Chapter 13

Anova and Elements of Statistical Design

**********anova.tex**********

In chapter 5 we introduced procedures for comparing means of two populations. However many real life experiments involve more than two populations. For example, an experimenter may be interested in comparing the test scores of students taught by several different methods.

At first glance it seems that we can apply t- test on all possible pairs of means. This "solution" would not be good since the probability of the error of first type for such procedure is unduly large.

The appropriate procedure for testing hypotheses of equality of several means is analysis of variances (ANOVA). ANOVA is probably one of the most frequently used statistical procedures and its reasoning is applicable in many other seemingly different problems.

The rationale of ANOVA is very simple. The variability between sample means is compared to the variability within the populations.

13.1 One-way analysis of variance

Suppose we are interested in testing the equality of means $\mu_1, \ldots, \mu_k$ belonging to the populations $\mathcal{P}_1, \ldots, \mathcal{P}_k$. From the ith population $\mathcal{P}_i$ we take a sample

$$y_{i1}, y_{i2}, \ldots, y_{in_i},$$

of size n_i. We model y_{ij} as

$$y_{ij} = \mu_i + \epsilon_{ij}, \ 1 \le i \le k, \ 1 \le j \le n_i. \tag{13.1}$$

The assumptions underlying the one-way ANOVA are the following:
(i) All populations are assumed to be normal;
(ii) The variance in all populations is the same;
(iii) different samples are independent;
That can be expressed by the requirement that all ϵ_{ij} in (13.1) are iid $N(0, \sigma^2)$.

	Population			
	1	2	...	k
Population Means	μ_1	μ_2	...	μ_k
Common Standard Deviation	σ	σ	...	σ

If the sample sizes are the same, i.e. $n_1 = n_2 = \ldots = n_k$, then the ANOVA is called balanced. It is often the case that many experiments are designed as balanced ANOVA. During the treatment it may happen that experimental unit is lost, leading to the unbalanced case.

In terms of the model (13.1) the null hypothesis is

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k, \tag{13.2}$$

And the alternative is

$$H_1 : (H_0)^c \quad (\text{or} \quad \mu_i \neq \mu_j, \text{ for at least one pair } i, j. \tag{13.3}$$

One can parameterize the means μ_i as $\mu_i = \mu + \tau_i$. An alternative (and sometimes preferred) form of the hypothesis H_0 is:

$$H_0 : \tau_1 = \tau_2 = \ldots = \tau_k = 0 \tag{13.4}$$

The alternative is H_1: Not all τ_is are equal to 0, of course.

The parameter μ is called *the grand mean* and $\mu = \frac{1}{k}\Sigma_{i=1}^{k}\mu_i$, and τ_i is the *effect of i-th population (treatment).* We also assume that $\Sigma\tau_i = 0$, for balanced designs. It is needed to ensure uniqueness of the decomposition $\mu_i = \mu + \tau_i$. For unbalanced designs the condition is more complicated.

Fundamental ANOVA Identity

$$\Sigma_{i=1}^{k}\Sigma_{j=1}^{n}(y_{ij} - \bar{y})^2 = n\Sigma_{i=1}^{k}(\bar{y}_i - \bar{y})^2 + \Sigma_{i=1}^{k}\Sigma_{j=1}^{n}(y_{ij} - \bar{y}_i)^2. \tag{13.5}$$

SST = SSTr + SSE.

Why ANOVA when we are dealing with means?

Test the homogeneity of variances first!

There are several procedures that test for fulfillment of the ANOVA condition of homoscedascity. A reasonably sensitive and simple test is Cochran's test applicable if the design is balanced.

Reject hypotheses that k populations have the same mean if $G = \frac{s^2_{max}}{s_1^2 + \ldots + s_k^2}$ is large. The table below gives the critical values for $\alpha = 0.05$ and n sample size in balanced design.

$\nu = n - 1$ k	1	2	3	4	5	6	7	10	36	144
2	0.9985	0.9750	0.9392	0.9057	0.8772	0.8534	0.8332	0.7880	0.6602	0.5813
3	0.9669	0.8709	0.7977	0.7457	0.7071	0.6771	0.6530	0.6025	0.4748	0.4031
4	0.9065	0.7679	0.6841	0.6287	0.5896	0.5598	0.5365	0.4884	0.3720	0.3093

4. **Effects of THC on Activity in Rats.** The nucleus accumbent is a forebrain structure that has been shown to be involved in locomotor activity in rats. Systemic administrations of low doses of tetrahydrocannabinol (THC, the major active ingredient in marijuana) is known to increase locomotor activity, whereas high doses lead to a decrease in activity. In an attempt to examine whether THC is acting within the nucleus accumbens to produce its effects on activity, Conti and Musty (1984) bilaterally injected either a placebo, or 0.1, 0.5, 1, or 2 micrograms (μg) of THC into the nucleus accumbens of rats. The investigators recorded the activity level of animals before and after the injection. Activity was recorded by placing the animal in a test chamber and suspending the chamber on rubber mounts. The vibrations of the chamber as the animal moved around were transduced by an accelerometer and converted to activity units, which were read off a meter. These units, then, were arbitrary, a point that will become relevant when we consider transformations. Conti and Musty took as their dependent variable the rat's activity for 10 minutes after the injection as a proportion of the rat's activity in the 10 minutes before the injection. Since animals generally decrease their activity as they become accustomed to an apparatus, most ratios were less than 1. However, it was expected that those rats with intermediate levels of THC would decrease their activity less (exhibit a higher postinjection ÷ preinjection ratio) than would those with either low or high levels. (Intermediate levels were expected to lead to the greatest activity, because very low doses should be insufficient to produce an effect and high doses should lead to decreases in activity.)

	Control	$0.1\mu g$	$0.5\mu g$	$1\mu g$	$2\mu g$	All Groups
	30	60	71	33	36	
	27	42	50	78	27	
	52	48	38	71	60	
	38	52	59	58	51	
	20	28	65	35	29	
	26	93	58	35	29	
	8	32	74	46	24	
	41	46	67	32	17	
	49	63	61		50	
	49	44			53	
Total	340	508	543	388	381	2160
Mean	34.00	50.80	60.33	48.50	38.10	45.96
S.D.	14.30	18.39	11.07	18.32	14.46	17.62
n	10	10	9	8	10	47

4. **Animal diet.** In 1968, Koch and Sen ('Some aspects of the statistical analysis of the mixed model', *Biometrics*, 24, 27-48), examined the results of experiments undertaken at the Department of Pathology, Duke University Medical Centre, North Carolina by Dr N. Kaufmann and Dr J. V. Klavins. In one of their experiments, 16 animals were randomly placed into one of two groups - an experimental group which received ethionine in their diets, and a control group. The liver of each animal was divided into two parts, one of which was treated with radioactive iron and oxygen, and the other with radioactive iron and nitrogen. The data consist of the amount of iron absorbed by the variously treated liver portions; in the table the treatments are denoted by EO (ethionine-oxygen), EN (ethionine- nitrogen), CO (control-oxygen) and CN (control-nitrogen).

Pair	EO	EN	CO	CN
1	38.43	31.47	36.09	32.53
2	36.09	29.89	34.01	27.73
3	34.49	34.50	36.54	29.51
4	37.44	38.86	39.87	33.03
5	35.53	32.69	33.38	29.88
6	32.35	32.69	36.07	29.29
7	31.54	31.89	35.88	31.53
8	33.37	33.26	34.17	30.16

Analyze these data to see whether there are significant differences in the amounts of iron absorbed in livers subjected to the four treatments.

```
Sol.
 ANALYSIS OF VARIANCE ON C5
 SOURCE     DF        SS        MS        F        P
 C6          3    131.19     43.73     8.62    0.000
 ERROR      28    142.09      5.07
 TOTAL      31    273.28
```

NBA Players. In the 1993 basketball players in the NBA from the four ACC schools were analyzed: Duke, North Carolina, North Carolina State, and Georgia Tech. Below are their 1993 season's PPG (points per game).

Duke	UNC	NCSU	GT
7.5	5.5	16.9	7.9
8.7	6.2	4.5	7.8
7.1	13.0	10.5	14.5
18.2	9.7	4.4	6.1
	12.9	4.6	4.0
	5.9	18.7	14.0
	1.9	8.7	
		15.8	

Formulate the question.

[Ans. $F = 0.41$. H_0 not rejected. Basketball players from these four schools are likely to score equal amounts regardless where they attended college.]

Insects. Some colors are more attractive to insects than others. In an experiment aimed at determining the best color for attracting cereal leaf beetles, six boards in each of four colors were placed in a field of oats in July. The following table gives data on the number of cereal leaf beetles trapped. (Modified from M. C. Wilson and R. E. Shade, "Relative attractiveness of various luminescent colors to the cereal leaf beetle and the meadow spittle-bug," *Journal of Economic Entomology*, 60 (1967, pp. 578-580.)

Board color	Insects trapped					
Lemon Yellow	45	59	48	46	38	47
White	21	12	14	17	13	17
Green	37	32	15	25	39	41
Blue	16	11	20	21	14	7

(a) Make a plot of the counts of insects trapped against board color (space the four colors equally on the horizontal axis). Compute the mean count for each color, add the means to your plot, and connect the means with line segments.

(b) Based on the data, state your conclusions about the attractiveness of these colors to the beetles.

(c) Does it make sense to speak of a positive or negative association between board color and insect count?

```
> yellow_scan()
1: 45 59 48 46 38 47
7:
> white_scan()
1: 21 12 14 17 13 17
7:
> green_scan()
1: 37 32 15 25 39 41
7:
> blue_scan()
1: 16 11 20 21 14 7
7:
> data_cbind(yellow, white, green, blue)
> data
      yellow white green blue
[1,]     45    21    37   16
[2,]     59    12    32   11
[3,]     48    14    15   20
[4,]     46    17    25   21
[5,]     38    13    39   14
[6,]     47    17    41    7

> Anova(data,r=2)

Source  SS      df      MS      F       p
---------------------------------------------------------
Treat   4218.46 3       1406.15 30.55   0
Error   920.5   20      46.02
Total   5138.96 23
---------------------------------------------------------
RootMSE= 6.78    R-squared= 0.82
---------------------------------------------------------
```

Nematodes. Some varieties of nematodes (roundworms that live in the soil and are frequently so small they are invisible to the naked eye) feed on the roots of lawn grasses and crops such as strawberries and tomatoes. This pest, which is particularly troublesome in warm climates, can be treated by the application of nematocides. However, because of size of the worms, it is very difficult to measure the effectiveness of these pesticides directly. To compare four nematocides, the yields of equal-size plots of one variety of tomatoes were collected. The data (yields in pounds per plot) are shown in the table.

Nematocide A	Nematocide B	Nematocide C	Nematocide D
18.6	18.7	19.4	19.0
18.4	19.0	18.9	18.8
18.4	18.9	19.5	18.6
18.5	18.5	19.1	18.7
17.9		18.5	

One-way ANOVA was run and MINITAB output is given below.

```
ANALYSIS OF VARIANCE ON YIELD
SOURCE     DF        SS        MS        F        p
BRAND       3    1.3094    0.4365     5.20    0.013
ERROR      14    1.1750    0.0839
TOTAL      17    2.4844
                                    INDIVIDUAL 95 PCT CI'S FOR MEAN
                                    BASED ON POOLED STDEV
LEVEL      N      MEAN     STDEV  ----+---------+---------+---------+--
    1      5    18.360     0.270   (-------*-------)
    2      4    18.775     0.222              (-------*--------)
    3      5    19.080     0.402                       (-------*-------)
    4      4    18.775     0.171              (-------*--------)
                                  ----+---------+---------+---------+--
POOLED STDEV =    0.290             18.20     18.55     18.90     19.25
```

(a) Write a statistical model for ANOVA and state H_0 and H_1 in terms of your model.
(b) What is your decision if $\alpha = 0.05$?
(c) For what values of α your decision will be different than that in (b)?

Maternal behavior in rats. To investigate maternal behavior of laboratory rats, researchers separated the rat pup from the mother and record the time required for the mother to retrieve the pup. The study was run with 5, 20 and 35 day old pups, six in each group. The pups were moved to a fixed distance from the mother and the time of retrieval (in seconds) was recorded.

5 days	15	10	25	15	20	18
20 days	30	15	20	25	23	20
35 days	40	35	50	43	45	40

Run a One-way ANOVA on the data and state your conclusions.

Density of bricks. An experiment was run to determine whether four specific firing temperatures affect the density of a certain type of brick. The experiment led to the following data.

Temperature	Density					
100	21.8	21.9	21.7	21.7	21.6	21.7
125	21.7	21.4	21.5	21.4		
150	21.9	21.8	21.8	21.8	21.6	21.5
175	21.9	21.7	21.8	21.4		

(a) Does the firing temperature affect the density of the bricks?
(b) Compare the means using Duncan's multiple range test.

Chemist. A chemist wishes to test the effect of four chemical agents on the strength of a particular type of cloth. Because there might be variability from one bolt to another, the chemist decides to use a randomized block design, with the bolts of cloth considered as blocks. She selects five bolts and applies all four chemicals in random order to each bolt. The resulting tensile strengths follow. Analyze the data and draw appropriate conclusions.

	Bolt				
Chemical	1	2	3	4	5
1	73	68	74	71	67
2	73	67	75	72	70
3	75	68	78	73	68
4	73	71	75	75	69

Oscilloscope. An experiment is conducted to study the influence of operating temperature and three types of face-plate glass in the light output of an oscilloscope tube. The following data are collected.

	Temperature		
Glass Type	100	125	150
	580	1090	1392
1	568	1087	1380
	570	1085	1386
	550	1070	1328
2	530	1035	1312
	579	1000	1299
	546	1045	867
3	575	1053	904
	599	1066	889

Fertilizer. It has been shown that the fertilizer magnesium ammonium phosphate, $MgNH_4PO_4$, is an effective supplier of the nutrients necessary for plant growth. The compounds supplied by this fertilizer are highly soluble in water, allowing the fertilizer to be applied directly on the soil surface or mixed with the growth substrate during the potting process. A study on the *"Effect of Magnesium Ammonium Phosphate on Height of Chrysanthemums"* was conducted at George Mason University in 1980 to determine a possible optimum level of fertilization, based on the enhanced vertical growth response of the chrysanthemums. Forty chrysanthemums seedlings were divided into 4 groups each containing 10 plants. Each was planted in a similar pot containing a uniform growth medium. To each group of plants an increasing concentration of $MgNH_4PO_4$, measured in grams per bushel, was added. The 4 groups of plants were grown under uniform conditions in a greenhouse for a period of four weeks. The treatments and the respective changes in heights, measured in centimeters, are shown in the following table:

Treatment			
50 gm/bu	100 gm/bu	200 gm/bu	400 gm/bu
13.2	16.0	7.8	21.0
12.4	12.6	14.4	14.8
12.8	14.8	20.0	19.1
17.2	13.0	15.8	15.8
13.0	14.0	17.0	18.0
14.0	23.6	27.0	26.0
14.2	14.0	19.6	21.1
21.6	17.0	18.0	22.0
15.0	22.2	20.2	25.0
20.0	24.4	23.2	18.2

can we conclude at the 0.05 level of significance that different concentration of $MgNH_4PO_4$, affect the average attained height of chrysanthemums?

Repeated measures design. Seven patients each underwent three different methods of kidney dialysis (Daugridas, 1982). The following values were obtained for weight change in kilograms between dialysis sessions:

Patient	Treatment 1	Treatment 2	Treatment 3
1	2.90	2.97	2.67
2	2.56	2.45	2.62
3	2.88	2.76	1.84
4	2.73	2.20	2.33
5	2.50	2.16	1.27
6	3.18	2.89	2.39
7	2.83	2.87	2.39

(i) Test the null hypothesis that there is no difference in mean weight change among treatments.

13.2 SPLUS program

Explanation of the program and use

Sharp Willies. After loosing the contract with Sherwood Rascals, Sharp Willie, an arrow producing enterprise, has sent a number of employees to four educational institutions for technical training. The company hoped that the training would improve employee productivity and product quality. At the end of program Sharp Willie tested 40 graduates. The scores are:

Program A:	95 88 90 99 89 93 95 97 85 90
Program B:	92 88 80 75 67 78 92 80 77 69
Program C:	85 81 86 91 78 81 86 90 75 83
Program D:	98 65 74 82 90 62 75 85 70 82

Fill-in the ANOVA table and test that all programs are the same at $\alpha = 1\%$.

```
Analysis of variance on score
Source      SS      df        MS        F
program                     382.3
error
total     3367.9
```

Elasticity of Billiard Balls. The billiard balls are made under three different conditions: Ten batches of melted plastic were prepared.(These are 10 blocks of material.) Each batch is divided into 3 equal portions. One portion was chosen at random and set aside as a control. The second portion was chosen at random from the remaining two, and was mixed with additive A. The third portion was mixed with additive B. In this way, the experimenters hoped to balance out any variations in the plastic from batch to batch. Elasticity was measured on a scale from 0 to 100, with higher number representing greater elasticity (desirable property).

Batch	Control	Additive A	Additive B
1	51	75	39
2	45	89	43
3	49	73	51
4	66	84	34
5	53	66	54
6	41	85	43
7	58	73	42
8	56	71	37
9	60	78	37
10	63	65	44

```
Source      DF         SS         MS       F      P
batch        9      82.30       9.14    0.12  0.999
additive     2    5654.87    2827.43   36.62  0.000
Error       18    1389.80      77.21
Total       29    7126.97
```

```
> Anova <-
function(data, factors = "Treat", repeat = "Error", r = 3)
{
      out <- list()
      out$means <- apply(data, 2, mean.na)
      out$data <- data
      out$fit <- matrix(out$means, nrow = dim(data)[1], ncol = dim(data)[2],
              byrow = T, dimnames = dimnames(data))
      out$fit <- out$fit + 0 * data
      out$res <- data - out$fit
      out$grand <- mean.na(data)
      out$ngroup <- apply(!is.na(data), 2, sum)
      out$sstr <- sum.na((out$fit - out$grand)^2)
      out$sse <- sum.na(out$res^2)
      out$sst <- sum.na((data - out$grand)^2)
      out$dftr <- (dim(data)[2] - 1)
      out$dfe <- (sum(!is.na(data)) - dim(data)[2])
      out$dft <- sum(!is.na(data)) - 1
      out$mstr <- out$sstr/out$dftr
      out$mse <- out$sse/out$dfe
      out$F <- out$mstr/out$mse
      out$p <- 1 - pf(out$F, out$dftr, out$dfe)
      cat("\n-------------------------------------------------------\n")
      cat("Source", "SS", "df", "MS", "F", "p", sep = "\t")
      cat("\n-------------------------------------------------------\n")
      cat(factors, round(out$sstr, r), round(out$dftr, r), round(out$mstr, r),
              round(out$F, r), round(out$p, r + 1), "\n", sep = "\t")
      cat(repeat, round(out$sse, r), round(out$dfe, r), round(out$mse, r),
              "\n", sep = "\t")
      cat("Total", round(out$sst, r), round(out$dft, r), "\n", sep = "\t")
      cat("-------------------------------------------------------\n")
      cat("RootMSE=", round(sqrt(out$mse), r), "\t R-squared=", round(out$
              sstr/out$sst, r), "\n")
      cat("-------------------------------------------------------\n")
      invisible(out)
}
```

13.2.1 Scheffe's method for comparison of contrasts

In ANOVA only testable linear combinations of means are contrasts. That means that only hypotheses of the form $H_0 : \Sigma c_i \mu_i = 0$ can be tested, where $\underset{\sim}{c} = (c_1, \ldots, c_a)$ is a vector that satisfies $\Sigma_{i=1}^{a} n_i c_i = 0$, for n_i being the cell sample sizes. If the design is balanced, i.e. $n_1 = n_2 = \ldots = n_a$ then the requirement for $\underset{\sim}{c}$ is $\Sigma c_i = 0$.

For a balanced ANOVA with $a = 5$ the following linear combinations are contrasts.

Hypothesis H_0	Linear Combination	Vector $\underset{\sim}{c}$
$\mu_1 = \mu_3$	$\mu_1 - \mu_3 = 0$	$(1, 0, -1, 0, 0)$
$\mu_1 + \mu_2 = \mu_3 + \mu_4$	$\mu_1 + \mu_2 - \mu_3 - \mu_4 = 0$	$(1, 1, -1, -1, 0)$
$\mu_1 = \frac{\mu_4 + \mu_5}{2}$	$\mu_1 - \frac{1}{2}\mu_4 - \frac{1}{2}\mu_5 = 0$	$(1, 0, 0, -\frac{1}{2}, -\frac{1}{2})$

```
#First Version of Scheffe
> Scheffe
function(data, contrast, alpha = 0.05, r = 3)
{
        means <- apply(data, 2, mean.na)
        fit <- matrix(means, nrow = dim(data)[1], ncol = dim(data)[2], byrow =
                T, dimnames = dimnames(data))
        fit <- fit + 0 * data
        res <- data - fit
        ngroup <- apply(!is.na(data), 2, sum)
        if((contrast %*% ngroup)[1, 1] != 0)
                stop("The contrast is", " untestable for this design.", (
                        contrast %*% ngroup)[1, 1])
        sse <- sum.na(res^2)
        dfe <- (sum(!is.na(data)) - dim(data)[2])
        mse <- sse/dfe
        Cu <- (contrast %*% means)[1, 1]
        SCu <- sqrt(mse * sum(contrast^2/ngroup))
        Salp <- SCu * sqrt((dim(data)[2] - 1) * qf(1 - alpha, dim(data)[2] - 1,
                dfe))
        cat("\n Cu=", round(Cu, r), " SCu=", round(SCu, r), " Salp=", round(
                Salp, r), "\n")
        cat("\n", 100 * (1 - alpha), "% confidence interval is:(", round(Cu -
                Salp, r), ",", round(Cu + Salp, r), ")\n")
        if(abs(Cu) < Salp) {
                cat("\n The contrast is tested to be 0 at the level", alpha,
                        "\n")
        }
        else {
                cat("\n The contrast is tested not to be 0 at the level", alpha,
                        "\n")
        }
}
>
```

Notation in the S-plus program:

$Cu = c_1\bar{y}_1 + \Phi c_2\bar{y}_2 + \ldots + c_a\bar{y}_a$.

$SCu = \sqrt{MSE\Sigma_{i=1}^{a}(c_i^2/n_i)}$, where n_i are cell sample sizes.

$Salp = SCu\sqrt{(a-1)F_{a-1,N-a,\alpha}}$

The Tensile Strength. The tensile strength of synthetic fiber used to make cloth for men's shirts is of interest to manufacturer. It is suspected that the strength is affected by the percentage of cotton in the fiber. Five levels of cotton percentage are of interest, 15 percent, 20 percent, 25 percent, 30 percent, and 35 percent. Five observations are taken at each level of cotton percentage. The table below describes measurements.

	15 %	20 %	25 %	30 %	35 %
obs 1	7	12	14	19	7
obs 2	7	17	18	25	10
obs 3	15	12	18	22	11
obs 4	11	18	19	19	15
obs 5	9	18	19	23	11

The ANOVA for the problem is

```
-------------------------------------------------------
Source  SS      df      MS      F       p
-------------------------------------------------------
Treat   475.76  4       118.94  14.757  0
Error   161.2   20      8.06
```

```
Total   636.96  24
-------------------------------------------------------
RootMSE= 2.839   R-squared= 0.747
-------------------------------------------------------
```

Test the hypothesis $H_0 : \mu_1 + \mu_3 = \mu_4 + \mu_5$. at the level $\alpha = 0.01$.

```
> data
      15 % 20 % 25 % 30 % 35 %
obs 1    7   12   14   19    7
obs 2    7   17   18   25   10
obs 3   15   12   18   22   11
obs 4   11   18   19   19   15
obs 5    9   18   19   23   11

> Scheffe(data, c(1,0,1,-1,-1), alpha=0.01)

 Cu= -5  SCu= 2.539  Salp= 10.69

 99 % confidence interval is:( -15.69 , 5.69 )

 The contrast is tested to be 0 at the level 0.01
```

1. Beetles. The following data were extracted from the more extensive study by Sokal and Karten[1] The data represent mean dry weights (in *mg*) of three genotypes of beetles, *Tribolium castaneum*, reared at density of 20 beetles per gram of flour. The four series of experiments represent replications.

	Genotypes		
Series	++	+*b*	*bb*
1	0.958	0.986	0.925
2	0.971	1.051	0.952
3	0.927	0.891	0.829
4	0.971	1.010	0.955

1. Test whether the genotypes differ in mean dry weight. Take $\alpha = 0.01$.
   The following partial ANOVA table may be useful for the test.

```
-------------------------------------------------------
Source  SS      df      MS      F       p
-------------------------------------------------------
Geno    0.01    2       _____   ____  _____
Error   0.026   9       _____
Total   0.035   11
-------------------------------------------------------
```

2. What assumptions are needed for ANOVA?

Beetles. The following data were extracted from the more extensive study by Sokal and Karten[2] The data represent mean dry weights (in *mg*) of three genotypes of beetles, *Tribolium castaneum*, reared at density of 20 beetles per gram of flour. The four series of experiments represent replications.

	Genotypes		
Series	++	+*b*	*bb*
1	0.958	0.986	0.925
2	0.971	1.051	0.952
3	0.927	0.891	0.829
4	0.971	1.010	0.955

[1]Sokal, R. and Karten, I. (1964). Competition among genotypes in *Tribolium castaneum* at varying densities and gene frequencies (the black locus). *Genetics,* **49** *195-211.*

[2]Sokal, R. and Karten, I. (1964). Competition among genotypes in *Tribolium castaneum* at varying densities and gene frequencies (the black locus). *Genetics,* **49** *195-211.*

1. Test whether the genotypes differ in mean dry weight. Take $\alpha = 0.01$.
   The following partial ANOVA table may be useful for the test.

Source	SS	df	MS	F	p
Geno	0.01	2	-----	----	-----
Error	0.026	9	-----		
Total	0.035	11			

2. What assumptions are needed for ANOVA?

Bees. The data for this problem are taken from Park (1932)[3] who investigated changes in the concentration of nectar in the honey sac of the bee. Syrup of approximately 40% concentration was fed to the bees. The concentration in their honey sacs was determined upon their arrival at the hive. The decreases recorded in the table are classified according to date, both day (September, 1931) and time of day being differentiated. The question to be answered is this: Were significant differences introduced by changes in the time of gathering the data, or may the six groups be considered random samples from a homogeneous population?

3 10:20	3 11:10	3 2:20	10 4:00	11 1:10	12 10:30
1.1	1.0	0.6	-1.6	1.1	2.5
1.0	0.6	0.3	0.8	0.5	0.6
0.9	1.0	-0.1	2.1	2.2	1.1
1.1	0.4	0.0	1.1	1.1	0.6
0.9	0.4	1.5	0.6	0.4	1.8
1.1	0.9	0.9	0.6	-2.0	0.6
0.6	0.6	0.3	0.6	1.4	1.2
0.5	0.4	0.2	0.2	-0.4	1.2
0.5	1.1	0.4	0.8	2.4	0.4
0.7	0.7	0.4	0.6	0.0	1.0

```
 ANALYSIS OF VARIANCE ON syrup
SOURCE     DF        SS        MS        F        p
time        5     2.538     0.508     0.93    0.470
ERROR      54    29.515     0.547
TOTAL      59    32.052
                                   INDIVIDUAL 95 PCT CI'S FOR MEAN
                                  BASED ON POOLED STDEV
 LEVEL      N      MEAN     STDEV  -+---------+---------+---------+-----
     1     10    0.8400    0.2459          (---------*--------)
     2     10    0.7100    0.2726        (--------*---------)
     3     10    0.4500    0.4649   (--------*--------)
     4     10    0.5800    0.9175     (---------*--------)
     5     10    0.6700    1.2936       (--------*---------)
     6     10    1.1000    0.6429                (--------*--------)
                                  -+---------+---------+---------+-----
POOLED STDEV =   0.7393           0.00      0.50      1.00      1.50
```

Clover Varieties. Six plots each of five varieties of Clover were planted at the Danbury Experiment Station in North Carolina. Yields in tons per acre were as follows:

[3]Park, W. (1932). Studies on the change in nectar concentration produced by the honeybee, *Apis mellifera.* Part I: Changes that occur between the flower and the hive. Iowa Agricultural Experiment Station Research Bulletin No 151, 1932.

Variety	*Yield*
Spanish	2.79, 2.26, 3.09, 3.01, 2.56, 2.82
Evergreen	1.93, 2.07, 2.45, 2.20, 1.86, 2.44
Commercial Yellow	2.76, 2.34, 1.87, 2.55, 2.80, 2.21
Madrid	2.31, 2.30, 2.49, 2.26, 2.69, 2.17
Wisconsin A46	2.39, 2.05, 2.68, 2.96, 3.04, 2.60

The following MINITAB output is obtained:

```
MTB > oneway c1 c2;
SUBC>   fisher.

ANALYSIS OF VARIANCE ON C1
SOURCE     DF        SS        MS        F        p
C2          4    1.2784    0.3196     3.53    0.020
ERROR      25    2.2619    0.0905
TOTAL      29    3.5403

                                   BASED ON POOLED STDEV
 LEVEL      N      MEAN     STDEV  -------+---------+---------+---------
     1      6    2.7550    0.3052                     (--------*-------)
     2      6    2.1583    0.2510   (-------*-------)
     3      6    2.4217    0.3549          (--------*-------)
     4      6    2.3700    0.1884         (-------*-------)
     5      6    2.6200    0.3671                 (-------*--------)
                                   -------+---------+---------+---------
POOLED STDEV =   0.3008                 2.10      2.40      2.70

FISHER'S multiple comparison procedure

        Nominal level = 0.0500
    Family error rate = 0.268
Individual error rate = 0.0500

Critical value = 2.060

Intervals for (mean of column group) - (mean of row group)

             1        2        3        4

    2   0.2389
        0.9544

    3  -0.0244  -0.6211
        0.6911   0.0944

    4   0.0273  -0.5694  -0.3061
        0.7427   0.1461   0.4094

    5  -0.2227  -0.8194  -0.5561  -0.6077
        0.4927  -0.1039   0.1594   0.1077
```

(i) Test the hypothesis that the mean yields for the five clover varieties are the same. Take $\alpha = 5\%$. What happens if your α is 1%.

(ii) Which means are different at 5% level?

(iii) Is the hypothesis $H_0 : 3(\mu_1 + \mu_5) = 2(\mu_2 + \mu_3 + \mu_4)$ a contrast? Why? If yes, test it against the two sided alternative, at $\alpha = 5\%$ level.

Promiscuity at Duke. Many students criticize the Greek fraternity and sorority systems for their loose sexual

attitudes, so Katie, Joshua, and Anna[4] decided to conduct a study that would compare the levels of promiscuity among Greek and non-Greek men and women. Based on their experiences at Duke University so far, Katie, Joshua, and Anna agreed with the prevalent view that Greek students are generally more promiscuous than non-Greeks (independents), and that men are generally more promiscuous than women in this campus. In order to conduct a statistical analysis, they defined "promiscuity" as the number of different people that particular subject at least kissed on the lips this semester. The data (given in the table) have been tested by the two-way ANOVA procedure.

	men	women
greek	1 1 2 2 6	1 1 0 3 2
	4 1 1 1 2	1 1 2 1 1
	7 8 3 4 1	2 0 7 2 1
	1 4 3 3 4	4 5 1 2 3
independent	2 2 1 1 1	1 1 1 0 5
	1 1 5 8 2	1 3 2 5 2
	1 8 6 2 2	2 1 2 4 1
	6 0 0 0 1	1 0 0 1 1

```
MTB > anova prom = greek gender greek*gender

Factor     Type Levels Values
greek     fixed      2     1     2
gender    fixed      2     1     2


Analysis of Variance for prom

Source          DF         SS         MS       F
greek            1      2.812      2.812   --------
gender           1     15.313     15.313   --------
greek*gender     1      0.112      0.112   --------
Error           76    310.150      4.081
Total           79    328.388
```

- (i) Explain what type of statistical analysis the above table refers to.
- (ii) What can you say about the effect of interaction `greek*gender`.
- (iii) Test for significance of factor `greek`. Use $\alpha = 0.1$.
- (iv) Test for significance of factor `gender`. Use $\alpha = 0.1$.
- (v) Explain in words your findings.

Beautify me!.

```
Marketing research contractors.
A marketing research consultant evaluated the effects
of fee schedule (Factor A), scope of work (Factor B), and
type of supervisory control (Factor C) on the quality of
work performed under contract by independent marketing
research agencies. The factor levels in the study were as
follows.
A fee level  i=1 high
             i=2 average
             i=3 low
B scope      j=1 all contract work performed in the house
             j=2 some work subcontracted out
C supervis   k=1 local supervisors
             k=2 traveling supervisors only
The quality of the work performed was measured by an index
taking into account several characteristics of the quality.
Data:
```

[4]Katie Anderson, Joshua Smith, and Anna Wulfsberg: Promiscuity at Duke, STA110E Project, Fall 1995.

```
#-------------------------------------------------------
#          k=1              k=2
#      j=1     j=2       j=1   j=2
#----------------------------------
 MTB > set c1
 DATA> 124.3  115.1     112.7  88.2
 DATA> 120.6  119.9     110.2  96.0
 DATA> 120.7  115.4     113.5  96.4
 DATA> 122.6  117.3     108.6  90.1
 DATA>
 DATA> 119.3  117.2     113.6  92.7
 DATA> 118.9  114.4     109.1  91.1
 DATA> 125.3  113.4     108.9  90.7
 DATA> 121.4  120.0     112.3  87.9
 DATA>
 DATA>  90.9   89.9      78.6  58.6
 DATA>  95.3   83.0      80.6  63.5
 DATA>  88.8   86.5      83.5  59.8
 DATA>  92.0   82.7      77.1  62.3
 DATA> end
 MTB > set c2
 DATA> 1 1  1 1
 DATA> 1 1  1 1
 DATA> 1 1  1 1
 DATA> 1 1  1 1
 DATA>
 DATA> 2 2  2 2
 DATA> 2 2  2 2
 DATA> 2 2  2 2
 DATA> 2 2  2 2
 DATA>
 DATA> 3 3  3 3
 DATA> 3 3  3 3
 DATA> 3 3  3 3
 DATA> 3 3  3 3
 DATA> end
 MTB > set c3
 DATA> 1 2  1 2
 DATA> 1 2  1 2
 DATA> 1 2  1 2
 DATA> 1 2  1 2
 DATA>
 DATA> 1 2  1 2
 DATA> 1 2  1 2
 DATA> 1 2  1 2
 DATA> 1 2  1 2
 DATA>
 DATA> 1 2  1 2
 DATA> 1 2  1 2
 DATA> 1 2  1 2
 DATA> 1 2  1 2
 DATA> end
 MTB > set c4
 DATA> 1 1  2 2
 DATA> 1 1  2 2
 DATA> 1 1  2 2
 DATA> 1 1  2 2
 DATA>
 DATA> 1 1  2 2
```

```
DATA> 1 1  2 2
DATA> 1 1  2 2
DATA> 1 1  2 2
DATA>
DATA> 1 1  2 2
DATA> 1 1  2 2
DATA> 1 1  2 2
DATA> 1 1  2 2
DATA> end
MTB > names c1 'qual' c2 'fee' c3 'scope' c4 'suprv'
MTB >  anova  qual = fee|scope|suprv

Factor     Type Levels Values
fee       fixed      3     1     2     3
scope     fixed      2     1     2
suprv     fixed      2     1     2


Analysis of Variance for qual

Source             DF         SS         MS       F      P
fee                 2    10044.3     5022.1  679.34  0.000
scope               1     1834.0     1834.0  248.08  0.000
suprv               1     3832.4     3832.4  518.40  0.000
fee*scope           2        1.6        0.8    0.11  0.898
fee*suprv           2        0.8        0.4    0.05  0.948
scope*suprv         1      574.8      574.8   77.75  0.000
fee*scope*suprv     2        3.9        2.0    0.27  0.767
Error              36      266.1        7.4
Total              47    16557.9


MTB >  anova  qual = fee scope suprv scope*suprv

Factor     Type Levels Values
fee       fixed      3     1     2     3
scope     fixed      2     1     2
suprv     fixed      2     1     2


Analysis of Variance for qual

Source         DF         SS         MS       F      P
fee             2    10044.3     5022.1  774.14  0.000
scope           1     1834.0     1834.0  282.70  0.000
suprv           1     3832.4     3832.4  590.75  0.000
scope*suprv     1      574.8      574.8   88.60  0.000
Error          42      272.5        6.5
Total          47    16557.9

#-----------------------------------------
```

SINGERS

`singers.dat` file contains heights of singers in New Your Choral Society.

```
attach("/daub4/local/brani/pub/datasets/.Data")
```

```
#---------------------------------
> !ls
93cars.dat       attachme         bass2.dat        fruitfly.dat     televisions.doc
93cars.doc       basepath.dat     blank.doc        fruitfly.doc     tenor1.dat
airport.dat      basepath.info    cigarettes.dat   singers.dat      tenor2.dat
airport.doc      basketball.dat   cigarettes.doc   sop1.dat         ushighway1.dat
alto1.dat        basketball.doc   fishcatch.dat    sop2.dat         ushighway1.doc
alto2.dat        bass1.dat        fishcatch.doc    televisions.dat  ushighway2.dat

#-----------------
>  sop1_scan("sop1.dat")
>  sop1
 [1] 64 62 66 65 60 61 65 66 65 63 67 65 62 65 68 65 63 65 62 65 66 62 65 63 65
[26] 66 65 62 65 66 65 61 65 66 65 62

#-----------------
> sop2_scan("sop2.dat")
> sop2
 [1] 63 67 60 67 66 62 65 62 61 62 66 60 65 65 61 64 68 64 63 62 64 62 64 65 60
[26] 65 70 63 67 66

#-----------------
> alto1_scan("alto1.dat")
> alto1
 [1] 65 62 68 67 67 63 67 66 63 72 62 61 66 64 60 61 66 66 66 62 70 65 64 63 65
[26] 69 61 66 65 61
#-----------------

>  alto2_scan("alto2.dat")
>  alto2
 [1] 70 65 65 65 64 66 64 70 63 70 64 63 67 65 63 66 66 64 64 70 70 66 66 66 69
[26] 67 65

#-----------------
> tenor1_scan("tenor1.dat")
> tenor1
 [1] 69 72 71 66 76 74 71 66 68 67 70 65 72 70 68 64 73 66 68 67 64 63 64 67 66
[26] 68

#-----------------
> tenor2_scan("tenor2.dat")
> tenor2
 [1] 68 73 69 71 69 76 71 69 71 66 69 71 71 71 69 70 69 68 70 68 69
#-----------------

>  bass1_scan("bass1.dat")
> bass1
 [1] 72 70 72 69 73 71 72 68 68 71 66 68 71 73 73 70 68 70 75 68 71
#-----------------

>  bass2_scan("bass2.dat")
>  bass2
 [1] 72 75 67 75 74 72 72 74 72 72 74 70 66 68 75 68 70 72 67 70 70
```

Descriptive Statistics

```
> summary(sop1)
 Min. 1st Qu. Median  Mean 3rd Qu. Max.
```

```
   60   62.75     65 64.25      65   68
>  summary(sop1)[4]-mean(sop1)
 Mean
    0

> Eda.shape
function(x)
{
        par(mfrow = c(2, 2))
        hist(x)
        boxplot(x)
        iqd <- summary(x)[5] - summary(x)[2]
        plot(density(x, width = 2 * iqd), xlab = "x", ylab = "", type = "l")
        qqnorm(x)
        qqline()
}

> Eda.shape(sop1)
> #Figure 1.


> HH_c(sop1, sop2, alto1, alto2, tenor1, tenor2, bass1, bass2)
> voices_rep(c("S1", "S2", "A1", "A2", "T1", "T2", "B1", "B2"),
+ c(length(sop1),length(sop2), length(alto1), length(alto2), length(tenor1)
+ , length(tenor2), length(bass1), length(bass2)) )
> voices_factor(voices)
> # making data frame
> HH.df_data.frame(voices, HH)
> HH.df
   voices HH
 1     S1 64
 2     S1 62
 3     S1 66
 .
 .
 .
209     B2 72
210     B2 67
211     B2 70
212     B2 70

> length(c(sop1, sop2, alto1, alto2))
[1] 123
> length(HH)
[1] 212
> gen_rep(c("F","M"),c(123, 212-123))
> gen_factor(gen)
> GEN.df_data.frame(gen, HH)

> par(mfrow=c(2,2))
> plot.design(GEN.df)
> plot.factor(GEN.df)
> plot.design(HH.df)
> plot.factor(HH.df)
> #Figure 2.
```

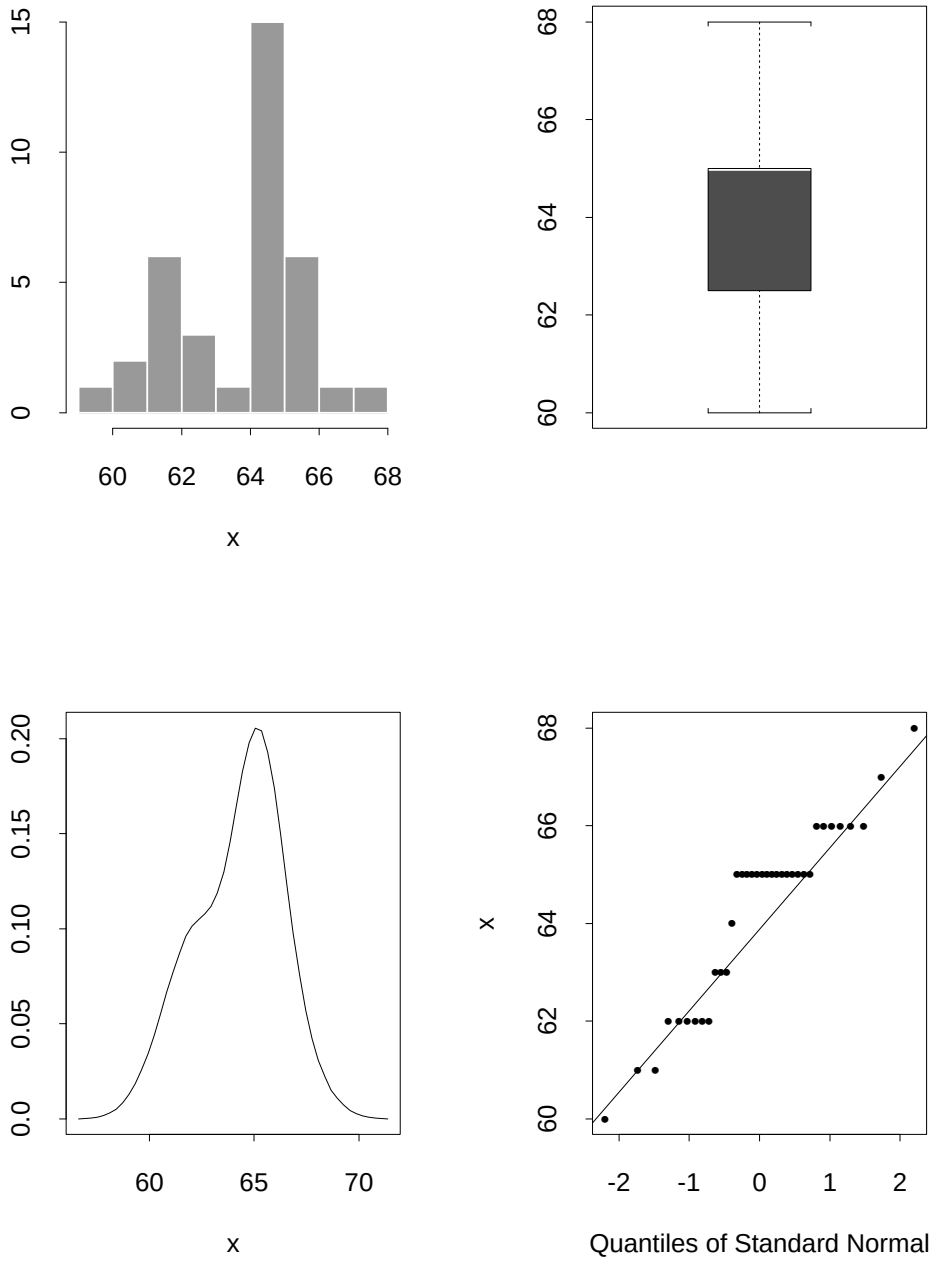

Figure 13.1: Eda.shape for `sop1` data set

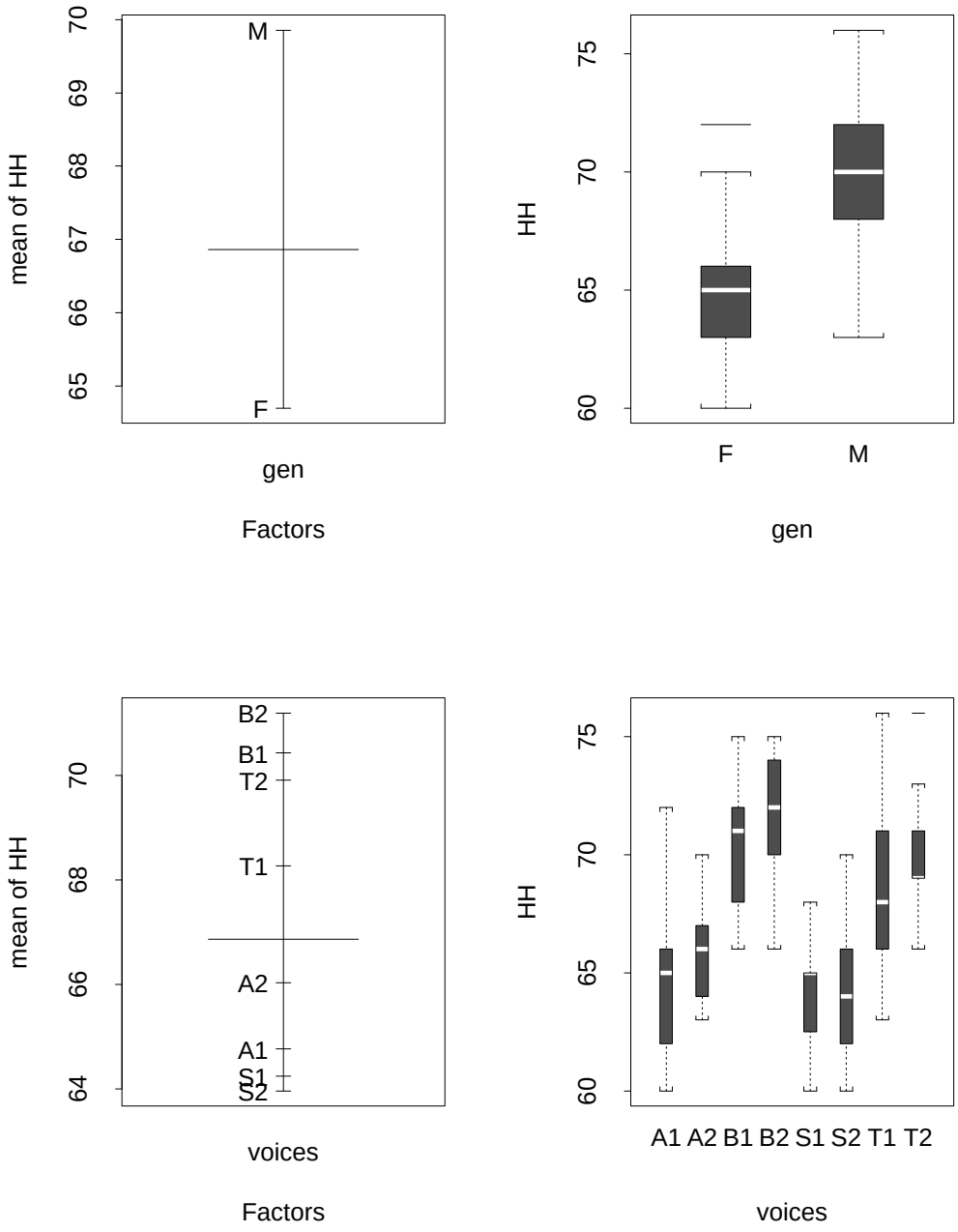

Figure 13.2: Design and Factorplots for GEN and HH data frames

Inference

```
> t.test(c(sop1, sop2, alto1, alto2),
+        c(tenor1, tenor2, bass1, bass2), alternative="less")

         Standard Two-Sample t-Test

data:  c(sop1, sop2, alto1, alto2) and c(tenor1, tenor2, bass1, bass2)
t = -13.8275, df = 210, p-value = 0
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
        NA -4.538845
sample estimates:
 mean of x mean of y
  64.69919  69.85393
```

Model: $y_{ij} = \mu_i + \epsilon_{ij} = \mu + \tau_i + \epsilon_{ij}, j = 1, \ldots, J_i;\ i = 1, \ldots, I.$

```
>  aov.HH_aov(HH ~ voices, HH.df)
>  summary(aov.HH)
           Df Sum of Sq  Mean Sq  F Value Pr(F)
voices      7  1553.681 221.9544 34.16352     0
Residuals 204  1325.352   6.4968

> fitted.values(aov.HH)
     1     2     3     4     5     6     7     8     9    10    11    12    13
 64.25 64.25 64.25 64.25 64.25 64.25 64.25 64.25 64.25 64.25 64.25 64.25 64.25
 .
 .
 .
  71.19048 71.19048 71.19048 71.19048 71.19048 71.19048 71.19048 71.19048
      207      208      209      210      211      212
 71.19048 71.19048 71.19048 71.19048 71.19048 71.19048
> Eda.shape( resid(aov.HH))
> #Figure 3.


> coef(aov.HH)
 (Intercept)   voices1  voices2  voices3    voices4    voices5   voices6
    67.35168 0.6351852 1.675573 1.028263 -0.7711376 -0.5613139 0.2137135
   voices7
 0.3647265
> contrasts(HH.df$voices)
   [,1] [,2] [,3] [,4] [,5] [,6] [,7]
A1   -1   -1   -1   -1   -1   -1   -1
A2    1   -1   -1   -1   -1   -1   -1
B1    0    2   -1   -1   -1   -1   -1
B2    0    0    3   -1   -1   -1   -1
S1    0    0    0    4   -1   -1   -1
S2    0    0    0    0    5   -1   -1
T1    0    0    0    0    0    6   -1
T2    0    0    0    0    0    0    7
>  as.vector(coef(aov.HH)) %*% t(as.matrix(cbind(rep(1,8),contrasts(HH.df$voices))))
           A1       A2       B1       B2    S1       S2       T1       T2
[1,] 64.76667 66.03704 70.42857 71.19048 64.25 63.96667 68.26923 69.90476
```

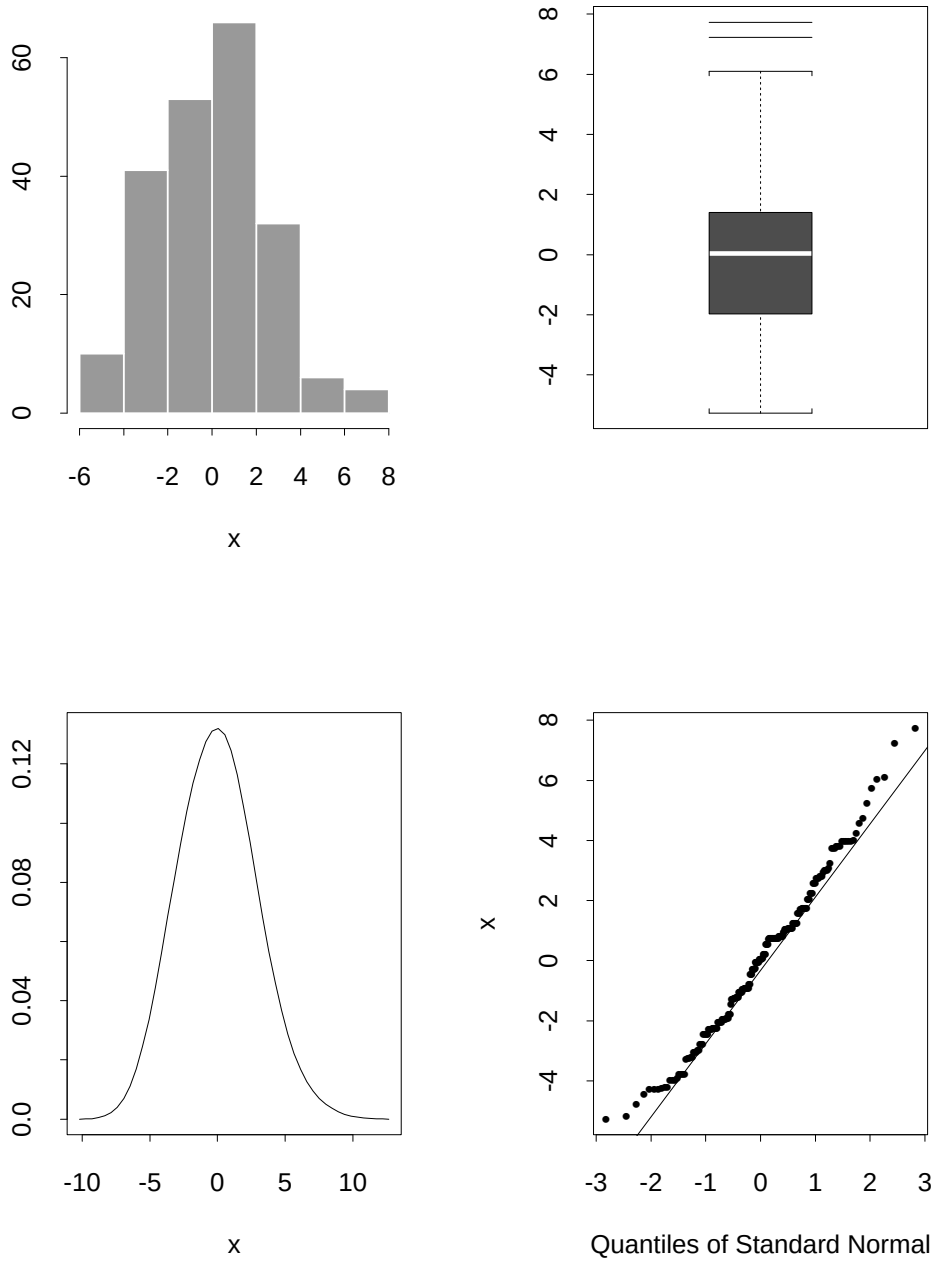

Figure 13.3: Eda.shape for the residuals

Pairwise Comparisons

```
> Scheffe
function(data, contrast, alpha = 0.05, r = 3)
{
        means <- apply(data, 2, mean.na)
        cat("\n means:", means, "\n")
        fit <- matrix(means, nrow = dim(data)[1], ncol = dim(data)[2], byrow =
                T, dimnames = dimnames(data))
        fit <- fit + 0 * data
        res <- data - fit
        ngroup <- apply(!is.na(data), 2, sum)
        if((contrast %*% ngroup)[1, 1] != 0)
                stop("The contrast is", " untestable for this design.", (
                        contrast %*% ngroup)[1, 1])
        cat("\n ngroup:", ngroup, "\n")
        sse <- sum.na(res^2)
        dfe <- (sum(!is.na(data)) - dim(data)[2])
        mse <- sse/dfe
        C <- (contrast %*% means)[1, 1] #dim(C) <- NULL
        cat("\n C=", round(C, r), "\n")
        SC <- sqrt(mse * sum(contrast^2/ngroup))
        Salp <- SC * sqrt((dim(data)[2] - 1) * qf(1 - alpha, dim(data)[2] - 1,
                dfe))
        cat("\n C=", round(C, r), " SC=", round(SC, r), " Salp=", round(Salp, r
                ), "\n")
        cat("\n", 1 - alpha, "*100% confidence interval is:(", round(C - Salp,
                r), ",", round(C + Salp, r), ")\n")
        if(abs(C) < Salp) {
                cat("\n The contrast is tested to be 0 at the level", alpha,
                        "\n")
        }
        else {
                cat("\n The contrast is tested not to be 0 at the level", alpha,
                        "\n")
        }
}


> cont_c(1/36, 1/30, -1/30, -1/27, -1/26, -1/21, 1/21, 1/21)
> Scheffe(data, cont)

 means: 64.25 63.9666666666667 64.7666666666667 66.037037037037 68.2692307692308
        69.9047619047619 70.4285714285714 71.1904761904762

 ngroup: 36 30 30 27 26 21 21 21

 C= 0.101

 C= 0.101  SC= 0.059  Salp= 0.222

 0.95 *100% confidence interval is:( -0.12 , 0.323 )

 The contrast is tested to be 0 at the level 0.05
```

$\Gamma = c_1\mu_1 + \ldots + c_I\mu_I$
$\Sigma c_i n_i = 0,\ \Sigma n_i = N.$
$C = c_1\bar{y}_1 + \ldots + c_I\bar{y}_I.$

$SC = \sqrt{MSE\Sigma_{i=1}^{I}(c_i^2/n_i)}$
$S_\alpha = SC\sqrt{(I-1)F_{\alpha,I-1,N-I}}$
Conf. Int: $[C - S_{alpha}, C + S_\alpha]$

Singers in the `trellis` library

The Trellis library is a collection of functions and datasets for creating Trellis displays which have multiple panels arranged in a regular grid-like structure. Each panel graphs a subset of the data. All panels in a Trellis display contain the same type of graph. Graph types include histograms, scatter plots, dot plots, contour plots, wireframe plots and 3-d point clouds. The data subsets for each panel are chosen in a regular manner conditioning on continuous or discrete variables in the data, thus providing a coordinated series of views of high dimensional data. The Trellis functions include control over axes and aspect ratio and contain *banking* computations that let the data select the aspect ratio.

To access the functions in the library correctly you must attach it with the `first=T` argument to library set, i.e.

```
> library(trellis,first=T)
> library(help="trellis")
```

After attaching the library, start a graphics device using the trellis.device function e.g.

```
> trellis.device(motif)
```

This will set appropriate graphics parameters for the specified device.

There is a collection of example functions in the library that illustrate its capabilities. See the trellis.examples help file.

`singer`: Data frame giving the heights of singers in the New York Choral Society. Components are named height (inches) and voice.part.

```
> singer
   height voice.part
 1     64  Soprano 1
 2     62  Soprano 1
 3     66  Soprano 1
 4     65  Soprano 1
 .
 .
 .
232     72     Bass 2
233     71     Bass 2
234     74     Bass 2
235     75     Bass 2

> boxplot(height ~ voice.part, data=singer, xlab="Height (inches)" )
> #Figure 4.


>  qqmath(height ~ qnorm | voice.part, data=singer)
>  qqmath(log(height)  ~ qnorm | voice.part, data=singer)
>  #Figure 5.


> qqmath(height ~ qnorm | voice.part, data=singer, aspect=1, layout=c(4,2), strip.name=FALSE,
       prepanel = prepanel.qqmathline, panel=function(x,y){
                                                 panel.grid()
                                                 panel.qqmathline(y)
                                                 panel.qqmath(x,y)
                                                 },
       xlab = "Unit Normal Quantile"
       )
> #Figure 6.
```

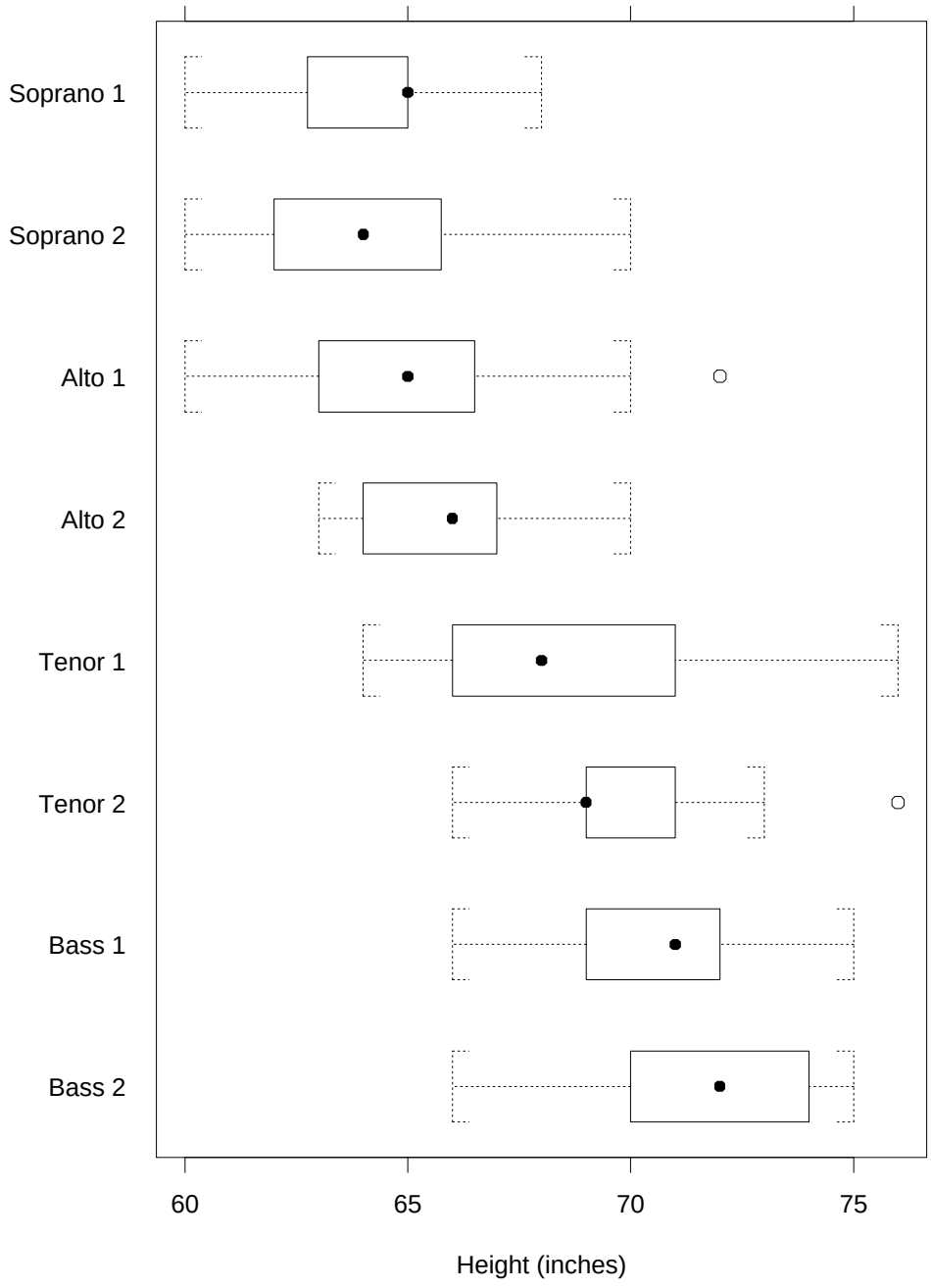

Figure 13.4: A boxplot showing heights of members of a choral group, arranged according to the part they sing.

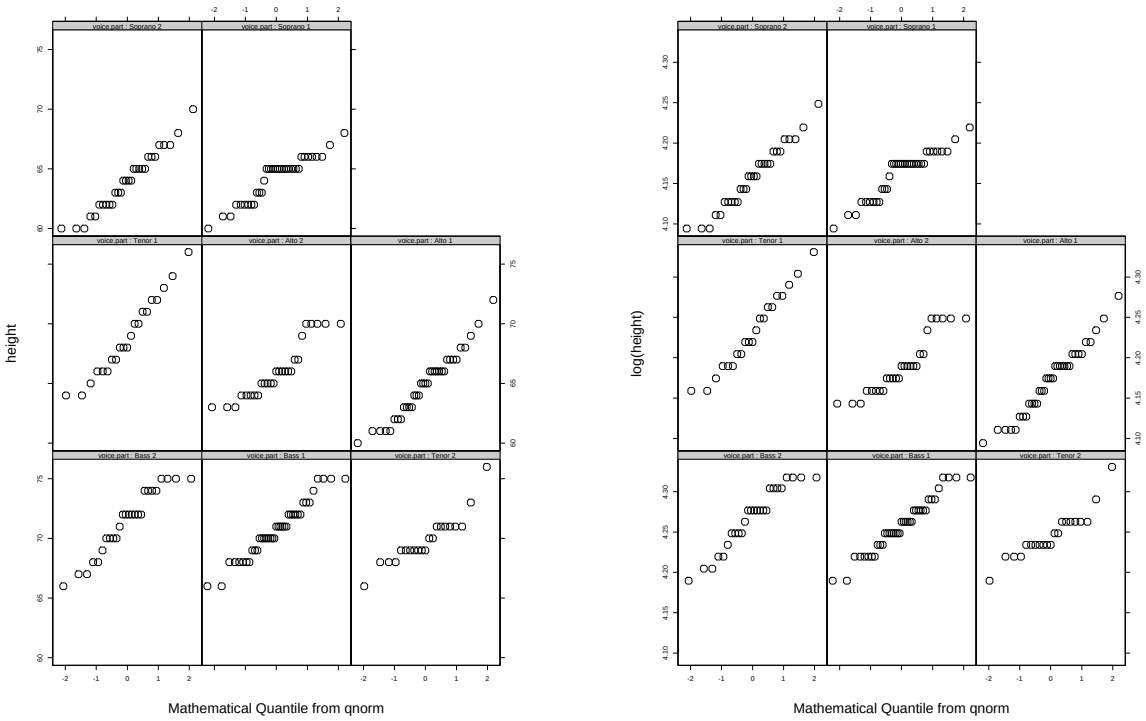

Figure 13.5: Mathematical quantiles: responses height (laft panel) and Log(height) (right panel).

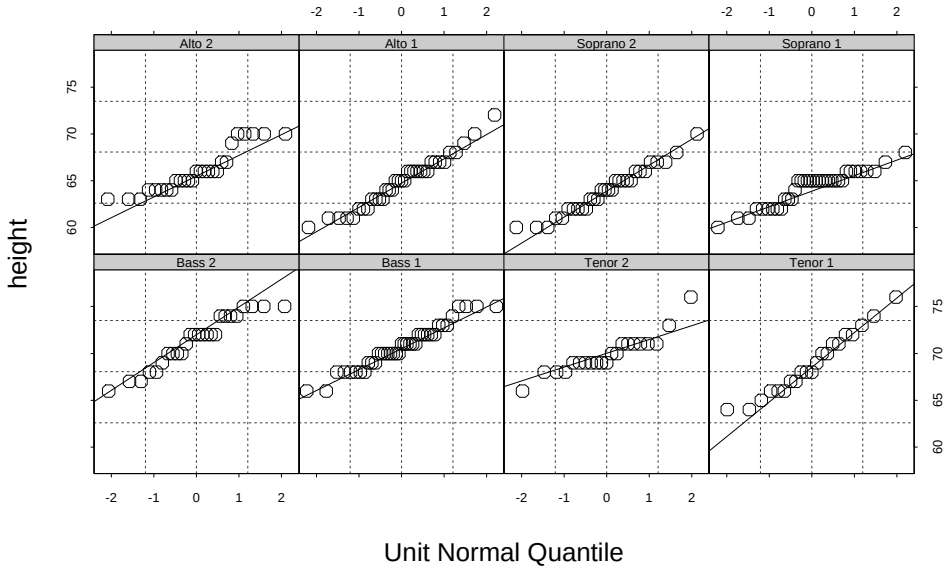

Figure 13.6: Improved display of `qqmath`.