WIKIPEDIA

Analysis of variance

Analysis of variance (**ANOVA**) is a collection of statistical models and their associated estimation procedures (such as the "variation" among and between groups) used to analyze the differences among group means in a sample. ANOVA was developed by statistician and evolutionary biologist Ronald Fisher. The ANOVA is based on the law of total variance, where the observed variance in a particular variable is partitioned into components attributable to different sources of variation. In its simplest form, ANOVA provides a statistical test of whether two or more population means are equal, and therefore generalizes the *t*-test beyond two means.

Contents

History

Motivating example

Background and terminology
- Design-of-experiments terms

Classes of models
- Fixed-effects models
- Random-effects models
- Mixed-effects models

Assumptions
- Textbook analysis using a normal distribution
- Randomization-based analysis
- Summary of assumptions

Characteristics

Logic
- Partitioning of the sum of squares
- The *F*-test
- Extended logic

For a single factor

For multiple factors

Worked numeric examples

Associated analysis
- Preparatory analysis
- Follow-up analysis

Study designs

Cautions

Generalizations
- Connection to linear regression

See also

Footnotes

Notes

References

Further reading

External links

History

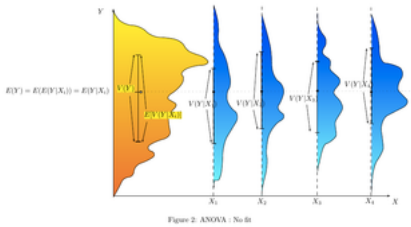
While the analysis of variance reached fruition in the 20th century, antecedents extend centuries into the past according to Stigler.[1] These include hypothesis testing, the partitioning of sums of squares, experimental techniques and the additive model. Laplace was performing hypothesis testing in the 1770s.[2] The development of least-squares methods by Laplace and Gauss circa 1800 provided an improved method of combining observations (over the existing practices then used in astronomy and geodesy). It also initiated much study of the contributions to sums of squares. Laplace knew how to estimate a variance from a residual (rather than a total) sum of squares.[3] By 1827, Laplace was using least squares methods to address ANOVA problems regarding measurements of atmospheric tides.[4] Before 1800, astronomers had isolated observational errors resulting from reaction times (the "personal equation") and had developed methods of reducing the errors.[5] The experimental methods used in the study of the personal equation were later accepted by the emerging field of psychology [6] which developed strong (full factorial) experimental methods to which randomization and blinding were soon added.[7] An eloquent non-mathematical explanation of the additive effects model was available in 1885.[8]

Ronald Fisher introduced the term variance and proposed its formal analysis in a 1918 article *The Correlation Between Relatives on the Supposition of Mendelian Inheritance*.[9] His first application of the analysis of variance was published in 1921.[10] Analysis of variance became widely known after being included in Fisher's 1925 book *Statistical Methods for Research Workers*.

Randomization models were developed by several researchers. The first was published in Polish by Jerzy Neyman in 1923.[11]

One of the attributes of ANOVA that ensured its early popularity was computational elegance. The structure of the additive model allows solution for the additive coefficients by simple algebra rather than by matrix calculations. In the era of mechanical calculators this simplicity was critical. The determination of statistical significance also required access to tables of the F function which were supplied by early statistics texts.

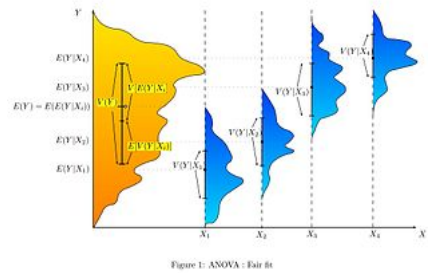
Motivating example



No fit.

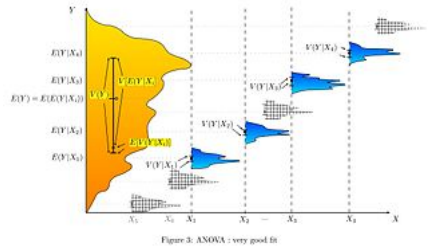
The analysis of variance can be used as an exploratory tool to explain observations. A dog show provides an example. A dog show is not a random sampling of the breed: it is typically limited to dogs that are adult, pure-bred, and exemplary. A histogram of dog weights from a show might plausibly be rather complex, like the yellow-orange distribution shown in the illustrations. Suppose we wanted to predict the weight of a dog based on a certain set of characteristics of each dog. One way to do that is to *explain* the distribution of weights by dividing the dog population into groups based on those characteristics. A successful grouping will split dogs such that (a) each group has a low variance of dog weights (meaning the group is relatively homogeneous) and (b) the mean of each group is distinct (if two groups have the same mean, then it isn't reasonable to conclude that the groups are, in fact, separate in any meaningful way).

In the illustrations to the right, groups are identified as X_1, X_2, etc. In the first illustration, the dogs are divided according to the product (interaction) of two binary groupings: young vs old, and short-haired vs long-haired (e.g., group 1 is young, short-haired dogs, group 2 is young, long-haired dogs, etc.). Since the distributions of dog weight within each of the groups (shown in blue) has a relatively large variance, and since the means are very similar across groups, grouping dogs by these characteristics does not produce an effective way to explain the variation in dog weights: knowing which group a dog is in doesn't allow us to predict its weight much better than simply knowing the dog is in a dog show. Thus, this grouping fails to explain the variation in the overall distribution (yellow-orange).



Fair fit

An attempt to explain the weight distribution by grouping dogs as *pet vs working breed* and *less athletic vs more athletic* would probably be somewhat more successful (fair fit). The heaviest show dogs are likely to be big, strong, working breeds, while breeds kept as pets tend to be smaller and thus lighter. As shown by the second illustration, the distributions have variances that are considerably smaller than in the first case, and the means are more distinguishable. However, the significant overlap of distributions, for example, means that we cannot distinguish X_1 and X_2 reliably. Grouping dogs according to a coin flip might produce distributions that look similar.



Very good fit

An attempt to explain weight by breed is likely to produce a very good fit. All Chihuahuas are light and all St Bernards are heavy. The difference in weights between Setters and Pointers does not justify separate breeds. The analysis of variance provides the formal tools to justify these intuitive judgments. A common use of the method is the analysis of experimental data or the development of models. The method has some advantages over correlation: not all of the data must be numeric and one result of the method is a judgment in the confidence in an explanatory relationship.

Background and terminology

ANOVA is a form of statistical hypothesis testing heavily used in the analysis of experimental data. A test result (calculated from the null hypothesis and the sample) is called statistically significant if it is deemed unlikely to have occurred by chance, *assuming the truth of the null hypothesis*. A statistically significant result, when a probability (*p*-value) is less than a pre-specified threshold (significance level), justifies the rejection of the null hypothesis, but only if the a priori probability of the null hypothesis is not high.

In the typical application of ANOVA, the null hypothesis is that all groups are random samples from the same population. For example, when studying the effect of different treatments on similar samples of patients, the null hypothesis would be that all treatments have the same effect (perhaps none). Rejecting the null hypothesis is taken to mean that the differences in observed effects between treatment groups are unlikely to be due to random chance.

By construction, hypothesis testing limits the rate of Type I errors (false positives) to a significance level. Experimenters also wish to limit Type II errors (false negatives). The rate of Type II errors depends largely on sample size (the rate is larger for smaller samples), significance level (when the standard of proof is high, the chances of overlooking a discovery are also high) and effect size (a smaller effect size is more prone to Type II error).

The terminology of ANOVA is largely from the statistical design of experiments. The experimenter adjusts factors and measures responses in an attempt to determine an effect. Factors are assigned to experimental units by a combination of randomization and blocking to ensure the validity of the results. Blinding keeps the weighing impartial. Responses show a variability that is partially the result of the effect and is partially random error.

ANOVA is the synthesis of several ideas and it is used for multiple purposes. As a consequence, it is difficult to define concisely or precisely.

"Classical" ANOVA for balanced data does three things at once:

1. As exploratory data analysis, an ANOVA employs an additive data decomposition, and its sums of squares indicate the variance of each component of the decomposition (or, equivalently, each set of terms of a linear model).
2. Comparisons of mean squares, along with an *F*-test ... allow testing of a nested sequence of models.
3. Closely related to the ANOVA is a linear model fit with coefficient estimates and standard errors.[12]

In short, ANOVA is a statistical tool used in several ways to develop and confirm an explanation for the observed data.

Additionally:

4. It is computationally elegant and relatively robust against violations of its assumptions.
5. ANOVA provides strong (multiple sample comparison) statistical analysis.
6. It has been adapted to the analysis of a variety of experimental designs.

As a result: ANOVA "has long enjoyed the status of being the most used (some would say abused) statistical technique in psychological research."[13] ANOVA "is probably the most useful technique in the field of statistical inference."[14]

ANOVA is difficult to teach, particularly for complex experiments, with split-plot designs being notorious.[15] In some cases the proper application of the method is best determined by problem pattern recognition followed by the consultation of a classic authoritative test.[16]

Design-of-experiments terms

(Condensed from the "NIST Engineering Statistics Handbook": Section 5.7. A Glossary of DOE Terminology.)[17]

Balanced design
: An experimental design where all cells (i.e. treatment combinations) have the same number of observations.

Blocking
: A schedule for conducting treatment combinations in an experimental study such that any effects on the experimental results due to a known change in raw materials, operators, machines, etc., become concentrated in the levels of the blocking variable. The reason for blocking is to isolate a systematic effect and prevent it from obscuring the main effects. Blocking is achieved by restricting randomization.

Design
: A set of experimental runs which allows the fit of a particular model and the estimate of effects.

DOE
: Design of experiments. An approach to problem solving involving collection of data that will support valid, defensible, and supportable conclusions.[18]

Effect
: How changing the settings of a factor changes the response. The effect of a single factor is also called a main effect.

Error
: Unexplained variation in a collection of observations. DOE's typically require understanding of both random error and lack of fit error.

Experimental unit
: The entity to which a specific treatment combination is applied.

Factors
: Process inputs that an investigator manipulates to cause a change in the output.

Lack-of-fit error
: Error that occurs when the analysis omits one or more important terms or factors from the process model. Including replication in a DOE allows separation of experimental error into its components: lack of fit and random (pure) error.

Model
: Mathematical relationship which relates changes in a given response to changes in one or more factors.

Random error
: Error that occurs due to natural variation in the process. Random error is typically assumed to be normally distributed with zero mean and a constant variance. Random error is also called experimental error.

Randomization
: A schedule for allocating treatment material and for conducting treatment combinations in a DOE such that the conditions in one run neither depend on the conditions of the previous run nor predict the conditions in the subsequent runs.[nb 1]

Replication
: Performing the same treatment combination more than once. Including replication allows an estimate of the random error independent of any lack of fit error.

Responses
: The output(s) of a process. Sometimes called dependent variable(s).

Treatment
: A treatment is a specific combination of factor levels whose effect is to be compared with other treatments.

Classes of models

There are three classes of models used in the analysis of variance, and these are outlined here.

Fixed-effects models

The fixed-effects model (class I) of analysis of variance applies to situations in which the experimenter applies one or more treatments to the subjects of the experiment to see whether the response variable values change. This allows the experimenter to estimate the ranges of response variable values that the treatment would generate in the population as a whole.

Random-effects models

Random-effects model (class II) is used when the treatments are not fixed. This occurs when the various factor levels are sampled from a larger population. Because the levels themselves are random variables, some assumptions and the method of contrasting the treatments (a multi-variable generalization of simple differences) differ from the fixed-effects model.[19]

Mixed-effects models

A mixed-effects model (class III) contains experimental factors of both fixed and random-effects types, with appropriately different interpretations and analysis for the two types.

Example: Teaching experiments could be performed by a college or university department to find a good introductory textbook, with each text considered a treatment. The fixed-effects model would compare a list of candidate texts. The random-effects model would determine whether important differences exist among a list of randomly selected texts. The mixed-effects model would compare the (fixed) incumbent texts to randomly selected alternatives.

Defining fixed and random effects has proven elusive, with competing definitions arguably leading toward a linguistic quagmire.[20]

Assumptions

The analysis of variance has been studied from several approaches, the most common of which uses a linear model that relates the response to the treatments and blocks. Note that the model is linear in parameters but may be nonlinear across factor levels. Interpretation is easy when data is balanced across factors but much deeper understanding is needed for unbalanced data.

Textbook analysis using a normal distribution

The analysis of variance can be presented in terms of a linear model, which makes the following assumptions about the probability distribution of the responses:[21][22][23][24]

- Independence of observations – this is an assumption of the model that simplifies the statistical analysis.
- Normality – the distributions of the residuals are normal.
- Equality (or "homogeneity") of variances, called homoscedasticity — the variance of data in groups should be the same.

The separate assumptions of the textbook model imply that the errors are independently, identically, and normally distributed for fixed effects models, that is, that the errors (ε) are independent and

$$\varepsilon \sim N(0, \sigma^2).$$

Randomization-based analysis

In a randomized controlled experiment, the treatments are randomly assigned to experimental units, following the experimental protocol. This randomization is objective and declared before the experiment is carried out. The objective random-assignment is used to test the significance of the null hypothesis, following the ideas of C. S. Peirce and Ronald Fisher. This design-based analysis was discussed and developed by Francis J. Anscombe at Rothamsted Experimental Station and by Oscar Kempthorne at Iowa State University.[25] Kempthorne and his students make an assumption of *unit treatment additivity*, which is discussed in the books of Kempthorne and David R. Cox.

Unit-treatment additivity

In its simplest form, the assumption of unit-treatment additivity[nb 2] states that the observed response $y_{i,j}$ from experimental unit i when receiving treatment j can be written as the sum of the unit's response y_i and the treatment-effect t_j, that is [26][27][28]

$$y_{i,j} = y_i + t_j.$$

The assumption of unit-treatment additivity implies that, for every treatment j, the jth treatment has exactly the same effect t_j on every experiment unit.

The assumption of unit treatment additivity usually cannot be directly falsified, according to Cox and Kempthorne. However, many *consequences* of treatment-unit additivity can be falsified. For a randomized experiment, the assumption of unit-treatment additivity *implies* that the variance is constant for all treatments. Therefore, by contraposition, a necessary condition for unit-treatment additivity is that the variance is constant.

The use of unit treatment additivity and randomization is similar to the design-based inference that is standard in finite-population survey sampling.

Derived linear model

Kempthorne uses the randomization-distribution and the assumption of *unit treatment additivity* to produce a *derived linear model*, very similar to the textbook model discussed previously.[29] The test statistics of this derived linear model are closely approximated by the test statistics of an appropriate normal linear model, according to approximation theorems and simulation studies.[30] However, there are differences. For example, the randomization-based analysis results in a small but (strictly) negative correlation between the observations.[31][32] In the randomization-based analysis, there is *no assumption* of a *normal* distribution and certainly *no assumption* of *independence*. On the contrary, *the observations are dependent*!

The randomization-based analysis has the disadvantage that its exposition involves tedious algebra and extensive time. Since the randomization-based analysis is complicated and is closely approximated by the approach using a normal linear model, most teachers emphasize the normal linear model approach. Few statisticians object to model-based analysis of balanced randomized experiments.

Statistical models for observational data

However, when applied to data from non-randomized experiments or observational studies, model-based analysis lacks the warrant of randomization.[33] For observational data, the derivation of confidence intervals must use *subjective* models, as emphasized by Ronald Fisher and his followers. In practice, the estimates of treatment-effects from observational studies generally are often inconsistent. In practice, "statistical models" and observational data are useful for suggesting hypotheses that should be treated very cautiously by the public.[34]

Summary of assumptions

The normal-model based ANOVA analysis assumes the independence, normality and homogeneity of variances of the residuals. The randomization-based analysis assumes only the homogeneity of the variances of the residuals (as a consequence of unit-treatment additivity) and uses the randomization procedure of the experiment. Both these analyses require homoscedasticity, as an assumption for the normal-model analysis and as a consequence of randomization and additivity for the randomization-based analysis.

However, studies of processes that change variances rather than means (called dispersion effects) have been successfully conducted using ANOVA.[35] There are *no* necessary assumptions for ANOVA in its full generality, but the *F*-test used for ANOVA hypothesis testing has assumptions and practical limitations which are of continuing interest.

Problems which do not satisfy the assumptions of ANOVA can often be transformed to satisfy the assumptions. The property of unit-treatment additivity is not invariant under a "change of scale", so statisticians often use transformations to achieve unit-treatment additivity. If the response variable is expected to follow a parametric family of probability distributions, then the statistician may specify (in the protocol for the experiment or observational study) that the

responses be transformed to stabilize the variance.[36] Also, a statistician may specify that logarithmic transforms be applied to the responses, which are believed to follow a multiplicative model.[27][37] According to Cauchy's functional equation theorem, the logarithm is the only continuous transformation that transforms real multiplication to addition.

Characteristics

ANOVA is used in the analysis of comparative experiments, those in which only the difference in outcomes is of interest. The statistical significance of the experiment is determined by a ratio of two variances. This ratio is independent of several possible alterations to the experimental observations: Adding a constant to all observations does not alter significance. Multiplying all observations by a constant does not alter significance. So ANOVA statistical significance result is independent of constant bias and scaling errors as well as the units used in expressing observations. In the era of mechanical calculation it was common to subtract a constant from all observations (when equivalent to dropping leading digits) to simplify data entry.[38][39] This is an example of data coding.

Logic

The calculations of ANOVA can be characterized as computing a number of means and variances, dividing two variances and comparing the ratio to a handbook value to determine statistical significance. Calculating a treatment effect is then trivial, "the effect of any treatment is estimated by taking the difference between the mean of the observations which receive the treatment and the general mean".[40]

Partitioning of the sum of squares

ANOVA uses traditional standardized terminology. The definitional equation of sample variance is $s^2 = \frac{1}{n-1}\sum(y_i - \bar{y})^2$, where the divisor is called the degrees of freedom (DF), the summation is called the sum of squares (SS), the result is called the mean square (MS) and the squared terms are deviations from the sample mean. ANOVA estimates 3 sample variances: a total variance based on all the observation deviations from the grand mean, an error variance based on all the observation deviations from their appropriate treatment means, and a treatment variance. The treatment variance is based on the deviations of treatment means from the grand mean, the result being multiplied by the number of observations in each treatment to account for the difference between the variance of observations and the variance of means.

The fundamental technique is a partitioning of the total sum of squares *SS* into components related to the effects used in the model. For example, the model for a simplified ANOVA with one type of treatment at different levels.

$$SS_{\text{Total}} = SS_{\text{Error}} + SS_{\text{Treatments}}$$

The number of degrees of freedom *DF* can be partitioned in a similar way: one of these components (that for error) specifies a chi-squared distribution which describes the associated sum of squares, while the same is true for "treatments" if there is no treatment effect.

$$DF_{\text{Total}} = DF_{\text{Error}} + DF_{\text{Treatments}}$$

See also Lack-of-fit sum of squares.

The *F*-test

The *F*-test is used for comparing the factors of the total deviation. For example, in one-way, or single-factor ANOVA, statistical significance is tested for by comparing the F test statistic

$$F = \frac{\text{variance between treatments}}{\text{variance within treatments}}$$

$$F = \frac{MS_{\text{Treatments}}}{MS_{\text{Error}}} = \frac{SS_{\text{Treatments}}/(I-1)}{SS_{\text{Error}}/(n_T - I)}$$

where *MS* is mean square, I = number of treatments and n_T = total number of cases

to the *F*-distribution with $I - 1$, $n_T - I$ degrees of freedom. Using the *F*-distribution is a natural candidate because the test statistic is the ratio of two scaled sums of squares each of which follows a scaled chi-squared distribution.

The expected value of F is $1 + n\sigma^2_{\text{Treatment}}/\sigma^2_{\text{Error}}$ (where n is the treatment sample size) which is 1 for no treatment effect. As values of F increase above 1, the evidence is increasingly inconsistent with the null hypothesis. Two apparent experimental methods of increasing F are increasing the sample size and reducing the error variance by tight experimental controls.

There are two methods of concluding the ANOVA hypothesis test, both of which produce the same result:

- The textbook method is to compare the observed value of F with the critical value of F determined from tables. The critical value of F is a function of the degrees of freedom of the numerator and the denominator and the significance level (α). If $F \geq F_{\text{Critical}}$, the null hypothesis is rejected.
- The computer method calculates the probability (p-value) of a value of F greater than or equal to the observed value. The null hypothesis is rejected if this probability is less than or equal to the significance level (α).

The ANOVA *F*-test is known to be nearly optimal in the sense of minimizing false negative errors for a fixed rate of false positive errors (i.e. maximizing power for a fixed significance level). For example, to test the hypothesis that various medical treatments have exactly the same effect, the *F*-test's *p*-values closely approximate the permutation test's p-values: The approximation is particularly close when the design is balanced.[30][41] Such permutation tests characterize tests with maximum power against all alternative hypotheses, as observed by Rosenbaum.[nb 3] The ANOVA *F*-test (of the null-hypothesis that all treatments have exactly the same effect) is recommended as a practical test, because of its robustness against many alternative distributions.[42][nb 4]

Extended logic

ANOVA consists of separable parts; partitioning sources of variance and hypothesis testing can be used individually. ANOVA is used to support other statistical tools. Regression is first used to fit more complex models to data, then ANOVA is used to compare models with the objective of selecting simple(r) models that adequately describe the data. "Such models could be fit without any reference to ANOVA, but ANOVA tools could then be used to make some sense of the fitted models, and to test hypotheses about batches of coefficients."[43] "[W]e think of the analysis of variance as a way of understanding and structuring multilevel models—not as an alternative to regression but as a tool for summarizing complex high-dimensional inferences ..."[43]

For a single factor

The simplest experiment suitable for ANOVA analysis is the completely randomized experiment with a single factor. More complex experiments with a single factor involve constraints on randomization and include completely randomized blocks and Latin squares (and variants: Graeco-Latin squares, etc.). The more complex experiments share many of the

complexities of multiple factors. A relatively complete discussion of the analysis (models, data summaries, ANOVA table) of the completely randomized experiment is available.

For multiple factors

ANOVA generalizes to the study of the effects of multiple factors. When the experiment includes observations at all combinations of levels of each factor, it is termed factorial. Factorial experiments are more efficient than a series of single factor experiments and the efficiency grows as the number of factors increases.[44] Consequently, factorial designs are heavily used.

The use of ANOVA to study the effects of multiple factors has a complication. In a 3-way ANOVA with factors x, y and z, the ANOVA model includes terms for the main effects (x, y, z) and terms for interactions (xy, xz, yz, xyz). All terms require hypothesis tests. The proliferation of interaction terms increases the risk that some hypothesis test will produce a false positive by chance. Fortunately, experience says that high order interactions are rare.[45] The ability to detect interactions is a major advantage of multiple factor ANOVA. Testing one factor at a time hides interactions, but produces apparently inconsistent experimental results.[44]

Caution is advised when encountering interactions; Test interaction terms first and expand the analysis beyond ANOVA if interactions are found. Texts vary in their recommendations regarding the continuation of the ANOVA procedure after encountering an interaction. Interactions complicate the interpretation of experimental data. Neither the calculations of significance nor the estimated treatment effects can be taken at face value. "A significant interaction will often mask the significance of main effects."[46] Graphical methods are recommended to enhance understanding. Regression is often useful. A lengthy discussion of interactions is available in Cox (1958).[47] Some interactions can be removed (by transformations) while others cannot.

A variety of techniques are used with multiple factor ANOVA to reduce expense. One technique used in factorial designs is to minimize replication (possibly no replication with support of analytical trickery) and to combine groups when effects are found to be statistically (or practically) insignificant. An experiment with many insignificant factors may collapse into one with a few factors supported by many replications.[48]

Worked numeric examples

Numerous fully worked numerical examples are available in standard textbooks and online. A simple case uses one-way (a single factor) analysis.

Associated analysis

Some analysis is required in support of the *design* of the experiment while other analysis is performed after changes in the factors are formally found to produce statistically significant changes in the responses. Because experimentation is iterative, the results of one experiment alter plans for following experiments.

Preparatory analysis

The number of experimental units

In the design of an experiment, the number of experimental units is planned to satisfy the goals of the experiment. Experimentation is often sequential.

Early experiments are often designed to provide mean-unbiased estimates of treatment effects and of experimental error. Later experiments are often designed to test a hypothesis that a treatment effect has an important magnitude; in this case, the number of experimental units is chosen so that the experiment is within budget and has adequate power, among other goals.

Reporting sample size analysis is generally required in psychology. "Provide information on sample size and the process that led to sample size decisions."[49] The analysis, which is written in the experimental protocol before the experiment is conducted, is examined in grant applications and administrative review boards.

Besides the power analysis, there are less formal methods for selecting the number of experimental units. These include graphical methods based on limiting the probability of false negative errors, graphical methods based on an expected variation increase (above the residuals) and methods based on achieving a desired confident interval.[50]

Power analysis

Power analysis is often applied in the context of ANOVA in order to assess the probability of successfully rejecting the null hypothesis if we assume a certain ANOVA design, effect size in the population, sample size and significance level. Power analysis can assist in study design by determining what sample size would be required in order to have a reasonable chance of rejecting the null hypothesis when the alternative hypothesis is true.[51][52][53][54]

Effect size

Several standardized measures of effect have been proposed for ANOVA to summarize the strength of the association between a predictor(s) and the dependent variable or the overall standardized difference of the complete model. Standardized effect-size estimates facilitate comparison of findings across studies and disciplines. However, while standardized effect sizes are commonly used in much of the professional literature, a non-standardized measure of effect size that has immediately "meaningful" units may be preferable for reporting purposes.[55]

Follow-up analysis

It is always appropriate to carefully consider outliers. They have a disproportionate impact on statistical conclusions and are often the result of errors.

Model confirmation

It is prudent to verify that the assumptions of ANOVA have been met. Residuals are examined or analyzed to confirm homoscedasticity and gross normality.[56] Residuals should have the appearance of (zero mean normal distribution) noise when plotted as a function of anything including time and modeled data values. Trends hint at interactions among factors or among observations. One rule of thumb: "If the largest standard deviation is less than twice the smallest standard deviation, we can use methods based on the assumption of equal standard deviations and our results will still be approximately correct."[57]

Follow-up tests

A statistically significant effect in ANOVA is often followed up with one or more different follow-up tests. This can be done in order to assess which groups are different from which other groups or to test various other focused hypotheses. Follow-up tests are often distinguished in terms of whether they are planned (a priori) or post hoc. Planned tests are determined before looking at the data and post hoc tests are performed after looking at the data.

Often one of the "treatments" is none, so the treatment group can act as a control. Dunnett's test (a modification of the *t*-test) tests whether each of the other treatment groups has the same mean as the control.[58]

Post hoc tests such as Tukey's range test most commonly compare every group mean with every other group mean and typically incorporate some method of controlling for Type I errors. Comparisons, which are most commonly planned, can be either simple or compound. Simple comparisons compare one group mean with one other group mean. Compound comparisons typically compare two sets of groups means where one set has two or more groups (e.g., compare average group means of group A, B and C with group D). Comparisons can also look at tests of trend, such as linear and quadratic relationships, when the independent variable involves ordered levels.

Following ANOVA with pair-wise multiple-comparison tests has been criticized on several grounds.[55][59] There are many such tests (10 in one table) and recommendations regarding their use are vague or conflicting.[60][61]

Study designs

There are several types of ANOVA. Many statisticians base ANOVA on the design of the experiment,[62] especially on the protocol that specifies the random assignment of treatments to subjects; the protocol's description of the assignment mechanism should include a specification of the structure of the treatments and of any blocking. It is also common to apply ANOVA to observational data using an appropriate statistical model.

Some popular designs use the following types of ANOVA:

- One-way ANOVA is used to test for differences among two or more independent groups (means), e.g. different levels of urea application in a crop, or different levels of antibiotic action on several different bacterial species,[63] or different levels of effect of some medicine on groups of patients. However, should these groups not be independent, and there is an order in the groups (such as mild, moderate and severe disease), or in the dose of a drug (such as 5 mg/mL, 10 mg/mL, 20 mg/mL) given to the same group of patients, then a linear trend estimation should be used. Typically, however, the one-way ANOVA is used to test for differences among at least three groups, since the two-group case can be covered by a t-test.[64] When there are only two means to compare, the t-test and the ANOVA *F*-test are equivalent; the relation between ANOVA and *t* is given by $F = t^2$.
- Factorial ANOVA is used when the experimenter wants to study the interaction effects among the treatments.
- Repeated measures ANOVA is used when the same subjects are used for each treatment (e.g., in a longitudinal study).
- Multivariate analysis of variance (MANOVA) is used when there is more than one response variable.

Cautions

Balanced experiments (those with an equal sample size for each treatment) are relatively easy to interpret; Unbalanced experiments offer more complexity. For single-factor (one-way) ANOVA, the adjustment for unbalanced data is easy, but the unbalanced analysis lacks both robustness and power.[65] For more complex designs the lack of balance leads to further complications. "The orthogonality property of main effects and interactions present in balanced data does not carry over to the unbalanced case. This means that the usual analysis of variance techniques do not apply. Consequently, the analysis of unbalanced factorials is much more difficult than that for balanced designs."[66] In the general case, "The analysis of variance can also be applied to unbalanced data, but then the sums of squares, mean squares, and *F*-ratios will depend on the order in which the sources of variation are considered."[43] The simplest techniques for handling unbalanced data restore balance by either throwing out data or by synthesizing missing data. More complex techniques use regression.

ANOVA is (in part) a test of statistical significance. The American Psychological Association holds the view that simply reporting statistical significance is insufficient and that reporting confidence bounds is preferred.[55]

While ANOVA is conservative (in maintaining a significance level) against multiple comparisons in one dimension, it is not conservative against comparisons in multiple dimensions.[67]

Generalizations

ANOVA is considered to be a special case of linear regression[68][69] which in turn is a special case of the general linear model.[70] All consider the observations to be the sum of a model (fit) and a residual (error) to be minimized.

The Kruskal–Wallis test and the Friedman test are nonparametric tests, which do not rely on an assumption of normality.[71][72]

Connection to linear regression

Below we make clear the connection between multi-way ANOVA and linear regression.

Linearly re-order the data so that k^{th} observation is associated with a response y_k and factors $Z_{k,b}$ where $b \in \{1, 2, \ldots, B\}$ denotes the different factors and B is the total number of factors. In one-way ANOVA $B = 1$ and in two-way ANOVA $B = 2$. Furthermore, we assume the bth factor has I_b levels, namely $\{1, 2, \ldots, I_b\}$. Now, we can one-hot encode the factors into the $\sum_{b=1}^{B} I_b$ dimensional vector v_k.

The one-hot encoding function $g_b : \{1, 2, \ldots, I_b\} \mapsto \{0, 1\}^{I_b}$ is defined such that the ith entry of $g_b(Z_{k,b})$ is

$$g_b(Z_{k,b})_i = \begin{cases} 1 & \text{if } i = Z_{k,b} \\ 0 & \text{otherwise} \end{cases}$$

The vector v_k is the concatenation of all of the above vectors for all b. Thus, $v_k = [g_1(Z_{k,1}), g_2(Z_{k,2}), \ldots, g_B(Z_{k,B})]$. In order to obtain a fully general B-way interaction ANOVA we must also concatenate every additional interaction term in the vector v_k and then add an intercept term. Let that vector be X_k.

With this notation in place, we now have the exact connection with linear regression. We simply regress response y_k against the vector X_k. However, there is a concern about identifiability. In order to overcome such issues we assume that the sum of the parameters within each set of interactions is equal to zero. From here, one can use *F*-statistics or other methods to determine the relevance of the individual factors.

Example

We can consider the 2-way interaction example where we assume that the first factor has 2 levels and the second factor has 3 levels.

Define $a_i = 1$ if $Z_{k,1} = i$ and $b_i = 1$ if $Z_{k,2} = i$, i.e. a is the one-hot encoding of the first factor and b is the one-hot encoding of the second factor.

With that,

$$X_k = [a_1, a_2, b_1, b_2, b_3, a_1 \times b_1, a_1 \times b_2, a_1 \times b_3, a_2 \times b_1, a_2 \times b_2, a_2 \times b_3, 1]$$

where the last term is an intercept term. For a more concrete example suppose that

$$Z_{k,1} = 2$$
$$Z_{k,2} = 1$$

Then,

$$X_k = [0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1]$$

See also

- One-way analysis of variance (**one-way ANOVA**)
- Two-way analysis of variance (**two-way ANOVA**)
- ANOVA on ranks
- ANOVA-simultaneous component analysis
- Analysis of covariance (**ANCOVA**)
- Multivariate analysis of variance (**MANOVA**)
- Multivariate analysis of covariance (**MANCOVA**)
- Analysis of molecular variance (AMOVA)
- Analysis of rhythmic variance (ANORVA)
- Explained variation
- Mixed-design analysis of variance
- Permutational analysis of variance
- Repeated measures ANOVA
- Variance decomposition

Footnotes

1. Randomization is a term used in multiple ways in this material. "Randomization has three roles in applications: as a device for eliminating biases, for example from unobserved explanatory variables and selection effects; as a basis for estimating standard errors; and as a foundation for formally exact significance tests." Cox (2006, page 192) Hinkelmann and Kempthorne use randomization both in experimental design and for statistical analysis.
2. Unit-treatment additivity is simply termed additivity in most texts. Hinkelmann and Kempthorne add adjectives and distinguish between additivity in the strict and broad senses. This allows a detailed consideration of multiple error sources (treatment, state, selection, measurement and sampling) on page 161.
3. Rosenbaum (2002, page 40) cites Section 5.7 (Permutation Tests), Theorem 2.3 (actually Theorem 3, page 184) of Lehmann's *Testing Statistical Hypotheses* (1959).
4. The *F*-test for the comparison of variances has a mixed reputation. It is not recommended as a hypothesis test to determine whether two *different* samples have the same variance. It is recommended for ANOVA where two estimates of the variance of the *same* sample are compared. While the *F*-test is not generally robust against departures from normality, it has been found to be robust in the special case of ANOVA. Citations from Moore & McCabe (2003): "Analysis of variance uses F statistics, but these are not the same as the F statistic for comparing two population standard deviations." (page 554) "The F test and other procedures for inference about variances are so lacking in robustness as to be of little use in practice." (page 556) "[The ANOVA *F*-test] is relatively insensitive to moderate nonnormality and unequal variances, especially when the sample sizes are similar." (page 763) ANOVA assumes homoscedasticity, but it is robust. The statistical test for homoscedasticity (the *F*-test) is not robust. Moore & McCabe recommend a rule of thumb.

Notes

1. Stigler (1986)
2. Stigler (1986, p 134)

3. Stigler (1986, p 153)
4. Stigler (1986, pp 154–155)
5. Stigler (1986, pp 240–242)
6. Stigler (1986, Chapter 7 – Psychophysics as a Counterpoint)
7. Stigler (1986, p 253)
8. Stigler (1986, pp 314–315)
9. *The Correlation Between Relatives on the Supposition of Mendelian Inheritance*. Ronald A. Fisher. *Philosophical Transactions of the Royal Society of Edinburgh*. 1918. (volume 52, pages 399–433)
10. On the "Probable Error" of a Coefficient of Correlation Deduced from a Small Sample. Ronald A. Fisher. Metron, 1: 3–32 (1921)
11. Scheffé (1959, p 291, "Randomization models were first formulated by Neyman (1923) for the completely randomized design, by Neyman (1935) for randomized blocks, by Welch (1937) and Pitman (1937) for the Latin square under a certain null hypothesis, and by Kempthorne (1952, 1955) and Wilk (1955) for many other designs.")
12. Gelman (2005, p 2)
13. Howell (2002, p 320)
14. Montgomery (2001, p 63)
15. Gelman (2005, p 1)
16. Gelman (2005, p 5)
17. "Section 5.7. A Glossary of DOE Terminology" (http://www.itl.nist.gov/div898/handbook/pri/section7/pri7.htm). *NIST Engineering Statistics handbook*. NIST. Retrieved 5 April 2012.
18. "Section 4.3.1 A Glossary of DOE Terminology" (http://www.itl.nist.gov/div898/handbook/pmd/section3/pmd31.htm). *NIST Engineering Statistics handbook*. NIST. Retrieved 14 August 2012.
19. Montgomery (2001, Chapter 12: Experiments with random factors)
20. Gelman (2005, pp. 20–21)
21. Snedecor, George W.; Cochran, William G. (1967). *Statistical Methods* (6th ed.). p. 321.
22. Cochran & Cox (1992, p 48)
23. Howell (2002, p 323)
24. Anderson, David R.; Sweeney, Dennis J.; Williams, Thomas A. (1996). *Statistics for business and economics* (6th ed.). Minneapolis/St. Paul: West Pub. Co. pp. 452–453. ISBN 978-0-314-06378-6.
25. Anscombe (1948)
26. Kempthorne (1979, p 30)
27. Cox (1958, Chapter 2: Some Key Assumptions)
28. Hinkelmann and Kempthorne (2008, Volume 1, Throughout. Introduced in Section 2.3.3: Principles of experimental design; The linear model; Outline of a model)
29. Hinkelmann and Kempthorne (2008, Volume 1, Section 6.3: Completely Randomized Design; Derived Linear Model)
30. Hinkelmann and Kempthorne (2008, Volume 1, Section 6.6: Completely randomized design; Approximating the randomization test)
31. Bailey (2008, Chapter 2.14 "A More General Model" in Bailey, pp. 38–40)
32. Hinkelmann and Kempthorne (2008, Volume 1, Chapter 7: Comparison of Treatments)
33. Kempthorne (1979, pp 125–126, "The experimenter must decide which of the various causes that he feels will produce variations in his results must be controlled experimentally. Those causes that he does not control experimentally, because he is not cognizant of them, he must control by the device of randomization." "[O]nly when the treatments in the experiment are applied by the experimenter using the full randomization procedure is the chain of inductive inference sound. It is *only* under these circumstances that the experimenter can attribute whatever effects he observes to the treatment and the treatment only. Under these circumstances his conclusions are reliable in the statistical sense.")
34. Freedman

35. Montgomery (2001, Section 3.8: Discovering dispersion effects)
36. Hinkelmann and Kempthorne (2008, Volume 1, Section 6.10: Completely randomized design; Transformations)
37. Bailey (2008)
38. Montgomery (2001, Section 3-3: Experiments with a single factor: The analysis of variance; Analysis of the fixed effects model)
39. Cochran & Cox (1992, p 2 example)
40. Cochran & Cox (1992, p 49)
41. Hinkelmann and Kempthorne (2008, Volume 1, Section 6.7: Completely randomized design; CRD with unequal numbers of replications)
42. Moore and McCabe (2003, page 763)
43. Gelman (2008)
44. Montgomery (2001, Section 5-2: Introduction to factorial designs; The advantages of factorials)
45. Belle (2008, Section 8.4: High-order interactions occur rarely)
46. Montgomery (2001, Section 5-1: Introduction to factorial designs; Basic definitions and principles)
47. Cox (1958, Chapter 6: Basic ideas about factorial experiments)
48. Montgomery (2001, Section 5-3.7: Introduction to factorial designs; The two-factor factorial design; One observation per cell)
49. Wilkinson (1999, p 596)
50. Montgomery (2001, Section 3-7: Determining sample size)
51. Howell (2002, Chapter 8: Power)
52. Howell (2002, Section 11.12: Power (in ANOVA))
53. Howell (2002, Section 13.7: Power analysis for factorial experiments)
54. Moore and McCabe (2003, pp 778–780)
55. Wilkinson (1999, p 599)
56. Montgomery (2001, Section 3-4: Model adequacy checking)
57. Moore and McCabe (2003, p 755, Qualifications to this rule appear in a footnote.)
58. Montgomery (2001, Section 3-5.8: Experiments with a single factor: The analysis of variance; Practical interpretation of results; Comparing means with a control)
59. Hinkelmann and Kempthorne (2008, Volume 1, Section 7.5: Comparison of Treatments; Multiple Comparison Procedures)
60. Howell (2002, Chapter 12: Multiple comparisons among treatment means)
61. Montgomery (2001, Section 3-5: Practical interpretation of results)
62. Cochran & Cox (1957, p 9, "[T]he general rule [is] that the way in which the experiment is conducted determines not only whether inferences can be made, but also the calculations required to make them.")
63. One-way/single factor ANOVA. Biomedical Statistics (http://www.biomedicalstatistics.info/en/multiplegroups/one-way-anova.html) Archived (https://web.archive.org/web/20141107211953/http://www.biomedicalstatistics.info/en/multiplegroups/one-way-anova.html) 7 November 2014 at the Wayback Machine
64. "The Probable Error of a Mean". *Biometrika*. **6**: 1–25. 1908. doi:10.1093/biomet/6.1.1 (https://doi.org/10.1093%2Fbiomet%2F6.1.1).
65. Montgomery (2001, Section 3-3.4: Unbalanced data)
66. Montgomery (2001, Section 14-2: Unbalanced data in factorial design)
67. Wilkinson (1999, p 600)
68. Gelman (2005, p.1) (with qualification in the later text)
69. Montgomery (2001, Section 3.9: The Regression Approach to the Analysis of Variance)
70. Howell (2002, p 604)
71. Howell (2002, Chapter 18: Resampling and nonparametric approaches to data)

72. Montgomery (2001, Section 3-10: Nonparametric methods in the analysis of variance)

References

- Anscombe, F. J. (1948). "The Validity of Comparative Experiments". *Journal of the Royal Statistical Society. Series A (General)*. **111** (3): 181–211. doi:10.2307/2984159 (https://doi.org/10.2307%2F2984159). JSTOR 2984159 (https://www.jstor.org/stable/2984159). MR 0030181 (https://www.ams.org/mathscinet-getitem?mr=0030181).
- Bailey, R. A. (2008). *Design of Comparative Experiments* (http://www.maths.qmul.ac.uk/~rab/DOEbook). Cambridge University Press. ISBN 978-0-521-68357-9. Pre-publication chapters are available on-line.
- Belle, Gerald van (2008). *Statistical rules of thumb* (2nd ed.). Hoboken, N.J: Wiley. ISBN 978-0-470-14448-0.
- Cochran, William G.; Cox, Gertrude M. (1992). *Experimental designs* (2nd ed.). New York: Wiley. ISBN 978-0-471-54567-5.
- Cohen, Jacob (1988). *Statistical power analysis for the behavior sciences* (2nd ed.). Routledge ISBN 978-0-8058-0283-2
- Cohen, Jacob (1992). "Statistics a power primer". *Psychological Bulletin*. **112** (1): 155–159. doi:10.1037/0033-2909.112.1.155 (https://doi.org/10.1037%2F0033-2909.112.1.155). PMID 19565683 (https://www.ncbi.nlm.nih.gov/pubmed/19565683).
- Cox, David R. (1958). *Planning of experiments*. Reprinted as ISBN 978-0-471-57429-3
- Cox, D. R. (2006). *Principles of statistical inference*. Cambridge New York: Cambridge University Press. ISBN 978-0-521-68567-2.
- Freedman, David A.(2005). *Statistical Models: Theory and Practice*, Cambridge University Press. ISBN 978-0-521-67105-7
- Gelman, Andrew (2005). "Analysis of variance? Why it is more important than ever". *The Annals of Statistics*. **33**: 1–53. arXiv:math/0504499 (https://arxiv.org/abs/math/0504499). doi:10.1214/009053604000001048 (https://doi.org/10.1214%2F009053604000001048).
- Gelman, Andrew (2008). "Variance, analysis of". *The new Palgrave dictionary of economics* (2nd ed.). Basingstoke, Hampshire New York: Palgrave Macmillan. ISBN 978-0-333-78676-5.
- Hinkelmann, Klaus & Kempthorne, Oscar (2008). *Design and Analysis of Experiments*. I and II (Second ed.). Wiley. ISBN 978-0-470-38551-7.
- Howell, David C. (2002). *Statistical methods for psychology* (5th ed.). Pacific Grove, CA: Duxbury/Thomson Learning. ISBN 978-0-534-37770-0.
- Kempthorne, Oscar (1979). *The Design and Analysis of Experiments* (Corrected reprint of (1952) Wiley ed.). Robert E. Krieger. ISBN 978-0-88275-105-4.
- Lehmann, E.L. (1959) Testing Statistical Hypotheses. John Wiley & Sons.
- Montgomery, Douglas C. (2001). *Design and Analysis of Experiments* (5th ed.). New York: Wiley. ISBN 978-0-471-31649-7.
- Moore, David S. & McCabe, George P. (2003). Introduction to the Practice of Statistics (4e). W H Freeman & Co. ISBN 0-7167-9657-0
- Rosenbaum, Paul R. (2002). *Observational Studies* (2nd ed.). New York: Springer-Verlag. ISBN 978-0-387-98967-9
- Scheffé, Henry (1959). *The Analysis of Variance*. New York: Wiley.
- Stigler, Stephen M. (1986). *The history of statistics : the measurement of uncertainty before 1900*. Cambridge, Mass: Belknap Press of Harvard University Press. ISBN 978-0-674-40340-6.
- Wilkinson, Leland (1999). "Statistical Methods in Psychology Journals; Guidelines and Explanations". *American Psychologist*. **5** (8): 594–604. CiteSeerX 10.1.1.120.4818 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.120.4818). doi:10.1037/0003-066X.54.8.594 (https://doi.org/10.1037%2F0003-066X.54.8.594).

Further reading

- Box, G. e. p. (1953). "Non-Normality and Tests on Variances". *Biometrika*. **40** (3/4): 318–335. doi:10.1093/biomet/40.3-4.318 (https://doi.org/10.1093%2Fbiomet%2F40.3-4.318). JSTOR 2333350 (https://www.jstor.org/stable/2333350).
- Box, G. E. P. (1954). "Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems, I. Effect of Inequality of Variance in the One-Way Classification". *The Annals of Mathematical Statistics*. **25** (2): 290. doi:10.1214/aoms/1177728786 (https://doi.org/10.1214%2Faoms%2F1177728786).
- Box, G. E. P. (1954). "Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems, II. Effects of Inequality of Variance and of Correlation Between Errors in the Two-Way Classification". *The Annals of*

Mathematical Statistics. **25** (3): 484. doi:10.1214/aoms/1177728717 (https://doi.org/10.1214%2Faoms%2F1177728717).

- Caliński, Tadeusz; Kageyama, Sanpei (2000). *Block designs: A Randomization approach, Volume **I**: Analysis*. Lecture Notes in Statistics. **150**. New York: Springer-Verlag. ISBN 978-0-387-98578-7.
- Christensen, Ronald (2002). *Plane Answers to Complex Questions: The Theory of Linear Models* (Third ed.). New York: Springer. ISBN 978-0-387-95361-8.
- Cox, David R. & Reid, Nancy M. (2000). *The theory of design of experiments*. (Chapman & Hall/CRC). ISBN 978-1-58488-195-7
- Fisher, Ronald (1918). "Studies in Crop Variation. I. An examination of the yield of dressed grain from Broadbalk" (https://web.archive.org/web/20010612211752/http://www.library.adelaide.edu.au/digitised/fisher/15.pdf) (PDF). *Journal of Agricultural Science*. **11** (2): 107–135. doi:10.1017/S0021859600003750 (https://doi.org/10.1017%2FS0021859600003750). hdl:2440/15170 (https://hdl.handle.net/2440%2F15170). Archived from the original (http://www.library.adelaide.edu.au/digitised/fisher/15.pdf) (PDF) on 12 June 2001.
- Freedman, David A.; Pisani, Robert; Purves, Roger (2007) *Statistics*, 4th edition. W.W. Norton & Company ISBN 978-0-393-92972-0
- Hettmansperger, T. P.; McKean, J. W. (1998). Edward Arnold (ed.). *Robust nonparametric statistical methods*. Kendall's Library of Statistics. Volume 5 (First ed.). New York: John Wiley & Sons, Inc. pp. xiv+467 pp. ISBN 978-0-340-54937-7. MR 1604954 (https://www.ams.org/mathscinet-getitem?mr=1604954).
- Lentner, Marvin; Thomas Bishop (1993). *Experimental design and analysis* (Second ed.). P.O. Box 884, Blacksburg, VA 24063: Valley Book Company. ISBN 978-0-9616255-2-8.
- Tabachnick, Barbara G. & Fidell, Linda S. (2007). *Using Multivariate Statistics* (5th ed.). Boston: Pearson International Edition. ISBN 978-0-205-45938-4
- Wichura, Michael J. (2006). *The coordinate-free approach to linear models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press. pp. xiv+199. ISBN 978-0-521-86842-6. MR 2283455 (https://www.ams.org/mathscinet-getitem?mr=2283455).
- Phadke, Madhav S. (1989). *Quality Engineering using Robust Design*. New Jersey: Prentice Hall PTR. ISBN 978-0-13-745167-8.

External links

- SOCR ANOVA Activity (http://wiki.stat.ucla.edu/socr/index.php/AP_Statistics_Curriculum_2007_ANOVA_1Way) and interactive applet (http://www.socr.ucla.edu/htmls/ana/ANOVA1Way_Analysis.html).
- Examples of all ANOVA and ANCOVA models with up to three treatment factors, including randomized block, split plot, repeated measures, and Latin squares, and their analysis in R (http://www.southampton.ac.uk/~cpd/anovas/datasets/index.htm) (University of Southampton)
- NIST/SEMATECH e-Handbook of Statistical Methods, section 7.4.3: "Are the means equal?" (http://www.itl.nist.gov/div898/handbook/prc/section4/prc43.htm)
- Analysis of variance: Introduction (https://web.archive.org/web/20150405053021/http://biostat.katerynakon.in.ua/en/multiplegroups/anova.html)
- One Way & Two Way ANOVA Calculator (https://getcalc.com/statistics-anova-calculator.htm)

Retrieved from "https://en.wikipedia.org/w/index.php?title=Analysis_of_variance&oldid=909827584"

This page was last edited on 7 August 2019, at 21:09 (UTC).

Text is available under the Creative Commons Attribution-ShareAlike License; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy. Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a non-profit organization.