Statistical Methods and Data Analysis I
Lecture 19: One Factor ANOVA.

Oleg Goldshmidt
oleg.goldshmidt@post.idc.ac.il

Arison School of Business
Interdisciplinary Center (IDC)
Herzliya, Israel

June 07, 2019

Multiple Samples: A Deeper Analysis

- Consider k samples. There are several types of analysis you can do:
  1. Compute the *sample mean and variance*, for each sample separately.
  2. *Combine the samples*, compute the *grand mean* and the *total variance* around that mean.
  3. Compute the *sample means*, and the *variance between those means*.
- Where does the *total variance* come from?
- Pick a single measurement X from one of the samples. Its deviation from the *grand mean* consists of 2 parts:
  1. how far it is from its sample's mean
  2. how far the sample mean is from the grand mean.
- Do the k samples come from the same population?
  - If yes then the sample means will probably be similar and the variation will be mostly in-sample — the sample variances will be larger than the variance between the sample means.
  - If not the the variation of sample means will be larger than the in-sample variations.
- This is an interesting thought: *analyzing variances gives us information about the means*!

One Factor ANOVA (ANalysis Of VAriance): Parameters

- So we have k samples, $i = 1, \ldots, k$ of n_i observations $X_{i,j}, j = 1, \ldots, n_i$ (with $\sum_{i=1}^{k} n_i = n$).
- Sample means $\overline{X}_i$ and variances σ_i^2 and the *grand mean* $\overline{X}$ are:

$$\overline{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{i,j}, \quad \sigma_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(X_{i,j} - \overline{X}_i)^2, \quad \overline{X} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{i,j} = \frac{1}{n}\sum_{i=1}^{k} n_i \overline{X}_i$$

- Variance *between* samples σ_b^2 and variances *within* samples σ_w^2 (*NB: pooled!*) are

$$\sigma_b^2 = \frac{1}{k-1}\sum_{i=1}^{k} n_i(\overline{X}_i - \overline{X})^2, \quad \sigma_w^2 = \frac{\sum_{i=1}^{k}(n_i - 1)\sigma_i^2}{\sum_{i=1}^{k}(n_i - 1)} = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{i,j} - \overline{X}_i)^2$$

One Factor ANOVA: F-Score

- Test statistic (*"F-score"*): $F = \sigma_b^2/\sigma_w^2$
- *F-score* has *F distribution with* $k-1, n-k$ *degrees of freedom* if samples are drawn from one or more normal distributions.
- H_0: all the sample means are equal
- H_A: the sample means are different
- We *reject* H_0 if $F \gg 1$ (variance *between* is much higher than variance *within*)
- Battle plan:
  1. Formulate H_0 and H_A and determine the desired confidence level
  2. Compute sample means and variances
  3. Compute the *grand mean*
  4. Compute *variance between* and *variance within*
  5. Compute F-score
  6. Compute the probability to obtain F-score as high as observed by chance (p-value)
  7. For small enough p-value we decide to reject H_0

Using ANOVA: Customer Service Scenario — Data

- Our data analyst asks for historical data on customers served on each day of the week.
- Over the last couple of months the data are:

```
> Sun <- c(311, 264, 300, 270, 282, 316, 266, 299, 303)
> Mon <- c(276, 323, 298, 256, 277, 309, 312, 265, 311)
> Tue <- c(243, 279, 301, 285, 274, 243, 228, 298, 255)
> Wed <- c(288, 292, 310, 267, 243, 293, 255, 273)
> Thu <- c(254, 279, 241, 227, 278, 276, 256, 262)
> load <- data.frame(day=c(rep("Sun",length(Sun)),
                           rep("Mon",length(Mon)),
                           rep("Tue",length(Tue)),
                           rep("Wed",length(Wed)),
                           rep("Thu",length(Thu))),
                   customers=c(Sun,Mon,Tue,Wed,Thu))

> load
   day customers
1  Sun       311
2  Sun       264
...
18 Mon       311
19 Tue       243
...
42 Thu       256
43 Thu       262
```

Using ANOVA: Customer Service Scenario — Analysis

```
> library(dplyr)
> samples <- group_by(load,day) %>%
      summarize(count=n(),mean=mean(customers),sd=sd(customers),var=var(customers))
> samples
# A tibble: 5 x 5
  day   count  mean    sd   var
  <fct> <int> <dbl> <dbl> <dbl>
1 Mon       9  292.  23.9  569.
2 Sun       9  290.  19.9  398.
3 Thu       8  259.  18.7  349.
4 Tue       9  267.  26.1  681.
5 Wed       8  278.  22.2  492.
> k <- length(samples)
> n <- sum(samples$count)
> k; n
[1] 5
[1] 43
> grand_mean <- sum(samples$count*samples$mean)/n
> grand_mean
[1] 277.6279
> var_between <- sum(samples$count*(samples$mean-grand_mean)^2)/(k-1)
> var_within <- sum((samples$count-1)*samples$var)/(n-k)
> f_score <- var_between/var_within
> p <- pf(f_score,k-1,n-k,lower.tail=FALSE)
> var_between; var_within; f_score; p
[1] 1731.38
[1] 501.856
[1] 3.449953
[1] 0.01683639
```

Using ANOVA: Customer Service Scenario — Analysis (`aov()`)

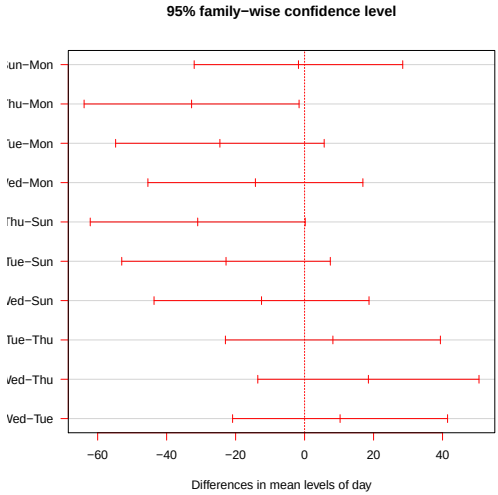
```
> results <- aov(customers ~ day,data=load)
> anova(results) # or summary(results)
Analysis of Variance Table

Response: customers
          Df  Sum Sq Mean Sq F value  Pr(>F)
day        4  6925.5 1731.38    3.45 0.01684 *
Residuals 38 19070.5  501.86
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> TukeyHSD(results,conf.level=0.95)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = customers ~ day, data = load)

$day
              diff       lwr        upr     p adj
Sun-Mon   -1.777778 -32.01305 28.4574948 0.9998112
Thu-Mon  -32.763889 -63.92969 -1.5980833 0.0351906
Tue-Mon  -24.555556 -54.79083  5.6797170 0.1592335
Wed-Mon  -14.263889 -45.42969 16.9019167 0.6865418
Thu-Sun  -30.986111 -62.15192  0.1796945 0.0519700
Tue-Sun  -22.777778 -53.01305  7.4574948 0.2180791
Wed-Sun  -12.486111 -43.65192 18.6796945 0.7806191
Tue-Thu    8.208333 -22.95747 39.3741389 0.9419779
Wed-Thu   18.500000 -13.56935 50.5693494 0.4750281
Wed-Tue   10.291667 -20.87414 41.4574723 0.8771901
```



Conclusion: there are more customers on Sundays and Mondays

Probably need more staff on those days