Chapter4ANOVA

January 23, 2023

1 Chapter 4 Analysis of Variance

1.1 4.1 One-way ANOVA

1.1.1 4.1.1 A motivating example: the Spock trial

In 1968 Dr. Benjamin Spock was tried in Boston for conspiring against the government for helping young men to escape the military draft. He was convicted by the Boston federal court, but the judgement was overturned by the Court of Appeals in 1969 for many reasons, one of which was cited as the bias of the presiding judge Francis Ford.

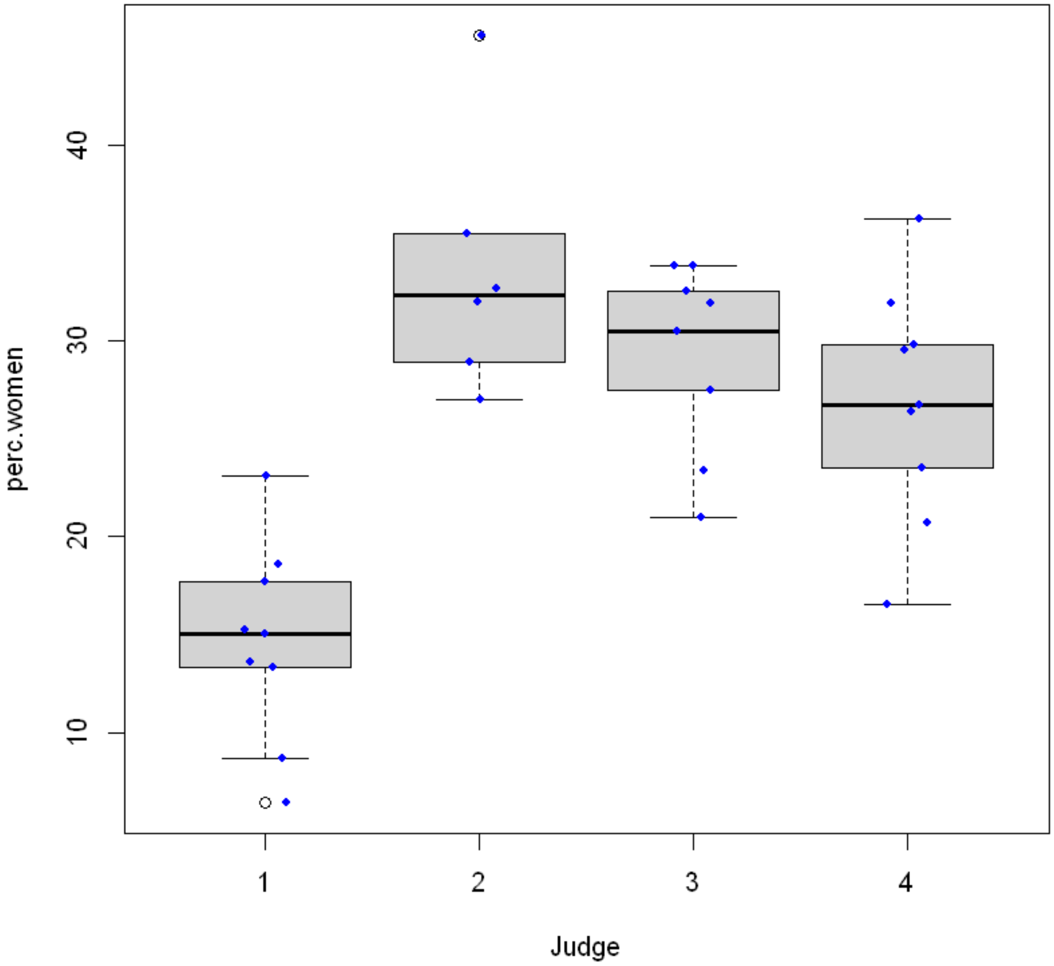
Dr. Spock, a pediatrician, was very famous for his books on rearing of children, and thus was widely admired by women. As a matter of fact, the jury in Spock trial has no women. Note that jury panels, though randomly selected, should reflect the demographics.

In any particular trial, there may not be any woman on the jury, but it is worthwhile to examine if the jury panels of Judge Ford had fewer women than other judges in Boston in few months before the trial. Data are available for jury panels for 7, but we investigate the data for only 4 judges including Judge Ford.

We can start our analysis with a visualization of the data set.

[2]:

```r
Spock <- read.csv(file="../Data/SpockTrial.csv", header=TRUE, sep=",")
Spock$Judge<-as.factor(Spock$Judge);
# Box plot with jittered points
boxplot(perc.women~Judge,data=Spock)
stripchart(perc.women~Judge, vertical = TRUE, data = Spock,
    method = "jitter", add = TRUE, pch = 20, col = 'blue')
```



1.1.2 4.1.2 One-way ANOVA

In the Spock trial data, let $r = 4$ denote the number of judges, Y_{ij} be the percentage of women in the jth panel for the ith judge. Let Judge 1 be the judge in the Spock trial. We can propose the following model, for $j = 1, \ldots, n_i, i = 1, \ldots, r$,

$$Y_{ij} = \mu_i + \epsilon_{ij},$$

where $\{\epsilon_{ij}\}$ are i.i.d. $N(0, \sigma^2)$. In the Spock trial data, $r = 4$, $n_2 = 6$, and $n_1 = n_3 = n_4 = 9$. This model is a one-way **an**alysis **o**f **va**riance model in its cell means form. We will discuss other forms later in this chapter.

Because n_1, n_2, n_3 and n_4 are not equal, this is an imbalanced ANOVA model. If $n_1 = n_2 = \cdots = n_r$, then the ANOVA model is **balanced**.

In this model, μ_i represents the mean percentage of women in the panels for the ith judge, and σ^2 represents the variance in the percentages across panels. It is easy to see that, by using one σ^2 across all judges, we assume the panels of all four judges have the same amount of variability.

The question of interest in the Spock trial data can now be translated to whether $\{\mu_i\}_{i=1}^{r}$ are the same, where $\{\mu_i\}$ and σ^2 are unknown.

The estimators for $\mu_i, i = 1, 2, \ldots, r$ are simply the within-group sample means, i.e., for $i = 1, \ldots, r$,

$$\hat{\mu}_i = \bar{Y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}.$$

We have two observations on our estimators $\hat{\mu}_i$, $i = 1, \ldots, r$. 1. The estimator $\hat{\mu}_i$ is also the maximum likelihood estimator for μ_i. 2. $\hat{\mu}_i$ is the best linear unbiased estimator if $\{Y_{ij}\}$ are mutually uncorrelated but not necessarily normally distributed.

We can estimate the variance of errors as

$$\hat{\sigma}^2 \equiv \frac{1}{n_T - r}\sum_{i=1}^{r}\sum_{j=1}^{n_i}\left(Y_{ij} - \hat{\mu}_i\right)^2,$$

where $n_T \equiv \sum_{i=1}^{r} n_i$ is the (total) sample size.

We can call the `aov()` function to fit a one-way ANOVA model in `R`.

[3]:

```
anova.fit<- aov(perc.women~Judge,data=Spock)

# Summary
summary(anova.fit)
```

```
Call:
   aov(formula = perc.women ~ Judge, data = Spock)

Terms:
                    Judge Residuals
Sum of Squares  1591.2779  873.5239
Deg. of Freedom         3        29

Residual standard error: 5.488307
Estimated effects may be unbalanced

            Df Sum Sq Mean Sq F value   Pr(>F)    
Judge        3 1591.3   530.4   17.61 1.06e-06 ***
Residuals   29  873.5    30.1                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1.1.3 4.1.3 Alternative forms of ANOVA

Factor effect form

Let $\mu = \sum_{i=1}^{r} w_i \mu_i$ and $\tau_i = \mu_i - \mu$. Then $\sum_{i=1}^{r} w_i \tau_i = 0$. We can rewrite the ANOVA model as

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij},$$

where $\{\epsilon_{ij}\}$ are i.i.d. $N(0, \sigma^2)$.

The least squares estimators are $\hat{\mu} = \sum_{i=1}^{r} w_i \bar{Y}_{i\cdot}$ and $\hat{\tau}_i = \bar{Y}_{i\cdot} - \hat{\mu}$.

Notes in `R`: 1. One of τ_i is set to zero by default. 2. `R` assumes $w_i = n_i/n_T$ for one-way ANOVA, but equal weights for higher-order ANOVAs.

Regression form

There are multiple equivalent forms to turn a cell-mean model into the typical linear regression form. In the Spock trial data, the regression equation takes the following form, for $j = 1, \ldots, n_i, \quad i = 1, \ldots, 4$,

$$Y_{ij} = \mu + \tau_1 X_{1,ij} + \tau_2 X_{2,ij} + \tau_3 X_{3,ij} + \epsilon_{ij}.$$

where $\{\epsilon_{ij}\}$ are i.i.d. $N(0, \sigma^2)$. Here we set τ_4 to be zero, and thus we ignore the the fourth term $ X_{4,ij}$.

There are multiple choices in the coding of $\{X_{1,ij}, \ldots, X_{4,ij}\}$. 1. *Dummy variables.* $X_{l,ij} = 1$ when $l = i$, and 0 otherwise. 2. *Unweighted effect coding.* For $l = 1, 2, 3$, $X_{l,ij} = 1$ when $i = l$, $X_{l,ij} = -1$ when $i = 4$, and $X_{l,ij} = 0$ otherwise. This is equivalent to an ANOVA model with equal weights. 3. *Weighted effect coding.* For $l = 1, 2, 3$, $X_{l,ij} = 1$ when $i = l$, $X_{l,ij} = -n_l/n_4$ when $i = 4$, and $X_{l,ij} = 0$ otherwise. This is equivalent to an ANOVA model with unequal weights.

It is easy to see that all three regression models are equivalent, since they are all equivalent forms of the same ANOVA model. However, the interpretations of the model parameters (i.e., τs) may differ slightly.

[5]:

```r
lm.fit<- lm(perc.women~as.factor(Judge),data=Spock)
summary(lm.fit)
```

```
Call:
lm(formula = perc.women ~ as.factor(Judge), data = Spock)

Residuals:
    Min      1Q  Median      3Q     Max 
-10.300  -3.300  -0.100   3.078  11.983 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)         14.622      1.829   7.993 8.16e-09 ***
as.factor(Judge)2   18.994      2.893   6.567 3.41e-07 ***
as.factor(Judge)3   14.478      2.587   5.596 4.85e-06 ***
as.factor(Judge)4   12.178      2.587   4.707 5.73e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.488 on 29 degrees of freedom
```

```
Multiple R-squared:  0.6456,	Adjusted R-squared:  0.6089 
F-statistic: 17.61 on 3 and 29 DF,  p-value: 1.057e-06
```

1.1.4 4.1.4 Sum of squares decomposition

Define the residuals $e_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \bar{Y}_{i\cdot}$, $j = 1, 2, \ldots, n_i$ and $i = 1, 2, \ldots, r$. From the model equation, we can see that e_{ij} is an estimator of ϵ_{ij}.

A simple identity holds that

$$Y_{ij} - \bar{Y}_{\cdot\cdot} = (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}) + (Y_{ij} - \bar{Y}_{i\cdot}).$$

Squaring both sides of the equation yields

$$\begin{aligned} \sum_{i=1}^{r}\sum_{j=1}^{n_i}\big(Y_{ij} - \bar{Y}_{\cdot\cdot}\big)^2 &= \sum_{i=1}^{r}\sum_{j=1}^{n_i}\big(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}\big)^2 + \sum_{i=1}^{r}\sum_{j=1}^{n_i}\big(Y_{ij} - \bar{Y}_{i\cdot}\big)^2 && (1) \\ &= \sum_{i=1}^{r} n_i\big(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}\big)^2 + \sum_{i=1}^{r}\sum_{j=1}^{n_i}\big(Y_{ij} - \bar{Y}_{i\cdot}\big)^2, && (2) \end{aligned}$$

where the first equality holds because $2\sum_{i=1}^{r}\sum_{j=1}^{n_i}\big(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}\big)\big(Y_{ij} - \bar{Y}_{i\cdot}\big) = 0$.

Some terminologies:

Acronym

Definition

d.f.

Total sum of squares

SSTO

$\sum_{i=1}^{r}\sum_{j=1}^{n_i}\big(Y_{ij} - \bar{Y}_{\cdot\cdot}\big)^2$

$n_T - 1$

Treatment sum of squares

SSTR

$\sum_{i=1}^{r} n_i\big(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}\big)^2$

$r - 1$

Residual sum of squares

SSE

$\sum_{i=1}^{r}\sum_{j=1}^{n_i}\big(Y_{ij} - \bar{Y}_{i\cdot}\big)^2$

$n_T - r$

We can see that

$$\text{SSTO} = \text{SSTR} + \text{SSE}, \quad \text{df(SSTO)} = \text{df(SSTR)} + \text{df(SSE)}.$$

We can define the mean squares accordingly,

$$
\begin{aligned}
\text{MSTR} &= \text{SSTR}/\text{df(SSTR)} = & \text{SSTR}/(r-1) \qquad (3)\\
\text{MSE} &= \text{SSE}/\text{df(SSE)} = & \text{SSE}/(n_T-r) \qquad (4)\\
\text{MSTO} &= \text{SSTO}/\text{df(SSTO)} = & \text{SSTO}/(n_T-1) \qquad (5)
\end{aligned}
$$

[6]:

```r
# Summary
summary(anova.fit)
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)    
Judge        3 1591.3   530.4   17.61 1.06e-06 ***
Residuals   29  873.5    30.1                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1.1.5 4.1.5 Gauss-Markov properties

We assume that $Y = X\beta + \epsilon$, where ϵ is an $n \times 1$ vector of mean zero, uncorrelated errors with a common variance σ^2. Assume that $X \in \mathbb{R}^{n\times r}$ is of rank r and $n > r$. The least squares estimator of β is $\hat{\beta} = (X^TX)^{-1}X^TY$. Consequently, $\hat{Y} = X\hat{\beta}$ and $e = Y - X\hat{\beta}$.

Properties: 1. $\mathbb{E}\big[\hat{Y}\big] = X\beta$ and $\text{cov}\big(\hat{Y}\big) = X(X^TX)^{-1}X^T\sigma^2 \equiv H\sigma^2$. 2. $\mathbb{E}[e] = 0$ and $\text{cov}(e) = (I-H)\sigma^2$. 3. $\text{cov}\big(\hat{Y}, e\big) = 0$. Furthermore, if ϵ is i.i.d. $N(0,\sigma^2)$, $\hat{Y}$ and e are independent. 4. $\mathbb{E}[\text{SSE}] = (n_T - r)\sigma^2$, where $\text{SSE} = e^Te$. Thus, $\mathbb{E}[\text{MSE}] = \sigma^2$.

How does the ANOVA model relate to the Gauss-Markov models? - Y: turn $\{Y_{ij}\}$ to $Y \in \mathbb{R}^{n_T}$

- X: an $n_T \times r$ matrix of 0's and 1's
- β: $\beta \equiv (\mu, \mu_2, \ldots, \mu_r)^T$

Corresponding properties of ANOVA: 1. $\big\{\hat{Y}_{ij}\big\}$ are independent of the residuals $\{e_{ij}\}$ 2. MSTR $\perp$ MSE 3. $\mathbb{E}[\text{MSE}] = \sigma^2$ 4. $\mathbb{E}[\text{MSTR}] = \sigma^2 + \sum n_i(\mu_i - \mu)^2/(r-1)$ where $\mu = \sum_{i=1}^r (n_i/n_T)\mu_i$ 5. $\text{SSE}/\sigma^2 \sim \chi^2_{n_T - r}$ and when $\mu_1 = \cdots = \mu_r$ and $\text{SSTR}/\sigma^2 \sim \chi^2_{r-1}$. 6. When $\mu_1 = \cdots = \mu_r$ then $F = \text{MSTR}/\text{MSE}$ has an F-distribution with degrees of freedom $(r-1, n_T - r)$.

1.1.6 4.1.6 F-test

Recall that the question of interest in the Spock trial data can now be translated to whether $\{\mu_i\}_{i=1}^r$ are the same. Formally, we want to test the null hypothesis

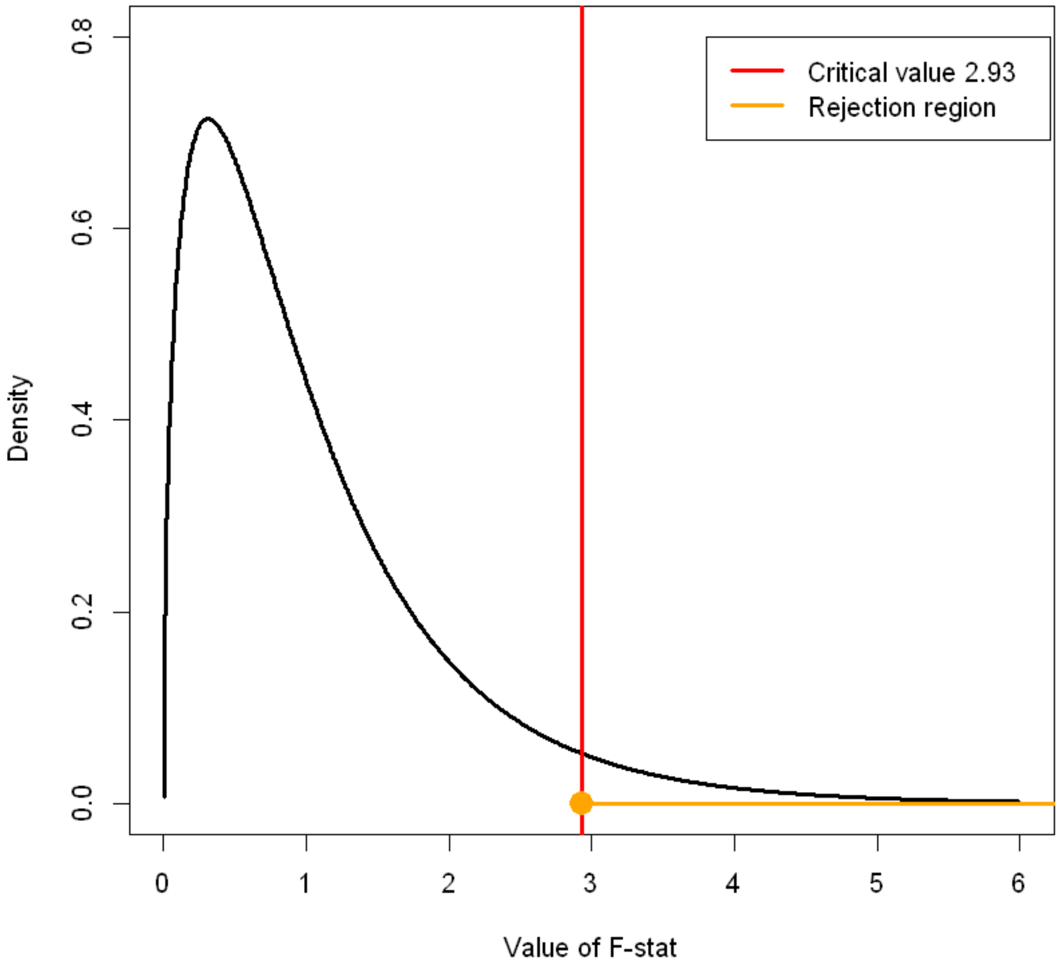
$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_r$$

against the alternative

$$H_1 : \text{not all } \mu_i \text{ are the same.}$$

The first test statistic we consider is the F-statistic $F^* \equiv \text{MSTR}/\text{MSE}$. Under the null, F^* follows an F-distribution with d.f.s $(r-1, n_T - r)$ when assuming ϵs are i.i.d. $N(0, \sigma^2)$. We reject the null hypothesis at the significance level α, if $F^* > F(1-\alpha; r-1, n_T - r)$. Here the value of $F(1-\alpha; r-1, n_T - r)$ is known as the *critical value*, and the set $(F(1-\alpha; r-1, n_T - r), +\infty)$ is the *rejection region*.

[34]:
```r
# Visualization of critical value, rejection region for F-test
r=4;n=33;alpha=0.05;
x.grid=seq(from=1e-5,to=6,length.out=1000);
density.grid=df(x=x.grid, df1=r-1, df2=n-r)
critical.value=qf(1-alpha,df1=r-1,df2=n-r);
plot(density.grid~x.grid,type='l',xlab="Value of␣
 ↪F-stat",ylab="Density",lwd=3,xlim=c(0,6),ylim=c(0,0.8))
abline(v=critical.value,lwd=3,col='red')
segments(x0=critical.value,x1=10,y0=0,y1=0,lwd=3,col="orange")
points(x=critical.value,y=0,pch=16,col="orange",cex=2)
legend(x=3.8,y=0.8,legend=c(paste0('Critical value ', round(critical.
 ↪value,digits=2)), 'Rejection region'),lty=1,lwd=3,col=c('Red','Orange'))
```



Example. In the Spock trial, we want to test the null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ against the alternative H_1 : not all μ_i's are equal. We can calculate the F-statistics $F^* = \text{MSTR}/\text{MSE} = 17.6$, when $F(0.95; 3, 29) = 2.93$. We can thus reject the null hypothesis at the nominal significance level 0.05.

Rationale. In real life, we reject a hypothesis if the null is *unlikely to be true* given our observations. We say the null hypothesis is *unlikely to be true* if the observed data is very **extreme** under the null hypothesis. We often calculate the p-value, which is *the probability of observing a more extreme test statistic F than the test statistic F^* calculated on the current sample.* In other words,

$$\text{pvalue } = \mathbb{P}(F \geq F^*).$$

We should see the p-value as a summary measure of the plausibility of the null hypothesis. Noting the subjectivity in "plausibility", we thus should specify the significance level (i.e., the acceptable type I error rate) before data analysis.

Distribution of F^* under the alternative. The numerator SSTR/σ^2 has a non-central Chisquare distribution with d.f. $r - 1$ and non-centrality parameter $\sum_{i=1}^r n_i(\mu_i - \mu)^2/\sigma^2$. We know that F^* follows a non-central F-distribution with $df = (r - 1, n - r)$ and the noncentrality parameter $\sum_{i=1}^r n_1(\mu_i - \mu)^2/\sigma^2$.

Expectation of F^*. To derive the expectation of F^*, we have

$$\mathbb{E}[F^*] = \mathbb{E}\left[\frac{\text{MSTR}}{\text{MSE}}\right] = \mathbb{E}[\text{MSTR}]\mathbb{E}\left[\frac{1}{\text{MSE}}\right],$$

where the second equality holds due to independence between MSTR and MSE. We know that the numerator $\mathbb{E}[\text{MSTR}] = (n_T - r)^{-1}\sum_{i=1}^r n_i(\mu_i - \mu)^2/\sigma^2$. It remains to derive the expectation of the denominator.

We have

$$\left[\frac{1}{\text{MSE}}\right] = \frac{1}{\sigma^2}(n_T - r)\mathbb{E}\left[\frac{1}{\chi^2_{n_T-r}}\right] = \frac{1}{\sigma^2}\frac{n_T - r}{n_T - r - 2},$$

because $(n_T - r)\text{MSE}/\sigma^2 \sim \chi^2_{n_T-r}$ and $\mathbb{E}[1/\chi^2_v] = 1/(v - 2)$ for $v > 2$.

Therefore,

$$\begin{aligned}\mathbb{E}[F] =& \frac{n_T - r}{n_T - r - 2}\left[1 + \sum_{i=1}^r \frac{n_i}{r - 1}\left(\frac{\mu_i - \mu}{\sigma}\right)^2\right] && (6)\\ \approx& 1 + \sum_{i=1}^r \frac{n_i}{r - 1}\left(\frac{\mu_i - \mu}{\sigma}\right)^2. && (7)\end{aligned}$$

Interpretation of F^*. 1. MSE is an estimator of σ^2, but the maximum likelihood estimator for σ^2 is SSE/n_T. 3. The ratio F^* fluctuates about

$$\frac{\mathbb{E}[\text{MSTR}]}{\mathbb{E}[\text{MSE}]} = 1 + \sum_{i=1}^r \frac{n_i}{r - 1}\left(\frac{\mu_i - \mu}{\sigma}\right)^2,$$

where the second term is a unit-free measure of the variablity among $\mu_1, \mu_2, \ldots, \mu_r$. Under the null, the second term equals zero and thus F fluctuates around one.

1.1.7 4.1.7 Testing linear combination

For any $c_i, i = 1, \ldots, r$, the least squares and maximum likelihood estimator of $L = \sum_{i=1}^{r} c_i \mu_i$ is $\hat{L} = \sum_{i=1}^{r} c_i \bar{Y}_{i\cdot}$, and further

$$\mathbb{E}\big[\hat{L}\big] = L, \text{ and } \operatorname{var}\big(\hat{L}\big) = \sum_{i=1}^{r} \frac{c_i^2}{n_i} \sigma^2.$$

Therefore, an unbiased estimator of $\operatorname{var}\big(\hat{L}\big)$ is

$$s^2\big(\hat{L}\big) = \sum_{i=1}^{r} \frac{c_i^2}{n_i} \text{MSE}.$$

Under the null hypothesis $H_0 : L = 0$ and the normality assumption on ϵ, we have

$$\frac{\hat{L} - L}{s\big(\hat{L}\big)} \sim t(n_T - r).$$

To test the hypothesis $H_0 : L = 0$ against $H_1 : L \neq 0$. We can calculate the t-statistic $t^* = \hat{L}/s(\hat{L})$. We can calculate the quantile of the t-distribution as before to finish the test.

Example. Consider the quantity $L = \mu_1 - (\mu_2 + \mu_3 + \mu_4)/3$. We can calculate that $\hat{L} = -15.217$, and $s^2(\hat{L}) = 4.648$. We can then test the null hypothesis or construct a 99% confidence interval with $t(1 - 0.01/2; n_T - r) = 2.756$.

[4]:
```
(Lhat=unname(round(sum(anova.fit[[1]][2:4])/3,digits=3)))
(s2=round(sum(anova.fit[[2]]^2)/anova.fit[[8]]*sum(c(1,-1/3,-1/3,-1/3)^2/
 ↪table(Spock[,1])),digits=3))
(tquantile=round(qt(1-0.01/2,anova.fit[[8]]),digits=3))
```

```
15.217
```

```
4.648
```

```
2.756
```

1.1.8 4.1.8 Simultaneous inference

Assume that we have a few linear combinations, $L_1, L_2, \ldots, L_m$. Let $L \equiv (L_1, \ldots, L_m)^T$. We can construct a simultaneous interval for L as $\hat{L} \mp Ks\big(\hat{L}\big)$ for some appropriate multiplier K, where $\hat{L}$ is an estimator of L and $s(\hat{L})$ is the standard deviation of $\hat{L}$.

[3]:
```
mse=sum(anova.fit$residuals^2)/anova.fit$df.residual;

# Create vectors for the first two linear combinations
```

```
comb.mat<-matrix(0,nrow=2,ncol=4)
comb.mat[1,]=c(1,-1,0,0);comb.mat[2,]=c(1,0,-1,0);

# Obtain the estimates
diff = numeric(dim(comb.mat)[1]);
diff.sd=diff;
mean.tmp=anova.fit$coefficients;mean.tmp[1]=0;
ns=as.numeric(table(Spock$Judge));
for(i in 1:length(diff)){
  diff[i]=sum(comb.mat[i,]*mean.tmp);
  diff.sd[i]=sqrt(sum(comb.mat[i,]^2/ns)*mse);
}
alpha=0.05;
```

1. Bonferroni method. The multiplier is $K = t\big(1-\alpha/(2m); n_T - r\big)$. Then, the simultaneous intervals cover the true parameters with probability at least $(1-\alpha)$.

[4]:
```
# Bonferroni correction:
m=6; # for all pairwise differences, although we only show two here
B.stat=qt(1-alpha/(2*m),anova.fit$df.residual);
```

2. Tukey-Kramer method. This approach only works for pairwise comparisons, e.g., $\mu_i - \mu_{i'}$. The $100(1-\alpha)\%$ confidence interval for $\{\mu_i - \mu_{i'} : i, i' \in \{1, \ldots, r\}, i \neq i'\}$ is

$$\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot} \mp Ts\big(\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot}\big),\ i \neq i',\ T = \frac{1}{\sqrt{2}} q(1-\alpha; r, n_T - r),$$

where q is the studentized range distribution. The coverage is exactly $1-\alpha$ for a balanced ANOVA model, and at least $1-\alpha$ for unbalanced cases.

[5]:
```
# Tukey-Kramer
T.stat=qtukey(1-alpha, nmeans=length(anova.fit$coefficients), df=anova.fit$df.
 ↪residual)/sqrt(2);
```

3. Scheffe method. This approach applies to all possible constrasts, where the coefficients sum up to zero for each L_j, i.e.,$L_j = \sum_{i=1}^r c_{ji}\mu_i$ and $\sum_{i=1}^r c_{ji} = 0$. The multiplier is $S = \big[(r-1)F(1-\alpha; r-1, n_T - r)\big]^{1/2}$. The coverage is exactly $1-\alpha$ for *all* possible contrasts. For finitely many contrasts, the coverage is at least $1-\alpha$.

[6]:
```
# Scheffe
S.stat=sqrt( (length(anova.fit$coefficients)-1)*qf(1-alpha,length(anova.
 ↪fit$coefficients)-1,anova.fit$df.residual))
```

Choosing which method - All pair comparison, (near) balanced design, then Tukey method is the best. - If testing m hypothesis, m is small use Bonferroni; If m large, use Scheffe method.

[7]:
```
table.stats=matrix(0,1,3);
table.stats[1,]=c(B.stat,T.stat,S.stat);
colnames(table.stats)=c('Bonferroni', 'Tukey', 'Scheffe')
```

```
table.stats
# Then, we can construct the confidence intervals as, e.g.,
CI.bonferroni =matrix(0,nrow=2,ncol=2);
for(i in 1:length(diff)){
  CI.bonferroni[i,]=diff[i]+c(1,-1)*B.stat*diff.sd[i];
}
```

A matrix: 1 × 3 of type dbl

Bonferroni	Tukey	Scheffe
2.831553	2.724509	2.966832

1.1.9 4.1.9 Diagnostics of one-way ANOVA

Recall that the (cell means) model for a one-way ANOVA is

$$Y_{ij} = \mu_i + \epsilon_{ij},\ j = 1, \ldots, n_i,\ i = 1, \ldots, r,$$

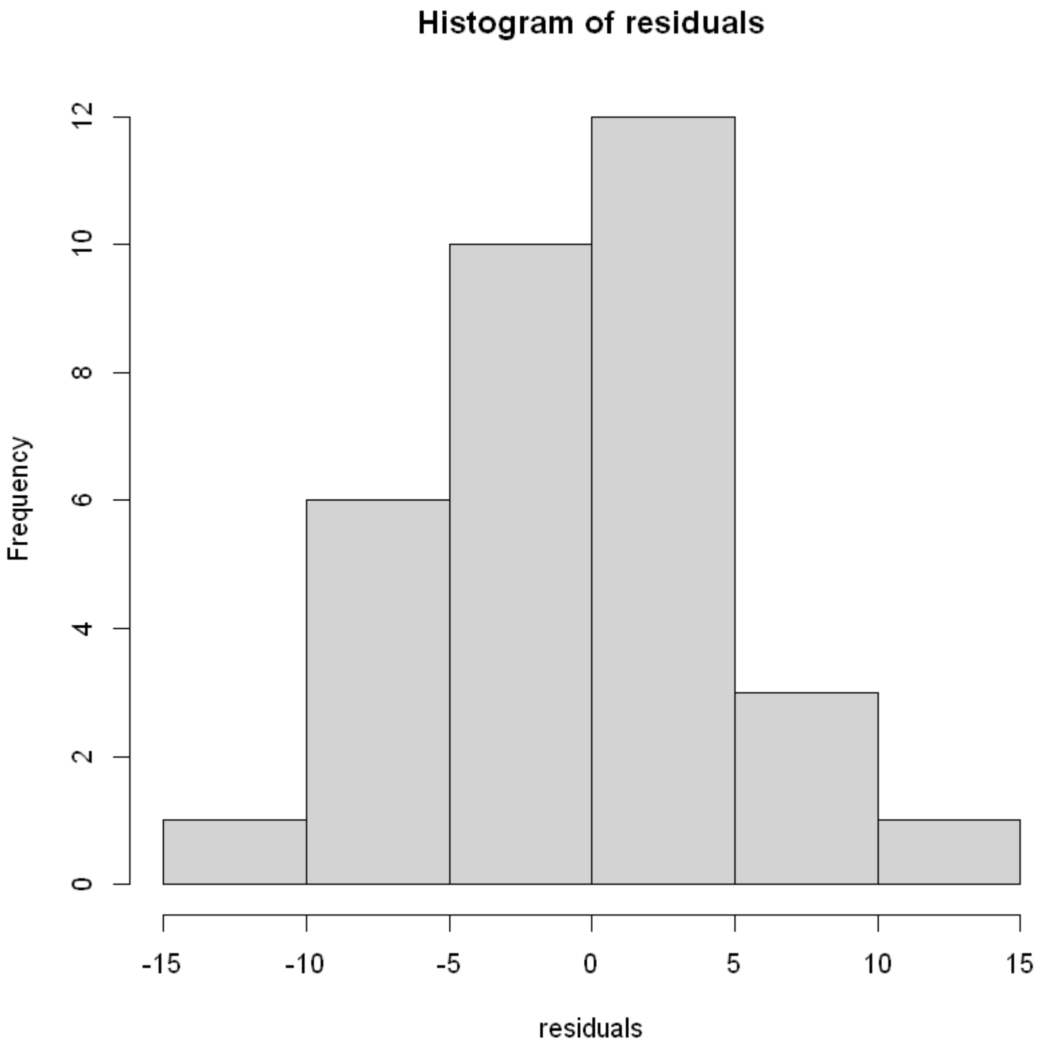
where ϵ_{ij} are i.i.d. $N(0, \sigma^2)$.

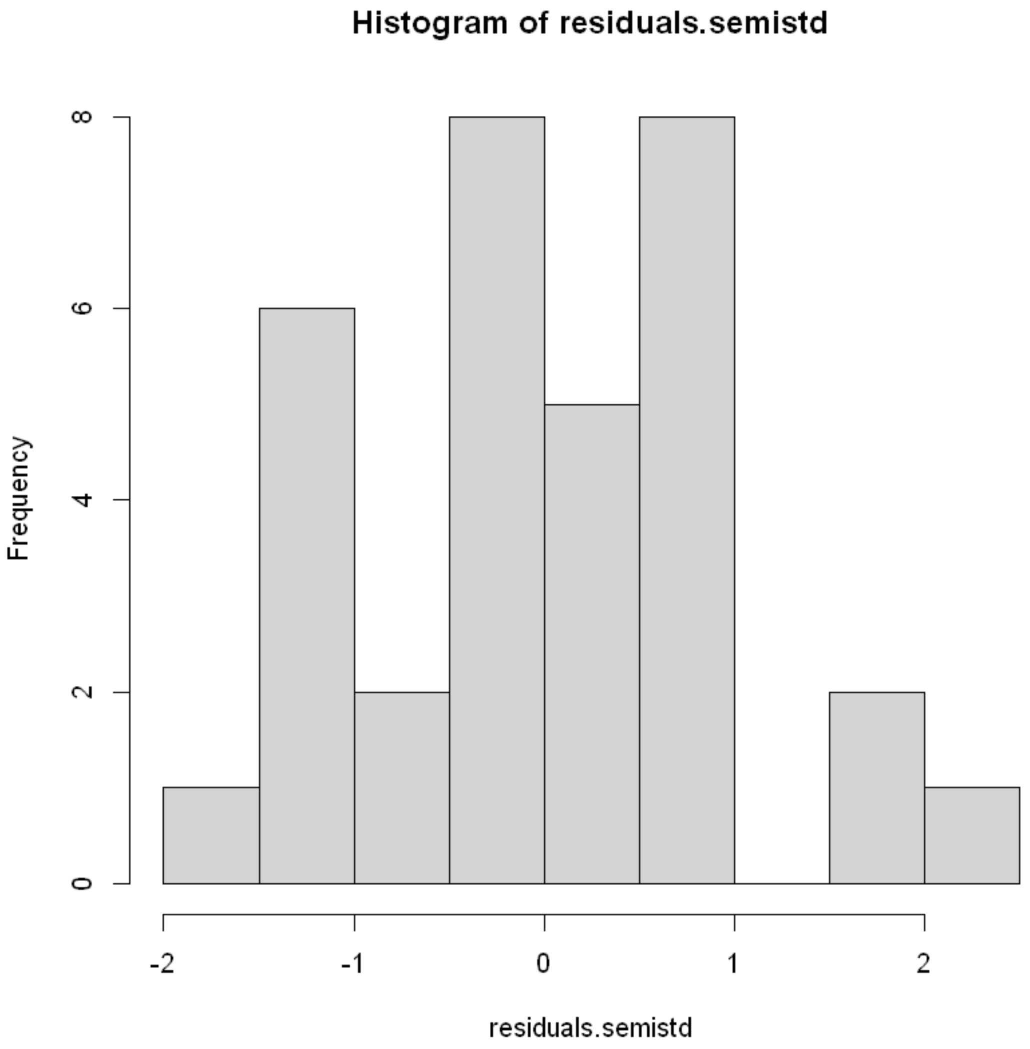
Possible departures from the model assumptions 1. Variances of error terms are unequal. 2. Error terms are not independent. 3. Error terms are not normally distributed. 4. Outliers (samples that do not follow this model). 5. Missing variables.

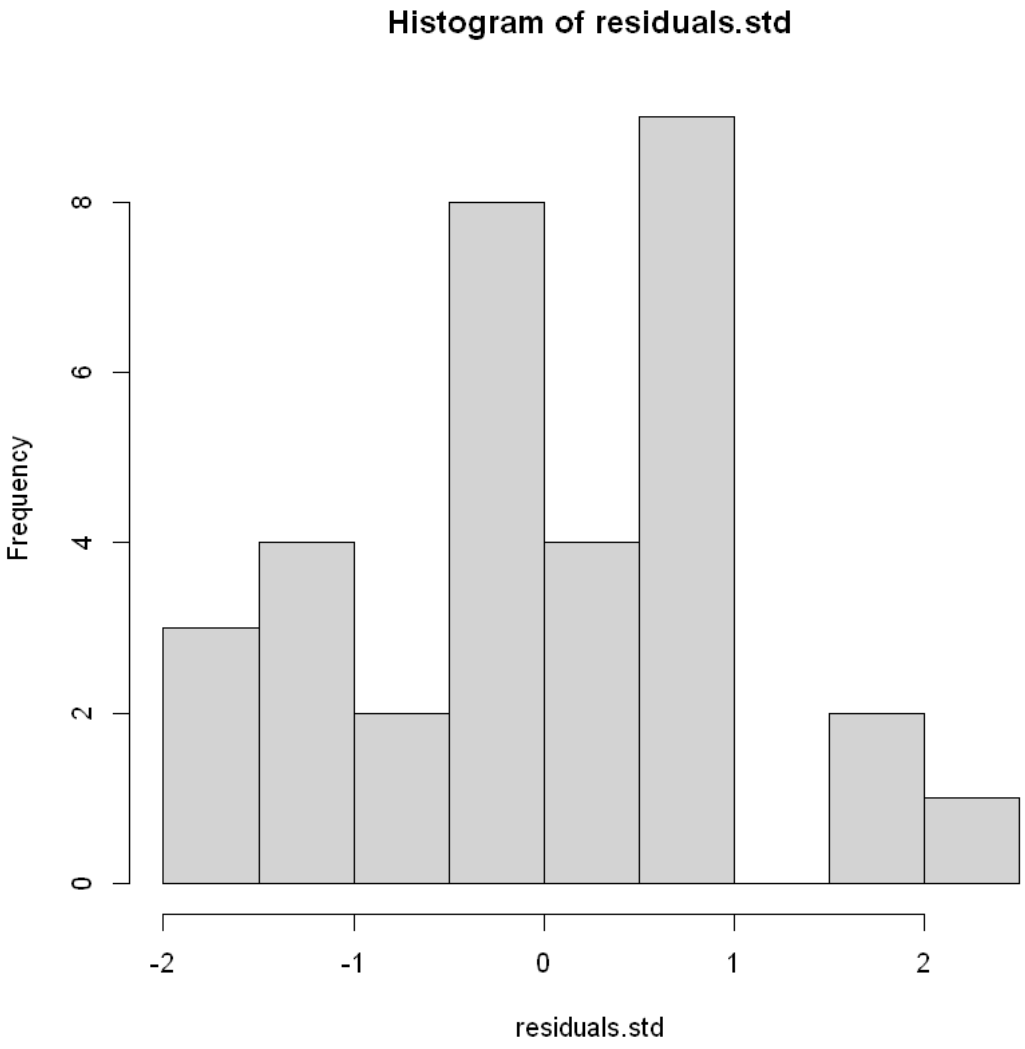
All diagnostics rely on the *residuals*. - Residuals $e_{ij} \equiv Y_{ij} - \bar{Y}_{i\cdot}$. - Semistudentized residuals $e^*_{ij} = e_{ij}/\sqrt{\text{MSE}}$ - Studentized residuals $r_{ij} = e_{ij}/s(e_{ij})$, where $s^2(e_{ij}) = (1 - 1/n_i)\text{MSE}$.

[8]:

```
# Obtain the residuals from the ANOVA fit
residuals=anova.fit$residuals;
hist(residuals)
# Semistudentized residuals
residuals.semistd=anova.fit$residuals/sqrt(mse);
hist(residuals.semistd)
# Studentized residuals
weights=1-1/ns[as.numeric(Spock$Judge)];
residuals.std=anova.fit$residuals/sqrt(mse)/sqrt(weights);
hist(residuals.std)
```


Histogram of residuals
Frequency
0
2
4
6
8
10
12
-15
-10
-5
0
5
10
15
residuals


Histogram of residuals.semistd
Frequency
0
2
4
6
8
-2
-1
0
1
2
residuals.semistd

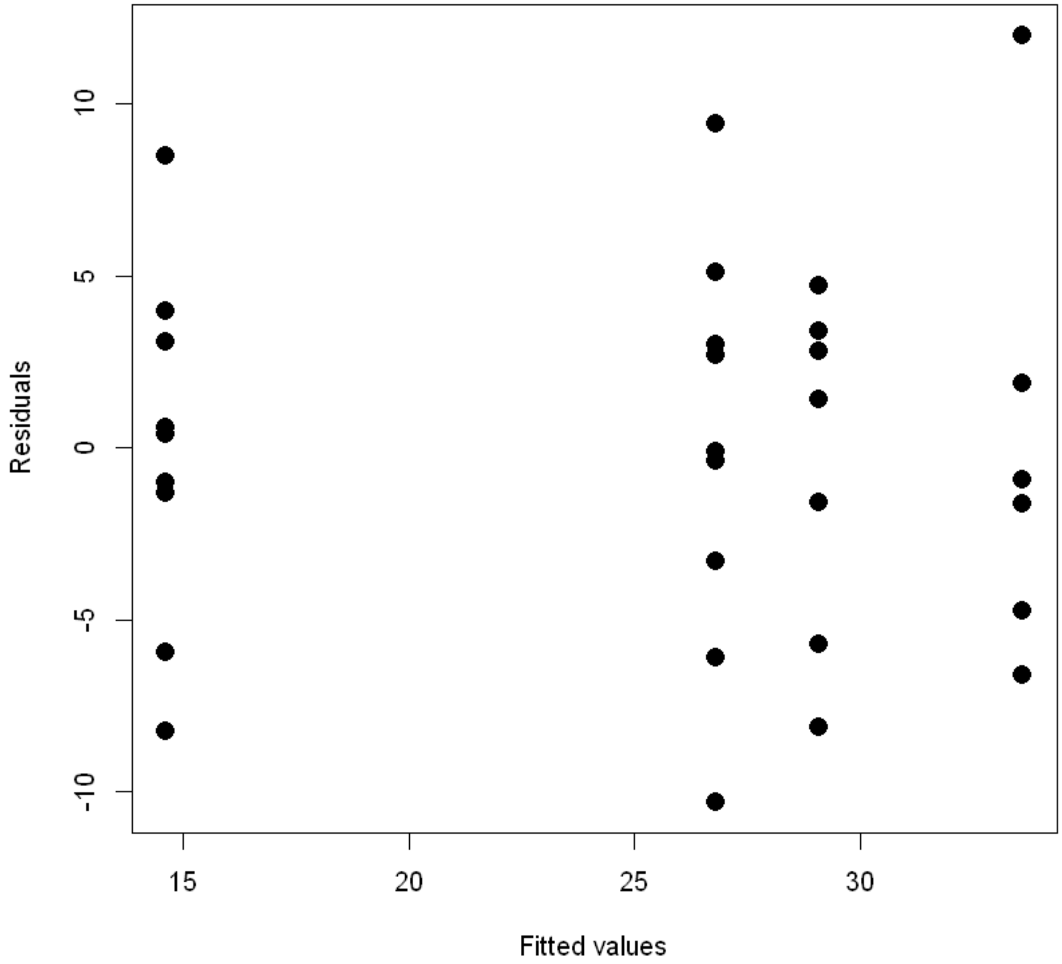


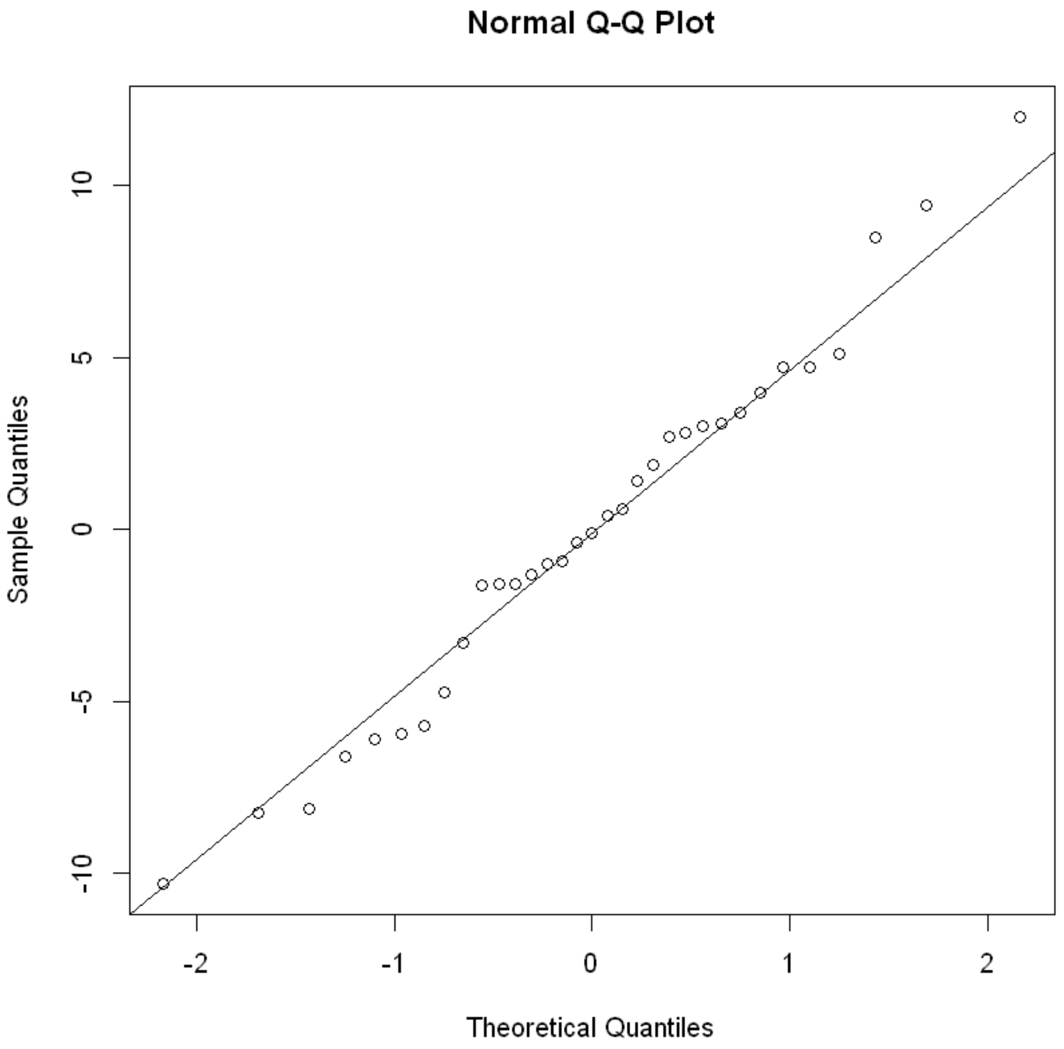
With residuals, we usually begin the diagnostics with visualizations to explore how the residuals are related with other vairbales.

1. e^* or r v.s. $\hat{Y}_{ij}$
2. e^* or r v.s. indices or other structures
3. Quantile-Quantile plot or stem-leaf plot
4. e^* or r v.s. missing variables.
5. …

[9]:
```
# Plot the residuals (or the other two versions) against fitted values
plot(residuals~anova.fit$fitted.values,type='p',pch=16,cex=1.5,xlab="Fitted␣
 ↪values",ylab="Residuals")

# Plot the residual against certain orders
# No clear orders make sense in the Spock trial data

# Stem-leaf plot  (or use histogram, or qq-plot )
stem(residuals)
qqnorm(residuals);qqline(residuals)

# Plot residuals against missing variables
# Not applicable on Spock trial data
```

```
 The decimal point is at the |

 -10 | 3
  -8 | 21
  -6 | 61
  -4 | 977
  -2 | 3
  -0 | 66630941
   0 | 4649
   2 | 78014
   4 | 0771
   6 |
   8 | 54
  10 |
  12 | 0
```


Residuals
10
5
0
-5
-10
15
20
25
30
Fitted values



1.1.10 4.1.10 Testing equal variance

Estimate $\sigma_1^2, \sigma_2^2, \ldots, \sigma_r^2$ separately as, for $i = 1, \ldots, r$,

$$s_i^2 = \sum_{j=1}^{n_i} \frac{\left(Y_{ij} - \bar{Y}_{i.}\right)^2}{n_i - 1}.$$

We want to test the null hypothesis $H_0 : \sigma_1 = \cdots = \sigma_r$ against the alternative H_a : not all σs are equal.

[10]:

```
# Calculate the variances for each group:
(vars = tapply(Spock$perc.women,Spock$Judge,var))
alpha=0.05;
```

1 25.3894444444444 **2** 43.3256666666667 **3** 21.095 **4** 35.6275

Hartley test. The test statistic is

$$H = \frac{\max(s_1^2, \ldots, s_r^2)}{\min(s_1^2, \ldots, s_r^2)}.$$

At significance level α, reject H_0 if $H > H(1-\alpha; r, n_i - 1)$ when all n_is are the same (balanced design).

[11]:

```
# Hartley test:
H.stat=max(vars)/min(vars);
#install.packages('SuppDists')
library(SuppDists) # The distribution is in this package
# Both df and k only take integers:
qmaxFratio(1-alpha,df=floor(sum(ns)/length(ns)-1),k=length(ns))
qmaxFratio(1-alpha,df=ceiling(sum(ns)/length(ns)-1),k=length(ns))
```

8.43999329265814

7.18532327249131

Bartlett test. The test statistics is

$$K^2 = (n_T - r)\log(\text{MSE}) - \sum_{i=1}^{r} (n_i - 1)\log(s_i^2).$$

We know that $K^2 \geq 0$ from Jensen's inequality. Under H_0, K^2 is approximately χ^2_{r-1} assuming that n_i are not small. Reject H_0 if $K^2 > \chi^2(1-\alpha; r-1)$ at significance level α. Related to the likelihood ratio test.

[12]:

```
# Bartlett test:
K.stat= (sum(ns)-length(ns))*log(mse)-sum( (ns-1)*log(vars) );
qchisq(1-alpha,df=length(ns)-1)
```

7.81472790325118

Levene test. 1. Create new data with $d_{ij} = |Y_{ij} - \bar{Y}_{i\cdot}|$. 2. Treat $\{d_{ij}\}$ as response variables 3. Calculate the F-statistic for $H_0 : \mathbb{E}[d_{1\cdot}] = \mathbb{E}[d_{2\cdot}] = \cdots = \mathbb{E}[d_{r\cdot}]$ Reject H_0 if $F^* > F(1-\alpha; r-1, n_T - r)$ at significance level α.

[13]:

```
# Levene test:
Spock$res.abs=abs(anova.fit$residuals);
summary(aov(res.abs~Judge,data=Spock))
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
Judge        3   5.64    1.88   0.173  0.914
Residuals   29 314.70   10.85
```

1.1.11 4.1.11 Remedies for departures from model assumptions

Weighted least squares

Idea: $\sqrt{w_i}\epsilon_{ij} \sim N(0,1)$ if $w_i = 1/\sigma_i^2$.

Find weighted least squares estimator by minimizing $\sum_{i=1}^r \sum_{j=1}^{n_i} w_i(Y_{ij} - \mu)^2$ for the common mean $\tilde{\mu} = \sum n_i w_i \bar{Y}_{i\cdot} / \sum n_i w_i$. We will have a new set of SSTR and SSE. In the end, we still have $F^* \sim F(r-1, n_T - r)$.

In practice, we plug in $w_i = 1/s_i^2$, and the null distribution remains the same.

Rank test

$$F^* = \frac{\mathrm{MSTR}(R)}{\mathrm{MSE}(R)} = \frac{\sum\sum(\bar{R}_{i\cdot} - \bar{R})^2/(r-1)}{\sum\sum(R_{ij} - \bar{R}_{i\cdot})^2/(n_T - r)} \sim F(r-1, n_T - r),$$

where R_{ij} is the rank of Y_{ij} among all n_T observations. Works when the sample size is large.

[14]:

```
# The rank test
Spock$rank.perc=rank(Spock$perc.women)
summary(aov(rank.perc~Judge,data=Spock))
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)    
Judge        3   1846   615.4    15.6 3.15e-06 ***
Residuals   29   1144    39.5                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Kruskal-Wallis test

$$F^* = (n_T - 1)\frac{\sum_{i=1}^r (\bar{R}_{i\cdot} - \bar{R}_{\cdot\cdot})^2}{\sum\sum(R_{ij} - \bar{R}_{\cdot\cdot})^2} \sim \chi^2_{r-1},$$

if n_T is large.

[15]:

```
# Krusal-Wallis test:
kruskal.test(perc.women~Judge,data=Spock)
```

```

	Kruskal-Wallis rank sum test

data:  perc.women by Judge
Kruskal-Wallis chi-squared = 19.757, df = 3, p-value = 0.0001906
```

Box-Cox transformation

$$Y(\lambda) = \frac{Y^\lambda - 1}{\lambda},$$

and $Y(0) \equiv \log(Y)$ for $\lambda = 0$.

To tune the parameter λ, we can calculate the likelihood $L(\lambda) \equiv \max_{\mu,\sigma} L(\lambda; \mu, \sigma)$. Then $\lambda^* = \arg\max L(\lambda)$.

It can be shown that $\max L(\lambda)$ is equivalent to $\min \mathrm{SSE}[Y^*(\lambda)]$, where

$$Y^*_{ij}(\lambda) \equiv \begin{cases} \frac{Y_{ij}^{\lambda}-1}{\lambda \dot{Y}^{\lambda}-1} & \lambda \neq 0 \\ \dot{Y}\log(Y_{ij}) & \lambda = 0 \end{cases},$$

where $\dot{Y}$ is the geometric mean of Y.

Box-Cox for equal variance amounts to minimize the Bartlett statistics, or other test statistics. We can use the `boxcox()` function in library `MASS` for Box-Cox transformation in `R`.

1.2 4.1.12 Power calculation

Recall that type II error is *accepting the null hypothesis when it is false.* The type II error rate is the probability of type II error given repeated samples from the true population, i.e., Prob(Not rejecting $H_0|H_0$ false). The **power** of a testing procedure is defined as *the probability of rejecting H_0 when it is false.* We have the following equality

$$\text{Power} = \text{Prob(Reject } H_0 \mid H_0 \text{ false)} = 1 - \text{type II error rate.}$$

The power of a test depends on multiple factors including

- Deviation of the alternative H_a from the null H_0, or the strength of signal.
- The noise lelve, or the degree of random fluctuation. For instance, a larger var(ϵ) in an ANOVA model leads to lower power of the F-test.
- Significance level. Recalling that the significance level is the type I error rate. A larger significance level results in larger power, which reflects the trade-off between type I and II error rates.
- Sample size. The more samples we have, the more powerful the test is.

Among the four factors above, we can only change the significance level during the analysis. However, given the relationship between type I and type II error rates, simply increasing the significance level will decrease the reliability of the test results. As an extreme case, a test at significance level 100% has a power of 100%, but this test is meaningless. Moreover, changing the significance level during analysis is considered as a form of p-hacking.

Prior to the study, we can determine the sample size of the study to achieve a desirable power at a given significance level. This is known as the **power calculation**. To this end, we need to know the other factors that influence the power. Here the significance level is pre-specified and hence known to us prior to the study. Deviation of the alternative and the noise level, however, are population quantities that are unknown, which is the reason why we need to conduct a study at the first place. In practice, we can either conduct a pilot study to collect a few samples to estimate these parameters, or to specific a level of clinical significance. Finally, we will need to use the property of the testing procedure to "predict" its power. Briefly, we need to obtain the distribution of the test statistic under the alternative hypothesis, and then to calculate the probability that the test statistic surpasses the critical value. The distribution of F-test in Section 4.1.6 will come in handy if we want to calculate the power before hand.