IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.53

Volume 4, Issue 2, December 2024

Study on Analysis of Variance in Statistics

Prof. Asane Bharat Baban
Lecturer, Department of Science
Amrutvahini Polytechnic, Sangamner, A.Nagar, Maharashtra, India
asanebb@amrutpoly.in

Abstract: *Variance is a key statistical measure that quantifies the spread of values within a dataset, offering insights into data variability. It evaluates the average squared deviation of data points from the mean and serves as the foundation for statistical methods such as ANOVA and regression analysis. Introduced by Ronald Fisher in 1918, variance builds upon earlier concepts by Gauss and is calculated differently for populations and samples, using Bessel's correction for unbiased estimation in samples. Unlike range, variance incorporates all data points, ensuring comparability across datasets of varying sizes. Its properties, including invariance to additive constants and scalability by squared factors, make it a robust metric for analyzing data spread. Variance is especially critical in experimental designs to distinguish between systematic (between-groups) and unsystematic (within-groups) variance, enabling researchers to assess and interpret variability in diverse contexts.*

Keywords: Variance, Data variability, Statistical analysis, ANOVA, Experimental design.

I. INTRODUCTION

Variability is a numerical description that refers to the spread values within a given distribution. Variability can be described using four common measures: range, the mean deviation score, standard deviation, and variance. Variance is most useful in determining the overall spread of a data set. Range simply takes two data points, the lowest and highest values, and measures the space between them. The interquartile range also interprets the amount of spread within a set of data by using only the middle 50% of scores. Unlike range, variance takes into account every data point within the set and measures each distance from the mean. Variance is calculated using a number of values from data and the associated distribution. Specifically, variance requires raw data scores and the sample size of the data.

The sum of squares (or sum of squared errors) is also used as a method to determine total spread or dispersion. The problem with sum of squares is this value cannot be compared across samples that differ in size. Variance deals with average spread, which is comparable across groups that may change sizes.

Variances are most impactful when comparing multiple distributions. Furthermore, this statistic becomes the basis for various statistical comparisons, including an analysis of variance (ANOVA). In experiments, people or groups undergo treatments to potentially elicit various responses. If all scores, or responses, are different, variability will be large. If all scores are exactly the same, variability will be zero.

Variance is used in many different experimental designs. After discussing a brief history, variance is described not only by its formulae but also by its conceptual properties. Variance is then applied to statistical testing, including ANOVA and multiple regression.

History

Although credit for the concept of variance is given to Ronald Fisher, it is apparent that the concept of variance existed long before Fisher, in the work of Carl Friedrich Gauss and his endeavor to estimate the locations of stars. Within his search, Gauss encountered a probability distribution with deviations that may have given way to the current concept of variance.

Fisher introduced the concept of variance in his 1918 paper "The Correlation Between Relatives on the Supposition of Mendelian Inheritance." Fisher was the first to introduce the test now known as ANOVA. Most of his later work involved significance and hypothesis testing.

IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.53

Volume 4, Issue 2, December 2024

Formulae

When research designs are considered, it is the goal of the researcher to collect as many real-world observations as possible, knowing that the collected observations will not account for all possible observations that exist. As mean and variance are calculated from these select observations, it can be assumed that the values will not perfectly match the values that would have been obtained using every possible observation. Therefore, the researcher must estimate the mean and variance using an equation to account for observational bias. Two formulae exist to calculate variance, one describing the population variance and the other describing the sample variance.

Population variance is calculated using the mean of the squared deviations from the distribution mean. Deviation is the difference between a specific value of data and the distribution's mean. The formula for population variance can be defined as

$$\sigma^2 = \frac{\sum (X-\mu)^2}{N},$$

where σ^2 represents the population variance, $X - \mu$ represents the deviation between each score and the population mean, and N is the population size. This formula is conceptual and represents a hypothetical population. A population refers to the entire group being defined. For example, if the experiment involves 20-year-old females, the population would be every 20-year-old female on the planet. Because the population in question is not always readily available, the researcher must obtain a sample or a proportion of the population. This sample is meant to provide a snapshot of what the population actually looks like.

If a population is extremely large, specifically when it is not possible to count every observation available, a sample of the population must be used to compute variance. This formula for sample variance is used most frequently and can be defined as:

$$S^2 = \frac{\sum (X-\overline{X})^2}{n-1},$$

where S^2 is the variance, $X - \overline{X}$ represents the deviation between each score and the sample mean, and n is the sample size. The numerator is often referred to as the sum of squared errors, or SS. Using $n - 1$ as the denominator for sample variance is called Bessel's correction. The result is an unbiased estimator of the population variance called the corrected sample variance or unbiased sample variance. By taking the sum of the deviations and dividing by Bessel's correction, what results is an average of all deviations. The sum of the deviations of any data set will always be zero; therefore, the deviations must be squared before being summed to account for negative values. If the distribution mean is not easily attainable, another formula may be used while employing raw data:

$$S^2 = \frac{\sum X^2}{N} - \frac{\left(\sum X\right)^2}{N^2}.$$

Both formulas yield identical answers but are dependent on the data and statistics available.

Once variance is calculated, variability can be further explored through another statistic, standard deviation. Standard deviation is another method of determining spread, specifically focusing on the data and its location from the mean. Variance is more about the average distance away from the mean and is reported in square units. Standard deviation can be calculated by taking the square root of the variance:

$$S = \sqrt{S^2}.$$

By taking the square root of the variance, the units are also restored to their original form. Standard deviation is more appropriate when units do not make sense to be squared (e.g., people).

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, December 2024

Properties of Variance

Variance will always yield a nonnegative value because squares are positives or zero. If a constant is applied to all values of a variable through addition, variance will remain unchanged. In other words, variance is invariant as it relates to the location parameter.

However, if a constant is applied through multiplication, or scaling, variance will adjust by the square of that constant.

Variance is often the preferred method of determining the spread of a given data set due to the fact that the variance of the sum of the uncorrelated (independent) random variables can be calculated using the sum of the variances. The Bienaymé formula is a derivation of the aforementioned concept. This derived formula was discovered in 1853 and is defined as:

$$\mathrm{Var}\left(\overline{X}\right) = \frac{\sigma^2}{n},$$

where $\overline{X}$ refers to the sample mean. The Bienaymé formula can be interpreted to indicate that the mean in every experiment with a sample size of n will fall within $\pm\sigma n$ of the true population 68% of the time. The 68% refers to the area under the normal curve within ±1 standard deviation from the mean. This formula also shows that the variance of the mean will decrease when the sample size, n, increases when the variables have the same variance.

Correlated variables are treated differently in terms of the variance of their sum. Instead of adding the variances as in uncorrelated variables, the variance of the sum of correlated variables is equal to the sum of their covariances. Because the variables are now related, the covariance must be used to determine how the variables change together. Variance of the mean is calculated differently when the variables are correlated and have equal variance:

$$\mathrm{Var}\left(\overline{X}\right) = \frac{\sigma^2}{n} + \frac{n-1}{n}\rho\sigma^2,$$

where ρ refers to the average correlation of distinct variables.

The variance sum law states that the variance of a difference between two independent variables can be calculated using the sum of their respective variances. For example, a test is given to fifth graders to determine level of mathematics ability based on gender. Therefore, boys and girls are split into two subgroups for testing. The two populations are fifth grade boys and fifth grade girls. For each group, the variance is calculated separately. When the two variables are independent, the variance can be added together to show the variance of the sum. The standard error of that variance can be calculated by taking the square root of the aforementioned sum.

Variance should always be reported in units squared. For example, if data points are measured in centimeters (cm), variance will be reported as cm^2. This reporting can be problematic, depending on the units. It may be reasonable to report cm^2 or kg^2, but, as noted earlier, when the units are people, it does not make sense to report people. When this is the case, standard deviation is often reported instead, by taking the square root of the variance and defining by the same units as the original variable.

Between Groups and Within Groups Variance

Within one set of data exists two types of variance: between-groups variance and within- groups variance. These types of variance are most relevant when conducting an ANOVA, specifically when multiple groups are being compared. Between-groups variance is frequently referred to as explained or systematic variance. Systematic variance is derived from the independent variable or among the groups in the study. Explained variance is accompanied by a direct manipulation from the researcher. Within-groups variance refers specifically to the error variance and comes from within each group. It does not come from the independent variable. This unsystematic variance is due to variations in performance, by factors outside of the manipulation directly from the researcher. This type of variance is not consistent and may fluctuate from one testing administration to the next. Comparing examinees in the same group would result in within-groups variance. Comparing examinees in two different groups would result in between-groups variance. Both of these variances will impact the validity and reliability of tests.

Homogeneity of Variance

The assumption of homogeneity of variance, or homoscedasticity, is used when several groups of examinees are tested for differences in group means. In such a design, it is assumed that all samples come from populations of equal variance. This assumption may also be applicable in a correlational design, where the variance of the dependent variable should be equal at all levels of the independent variable. Any fluctuation of the variance would be a violation of the assumption of homogeneity of variance.

If the spread around the scores are roughly the same from each sample taken, the assumption has been satisfied. If the spread fluctuates greatly from sample to sample, this assumption has been violated and is called heterogeneity of variance or heteroscedasticity. The violation of this assumption is most important when group sizes are unequal.

Homogeneity of variance is most important when using the method of least squares to estimate parameters. Unequal variances will create a bias and an inconsistent estimate of the standard error. Inconsistent standard error will directly affect confidence intervals and significance tests. Levene's test will test the null hypothesis, stating different groups have equal variances. This test is executed through a one-way ANOVA, where a significant *p* value indicates the variances among groups are significantly different. This significant *p* value from Levene's test, therefore, indicates a violation of the assumption of homogeneity of variance. However, a large sample size may result in an incorrect reporting of significance. Small differences among group variances when the sample size is large will generate a significant Levene's test. It is important to note that Levene's test works best with equal group sizes.

Another way to compare groups is to use the variance ratio or Hartley's $F_{\max}$. Hartley's $F_{\max}$ is equal to the ratio of the largest group variance to the smallest group variance. This value is then compared to critical values that are dependent on the group size and the number of variances being evaluated. The variance ratio is expected to be smaller than the critical values specified by Hartley's table to be nonsignificant.

Homogeneity of variance should be evaluated by both Levene's test and the variance ratio, as false significance may by reported by one test.

Variance in Multiple Regression

In a multiple regression design, fit of the regression line is measured by correlation (r) and r^2. In relation to variance, however, R^2 is often reported specifically in reference to explained variance. R^2, often called the coefficient of determination, may be reported as a decimal ($R^2 = .71$) or a percent. When reported as a percentage, it is often stated as the percentage of variance in the independent variable as explained by the model. This value does not make the implication that causation occurs. Using the output from statistical analysis software, or from hand calculations, R^2 can be calculated using the following formula:

$$R^2 = \frac{SS_M}{SS_T},$$

where SS_M refers to the model sum of squares and SS_T refers to the total sum of squares. The square root of this value yields Pearson's correlation coefficient, r. Adjusted R^2 will often be reported in addition to R^2, which removes the bias associated with the latter by diminishing its value.

II. CONCLUSION

Variance is a fundamental statistical tool that measures the spread of data within a distribution, providing critical insights into variability across datasets. By accounting for every data point and its deviation from the mean, variance enables accurate comparisons across groups and lays the groundwork for advanced statistical methods like ANOVA and regression analysis. Its application in experimental design helps distinguish between systematic and unsystematic variability, offering a deeper understanding of data patterns and relationships. As a versatile measure, variance continues to be indispensable in research, enabling precise interpretation and informed decision-making in diverse fields.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, December 2024

REFERENCES

[1]. Fisher, R. A. (1918). The Correlation Between Relatives on the Supposition of Mendelian Inheritance. Transactions of the Royal Society of Edinburgh, 52, 399–433.

[2]. Fisher, R. A. (1925). Statistical Methods for Research Workers. Edinburgh: Oliver and Boyd.

[3]. Snedecor, G. W., & Cochran, W. G. (1989). Statistical Methods (8th ed.). Iowa State University Press.

[4]. Montgomery, D. C. (2019). Design and Analysis of Experiments (10th ed.). Wiley.

[5]. Hogg, R. V., & Tanis, E. A. (2014). Probability and Statistical Inference (9th ed.). Pearson.

[6]. Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). Applied Linear Statistical Models (5th ed.). McGraw-Hill.

[7]. Scheffé, H. (1959). The Analysis of Variance. Wiley.

[8]. Winer, B. J., Brown, D. R., & Michels, K. M. (1991). Statistical Principles in Experimental Design (3rd ed.). McGraw-Hill.

[9]. Searle, S. R. (1971). Linear Models. Wiley.

[10]. Box, G. E. P. (1954). Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems. The Annals of Mathematical Statistics, 25(2), 290–302.

[11]. Dean, A., & Voss, D. (1999). Design and Analysis of Experiments. Springer.

[12]. Zar, J. H. (2010). Biostatistical Analysis (5th ed.). Pearson.

[13]. Steel, R. G. D., & Torrie, J. H. (1980). Principles and Procedures of Statistics: A Biometrical Approach (2nd ed.). McGraw-Hill.

[14]. Rasmussen, P. H. (1986). Analysis of Variance and Covariance: How to Choose and Construct Models for the Life Sciences. Springer-Verlag.

[15]. Christensen, R. (2011). Plane Answers to Complex Questions: The Theory of Linear Models (4th ed.). Springer.

[16]. Oehlert, G. W. (2000). A First Course in Design and Analysis of Experiments. W. H. Freeman.

[17]. Milliken, G. A., & Johnson, D. E. (2009). Analysis of Messy Data Volume 1: Designed Experiments (2nd ed.). CRC Press.

[18]. Cochran, W. G., & Cox, G. M. (1957). Experimental Designs (2nd ed.). Wiley.

[19]. Draper, N. R., & Smith, H. (1998). Applied Regression Analysis (3rd ed.). Wiley.

[20]. Hocking, R. R. (2013). Methods and Applications of Linear Models: Regression and the Analysis of Variance (3rd ed.). Wiley.

[21]. Bock, R. D. (1975). Multivariate Statistical Methods in Behavioral Research. McGraw-Hill.

[22]. Sokal, R. R., & Rohlf, F. J. (1995). Biometry (3rd ed.). W. H. Freeman.

[23]. Kleinbaum, D. G., Kupper, L. L., Nizam, A., & Rosenberg, E. S. (2013). Applied Regression Analysis and Other Multivariable Methods (5th ed.). Cengage Learning.

[24]. McCullagh, P., & Nelder, J. A. (1989). Generalized Linear Models (2nd ed.). Chapman and Hall/CRC.

[25]. Rosenthal, R., & Rosnow, R. L. (2008). Essentials of Behavioral Research: Methods and Data Analysis (3rd ed.). McGraw-Hill.

[26]. Maxwell, S. E., & Delaney, H. D. (2004). Designing Experiments and Analyzing Data: A Model Comparison Perspective (2nd ed.). Routledge.

[27]. Moore, D. S., McCabe, G. P., & Craig, B. (2012). Introduction to the Practice of Statistics (8th ed.). W. H. Freeman.

[28]. Altman, D. G. (1991). Practical Statistics for Medical Research. Chapman and Hall/CRC.

[29]. Field, A. P. (2018). Discovering Statistics Using IBM SPSS Statistics (5th ed.). SAGE Publications.

[30]. Howell, D. C. (2012). Statistical Methods for Psychology (8th ed.). Wadsworth.

[31]. Neter, J., Wasserman, W., & Kutner, M. H. (1990). Applied Linear Statistical Models. Irwin.

[32]. Hays, W. L. (1994). Statistics (5th ed.). Harcourt Brace College Publishers.

[33]. Agresti, A., & Franklin, C. (2018). Statistics: The Art and Science of Learning from Data (4th ed.). Pearson.

[34]. Venables, W. N., & Ripley, B. D. (2002). Modern Applied Statistics with S (4th ed.). Springer.

[35]. Kirk, R. E. (2012). Experimental Design: Procedures for the Behavioral Sciences (4th ed.). SAGE Publications.