Two-way analysis of variance

From Wikipedia, the free encyclopedia

‹ The template below (*Expert needed*) is being considered for deletion. See templates for discussion to help reach a consensus. ›

This article **needs attention from an expert in statistics**. Please add a *reason* or a *talk* parameter to this template to explain the issue with the article. WikiProject Statistics may be able to help recruit an expert. *(January 2012)*

In statistics, the **two-way analysis of variance** (**ANOVA**) is an extension of the one-way ANOVA that examines the influence of two different categorical independent variables on one continuous dependent variable. The two-way ANOVA not only aims at assessing the main effect of each independent variable but also if there is any interaction between them.

Contents [hide]

History [edit]

In 1925, Ronald Fisher mentions the two-way ANOVA in his celebrated book, *Statistical Methods for Research Workers* (chapters 7 and 8). In 1934, Frank Yates published procedures for the unbalanced case.[1] Since then, an extensive literature has been produced. The topic was reviewed in 1993 by Yasunori Fujikoshi.[2] In 2005, Andrew Gelman proposed a different approach of ANOVA, viewed as a multilevel model.[3]

Data set [edit]

Let us imagine a data set for which a dependent variable may be influenced by two **factors** which are potential sources of variation. The first factor has I levels ($i \in \{1, \dots, I\}$) and the second has J levels ($j \in \{1, \dots, J\}$). Each combination (i, j) defines a **treatment**, for a total of $I \times J$ treatments. We represent the number of **replicates** for treatment (i, j) by n_{ij}, and let k be the index of the replicate in this treatment ($k \in \{1, \dots, n_{ij}\}$).

From these data, we can build a contingency table, where $n_{i+} = \sum_{j=1}^{J} n_{ij}$ and $n_{+j} = \sum_{i=1}^{I} n_{ij}$, and the total number of replicates is equal to $n = \sum_{i,j} n_{ij} = \sum_{i} n_{i+} = \sum_{j} n_{+j}$.

The experimental design is **balanced** if each treatment has the same number of replicates, K. In such a case, the design is also said to be **orthogonal**, allowing to fully distinguish the effects of both factors. We hence can write $\forall i, j \; n_{ij} = K$, and $\forall i, j \; n_{ij} = \frac{n_{i+} \cdot n_{+j}}{n}$.

Model [edit]

Upon observing variation among all n data points, for instance via a histogram, "probability may be used to describe such variation".[4] Let us hence denote by Y_{ijk} the random variable which observed value y_{ijk} is the k-th measure for treatment (i, j). The **two-way ANOVA** models all these variables as varying independently and normally around a mean, μ_{ij}, with a constant variance, σ^2 (homoscedasticity):

$$Y_{ijk} \mid \mu_{ij}, \sigma^2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_{ij}, \sigma^2).$$

Specifically, the mean of the response variable is modeled as a linear combination of the explanatory variables:

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij},$$

where μ is the grand mean, α_i is the additive main effect of level i from the first factor (*i*-th row in the contigency table), β_j is the additive main effect of level j from the second factor (*j*-th column in the contingency table) and γ_{ij} is the non-additive interaction effect of treatment (i, j) from both factors (cell at row *i* and column *j* in the contingency table).

Another equivalent way of describing the two-way ANOVA is by mentioning that, besides the variation explained by the factors, there remains some statistical noise. This amount of unexplained variation is handled via the introduction of one random variable per data point, ϵ_{ijk}, called error. These n random variables are seen as deviations from the means, and are assumed to be independent and normally distributed:

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk} \text{ with } \epsilon_{ijk} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2).$$

Assumptions [edit]

Following Gelman and Hill, the assumptions of the ANOVA, and more generally the general linear model, are, in decreasing order of importance:[5]

1. the data points are relevant with respect to the scientific question under investigation;
2. the mean of the response variable is influenced additively (if not interaction term) and linearly by the factors;
3. the errors are independent;
4. the errors have the same variance;
5. the errors are normally distributed.

Parameter estimation [edit]

To ensure identifiability of parameters, we can add the following "sum-to-zero" constraints:

$$\sum_i \alpha_i = \sum_j \beta_j = \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0$$

Hypothesis testing [edit]

In the classical approach, testing null hypotheses (that the factors have no effect) is achieved via their significance which requires calculating sums of squares.

Testing if the interaction term is significant can be difficult because of the potentially-large number of degrees of freedom.[6]

See also [edit]

- Analysis of variance
- F test (*Includes a one-way ANOVA example*)
- Mixed model
- Multivariate analysis of variance (MANOVA)
- One-way ANOVA
- Repeated measures ANOVA
- Tukey's test of additivity

Notes [edit]

1. ^ Yates, Frank (March 1934). "The analysis of multiple classifications with unequal numbers in the different classes". *Journal of the American Statistical Association*. **29** (185): 51–66. doi:10.1080/01621459.1934.10502686. JSTOR 2278459.
2. ^ Fujikoshi, Yasunori (1993). "Two-way ANOVA models with unbalanced data". *Discrete Mathematics*. **116** (1): 315–334. doi:10.1016/0012-365X(93)90410-U.
3. ^ Gelman, Andrew (February 2005). "Analysis of variance? why it is more important than ever". *The Annals of Statistics*. **33** (1): 1–53. arXiv:math/0508526. doi:10.1214/009053604000001048.
4. ^ Kass, Robert E (1 February 2011). "Statistical inference: The big picture". *Statistical Science*. **26** (1): 1–9. arXiv:1106.2895. doi:10.1214/10-sts337. PMC 3153074. PMID 21841892.
5. ^ Gelman, Andrew; Hill, Jennifer (18 December 2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press. pp. 45–46. ISBN 978-0521867061.
6. ^ Yi-An Ko; et al. (September 2013). "Novel Likelihood Ratio Tests for Screening Gene-Gene and Gene-Environment Interactions with Unbalanced Repeated-Measures Data". *Genetic Epidemiology*. **37** (6): 581–591. doi:10.1002/gepi.21744. PMC 4009698. PMID 23798480.

References [edit]

- George Casella (18 April 2008). *Statistical design*. Springer Texts in Statistics. Springer. ISBN 978-0-387-75965-4.

Categories: Analysis of variance