MATH 5910

Multiple Comparisons

Multiple Comparisons

- Simultaneous inference.
- Corrections of multiplicity.
- Identify the significant components.
- Introduce some modern ideas.

More Than 2 Groups

- Example: 3 groups (populations).
- Data:

A	B	C
4, 2, 5, 2, 3	6, 3, 5, 3, 6	5, 4, 6, 7, 6, 5, 7, 2, 3, 5

- Want to see if all 3 group means are equal.
- Already know: Use ANOVA
- Other ways to look at this?

Multiple Comparisons

- There are different ways to do this.
- Can perform pairwise comparisons.
- How many comparisons?

Multiple Comparisons

- Compare: A-B, A-C, B-C
- There are 3 comparisons.
- Or mathematically,

$$\binom{3}{2} = \frac{3!}{2!1!} = 3$$

- In general, if there are k groups to compare, then there are

$$\binom{k}{2} = \frac{k!}{2!(k-2)!} \quad \text{comparisons.}$$

Example

Enter data:

```
> Y<-c(4,2,5,2,3,6,3,5,3,6,5,4,6,7,6,5,7,2,3,5)
> X<-c(rep("A",5),rep("B",5),rep("C",10))
```

Plot:

```
> boxplot(Y~X)
```

Example

Try

```
> t.test(Y~X)
Error in t.test.formula(Y ~ X) :
  grouping factor must have exactly 2 levels
```

Doesn't work.

Example

Do pairwise comparisons (with suppressed output)

```
# A vs. B
> t.test(Y[1:10]~X[1:10])
t = -1.5652, df = 7.824, p-value = 0.157

# A vs. C
> t.test(Y[c(1:5,11:20)]~X[c(1:5,11:20)])
t = -2.311, df = 10.001, p-value = 0.04343

# B vs. C
> t.test(Y[6:20]~X[6:20])
t = -0.4692, df = 8.685, p-value = 0.6505
```

Example

Same as before

```
t.test(Y[X=='A'|X=='B']~X[X=='A'|X=='B'])
t.test(Y[X=='A'|X=='C']~X[X=='A'|X=='C'])
t.test(Y[X=='B'|X=='C']~X[X=='B'|X=='C'])
```

Example

- Comparing A vs C yields a p-value of 0.04.
- Conclude significant difference between A and C at $\alpha = 0.05$?
- And conclude significant difference of A-B-C?
- NO to both
- Need adjustment.

Bonferroni

Bonferroni adjustment

- Simplest adjustment.
- Divide α by the number of comparisons.
- In this case, $\alpha/3$, or $0.05/3 = 0.0167$
- Now, all 3 p-values greater than 0.0167.
- No significant difference

Bonferroni

Equivalently, can also do:

- Multiply p-value by the number of comparisons (3 in this case)
- Then compare it with α.
- The adjusted p-value: $p = 3 \times 0.04 = 0.12$.
- Same conclusion.
- If the adjusted p-value bigger than 1: usually set to 1.

Bonferroni

If there are k groups.

- Divide α by $\binom{k}{2}$
- Examples:

```
> choose(3,2)
[1] 3

> choose(4,2)
[1] 6

> choose(10,2)
[1] 45
```

Bonferroni

- Bonferroni: Simple, but can be problematic (why?)
- Other solutions to this issue

ANOVA

A proper way to assess more than 2 groups

```
# Same

> anova(lm(Y~X))

> summary(aov(Y~X))
```

May still need a follow up...

Follow Up

- ANOVA, follow-up.
- High-dimensional/massive variable dataset (concepts).

Idea

- Once you determine significance, overall.
- Natural to look for where the difference occurs.
- But needs correction because you are testing many things at once.

ANOVA Example

- Look back to ANOVA.
- If we have Y (continuous response) and X (factor with several levels), run ANOVA.
- Suppose that X is significant, meaning there is difference in mean between levels of X.
- Which levels are different?

ANOVA Example

- Many ways to solve this multiple comparison.
- Easier to do in SAS.
- Can still do in R.
  - Tukey's method (all pairwise comparisons).
  - Function: `TukeyHSD()`

ANOVA Example

In R

```
> warpbreaks
   breaks wool tension
1      26    A       L
2      30    A       L
3      54    A       L
4      25    A       L
5      70    A       L
6      52    A       L
7      51    A       L
8      26    A       L
9      67    A       L
10     18    A       M
:
54     28    B       H
```

ANOVA Example

From help file `?TukeyHSD`

```
> summary(fm1 <- aov(breaks ~ wool + tension,
                     data = warpbreaks))
            Df Sum Sq Mean Sq F value   Pr(>F)
wool         1  450.7  450.67  3.3393 0.073614 .
tension      2 2034.3 1017.13  7.5367 0.001378 **
Residuals   50 6747.9  134.96
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Note the `aov()` function.

ANOVA Example

Same as

```
> anova(lm(breaks ~ wool + tension, data = warpbreaks))
Analysis of Variance Table

Response: breaks
          Df Sum Sq Mean Sq F value   Pr(>F)
wool       1  450.7  450.67  3.3393 0.073614 .
tension    2 2034.3 1017.13  7.5367 0.001378 **
Residuals 50 6747.9  134.96
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Either case, `tension` is significant.

ANOVA Example

Note that `tension` has 3 levels (L-M-H), so where different?

ANOVA Example

```
> TukeyHSD(fm1, "tension", ordered = TRUE)
  Tukey multiple comparisons of means
    95% family-wise confidence level
    factor levels have been ordered

Fit: aov(formula = breaks ~ wool + tension, data = warpbreaks)

$tension
         diff        lwr      upr     p adj
M-H  4.722222 -4.6311985 14.07564 0.4474210
L-H 14.722222  5.3688015 24.07564 0.0011218
L-M 10.000000  0.6465793 19.35342 0.0336262
```

Difference in L-H and L-M.

ANOVA Example

- See `?TukeyHSD` for more details.
- The help example also gives you a graph.

ANOVA Example

Others

- Simultaneous C.I.
  - Scheffé.
  - Tests for all "contrasts" - flexible.
  - Inferior to Tukey for pairwise comparison.
- Bonferroni
- Other methods.

Modern Usage

- Previous examples: Good for several comparisons.
- Modern: Massive number of comparisons.
- Algorithm

Modern Usage

So far

- Bonferroni, Tukey, Scheffé - controls for family-wise error rate (FWER).
- FWER: This is the probability of (falsely) rejecting at least one true null hypothesis, so we want this to be small.
- Potential problems (why?)

FWER

For example

- Let m be the number of total comparisons (e.g., the number of genes).
- If we set α, the probability of false rejection of H_0 (the Type I error).
- Then the expected number of false rejections is $\alpha \cdot m$
- So, if $\alpha = 0.05$ and $m = 10,000$, then you would reject $\alpha \cdot m = 500$ components even if all H_0 is true.
- The Bonferroni correction for FWER: Divide α by m to preserve the FWER of α (or multiply p-values by m).

FWER

Consider the following table:

	"Accept" H_0	Reject H_0	
H_0 True	U	V	m_0
H_0 Not true	T	S	$m - m_0$
	$m - R$	R	m

FWER

- m: # of total comparisons.
- m_0: # of true null hypotheses (unknown).
- R: # of total rejections (random but observable).
- V: # of Type I errors.
- T: # of Type II errors.
- U, S: Correct decisions, want to minimize V and T.

FWER

So

- $\boxed{\text{FWER} = P(V \geq 1)}$
- That is, the probability of (falsely) rejecting at least one true null hypothesis.
- We control FWER by setting FWER $= P(V \geq 1) \leq \alpha$.

FDR

False discovery rate (FDR).

- Modern method.
- More comlpicated.
- Concept.

FDR

Go back to

	"Accept" H_0	Reject H_0	
H_0 True	U	V	m_0
H_0 Not true	T	S	$m - m_0$
	$m - R$	R	m

FDR

- Whereas FWER $= P(V \geq 1)$.
- $\boxed{\text{FDR} = E(V/R)}$
- i.e., FDR is the expected proportion of false rejections.
- So FDR only concerns the rejected hypotheses.

FDR

Comparison with FWER:

- FWER tends to give too many non-rejections, esp. if m is large (conservative).
- FDR is more powerful (more power to detect the true difference) compared to FWER.
- But, FDR is more complicated (algorithmic) and limited/misunderstood.

More Information

Some websites

https://www.itl.nist.gov/div898/handbook/prc/section4/prc47.htm

https://www.publichealth.columbia.edu/research/population-health-methods/false-discovery-rate

https://www.stat.cmu.edu/~genovese/talks/hannover1-04.pdf