Experimental Design and Biostatistics (TAMS38)
Lecture 2 – One-Way Analysis

Martin Singull

Department of Mathematics
Mathematical Statistics
Linköping University, Sweden

LINKÖPING UNIVERSITY

Content

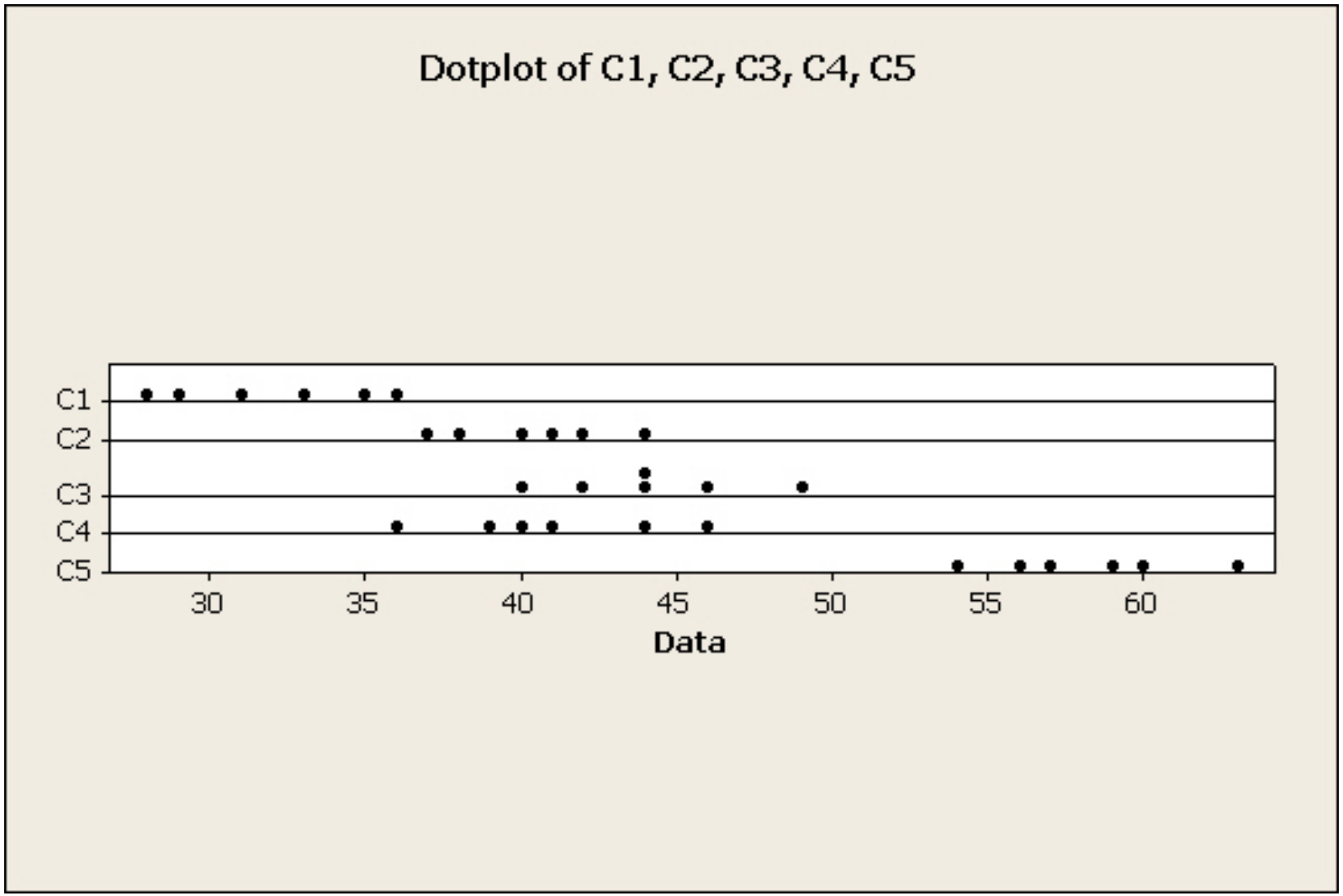
Example 1 – Strontium in watercourses

We have measured the amount strontium in five different watercourses. Resultat (in: mg/ml):

Grayson's Pond:	28.2	33.2	36.4	34.6	29.1	31.0;	$\bar{x}_1 = 32.1$
Beaver Lake:	39.6	40.8	37.9	37.1	43.6	42.4;	$\bar{x}_2 = 40.2$
Angler's Cove:	46.3	42.1	43.5	48.8	43.7	40.1;	$\bar{x}_3 = 44.1$
Appletree Lake:	41.0	44.1	46.4	40.2	38.6	36.3;	$\bar{x}_4 = 41.1$
Rock River:	56.3	54.1	59.4	62.7	60.0	57.3;	$\bar{x}_5 = 58.3$

- Are there any difference between the watercourses?
- Statistical model?
- How can we analyze the data?

Plot of the data.



One-Way Classification

In a design with complete randomization we want to compare the effect of a treatments.

For treatment i we have observation y_{il}, i.e.,

	Observations
Treatment 1:	$y_{11},\ y_{12}, \ldots, y_{1n_1}$
$\vdots$	$\vdots$
Treatment a:	$y_{a1},\ y_{a2}, \ldots, y_{an_a}$

If $n_1 = n_2 = \ldots = n_a$ we say that the design is **balanced**.

Model:

$$Y_{il} = \mu + \tau_i + \varepsilon_{il},$$

where $\varepsilon_{il} \sim N(0,\ \sigma)$, for $l = 1, \ldots, n_i$ and $i = 1, \ldots, a$.
Furthermore, the random variables ε_{il} are independent.

Case 1: $\tau_1, \ldots, \tau_a$ are fix (not random). Hence, we have $\mu_i = \mu + \tau_i$ for each treatment i.

Case 2: $\tau_1, \ldots, \tau_a$ are random variables and we have a random effects model (varianskomponentmodell).

Case 1: $\tau_1, \ldots, \tau_a$ are fix (not random)

In the model above μ is the general mean for Y_{il}.

The model is overparameterized and to get unique estimators for the parameters $\tau_1, \ldots, \tau_a$ we need an extra constraint, for example

$$\sum_{1}^{a} n_i \tau_i = 0 \tag{1}$$

In total we now have $1 + (a - 1) = a$ free mean parameters.

To compare the a treatments, we would like to test the hypothesis:

$$H_0: \ \mu_1 = \ldots = \mu_a \ \Leftrightarrow \ \tau_1 = \tau_2 = \ldots = \tau_a = 0$$

versus

$$H_1: \ \mu_i \neq \mu_j \text{ at least one } (i,j) \ \Leftrightarrow \ \text{ at least one } \tau_i \neq 0$$

at level α.

Estimators

The estimators are given by

- μ_i is estimated by $\hat{\mu}_i = \bar{y}_{i\cdot} = \frac{1}{n_i} \sum_{l=1}^{n_i} y_{il}$,
- μ is estimated by $\hat{\mu} = \bar{y}_{\cdot\cdot} = \frac{1}{N} \sum_{i=1}^{a} \sum_{l=1}^{n_i} y_{il} = \frac{1}{N} \sum_{i=1}^{a} n_i \bar{y}_{i\cdot}$,

  where $N = \sum_{i=1}^{a} n_i$, and
- τ_i is estimated by $\hat{\tau}_i = \hat{\mu}_i - \hat{\mu} = \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}$ since $\mu_i = \mu + \tau_i$.

Using constraint (1) above on can show that these estimators are unbiased estimators of the parameters.

$$\mathsf{E}(\hat{\mu}_i) = \mathsf{E}\left[\frac{1}{n_i}\sum_{l=1}^{n_i} Y_{il}\right] = \frac{1}{n_i}\sum_{l=1}^{n_i}\mu_i = \mu_i$$

and

$$\begin{aligned}\mathsf{E}(\hat{\mu}) &= \mathsf{E}\left[\frac{1}{N}\sum_{i=1}^{a}\sum_{l=1}^{n_i} Y_{il}\right] = \frac{1}{N}\sum_{i=1}^{a}\sum_{l=1}^{n_i}(\mu+\tau_i)\\ &= \frac{1}{N}\sum_{i=1}^{a} n_i(\mu+\tau_i) = \frac{1}{N}\left[\sum_{i=1}^{a} n_i\mu + \sum_{i=1}^{a} n_i\tau_i\right]\\ &= \frac{1}{N}\cdot N\mu + 0 = \mu.\end{aligned}$$

Sum of Squares

As in regression, the total sum of squares can be written as

$$
\begin{aligned}
SS_T &= \sum_{i=1}^{a}\sum_{l=1}^{n_i}(y_{il}-\bar{y}_{\cdot\cdot})^2 = \sum_{i=1}^{a}\sum_{l=1}^{n_i}[(y_{il}-\bar{y}_{i\cdot})+(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})]^2 \\
&= \sum_{i=1}^{a}\sum_{l=1}^{n_i}(y_{il}-\bar{y}_{i\cdot})^2 + \sum_{i=1}^{a}\sum_{l=1}^{n_i}(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2 \\
&= SS_E + SS_{TREAT},
\end{aligned}
$$

since

$$
\begin{aligned}
2\sum_{i=1}^{a}\sum_{l=1}^{n_i}(y_{il}-\bar{y}_{i\cdot})(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot}) &= 2\sum_{i=1}^{a}(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})\sum_{l=1}^{n_i}(y_{il}-\bar{y}_{i\cdot}) \\
&= 2\sum_{i=1}^{a}(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})\left[\sum_{l=1}^{n_i}y_{il}-n_i\bar{y}_{i\cdot}\right] = 0
\end{aligned}
$$

- Between sum of squares

$$SS_{TREAT} = \sum_{i=1}^{a} n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 = \sum_{i=1}^{a} n_i(\hat{\mu}_i - \hat{\mu})^2 = \sum_{i=1}^{a} n_i\hat{\tau}_i^2$$

  SS_{TREAT} is small if $\tau_1 = \tau_2 = \ldots = \tau_a = 0$.

- Within sum of squares

$$SS_E = \sum_{i=1}^{a}\sum_{l=1}^{n_i}(y_{il} - \bar{y}_{i\cdot})^2 = \sum_{i=1}^{a}(n_i - 1)s_i^2 = (N - a)s^2$$

  where $s_i^2 = \frac{1}{n_i - 1}\sum_{l=1}^{n_i}(y_{il} - \bar{y}_{i\cdot})^2$ is the sample variance for treatment i and

$$s^2 = \frac{(n_1 - 1)s_1^2 + \ldots + (n_a - 1)s_a^2}{N - a}$$

  is the pooled variance for the a treatments.

The means for SS_E and SS_{TREAT} are given by

$$\mathrm{E}(SS_E) = (N - a)\sigma^2 \tag{2}$$

and

$$\mathrm{E}(SS_{TREAT}) = (a - 1)\sigma^2 + \sum_{i=1}^{a} n_i \tau_i^2 \tag{3}$$

Proof of (3) is given in Appendix in the end of this lecture.

ANOVA

Given the model $Y_{il} \sim N(\mu + \tau_i,\ \sigma)$ independent for $l = 1, \ldots, n_i$ and $i = 1, \ldots, a$, where $\sum_{i=1}^{a} n_i \tau_i = 0$, we have

(i) SS_{TREAT} and SS_E are independent,

(ii) $\frac{SS_E}{\sigma^2} \sim \chi^2(N - a)$, where $N = \sum_{i=1}^{a} n_i$,

(iii) if $\tau_1 = \tau_2 = \ldots = \tau_a = 0$, then $\frac{SS_{TREAT}}{\sigma^2} \sim \chi^2(a - 1)$.

We also know that the random variables $\overline{Y}_{1\cdot}, \ldots, \overline{Y}_{a\cdot}$ and SS_E are independent. We will use this when we construct confidence intervals.

ANOVA – $H_0: \ \mu_1 = \ldots = \mu_a$

We test the hypothesis

$$H_0: \ \mu_1 = \ldots = \mu_a \ \Leftrightarrow \ \tau_1 = \tau_2 = \ldots = \tau_a = 0$$

vs.

$$H_1: \ \mu_i \neq \mu_j \text{ for at least one } (i,j) \ \Leftrightarrow \ \text{ at least one } \tau_i \neq 0,$$

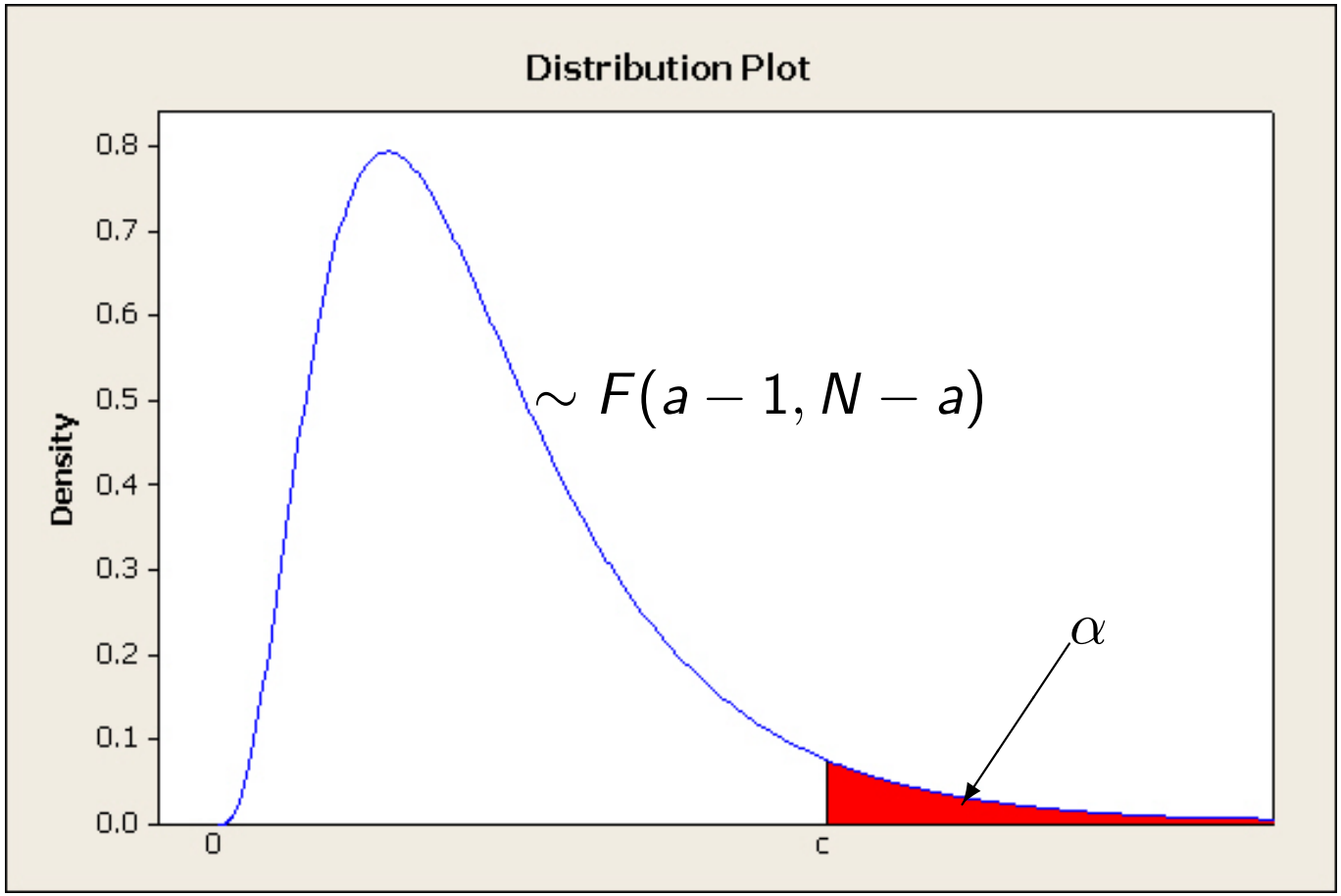
on level α using the test statistic

$$v = \frac{SS_{TREAT}/(a-1)}{SS_E/(N-a)}.$$

If H_0 is true, then $V \sim F(a-1, N-a)$.

Reject H_0 at level α if $v > c$, where $c = F_{1-\alpha}(a-1, N-a)$.

F-distribution



$$1 - \alpha = F_v(c) = \mathrm{P}(v \leq c) \Leftrightarrow c = F_{1-\alpha}(a-1, N-a).$$

Example 2 (Ex. 2, Lecture 1)

Observations from the four laboratories:

A	B	C	D
0.25	0.18	0.19	0.23
0.27	0.28	0.25	0.30
0.22	0.21	0.27	0.28
0.30	0.23	0.24	0.28
0.27	0.25	0.18	0.24
0.28	0.20	0.26	0.34
0.32	0.27	0.28	0.20
0.24	0.19	0.24	0.18
0.31	0.24	0.25	0.24
0.26	0.22	0.20	0.28
0.21	0.29	0.21	0.22
0.28	0.16	0.19	0.21

Model: For laboratoy i, $Y_{il} = \mu_i + \varepsilon_{il}$ and $\varepsilon_{il} \sim N(0, \sigma)$ for $i = 1, \ldots, 4, \;\; l = 1, \ldots, 12$.

$$SS_T = \sum_{i=1}^{4} \sum_{l=1}^{12} (y_{il} - \bar{y}_{..})^2 = 47 s^2_{TOT} = 0.08090,$$

where $47 = 48 - 1$ and $s^2_{TOT} =$ is the sample variance for all the 48 observations.

$SS_E = (12-1)s_1^2 + 11s_2^2 + 11s_3^2 + 11s_4^2 = 0.06789$ with 44 df.

$SS_{TREAT} = SS_T - SS_E = 0.01301$ with $4 - 1 = 3$ df.

$H_0 : \mu_1 = \ldots = \mu_4$ (all laboratories are equal), vs. H_1 : not all μ_i equal,

$$v = \frac{0.01301/3}{0.06789/44} = 2.811$$

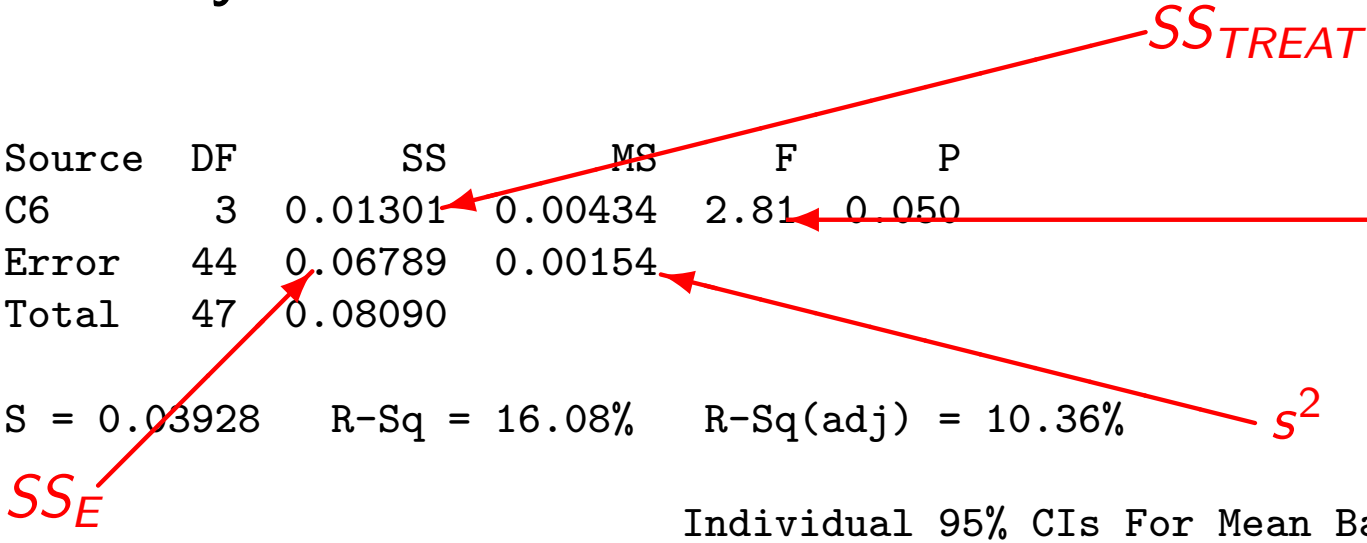
Reject H_0 at level 0.05 if $v > c = F_{0.95}(3, 44) \approx 2.82$

Hence, $v < c$ and we can not reject H_0. The laboratories can be equal.

```
MTB > Stack c1-c4 c5;
SUBC>   Subscripts c6;
SUBC>   UseNames.
MTB > Name c7 "RESI1" c8 "FITS1"
MTB > Oneway C5 C6;
SUBC>   Residuals 'RESI1';
SUBC>   Fits 'FITS1'.
```

- Observations are stored in columns c1-c4 in Minitab.
- `Stack` will stack all the columns in column c5.
- `Subscripts` will put the numbers for the laboratories in column c6.
- `Oneway` analyse the data using a one-factor model.

One-way ANOVA: C5 versus C6



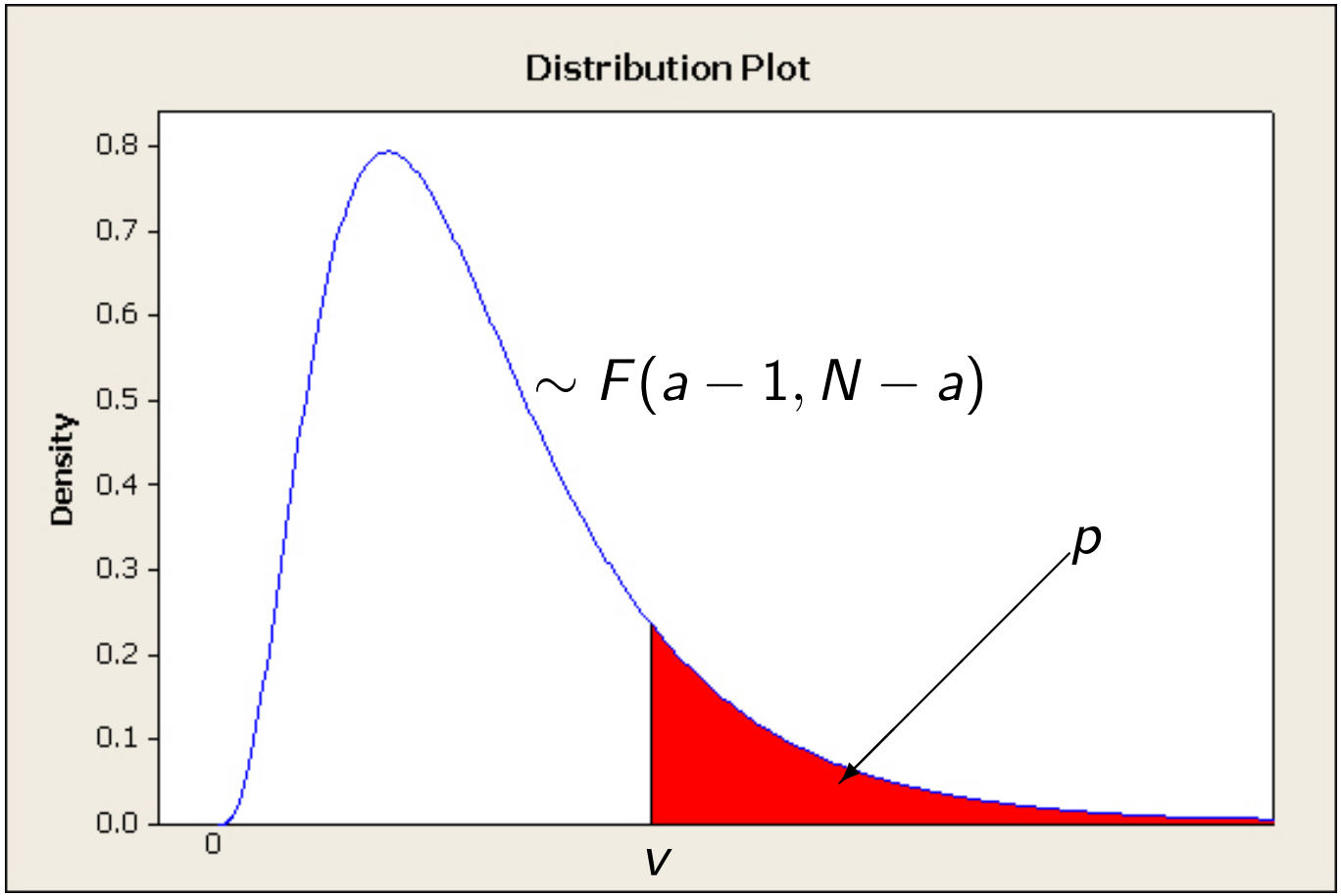

```
Source  DF       SS       MS      F      P
C6       3  0.01301  0.00434  2.81  0.050
Error   44  0.06789  0.00154
Total   47  0.08090

S = 0.03928   R-Sq = 16.08%   R-Sq(adj) = 10.36%

                                Individual 95% CIs For Mean Based on
                                Pooled StDev
Level   N     Mean    StDev     --------+---------+---------+---------+-
A      12  0.26750  0.03388                    (--------*--------)
B      12  0.22667  0.04097   (--------*--------)
C      12  0.23000  0.03438    (--------*--------)
D      12  0.25000  0.04651           (--------*--------)
                                --------+---------+---------+---------+-
                                      0.225     0.250     0.275     0.300

Pooled StDev = 0.03928
```

$= \sqrt{\frac{SS_E}{44}} = s$

SS = sum of squares, MS = mean square = $\frac{SS}{DF}$

DF = degrees of freedom

p-value



$$p = \mathrm{P}(v < V | H_0 \text{ is true}) = \begin{cases} < \alpha & \Rightarrow c < v \Rightarrow \text{ Reject } H_0 \\ > \alpha & \Rightarrow v < c \Rightarrow \text{ Do not reject } H_0 \end{cases}$$

Example 1, cont.

We have measured the amount strontium in five different watercourses. Resultat (in: mg/ml):

Grayson's Pond:	28.2	33.2	36.4	34.6	29.1	31.0;	$\bar{x}_1 = 32.1$
Beaver Lake:	39.6	40.8	37.9	37.1	43.6	42.4;	$\bar{x}_2 = 40.2$
Angler's Cove:	46.3	42.1	43.5	48.8	43.7	40.1;	$\bar{x}_3 = 44.1$
Appletree Lake:	41.0	44.1	46.4	40.2	38.6	36.3;	$\bar{x}_4 = 41.1$
Rock River:	56.3	54.1	59.4	62.7	60.0	57.3;	$\bar{x}_5 = 58.3$

- Are there any difference between the watercourses?
- Statistical model?
- How can we analyze the data?

Let y_{ij} be observation j for watercourse i.

Model: y_{ij} is an observation of

$$Y_{ij} = \underbrace{\mu + \tau_i}_{=\mu_i} + \varepsilon_{ij},$$

where μ_i is the mean for watercourse i and $\varepsilon_{ij} \sim N(0, \sigma)$ independently.

Minitab-analysis:

```
MTB > Stack c1-c5 c6;
SUBC>   Subscripts c7.
```

```
MTB > Oneway C6 C7;
SUBC>   Residuals 'RESI1';
SUBC>   Fits 'FITS1'.
```

One-way ANOVA: C6 versus C7

```
Source  DF       SS      MS      F      P
C7       4  2193.44  548.36  56.15  0.000
Error   25   244.13    9.77
Total   29  2437.57

S = 3.125   R-Sq = 89.98%   R-Sq(adj) = 88.38%

                          Individual 95% CIs For Mean Based on
                          Pooled StDev
Level  N    Mean  StDev  ---+---------+---------+---------+------
1      6  32.083  3.205  (--*--)
2      6  40.233  2.530           (--*---)
3      6  44.083  3.081                (--*--)
4      6  41.100  3.666            (--*---)
5      6  58.300  3.036                                  (--*--)
                          ---+---------+---------+---------+------
                          32.0      40.0      48.0      56.0
Pooled StDev = 3.125
```

We have $SS_{TREAT} = 2193.44$ with $DF = 4$ degrees of freedom.

Furthermore, $SS_E = 244.13$ with $DF = 25$.

The hypothesis

$$H_0 : \ \mu_1 = \ldots = \mu_5$$

vs.

$$H_1 : \ \mu_i \neq \mu_j \text{ for at least one } (i, j)$$

is tested with the test statistic

$$v = \frac{SS_{TREAT}/4}{SS_E/25} = 56.15 \quad \text{(F in the Minitab-analysis)}$$

and we have the random variable $V \sim F(4, 25)$ if H_0 is true.

Reject H_0 if $v > c$.

Table for the F-distribution gives $c = 4.2$ for $\alpha = 0.01$.

$v = 56.15 > 4.2 = c$ hence H_0 is rejected.

There is significant differences in the amount of strontium.

(Pairwise comparisons of the μ_i's in next lecture.)

Exampe 3 – Random effects model

Suppose that a clinical trial was conducted at 7 different clinics in Rwanda. The patients receiving the same dose level. The model equation for y_{ij}, which represents the jth patient at the ith clinic, could then be

$$\mathrm{E}(Y_{ij}) = \mu + \tau_i,$$

with $i = 1, 2, \ldots, 7$ for the 7 clinics.

It is not unreasonable to think of those 7 clinics as a random sample of clinics from some distribution of clinics, perhaps all the clinics in Rwanda.

Thus τ_i is here a *random effect*, i.e., τ_i can be assumed to be a random variable.

Case 2: τ_i:s are random variables – Random effects model

Model: For clinic i we have the responses $y_{i1}, \ldots, y_{in}$ with

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

where μ is the general mean,

$$\tau_i \sim N(0,\ \sigma_\tau); \quad \varepsilon_{ij} \sim N(0,\ \sigma)$$

and the random variables τ_i and ε_{ij}, $i = 1, \ldots, a$, $j = 1, \ldots, n$ are independent.

Here is $\mu + \tau_i$ a random variable, since the clinic is chosen at random.

The parameter σ_τ^2 gives the variation between the clinics.

Note that $\mathrm{E}(Y_{ij}) = \mu$ and $\mathrm{var}(Y_{ij}) = \sigma_\tau^2 + \sigma^2$ for all i, j.

The variances σ_τ^2 and σ^2 are called the *variance components*.

Now, we want to do two things,

(i) estimate μ, and

(ii) test if the individual treatment effects is meaningless, i.e., test the hypothesis

$$H_0 : \sigma_\tau^2 = 0.$$

ANOVA - table

The basic ANOVA sum of squares is still valid,

$$SS_T = SS_{TREAT} + SS_E,$$

where

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(Y_{ij} - \overline{Y}_{..})^2 = \sum_{i=1}^{a} n(\overline{Y}_{i.} - \overline{Y}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}(Y_{ij} - \overline{Y}_{i.})^2.$$

That is, we partition the total variability in the observations into component that measure the variation between treatments (SS_{TREAT}) and a component that measures the variation within treatments (SS_E).

Estimation of σ^2

Furthermore,

$$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{n}(Y_{ij} - \overline{Y}_{i\cdot})^2 = \sum_{i=1}^{a}\sum_{j=1}^{n}(\mu + \tau_i + \varepsilon_{ij} - \mu - \tau_i - \bar{\varepsilon}_{i\cdot})^2$$
$$= \sum_{i=1}^{a}\sum_{j=1}^{n}(\varepsilon_{ij} - \bar{\varepsilon}_{i\cdot})^2.$$

Hence, we have

$$\frac{SS_E}{\sigma^2} \sim \chi^2(N-a) \text{ where } N = a \cdot n,$$

and the estimator for σ^2 as

$$s^2 = \frac{SS_E}{N-a}.$$

For SS_{TREAT}, we have

$$SS_{TREAT} = \sum_{i=1}^{a} n(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2$$
$$= \sum_{i=1}^{a} n(\mu + \tau_i + \bar{\varepsilon}_{i\cdot} - \mu - \bar{\tau}_{\cdot} - \bar{\varepsilon}_{\cdot\cdot})^2 = \sum_{i=1}^{a} n(\xi_i - \bar{\xi})^2,$$

where $\xi_i = \tau_i + \bar{\varepsilon}_{i\cdot} \sim N\left(0, \sqrt{\sigma_\tau^2 + \frac{\sigma^2}{n}}\right)$.

Hence,

$$\frac{(a-1)s_\xi^2}{\sigma_\tau^2 + \frac{\sigma^2}{n}} = \frac{\sum_{i=1}^{a}(\xi_i - \bar{\xi})^2}{\sigma_\tau^2 + \frac{\sigma^2}{n}} \sim \chi^2(a-1)$$

which give

$$\frac{SS_{TREAT}}{n\sigma_\tau^2 + \sigma^2} \sim \chi^2(a-1).$$

Estimation of σ_τ^2

The expectation is given by $\mathrm{E}\left[\frac{SS_{TREAT}}{n\sigma_\tau^2+\sigma^2}\right] = a-1$, i.e.,

$$\mathrm{E}(SS_{TREAT}) = (a-1)\left(n\sigma_\tau^2+\sigma^2\right).$$

Hence, $\frac{SS_{TREAT}}{(a-1)n} - \frac{s^2}{n}$ is an unbiased estimator of σ_τ^2

since

$$E\left(\frac{SS_{TREAT}}{(a-1)n} - \frac{S^2}{n}\right) = \frac{1}{(a-1)n}\cdot(a-1)\left(n\sigma_\tau^2+\sigma^2\right) - \frac{\sigma^2}{n} = \sigma_\tau^2$$

Problem: What if the estimate of σ_τ^2 is negative? See the discussion in the book chapter 3.9.3.

Test the hypotesis $H_0 : \sigma_\tau^2 = 0$

To test the hypothesis

$$H_0 : \sigma_\tau^2 = 0 \quad \text{vs.} \quad H_1 : \sigma_\tau^2 \neq 0,$$

use the test statistic

$$v = \frac{SS_{TREAT}/(a-1)}{SS_E/(N-a)}.$$

Reject H_0 if $v > F_{1-\alpha}(a-1, N-a)$.

Confidence intervals for σ^2 and $\sigma_\tau^2/(\sigma^2+\sigma_\tau^2)$

Using the distributions

$$\frac{SS_E}{\sigma^2} \sim \chi^2(N-a)$$

and

$$\frac{SS_{TREAT}/((a-1)(n\sigma_\tau^2+\sigma^2))}{SS_E/((N-a)\sigma^2)} \sim F(a-1, N-a),$$

we can construct confidence interval for σ^2 and $\dfrac{\sigma_\tau^2}{\sigma^2+\sigma_\tau^2}$.

Confidence interval for μ

An estimator of μ is $\hat{\mu} = \bar{y}_{\cdot\cdot} = \frac{1}{a}\sum_{i=1}^{a} \bar{y}_{i\cdot}$.

Note that the random variables $Y_{i1}, \ldots, Y_{in}$ are dependent (same τ_i) but the random variables $\overline{Y}_{1\cdot}, \ldots, \overline{Y}_{a\cdot}$ are independent.

We will use this knowledge to construct a confidence interval for μ.

Let $z_i = \bar{y}_{i\cdot}$. Hence, the random variables $Z_1, \ldots, Z_a$ are independent and

$$Z_i \sim N(\mu, \sigma_z)$$

which gives

$$\hat{\mu} = \bar{Z} \sim N\left(\mu, \frac{\sigma_z}{\sqrt{a}}\right).$$

Use the random variable

$$\frac{\overline{Z} - \mu}{S_z/\sqrt{a}} \sim t(a-1)$$

where $s_z^2 = \frac{1}{a-1}\sum_{i=1}^{a}(z_i - \bar{z})^2$ is the sample variance for $z_1, \ldots, z_a$, for a confidence interval for μ.

Hence, a two sided interval for μ is given by

$$I_\mu = \left(\bar{z} \mp t_{1-\alpha/2}(a-1)\frac{s_z}{\sqrt{a}}\right).$$

Example 4

A certain infection is treated with the drug X.

For $a = 15$ random chosen patients we have $n = 5$ random observations how well the infection is treated with the drug.

- What can we say about the general mean of the treatment?
- What kinds of variations is there if we consider a new patient?

```
MTB > print c1-c15
```

Data Display

Row	C1	C2	C3	C4	C5	C6	C7	C8
1	19.3595	22.9155	26.1343	21.1629	21.6758	21.6846	20.5677	19.1862
2	19.7363	21.2579	26.6294	16.3660	20.6266	22.0661	19.2197	17.7990
3	20.5238	21.6356	25.9448	19.9420	23.7262	21.7052	18.9946	18.0906
4	21.5219	21.2248	25.2249	18.9539	20.8541	19.2029	20.7226	18.5292
5	22.4577	19.8331	26.2513	19.5058	21.1182	20.6110	22.0145	18.7037

Row	C9	C10	C11	C12	C13	C14	C15
1	20.8086	20.1142	15.1426	19.1095	24.3115	18.5970	23.5008
2	22.1398	21.2296	19.5518	19.4004	22.1113	19.5598	21.0335
3	20.7904	20.8418	20.1035	18.5683	23.6206	17.0910	21.5598
4	22.9082	22.8426	21.6458	19.4689	26.7137	20.1738	19.0197
5	21.4469	19.2864	18.8660	18.7319	20.2166	20.1907	22.9414

Model: Let y_{ij} be the response for observation j for patient i, i.e.,

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

where $\tau_1, \ldots, \tau_{15}$ are independent and $N(0,\ \sigma_\tau)$, and $\varepsilon_{11},\ \varepsilon_{12}, \ldots,$ $\varepsilon_{15,5}$ are independent and $N(0,\ \sigma)$.

- $\mu =$ "theoretical" level of the treatment (the general mean)
- $\mu + \tau_i =$ characteristic level for patient i.

Using the random variable

$$\frac{\overline{Z} - \mu}{S_z/\sqrt{15}} \sim t(14)$$

we will have the confidence interval for μ as

$$\overline{z} - t_{0.975}(14)\frac{s_z}{\sqrt{15}} < \mu < \overline{z} + t_{0.975}(14)\frac{s_z}{\sqrt{15}},$$

where $t_{0.975}(14) = 2.14$.

For a new patient we have

$$y_0 = \mu + \tau_0 + \varepsilon_0 \sim N\left(\mu, \sqrt{\sigma_\tau^2 + \sigma^2}\right)$$

with $\mathrm{P}\left(\mu - 1.96\sqrt{\sigma_\tau^2 + \sigma^2} < y_0 < \mu + 1.96\sqrt{\sigma_\tau^2 + \sigma^2}\right) \approx 0.95$.

We need to estimate the parameters. Minitab gives

```
MTB > stack c1-c15 c16;
SUBC> subscripts c17.
MTB > Oneway 'Y' 'person'.
```

One-way ANOVA: Y versus person

```
Source  DF      SS     MS     F      P
person  14  267.92  19.14  9.46  0.000
Error   60  121.32   2.02
Total   74  389.24

S = 1.422   R-Sq = 68.83%   R-Sq(adj) = 61.56%
```

Estimates are given by

$$\hat{\mu} = \bar{y}_{\cdot\cdot} = \bar{z} = 20.898,$$
$$s_z = \sqrt{3.8274} = 1.9564,$$
$$\hat{\sigma}^2 = \frac{SS_E}{N-a} = \frac{121.32}{75-15} = 2.022,$$
$$\hat{\sigma}_\tau^2 = \frac{SS_{TREAT}}{n(a-1)} - \frac{SS_E}{n(N-a)} = \frac{267.91}{5 \cdot 14} - \frac{2.022}{5} = 3.4229.$$

Hence,

$$I_\mu = \left(\bar{z} - 2.14\frac{s_z}{\sqrt{15}},\ \bar{z} + 2.14\frac{s_z}{\sqrt{15}} \right) = (19.8,\ 22.0), \quad \text{and}$$
$$I_{y_0} = \left(\hat{\mu} - 1.96\sqrt{\hat{\sigma}_\tau^2 + \hat{\sigma}^2},\ \hat{\mu} + 1.96\sqrt{\hat{\sigma}_\tau^2 + \hat{\sigma}^2} \right) = (16.3,\ 25.5).$$

Appendix: E(SS_{TREAT})

We need to use

$$\text{var}(Z) = \text{E}(Z^2) - [\text{E}(Z)]^2 \iff \text{E}(Z^2) = \text{var}(Z) + [\text{E}(Z)]^2$$

in the derivation for the expectation of S_{TREAT} as

$$\begin{aligned}
\text{E}(SS_{TREAT}) &= E\left[\sum_{i=1}^{a} n_i(\bar{Y}_{i\cdot}^2 - 2\bar{Y}_{i\cdot}\cdot\bar{Y}_{\cdot\cdot} + \bar{Y}_{\cdot\cdot}^2)\right] \\
&= \text{E}\left[\sum_{i=1}^{a} n_i\bar{Y}_{i\cdot}^2 - 2\bar{Y}_{\cdot\cdot}\sum_{i=1}^{a} n_i\bar{Y}_{i\cdot} + \sum_{i=1}^{a} n_i\bar{Y}_{\cdot\cdot}^2\right] \\
&= \text{E}\left[\sum_{i=1}^{a} n_i\bar{Y}_{i\cdot}^2 - 2\bar{Y}_{\cdot\cdot}\cdot N\cdot\bar{Y}_{\cdot\cdot} + N\cdot\bar{Y}_{\cdot\cdot}^2\right] \\
&= \text{E}\left[\sum_{i=1}^{a} n_i\bar{Y}_{i\cdot}^2 - N\bar{Y}_{\cdot\cdot}^2\right] = \sum_{i=1}^{a} n_i\,\text{E}(\bar{Y}_{i\cdot}^2) - N\,\text{E}(\bar{Y}_{\cdot\cdot}^2) \\
&= \ldots
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{a} n_i \left[\mathrm{var}(\bar{Y}_{i\cdot}) + (\mathrm{E}(\bar{Y}_{i\cdot}))^2\right] - N\left[\mathrm{var}(\bar{Y}_{\cdot\cdot}) + (\mathrm{E}(\bar{Y}_{\cdot\cdot}))^2\right] \\
&= \sum_{i=1}^{a} n_i \left[\frac{\sigma^2}{n_i} + (\mu+\tau_i)^2\right] - N\left(\frac{\sigma^2}{N} + \mu^2\right) \\
&= a\cdot\sigma^2 + \sum_{i=1}^{a} n_i(\mu^2 + 2\mu\tau_i + \tau_i^2) - \sigma^2 - N\mu^2 \\
&= (a-1)\sigma^2 + N\mu^2 + 2\mu\sum_{i=1}^{a} n_i\tau_i + \sum_{i=1}^{a} n_i\tau_i^2 - N\mu^2 \\
&= (a-1)\sigma^2 + \sum_{i=1}^{a} n_i\tau_i^2,
\end{aligned}
$$

since $\sum_{i=1}^{a} n_i\tau_i = 0$ as in (1).

Linköping University - Research that makes a difference