Multiple comparison procedures

Cavan Reilly

September 27, 2013

Table of contents

Overview

Power

Test multiplicity

Family-wise error rate

False discovery rate

Single step approaches

- Bonferroni adjustment
- Multiple comparisons in ANOVA

Test multiplicity

In the traditional application of statistical methods to data analysis one has a single primary outcome and the goal is to define a procedure for testing a hypothesis about that outcome.

For example, we may hypothesize that some drug lowers mortality, so we would design a study to test this hypothesis.

In practice there are also typically secondary outcomes which are also assessed but whose role is not as important as the primary outcome.

A positive finding for a secondary outcome would suggest another study whose primary outcome would be the past secondary outcome.

One could use the previous data for the secondary outcome to design a well powered study for this secondary outcome.

Power

Power is the probability that the null hypothesis will be rejected if the alternative hypothesis is true.

For example, if we simulate data so that the null hypothesis is true then we will reject the null hypothesis sometimes: if we use a significance level of 0.05 this will happen 5% of the time.

We can check this in R using a for loop.

Power

```
> pval <- rep(NA,1000)
> for(i in 1:1000){
+   y1 <- rnorm(20)
+   y2 <- rnorm(20)
+   pval[i] <- t.test(y1, y2, var.equal=TRUE)$p.value
+ }
> sum(pval<.05)/1000
[1] 0.046
```

Which is about 5%.

We can use the same approach to determine the power when the null hypothesis is not true.

Power

```
> for(i in 1:1000){
+   y1=rnorm(20)
+   y2=rnorm(20, mean=1)
+   pval[i]=t.test(y1, y2, var.equal=TRUE)$p.value
+ }
> sum(pval<.05)/1000
[1] 0.87
```

So this means that a study with 20 subjects per group and a primary outcome with mean 0 in one group and mean 1 in the other group (and a standard deviation of 1 in both groups) has about a 90% chance of finding a significant difference if one tests for a difference using a 2 sample *t*-test.

The more subjects one has in the study, or the greater difference between the 2 groups, the larger the power of the study.

Test multiplicity

If one has many outcomes and tests them all it is possible that one will mistakenly reject a null hypothesis when in fact it is true.

This is the primary reason why studies are designed with a single primary outcome.

In this context we are controlling what is called the *family-wise error rate*.

A type I error occurs when we falsely reject the null hypothesis, i.e. we say there is an effect or difference when in truth none exists.

A type II error occurs when we falsely fail to reject the null hypothesis, i.e. we fail to detect a difference.

Family-wise error rate

Suppose we are interested in testing m null hypotheses, denoted $H_0^1, \ldots, H_0^m$.

	Non-significant	Significant	
H_0	U	V	m_0
H_A	T	S	$m - m_0$
	$m - R$	R	m

Here V is the number of times a type I error is made while T is the number of times a type II error is made.

The family-wise error rate (FWER) is $P(V \geq 1)$.

Family-wise error rate

There are stricter definitions that condition on the complete set of nulls or only subsets.

In particular the *FWER under the complete null* is

$$P(V \geq 1 | H_0^1, \ldots, H_0^m)$$

whereas the *FWER under a partial null* is $P(V \geq 1)$ conditioning on some subset of the null hypotheses.

A test procedure has *strong control* of the FWER if the FWER is less than or equal to the level of the test under all partial nulls.

It is said to have *weak control* if the FWER is less than or equal to the level of the test conditioning on the complete set of nulls.

False discovery rate

The *false discovery rate* is the proportion of falsely rejected nulls among the set of nulls that are rejected.

In terms of the table we have that the FDR is just $\mathrm{E}\left(\frac{V}{R}\right)$.

Being more careful, we note that since $R = 0$ with positive probability we must define the FDR to be 0 when $R = 0$, hence we find that

$$\begin{aligned} \mathrm{FDR} &= \mathrm{E}\left(\frac{V}{R}\,\middle|\, R > 0\right) P(R > 0) + \mathrm{E}\left(\frac{V}{R}\,\middle|\, R = 0\right) P(R = 0) \\ &= \mathrm{E}\left(\frac{V}{R}\,\middle|\, R > 0\right) P(R > 0) \end{aligned}$$

False discovery rate

If we assume that all nulls are true then $V = R$ and so $\frac{V}{R}$ is 0 if $V = 0$ and it is 1 if $V > 0$, hence

$$\begin{aligned} \mathrm{E}\left(\frac{V}{R}\right) &= 0 \times P(V = 0) + 1 \times P(V \geq 1) \\ &= P(V \geq 1) \\ &= \mathrm{FWER} \end{aligned}$$

so that if all nulls are true then the FDR equals the FWER.

False discovery rate

Thus if we control the FDR (i.e. keep it less than some value) then we are controlling the FWER in the weak sense.

If not all of the null hypotheses are true, so that $V < R$, then $V/R < 1$ and so $\mathrm{E}(\frac{V}{R}|V \geq 1) < 1$ which gives us

$$\begin{aligned} \mathrm{E}\left(\frac{V}{R}\right) &= \mathrm{E}\left(\frac{V}{R} \,\middle|\, V = 0\right) P(V = 0) + \mathrm{E}\left(\frac{V}{R} \,\middle|\, V \geq 1\right) p(V \geq 1) \\ &= 0 \times P(V = 0) + \mathrm{E}\left(\frac{V}{R} \,\middle|\, V \geq 1\right) p(V \geq 1) \\ &< \mathrm{FWER}. \end{aligned}$$

So that the false discovery rate is less than the FWER in general.

Control of the FDR versus the FWER in the strong sense depends on the application: exploratory versus confirmatory.

Single step approaches

There are 2 types of algorithms that are used to control the FWER: single step procedures and step down procedures.

In a single step procedure the same criterion is used for all tests.

In a step down procedure the p-values are sorted and a different criterion is used for each test in the sorted arrangement.

The Bonferroni adjustment is the most widely used single step procedure.

To motivate this procedure, first assume that we control the level of each of our m tests by requiring

$$P(\text{reject } H_0^i | H_0^i \text{ true}) \leq \alpha$$

for some α, the level of the individual tests.

Single step approaches

Then we have

$$\begin{aligned}\text{FWER} &= P(V \geq 1 | H_0^1, \ldots, H_0^m)\\ &= 1 - P(V = 0 | H_0^1, \ldots, H_0^m).\end{aligned}$$

If we then assume the tests are independent

$$\begin{aligned}\text{FWER} &= 1 - \prod_{i=1}^{m} P(\text{do not reject } H_0^i | H_0^i)\\ &= 1 - \prod_{i=1}^{m} [1 - P(\text{reject } H_0^i | H_0^i)]\\ &\leq 1 - \prod_{i=1}^{m} (1 - \alpha)\\ &= 1 - (1 - \alpha)^m\end{aligned}$$

Single step approaches

Note that with $m = 2$ and $\alpha = 0.05$ we get a FWER of 0.0975, so that with just 2 independent tests each controlled at the conventional significance level we have almost doubled our chances of making a type I error!

With $m = 14$ this probability becomes over 0.50.

The Bonferroni adjustment changes the significance level for all tests.

Instead of using a significance level of α one just uses $\alpha' = \alpha/m$.

This is because $1 - (1 - \alpha)^m \approx m\alpha$ by a first order Taylor expansion for α near zero.

Single step approaches

As an example, we will test for associations between mutations in the protease gene and indinavir and nelfinavir fold resistance using the virco data set.

We will dichotomize the genotype data and only consider amino acids with at least 5% of the strains have an observed mutation.

```
> attach(virco)
> PrMut <- virco[,23:121]!="-" & virco[,23:121]!="."
> dim(PrMut)
[1] 1066 99
> NObs <- dim(virco)[1]
> PrMutSub <- data.frame(PrMut[,apply(PrMut,2,sum) > NObs*0.05])
> dim(PrMutSub)
[1] 1066 47
> Trait <- IDV.Fold-NFV.Fold
```

Bonferroni adjustment

Now we will write a function to get the the *p*-value from a 2 sample *t*-test and apply this to our data set.

```
> TtestP <- function(Geno){
+ return(t.test(Trait[Geno==1],Trait[Geno==0],na.rm=T)$"p.value")
+ }
> Pvec <- apply(PrMutSub,2,TtestP)
> sort(Pvec)
         P30          P76          P88          P55          P48          P89
3.732500e-12 9.782323e-10 1.432468e-06 2.286695e-06 5.749467e-06 8.924013e-05
         P11          P82          P60          P85          P54          P43
4.171618e-04 9.500604e-04 1.115441e-03 1.219064e-03 1.489381e-03 2.025621e-03
...
         P58          P62          P41          P12          P57
5.440101e-01 6.677043e-01 6.998280e-01 8.050362e-01 9.938846e-01
```

Bonferroni adjustment

So if we just use a significance level of 0.05 for each test we get the following sets of mutations.

```
> names(PrMutSub)[Pvec < 0.05]
 [1] "P11" "P14" "P30" "P32" "P33" "P35" "P43" "P46" "P47" "P48" "P54" "P55"

[13] "P60" "P61" "P67" "P69" "P76" "P82" "P84" "P85" "P88" "P89"
```

However if we use the Bonferroni adjustment we find fewer differences.

```
> PvecAdj <- p.adjust(Pvec,method="bonferroni")
> names(PrMutSub)[PvecAdj < 0.05]

[1] "P11" "P30" "P48" "P55" "P76" "P82" "P88" "P89"
```

Multiple comparisons in ANOVA

Historically, the first investigations into multiple hypothesis testing were motivated by *post-hoc* comparisons in ANOVA.

Recall, in ANOVA one tests the null hypothesis of no difference between the groups.

So if that hypothesis is rejected the natural question is, which groups differ and how.

Tukey addressed this question by determining the sampling distribution of the largest difference between means.

Scheffe took a different approach in which he considered every possible linear combination of the means, so his approach is more general.