Statistics 5401
13. Profile Analysis

Gary W. Oehlert
School of Statistics
313B Ford Hall
612-625-1557
gary@stat.umn.edu

Let me add a few more things about simultaneous inference before going on to profile analysis.

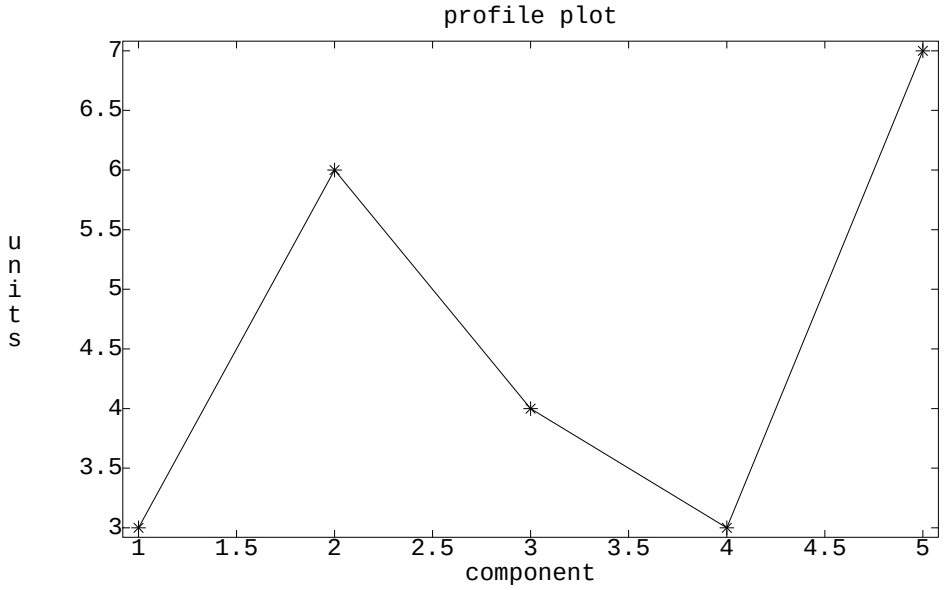
Advantages of the Bonferroni approach.

- Easy to compute and understand.
- If we do reject H_0, we have information about how it was violated (which variables differ).
- Only requires univariate assumptions, versus multivariate assumptions for T^2.
- Can use unpooled t when variances differ. No simple small sample fix for T^2.
- Will give shorter confidence intervals for a prechosen set of linear combinations (including the coordinates themselves).

Advantages of T^2.

- T^2 is invariant under nonsingular linear transformations. The Bonferroni approach is not.
- T^2 can have greater power than the Bonferroni approach, particularly when correlations are strong.
- The area (or volume) of a T^2 confidence region can be much less than that of the corresponding Bonferroni region.
- The T^2 confidence region can be used for linear combinations suggested by the data.

Consider a situation where we have p variables, all expressed in the same units, so that comparisons amongst them make sense. We have $\mu' = (\mu_1, \mu_2, \ldots, \mu_p)$. The *profile* is a connect-the-dots plot of the means.



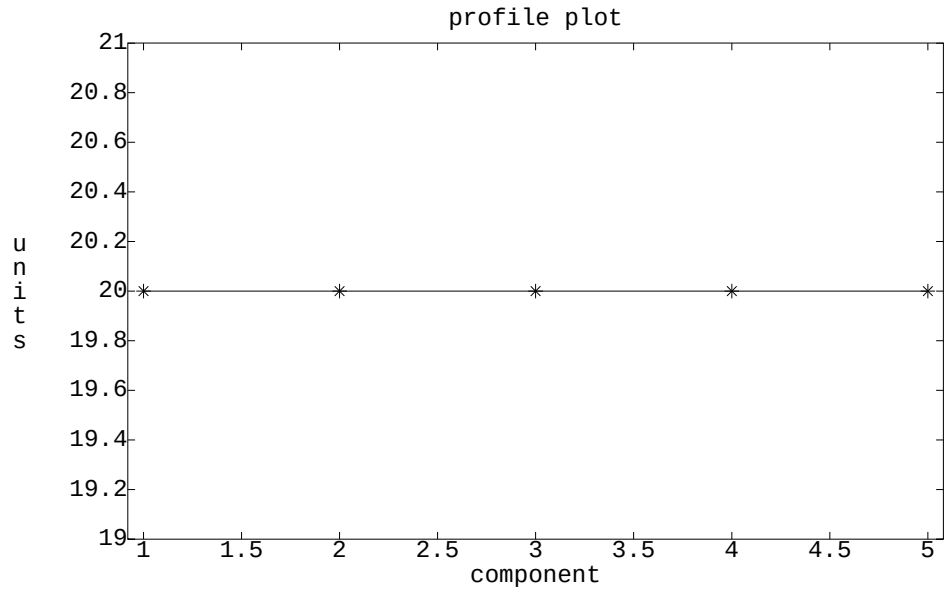
In many cases, each component corresponds to a different treatment, and the variables may be responses to the treatments.

Drug concentration in blood 1, 2, 3, 4, and 5 hours after administration.

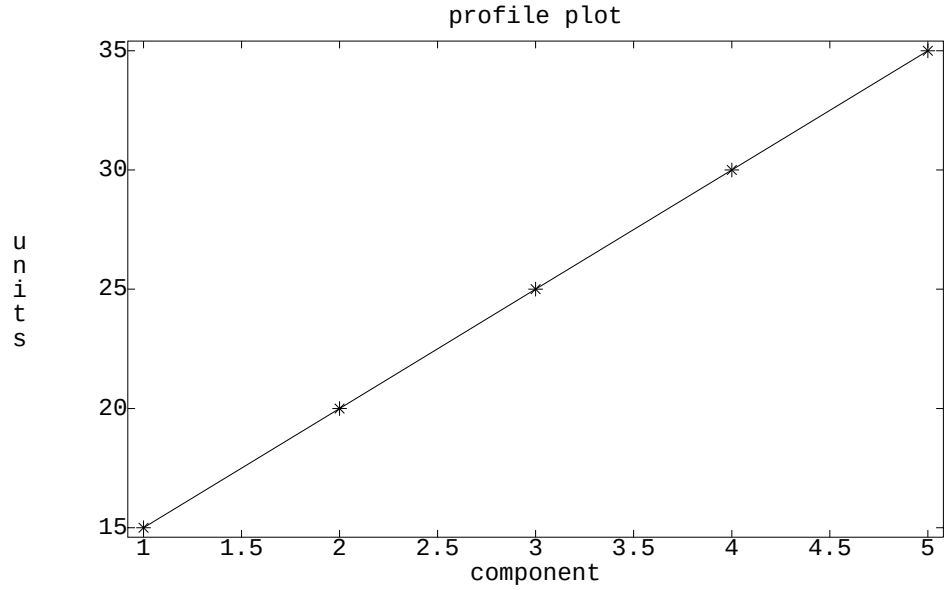
Corn yield using five different varieties.

IQ measured using four different techniques.

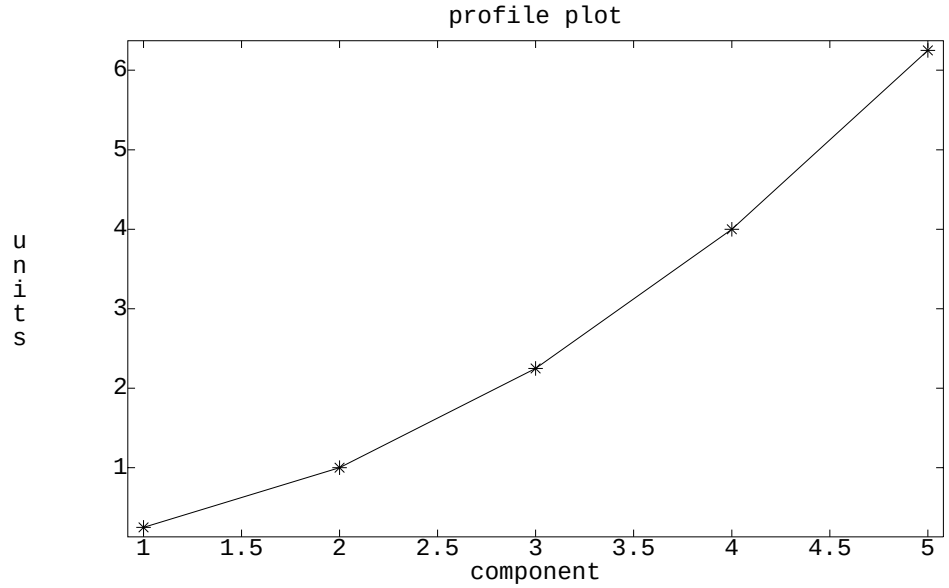
I will use the treatment/response terminology, even though there are other possibilities.
What might we like to know about profiles?
Does $\mu_1 = \mu_2 = \ldots = \mu_p$? Is the profile flat?



Does $\mu_i = a + bi$? Is the profile linear?

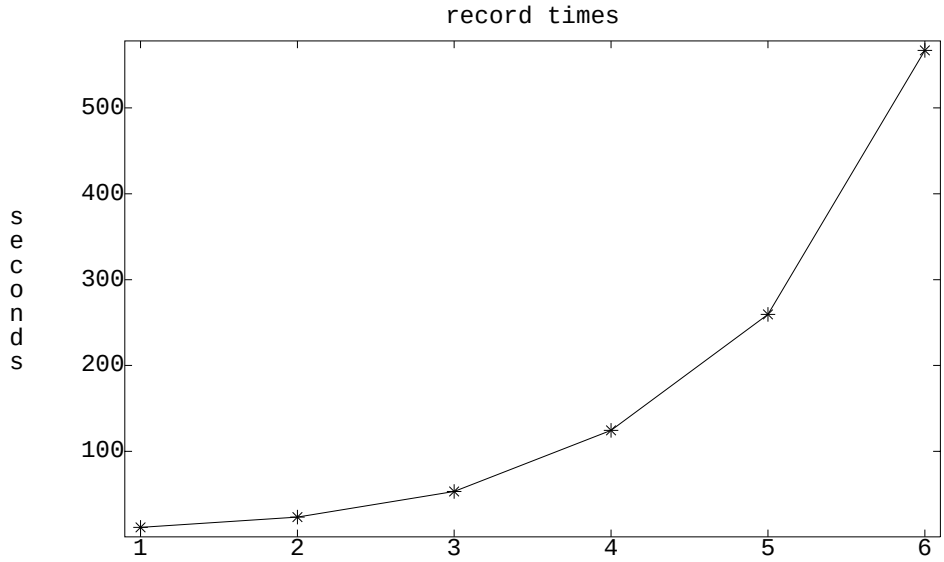


Does $\mu_i = a + bi + ci^2$? Is the profile quadratic?

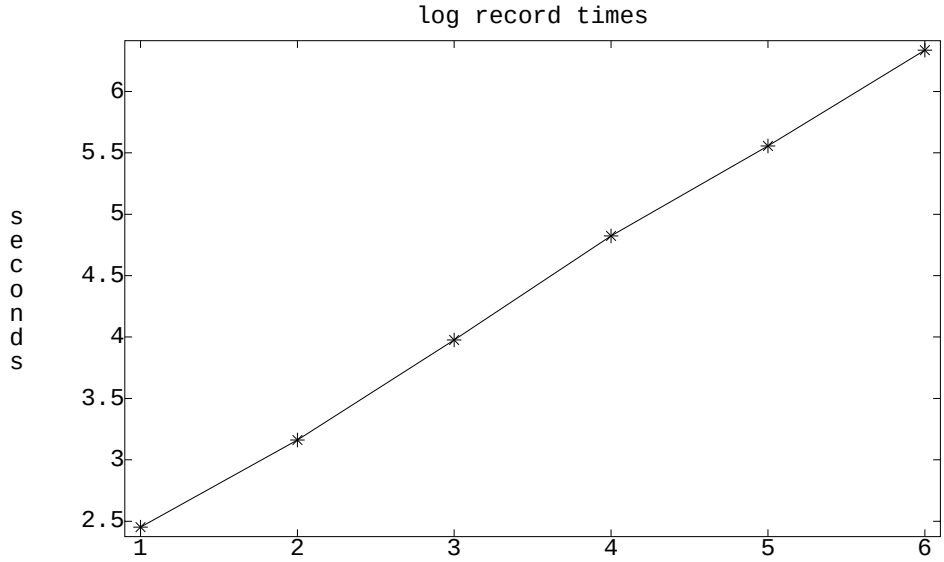


Women's track records from Table 1.9 of text.

```
Cmd> readdata("")
Read from file "/HOME/faculty/gary/classes/5401/JW5data/T1-9b.dat"
Column 1 saved as REAL vector x100
Column 2 saved as REAL vector x200
Column 3 saved as REAL vector x400
Column 4 saved as REAL vector x800
Column 5 saved as REAL vector x1500
Column 6 saved as REAL vector x3000
Column 7 saved as REAL vector xm
Column 8 saved as factor country

Cmd> x800 <- x800*60

Cmd> x1500 <- x1500*60

Cmd> x3000 <- x3000*60

Cmd> X <- hconcat(x100,x200,x400,\
x800,x1500,x3000)

Cmd> xbar <- tabs(X,mean:T)

Cmd> plot(1,xbar,lines:T,\
ylab:"seconds",title:"record times")
```




```
Cmd> xbarl <- tabs(log(X),mean:T)

Cmd> plot(1,xbarl,lines:T,\
ylab:"seconds",title:"log record times")
```



In each case, we can find a matrix $\mathbf{C}$ $(q \times p)$ and change the question to does $\mathbf{C}\mu = 0$.
Often there are many choices for $\mathbf{C}$. For concreteness, assume $p = 5$.
For a flat profile?

$$\mathbf{C}_1 = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

$$\mathbf{C}_2 = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 \\ 1 & 0 & 0 & 0 & -1 \end{bmatrix}$$

$$\mathbf{C}_3 = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ .5 & .5 & -1 & 0 & 0 \\ .333 & .333 & .333 & -1 & 0 \\ .25 & .25 & .25 & .25 & -1 \end{bmatrix}$$

$$\mathbf{C}_4 = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 \\ 1 & 0 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

The first three of these $\mathbf{C}$s have full rank $(p-1)$. $\mathbf{C}_4$ also has rank $(p-1)$, but it is not itself of full rank.
$\mathbf{C}_1$ looks at pairs of successive means:

$$\mathbf{C}_1\mu = \begin{bmatrix} \mu_1 - \mu_2 \\ \mu_2 - \mu_3 \\ \vdots \\ \mu_{p-1} - \mu_p \end{bmatrix}$$

A nonzero difference indicates a *change point* in the means.
$\mathbf{C}_2$ compares the first component to successive components

$$\mathbf{C}_2\mu = \begin{bmatrix} \mu_1 - \mu_2 \\ \mu_1 - \mu_3 \\ \vdots \\ \mu_1 - \mu_p \end{bmatrix}$$

This might be helpful if the first treatment represented a control or a standard, and we were interested in which treatments differed from control.
$\mathbf{C}_3$ compares the kth component to average of the preceding components

$$\mathbf{C}_3\mu = \begin{bmatrix} \mu_1 - \mu_2 \\ (\mu_1 + \mu_2)/2 - \mu_3 \\ \vdots \\ (\mu_1 + \ldots + \mu_{p-1})/(p-1) - \mu_p \end{bmatrix}$$

This could also be used to look for a change point.
$\mathbf{C}_4$ looks at all pairwise comparisons between treatments. There is reduncancy in this set of comparisons.
What if we were interested in linearity of the means? Linearity means that the increment from treatment i to $i+1$ is the same as the increment from $i+1$ to $i+2$. Thus we might consider looking at the difference of these increments:

$$(\mu_{i+2} - \mu_{i+1}) - (\mu_{i+1} - \mu_i) = \mu_{i+2} - 2\mu_{i+1} + \mu_i$$

$$\mathbf{C}_5 = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 1 & -2 & 1 \end{bmatrix}$$

Let $\mathbf{D}_p$ be the $(p-1) \times p$ matrix that computes $p-1$ successive differences from a p-vector. Note that $\mathbf{C}_1$ above is the same as $\mathbf{D}_5$.
To test linearity of a p-vector, use

$$\mathbf{C} = \mathbf{D}_{p-1}\mathbf{D}_p$$

Namely, $\mathbf{C}_5 = \mathbf{D}_4\mathbf{D}_5$.
If we are interested in the quadratic nature of the profile, we want to know if the change in increment from $i, i+1$ to $i+1, i+2$ is the same as the change in increment from $i+1, i+2$ to $i+2, i+3$. In differences:

$$(\mu_{i+2} - 2\mu_{i+1} + \mu_i) - (\mu_{i+3} - 2\mu_{i+2} + \mu_{i+1}) =$$

$$\mu_i - 3\mu_{i+1} + 3\mu_{i+2} - \mu_{i+3}$$

$$\mathbf{C}_6 = \begin{bmatrix} 1 & -3 & 3 & -1 & 0 \\ 0 & 1 & -3 & 3 & -1 \end{bmatrix}$$

Note, $\mathbf{C}_6 = \mathbf{D}_3\mathbf{D}_4\mathbf{D}_5$.
For a constant μ, $\mathbf{C}_1\mu = 0$.
For a linear μ, $\mathbf{C}_1\mu \neq 0$, but $\mathbf{C}_5\mu = 0$.
For a quadratic μ, $\mathbf{C}_1\mu \neq 0$, and $\mathbf{C}_5\mu \neq 0$, but $\mathbf{C}_6\mu = 0$.

```
Cmd> C1 %*% xbar
(1,1)      -12.023
(2,1)      -29.764
(3,1)      -71.176
(4,1)      -134.95
(5,1)      -307.33

Cmd> C5 %*% xbar
(1,1)       17.741
(2,1)       41.412
(3,1)       63.769
(4,1)       172.39

Cmd> C6 %*% xbar
(1,1)      -23.671
(2,1)      -22.358
```

```
(3,1)      -108.62

Cmd> C1 %*% xbarl
(1,1)     -0.71007
(2,1)     -0.81475
(3,1)     -0.84696
(4,1)     -0.73253
(5,1)     -0.78049

Cmd> C5 %*% xbarl
(1,1)      0.10468
(2,1)     0.032207
(3,1)     -0.11443
(4,1)     0.047964

Cmd> C6 %*% xbarl
(1,1)     0.072471
(2,1)      0.14664
(3,1)      -0.1624
```

How do we test $H_0 : \mathbf{C} = 0$?
We can use T^2 or Bonferroni t-tests.
Note: we had $\mathbf{C}_1$, $\mathbf{C}_2$, and $\mathbf{C}_3$ above, all describing equality of component means. Which should we use for testing?
It *does not* matter for T^2.
It *does* matter for Bonferroni t, which will be more sensitive if deviations from the null match the pattern exemplified by $\mathbf{C}$.

```
Cmd> T2 <- macro("
@X <- $1
@xb <- tabs(@X,mean:T)
@s <- tabs(@X,covar:T)
@p <- ncols(@X)
@n <- nrows(@X)
@T2 <- @xb'%*%solve(@s)%*%@xb*@n
@T2s <- @T2/@p*(@n-@p)/(@n-1)
@pv <- 1-cumF(@T2s,@p,@n-@p)
structure(T2:@T2,df:vector(@p,@n-@p),pval:@pv)")

Cmd> T2(X%*%C1')
component: T2
(1,1)        27243
component: df
(1)            5           50
component: pval
(1,1)            0
```

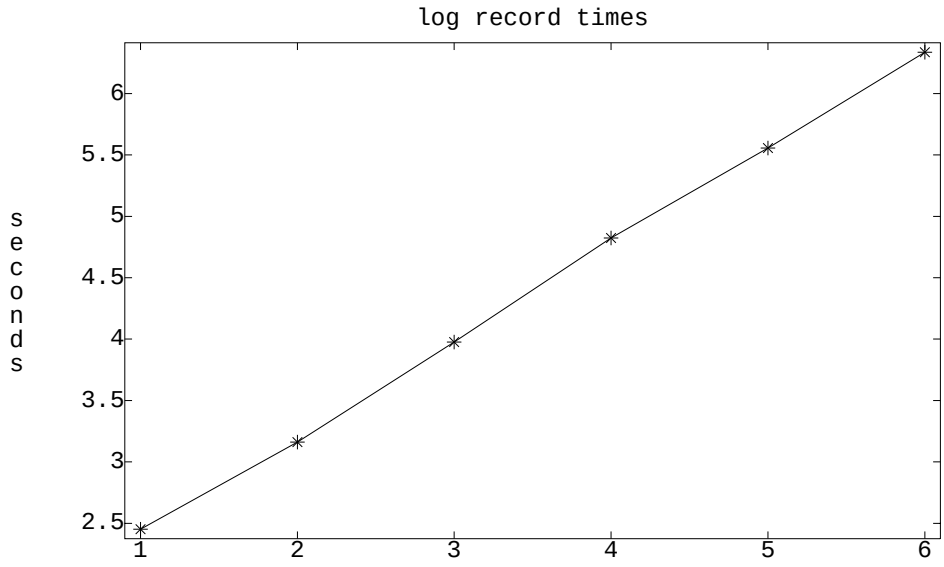
The original times aren't constant over distance, but we knew that anyway.

```
Cmd> T2(Xl%*%C1')
component: T2
(1,1)   4.0121e+05
component: df
(1)            5           50
component: pval
(1,1)            0
```

The logarithmic times are not constant over distance.

```
Cmd> T2(Xl%*%C5')
component: T2
(1,1)       1457.4
component: df
(1)            4           51
component: pval
(1,1)            0
```

The slopes are not constant using logarithmic data, I had high hopes here.




```
Cmd> T2(Xl%*%C6')
component: T2
(1,1)       1037.2
component: df
(1)            3           52
component: pval
(1,1)            0
```

Not quadratic either.

```
Cmd> tabs(Xl%*%C5',stddev:T)
(1)   0.0321    0.0324   0.0357   0.0340

Cmd> tabs(Xl%*%C5',mean:T)/\
tabs(Xl%*%C5',stddev:T)*sqrt(55)
(1)   24.179    7.375  -23.761   10.436
```

Bonferroni makes clear that each individual change in slope statistically different from 0.
There are two-sample analogs of these profile procedures. Suppose μ_1 and μ_2 are the means of the two populations.
Are the profiles parallel? Ie, does $\mathbf{C}_1\mu_1 = \mathbf{C}_1\mu_2$?
If μ_1 and μ_2 are parallel, we can test for equal by testing if $\mathbf{1}'\mu_1 = \mathbf{1}'\mu_2$. For parallel profiles, this is more powerful than testing $\mu_1 = \mu_2$.
For equal profiles, we can test for flatness via $\mathbf{C}_1\mu = 0$.
Let $\mathbf{S}$ be the *pooled* estimate of variance from two independent samples of size n and m.
Test for parallel profiles

$$T^2 = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)'\mathbf{C}' \left[(\frac{1}{n} + \frac{1}{m})\mathbf{C}\mathbf{S}\mathbf{C}' \right]^{-1} \mathbf{C}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)$$

compared with

$$\frac{(n+m-2)(p-1)}{n+m-p} \, F_{(p-1),\,(n+m-p)}$$

We can use $\mathbf{C}_1$, $\mathbf{C}_2$, $\mathbf{C}_3$ or any equivalent form for $\mathbf{C}$.
Test of equality *given* parallel profiles

$$T^2 = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)'\mathbf{1}' \left[(\frac{1}{n} + \frac{1}{m})\mathbf{1}'\mathbf{S}\mathbf{1} \right]^{-1} \mathbf{1}'(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)$$

compared with

$$F_{1,\,n+m-2}$$

Test of flatness *given* equal profiles. Let

$$\overline{\mathbf{x}} = \frac{n}{n+m}\overline{\mathbf{x}}_1 + \frac{m}{n+m}\overline{\mathbf{x}}_2$$

$$T^2 = (n+m)\overline{\mathbf{x}}'\mathbf{C}' \left[\mathbf{C}\mathbf{S}\mathbf{C}' \right]^{-1} \mathbf{C}\overline{\mathbf{x}}$$

compared with

$$\frac{(n+m-1)(p-1)}{n+m-p+1} \, F_{(p-1),\,(n+m-p+1)}$$

There is some muddle over multiple testing here. We may want to run each test at a smaller error rate so that the accumulated error rate does not get too large.
Also note that this approach is not unique. To test for equal, flat profiles, we could do as above, or we could test for equality of means, and then test for flatness of the common mean if equality is not rejected. We don't *have* to go through the parallel stage.
Split times data into sets of 25 and 30.

```
Cmd> X1 <- X[run(25),]

Cmd> X2 <- X[run(26,55),]

Cmd> S1 <- tabs(X1,covar:T)

Cmd> S2 <- tabs(X2,covar:T)

Cmd> Sp <- (24*S1+29*S2)/53

Cmd> xb1 <- tabs(X1,mean:T)

Cmd> xb2 <- tabs(X2,mean:T)
```

Test for parallel profiles

```
Cmd> (xb1-xb2)'%*%C1'%*%\
solve((1/25+1/30)*C1%*%Sp%*%C1')\
%*%C1%*%(xb1-xb2)
(1,1)       3.6987

Cmd> 3.6987/5/53*49
(1)      0.68391

Cmd> 1-cumF(.68,5,49)
(1)      0.64068
```

Parallel looks OK.
Test for equal, given parallel.

```
Cmd> sum(xb1-xb2)^2/\
((1/25+1/30)*sum(vector(Sp)))
(1)     0.080678

Cmd> 1-cumF(.08,1,53)
(1)       0.7784
```

Equal looks OK. We could have gone straight to equal with a two-sample T^2.
Test for flat.

```
Cmd> xb <- (25*xb1+30*xb2)/55

Cmd> 55*xb'%*%C1'%*%\
solve(C1%*%Sp%*%C1')%*%C1%*%xb
(1,1)        26905

Cmd> 26905/54/5*50
```

```
(1)       4982.4

Cmd> 1-cumF(4982,5,50)
(1)            0
```

Not flat, but we didn’t expect it to be.