PSY 201: Statistics in Psychology

Lecture 33
Analysis of Variance
Some thing versus which thing.

Greg Francis

Purdue University

Fall 2019

ANOVA

- Test statistic:

$$F = \frac{MS_B}{MS_W}$$

$$F = \frac{\text{Estimated variability from noise and mean differences}}{\text{Estimated variability from noise}}$$

- if H_0 is true, and F is sufficiently larger than 1, then a rare event has happened. Since rare events are rare, when $F >> 1$ we suppose that H_0 is not true
- Rareness is established by the p value, which is gotten from an F distribution with $K - 1$ *df* in the numerator and $N - K$ *df* in the denominator

HYPOTHESES

- The null is an *omnibus* hypothesis. It supposes no difference between any population means

$$H_0 : \mu_i = \mu_j \; \forall \; i, j$$

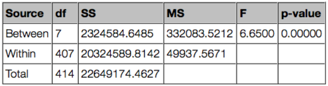
- the alternative is the complement

$$H_0 : \mu_i \neq \mu_j \;\; \text{form some } i, j$$

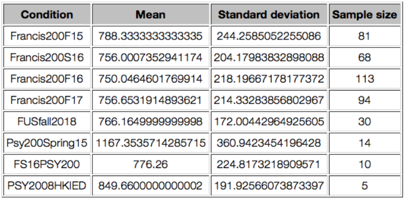
- Note, there is no *one-tailed* version of ANOVA

INTERPRETING

- I happen to have data from 8 different classes that all completed an experiment where subjects responded as quickly as possible whether a set of letters formed a word or not



Source	df	SS	MS	F	p-value
Between	7	2324584.6485	332083.5212	6.6500	0.00000
Within	407	20324589.8142	49937.5671		
Total	414	22649174.4627			

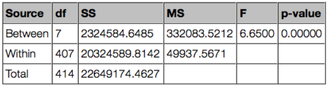


Condition	Mean	Standard deviation	Sample size
Francis200F15	788.3333333333335	244.2585052255086	81
Francis200S16	756.0007352941174	204.17983832898088	68
Francis200F16	750.0464601769914	218.19667178177372	113
Francis200F17	756.6531914893621	214.33283856802967	94
FUSfall2018	766.1649999999998	172.00442964925605	30
Psy200Spring15	1167.3535714285715	360.9423454196428	14
FS16PSY200	776.26	224.8173218909571	10
PSY2008HKIED	849.6600000000002	191.92566073873397	5

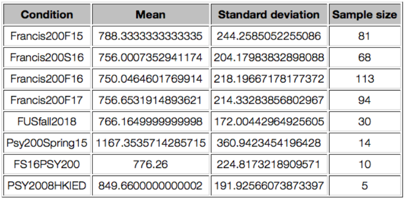
- The conclusion is that *at least one* population mean seems to be different from the other population means. **Something** is different
- The ANOVA does not tell you **which** mean is different from the others; or if more than one mean is different from others.

INTERPRETING

- It might be tempting to just look at the data and "wing it"
- For example, looking at the means, it seems that class *Psy200Spring15* has a much larger mean than any other class



Source	df	SS	MS	F	p-value
Between	7	2324584.6485	332083.5212	6.6500	0.00000
Within	407	20324589.8142	49937.5671		
Total	414	22649174.4627			

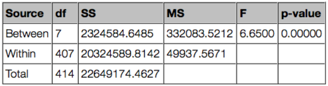


Condition	Mean	Standard deviation	Sample size
Francis200F15	788.3333333333335	244.2585052255086	81
Francis200S16	756.0007352941174	204.17983832898088	68
Francis200F16	750.0464601769914	218.19667178177372	113
Francis200F17	756.6531914893621	214.33283856802967	94
FUSfall2018	766.1649999999998	172.00442964925605	30
Psy200Spring15	1167.3535714285715	360.9423454196428	14
FS16PSY200	776.26	224.8173218909571	10
PSY2008HKIED	849.6600000000002	191.92566073873397	5

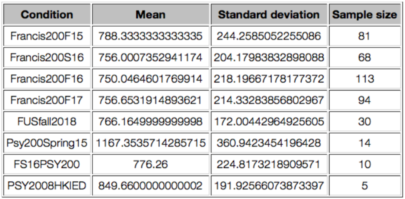
- But that class also has a small number of students ($n = 14$), and a large standard deviation ($s = 360.9$), so we would expect quite a bit of variability in the mean value. Maybe this big mean is not so rare, given the variability due to random sampling

INTERPRETING

- More than one mean might differ from other means
- Even if the mean for *Psy200Spring15* is different from the others, might other means also be different?



Source	df	SS	MS	F	p-value
Between	7	2324584.6485	332083.5212	6.6500	0.00000
Within	407	20324589.8142	49937.5671		
Total	414	22649174.4627			



Condition	Mean	Standard deviation	Sample size
Francis200F15	788.3333333333335	244.2585052255086	81
Francis200S16	756.0007352941174	204.17983832898088	68
Francis200F16	750.0464601769914	218.19667178177372	113
Francis200F17	756.6531914893621	214.33283856802967	94
FUSfall2018	766.1649999999998	172.00442964925605	30
Psy200Spring15	1167.3535714285715	360.9423454196428	14
FS16PSY200	776.26	224.8173218909571	10
PSY2008HKIED	849.6600000000002	191.92566073873397	5

- We would really like to know which means seem to be different from which other means

TYPE I ERROR

- Multiple testing problem
- To motivate ANOVA, we mentioned that it is problematic to just test all pairwise comparisons of group means. With 8 means, there would be 28 tests. So the Type I error rate would be around

$$1-(1-\alpha)^{28}=(1-0.95^{28})=0.76$$

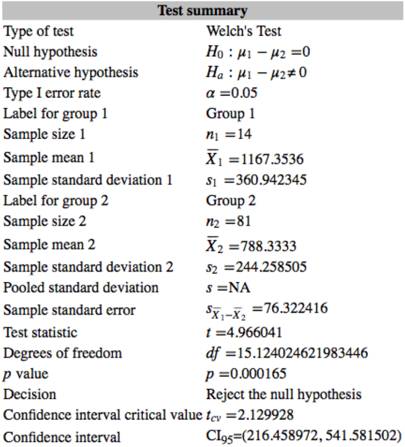
- Instead of just testing all possible comparisons, suppose we first require that the ANOVA produces a significant result. If H_0 is true, the ANOVA should only conclude that some difference exists with a probability of 0.05 (or whatever you choose as α)

TYPE I ERROR

- Thus, we can control the overall Type I error rate by insisting that our data produce a significant ANOVA *before* we start testing different means
- We want to check that something is different before we check which means are different!
- If we test *Psy200Spring15* against each of the other seven means, the Type I error rate can be no bigger than what it was for the ANOVA
- In fact, it has to be a bit smaller than the α used for the ANOVA because we have to satisfy two criteria
- If H_0 is true, 95% of the time, we never compare the means to each other

t tests

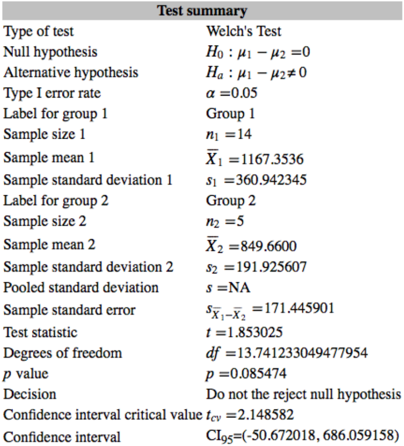
- One approach is to just run *t* tests (Welch's test) to compare different means
- For example, we can test *Psy200Spring15* against *Francis200F15*



Test summary	
Type of test	Welch's Test
Null hypothesis	$H_0 : \mu_1 - \mu_2 = 0$
Alternative hypothesis	$H_a : \mu_1 - \mu_2 \neq 0$
Type I error rate	$\alpha = 0.05$
Label for group 1	Group 1
Sample size 1	$n_1 = 14$
Sample mean 1	$\overline{X}_1 = 1167.3536$
Sample standard deviation 1	$s_1 = 360.942345$
Label for group 2	Group 2
Sample size 2	$n_2 = 81$
Sample mean 2	$\overline{X}_2 = 788.3333$
Sample standard deviation 2	$s_2 = 244.258505$
Pooled standard deviation	$s =$NA
Sample standard error	$s_{\overline{X}_1 - \overline{X}_2} = 76.322416$
Test statistic	$t = 4.966041$
Degrees of freedom	$df = 15.124024621983446$
p value	$p = 0.000165$
Decision	Reject the null hypothesis
Confidence interval critical value	$t_{cv} = 2.129928$
Confidence interval	CI_{95}=(216.458972, 541.581502)

t tests

- One approach is to just run *t* tests (Welch's test) to compare different means
- For example, we can test *Psy200Spring15* against *PSY2008HKIED*



Test summary	
Type of test	Welch's Test
Null hypothesis	$H_0 : \mu_1 - \mu_2 = 0$
Alternative hypothesis	$H_a : \mu_1 - \mu_2 \neq 0$
Type I error rate	$\alpha = 0.05$
Label for group 1	Group 1
Sample size 1	$n_1 = 14$
Sample mean 1	$\overline{X}_1 = 1167.3536$
Sample standard deviation 1	$s_1 = 360.942345$
Label for group 2	Group 2
Sample size 2	$n_2 = 5$
Sample mean 2	$\overline{X}_2 = 849.6600$
Sample standard deviation 2	$s_2 = 191.925607$
Pooled standard deviation	$s =$NA
Sample standard error	$s_{\overline{X}_1 - \overline{X}_2} = 171.445901$
Test statistic	$t = 1.853025$
Degrees of freedom	$df = 13.741233049477954$
p value	$p = 0.085474$
Decision	Do not the reject null hypothesis
Confidence interval critical value	$t_{cv} = 2.148582$
Confidence interval	CI_{95}=(-50.672018, 686.059158)

CONTRASTS

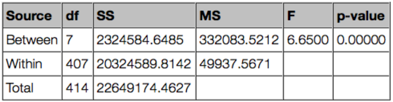
- There is a better (and more general approach)
- ANOVA assumes/requires homogeneity of variance

$$\sigma_i^2 = \sigma_j^2 \ \forall \ i, j$$

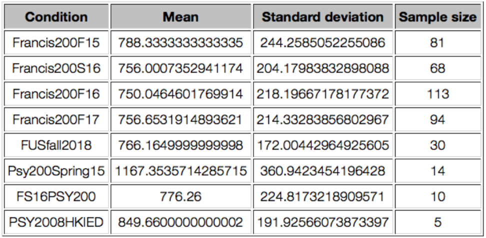
- For the *t*-test we pooled variances/standard deviations to get a better estimate of σ
- With more populations, we can pool all of the sample variances and thereby get a still better estimate
- Thus, even when we compare *Psy200Spring15* against *Francis200F15*, we can use the data from the other samples to get a better estimate of σ

POOLED ESTIMATE

- Fortunately, the pooled estimate of variance is easy to find
- We computed it in the ANOVA, it is MS_W



Source	df	SS	MS	F	p-value
Between	7	2324584.6485	332083.5212	6.6500	0.00000
Within	407	20324589.8142	49937.5671		
Total	414	22649174.4627			



Condition	Mean	Standard deviation	Sample size
Francis200F15	788.3333333333335	244.2585052255086	81
Francis200S16	756.0007352941174	204.17983832898088	68
Francis200F16	750.0464601769914	218.19667178177372	113
Francis200F17	756.6531914893621	214.33283856802967	94
FUSfall2018	766.1649999999998	172.00442964925605	30
Psy200Spring15	1167.3535714285715	360.9423454196428	14
FS16PSY200	776.26	224.8173218909571	10
PSY2008HKIED	849.6600000000002	191.92566073873397	5

- Thus, the standard error that we use for the t test is

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{MS_W \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

CONTRASTS

- For example, we can test *Psy200Spring15* against *Francis200F15*

$$s_{\overline{X}_1-\overline{X}_2} = \sqrt{MS_W\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = \sqrt{(49937.5671)\left(\frac{1}{14}+\frac{1}{81}\right)} = 64.67984425$$

- Compare to the traditional t test, where

$$s_{\overline{X}_1-\overline{X}_2} = 76.322416$$

- So with the pooled variance, we get

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1-\overline{X}_2}} = \frac{379.02}{76.322} = 5.8599$$

- Compare to $t = 4.966$ for the traditional t test
- The degrees of freedom is based on how many scores contribute to the variance calculation, so we get

$$df = N - K = 415 - 8 = 407$$

- compare to $df = n_1 + n_2 - 2 = 14 + 81 - 2 = 93$, for traditional t test (smaller with Welch's test)

CONTRASTS

- For example, we can test *Psy200Spring15* against *PSY2008HKIED*

$$s_{\overline{X}_1-\overline{X}_2} = \sqrt{MS_W\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = \sqrt{(49937.5671)\left(\frac{1}{14}+\frac{1}{5}\right)} = 116.423$$

- Compare to the traditional t test, where

$$s_{\overline{X}_1-\overline{X}_2} = 171.446$$

- So with the pooled variance, we get

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1-\overline{X}_2}} = \frac{317.69}{171.446} = 2.7288$$

- The degrees of freedom is based on how many scores contribute to the variance calculation, so we get

$$df = N - K = 415 - 8 = 407$$

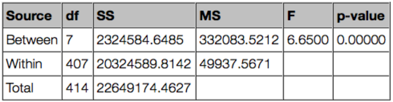
- so $p = 0.0066$
- Compare to $t = 1.853$ for the traditional t test
- compare to $df = n_1 + n_2 - 2 = 14 + 5 - 2 = 17$, and $p = 0.08$

BETTER IS BETTER

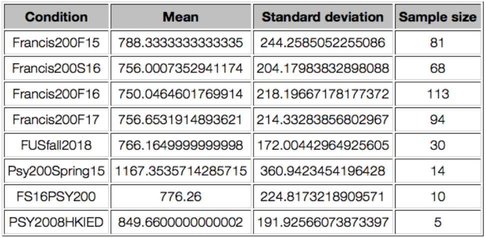
- With a contrast, we get a better estimate of $s_{\overline{X}_1 - \overline{X}_2}$, which sometimes means we can reject H_0. Not always, though.
- It is possible for a standard t test to reject H_0, but the corresponding contrast test does not reject H_0 (because the sample s^2 is smaller than MS_W)
- We do not have any cases like that in our current data set
- Generally speaking, using MS_W is better than using the pooled s^2 because more data contributes to the estimate

OTHER CONTRASTS

- Comparing two means is actually a special case of using contrasts
- We can also compare various *combinations* of means



Source	df	SS	MS	F	p-value
Between	7	2324584.6485	332083.5212	6.6500	0.00000
Within	407	20324589.8142	49937.5671		
Total	414	22649174.4627			



Condition	Mean	Standard deviation	Sample size
Francis200F15	788.3333333333335	244.2585052255086	81
Francis200S16	756.0007352941174	204.17983832898088	68
Francis200F16	750.0464601769914	218.19667178177372	113
Francis200F17	756.6531914893621	214.33283856802967	94
FUSfall2018	766.1649999999998	172.00442964925605	30
Psy200Spring15	1167.3535714285715	360.9423454196428	14
FS16PSY200	776.26	224.8173218909571	10
PSY2008HKIED	849.6600000000002	191.92566073873397	5

- For example, we might wonder if the mean for classes taught by Dr. Francis differs from the mean for classes not taught by Dr. Francis

OTHER CONTRASTS

- We set up *contrast weights*, c_i, for each class' mean
- Our null hypothesis will be

$$H_0 : \sum_{i=1}^{K} (c_i \mu_i) = 0$$

- and we require that the contrast weights sum to 0:

$$\sum_{i=1}^{K} c_i = 0$$

- Our alternative hypothesis is

$$H_a : \sum_{i=1}^{K} (c_i \mu_i) \neq 0$$

- (one-tailed tests are also possible)

TEST STATISTIC

- We compute the weighted sum of means

$$L = \sum_{i=1}^{K} (c_i \overline{X}_i)$$

- which has a standard error of:

$$s_L = \sqrt{MS_W \sum_{i=1}^{K} \frac{c_i^2}{n_i}}$$

- and our test statistic is

$$t = \frac{L}{s_L}$$

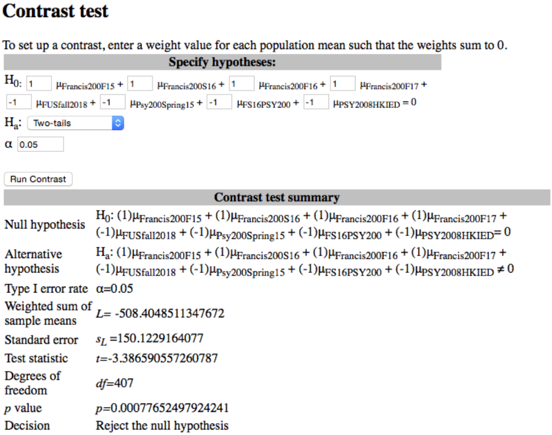
- which follows a t distribution with

$$df = N - K$$

  - where N is the sum of sample sizes across all groups and K is the number of groups

ONLINE CALCULATOR

- To compare the mean of the four classes taught by Dr. Francis to the mean of other four classes, we use contrast weights of ± 1



Contrast test

To set up a contrast, enter a weight value for each population mean such that the weights sum to 0.

Specify hypotheses:

H_0: 1 $\mu_{Francis200F15}$ + 1 $\mu_{Francis200S16}$ + 1 $\mu_{Francis200F16}$ + 1 $\mu_{Francis200F17}$ +
-1 $\mu_{FUSfall2018}$ + -1 $\mu_{Psy200Spring15}$ + -1 $\mu_{FS16PSY200}$ + -1 $\mu_{PSY2008HKIED}$ = 0

H_a: Two-tails

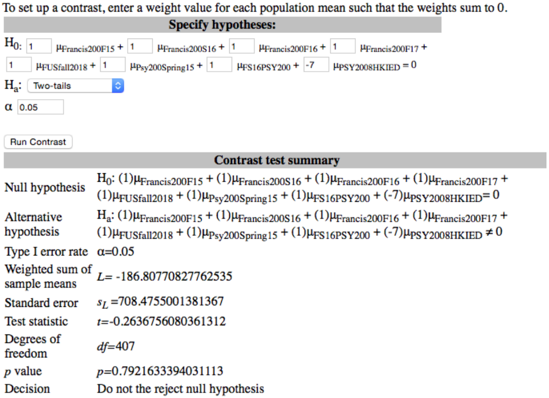
α 0.05

Contrast test summary

Null hypothesis	H_0: $(1)\mu_{Francis200F15} + (1)\mu_{Francis200S16} + (1)\mu_{Francis200F16} + (1)\mu_{Francis200F17} + (-1)\mu_{FUSfall2018} + (-1)\mu_{Psy200Spring15} + (-1)\mu_{FS16PSY200} + (-1)\mu_{PSY2008HKIED} = 0$
Alternative hypothesis	H_a: $(1)\mu_{Francis200F15} + (1)\mu_{Francis200S16} + (1)\mu_{Francis200F16} + (1)\mu_{Francis200F17} + (-1)\mu_{FUSfall2018} + (-1)\mu_{Psy200Spring15} + (-1)\mu_{FS16PSY200} + (-1)\mu_{PSY2008HKIED} \neq 0$
Type I error rate	α=0.05
Weighted sum of sample means	L= -508.4048511347672
Standard error	s_L =150.1229164077
Test statistic	t=-3.386590557260787
Degrees of freedom	df=407
p value	p=0.00077652497924241
Decision	Reject the null hypothesis

ONLINE CALCULATOR

- Other sets of contrast weights compare other combinations. For example, to contrast the mean of the non-US based class, *PSY2008HKIED*, against all the other classes, we could use:



To set up a contrast, enter a weight value for each population mean such that the weights sum to 0.

Specify hypotheses:

H_0: 1 $\mu_{Francis200F15}$ + 1 $\mu_{Francis200S16}$ + 1 $\mu_{Francis200F16}$ + 1 $\mu_{Francis200F17}$ + 1 $\mu_{FUSfall2018}$ + 1 $\mu_{Psy200Spring15}$ + 1 $\mu_{FS16PSY200}$ + -7 $\mu_{PSY2008HKIED}$ = 0

H_a: Two-tails

α 0.05

Run Contrast

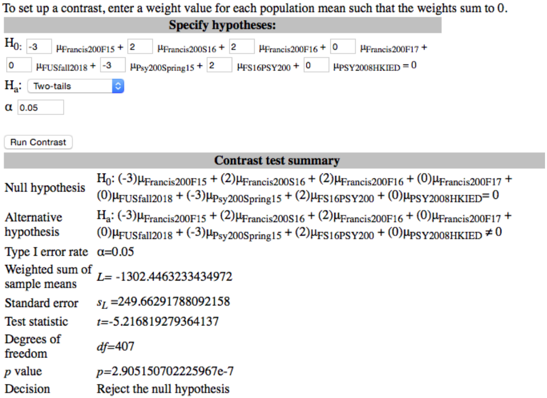
Contrast test summary

Null hypothesis	H_0: (1)$\mu_{Francis200F15}$ + (1)$\mu_{Francis200S16}$ + (1)$\mu_{Francis200F16}$ + (1)$\mu_{Francis200F17}$ + (1)$\mu_{FUSfall2018}$ + (1)$\mu_{Psy200Spring15}$ + (1)$\mu_{FS16PSY200}$ + (-7)$\mu_{PSY2008HKIED}$= 0
Alternative hypothesis	H_a: (1)$\mu_{Francis200F15}$ + (1)$\mu_{Francis200S16}$ + (1)$\mu_{Francis200F16}$ + (1)$\mu_{Francis200F17}$ + (1)$\mu_{FUSfall2018}$ + (1)$\mu_{Psy200Spring15}$ + (1)$\mu_{FS16PSY200}$ + (-7)$\mu_{PSY2008HKIED}$ $\neq$ 0
Type I error rate	α=0.05
Weighted sum of sample means	L= -186.80770827762535
Standard error	s_L =708.4755001381367
Test statistic	t=-0.2636756080361312
Degrees of freedom	df=407
p value	p=0.7921633394031113
Decision	Do not the reject null hypothesis

- You do not have to use integer values for the c_i terms, but it helps to avoid rounding issues.

ONLINE CALCULATOR

- It can be appropriate to set some weights equal to 0. For example, if you want to compare the mean from two classes in 2015 against the mean from three classes in 2016, you can set weights as:



To set up a contrast, enter a weight value for each population mean such that the weights sum to 0.

Specify hypotheses:

H_0: -3 $\mu_{\text{Francis200F15}}$ + 2 $\mu_{\text{Francis200S16}}$ + 2 $\mu_{\text{Francis200F16}}$ + 0 $\mu_{\text{Francis200F17}}$ + 0 $\mu_{\text{FUSfall2018}}$ + -3 $\mu_{\text{Psy200Spring15}}$ + 2 $\mu_{\text{FS16PSY200}}$ + 0 $\mu_{\text{PSY2008HKIED}} = 0$

H_a: Two-tails

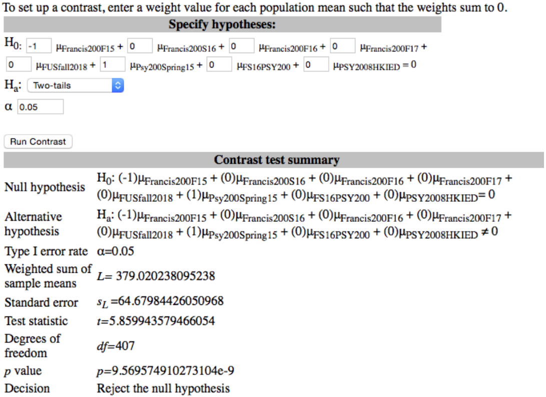
α 0.05

Contrast test summary

Null hypothesis	H_0: $(-3)\mu_{\text{Francis200F15}} + (2)\mu_{\text{Francis200S16}} + (2)\mu_{\text{Francis200F16}} + (0)\mu_{\text{Francis200F17}} + (0)\mu_{\text{FUSfall2018}} + (-3)\mu_{\text{Psy200Spring15}} + (2)\mu_{\text{FS16PSY200}} + (0)\mu_{\text{PSY2008HKIED}} = 0$
Alternative hypothesis	H_a: $(-3)\mu_{\text{Francis200F15}} + (2)\mu_{\text{Francis200S16}} + (2)\mu_{\text{Francis200F16}} + (0)\mu_{\text{Francis200F17}} + (0)\mu_{\text{FUSfall2018}} + (-3)\mu_{\text{Psy200Spring15}} + (2)\mu_{\text{FS16PSY200}} + (0)\mu_{\text{PSY2008HKIED}} \neq 0$
Type I error rate	α=0.05
Weighted sum of sample means	L= -1302.4463233434972
Standard error	s_L =249.66291788092158
Test statistic	t=-5.216819279364137
Degrees of freedom	df=407
p value	p=2.905150702225967e-7
Decision	Reject the null hypothesis

SPECIAL CASE

- Comparing two means is just a special case where the contrast weights for those means are set to ±1 and the other weights are set to 0:



To set up a contrast, enter a weight value for each population mean such that the weights sum to 0.

Specify hypotheses:

H_0: -1 $\mu_{Francis200F15}$ + 0 $\mu_{Francis200S16}$ + 0 $\mu_{Francis200F16}$ + 0 $\mu_{Francis200F17}$ + 0 $\mu_{FUSfall2018}$ + 1 $\mu_{Psy200Spring15}$ + 0 $\mu_{FS16PSY200}$ + 0 $\mu_{PSY2008HKIED}$ = 0

H_a: Two-tails

α 0.05

Run Contrast

Contrast test summary

Null hypothesis	H_0: $(-1)\mu_{Francis200F15} + (0)\mu_{Francis200S16} + (0)\mu_{Francis200F16} + (0)\mu_{Francis200F17} + (0)\mu_{FUSfall2018} + (1)\mu_{Psy200Spring15} + (0)\mu_{FS16PSY200} + (0)\mu_{PSY2008HKIED} = 0$
Alternative hypothesis	H_a: $(-1)\mu_{Francis200F15} + (0)\mu_{Francis200S16} + (0)\mu_{Francis200F16} + (0)\mu_{Francis200F17} + (0)\mu_{FUSfall2018} + (1)\mu_{Psy200Spring15} + (0)\mu_{FS16PSY200} + (0)\mu_{PSY2008HKIED} \neq 0$
Type I error rate	α=0.05
Weighted sum of sample means	L= 379.020238095238
Standard error	s_L =64.67984426050968
Test statistic	t=5.859943579466054
Degrees of freedom	df=407
p value	p=9.569574910273104e-9
Decision	Reject the null hypothesis

- This gives the same result as we computed previously

MULTIPLE TESTING

- There are an *enormous* number of different contrasts that you could create
- If you require a significant ANOVA before running any contrasts, then you can control the Type I error rate to be no higher than α
- However, we have a new kind of “conditional” Type I error
- Given that the ANOVA indicates there is some difference in means, what means (or combinations of means) differ? For some contrasts the H_0 is true, but, just due to random sampling, they indicate that there is a difference

MULTIPLE TESTING

- Thus, we have a new multiple testing problem for identifying the differences; even though we only get to that situation with probability α if the ANOVA omnibus H_0 is true
- Worse, it could be that $\mu_7 \neq \mu_8$, so you reject the ANOVA H_0
- but then you run contrasts for other means where $\mu_i = \mu_j$
- Generally, it is not a good idea to try all possible contrasts. Contrasts (and hypothesis testing in general) make the most sense when you have some specific plans to compare combinations of means

CONCLUSIONS

- interpreting an ANOVA
- identifying differences
- contrast tests

NEXT TIME

- power for ANOVA
- power for contrasts

Keep it simple!