# BIOSTATISTICAL EVALUATONS

Editör: Prof.Dr. Sıddık KESKİN

yaz
yayınları

# Biostatistical Evaluations

**Editör**

Prof.Dr. Sıddık KESKİN

yaz
yayınları

**2025**

yaz
yayınları

# Biostatistical Evaluations

Editör: Prof.Dr. Sıddık KESKİN

**© YAZ Yayınları**

Bu kitabın her türlü yayın hakkı Yaz Yayınları’na aittir, tüm hakları saklıdır. Kitabın tamamı ya da bir kısmı 5846 sayılı Kanun’un hükümlerine göre, kitabı yayınlayan firmanın önceden izni alınmaksızın elektronik, mekanik, fotokopi ya da herhangi bir kayıt sistemiyle çoğaltılamaz, yayınlanamaz, depolanamaz.

E_ISBN 978-625-8508-70-3

Ekim 2025 – Afyonkarahisar

Dizgi/Mizanpaj: YAZ Yayınları

Kapak Tasarım: YAZ Yayınları

YAZ Yayınları. Yayıncı Sertifika No: 73086

M.İhtisas OSB Mah. 4A Cad. No:3/3
İscehisar/AFYONKARAHİSAR

www.yazyayinlari.com

yazyayinlari@gmail.com

info@yazyayinlari.com

# İÇİNDEKİLER

*"Bu kitapta yer alan bölümlerde kullanılan kaynakların, görüşlerin, bulguların, sonuçların, tablo, şekil, resim ve her türlü içeriğin sorumluluğu yazar veya yazarlarına ait olup ulusal ve uluslararası telif haklarına konu olabilecek mali ve hukuki sorumluluk da yazarlara aittir."*

# BIOSTATISTICAL APPROACHES IN PSYCHOMETRIC SCALE DEVELOPMENT AND VALIDATION PROCESSES

**Hakan ÖZTÜRK[1]**

## 1. INTRODUCTION

Psychometric scales are tools that enable abstract concepts to be made measurable. Many psychosocial constructs such as depression, anxiety, stress, quality of life, pain perception, or sleep patterns cannot be directly observed; measuring these constructs is only possible with valid and reliable scales. In health sciences, scales play a critical role in diagnosis and screening processes, in evaluating treatment effectiveness, and in epidemiological research (DeVellis & Thorpe, 2021).

Validity and reliability are two fundamental characteristics that determine the scientific value of scales. Validity refers to whether the scale actually measures the construct it intends to measure, while reliability indicates the consistency and reproducibility of the measurement. A reliable but invalid scale does not provide accurate information, just as a valid but low-reliability scale does not produce consistent results. Therefore, both characteristics must be present (Carmines & Zeller, 1979; Sullivan & Artino, 2011).

Biostatistics is not merely a technical tool in the scale development process, but also a methodological guide. Statistical analyses are used at every stage, from the creation of the item pool

[1] Araş. Gör. Dr., Aydın Adnan Menderes Üniversitesi, Tıp Fakültesi, Biyoistatistik, hakan.ozturk@adu.edu.tr, ORCID: 0000-0001-8112-4934.

to the testing of the factor structure, from the calculation of reliability coefficients to advanced modeling. In particular, factor analysis (exploratory and confirmatory), Cronbach's alpha, Kuder-Richardson 20, ICC, content validity indices (CVI, CVR), and fit indices (CFI, TLI, RMSEA, SRMR) are indispensable statistical tools in the scale development process (Boateng, Neilands, Frongillo, Melgar-Quiñonez, & Young, 2018).

Statistical errors made during the scale development and validation process can seriously undermine the reliability and generalizability of research in the health field. Therefore, researchers must have not only a strong grasp of psychometric principles but also a solid background in biostatistics. The quality of scales developed in fields that directly impact human life, such as health sciences, directly affects not only research results but also clinical decision-making processes (Setia, 2017).

The purpose of this section is to systematically examine the psychometric scale development and validation process in light of biostatistical approaches. First, the conceptual framework and scale development steps will be addressed, followed by a detailed discussion of validity and reliability analyses. Finally, the process will be explained through a practical example.

## 2. SCALE DEVELOPMENT PROCESS

Psychometric scale development is a systematic process that progresses through specific methodological stages, not merely the random assembly of items. The fundamental goal of this process is to create a tool that can measure the structure to be assessed (e.g., depression, quality of life, sleep patterns) in a conceptually grounded, valid, and reliable manner. Therefore, the scale development process consists of the following stages (DeVellis & Thorpe, 2021; Boateng et al., 2018).

## 2.1. Determining the Conceptual Framework

The first step in the scale development process is the theoretical definition of the construct to be measured. Literature review plays a critical role at this point. If the conceptual framework is not clearly defined, the scale items developed may fail to adequately represent the construct to be measured.

From a biostatistical perspective, no direct analysis is performed at this stage; however, the foundation for the subsequent validity and reliability stages is laid here. Therefore, determining the dimensions of the scale (e.g., unidimensional or multidimensional structures) in advance increases the interpretability of factor analyses to be performed later (Worthington & Whittaker, 2006).

## 2.2. Developing an Item Pool

Once the conceptual framework has been established, a broad pool of items representing this framework is created. Items can be obtained from the literature, similar scales, expert opinions, and focus group discussions.

Points to consider at this stage:

- Items should be clear and understandable.
- Each item should measure only one concept.
- Biased or leading statements should be avoided.

Statistical analysis is not performed directly at this stage; however, sufficient diversity must be ensured for the item analysis to be performed in the next step.

## 2.3. Expert Review and Content Validity

After the item pool is created, content experts are consulted to assess content validity. One of the most commonly used methods for this purpose is Lawshe's Content Validity Ratio

(CVR) method. Experts are asked to evaluate each item as “necessary,” “useful but not necessary,” or “unnecessary.” Then, using the CVR formula, a value is calculated for each item. Items falling below a certain threshold value are eliminated (Lawshe, 1975).

After creating the item pool, content experts are consulted to evaluate content validity. One of the most commonly used methods for this purpose is Lawshe's Content Validity Ratio (CVR) method. Experts are asked to evaluate each item as “necessary,” “useful but not necessary,” or “unnecessary.” Then, the CVR formula is used to calculate the value for each item. For an item to be removed from the scale, its calculated CVR value must be below the critical value determined based on the number of experts. The critical threshold value is determined according to the table proposed by Lawshe (1975); for example, the minimum CVR value is 0.99 for 5 experts, 0.62 for 10 experts, and 0.42 for 20 experts. Therefore, as the number of experts increases, the acceptable minimum CVR value decreases, but if the calculated CVR falls below this value, the item is eliminated (Lawshe, 1975).

Additionally, the Content Validity Index (CVI) can also be used. The CVI is based on assessments where experts rate the appropriateness of items. Content validity is generally accepted when this value is above 0.80 (Polit & Beck, 2006).

### 2.4. Pilot Study

After establishing content validity, a pilot study is conducted on a small sample to test the scale's understandability, applicability, and duration. The sample size for a pilot study is generally recommended to be between 30 and 50 people (Johanson & Brooks, 2010).

The statistical outputs of the pilot study include:

- Participant feedback on the comprehensibility of the items,
- Examination of item distributions (mean, standard deviation, skewness, kurtosis),
- Preliminary assessment of item-total correlations.

As a result of this stage, items with low statistical discriminability can be eliminated or revised.

### 2.5. Common Mistakes in the Scale Development Process

The scale development process is quite sensitive from a methodological perspective, and any mistakes made can seriously undermine the scientific value of the scale. The most common mistakes encountered in the literature are as follows:

**1. Inadequacy of the Conceptual Framework**

- Inadequate definition of the structure to be measured leads to random creation of the item pool.
- As a result, the scale fails to adequately measure the targeted psychological or clinical structure (Clark & Watson, 1995).

**2. An Inadequate or Biased Item Pool**

- Attempting to develop a scale with very few items or using leading statements causes problems in factor analyses.
- The reliability of the scale decreases when items with low item-total correlations are not eliminated (Worthington & Whittaker, 2006).

### 3. Inadequate Sample Size

- Sample size is critically important for multivariate methods such as factor analysis. Although the 5–10 times participant/item rule is generally recommended (MacCallum, Widaman, Zhang, & Hong, 1999), factor analysis is sometimes performed with very small samples in some studies. This can lead to misidentification of the structure.

### 4. Inadequate Reporting of Validity and Reliability

- Some studies claim that a scale is “valid and reliable” based solely on Cronbach's Alpha. However, alpha alone is not sufficient; additional analyses such as factor analysis, test-retest reliability, and criterion validity must be performed (Tavakol & Dennick, 2011).

### 5. Incorrect Use of Statistical Methods

A significant portion of methodological errors made in scale development studies stem from the incorrect or incomplete application of statistical methods.

- For example, common errors include not reporting KMO and Bartlett tests in factor analysis, selecting inappropriate rotation methods, or disregarding fit indices.
- Furthermore, applying parametric assumptions without questioning them in Likert-type data weakens validity.

To prevent these errors, it is recommended that biostatistical consultation be sought at every stage of the scale development process. Considering that scales developed in health sciences directly influence clinical decision-making processes, the importance of methodological robustness increases further (Boateng et al., 2018).

## 3. VALIDITY ANALYSES

For a scale to be scientifically valuable, it is not enough for it to be reliable; it must also accurately measure the construct it is intended to measure. This characteristic is called validity. Validity, in general terms, refers to the degree to which a scale or measurement tool accurately measures the concept it targets (Messick, 1995). In psychometric literature, validity is a multidimensional concept and has different types.

### 3.1. Content Validity

Content validity determines the extent to which scale items represent the structure being measured. It is generally based on evaluations made by experts in the field.

**Lawshe Method:** Experts are asked to classify each item as “necessary,” “useful but not necessary,” and “unnecessary.” The Content Validity Ratio (CVR) is then calculated. For example, in a study with 10 experts, if 8 mark an item as “necessary,” the CVR value is found using the formula. Items below the specified threshold values are removed from the scale (Lawshe, 1975).

**Content Validity Index (CVI):** Experts evaluate items using a 4-point rating scale. If each item's CVI value is above 0.80, the content is considered valid (Polit & Beck, 2006).

### 3.2. Construct Validity

Construct validity indicates whether the scale actually measures the theoretical construct it intends to measure. In other words, it is a type of validity that assesses whether the scores obtained from the scale are consistent with the relevant theoretical construct (Cronbach & Meehl, 1955). Therefore, construct validity is considered one of the most fundamental and comprehensive validity criteria for psychometric tests.

The most commonly used methods are factor analyses. Factor analysis aims to reveal a smaller number of latent factors that explain the relationships between observed variables (Fabrigar, Wegener, MacCallum, & Strahan, 1999). Exploratory factor analysis (EFA) is used to discover the structure of the scale, while confirmatory factor analysis (CFA) is used to test the theoretically predicted structure. These analyses reveal the dimensional structure of the scale and show the degree to which the structure to be measured corresponds to the theoretical framework (Brown, 2015).

### 3.2.1. Exploratory Factor Analysis

Exploratory factor analysis is one of the most commonly used methods, particularly in the early stages of scale development, to determine which factors items cluster under and to discover the underlying structure of the scale. Key points to consider in EFA applications are summarized below:

- It is used to discover the factor structure of the scale.
- Prerequisites: Kaiser-Meyer-Olkin (KMO) ≥ 0.60 and Bartlett's Test of Sphericity should be significant (Field, 2018).
- Factor loadings are generally preferred to be ≥ 0.40.
- Rotation methods: Varimax (independent factors) or Oblimin (related factors).

### 3.2.2. Confirmatory Factor Analysis

Confirmatory factor analysis is used to statistically validate the factor structure identified by exploratory factor analysis or predicted based on theoretical foundations. This method utilizes various fit indices to assess the extent to which the scale's dimensions conform to the expected theoretical model. The following criteria (fit indices) are generally considered to demonstrate that the scale has a valid structure:

- $\chi^2/df < 2$ indicates excellent fit, $2 < \chi^2/df < 3$ indicates acceptable fit,
- $CFI \geq 0.90$, $TLI \geq 0.90$,
- $RMSEA \leq 0.08$, $SRMR \leq 0.08$ (Kelloway, 1998; Hu & Bentler, 1999).

### 3.3. Criterion Validity

It shows the degree to which the scale's results correlate with a measurement accepted as the gold standard. Criterion validity is generally examined through two main approaches, which are concurrent validity and predictive validity.

**Concurrent validity:** The measurement results of the new scale are compared with those of an existing valid scale at the same time.

**Predictive validity:** The scale is used to predict a future situation (e.g., disease development).

### 3.4. Face Validity

Face validity refers to the extent to which a scale appears to measure the construct it is intended to measure. It is not measured by a statistical test, but rather evaluated through participant and expert opinions. Face validity alone is not sufficient; however, it is important in terms of participants finding the scale acceptable (Holden, 2010).

## 4. RELIABILITY ANALYSES

The reliability of a measurement tool means that it produces consistent results when repeated under the same conditions. Reliability is one of the most fundamental indicators in the scale development process. If a scale is not reliable, it indicates that the scores obtained are largely affected by random errors and therefore the scientific value of the measurements is low (DeVellis & Thorpe, 2021).

Reliability can be assessed using different methods in psychometric studies. The most commonly used approaches are internal consistency, test-retest reliability, parallel form reliability, and inter-rater reliability.

### 4.1. Internal Consistency

Internal consistency determines whether the items on the scale measure the same concept.

**Cronbach's Alpha Coefficient (α):** This is the most commonly used measure for Likert-type multi-category scales. Generally, $\alpha \geq 0.70$ values are considered acceptable, $\alpha \geq 0.80$ is considered good, and $\alpha \geq 0.90$ is considered an excellent indicator of internal consistency (Nunnally & Bernstein, 1994). However, an excessively high alpha coefficient (>0.95) may suggest excessive similarity (redundancy) between items.

**Kuder-Richardson 20 (KR-20):** Used as an alternative to Cronbach's Alpha for scales with binary responses such as Yes/No.

### 4.2. Test-Retest Reliability

It assesses whether the scale produces stable measurements over time. The same scale is administered to the same participants at specific intervals (e.g., 2–4 weeks), and the correlation between the scores is calculated.

- Pearson correlation or Intraclass Correlation Coefficient (ICC) can be used.
- $r \geq 0.70$ is generally considered an acceptable level (Streiner, Norman, & Cairney, 2015).

### 4.3. Parallel Forms Reliability

In this method, two equivalent forms developed to measure the same construct are administered to the same participants. The high correlation between the two forms indicates that the scale is reliable. It is frequently used, particularly in educational measurements (Anastasi & Urbina, 1997).

### 4.4. Inter-rater Reliability

It tests whether multiple evaluators assess the same case in a similar manner.

- Cohen's Kappa (κ): Used for binary categorical variables; $\kappa \geq 0.60$ is considered good, $\kappa \geq 0.80$ is considered very good agreement (Landis & Koch, 1977).
- Intraclass Correlation Coefficient (ICC): Assesses inter-observer agreement for continuous variables.

## 5. STATISTICAL APPROACHES AND APPLICATIONS

The psychometric scale development process is not only theoretical but also requires intensive statistical analysis. The methods used cover a wide range, from sample size to factor analysis, and from item statistics to advanced modeling. This section summarizes the most frequently used statistical approaches in the scale development process.

### 5.1. Sample Size and Power Analysis

In scale development studies, sample size plays a critical role. An insufficient sample may reduce the validity of factor analyses and negatively influence reliability coefficients. Therefore, it is generally recommended that each item in a scale be represented by at least 5–10 participants (MacCallum, Widaman, Zhang, & Hong, 1999). In addition, in most cases, a sample size of $n \geq 200$ is considered adequate for factor analyses (Comrey & Lee, 2013). Nevertheless, in order to more accurately determine the required sample size for statistical analyses such as Cronbach’s alpha, correlation coefficients, and factor loadings, conducting a power analysis using G*Power or similar software is advised.

### 5.2. Prerequisites for Factor Analysis

Before conducting factor analysis, the suitability of the data for factor analysis should be tested.

**Kaiser-Meyer-Olkin (KMO):** The KMO test is a criterion used to assess sample adequacy. A KMO value of 0.60 and above is acceptable, 0.80 and above indicates good, and 0.90 and above indicates excellent sample adequacy (Field, 2018).

**Bartlett’s Test of Sphericity:** This test tests whether the correlation matrix is not an identity matrix. A significant result ($p < 0.05$) obtained in this test indicates that the correlations between variables are suitable for factor analysis (Field, 2018).

### 5.3. Item Analyses

The contribution of each item to the scale is assessed through item analyses.

**Item–total correlation:** Items with correlations below 0.30 are considered to have low discriminative power and may be recommended for removal from the scale (DeVellis & Thorpe, 2021).

**Factor loadings:** In both EFA and CFA, factor loadings are generally expected to be ≥ 0.40.

**Contribution to internal consistency:** The contribution of each item to the overall reliability of the scale can be examined using Cronbach’s alpha “if item deleted” analysis.

### 5.4. Advanced Methods

Beyond basic analysis, advanced statistical methods are also used in the scale validation process:

**Multi-group Confirmatory Factor Analysis (CFA):** This technique examines whether the scale measures the same construct equivalently across different groups (e.g., gender or cultural backgrounds).

**Structural Equation Modeling (SEM):** SEM allows for the modeling of relationships among latent factors derived from the scale items, providing a comprehensive framework for testing theoretical models (Byrne, 2016).

**Cross-validation:** This approach involves testing the scale on different samples to enhance the generalizability and stability of the findings.

**Item Response Theory (IRT):** IRT evaluates each item’s measurement power and difficulty level, offering detailed insights into item performance. It has become increasingly prominent in fields such as educational measurement and clinical psychometrics (Embretson & Reise, 2000).

## 6. APPLICATION EXAMPLE: A SCALE DEVELOPMENT STUDY

This section demonstrates how the theoretical framework of scale development is reflected in practice. The example presented here has been constructed for illustrative purposes and is not based on an actual dataset. The aim is to concretize the statistical steps of the scale development process and provide guidance to the reader.

### 6.1. Purpose of the Research and Scale Subject

The aim is to develop an original scale to assess symptoms of depression in fathers of infants aged 3–12 months and to test its psychometric properties. Most of the existing instruments in the literature have been developed for mothers, and the limited availability of father-specific scales has created the need for such a measure (Matthey et al., 2001).

### 6.2. Method and Sample

In the first stage of the study, existing scales in the literature were reviewed, and a conceptual framework for paternal depression was established. Based on this framework and with input from subject-matter experts, a draft scale consisting of 28 items was developed. The sample size for the scale development process was determined in accordance with recommended psychometric criteria, with the aim of including at least ten participants per item. Consistent with this principle, the study sample comprised 350 fathers who voluntarily participated through various family health centers and pediatric outpatient clinics. The administration of the draft form to participants constituted the preliminary phase of the scale development process. Prior to data collection, approval was obtained from the relevant University Ethics Committee, and written informed consent was secured from all participants.

### 6.3. Analysis Processes

#### 1) Pilot Study:

The scale form was first administered to a small pilot sample (n = 30–50). At this stage, the clarity of the items, the response time, and the overall feasibility of the scale were evaluated. Based on participant feedback, revisions to wording and phrasing were made where necessary. The data obtained from the pilot study served as a foundation for administering the scale to the main sample.

### 2) Content Validity:

Expert opinions were obtained from five specialists, and the CVR was calculated using the Lawshe method. Three items with CVR values below 0.62 were eliminated, reducing the scale to 25 items.

### 3) Exploratory Factor Analysis:

- KMO = 0.89, Bartlett's test $\chi^2(300) = 2156.42$, $p < 0.001$.
- A four-factor structure was identified, explaining 62.4% of the total variance.
- Factor loadings ranged from 0.45 to 0.78.

**Table 1. Exploratory Factor Analysis Factor Loadings (Example)**

| Item | Factor 1 | Factor 2 | Factor 3 | Factor 4 |
|---|---|---|---|---|
| **Item 1** | 0.65 | - | - | - |
| **Item 2** | 0.72 | - | - | - |
| **Item 5** | - | 0.58 | - | - |
| **Item 9** | - | - | 0.64 | - |
| **Item 12** | - | - | - | 0.75 |
| … | … | … | … | … |

Note: Table is for illustrative purposes only.

### 4) Confirmatory Factor Analysis:

The model fit indices indicated an acceptable to good fit: $\chi^2/df = 2.15$, CFI = 0.93, TLI = 0.91, RMSEA = 0.056, SRMR = 0.047.

**Table 2. Confirmatory Factor Analysis Fit Indices (Example)**

| Fit Index | Value | Criterion | Interpretation |
|---|---|---|---|
| $\chi^2$/df | 2.15 | < 3 | Acceptable |
| CFI | 0.93 | ≥ 0.90 | Acceptable |
| TLI | 0.91 | ≥ 0.90 | Acceptable |
| RMSEA | 0.056 | ≤ 0.08 | Good fit |
| SRMR | 0.047 | ≤ 0.08 | Good fit |

**5) Reliability Analyses:**

- Cronbach's Alpha: The internal consistency of the total scale was α = 0.88, with subscale values ranging between 0.79 and 0.86.
- Test–retest reliability: Assessed with a subsample of 60 participants over a three-week interval, yielding r = 0.82.
- ICC = 0.84 (95% CI: 0.79–0.88), indicating high stability across measurements.

### 6.4. Interpretation of Findings

In this illustrative example, the developed scale consisted of four subdimensions and a total of 25 items. Factor loadings were found to be within acceptable ranges, and the CFA fit indices demonstrated that the model achieved a good level of fit. Reliability analyses further indicated that the scale provides stable and consistent measurements.

In conclusion, the example of the Paternal Depression Scale illustrates how validity and reliability analyses are applied in the process of scale development. The data presented here are entirely fictional; however, similar methodological steps should be followed in an actual scale development study.

## 7. CONCLUSION

The process of psychometric scale development is a multi-stage methodological framework that requires not only theoretical knowledge but also a strong biostatistical foundation. Ensuring the reliability and validity of scales enhances the scientific value of the resulting measurements and supports accurate decision-making in healthcare. In particular, statistical approaches such as factor analyses, reliability coefficients, content validity, and fit indices establish a robust methodological basis for scale construction.

Biostatistical methods play a critical role that extends beyond treating scales merely as measurement tools in clinical research. They contribute to diagnostic accuracy, the evaluation of treatment effectiveness, and the reliable reporting of epidemiological indicators. Given that poorly designed or inadequately validated instruments may lead to erroneous clinical decisions, methodological rigor in this field is directly related to patient safety.

For future research, several methodological recommendations are emphasized:

- Sample size determination in scale development should not rely solely on practical rules of thumb (e.g., ten times the number of items), but should also incorporate power analyses.
- Reporting of validity and reliability should extend beyond Cronbach’s alpha to include factor analyses, test–retest reliability, intraclass correlation coefficients (ICC), and criterion validity.
- Cross-validation and multi-group CFA across different cultures and populations are crucial to ensure generalizability.

- Advanced statistical approaches (e.g., Structural Equation Modeling, Item Response Theory) can enhance the precision of measurement.

In conclusion, the rigorous and appropriate application of biostatistical methods in the development and validation of psychometric scales not only improves research quality but also directly contributes to more reliable decision-making in healthcare practice.

## REFERENCES

Anastasi, A., & Urbina, S. (1997). Psychological testing (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. Frontiers in Public Health, 6, 149. https://doi.org/10.3389/fpubh.2018.00149

Brown, T. A. (2015). Confirmatory factor analysis for applied research (2nd ed.). New York: Guilford Press.

Byrne, B. M. (2016). Structural equation modeling with AMOS: Basic concepts, applications, and programming (3rd ed.). New York: Routledge.

Carmines, E. G., & Zeller, R. A. (1979). Reliability and validity assessment. Beverly Hills, CA: Sage.

Comrey, A. L., & Lee, H. B. (2013). A first course in factor analysis (2nd ed.). New York: Psychology Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52(4), 281–302. https://doi.org/10.1037/h0040957

DeVellis, R. F., & Thorpe, C. T. (2021). Scale development: Theory and applications (5th ed.). Los Angeles, CA: Sage.

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Mahwah, NJ: Lawrence Erlbaum Associates.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. Psychological Methods, 4(3), 272–299. https://doi.org/10.1037/1082-989X.4.3.272

Field, A. (2018). Discovering statistics using IBM SPSS statistics (5th ed.). London: Sage.

Holden, R. R. (2010). Face validity. In I. B. Weiner & W. E. Craighead (Eds.), The Corsini encyclopedia of psychology (4th ed.). Hoboken, NJ: John Wiley & Sons.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling, 6(1), 1–55. https://doi.org/10.1080/10705519909540118

Johanson, G. A., & Brooks, G. P. (2010). Initial scale development: Sample size for pilot studies. Educational and Psychological Measurement, 70(3), 394–400. https://doi.org/10.1177/0013164409355692

Kelloway, E. K. (1998). Using LISREL for structural equation modeling. Thousand Oaks, CA: Sage Publications.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159–174. https://doi.org/10.2307/2529310

Lawshe, C. H. (1975). A quantitative approach to content validity. Personnel Psychology, 28(4), 563–575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. Psychological Methods, 4(1), 84–99. https://doi.org/10.1037/1082-989X.4.1.84

Matthey, S., Barnett, B., Ungerer, J., & Waters, B. (2001). Paternal and maternal depressed mood during the transition to parenthood. Journal of Affective Disorders, 64(2–3), 93–103.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist, 50(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric theory (3rd ed.). New York: McGraw-Hill.

Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. Research in Nursing & Health, 29(5), 489–497. https://doi.org/10.1002/nur.20147

Setia, M. S. (2017). Methodology series module 5: Sampling strategies. Indian Journal of Dermatology, 62(1), 39–44. https://doi.org/10.4103/0019-5154.198646

Streiner, D. L., Norman, G. R., & Cairney, J. (2015). Health measurement scales: A practical guide to their development and use (5th ed.). Oxford: Oxford University Press.

Sullivan, G. M., & Artino, A. R. (2011). Analyzing and interpreting data from Likert-type scales. Journal of Graduate Medical Education, 3(4), 541–542. https://doi.org/10.4300/JGME-5-18

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. The Counseling Psychologist, 34(6), 806–838. https://doi.org/10.1177/0011000006288127

# POWER ANALYSIS AND SAMPLE SIZE CALCULATION IN HEALTH RESEARCH

**Hakan ÖZTÜRK**[1]

**Elvan HAYAT**[2]

## 1. INTRODUCTION

Hypothesis tests are widely used in scientific research; however, the outcome of any statistical test depends not only on the data set but also to a large extent on the research design—particularly the sample size. In this regard, statistical power analysis is a planning tool that is often ignored in research but is critical for the reliability and validity of results.

The power is the likelihood of a statistical test to correctly refute the null ($H_0$) when an alternative condition ($H_1$) actually holds, measured as $1 - \beta$. Beta (ß) represents the probability of committing a type II error, or that there is not a detectable difference (Cohen, 1988). In simpler terms, power is an estimate of the likelihood a study will find a true effect if one really exists.

Researchers use power analysis in three main ways:

1. A priori analysis, to determine the necessary sample size for a planned effect size, significance level ($\alpha$), and target power;
2. Post hoc (retrospective) analysis, to evaluate the achieved power of a completed study; and

[1] Dr., Aydın Adnan Menderes University, Faculty of Medicine, Department of Biostatistics, hakan.ozturk@adu.edu.tr, ORCID: 0000-0001-8112-4934.

[2] Assoc. Prof., Aydın Adnan Menderes University, Faculty of Political Sciences, Department of Econometrics, elvan.hayat@adu.edu.tr, ORCID: 0000-0001-8200-8046.

3. Sensitivity analysis, to estimate the smallest effect that can be detected with the available data (Suresh, 2011; Thomas & Krebs, 1997).

In health science especially clinical studies the significance of power analysis remain ethical as well as economy. The collection of an overly large sample size may result in wastage of resources and ethical concerns, while a small sample size might prevent detection of potential effects. Thus, a carefully planned power calculation could play an important role contributing to resources efficiency and the validity of research results (Biau et al., 2008).

Two types of error underlie all hypothesis tests. A Type I error (α) occurs when a true null hypothesis is wrongly rejected, commonly controlled at 0.05 or 0.01. A Type II error (β) occurs when a false null hypothesis is not rejected, producing a false-negative result. These errors are inversely related: increasing α lowers β and increases power but also heightens the risk of false positives. The researcher must balance these risks according to disciplinary norms and the practical consequences of each (Cohen, 1988; Suresh, 2011).

The power of a test, 1 – β, is dependent on factors like direct effect size and sample size and is inversely related to factors like data variance, such as measurement error and sample heterogeneity. If the effect size and variance is large, one can achieve high power even with a smaller sample: on the contrary if the effect size is small or measurement error is large, a larger sample size is required to achieve sufficient power (Cohen, 1988).

There are some methodological peculiarities in power analysis in health sciences. The first one is that interpretational issues do not relate only to statistical significance, but clinical significance is also to be considered while determining the effect size. It means that a difference between treatment groups should be presented on a clinical scale, such as risk ratio , odds ratio, and hazard ratio. These usually can be taken from the literature or the results of similar studies, for example, meta-analyses . The second peculiar feature of health sciences studies is that Bonferroni or FDR corrections are used in studies involving multiple tests. Since these corrections reduce power, they should be considered while calculating the number of samples.

In recent years, Bayesian power analyses and Monte Carlo simulation-based methods have become increasingly prevalent alongside classical frequency-based approaches. These approaches offer more flexible and realistic estimates, particularly for complex data structures such as mixed-effects models or longitudinal data (Gelman, Hill, & Vehtari, 2020).

In short, power analysis is not only a mathematical calculation but also a critical component of research design. Researchers should reconsider many decisions within the power analysis framework, from hypothesis formulation to variable selection, data collection plans, and analysis strategies. This approach reduces potential errors in the research process and enhances the scientific reliability of the results obtained (UCLA Statistical Consulting Group, n.d.).

Consequently, the concept of “power” forms the basis of research design in health sciences. A well-planned power analysis ensures the validity and ethical reliability of scientific findings through accurate sample size, balanced error levels, and meaningful effect estimates. This book chapter aims to provide researchers with a practical guide by addressing the theoretical foundations, application examples, and interpretation principles of power analysis in health research.

## 2. BASIC CONCEPTS

### 2.1. Power, Significance Level and Error Types

The purpose of statistical tests in a study is to make inferences about the population based on data obtained from the sample. Two types of errors can occur in this process:

**Type I error (α):** Incorrectly rejecting the null hypothesis ($H_0$) when it is actually true.

**Type II error (β):** Failure to reject $H_0$ when it is actually false, i.e., a false negative result.

Statistical power (1 – β) is the probability that the test will detect the alternative hypothesis ($H_1$) when it is true (Cohen, 1988). In research, a power of 80% (β = 0.20) or 90% (β = 0.10) is often targeted (Biau, Kernéis, & Porcher, 2008).

The α value (usually 0.05) represents the significance level. α and β are inversely related: when the α value is reduced (a stricter limit is set), the power is likely to decrease. Therefore, when planning a study, a balance must be struck between α, β, and sample size.

### 2.2. Effect Size

Effect size is a quantitative measure of the relationship between two variables or the difference between two groups. Unlike statistical significance, effect size indicates the clinical or practical importance of the finding (Sullivan & Feinn, 2012). Because the p-value only provides a binary classification of “significant” or “insignificant,” effect size is a complementary measure for evaluating the real-world impact of research findings.

The most commonly used effect size indicators in health sciences research vary depending on the type of analysis.

When comparing the means of two groups, Cohen's d measure is used, which expresses the ratio of the mean difference between the two groups to the pooled standard deviation. Cohen (1988) classified $d = 0.20$ as a small effect, 0.50 as a medium effect, and 0.80 as a large effect.

In variance analyses (ANOVA) comparing three or more groups, effect size is generally assessed using $\eta^2$ (eta squared) or the f value derived from it; $\eta^2$ values are interpreted with thresholds of 0.01 (small), 0.06 (medium), and 0.14 (large) (Cohen, 1988).

In correlation analyses, the effect size is directly the r coefficient. Cohen (1988) suggested $r = 0.10$ as a small, 0.30 as a medium, and 0.50 as a large relationship.

In logistic regression models, effect size is generally assessed using the Odds Ratio (OR); it is calculated as the exponential value of the $\beta$ coefficient ($e^{\beta}$) and interpreted within the context of the literature.

In risk analysis or survival studies, the Risk Ratio (RR) or Hazard Ratio (HR) is used; these ratios are evaluated according to the clinical context (Matthay et al., 2021).

Accurate estimation of effect size is critical for the reliability of power analysis. This estimate is typically obtained from three sources:

1. Results from previously published similar studies,
2. Systematic reviews or meta-analyses,
3. Findings from small-scale pilot studies.

Detecting small effect sizes requires larger sample sizes. Therefore, in clinical research, clinical significance must be evaluated alongside statistical significance ($p < 0.05$). A statistically significant difference does not necessarily mean that it is clinically important; similarly, a clinically valuable difference may not be statistically significant due to low power (Biau, Kernéis, & Porcher, 2008).

### 2.3. Sample Size, Variance and Power Relationship

The statistical power of a study is impacted by four main components (Cohen, 1988; Faul et al., 2007):

1. Effect size
2. Significance level (α)
3. Sample size (n)
4. Variance ($\sigma^2$)

The power of a statistical test varies depending on the interaction of various parameters. When one of these parameters is changed while the others are held constant, the power also varies accordingly. As the effect size increases, the power of the test increases, while an increase in variance reduces the power. Similarly, an increase in sample size increases the power. Furthermore, when the significance level (α) increases, i.e., when the threshold becomes more lenient, the power of the test increases, but the risk of false positives (Type I error) also rises. These relationships are often visualized using “power curve” graphs. For example, while 30 participants may be sufficient to detect an effect size of 0.5, approximately 200 participants may be needed to detect a smaller effect size of 0.2 (Cohen, 1988).

### 2.4. Types of Power Analysis

Power analyses can be examined under three main categories (Biau et al., 2008):

**A priori (power before data collection):** Before data collection, the sample size is calculated based on the targeted power (e.g., 0.80), the expected effect size, and the α level.

**Post hoc (power after data collection):** After the study is completed, the power achieved is calculated based on the effect size obtained. However, interpretation is limited because it is based on the observed effect (Hoenig & Heisey, 2001).

**Sensitivity analysis:** While the sample size is fixed, the minimum effect size that can be detected with a certain power is examined.

These concepts are important in determining which strategy the researcher will adopt during the design phase.

### 2.5. Interpretation of Effect Size in Health Sciences

The evaluation of effect size in health sciences must go beyond just statistical significance. It should also consider clinical relevance. A finding can be statistically significant but still lack real-world importance. For example, a 2 mmHg difference in mean systolic blood pressure may be statistically significant; however, it may not matter clinically if it doesn’t affect treatment, quality of life, or patient outcomes. Therefore, researchers should focus more on the clinical significance of their findings instead of only stressing p-values.

Effect size links statistical findings with clinical reality. Clinical significance looks at the real impact of an intervention on patients and whether this impact creates a meaningful difference in healthcare. Therefore, when performing power analysis in health research, it is crucial to evaluate both the statistical aspect of effect size and the ability to detect clinically significant changes (Sullivan & Feinn, 2012; Matthay et al., 2021).

Clinical significance thresholds are generally determined by considering clinical experience, previous research in the literature, and meaningful differences from the patient's perspective. For example, in depression treatments, a 2-point decrease in the Beck Depression Inventory score may be statistically significant, but it may not result in a noticeable improvement in the patient's quality of life. Similarly, a drug that extends life expectancy by an average of 10 days may yield a statistically robust result; however, this difference may not be considered clinically meaningful if it involves serious side effects or high costs. Therefore, researchers should consider both patient benefit and treatment costs and risks when determining clinical thresholds.

## 3. POWER ANALYSIS BY RESEARCH TYPE

The type of statistical test to be used in the research directly determines the structure of the power analysis. Since the definition of effect size, variance structure, and degrees of freedom differ for each test type, sample calculations also vary accordingly (Biau, Kernéis, & Porcher, 2008). Therefore, power analysis should be planned in conjunction with the selection of statistical tests for the research.

### 3.1. Power Analysis for the Difference Between Two Means (t-Test)

The t-test comparing the means of two independent groups (e.g., “treatment” vs. “control”) is one of the most commonly used methods in health research.

Here, power analysis is usually performed using the following four parameters:

- **α:** Significance level
- **1 – β**: Targeted power (usually 0.80 or 0.90)
- **σ**: Measurement variance
- **Δ**: Expected difference between groups (effect size)

Cohen's d effect size, defined by Cohen (1988), is used for this test:

$$d = \frac{\bar{X}_1 - \bar{X}_2}{s_{pooled}}$$

where $\bar{X}_1$ and $\bar{X}_2$ are the means of the two groups, $s_{pooled}$ is the pooled standard deviation of both groups.

For example, if a difference of 0.5 standard deviation is expected between the means of two groups and α=0.05, power = 0.80 is selected, approximately 64 participants are required in each group (Faul et al., 2007).

If the variance is high or the expected difference is small, the sample size increases rapidly.

In health research, this test is frequently used for blood parameters, quality of life scores, or biochemical measurements.

In sample planning, the clinically meaningful minimum difference (e.g., a 0.5% decrease in HbA1c) should be taken from the literature (Suresh, 2011).

### 3.2. Power Analysis for Analysis of Variance (ANOVA)

One-way ANOVA is used when comparing three or more groups.

Effect size is usually expressed as $\eta^2$ or f (Cohen's f effect size for ANOVA):

$$f = \sqrt{\frac{\eta^2}{1 - \eta^2}}$$

where $\eta^2$ is a measure of the effect size's variance ratio ($\eta^2$ = Explained variance / Total variance).

Cohen (1988) classified these values as small = 0.10, medium = 0.25, large = 0.40.

In ANOVA, as the number of groups increases, the degrees of freedom also increase, so more participants are needed for a fixed α and effect size.

For example, in a study comparing three treatment groups (α = 0.05, power = 0.80, f = 0.25), a total sample of 159 (53 people in each group) is required (Faul et al., 2007).

### 3.3. Power Analysis for Categorical Data (Chi-Square Test)

Chi-square tests evaluate the relationship between two or more categorical variables.

Power analysis is usually based on the magnitude of the difference between expected and observed frequencies.

Effect size is denoted by w:

$$w = \sqrt{\sum \frac{(p_0 - p_1)^2}{p_0}}$$

where $p_0$ is the expected proportions and $p_1$ is the observed proportions.

Cohen (1988) classified small = 0.10, medium = 0.30, large = 0.50.

For example, assuming a medium-level relationship (w = 0.30) in a two-category table (2×2), with α = 0.05 and power = 0.80 targeted, a total of 88 participants are required.

This approach is suitable for proportional outcomes in clinical trials, such as the frequency of side effects or success rates.

### 3.4. Power Analysis in Correlation and Regression Analysis

Correlation analyses measure the strength of the linear relationship between two continuous variables. The effect size is directly the r correlation coefficient.

According to Cohen (1988), r = 0.10 (small), 0.30 (medium), 0.50 (large). For example, for r = 0.30, $\alpha = 0.05$, and power = 0.80, the required sample size is 84.

A similar calculation is performed in simple linear regression; however, in multiple regression, the proportion of variance explained ($R^2$) is used as a basis. For example, in a model with 5 independent variables, approximately 92 samples are required to detect a 0.10 increase in $R^2$ (Faul et al., 2007).

### 3.5. Power Analysis in Clinical Trials: Non-Inferiority and Equivalence Tests

In classical hypothesis testing, the goal is to test the hypothesis that "there is a difference"; however, in many clinical trials, the objective is to demonstrate that the new treatment is not worse than the current standard (non-inferiority) or that it is equivalent (equivalence).

Power analysis is more complex in these types of designs because the difference margin (Δ, "non-inferiority margin") is predetermined (Piaggio et al., 2006). For example, if a new antibiotic is allowed to have a recovery rate no more than 10% lower than the standard treatment ($\Delta = 0.10$), the power analysis is performed based on this margin.

Non-inferiority studies generally require a larger sample size than classical equivalence tests (Chow et al., 2017).

### 3.6. Simulation Based Power Analyses

Analytical solutions can be difficult in complex structures such as mixed models, repeated measurements, or survival analyses. In such cases, data is generated using Monte Carlo simulation, and the power estimate is calculated by determining the rejection rate for each scenario. This method is particularly preferred for mixed-effects models and Bayesian frameworks (Green & MacLeod, 2016; Lakens & Caldwell, 2021; Muthén & Muthén, 2002).

## 4. APPLICATIONS OF POWER ANALYSIS

### 4.1. Power Analysis with G*Power Software

G*Power is a free power analysis software widely used in social, behavioral, and health sciences (Faul et al., 2007). The program supports both a priori (sample size calculation) and post hoc (actual power) as well as sensitivity analyses. Usage steps:

**Step 1:** Select the test family.

For example, if the means of two independent groups are to be compared:

Test family → t tests, then Statistical test → Means: Two independent groups (two-tailed) is selected.

**Step 2:** Enter effect size.

Cohen's recommended d values can be used (0.2 small, 0.5 medium, 0.8 large).

For greater realism, effect sizes from previous studies should be preferred (Sullivan & Feinn, 2012).

**Step 3:** Enter α, power (1–β), and the estimated d value.

For example: When α = 0.05, power = 0.80, and d = 0.5 are entered, G*Power recommends approximately 64 participants per group.

**Step 4:** On the results screen, G*Power provides the following outputs:

- Total sample size required
- Critical t value
- Noncentrality parameter (δ)
- Targeted power (1–β)

These results can be visualized in both tabular and graphical form with a “Power vs. Sample Size” curve.

G*Power is an easy-to-use software; however, it may be insufficient for power analysis of complex models such as multilevel analyses and mixed-effects models. For such analyses, it is recommended to use R or simulation-based methods (Gelman, Hill, & Vehtari, 2020).

### 4.2. Power Analysis in R

R provides more flexible and reproducible analyses with its open-source structure and extensive package support.

The most commonly used packages for power analysis are *pwr*, *simr*, and *WebPower*.

The pwr package (Champely, 2020) is based on Cohen's formulas. Below are sample R codes used in the *pwr* package to compare the means of two independent groups and calculate the minimum sample size required for correlation analysis.

```
–.Package.installation
install.packages("pwr")
library(pwr)

–.To.compare.the.means.of.two.independent.groups
pwr.t.test(d = 0.5, sig.level = 0.05, power = 0.80, type =
"two.sample")

–.For.correlation.analysis
pwr.r.test(r = 0.3, sig.level = 0.05, power = 0.80)
```

The output includes the n value (required sample size for each group).

For example, the second command shows that approximately 84 observations are required for r = 0.3; this is consistent with Cohen's (1988) tables.

Classical formulas are not valid for mixed models or repeated measures data. In this case, simulation-based power analysis can be performed using the *simr* package (Green & MacLeod, 2016).

```
library(lme4)
library(simr)

–.Example.of.a.mixed.model
Model <- lmer(outcome ~ group + (1|subject), data =
mydata )

–.Simulation_based.power.analysis
powerSim(model, nsim = 100)
```

This method calculates the model's rejection rate of $H_0$ in each scenario; this rate represents the estimated power. It is particularly useful in health research for longitudinal data (e.g., blood pressure monitoring).

The *WebPower* package is integrated with a web interface and supports power calculations in more advanced analyses such as multiple regression and structural equation modeling (SEM) (Zhang & Yuan, 2018).

### 4.3. An Example of Power Calculation for the Mean Difference in Clinical Research

Research Question: Does a new physical therapy protocol reduce pain scores (VAS) by an average of 2 points compared to conventional treatment?

Assumptions:

- Standard deviation (σ) = 3
- Target difference (Δ) = 2
- α = 0.05, power = 0.80

Cohen's d = Δ / σ = 0.67

When using the G*Power or pwr.t.test(d = 0.67, power = 0.80) command, approximately 36 participants per group are sufficient. However, considering a 10–15% potential attrition rate in clinical trials, at least 40 participants per group are recommended (Chow et al., 2017).

### 4.4. Interpretation and Reporting of Results

When doing applied power analyses, it's important to be clear about the assumptions you make about the input (like effect size and variance). To make sure that the analyses can be repeated, you should include the name and version of the calculation software you used (e.g., G*Power, R, etc.). Also, adding graphs like power curves makes it easier for the reader to understand how sample size and statistical power are related. Nonetheless, the results must be interpreted both quantitatively and clinically (Sullivan & Feinn, 2012).

Focusing only on the p-value in clinical research limits how we understand the results. In contrast, reporting extra measures like effect size and power value along with power analysis leads to a better evaluation of findings. Reputable medical journals and research guidelines, such as CONSORT and STROBE, also recommend including power analysis. This analysis should show the statistical strength of the study (Schulz, Altman, & Moher, 2010; von Elm et al., 2007).

### 4.5. Common Mistakes and Misinterpretation

**1. Overreliance on post hoc power analysis:**

The power analysis performed after the study is completed is often misleading because it is based on the observed effect size (Hoenig & Heisey, 2001). Therefore, it is recommended to perform calculations during the planning phase (a priori).

2**. Failure to report effect size:**

Reporting only the p-value may indicate the presence of a statistically significant difference; however, it does not provide information about the magnitude, direction, or clinical importance of this difference. Failure to specify the effect size makes it difficult to assess the practical or clinical significance of the results (Sullivan & Feinn, 2012).

**3. Missing assumptions:**

Simply stating that a “power analysis was performed” without clearly specifying the basic assumptions of the power analysis—such as the significance level (α), error probability (β), effect size, and variance—is insufficient. Such incomplete reporting reduces the reproducibility of the analyses and the reliability of the results.

### 4. Failure to account for missing data:

Power analyses conducted without considering the possibility of sample loss (dropout) in clinical trials often result in lower-than-expected power because the trial is completed with a smaller sample size than planned. Therefore, potential dropout rates should be estimated in advance in the power analysis, and the sample size should be adjusted accordingly.

### 5. Failure to adjust for multiple comparisons:

When multiple statistical tests are applied, failure to adjust for multiple comparisons using methods such as Bonferroni or FDR (False Discovery Rate) increases the significance level (α) and leads to α-inflation. This situation negatively affects the reliability and statistical power of the analyses by increasing the likelihood of false positive results (Type I error) (Bender & Lange, 2001).

## 4.6. Clinical and Ethical Perspective

Power analysis is not just a statistical requirement; it is also an ethical necessity in clinical research (Biau et al., 2008). Studies with insufficient sample sizes may miss meaningful differences and overlook results that could be valuable for clinical practice. This creates an ethical issue by wasting participants' time and effort and exposing them to unnecessary risks.

Conversely, studies with overly large sample sizes can result in unethical consequences, such as wasting resources, creating economic burdens, and subjecting participants to unnecessary interventions. Thus, proper power analysis is crucial for enhancing scientific validity and ensuring participant safety, making good use of resources, and following research ethics.

In health-related clinical research, including information about power analysis in ethics committee applications is a required evaluation criterion. A well-planned power analysis shows both the statistical soundness of the study and ethical responsibility.

## 5. CONCLUSIONS AND RECOMMENDATIONS

Power analysis is not just a statistical technique in health sciences research; it plays a key role in ensuring scientific validity and ethical responsibility. Making sure a study has enough power allows findings to be interpreted reliably in both scientific and clinical contexts.

Statistical power is central to research design. Not finding a significant difference in a study does not always mean there is no difference; real effects could be missed due to low power (Biau, Kernéis, & Porcher, 2008). Therefore, power analysis is essential for preventing false negative results and ensuring the effective use of research resources. Since Cohen's (1988) classic framework, the key components of power analysis—effect size, sample size, variance, and significance level—have not changed; however, how we apply them has progressed. Today, tools like G*Power, R, and simulation-based methods allow for practical solutions for both traditional and complex models (Faul, Erdfelder, Lang, & Buchner, 2007; Green & MacLeod, 2016).

Thorough analysis during the planning stage in health sciences ensures both statistical accuracy and ethical soundness. Low power can make data collected from participants meaningless, while overly large samples can lead to ethical concerns, such as unnecessary use of resources and subjecting participants to extra interventions (Biau et al., 2008). Thus, power analysis is closely tied to research ethics principles.

A priori power analysis should always be done in research; the expected effect size, target power, and significance level (α) should be clearly stated before data collection and included in the ethics committee application. When choosing effect size, literature-based approaches should be prioritized; results from meta-analyses or earlier similar studies provide realistic estimates (Sullivan & Feinn, 2012). Variance estimates and potential sample losses must also be considered. Since a 10 to 20 percent sample loss is common in clinical trials, it is advisable to plan for an additional sample that reflects this rate (Chow et al., 2017).

Power analysis results should be reported in a clear, transparent, and reproducible way. Stating the type of test used, effect size, significance level (α), error probability (β), software, and references increases the reliability of the methods (APA 7; CONSORT, 2010; STROBE, 2007). Post hoc power analyses should be approached with caution. These analyses, which rely on observed effects, can often be misleading and should only be considered as explanatory information (Hoenig & Heisey, 2001).

With the development of research methods, simulation-based power analyses have also gained importance. More flexible and realistic power estimates can be made using simr or similar R packages in mixed models, repeated measures, and Bayesian analyses (Green & MacLeod, 2016). Furthermore, Bayesian power analysis allows uncertainty to be modeled more realistically by incorporating prior information into the model. Adaptive designs offer both ethical and economic advantages by allowing the sample size to be updated based on interim analysis results. Machine learning-supported power estimates add a new dimension to model-based simulations by using real-world data (e.g., EHR, hospital records) (Gelman, Hill, & Vehtari, 2020).

Finally, we must always consider clinical significance. Statistical significance ($p < 0.05$) alone is not enough; we also need to evaluate the clinical importance and effect size of the finding (Sullivan & Feinn, 2012). To raise this awareness among clinical researchers, we should create training programs, online modules, and open-access guides on statistical power analysis.

In conclusion, creating reliable information in health sciences depends on well-designed and properly powered studies. A well-planned power analysis is not just a statistical requirement; it also shows scientific and ethical responsibility. For researchers, this analysis is a key step in understanding the meaning of the data and the importance of the findings.

## REFERENCES

Bender, R., & Lange, S. (2001). Adjusting for multiple testing—when and how?. Journal of clinical epidemiology, 54(4), 343-349. https://doi.org/10.1016/S0895-4356(00)00314-0

Biau, D. J., Kernéis, S., & Porcher, R. (2008). Statistics in brief: the importance of sample size in the planning and interpretation of medical research. Clinical orthopaedics and related research, 466(9), 2282-2288. https://doi.org/10.1007/s11999-008-0346-9

Champely, S. (2020). pwr: Basic functions for power analysis (R package version 1.3-0). Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=pwr

Chow, S. C., Shao, J., Wang, H., & Lokhnygina, Y. (2017). Sample size calculations in clinical research. chapman and hall/CRC.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior research methods, 39(2), 175-191. https://doi.org/10.3758/BF03193146

Gelman, A., Hill, J., & Vehtari, A. (2020). Regression and other stories. Cambridge University Press. https://doi.org/10.1017/9781139161879

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. Methods in Ecology and Evolution, 7(4), 493-498. https://doi.org/10.1111/2041-210X.12504

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: the pervasive fallacy of power calculations for data analysis. The American Statistician, 55(1), 19-24. https://doi.org/10.1198/000313001300339897

Lakens, D., & Caldwell, A. R. (2021). Simulation-based power analysis for factorial analysis of variance designs. Advances in Methods and Practices in Psychological Science, 4(1), 1–17. https://doi.org/10.1177/2515245920951503

Matthay, E. C., Hagan, E., Gottlieb, L. M., Tan, M. L., Vlahov, D., Adler, N., & Glymour, M. M. (2021). Powering population health research: considerations for plausible and actionable effect sizes. SSM-population health, 14, 100789. https://doi.org/10.1016/j.ssmph.2021.100789

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. Structural equation modeling, 9(4), 599-620. https://doi.org/10.1207/S15328007SEM0904_8

Piaggio, G., Elbourne, D. R., Altman, D. G., Pocock, S. J., Evans, S. J., & CONSORT Group, F. T. (2006). Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement. Jama, 295(10), 1152-1160. https://doi.org/10.1001/jama.295.10.1152

Schulz, K. F., Altman, D. G., Moher, D., & Consort Group. (2010). CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. Journal of clinical epidemiology, 63(8), 834-840. https://doi.org/10.1136/bmj.c332

Suresh, K. P. (2011). An overview of randomization techniques: an unbiased assessment of outcome in clinical research.

Journal of human reproductive sciences, 4(1), 8-11. https://doi.org/10.4103/0974-1208.82352

Sullivan, G. M., & Feinn, R. (2012). Using effect size—or why the P value is not enough. Journal of graduate medical education, 4(3), 279-282. https://doi.org/10.4300/JGME-D-12-00156.1

Thomas, L., & Krebs, C. J. (1997). A review of statistical power analysis software. Bulletin of the ecological society of America, 78(2), 126-138.

UCLA Statistical Consulting Group. (n.d.). Introduction to power analysis. UCLA Institute for Digital Research and Education (IDRE). Retrieved October 20, 2025, from https://stats.oarc.ucla.edu/other/mult-pkg/seminars/intro-power/

Von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., & Vandenbroucke, J. P. (2007). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. The lancet, 370(9596), 1453-1457. https://doi.org/10.1371/journal.pmed.0040296

Zhang, Z., & Yuan, K. H. (2018). Practical Statistical Power Analysis. ISDSA Press, Granger, Indiana.

BIOSTATISTICAL EVALUATONS

yaz
yayınları

YAZ Yayınları
M.İhtisas OSB Mah. 4A Cad. No:3/3
İscehisar / AFYONKARAHİSAR
Tel : (0 531) 880 92 99
yazyayinlari@gmail.com • www.yazyayinlari.com