One-way Analysis of Variance (One-way ANOVA, 單因子變異數分析)

- ANOVA is used to determine whether the means for different populations are the same. When the different populations represent responses corresponding to different levels of a treatment, ANOVA can be used to determine whether the different treatment levels have the same effect on the mean of the response.
  - Example.
    - Treatment levels: receiving different medicines.
    - Response: blood pressure.
- Notation and assumptions for data.
  - k: total number of treatment levels.
  - n_i: number of observations for the response for treatment level i.
  - $n = \sum_{i=1}^{k} n_i$.
  - $X_{i,j}$: the j-th observation for the response under treatment level i.
  - It is assumed that all $X_{i,j}$'s are independent and for each i,

$$X_{i,1}, \ldots, X_{i,n_i} \text{ are IID } N(\mu_i, \sigma^2).$$

- The problem of interest is to test

$$H_0\text{: } \mu_1 = \cdots = \mu_k \text{ versus } H_1\text{: not all the } \mu_i\text{'s are the same.} \qquad (1)$$

- Some relevant statistics for testing (1).
  - The grand mean (overall sample mean) is $\bar{X}_G = \dfrac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{i,j}$.
  - For the treatment level i group, the sample mean and sample standard deviation are $\bar{X}_i = \dfrac{1}{n_i} \sum_{j=1}^{n_i} X_{i,j}$ and $S_i = \sqrt{\dfrac{\sum_{j=1}^{n_i} (X_{i,j} - \bar{X}_i)^2}{n_i - 1}}$ respectively.
  - The sum of squares due to treatment is $\mathrm{SS}_{\mathrm{treat}} = \sum_{i=1}^{k} n_i (\bar{X}_i - \bar{X}_G)^2$.
  - The sum of squares due to error is $\mathrm{SSE} = \sum_{i=1}^{k} (n_i - 1) S_i^2$.
- Note. In this class, we use $\mathrm{SS}_{\mathrm{treat}}$ or SST to denote the sum of squares due to treatment. In some other places, SST may denote the toal sum of squares. The toal sum of squares, denoted by $\mathrm{SS}_{\mathrm{total}}$, is

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{i,j} - \bar{X}_G)^2.$$

  It can be shown that

$$\mathrm{SS}_{\mathrm{total}} = \mathrm{SS}_{\mathrm{treat}} + \mathrm{SSE}.$$

- The ANOVA F test uses the following statistic for testing (1):

$$F = \frac{\text{SS}_{\text{treat}}/(k-1)}{\text{SSE}/(n-k)} = \frac{\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X}_G)^2/(k-1)}{\sum_{i=1}^{k}(n_i-1)S_i^2/(n-k)}.$$

  - $F \sim F(k-1, n-k)$ under H_0 since under H_0, $\text{SS}_{\text{treat}}/\sigma^2 \sim \chi^2(k-1)$ and $\text{SSE}/\sigma^2 \sim \chi^2(n-k)$ and SS_{treat} and SSE are independent.
  - The ANOVA F test rejects H_0 at level a if $F > f_{a,k-1,n-k}$.
  - p-value for the ANOVA F test: $P(F(k-1, n-k) >$ observed value for $F)$.
  - R command for computing the p-value: `1-pf(obs.F, k-1, n-k)`, where `obs.F` denotes the observed value for F.

Example 1. Suppose that we want to compare the effects of three weight loss drugs A, B and C. 12 participants are assigned to receive one of the three medicines. Their weight losses (in kilograms) are given below.

A	B	C
5.5	6.6	5.1
5.4	7.5	4.6
5.6	6.7	5.5
	7.1	4.8
	5.8	

Are the three drugs equally effective? Use the 0.01 significance level.

- Solution to Example 1 using R.
  - R codes

```
drug <- factor(c( rep("A",3), rep("B",5), rep("C",4)))
weight.loss <- c(5.5,5.4,5.6,  6.6,7.5,6.7,7.1,5.8,  5.1,4.6,5.5,4.8)
anova(lm(weight.loss~drug))
```

  - R output

```
Analysis of Variance Table

Response: weight.loss
          Df Sum Sq Mean Sq F value   Pr(>F)
drug       2 7.2180  3.6090  15.526 0.001209 **
Residuals  9 2.0920  0.2324
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

  - $0.001209 < 0.01$, so we conclude that the effects of the three drugs are not all the same at the 0.01 significance level.
- Solution to Example 1 by direct computation. The sample means and sample standard deviations for each drug group can be computed directly. The results are:

	A	B	C
sample mean	5.5	6.74	5
sample standard deviation	0.1	$\sqrt{0.403}$	$\sqrt{0.46/3}$

The grand mean can be computed from the sample means:

$$\bar{X}_G = \frac{1}{n}\sum_{i=1}^{k} n_i \bar{X}_i = \frac{3\times 5.5 + 5\times 6.74 + 4\times 5}{12} = 5.85.$$

From the above calculation,

$$\begin{aligned} \mathrm{SS}_{\mathrm{treat}} &= \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X}_G)^2 \\ &= 3\times(5.5-5.85)^2 + 5\times(6.74-5.85)^2 + 4\times(5-5.85)^2 = 7.218 \end{aligned}$$

and

$$\begin{aligned} \mathrm{SSE} &= \sum_{i=1}^{k}(n_i-1)S_i^2 \\ &= 2\times(0.1)^2 + 4\times(\sqrt{0.403})^2 + 3\times(\sqrt{0.46/3})^2 = 2.092, \end{aligned}$$

so the F statistic is

$$\frac{\mathrm{SS}_{\mathrm{treat}}/(k-1)}{\mathrm{SSE}/(n-k)} = \frac{7.218/(3-1)}{2.092/(12-3)} = 15.52629.$$

From the table "0.99 quantiles for F distributions", $f_{0.01,2,9} = 8.02 < 15.52629$, so we conclude that the effects of the three drugs are not all the same at the 0.01 significance level.

Example 2. 假設農夫收集了使用數種不同肥料組合下, 某種作物產量的資料. 假設使用 R 分析資料, 而作物產量和肥料組合資料已分別儲存於 R 中兩變數 `yield` 和 `fertilizer`, 其中`fertilizer`爲 factor 型態的變數. 假設執行 R 指令 `anova(lm(yield~fertilizer))`的結果爲

```
Analysis of Variance Table

Response: yield
           Df Sum Sq Mean Sq F value    Pr(>F)
fertilizer  2 107.50  53.749  15.213 1.627e-05 ***
Residuals  36 127.19   3.533
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

(a) 在0.001的顯著水準下, 是否可推論不同肥料組合對產量的影響不全相同?

(b) 資料中的肥料組合方式有幾種?

(c) 共有幾筆資料?

解: (a) 是, p-value $= 1.627 \times 10^{-5} < 0.001$. (b) 資料中的肥料組合方式有3 種 ($k - 1 = 2 \Rightarrow k = 3$). (c) 資料總共有39筆 ($n - k = n - 3 = 36 \Rightarrow n = 39$).

- $k = 2$ 時, one-way ANOVA F test 和 pooled t test 的檢定結果是一致的.
- An experiment for checking the distribution of $\mathrm{SS}_{\mathrm{treat}}/\sigma^2$ under H_0. Experiment setup:
  - $k = 3$, $n_1 = 100$, $n_2 = 200$, $n_3 = 300$.
  - $X_{i,j} \sim N(0, 1)$ for $j = 1, \ldots, n_i$, $i = 1, 2, 3$.

  Experiment steps:

  (i) In each trial, generate $X_{i,j}$s and compute $\mathrm{SS}_{\mathrm{treat}}/\sigma^2 = \mathrm{SS}_{\mathrm{treat}}$. Run 10^4 trials to obtain 10^4 $\mathrm{SS}_{\mathrm{treat}}$ values.

  (ii) Plot the histogram (normalized) of the 10^4 $\mathrm{SS}_{\mathrm{treat}}$ values.

  (iii) Add the graph of the PDF of $\chi^2(2)$ to the histogram plot.

  The result agrees with the fact that the distribution of $\mathrm{SS}_{\mathrm{treat}}/\sigma^2$ under $H_0 : \mu_1 = \mu_2 = \mu_3$ is $\chi^2(2)$. R commands

```
n1 <- 100
n2 <- 200
n3 <- 300
m <- 10^4

sst <- rep(0, m)
for (i in 1:m){
  x1 <- rnorm(n1, mean=0, sd=1)
  x2 <- rnorm(n2, mean=0, sd=1)
  x3 <- rnorm(n3, mean=0, sd=1)
  x <- c(x1, x2, x3)
  xbar <- mean(x)
  sst[i] <- n1*(mean(x1)-xbar)^2 + n2*(mean(x2)-xbar)^2 + n3*(mean(x3)-xbar)^2
}
hist(sst, nclass="scott", freq=FALSE)

f.fun <- function(x){
   dchisq(x, 2)   #The PDF of chi-squared distribution of 2 degrees of freedom
  }

curve(f.fun, add=TRUE)
```

- Fact 1 Suppose that $Z_1, \ldots, Z_k$ are IID $N(0,1)$ random variables, $U_1, \ldots, U_m$ are linear combinations of $Z_1, \ldots, Z_k$ such that $Cov(U_i, U_j) = 0$ for $i \neq j$ and $Var(U_i) = 1$ for all i, where $m < k$. Then,

$$Z_1^2 + \cdots + Z_k^2 - (U_1^2 + \cdots + U_m^2) \sim \chi^2(k-m).$$

  Moreover, $Z_1^2 + \cdots + Z_k^2 - (U_1^2 + \cdots + U_m^2)$ and $(U_1, \ldots, U_m)$ are independent.

  - $Cov(X,Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y)$.