Invited Review / Davetli Derleme
doi: 10.18678/dtfd.1604460

Duzce Med J, 2024;26(3):183-190
Düzce Tıp Fak Derg, 2024;26(3):183-190

Conscious and Correct Use of Biostatistical Methods in Medical Researches: From Planning to Reporting the Results - Part II

Tıbbi Araştırmalarda Biyoistatistiksel Yöntemlerin Bilinçli ve Doğru Kullanımı: Planlamadan Sonuçların Raporlanmasına - Bölüm II

Mehmet Ali SUNGUR[1]
0000-0001-5380-0819
Erdem KARABULUT[2]
0000-0002-7811-8215

[1]Department of Biostatistics and Medical Informatics, Düzce University Faculty of Medicine, Düzce, Türkiye
[2]Department of Biostatistics, Hacettepe University School of Medicine, Ankara, Türkiye
Chair of the Biostatistics Association

Corresponding Author
Sorumlu Yazar
Erdem KARABULUT
ekarabul@hacettepe.edu.tr

Received / Geliş Tarihi : 04.11.2024
Accepted / Kabul Tarihi : 15.12.2024
Available Online / Çevrimiçi Yayın Tarihi : 20.12.2024

ABSTRACT

In this part of the review, statistical tests utilized to examine hypotheses regarding population parameters on a representative sample, which forms the fundamentals of inferential statistics are discussed. The selection of an appropriate statistical test by verifying its assumptions and interpreting the results objectively is crucial for obtaining accurate conclusions. Understanding the terms related to type I and type II errors, p-value, power of the study, effect size, and confidence interval will contribute to the correct interpretation of both the results obtained from statistical tests in scientific research and the findings of articles read from the literature. In addition to univariate tests, the three most commonly employed multiple regression models are also addressed to control for the effect of potential confounding factors and other independent variables utilized in the study. Statistical computing has become much more accessible in recent times, with researchers having access to freeware packages or web applications to perform basic and advanced statistical analyses. Researchers frequently focus on the calculation of statistical tests used in data analysis, whereas, understanding the rationale behind statistical methods should be the primary goal. Therefore, this review emphasizes the logic of selecting appropriate statistical methods and interpreting the results rather than mathematical calculations. It is essential to recognize that biostatistical principles should be considered not only in the data analysis phase but also in all phases of research, from planning to report writing. It should be note that, no statistical analysis method can correct erroneous data obtained from a poorly designed study.
Keywords: Hypothesis tests; regression model; ANOVA; parametric tests; nonparametric tests; p-value.

ÖZ

Derlemenin bu bölümünde, çıkarımsal istatistiğin temelini oluşturan, popülasyonu temsil eden bir örneklem üzerinde popülasyon parametrelerine ilişkin hipotezleri incelemekte kullanılan istatistiksel testler ele alınmıştır. Varsayımlarının karşılanıp karşılanmadığı kontrol edilerek uygun istatistiksel testin seçilmesi ve bulguların yansız bir şekilde yorumlanması doğru sonuçlara ulaşılmasında çok önemlidir. Tip I ve tip II hatalar, p-değeri, çalışmanın gücü, etki büyüklüğü ve güven aralığı ile ilgili kavramların ne anlama geldiğini anlamak, hem bilimsel araştırmalarda istatistiksel testlerden elde edilen sonuçların hem de literatürden okunan makalelerin bulgularının doğru yorumlanmasına katkı sağlayacaktır. Tek değişkenli testlerin yanı sıra, olası etki karıştırıcı faktörlerin ve çalışmada ele alınan diğer bağımsız değişkenlerin etkisini kontrol etmek amacıyla en yaygın olarak kullanılan üç çoklu regresyon modeline de değinilmiştir. Araştırmacılar, temel ve ileri düzey istatistiksel analizleri yapabilecekleri ücretsiz yazılımlara ya da web uygulamalarına erişebildiklerinden, istatistiksel hesaplamaların yapılması günümüzde çok daha erişilebilir hale gelmiştir. Araştırmacılar daha çok veri analizinde kullanılan istatistiksel testlerin nasıl hesaplanacağı üzerinde durmaktadır, oysaki istatistiksel yöntemlerin arkasındaki mantığı anlamak birincil hedef olmalıdır. Bu nedenle, bu derlemede matematiksel hesaplamalardan ziyade, uygun istatistiksel yöntemlerin seçilmesi ve bulguların yorumlanması mantığı üzerinde durulmuştur. Biyoistatistik prensiplerin sadece veri analizi aşamasında değil, planlamadan rapor yazımına kadar araştırmanın tüm aşamalarında dikkate alınması gerektiğinin farkında olunması önemlidir. Unutulmamalıdır ki, iyi düzenlenmemiş bir araştırmadan elde edilen hatalı verileri düzeltecek bir istatistiksel analiz yöntemi bulunmamaktadır.
Anahtar kelimeler: Hipotez testleri; regresyon modeli; ANOVA; parametrik testler; parametrik olmayan testler; p-değeri.

INTRODUCTION

Studies that incorporate appropriate biostatistical support, from design to analysis and reporting, are the most effective experimental studies; therefore, biostatistics plays a critical role in health research. The most valuable insights into complex causal relationships can be obtained through the application of various statistical methods. As a result, as in other research fields, researchers increasingly need to use biostatistics in their study of medical sciences.

The impact of statistical sciences on medical and biological sciences has increased rapidly over the past few decades. Physicians must have knowledge of statistical principles and techniques, as the information they use in clinical decision-making is always based on the results of statistical analyses. However, conclusions drawn from statistical evidence can be incorrect or misleading, and without sufficient knowledge and understanding of statistics, physicians may not be able to make the most appropriate decisions (1-3).

When research is conducted on all subjects in the population, the population parameters are calculated without error; however, as is known, research is usually conducted on a small sample representing the population of all subjects instead of the population. When the sample is drawn to represent the population, inferences are made about the population with the help of statistics calculated from the sample, and the obtained result is generalized to the population. The sample can only be representative of the population if the sample is drawn using probability sampling methods and the sample size is sufficient.

If the sample is not representative of the population, misleading results will be obtained and this cannot be corrected by statistical procedures. On the other hand, even in a well-designed study, incorrect results may be obtained due to random variation in the sample. Thus, results from a single sample have statistical uncertainty, which is strongly related to the sample size (4).

In this part, hypothesis testing and statistical modeling will be discussed. Firstly, how to determine the appropriate test from univariate statistical tests is emphasized, and then how to select appropriate multiple analyses to eliminate the effect of other possible confounders is discussed. In addition, concepts such as standard error, confidence interval, type I error, type II error, p-value, power, etc. are also explained.

Standard Error

The standard error is a measure of the uncertainty of a test statistic and is an estimate of how much the value of the test statistic varies from sample to sample. The standard error is also defined as the difference between the population parameter (e.g. population mean μ) and the statistic calculated from the sample (e.g. sample mean $\bar{x}$) and is also referred to as sampling error. The standard error is obtained by taking the standard deviation of the sampling distribution of the statistics calculated from all possible samples of size n that can be drawn from the population.

The formula for estimating standard error is

$$Standard\ Error = Standard\ Deviation / \sqrt{Sample\ Size}$$

The equation demonstrates that the standard error decreases as the sample size increases. The standard error is used for hypothesis testing, confidence interval calculations, etc.

Confidence Interval

In inferential statistics, there are two different approaches to estimating population parameters: point estimation and interval estimation. In point estimation, a statistic such as mean, standard deviation, ratio, correlation coefficient, etc. calculated from the sample is used to estimate the population parameter. The probability that point estimation will precisely determine the population parameter is quite low. On the other hand, an interval estimation is a range of values that most likely contains the parameter being estimated with a given confidence level. Therefore, interval estimation is preferred in inferential statistics.

In order to calculate confidence intervals, statistics obtained from the sample, standard error, and two-tailed table statistics values obtained from the relevant theoretical distribution for the specified confidence level are needed. In statistics, a 95% confidence level is most commonly used. Confidence intervals for the population parameter are obtained using the following general notation.

$$Statistic \pm Table\ Value \times Standard\ Error$$

For example, the 95% confidence interval for an unknown population mean is given as follows:

$$\bar{x} - t_{1-\alpha/2} \times Standard\ Error \leq \mu \leq \bar{x} + t_{1-\alpha/2} \times Standard\ Error$$

Confidence intervals are directly related to hypothesis testing and sample size (or power) calculation. If, for example, the confidence intervals for the mean difference between two independent groups include 0, it indicates that there is no significant difference between the group means, whereas if they do not include 0, it indicates a significant difference between the group means. In addition, the width of the confidence intervals provides information about the sample size of the study. If the sample size in the study is small, this will cause the standard error to increase and the confidence interval to widen. Although confidence intervals are mostly symmetric, non-symmetric confidence intervals are obtained for some statistics. For example, the confidence intervals of odds ratio and relative risk used to calculate risk in health sciences are not symmetric.

HYPOTHESIS TESTING

Hypotheses are propositions concerning the population parameters from which samples are drawn. One of the objectives of inferential statistics is to test hypotheses about unknown phenomena within a population. Hypothesis testing aims to make inferences about the population parameters of interest using data obtained from a representative sample of the population.

Hypothesis testing plays a crucial role in evidence-based medicine, wherein clinical decisions are based on research findings. This contributes to increasing the accuracy of clinical decisions by ensuring that conclusions are derived objectively from the research data.

Due to the utilization of data from randomly selected samples in hypothesis testing, each result inherently carries the potential for error (type I error or type II error). Consequently, these tests cannot be employed to verify any phenomenon definitively. For instance, depending on the sample selected, one study may conclude that influenza vaccination is efficacious in disease prevention, while

another study may determine it is not. Therefore, hypothesis testing cannot be utilized to conclusively prove that influenza vaccination prevents (or does not prevent) the disease, but can be employed to support this hypothesis. However, when multiple studies on a specific topic yield similar results, they are generally considered to be valid. As a result, the findings of systematic reviews and meta-analyses are utilized in the development of diagnostic and treatment protocols, rather than relying on the results of a single study.

Parametric and Nonparametric Hypothesis Tests

Hypothesis tests are divided into parametric and nonparametric (or distribution-free) tests. Which test will be applied in the studies is decided according to whether the parametric test assumptions are met or not.

Methods that require assumptions about the distribution of variables in the population are called parametric methods. For example, in the independent samples t-test, it is assumed that the observations in the two sample groups are drawn from normally distributed populations. All parametric tests of continuous numerical variables require the assumption of a normal distribution.

Methods that do not assume the distribution of variables in the population are called nonparametric methods. Instead of using the original measurements in the data set, these methods use the rank or sign of data. When the data type is ordinal, parametric methods should not be used, considering the data as continuous numerical data.

When parametric test assumptions are satisfied, parametric tests are known to have higher power than their alternative nonparametric tests. Therefore, when the assumptions are not satisfied in studies with a large number of observations, appropriate transformations are applied to the data before using nonparametric methods. If the assumptions are satisfied after the transformation, parametric tests are applied, and if not, nonparametric methods are applied (5,6).

Formulation of Hypotheses and Steps of Hypothesis Testing

The first stage of hypothesis testing is the formulation of statistical hypotheses in accordance with the aim of the study. There are two hypotheses in hypothesis testing; the null and alternative hypotheses denoted by H_0 and H_1. The H_0 hypothesis is called the null hypothesis because it is always stated that there is no difference between groups or there is no relationship between variables, etc. The H_1 hypothesis contains inequality and can be one or two-tailed. The conclusion drawn from the hypothesis test depends on whether the hypothesis is one-tailed or two-tailed. A null hypothesis that can not be rejected in a two-tailed test can be rejected in a one-tailed test. Therefore, it is important for researchers to set up their hypotheses appropriately at the beginning of the study. If there is an expectation that the parameter related to one group will be larger or smaller than the other, a one-tailed hypothesis should be formed, and if it is only intended to examine whether there is a difference, a two-tailed hypothesis should be formed.

In the second stage of hypothesis testing, the appropriate test statistic is decided and calculations are performed. The appropriate hypothesis test is decided according to the data type of the variable of interest (quantitative or qualitative), the number of groups, whether the groups are dependent or independent, and whether the parametric test assumptions are met. Figure 1, Figure 2, and Figure 3 illustrate which statistical test is appropriate in which situation. After deciding on the appropriate test, test statistics are calculated with different formulas for each hypothesis test (5,7-9).

In the last stage, it is assessed whether the acquired result is significant or not using the predetermined significance level. At the end of each test, a test statistic is calculated that fits a theoretical statistical distribution such as z, t, F, chi-square, U, etc. Based on the value of the calculated test statistic and the corresponding critical value or p-value derived from the theoretical statistical distribution, the null hypothesis H_0 is either rejected or not rejected (9).

When statistically significant differences are observed in One-way ANOVA, Kruskal-Wallis, ANOVA in repeated measures, or Friedman tests, post hoc (pairwise or multiple comparisons) tests should be performed to determine the group(s) that cause the difference.

In univariate hypothesis testing, only the relationship or difference between the dependent variable and the independent variable is examined. The effect of potential confounding variables is not considered. Consequently, interpreting findings based solely on the results of univariate analyses may lead to erroneous conclusions. Interpreting the results of appropriate regression models that incorporate variables found to be significant in univariate analyses, along with confounding variables, will yield more accurate results.

Type I error (α) and Type II error (β)

There are two types of errors in hypothesis testing;

• *Type I error (α):* reject a null hypothesis H_0 that is actually true in the population

• *Type II error (β):* fails to reject a null hypothesis H_0 that is actually false in the population

The values α and β are probabilities and take values between 0 and 1. The difference of these probabilities from 1 is called the confidence level (1-α) and the power of the test (1-β), respectively. In general, a maximum value of 0.05 for α, and a minimum value of 0.20 for β (or 0.80 for power, 1-β) are used in research.

Although type I and type II errors can never be completely eliminated, the researcher can reduce their likelihood by increasing the sample size (10). Type II error usually occurs when the statistical power of a test is low. In the hypothesis testing process, since all analyses are performed under the assumption that the null hypothesis H_0 is true, the probability of committing a type I error is calculated. The researcher has no control over type II error (β) in the hypothesis testing process. Type II error is taken into account when calculating the sample size at the beginning of the study.

p-value

Since all hypothesis tests are performed under the assumption that the null hypothesis, H_0, is true, we are interested in the probability of committing a type I error. While the α value is determined by the researcher at the beginning of the study, the p-value is obtained based on the test statistic calculated at the end of the study. The p-value indicates the probability of falsely rejecting the true null hypothesis. In other words, the p-value indicates the probability that the observed outcomes were obtained by chance, assuming the null hypothesis H_0 is true. Since a value of 0.05 is generally considered a type I error (α), the H_0 hypothesis is rejected when the p-value obtained from the test is less than 0.05.

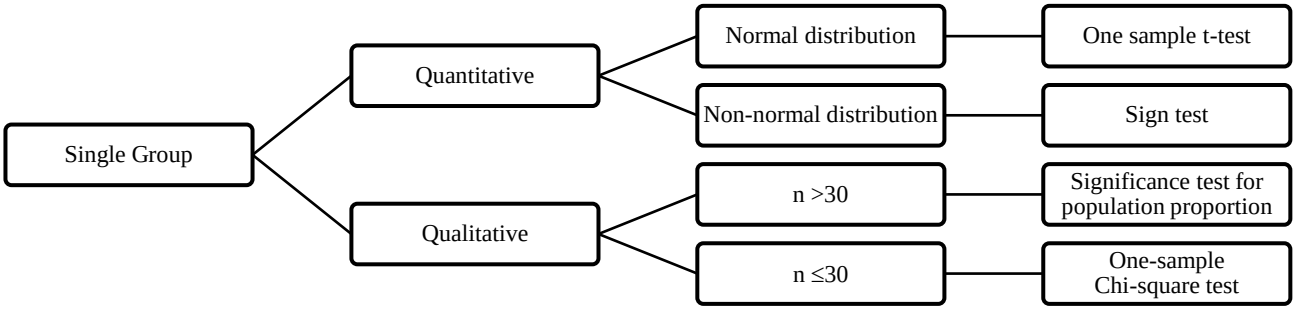

Figure 1. Flowchart for choosing appropriate statistical test for one group

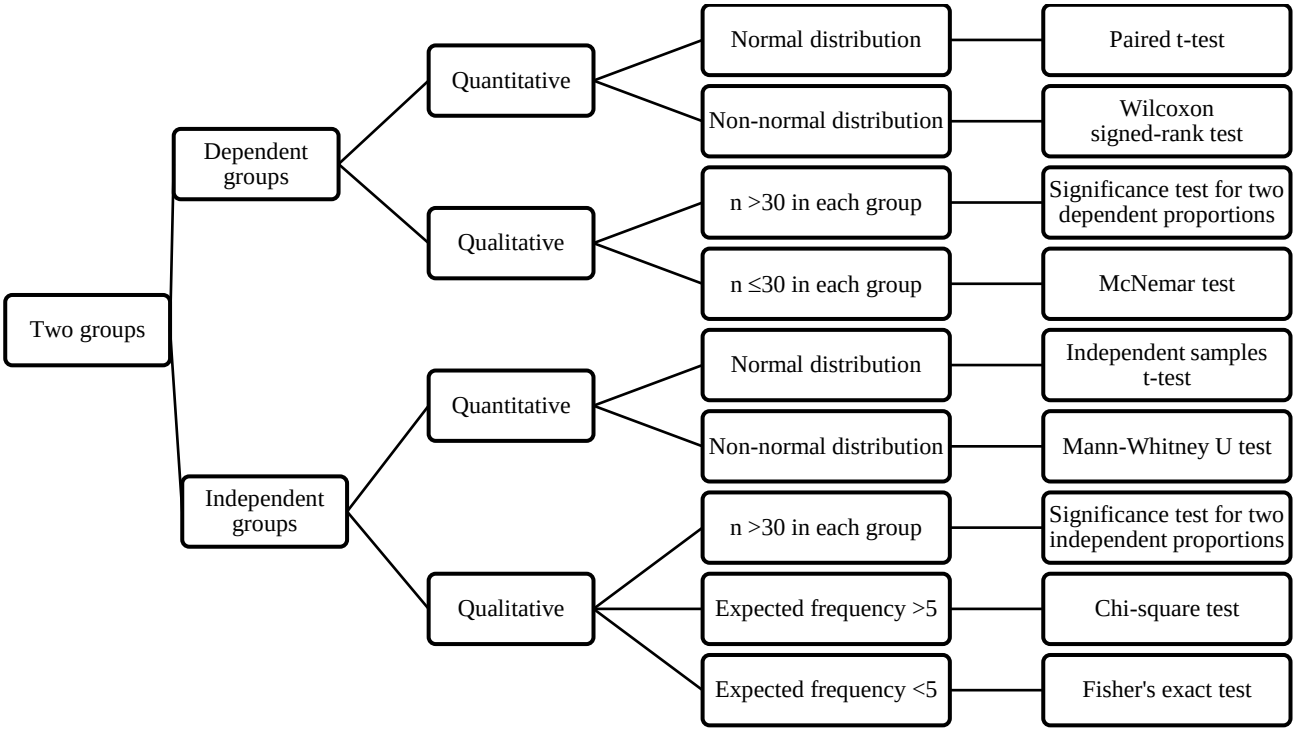

Figure 2. Flowchart for choosing appropriate statistical test for two groups

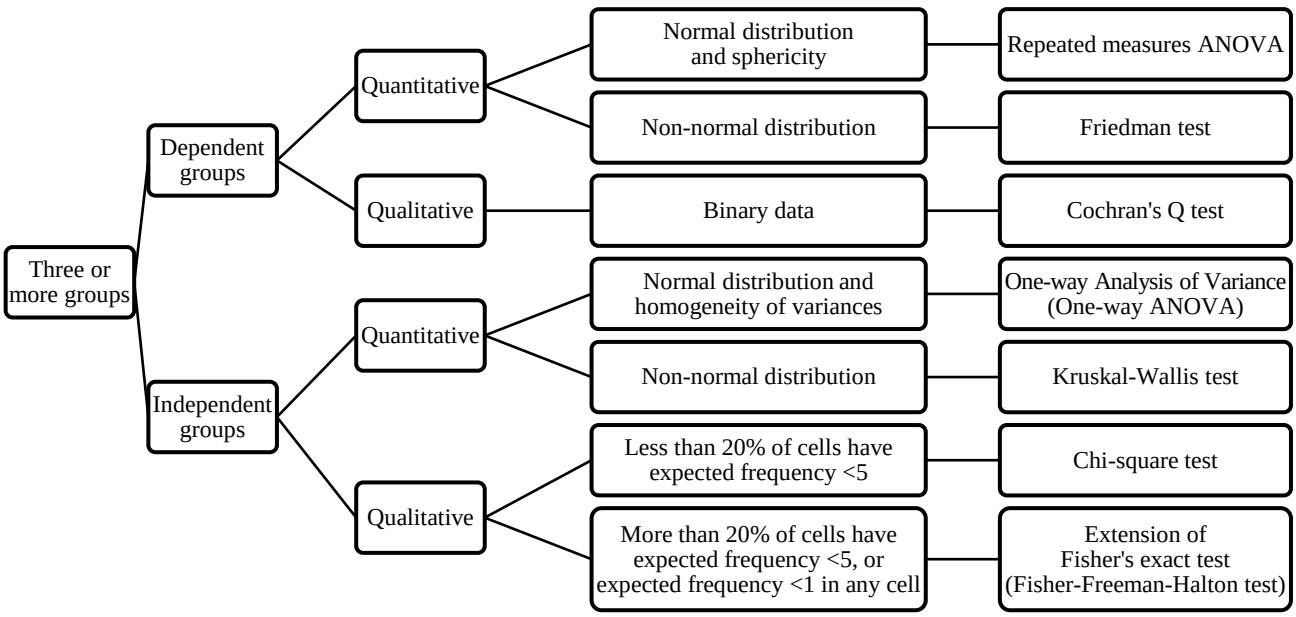

Figure 3. Flowchart for choosing appropriate statistical test for three or more groups

When reporting the p-value obtained from the test, the exact p-value (e.g. p=0.180, p=0.018, p=0.108, p=0.810) should be written in detail in the tables and the text. However, when the obtained p-value is small, it should be reported as p<0.001 instead of p=0.000.

Even small differences can be found to be significant when the number of observations in the study is too large. In this case, researchers should determine whether the statistically significant difference is also clinically significant. On the other hand, when the number of observations in the study is small, the difference or relationship that actually exists may not be determined since not enough samples are included in the study. Here, the power of the test will decrease as the probability of making a β-type error will increase (11).

Effect Size

One of the issues with hypothesis tests is that significance does not indicate the clinical or practical significance of an effect. To address this drawback, the size of the effect being investigated should be measured in a standardized method.

Effect size is a statistical measure of the magnitude of the difference between groups or association between variables observed in a study. In addition to the statistical significance decided by the p-value, it allows the clinical significance of an outcome to be assessed. While the p-value indicates whether an effect exists, the effect size indicates how large or practically important this effect is (5,9,12).

The most commonly used type of effect size in the literature are Cohen's d, Pearson correlation coefficient r, partial eta-square, and odds ratio. In recent years, it has become common to present effect size and/or confidence intervals in tables in scientific articles in the field of health sciences (12).

TYPES OF ANOVA MODELS

An ANalysis Of VAriance (ANOVA) is an appropriate statistical technique for determining the differences between independent groups on a continuous measurement. Different ANOVA models are used depending on the purpose of the study and the research design. The most commonly used ANOVA model is the one-way ANOVA (one-way between-subjects ANOVA) model, which is used to examine the effect of a factor with three or more categories on a quantitative dependent variable (13).

One-way ANOVA requires the assumption of normal distribution and homogeneity of variances. If the assumptions of normal distribution and homogeneity of variances are not met, the nonparametric Kruskal-Wallis test and Welch's ANOVA are used, respectively.

Factorial between-factor ANOVA is used to examine the effect of two or more factors on a quantitative dependent variable. When groups are created based on multiple dimensions, variations in the means can be attributed to multiple factors. For example, a three-way ANOVA can be used to analyze how treatment type, age group, and gender influence the length of hospital stay. When conducting a factorial ANOVA to examine the effects of two or more factors, it is important to include interaction terms along with the main effects. The use of one-way or two-way ANOVA models is common because the increase in number of factors makes it hard to interpret the model.

In certain designs, the tested means come from the same subjects measured at different occasions, rather than from independent groups of subjects. A within-subjects ANOVA, also known as a repeated measures ANOVA, is commonly applied in pretest-posttest designs but is not restricted to just two time points. This analysis can be used to assess differences across two or more time periods. In repeated measures analysis of variance, in the presence of independent groups, both between-group and within-group differences can be examined simultaneously. When the group × time interaction is statistically significant in this analysis, the main effects of group and time should not be interpreted directly. The change of each group over time, and the difference between groups at each time point should be examined separately and p-values are corrected for multiple comparisons (14).

An analysis of covariance (ANCOVA) is suitable for evaluating the differences in terms of a continuous dependent variable among groups while accounting for the effect of covariates. Covariates have an effect on the dependent variable. There are two ways to control their effects, either by selecting samples during the research design to make the groups similar in terms of covariates, or by statistical methods during the data analysis process. ANCOVA is a type of ANOVA that enables the effects of covariates to be controlled through the use of regression analysis. In ANCOVA, in addition to the assumptions of the analysis of variance, there are assumptions that the independent variable and the covariate are independent, the relationship between the covariate and the dependent variable is linear, and the slope of the regression lines is similar in all groups (12,13).

Multivariate analysis of variance (MANOVA) is a statistical technique used to examine the effect of one or more factors on two or more numerical dependent variables. In other words, by comparing the mean vectors between groups, it is investigated whether there is a difference in terms of two or more dependent variables. The general assumptions of MANOVA are that the variance-covariance matrices of all groups are homogeneous and that the variables in each group have a multivariate normal distribution (13,14).

CORRELATION ANALYSIS

Correlation analysis examines whether there is a relationship between two variables, and if so, the strength and direction of the relationship. Correlation coefficients generally range between -1 and +1; as they approach -1 or +1, the strength of the relationship increases, while as they approach 0, the strength of the relationship decreases. If the sign of the coefficient is negative, it indicates that as the value of one variable increases, the value of the other variable decreases. The significance of the correlation value obtained is tested and if the result is significant, the strength of the relationship is interpreted. Otherwise, it is stated that there is no relationship between the variables.

Depending on the type of data and whether the assumptions are met, different correlation coefficients are calculated to determine the relationship between variables. Pearson and Spearman rank-order correlation coefficients are commonly used to examine the relationship between two quantitative variables. Pearson correlation coefficient is used if the relationship between variables is linear and

the assumption of normal distribution is met. There are also correlation coefficients for qualitative variables such as Phi, Cramer's V, Somer's d, etc. However, these correlation coefficients are rarely presented in studies since the chi-square test is often used to examine the relationship between qualitative variables (9).

In order to examine the relationship between numerical variables, the first step is to draw a scatter plot. The linearity of the relationship between the variables and the presence of outliers in the data that would cause misinterpretation of the relationship should be checked. A correlation coefficient does not indicate a cause-and-effect relationship between two variables. It only indicates whether the variables change together in the same direction or the opposite direction.

The relationship between more than two variables can also be examined. For example, the relationship between one dependent and two or more independent variables can be examined by multiple correlation coefficients, and the relationship between more than two sets of dependent and independent numerical variables can be examined by canonical correlation.

Due to indirect relationships, a significant correlation can be found between two variables that are not actually correlated. In this case, the partial correlation coefficient is used to examine whether there is a relationship between the variables by controlling for the effect of the covariate that has an effect on both variables.

REGRESSION MODELS

Regression analysis is a useful method for determining the relationship between a dependent variable and a group of independent variables. The overall goal of regression analysis is to predict the value of an outcome variable of interest using variables that are simple to measure, provide minimal risk to participants, can be collected in a shorter period or at a lower cost, etc. The variable whose value is being estimated in the model is called the dependent variable (response or outcome variable) and the variables used to explain the dependent variable are called independent variables (predictors or explanatory variables).

In addition to making predictions, regression analysis is also used to determine the important factors affecting the dependent variable or to determine which variables have a greater effect on the dependent variable and to make adjustments for covariates.

Depending on the type of dependent variable, the nature of the relationship between variables, the presence of censored observations, etc., many different regression analysis methods have been developed: linear regression, non-linear regression, logistic regression, Poisson regression, Cox regression, etc.

Linear Regression

Linear regression analysis is the most widely used regression analysis method in the literature. In order to use linear regression, the dependent variable must be a continuous numerical data type and the relationship between the dependent and independent variables should be linear.

The dependent variable should always be a continuous numerical variable with a normal distribution. The type of independent variables can be either quantitative (continuous numeric or discrete numeric) or qualitative (ordinal or nominal). Nominal qualitative variables are included directly in the model if they have only two categories (e.g. yes-no), but if they have more than two categories, they should be included in the model after being coded as a dummy variable. In dummy variable coding, the nominal variable is converted into a set of binary variables. Ordinal variables can be included in the model as numerical variables if there is an equal distance between categories, otherwise, they can be included as nominal variables (after coding as dummy variables).

The regression coefficients obtained for independent variables of the continuous numeric variable type indicate the average change in the dependent variable caused by a 1-unit change in the independent variable. When the independent variable is qualitative data, it shows how much the other categories differ on average from the reference category.

In articles in the field of health sciences, only the regression coefficients, confidence intervals, and their significance are commonly presented. The results on whether the assumptions of the model are met and the goodness of fit of the model are not sufficiently presented. Therefore, to evaluate the performance of the model, statistics such as mean squared error, root mean squared error, information criterion, or coefficient of determination (R^2) should also be given. Furthermore, in multiple linear regression, it is essential to assess the normality of residuals, identify influential and outlier observations, and examine for heteroscedasticity and multicollinearity.

The coefficient of determination (R^2) is often used to assess model fit. The coefficient of determination takes values between 0 and 1. It shows the percentage of the change in the dependent variable explained by the independent variables. If the regression model is to be used for prediction, a high R^2 value is required. In addition to assessing the overall fit of the model using measures such as the F test and R^2, the data should be examined for outliers and influential observations. To identify outliers, scatter plots of the standardized residuals and the estimated values (or the dependent variable) are drawn. Observations with values outside the ±2 limits in the scatter plot are usually identified as outlier observations. If the residuals are randomly scattered around 0 in this scatter plot, it indicates that the linear model is appropriate and there is no problem of heteroscedasticity. Cook's distance is used to identify influential observations that have an effect on the regression coefficients. Observations with a Cook's distance value above 1 are observations that are highly influential on the regression coefficients.

A multicollinearity problem occurs when a high degree of correlation exists among the independent variables in a multiple linear regression model. The variance inflation factor (VIF) and tolerance statistics can be utilized to determine the presence of a multicollinearity issue. A multicollinearity problem is indicated when the VIF value for the variables in the model exceeds 10 or the tolerance value is less than 0.10. This problem can be overcome by choosing one of the highly correlated variables in the model or by using special regression methods such as Ridge regression or lasso regression.

The primary objective in multiple linear regression and other multivariate analyses is to construct a model with the minimum number of variables necessary. In instances

where a substantial number of independent variables are present, univariate tests may be utilized initially to eliminate variables that do not demonstrate a significant relationship with the outcome variable. Subsequently, a multiple linear regression model can be fitted using the independent variables that have been identified as potentially relevant. In practice, variable selection is typically conducted through stepwise methods, such as stepwise, forward, or backward selection, using specific statistical criteria. The hierarchical regression method is another approach. In this method, the researcher determines the order in which variables are introduced into the model (5).

Logistic Regression

The goal of a logistic regression analysis is the same as for any other regression model used in statistics: to find the best fitting, most parsimonious, clinically interpretable model to describe the relationship between a dependent variable and a set of independent variables. Logistic regression is used when the dependent variable is a qualitative data type. Logistic regression analysis is primarily concerned with estimating the probability of the occurrence of a specified event, rather than predicting the value of the dependent variable.

The distinction between logistic and linear regression can be seen in both the form of the model and the assumptions used. In linear regression, the residuals are normally distributed, while in logistic regression the residuals follow a binomial distribution (15).

The reason why logistic regression is widely used is that it requires fewer assumptions than linear regression. Logistic regression only assumes that the relationship between the logit and the numerical independent variables is linear. In addition, when the obtained regression coefficients are taken as exponential, the interpretation of the results becomes easier since the odds ratio commonly used in health sciences is obtained.

Binary logistic regression analysis is the most commonly used model. In this model, the dependent variable should be a qualitative variable with two categories, such as "present-absent" or "diseased-healthy". Both quantitative and qualitative data types can be used as independent variables in the model.

In evaluating the model fit, statistics such as the Pearson chi-square test, Hosmer-Lemeshow test, correct classification rates, and area under the receiver operating characteristic curve are used. Correct classification rates, which show the fit between the values predicted from the model and the observed values, are widely used (14).

Cox Regression

Cox regression or Cox proportional hazards model is a regression method used in studies where the outcome of interest is not observed in all individuals in the study, in other words, in studies with censored observations.

In survival studies, individuals are monitored from a defined starting point until the occurrence of the event of interest, and factors contributing to the event are investigated. However, researchers cannot always follow all patients until the event of interest occurs. Censored observations occur because some patients are lost to follow-up, patients die from another cause, the length of follow-up is different for all patients when the study is terminated, etc.

The hazard function plays an important role in Cox regression. The hazard function gives the probability that someone who lived until the beginning of the time interval of interest will die in the next time interval.

After the Cox regression model is fitted, it is essential to examine whether the proportional risk assumption is met. For this purpose, the time × covariate interaction term is added to the model and its significance is tested. If the interaction term is significant, it indicates the proportional risk assumption is not met. Another approach to examining the assumption is plotting the ln(-ln) (survival curves). In the graph, the survival curves corresponding to the categories of the qualitative variable being parallel over time indicate that the proportional hazards assumption is satisfied (16,17).

Direct interpretation of beta coefficients in Cox regression is complicated because they correspond to changes in the log-hazard rate. The results are interpreted with hazard ratios obtained by taking the exponential of the regression coefficients. A hazard ratio quantifies how many times the hazard rate increases (or decreases) when the associated predictor variable is incremented by one unit.

CONCLUSION

The increase in the number of commercial or freeware statistical software, as well as the ability to perform analyses through web platforms, has made it easier for researchers to access these tools. Researchers who are not aware of the conditions in which statistical methods should be used and how to interpret the results are likely to make errors. Therefore, choosing the appropriate statistical methods and interpreting the results correctly is crucial when analyzing data obtained from well-designed research. Even if univariate statistical methods are used appropriately, the effects of possible confounding factors are not considered, and multiple or multivariate statistical analyses should be performed to achieve unbiased and accurate results.

Ethics Committee Approval: Since our study was not an experimental study including human or animal subject, ethics committee approval was not required.

Conflict of Interest: None declared by the authors.

Financial Disclosure: None declared by the authors.

Acknowledgments: None declared by the authors.

Author Contributions: Idea/Concept: MAS; Design: EK; Data Collection/Processing: MAS, EK; Analysis/Interpretation: EK; Literature Review: MAS, EK; Drafting/Writing: MAS, EK; Critical Review: EK.

Peer Review: Externally peer reviewed.

Editorial Board Note: Mehmet Ali Sungur is the Editor of Duzce Medical Journal, however he did not take place at any stage on the decision of this article.

REFERENCES

1. Sungur MA. Importance of biostatistics in medical researches. Duzce Med J. 2021;23(3):232-3.
2. Karahan S, Karaağaoğlu AE. Development of biostatistics: from past to future. Duzce Med J. 2021;23(3):234-8.
3. Karabulut E. Conscious and correct use of biostatistical methods in medical researches: from planning to reporting the results-I. Duzce Med J. 2022;24(1):1-6.
4. Altman DG, Machin D, Bryant TN, Gardner MJ. Statistics with confidence: Confidence intervals and statistical guidelines. 2nd ed. London: BMJ Books; 2000.
5. Alpar R. Spor, sağlık ve eğitim bilimlerinden örneklerle uygulamalı istatistik ve geçerlik-güvenirlik. 6th ed. Ankara: Detay; 2020. Turkish.
6. Grech V, Calleja N. WASP (Write a Scientific Paper): Parametric vs. non-parametric tests. Early Hum Dev. 2018;123:48-9.
7. Dawson B, Trapp RG. Basic & clinical biostatistics. 4th ed. New York: McGraw-Hill; 2004.
8. Daniel WW, Cross CL. Biostatistics: A foundation for analysis in the health sciences. 11th ed. Hoboken, NJ: Wiley; 2019.
9. Özdemir P, Karabulut E, Alpar R. Biyoistatistik. In: Günalp S, editor. Temel Kadın Hastalıkları ve Doğum Bilgisi. 4th ed. Ankara: Güneş Tıp Kitapevleri; 2024. Turkish.
10. Banerjee A, Chitnis UB, Jadhav SL, Bhawalkar JS, Chaudhury S. Hypothesis testing, type I and type II errors. Ind Psychiatry J. 2009;18(2):127-31.
11. Cohen HW. P values: use and misuse in medical literature. Am J Hypertens. 2011;24(1):18-23.
12. Field A. Discovering statistics using IBM SPSS statistics. 6th ed. London: SAGE; 2024.
13. Tabachnick BG, Fidell LS. Using multivariate statistics. 6th ed. Boston: Pearson; 2013.
14. Alpar R. Uygulamalı çok değişkenli istatistiksel yöntemler. 6th ed. Ankara: Detay; 2021. Turkish.
15. Hosmer DW, Lemeshow S, Sturdivant RX. Applied logistic regression. 3rd ed. Hoboken, NJ: Wiley; 2013.
16. Hosmer DW, Lemeshow S. Applied survival analysis: Regression modeling of time-to-event data. New York: John Wiley & Sons; 1999.
17. Kleinbaum DG, Klein M. Survival analysis: A self-learning text, 2nd ed. New York: Springer; 2005.