Multiple Comparisons

David M. Rocke

November 12, 2024

Multiple Comparisons

Generally speaking, *multiple comparisons* involves statistical techniques for investigating the properties of multiple hypothesis tests considered as a group. We will consider two commonly encountered cases:

- multiple hypothesis tests on a list of responses or on a list of covariates;
- a group of hypotheses on a levels of a factor usually hypothesizing equality of the coefficients of several levels or more generally that a linear function of the coefficients is zero.

Multiple Testing example

The Bottomly et al. mouse gene expression data contains gene expression by RNA-Seq for brain tissue in two strains of mice, 10 from the C57BL/6J strain and 11 from the DBA/2J strain. Fragments from the RNA were mapped to mouse genes, resulting in counts for 11,870 genes. If we conduct tests for difference between the counts or relative counts for the two strains for each of the genes, we have potentially 11,870 tests, though this could be reduced by eliminating genes whose total count was so small that there was little information.

If we conduct 11,870 tests at the 5% level, and all of the null hypotheses are true, then the expected number of false positives is $11870 \times 0.05 = 593.5$, which will cause considerable trouble in interpretation. A common proposal for adjusting the p-values of tests uses the Bonferroni inequality, in which when conducting k hypothesis tests at level α, the chance of at least one false positive is less than or equal to $k\alpha$. If we conduct the hypothesis tests each at level $0.05/k$, then the chance of any false positive is less than or equal to 0.05.

What kind of effects could be detected with $\alpha = 0.05/11870 = 4.2 \times 10^{-6}$? This would require (under normality with 11 in each group) that the means would be separated by 3.13 standard deviations, which would eliminate any real possibility of detecting an effect. Tests with $p = 0.00001$ would not be rejected! This would be difficult for researcher to accept, and so a different approach has come to be common in these areas of investigation.

If group 1 had mean 10 and standard deviation 1, and group 2 had mean 13.13 with standard deviation 1, then the denominator of the two sample t-test would be $\sqrt{s_1^2/11 + s_2^2/11} \approx \sqrt{2/11} = 0.426$ and the expected value of the difference of the means would be 3.13, so a typical z-score when the difference would be detected would be 7.34!

Requiring zero false positives in 11,870 test would be widely considered unreasonable. So how many false positives should we allow?

False Discovery Rate

Instead of trying to control the false positive number, we might instead choose to control the false discovery rate (FDR), which is the fraction of tests in a group of tests that are false positives. The most commonly used such procedure is that of Benjamini and Hochberg (1995). To use this, we can run all the hypothesis tests in the usual way and construct the vector of p-values, then apply the R function `p.adjust`.

```
p.adjust(p, method = p.adjust.methods, n = length(p))

p.adjust.methods
# c("holm", "hochberg", "hommel", "bonferroni", "BH", "BY",
#   "fdr", "none")
Arguments
p
numeric vector of p-values (possibly with NAs). Any other R object is coerced
  by as.numeric.

method
correction method, a character string. Can be abbreviated.

n
number of comparisons, must be at least length(p);
  only set this (to non-default) when you know what you are doing!
```

Note that declaring significant any test with FDR-adjusted p-value less than 0.05 is conceptually completely different from using the 0.05 threshold on an unadjusted p-value. It seems completely reasonable to set a higher threshold such as 0.10.

Any time you conduct hypothesis tests, some of the “significant” ones may not be truly different, and 5 or 10 percent seems a modest penalty to wade through. The Bonferroni method tries to limit the false positives to zero, which reduces the power to a large extent.

Multiple Comparisons

The other kind of multiple comparison adjustment occurs especially when we have a factor with multiple levels in a linear predictor. Suppose we have three disease types, “ALL”, “AML-High”, and “AML-Low” with estimates of the log hazard ratio of AML-High vs. ALL and of the log hazard ratio of AML-Low vs. ALL. The coefficient table has hypothesis tests for whether these each are zero, but not of the comparison of AML-High and AML-Low.

If we have a factor describing the type of burn as chemical, scald, electric, and flame, there are six distinct pairwise comparisons, and well as possibly other linear hypotheses, and if we conduct lots of tests on the three coefficients in the table, we may have false positives. Of course a test of whether all the rates are equal is obtainable by a LR test of the models with and without that factor. Note that Wald tests of any linear hypothesis can be conducted for any type of regression model for which asymptotically valid covariance matrices can be derived.

A linear hypothesis on a vector of coefficients β of length p with estimates $\hat{\beta}$ is of the form

$$H_0 : L^\top \beta = k,$$

where L is a vector of numbers of length p; often L is a contrast meaning that the sum of the entries is zero and k is also often zero. If $\hat{\beta}$ has estimated covariance matrix $\hat{V}$, then the estimated variance of $L^\top \hat{\beta}$ is $L^\top \hat{V} L$ and an approximate z-statistic for the hypothesis as stated is

$$z = \frac{L^\top \hat{\beta} - k}{\sqrt{L^\top \hat{V} L}}.$$

Comparison with a Control

```
recovery {multcomp}
Recovery time after surgery.

This data frame contains the following variables

blanket
blanket type, a factor at four levels: b0, b1, b2, and b3.

minutes
response variable: recovery time after a surgical procedure.

Details
A company developed specialized heating blankets designed to help the body heat
following a surgical procedure. Four types of blankets were tried on surgical
patients with the aim of comparing the recovery time of patients.
One of the blanket was a standard blanket that had been in use already
in various hospitals.

Source
P. H. Westfall, R. D. Tobias, D. Rom, R. D. Wolfinger, Y. Hochberg (1999).
Multiple Comparisons and Multiple Tests Using the SAS System. Cary,
NC: SAS Institute Inc., page 66.
```

```
> library(multcomp)
> data(recovery)
> recovery.lm <- lm(minutes~blanket,data=recovery)
> summary(recovery.lm)

Call:
lm(formula = minutes ~ blanket, data = recovery)

Residuals:
   Min     1Q Median     3Q    Max
-6.133 -1.800  0.200  2.200  4.867

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.8000     0.5792  25.552  < 2e-16 ***
blanketb1    -2.1333     1.6038  -1.330   0.1916
blanketb2    -7.4667     1.6038  -4.656 4.07e-05 ***
blanketb3    -1.6667     0.8848  -1.884   0.0675 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.59 on 37 degrees of freedom
Multiple R-squared:  0.3797,    Adjusted R-squared:  0.3294
F-statistic:  7.55 on 3 and 37 DF,  p-value: 0.0004619
```

It looks like blanket b2 is better than b0, but we did conduct three hypothesis tests to obtain that finding. The F-test shows that not all the blankets are the same, so it might be reasonable to attribute that only to b2, but we can test that allowing for the multiple comparisons and the correlations between the tests using the Dunnett procedure and also obtain confidence intervals adjusted for multiple comparisons. This is based on the multivariate t distribution of the coefficients and is implemented in the `glht()` command in the R package `multcomp`.

```
> recovery.mc <- glht(recovery.lm,linfct=mcp(blanket="Dunnett"))
> summary(recovery.mc)

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Dunnett Contrasts


Fit: lm(formula = minutes ~ blanket, data = recovery)

Linear Hypotheses:
             Estimate Std. Error t value Pr(>|t|)
b1 - b0 == 0  -2.1333     1.6038  -1.330    0.456
b2 - b0 == 0  -7.4667     1.6038  -4.656   <0.001 ***
b3 - b0 == 0  -1.6667     0.8848  -1.884    0.182
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)
```

```
> names(recovery.mc)
[1] "model"       "linfct"      "rhs"         "coef"        "vcov"        "df"
[7] "alternative" "type"        "focus"

> recovery.mc$linfct

        (Intercept) blanketb1 blanketb2 blanketb3
b1 - b0           0         1         0         0
b2 - b0           0         0         1         0
b3 - b0           0         0         0         1
attr(,"type")
[1] "Dunnett"

> recovery.mc$rhs
[1] 0 0 0
> recovery.mc$focus
[1] "blanket"
```

Some attributes of an object have extractor functions, including `coef` and `vcov`. All the components can be accessed as attributes of the object. The three linear hypotheses require the linear vectors L and the right-hand sides k.

```
contrMat(n, type = c("Dunnett", "Tukey", "Sequen", "AVE",
                     "Changepoint", "Williams", "Marcus",
                      "McDermott", "UmbrellaWilliams", "GrandMean"),
                    base = 1)
Arguments
n                    a (possibly named) vector of sample sizes for each group.
type                 type of contrast.
base                 an integer specifying which group is considered the baseline
                     group for Dunnett contrasts.
```

This lists the types of pre-specified contrasts. Any set of linear hypotheses can also be specified just as a matrix `linfct` and right-hand side vector `rhs`. A base level can be given for Dunnett comparisons, which for general hypotheses is the `focus` attribute.

```
recovery.lm
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.8000     0.5792  25.552  < 2e-16 ***
blanketb1    -2.1333     1.6038  -1.330   0.1916
blanketb2    -7.4667     1.6038  -4.656 4.07e-05 ***
blanketb3    -1.6667     0.8848  -1.884   0.0675 .

recovery.mc
Linear Hypotheses:
             Estimate Std. Error t value Pr(>|t|)
b1 - b0 == 0  -2.1333     1.6038  -1.330    0.456
b2 - b0 == 0  -7.4667     1.6038  -4.656   <0.001 ***
b3 - b0 == 0  -1.6667     0.8848  -1.884    0.182
```

Note that the t-scores are the same, but the p-values are adjusted for multiple comparisons so that the chance that one or more is significant at level α in the null case is less than or equal to α

```
> summary(recovery.mc,test  = adjusted(type="bonferroni"))

	 Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Dunnett Contrasts


Fit: lm(formula = minutes ~ blanket, data = recovery)

Linear Hypotheses:
             Estimate Std. Error t value Pr(>|t|)    
b1 - b0 == 0  -2.1333     1.6038  -1.330 0.574796    
b2 - b0 == 0  -7.4667     1.6038  -4.656 0.000122 ***
b3 - b0 == 0  -1.6667     0.8848  -1.884 0.202439    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- bonferroni method)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.8000     0.5792  25.552  < 2e-16 ***          lm
blanketb1    -2.1333     1.6038  -1.330   0.1916
blanketb2    -7.4667     1.6038  -4.656 4.07e-05 ***
blanketb3    -1.6667     0.8848  -1.884   0.0675 .

Linear Hypotheses:
             Estimate Std. Error t value Pr(>|t|)
b1 - b0 == 0  -2.1333     1.6038  -1.330    0.456             Dunnett
b2 - b0 == 0  -7.4667     1.6038  -4.656   <0.001 ***
b3 - b0 == 0  -1.6667     0.8848  -1.884    0.182

Linear Hypotheses:
             Estimate Std. Error t value Pr(>|t|)
b1 - b0 == 0  -2.1333     1.6038  -1.330 0.574796             Bonferroni
b2 - b0 == 0  -7.4667     1.6038  -4.656 0.000122 ***
b3 - b0 == 0  -1.6667     0.8848  -1.884 0.202439
```

Both Dunnett and Bonferroni protect the familywise error rate, but Dunnett has smaller p-values because it uses the correlations of the tests.

```
> confint(recovery.mc)

         Simultaneous Confidence Intervals

Multiple Comparisons of Means: Dunnett Contrasts


Fit: lm(formula = minutes ~ blanket, data = recovery)

Quantile = 2.489
95% family-wise confidence level


Linear Hypotheses:
             Estimate lwr      upr
b1 - b0 == 0  -2.1333  -6.1251   1.8584
b2 - b0 == 0  -7.4667 -11.4584  -3.4749
b3 - b0 == 0  -1.6667  -3.8688   0.5355
```

There is at least a 95% chance that all the true values of the contrasts lie in their stated intervals.

```
> confint(recovery.lm)
                 2.5 %     97.5 %
(Intercept)  13.626389 15.9736107
blanketb1    -5.382914  1.1162474
blanketb2   -10.716247 -4.2170859
blanketb3    -3.459387  0.1260532
```

```
Linear Hypotheses:
             Estimate lwr      upr
b1 - b0 == 0  -2.1333  -6.1251   1.8584
b2 - b0 == 0  -7.4667 -11.4584  -3.4749
b3 - b0 == 0  -1.6667  -3.8688   0.5355
```

The confidence intervals from `lm` are individually valid, but if we consider them to be independent the chance that at least one does not contain the true value is $1 - (0.95)^3 = 0.14$. We could use Bonferroni confidence intervals at 98.3% confidence, but the Dunnett ones will be narrower because they use the correlations of the variables.

All Pairs Comparisons

```
immer {MASS}
Yields from a Barley Field Trial
Description
The immer data frame has 30 rows and 4 columns. Five varieties of barley were
grown in six locations in each of 1931 and 1932.

This data frame contains the following columns:

Loc
The location.

Var
The variety of barley ("manchuria", "svansota", "velvet", "trebi" and "peatland").

Y1
Yield in 1931.

Y2
Yield in 1932.
```

```
> library(MASS)
> data(immer)
> immer1 <- data.frame(immer,Yield = (immer$Y1+immer$Y2))
> summary(immer.lm)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  204.403     12.156  16.815 2.88e-13 ***
VarP          16.300     12.156   1.341 0.194983    
VarS          -6.517     12.156  -0.536 0.597810    
VarT          47.617     12.156   3.917 0.000854 ***
VarV           9.583     12.156   0.788 0.439728    
LocD         -52.120     13.316  -3.914 0.000860 ***
LocGR        -56.680     13.316  -4.256 0.000386 ***
LocM          -7.180     13.316  -0.539 0.595705    
LocUF        -32.020     13.316  -2.405 0.025996 *  
LocW          54.280     13.316   4.076 0.000589 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.05 on 20 degrees of freedom
Multiple R-squared:  0.8568,    Adjusted R-squared:  0.7924 
F-statistic:  13.3 on 9 and 20 DF,  p-value: 1.216e-06
```

```
> drop1(immer.lm,test="F")
Single term deletions

Model:
Yield ~ Var + Loc
       Df Sum of Sq   RSS    AIC F value    Pr(>F)
<none>                 8866 190.66
Var     4     10620 19486 206.29  5.9891  0.002453 **
Loc     5     42442 51308 233.33 19.1480 5.212e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> tapply(immer1$Yield,immer1$Var,mean)
       M        P        S        T        V
188.7833 205.0833 182.2667 236.4000 198.3667
> sort(tapply(immer1$Yield,immer1$Var,mean))
       S        M        V        P        T
182.2667 188.7833 198.3667 205.0833 236.4000
```

Both variety and location are significant, but it is not clear which pairs of varieties are shown to differ.

```
       S        M        V        P        T 
182.2667 188.7833 198.3667 205.0833 236.4000 

> immer.mc <- glht(immer.lm,linfct=mcp(Var = "Tukey"))
> summary(immer.mc)

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts

Fit: lm(formula = Yield ~ Var + Loc, data = immer1)

Linear Hypotheses:
           Estimate Std. Error t value Pr(>|t|)   
P - M == 0   16.300     12.156   1.341  0.67008   
S - M == 0   -6.517     12.156  -0.536  0.98242   
T - M == 0   47.617     12.156   3.917  0.00675 **
V - M == 0    9.583     12.156   0.788  0.93102   
S - P == 0  -22.817     12.156  -1.877  0.36064   
T - P == 0   31.317     12.156   2.576  0.11336   
V - P == 0   -6.717     12.156  -0.553  0.98035   
T - S == 0   54.133     12.156   4.453  0.00201 **
V - S == 0   16.100     12.156   1.324  0.67981   
V - T == 0  -38.033     12.156  -3.129  0.03773 * 
```

```
> confint(immer.mc)

         Simultaneous Confidence Intervals

Multiple Comparisons of Means: Tukey Contrasts

Fit: lm(formula = Yield ~ Var + Loc, data = immer1)

Quantile = 2.9932
95% family-wise confidence level


Linear Hypotheses:
           Estimate lwr      upr
P - M == 0  16.3000 -20.0850  52.6850
S - M == 0  -6.5167 -42.9016  29.8683
T - M == 0  47.6167  11.2317  84.0016
V - M == 0   9.5833 -26.8016  45.9683
S - P == 0 -22.8167 -59.2016  13.5683
T - P == 0  31.3167  -5.0683  67.7016
V - P == 0  -6.7167 -43.1016  29.6683
T - S == 0  54.1333  17.7484  90.5183
V - S == 0  16.1000 -20.2850  52.4850
V - T == 0 -38.0333 -74.4183  -1.6484
```

The confidence intervals and tests that result from uncorrected `lm` and other regression models are often called *Least Significant Difference* $=$ *LSD* tests and intervals. When there are many levels of a factor, this can result in false positives. One possible intermediate choice is to use the LSD tests and intervals, but only if the anova test for the factor is significant. This method is sometimes called the *Protected LSD*. This protects against the case where all the levels have equal effect, but not against partial equalities. However, for some applications this may be enough.

Another Example

```
> summary(burn1$BurnType)
Chemical    Scald Electric    Flame
       9       18       11      116
> summary(coxph(burn1.surv˜Treatment+BurnType,data=burn1))

                     coef exp(coef) se(coef)      z Pr(>|z|)
TreatmentCleansing -0.5958    0.5511   0.2968 -2.008   0.0447 *
BurnTypeScald       1.1328    3.1044   1.0828  1.046   0.2955
BurnTypeElectric    2.2660    9.6407   1.0837  2.091   0.0365 *
BurnTypeFlame       0.9888    2.6879   1.0160  0.973   0.3305

> drop1(coxph(burn1.surv˜Treatment+BurnType,data=burn1),test="Chisq")
          Df    AIC    LRT Pr(>Chi)
<none>       435.03
Treatment  1 437.14 4.1021  0.04283 *
BurnType   3 436.84 7.8095  0.05012 .
```

```
> summary(burn1$BurnType)
Chemical    Scald Electric    Flame
       9       18       11      116
> burntype.mc <- glht(coxph(burn1.surv~Treatment+BurnType,data=burn1),
    linfct=mcp(BurnType="Tukey"))
> summary(burntype.mc)

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts

Fit: coxph(formula = burn1.surv ~ Treatment + BurnType, data = burn1)

Linear Hypotheses:
                         Estimate Std. Error z value Pr(>|z|)
Scald - Chemical == 0      1.1328     1.0828   1.046   0.7013
Electric - Chemical == 0   2.2660     1.0837   2.091   0.1406
Flame - Chemical == 0      0.9888     1.0160   0.973   0.7460
Electric - Scald == 0      1.1332     0.5902   1.920   0.1999
Flame - Scald == 0        -0.1441     0.4456  -0.323   0.9870
Flame - Electric == 0     -1.2772     0.4521  -2.825   0.0212 *
```

```
> confint(burntype.mc)

         Simultaneous Confidence Intervals

Multiple Comparisons of Means: Tukey Contrasts


Fit: coxph(formula = burn1.surv ~ Treatment + BurnType, data = burn1)

Quantile = 2.5176
95% family-wise confidence level


Linear Hypotheses:
                         Estimate lwr     upr
Scald - Chemical == 0     1.1328  -1.5931  3.8587
Electric - Chemical == 0  2.2660  -0.4622  4.9942
Flame - Chemical == 0     0.9888  -1.5691  3.5466
Electric - Scald == 0     1.1332  -0.3527  2.6190
Flame - Scald == 0       -0.1441  -1.2658  0.9777
Flame - Electric == 0    -1.2772  -2.4153 -0.1391
```

Scheffé Tests and Intervals

In the last example, allowing for the fact that six tests were conducted, one of them was still significant and yet the anova test for the type factor was not (quite). One reason might be that the LR test and the Wald test are only asymptotically equivalent. But another is that the hypothesis that a factor has no effect means that *any* linear combination of levels has no associated effect because the effect associated with any factor level is zero. Tests and intervals can be based on this idea, that we need to be protected from false positives in any (linear) test suggested by the results of an analysis.

Suppose we have a factor with r levels and an effect μ_i associated with each level. This could be coefficients in a regression in which the coefficient for level i is already a comparison between level i and level 1. The assertion that the factor has no effect in either case is the hypothesis that $\mu_1 = \mu_2 = \cdots = \mu_r$. In the coefficient case, $\mu_1 = 0$ so then all the values of $\mu_i = 0$, but in any case, if we have a contrast L, then $L^\top M = 0$, where M is the vector $(\mu_1, \mu_2, \ldots, \mu_r)$. We have the (infinite) collection of contrasts and we want a test/interval such that when the total null hypothesis on the factor is true, then the chance that any test will be significant is less than or equal to α.

The method uses the estimated value of the contrast and the standard error, but instead using the t-statistic, one uses instead $\hat{C} \pm s_{\hat{C}}\sqrt{(r-1)\,F_{\alpha;r-1;df}}$, where df is the residual degrees of freedom. So with six types of barley in an experiment with 30 data points, the multipier is $\sqrt{5F_{.05,5,20}} = 3.68$ instead of $t_{20} = 2.086$. Generally, this level of protection is achieved at too high a cost. If differences are the inferential target, the Tukey HSD is better (with multiplier 2.518).