BAREKENG: Journal of Mathematics and Its Applications
December 2024 Volume 18 Issue 4 Page 2413–2422
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol18iss4pp2413-2422

APPLICATION OF YATES METHOD FOR MISSING DATA ESTIMATION IN YOUDEN SQUARE DESIGN AND ANALYSIS

Ratna Melati Hadiputri [1], Syifa Nayla Alifaa[2], Restu Arisanti[3], Sri Winarni[4*]

[1,2] Bachelor Programme of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran
[3,4]Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran
Jln. Raya Bandung-Sumedang, Jatinangor, West Java, 45363, Indonesia

*Corresponding author's e-mail: *sri.winarni@unpad.ac.id*

ABSTRACT

Article History:

Received: 25th, March 2024
Revised: 30th, May 2024
Accepted: 26th, July 2024
Published:14th, October 2024

Keywords:

Yates Method;
Youden Square Design;
Analysis of Variance;
Missing Data;
Chemical Substances.

The Youden square design is widely used in experimental research to control two sources of variability, but missing data can compromise the results. Addressing missing data is critical to maintaining the integrity and reliability of such experiments. This paper proposes to adapt the Yates method to handle missing data specifically in Youden square designs. We begin by outlining the structure of the Youden square design and the challenges posed by missing data. The Yates method, known for its robustness in estimating missing data, is adapted to fit this design. We demonstrate its effectiveness through simulations and real-world case studies. The simulation involved generating experimental data with one missing value, and the case study analyzed chemical process research with critical missing data points. The results show that the Yates method maintains statistical validity and improves data completeness compared to traditional methods. Its advantage lies in utilizing Youden's quadratic structure for more accurate estimation. This study highlights the Yates method as a solution to handle missing data, improving the quality and reliability of experimental research.

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-ShareAlike 4.0 International License.

How to cite this article:

R. M. Hadiputri, S. N. Alifaa, R. Arisanti and S. Winarni., "APPLICATION OF YATES METHOD FOR MISSING DATA ESTIMATION IN YOUDEN SQUARE DESIGN AND ANALYSIS," *BAREKENG: J. Math. & App.*, vol. 18, iss. 4, pp. 2413-2422, December, 2024.

Copyright © 2024 Author(s)
Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

Scientific planning is constantly being done for the betterment of the world. Generally, scientific planning is known as experimental design. An experimental design is a well-defined plan, with each action step clearly defined, so that information is relevant or necessary to reveal the problem. The existence of an experimental design aims to obtain as much information as possible that is necessary and useful in carrying out research. The fundamental form of experimental design is the perfectly randomized design, which involves completely random assignment of treatments to experimental units, ideally possessing homogeneity. If the experimental units are not relatively homogeneous, a randomized block design can be used where the treatments to the experimental units cannot be fully assigned due to restrictions or confounding so the experimental units are grouped in blocks so that the units within a block are relatively homogeneous. Many studies use an experimental design to determine the effect of experimental factors on the response.

In cases where two factors are organized in rows and columns, a Latin square design is employed. This design ensures random placement of all treatments within the experimental unit, with each treatment appearing an equal number of times in every row and column [1]. Should the treatments not be evenly distributed across rows and columns, a Youden square design is utilized. W.J. Youden devised the Youden square design to address the deficiencies of the Latin square design's incompleteness. This design is frequently employed in research to mitigate block effects and enhance the internal validity of experiments [2]. Other studies utilize the Youden experimental design to evaluate the influence of various factors on experimental outcomes. The Youden design is often chosen due to its ability to provide comprehensive information about the effects of each factor and their interactions while minimizing the number of experiments required [3]–[8].

Incomplete data is often found in experimental results due to confounding factors or errors that occur when experiments are in progress. Generally, incomplete data or missing data is caused by damage to the object of research, improper treatment, and irrational data. As a result of incomplete data, it causes imbalance, loss of symmetry, or orthogonal properties due to no longer equal to zero values of $\Sigma\beta_i$ or $\Sigma\pi_j$ [9]. The study on the design and analysis of experimental data using the Youden squares method, with the data studied using complete Youden squares, concludes that there is an effect of using different chemical substances on the length of reaction time [10]. In previous studies, it was found that the data used in the Youden design were complete, but there has not been much discussion about the case of missing data [11].

Therefore, the purpose of this study is to examine data analysis in Youden design when facing the case of missing data. We used the complete data of the Youden design as an illustrative example and then simulated the loss of a single data point. The Yates method, which was chosen for its robustness and simplicity, was used to predict the missing data. This experimental case explores the impact of different chemical substances on reaction time. In addition, this study compares the variance analysis results between data sets with missing data and complete data, evaluating the consistency of their conclusions. While previous studies have addressed missing data in Youden Square Designs, often considering the loss of up to two data points, our focus on the Yates method is due to its specific adaptability to the Youden square structure. Our findings demonstrate the applicability of the Yates method in maintaining statistical validity and data completeness, making it a valuable tool for analyzing other Youden square designs with missing data.

2. RESEARCH METHODS

2.1 Data Description

In this study, the data used is secondary data, which has been modified to include five rows, four columns, five treatments, and one missing data point. The secondary data pertains to the effects of five different chemicals (A, B, C, D, and E) on the duration of chemical processes. Additionally, the secondary data has two factors: the first factor is batch, consisting of five batches with four processes each, and the second factor is day, where each of the four days runs five different chemicals. The variables used in this study are the type of chemical and its effect on the duration of chemical processes. There are also two other variables acting as factors: batch variable and day variable. The secondary data is presented in Table 1.

Table 1. Data Regarding the Impact of Various Chemicals on Chemical Process Reaction Times

Batch (Row)	Day (Column)			
	1	2	3	4
1	A = 8	B = 7	D = 1	C = 7
2	C = 11	E = 2	A = 7	D = 3
3	B = 4	A = 9	C = 10	E = 1
4	D = 6	C = ...	E = 6	B = 6
5	E = 4	D = 2	B = 3	A = 8

***Data source*: [9]**

2.2 Youden Square Design

Youden Square Design is a method of designing experiments to identify the effects of two or more factors. The design was first developed by George Youden. Youden Square Design has the equilibrium property of the Incompletely Balanced Group Random Design where all treatments are tried in equal numbers in each row or column **[12]**. This design is a combination of Latin Square Design and incompletely Balanced Group Random Design **[13]**. If the number of rows equals the number of treatments, then the linear model for the Youden square design with a row, b column and c treatment is:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \tau_k + \varepsilon_{ijk} \quad (1)$$

$$\text{with } i = 1, 2, \dots, a \; ; \; j = 1, 2, \dots, b \; ; \; k = 1, 2, \dots, c$$

Where:

Y_{ijk} : Observations on i-th row, j-th column, and k-th treatment
μ : General Average
α_i : Influence of the i-th row
β_j : Influence of the j-th column
$\tau_{(k)}$: Influence of the k-th row treatment
$\varepsilon_{ij(k)}$: The effect of the k-th error, in the i-th row, and the j-th column

The hypothesis in the Youden Square Design to test the treatment is as follows.

$H_0 : \tau_1 = \tau_2 = \cdots = \tau_a$ (no effect of treatment on response was observed)
H_1 : at least one k with $\tau_k \neq 0$, $k = 1,2, \dots, a$ (there is an effect of treatment on the observed response)

2.3 Yates Method for Missing Data

The Yates method is a method of estimating missing data on the design of experiments by minimizing the Number of Sum Squares Error (SSE) **[14]**. This method was developed by Frank Yates and is useful for correcting standard error estimates in chi-square tests when the sample size is relatively small. According to Little and Rubin, the Yates method consists of three steps:

1. Expected missing data
2. Filling missing data with suspected data
3. Analyzing complete data

There is also the formulation of the expected value of missing data using the Yates method as follows.

$$\hat{\boldsymbol{P}} = \left(\frac{(\boldsymbol{r} \times \boldsymbol{R_a}) + (\boldsymbol{t} \times \boldsymbol{T_b}) - \boldsymbol{D}}{(\boldsymbol{t} - \boldsymbol{1})(\boldsymbol{r} - \boldsymbol{1})} \right) \quad (2)$$

Where:

$\hat{\boldsymbol{P}}$: Expected value of missing data
$\boldsymbol{r}$: The amount of treatment for each factor
$\boldsymbol{t}$: The amount of treatment
$\boldsymbol{R_a}$: The number of responses in the row/column of a factor containing missing data
$\boldsymbol{T_b}$: Number of responses in treatment b that contain missing data
$\boldsymbol{D}$: The amount of overall data that is not missing

2.4 Assumption Test

a. Residual Normality Test

Testing for normally distributed data can be done using the Shapiro-Wilk test. The Shapiro-Wilk test is used if the sample data used is smaller or less than 50. The following is the hypothesis.

H_0 : Normally distributed residual.
H_1 : Residuals are not normally distributed.

There is also a decision based on normality testing. If the $P_{value} \geq 0.05$, then the research data is normally distributed. On the other hand, if the $P_{value} < 0.05$ then the research data is not normally distributed.

b. Variance Homogeneity Test

The homogeneity test is a statistical test procedure that aims to show that two or more groups of data samples are taken from a population that has the same variance [15]. In the variance homogeneity test, the null hypothesis (H_0) and alternative hypothesis (H_1) can be formulated as follows.

H_0 : Homogeneous residual variance.
H_1 : At least one pair of groups has a non-homogeneous variance.

There is also a decision based on the homogeneity of variance test. If the $P_{value} \geq 0.05$, then the homogeneity of variance assumption test is fulfilled because the residual variance is homogeneous. On the other hand, if the $P_{value} < 0.05$, then the assumption of homogeneity of variance test is not fulfilled because there is at least one pair of variance groups that are not homogeneous.

2.5 Analysis of Variance (ANOVA) on the Youden Square Design

To prepare an ANOVA table, several calculations with the following formula are required [10].

$$\text{Total Sum of Squares } (TSS) = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{a} y_{ijk}^2 - \frac{y^2}{N} \quad (3)$$

$$\text{Sum of Row Squares } (SSR) = \frac{1}{b}\sum_{i=1}^{a} y_{i..}^2 - \frac{y^2}{N} \quad (4)$$

$$\text{Sum of Column Squares } (SSC) = \frac{1}{a}\sum_{j=1}^{b} y_{i..}^2 - \frac{y^2}{N} \quad (5)$$

$$\text{Sum of Squares Treatment } (SST) = \frac{b}{\lambda a}\sum_{k=1}^{a} Q_k^2 \quad (6)$$

$$\text{Sum of Squares of Error } (SSE) = TSS - SSR - SSC - SST \quad (7)$$

$$\text{Mean Square of Rows } (MSR) = \frac{SSR}{a-1} \quad (8)$$

$$\text{Mean Square of Column } (MSC) = \frac{SSC}{b-1} \quad (9)$$

$$\text{Mean Square of Treatment } (MST) = \frac{SST}{a-1} \quad (10)$$

$$\text{Mean Square of Error } (MSE) = \frac{SSE}{(a-1)(b-2)} \quad (11)$$

Table 2. Calculations in the ANOVA table

Sources of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F_{value}	F_{table}
Treatment	$a-1$	SST	MST	$\frac{MST}{MSE}$	$F_{(a-1);(a-1)(b-2)(\alpha)}$
Row	$a-1$	SSR	MSR		
Column	$b-1$	SSC	MSC		

Sources of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F_{value}	F_{table}
Error	$(a-1)(b-2)$	SSE	MSE		
Total	$ab-1$	TSS			

With the test criteria, reject H_0 if $F_{value} > F_{table}$ or if $P_{value} < \alpha$.

2.6 Duncan Test

Duncan's further test is one of the further tests carried out if the results of the variance analysis are significant. The purpose of the Duncan test is to find out which treatment has the greatest effect to the smallest effect. The steps for carrying out the Duncan Advanced Test are as follows.

a. Treatment averages are sorted from smallest to largest
b. Calculate Duncan's value with the formula

$$D = d_{\alpha,p,dfe}\sqrt{\frac{MSE}{r}} \quad (12)$$

Where:

D	:	Duncan value
α	:	Significance level
p	:	Treatment rank distance
dfe	:	Degree of freedom error
MSE	:	Mean Square Error
r	:	number of observations in each group

c. Treatment is said to be significant if the Duncan value < the difference in the treatment average

3. RESULTS AND DISCUSSION

3.1 Descriptive Statistics

There are also descriptive statistics of the response, namely the length of the chemical process reaction time, presented in **Table 3** as follows.

Table 3. Descriptive Statistics of Response (Chemical Process Reaction Time)

Descriptive Statistics	Value
Minimum	0
Quartile 1	2.75
Median	6
Mean	5.25
Quartile 3	7.25
Maximum	11

From **Table 3** above, it can be seen that the minimum value of the chemical process reaction time is 0. This occurs because of missing data in the data used. In addition, the maximum value is 11. There is also a middle value between the smallest value and the median (middle value) of the production capacity which is 2.75 where the middle value is 6. There is also a middle value between the median and the highest value of the length of time of the chemical process reaction which is 7.25. In addition, the average (mean) of the chemical process reaction time is 5.25.

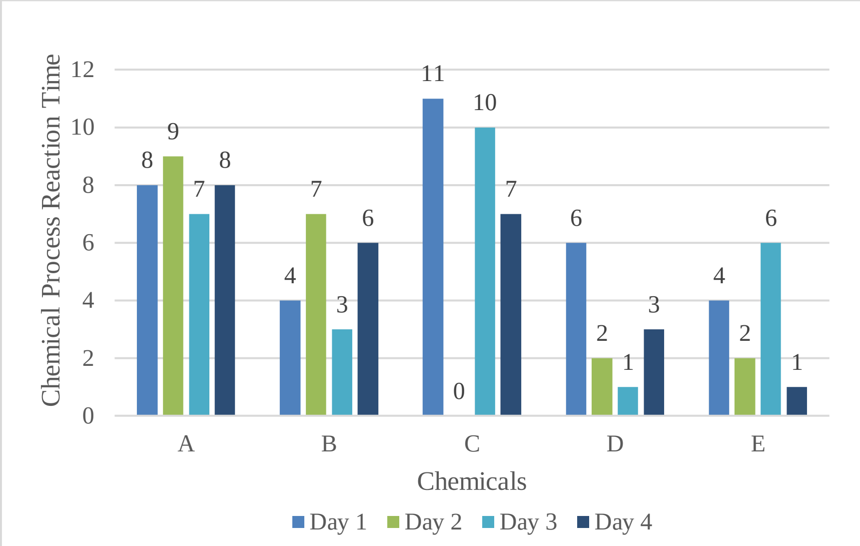

Figure 1. Chemical Process Reaction Time Chart of Each Chemical Per Day

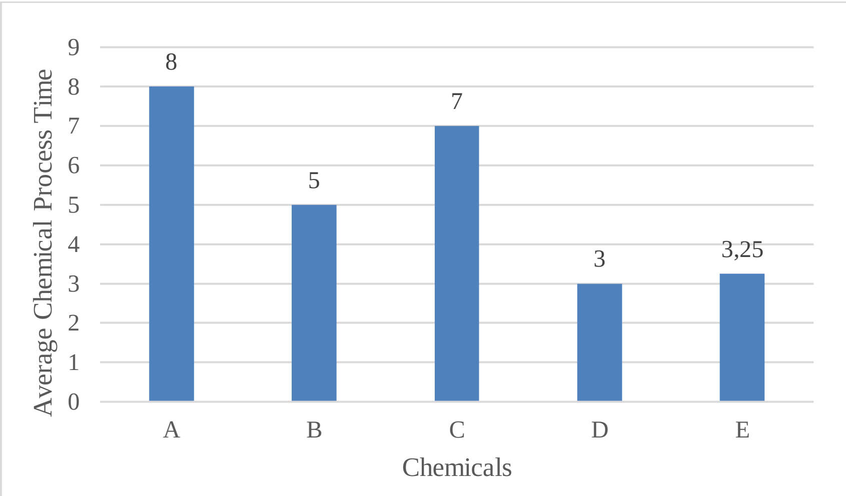

Figure 2. Average Chemical Process Time of Each Chemical

In **Figure 1**, it can be seen that each chemical shows fluctuations, which means that the length of the chemical process reaction time varies every day (up and down within 4 days). It can also be seen that the chemical with the highest chemical process reaction time is chemical C on the first day of 11. In addition, it can also be seen visually in **Figure 2** that the average of chemical A is the highest compared to other chemicals, which is 8. Further statistical analysis needs to be carried out to eliminate the element of subjectivity based on visual analysis.

3.2 Missing Data Estimation Using the Yates Method

In this research, a development was carried out by eliminating one piece of data randomly. The result of missing data randomization is located in the fourth row, second column, and chemical treatment C (x_{423}). The number of responses per row and the number of responses per column were calculated.

Table 4. Calculation of the Number of Responses for Each Row and Column

Batch (Row)	Day (Column)				The number of responses in each row ($y_{i..}$)
	1	2	3	4	
1	A = 8	B = 7	D = 1	C = 7	23
2	C = 11	E = 2	A = 7	D = 3	23

Batch (Row)	Day (Column)				The number of responses in each row ($y_{i\cdot\cdot}$)
	1	2	3	4	
3	B = 4	A = 9	C = 10	E = 1	24
4	D = 6	C = ...	E = 6	B = 6	18
5	E = 4	D = 2	B = 3	A = 8	17
The number of responses in each column ($y_{\cdot j\cdot}$)	33	20	27	25	105

Some treatments contain missing data (chemical C) of 28. The calculation results in **Table 4** can help in calculating missing data using the Yates method. Because there are two factors in the Youden Square Design data, the Yates data calculation is carried out twice for each factor, namely the batch factor and the day factor.

- Calculation of missing data using the Yates method on batch factors

$$\hat{P} = \left(\frac{(4 \times 18) + (4 \times 28) - 105}{(4-1)(4-1)}\right) = 8.778$$

- Calculation of missing data using the Yates method on day factors

$$\hat{P} = \left(\frac{(5 \times 20) + (4 \times 28) - 105}{(5-1)(4-1)}\right) = 8.9167$$

After calculating the missing data using the Yates method on both factors, the average of the two missing data values is taken. Then the missing data value x_{423} is obtained, which is 8.85. The value obtained is different from the original value in the complete data, which is 8. From these results, it can be concluded that the calculation used can only estimate or estimate the value of missing data and this value can help further analysis.

3.3 Assumption Tests

3.3.1 Residual Normality Test

a. Hypothesis
   H_0 : residuals are normally distributed
   H_1 : residuals are not normally distributed

b. Significance level

   $\alpha = 5\% = 0.05$

c. Statistics Test

   Shapiro-Wilk Test

Table 5. Residual Normality Assumption Test Results

	Shapiro-Wilk Test
P-Value	0.4025

d. Criteria Test

   Reject H_0 if the $P_{value} < \alpha$, accept in other cases.

e. Decision

   Since the $P_{value} = 0.4025 > \alpha = 0.05$, H_0 is accepted.

f. Conclusion

With a significance level of 5%, it can be concluded that the residuals are normally distributed so that the assumption of residual normality is fulfilled.

3.3.2 Variance Homogeneity Test

a. Hypothesis

$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_5^2$ (homogeneous residual variance)
H_1 : there is at least one pair of σ^2 that is not equal (residual variance is not homogeneous)

b. Significance level

$\alpha = 5\% = 0.05$

c. Statistics Test

Bartlett's test uses the chi-square distribution approach with $a - 1$ degrees of freedom.

Table 6. Bartlett Test Results

Bartlett's K-squared	Degrees of Freedom (df)	P-Value
3.2295	4	0.5202

d. Criteria Test

Reject H_0 if the $p_{value} < \alpha$, accept in other cases.

e. Decision

Since the $p_{value} = 0.5202 > \alpha = 0.05$, H_0 is accepted.

f. Conclusion

With a significance level of 5%, it can be concluded that the residual variance is homogeneous so that the assumption of homogeneity of variance is fulfilled.

3.4 Analysis of Variance

By calculating the analysis of variance, the results presented in **Table 7** show the average response of treatment (chemicals), each row (batch), and each column (day). In **Table 7** it can be seen that the p_{value} of chemicals (treatment) is 0.00496. Because the $p_{value} = 0.00496 < \alpha = 0.05$, it can be concluded that H_0 is rejected, which means there is a significant effect. So it is concluded that there is an effect of the type of chemical on the length of time of the chemical process.

The conclusion obtained from the analysis of the variance of missing data with the analysis of the variance of complete data has the same conclusion that the length of time of the chemical process is significantly different, so it can be concluded that the estimation of missing data can be used to assist the analysis of variance and obtain the same analysis conclusion as the complete data. In addition, because there is an effect of treatment on the response, further tests are needed to see different treatments.

Table 7. Bartlett Test Results

Sources of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F_{value}	P_{value}
Chemicals	4	125.64	31.41	8.831	0.00496
Batch	4	14.16	3.541	0.995	0.46292
Day	3	6.97	2.324	0.653	0.60289
Error	8	28.46	3.557		
Total	19				

3.5 Further Test

The effect of treatment on the response can be further analyzed to determine the type of chemical that has the greatest effect or gives the greatest effect on the length of time of the chemical process. The method that will be used for further tests is the Duncan test.

Table 8. Results of the Duncan Test Analyzing the Treatment's Effect on Response

Factors (Treatment)	Average Response	Grouping
C	9.21	a
A	8	ab
B	5	bc
E	3.25	c
D	3	c

From Table 8 above, it can be seen that chemical C has a different effect than chemical B, chemical E, and chemical D. Chemical C and chemical A have the same effect on the length of time of the chemical process. However, chemical C gives a length of chemical process time of 9.2125 where the value is greater than the length of chemical process time by chemical treatment B. Thus, it can be concluded that chemical C has the greatest effect on the length of chemical process time with chemical A can be used as an alternative.

The conclusions generated from the further test with missing data yielded similar conclusions to the further test with complete data. Thus, the estimated value of missing data can be used for further analysis by producing similar conclusions to the complete data.

4. CONCLUSIONS

Based on the results of the analysis and discussion of estimating missing data using the Yates method in the Youden Square Design, an estimated value of 8.85 was obtained, where this value is different from the original value of the data so it can be concluded that the Yates method can only estimate the value of missing data. However, the results of missing data estimation provide similar conclusions to the analysis of complete data starting from the conclusion of assumption tests, variance analysis, to further tests. Therefore, the Yates method can be used to estimate missing data in the Youden Square Design.

From the results of the analysis, it was also concluded that there was an influence of the type of chemical on the length of time of the chemical process with the assumption tests that had previously been fulfilled, namely the normality assumption test and the homogeneity assumption test. Because of the influence of the type of chemical, further tests were carried out to determine the type of chemical that had the greatest effect on the length of the chemical process. The results showed that chemical C had the greatest effect on the length of time of the chemical process, with chemical A being used as an alternative.

REFERENCES

[1] A. Dean, D. Voss, and D. Draguljic, "Design and Analysis of Experiments 2nd ed," New York, NY: Springer New York, 2017, doi: 10.1007/978-3-319-52250-0.

[2] I. Munawwaroh, L. Saliaputri, S. M. Herdiyani, T. T. Puspasari, and S. Winarni, "Implementasi Analisis Variansi Pada Desain Bujur Sangkar Youden Untuk Eksperimen," *Equator J. Math. Stat. Sci.*, vol. 2, no. 1, p. 10, 2023, doi: 10.26418/ejmss.v2i1.64598.

[3] E. Karageorgou and V. Samaniduo, "Youden Test Application in Robustness Assays During Method Validation," *Journal of Chromatography A.*, vol. 1353, pp. 131-139, 2014, doi: 10.1016/j.chroma.2014.01.050.

[4] A. R. J. Degray, "American Society for Quality The Sequential Youden Square American Society for Quality Stable URL : http://www.jstor.org/stable/1267359," vol. 12, no. 1, pp. 147–152, 2016.

[5] Y. Hua, A. S. Hedayat, M. Yang, and S. Altan, "An Extended Youden Design in Biological Assays," *Stat. Biopharm. Res.*, vol. 13, no. 3, pp. 330–336, 2021, doi: 10.1080/19466315.2020.1841023.

[6] T. Taner and J. Antony, "The assessment of quality in medical diagnostic tests: A comparison of ROC/Youden and Taguchi methods," *Int. J. Health Care Qual. Assur.*, vol. 13, no. 7, pp. 300–307, 2000, doi: 10.1108/09526860010378744.

[7] E. W. Yunitasari, "Desain dan Analisis Eksperimen Menggunakan Bujursangkar Youden di PT. XYZ," vol. 2, no.2, 2019.
[8] H. M. Uti, S. Alisya, M. A. Rahmah, "Analisis Perakitan Komponen Televisi dengan Rancangan Bujur Sangkar Youden," *BIAStatistics Journal of Statistics Theory and Aplication*, vol. 17, no. 2, pp. 47-53, 2023.
[9] D. C. Montgomery, *Design and analysis of experiments*. John Wiley & Sons, Inc., 2017.
[10] T. Yaumira, C. Firmansyah, and R. M. Hadiputri, “Desain Youden Sebagai Desain Bujur Sangkar Latin Tak Lengkap,” vol. 6274, pp. 1–13, 2023.
[11] A. K. Kaushik and P. Kaushik, “Analysis of Youden square design with several missing observations,” vol. 4, no. 5, pp. 104–111, 2019.
[12] A. S. Dewi, T. Widiharih, and R. Rahmawati, “Analisis Variansi Pada Rancangan Bujur Sangkar Youden Dengan Dua Data Hilang,” *J. Gaussian*, vol. 8, no. 3, pp. 356–365, 2019, doi: 10.14710/j.gauss.v8i3.26680.
[13] G. W. Oehlert, *A First Course in Design and Analysis of Experiments*, vol. 1, no. 1. 2000.
[14] M. Susilawati and K. Sari, “Pendugaan Data Hilang Dengan Metode Yates Dan Algoritma Em Pada Rancangan Lattice Seimbang,” *E-Jurnal Mat.*, vol. 4, no. 2, p. 74, 2015, doi: 10.24843/mtk.2015.v04.i02.p092.
[15] Y. Zhou, Y. Zhu, and W. K. Wong, "Statistical Tests for Homogeneity of Variance for Clinical Trials and Recommendations," *Contemporary clinical trials communications*, vol. 33, 2023, doi : 10.1016/j.conctc.2023.101119