Chapter 2

One-way ANOVA: Fixed factor levels

2.1 Questions, assumptions and model

1. What does *one-way* mean?

   One-way means that a single factor is of interest.

2. What are *factor levels*?

   A factor level is a particular form or value of the factor.
   (To understand what this really mean, we need examples.)

3. What are the major questions on this topic?

   (1) Determine whether or not the factor level means are the same.

   (2) If the factor level means differ, examine

   (i) how they differ
   (ii) what the implications of the difference are

2.1.1 Assumptions

(i) r independent populations;

(ii) Each population follows a normal distribution $\mathcal{N}(\mu_i, \sigma_i^2), i = 1, \ldots, r$;

(iii) Equal variance (but unknown):

$$\sigma_1^2 = \ldots = \sigma_r^2 = \sigma^2;$$

(iv) r independent random samples

$$Y_{11}, \ldots, Y_{1n_1} \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mu_1, \sigma^2)$$

$$Y_{21}, \ldots, Y_{2n_2} \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mu_2, \sigma^2)$$

$$\ldots\ldots$$

$$Y_{r1}, \ldots, Y_{rn_r} \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mu_r, \sigma^2)$$

Table 2.1: Format of data set

Factor level	Population mean	Sample	
		Observations	Sample mean
level 1	μ_1	Y_{11} $\vdots$ Y_{1n_1}	$\bar{Y}_{1\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
level i	μ_i	Y_{i1} $\vdots$ Y_{in_i}	$\bar{Y}_{i\cdot} \triangleq \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
level r	μ_r	Y_{r1} $\vdots$ Y_{rn_r}	$\bar{Y}_{r\cdot}$

Balanced data: $n_1 = \ldots = n_r$ (equal sample sizes)
Unbalanced data: unequal sample sizes

2.1.2 Major question

Compare population means $\mu_1, \ldots, \mu_r$

2.1.3 Cell means model

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

$\mu_1, \ldots, \mu_r$ are mean parameters;

Error terms $\varepsilon_{ij}, \quad i = 1, \ldots, r, \quad j = 1, \ldots, n_i$, are independent $\mathcal{N}(0, \sigma^2)$

2.1.4 Notations

$$n_T \triangleq \sum_{i=1}^{r} n_i \qquad \text{(The total number of cases in the study)}$$

$$\bar{Y}_{i\cdot} \triangleq \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} \qquad \text{(Sample mean of the i-th factor level)}$$

$$\bar{Y}_{\cdot\cdot} \triangleq \frac{1}{n_T}\sum_{i=1}^{r}\sum_{j=1}^{n_i} Y_{ij} = \frac{1}{n_T}\sum_{i=1}^{r} n_i \bar{Y}_{i\cdot} \qquad \text{(The overall mean for all responses)}$$

$$SSTR \triangleq \sum_{i=1}^{r} n_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 \qquad \text{(Treatment sum of squares)}$$

$$SSE \triangleq \sum_{i=1}^{r}\sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2 \qquad \text{(Error sum of squares)}$$

$$SSTO \triangleq \sum_{i=1}^{r}\sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{\cdot\cdot})^2 \qquad \text{(Total sum of squares)}$$

Formal development of partitioning.

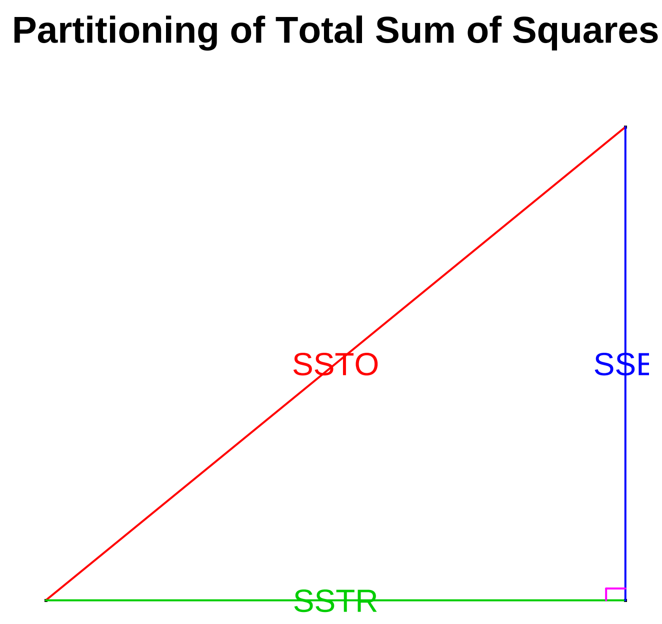
The total deviation $Y_{ij} - \bar{Y}_{\cdot\cdot}$, used in the measure of the total variation of the observations Y_i without using any information about factor levels, can be decomposed into two components:

$$\underbrace{Y_{ij} - \bar{Y}_{\cdot\cdot}}_{\text{Total deviation}} = \underbrace{Y_{ij} - \bar{Y}_{i\cdot}}_{\substack{\text{Deviation around} \\ \text{estimated factor levelmean} \\ \text{(within group)}}} + \underbrace{\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}}_{\substack{\text{Deviation of} \\ \text{estimated factor level mean} \\ \text{around overall mean} \\ \text{(between group)}}}$$

The sums of these squared deviations have the same relationship:

$$\underbrace{\text{SSTO}}_{(n_T-1 \text{ degrees of freedom})} = \underbrace{\text{SSE}}_{(n_T-r \text{ degree of freedom})} + \underbrace{\text{SSTR}}_{(r-1 \text{ degrees of freedom})}$$

It can be shown that SSE and $SSTR$ are independent.



2.2 F test equality of factor level means

Hypothesis testing problem:

$$\mathcal{H}_0: \ \mu_1 = \ldots = \mu_r \qquad \longleftrightarrow \qquad \mathcal{H}_a: \text{not all } \mu_i \text{ are equal.}$$

Test statistic:

$$F \ \triangleq \ \frac{MSTR}{MSE} \qquad\qquad (\text{ F test})$$

When $\mathcal{H}_0$ is true, $F \sim F_{r-1, n_T - r}$.

Two ways to make a decision:

(i) When controlling the level of significance at α, the decision rule is

$$\begin{cases} \text{If } \ observed\ F \leq qf(1-\alpha, r-1, n_T - r), & \text{conclude } \mathcal{H}_0 \\ \text{If } observed\ F > qf(1-\alpha, r-1, n_T - r), & \text{conclude } \mathcal{H}_a \end{cases}$$

(ii) Report $p - value$, which may be obtained via

$$p - value \ \triangleq \ 1 - pF(\ observed\ F, \ r-1, n_T - r)$$

A smaller $p - value$ leads us to conclude $\mathcal{H}_a$.

2.2.1 One-way ANOVA table

Table 2.2: ANOVA table

Source of Variation	df	Sum of Squares (SS)	Mean Square (MS)	F-value	$Pr(>F)$ (p-value)
Between treatments	$r-1$	SSTR	MSTR=$\frac{SSTR}{r-1}$	$F=\frac{MSTR}{MSE}$	
Error or Residuals (within treatments)	n_T-r	SSE	MSE=$\frac{SSE}{n_T-r}$		
Total	n_T-1	SSTO			

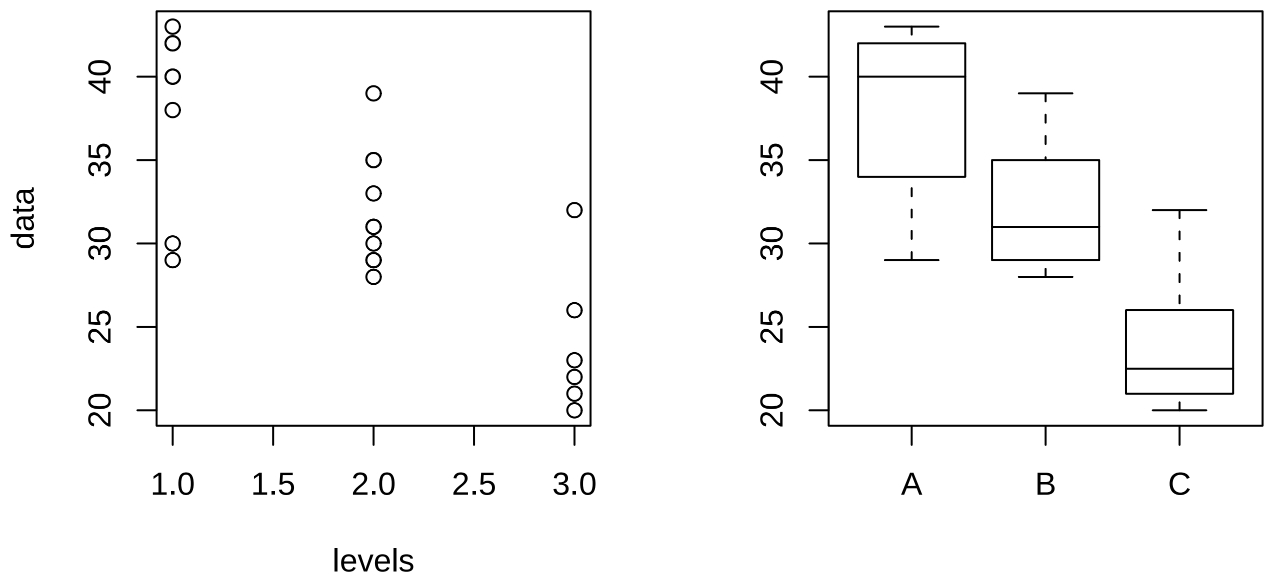
1. Intuitive idea about F-test

2. Derive F-test via the likelihood -ratio method

Example. Data set: Ch.16, Problem16.12

```
> x <- read.table("CH16PR12.DAT")
         V1  V2  V3
     1   29   1   1
        ......
     8   42   1   8
     9   30   2   1
        ......
     18  33   2  10
     19  26   3   1
        ......
     24  22   3   6

 > mean.all <- mean(x[,1])
   [1] 32
 > ssto <- (length(x[,1])-1)*var(x[,1])
   [1] 1088
 > new <- cbind(x, mean.all, x[,1])            # add two extra columns

 >  new[,5][x[,2]==1] <- mean(x[1:8,1])        # replace 5th column if 2nd column is 1
 >  new[,5][x[,2]==2] <- mean(x[9:18,1])
 >  new[,5][x[,2]==3] <- mean(x[19:24,1])

 >  new
        V1  V2  V3  mean.all  x[, 1]
    1   29   1   1       32     38
          ......
    8   42   1   8       32     38
    9   30   2   1       32     32
         ......
    18  33   2  10       32     32
    19  26   3   1       32     24
         ......
    24  22   3   6       32     24

  > sse <- sum((new[,1]-new[,5])^2)           # by definition
    [1] 416
  > sstr <- sum((new[,5]-new[,4])^2)
    [1] 672
```

```
> BelowAverage <- x[,1][x[,2]==1]
>  Average     <- x[,1][x[,2]==2]
> HighAverage  <- x[,1][x[,2]==3]
```



Define the *data frame*

```
> data   <- x[,1]
> levels <- factor(rep(LETTERS[1:3],
                  c(length(BelowAverage), length(Average), length(HighAverage))))

> therapy.df <- data.frame(levels, data)
    levels  data
 1       A   29
    ...
 9       B   30
    ...
 24      C   22
```

The first step is often to look at the data graphically.

```
 > plot(therapy.df)               # dot plot
```

```
> plot(levels, data)            # boxplot
```

To obtain the ANOVA table, type `aov`, and then `summary`.

```
> anova <- aov(data~levels, therapy.df)

> summary(anova)
                Df   Sum Sq   Mean Sq  F value    Pr(>F)
   levels        2   672.00   336.00   16.962     4.129e-05
   Residuals    21   416.00    19.81
```

Various treatment means can be easily obtained as

```
> model.tables(anova, type="means")
   Tables of means
   Grand mean
    32

   levels
         A   B   C
        38  32  24
   rep   8  10   6


    ##  Alternative form  -- Factor effects model
         (cf. Section 16.10 of the textbook)

>  model.tables(anova)
    Tables of effects

     levels
       A          B  C
       6 -1.700e-15 -8
   rep 8  1.000e+01  6


 >  fitted(anova)                      # fitted values
                                       # (DON'T print it)
```

What are the *fitted values*?

2.2.2 Two samples with equal variances: t-Test $\longleftrightarrow$ ANOVA

Assumptions:

(1) Two independent samples $Y_{11}, \ldots, Y_{1n_1}$ and $Y_{21}, \ldots, Y_{2n_2}$;

(2) $Y_{i1}, \ldots, Y_{in_i} \stackrel{\text{i.i.d}}{\sim} \mathcal{N}(\mu_i, \sigma_i^2) \quad (i = 1, 2)$;

(3) $\sigma_1 = \sigma_2$, but unknown.

Hypothesis testing problem:

$\mathcal{H}_0 : \ \mu_1 = \mu_2$,

$\mathcal{H}_a : \ \mu_1 \neq \mu_2$ (two-sided) or $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$ (one-sided).

Student's t-test:

$$t = \frac{\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where $S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$ is the pooled variance.

When $\mathcal{H}_0$ is true, $t \sim t_{n_1+n_2-2}$.

F-test for one-way ANOVA:

$$F = \frac{MSTR}{MSE} = \frac{SSTR}{SSE/(n_1+n_2-2)}.$$

When $\mathcal{H}_0$ is true, $F \sim F_{1,n_1+n_2-2}$.

Relationship between t-value and F-value: $F = t^2$

To see this, note that

$$\begin{aligned} SSTR &= n_1\Big(\bar{Y}_{1\cdot} - \frac{n_1\bar{Y}_{1\cdot} + n_2\bar{Y}_{2\cdot}}{n_1+n_2}\Big)^2 + n_2\Big(\bar{Y}_{2\cdot} - \frac{n_1\bar{Y}_{1\cdot} + n_2\bar{Y}_{2\cdot}}{n_1+n_2}\Big)^2 = \frac{n_1 n_2}{n_1+n_2}(\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot})^2 \\ SSE &= (n_1+n_2-2)S_p^2 \end{aligned}$$

2.3 Diagnostics and remedial measures

2.3.1 Basic assumptions, departures and remedial measures

Assumption	Departure
Constancy of error variance	Nonconstancy of error variance
Independence of error terms	Nonindependence of error terms
Normality of error terms	Nonnormality of error terms
	Presence of outliers
	Omission of important predictor variables

- Two methods for studying the appropriateness of a model

  (1) Graphic diagnostics;

  (2) Formal statistical tests with the null hypothesis being a basic assumption.

- (Remedial measures) Two choices if a model is not appropriate for a data set

  (1) Abandon the model and develop and use a more appropriate model;

  (2) Make some transformation on the data.

2.3.2 Residual analysis

- The *ij*th *residual* is the difference between the observed value y_{ij} and the corresponding fitted value $\bar{y}_{i\cdot}$.

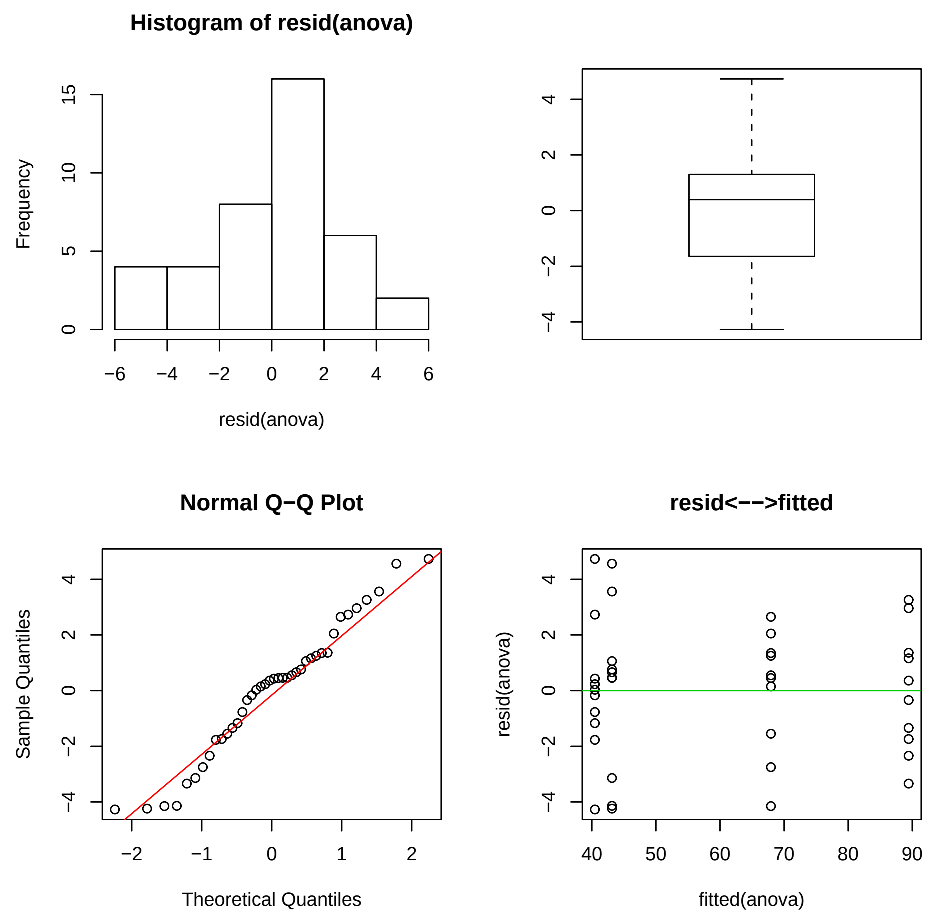
$$e_{ij} \triangleq y_{ij} - \bar{y}_{i\cdot}$$

It can be easily obtained via `resid`

```
> x <- read.table("CH17TA02.DAT")
> data <- x[,1]
> brands <- factor(rep(LETTERS[1:4],c(10,10,10,10)))       # 4 factor levels
> rust.df <- data.frame(brands,data)                         # data frame
> anova <- aov(data~brands, rust.df)

> resid(anova)                 # DON'T print it!
```

```
> hist(resid(anova))

> boxplot(resid(anova))

> qqnorm(resid(anova))
> qqline(resid(anova))

> plot(fitted(anova),resid(anova),main="resid<-->fitted")
> abline(h=0)
```

Figure 2.1: Various plots for the residuals



2.3.3 Tests for constancy of error variance

Assumptions

(i) r independent populations;

(ii) Each population follows a normal distribution $\mathcal{N}(\mu_i, \sigma_i^2), i = 1, \ldots, r$;

(iii) r independent random samples

$$\begin{aligned} Y_{11}, \ldots, Y_{1n} &\overset{\text{i.i.d}}{\sim} \mathcal{N}(\mu_1, \sigma^2) \\ Y_{21}, \ldots, Y_{2n} &\overset{\text{i.i.d}}{\sim} \mathcal{N}(\mu_2, \sigma^2) \\ &\cdots\cdots \\ Y_{r1}, \ldots, Y_{rn} &\overset{\text{i.i.d}}{\sim} \mathcal{N}(\mu_r, \sigma^2) \end{aligned}$$

Hypothesis testing problem

$$\mathcal{H}_0 : \sigma_1^2 = \ldots = \sigma_r^2 \qquad \longleftrightarrow \qquad \mathcal{H}_a : \text{ not all } \sigma_i^2 \text{ are equal}$$

- **The Hartley test** (only for the balanced case)

Hartley test statistic:

$$H \triangleq \frac{\max\{S_i^2\}}{\min\{S_i^2\}}$$

Values of H near 1 support $\mathcal{H}_0$, and large values of H support $\mathcal{H}_a$.

For a given α, use Table B.10 to obtain the threshold.

It reduces to the F-test when $r = 2$. (Explain the test intuitively).

Solder joint pull strengths - ABT electronics example (p.765).

```
>  x <- read.table("CH18TA02.DAT")
>  data <- x[,1]

> var.types <- c(var(data[1:8]),var(data[9:16]),var(data[17:24]),
                 var(data[25:32]),var(data[33:40]))

> hartley <- max(var.types)/min(var.types)   #  The Hartley test is available (why?)
   [1]   10.44493
```

- **The modified Levene test**

The modified Levene test is essentially the F-test for an ANOVA table with the data set based on the absolute deviations of the Y_{ij} observations about their respective factor level medians `median(Y_i.)`

$$d_{ij} \triangleq |Y_{ij} - median(Y_{i.})|$$

The modified Levene test statistic:

$$F_L^* \triangleq \frac{MSTR}{MSE} = \frac{MSTR/(r-1)}{MSE/(n_T - r)}$$

where

$$SSTR \triangleq \sum_{i=1}^{r} n_i(\bar{d}_{i.} - \bar{d}_{..})^2$$

$$SSE \triangleq \sum_{i=1}^{r}\sum_{j=1}^{n_i}(d_{ij} - \bar{d}_{i.})^2$$

$$\bar{d}_{i.} \triangleq \frac{1}{n_i}\sum_{j=1}^{n_i} d_{ij}$$

$$\bar{d}_{..} \triangleq \frac{1}{n_T}\sum_{i=1}^{r}\sum_{j=1}^{n_i} d_{ij} = \frac{1}{n_T}\sum_{i=1}^{r} n_i\bar{d}_{i.}$$

```
>  x <- read.table("CH18TA02.DAT")
```

To find group medians, type

```
>  group1.med <- median(x[,1][x[,2]==1])
>  group2.med <- median(x[,1][x[,2]==2])
>  group3.med <- median(x[,1][x[,2]==3])
>  group4.med <- median(x[,1][x[,2]==4])
>  group5.med <- median(x[,1][x[,2]==5])
```

or alternatively,

```
> med <- for(i in 1:5){
         med[i]<-median(x[,1][x[,2]==i])
         med}
```

Now d_{ij} is defined by

```
>  d <- x[,1]-c(rep(group1.med,8),
                rep(group2.med,8),
                rep(group3.med,8),
                rep(group4.med,8),
                rep(group5.med,8))
```

or alternatively,

```
>  d <- x[,1]-c(rep(med, c(8,8,8,8,8)))

>  d <- abs(d)                    # absolute value
```

Set up the date frame and then apply `aov`

```
>  types <- factor(rep(LETTERS[1:5],c(8,8,8,8,8)))
>  d.df <- data.frame(types, d)

>  anova <- aov(d~types, d.df)

> summary(anova)
            Df  Sum Sq Mean Sq F value  Pr(>F)
  types      4  9.3477  2.3369  2.9358 0.03414
  Residuals 35 27.8606  0.7960
```

What is your conclusion?

2.3.4 Transformations of response variable

Time between computer failures at three locations (in hours) – Servo-Data, Inc. Example, page 773

```
> y <- read.table("CH18TA05.DAT")
        V1 V2 V3
1     4.41  1  1
       ...
5    85.21  1  5
6     8.24  2  1
       ...
10    1.61  2  5
11  106.19  3  1
       ...
15   44.33  3  5
```

There are 3 factor levels in this case. Sample sizes are the same (balanced case).

```
> y1 <- y[,1][y[,2]==1]
> y2 <- y[,1][y[,2]==2]
> y3 <- y[,1][y[,2]==3]

> location <- factor(rep(1:3,c(5,5,5)))
> data <- y[,1]

> time.df <- data.frame(location, data)
```

Look at the data graphically. Outliers exist in groups 2 and 3.

```
> plot(location, data)

> plot(time.df)
```

The variances may not be the same.

```
> sd(y1)                    # check group sd
     [1] 42.29353
> sd(y2)
     [1] 33.21828
> sd(y3)
     [1] 127.1513           # sd(y3) is almost 4 times sd(y2)!
                            # Constancy of error variance isn't true.
```

It's necessary to perform the Hartley test.

```
> hartley <- max(var(y1), var(y2), var(y3))/min(var(y1), var(y2), var(y3))
    [1] 14.65167                  # From Table B.10,
                                  # H(1-0.05; r=3,df=5)=10.8
                                  # H-value > H(1-0.05; r=3,df=5)
```

Constancy of the error variance is violated. Don't apply the ANOVA method!

The Box-Cox procedure is a useful tool. It identifies a transformation from the family of power transformations on Y. The family of power transformations is of the form

$$Y' = Y^{\lambda},$$

where λ is a parameter to be determined from the data.

Define the Box-Cox transformation as below,

```
> box.cox <- function(lambda, x){
              if(lambda !=0)    box.cox <- x^lambda
              if(lambda==0)     box.cox <- log(x)
              box.cox
              }
```

Now search for an appropriate λ.

```
> box.cox(.5, y1)           # lambda=.5, y1
  [1]  2.100000 10.032447  3.801316  6.865129  9.230926

> sd(box.cox(.5, y1))       # check sd for transformed data
  [1] 3.415704
> sd(box.cox(.5, y2))
  [1] 2.985328
> sd(box.cox(.5, y3))        # lambda=.5 is not good.
  [1] 5.062526

> sd(box.cox(0,y1))          # try the log-transform
  [1] 1.319857
> sd(box.cox(0,y2))
  [1] 1.404844
> sd(box.cox(0,y3))          # seems OK.
  [1] 0.9044658
```

Consider the transformed data.

```
> y1 <- box.cox(0,y1)
> y2 <- box.cox(0,y2)
> y3 <- box.cox(0,y3)
```

Perform the Hartley test.

```
> hartley <- max(var(y1), var(y2), var(y3))/min(var(y1), var(y2), var(y3))
   [1]  2.412525            # From Table B.10,
                            # H(1-0.05; r=3,df=5)=10.8
                            # H-value < H(1-0.05; r=3,df=5)
```

Constancy of the error variance is satisfied. Now apply the ANOVA method for the transformed data.

```
> data.trans <- c(y1,y2,y3)
> trans.df <- data.frame(location, data.trans)

> anova <- aov(data.trans~location, trans.df)
> summary(anova)

                Df   Sum Sq   Mean Sq  F value   Pr(>F)
    location     2   11.4522  5.7261   3.7891    0.05302
    Residuals   12   18.1347  1.5112
```

The residual Analysis is needed.

Question: which λ is the best choice?

It's hard to answer such a question. The more transformation you do, the better fitting model you may find. However, you may have difficulty to interpret the transformation, $Y' = Y^{1.003}$, say.

2.4 Statistical strategy

In this chapter we have learnt various tactics

(i) *ANOVA table*

(ii) *Diagnostics:* checking of assumptions: constancy of error variance, normality, et al.

(iii) *Transformation*

A natural question arises: what order should these be done in?

A recommended strategy is

$$Diagnostics \longrightarrow Transformation \longrightarrow ANOVA\ table$$

The SENIC data set `APC1.DAT` page 1365-1366. It is a large data set with 12 variables.

```
>  senic <- read.table("APC1.DAT")
```

In project 18.30, a test of whether or not mean length of stay (variable 2) is the same in the four geographic regions (variable 9) is desired. Before starting, you have to answer the following questions:

Question 1. What is the factor?

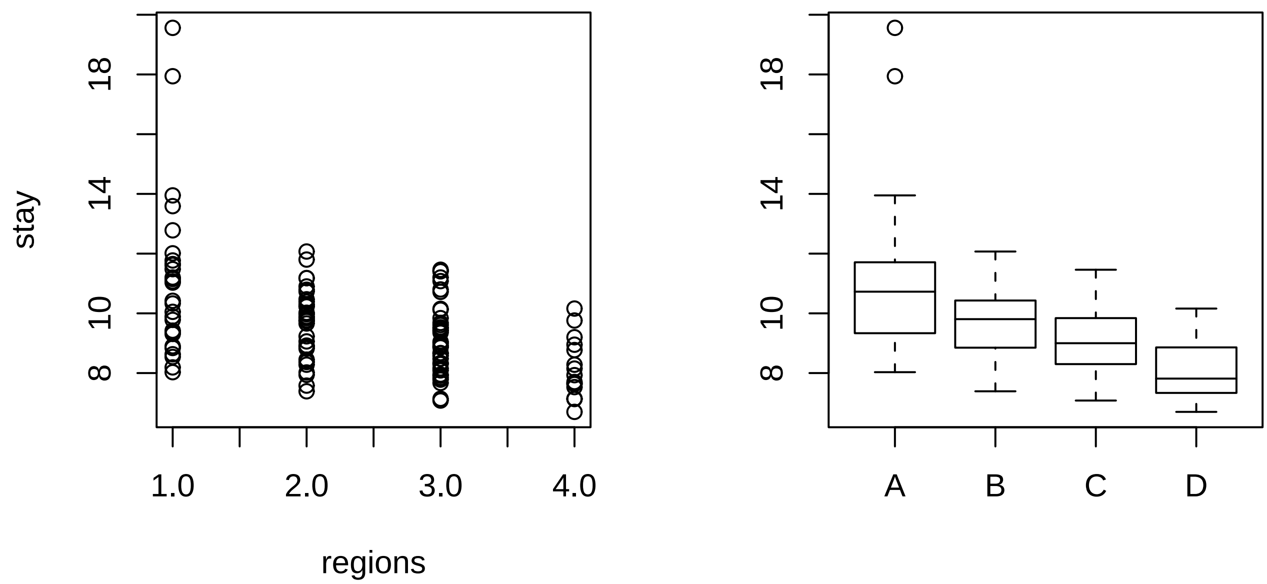
Question 2. What are the factor levels?

Question 3. Is this a balanced or unbalanced case?

```
> y <- cbind(senic[,2], senic[,9])

> y1 <- y[,1][y[,2]==1]
> y2 <- y[,1][y[,2]==2]
> y3 <- y[,1][y[,2]==3]
> y4 <- y[,1][y[,2]==4]
```

Set up the data frame.

```
> stay <- c(y1,y2,y3,y4)       # Is this equivalent to  y[,1]?
> regions <- factor(rep(LETTERS[1:4],c(length(y1),length(y2),length(y3),length(y4))))
> stay.df <- data.frame(regions, stay)
```



Look at the data set graphically.

```
> par(mfrow=c(1,2))
> plot(stay.df)

> plot(regions, stay)
> abline(h=mean(y1),col=1)     # The bar in the middle of a boxplot is the sample median,
> abline(h=mean(y2),col=2)     # not the sample mean
> abline(h=mean(y3),col=3)
> abline(h=mean(y4),col=4)
```

What do you see? What are you going to do next?