DUNCAN MULTIPLE-RANGE TEST

Authored by

mohammad looti

October 25, 2025

DUNCAN MULTIPLE-RANGE TEST

Primary Disciplinary Field(s): Statistics, Experimental Design, Quantitative Psychology

1. Core Definition and Purpose

The **Duncan Multiple-Range Test** (DM-RT), often simply referred to as Duncan's Test, is a post-hoc statistical procedure designed for performing multiple comparisons among sets of means following a significant omnibus test, typically an Analysis of Variance (ANOVA). Its primary function is to systematically determine which specific pairs of treatment means are statistically different from one another, thereby locating the source of the variance observed in the initial ANOVA.

The necessity for the DM-RT arises directly from the statistical issue known as the multiple comparisons problem. If a researcher conducts numerous pairwise t-tests without adjusting for the increased number of comparisons, the overall probability of committing a **Type I error**--falsely rejecting a true null hypothesis--skyrockets. The Duncan Multiple-Range Test attempts to manage this inflation of the familywise error rate (FWER) while maintaining a relatively high degree of statistical power, which distinguishes it from more conservative tests like the Tukey Honestly Significant Difference (HSD) test.

In essence, the DM-RT serves as a bridge, allowing researchers to transition from the general conclusion that "at least one mean differs" (provided by ANOVA) to the specific conclusion that "Mean A differs significantly from Mean C, but not from Mean B." It achieves this by comparing the observed differences between means against a series of critical range values that depend on the number of means spanned by the comparison, incorporating an element of sequential testing that is central to its operation.

2. Historical Development and Proponent

The Duncan Multiple-Range Test was developed by statistician David B. Duncan (1916-2006) and was formally introduced in 1955. Duncan's work built upon the foundation laid by statisticians working on range tests, particularly those involving the studentized range distribution, such as Newman and Keuls. The primary impetus behind Duncan's creation was the perceived overly conservative nature of existing multiple comparison procedures at the time, which often resulted in a high risk of committing **Type II errors** (falsely failing to detect a real difference).

Duncan argued that the error rate for a specific comparison should not be based on the total number of means in the experiment (as in FWER control), but rather on the number of steps separating the specific pair of means being compared after they have been ranked. This concept led to the development of the "protection level" approach. Historically, the DM-RT gained immense popularity quickly, especially within agricultural, biological, and psychological research fields, due

to its ability to identify more significant differences than its contemporaries, thus appealing to researchers seeking greater statistical power.

However, the DM-RT's unique method of error control soon became a major point of academic contention. While Duncan aimed to balance Type I and Type II errors, critics argued that the procedure did not adequately control the familywise error rate, particularly as the number of group means increased. Despite these methodological debates, the DM-RT remains a historically significant procedure and is still utilized in certain contexts where high sensitivity to detecting differences is prioritized over stringent FWER control.

3. Methodological Approach: Sequential Testing and Range Calculation

The core methodology of the Duncan Multiple-Range Test relies on the calculation of a series of minimum significant ranges. Unlike procedures that use a single critical value (like Tukey's HSD), the DM-RT determines a unique critical range for every possible subset of means being compared. This is why it is classified as a sequential range test. The first step involves ranking all group means from smallest to largest.

Once ranked, the difference between any two means is compared against a critical range whose value depends on the number of steps (or means) separating them in the ranked order. For example, the critical range required to declare a difference between Mean 1 and Mean 4 (spanning four means) will be larger than the critical range required for Mean 1 and Mean 2 (spanning two means). This approach is based on the studentized range statistic, which incorporates the mean square error (MSE) from the ANOVA--a pooled estimate of variance--and the sample sizes.

The formula for calculating the significant range (R_p) for a comparison spanning p means is defined by $R_p = q_{alpha, p, nu} sqrt{MSE/n}$, where $q_{alpha, p, nu}$ is the value of the studentized range statistic corresponding to the significance level $alpha$, the number of means spanned p, and the degrees of freedom for the error term nu. If the absolute difference between two means exceeds this calculated significant range ($|bar{X}_i - bar{X}_j| > R_p$), the difference is declared statistically significant. This sequential, step-down process allows for a more nuanced assessment of pairwise differences than non-sequential methods.

4. Error Control Strategy: The "Protection Level"

The fundamental distinction of the DM-RT lies in its approach to controlling the Type I error rate, often referred to as the **"protection level"** or the "shortest significant range." Duncan specifically structured the test so that the significance level for a given comparison is conditional on the results of the comparisons of all smaller subsets of means contained within that range. This conditional nature means that the test does not control the overall familywise error rate (FWER) at the nominal alpha level ($alpha$) across all comparisons.

Instead, the DM-RT controls the comparison-wise error rate at $alpha$ for any single comparison, but more importantly, it ensures that the probability of declaring a significant difference between two means that are truly equal is $alpha$ only if all intermediate means are also equal. For p means, the overall significance level used is $alpha' = 1 - (1 - alpha)^{p-1}$. This adjustment makes the test increasingly liberal as the number of means increases. For instance, if $alpha = 0.05$, a comparison spanning five means effectively operates at an error rate closer to $1 - (0.95)^4 approx 0.185$, meaning the probability of one or more false positives across the family of comparisons is much higher than 0.05.

This calculated risk is precisely what Duncan intended, aiming for a test that had a higher chance of detecting true differences. By accepting a higher risk of FWER inflation, the test maximizes **power**. However, this methodological choice has led to the DM-RT being largely superseded in many scientific fields by tests that strictly control the FWER, as conservative FWER control is often prioritized, especially in medical or regulatory research.

5. Key Characteristics and Assumptions

Dependence on ANOVA Significance: Like most post-hoc tests, the DM-RT should theoretically only be applied if the overall ANOVA F-test yields a statistically significant result. This pre-condition helps ensure that there is at least some detectable difference among the group means to justify further pairwise exploration.

Equal Sample Sizes (Balanced Design): While variations exist for unbalanced designs, the original and most straightforward application of the DM-RT assumes equal sample sizes (n) across all treatment groups. Unequal sample sizes require complex adjustments, typically using the harmonic mean of the sample sizes, which can slightly reduce the test's power.

Assumptions of Parametric Tests: As a parametric procedure, the DM-RT inherits the standard assumptions of ANOVA: **normality** of the underlying population distributions for each group, and **homogeneity of variances** (the variances of the groups being compared must be approximately equal). Violations of these assumptions, especially homogeneity of variance, can compromise the validity of the pooled mean square error used in the range calculation.

Sequential Logic: The application requires a rigid sequential process. Once a range is declared non-significant, all comparisons nested within that range must also be considered non-significant, regardless of their calculated differences. This sequential rule structure maintains the conditional nature of the test's error control.

6. Significance and Application in Research

Historically, the Duncan Multiple-Range Test held a position of prominence, particularly in

disciplines focused on optimizing conditions or comparing specific experimental treatments where the detection of even small effects was highly valued. Its applications were widespread in agricultural research (e.g., comparing crop yields across different fertilizer treatments), pharmaceutical testing (comparing drug efficacies), and educational studies (comparing learning outcomes across teaching methodologies).

The significance of the DM-RT lies in its contribution to the evolution of post-hoc analysis. It represented an important attempt to address the power deficit inherent in extremely conservative tests. For researchers studying phenomena with potentially small effect sizes, the higher power offered by Duncan's test was a crucial advantage, allowing them to identify differences that might have been masked by procedures with stricter FWER control.

Despite modern statistical packages offering more robust alternatives, the DM-RT persists in certain applied fields, often due to established departmental traditions or standardized reporting procedures. In these areas, the explicit trade-off between increased power and relaxed FWER control is consciously accepted, provided the research context minimizes the practical consequences of an inflated Type I error rate.

7. Debates and Criticisms

The Duncan Multiple-Range Test has been a consistent subject of statistical debate since its introduction, primarily centered on its error rate control. The central criticism is that the DM-RT fails to adequately control the **familywise error rate** (FWER). Statisticians who advocate for strict FWER control argue that researchers must have confidence that the probability of making at least one false discovery across the entire experiment does not exceed the nominal $alpha$ level (e.g., 0.05).

Because the DM-RT uses an increasingly liberal critical value as the range of means increases, it guarantees only that the comparison-wise error rate is controlled, leading to an FWER that can be significantly higher than the nominal $alpha$ level, especially when comparing a large number of groups (e.g., $k>6$). This makes the test prone to declaring differences as significant when they are, in reality, due to chance variation.

Consequently, in many academic disciplines, the DM-RT has been largely replaced by more rigorous procedures such as the Tukey HSD test (which controls FWER conservatively) or the Holm-Bonferroni method (which is sequential but controls FWER more accurately). Modern guidelines often recommend against the use of the DM-RT in contexts where minimizing false positives is critical, favoring methods that provide clearer control over the total number of erroneous conclusions drawn from the dataset.

8. Further Reading

Wikipedia: David B. Duncan
Wikipedia: Multiple comparisons problem
Purdue University Statistics: Post Hoc Procedures (PDF)
Duncan, D. B. (1955). Multiple range and multiple F tests. Biometrics, 11(1), 1-42.