Multiple Testing

gustavoc@msu.edu

In this note I briefly review the concepts of power and type-I error rate. I will first consider the problem of conducting a single test and then extend the framework to multiple tests conducted simultaneously. The recommended reading for this module is Chapter 15 of the book Computer Age Statistical Inference by Efron and Hastie (2017).

(1) Power & Type-I error rate with a single test

Consider a single test for which there are two possible states of nature (H_0 or H_a) and two possible decisions (reject/do not reject, **Table 1**).

Table 1: Classification of decisions in hypothesis testing.

	Do not reject H_0	Reject H_0
H_0 holds	True Negative (N_1)	False Positive (Type-I error) (N_2)
H_a holds	False Negative (Type-II error) (N_3)	True positive (N_4)

Suppose we repeat our experiment a large number of times, each time re-sampling data from the population and recording whether we reject or don't reject.

If H_0 holds, the proportion of times we wrongly reject H_0 (type-I errors) is the type-I error rate:

- **Type-I error rate**: $P(reject\ H_0|H_0\ holds) = E[\frac{N_2}{N_1+N_2}]$

On the other hand, if H_A holds, the proportion of time we reject is the power of the experiment

- **Power**: $P(reject\ H_0|Ha\ holds)E[\frac{N_4}{N_3+N_4}]$

In hypothesis testing, we tune our decision rule (e.g., reject if p-values is smaller than some threshold, α) to control Type-I error rate at a low level (say $\alpha = 0.05$).

If p-values are correct (i.e., if the assumptions used to derive p-values hold) and we conduct a single test each time rejecting if the p-value is smaller than 0.05, we expect a Type-I error rate of 5%.

(2) Family-Wise Error Rate (FWER)

Suppose now that we test p hypothesis (H_{0_1} vs H_{A_1}, H_{0_2} vs H_{A_2},..., H_{0_p} vs H_{A_p}). Out of these hypothesis, typically only a small fraction H_A's will hold. Our goal is to identify for which of the hypotheses that we test we have evidence to reject H_0.

When testing multiple hypotheses, the Family-Wise Error Rate (**FWER**, aka Experiment-wise error rate) is the probability of wrongly rejecting at least one of the null hypotheses tested.

Consider for example testing two hypotheses (H_{0_1} vs H_{A_1} and H_{0_2} vs H_{A_2}). Recall that the probability of the union of two events is $P(A\ or\ B) = P(A) + P(B) - P(A \cap B)$. Therefore, if H_{0_1} and H_{0_2} hold and we test each of these hypotheses with a significance level equal to α, the probability of making at least one mistake is:

$P(reject\ H_{0_1}\ or\ H_{0_2}) = P(reject\ H_{0_1}) + P(reject\ H_{0_2}) - P(reject\ H_{0_1}\ and\ H_{0_2}) = 2\alpha - P(reject\ H_{0_1}\ and\ H_{0_2}) \leq 2\alpha$ **[1]**

If both tests are independent, $P(reject\ H_{0_1}\ and\ H_{0_2}|H_{0_1}\ and\ H_{0_2}) = \alpha^2$ leading to $P(reject\ H_{0_1}\ or\ H_{0_2}|H_{0_1}\ and\ H_{0_2}) = 2\alpha - \alpha^2$. Thus, in general, if we reject at a significance level $\tilde{\alpha} = \alpha/2$, we would be controlling the probability of making at least one mistake at a rate $\leq \alpha$.

Bonferroni correction

To keep the FWER at a level smaller than α, the Bonferroni correction method rejects each hypothesis at a level $\tilde{\alpha} = \alpha/p$ where p is the number of hypothesis tested.

This procedure is conservative because, by eq. [1], the type-I error rate is expected to be lower than α, this can be particularly conservative if the test-statistics are positively correlated (i.e., if the probability of rejecting both tests is non negligable).

The following example illustrate the use of Boferroni correction for two un-correlated tests.

```
n=500
Rsq=0 # use 0 (1) for two independent ((perfectly correlated) tests
nRep=100 # number of MC reps

PVAL=matrix(nrow=nRep,ncol=2,NA)

for(i in 1:nRep){
  x1=rnorm(n)
  x2=x1*sqrt(Rsq)+rnorm(n,sd=sqrt(1-Rsq))

  y=rnorm(n) # both nulls are true

  fm1=lm(y~x1)
  fm2=lm(y~x2)
  PVAL[i,1]=summary(fm1)$coef[2,4]
  PVAL[i,2]=summary(fm2)$coef[2,4]
}
```

Say our target FWER is 0.1, if we do not adjust by multiple testing, we have that the FWER is larger than 0.1

```
reject1=PVAL[,1]<0.1
reject2=PVAL[,2]<0.1

table(reject1,reject2)
```

```
##         reject2
## reject1 FALSE TRUE
##   FALSE    75   11
##   TRUE     14    0
```

```
FWER=mean(reject1|reject2)
FWER
```

```
## [1] 0.25
```

However, if we use Bonferroni's method the FWER is controled at a level below 0.1.

```
reject1=PVAL[,1]<0.1/2
reject2=PVAL[,2]<0.1/2
```

```
  FWER=mean(reject1|reject2)
  FWER
```

```
## [1] 0.11
```

If we re-run the example, this time using postively correlated tests (Rsq between predictors of 0.8 in the example below), we have that the FWER is much lower than the target FWER (0.1), this implies that our test could be overly conservative.

```
n=500
 Rsq=0.8 # use 0 (1) for two independent ((pefectly correlated) tests

 PVAL=matrix(nrow=nRep,ncol=2,NA)
 P.ADJ=PVAL
 P.ADJ.holm=PVAL
 for(i in 1:nRep){
   x1=rnorm(n)
   x2=x1*sqrt(Rsq)+rnorm(n,sd=sqrt(1-Rsq))

   y=rnorm(n) # both nulls are true

   fm1=lm(y~x1)
   fm2=lm(y~x2)
   PVAL[i,1]=summary(fm1)$coef[2,4]
   PVAL[i,2]=summary(fm2)$coef[2,4]

   P.ADJ[i,]=p.adjust(PVAL[i,],method='bonferroni')
   P.ADJ.holm[i,]=p.adjust(PVAL[i,],method='holm')

 }

 reject1=PVAL[,1]<0.1/2
 reject2=PVAL[,2]<0.1/2
 FWER=mean(reject1|reject2)
 FWER
```

```
## [1] 0.13
```

There are two equivalent ways of applying Bonferroni's method:

- Used un-adjusted p-values and reject whenever $p-value < \tilde{\alpha} = \alpha/p$
- Adjust the p-values using $pAdj = min(1, pval \times p)$ and reject if $pAdj < \alpha$.

Both procedures are of-course equivalent. The second approachis implemented in the `p.adjust()` function. The use of this function is illustrated in the example presented above.

To avoid having an overly conservative decision rule, one possiblity is to replace p in $\tilde{\alpha} = \alpha/p$ with an estimate of the number of independent tests.

Holm's method

A slighlty less conservative method (Holm's method) works as follows:

- Sort p-values from smallest to largest.
- Compare each p-value with the significance level $\tilde{\alpha}_i = \alpha/(p-i)$ (p here is the number of tests conducted, and i is the order (from sallest to largest) of the p-value).

This method is also implemented in `p.adjust(,method="holm")`.

```r
pVals=c(.1,.2,.015,.01)
cbind( p.adjust(pVals,method='bonferroni') , p.adjust(pVals,method='holm'))
```

```
##      [,1]  [,2]
## [1,] 0.40 0.200
## [2,] 0.80 0.200
## [3,] 0.06 0.045
## [4,] 0.04 0.040
```

```r
# Reject?
cbind( p.adjust(pVals,method='bonferroni') , p.adjust(pVals,method='holm'))<0.05
```

```
##       [,1]  [,2]
## [1,] FALSE FALSE
## [2,] FALSE FALSE
## [3,] FALSE  TRUE
## [4,]  TRUE  TRUE
```

In practice, Holm's method provides only a small power improvement relative to Bonferroni, while preserving FWER control.

(3) False discovery rate

Many modern statistical analyses requires conducting a very large number of tests. For instance, in genetic studies we may need to test the association between a phenotype and potentially millions of genetic markers (e.g., single nucleotide polymorphisms, SNPs). When the number of tests is very large, controlling Family-Wise Error Rate (i.e., targeting a very low probability of making at most one mistake) leads to overly conservative tests, thus reducing power. Thus, an alternative is to use a decision rule for rejection that control the expected proportion of mistakes among the discoveries.

Suppose we test $N = N_1 + N_2 + N_3 + N_4$ hypothesis, and assume the number of true negatives, false positives, false negatives, and true positives are given by N_1, N_2, N_3, and N_4, respectively (see **Table 1**). The total number of discoveries is $N_2 + N_4$, of these, N_2 are false discoveries; thereore, we can define the false discovery proportion as

- **False Discovery Proportion** is $FDP = N_2/(N_2 + N_4)$.

The false discovery rate, is the expected value of FDP over conceptual repeated sampling, that is

- **False Discovery Rate (FDR)** $FDR = E[N_2/(N_2 + N_4)]$.

For any given decision rule and data we can obsreve the total number of rejections $N_2 + N_4$. At first glance, estimating how many of these are likely to be rejections of true nulls (N_2) is not straightoforward because we don't know the true state of nature. However, a simple procedure by Benjaminy and Hocberg (1995) can be used to define a decision rule with adequate FDR control.

Benjamini-Hocberg (BH) procedure:

- Sort the p-values from smallest to highest.
- For the sorted p-vaues compute $[i/p] \times \alpha$, where i is the order of the p-value ($i = 1$ for the smallest p-value), p is the number of tests conducted and α is the FDR-threshold ($\alpha \in [0, 1]$).
- Reject if $pValue[i] < [i/p] \times \alpha$.

If all the tests conducted are independent and all originate from null hypothesis, the BH procedure controls FDR at the desired level (α); in practice some proportion (typically a very small one when the number of hypothesis tested is very large) of the tests originate from alternative hypothesis. If the proportion of tests that originates from null hypothesis is π_0 and tests are independent, the BH procedure controls FDR at $\pi_0\alpha < \alpha$. The function `p.adjust(,method='fdr')` adjust p-values using the BH procedure, after adjustment

we simply reject for all adjuste p-values $< \alpha$ (e.g., $q = 0.05$). The following example illustrates the use of the BH procedure, in the example 5% of the hypothesis originate from H_a's.

```
pH0=0.95
nTests=5000
n=1000 # sample size
pVals=rep(NA,nTests)
isHA=runif(nTests)>pH0
varB=.03 #  variance explained if Ha holds

for(i in 1:nTests){
  x=rnorm(n)
  y=rnorm(n)
  if(isHA[i]){
    y=y+x*rnorm(1,sd=sqrt(varB)) # adding an effect if Ha
  }
  pVals[i]=summary(lm(y~x))$coef[2,4]
}

pADJ.Bonf=p.adjust(pVals,method='bonferroni')
pADJ.Holm=p.adjust(pVals,method='holm')
pADJ.FDR=p.adjust(pVals,method='fdr')
```

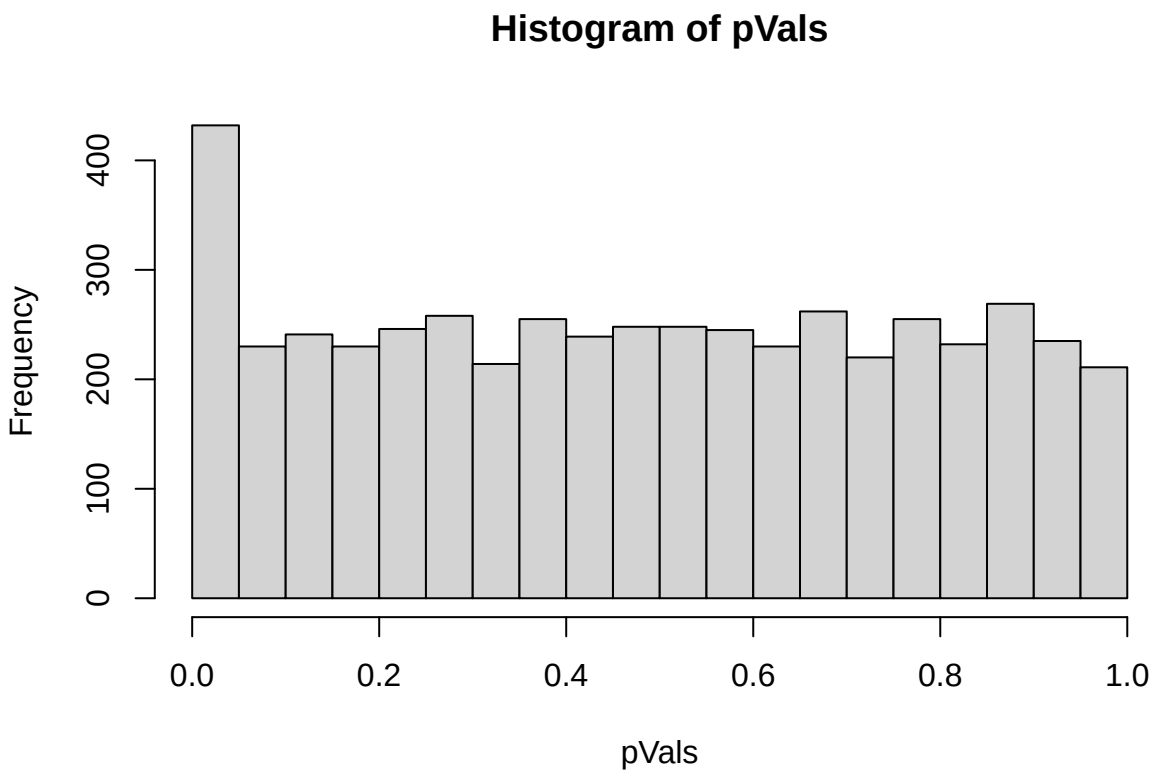
If we use a FDR for rejection equal to 0.05, then the false discovery proportion in the single monte carlo replicate was

```
mean((pADJ.FDR<0.05)[!isHA])
```

```
## [1] 0.001260769
```

Recall that under the null hypothesis, the distribution of the p-values is uniform. What is the empirical distribution of the p-values when some proprotion of the tests originate from alternative hypotheses? Instead of uniform in this case we see an increase in the frequency of low-pvalues, the majority of which originate from alternative hypothesis.

```
hist(pVals,30)
```

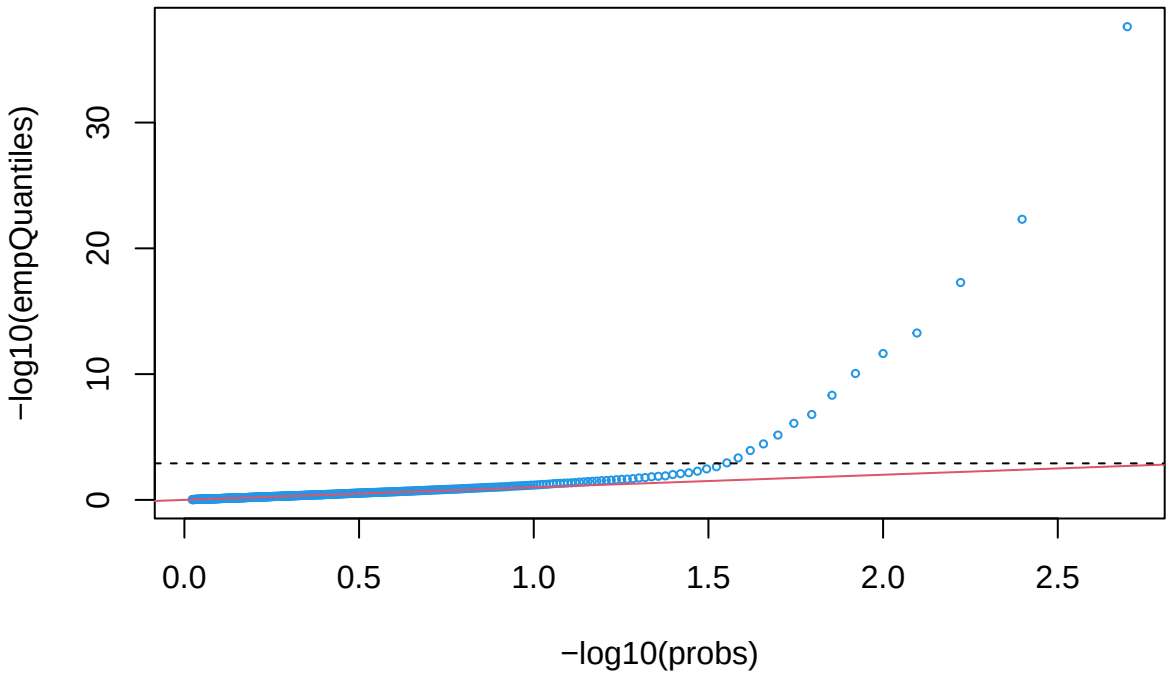


(4) Quantile-Quantile plots (QQ-plot)

A standard approach for exhamining p-values in large-scale hypothesis testing problems is to plot the empirical quantiles of the p-values agains the empirical quantiles of the uniform distribution (the distribution of pvaleus under the null). If all the tests originate from the null, the two quantiles should be similar (i.e., align in the 45-degree line when we plot one against the other). When a number of test originate from the alternative we expect that the p-values obtained are 'enriched' for low pavalues, thus we expect some departure from the 45 degree line in the QQ-plot. Because we want to zoom-in for enrichemnt of small pvalues, QQ-plots are often done in the -log10() scale (large value is indicative of small pvalue). The following example uses the qqplot() function to display the pvalues of the previous example.

```
probs=seq(from=1/500,to=.95,by=1/500)
empQuantiles=quantile(pVals,probs)
plot(-log10(empQuantiles),x= -log10(probs),cex=.5,col=4)
abline(a=0,b=1,col=2)

abline(h=-log10(max(pVals[pADJ.FDR<.05])),lty=2) # this is the threshold (in -log10 scale) implied by t
```


−log10(empQuantiles)
0
10
20
30
0.0
0.5
1.0
1.5
2.0
2.5
−log10(probs)