STA221

Neil Montgomery

Last edited: 2017-03-08 15:05

pairwise comparisons

pairwise comparisons

A pairwise comparison will be a pooled two-sample t procedures, using the overall MSE in place of the usual pooled variance:

$$\frac{\overline{y}_i - \overline{y}_j}{\sqrt{MSE}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} \sim t_{N-k}$$

The usual technique is to produce confidence intervals for each desired pair.

pairwise comparisons

A pairwise comparison will be a pooled two-sample t procedures, using the overall MSE in place of the usual pooled variance:

$$\frac{\overline{y}_i - \overline{y}_j}{\sqrt{MSE}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} \sim t_{N-k}$$

The usual technique is to produce confidence intervals for each desired pair.

We will adjust the confidence level $100 \cdot (1 - \alpha)\%$ when multiple comparisons take place.

pairwise comparisons

A pairwise comparison will be a pooled two-sample *t* procedures, using the overall *MSE* in place of the usual pooled variance:

$$\frac{\overline{y}_i - \overline{y}_j}{\sqrt{MSE}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} \sim t_{N-k}$$

The usual technique is to produce confidence intervals for each desired pair.

We will adjust the confidence level $100 \cdot (1 - \alpha)\%$ when multiple comparisons take place.

Recall: a lower α gives a *wider* confidence interval. In the *t* case the full formula is:

$$(\overline{y}_i - \overline{y}_j) \pm t_{N-k,\alpha/2}\sqrt{MSE}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

if you torture your data long enough, it will tell you anything

From "classical" hypothesis testing there was the notion of a "Type I Error", which is *rejecting* H_0 *when it is true*, aka "false positive".

if you torture your data long enough, it will tell you anything

From "classical" hypothesis testing there was the notion of a "Type I Error", which is *rejecting H_0 when it is true*, aka "false positive".

The probability of a Type I Error is often called α and is set to be something low like 0.05. This is also called the "level" of the test.

if you torture your data long enough, it will tell you anything

From "classical" hypothesis testing there was the notion of a "Type I Error", which is *rejecting* H_0 *when it is true*, aka "false positive".

The probability of a Type I Error is often called α and is set to be something low like 0.05. This is also called the "level" of the test.

When you subject one dataset to multiple hypothesis tests at the same "level" α, you are exposing yourself to that α probability over and over again.

if you torture your data long enough, it will tell you anything

From “classical” hypothesis testing there was the notion of a “Type I Error”, which is *rejecting H_0 when it is true*, aka “false positive”.

The probability of a Type I Error is often called α and is set to be something low like 0.05. This is also called the “level” of the test.

When you subject one dataset to multiple hypothesis tests at the same “level” α, you are exposing yourself to that α probability over and over again.

Also, all the different hypothesis tests are usually not independent, so calculating the overall effect of multiple hypothesis tests is usually impossible.

if you torture your data long enough, it will tell you anything

From “classical” hypothesis testing there was the notion of a “Type I Error”, which is *rejecting* H_0 *when it is true*, aka “false positive”.

The probability of a Type I Error is often called α and is set to be something low like 0.05. This is also called the “level” of the test.

When you subject one dataset to multiple hypothesis tests at the same “level” α, you are exposing yourself to that α probability over and over again.

Also, all the different hypothesis tests are usually not independent, so calculating the overall effect of multiple hypothesis tests is usually impossible.

But, it is possible to put an *upper bound* on the overall effect.

if you torture your data long enough, it will tell you anything

From “classical” hypothesis testing there was the notion of a “Type I Error”, which is *rejecting H_0 when it is true*, aka “false positive”.

The probability of a Type I Error is often called α and is set to be something low like 0.05. This is also called the “level” of the test.

When you subject one dataset to multiple hypothesis tests at the same “level” α, you are exposing yourself to that α probability over and over again.

Also, all the different hypothesis tests are usually not independent, so calculating the overall effect of multiple hypothesis tests is usually impossible.

But, it is possible to put an *upper bound* on the overall effect.

Definition: the *experimentwise error rate* is the probability of *any* Type I Errors among all tests done on the dataset from one experiment.

bounding the experimentwise error rate - I

Suppose we're going to do m hypothesis tests on a dataset.

Denote by $A_1, A_2, \ldots, A_m$ the events where A_i means "a Type I Error occurred when hypothesis test i took place", and $P(A_i) = \alpha$.

bounding the experimentwise error rate - I

Suppose we're going to do m hypothesis tests on a dataset.

Denote by $A_1, A_2, \ldots, A_m$ the events where A_i means "a Type I Error occurred when hypothesis test i took place", and $P(A_i) = \alpha$.

The goal is to put one overall upper bound on the experimentwise error rate, which we'll call α^*.

bounding the experimentwise error rate - I

Suppose we're going to do m hypothesis tests on a dataset.

Denote by $A_1, A_2, \ldots, A_m$ the events where A_i means "a Type I Error occurred when hypothesis test i took place", and $P(A_i) = \alpha$.

The goal is to put one overall upper bound on the experimentwise error rate, which we'll call α^*.

You might recall the expression $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

bounding the experimentwise error rate - I

Suppose we're going to do m hypothesis tests on a dataset.

Denote by $A_1, A_2, \ldots, A_m$ the events where A_i means "a Type I Error occurred when hypothesis test i took place", and $P(A_i) = \alpha$.

The goal is to put one overall upper bound on the experimentwise error rate, which we'll call α^*.

You might recall the expression $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

This implies $P(A \cup B) \leq P(A) + P(B)$. You can extend this to any number of events, i.e.:

$$P(A_1 \cup A_2 \cup \cdots \cup A_m) \leq P(A_1) + P(A_2) + \cdots + P(A_m)$$

bounding the experimentwise error rate - II

How could the individual tests all be adjusted so that $\alpha^* = \alpha$?

An easy method uses:

$$\alpha^* = P(\text{any Type I Errors})$$

bounding the experimentwise error rate - II

How could the individual tests all be adjusted so that $\alpha^* = \alpha$?

An easy method uses:

$$\begin{aligned} \alpha^* &= P(\text{any Type I Errors}) \\ &= P(A_1 \cup A_2 \cup \cdots A_m) \end{aligned}$$

bounding the experimentwise error rate - II

How could the individual tests all be adjusted so that $\alpha^* = \alpha$?

An easy method uses:

$$
\begin{aligned}
\alpha^* &= P(\text{any Type I Errors}) \\
&= P(A_1 \cup A_2 \cup \cdots A_m) \\
&\leq P(A_1) + P(A_2) + \cdots + P(A_m)
\end{aligned}
$$

bounding the experimentwise error rate - II

How could the individual tests all be adjusted so that $\alpha^* = \alpha$?

An easy method uses:

$$\begin{aligned} \alpha^* &= P(\text{any Type I Errors}) \\ &= P(A_1 \cup A_2 \cup \cdots A_m) \\ &\leq P(A_1) + P(A_2) + \cdots + P(A_m) \end{aligned}$$

So can achieve $\alpha^* = \alpha$ simply by dividing each of these α by m.

bounding the experimentwise error rate - II

How could the individual tests all be adjusted so that $\alpha^* = \alpha$?

An easy method uses:

$$\begin{aligned} \alpha^* &= P(\text{any Type I Errors}) \\ &= P(A_1 \cup A_2 \cup \cdots A_m) \\ &\leq P(A_1) + P(A_2) + \cdots + P(A_m) \end{aligned}$$

So can achieve $\alpha^* = \alpha$ simply by dividing each of these α by m.

This is called a "Bonferroni correction".

bounding the experimentwise error rate - II

How could the individual tests all be adjusted so that $\alpha^* = \alpha$?

An easy method uses:

$$\begin{aligned}\alpha^* &= P(\text{any Type I Errors}) \\ &= P(A_1 \cup A_2 \cup \cdots A_m) \\ &\leq P(A_1) + P(A_2) + \cdots + P(A_m)\end{aligned}$$

So can achieve $\alpha^* = \alpha$ simply by dividing each of these α by m.

This is called a "Bonferroni correction".

It's not a bad idea to apply a Bonferroni correct to any situation in which you are subjecting a dataset to lots of hypothesis tests.

full example, including some pairwise comparisons

From the “Hearing” example there were four lists of words and 96 people. The % of words understood by each person was recorded.

full example, including some pairwise comparisons

From the "Hearing" example there were four lists of words and 96 people. The % of words understood by each person was recorded.

Suppose `List4` was some sort of "default list", and the other three lists were new word lists being evaluated. So it will be particularly interesting to investigate these three pairwise differences:

$$\mu_1 - \mu_4$$
$$\mu_2 - \mu_4$$
$$\mu_3 - \mu_4$$

full example, including some pairwise comparisons

From the "Hearing" example there were four lists of words and 96 people. The % of words understood by each person was recorded.

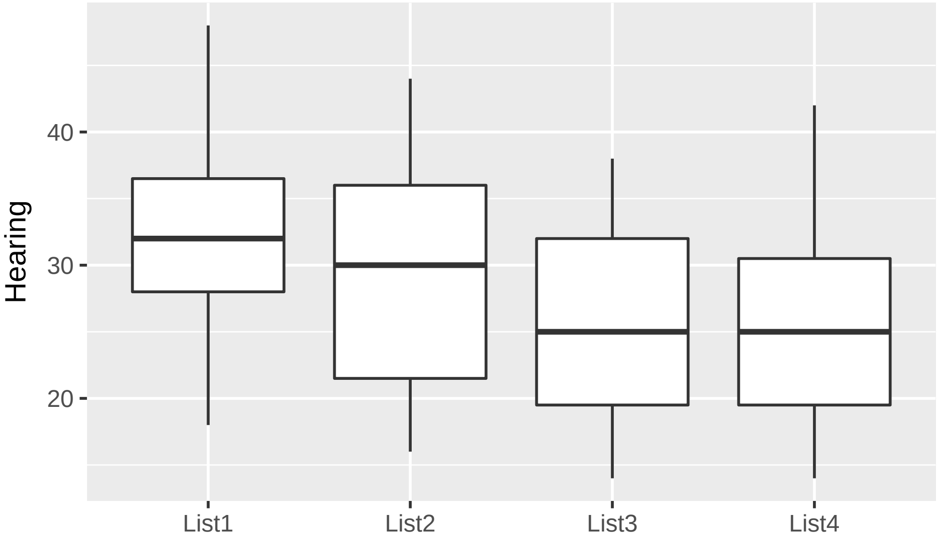
Suppose `List4` was some sort of "default list", and the other three lists were new word lists being evaluated. So it will be particularly interesting to investigate these three pairwise differences:

$$\mu_1 - \mu_4$$
$$\mu_2 - \mu_4$$
$$\mu_3 - \mu_4$$

We will fix the experimentwise error rate at $\alpha = 0.05$ for the multiple comparisons.

hearing full example - I

First, look at a plot:



hearing full example - II

Next, verify the model assumptions starting with Levene's test:

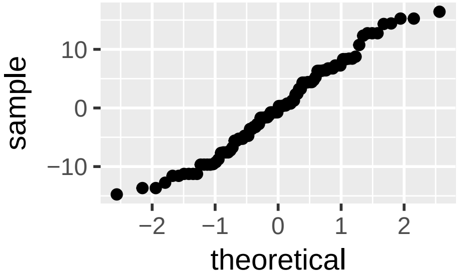
```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  3  0.5701 0.6361
##       92
```

hearing full example - II

Next, verify the model assumptions starting with Levene's test:

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group  3  0.5701 0.6361
##        92
```

. . . followed by the normal quantile plot of the residuals:



hearing full example - III

Next we do the overall ANOVA F test:

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## ListID        3    920  306.82   4.919 0.00325
## Residuals    92   5738   62.37
```

hearing full example - III

Next we do the overall ANOVA F test:

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## ListID        3    920  306.82   4.919 0.00325
## Residuals    92   5738   62.37
```

And since the p-value is low **we may proceed with the pairwise comparisons.**

hearing full example - IV

hearing full example - IV

To make the three confidence intervals we need the estimated mean differences and the group sample sizes:

```
## # A tibble: 4 × 3
##   ListID     n     mean
##   <fctr> <int>    <dbl>
## 1  List1    24 32.75000
## 2  List2    24 29.66667
## 3  List3    24 25.25000
## 4  List4    24 25.58333
```

We are doing three comparisons at an experimentwise error rate of 0.05, so we'll produce the $(1 - 0.05/3) \cdot 100\% = 98.33\%$ confidence intervals.

hearing full example - IV

To make the three confidence intervals we need the estimated mean differences and the group sample sizes:

```
## # A tibble: 4 × 3
##   ListID     n      mean
##   <fctr> <int>     <dbl>
## 1  List1    24  32.75000
## 2  List2    24  29.66667
## 3  List3    24  25.25000
## 4  List4    24  25.58333
```

We are doing three comparisons at an experimentwise error rate of 0.05, so we'll produce the $(1 - 0.05/3) \cdot 100\% = 98.33\%$ confidence intervals.

The value of $t_{92,0.0167}$ is 2.1604869.

hearing full example - V

The three pairwise comparisons of interest can be made using these confidence intervals:

Comparison	Estimate	Margin of Error	Lower	Upper
$\mu_1 - \mu_4$	7.167	4.926	2.241	12.092
$\mu_2 - \mu_4$	4.083	4.926	-0.842	9.009
$\mu_3 - \mu_4$	-0.333	4.926	-5.259	4.592

post-hoc comparison trick

Dangerous territory: perform a comparison *after looking at the data*.

post-hoc comparison trick

Dangerous territory: perform a comparison *after looking at the data*.

Trick: use Bonferroni's correction *assuming you were going to look at all the comparisons in advance.*

post-hoc comparison trick

Dangerous territory: perform a comparison *after looking at the data*.

Trick: use Bonferroni's correction *assuming you were going to look at all the comparisons in advance.*

With k groups there will be $k(k-1)/2$ such comparisons.