Two-way ANOVA, II

9.07

4/29/2004

Post-hoc comparisons & two-way analysis of variance

Post-hoc testing

- As before, you can perform post-hoc tests whenever there's a significant F_{obt}
  - But don't bother if it's a main effect and has only two levels – you already know the answer
- We'll just talk about the Tukey's HSD procedure
  - Requires that the n's in all levels of a factor are equal

Post-hoc testing for main effects

- This is just like post-hoc testing for the one-way ANOVA

Post-hoc testing for main effects is just like what we did for one-way ANOVA

$$HSD_\alpha = q_\alpha \sqrt{\frac{MS_{wn}}{n}}$$

α is the Type I error rate (.05).

q_α Is a value from a table of the studentized range statistic based on alpha, df_W, and *k*, the number of levels *in the factor you are testing*

MS_{wn} Is the mean square within groups.

n Is the number of people in each group. I.E. *how many numbers did you average to get each mean you are comparing?*

Our example from last time

- What effect do a workbook and coffee consumption have on exam performance?
- Both main effects and the interaction were significant
- Factor A (the workbook) had only two levels. No post-hoc testing required. The workbook helps.
- Factor B (the coffee) had three levels. We need to do post-hoc testing.

Numbers from our example last time

- $MS_{wn} = 205.56$
- $n = 6$
- q_k is a function of df_{wn} and k
  - $df_{wn} = 12$
  - $k = 3$
  - So, from the table, $q_k = 3.77$ for α=0.05

Cups of coffee (Factor B)	0	10, 30, 20	20, 45, 55	$\Sigma x = 180$, $n_{B1} = 6$
	1	45, 50, 85	40, 60, 65	$\Sigma x = 345$, $n_{B2} = 6$
	2	30, 40, 20	90, 85, 75	$\Sigma x = 340$, $n_{B3} = 6$

HSD for this example

- HSD = q_k sqrt(MS_{wn}/n)
  = 3.77 sqrt(205.56/6) = 22.07
- Differences in means:

Level 1: 0 cups	Level 2: 1 cup	Level 3: 2 cups
m_1=30	m_2= 57.5	m_3= 56.7

Level 1 vs Level 2: 27.5; Level 2 vs Level 3: 0.9; Level 1 vs Level 3: 26.7

- 0 cups of coffee differ significantly from both 1 and 2 cups of coffee

Post-hoc testing for the interaction

- Involves comparing cell means
- But we don't compare every possible pair of cell means…

		Workbook (Factor A) No	Workbook (Factor A) Yes
Cups of coffee (Factor B)	0	m = 20	m = 40
	1	m = 60	m = 55
	2	m = 30	m =83.33

Confounded & unconfounded comparisons

		Workbook (Factor A) No	Workbook (Factor A) Yes
Cups of coffee (Factor B)	0	m = 20	m = 40
	1	m = 60	m = 55
	2	m = 30	m =83.33

Confounded comparison, because the cells differ along more than one factor.
If there's a difference, what's the explanation?
Is it because of factor A or B? We can't tell, because there's a *confound*.

Confounded & unconfounded comparisons

		Workbook (Factor A) No	Workbook (Factor A) Yes
Cups of coffee (Factor B)	0	m = 20	m = 40
	1	m = 60	m = 55
	2	m = 30	m =83.33

(1) (2)

Unconfounded comparisons. The cells differ only in one factor. We can test these with post-hoc tests.

Tukey's HSD for interactions

- 1. Compute HSD = q_k sqrt(MS_{wn}/n)
  - Before, q_k was a function of df_{wn} and k, the number of levels in the factor of interest = # of means being compared
  - For the interaction, we use an *adjusted* k to account for the actual number of *unconfounded* comparisons (as opposed to all comparisons of cell means, some of which are confounded)
- 2. Compare with unconfounded differences in means

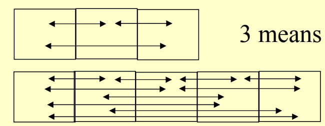
Table from the handout

TABLE 14.8 - Values of Adjusted k		
Design of Study	Number of Cell Means in Study	Adjusted Value of k
2 x 2	4	3
2 x 3	6	5
2 x 4	8	6
3 x 3	9	7
3 x 4	12	8
4 x 4	16	10
4 x 5	20	12

Figure by MIT OCW.

What's going on here?

- k is sort of short hand for the number of means you'd like to compare
- In one-way ANOVA or main effects analysis, e.g:

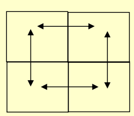


3 means -> 2+1 = 3 comparisons

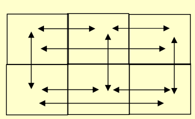
5 means -> 4+3+ 2+1 = 10 comparisons

What's going on here?

- Two-way interactions



2x2 -> 4 comparisons, k=3 is closest



2x3 -> 9 comparisons, k=5 is closest

Note

- Not all stat books bother with this adjusted value of k – many just use k = # cell means

Back to our example

- We had a 3x2 design, so the adjusted value of k = 5. df_{wn} = 12. So q_k = 4.51 for α=0.05
- MS_{wn} = 205.56, n = # in each mean = 3, so HSD = 4.51 sqrt(205.56/3) = 37.33
- What unconfounded comparisons lead to differences larger than 37.33?

		Workbook (Factor A)				
		No		Yes		
Cups of coffee (Factor B)	0	m = 20	20	m = 40	15	
	1	m = 60	15	m = 55		43.33
	2	m = 30	53.33	m =83.33	28.33	

Differences within No: 40 (0 vs 1), 30 (1 vs 2), 10 (0 vs 2)

		Workbook (Factor A)	
		No	Yes
Cups of coffee (Factor B)	0	m = 20	m = 40
	1	m = 60	m = 55
	2	m = 30	m =83.33

		Workbook (Factor A)		
		No	Yes	
Cups of coffee (Factor B)	0	m = 20	m = 40	m=30
	1	m = 60	m = 55	m=57.5
	2	m = 30	m =83.33	m=56.7
		m=36.67	m=59.44	

All significant effects shown (blue = interaction, green = main).

What is the interpretation of these results?

		Workbook (Factor A)		
		No	Yes	
Cups of coffee (Factor B)	0	m = 20	m = 40	m=30
	1	m = 60	m = 55	m=57.5
	2	m = 30	m =83.33	m=56.7
		m=36.67	m=59.44	

Interpretation:

1. If the interaction is not significant, interpretation is easy – it's just about what's significant in the main effects.

In this case, with no significant interaction, we could say that 1 or 2 cups of coffee are significantly better than 0 cups, and using the workbook is significantly better than not using it.

Workbook (Factor A)

Cups of coffee (Factor B)		No	Yes	
	0	m = 20	m = 40	m=30
	1	m = 60	m = 55	m=57.5
	2	m = 30	m =83.33	m=56.7
		m=36.67	m=59.44	

Interpretation:

2. However, if there is a significant interaction, then the main interpretation of the experiment has to do with the interaction.

Would we still say that 1 or 2 cups of coffee are better than 0? That using the workbook is better than not using it?

NO. It depends on the level of the other factor.

Workbook (Factor A)

Cups of coffee (Factor B)		No	Yes	
	0	m = 20	m = 40	m=30
	1	m = 60	m = 55	m=57.5
	2	m = 30	m =83.33	m=56.7
		m=36.67	m=59.44	

Interpretation:

- Increasing coffee consumption improves exam scores, where without the workbook there's an improvement going from 0 to 1 cups, and with the workbook there's an improvement in going from 0 to 2 cups.
- The workbook leads to significant improvement in exam scores, but only for students drinking 2 cups of coffee.

Within-subjects (one-way) ANOVA

Within-subjects experimental design

- Also known as "repeated-measures"
- Instead of having a bunch of people each try out one tennis racket, so you can compare two kinds of racket (between-subjects), you instead have a bunch of people each try out both rackets (within-subjects)

Why within-subjects designs can be useful

- Subjects may differ in ways that influence the dependent variable, e.g. some may be better tennis players than others
- In a between-subjects design, these differences add to the "noise" in the experiment, i.e. they increase the variability we cannot account for by the independent variable. As a result, it can be more difficult to see a significant effect.
- In a within-subjects design, we can discount the variability due to subject differences, and thus perhaps improve the power of the significance test

How to do a within-subjects ANOVA (and why we didn't cover it until now)

- A one-way within-subjects ANOVA looks an awful lot like the two-way ANOVA we did in (my) last lecture
- We just use a different measure for MS_{error}, the denominator of our F_{obt}, and a corresponding different df_{error}

An example

- How does your style of dress affect your comfort level when you are acting as a "greeter" in a social situation?
- 3 styles of dress: casual, semiformal, and formal.
- 5 subjects. Each subject wears each style of dress, one on each of 3 days. Order is randomized.
- Comfort level is measured by a questionnaire

The data

Factor A: Type of dress

	Casual	Semi-formal	Formal	
Subj1	5	8	4	$\Sigma x=17$
Subj2	7	11	6	$\Sigma x=24$
Subj3	5	9	2	$\Sigma x=16$
Subj4	5	9	3	$\Sigma x=17$
Subj5	3	8	1	$\Sigma x=12$
	$\Sigma x=25$ $\Sigma x^2=133$	$\Sigma x=45$ $\Sigma x^2=411$	$\Sigma x=16$ $\Sigma x^2=66$	Total: $\Sigma x=86$ $\Sigma x^2=610$

Two-way between-subjects vs. one-way within-subjects

- The table on the previous slide looks a lot like we're doing a two-way ANOVA, with subject as one of the factors
- However, cell (i, 1) is not necessarily independent of cell (i, 2) and cell (i, 3)
- Also, there is only, in this case, one data point per cell – we can't calculate $MS_{error} = MS_{wn}$ the way we did with two-way ANOVA
  - Sum of squared differences between the scores in each cell and the mean for that cell

We have to estimate the error variance in some other way

- Error variance is the variation we can't explain by one of the other factors
  - So it's clearly not variance in the data for the different levels of factor A, and it's not the variance in the data due to the different subjects
- We use as our estimate of the error variance the MS for the *interaction* between subject and factor A
  - The difference between the cell means not accounted for by the main effects

Steps

- 1. Compute SS_A, as before (see other lectures for the equation) =
  $25^2/5 + 45^2/5 + 16^2/5 - 86^2/15 = 88.13$
- 2. Similarly, compute SS_{subj} =
  $17^2/3 + 24^2/3 + 16^2/3 + 17^2/3 + 12^2/3 - 86^2/15 = 24.93$
- 3. Compute SS_{tot} as usual, =
  $610 - 86^2/15 = 116.93$

Steps

- 4. $SS_{tot} = SS_A + SS_{subj} + SS_{Axsubj}$ ->
  $SS_{Axsubj} = SS_{tot} - SS_A + SS_{subj}$
  $= 116.93 - 88.13 - 24.93 = 3.87$
- 5. Compute degrees of freedom:
  - $df_A = k_A - 1 = 2$
  - $df_{Axsubj} = (k_A - 1)(k_{subj} - 1) = (2)(4) = 8$

Steps

- We are doing this to check whether there's a significant effect of factor A, so:
- 6. $MS_A = SS_A/df_A = 88.13/2 = 44.07$
- 7. $MS_{error} = MS_{Axsubj} = SS_{Axsubj}/df_{Axsubj}$ $= 3.87/8 = 0.48$
- 8. Compute $F_{obt} = MS_A/MS_{error} = 91.08$
- 9. Compare with F_{crit} for df = $(df_A, df_{error}) = (2, 8)$. In this case, we wont bother, because it's clearly significant.

What if we had done this the between-subjects way?

- $SS_{tot} = 116.92$, $SS_{bn} = SS_A = 88.13$
- $SS_{wn} = SS_{tot} - SS_{bn} = 28.79$
- $df_{bn} = 2$, $df_{wn} = 15 - 3 = 12$
- $MS_{bn} = 88.13/2 = 44.07$
- $MS_{wn} = 28.79/12 = 2.40$
- $F_{obt} = 18.37$
  Still, no doubt significant, but not as huge of an F value as before.
- The extent to which the within-subjects design will have more statistical power is a function of how dependent the samples are for the different conditions, for each subject

(Some of) what this course did not cover

(This will not be on the exam; I just think it can be helpful to know what other sorts of tests are out there.)

Other two-sample parametric tests

- We talked about z- and t-tests for whether or not two means differ, assuming that the underlying distributions were approximately normal
- Recall that only two parameters are necessary to describe a normal distribution: mean and variance
- F-tests (which we used in ANOVA) can test whether the *variances* of two distributions differ significantly

Multiple regression and correlation

- We've talked about regression and correlation, in which we looked at linear prediction of Y given X, and how much of the variance in Y is accounted for by X (or vice versa)
- Sometimes *several* X variables help us more accurately predict Y
  - E.G. height and practise both affect a person's ability to shoot baskets in basketball
- This is like fitting a best fit *plane* instead of a best fit line

Multiple regression and correlation

- Put another way, sometimes we want to know the strength of relationship between 3 or more variables
- If we want to simultaneously study how several X variables affect the Y variable, we use *multiple regression & multiple correlation*

Non-linear regression

- And, as mentioned, you can fit curves other than lines and planes to the data

Correlation for ranked data

- To what extent do two rankings agree?
- Spearman's rank correlation coefficient, or
- Kendall's Tau

Other variants on ANOVA

- We talked about one-way and two-way between subjects ANOVA, and one-way within-subjects ANOVA
- You can, of course, also do two-way within-subjects ANOVA, and n-way ANOVA (though this gets complicated to interpret after n>3)
- Designs can also be *mixed-design*, meaning some factors are within-subjects factors, and others are between-subjects factors
- And there are all sorts of other complications as well…

What if our data don't meet the requirements for ANOVA?

- Recall for t-tests we talked about what to do when the data violate the assumption that the two groups have equal variance – we adjusted the degrees of freedom to account for this
- For ANOVA, there is a similar adjustment if the equivalent *sphericity assumption* is violated

Non-parametric procedures like t-tests

- The chi-square test was, in a sense, a non-parametric version of a t-test
  - A t-test tested whether a mean differed from what was expected, or whether two means were significantly different
  - A chi-square test tests whether cell values differ significantly from predicted, or whether two distributions were significantly different

Other non-parametric procedures like t-tests

- The equivalent of a t-test for ranked data is either the Mann-Whitney U test, or the rank sums test
- The Wilcoxon t-test is a test for *related samples* of ranked data
  - E.G. rank subjects reaction times on each of two tasks. Each subject participates in both tasks.

- In addition, in some special cases there are parametric techniques like t-tests that assume some distribution *other than a normal distribution*

Non-parametric version of ANOVA

- Kruskal-Wallis H test
  - Like a one-way, between-subjects ANOVA for ranked data
- Friedman χ^2 test
  - Like a one-way, within-subjects ANOVA for ranked data

There are also more advanced techniques

ANCOVA: ANalysis of COVAriance

- Provides a type of after-the-fact control for one or more variables that may have affected the dependent variable in an experiment
- The aim of this technique is to find out what the analysis of variance results might have been like if these variables had been held constant

More on ANCOVA

- Suppose factor A affects the response Y, but Y is also affected by a nuisance variable, X
- Ideally, you'd have run your experiment so that groups for different levels of factor A all had the same value of X
- But sometimes this isn't feasible, for whatever reason, & under certain conditions you can use ANCOVA to adjust things after the fact, *as if* X had been held constant
  - E.G. After the fact, adjust for the effects of intelligence on a training program

Non-parametric "bootstrapping" techniques

- "Pulling yourself up by your bootstraps"
- Use information gained in the experiment (e.g. about the distribution of the data) to create a non-parametric test that's basically designed for the distribution of your data
- These are basically *Montecarlo* techniques – they involve estimating the distribution of the data, and then generating multiple samples of new data with that distribution
- (Montecarlo techniques are what you did in MATLAB in the beginning of class)

Why bootstrapping?

- Because now we can
  - This is a recent technique (1970's), made feasible by computers
- It's conceptually and computationally simple
- The distribution assumptions depend upon the observed distribution of actual data, instead of upon large sample approximations like most of our parametric tests
- Why not: distribution estimates will change from one run of the experiment to another, and if the data does closely follow, say, a normal distribution, this technique will not do as well