Pairwise Comparisons

Andrew Zieffler

Educational Psychology

University of Minnesota

Driven to Discover℠

Unadjusted Mean Differences Between Ethnic Groups

Contrast	ΔM	p
Asian – Black	–0.29	0.938
Asian – Hispanic	4.22	0.213
Asian – Other	4.10	0.306
Asian – White	–3.13	0.326
Black – Hispanic	4.51	0.090
Black – Other	4.39	0.198
Black – White	–2.84	0.237
Hispanic – Other	–0.12	0.967
Hispanic – White	–7.36	0.00003
Other – White	–7.23	0.009

- The difference between the average reading score for white students and that for hispanic students is statistically reliable, ($p < 0.001$).
- The difference between the average reading score for white students and that for the ethnicity of "other" is statistically reliable, ($p = 0.009$).

Consider Examining the Following Regression Summary

```
# reading ~ gender + momEd + ses


Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  47.7286     2.6457  18.040  < 2e-16 ***
genderMale    0.1147     1.2986   0.088  0.92969    
momEd         0.7828     0.5956   1.314  0.19026    
ses           4.3119     1.4467   2.980  0.00324 ** 
---
```

In this model, each predictor is evaluated by comparing the coefficient's *p*-value to 0.05 (or some other alpha value).

This sets the probability of making a type I error 0.05 for each predictor….we will falsely reject the null hypothesis of no effect in 5 out of 100 samples that we may have drawn.

Now add ethnicity….

```
# reading ~ gender + momEd + ses + black + hispanic + other + white


Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 47.590683   3.713065  12.817  < 2e-16 ***
genderMale   0.001199   1.318386   0.001  0.99928    
momEd        0.516331   0.606508   0.851  0.39565    
ses          4.009769   1.478448   2.712  0.00729 ** 
black        2.029142   3.672466   0.553  0.58123    
hispanic    -1.445867   3.264705  -0.443  0.65835    
other       -0.987540   3.837160  -0.257  0.79717    
white        2.811233   3.043539   0.924  0.35682    
---
```

The effect for ethnicity should be evaluated at 0.05.

The effect of ethnicity is represented by four dummy predictors. To keep our type I error rate for ethnicity at 0.05, we should not evaluate each of these dummy predictors by comparing the coefficient's *p*-value to 0.05.

The Effect of Ethnicity

If any two ethnicities have a difference between their average reading score, we would say there is an effect of ethnicity.

Thus the effect of ethnicity is really composed of 10 unique **pairwise comparisons**, or contrasts.

- Asian vs. Black
- Asian vs. Hispanic
- Asian vs. Other
- Asian vs. White
- Black vs. Hispanic
- Black vs. Other
- Black vs. White
- Hispanic vs. Other
- Hispanic vs. White
- Other vs. White

All 10 of these comparisons are referred to as a family or an ensemble.

There are two schools of thought about evaluating mean differences that are part of a family or ensemble:

- Use the unadjusted p-values
- Adjust the p-values based on the number of comparisons in the ensemble.

In the social sciences, we tend to adjust the p-values when the mean difference is part of an ensemble.

Methods of adjustment invariably adjust the unadjusted p-value upwards, making them **larger** than unadjusted p-values.

When an unadjusted p-value is very small or very large, rarely will an adjustment change the judgment regarding statistical reliability (significance).

For unadjusted p-values that hover around 0.05, adjustment might lead to a different judgment about the group differences.

Familywise Type I Error

Is there an effect of
ethnicity?

$$\alpha = 0.05$$

When we consider the entire family of tests, we are interested in limiting the possibility of making a type I error across **all** the comparisons.

With more comparisons, there are more ways to make a type I error. For example it is possible to make a type I error on any one of the 10 comparisons. It is also possible to make a type I error on 2 of the 10 comparisons (there are 45 different ways to do this!). It is also possible to make a type I error on 3 of the 10 comparisons (120 ways to do this). etc.

Computing Familywise Type I Error

The type I error rate for each test is referred to as the the per-contrast error rate. This is the probability of making a type I error for that particular test, or contrast.

Say that we use 0.05 as the testwise type I error rate on each of our comparisons…

Is there a difference between Asians and Blacks? $\alpha = 0.05$

Is there a difference between Asians and Hispanics? $\alpha = 0.05$

$\vdots$ $\vdots$

Is there a difference between Others and Whites? $\alpha = 0.05$

The probability of making **at least one type I error**, the familywise error rate, is quite a bit higher than 0.05!

$$\alpha_{\text{FW}} = 1 - (1 - \alpha)^k$$

$$\alpha_{\text{FW}} = 1 - (1 - 0.05)^{10} = 0.401$$

The probability of making at least one type I error in these 10 tests is 0.401.

Big question: What significance-level should we use to test each of the pairwise hypotheses if **we want the familywise error rate to be 0.05**?

Is there a difference between Asians and Blacks? $\alpha = ?$

Is there a difference between Asians and Hispanics? $\alpha = ?$

$\vdots$ $\vdots$

Is there a difference between Others and Whites? $\alpha = ?$

$$\alpha_{\mathrm{FW}} = 1 - (1 - \alpha)^k$$

$$0.05 = 1 - (1 - \alpha)^{10}$$

We could use algebra to solve for the per-contrast error rate, but you would have to recall how to solve a polynomial.

Bonferroni solved this type of algebra problem to find the value for alpha that gives an upper-bound for the familywise error rate of 0.05.

Olive Jean Dunn then used Bonferroni's solution in practice.

?

Carlo Emilio Bonferroni

Olive Jean Dunn

This ensemble-adjustment procedure, the most common adjustment method used in the social sciences, is known as the **Dunn-Bonferroni adjustment**.

$$\alpha_{adj.} = \frac{\alpha_{FW}}{k}$$

$$\alpha_{adj.} = \frac{0.05}{10} = 0.005$$

The per-contrast alpha value for each pairwise contrast is 0.005. This means the *p*-value for each pairwise contrast should be evaluated against 0.005 rather than 0.05.

Contrast	ΔM	p	Statistically Reliable compared to 0.05?	Statistically Reliable compared to 0.005?
Asian – Black	–0.29	0.938	No	No
Asian – Hispanic	4.22	0.213	No	No
Asian – Other	4.10	0.306	No	No
Asian – White	–3.13	0.326	No	No
Black – Hispanic	4.51	0.090	No	No
Black – Other	4.39	0.198	No	No
Black – White	–2.84	0.237	No	No
Hispanic – Other	–0.12	0.967	No	No
Hispanic – White	–7.36	0.00003	✓	✓
Other – White	–7.23	0.009	✓	No

To make it easier to report and interpret ensemble-adjusted results, conventionally we **adjust the *p*-values rather than the alpha-values.**

The Dunn-Bonferroni adjustment to the *p*-value is

$$p_{adj.} = p_k \times k$$

Contrast	ΔM	p	Dunn–Bonferroni adjusted p-value	Statistically Reliable (compare to 0.05)
Asian – Black	–0.29	0.938	1	No
Asian – Hispanic	4.22	0.213	1	No
Asian – Other	4.10	0.306	1	No
Asian – White	–3.13	0.326	1	No
Black – Hispanic	4.51	0.090	0.900	No
Black – Other	4.39	0.198	1	No
Black – White	–2.84	0.237	1	No
Hispanic – Other	–0.12	0.967	1	No
Hispanic – White	–7.36	0.00003	0.0003	✓
Other – White	–7.23	0.009	0.090	No

Adjusted p-values > 1 are changed to 1

```
# Enter the unadjusted p-values into a vector
> p.values = c(
    0.938, # asian vs. black
    0.213, # asian vs. hispanic
    0.306, # asian vs. other
    0.326, # asian vs. white
    0.0899, # black vs. hispanic
    0.1976, # black vs. other
    0.2365, # black vs. white
    0.9674, # hispanic vs. other
    0.0000281, # hispanic vs. white
    0.0088 # other vs. white
    )

# Get the Dunn-Bonferroni adjusted p-values
> p.adjust(p.values, method = "bonferroni")

 [1] 1.000000 1.000000 1.000000 1.000000 0.899000 1.000000 1.000000
1.000000 0.000281 0.088000‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- bonferroni method)
```

There are many, many, many, many ensemble-adjustment methods.

- Hommel procedure
- Benjamani-Hochberg procedure
- Fisher's Least Significant Difference (LSD) procedure
- Benjamani-Yekutieli procedure
- Tukey's Honestly Significant Difference (HSD) procedure
- Tukey-Kramer procedure
- Scheffé's procedure
- Neuman-Keuls procedure
- Holm procedure
- Waller-Duncan procedure
- Hochberg procedure
- Miller-Winer procedure

Some of these are less conservative (produce smaller p-values) others are more conservative (produce higher p-values)

The Dunn-Bonferroni procedure is not generally recommended due to the conservative (high) p-values it produces, especially with large numbers of contrasts.

Yoav Benjamani

Yosef Hochberg

The Benjamini–Hochberg procedure is an ensemble method based on **false discovery rate** (FDR).

FDR is a relatively new approach to the multiple comparisons problem. Instead of controlling the chance of at least one type I error, FDR **controls the expected proportion of type I errors** making these methods less prone to over-adjustment of the p-values.

Growing pool of evidence showing that this method may be the best solution to the multiple comparisons problem in many practical situations (Williams, Jones, & Tukey, 1999)

Because of its usefulness, the I*nstitute of Education Sciences* has recommended this procedure for use in its **What Works Clearinghouse** handbook of standards (Institute of Education Sciences, 2008).
http://ies.ed.gov/ncee/wwc/DocumentSum.aspx?sid=19

To understand how the Benjamani–Hochberg procedure makes adjustments, we first rank order the unadjusted *p*-values (from lowest to highest).

Contrast	p	Rank
Hispanic – White	0.00003	1
Other – White	0.009	2
Black – Hispanic	0.090	3
Black – Other	0.198	4
Asian – Hispanic	0.213	5
Black – White	0.237	6
Asian – Other	0.306	7
Asian – White	0.326	8
Asian – Black	0.938	9
Hispanic – Other	0.967	10

Start with the largest p-value, and adjust as follows:

$$p_{adj.} = \frac{k \times p_k}{\text{Rank}}$$

Hispanic – White

$$p_{adj.} = \frac{10 \times 0.00003}{1} = 0.0003$$

Other – White

$$p_{adj.} = \frac{10 \times 0.009}{2} = 0.045$$

Asian – Black

$$p_{adj.} = \frac{10 \times 0.938}{9} = 1.04$$

Hispanic – Other

$$p_{adj.} = \frac{10 \times 0.967}{10} = 0.967$$

The Benjamini–Hochberg adjusted *p*-value for a test is either the raw *p*-value times *k*/*Rank* (the adjusted *p*-value) **or** the adjusted *p*-value for the next higher raw *p*-value, whichever is smaller.

Contrast	p	Rank	$\frac{k \times p_k}{\text{Rank}}$	Benjamini–Hochberg adjusted *p*-value
Hispanic – White	0.00003	1	0.0003	0.0003
Other – White	0.009	2	0.0440	0.0440
Black – Hispanic	0.090	3	0.2997	0.2997
Black – Other	0.198	4	0.4940	0.3942
Asian – Hispanic	0.213	5	0.4260	0.3942
Black – White	0.237	6	0.3942	0.3942
Asian – Other	0.306	7	0.4371	0.4075
Asian – White	0.326	8	0.4075	0.4075
Asian – Black	0.938	9	1.0422	0.9674
Hispanic – Other	0.967	10	0.9674	0.9674

```
# Get the Benjamani-Hochberg adjusted p-values
> p.adjust(p.values, method = "BH")



 [1] 0.9674000 0.3941667 0.4075000 0.4075000 0.2996667 0.3941667
 [7] 0.3941667 0.9674000 0.0002810 0.0440000
```

Contrast	ΔM	p	Benjamani–Hochberg adjusted p-value	Statistically Reliable
Asian – Black	–0.29	0.938	0.9674	No
Asian – Hispanic	4.22	0.213	0.3942	No
Asian – Other	4.10	0.306	0.4075	No
Asian – White	–3.13	0.326	0.4075	No
Black – Hispanic	4.51	0.090	0.2997	No
Black – Other	4.39	0.198	0.3942	No
Black – White	–2.84	0.237	0.3942	No
Hispanic – Other	–0.12	0.967	0.9674	No
Hispanic – White	–7.36	0.00003	0.0003	✓
Other – White	–7.23	0.009	0.0440	✓

Contrast	ΔM	p	Dunn–Bonferroni adjusted p-value	Benjamani–Hochberg adjusted p-value
Asian – Black	–0.29	0.938	1	0.9674
Asian – Hispanic	4.22	0.213	1	0.3942
Asian – Other	4.10	0.306	1	0.4075
Asian – White	–3.13	0.326	1	0.4075
Black – Hispanic	4.51	0.090	0.900	0.2997
Black – Other	4.39	0.198	1	0.3942
Black – White	–2.84	0.237	1	0.3942
Hispanic – Other	–0.12	0.967	1	0.9674
Hispanic – White	–7.36	0.00003	0.0003	0.0003
Other – White	–7.23	0.009	0.090	0.0440

In general, smaller p-values than the Benjamani–Hochberg adjustment than with the Dunn–Bonferroni adjustment