



(RESEARCH ARTICLE)



## Evaluation of Analysis of means, Wald, and Log-likelihood ratio tests in comparing independent proportions

SONER YİĞİT \*

*Faculty of Agriculture, Animal Science Department, Biometry and Genetics Unit, Canakkale Onsekiz Mart University, Canakkale, TURKIYE.*

International Journal of Science and Research Archive, 2025, 16(01), 1180-1187

Publication history: Received on 02 June 2025; revised on 13 July 2025; accepted on 15 July 2025

Article DOI: <https://doi.org/10.30574/ijrsra.2025.16.1.2094>

### Abstract

In this study, in experiments where groups are compared based on a binary variable, the Analysis of Means (ANOM), Wald, and Log-Likelihood Ratio tests for proportions were compared under comprehensive experimental conditions in terms of type I error rate and test power. As a result of 50,000 simulations, it was observed that the type I error rate for ANOM stayed within the 4.00% to 6.00% range under all experimental conditions. For the LR test, the type I error rate was negatively affected by small sample sizes ( $n=10$ , sometimes  $n=20$ ). For the Wald test, the type I error rate generally fell outside the 4.00%-6.00% limits and was quite irregular. In terms of test power, ANOM and the LR test were similar, although ANOM performed slightly better. Considering both type I error rate and test power together, ANOM was found to be considerably superior to the other two tests.

**Keywords:** ANOM; Monte Carlo Simulation; Type I Error; Test Power; Binary Outcome

### 1. Introduction

In group comparisons, the ANOVA-F test is commonly employed to assess differences among independent group means [1,2]. However, in practice, the dependent variable is not always quantitative. This is particularly true in studies involving human or animal behavior, where the data often follow a binomial distribution. In such cases, it is more appropriate to compare groups based on proportions rather than means, rendering the use of the ANOVA-F test unsuitable. For proportion comparisons, Z and t-tests are widely used. Nevertheless, these tests are limited to comparisons between only two independent groups [3]. When there are more than two groups, the Wald and the Log-Likelihood Ratio (LR) tests are more commonly applied [4, 5]. One alternative method suitable for such comparisons is the Analysis of Means (ANOM) [6]. The ANOM is a statistical hypothesis test that allows for the comparison of independent groups in terms of means, proportions [6], variances [7], correlation coefficients [8], and regression coefficients [9] and it presents the results graphically. Thanks to its graphical nature, interpreting the outcomes is relatively straightforward. Moreover, ANOM charts provide information about practical significance of observed differences. Despite its many advantages, the ANOM method remains relatively underutilized and unfamiliar to many researchers. Therefore, the primary objective of this study is to demonstrate the applicability of the ANOM method in proportion comparisons, and to evaluate and compare the performance of the ANOM, Wald, and LR tests under a variety of detailed conditions.

### 2. Material and Methods

The material of this study consists of random numbers generated from the Bernoulli distribution using the Monte Carlo simulation technique. In this study, the ANOM, Wald, and LR tests were compared in terms of Type I error rate and test

\* Corresponding author: SONER YİĞİT

power under varying conditions such as success probability (p), number of replications (n), number of groups (k), and effect size (δ). The experimental conditions considered are summarized in Table 1. Each condition was run 50,000 times. The Type I error rate was calculated by dividing the number of falsely rejected null hypotheses ( $H_0$ ) by 50,000. To compute the test power of the tests, one of the population proportions was made different from the others at varying degrees ( $\delta = 0.10, 0.15, \text{ and } 0.20$ ). Subsequently, the number of rejected null hypotheses divided by 50,000 was taken as the test power. The significance level was set at 5.00%. The Type I error rates of the tests were evaluated according to Cochran's [10] criterion (acceptable range: 4.00%–6.00%). The simulations were conducted using code written in the R [11] programming language.

**Table 1** Experimental conditions considered in the study

p	0.30, 0.50 and 0.70
n	10, 20, 30, 40, 50, 100, 200 and 400
k	3, 4, 5 and 10
δ	0.10, 0.15 and 0.20
Number of simulations	50000

**2.1. Statistical Tests**

*2.1.1. Wald Test*

$$W = (C\hat{\beta} - c)^T (C\widehat{Var}(\hat{\beta})C^T)^{-1} (C\hat{\beta} - c) \sim \chi_q^2$$

*2.1.2. were*

- $\hat{\beta}$  is the vector of estimated regression coefficients,
- $\widehat{Var}(\hat{\beta})$  is the estimated variance-covariance matrix of the coefficients,
- C is a  $q \times p$  contrast matrix specifying the linear hypotheses to be tested,
- c is a  $q \times 1$  constant vector (typically a zero vector, i.e., testing  $C\beta^{\wedge} = 0$ ),
- q is degree of freedom for independent variable.
- p is number of groups for independent variable.
- Under the null hypothesis, the test statistic W asymptotically follows a chi-squared distribution with q degrees of freedom [12].

*2.1.3. Log-Likelihood Ratio Test*

$$G^2 = -2[\log L_0 - \log L_1] \sim \chi_q^2$$

*2.1.4. were*

- $\log L_0$  is the log-likelihood of the reduced (null) model,
- $\log L_1$  is the log-likelihood of the full model.
- q is degree of freedom for independent variable.
- $G^2$  follows a chi-squared distribution with q degrees of freedom [4].

*2.1.5. ANOM for Proportions*

The ANOM produces a chart with upper (UDL) and lower decision (LDL) lines. If the proportion for any group falls outside these lines, it suggests that the group's proportion is significantly different from the overall proportion.

$$UDL = \bar{p} + h(\alpha, k, \infty) \sqrt{\bar{p}(1 - \bar{p})} \sqrt{(k - 1)/N}$$

$$LDL = \bar{p} - h(\alpha, k, \infty) \sqrt{\bar{p}(1 - \bar{p})} \sqrt{(k - 1)/N}$$

$$\bar{p} = \frac{\sum_{i=1}^k \hat{p}_i}{k}$$

$$\hat{p}_i = \frac{\sum_{j=1}^n X_i}{n}$$

2.1.6. were

- $h(\alpha, k, \infty)$  is critical table value of ANOM
- $k$  is number of groups,
- $N$  is total sample size,
- $\alpha$  is significance level,
- $\bar{p}$  is overall proportion,
- $\hat{p}_i$  is the proportion for the  $i$ th group.
- If the conditions  $n\hat{p}_i > 5$  and  $(1 - \hat{p}_i)n > 5$  are satisfied for all groups, the ANOM procedure designed for normally distributed data can be applied [6].

### 2.2. Example of ANOM for proportions

Three different bean varieties (A, B, and C) are compared in terms of their germination rates. For each variety, 20 seeds are planted. The germination rates for varieties A, B, and C are set at 0.20, 0.25, and 0.60, respectively. Let us compare the germination rates of these varieties using the ANOM (Analysis of Means) method.

$$\bar{p} = \frac{0.20 + 0.25 + 0.60}{3} = 0.35$$

$$h(0.05, 3, \infty) = 2.34$$

$$UDL = 0.35 + 2.34\sqrt{0.35(1 - 0.35)}\sqrt{(3 - 1)/60} = 0.5538$$

$$LDL = 0.35 - 2.34\sqrt{0.27(1 - 0.27)}\sqrt{(3 - 1)/60} = 0.1462$$

The ANOM chart shows that variety C falls outside the decision lines. The germination rate of variety C is statistically significantly higher than the overall proportion, whereas varieties A and B do not differ significantly from the overall rate.

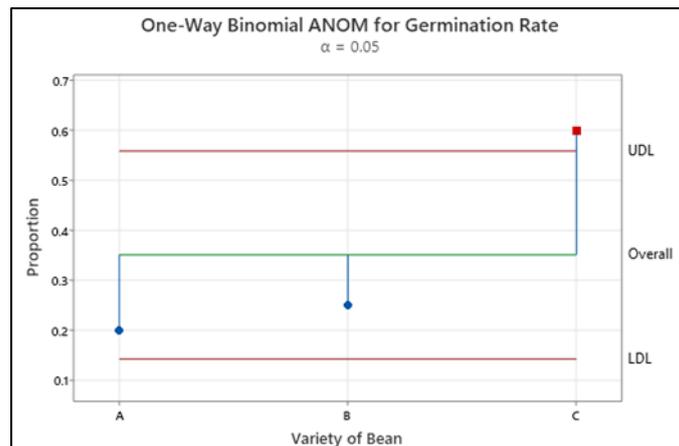


Figure 1 ANOM Chart for Germination Rate

### 3. Results and Discussion

The type I error rates for all tests were given in table 2. The type I error rates for the ANOM, LR and Wald tests were 4.05-5.97%, 4.57-8.86% and 0.22-5.35%, respectively. The type I error rates for the ANOM test were satisfy Cochran's criterion in all experimental conditions. For the LR test, Type I error rates exceeded the Cochran limits under all conditions when  $n = 10$ . Moreover, when  $k = 5$  and  $p = 0.7$ , as well as under all  $p$  values when  $k = 10$  and  $n = 20$ , the Type I error rates for the LR test consistently failed to fall within the acceptable 4.00%-6.00% range. For the Wald test, the Type I error rates were highly irregular and generally did not satisfy the Cochran criterion. The ANOM test satisfied the

Cochran criterion in all experimental conditions (100%), whereas the LR test did so in 83% and the Wald test in only 57% of the cases. This clearly indicates that ANOM is considerably more reliable compared to the other two tests.

**Table 2** Type I error rates for all conditions

p	n	k=3			k=4			k=5			k=10		
		ANOM	LR	Wald	ANOM	LR	Wald	ANOM	LR	Wald	ANOM	LR	Wald
0.3	10	5.58	8.46	1.21	4.05	8.40	0.50	4.12	8.20	0.63	4.70	8.86	0.28
	20	4.83	5.41	2.96	5.42	5.82	2.63	5.09	5.92	2.63	4.55	6.65	2.10
	30	5.05	5.69	4.20	5.30	5.74	3.73	4.67	5.38	3.36	4.37	5.72	2.94
	40	4.80	5.02	4.16	4.85	5.15	3.93	4.75	5.31	3.78	4.47	5.78	3.67
	50	5.16	5.40	4.56	5.12	5.66	4.58	4.66	5.69	4.60	4.66	5.10	3.40
	100	5.02	5.15	4.82	4.74	5.08	4.58	4.84	5.32	4.73	4.53	5.18	4.41
	200	5.07	5.03	4.79	4.80	4.91	4.63	4.72	4.90	4.68	4.95	4.93	4.61
	400	4.57	4.59	4.44	5.13	5.28	5.16	4.97	5.12	5.01	5.37	5.56	5.35
0.5	10	5.54	6.03	2.65	4.28	6.49	1.24	4.67	7.01	1.14	4.09	6.60	0.70
	20	5.97	5.33	5.04	4.56	5.32	3.35	4.87	5.78	3.66	4.57	6.10	2.96
	30	5.47	5.44	4.64	4.94	5.42	4.21	4.59	5.57	4.15	4.64	5.45	3.49
	40	4.32	5.46	4.49	4.52	4.95	4.33	4.54	5.18	4.19	4.24	4.90	3.70
	50	4.85	5.27	4.57	4.64	5.30	4.60	4.81	5.33	4.47	5.19	5.43	4.11
	100	5.22	5.04	4.82	5.00	5.20	4.87	4.56	4.98	4.65	4.88	5.14	4.50
	200	5.06	4.87	4.80	4.94	4.90	4.78	4.64	4.57	4.47	4.62	5.12	4.87
	400	4.87	4.93	4.90	4.98	4.97	4.91	4.46	4.92	4.89	4.90	5.06	4.93
0.7	10	5.61	8.40	1.32	4.07	8.50	0.48	4.09	8.51	0.67	4.99	8.59	0.22
	20	4.98	5.64	3.17	5.02	5.84	2.93	4.99	6.26	2.60	4.48	6.16	1.94
	30	5.02	5.54	3.98	5.12	5.55	3.33	4.26	5.16	3.16	4.52	5.70	2.95
	40	4.54	4.94	3.77	4.71	5.18	3.89	4.73	5.36	3.87	5.21	5.81	3.65
	50	5.10	5.33	4.44	4.99	5.46	4.40	5.01	5.65	4.46	4.74	5.24	3.47
	100	4.85	4.96	4.60	4.57	4.88	4.39	4.85	4.98	4.50	4.76	5.25	4.41
	200	5.39	5.20	4.98	4.88	4.93	4.70	5.01	4.82	4.62	5.00	5.14	4.87
	400	4.95	4.80	4.70	4.94	4.90	4.84	5.17	5.26	5.07	5.00	5.18	4.96

Test powers for all tests were given in table 3-5. It is known that test power depends on the magnitude of the differences between group proportions and the sample size. This situation is clearly observed in the present study as well. However, not all tests were equally affected by this situation. In general, the test power values for the ANOM and LR tests were similar, whereas those for the Wald test were highly irregular. An increase in the number of groups slightly reduced test power for all tests. When the maximum difference between group proportions was 0.10, and at  $p = 0.30$  and  $p = 0.70$ , all tests exceeded the 80.00% test power threshold at  $n = 400$ , regardless of the number of groups. At  $p = 0.50$  and  $n = 400$ , the test power for ANOM exceeded 80.00%, regardless of the number of groups. However, under the same conditions, neither the LR nor the Wald test reached the 80.00% test power level when  $k = 10$ , regardless of the number of replications (Table 3).

**Table 3** Test Power for  $\delta=0.10$

p	n	k=3			k=4			k=5			k=10		
		ANOM	LR	Wald	ANOM	LR	Wald	ANOM	LR	Wald	ANOM	LR	Wald
0.3	10	7.41	9.99	2.18	6.15	10.47	1.24	5.84	10.18	1.00	5.41	10.17	0.30
	20	9.84	10.27	7.61	9.39	9.97	6.37	8.58	9.91	5.48	7.34	9.17	3.46
	30	12.41	13.48	10.82	11.75	12.63	9.75	10.80	11.93	8.80	9.05	10.07	5.92
	40	15.28	15.45	13.88	14.05	14.94	12.89	13.60	13.55	11.23	11.36	11.44	8.38
	50	17.79	18.12	17.02	16.27	16.63	15.24	15.85	16.41	14.46	13.51	13.38	11.02
	100	32.12	32.12	31.46	31.92	31.40	30.72	30.78	29.53	28.67	26.59	22.91	21.45
	200	58.05	58.18	58.03	58.14	56.22	56.30	58.82	56.08	56.23	55.84	45.94	46.08
	400	87.68	87.70	87.79	89.47	88.06	88.15	89.45	86.97	87.11	90.01	82.01	82.44
0.5	10	7.47	8.21	3.68	5.85	8.53	1.75	5.71	8.40	1.45	4.55	8.41	0.98
	20	10.35	9.56	8.98	7.92	9.17	6.29	7.89	9.70	6.16	5.65	8.45	4.14
	30	11.84	11.92	10.53	10.11	11.45	9.65	9.39	10.23	8.19	7.31	8.92	5.78
	40	13.68	15.24	13.66	12.19	13.08	11.72	11.81	12.74	10.67	9.52	10.48	8.07
	50	15.64	16.72	15.20	14.48	15.90	14.33	13.77	15.04	13.19	11.44	12.36	10.17
	100	29.96	30.04	29.32	28.92	28.97	27.67	25.27	26.06	24.71	22.72	21.22	19.35
	200	53.12	53.11	52.63	52.48	52.77	52.13	54.03	51.88	51.00	47.81	40.55	39.07
	400	84.29	84.53	84.49	86.00	84.85	84.61	86.19	83.91	83.62	85.67	76.63	76.08
0.7	10	7.97	12.67	1.35	5.52	12.13	0.54	5.09	11.07	0.63	5.37	10.62	0.17
	20	10.48	11.49	5.60	9.74	11.73	5.03	8.06	10.01	3.64	6.02	10.10	2.87
	30	13.05	14.71	10.45	12.00	13.83	8.43	9.94	12.67	7.25	7.74	11.19	5.40
	40	16.37	17.49	13.81	13.87	16.28	11.99	12.85	15.13	10.47	9.58	12.92	7.81
	50	19.00	20.28	17.06	16.63	19.02	15.14	16.28	18.36	13.52	11.75	14.69	9.88
	100	36.28	37.24	35.16	34.38	35.66	32.70	31.76	33.48	29.63	25.60	26.03	20.68
	200	65.95	66.96	65.79	64.40	64.56	62.47	64.31	63.43	60.73	59.84	53.66	48.82
	400	93.09	93.20	93.02	94.22	93.61	93.02	94.67	93.12	92.34	94.29	88.86	86.84

When the difference between group proportions was 0.15 and n = 200, the test power values exceeded 80.00% for all tests regardless of the experimental conditions (Table 4).

**Table 4** Test Power for  $\delta=0.15$

p	n	k=3			k=4			k=5			k=10		
		ANOM	LR	Wald	ANOM	LR	Wald	ANOM	LR	Wald	ANOM	LR	Wald
0.3	10	11.02	13.59	4.41	8.44	13.40	2.31	7.77	12.8	1.72	7.80	12.49	0.51
	20	16.14	16.92	13.94	15.11	15.27	10.56	14.60	15.55	9.91	11.5	13.58	6.25
	30	21.80	22.63	19.98	21.32	21.53	18.44	19.84	19.96	16.51	17.31	16.41	11.24
	40	28.53	28.75	26.91	27.99	28.49	25.97	26.43	26.34	23.22	23.96	21.48	17.44
	50	35.59	35.61	34.36	35.34	34.56	32.86	33.51	32.53	30.61	30.07	25.64	22.52
	100	62.07	62.11	61.54	63.73	61.6	61.14	63.25	59.44	59.09	61.59	51.73	51.16
	200	90.86	90.57	90.53	92.42	91.33	91.44	92.80	90.25	90.42	92.79	85.61	85.88
	400	99.74	99.75	99.76	99.86	99.79	99.79	99.92	99.81	99.81	99.91	99.57	99.62
0.5	10	9.67	10.67	4.87	8.02	11.43	2.67	7.71	11.46	2.18	5.81	10.57	0.93
	20	14.92	14.49	13.33	13.57	15.37	11.42	12.67	15.38	10.21	8.83	12.15	6.28
	30	20.68	21.06	18.79	18.11	19.83	16.97	16.62	18.50	15.04	14.16	15.73	10.31
	40	26.55	28.79	26.73	24.29	25.15	22.54	23.05	24.45	21.17	19.84	20.03	15.64
	50	31.97	33.88	31.08	30.53	31.90	29.35	29.00	29.72	26.56	24.80	23.53	19.69
	100	59.54	59.81	59.14	60.15	59.3	57.71	58.86	57.14	55.36	54.51	46.32	43.36
	200	89.12	89.33	89.01	90.34	89.35	88.97	90.99	88.93	88.3	90.19	83.16	81.78
	400	99.58	99.59	99.59	99.80	99.69	99.69	99.80	99.63	99.62	99.90	99.31	99.26
0.7	10	10.67	17.43	1.25	7.17	16.35	0.96	6.75	14.96	0.61	6.33	13.42	0.24
	20	18.76	21.35	8.75	15.93	20.43	6.78	13.56	18.01	5.39	8.92	15.97	3.94
	30	26.06	29.37	19.98	23.99	27.35	15.30	19.37	25	12.35	14.41	20.29	8.32
	40	34.54	37.43	30.28	30.93	35.23	25.60	29.49	33.55	21.58	21.08	25.76	13.29
	50	42.28	45.14	39.49	39.02	43.05	34.90	37.64	40.83	31.11	29.55	31.85	18.89
	100	74.19	75.66	73.42	74.96	75.91	72.13	74.14	73.57	68.49	69.47	64.11	54.33
	200	96.94	97.16	96.82	97.71	97.55	97.12	97.98	97.62	97.07	98.01	95.10	92.91
	400	99.99	99.99	99.98	99.98	99.98	99.98	99.99	99.99	99.99	100.00	99.98	99.98

When the maximum difference was 0.20 and  $n = 100$ , the test power values for ANOM exceeded 80.00%, regardless of the number of groups and  $p$ . Under the same conditions, for  $p = 0.5$  and  $k = 10$ , the test power values for the LR and Wald tests exceeded 80.00% when  $n \geq 200$  (Table 5).

**Table 5** Test Power for  $\delta=0.20$

p	n	k=3			k=4			k=5			k=10		
		ANOM	LR	Wald	ANOM	LR	Wald	ANOM	LR	Wald	ANOM	LR	Wald
0.3	10	15.44	17.72	7.30	12.11	17.37	3.79	11.73	17.14	3.37	11.25	15.49	0.85
	20	24.60	25.08	21.86	25.44	25.45	20.20	23.23	24.01	17.69	21.20	19.49	10.51
	30	35.75	37.34	34.11	36.14	35.63	32.06	35.24	33.66	29.41	31.62	27.33	20.82
	40	47.10	47.02	45.53	45.23	45.26	42.65	46.52	44.53	41.52	43.74	36.43	32.14
	50	55.85	55.79	54.63	57.01	55.48	53.92	56.24	53.22	51.26	54.95	44.75	41.47
	100	86.48	86.24	86.02	88.26	86.69	86.48	88.67	86.06	85.85	89.34	80.57	80.46
	200	99.32	99.29	99.30	99.60	99.39	99.38	99.62	99.32	99.35	99.78	99.00	99.04
	400	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
0.5	10	14.28	15.72	7.06	11.72	16.35	3.60	10.40	15.09	2.68	7.75	13.77	1.60
	20	25.28	25.14	22.92	22.24	23.80	17.44	19.57	22.24	14.44	14.38	18.49	8.98
	30	35.41	36.28	32.47	31.98	33.69	29.10	31.09	33.08	27.23	25.72	26.14	17.84
	40	44.35	46.42	43.60	43.93	44.89	41.17	41.58	42.49	38.02	37.83	34.31	27.09
	50	54.28	55.62	52.50	53.98	54.72	51.62	52.56	52.07	47.53	49.14	42.78	36.14
	100	86.11	86.56	85.86	87.33	86.57	85.58	86.48	84.54	83.30	86.04	78.12	75.00
	200	99.35	99.38	99.32	99.53	99.44	99.38	99.68	99.34	99.30	99.77	98.77	98.54
	400	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
0.7	10	14.82	25.60	1.21	10.78	25.53	0.82	8.13	21.86	0.77	7.33	17.84	0.30
	20	31.70	36.58	11.51	28.41	35.81	9.17	25.30	33.26	7.65	14.59	26.64	4.58
	30	48.06	52.15	33.27	45.76	51.11	24.73	39.47	47.38	19.29	30.92	38.18	11.69
	40	60.90	65.25	53.12	58.49	63.59	46.35	57.98	62.40	39.69	47.87	50.35	22.39
	50	73.84	76.53	70.17	71.61	75.20	63.58	70.79	72.77	58.25	65.40	62.13	34.64
	100	96.96	97.33	96.66	97.28	97.39	96.25	97.93	97.54	96.05	97.50	94.81	89.30
	200	99.98	99.98	99.98	100.00	100.00	100.00	100.00	99.99	99.98	100.00	99.99	99.96
	400	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

**4. Conclusion**

According to the simulation results, considering both the Type I error rates and test power values, the superiority of ANOM over the LR and Wald tests is quite evident. In addition, since it is a graphical method, its results are very easy to interpret. Moreover, it provides information not only about statistical significance but also about practical significance. Therefore, when all these factors are considered together, the use of the ANOM method appears to be highly advantageous in experiments where groups are compared in terms of a binary variable.

**References**

[1] Zar, J.H. (1999). *Biostatistical Analysis*. Prentice-Hall Inc., Simon and Schuster/A Viacom Company, New Jersey, NJ.

[2] Yiğit, S. (2019). ANOM ve ANOMTR testlerinin performansları bakımından karşılaştırılması. *Türkiye Tarımsal Araştırmalar Dergisi*, 6(2), 193-198. <https://doi.org/10.19159/tutad.524099>

- [3] Sheskin, D.J. (2000). *Handbook of Parametric and Nonparametric Statistical Procedures* (2nd ed.). Chapman and Hall/CRC, New York.
- [4] McCullagh, P., and Nelder, J.A. (1989). Binary data. In *Generalized Linear Models* (pp. 98–148). Springer US.
- [5] Dobson, A.J. (2013). *Introduction to Statistical Modelling*. Springer.
- [6] Nelson, P.R., Wludyka, P.S., and Copeland, K.A.F. (2005). *The Analysis of Means: A Graphical Method for Comparing Means, Rates and Proportions*. SIAM, Philadelphia.
- [7] Bernard, A.J., and Wludyka, P.S. (2001). Robust I-sample analysis of means types randomization tests for variances. *Journal of Statistical Computation and Simulation*, 69(1), 57–88.
- [8] Rao, C.V., and Kumar, M.P. (2002). ANOM-type graphical methods for testing the equality of several correlation coefficients. *Gujarat Statistical Review*, 29, 47–56.
- [9] Rao, C.V., and Deva Raaj, V.J. (2006). ANOM for testing the equality of several intercepts in a k-sample regression model. *Statistical Methods*, 8(1), 60–72.
- [10] Cochran, W. G. (1952). The X<sup>2</sup> test of goodness of fit. *Annals of Mathematical Statistics*, 23, 315–345.
- [11] R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
- [12] Pustejovsky, J.E., and Tipton, E. (2018). Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business and Economic Statistics*, 36(4), 672–683. <https://doi.org/10.1080/07350015.2016.1247004>