Lab 4 - Simulation from chisquared, t and F-distribution

Here we will just go over the two problems posed in the lab.

First, using the Grass example:

```
Grass = read.table('Grass.csv',head=TRUE,sep=',')
Grass$Region = as.factor(Grass$Region)
```

In these data, we have coded Cultivar with two indicator functions, but left Region as a factor in R. Our basic model is

```
mod = lm(Speed~.,data=Grass)
summary(mod)
```

```
##
## Call:
## lm(formula = Speed ~ ., data = Grass)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.23817 -0.08706 -0.01128  0.04992  0.29325
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.421762   0.169847  49.584 3.34e-16 ***
## Humidity    -0.022765   0.002453  -9.281 4.25e-07 ***
## Region2     -0.072989   0.155935  -0.468   0.6475
## Region3     -0.084832   0.155846  -0.544   0.5954
## Region4     -0.186642   0.168924  -1.105   0.2892
## Region5      0.434006   0.164553   2.637   0.0205 *
## Region6      0.340397   0.158506   2.148   0.0512 .
## Region7      0.433041   0.164995   2.625   0.0210 *
## Region8      0.252458   0.155974   1.619   0.1295
## C2           0.917971   0.095581   9.604 2.87e-07 ***
## C3           1.885567   0.095644  19.714 4.55e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1904 on 13 degrees of freedom
## Multiple R-squared:  0.975,  Adjusted R-squared:  0.9559
## F-statistic:  50.8 on 10 and 13 DF,  p-value: 9.257e-09
```

Problem construct a test for Region.

Here we will make the observation that changes in sums of squares can also be obtained from a change in sum of squared errors. That is

$$H - H_{-j} = (I - H_{-j}) - (I - H)$$

So in this case we'll take a shortcut and look at a model without region

```
mod2 = lm(Speed~.-Region,data=Grass)
```

and we can now look at the difference between fitted values

```r
SS.R = sum( mod2$residuals^2 ) - sum( mod$residuals^2 )
SS.R
```

```
## [1] 1.304475
```

We still need to get a mean-square, given by the 7 region effects we drop. And we can now plug in MSE from the full model, too.

```r
MS.R = SS.R/7
sighat2 = (summary(mod)$sigma)^2
```

Now we can obtain the F statistic which we test with

```r
F.R = MS.R/sighat2
1-pf(F.R,7,mod$df.residual)
```

```
## [1] 0.005481006
```

Lets see how this compares to

```r
library(car)
```

```
## Warning: package 'car' was built under R version 3.5.1
```

```
## Loading required package: carData
```

```r
Anova(mod)
```

```
## Anova Table (Type II tests)
##
## Response: Speed
##            Sum Sq Df  F value    Pr(>F)
## Humidity   3.1225  1  86.1339 4.247e-07 ***
## Region     1.3045  7   5.1406  0.005481 **
## C2         3.3438  1  92.2396 2.869e-07 ***
## C3        14.0893  1 388.6571 4.554e-11 ***
## Residuals  0.4713 13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Problem perform an experiment to calculate the number of data points necessary to achieve 80% power to find the effect of C2 and C3. To do this a) assume that for a given n, SS.C is given by nK where you estimate K from the data by $SS.C/n$ using our SS.C above. b) remember that the denominator degrees of freedom change with n, but the numerator remains the same at 2.

Plot the power as n increases (what is the minimum n you could have?) and find where it crosses 80%.

Here we need to re-create SS.C (since I'm in a separate file)

```r
mod3 = lm(Speed~.-C2-C3,data=Grass)
SS.C = sum(mod3$residuals^2) - sum(mod$residuals^2)
K = SS.C/nrow(Grass)
```

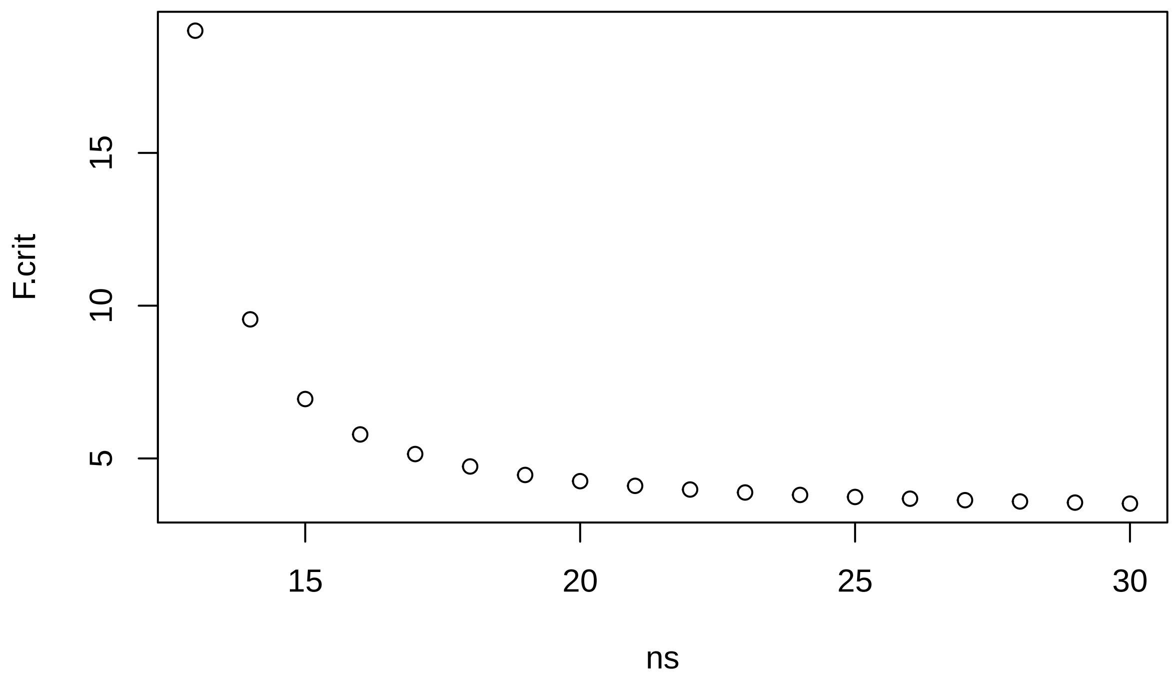
Let's first look at power with just 13 observations. That is, we have a threshold $F_2^{2,1-\alpha}$ (since we use 11 degrees of freedom on the model) and look at the probability that a non-central $F_2^2(K)$ exceeds it

```r
F.crit = qf(0.95,2,2)
1-pf(F.crit,2,2,K)
```

```
## [1] 0.06384613
```

*We're barely over 0.05, but we'd like to get to 0.8. So here let's try up to 30, remembering that the denominator degrees of freedom is always n-11. To do that we'll set up

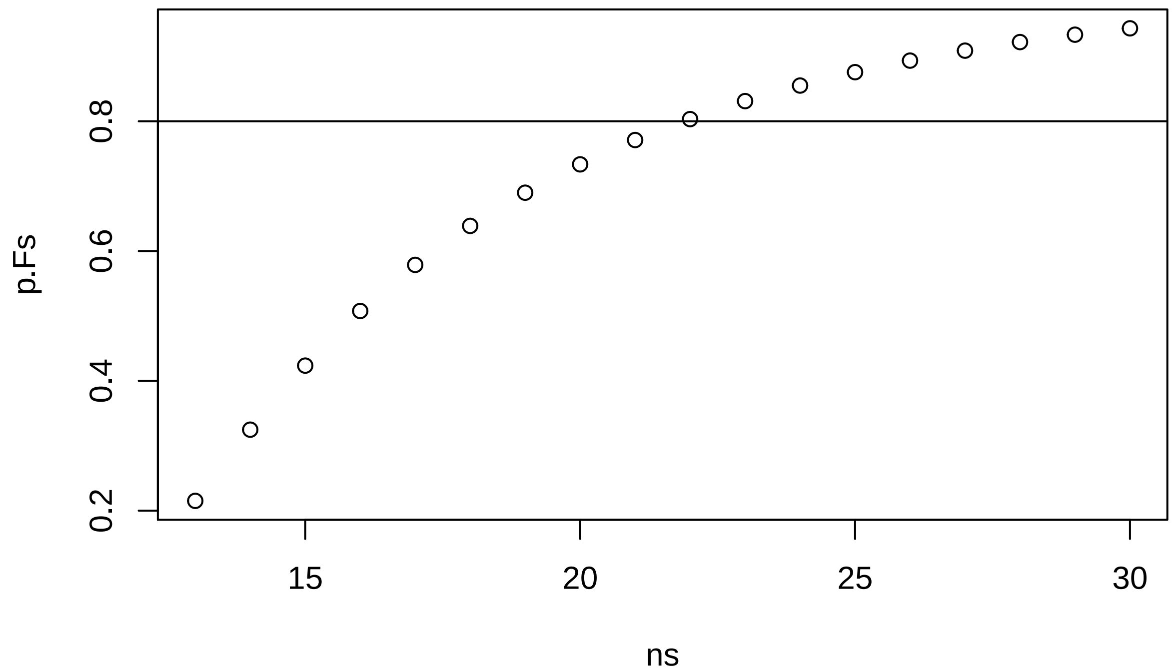
```
ns = 13:30  # potential number of observations
F.crit = qf(0.95,2, ns - 11)  # Sequence of critical values
plot(ns,F.crit)
```



where the critical value reduces towards about 3.

Now our non-centrality parameter also increases: its nK as well as the degrees of freedom for the alternative distribution changing. This gives us

```
p.Fs = 1-pf(F.crit, 2, ns-11,ncp = ns*K)
plot(ns,p.Fs)
abline(h= 0.80)
```




```
p.Fs
```

```
##  [1] 0.2150702 0.3247865 0.4234726 0.5076418 0.5786952 0.6388142 0.6899403
##  [8] 0.7336287 0.7711021 0.8033298 0.8310928 0.8550315 0.8756794 0.8934866
## [15] 0.9088367 0.9220594 0.9334393 0.9432232
```

Where we see that the first time we cross 80% is at 23 observations.