observations.

ANOVA for the NPK experiment

```
summary(aov(yield ~ N+P+K,data=npk))
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
N            1  189.3  189.28   6.488 0.0192 *
P            1    8.4    8.40   0.288 0.5974
K            1   95.2   95.20   3.263 0.0859 .
Residuals   20  583.5   29.17
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From these results we would conclude that the application of nitrogen affects the yield (p=0.02), whereas there is no evidence that phosphorous or potassium have effects on yield (p=0.60 and p=0.09, respectively).

Further analyses of these data would consider **interactions** between the 3 treatments (we will come back to this later).

Randomized Block Designs

Blocking induces dependence between observations.

In an experimental situation, sometimes we use *blocking* (also called stratifying) to control variability and increase the precision of our experiment. For example, in multi-center clinical trials, each "center" (e.g., a clinic or hospital) is treated as a homogeneous block and we would randomize patients to treatment groups using a separate random number list within each block. This means that the observations are no longer independent, e.g., outcomes for patients in the same center would be expected to be correlated (not independent).

In a quality testing experiment, we might use 10 different batches of raw material and for each batch make one run of each process using that same batch of material. Thus, the batches form the blocks and the outcomes for two units produced using the same batch might be expected to be correlated (not independent).

Clustered Designs

Clustering is another feature of experiments that induces dependence between observations. However, it induces a different kind of dependence that must be accommodated using different methods than for blocking.

In a medical research setting, it might be impossible to randomly assigne patients to different treatments within each hospital. If the intervention is at the hospital-level (e.g., new strategies for treating emergency trauma victims), then all patients in the same hospital must get the same treatment.

In this situation, we **randomize the hospitals** to different treatments. However, the outcome is measured at the individual patient level. Therefore, we have *clusters* of patients (all the patients) at the same hospital. Patients at the same hospital will likely have more similar outcomes than patients at different hospitals.

Contrast between blocked and clustered designs

In a blocked design experimental units are assigned to different treatments within each block. Thus, different units within the same *block* can have different treatments.

In a clustered design, entire clusters of experimental units are assigned to a treatment. Therefore, all units in the same *cluster* have the same treatment.

We will come back to the question of how to analyze data from designs that use blocking. For now, we will assume independence assumptions are satisfied and consider how to assess the other two assumptions of ANOVA.

Analysis of Randomized Block Designs

Special case: The Paired T-test

The simplest type of blocking is pairing, i.e., when experimental units are divided into blocks of size 2 and within each pair one unit is randomly assigned to receive treatment A while the other unit in the pair gets treatment B.

Examples of this include measures of a change in weight after patients take a drug, differences between pollutant levels in fresh versus aged water samples. A non-experimental (observational) example is comparison of birth weights of first and second born children in a family.

In these situations there are two sets of measurements but they are linked ("paired") in some way. In this case, the inference is *based on the differences between paired measures* using the paired t-test.

The paired t-test

Suppose that we have paired measurements of weight on 5 patients before and after a drug is given.

```
x1=c(130,150,160,125,170)
x2=c(140,150,170,130,165)
t.test(x1,x2,paired=T)
```

```
    Paired t-test

data:  x1 and x2
t = -1.372, df = 4, p-value = 0.242
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -12.094659   4.094659
sample estimates:
mean of the differences
                     -4
```

The paired t-test as a 1-sample test

The paired t-test is just the 1-sample t-test applied to the differences between the paired measurements.

```
d=x1-x2
n=length(d)
m=mean(d)
s=sd(d)
se=s/sqrt(n)
z=m/se
p=2*(1-pt(abs(z),df=n-1))
t.05=qt(0.975,df=n-1)
data.frame(mean.diff=m,sd.diff=s,df=n-1,z,p,lower=m-t.05*se,upper=m+t.05*se)
```

```
  mean.diff  sd.diff df         z         p     lower    upper
1        -4 6.519202  4 -1.371989 0.2419815 -12.09466 4.094659
```

The NPK experiment as a randomized block design

The 24 plots of land were arranged in 6 blocks with 4 plots per block.

```
table(npk$block,npk$combination)
```

```
  000 001 010 011 100 101 110 111
1   1   0   0   1   0   1   1   0
2   0   1   1   0   1   0   0   1
3   0   1   1   0   1   0   0   1
4   0   1   1   0   1   0   0   1
5   1   0   0   1   0   1   1   0
6   1   0   0   1   0   1   1   0
```

Note: *within each plot* this is called a *partial factorial* design.

Therefore, the previous analysis of the data was invalide because the observations are not independent: yields on plots within the same block are likely to be more similar to each other than yields on plots in different blocks. Thus, the yields are correlated (which means that the errors in the regression model are correlated). Let's account for the blocking.

Correct analysis of the NPK experiment (accounting for blocks)

```
              Df Sum Sq Mean Sq F value  Pr(>F)
factor(block)  5  343.3   68.66   4.288 0.01272 *
N              1  189.3  189.28  11.821 0.00366 **
P              1    8.4    8.40   0.525 0.47999
K              1   95.2   95.20   5.946 0.02767 *
Residuals     15  240.2   16.01
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

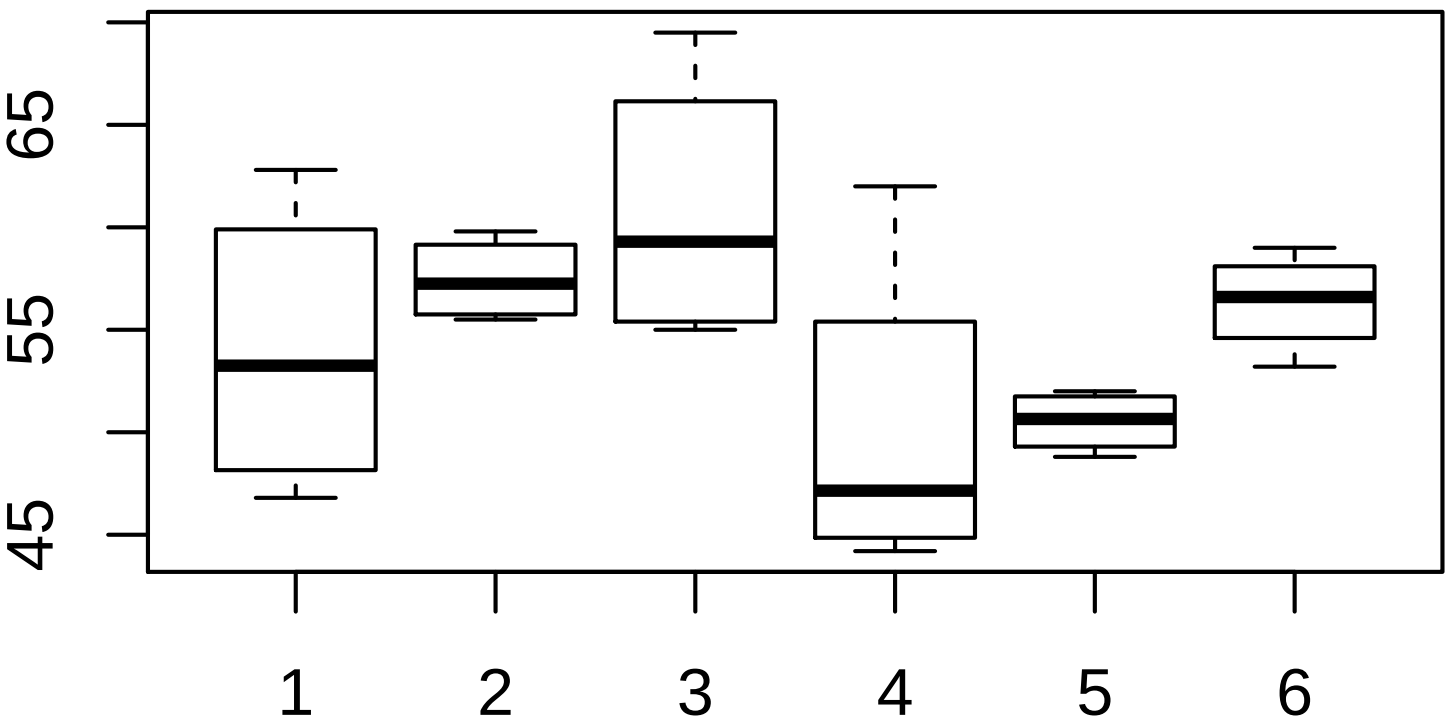
By accounting for the block effects, the assumption of independent errors is reasonable because the differences between blocks have been removed.

The p-values are now 0.004, 0.48 and 0.028. Thus, we would now conclude that there is evidence for an effect of potassium on crop yield, as well as nitrogen.

Examining the block effects

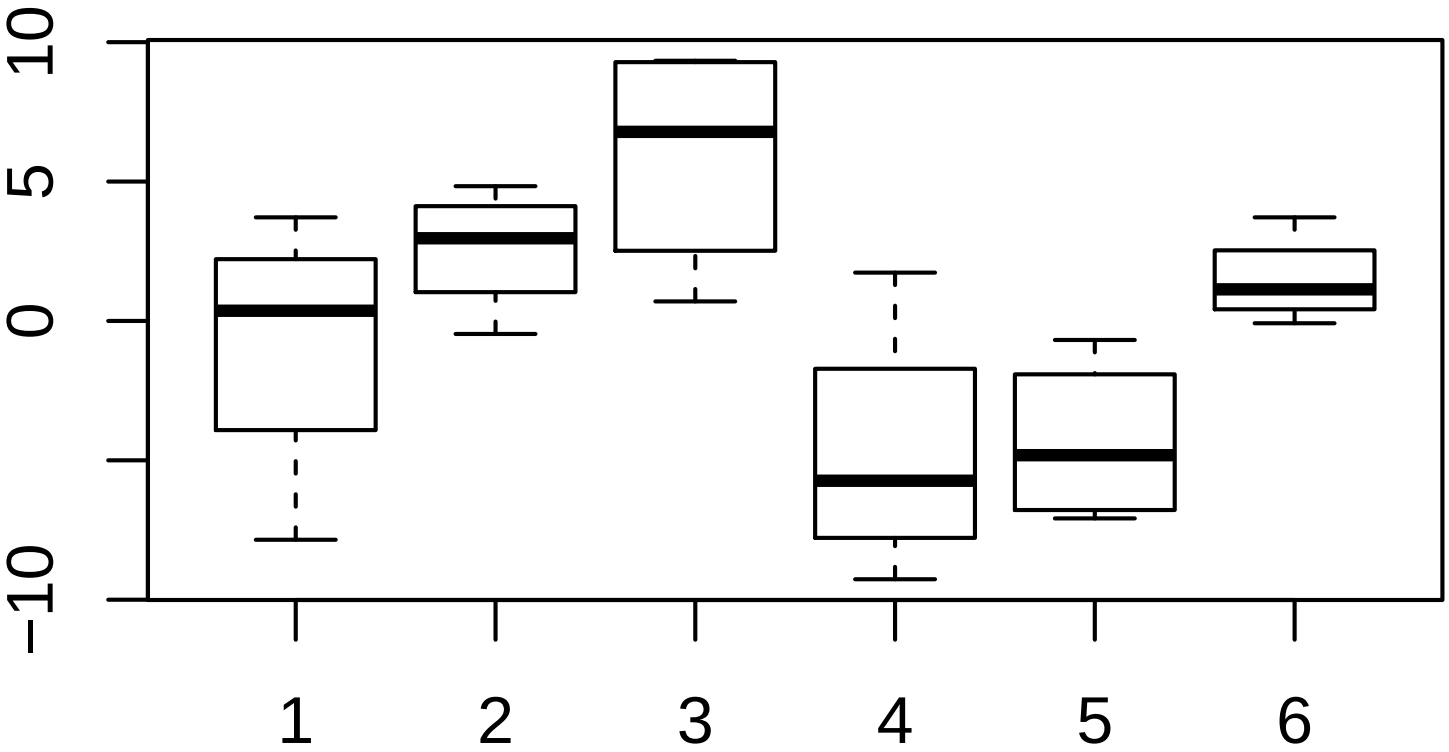
We can illustrate the block effects using boxplots of the yield by block, which suggest very large differences between blocks. However, these differences might be confounded by which combinations of N, P and K were used in the different blocks. Thus, we need to remove the N, P and K effects to properly examine the block effects.

```
boxplot(yield ~ block,data=npk)
```



Examining the block effects using residuals

Remove the effect of N, P, and K to assess block effects. The conclusion is not changed very much - there are large differences between blocks.



Efficiency gain due to blocking

We can quantify the effect of blocking by comparing the precision of the treatment effect estimates from the two analyses, one ignoring blocks and one accounting for blocks. In this example, the variances of the treatment effect estimates are 82% higher when blocking is ignored compared to when blocking is accounted for.[2]

One way to interpret this is by considering the reduction in sample size needed for the experiment if blocking is used, compared to an experiment that did not use blocking. Recall that variances of treatment effect estimates are inversely proportional to sample size ($\text{var}(\bar{X}) = \sigma^2/n$). A ratio of variances equal to 1.82 implies a ratio of required sample sizes of $1/1.82 = 0.55$. Thus, we could say that the blocked design requires 45% fewer observations than the unblocked design.

[2] Details of these calculations will be discussed later.