Displays for Statistics 5303

Lecture 9

September 23, 2002

Christopher Bingham, Instructor

612-625-7023 (St. Paul)
612-625-1024 (Minneapolis)

Class Web Page
http://www.stat.umn.edu/~kb/classes/5303

Error Rates

This has to do when you are testing a family of null hypotheses H_{0i}, i = 1, ... K.

Examples:

- All possible pairwise comparisons
  $\alpha_i - \alpha_j$, $1 \leq i < j \leq g$
  $K = g(g-1)/2$
- All comparisons with a control, say treatment 1
  $\alpha_i - \alpha_1$, $i = 2, \ldots, g$
  $K = g - 1$
- All polynomial contrasts
  $K = g - 1$

Assume that

- For each H_{0i} in the family you have a test statistic, say T_i
- A definite procedure based of the values of T_1, T_2, ..., T_K to determine, for each H_{0i} in the family, whether you should reject it or not reject it.

The T_i's might be t-statistics, F-statistics, χ^2 statistic or any other appropriate statistic.

Let p_i be the P-value associated with T_i.

We have considered two procedures:

- **Naive method**: Reject H_{0i} if T_i is significant at significance level α. Equivalently, reject H_{0i} if $p_i \leq \alpha$.
- **Bonferroni method**: Reject H_{0i} if T_i is significant at significance level α/K. Equivalently, reject H_{0i} if $Kp_i \leq \alpha$.

Neither of these makes any use of the values of T_j for other hypotheses.

A test of a single hypothesis is characterized by two **error rates**, the type I error rate

$$\alpha = P(\text{reject } H_0 \mid H_0 \text{ true})$$

and the type II error rate

$$\beta = P(\text{not reject } H_0 \mid H_0 \text{ false})$$

Of these, only α is under direct control.

When there are K hypotheses, the situation is more complicated since several type I error rates can be defined.

Per comparison or **per hypothesis error rate**

This is the probability ε of a type I error relative to any specific H_{0i}:

$$\varepsilon = P(\text{reject } H_{0i} \mid H_{0i} \text{ true})$$

- For the α level naive method the per comparison error rate is $\varepsilon = \alpha$
- For the α level Bonferroni method, the per comparison error rate is $\varepsilon = \alpha/K$

Per Experiment or **familywise error rate**

This is the probability ε of rejecting at least one H_{0i}, when all H_{0i} are true.

- Naive method: $\alpha \leq \varepsilon < K\alpha$, usually much greater than α
- Bonferroni method: $\alpha/K \leq \varepsilon < \alpha$, often close to α when α is small

An important part of this definition is the condition that *all* H_{0i} are true.

In many situations, this is very far fetched since you may have very strong evidence that some of the H_{0i} are false.

But this is probably the most commonly referred to type of error rate when testing a family of hypotheses.

False Discovery Rate FDR

When you reject H_{0i}, you want think you have *discovered* something, some effect, some difference of effects. For this reason, Oehlert calls a hypothesis rejection a *discovery.*

- A **true discovery** occurs when the tested hypothesis actually is false.
- A **false discovery** occurs when you reject a true null hypothesis, that is, commit a type I error.

When testing a family of hypotheses, you will make k discoveries where $0 \le k \le K$ ($k = 0$ means no H_{0i} are rejected; $k = K$ means all H_{0i} are rejected.

Some unknown number ℓ, $0 \le \ell \le k$, of these discoveries will be false discoveries, because the test hypothesis is true.

The **p**roportion of **f**alse **d**iscoveries is

$pfd = 0$, when $k = 0$

$pfd = \ell/k \le 1$ when $k > 0$.

pfd is an unobservable random variable

The *false discovery rate* is

$$FDR = \mu_{pfd} = E[pfd]$$

When all H_{0i} are true, pfd is either 0 or 1 and

$$FDR = P(\text{reject any } H_{0i}) = \text{Experimentwise error rate.}$$

Some multiple testing methods are designed to control type I errors so that $FDR \leq \varepsilon$. For such a method clearly limits the experimentwise error rate to ε, but also controls the error rate when some H_{0i} are false.

The actual value of FDR depends on how many H_{0i}'s are false. When all are false, FDR = 0. If all H_{0i} are false except 1, say, H_{01}, then FDR = P(reject H_{01}).

Reminder: FDR and other error rates are computed on the basis of the testing procedure which tests all H_{0i}. The decision for any particular H_{0i} may depend on all the P-values, not just on p_i.

This means that the FDR and other error rates may depend on how "untrue" H_{0i} is, for instance how different α_1 and α_2 are..

Strong familywise error rate

This is P(make any false discoveries)
= P(reject at least 1 true H_{0i})

This probability is not based on the assumption that all H_{0i} are true.

The actual value depends on how many H_{0i}'s are false and may depend on how false they are.

- If all H_{0i} are false, the strong familywise error rate = 0.
- If all H_{0i} are true, the strong familywise error rate = ordinary familywise error rate

The Bonferroni method (reject if $Kp_i \leq \alpha$) has strong familywise error rate $\leq \varepsilon$, but if H_{01} is true and H_{0i} is false, i = 2, ..., K, it has strong familywise error rate = ε/K.

Suppose

- Each H_{0i}: concerns one parameter θ_i. often defined in terms of other parameters. For example, with 3 groups, and H_{0i}: $\theta_i = 0$, i = 1, 2, 3, where $\theta_1 = \alpha_1 - \alpha_2$, $\theta_2 = \alpha_1 - \alpha_3$, $\theta_3 = \alpha_2 - \alpha_3$
- You have a procedure for computing a <u>simultaneous confidence intervals</u> for every θ_i

A set of intervals CI_1, CI_2, ..., CI_K are **simultaneous confidence intervals** provided:

P(*all* CI_1, CI_2, ...,CI_k **cover** the true θ_i's) = $1 - \varepsilon$ or at least $\geq 1 - \varepsilon$ for some specified ε. A CI "covers" its parameter θ if θ is in the CI (a random event).

The **simultaneous confidence level** is $1 - \varepsilon$.

You can base a test procedure for the family $\{H_{0i}\}$ where $H_{0i}: \theta_i = \delta_i$ (often $\delta_i = 0$) as follows:

1. Calculate all CI_i, $i = 1, \ldots, K$
2. Reject H_{0i} if CI_i does not cover δ_i, that is δ_i is not in CI_i.

If $1 - \varepsilon$ is the simultaneous confidence level, then the strong familywise error rate $\leq \varepsilon$.

When all the H_{0i} concern a hypothesis that can be tested by a t-statistic, the Bonferroni method is a method based on simultaneous confidence intervals

$$\theta_i = \hat{\theta}_i \pm t_{(\alpha/2)/K} \hat{SE}[\hat{\theta}_i]$$

There are at least two other procedures based on the Bonferroni approach, the **Holm** and **FDR** procedures.

They are similar in that the first step is to order the H_{0i} in order of increasing p-values p_i.

The Holm procedure works from the smallest p on up, rejecting H_0's until it finds a p value too big to reject at which point no further H_0's are rejected, using a Bonferroni-based procedure at each stage.

The FDR procedure works from the *largest* p *down*, checking each using a Bonferroni-based criterion until a small enough one is found largest. The corresponding hypothesis and all hypotheses associated with smaller p's are also rejected.

Holm procedure

0. Find ordinary P-values for the K hypotheses and sort them in increasing order $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(K)}$, and reorder the H_{0i} the same way as $H_{0(1)}$, $H_{0(2)}$, ..., $H_{0(K)}$
1. If $Kp_{(1)} > \varepsilon$, stop, rejecting no H_{0i}. Otherwise reject $H_{0(1)}$ and continue testing the remaining K-1 hypotheses.
2. If $(K - 1)p_{(2)} > \varepsilon$ stop. Otherwise reject $H_{0(2)}$ and continue testing the remaining K-2 hypotheses.

. . . . . .

j. Continue in a similar way until you find a $(K-j+1)p_{(j)} > \varepsilon$, at which point you declare $H_{0(i)}$ is not rejected, $i \geq j$

Thus at each stage, you apply the Bonferroni inequality to the set of hypotheses not yet rejected.

FDR procedure

0. Order the P-values and H_0's as for the Holm method.
1. If the largest P-value $p_{(K)} \leq \varepsilon$, stop and reject all H_0's.
2. If not, look at $p_{(K-1)}$. If $Kp_{(K-1)} \leq (K-1)\varepsilon$, stop and reject all $H_{0(i)}$, $i \leq K-1$, that is all with as small or smaller P-values

. . . . . .

j. If you have not rejected any $H_{0(\ell)}$, $\ell > j$, look at $p_{(j)}$. If $Kp_{(j)} \leq j\varepsilon$, stop and reject all $H_{0(i)}$, $i \leq j$

That is, the more p-values too large to reject that you find, the lower the cut point for a Bonferronized p-value. If you reach the final stage and the smallest $p_{(1)}$ leads to reject only if $Kp_{(1)} \leq \varepsilon$, that is $p_{(1)} \leq \varepsilon$.

Here I apply these four methods, naive, Bonferroni, Holm and FDR to the Problem 3.3 data.

```
Cmd> data33 <- read("","pr3.3",quiet:T) # Problem 3.3 data
Read from file "TP1:Stat5303:Data:OeCh03.dat"

Cmd> treat <- factor(data33[,1]) # create treatment factor

Cmd> longevity <- vector(data33[,2]) # create response vector

Cmd> anova("longevity=treat",fstat:T)
Model used is longevity=treat
              DF          SS          MS           F     P-value
CONSTANT       1      2782.4      2782.4  1349.49826     < 1e-08
treat          4      243.16       60.79    29.48371  5.9878e-07
ERROR1        15      30.928      2.0618

Cmd> W1 <- matrix(enter(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), 5)

Cmd> print(W1,format:"4.0f") # all 5*(5-1)/2 comparisons
W1:
(1,1)    1    1    1    1    0    0    0    0    0    0
(2,1)   -1    0    0    0    1    1    1    0    0    0
(3,1)    0   -1    0    0   -1    0    0    1    1    0
(4,1)    0    0   -1    0    0   -1    0   -1    0    1
(5,1)    0    0    0   -1    0    0   -1    0   -1   -1

Cmd> K <- ncols(W1); K # (5*4/2)
(1)          10

Cmd> tstats <- rep(0,K) # place to put t-statistics

Cmd> for(i,1,K){
        result <- contrast(treat,W1[,i]) # uses column i of W1
        tstats[i] <- result$estimate/result$se
    ;;}

Cmd> hypes <- vector("mu1-mu2","mu1-mu3","mu1-mu4","mu1-mu5",\
        "mu2-mu3","mu2-mu4","mu2-mu5","mu3-mu4","mu3-mu5",\
        "mu4-mu5")
```



```
Cmd> pvals <- twotailt(tstats,15)

Cmd> setlabels(pvals,hypes) # add labels to pvals

Cmd> pvals
     mu1-mu2     mu1-mu3     mu1-mu4     mu1-mu5     mu2-mu3
     mu2-mu4     mu2-mu5     mu3-mu4     mu3-mu5     mu4-mu5
  2.8671e-05  2.7416e-05  2.3821e-07  6.1028e-08     0.98068
   0.0098395   0.0013111    0.010344   0.0013786      0.3403

Cmd> J <- grade(pvals) # indices of increasing pvalues

Cmd> J # smallest is p[4], next is p[3], ..., largest is p[5]
(1)           4           3           2           1           7
(6)           9           6           8          10           5

Cmd> p_sorted <- pvals[J]; p_sorted
     mu1-mu5     mu1-mu4     mu1-mu3     mu1-mu2     mu2-mu5
     mu3-mu5     mu2-mu4     mu3-mu4     mu4-mu5     mu2-mu3
  6.1028e-08  2.3821e-07  2.7416e-05  2.8671e-05   0.0013111
   0.0013786   0.0098395    0.010344      0.3403     0.98068
```

The first eight $p_{(i)} \leq .05$ so the naive method finds all differences significant at the 5% level except $\mu_3 - \mu_5$ and $\mu_4 - \mu_5$.

```
Cmd> K*p_sorted # Bonferronized p-values
     mu1-mu5     mu1-mu4     mu1-mu3     mu1-mu2     mu2-mu5
     mu3-mu5     mu2-mu4     mu3-mu4     mu4-mu5     mu2-mu3
  6.1028e-07  2.3821e-06  0.00027416  0.00028671    0.013111
    0.013786    0.098395     0.10344       3.403      9.8068
```

Now only 6 are less than .05, and $\mu_2 - \mu_4$ and $\mu_3 - \mu_4$ are significant in addition to $\mu_3 - \mu_5$ and $\mu_4 - \mu_5$.

Here are the modified P-values for the Holm procedure.

```
Cmd> run(K,1)*p_sorted # Holmized p-values
     mu1-mu5     mu1-mu4     mu1-mu3     mu1-mu2     mu2-mu5
     mu3-mu5     mu2-mu4     mu3-mu4     mu4-mu5     mu2-mu3
  6.1028e-07  2.1439e-06  0.00021933   0.0002007   0.0078666
   0.0068929    0.039358    0.031033      0.6806     0.98068
```

The 9th is the first > .05 so the the first 8 are rejected. This is the same result as for the naive method, but controls the strong family wise error rate.

Finally, here are modified P-values for the FDR method.

```
Cmd> K*p_sorted/run(K,1) # "FDR-ized" p-values
     mu1-mu5     mu1-mu4     mu1-mu3     mu1-mu2     mu2-mu5
     mu3-mu5     mu2-mu4     mu3-mu4     mu4-mu5     mu2-mu3
  6.1028e-08  2.6468e-07   3.427e-05  4.0958e-05   0.0021852
   0.0027571    0.024599    0.034481      1.7015      9.8068
```

Starting at the high end, the third largest $p_{(8)} < .05$, so you again reject $H_{0(1)}, \ldots, H_{0(8)}$.