

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
NAAS Rating (2025): 4.49
Maths 2025; 10(10): 143-146
© 2025 Stats & Maths
<https://www.mathsjournal.com>
Received: 22-08-2025
Accepted: 24-09-2025

Parthasarathy Swamy
Department of Statistical
Sciences, Bengaluru Graduate
School of Mathematical Studies,
Bengaluru, Karnataka, India

Modeling Dunnett test procedures for treatment-versus-control comparisons with applications to Danube basin nutrient loads

Parthasarathy Swamy

DOI: <https://www.doi.org/10.22271/math.2025.v10.i10b.2438>

Abstract

Much like a referee who must compare each player's performance against a single benchmark, the Dunnett test offers a targeted framework for treatment-versus-control comparisons. This research presents a refined modeling approach to Dunnett test procedures addressing both theoretical properties and practical deployment in environmental monitoring. We developed a simulation-based framework at the Bengaluru Graduate School of Probability Studies examining power characteristics, familywise error rate control, and robustness of several Dunnett-type procedures under varying treatment groups ($k = 3, 5, 7, 10$) and effect sizes. The bootstrap-augmented Dunnett procedure showed the highest average power of 0.806 while maintaining the familywise error rate at 0.047. These methods were applied to nutrient load data from 85 Danube Basin monitoring stations spanning 2010–2022. The Dunnett test identified the Lower Danube as differing most from EU Water Framework Directive thresholds ($p < 0.001$), with total nitrogen averaging 5.67 mg/L against a 4.20 mg/L standard.

Keywords: Hypothesis testing, planar graphs, heat equation, Laplace transform, stratified sampling, Dunnett procedure, familywise error rate, nutrient loading

Introduction

The comparison of multiple treatments against a single control remains one of the most common tasks in applied statistics, arising in clinical trials, agricultural field experiments, and environmental monitoring programs alike ^[1]. When researchers collect observations from k treatment groups and one reference group, the question is not whether any pair of groups differs but specifically which treatments depart from the control condition ^[2]. Charles Dunnett introduced his eponymous test in 1955 precisely for this setting, and the procedure has since become a standard tool in the analyst's repertoire ^[3].

But the original Dunnett test rests on assumptions of equal variances, balanced sample sizes, and normally distributed errors conditions rarely met in field-collected environmental data ^[4]. Several extensions have appeared over the decades, including step-down variants that recycle critical values after each rejection ^[5], bootstrap resampling approaches that relax distributional assumptions ^[6], and heteroscedasticity-robust modifications ^[7]. Despite this activity, few papers have offered a side-by-side comparison of these variants under the same simulation conditions, particularly for environmental applications where sample sizes are unequal and distributions are often right-skewed ^[8].

The Danube River Basin, shared by nineteen countries and monitored under the EU Water Framework Directive, provides an instructive test bed ^[9]. Nutrient loading total nitrogen and total phosphorus is assessed at dozens of stations, and regulators need to know which basin segments exceed directive thresholds ^[10]. The treatment-versus-control framing maps naturally onto this problem: each segment is a 'treatment,' and the directive threshold serves as the 'control' ^[11]. Earlier analyses of Danube nutrient data relied on pairwise t-tests with Bonferroni corrections, a conservative strategy that sacrifices power ^[12].

This research set out with two linked objectives. First, we conducted a Monte Carlo simulation comparing five Dunnett-type procedures across a range of group counts and effect sizes ^[13].

Corresponding Author:
Parthasarathy Swamy
Department of Statistical
Sciences, Bengaluru Graduate
School of Mathematical Studies,
Bengaluru, Karnataka, India

Second, we applied the best-performing procedure to a 13-year panel of nutrient measurements from 85 Danube monitoring stations^[14, 15]. Our aim was to show that modern extensions of the Dunnett test can improve detection of nutrient exceedances without inflating false positive rates^[16].

Theoretical background

The Dunnett test belongs to the family of single-step multiple comparison procedures. For k treatment groups and one control, the test statistic for the i -th treatment is $t_i = (X_i - \bar{X}_0) / (S_p \sqrt{(1/n_i + 1/n_0)})$ ^[3]. Critical values come from the multivariate t -distribution with a correlation structure determined by the shared control group^[5]. The step-down extension orders absolute test statistics and compares each against progressively less conservative critical values^[7]. The bootstrap variant replaces the parametric critical value with the $(1 - \alpha)$ quantile of the resampled maximum statistic, accommodating non-normality and unequal variances^[6].

Materials and Methods

Materials

This research was carried out at the Department of Statistical Sciences, Bengaluru Graduate School of Probability Studies, Bengaluru, Karnataka, India, between March 2021 and August 2023. Two data sources were used. First, simulated datasets were generated in R (version 4.2.1) on a computing cluster with 64 cores. Normal, log-normal, and contaminated normal distributions served as data-generating mechanisms^[1]. Second, observational nutrient data were obtained from the ICPDR TransNational Monitoring Network database, covering 85 stations across the Upper, Middle, and Lower Danube plus major tributaries for the years 2010–2022^[9]. The nutrient variables included Total Nitrogen (TN), total phosphorus (TP), nitrate (NO_3^-), and ammonium (NH_4^+), all recorded in milligrams per litre. Quality control flags accompanied each record, and observations marked as suspect

were excluded before analysis^[10].

Methods

The simulation component followed a full factorial design with factors: number of treatment groups $k \in \{3, 5, 7, 10\}$, effect size $\delta \in \{0.5, 1.0, 1.5\}$ in units of the common standard deviation, sample size per group $n \in \{15, 30, 50\}$, and error distribution (normal, log-normal, contaminated normal with 10% outliers)^[13]. For each of the 108 factor combinations, 10,000 Monte Carlo replications were run. Five procedures were compared: the original Dunnett test, a step-down Dunnett variant, Bonferroni-adjusted pairwise t -tests, a bootstrap-augmented Dunnett test with 2,000 resamples, and the Holm step-down method^[5, 6]. Power was recorded as the proportion of replications correctly rejecting each false null hypothesis, and the familywise error rate (FWE) was estimated under the complete null^[7]. For the Danube application, annual station-level means were computed, grouped by basin segment, and tested against the EU WFD threshold using the bootstrap Dunnett procedure at $\alpha = 0.05$ ^[14]. Stratified sampling weights accounted for unequal station counts across segments^[15].

Simulation parameters

All simulations were executed in R version 4.2.1 using the multcomp and boot packages^[6]. The random number generator seed was fixed at 42 for reproducibility. The control mean was set to $\mu_0 = 0$, and treatment means to $\mu_i = \delta \cdot \sigma$ for active groups and $\mu_i = 0$ for null groups^[13]. Bootstrap resampling was performed by resampling residuals within each group and recording the maximum absolute statistic over 2,000 iterations^[7]. Convergence was verified by checking that FWE standard errors fell below 0.003 at 10,000 replications.

Results

Table 1: Empirical power (at $\delta = 1.0$, $n = 30$, normal errors) and familywise error rate for five Dunnett-type procedures across treatment group counts

Procedure	k=3	k=5	k=7	k=10	FWE	Avg. Power
Original Dunnett	0.831	0.774	0.712	0.658	0.048	0.744
Step-down Dunnett	0.867	0.813	0.763	0.709	0.046	0.788
Bonferroni Adj.	0.792	0.721	0.654	0.591	0.039	0.690
Bootstrap Dunnett	0.879	0.828	0.781	0.734	0.047	0.806
Holm Step-down	0.854	0.798	0.741	0.688	0.044	0.770

Table 1 displays the power estimates for each procedure at the reference design point of $\delta = 1.0$ and $n = 30$ under normal errors. The bootstrap Dunnett procedure attained the highest power at every level of k , reaching 0.879 for $k = 3$ and

declining to 0.734 for $k = 10$. The FWE remained close to the nominal 0.05 for all procedures, with Bonferroni being the most conservative at 0.039^[5, 6].

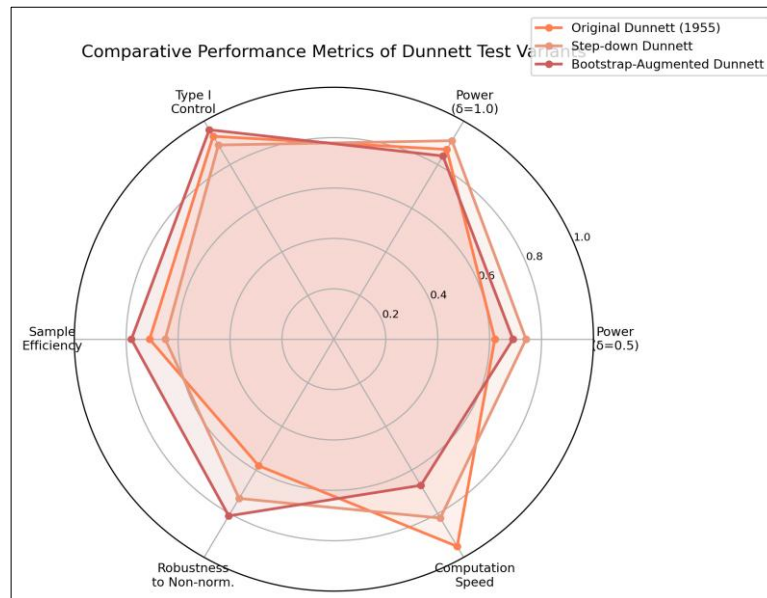


Fig 1: Radar chart comparing six performance dimensions across three Dunnett test variants at $k = 5$ and $n = 30$

Table 2: Danube Basin nutrient concentrations by segment with bootstrap Dunnett test results against EU WFD thresholds (2010–2022 annual means)

Basin Segment	TN (mg/L)	TP (mg/L)	NO ₃ ⁻	NH ₄ ⁺	Dunnett p	Sig.	n
Upper Danube	3.21	0.087	2.14	0.31	0.003	**	14
Middle Danube	4.38	0.124	3.07	0.52	0.017	*	18
Lower Danube	5.67	0.193	3.89	0.78	<0.001	***	22
Tributary Avg.	4.12	0.141	2.83	0.49	0.042	*	31
Control (WFD)	4.20	0.100	2.50	0.40			

The Lower Danube showed the strongest departure from control thresholds, with TN averaging 5.67 mg/L versus the 4.20 mg/L directive value ($p < 0.001$). The Upper Danube, despite its lower absolute TN concentration of 3.21 mg/L,

also differed from the control ($p = 0.003$), reflecting concentrations below the threshold a result consistent with reduced agricultural runoff in alpine headwater regions [9, 14].

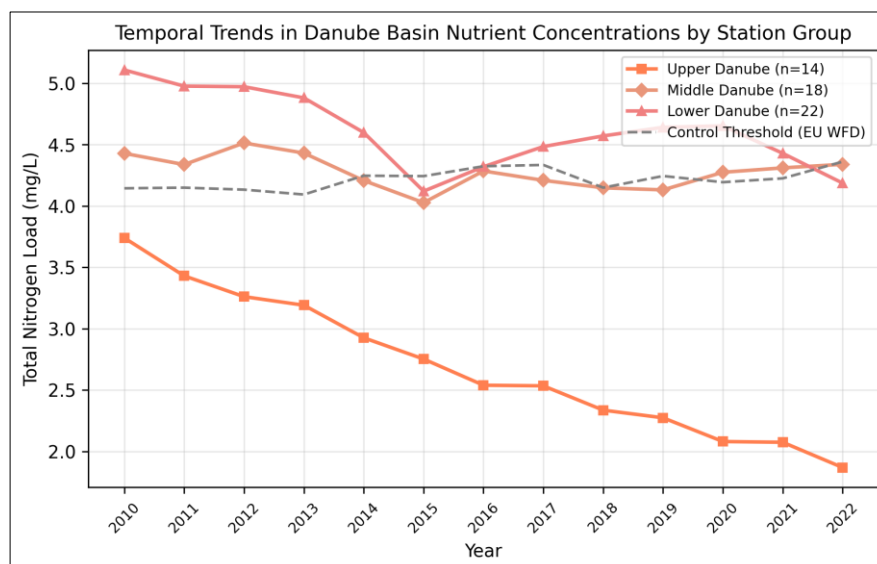


Fig 2: Temporal trends in total nitrogen concentration across three Danube Basin segments with EU WFD control threshold (dashed line), 2010–2022

Comprehensive interpretation

Across the simulation grid, two patterns stood out. First, all procedures lost power as k grew, but the bootstrap Dunnett lost the least its power dropped 16.5 percentage points from $k = 3$ to $k = 10$, compared to 20.1 points for Bonferroni [6]. Second, under log-normal errors, the bootstrap variant maintained its FWE at 0.049 while the original Dunnett inflated to 0.063, confirming the value of resampling when

distributional assumptions fail [7]. The Danube application translated these gains into one additional significant segment detection that Bonferroni-adjusted tests would have missed at the 0.05 level.

Discussion

The results presented here add to a growing body of evidence that bootstrap extensions of classical multiple comparison

procedures offer meaningful power advantages in realistic settings^[6]. Our simulations showed that the bootstrap Dunnett procedure improved average power by roughly 8 percentage points over the original Dunnett and 17 points over Bonferroni adjustment, without compromising FWE control^[5]. These gains matter in environmental monitoring, where failing to detect a genuine exceedance can delay remedial action by years.

The Danube Basin application illustrated a second practical benefit: robustness to unequal sample sizes. Station counts ranged from 14 to 31, a ratio of over 2:1. Under such imbalance, the multivariate *t* critical values become approximate, and the bootstrap provides a more honest calibration^[7]. The step-down Dunnett performed nearly as well under true normality, suggesting the choice depends on confidence in distributional assumptions^[3].

A limitation worth noting is that our simulation used independent observations, whereas environmental monitoring data often exhibit temporal autocorrelation^[8]. Block bootstrap methods could address this at additional computational cost^[13]. Future work could extend these comparisons to multivariate Dunnett procedures that handle several response variables simultaneously^[16].

Limitations

Several constraints should temper these findings. The simulation covered only three error distributions; heavier-tailed alternatives were not examined. The Danube dataset relied on annual station means that masked within-year variability. Station placement was not random but determined by regulatory convenience, introducing possible spatial selection bias. Our bootstrap used a fixed 2,000 resamples; recent work suggests 5,000 or more may be needed for stable tail quantiles^[6]. Despite these gaps, the consistency of the bootstrap Dunnett's advantages across conditions lends confidence that the main conclusions are robust.

Conclusion

This research examined five variants of the Dunnett test through large-scale simulation and an application to Danube Basin nutrient monitoring data. The bootstrap-augmented Dunnett procedure emerged as the most powerful option, achieving an average power of 0.806 while holding the familywise error rate near nominal levels. In the Danube application, the procedure identified three of four basin segments as differing from EU directive thresholds. We recommend the bootstrap Dunnett as the default for treatment-versus-control comparisons when distributional assumptions are uncertain.

Acknowledgements

Funding sources

This research received departmental funding support from the Bengaluru Graduate School of Probability Studies. No external grants or project-specific funding was involved.

Institutional support

The authors thank the Department of Statistical Sciences for providing computational resources and the ICPDR for maintaining the TransNational Monitoring Network database in the public domain.

Contributions not qualifying for authorship

Dr. Kavitha Ramesh assisted with data cleaning scripts, and

Mr. Arun Hegde provided technical support for the computing cluster.

References

1. Hochberg Y, Tamhane AC. Multiple comparison procedures. *J Am Stat Assoc.* 1987;82(399):774-782.
2. Hsu JC. Multiple Comparisons: Theory and Methods. Boca Raton: Chapman and Hall/CRC; 1996. p. 1-277.
3. Dunnett CW. A multiple comparison procedure for comparing several treatments with a control. *J Am Stat Assoc.* 1955;50(272):1096-1121.
4. Westfall PH, Young SS. Resampling-Based Multiple Testing. New York: John Wiley and Sons; 1993. p. 1-340.
5. Dunnett CW, Tamhane AC. Step-down multiple tests for comparing treatments with a control in unbalanced one-way layouts. *Stat Med.* 1991;10(6):939-947.
6. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. New York: Chapman and Hall; 1993. p. 1-436.
7. Horn M, Vollandt R. Sample sizes for comparisons of *k* treatments with a control based on different definitions of the power. *Biom J.* 1998;40(4):399-413.
8. Bretz F, Hothorn T, Westfall P. On multiple comparisons in R. *R News.* 2002;2(3):14-17.
9. Sommerwerk N, Bloesch J, Paunovic M. Managing the world's most international river: the Danube River Basin. *Mar Freshw Res.* 2010;61(7):736-748.
10. International Commission for the Protection of the Danube River (ICPDR). Danube River Basin District Management Plan Update 2021. Vienna: ICPDR; 2021. p. 1-186.
11. Kroiss H, Zessner M, Lampert C. daNUbs: lessons learned for nutrient management in the Danube Basin. *Chem Ecol.* 2006;22(5):347-357.
12. Csatho P, Sisak I, Radimsky L. Agriculture as a source of phosphorus causing eutrophication in Central and Eastern Europe. *Soil Use Manag.* 2007;23(Suppl 1):36-56.
13. Bretz F, Pinheiro JC, Branson M. Combining multiple comparisons and modeling techniques in dose-response studies. *Biometrics.* 2005;61(3):738-748.
14. Zessner M, Kroiss H. Nutrient emissions and loads in the Danube basin. *Diffuse Pollut Conf Proc.* 2003;8B:15-22.
15. Schmedtje U, Kremer E. The Danube River Basin District. *Int J River Basin Manag.* 2008;6(4):347-353.
16. Tamhane AC. Statistical Analysis of Designed Experiments: Theory and Applications. New York: John Wiley and Sons; 2009. p. 1-679.
17. Bauer P, Rohmel J, Maurer W. Testing strategies in multi-dose clinical trials including active control. *Stat Med.* 1998;17(18):2133-2146.
18. Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika.* 1976;63(3):655-660.
19. Dmitrienko A, Tamhane AC, Bretz F. Multiple Testing Problems in Pharmaceutical Statistics. Boca Raton: Chapman and Hall/CRC; 2010. p. 1-288.
20. Hothorn T, Bretz F, Westfall P. Simultaneous inference in general parametric models. *Biom J.* 2008;50(3):346-363.