

How to Control False Discovery Rate with the Benjamini-Hochberg Procedure

Authored by
stats writer

December 29, 2025

RECOMMENDED CITATION

stats writer (2025). *How to Control False Discovery Rate with the Benjamini-Hochberg Procedure*. PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=109701>

The Benjamini-Hochberg Procedure (B-H Procedure) stands as a foundational method in modern statistics, specifically designed to address the challenges inherent in large-scale data analysis. Its primary purpose is to rigorously control the False Discovery Rate (FDR) when researchers are evaluating multiple hypotheses simultaneously. Unlike traditional methods that focus on controlling the family-wise error rate, the B-H Procedure provides a powerful balance, allowing for more discoveries while maintaining strict control over the proportion of significant findings that turn out to be false positives. This adjustment technique operates by modifying the individual p-values derived from each test, ensuring that the cumulative FDR remains below a predetermined threshold, thereby significantly enhancing the overall power of the analytical framework.

In fields ranging from genomics to finance, researchers routinely perform hundreds or even thousands of statistical tests within a single study. When conducting such extensive parallel testing, the risk of committing a Type I error--a false positive--accumulates rapidly. The B-H procedure offers a scientifically sound methodology to manage this inflation of error, effectively reducing the number of erroneous declarations of significance and ensuring the reliability of high-throughput experimental results. Understanding and correctly applying this procedure is paramount for any researcher engaged in large-scale multiple testing, ensuring that reported discoveries are robust and reproducible.

The Challenge of Multiple Comparisons

The core premise of hypothesis testing rests on the calculated risk of error. When conducting any single statistical test, researchers establish a significance level, often denoted as alpha (α), conventionally set at 0.05. This means that, even if the null hypothesis (H_0) is genuinely true--meaning there is no underlying effect--there is still a 5% chance of observing a p-value below 0.05 purely due to random chance. This accidental rejection of a true null hypothesis is known as a Type I error, or a **false positive**. This inherent statistical risk is manageable when only one test is performed, but it becomes statistically untenable when the volume of tests increases.

Consider a simple example illustrating this fundamental concept. Suppose a research team is investigating whether a specific plant species has a mean height significantly greater than 10 inches. They define their hypotheses formally as:

H_0 : $\mu = 10$ inches (The mean height is 10 inches)

H_A : $\mu > 10$ inches (The mean height is greater than 10 inches)

To test this hypothesis, the researchers collect a sample, perhaps measuring 20 plants. Even if the population's true mean height is exactly 10 inches, the random sampling process might inadvertently select 20 plants that are unusually tall simply by luck. If this highly unrepresentative

sample produces a sufficiently small p-value (e.g., less than 0.05), the team would be led to reject the null hypothesis. In this scenario, they would conclude that the mean height is greater than 10 inches, despite the fact that the underlying truth--the **null hypothesis**--was correct. This incorrect rejection is precisely what statisticians define as a **false discovery**: a claim of a significant result that is statistically spurious.

Scaling Statistical Risk: Why Standard Alpha Fails

The crucial issue arises when researchers move from conducting a single test to performing a vast number of parallel tests. If an individual test has a 5% chance of resulting in a false discovery (Type I error), this risk accumulates dramatically across multiple tests. Consider a study where 100 independent statistical tests are conducted, all using the conventional alpha level (α) of 0.05. While the probability of error for any single test remains 5%, the expectation is that approximately 5 of those 100 tests will yield statistically significant results that are, in fact, false positives. The vast increase in the number of tests performed drives an unacceptable inflation of the overall error rate.

In contemporary research environments, driven by advanced technology and massive datasets, this problem has become pervasive. Fields like genomics, neuroimaging, and machine learning often require analyzing thousands or even tens of thousands of variables simultaneously. For instance, medical researchers might run statistical comparisons across thousands of genes to find markers associated with a disease. If 10,000 tests are performed with a standard significance threshold, one could anticipate 500 spurious associations--500 false discoveries--being reported as significant findings. Such a high volume of false positives can severely mislead subsequent research, waste resources, and undermine scientific credibility.

The traditional method for controlling error in multiple testing, known as the Family-Wise Error Rate (FWER) control (such as the Bonferroni correction), is overly conservative. While FWER methods ensure the probability of making even one false discovery across the entire set of tests is low, they severely reduce the power to detect genuine effects. This conservative nature leads to a high rate of Type II errors (false negatives), causing researchers to miss true discoveries. Therefore, a more robust and statistically powerful solution is required to balance these two types of errors effectively, leading us directly to the concept of the False Discovery Rate (FDR) and its control mechanism: the **Benjamini-Hochberg Procedure**.

Executing the Benjamini-Hochberg Procedure

The Benjamini-Hochberg Procedure provides a streamlined, four-step approach to controlling the expected proportion of false discoveries among all rejected null hypotheses. This method, often referred to as the BH-FDR procedure, is inherently sequential and requires careful calculation of specific critical values based on the rank of each individual p-value. It is crucial that the desired

False Discovery Rate (Q) is determined and fixed before any data analysis or calculation begins.

The steps for calculating the adjusted significance threshold are as follows:

Step 1: Conduct Testing and Extract P-values. First, perform all required statistical tests (e.g., m tests) and meticulously record the raw p-value associated with each test.

Step 2: Rank the P-values. The observed p-values must be sorted in ascending order, from the smallest value to the largest value. Each p-value is then assigned a rank, denoted by i , where the smallest p-value receives a rank of 1, the next smallest receives a rank of 2, and so on, up to the total number of tests, m .

Step 3: Calculate the Critical Threshold. For every ranked p-value, calculate the corresponding Benjamini-Hochberg critical value. This adjusted critical threshold, which changes for each rank, is determined by the formula:

$$\text{Critical Value} = \frac{i}{m} \times Q$$

Where:

i = The specific rank of the p-value (from 1 to m).

m = The total number of hypotheses tested.

Q = The researcher's pre-selected maximum acceptable **False Discovery Rate (FDR)**.

Step 4: Determine Significance via Backward Procedure. Starting from the largest p-value (rank m) and moving backward sequentially, identify the highest rank k such that the observed p-value (p_k) is less than or equal to its calculated critical value. Once this critical point k is identified, **all hypotheses corresponding to p-values ranked from 1 up to rank k are declared statistically significant.**

This powerful sequential procedure ensures that the overall proportion of Type I errors remains controlled at the specified level Q . The following practical example demonstrates how to apply these calculations using concrete data.

Detailed Example: Applying the Procedure to Medical Research

To solidify the understanding of the Benjamini-Hochberg Procedure, let us walk through a practical scenario. Imagine a team of medical researchers investigating the association between 20 different genetic or environmental factors (variables) and the risk of heart disease. They perform 20 distinct hypothesis tests ($m=20$) concurrently. The first step involves collecting the raw p-values from these 20 tests and arranging them in ascending order, as mandated by Step 2 of the procedure. The resulting ranked p-values are displayed below, showing the variable and its associated rank i .

Variable	P-Value	Rank
Variable #12	0.001	1
Variable #17	0.034	2
Variable #3	0.035	3
Variable #11	0.039	4
Variable #1	0.094	5
Variable #10	0.106	6
Variable #4	0.145	7
Variable #13	0.145	8
Variable #16	0.223	9
Variable #9	0.241	10
Variable #18	0.315	11
Variable #20	0.338	12
Variable #7	0.432	13
Variable #14	0.478	14
Variable #6	0.497	15
Variable #8	0.566	16
Variable #15	0.654	17
Variable #2	0.876	18
Variable #5	0.905	19
Variable #19	0.965	20

For this study, the researchers have decided that they are willing to accept a maximum 20% False Discovery Rate (FDR), meaning $Q = 0.20$. Based on the total number of tests ($m=20$) and the selected FDR, the formula used to calculate the B-H critical value for any given rank i simplifies to: $\text{Critical Value} = (i/20) \times 0.20$. This calculation is performed for all 20 ranks to establish the dynamic significance thresholds.

The table below presents the results of these calculations, listing the ranked p-value alongside its calculated B-H critical value. It is important to note that the critical value increases linearly with the rank, providing a less stringent threshold for larger p-values. This increasing threshold is what distinguishes the B-H procedure from the constant, highly conservative threshold used in methods like the Bonferroni correction.

Variable	P-Value	Rank	$(i/m)*Q$
Variable #12	0.001	1	0.010
Variable #17	0.034	2	0.020
Variable #3	0.035	3	0.030
Variable #11	0.039	4	0.040
Variable #1	0.094	5	0.050
Variable #10	0.106	6	0.060
Variable #4	0.145	7	0.070
Variable #13	0.145	8	0.080
Variable #16	0.223	9	0.090
Variable #9	0.241	10	0.100
Variable #18	0.315	11	0.110
Variable #20	0.338	12	0.120
Variable #7	0.432	13	0.130
Variable #14	0.478	14	0.140
Variable #6	0.497	15	0.150
Variable #8	0.566	16	0.160
Variable #15	0.654	17	0.170
Variable #2	0.876	18	0.180
Variable #5	0.905	19	0.190
Variable #19	0.965	20	0.200

Determining the Significance Threshold (Step 4)

The final, crucial step of the Benjamini-Hochberg Procedure involves identifying the cut-off point for significance by comparing the calculated B-H critical value against the observed p-value at each rank, moving backward from the largest rank ($i=20$). We search for the largest rank, denoted as k , where the observed p-value (p_k) is less than or equal to its corresponding critical value. Scanning the results from the previous table, we observe that the test corresponding to Variable #11 (Rank $i=4$) is the largest p-value that satisfies this condition. Its p-value is 0.039, which is less than its B-H critical value of 0.040.

The application of the B-H rule requires working backward from the highest rank ($i=20$). We search for the highest rank k where the p-value p_k is less than or equal to the critical value CV_k . In this dataset, the cutoff point is established by Variable #11, which holds rank $k = 4$. With a p-value of 0.039 and a corresponding B-H critical value of 0.040, this condition is met ($0.039 \leq 0.040$). Since Variable #11 is the highest-ranked test satisfying the condition, we declare Variable #11 and all tests with smaller ranks ($i=1, 2$, and 3) as statistically significant, resulting in four total discoveries.

	Variable	P-Value	Rank	$(i/m)*Q$
Significant	Variable #12	0.001	1	0.010
	Variable #17	0.034	2	0.020
	Variable #3	0.035	3	0.030
	Variable #11	0.039	4	0.040
	Variable #1	0.094	5	0.050
	Variable #10	0.106	6	0.060
	Variable #4	0.145	7	0.070
	Variable #13	0.145	8	0.080
	Variable #16	0.223	9	0.090
	Variable #9	0.241	10	0.100
	Variable #18	0.315	11	0.110
	Variable #20	0.338	12	0.120
	Variable #7	0.432	13	0.130
	Variable #14	0.478	14	0.140
	Variable #6	0.497	15	0.150
	Variable #8	0.566	16	0.160
	Variable #15	0.654	17	0.170
	Variable #2	0.876	18	0.180
	Variable #5	0.905	19	0.190
	Variable #19	0.965	20	0.200

The selection of $k=4$ means that although other tests with higher ranks (like Variable #3 at Rank 5) might also have p-values below their respective critical values, they are not declared significant because the backward procedure stops at the critical rank k . Only those tests ranked 1 through 4 are confirmed as significant findings under the 20% False Discovery Rate.

Strategic Selection of the False Discovery Rate (Q)

The selection of the maximum acceptable False Discovery Rate (Q) is perhaps the most critical subjective decision in the Benjamini-Hochberg Procedure. Researchers must establish this value **a priori**--before any data collection or hypothesis testing takes place--to maintain the integrity and objectivity of the analysis. The choice of Q reflects the scientific context and the specific consequences associated with making a false discovery versus missing a true one (Type I vs. Type II error trade-off).

In practice, a higher False Discovery Rate (e.g., $Q=0.20$, as used in the example) is often acceptable during the **exploratory phase** of research, particularly when the initial analysis involves screening hundreds or thousands of variables. A higher Q increases the power of the study, meaning more potentially significant associations are identified. This approach is justified if the research design includes subsequent validation steps, such as low-cost follow-up experiments or specialized confirmatory statistical tests. If the downstream cost of correcting a false discovery is low, researchers often prefer a higher FDR to maximize the initial detection yield.

Conversely, a lower Q (e.g., $Q=0.01$ or $Q=0.05$) is necessary when the cost of a false discovery is extremely high. For instance, in clinical trials where a false positive could lead to unnecessary treatment or dangerous medical recommendations, maintaining a strict control over error is essential. Furthermore, if the cost of missing a genuinely important discovery (Type II error) is deemed extremely high, setting a slightly higher FDR may be warranted to improve sensitivity. Ultimately, the optimal False Discovery Rate is determined by a careful consideration of the research costs, ethical implications, and the scientific importance of the discoveries being sought.

Summary of Advantages

The Benjamini-Hochberg procedure provides significant advantages over older methods like the Bonferroni correction, primarily because it directly controls the expected proportion of false positives among the list of significant findings, rather than controlling the probability of making *any* false positive. This focus on the False Discovery Rate allows researchers to achieve a superior balance between minimizing Type I errors and maximizing statistical power, making it the preferred method for modern high-throughput studies. By adjusting the critical value based on the rank of the observed p-value, it maintains robustness while promoting more true discoveries.

The structured, sequential nature of the Benjamini-Hochberg Procedure ensures that the control over false discoveries is rigorous and transparent, irrespective of the dependence structure among the hypotheses being tested (a critical advantage in complex datasets). Its widespread adoption across scientific disciplines underscores its role as an indispensable tool for reliable interpretation of results derived from simultaneous hypothesis testing.