# Multiple Testing

*Chapter 5, Lab 4: Solutions*

*OpenIntro Biostatistics*

**Topics**

- Experiment-wise error
- Controlling experiment-wise error for correlated data

The previous lab in this chapter introduced the Bonferroni correction as a method for controlling Type I error when conducting pairwise $t$-tests following an ANOVA. This lab more formally introduces the multiple testing problem and discusses one specific approach for controlling Type I error: controlling the **experiment-wise error rate**, the probability of making at least one Type I error in a set of hypothesis tests.

The material in this lab is an extension to Section 5.5 of *OpenIntro Biostatistics*.

## Introduction

Recall that making a Type I error, rejecting the null hypothesis when the null hypothesis is true, occurs with probability $\alpha$. The Type I error rate is controlled by rejecting the null hypothesis only when the $p$-value of a test is smaller than $\alpha$. For a single hypothesis test conducted at significance level $\alpha = 0.05$, there is a 5% chance of incorrectly rejecting the null hypothesis.

1. With each iteration of the following code, 100 values are randomly sampled from a standard normal distribution and a one-sample $t$-test of $H_0 : \mu = 0$ is conducted, with the $p$-value stored. The vector `reject` records whether the $p$-value for a test is smaller than $\alpha$.

   Run the simulation to obtain an empirical estimate of the Type I error rate when one hypothesis test is conducted. Confirm that the Type I error rate is approximately 5%.

   Out of 1,000 tests, the $p$-value was less than $\alpha$ 56 times, which represents a Type I error rate of 5.6%. This is reasonably close to 5%. The number of rejections represents Type I error because the sampled values are drawn from a distribution with $\mu = 0$; rejecting $H_0 : \mu = 0$ represents a decision error.

```
#set parameters
num.iterations = 1000
num.obs = 100
alpha = 0.05

#set seed
set.seed(2018)

#create empty list
p.values.A = vector("numeric", num.iterations)
```

```
#run simulation
for(k in 1:num.iterations){

  A = rnorm(num.obs)

  p.values.A[k] = t.test(A, mu = 0)$p.val

}

#view results
reject = (p.values.A <= alpha)
table(reject)
```

```
## reject
## FALSE  TRUE
##   944    56
```

When conducting more than one $t$-test, the significance level $\alpha$ used in each test controls the error rate for that test. The **experiment-wise error rate** is the chance that at least one test will incorrectly reject $H_0$ when all tested null hypotheses are true.

2. When two hypothesis tests are conducted at $\alpha = 0.05$, is the probability of making at least one Type I error equal to 0.05?

   a) Modify the simulation code to estimate the experiment-wise error rate (i.e., the 'overall' Type I error rate) when two hypothesis tests are conducted; let the vector `B` contain a set of 100 values randomly drawn from a standard normal distribution.

   The probability of making at least one Type I error is the probability that, for a specific iteration, either the $p$-value of the test conducted on `A` is less than $\alpha$ or the $p$-value of the test conducted on `B` is less than $\alpha$. The estimated experiment-wise error rate is 11%.

   The lab notes for this chapter demonstrate a more flexible approach to conducting this simulation that allows the number of tests to be specified as a parameter.

```
#set parameters
num.iterations = 1000
num.obs = 100
alpha = 0.05

#set seed
set.seed(2018)

#create empty list
p.values.A = vector("numeric", num.iterations)
p.values.B = vector("numeric", num.iterations)

#run simulation
for(k in 1:num.iterations){
  A = rnorm(num.obs)
  B = rnorm(num.obs)

  p.values.A[k] = t.test(A, mu = 0)$p.val
  p.values.B[k] = t.test(B, mu = 0)$p.val

}

#view results
reject = (p.values.A <= alpha | p.values.B <= alpha)
table(reject)
```

```
## reject
## FALSE  TRUE
##   890   110
```

b) Using an algebraic approach, calculate the probability of making at least one Type I error when conducting two hypothesis tests, if the null hypotheses in both cases are true. Assume independence between the tests.

Let $A$ represent the event of making a Type I error on one test, and $B$ represent the event of making a Type I error on the other test, where $P(A) = P(B) = \alpha = 0.05$. The probability of making at least one error is equal to the complement of the event that a Type I error is not made with either test. Mathematically, this is $1 - [P(A^C) \times P(B^C)] = 1 - (1 - 0.05)^2 = 0.0975$.

```
#use r as a calculator
alpha = 0.05
num.tests = 2

experiment.wise.error = 1 - (1 - alpha)^num.tests
experiment.wise.error
```

```
## [1] 0.0975
```

3. Suppose that 100 independent two-sample $t$-tests are conducted. What is the probability of at least one incorrect rejection of $H_0$ at $\alpha = 0.05$, given that in all cases there is no difference between the population group means?

When 100 independent $t$-tests are conducted, given that in all cases the null hypothesis of no difference is true, the experiment-wise error rate is 99.4%! Note that the calculation of experiment-wise error rate is identical between two-sample and one-sample tests: the key details are that the calculation is made under the null hypothesis for a specific value of $\alpha$, with a set number of tests.

```
#use r as a calculator
alpha = 0.05
num.tests = 100

experiment.wise.error = 1 - (1 - alpha)^num.tests
experiment.wise.error
```

```
## [1] 0.9940795
```

### Multiple testing in the Golub leukemia dataset

The Golub leukemia dataset was introduced in Chapter 1 (Lab 3). To investigate whether gene expression profiling could be a tool for classifying acute leukemia type, Golub and co-authors used DNA microarrays to measure the expression level of 7,129 genes from children known to have either acute myeloblastic leukemia (AML) or acute lymphoblastic leukemia (ALL). The goal of the experiment was to identify genes that are differentially expressed between individuals with AML versus ALL.

The analysis from Chapter 1 used a "data-driven" approach, searching for genes that appeared substantially differentially expressed relative to the distribution of differences in mean expression levels between AML and ALL patients. No claims were made regarding whether observed differences were more extreme than expected by chance alone.

In this lab, a hypothesis testing approach will be used to assess whether, for a particular gene $i$, there is significant evidence that the mean expression level among ALL patients is different from the mean expression level among AML patients. For simplicity, the analysis will be conducted on a subset of 100 genes rather than the full set of 7,129 genes.

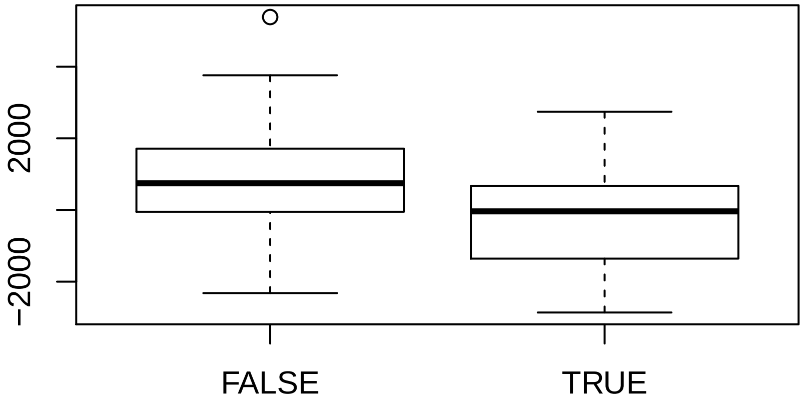
4. Run the following code to load the Golub data and prepare it for analysis; the code was introduced in Chapter 1, Lab 3.

   ```r
   #load the data
   library(oibiostat)
   data(golub)

   #remove phenotype information (in the first 6 columns)
   gene.matrix = as.matrix(golub[ , -(1:6)])

   #set parameters
   num.genes = ncol(gene.matrix)
   num.genes.subset = 100

   #create matrix with expression data from subset of 100 genes
   set.seed(2401)
   gene.index.set = sample(1:num.genes, size = num.genes.subset, replace = FALSE)
   gene.matrix.sample = gene.matrix[ , gene.index.set]

   #create logical variable for cancer type
   leuk.type = (golub$cancer == "aml")
   ```

   a) Produce a plot to show the association between cancer type and expression levels for the gene in the first column of `gene.matrix.sample`. Describe what you see. What do the values `TRUE` and `FALSE` mean?

   The value `TRUE` correspondes to AML patients, while `FALSE` corresponds to ALL patients. The boxplot shows that the gene in the first column has lower median expression in AML patients relative to ALL patients.

```r
#make a plot
boxplot(gene.matrix.sample[, 1] ~ leuk.type)
```



b) Conduct a $t$-test comparing expression levels between AML and ALL patients for the gene in the first column of `gene.matrix.sample`. Summarize the results.

The null hypothesis is that expression level for this gene does not differ between AML and ALL patients, $H_0 : \mu_{AML} = \mu_{ALL}$. The alternative is $H_A : \mu_{AML} \neq \mu_{ALL}$. Let $\alpha = 0.05$. The $p$-value is smaller than $\alpha$; there is sufficient evidence to reject the null hypothesis of no difference. The observed difference between the mean expression levels suggests that this gene is more highly expressed in ALL patients than AML patients.

```r
#conduct a t-test
t.test(gene.matrix.sample[, 1] ~ leuk.type, mu = 0, paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  gene.matrix.sample[, 1] by leuk.type
## t = 2.9239, df = 55.917, p-value = 0.004983
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   337.8189 1808.0940
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            844.9629           -227.9936
```

5. Run the following code to conduct 100 $t$-tests, comparing expression levels between AML and ALL patients for each of the 100 genes. The $p$-values are stored in the vector `p.vals`.

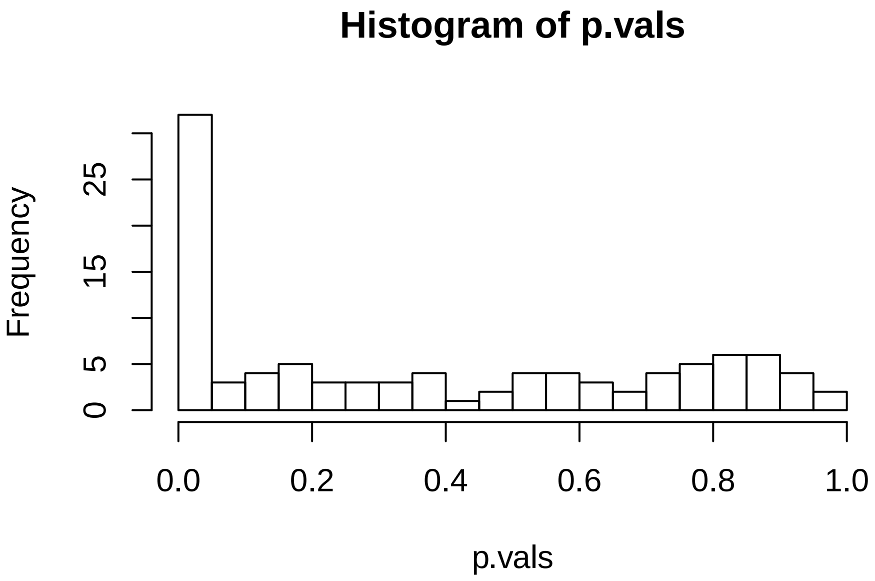
```
#set parameters
alpha = 0.05

#create empty vector to store results
p.vals = vector("numeric", num.genes.subset)

#conduct a t-test for each gene
for(k in 1:num.genes.subset){

  p.vals[k] = t.test(gene.matrix.sample[, k] ~ leuk.type,
                     alternative = "two.sided", mu = 0,
                     var.equal = FALSE, conf.level = 0.95)$p.val

}

#create table of results
gene.names = colnames(gene.matrix.sample)
results = cbind(gene.names, p.vals)

#is the p-value smaller than alpha?
reject = (p.vals <= alpha)

#view results
table(reject)
```

```
## reject
## FALSE  TRUE
##    68    32
```

a) Plot a histogram of the *p*-values. Describe what you see.

Out of the 100 genes, there are over 30 with a *p*-value lower than 0.05. The rest of the *p*-values are scattered somewhat evenly between 0.05 and 1.

```
#plot a histogram
hist(p.vals, breaks = 15)
```



b) The logical vector `reject` has value `TRUE` if the $p$-value of a test is smaller than $\alpha = 0.05$.

i. Of the 100 $t$-tests conducted, how many result in rejecting the null hypothesis at $\alpha = 0.05$? In other words, how many of the 100 genes are identified as being significantly differently expressed between AML and ALL patients?

Of the 100 $t$-tests conducted, 32 resulted in rejecting the null hypothesis at $\alpha = 0.05$; 32 genes are identified as having mean expression levels that are significantly different between AML and ALL patients.

ii. Do you think it is reasonable to expect that for all the genes identified as significantly differently expressed between AML and ALL patients, there is actually an association between expression level and leukemia type? In other words, is it reasonable to expect that for each rejection, the null hypothesis was correctly rejected?

No, it seems likely that at least some of these observed differences occurred by chance, and that the null hypothesis was incorrectly rejected for some of the tests.

6. One approach to controlling the experiment-wise error rate is the Bonferroni correction. Using the Bonferroni correction, how many genes out of the 100 are identified as being significantly differently expressed between AML and ALL patients?

In the Bonferroni correction, each test is conducted at the $\alpha^\star = \alpha/K$ level, where $K$ is the number of comparisons. With this stricter significance level, only 7 of the 100 genes are identified as being significantly differentially expressed between AML and ALL patients.

```
#define parameter
alpha.star = alpha/num.tests

#is the p-value smaller than alpha.star?
reject.bonf = (p.vals <= alpha.star)
table(reject.bonf)
```

```
## reject.bonf
## FALSE  TRUE
##    93     7
```

From theory, it can be shown that the Bonferroni correction is too strict when comparisons are not independent, inflating the rate of false negatives (i.e., failing to reject when there is a difference between groups). In this setting, independence is not a realistic assumption; from a biological perspective, the expression level of each gene is unlikely to be completely independent of the expression level of another gene.

7. The following simulation estimates the experiment-wise error rate for 100 tests conducted on the Golub gene expression data under the null hypothesis that there is no difference between the population mean expression of AML versus ALL patients.

   The simulation is based on randomly sampling values from a distribution related to the normal distribution (specifically, the multivariate normal) in which the variance is described not by a single value, but a matrix of values known as a covariance matrix. This matrix can be thought of as measuring the pairwise correlation between expression levels of each of the 100 genes being tested. Further details of the multivariate normal distribution are beyond the scope of this course.

```
#load package for using multivariate normal distribution
library("MASS")

#set seed
set.seed(2401)

#set parameters
num.patients = nrow(golub)
num.replicates = 1000
alpha = 0.05

#set parameters for multivariate normal distribution
theoretical.means = rep(0, len = num.genes.subset)
estimated.correlation = cov(gene.matrix.sample)

#create empty vectors to store results
gene.p.vals = vector("numeric", num.genes.subset)
min.p.vals = vector("numeric", num.replicates)

#run the simulation
for(k in 1:num.replicates){

  gene.expression.matrix = mvrnorm(num.patients, mu = theoretical.means,
                              Sigma = estimated.correlation)

  for(j in 1:num.genes.subset){

gene.p.vals[j] = t.test(gene.expression.matrix[, j] ~ leuk.type)$p.val
```

```
  }

  min.p.vals[k] = min(gene.p.vals)

}

#view results
reject = (min.p.vals <= alpha)
table(reject)
```

```
## reject
## FALSE  TRUE
##    58   942
```

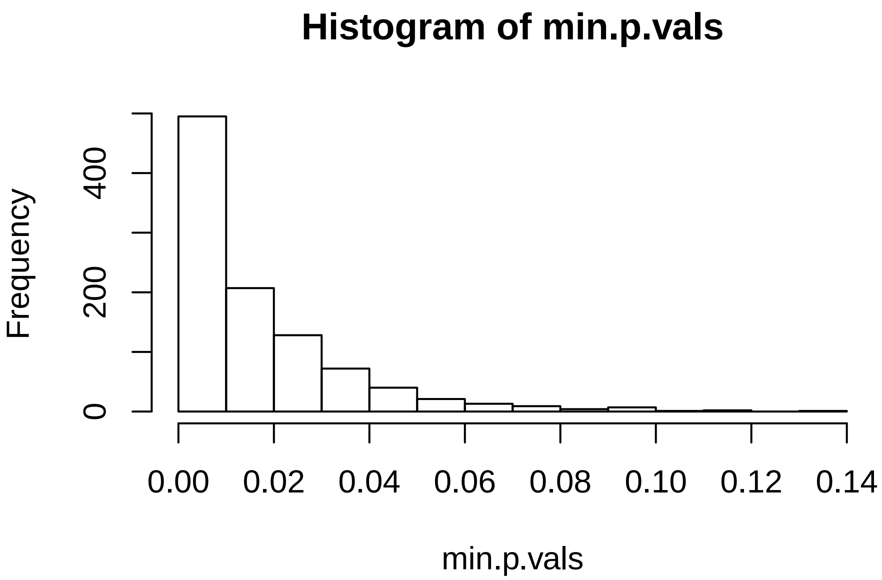
a) Run the simulation. What does the vector `gene.p.vals` contain, versus `min.p.vals`?

The vector `gene.p.vals` contains the *p*-values for each set of 100 tests conducted on the simulated expression data. The vector `min.p.vals` contains the minimum *p*-value of the 100 *p*-values for each iteration.

b) Create a histogram of `min.p.vals`. Describe the section of the histogram that denotes the estimated experiment-wise error.

If the smallest *p*-value out of a set of 100 tests is less than $\alpha$, this represents an instance of experiment-wise error, since experiment-wise error refers to at least one incorrect rejection of the null. The section of the histogram with values below 0.05 represent the estimated experiment-wise error.

```
#create a histogram
hist(min.p.vals)
```



c) From the simulation, what is the estimated experiment-wise error rate? Does this differ from the answer to Question 3?

From the simulation, the estimated experiment-wise error rate is 94.2%. This is lower than the experiment-wise error rate calculated under independence in Question 3, which is 99.4%.

d) How does the estimated experiment-wise error change as $\alpha$ decreases from 0.05? Test values of 0.005 and 0.001 for $\alpha$.

As $\alpha$ decreases from 0.05, the estimated experiment-wise error rate decreases. When $\alpha$ is 0.005, the experiment-wise error rate is 30.8%. When $\alpha$ is 0.001, the experiment-wise error rate is 7.6%.

```
## reject
## FALSE  TRUE
##   692   308
```

```
## reject
## FALSE  TRUE
##   924    76
```

e) Run the following code to find the value for $\alpha$ that will yield an experiment-wise error of approximately 0.05. Compare this value to the $\alpha^\star$ value recommended by the Bonferroni correction.

The value for $\alpha$ that yields an experiment-wise error of approximately 0.05 is 0.0006. This value is larger than the $\alpha^\star$ value from the Bonferroni correction.

```
sim.based.correction = quantile(min.p.vals, 0.05)
sim.based.correction
```

```
##           5%
## 0.0006411709
```

f) Why is it not advisable to use an overly strict correction for $\alpha$?

Using an overly strict correction for $\alpha$ necessarily increases the Type II error rate, the rate of incorrectly failing to reject the null hypothesis.

g) Using the simulation-based adjustment to $\alpha$, how many genes out of the 100 are identified as being significantly differentially expressed between AML and ALL patients?

Using the simulation-based adjustment, 8 genes out of the 100 are identified as being significantly differentially expressed between AML and ALL patients.

```
#is the p-value smaller than sim.based.correction?
reject.sim = (p.vals <= sim.based.correction)
table(reject.sim)
```

```
## reject.sim
## FALSE  TRUE
##    92     8
```