MATH 7343 Applied Statistics

Prof. (Aidong) Adam Ding

Northeastern University

Review

- Last time, we started on ANOVA
- (1)Test for an overall difference. $\mathbf{H_0}$: $\mu_1=\cdots=\mu_k=\mu$
- F-test: Reject H_0 if $F_{obs}=\frac{MSB}{MSE} > F_{k-1,n-k,\alpha}$
- (2) Pinpoint the difference. $\mathbf{H_0}$: contrast $L=\lambda_1\mu_1+\ldots+\lambda_k\mu_k=0$
- T-test: Reject H_0 if $T_{obs}=\left|\frac{\hat{L}}{\sqrt{MSE\sum_{i=1}^{k}\frac{\lambda_i^2}{n_i}}}\right| > t_{n-k,\alpha/2}$
- For multiple contrasts, we need multiple testing adjustment. We continue from here.

Need for Multiple testing adjustments

- How to do multiple contrasts? Each one can use t-test. However, multiple testing adjustments are needed if we are doing more than one contrast.
- If I test two hypothesis $\boldsymbol{H_0^1}$: L_1=0 and $\boldsymbol{H_0^2}$: L_2=0 on the same data set each at α level. When both $\boldsymbol{H_0}$'$\boldsymbol{s}$ are true, we have α probability to reject each one. But

P(Falsely reject at least one of them) = P(Reject L_1=0 or Reject L_2=0)

= P(Reject L_1=0) + P(Reject L_2=0 but not L_1=0)

= α + P(Reject L_2=0 but not L_1=0) **> α**

Need for Multiple testing adjustments

- For two independent $H_0{}'s$ (Orthogonal contrasts) each tested at α=0.05 level, the overall error rate is

P(Reject L_1=0 or Reject L_2=0)

= 1- P(Not reject L_1=0 and Not reject L_2=0)

= 1- P(Not reject L_1=0) P(Not reject L_2=0)

= 1- (0.95)(0.95) = 1-0.09025

= 0.0975 **> 0.05**

- With more hypothesis, the error rate becomes bigger.
- Generally, without independence, no exact formula.

Family-wise error rate (FWER)

- Example: we conduct 5 experiments and do 20 hypothesis tests in each. All H_0's are in fact true.

Experiments	**1**	**2**	**3**	**4**	**5**	**Total**
Number of tests	20	20	20	20	20	100
Number of rejections	1	3	2	4	0	10

- Total 10 out of 100 tests reject. So the comparison wise error rate is 10% (significance level).
- But 4 out 5 experiments report wrong conclusions!
- FWER = 80% here!
- We need to control FWER.

Family-wise error rate (FWER)

- The familiar-wise error rate (FWER) is also called Tukey's experimental error rate.
- In above example, we can see that if we control each test level at 10%, 80% of experiments can report something wrong!
- Nowadays big data especially need to consider this risk. (E.g. Genome analysis often scans millions of SNPs in thousands of genes for disease markers.)
- We need methods to control the FWER!

Multiple testing adjustment for FWER

- (1) **Bonferroni** correction.
- If we plan to test m hypothesis, then test each one at α/m level instead.

P(Reject $\boldsymbol{H_0^1}$ or $\boldsymbol{H_0^2}$ or… or $\boldsymbol{H_0^m}$)

≤ P(Reject $\boldsymbol{H_0^1}$) + P(Reject $\boldsymbol{H_0^2}$) +…+ P(Reject $\boldsymbol{H_0^m}$)

= α/m + α/m +…+ α/m

= α

- Notice: The m tests need to be pre-planned, not decided post-hoc (i.e., based on data).

Multiple testing adjustment for FWER

- (2) **Scheffé**'s method (control FWER for all contrasts)
- For each contrast L=$\lambda_1\mu_1+\ldots+\lambda_k\mu_k$,

  Reject H_0 if $T_{obs} = \left|\frac{\lambda_1\hat{\mu}_1+\ldots+\lambda_k\hat{\mu}_k}{\sqrt{MSE\sum_{i=1}^{k}\frac{\lambda_i^2}{n_i}}}\right| > \sqrt{(k-1)F_{k-1,n-k,\alpha}}$
- Notice that, without any multiple testing adjustment, the original ANOVA t-test for the contrast use $t_{n-k,\alpha/2}$ as cutoff points. And $t_{n-k,\alpha/2} = \sqrt{F_{1,n-k,\alpha}}$.

Multiple testing adjustment for FWER

- (3)* **Tukey**'s **HSD** (Honestly significant difference) for all pairwise comparisons.
- The difference between two group means is real if

$$|\bar{X}_i - \bar{X}_j| > Q_{v,k,\alpha}\sqrt{\frac{s^2}{2}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$ where v is the degree of freedom in s, usually = n-k.

- While the textbook did not cover the Tukey's Q distribution. We can use R to find its probability by ptukey(. , nmeans=k, df=v), and its quantile by qtukey(. , nmeans=k, df=v)

Multiple testing adjustment for FWER

- (4)* **LSD** (Least significant difference) for pairwise comparisons.
- The difference between two group means is real if

$$|\bar{X}_i - \bar{X}_j| > t_{v,\alpha/2}\sqrt{s^2\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$ where v is the degree of freedom in s, usually = n-k.

- This is in fact the ANOVA t-test <u>without adjustment</u>.

- This t-test without adjustment could be used in exploratory studies (not confirmatory studies) which looks for possible group differences for further investigation.

Multiple testing adjustment

- (5) Control false discovery rate (FDR) instead of FWER.
- FDR = proportion of false discoveries among all discoveries. (For hypothesis test, a discovery means that H_0 is rejected.)
- Test m hypothesis. Put the numbers in following table

Decision	H_0 true	H_A true	Total
Fail to reject	U	T	m-R
Reject	V	S	R
	m_0	$m-m_0$	m

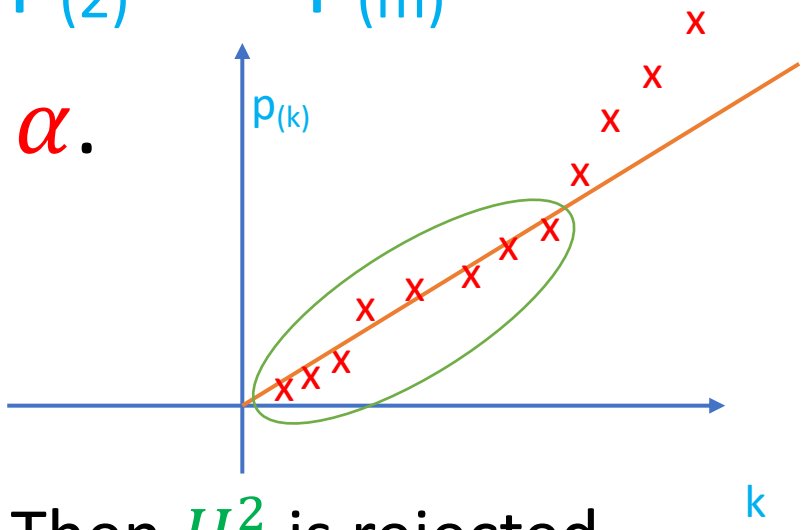
- Then FDR = E(V/R). In contrast, FWER=P(V≥1)
- Notice FDR=FWER when $m=m_0$.

Multiple testing adjustment

- (5)^ Control FDR instead of FWER.
- **Benjamini-Hochberg** procedure:

  m planned hypothesis tests H_0^1, H_0^2 , …, H_0^m at FDR=α.
- (a) Sort the p-values p_1, …, p_m as $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$
- (b) Find the largest k such that $p_{(k)} \leq \frac{k}{m}\alpha$.

Reject all H_0^i for i=1,…,k.



- Notice it is possible that $p_{(2)} > \frac{2}{m}\alpha$ but $p_{(3)} \leq \frac{3}{m}\alpha$. Then H_0^2 is rejected.

Using R to do ANOVA with multiple testing adjustment

- Example in ANOVA.pdf handout. Last time we got R outputs for three contrasts: μ_1-μ_2, μ_2-μ_3 and 2μ_1-(μ_2+μ_3)

```
             Estimate Std. Error t value Pr(>|t|)
centercontr1 -0.40631    0.16726  -2.429   0.0183 *
centercontr2  0.15380    0.16408   0.937   0.3525
centercontr3 -0.65881    0.27443  -2.401   0.0197 *
```

For Bonferroni:
m=3 contrasts, so compare p-values with α/m instead. At α=0.05 level, we need to compare with 0.05/3=0.0167.
Now all 3 contrasts p-values are bigger than 0.0167, thus none is significant.

For Scheffe:
Calculate the cutoff (α=0.05) as

```
> sqrt((k-1)*qf(0.95,df1=k-1,df2=n-k)) #Scheffe coeficient at alpha=0.05
[1] 2.513501
```

None of the t statistics exceed 2.51, thus none is significant. The answer is same as that using Bonferroni adjustment.

Using R to do ANOVA with multiple testing adjustment

For planned contrasts, say these three contrasts here. Check the output

```
             Estimate Std. Error t value Pr(>|t|)
centercontr1 -0.40631    0.16726  -2.429   0.0183 *
centercontr2  0.15380    0.16408   0.937   0.3525
centercontr3 -0.65881    0.27443  -2.401   0.0197 *
```

For Tukey:

The first two contrasts are in fact pairwise comparisons and can use the Tukey's correction. So calculate the cutoff (α=0.05) as

```
> qtukey(0.95,nmeans=k, df=n-k)*contr1.se/sqrt(2)
[1] 0.4024881
```

Since the difference is 0.40631, bigger than the cutoff, the first contrast is significant. This differs from the conclusion from the Bonferroni and Scheffe's adjustment above. Notice that the Tukey's adjustment assume that only pairwise comparisons are considered, thus excluding the third contrast above.

Using R to do ANOVA with multiple testing adjustment

- For **all** pair-wise comparisons, R have pre-programmed these adjustments. Use pairwise.t.test() and TukeyHSD().

```
> ## Pairwise comparisons
> # No adjustment or LSD
> pairwise.t.test(PF.data$fev1, g=PF.data$center, p.adjust.method = 'none')
        Pairwise comparisons using t tests with pooled SD
data:  PF.data$fev1 and PF.data$center
  1     2
2 0.018 -
3 0.102 0.353
P value adjustment method: none
> # Bonferroni for the k(k-1)/2 pairs
> pairwise.t.test(PF.data$fev1, g=PF.data$center, p.adjust.method = 'bonferroni')
        Pairwise comparisons using t tests with pooled SD

data:  PF.data$fev1 and PF.data$center

  1     2
2 0.055 -
3 0.307 1.000

P value adjustment method: bonferroni
```

- LSD: $\mu_1 \neq \mu_2$. Bonferroni: no pair are significantly different.

Using R to do ANOVA with multiple testing adjustment

- For pair-wise comparisons, R have pre-programmed these adjustments. Use pairwise.t.test() and TukeyHSD().

```
> # Tukey's HSD
> TukeyHSD(PF.fit, confidence.level=0.95)
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = fev1 ~ center, data = PF.data)

$center
           diff           lwr       upr     p adj
2-1  0.4063095  0.003821453 0.8087976 0.0473852
3-1  0.2525052 -0.113573610 0.6185840 0.2294901
3-2 -0.1538043 -0.548652567 0.2410439 0.6191128
```

At α=0.05 level, the difference between first and second groups (centers here) is significant under the Tukey's adjustment, but not significant under Bonferroni

Using R to do ANOVA with multiple testing adjustment

- For pair-wise comparisons, R have pre-programmed these adjustments. Use pairwise.t.test() and TukeyHSD().

```
> # FDR or BH
> pairwise.t.test(PF.data$fev1, g=PF.data$center, p.adjust.method = 'fdr')
        Pairwise comparisons using t tests with pooled SD

data:  PF.data$fev1 and PF.data$center

  1     2
2 0.055 -
3 0.154 0.353
P value adjustment method: fdr
```

- No pair are significantly different under FDR=0.05 level.
- Notice p.adjust.method =‘fdr’ and =‘bh’ gives exactly same output

Other R commands in the handout (anova.pdf)

```
We extracted the MSE using
MSE<-summary(PF.fit)[[1]][2,'Mean Sq'] #extract MSE

How do I know to extract it this way? Look at what is contained in summary(PF.fit)
> sum1<-summary(PF.fit)
> str(sum1)
List of 1
 $ :Classes ‘anova’ and 'data.frame': 2 obs. of  5 variables:
  ..$ Df     : num [1:2] 2 57
  ..$ Sum Sq : num [1:2] 1.58 14.48
  ..$ Mean Sq: num [1:2] 0.791 0.254
  ..$ F value: num [1:2] 3.12 NA
  ..$ Pr(>F) : num [1:2] 0.052 NA
 - attr(*, "class")= chr [1:2] "summary.aov" "listof"
 - attr(*, "na.action")=Class 'omit'  Named int [1:3] 12 30 37
  .. ..- attr(*, "names")= chr [1:3] "12" "30" "37"

It is a list, the first element in a list is indexed as [[1]], which is a data.frame here:
> sum1[[1]]
            Df  Sum Sq Mean Sq F value Pr(>F)
center       2  1.5828 0.79142  3.1153  0.052 .
Residuals   57 14.4803 0.25404
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
The MSE is in the second row (Residuals) and in the third column, so we can extract it using either of the two w
ays below:
> sum1[[1]][2,'Mean Sq']
[1] 0.2540396
> sum1[[1]]['Residuals','Mean Sq']
[1] 0.2540396
```

ANOVA with multiple testing adjustment

- Which adjustment to use in practice?
- Notice that (1) Bonferroni and (2) Scheffé are for contrasts; but (3)* Tukey's HSD is for only pairwise comparisons.
- (4)* LSD is in fact no adjustment. It is out if need adjustment.
- If multiple hypothesis include contrasts other than pairwise comparison, then use (1) Bonferroni or (2) Scheffé. Either is fine, however for m big, (2) Scheffé is usually better at detecting the difference.
- FDR can be used. But it is controlling a different quantity than FWER, not in the same category as other methods.

When to use multiple testing adjustment?

- Example: **After seeing data**, I observe that group 1 and group 2 means are very different and decides to test $\mathbf{H_0}$:μ_1=μ_2.

  Do I have to adjust? And How? If I use Bonferroni with m=1, same as no adjustment. So it appears that we do not need adjustment here?

- Answer: We DO need to do multiple testing adjustment and m is NOT one here. Since you saw the data and then choose the hypothesis, potentially you are testing all (infinitely many) hypothesis!
- Tests chosen after seeing data is called post-hoc contrasts.
- We can only use *Scheffé* or *Tukey* to control FWER for *post-hoc testing*.

Post Hoc testing need adjustments

- Sweepstake Example: Jeff suspects that there is some cheating arrangement so that Wonderboy will get the jackpot in the end. He decides to do a statistical test $\mathbf{H_0}$: no cheating versus $\mathbf{H_A}$: cheating.
- Data: Wonderboy does win in the next drawing with just one entry.

  p-value = P(Wonderboy wins | equal chance for everyone)

  = 1/(25 million)

  So we reject $\mathbf{H_0}$. This is a reasonable *pre-planned* hypothesis test.
- However, if after each sweepstake drawing you decide to test $\mathbf{H_0}$: the winner did not cheat. The p-value without adjustment is always 1/(25 million).

  So all winners are cheaters!

- Post hoc testing p-value need multiple testing adjustment!
  Need methods which adjust for ALL possible hypothesis (Scheffé/Tukey)

Basic experimental design

- ANOVA assumption: k independent groups; within each group have i.i.d data ~ N(μ_i,σ^2).
- Say we have k·n experimental units, how to get ANOVA data?
- **Completely randomized (CR) design**: Randomly allocate n observations to each treatment level.
- **Randomized blocks (RB) design**: groups of similar experimental units (blocks) are formed, and the experimental units are randomly allocated to treatment levels within a block. That is, n blocks each with k observations, randomly put the k observations one each to the k treatment levels.

ANOVA with blocking

- Example: We study if preheating milk increases cheese yield.

Lot number	20°C	60°C	70°C	80°C	Lot means
1	2.89	2.95	3.10	3.23	3.04
2	2.86	3.20	3.03	3.18	3.07
3	3.18	3.06	3.15	3.18	3.14
4	2.92	3.15	3.26	3.32	3.16
5	3.09	3.25	3.22	3.26	3.21
Treatment means	2.99	3.12	3.15	3.23	

Overall mean $\bar{X} = \frac{1}{kn}\sum_{i,j} X_{i,j} = 3.12$

- This data comes from a randomized blocks design with the lots as blocks. The yields in each block are more similar.

ANOVA with blocking

- Cheese example. One-way ANOVA(ignoring blocks) CR design

Sources	DF	SS	MS	F	Pr (>F)
Treatment	3	0.1569	0.0523	4.589	0.0168
Residuals (Error)	16	0.1824	0.114		
Total	19	0.3393			

Cutoff $F_{3,16}$ =3.24 < F_{obs}. So reject $\mathbf{H_0}$.

- ANOVA with blocks. RB design

Sources	DF	SS	MS	F	Pr (>F)
Treatment	3	0.1569	0.0523	5.73	0.0114
Block	4	0.0729			
Error (Residuals)	12	0.1095	0.0091		
Total	19	0.3393			

Cutoff$F_{3,12}$ =3.49 < F_{obs}. So reject $\mathbf{H_0}$.

ANOVA with blocking

- Compare the decomposition for sums of squares in one-way ANOVA without and with blocking:

  a treatment levels and b blocks

One-way (CR)	RB	DF	
Treatment	Treatment	a-1	
Error	Blocks	b-1	ab-a
	Error	(a-1) (b-1)	
Total	Total	ab-1	

- In the cheese example, $\frac{MSE_{CR}}{MSE_{RB}}=\frac{0.114}{0.0091}=1.25$

Without blocking, 25% more variance. Thus C.I. will be longer by ≈11%.

ANOVA with blocking

- The ANOVA with blocking basically analyze the variance due to two factors: treatment and blocks, but then ignore blocking variable and focus on treatment variable only. If the blocking variable is of equal importance as the treatment variable, then we should do the full two-way ANOVA:

Treatment A: a levels

Treatment B: b levels

Each combination (ab in total) has n measurements.

Source		DF
Treatment	A (factor effect)	a-1
	B (factor effect)	b-1
	A x B (interaction)	(a-1) (b-1)
Error		(n-1)ab
Total		nab-1

- In the cheese example, n=1 and no degree of freedom left for MSE. Then use the AxB part of MS as the error MS.

Summary

Today, we went over more ANOVA

- Multiple testing adjustment: Bonferroni, Scheffé (all contrasts) and Tukey's HSD (all pairwise comparisons) controls FWER. The BH procedure controls FDR.
- Pre-planned tests, all above adjustment methods can be used.
- Post-hoc tests, need Scheffé or Tukey's HSD.
- ANOVA with blocking. To get ANOVA with blocking, using the two-way ANOVA and focus on the treatment effects only.
- How to do multiple adjustment with R. For ANOVA with blocking R commands see twoanova.pdf (next lecture)
- Homework 5 is due on Monday the week after next.