hw-lect6-1

Show that $\sum_{j}^{n} (y_{ij} - \bar{y}_{i.}) = 0$

$$\sum_j (y_{ij} - \bar{y}_{i.}) = \sum_j y_{ij} - \sum_j \bar{y}_{i.}$$

$$= y_{i.} - \bar{y}_{i.} \underbrace{\sum_j^n 1}_{n} = y_{i.} - n\bar{y}_{i.} = y_{i.} - y_{i.} = 0$$

---

hw-lect6-2

We showed that $SS_{Treatment} = n \sum_{i}^{a} (\bar{y}_{i.} - \bar{y}_{..})^2$.

But the book, in example 3.1 (for example) uses

$$SS_{Treatment} = \frac{1}{n} \sum_{i=1}^{a} y_{i.}^2 - \frac{y_{..}^2}{N}$$

prove that the two expressions are equal.

$$SS_{Tr} = n \sum_{i}^{a} (\bar{y}_{i.} - \bar{y}_{..})^2 = n \sum_{i}^{a} \left[(\bar{y}_{i.})^2 + (\bar{y}_{..})^2 - 2\bar{y}_{i.}\bar{y}_{..} \right]$$

$$= n \sum_i \left[\left(\frac{1}{n} y_{i.}\right)^2 + \left(\frac{1}{an} y_{..}\right)^2 - 2\frac{1}{n} y_{i.} \frac{1}{an} y_{..} \right]$$

$$= \frac{1}{n} \sum_i (y_{i.})^2 + \frac{1}{a^2 n} (y_{..})^2 \underbrace{\sum_i 1}_{a} - \frac{2}{an} y_{..} \underbrace{\sum_i y_{i.}}_{y_{..}}$$

$$= \frac{1}{n} \sum_i (y_{i.})^2 - \frac{1}{an} (y_{..})^2$$

hw-lect6-3

Consider the following data on y involving a factor x with 4 levels, and 10 replications:

```
y = matrix(nrow=4,ncol=10)
  y[1,] = c( -2.10552316, 1.89491371, -1.52919682, -0.99265143, -0.45911960, 1.09271028, -1.54680778,
0.13890677, 0.06240357, -1.09273045)
  y[2,] = c( 4.25667943, 4.36518096, 4.42108835, 3.77229146, 2.22264903, 3.95354759, 6.29377745, 3.58501081,
3.12457306, 3.04360597)
  y[3,] = c( 3.64209745, 2.76932242, 1.46001019, 0.23739519, 0.27629510, 2.83897173, 2.99999590, 3.54657820,
2.03955378, 1.28515784)
  y[4,] = c( 3.18088593, 5.44976665, 5.87116946, 4.01275036, 6.00826692, 5.19220036, 6.17338313, 4.88846073,
4.49330445, 4.83224707)
```

a) Make a comparative boxplot of y for the 4 levels of x. Based on this plot, do you think that x has an effect on y? Explain.

```
a = 4
n = 10
x = rep(1:a,n)

plot(x, y)       # scatterplot
boxplot(t(y))    # comparative boxplots
```

Based on these plots, it's likely that x has an effect on y, because at least two of the treatment levels appear to
have y-values coming from distributions with different mu values. For example, X=1 and X=4 appear to come from
dists with different means.

b) Compute the sample grand mean of all y_{ij} observations, and subtract it from all y measuremets, i.e. y_{ij} - grand mean. Recall, i goes from 1 to the number of levels in x, and j goes from 1 to the number of replications in your data. Then plot the comparative boxplots of this "new" data, still for the different levels of x. What is your conclusion now - do you think that x has an effect on (y - grand mean)?

```
y_new = y - mean(y)
plot(x, y_new)    # scatterplot
boxplot(t(y_new))    # comparative boxplots
```

The whole plot has simply shifted down to around 0. Otherwise, the conclusion is the same as in part a.

c) Compute the sample (conditional) mean of y, for each of the levels of x. Call these y_bar. Then, subtract the grand mean from these means, i.e., (y_bar[1] - grand mean), (y_bar[2] - grand mean), ...Each of these is called an effect.

```
y_bar = apply(y,1,mean)              # y_bar1, y_bar2, ...
effects = y_bar-mean(y)
effects                              # -3.0961875  1.2613624 -0.5329403  2.3677655
```

These quantities are estimates of the *effects* I talked about in class. We'll see that later.

d) Now compute the standard error of the conditional means. For now, approximate it by the sd of the y's in each treatement level, divided by sqrt(n), where n is the number of observations at each treatment level.

```
y_std_err = apply(y,1,sd) /sqrt(n)
```

e) You now have 4 effects, each accompanied by a standard error. Based on these (mean +- std.err) values, without doing any other calculation, does it look like any of the true effects can be zero? Is your conclusion here consistent with the conclusions above?

```
# -3.0961875  +- 0.3986627
#  1.2613624  +- 0.3433781
# -0.5329403  +- 0.3964595
#  2.3677655  +- 0.2976147
```

The third effect may be zero, because 0 is included in the interval, but all of the other effects are nonzero.
So, yes, the conclusion is the same, that at least one of the effects is nonzero.

f) Now perform a 1-way ANOVA on x and y, by hand (i.e. using the formulas we have developed), for testing whether x has an effect on y. Use the rejection region method. State your conclusion "in English." .

```
vars = apply(y,1,var)
grand.mean = mean(y_bar)
SS_between = n*sum( effects^2 )
SS_within = (n-1)*sum(vars)
F_ratio = (SS_between/(a-1))/(SS_within/(n*a-a))   # 43.54614
# At alpha = 0.05, according to Table IV, the critical value of F is about 2.84,
# with df =(a-1. n*a-a), Since 43.54 > 2.84, we reject H0 in favor of H1,
# i.e., at least two of the means are different.
# In English: x has an effect on y.

# This part of the problem didn't ask for a p-value, but if it had asked, it would be:
pf(F_ratio,a-1,n*a-a, lower.tail=F)
```

g)) Now perform a 1-way ANOVA on x and y, by R.. Find the p-value, and again state your conclusion "in English."

```
y.vector = as.numeric(y) # visually confirm y.vector is the correct y in vector form.
summary.aov(lm(y.vector ~ as.factor(x)) )
```

Same answers as part f.

hw-lect6-4

The proof that $E[MSE]=\sigma_\epsilon^2$ is a bit complex; but as a warm-up exercise show that if y_i are iid, then $E[\overline{y^2}]-E[\bar{y}^2]=\frac{n-1}{n}V[y]$. Show where each of the i's in "iid" is used. Hint: This is similar to the proof $E[s^2]=\sigma_y^2$

$$E[\overline{y^2}]-E[\bar{y}^2] = E\left[\frac{1}{n}\sum_i^n y_i^2\right] - E\left[\frac{1}{n}\sum_i y_i \frac{1}{n}\sum_j y_j\right]$$

$$= \frac{1}{n}\sum_i \underbrace{E[y_i^2]}_{\substack{\text{identically dist.}\\ E[y^2]}} - \frac{1}{n^2}\underbrace{\sum_{ij} E[y_i y_j]}$$

$$\sum_i \underbrace{E[y_i y_i]}_{\substack{\text{identically dist}\\ E[y^2]}} + 2\sum_{i<j}\underbrace{E[y_i y_j]}_{\substack{\text{independently}\\ E[y_i]E[y_j]\\ \text{identically}\\ E[y]E[y]}}$$

$$= \frac{1}{n}nE[y^2] - \frac{1}{n^2}\left(nE[y^2]+2\frac{n(n-1)}{2}E^2[y]\right)$$

$$= E[y^2]-\frac{1}{n}E[y^2]-\frac{n-1}{n}E^2[y]$$

$$= \frac{(n-1)}{n}E[y^2]-\frac{n-1}{n}E^2[y] = \left(\frac{n-1}{n}\right)\left(E[y^2]-E^2[y]\right)$$

$$= \frac{n-1}{n}V[y]$$

hw-lect7-1

The proof that $E[MS_{Tr}] = \sigma_\epsilon^2 + \frac{n}{a-1}\sum_i^a \tau_i^2$ tricky. So, I will walk you through it. First, go over this proof of $E[MSE] = \sigma_\epsilon^2$.

$$E[MSE] = \frac{1}{N-a} E[SSE] = \frac{1}{N-a} E\left[\sum_{ij}(y_{ij}-\bar{y}_{i\cdot})^2\right] \quad (*)$$

Model: $y_{ij} = \mu + \tau_i + \epsilon_{ij} \xrightarrow{\frac{1}{n}\sum_j} \bar{y}_{i\cdot} = \mu + \tau_i + \bar{\epsilon}_{i\cdot}$

$$= \frac{1}{N-a} E\left[\sum_{ij}(\cancel{\mu}+\cancel{\tau_i}+\epsilon_{ij} - \cancel{\mu} - \cancel{\tau_i} - \bar{\epsilon}_{i\cdot})^2\right]$$

$$= \frac{1}{N-a} E\left[\sum_{ij}(\epsilon_{ij}-\bar{\epsilon}_{i\cdot})^2\right] = \frac{1}{N-a}\sum_i^a E\left[\sum_j^n(\epsilon_{ij}-\bar{\epsilon}_{i\cdot})^2\right]$$

In lect3

Check the derivation of $E[s^2] = \sigma_y^2$, which can be written as

$$E\left[\sum_i^n (y_i-\bar{y})^2\right] = (n-1)\,\sigma_y^2 . \quad \mathrm{I}$$

Note that the only thing we assumed was $y_i \sim$ iid. No normality!

(*) Then for each term in the $\sum_i$ sum we can write:

var. of ϵ in ith group.

$$E\left[\sum_j^n(\epsilon_{ij}-\bar{\epsilon}_{i\cdot})^2\right] = (n-1)\underbrace{(\sigma_\epsilon^2)_i}_{} = \sigma_\epsilon^2 \;\forall i$$

$$\therefore E[MSE] = \frac{1}{an-a}\sum_i^a (n-1)\sigma_\epsilon^2 = \frac{a(n-1)}{a(n-1)}\sigma_\epsilon^2 = \sigma_\epsilon^2 .$$

homoscedasticity

(*) You could also say, for each term in the sum $\sum_i$

$$E\left[\sum_j^n (y_{ij}-\bar{y}_{i\cdot})^2\right] = (n-1)\underbrace{(\sigma_y^2)_i}_{} = (\sigma_\epsilon^2)_i = \sigma_\epsilon^2$$

above model. homosc....

$$\therefore E[MSE] = \frac{1}{an-a}\sum_i^a (n-1)\sigma_\epsilon^2 = \sigma_\epsilon^2 .$$

Now, let's find $E[MS_{Tr}]$:

$$E[MS_{Tr}] = E\left[\frac{n}{a-1}\sum_{i}^{a}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2\right]$$

Model: $Y_{ij} = \mu + \tau_i + \epsilon_{ij} \xrightarrow{\frac{1}{n}\sum_j^n} \bar{Y}_{i\cdot} = \mu + \tau_i + \bar{\epsilon}_{i\cdot} \longrightarrow \bar{Y}_{\cdot\cdot} = \mu + \bar{\tau}_{\cdot} + \bar{\epsilon}_{\cdot\cdot}$

$$= \frac{n}{a-1} E\left[\sum_{i}^{a}(\cancel{\mu} + \tau_i + \bar{\epsilon}_{i\cdot} - \cancel{\mu} - \bar{\tau}_{\cdot} - \bar{\epsilon}_{\cdot\cdot})^2\right]$$

$$= \frac{n}{a-1} E\left[\sum_{i}^{a}(\tau_i - \bar{\tau}_{\cdot} + \bar{\epsilon}_{i\cdot} - \bar{\epsilon}_{\cdot\cdot})^2\right]$$

$$= \frac{n}{a-1}\left(\underbrace{E\left[\sum_{i}^{a}(\bar{\epsilon}_{i\cdot} - \bar{\epsilon}_{\cdot\cdot})^2\right]}_{(a-1)(\sigma^2_{\bar{\epsilon}_{i\cdot}}) \,=\, (a-1)\frac{\sigma_\epsilon^2}{n}} + \underbrace{E\left[\sum_{i}^{a}(\tau_i - \bar{\tau}_{\cdot})^2\right]}_{\sum_i^a (\tau_i - \bar{\tau}_{\cdot})^2} + \underbrace{2E\left[\sum_i(\tau_i - \bar{\tau}_{\cdot})(\bar{\epsilon}_{i\cdot} - \bar{\epsilon}_{\cdot\cdot})\right]}_{2\sum_i(\tau_i - \bar{\tau}_{\cdot})E[\bar{\epsilon}_{i\cdot} - \bar{\epsilon}_{\cdot\cdot}]}\right)$$

(*) $\sigma^2_{\bar{\epsilon}_{i\cdot}}$ ← Var. of the mean of ϵ's in ith group

$$E[\bar{\epsilon}_{i\cdot} - \bar{\epsilon}_{\cdot\cdot}] = E[\bar{\epsilon}_{i\cdot}] - E[\bar{\epsilon}_{\cdot\cdot}] = \frac{1}{n}\sum_j \underbrace{E[\epsilon_{ij}]}_{=0} - \frac{1}{an}\sum_{ij}\underbrace{E(\epsilon_{ij})}_{=0}$$

$$E[MS_{Tr}] = \frac{n}{a-1}\left(\frac{a-1}{n}\sigma_\epsilon^2 + \sum_i^a(\tau_i - \bar{\tau}_{\cdot})^2\right)$$

$$E[MS_{Tr}] = \sigma_\epsilon^2 + \frac{n}{a-1}\sum_i^a(\tau_i - \bar{\tau}_{\cdot})^2$$

4. $E\left[\sum_i^a (\bar{\epsilon}_{i.} - \bar{\epsilon}_{..})^2\right] = \sum_i^a E\left[(\bar{\epsilon}_{i.} - \bar{\epsilon}_{..})^2\right] = \sum_i \left(V[\bar{\epsilon}_{i.} - \bar{\epsilon}_{..}] + \cancel{E^2[\bar{\epsilon}_{i.} - \bar{\epsilon}_{..}]} \right)$

$$= \sum_i \left(\underbrace{V[\bar{\epsilon}_{i.}]} + \underbrace{V[\bar{\epsilon}_{..}]} - 2 \underbrace{Cov[\bar{\epsilon}_{i.}, \bar{\epsilon}_{..}]} \right)$$

$V[\frac{1}{n}\sum_j \epsilon_{ij}]$

$\frac{1}{n^2} V[\sum_j \epsilon_{ij}]$

$\frac{1}{n^2} \sum_j V[\epsilon_{ij}] + 0$

$\frac{1}{n}\sigma_\epsilon^2$

$\frac{1}{an}\sigma_\epsilon^2$

$Cov[\bar{\epsilon}_{i.}, \frac{1}{a}\sum_l^a \bar{\epsilon}_{l.}]$

$= \frac{1}{a}\underbrace{\sum_l^a Cov[\bar{\epsilon}_{i.}, \bar{\epsilon}_{l.}]}_{Cov[\bar{\epsilon}_{i.}, \bar{\epsilon}_{i.}] = V[\bar{\epsilon}_{i.}] = \frac{\sigma_\epsilon^2}{n}} = \frac{\sigma_\epsilon^2}{an}$

All terms 0, except when $i = l$

$$= \sum_i^a \left[\frac{1}{n}\sigma_\epsilon^2 + \frac{1}{an}\sigma_\epsilon^2 - \frac{2\sigma_\epsilon^2}{an} \right]$$

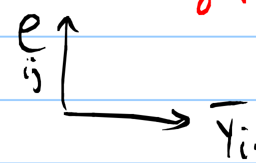
$$= \frac{(a-1)}{n}\sigma_\epsilon^2 .$$

hw - lect 7-2

Consider The PEMF data in problem 3.6 ; ignore The "Sham" col.

a) perform a 1-way ANOVA F-test to see if The number of hours has an effect on bone density.
Report The p-value and The conclusion in English. By R

b) Make The following residual plots, and interpret each: By R



c) Make qq plots for each of The 3 levels of x, and interpret The results. By R

d) Compute The C.I. for μ_1 By hand

e) " " " " $\mu_2 - \mu_3$ By hand

f) Use The contrast method to test

$H_0: \mu_1 = \frac{1}{2}(\mu_2 + \mu_3)$

$H_1: \mu_1 \neq \frac{1}{2}(\mu_2 + \mu_3)$

By hand.

Use The rejection region method, and state your conclusion "In English." use $\alpha = .05$.

g) Compute a 95% C.I for The contrast in part f.

```
rm(list=ls(all=TRUE))

y = matrix( c(
4.51, 5.32, 4.73, 7.03,
7.95, 6.00, 5.81, 4.65,
4.97, 5.12, 5.69, 6.65,
3.00, 7.08, 3.86, 5.49,
7.97, 5.48, 4.06, 6.98,
2.23, 6.52, 6.56, 4.85,
3.95, 4.09, 8.34, 7.26,
5.64, 6.28, 3.01, 5.92,
9.35, 7.77, 6.71, 5.58,
6.52, 5.68, 6.51, 7.91,
4.96, 8.47, 1.70, 4.90,
6.10, 4.58, 5.89, 4.54,
7.19, 4.11, 6.55, 8.18,
4.03, 5.72, 5.34, 5.42,
2.72, 5.91, 5.88, 6.03,
9.19, 6.89, 7.50, 7.04,
5.17, 6.99, 3.28, 5.17,
5.70, 4.98, 5.38, 7.60,
5.85, 9.94, 7.30, 7.90,
6.45, 6.38, 5.46, 7.91), ncol=4, byrow=T)

# a)

Y = y[,2:4]
x = rep(c(1:3), nrow(Y))
Y.vector = c(t(Y))
lm.1 = lm( Y.vector ~ as.factor(x) )
summary.aov(lm.1)

#              Df Sum Sq Mean Sq F value Pr(>F)
# as.factor(x)  2   8.45   4.227    2.01  0.143
# Residuals    57 119.85   2.103

# Not significant. Consistent with
boxplot(Y)

# b)

#   pred = rep ( apply(Y, 2, mean), nrow(Y))  # or  pred = predict(lm.1)
#   resid = Y.vector - pred

pred = predict(lm.1)
resid = lm.1$residuals
plot(pred, resid)
abline(h=0, lt=2)

# Looks nice and random. However, there may be a violation of the equal-var assumption.
```

```
# c) Use the code called qq_by_hand.R we used in a previous hw

   n = nrow(Y)
   X = seq(.5/n,1-.5/n,length=n)
   Q = qnorm(X,0,1)
   plot(Q, sort(Y[,1]),col=1, type="b", ylim=range(Y))
   points(Q, sort(Y[,2]),col=2, type="b")
   points(Q, sort(Y[,3]),col=4, type="b")

# The 3 qqplots are all mostly linear, and so normality is not violated, not even within each of the 3 populations. The slopes are
# mostly the same, and so equality of variance is probably not violated.  Note that the y-intercepts (i.e., estimates of the
# conditional mean of Y for each of the 3 populations), are similar; and that is again consistent with the null result found in the
anova test.

# d) CI for mu1

   a = 3
   n = nrow(Y)
   SSE = sum( t((t(Y) - (apply(Y,2,mean))))^2 )    # pay attention to t()
   MSE = SSE/(n*a-a)                               # = MSE in anova table from lm()
   B =  qt(.05/2, n*a - a, lower.tail=F) * sqrt(MSE/n)
   c( mean(Y[,1]) - B, mean(Y[,1]) + B)            #  5.516 6.815

# e) CI for mu2 - mu3

   B = qt(.05/2, n*a - a, lower.tail=F)*sqrt(2*MSE/n)
   c( mean(Y[,2]) - mean(Y[,3]) - B,  mean(Y[,2]) - mean(Y[,3]) + B)   # -1.797  0.0457 CORRECTION (wrong difference)

# f) contrast method for 2-sided test of  mu1 = (mu2+mu3)/2, using rejection region.

  C = c(1,-.5,-0.5)
  ybars = apply(Y,2,mean)
  gamma_hat = sum(C*ybars)
  t_obs = gamma_hat/sqrt(MSE * sum(C^2)/n)          # 0.632705

# The Rejection Region (RR) for a 2-sided t-test are the regions beyond +- 2:
  qt(.05/2, n*a - a, lower.tail=F)            # 2.002465
# Because t_obs is not in RR, we cannot reject H0: gamma=0 in favor of
#  H1: gamma != 0 . In English, there is no evidence from data that gamma is nonzero.

# g) CI for contrast in part f
  B = qt(.05/2, n*a - a, lower.tail=F) * sqrt(MSE * sum(C^2)/n)
  c( gamma_hat - B, gamma_hat + B)                  # -0.544 1.046

# FYI: (not required for hw)
# Interpretation: We are 95% confident that gamma is between -0.544 and 1.046
# I.e.,    -0.544 < mu1 - 0.5 mu2 - 0.5 mu3 < 1.046
# So, all combinations of mu1, mu2, and mu3 in the above interval are also plausible.
# Specifically, note that this CI includes zero, and so it's plausible that  mu1 = (mu2+mu3)/2 .
```

hw-lect8-1

```
# For the data in problem 3.20 (7th ed) = 3.22 (8th ed.)
rm(list=ls(all=TRUE))
a = 3
n = 5
N = a*n
y.m = matrix(nrow=a,ncol=n)      # rows = factor, col = replicates
y.m[1,] = c(9, 12, 10, 8, 15)
y.m[2,] = c(20, 21, 23, 17, 30)
y.m[3,] = c(6, 5, 8, 16, 7)
y = t(y.m)
y = as.vector(y)

# a) Do the means vary across the levels of y? Report a p-value. (By hand)

means = apply(y.m,1,mean)       # 10.8 22.2  8.4
vars = apply(y.m,1,var)         #  7.7 23.7 19.3

SS_treatment = n*sum((means - mean(means))^2)        # 543.6
SS_E = (n-1)*sum(vars)                               # 202.8
MS_treatment = SS_treatment/(a-1)                    # 271.8
MS_E = SS_E/(N-a)                                    # 16.9
F_obs = MS_treatment/MS_E                            # 16.08284
pf(F_obs, df1=a-1, df2=N-a, 0, lower.tail = FALSE)   # 0.0004023258

# Given that p-value < alpha, we reject H0 (equal means) in favor
# of H1 (at least 2 mus are different).

# b) Find SS_treatement, SS_E, F, and the corresponding p-value, but by R.
#    Confirm that these are equal to those in part a.

A = as.factor(rep(1:a, each = n))
lm.1 = lm(y~A)
summary.aov(lm.1)

# A            2  543.6   271.8   16.08 0.000402 ***
# Residuals   12  202.8    16.9

# c) Compute a 95% CI for mu1-mu3 (by hand)

means[1] - means[3] + qt(.05/2,N-a, lower.tail=T) *sqrt(2*MS_E/n)
means[1] - means[3] + qt(.05/2,N-a, lower.tail=F) *sqrt(2*MS_E/n)

# CI for mu1-mu3: (-3.264913, 8.064913)
```

d) Compute the p-value for testing whether mu1 and mu3 are different. State your conclusion. By hand. Note that
eventhough you are testing only 2 means, the estimate of sigma in the t statistic is the mse of the full model with the
treatment factor having (a) levels.

```
t_obs =  (means[1] - means[3])/sqrt(2*MS_E/n)       # 0.9230769
2*pt(t_obs,N-a,lower.tail=F)                        # 0.374155 (NOTE tail=upper)

# p-value > alpha ---> cannot reject H0 (mu1=m3) in favor of
# H1 (mu1 != mu3) . Note: this is consistent with the CI above.
```

e) The comparison in part d can be written as a contrast. Construct another contrast which is orthogonal, and confirm that their SS_C (contrast sum-of-squares) add-up to SS_treatment found in part b. By hand.

```
contrast1 = c(1,-2,1)   # "New" contrast vector, such that it is orthogonal to
contrast2 = c(1,0,-1)   # the "original" contrast vector in part d.
gamma1_hat = sum(contrast1 * means)
gamma2_hat = sum(contrast2 * means)
SS_C1 = n*gamma1_hat^2/sum(contrast1^2)    # 529.2
SS_C2 = n*gamma2_hat^2/sum(contrast2^2)    # 14.4
# Sum                                      # 543.6 = SS_treatment
```

f) Which of the two orthogonal contrasts is contributing more to statistical significance found in part a.

The new one is contributing more.

g) The two SS_C terms in part e are each chi-squres with df = 1, but only after dividing by sima_epsilon, which we don't know. However, each of the SS_C terms is still associated with a contrast; and we do have t-tests for testing contrasts. So, perform a t-test on each of the two contrasts testing whether they are non-zero. By hand.

```
t1_obs = gamma1_hat/sqrt(MS_E*sum(contrast1^2)/n) # -5.595856
2*pt(t1_obs,N-a,lower.tail=T)                    # 0.0001169069 ( tail = lower )

# p-value < alpha  --> 2nd differs from the average of 1st and 3rd.

t2_obs = gamma2_hat/sqrt(MS_E*sum(contrast2^2)/n)   # 0.9230769
2*pt(t2_obs,N-a,lower.tail=F)                       # 0.374155

#  p-value > alpha  --> there is no evidence that 1st and 3rd means are different.   Note this is the same as the t-test in d.
```

Moral: The ANOVA F-test above tells us that there is some kind of difference between the three treatment means,
but we don't know what kind. For example, we ask if the 2nd treatment is different from the average of the other two.
This is where orthogonal contrasts are useful. The contribution of each of the (orthogonal) contrasts to
SS_treatment tells us which of the corresponding hypotheses are more *significant*. But you do have to start with
at least one specific contrast, and then find more contrasts that are orthogonal to it.

hw_lect 8-2

Consider the data in 3.16 (7^{th} ed) = 3.18 (8^{th} ed.) $a=4$, $n=4$

The ANOVA Table is

	DF	SS	MS	F	P
x	3	844.7	281.56	14.3	0.0003
Residual	12	236.3	19.69		
Total	15	1080.93			

So, $SST = SSTr + SSE \implies 1080.93 = 844.7 + 236.3$

we can also see that p-value = very small

$\therefore$ Reject $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$ in favor of H_1: At least 2 μ_i are diff.

p-value is "small," because F is "large".

And F is "large", because SSTr is "large".

Let's find out "why" SSTr is large, by checking some contrasts.

There are $a-1 = 3$ orthog. contrast vectors.

Consider $\vec{c} = (1,1,-1,-1)$ which represents $H_0: \mu_1 + \mu_2 = \mu_3 + \mu_4$

You can confirm that one set of orthogonal contrasts is

$\vec{c} = (1,1,-1,-1)$ $\vec{d} = (1,-1,0,0)$ $\vec{e} = (0,0,1,-1)$

a) compute the contrast sum of squares for each contrast vector; and confirm that they sum to SSTr.

b) Choose another set of orthogonal contrast vectors and repeat part a). Make sure that the new set you pick does include one vector that is a permutation of $\vec{c}$.

c) Repeat part b), but now make sure your set does not include any permutation of $\vec{c}$.

```
rm(list=ls(all=TRUE))
a = 4
n = 4
y.m = matrix(nrow=a,ncol=n)      # rows = factor, col = replicates
y.m[1,] = c(143, 141, 150, 146)
y.m[2,] = c(152, 149, 137, 143)
y.m[3,] = c(134, 136, 132, 127)
y.m[4,] = c(129, 127, 132, 129)
y = t(y.m)                       # Better/easier for R.
y.vector = as.vector(y)
x = as.factor(rep(1:a, each = n))
plot(c(x),y.vector)
boxplot(y)
ybars = apply(y.m,1,mean)

X = as.factor(rep(1:a, each = n))
lm.1 = lm(y.vector~X)
summary.aov(lm.1)
SStr = 844.67                    # From anova table.

#            Df Sum Sq Mean Sq F value   Pr(>F)
# A           3  844.7  281.56    14.3 0.000288 ***
# Residuals  12  236.3   19.69

# a)
C1 = c(1,1,-1,-1)
C2 = c(1,-1,0,0)
C3 = c(0,0,1,-1)

# b)
#   C1 = c(1,-1,1,-1)
#   C2 = c(-1,0,1,0)
#   C3 = c(0,1,0,-1)

# c)
#   C1 = c(1,1,-2,0)
#   C2 = c(1,-1,0,0)
#   C3 = c(1,1,1,-3)

c(sum(C1), sum(C2), sum(C3))                 # check zero-sum.
c(sum(C1*C2), sum(C2*C3), sum(C1*C3))        # check orthogonality.

gamma_hat1 = sum(C1*ybars)
gamma_hat2 = sum(C2*ybars)
gamma_hat3 = sum(C3*ybars)

SS1 = (gamma_hat1)^2/(sum(C1^2)/n)
SS2 = (gamma_hat2)^2/(sum(C2^2)/n)
SS3 = (gamma_hat3)^2/(sum(C3^2)/n)
c(SS1, SS2, SS3)                        # for part a) 826.5625   0.1250  18.0000
c(SS1 + SS2 + SS3 , SStr)                  # Confirm equality.

# For part a, the three contrast sums-of squares are 826.5625   0.1250  18.0000. This tells us that the "reason" the anova F-test was
# significant (i.e. at least two of the means are different) is because the average of the 1st two means is different from the average of
# the last two (SS1 = 826.5625). On the other hand, the first two means are not too different (SS2 - 0.125). Similar interpretations are
# allowed for parts b and c.
```

```
rm(list=ls(all=TRUE))
 a = 4
 n = 4
 y.m = matrix(nrow=a,ncol=n)      # rows = factor, col = replicates
 y.m[1,] = c(143, 141, 150, 146)
 y.m[2,] = c(152, 149, 137, 143)
 y.m[3,] = c(134, 136, 132, 127)
 y.m[4,] = c(129, 127, 132, 129)
 y = t(y.m)                       # Better/easier for R.
 y.vector = as.vector(y)
 x = as.factor(rep(1:a, each = n))
 plot(c(x),y.vector)
 boxplot(y)
 ybars = apply(y.m,1,mean)

 X = as.factor(rep(1:a, each = n))
 lm.1 = lm(y.vector~X)
 summary.aov(lm.1)
 SStr = 844.67                    # From anova table.

#            Df Sum Sq Mean Sq F value   Pr(>F)
# A           3  844.7  281.56    14.3 0.000288 ***
# Residuals   12  236.3   19.69

# a)
 C1 = c(1,1,-1,-1)
 C2 = c(1,-1,0,0)
 C3 = c(0,0,1,-1)

# b)
#   C1 = c(1,-1,1,-1)
#   C2 = c(-1,0,1,0)
#   C3 = c(0,1,0,-1)

# c)
#   C1 = c(1,1,-2,0)
#   C2 = c(1,-1,0,0)
#   C3 = c(1,1,1,-3)

 c(sum(C1), sum(C2), sum(C3))                        # check zero-sum.
 c(sum(C1*C2), sum(C2*C3), sum(C1*C3))               # check orthogonality.

 gamma_hat1 = sum(C1*ybars)
 gamma_hat2 = sum(C2*ybars)
 gamma_hat3 = sum(C3*ybars)

 SS1 = (gamma_hat1)^2/(sum(C1^2)/n)
 SS2 = (gamma_hat2)^2/(sum(C2^2)/n)
 SS3 = (gamma_hat3)^2/(sum(C3^2)/n)
 c(SS1, SS2, SS3)
 c(SS1 + SS2 + SS3 , SStr)                           # Confirm equality.
```