Day02_ExercisesANOVA

February 28, 2018

0.0.1 Example

Read in the file "Rhizobium.csv"

For each column (i), we need to compute 1. Sum: $\sum_j Y_{ij} = Y_{i.}$ 2. Sum of squares: $\sum_j Y_{i,j}^2$ 3. Squared sum divided by r (replicates): $\frac{(Y_{i.})^2}{r}$ 4. Sum of squared deviants: $\sum_j (Y_{ij} - \bar{Y}_{i.})^2$ 5. Mean: $\bar{Y}_{i.}$

In [2]:
```
import pandas as pd
import numpy as np
from scipy.stats import f

aovData =  pd.read_csv('static/Data/Rhizobium.csv')
```

In [3]:
```
# Create a container for the results
aovRes=pd.DataFrame(np.zeros(shape=(5,6)))
aovRes.columns = aovData.columns
aovRes.index = ['Sum','SS','SqSums','SqDev','Mean']
aovRes
```

Out[3]:

	3DOk1	3DOk5	3DOk4	3DOk7	3DOk13	Composite
Sum	0.0	0.0	0.0	0.0	0.0	0.0
SS	0.0	0.0	0.0	0.0	0.0	0.0
SqSums	0.0	0.0	0.0	0.0	0.0	0.0
SqDev	0.0	0.0	0.0	0.0	0.0	0.0
Mean	0.0	0.0	0.0	0.0	0.0	0.0

In [4]:
```
# Compute the row statistics
#########
# We need the row count, no need to recompute

nrow = aovData.shape[0]
ncol = aovData.shape[1]

for column in aovData.columns:
    print(column)
    aovRes.loc[['Sum'],[column]]    = aovData[column].sum()
    aovRes.loc[['SS'],[column]]     = sum(aovData[column]**2)
    aovRes.loc[['SqSums'],[column]] = aovData[column].sum()**2/nrow
```

```
            aovRes.loc[['SqDev'],[column]]  = sum((aovData[column]-aovData[column].mean())**2)
            aovRes.loc[['Mean'],[column]]   = aovData[column].mean()

        aovRes
```

```
3DOk1
3DOk5
3DOk4
3DOk7
3DOk13
Composite
```

```
/usr/local/lib/python3.5/dist-packages/pandas/core/computation/check.py:17: UserWarning: The i
The minimum supported version is 2.4.6

  ver=ver, min_ver=_MIN_NUMEXPR_VERSION), UserWarning)
```

Out[4]:

	3DOk1	3DOk5	3DOk4	3DOk7	3DOk13	Composite
Sum	144.100	119.900	73.200	99.600	66.300	93.50
SS	4287.530	2932.270	1139.420	1989.140	887.290	1758.71
SqSums	4152.962	2875.202	1071.648	1984.032	879.138	1748.45
SqDev	134.568	57.068	67.772	5.108	8.152	10.26
Mean	28.820	23.980	14.640	19.920	13.260	18.70

```
In [5]: # Compute the row totals in aovRes

        # add a column titled 'Total'
        aovRes['Total']=0

        aovRes['Total']=aovRes.sum(axis=1)

        aovRes
```

Out[5]:

	3DOk1	3DOk5	3DOk4	3DOk7	3DOk13	Composite	Total
Sum	144.100	119.900	73.200	99.600	66.300	93.50	596.600
SS	4287.530	2932.270	1139.420	1989.140	887.290	1758.71	12994.360
SqSums	4152.962	2875.202	1071.648	1984.032	879.138	1748.45	12711.432
SqDev	134.568	57.068	67.772	5.108	8.152	10.26	282.928
Mean	28.820	23.980	14.640	19.920	13.260	18.70	119.320

0.0.2 Compute the different SS statistics

First: CF (C)

$$CF = \frac{Y_{..}^2}{rt} = \frac{(\sum_{i,j} Y_{i,j})^2}{rt}$$

```
In [12]: # The term above the line contains the row total of 'Sum'
```

```
CF = (aovRes.loc[['Sum'],['Total']]**2/ (nrow*ncol)).iloc[0]
```

Next SStot

$$SS(total) = \sum_{i,j} Y_{i,j}^2 - CF$$

```
In [7]: SStot = (aovRes.loc[['SS'],['Total']]).iloc[0] - CF
```

Next SStreat

$$SS(treatment) = \frac{\sum_{i=1}^{t} Y_{i,\cdot}^2}{r} - CF$$

$$= \frac{Y_{1\cdot}^2 + Y_{2\cdot}^2 + \cdots + Y_{t\cdot}^2}{r} - CF$$

```
In [8]: sumR = (aovRes.loc[['Sum'],aovRes.columns.difference(['Total'])]).iloc[0]
        SStreat = ( np.sum( sumR**2) / nrow) - CF
```

Last, SSerror

$$SS(error) = \sum_i (\sum_j Y_{ij}^2 - \frac{Y_{i\cdot}^2}{r})$$

This is aovResult['SqDev']['Total']

```
In [9]: SSerr = (aovRes.loc[['SqDev'],['Total']]).iloc[0]
```

Summarize what we have

```
In [10]: print("CF is:      %10.2f" % (CF) )
         print("SStot is:   %10.2f" % (SStot) )
         print("SStreat is: %10.2f" % (SStreat) )
         print("SSerr is:   %10.2f" % (SSerr) )
```

```
CF is:        11864.39
SStot is:      1129.97
SStreat is:     847.05
SSerr is:       282.93
```

Compute the F and interpret

- df among is $N_{treat} - 1 = t - 1$
- df between is $(N_{treat}(N_{rep} - 1) = t(r - 1)$
- df total is $N_{treat} * Nrep - 1 = rt - 1$

Mean square is SS/df for the row
F is the ratio of mean square among œ mean square within - $F = \frac{SS_{between}}{SS_{within}}$

```
In [11]: #
         F =  (SStreat/(ncol-1))/(SSerr/(ncol*(nrow-1)))
         print("The F for the ANOVA is: %8.4f\nand the P-value is:       %6.4f" \
               % (F,f.sf(F,dfd=(ncol-1),dfn=(ncol*(nrow-1)))))

The F for the ANOVA is:  14.3705
and the P-value is:       0.0038
```