Statistics and learning

Analysis of variance (ANOVA)

Emmanuel Rachelson and Matthieu Vignes

ISAE SupAero

Friday 23rd October 2013

ANOVA: presentation

- Allows to evaluate and compare the effect of one or several controlled factors on a population from the point of view of a given variable.
- Under the hypothesis of Gaussian distribution, ANOVA is just a global test to compare the means of subpopulations associated to the levels of the considered factors.

1 way-ANOVA

- a factor can take k different values. To each level is associated $X_i \sim \mathcal{N}(\mu_i, \sigma^2)$.
- μ_i's are unknown, σ is known.
- $\forall 1 \leq i \leq k$, a sample of size n_i is taken from subpopulation i (we write $n = \sum n_i$):

$$(X_i^1 = x_i^1, \ldots, X_i^{n_i} = x_i^{n_i})$$

1 way-ANOVA

- a factor can take k different values. To each level is associated $X_i \sim \mathcal{N}(\mu_i, \sigma^2)$.
- μ_i's are unknown, σ is known.
- $\forall 1 \leq i \leq k$, a sample of size n_i is taken from subpopulation i (we write $n = \sum n_i$):

$$(X_i^1 = x_i^1, \ldots, X_i^{n_i} = x_i^{n_i})$$

- Finally the ANOVA is a test:

ANOVA = test of equality for all means

(H0) $m_1 = m_2 = \ldots = m_k$ and
(H1) $\exists p, q$ such that $m_p \neq m_q$

1 way-ANOVA explained

- Variable X_i^j associated to the j^{th} draw can be decomposed into

$$X_i^j = \mu + \alpha_i + \epsilon_i^j,$$

- where μ is the mean of all X, α_i is the mean effect due to level i of the considered factor and ϵ is the residual, with $\mathcal{N}(0, \sigma^2)$ distribution.
- Note that $\mu + \alpha_i$ is the mean of X on population i which corresponds to level i of the factor.
- Some notations: $\bar{X} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} X_i^j}{n}$, $\bar{X}_i = \frac{\sum_j X_i^j}{n_i}$ and more specifically:
- $S_A^2 = \frac{1}{n} \sum_i n_i (\bar{X}_i - \bar{X})^2$ (variance between),
  $S_R^2 = \frac{1}{n} \sum_i \sum_j (X_i^j - \bar{X}_i)^2$ (residual variance) and
  $S = \frac{1}{n} \sum_i \sum_j (X_i^j - \bar{X})^2$ (total variance)

1 way-ANOVA: theory

Theorem (1 way-ANOVA formula)

$$S^2 = S_A^2 + S_R^2$$

Theorem (Useful "cooking recipe" for the test)

1. $nS_R^2/\sigma^2 \sim \chi^2(n-k)$.
2. *Under (H0),* $nS^2/\sigma^2 \sim \chi^2(n-1)$ *and* $nS_A^2/\sigma^2 \sim \chi^2(k-1)$.

So that under (H0), $\frac{S_A^2/(k-1)}{S_R^2/(n-k)} \sim F(k-1; n-k)$, *a Fisher Snedecor distribution with* $(k-1; n-k)$ *dof.*

Morality: we just test whether S_A^2 is small compared to S_R^2: is the between dispersion small as compared to the inner dispersion ?

2 way-ANOVA

- We just want to generalise that to 2 factors A and B with resp. p and q levels.
- to the (i, j) couple of levels for both factors correspond a sample of size $n_{i,j}$ for measured variable X.
- The statistical model is balanced if $n_{i,j} = r$, $\forall (i, j)$. We restrict the presentation in this framework to keep notations more simple.
- So to any couple of levels (i, j) is associated sample $(X^1_{i,j} = x^1_{i,j}, \dots, X^r_{i,j} = x^r_{i,j})$.
- $X_{i,j}$ is assumed to be $\mathcal{N}(\mu_{i,j}, \sigma^2)$ and we can decompose...

2-way ANOVA decomposition

- $$\mu_{i,j} = \mu + \alpha_i + \beta_j + \gamma_{i,j},$$
- with resp. effects for A, B and the $A \times B$ interaction.
- We adapt previous notations: $\bar{X} = \frac{\sum_{i=1}^{p} \sum_{j=1}^{q} \sum_{k=1}^{r} X_{i,j}^k}{pqr}$, $\bar{X_{i,j}} = \frac{\sum_k X_{i,j}^k}{r}$, $\bar{X_{i,\bullet}} = \frac{\sum_j \sum_k X_{i,j}^k}{qr}$ and $\bar{X_{\bullet,j}} = \frac{\sum_i \sum_k X_{i,j}^k}{pr}$ and for variances:
- $S_A^2 = qr \sum_i (\bar{x_{i,\bullet}} - \bar{x})^2$, $S_B^2 = pr \sum_j (\bar{x_{\bullet,j}} - \bar{x})^2$,
  $S_{AB}^2 = r \sum_u \sum_j (\bar{x_{i,j}} - \bar{x_{i,\bullet}} - \bar{x_{\bullet,j}} + \bar{x})^2$,
  $S_R^2 = \sum_i \sum_j \sum_k (x_{i,j}^k - \bar{x_{i,j}})^2$ and $S^2 = \sum_i \sum_j \sum_k (x_{i,j}^k - \bar{x})^2$.
  Whooosh !

2 way-ANOVA: theory

Theorem (Formula for 2 way ANOVA)

$$S^2 = S_A^2 + S_B^2 + S_{AB}^2 + S_R^2$$

Proof is tedious and does not have that much interest.

Instead of listing all distributions, we summarise all of that in the table on the next slide...

2 way-ANOVA analysis table

Variat. origin	$\sum$(squares)	d.o.f.	Mean squares	F-variable
A	S_A^2	$p-1$	$S_A^2/(p-1) = S_{Am}^2$	S_{Am}^2/S_{Rm}^2
B	S_B^2	$q-1$	$S_B^2/(q-1) = S_{Bm}^2$	S_{Bm}^2/S_{Rm}^2
$A \times B$	S_{AB}^2	$(p-1)(q-1)$	$\frac{S_{AB}^2}{(p-1)(q-1)} = S_{ABm}^2$	S_{ABm}^2/S_{Rm}^2
Residual	S_R^2	$pq(r-1)$	$S_R^2/(p-1) = S_{Rm}^2$	
Total	S^2	$pqr-1$		

That's all

For today: next time $\rightarrow$ regression !!