MATH 7343 Applied Statistics

Prof. (Aidong) Adam Ding

Northeastern University

Review

- Last time, we finished Module 5 Two Population Means Comparison

- Today we will first have a brief review of the materials covered up to now. Then we start on the next topic Analysis of Variance (ANOVA)

Review

- (1) Basic concepts: Population (targeted or sample) vs sample, parameter vs statistics

Example: We want to know the mean body length of squirrels. We catch 20 squirrels in Boston and measure their lengths.

Target population:

Sampled population:

Sample:

Parameter:

Statistic:

Review

- (2) Observational study versus controlled study

  Conclusion: association vs cause-effect.

- Example: In an animal study, we wish to compare brain scans between the sleeping mice versus awake mice.

If we simply scan those who happen at the time to be awake or be sleeping, this is an observation study.

Can we make it a controlled study? How?

Review

- (3) Descriptive (sample) statistics

  center of data: mean versus median

  spread of data: variance, standard deviation, range, IQR, etc.

- (4) Inference on population mean

a) Point estimation $\overline{X}$

b) Confidence intervals:

σ known: 2-sided $\overline{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$, and 1-sided variation.

σ unknown: 2-sided $\overline{X} \pm t_{n-1,\, \alpha/2} \frac{s}{\sqrt{n}}$, and 1-sided variation

Review

- (4) Inference on population mean

c) Hypothesis testing

$\mathbf{H_0}$: $\mu = \mu_0$ versus $\mathbf{H_A}$: $\mu \neq \mu_0$:

when σ known, use z-test; when σ unknown, use t-test.

Also, one-sided test $\mathbf{H_0}$: $\mu = \mu_0$ versus $\mathbf{H_A}$: $\mu < \mu_0$.

d) P-value = the maximum level at which $\mathbf{H_0}$ is rejected for the observed data.

practical significance versus statistical significance (p-value)

- (5) Power and sample size calculations.

Review

- (6) Two population means comparison

Data $X_{1,1},\ldots,X_{1,n1}$ versus $X_{2,1},\ldots,X_{2,n2}$

(A) Paired data ($n_1=n_2$). Then $\mathbf{H_0}$: $\mu_1=\mu_2 \Leftrightarrow \mathbf{H_0}$: $\mu_1-\mu_2=0$

Use univariate inference methods on differences $D_i = X_{1,i} - X_{2,i}$

(B) Unpaired independent two samples

(a) $\sigma_1=\sigma_2=\sigma$ but value unknown, $\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{s_p\sqrt{1/n_1+1/n_2}} \sim t_{n_1+n_2-2}$ exactly.

(b) generally, $\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{s_1^2/n_1+s_s^2/n_2}} \sim t_{df}$ approximately with

$$df = \frac{(s_1^2/n_1)^2+(s_2^2/n_2)^2}{\sqrt{(s_1^2/n_1)^2/(n_1-1)+(s_2^2/n_2)^2/(n_2-1)}}$$

One-way Analysis of Variance (ANOVA) Chap12

The problem is to compare means among several populations (or treatment groups). We want

- (1) Test for an overall difference.
- (2) Pinpoint the difference.

Assumptions:

1. The data within one group represents a random sample from a population.
2. The populations are independent;
3. Within the i-th group, the observations are normally distributed with mean μ_i
4. Variance σ^2 is the same for all groups.

One-way ANOVA

- Example: FEV for patients in three hospitals. (P287 of the textbook)

The summary statistics for each group are calculated below

$n_1 = 21$ $\bar{x}_1 = 2.63$ liters $s_1 = 0.496$ liters	$n_2 = 16$ $\bar{x}_2 = 3.03$ liters $s_2 = 0.523$ liters	$n_3 = 23$ $\bar{x}_3 = 2.88$ liters $s_3 = 0.498$ liters

TABLE 12.1
Forced expiratory volume in one second for patients with coronary artery disease sampled at three different medical centers

Johns Hopkins	Rancho Los Amigos	St. Louis
3.23	3.22	2.79
3.47	2.88	3.22
1.86	1.71	2.25
2.47	2.89	2.98
3.01	3.77	2.47
1.69	3.29	2.77
2.10	3.39	2.95
2.81	3.86	3.56
3.28	2.64	2.88
3.36	2.71	2.63
2.61	2.71	3.38
2.91	3.41	3.07
1.98	2.87	2.81
2.57	2.61	3.17
2.08	3.39	2.23
2.47	3.17	2.19
2.47		4.06
2.74		1.98
2.88		2.81
2.63		2.85
2.53		2.43
		3.20
		3.53

One-way ANOVA

- (1) Test for an overall difference. $\mathbf{H_0}$: $\mu_1 = \cdots = \mu_k = \mu$
- Data $X_{ij} \sim N(\mu_i, \sigma^2)$, $i=1,\ldots,k$, $j=1,\ldots,n_i$.

Point estimates for group means $\hat{\mu}_i = \bar{X}_i = \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i}$

Estimate σ^2 by pooling $s^2_{within} = \frac{(n_1-1)s_1^2+\cdots+(n_k-1)s_k^2}{n_1+\cdots+n_k-k}$

Short hand notation $n = n_1 + \cdots + n_k$, the overall mean $\bar{X} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}}{n} = \frac{\sum_{i=1}^{k} n_i \bar{X}_i}{n}$

And the between group variance is $s^2_{between} = \frac{\sum_{i=1}^{k} n_i(\bar{X}_i-\bar{X})^2}{n}$

One-way ANOVA

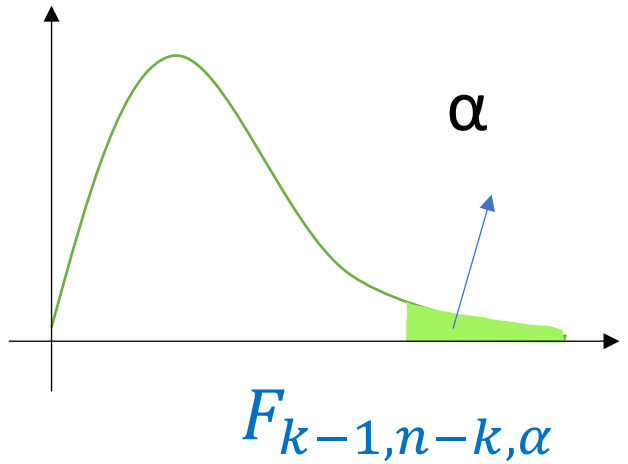
- (1) Test for an overall difference. **H_0**: $\mu_1 = \cdots = \mu_k = \mu$

Theoretically, $s^2_{within} \sim \frac{\sigma^2 \chi^2_{n-k}}{n-k}$ always, and

under H_0, $s^2_{between} \sim \frac{\sigma^2 \chi^2_{k-1}}{k-1}$. Thus

$F_{obs} = \frac{s^2_{between}}{s^2_{within}} \sim F_{k-1,n-k}$ under H_0

- Reject H_0 if $F_{obs} > F_{k-1,n-k,\alpha}$ (from Table A.5)



One-way ANOVA

- (1) Test for an overall difference. $\mathbf{H_0}$: $\mu_1 = \cdots = \mu_k = \mu$

FEV example:

$n_1 = 21$	$n_2 = 16$	$n_3 = 23$
$\bar{x}_1 = 2.63$ liters	$\bar{x}_2 = 3.03$ liters	$\bar{x}_3 = 2.88$ liters
$s_1 = 0.496$ liters	$s_2 = 0.523$ liters	$s_3 = 0.498$ liters

$$\bar{X} = \frac{21*2.63+16*3.03+23*2.88}{60} = 2.83$$

$$s^2_{within} = \frac{20*0.496^2+15*0.523^2+22*0.498^2}{60-3} = 0.254$$

$$s^2_{between} = \frac{21*(2.63-2.83)^2+16*(3.03-2.83)^2+23*(2.88-2.83)^2}{60} = 0.769$$

$$F_{obs} = \frac{0.769}{0.254} = 3.03.$$

k-1=2, n-k=60-3=57. From Table A.5, $F_{2,57,0.10}$ =2.39, $F_{2,57,0.05}$ =3.15.

Hence 0.05< **p-value** < 0.10.

One-way ANOVA Table

Above analysis is usually summarized in a ANOVA table.

Sources of variation	Degrees of freedom (DF)	Sum of squares (SS)	Mean of squares (MS)	F	P-value
Treatment	$DF_{Trt} = k-1$	$SSB = SS_{Between}$	SSB/DF_{Trt}	MSB/MSE	$P(F_{k-1,n-k} > F_{obs})$
Error (Residuals)	$DF_{Err} = n-k$	$SSE = SS_{within}$	SSB/DF_{Err}		
Total	$n-1$	SST =			

- For the FEV example,

	DF	SS	MS	F	Pr (>F)
Center	2	1.5825	0.79142	3.1153	0.052
Residuals	57	14.4803	0.25404		
~~Total~~	~~59~~	~~16.0631~~			

- The last line is not in R outputs but is outputted by some other statistical software.

Basic ANOVA Equation

The ANOVA use a fundamental decomposition:

Total Variation (SST) = Between Variation(SSB) + Within Variation (SSE)

- SST = $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\bar{X})^2$ $= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}^2-2X_{ij}\bar{X}+\bar{X}^2)$

  $= \sum_{i=1}^{k}\sum_{j=1}^{n_i}X_{ij}^2-2\bar{X}(\sum_{i=1}^{k}\sum_{j=1}^{n_i}X_{ij})+\mathrm{n}\bar{X}^2$ $= \sum_{i=1}^{k}\sum_{j=1}^{n_i}X_{ij}^2-\mathrm{n}\bar{X}^2$

  (note $\mathrm{n}\bar{X}=\sum_{i=1}^{k}\sum_{j=1}^{n_i}X_{ij}$, so this becomes -2 $\mathrm{n}\bar{X}^2$+ $\mathrm{n}\bar{X}^2$)

- SSB = $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{X}_i-\bar{X})^2$ $=\sum_{i=1}^{k}n_i(\bar{X}_i-\bar{X})^2$

  $=\sum_{i=1}^{k}n_i\bar{X}_i^2-2\bar{X}(\sum_{i=1}^{k}n_i\bar{X}_i)+(\sum_{i=1}^{k}n_i)\bar{X}^2$ $=\sum_{i=1}^{k}n_i\bar{X}_i^2-\mathrm{n}\bar{X}^2$

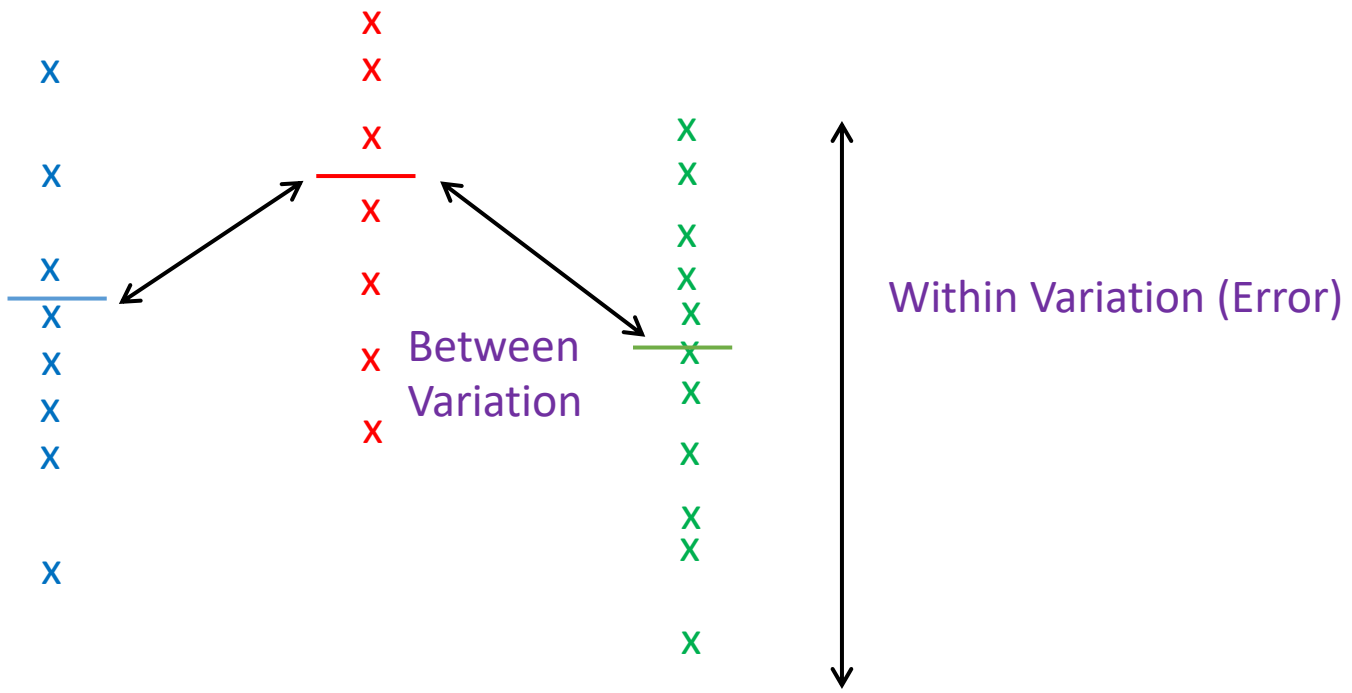
- SSE = $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_i)^2$ $=\sum_{i=1}^{k}\sum_{j=1}^{n_i}X_{ij}^2-2\sum_{i=1}^{k}\bar{X}_i\sum_{j=1}^{n_i}X_{ij}+\sum_{i=1}^{k}n_i\bar{X}_i^2$

  $=\sum_{i=1}^{k}\sum_{j=1}^{n_i}X_{ij}^2-2\sum_{i=1}^{k}\bar{X}_in_i\bar{X}_i+\sum_{i=1}^{k}n_i\bar{X}_i^2$ $=\sum_{i=1}^{k}\sum_{j=1}^{n_i}X_{ij}^2-\sum_{i=1}^{k}n_i\bar{X}_i^2$

- Combine the above three expressions, **SST=SSB+SSE**

One-way ANOVA idea

- (1) Test for an overall difference. $\mathbf{H_0}$: $\mu_1 = \cdots = \mu_k = \mu$



- Basic idea: Compare the ratio Between Variation/ Within Variation.
  If the ratio big, the differences among groups are real.

One-way ANOVA

- (1) Test for an overall difference. $\mathbf{H_0}$: $\mu_1 = \ldots = \mu_k = \mu$
- Basic idea: Compare the ratio Between Variation/ Within Variation.
- How to quantify these two variations? Use Mean Squares: $\frac{s^2_{between}}{s^2_{within}} = \frac{MSB}{MSE} = \frac{SSB/DF_B}{SSE/DF_E}$.
- $F_{obs} = \frac{MSB}{MSE} \sim F_{DF_B, DF_E}$ under H_0.
- So we use the F-test (cutoff from Table A.5).
- Reject H_0 if $F_{obs} > F_{k-1, n-k, \alpha}$

R commands to do ANOVA

- aov(Y~group). Example in ANOVA.pdf handout.

```
> # Import data set. This is formated two columns/variables
> PF.data <- read.table(file="pulmonary function.txt", header=TRUE, na.strings =
".")
> # Display a few lines
> head(PF.data)
  center fev1
1      1 3.23
2      1 3.47
3      1 1.86
4      1 2.47
5      1 3.01
6      1 1.69
>
> # center is a categorical (factor) variable
> #We need to do this for the grouping variable
> PF.data$center<-as.factor(PF.data$center)
>
> # Fit by aov(), then prodcue the ANOVA table
> PF.fit <- aov(fev1~center, data=PF.data)
> summary(PF.fit)
             Df Sum Sq Mean Sq F value Pr(>F)
center        2  1.583  0.7914   3.115  0.052 .
Residuals    57 14.480  0.2540
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
3 observations deleted due to missingness
```

One-way ANOVA

- (2) Pinpoint the difference. After F-test for (1)

  $\mathbf{H_0}$: $\mu_1 = \cdots = \mu_k = \mu$ vs. $\mathbf{H_A}$: some of the '=' are violated.

- If the F-test reject $\mathbf{H_0}$, which groups are different?

We may be interested in

1. A particular (set of) hypothesis about the means.
2. Find the biggest means.
3. Difference among pairs of means.

One-way ANOVA

- (2) Pinpoint the difference.
- FEV Example: We are interested in the difference between John Hopkins and Rancho Los Amigos patients

Johns Hopkins	Rancho Los Amigos	St. Louis
3.23	3.22	2.79
3.47	2.88	3.22

- Parameter: μ_1-μ_2
- Point estimator: $\bar{X}_1$ -$\bar{X}_2$ =2.63-30.3 = -0.4

$n_1 = 21$	$n_2 = 16$	$n_3 = 23$
$\bar{x}_1 = 2.63$ liters	$\bar{x}_2 = 3.03$ liters	$\bar{x}_3 = 2.88$ liters
$s_1 = 0.496$ liters	$s_2 = 0.523$ liters	$s_3 = 0.498$ liters

Var($\bar{X}_1$ -$\bar{X}_2$)= $\sigma^2(1/n_1 + 1/n_2)$,

Estimate σ^2 by the MSE= s^2 =0.254, which has df =n-k =57

- The inferences are based on $\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{s\sqrt{1/n_1+1/n_2}} \sim t_{n-k}$

Notice the degree of freedom here is n-k. **NOT** $n_1 + n_2 - 2$ as in last chapter.

One-way ANOVA

- FEV Example: We are interested in the difference between John Hopkins and Rancho Los Amigos patients
- To test $\mathbf{H_0}$:μ_1=μ_2 $\Leftrightarrow$ $\mathbf{H_0}$:$\mu_1-\mu_2$=0

$$T_{obs} = \left|\frac{\bar{X}_1-\bar{X}_2}{s\sqrt{1/n_1+1/n_2}}\right| = \left|\frac{-0.4}{\sqrt{0.254}\sqrt{1/21+1/16}}\right| = 2.43$$

Compare with $t_{n-k,\alpha/2} = t_{57,\alpha/2}$ from TableA.4, $t_{60,0.01}$=2.39

we see that p-value < 2(0.01) = 0.02.

Reject $\mathbf{H_0}$ at α=0.05 level and thus conclude $\mu_1 \neq \mu_2$.

There are differences in the FEV for patients in these two medical centers.

One-way ANOVA

- (2) Pinpoint the difference.
- In the above example, we made the comparison on μ_1-μ_2

  May do other comparisons: μ_2-μ_3, μ_1-0.5(μ_2+μ_3) or 2μ_1-(μ_2+μ_3)
- Comparison: A linear combination of group means

$$L=\lambda_1\mu_1 + \lambda_2\mu_2 + \ \dots + \lambda_k\mu_k$$

Estimator is $\hat{L}=\lambda_1\bar{X}_1 + \lambda_2\bar{X}_2 + \ \dots + \lambda_k\bar{X}_k \sim N(L,\sum_{i=1}^{k}\lambda_i^2\frac{\sigma^2}{n_i})$

- Contrast: A comparison L with $\lambda_1 + \lambda_2 + \ \dots + \lambda_k=0$.
- We generally are interested in $\mathbf{H_0}$: L=0 for a contrast.

Since $\hat{L} \sim N(0, \sigma^2 \ \sum_{i=1}^{k}\frac{\lambda_i^2}{n_i})$ under $\mathbf{H_0}$, we can use t-test for testing.

One-way ANOVA: R commands to do contrast

- See the handout (anova.pdf) on the FEV example
- To test the contrast μ_1-μ_2 (John Hopkins versus Rancho Los Amigos)

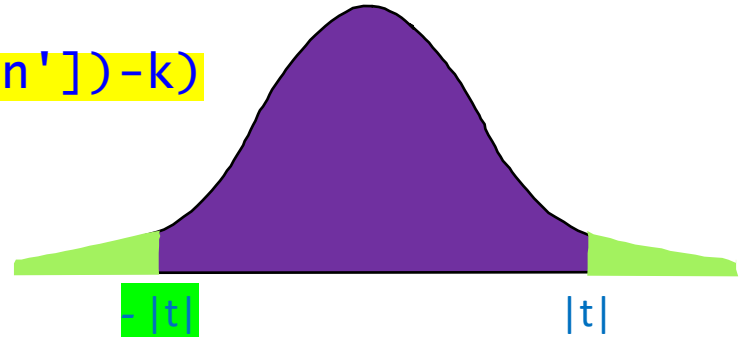
```
> #Contrast of John Hopkins (center 1) versus Rancho Los Amigos (2)
> contr1<-c(1, -1, 0)                              # 1(μ1)+(-1) μ2 + 0(μ3)
> #Find groups means and sds by center
> require(psych)
> stat.fev1<-describeBy(PF.data$fev1,PF.data$center)
> require(data.table)
> stat.fev2<-data.frame(rbindlist(stat.fev1))[,c('n','mean','sd')]
> stat.fev2 #display
   n     mean        sd
1 21 2.626190 0.4961701
2 16 3.032500 0.5232399
3 23 2.878696 0.4977157
> # Estimate of contrast
> contr1.est<-sum(contr1*stat.fev2[,'mean'])    #contr1·group.means
```

One-way ANOVA: R commands to do contrast

- See the handout (anova.pdf) on the FEV example
- To test the contrast μ_1-μ_2 (John Hopkins versus Rancho Los Amigos)

```
> contr1.se<- sqrt(sum(contr1^2/stat.fev2[,'n'])*MSE)   # sqrt(sum_{i=1}^{k} lambda_i^2 sigma^2/n_i)
> contr1.t<- contr1.est/contr1.se
> k<-length(levels(PF.data$center)) #number of groups
> contr1.p<- 2*pt( -abs(contr1.t),df=sum(stat.fev2[,'n'])-k)

> #Display estimate, se, t-statistic, p-value
> c(contr1.est, contr1.se, contr1.t, contr1.p)
[1] -0.40630952  0.16725608 -2.42926607  0.01830437
```



Here t = -2.43. p-value = 0.0183.

Reject $\mathbf{H_0}$ at α=0.05 level.

Fail to reject $\mathbf{H_0}$ at α=0.01 level.

One-way ANOVA

- (2) Pinpoint the difference.
- How to do multiple contrasts?
- Orthogonal contrasts: L=$\lambda_1\mu_1 + \lambda_2\mu_2 + \ \ldots + \lambda_k\mu_k$ and L*=$\lambda_1^*\mu_1 + \lambda_2^*\mu_2 + \ \ldots + \lambda_k^*\mu_k$ are orthogonal if and only if $\lambda_1\lambda_1^* + \lambda_2\lambda_2^* + \ \ldots + \lambda_k\lambda_k^*$=0.
- If L and L* are orthogonal, then $\hat{L}$ and $\hat{L}^*$ are independent. (So easy to derive joint probability.)
- If we have (k-1) orthogonal contrasts, then all other contrasts are linear combinations of the (k-1) contrasts.

R commands to do contrasts

- Example in ANOVA.pdf handout.
- We will do three contrasts: μ_1-μ_2, μ_2-μ_3 and 2μ_1-(μ_2+μ_3)
- Can do them one by one as shown above. Quicker to do them using a design matrix, which will need (k-1) non-linear-dependent contrasts. Here we have k=3 groups, so that means we can do 2 contrasts at a time.

```
> contr1<-c(1, -1, 0)    #compare the first two centers
> contr2<-c(0,1,-1)      #compare the last two centers
> contr3<-c(2,-1,-1)     #compare the first center with the last two together.
> contr.mat <- rbind(rep(1/3, 3), contr1, contr2)  #base vector (1,1,1)/3 is (μ1+μ2+μ3)/3
> my.contr <- solve(contr.mat)[,-1] ## Get the inverse matrix, put into the contrasts in lm()
> contrasts(PF.data$center)<-my.contr
> summary(lm(fev1~center, data=PF.data) )
```

R commands to do contrast

- Using the design matrix to do 1st and 2nd contrasts: μ_1-μ_2 and μ_2-μ_3

```
> summary(lm(fev1~center, data=PF.data))
Call:
lm(formula = fev1 ~ center, data = PF.data)
Residuals:
     Min       1Q   Median       3Q      Max
-1.32250 -0.32250 -0.02244  0.32630  1.18130
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.84580    0.06584  43.220   <2e-16 ***
centercontr1 -0.40631    0.16726  -2.429   0.0183 *
centercontr2  0.15380    0.16408   0.937   0.3525
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.504 on 57 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.09854,  Adjusted R-squared:  0.06691
F-statistic: 3.115 on 2 and 57 DF,  p-value: 0.052
```

- Compare with the results we did above for the first contrast, we can see the results are the same.

```
> c(contr1.est, contr1.se, contr1.t, contr1.p)
[1] -0.40630952  0.16725608 -2.42926607  0.01830437
```

R commands to do contrast

To get the third contrast, use a new design matrix

```
> contr.mat <- rbind(rep(1/3, 3), contr2, contr3) #contrasts 2 and 3
> my.contr <- solve(contr.mat)[,-1] ## Get the inverse matrix, put into the contra
sts in lm()
> contrasts(PF.data$center)<-my.contr
> summary(lm(fev1~center, data=PF.data))
Call:
lm(formula = fev1 ~ center, data = PF.data)
Residuals:
     Min       1Q   Median       3Q      Max
-1.32250 -0.32250 -0.02244  0.32630  1.18130
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.84580    0.06584  43.220   <2e-16 ***
centercontr2  0.15380    0.16408   0.937   0.3525
centercontr3 -0.65881    0.27443  -2.401   0.0197 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.504 on 57 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.09854,  Adjusted R-squared:  0.06691
F-statistic: 3.115 on 2 and 57 DF,  p-value: 0.052
```

One-way ANOVA

- (2) Pinpoint the difference.

- Here and in next lecture, we are using this FEV example to show the R commands to conduct contrasts (t-tests). However, if we are doing real data analysis on this data set, we would never do the stage 2 analysis of pinpointing the difference, since the F-test in first stage did not reject. So know that this is a just an example showing the technical commands.

R commands to do contrast

- Summarize above R outputs

	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	*

- This output allow us to do inference for each contrast.
- However, we also need to do multiple testing adjustment.

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**

Summary

Today, we introduced ANOVA

- The basic setup: comparing several population means.
- F-test for overall difference
- t-test for contrasts/comparisons
- Adjustment is needed for multiple contrasts testing.

- Next lecture we discuss how to do the multiple testing adjustments and other extensions of ANOVA.