Lecture 16.
Simultaneous t-tests – the f-test.

There are many occasions in data analysis when we want to perform several t-tests simultaneously.

Example:
A large company carried out a survey as how much it cost them to print a page of a document. Three types/makes of colour laser printer as used across the company.
Various department use different printers and print different material.
The sampling unit was an individual printer, number of pages printed over 3 months and the associated costs (paper, cartridges, maintenance, depreciation etc) was recorded. The data are costs per page in cent.

	Type of printer		
	1	2	3
	2.08	1.96	1.47
	1.89	1.98	1.45
	2.23	1.96	1.77
	1.99	2.31	2.05
	2.01		
average	2.040	2.053	1.685
sd.	0.126	0.172	0.284

How do we proceed here? It would be inappropriate to conduct the three tests:
$\mu_1 = \mu_2$ then $\mu_1 = \mu_3$ and finally $\mu_2 = \mu_3$. Why?

What significance level (p-value) do we attach to this? The tests are not independent.

We approach the problem by constructing a *linear model*. Linear models are a general approach to analysing data that are more complicated than simple samples taken to estimate means.

Notation:

Y_{ij} - data (yield)
μ_i - group mean
e_{ij} - error - unexplainable variation.

In a more general model Y and e have more subscripts and we have other factors b, c.. etc.

$$Y_{ij} = \mu_i + e_{ij}$$

μ_i are parameters to be estimated. e_{ij} are assumed to be $N(0,\sigma^2)$.
In the specific example:

$2.08 = \mu_1 + e_{11}$ printer 1 of type 1
$1.89 = \mu_1 + e_{12}$ printer 2 of type 1
...
$1.96 = \mu_2 + e_{21}$ printer 1 of type 2 ... and so on.

The model can be fitted by using regression:

Define indicator variables for printers:

P1 = 1 if printer type 1, 0 otherwise.
P2 =1 if printer type 2, 0 otherwise.

We do not need an indictor variable for a printer of type 3- because type 3 is implied by P1=0, P2=0.

If we now fit the multiple regression:

$$Y = \beta_0 + \beta_1 P1 + \beta_2 P2$$

The fitted value is

for type 1 printers	$Y = \beta_0 + \beta_1$	(P1=1, P2=0)
for type 2	$Y = \beta_0 + \beta_2$	(P1=0, P2=1)
for type 3	$Y = \beta_0$	(P1=0, P2=0)

Thus β_0 is the estimate of the mean for type 3.

β_1 measures the difference between type 1 and type 3.

β_2 measures the difference between type 2 and type 3.

Thus if $\beta_1 = \beta_2 = 0$ all the means are equal.

So we want a test to see if we can drop P1 and P2 from the model simultaneously.

The t-tests that the regression output gives are one variable at a time tests (so called marginal).

We can test the effect of dropping more than one variable at the same time by the change in the residual sum of squares = SSE.

When one or more variables is dropped from the model =regression formula the SSE always increases (or stays the same).

Notation:

M1 – model one the larger model (more terms)

M2 – smaller model, *obtained by dropping variables from M1*.
M2 must **never** involve variables that are not in M1.
SSE(M1) is the SSE from model 1, SSE(M2) SSE from model 2.

$$f = \frac{\frac{SSE(M2) - SSE(M1)}{df1}}{\frac{SSE(M1)}{df2}}$$

$df1$ = number of variables dropped
= #vars M1 - #vars M2
= # degrees of freedom M2 - #deg. of fr. M1.

$df2$ = degrees of freedom M1.

The value of f is compared with the F - distribution.

If f is too big, we reject the hypothesis that all the β belonging to the dropped variables are 0. i.e. conclude that at least one of them is **not 0.**

The data:

Y	P1	P2
2.08	1	0
1.89	1	0
2.23	1	0
1.99	1	0
2.01	1	0
1.96	0	1
1.98	0	1
1.96	0	1
2.31	0	1
1.47	0	0
1.45	0	0
1.77	0	0
2.05	0	0

Regression output:
Dependent variable is: Y
No Selector

R squared = 47.8% R squared (adjusted) = 37.3%
s = 0.1985 with 13 - 3 = 10 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	0.360348	2	0.180174	4.57
Residual	**0.394175**	**10**	0.0394175	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	1.685	0.09927	17	< 0.0001
P1	0.355	0.1332	2.67	0.0237
P2	0.3675	0.1404	2.62	0.0257

All we actually want from this is

SSE =0.394175 and $df\,2=10$

Note however that:

the estimate β_0=constant = 1.685 = printer 3 average.
and β_1 = 0.355 = 2.040 –1.685
With just indicator variables the β estimates are just averages – the regression is very simple.

The f-test however applies to any regression.

We also need SSE associated with the model with β_1 and β_2 dropped.

You can fit this degenerate regression (no variables) in Data Desk by "dropping dimensions".

SSE(M2)=0.754523

$$f = \frac{\frac{0.754523 - 0.394175}{2}}{\frac{0.394175}{10}} = 4.57$$

This now must be compared to the F(2,10) distribution.

The F distribution has two parameters:

df1=numerator degrees of freedom

df2=denominator degrees of freedom.

Thus for its tabulation it requires a 2 dimensional table for each percentage point.

In examinations I only provide the 5% percentage point.

F critical values

p=5%

			numerator df									
denominator df	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**24**
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	249.1
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.5
3	10.1	9.6	9.3	9.1	9.0	8.9	8.9	8.8	8.8	8.8	8.7	8.6
4	7.7	6.9	6.6	6.4	6.3	6.2	6.1	6.0	6.0	6.0	5.9	5.8
5	6.6	5.8	5.4	5.2	5.1	5.0	4.9	4.8	4.8	4.7	4.7	4.5
6	6.0	5.1	4.8	4.5	4.4	4.3	4.2	4.1	4.1	4.1	4.0	3.8
7	5.6	4.7	4.3	4.1	4.0	3.9	3.8	3.7	3.7	3.6	3.6	3.4
8	5.3	4.5	4.1	3.8	3.7	3.6	3.5	3.4	3.4	3.3	3.3	3.1
9	5.1	4.3	3.9	3.6	3.5	3.4	3.3	3.2	3.2	3.1	3.1	2.9
10	5.0	4.1	3.7	3.5	3.3	3.2	3.1	3.1	3.0	3.0	2.9	2.7
12	4.7	3.9	3.5	3.3	3.1	3.0	2.9	2.8	2.8	2.8	2.7	2.5
15	4.5	3.7	3.3	3.1	2.9	2.8	2.7	2.6	2.6	2.5	2.5	2.3
20	4.4	3.5	3.1	2.9	2.7	2.6	2.5	2.4	2.4	2.3	2.3	2.1
24	4.3	3.4	3.0	2.8	2.6	2.5	2.4	2.4	2.3	2.3	2.2	2.0
30	4.2	3.3	2.9	2.7	2.5	2.4	2.3	2.3	2.2	2.2	2.1	1.9
40	4.1	3.2	2.8	2.6	2.4	2.3	2.2	2.2	2.1	2.1	2.0	1.8
60	4.0	3.2	2.8	2.5	2.4	2.3	2.2	2.1	2.0	2.0	1.9	1.7
120	3.9	3.1	2.7	2.4	2.3	2.2	2.1	2.0	2.0	1.9	1.8	1.6

Excel is more flexible giving FDIST(x,df1,df2) = p value

From the table F(2,10) = 4.1 as we observed 4.57 we reject $\beta_1 = \beta_2 = 0 \Rightarrow$ reject equal means.

FDIST(4.57,2,10)=0.038.. thus the means differ at a significance of 3.8%.

Addendum

It should be pointed out that for this particular problem:

a simple comparison of k (=3) means =
a regression involving *only* indicator variables ,

a different method of estimation and computation of sums of squares can be used. – formulas are available.

But this method gives exactly the same F-test.

I have chosen to frame this as in a more general form - as a model testing in regression problem.

The simple t-test, comparison of 2 means, can also be treated in this way (just one indicator variable).
The test you get is F(x,1,df), this test is identical to the t-test because of the identity:

$$FDIST(x,1,df) = TDIST(x^2,df,2)$$