One-way ANOVA, II

9.07

4/22/2004

Your schedule of coming weeks

- Today: One-way ANOVA, part II
- Next week:
  Two-way ANOVA, parts I and II.
  One-way ANOVA HW due Thursday
- Week of May 4
  Teacher out of town all week
  No class on Tuesday, 5/4
  Thursday class: TA's talk about statistical learning and other uses for statistics outside of the scope of this course, and do a bit of a review.
  Two-way ANOVA HW due Thursday (last HW!)

Review from last time

- Create ANOVA table

Source	Sum of squares	df	Mean square	F	P
Between	SS_{bn}	df_{bn}	MS_{bn}	F_{obt}	p-value
Within	SS_{wn}	df_{wn}	MS_{wn}		
Total	SS_{tot}	df_{tot}			

Review from last time

$$SS_{tot} = (\sum x^2)_{tot} - \frac{(\sum x)^2_{tot}}{N_{tot}}, \quad \text{df} = \text{N - 1}$$

$$SS_{bn} = \sum_{i=1}^{\#\text{condits}} n_i (m_i - M)^2 = \sum n_i m_i^2 - M^2 N, \quad \text{df} = \text{k - 1}$$

$$SS_{wn} = SS_{tot} - SS_{bn} = \sum_{i=1}^{\#\text{condits}} \sum_{j=1}^{n_i} (x_{ij} - m_i)^2, \quad \text{df} = \text{N - k}$$

MS = SS/df

$F_{obt} = MS_{bn} / MS_{wn}$

Compare F_{obt} to F_{crit}, df = (df_{bn}, df_{wn})

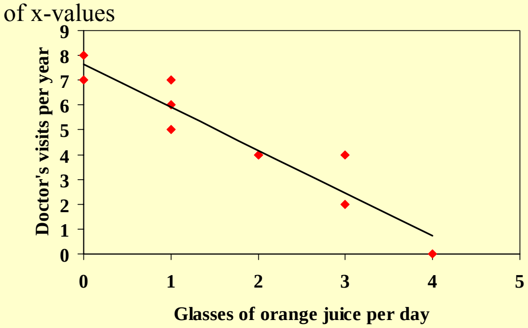
Relationship to proportion of variance accounted for

- MS_{bn} / MS_{tot} = proportion of variance accounted for by the systematic effect
- MS_{wn} / MS_{tot} = proportion of variance not accounted for
- So, another way of looking at F_{obt} is:

$$F_{obt} = \frac{\text{proportion of variance accounted for}}{\text{proportion of variance not accounted for}}$$

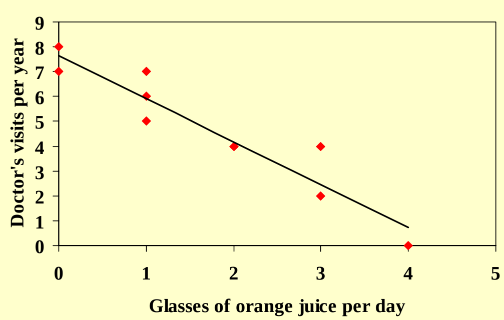
Relationship to correlation analysis

- In correlation analysis, we weren't interested in whether there was a difference in y-values between a particular pair of x-values



Relationship to correlation analysis

- Instead, we were interested in whether there was a significant relationship between the x-values and y-values



Relationship to correlation analysis

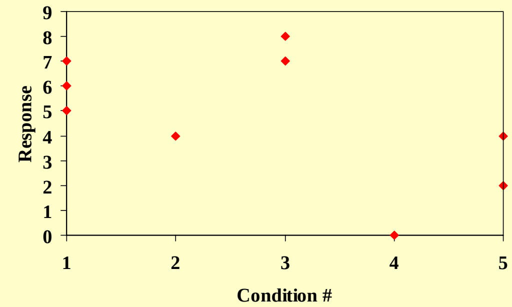
- Similarly, in ANOVA we are not asking whether there is a difference between a particular pair of conditions ("x-values")
- Instead, we are testing whether there is a significant effect of x-values on y-values

Relationship to correlation analysis

- The main difference between correlation analysis and ANOVA is that in ANOVA there may not be a natural ordering to the conditions (x-values)
- So, one shouldn't expect that the relationship is linear (as in correlation analysis), since it's not clear what this would mean

ANOVA vs. correlation analysis

- ANOVA:



Relationship between ANOVA and correlation analysis

- In both cases, since we're not interested in the difference in mean between a particular pair of x-values, we look for an overall effect of the x-values on the y-values by looking at how much knowing the x-value of a data point reduces your uncertainty in its y-value

The tests for ANOVA and correlation analysis are very similar

- In fact, some people treat correlation analysis as a form of ANOVA
- ANOVA:

$$F_{obt} = \frac{\text{proportion of variance accounted for}}{\text{proportion of variance not accounted for}}$$

- Correlation:

$$t_{obt}^2 = \frac{r^2 (N-2)}{(1-r^2)}$$

Proportion of variance accounted for (r^2)

Proportion of variance not accounted for ($1-r^2$)

Post-hoc comparisons

So, suppose you get a significant results from a one-way ANOVA

- If there are only two conditions, you know that there's a significant difference in the means of those two conditions
- If there are more than two conditions, you just know that there's *at least one* significant difference between the means for each condition
  - A=B>C A>B=C B=A>C=B A>B>C
- We need to do a *post-hoc comparison* to determine which means differ significantly

Post-hoc comparisons

- There are a whole bunch of these
- Your handout covers two of them
  - Tukey's HSD (very popular)
  - Fisher's protected t-test (also called Fisher's LSD)

Post-hoc comparison factoids

- Require significant F_{obt}
- Main idea behind these tests:
  - Before we talked about how multiple comparisons are problematic because they inflate the experiment-wise Type I error
  - These tests are designed to adjust Type I error by considering all possible pairwise comparisons
  - As a result, post-hoc comparisons are less powerful than *planned comparisons*, i.e. tests that were planned before the data was collected. We'll talk about these next.

Tukey's HSD

- Tukey's **H**onestly **S**ignificant **D**ifference
- *Requires that the n_i's in all levels of the factor (all conditions) are equal*

The basic idea in Tukey's HSD

- Find out how big the difference between two means needs to be in order for the difference to be significant
  - This difference depends upon your desired α, the amount of "noise", MS_{wn}, the number of means to compare, and n_i
  - Critical difference = HSD
- Compare all differences in mean to HSD, to determine which are significant

Tukey's HSD

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

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 groups.

MS_{wn} Is the mean square within groups.

n Is the number of people in each group.

Back to our example from last time

- Does perceived difficulty affect performance on a math test?

Level 1: easy	Level 2: medium	Level 3: difficult
9	4	1
12	6	3
4	8	4
8	2	5
7	10	2
Σx=40	Σx=30	Σx=15
Σx^2=354	Σx^2=220	Σx^2=55
n_1=5	n_2=5	n_3=5
m_1=8	m_2=6	m_3=3

ANOVA table for this example

Source	Sum of squares	df	Mean square	F	P
Between	63.33	2	31.67	4.52	$p<0.05$
Within	84.00	12	7.00		
Total	147.33	14			

1. Get q_α from the studentized range statistic table

- Table looks like this:

df_{wn}	α	k=# means being compared		
		2	3	4
11	0.05	3.11	3.82	4.26
	0.01	4.39	5.14	5.62
12	0.05	3.08	3.77	4.20
	0.01	4.32	5.04	5.50
13	0.05	3.06	3.73	4.15
	0.01	4.26	4.96	5.40

In our case, k=3, df_{wn}=12, and take α=0.05

- So, it looks like $q_a = 3.77$

df_{wn}	α	k=# means being compared		
		2	3	4
11	0.05	3.11	3.82	4.26
	0.01	4.39	5.14	5.62
12	0.05	3.08	3.77	4.20
	0.01	4.32	5.04	5.50
13	0.05	3.06	3.73	4.15
	0.01	4.26	4.96	5.40

2. Compute the HSD

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

- From the preceding slides, q_a= 3.77, MS_{wn} = 7, and n = 5, so

  HSD = 3.77 sqrt(7/5) = 4.46

3. Compare each difference in mean to the HSD

- Any differences > HSD are significant at the α level
  - These are absolute differences – ignore the sign

Level 1: easy	Level 2: medium	Level 3: difficult
$m_1=8$	$m_2=6$	$m_3=3$

m_1 to m_2: 2.0; m_2 to m_3: 3.0; m_1 to m_3: 5.0

  - HSD = 4.46, so the difference between the easy and difficult condition is significant, but neither of the other differences is.

Results of the ANOVA and Tukey's HSD

- A one-way, between subjects ANOVA demonstrated a significant effect of instruction on performance on the math test ($F(2, 12) = 4.52$, $p<0.05$). Post-hoc comparisons with Tukey's HSD test indicated a significant difference only between subjects told the problems would be easy (M=8), and those told the problems would be difficult (M=3), $p<0.05$.

Aside: a message from your TA's

- When you report results on your homework (e.g. when you are asked "what do you conclude?") do *not* just conclude "we reject the null hypothesis"!
- Make a conclusion about the *science*: e.g. "there seems to be a systematic effect of instruction on test performance."
- If research papers all reported only "we rejected the null hypothesis," they'd be very difficult to read (and half the time you're not entirely clear on what exactly were the authors' null and alternative hypotheses)!

Fisher's LSD (protected t-test)

- Fisher's **L**east **S**ignificant **D**ifference
- Does *not* require that all the n_i's be the same
- But is considerably more cumbersome than Tukey's HSD, because you essentially carry out a whole bunch of t-tests
- This is a liberal test, i.e. it gives you high power to detect a difference if there is one, but at somewhat increased risk of a Type I error

The basic idea in Fisher's LSD

- This test is very much like doing a t-test between each pair of means
- We get our df and SE for the tests from the ANOVA table

Fisher's LSD

$$LSD = \frac{m_i - m_j}{\sqrt{MS_{wn}\left[\frac{1}{n_i} + \frac{1}{n_j}\right]}}$$

m_i, m_j	Group means (i.e. means for conditions i & j)
n_i, n_j	Group sample sizes
MS_{wn}	Mean square within

$df_{crit} = df_{wn}$

Fisher's LSD test

- Compare LSD to t_{crit} from a standard t-table, with the usual α, and df = df_{wn}
- If LSD > t_{crit}, the difference is significant

An example

- We can run Fisher's LSD even if n_i are all equal, so let's run it on the same example we've been using.
  - $MS_{wn} = 7.00$, $df_{wn} = 12$

Level 1: easy	Level 2: medium	Level 3: difficult
n_1=5	n_2=5	n_3=5
m_1=8	m_2=6	m_3=3

An example

$$LSD = \frac{m_i - m_j}{\sqrt{MS_{wn}\left[\frac{1}{n_i} + \frac{1}{n_j}\right]}}$$

Level 1: easy	Level 2: medium	Level 3: difficult
n_1=5	n_2=5	n_3=5
m_1=8	m_2=6	m_3=3

- 1. In this case, the denominator of LSD is always sqrt(7(2)(1/5)) ≈ 1.67
  - MS_{wn} = 7.00

An example

- Denom ≈ 1.67
- 2. Compute differences in mean between each pair of conditions
- 3. Compute LSD for each pair of conditions.

Level 1: easy	Level 2: medium	Level 3: difficult
n_1=5	n_2=5	n_3=5
m_1=8	m_2=6	m_3=3

2.0 (Level 1 – Level 2): LSD≈1.20

3.0 (Level 2 – Level 3): LSD≈1.79

5.0 (Level 1 – Level 3): LSD≈2.99

An example

- Now, compare the LSD's to t_{crit}
  - $t_{crit,0.05}$(df=12) = 2.179
- Once again, only the difference in mean between the easy and difficult levels is significant, so we have the same conclusion as in the Tukey HSD test

Review

- One-way ANOVA is used for one-factor experiments, in which there are ≥ 2 conditions
- Use ANOVA to test whether there are *any* differences in means between the conditions
- Then if the ANOVA gives a significant result, run post-hoc tests to find out which differences are significant
  - We talked about Tukey's HSD, and Fisher's LSD.
  - There are a bunch of other post-hoc tests

An alternative: planned comparisons

- In certain circumstances, it makes sense to compare the means in the various conditions using *planned comparisions*
- This is done *instead of* the ordinary analysis of variance

Planned comparisons

- As the name suggests, this applies only when the experimenter has specific questions they want to ask *before* they ever collect and analyze the data
- It is often worth our while to do planned comparisons when our experiment contains a number of pairs of means that we might compare, but we are only interested in a small number of those comparisons

An example

- Suppose you are interested in studying changing attitudes toward some institution or group, e.g. changing attitudes toward a minority group
- In particular, you want to know if there is a difference between the following modes of persuasion, in terms of changing people's attitudes:
  - A movie favorable to the minority group
  - A lecture on the same topic, also favorable to the group
  - A combination of the movie and the lecture

The experiment

- Since you're studying changing attitudes, you measure the attitude of the subjects both before and after the "treatment" (the movie and/or lecture)
- Just asking people about their attitudes may cause them to answer differently the second time. And perhaps even watching a neutral movie or lecture in between the attitude questionnaires may have some effect. So, as a control, you look at changing attitudes among control groups that:
  - Get no treatment,
  - Watch a neutral movie, or
  - Hear a neutral lecture

The experimental design

- The experiment, then, has one factor (the treatment), with the following 6 levels:

Experimental groups			Control groups		
I	II	III	IV	V	VI
Movie	Lecture	Movie & lecture	Nothing	Neutral movie	Neutral lecture

Paired comparisons

- There are a total of 15 possible simple comparisons between the means for two of the conditions
- Analysis of variance will test whether or not there are differences in any of these 15 comparisons
- However, the experimenter may only be interested in answering a relatively small set of questions

Questions we might be interested in answering

- 1. Do the experimental groups as a whole tend to differ from the control groups?
- 2. Is the effect of the experimental movie+lecture combination different from the average effect of either experimental movie or the experimental lecture?
- 3. Is the effect of the experimental lecture different from the effect of the experimental movie?
- 4. Within the control groups, is there any effect of the neutral movie or lecture compared with the group getting no treatment?

Why do planned comparisons, I

- ANOVA+post-hoc testing effectively tests all possible pairs of means, while keeping the experiment-wise Type I error rate under control.
- This means that ANOVA+post-hoc testing is overly conservative if we're only interested in a few of those comparisons
- E.G. in our example, we are only interested in 4 out of 15 comparisons

Why do planned comparisons, II

- Some researchers seem to like ANOVA because it allows them to find at least *some* significant result from their experiment
- But, this sort of "fishing" should generally be discouraged
- One benefit of planned comparisons is that you're forced to really think about what questions you want to ask

Why do planned comparisons, II

- Among other things, the over-reliance on ANOVA tends to lead to confusing presentation of the results
  - People tend to report *any* significant difference they find, even if it's not interesting, and they can't make any sense of it
  - Sometimes there are a lot of these little confusing significant differences…
  - In our example:
    What if the only significant difference is neutral movie > neutral lecture? What would that tell us? Not that we should show neutral movies – that's no better than doing nothing at all!

A different way of looking at comparisons

- To talk about planned comparisons, we need a different way of looking at the sorts of comparisons we might want to conduct
- Typical comparison: is the mean of group 1 (μ_1) different from the mean of group 2 (μ_2)?

Is μ_1 different from μ_2?

- Recall the logic of a two-sample t-test:
- Distribution of sample mean m_1 is normal
- Distribution of sample mean m_2 is normal
- Therefore the distribution of $m_1 - m_2$ is approximately normal
  - *Linear combinations of normal random variables are normal*
- Therefore we can use the normal approximation to test whether our observed $m_1 - m_2$ was likely to occur, if actually $\mu_1 - \mu_2 = 0$

Comparisons as vectors

- We can think of the desired population comparison as
  $\Psi = \mu_1 - \mu_2 = c_1 \mu_1 + c_2 \mu_2$,
  where $c_1 = 1$, $c_2 = -1$
- Similarly, our test statistic, ψ':
  $\Psi' = m_1 - m_2 = c_1 m_1 + c_2 m_2$,
  where $c_1 = 1$, $c_2 = -1$
- So, we can think of the comparison as represented by the vector [1 -1]

Comparisons can be any linear combination of means, not just μ_i - μ_j

- More generally, by the same logic as for the $\mu_i - \mu_j$ comparison, we can test other linear combinations of means, and represent the comparison by the vector of coefficients c_i
- When there are k conditions in our single-factor experiment, there are k coefficients c_i, per comparison
- Typically we are interested in comparisons satisfying the following constraint: $\sum_{j=1}^{J} c_j = 0$
- Such comparisons have some nice properties, so we will restrict ourselves to comparisons satisfying that constraint

Let's put some of our desired comparisons into this notation

- 1. Do the experimental groups as a whole tend to differ from the control groups?
- I take this to mean " is the average of the means for the 3 experimental groups different from the average of the means for the 3 control groups?"
- Average of the means for the experimental groups =

Experimental groups			Control groups		
I	II	III	IV	V	VI
Movie	Lecture	Movie & lecture	Nothing	Neutral movie	Neutral lecture

- 1. Do the experimental groups as a whole tend to differ from the control groups?
- Mean of experimental groups =
  $(\mu_1 + \mu_2 + \mu_3)/3 = 1/3\ \mu_1 + 1/3\ \mu_2 + 1/3\ \mu_3$
- So, the difference in mean =
  $1/3\ \mu_1 + 1/3\ \mu_2 + 1/3\ \mu_3 - 1/3\ \mu_4 - 1/3\ \mu_5 - 1/3\ \mu_6$
- We can represent this by
  C_1 = [1/3 1/3 1/3 -1/3 -1/3 -1/3]

2. Is the effect of the experimental movie+lecture combination different from the average effect of either experimental movie or the experimental lecture?

- This is somewhat ambiguously worded
  - Are we talking about two comparisons, movie+lecture vs. movie, and movie+lecture vs. lecture?
  - Or are we talking about movie+lecture vs. the pooled mean from movie and lecture separately
- Convention, in this class: in a list of numbered comparisons, *each item in the list corresponds to a single comparison*
  - Is movie+lecture different from the mean of (movie, lecture)?

Experimental groups			Control groups		
I	II	III	IV	V	VI
Movie	Lecture	Movie & lecture	Nothing	Neutral movie	Neutral lecture

- 2. Is the mean effect of movie+lecture equal to the average of presenting only the experimental movie and presenting only the experimental lecture?
- Mean of experimental movie and experimental lecture = $(\mu_1 + \mu_2)/2 = ½\ \mu_1 + ½\ \mu_2$
- So, this comparison = $½\ \mu_1 + ½\ \mu_2 - \mu_3$
- We can represent this by
  $C_2 = [½\ ½\ \text{-}1\ 0\ 0\ 0]$

3. Is the effect of the experimental lecture different from the effect of the experimental movie?

- This is just like our usual comparison between two means
- Comparison: $\mu_1 - \mu_2$
- We can represent this by $C_3 = [1\ \text{-}1\ 0\ 0\ 0\ 0]$

4. Within the control groups, is there any effect of the neutral movie or lecture compared with the group getting no treatment?

- Again, we take this to refer to a single comparison, in this case between no treatment, and the average effect of a neutral movie or a neutral lecture

Experimental groups			Control groups		
I	II	III	IV	V	VI
Movie	Lecture	Movie & lecture	Nothing	Neutral movie	Neutral lecture

- 4. Is the mean effect of doing nothing equal to the average of presenting the neutral movie and presenting the neutral lecture?
- This comparison = $1\ \mu_4 - ½\ \mu_5 - ½\ \mu_6$
- We can represent this by C_4 = [0 0 0 1 -½ -½]

Our comparisons

	I	II	III	IV	V	VI
Comparison 1	1/3	1/3	1/3	-1/3	-1/3	-1/3
Comparison 2	½	½	-1	0	0	0
Comparison 3	1	-1	0	0	0	0
Comparison 4	0	0	0	1	-1/2	-1/2

Note each row sums to 0.

Carrying out the comparisons

- Apply the weights given in the previous table to the sample means by doing a dot product between the comparison vector C_i and the vector of sample means
- E.G. suppose our 6 sample means = [27 15 32 -7 4 -3]
- Then comparison 3 = [1 -1 0 0 0 0] . [27 15 32 -7 4 -3] = 27 – 15 = 12
- Comparison 4 = [0 0 0 1 -1/2 -1/2] . [27 15 32 -7 4 -3] = (1)(-7) + (-1/2)(4) + (-1/2)(-3) = -7.5

Statistical tests on these comparisons

- To conduct a two-sample t-test, we needed to know the mean and variance of the sampling distribution of $m_1 - m_2$, given the null hypothesis that the means are equal
- The same is true here, except we have a more complicated linear combination of sample means

The mean of the sampling distribution of the comparison is 0

- $E(\psi') = E(\Sigma c_j m_j) = \Sigma c_j E(m_j) = \mu \Sigma c_j = 0$

The variance of a linear combination of independent random variables is easy to find

- Assume the sample means m_j are based upon independent samples of size n_j from a population with variance σ^2
  - We assume that the variances in the different populations are all equal
- Then the sampling distribution for m_j has variance σ^2/n_j

The variance of a linear combination of independent random variables is easy to find

- Recall that if we multiply a random variable by a scaling constant, say c_j, the variance of that random variable is multiplied by c_j^2
- And, the variance of the sum is the sum of the variances
- From this, it should be clear that the variance of the sampling distribution of a comparison represented by vector $C = [c_1\ c_2\ \ldots\ c_k]$ is:

$$\mathrm{var}(\psi') = \sigma^2 \sum_{j=1}^{k} \frac{c_j^2}{n_j}$$

Estimating σ^2

- Of course, as usual, we don't actually know the population variance
- Estimate it with $MS_{wn} = MS_{error}$, computed as for the ANOVA

Conducting a planned comparison

$$t_{obt} = (\psi' - \psi) / \sqrt{MS_{wn} \sum_{j=1}^{k} \frac{c_j^2}{n_j}}$$

- Compare with t_{crit} from a standard t-table, with the df_{wn} = N-k degrees of freedom
- Alternative for two-tailed tests: $F_{obt} = (t_{obt})^2$. Compare with F_{crit} from your F-table, with (1, N-k) degrees of freedom

Multiple planned comparisons

- This is all fine and good if you do only one comparison
- For multiple comparisons, you need to worry about the experiment-wise error rate
  - Adjust α to compensate for multiple comparisons, e.g. using Bonferroni adjustment
  - This is really only well behaved if your comparisons are *independent*
    - You can still do the tests if they aren't, however

How to tell if your comparisons are independent

- Take the dot product of the two comparison vectors. If this equals 0, the pair is independent.
- E.G. if $C_1 = [c_{11}\ c_{12}\ c_{13}\ c_{14}]$, and $C_2 = [c_{21}\ c_{22}\ c_{23}\ c_{24}]$, then C_1 and C_2 are independent if and only if

  $c_{11}c_{21} + c_{12}c_{22} + c_{13}c_{23} + c_{14}c_{24} = 0$
- Note in MATLAB the dot product is C1*C2'

Testing the comparisons for our example

	I	II	III	IV	V	VI
Comparison 1	1/3	1/3	1/3	-1/3	-1/3	-1/3
Comparison 2	½	½	-1	0	0	0
Comparison 3	1	-1	0	0	0	0
Comparison 4	0	0	0	1	-1/2	-1/2

- C1.C2 = (1/3)(1/2)+(1/3)(1/2) – 1/3 = 0
- C1.C3 = 1/3 – 1/3 = 0
- C1.C4 = -1/3 + (1/3)(1/2) + (1/3)(1/2) = 0
- C2.C3 = ½ - ½ = 0
- C2.C4 = 0
- C3.C4 = 0

Suppose we were to test these 4 comparisons – what should α be?

- Suppose you want the experiment-wise error rate to be 0.05
- Then the Bonferroni adjustment says to use α=0.05/(number of comparisons) as your per-comparison error rate
- So, use α=0.05/4 = 0.0125
- For this example, we'll use 0.01, since it's in our tables

Example – here's the data

	Movie	Lecture	Movie & lecture	Nothing	Neutral movie	Neutral lecture
	6	3	7	-6	5	-1
	10	6	9	0	-5	3
	1	-1	4	-5	3	2
	6	5	9	2	-4	-1
	4	2	3	2	5	-6
Mean	5.4	3	6.4	-1.4	0.8	-0.6

Standard one-way ANOVA calculations yield MS_{wn} = 12.90, df_{wn} = 5*6 – 6 = 24

Check comparison 1

- C = [1/3 1/3 1/3 -1/3 -1/3 -1/3]
- ψ' = C . (vector of means) = (1/3)(5.4) + (1/3)(3) + (1/3)(6.4) + (-1/3)(-1.4) + (-1/3)(0.8) + (-1/3)(-0.6) = 5.33
- For the standard error, we also need $\Sigma c_j^2/n_j$ = 1/5 (6 · 1/9) = 2/15
- SE = sqrt(12.90 · 2/15) = 1.31

So, testing comparison 1

- We are testing against the null hypothesis that there is no difference between the experimental and control groups, i.e. ψ=0
- t_{obt} = ψ'/SE = 5.33/1.31 = 4.07
- t_{crit} = 2.797 (df = 24, α=0.01)
- So, there is a significant difference in change of attitudes toward a minority group between the experimental group, which sees a movie and/or lecture in favor of the minority group, and the control group, which sees either nothing or a neutral movie/lecture.

And so on for the other 3 comparisons

- We don’t have time to go through the rest of the comparisons today, but the calculations are much the same.
- My quick back-of-the-envelope calculations suggest that none of the rest of the comparisons yield significant results, but you should try them for yourself