

Notes for Stat 430 – Design of Experiments – Lab #5 Oct. 15, 2012.

Agenda:

ANOVA using `aov()`

ANOVA using `lm()` and `anova()`, Using exercise 3.14

ANOVA using `lm()` and `anova()`, Using exercise 3.21

Contrasts – 3 groups vs. 3 groups

Contrasts – 1 group vs. 5 groups

Useful files: http://www.sfu.ca/~jackd/Stat430/Wk05_Code.txt

All lab material found at : <http://www.sfu.ca/~jackd/Stat430/>

ANOVA with `aov()`

First, we need some data. Here is the data from exercise 3.14 in the 8th edition hardcover. It's the golf scores of a player in three seasons. "Shoulder" means both fall and spring. Lower scores are better in golf.

```
season = c(rep("Summer",10), rep("Shoulder",7), rep("Winter",8))
score = c(83,85,85,87,90,88,88,84,91,90,91,87,84,87,85,86,
83,94,91,87,85,87,91,92,86)
obs = 1:length(score)
data314 = data.frame(obs,season,score)
```

"aov" stands for Analysis of Variance. It's the fastest way to get values like the SSE (sum of squares error) and SST (sum of squares total)

```
aov(score ~ season, data=data314)
```

aov should compute. **35.60** for SS_{Season} and **184.63** for SS_{Error}

Which makes a total of 220.24 for SS_{Total}

Variance is mean squared error without considering factors, which is MS_{Total}

We can exploit this to verify the sum of squares total.

$MST = SST / (n-1) \quad \rightarrow \quad SST = MST * (n-1)$

```
n = length(data314$score)
var(data314$score) * (n - 1)
```

This also gives 220.24.

ANOVA using lm() and anova(), Using Exercise 3.14

A more user-friendly command is `anova()` but it requires `lm()` to be used first.

`lm()` is the linear model function we can use it to compare these groups

```
lm(score ~ season, data=data314)
```

That's not very informative, so we can use `summary()` to get more out of it.

```
mod = lm(score ~ season, data=data314)
summary(mod)
```

The (intercept) is the mean response of the first level of the factor alphabetically in this case, that's "Shoulder". If there were multiple factors, (intercept) would use first level of each one.

`seasonSummer` is the additional mean score of games in summer rather than during shoulder

`seasonWinter` is the same, but for winter rather than shoulder

to verify these coefficients

```
mean(score[which(season == "Shoulder")])
mean(score[which(season == "Summer")]) ## 0.9571 more than Shoulder
mean(score[which(season == "Winter")]) ## 2.9821 more than Shoulder
```

Now that we have the linear model, we can put in the `anova()`

```
mod = lm(score ~ season, data=data314)
anova(mod)
```

`anova()` gives you F-statistic and the p-value against the null that all the groups have equal means. Summary also gives you some of this information.

You can get other information from the linear model, for example you can...

```
mod$residuals ## See the residuals
plot(mod$residuals) ## plot them
shapiro.test(mod$residuals) ## Check for normality
qqnorm(mod$residuals)
```

Other things the `lm()` linear model can produce include the fitted values, coefficients, and df. You can also see what the model was without reading the whole summary with `$call`

```
mod$fitted.values
mod$coefficients
mod$df.residual
mod$call
```

ANOVA using `lm()` and `anova()`, Using Exercise 3.21

“Data, data, data. I cannot make bricks without clay.” – Sherlock Holmes

Exercise 3.21 is about the percentage of radon (0-100) that is passing through something as a function of the orifice diameter.

```
orifice = rep( c(0.37,0.51,0.71,1.02,1.40,1.99), each=4)
radon = c(80,83,83,85,75,75,79,79,74,73,76,77,67,72,
74,74,62,62,67,69,60,61,64,66)
obs = 1:length(radon)
data321 = data.frame(obs,radon,orifice)
```

There are six levels to our factor now, different sizes of orifice diameter.

If we take an `aov()`

```
aov(radon ~ orifice, data=data321)
```

Orifice is only taking 1 degree of freedom.. why?

A linear model tells us more

```
mod1 = lm(radon ~ orifice, data=data321) #intercept = 84.22
```

Compared to the mean radon passage with the first orifice level

```
mean( radon[which(data321$orifice == 0.37)]) # 82.75
```

The intercept isn't giving the mean response of the 1st level of orifice it's giving the expected response for an orifice of diameter 0. `lm()` is treating this data as if orifice is a linear explanatory variable. The coefficient for orifice is the slope; the reduction in the response for every unit increase in orifice. If we're looking to compare specific levels of orifice diameter, this might not be what we want.

Orifice can be treated as a factor

```
data321.f = data321 ## I've added .f for factor, just we can tell the difference
data321.f$orifice = as.factor(data321$orifice)
```

Now to try the aov again. It should need 5 df for 6 levels – 1, and 24-6=18 residual.

```
aov(radon ~ orifice, data=data321.f)
```

Now the linear model will show five coefficients. The intercept is the mean response for orifice size 82.750; the mean response for orifice size is 5.750 less than that.

```
mod2 = lm(radon ~ orifice, data=data321.f)
```

We can verify this difference of 5.750.

```
mean( radon[which(data321.f$orifice == 0.37)])
mean( radon[which(data321.f$orifice == 0.51)])
```

ANOVA assumes that the variances within each group are equal and can be pooled.

We can check this with the Bartlett test

```
bartlett.test(data321.f$radon, data321.f$orifice)
```

This should fail to reject the null of equal variances.

Finally, the anova()

```
anova(mod2)
```

Contrasts – 3 groups vs. 3 groups

t-tests determine if two group means are significantly different

ANOVA determines if any of 3 or more group means are significantly different

contrasts are used to determine more complex differences

Say, with the radon example, the only difference we cared about was whether the orifice diameter was greater than 1.00 or not. The first three groups have less than 1, the last three more than 1. So we could be interested in the contrast

$$(\mu_1 + \mu_2 + \mu_3) - (\mu_4 + \mu_5 + \mu_6)$$

Which is easy to find

```
u1 = mean( radon[which(data321.f$orifice == 0.37)])  
u2 = mean( radon[which(data321.f$orifice == 0.51)])  
u3 = mean( radon[which(data321.f$orifice == 0.71)])  
u4 = mean( radon[which(data321.f$orifice == 1.02)])  
u5 = mean( radon[which(data321.f$orifice == 1.40)])  
u6 = mean( radon[which(data321.f$orifice == 1.99)])
```

```
contrast.1 = (u1 + u2 + u3) - (u4 + u5 + u6)  
c1 = c(1,1,1,-1,-1,-1) # +1 for each positive u1,u2... -1 for each negative
```

We have a contrast measure. If the first groups 1,2, and 3 are the same as groups 4,5, and 6 then this contrast should be near zero. How near? For that we need a variance measure.

σ^2 is the mean squared error

From anova(mod2) we know an unbiased estimate of this is the mse, 7.437

```
mse = 7.347  
n = length(data321.f$radon)
```

From 3.26 on p.92

```
var.contr1 = mse/n * sum(c1^2)
```

If the first three means equal the last three, contrast.1 should be distributed about zero with a t-distribution. It's based on a sum of 24 values, which are described with 6 means.

We use t with $(24-6) = 18$ degrees of freedom.

From 3.27 on p.93

```
t = value / sqrt(var)  
t.contr1 = contrast.1 / sqrt(var.contr1)  
t.contr1  
1 - pt(t.contr1,df=18)
```

Contrasts – 1 group vs. 5 groups

Another common case is where everything is being compared to the mean of single group
In medicine this happens when multiple treatments are being assigned, and their effect is being compared to a control.

$$5\mu_1 - (\mu_2 + \mu_3 + \mu_4 + \mu_5 + \mu_6)$$

To compensate, the side with fewer groups is multiplied.

```
contrast.2 = 5*u1 - (u2 + u3 + u4 + u5 + u6)
c2 = c(5,-1,-1,-1,-1,-1)
```

This isn't perfect. With so much riding on one group, the variance of that group matters more.

```
var.contr2 = mse/n * sum(c2^2)
```

The t-score also depends a lot on that one group

```
t.contr2 = contrast.2 / sqrt(var.contr2)
t.contr2
1 - pt(t.contr2,df=18)
```