

AOV, continued

The AOV table

Source	df	Sum of squares	Mean square	F test	
Model	$t - 1$	SSB	$SSB/(t - 1)$	f	P -val
Error	$n_T - t$	SSW	$SSW/(n_T - t)$		
Corrected total	n_T	SST			

The AOV model

$$Y_{ij} \sim \text{normal}(\mu_i, \sigma^2)$$

where $i = 1, 2, \dots, t$ and $j = 1, 2, \dots, n_i$.

Different “parameterization”: $\mu_i = \mu + \alpha_i$, where $\alpha_1 + \alpha_2 + \dots + \alpha_t = 0$ (α_i is the “effect” of the i th treatment level). There are still t parameters: $\mu, \alpha_1, \alpha_2, \dots, \alpha_{t-1}$ (and $\alpha_t = -\alpha_1 - \alpha_2 - \dots - \alpha_{t-1}$). Note that $\mu = (\mu_1 + \mu_2 + \dots + \mu_t)/t$ and $\alpha_i = \mu_i - \mu$.

Another way of writing the model:

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

where $\epsilon_{ij} \sim \text{normal}(0, \sigma^2)$.

Model evaluation

Statistics is the study of how to draw conclusions from counts and measurements. Statistical analysis proceeds by building probabilistic (stochastic) models of the processes that produce the variability in the counts and measurements.

You should think of statistics as focused on *models* rather than *methods*.

“Fitting a model to data” \equiv estimating the unknown model parameters with data.

Model parameters in 1-way AOV: $\mu_1, \mu_2, \dots, \mu_t, \sigma^2$ (or $\mu, \alpha_1, \alpha_2, \dots, \alpha_{t-1}, \sigma^2$). Estimates usually denoted with “hats”, for instance, $\hat{\mu}_i = \bar{y}_{i.}$, and so on.

Evaluating the fitted AOV model centers around the key ingredients of the model:

1. The variance σ^2 is constant among the t populations (**homoskedastic**)
2. The Y_{ij} s within each population i have a normal distribution
3. The Y_{ij} s are independent random variables

Evaluation uses the following quantities.

Predicted value for y_{ij} under the model:

$$\hat{y}_{ij} = \hat{\mu}_i = \bar{y}_i.$$

Residual for y_{ij} under the model:

$$e_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \hat{\mu}_i$$

Recall the model: $Y_{ij} - \mu_i = \epsilon_{ij}$. Residual can be thought of as an estimate of the “noise” amount ϵ_{ij} for that observation. If the AOV model is adequate, the residuals should be similar to a random sample from a normal distribution with constant variance. (Note: the residuals are in fact dependent, but the amount of dependence is small in adequate-sized samples.)

(For example, if there is just one normal population: $Y_j \sim \text{normal}(\mu, \sigma^2)$, with random sample Y_1, Y_2, \dots, Y_n , where $n \geq 2$, estimate $\hat{\mu} = \bar{Y}$, residual $E_j = Y_j - \bar{Y}$; one can show that $\text{Corr}(E_k, E_l) = -1/(n-1)$)

1. Evaluating constant variance assumption (important)

Side-by-side box plots of residuals from each group

Scatter plot of residuals (vertical) vs predicted values

Hartley test (very sensitive to normality departures) or
Levine test (text, section 7.4 pp. 365-371)

2. Evaluating normality (not quite as important)

Normal probability plot of residuals

Test of normality for residuals (Kolmogorov-Smirnov
etc.)

Note: can obtain both of these from PROC
UNIVARIATE (with the PLOT and
NORMAL options)

3. Independence: assured by design of sampling
or experiment

Fixes for non-normal data

1. Nonparametric (“model free”) statistical methods (Stat 514)

These approaches use weaker assumptions, such as assuming that the distributions are symmetric

one sample: Wilcoxon test for a median

two sample: Mann-Whitney-Wilcoxon test for comparing two distributions

1-way AOV: Kruskal-Wallis test

These are all examples of methods based on **linear rank statistics**. In fact, one can perform the equivalent of all these tests by calculating the **ranks** of the data (ties get averaged ranks) and performing the usual normal-based methods on the ranks!

2. Transformations

Data can often be transformed so that the transformed observations are approximately normal, and/or have their variances stabilized.

Multiplicative processes (finances/income, biological growth) are often normalized by the logarithmic transformation: $Y_{ij} = \log(X_{ij})$

Poisson observations can be normalized and variance-stabilized with the square root transformation:

$$Y_{ij} = \sqrt{X_{ij} + (3/8)}$$

Binomial observations can be normalized and variance-stabilized with the arcsin (inverse sin) transformation:

$$Y_{ij} = \arcsin(X_{ij})$$

The Box-Cox transformation is a general family of transformations (one for each value of λ):

$$Y_{ij} = \begin{cases} \frac{X_{ij}^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \log(X_{ij}), & \lambda = 0 \end{cases}$$

3. Non-normal models