Multiple Testing

by

AINDRILA GARAI

MSC STATISTICS, IIT KANPUR

aindrilag22@iitk.ac.in

Introduction:

We are often faced with huge amounts of data and consequently may wish to test many null hypotheses. When conducting multiple testing, we need to be very careful about how we interpret the results, in order to avoid erroneously rejecting far too many null hypotheses.

Eg. We might want to test m null hypotheses, $H_{01}, \ldots, H_{0m}$, where H_{0j} : the mean value of the jth biomarker among mice in the control group equals the mean value of the jth biomarker among mice in the treatment group.

Hypothesis Testing:

Hypothesis tests provide a rigorous statistical framework for answering simple "yes-or-no" questions about data.

Null and Alternative Hypotheses:

- First we define the null and alternative hypotheses. Null hypothesis is constructed in such a way so that we can reject it. The alternative hypothesis represents something different and unexpected.
- We use "we fail to reject H_0" means H_0 really holds or due to small sample size we fail to reject H_0 in which case testing H_0 again on a larger or higher-quality dataset might lead to rejection.

Test Statistic and p-value:

- Next, we construct a test statistic that summarizes the strength of evidence against the null hypothesis.
- Test statistics follow a well-known statistical distribution under the null hypothesis — such as a normal distribution, a t-distribution, a χ^2-distribution, or an F-distribution, provided that the sample size is sufficiently large.
- A large (absolute) value of a test statistic provides evidence against H_0.
- We then compute a p-value that quantifies the probability of having obtained a comparable or more extreme value of the test statistic under the null hypothesis. The p-value is defined as the probability of observing a test statistic equal to or

more extreme than the observed statistic, under the assumption that H_0 is in fact true.

- Finally, based on the p-value, we decide whether to reject the null hypothesis. A small p-value provides evidence against H_0.

Type I and Type II Errors:

Type I error occurs when H_0 is true but we reject it and Type II error occurs when H_0 is false but we do not reject it. The power of the hypothesis test is defined as the probability of not making a Type II error given that H_a holds, i.e., the probability of correctly rejecting H_0.

- Type I and Type II error rates can't be simultaneously small.
- In practice, we typically view Type I errors as more "serious" than Type II errors because the former involves declaring a scientific finding that is not correct.
- **Problem:** Rejecting a null hypothesis if the p-value is below α controls the probability of falsely rejecting that null hypothesis at level α. However, if we do this for m null hypotheses, then the chance of falsely rejecting at least one of the m null hypotheses is quite a bit higher.

```
x <- matrix ( rnorm (10 * 100), 10, 100)
x[,1:10] <- x[, 1:10] + 2
t.test (x[, 1], mu = 0) # One Sample t-test

##
##  One Sample t-test
##
## data:  x[, 1]
## t = 4.5032, df = 9, p-value = 0.001482
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.7448936 2.2487436
## sample estimates:
## mean of x
##  1.496819

p.values <- rep (0, 100)
for (i in 1:100)
p.values[i] <- t.test (x[, i], mu = 0)$p.value
decision <- rep ("Do not reject H0", 100)
decision[p.values <= .05] <- " Reject H0"
table (decision , c( rep ("H0 is False ", 50), rep ("H0 is True ", 50)))

##
## decision           H0 is False  H0 is True
##    Reject H0                10           3
##   Do not reject H0          40          47
```

The Family-Wise Error Rate:

The family-wise error rate (FWER) generalizes this notion to the setting of m null hypotheses, $H_{01}, \ldots, H_{0m}$, and is defined as the probability of making at least one Type I error.

A strategy of rejecting any null hypothesis for which the p-value is below α

$$\text{FWER}(\alpha) = 1 - \Pr\left(\bigcap_{j=1}^{m} \{\text{ do not falsely reject } H_{0j}\}\right)$$

If two events A and B are independent,

$$\text{FWER}(\alpha) = 1 - \prod_{j=1}^{m}(1-\alpha) = 1 - (1-\alpha)^m$$

Approaches to Control the Family-Wise Error Rate:

The Bonferroni Method:

$$\text{FWER}(\alpha) \leq \sum_{j=1}^{m} \Pr(A_j)$$

The Bonferroni method sets the threshold for rejecting each hypothesis test to α/m, so that $\Pr(A_j) \leq \alpha/m$. then, $\text{FWER}(\alpha) \leq m \times \frac{\alpha}{m} = \alpha$ so this procedure controls the FWER at level α

Holm's Step-Down Procedure:

In Holm's method it will reject more null hypotheses resulting in fewer Type II errors and hence greater power. Holm's method makes no independence assumptions about the m hypothesis tests and is uniformly more powerful than the Bonferroni method — it will always reject at least as many null hypotheses as Bonferroni — and so it should always be preferred.

Algorithm:

1. Specify α, the level at which to control the FWER.
2. Compute p-values, $p_1, \ldots, p_m$ for the m null hypotheses $H_{01}, \ldots, H_{0m}$.
3. Order the mp-values so that $p(1) \leq p(2) \leq \cdots \leq p(m)$.
4. $L = \min\left\{j : p_{(j)} > \frac{\alpha}{m+1-j}\right\}$
5. Reject all null hypotheses H_{0j} for which $p_j < p_{(L)}$.

Note:

- Controlling the FWER at level α guarantees that the data analyst is very unlikely (with probability no more than α) to reject any true null hypotheses, i.e. to have any false positives. When m is large, we may be willing to tolerate a few false positives i.e. more rejections of the null hypothesis.

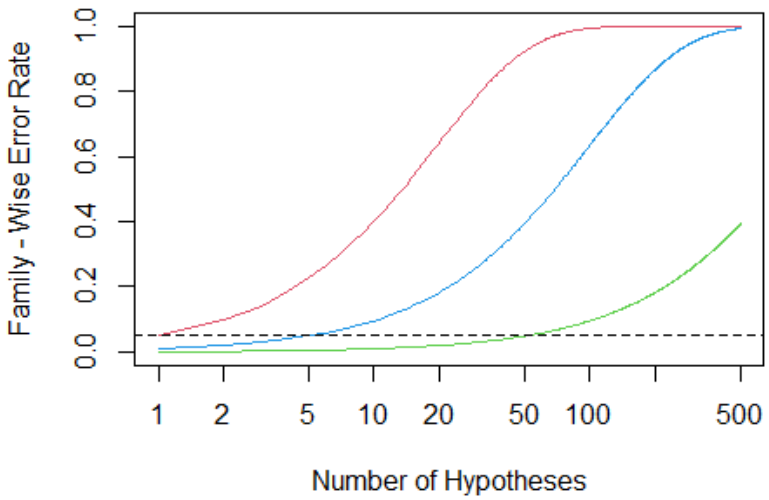
Tukey's method:

When performing $m = G(G-1)/2$ pairwise comparisons of G means, it allows us to control the FWER at level α while rejecting all null hypotheses for which the p-value falls below α_T , for some $\alpha_T > \alpha/m$.

Scheff´e's method:

It allows us to compute a value α_S such that rejecting the null hypothesis H_0 if the p-value is below α_S will control the Type I error at level α.

```
m <- 1:500
fwe1 <- 1 - (1 - 0.05)^m # FWER
fwe2 <- 1 - (1 - 0.01)^m
fwe3 <- 1 - (1 - 0.001)^m
plot (m, fwe1 , type = "l", log = "x", ylim = c(0, 1), col = 2,ylab = " Famil
y - Wise Error Rate ",
xlab = " Number of Hypotheses ")
lines (m, fwe2 , col = 4)
lines (m, fwe3 , col = 3)
abline (h = 0.05, lty = 2)
```



```r
library (ISLR2)
fund.mini <- Fund[, 1:5]
t.test (fund.mini[, 1], mu = 0)
```

```
##
##  One Sample t-test
##
## data:  fund.mini[, 1]
## t = 2.8604, df = 49, p-value = 0.006202
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.8923397 5.1076603
## sample estimates:
## mean of x
##         3
```

```r
fund.pvalue <- rep (0, 5)
for (i in 1:5)
fund.pvalue[i] <- t.test (fund.mini[, i], mu = 0)$p.value
round(fund.pvalue,2)
```

```
## [1] 0.01 0.92 0.01 0.60 0.76
```

```r
p.adjust (fund.pvalue , method = "bonferroni") # or "holm"
```

```
## [1] 0.03101178 1.00000000 0.05800491 1.00000000 1.00000000
```

```r
t.test (fund.mini[, 1], fund.mini[, 2], paired = T) # paired t-test
```

```
##
##  Paired t-test
##
## data:  fund.mini[, 1] and fund.mini[, 2]
## t = 2.128, df = 49, p-value = 0.03839
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  0.1725378 6.0274622
## sample estimates:
## mean difference
##             3.1
```

```r
returns <- as.vector ( as.matrix (fund.mini))
manager <- rep (c("1", "2", "3", "4", "5") , rep (50, 5))
a1 <- aov (returns ~ manager)
a1
```

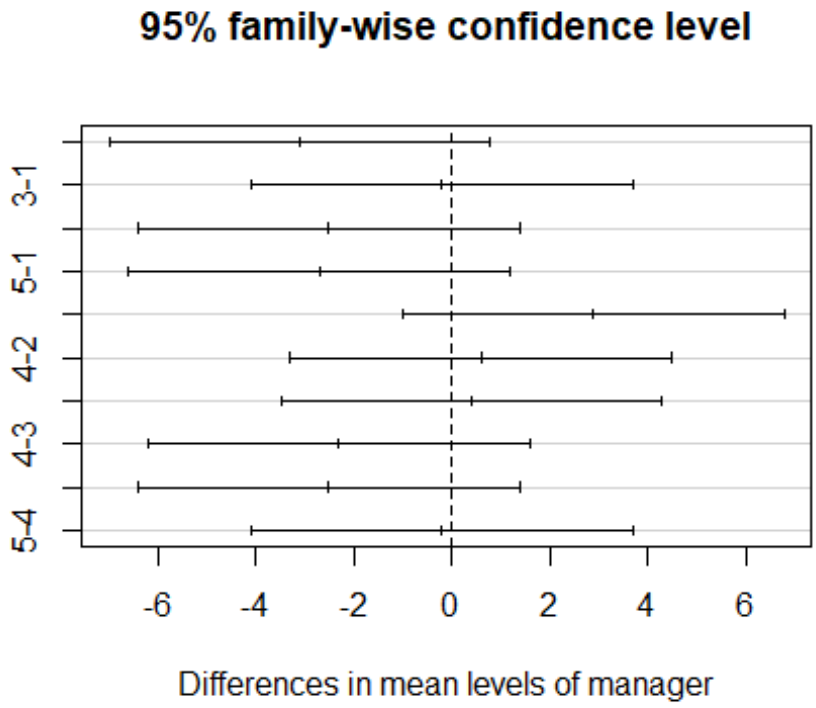
```
## Call:
##    aov(formula = returns ~ manager)
##
## Terms:
##                 manager Residuals
## Sum of Squares      437     12250
```

```
## Deg. of Freedom       4       245
##
## Residual standard error: 7.071068
## Estimated effects may be unbalanced

TukeyHSD (x = a1) #Tukey multiple comparisons of means

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = returns ~ manager)
##
## $manager
##     diff        lwr       upr     p adj
## 2-1 -3.1 -6.9865435 0.7865435 0.1861585
## 3-1 -0.2 -4.0865435 3.6865435 0.9999095
## 4-1 -2.5 -6.3865435 1.3865435 0.3948292
## 5-1 -2.7 -6.5865435 1.1865435 0.3151702
## 3-2  2.9 -0.9865435 6.7865435 0.2452611
## 4-2  0.6 -3.2865435 4.4865435 0.9932010
## 5-2  0.4 -3.4865435 4.2865435 0.9985924
## 4-3 -2.3 -6.1865435 1.5865435 0.4819994
## 5-3 -2.5 -6.3865435 1.3865435 0.3948292
## 5-4 -0.2 -4.0865435 3.6865435 0.9999095

plot ( TukeyHSD (x = a1))
```



Discovery Rate:

We might try to make sure that the ratio of false positives (V) to total positives (V + S = R) is sufficiently low, so that most of the rejected null hypotheses are not false positives. The ratio V /R is known as the false discovery proportion (FDP).

Controlling the FDP is an impossible task, since one doesn't know which hypotheses are true and which are false.

So, we define,

$$\text{FDR} = \text{E}(\text{FDP}) = \text{E}(V/R)$$

When we control the FDR at (say) level q = 20%, we are rejecting as many null hypotheses as possible while guaranteeing that no more than 20% of those rejected null hypotheses are false positives, on average.

The Benjamini-Hochberg Procedure:

We now focus on the task of controlling the FDR: that is, deciding which null hypotheses to reject while guaranteeing that the FDR, $E(V/R)$, is less than or equal to some pre-specified value q.

1. Specify q, the level at which to control the FDR.
2. Compute p-values, $p_1, \ldots, p_m$ for the m null hypotheses $H_{01}, \ldots, H_{0m}$.
3. Order the m p-values so that $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$.
4. $L = \max\{j: p_{(j)} < qj/m\}$.
5. Reject all null hypotheses H_{0j} for which $p_j \leq p_{(L)}$.

The rejection threshold used in the BenjaminiHochberg procedure is more complicated: we reject all null hypotheses for which the p-value is less than or equal to the Lth smallest p-value, where L is itself a function of all m p-values.

```
fund.pvalues <- rep (0, 2000)
for (i in 1:2000)
fund.pvalues[i] <- t.test (Fund[, i], mu = 0)$p.value
q.values.BH <- p.adjust (fund.pvalues , method = "BH")
round(q.values.BH[1:10],3)

##  [1] 0.090 0.991 0.122 0.923 0.956 0.075 0.077 0.075 0.075 0.075

sum (q.values.BH <= .1)

## [1] 146

ps <- sort (fund.pvalues)
m <- length (fund.pvalues)
q <- 0.1
```

```
wh.ps <- which (ps < q * (1:m) / m)
if ( length (wh.ps) >0) {
wh <- 1: max (wh.ps)
} else {
wh <- numeric (0)
}
wh.ps #less than or equal to the largest p-value
```

```
##  [1]  52  53  54  55  56  57  58  59  60  61  62  63  64  65  66  67  68
69  70
## [20]  71  72  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87
88  89
## [39]  90  91  92  93  94  95  96  97  98  99 100 101 102 103 104 105 106 1
07 108
## [58] 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 1
26 127
## [77] 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 1
45 146
```

A Re-Sampling Approach:

If our null hypothesis H0 or test statistic T is somewhat unusual, it may be the case that no theoretical null distribution is available. Even if a theoretical null distribution exists, then we may be wary of relying upon it, perhaps because some assumption that is required for it to hold is violated or the sample size is too small.

- **Re-Sampling p-Value for a Two-Sample t-Test:**

1. Compute $T = \frac{\hat{\mu}_X - \hat{\mu}_Y}{s\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}}$, on the original data $x_1, \ldots, x_{n_X}$ and $y_1, \ldots, y_{n_Y}$

2. For $b = 1, \ldots, B$, where B is a large number (e.g. $B = 10{,}000$):

(a) Permute the $n_X + n_Y$ observations at random. Call the first n_X permuted observations $x_1^*, \ldots, x_{n_X}^*$ and call the remaining n_Y observations

$$y_1^*, \ldots, y_{n_Y}^*$$

(b) Compute T on the permuted data $x_1^*, \ldots, x_{n_X}^*$ and $y_1^*, \ldots, y_{n_Y}^*$, and call the result T^{*b}.

3. The p-value is given by $\frac{\sum_{b=1}^{B} 1_{(|T^{*b}| \geq |T|)}}{B}$.

Note:

- In settings with a smaller sample size or a more skewed data distribution there is a substantial difference between the theoretical and re-sampling null distributions, which results in a difference between their p-values.

- The estimated FDR associated with the threshold c is $\hat{V}/R$. where $R = \sum_{j=1}^{m} 1_{(|T^{(j)}| \geq c)}$ and $\hat{V} = \frac{\sum_{b=1}^{B} \sum_{j=1}^{m} 1(|T^{(j),*b}| \geq c)}{B}$ these all are calculated on permuted data.

```
attach (Khan)
x <- rbind (xtrain , xtest)
y <- c(as.numeric (ytrain), as.numeric (ytest))
x <- as.matrix (x)
x1 <- x[ which (y == 2), ]
x2 <- x[ which (y == 4), ]
n1 <- nrow (x1)
n2 <- nrow (x2)
t.out <- t.test (x1[, 11], x2[, 11], var.equal = TRUE)
TT <- t.out$statistic
TT

##         t
## -2.093633

set.seed (1)
B <- 10000
Tbs <- rep (NA, B)
for (b in 1:B) {
dat <- sample (c(x1[, 11], x2[, 11]))
Tbs[b] <- t.test (dat[1:n1], dat[(n1 + 1):(n1 + n2)],
var.equal = TRUE
)$statistic
 }
mean (( abs (Tbs) >= abs (TT)))

## [1] 0.0416

hist (Tbs , breaks = 100, xlim = c(-4.2, 4.2), main = "", xlab = " Null Distr
ibution of Test Statistic ", col = 7)
```