HIGH SCHOOL EDITION
Journal of Student Research
Volume 14 Issue 1 (2025)

P-Value or False Discovery Rate When Two-Sample t-Test is Employed Instead of Paired t-Test

Meryem Bourget[1] and Cyril Rakovski[#]

[1]Troy High School, USA
[#]Advisor

ABSTRACT

When performing several tests on the same dataset (i.e., multiple comparison testing), the False Discovery Rate (FDR) is commonly used instead of p-values. If all the assumptions of the two-sample independent t-test are met, the p-value is commonly used to test if there is a difference between the population means. What if the two samples are correlated, and a two-sample independent test is carried out instead of the paired t-test? Should we consider p-values or FDRs to make decisions regarding hypotheses? This research explores two methodologies (p-values and FDR) for different magnitudes of correlation between two samples and sample sizes when one hypothesis test (not multiple) is performed. The two strategies are tested using a 100,000-run simulation with correlation coefficients ranging from -0.9 to 0.9 for various sample sizes. The simulation results reveal that if there is no correlation, both approaches are equally valid, which is predicted. However, if a correlation exists, the FDR is recommended to avoid making less erroneous decisions.

Introduction

William Sealy Gosset (1908) introduced a new distribution and test for small sample sizes because existing methods were developed for large sample sizes and normal distribution. He published his findings under the pseudonym "Student" since his company, the Guinness Brewery in Dublin, Ireland, prohibited employees from using their actual names in publications. Gosset observed that the new distribution (the t-distribution) had a symmetric bell shape like the normal distribution but with a heavier tail (Encyclopedia of Britannica). Ronald Fisher (1925) was the first to use the terms Student's distribution and t-tests in literature.

The significance level (α) determines whether an effect is statistically significant. Ronald Fisher (1925) introduced the p-value as a statistical significance test. Assuming the null hypothesis is correct, the p-value is defined as the probability of finding a test statistic result that is extreme or more. Although the p-value and α are closely related, there is a small difference. The p-value for the test is computed based on the data, while the value of α is determined before the experiment.

The data analysis includes the steps of formulating the research question, identifying null and alternative hypotheses, determining appropriate test statistics and their distribution under the null hypothesis, specifying α, calculating the test statistics from the data, computing the p-value for the test statistics based on the region determined by the alternative hypothesis, and finally rejecting the null hypothesis when the p-value is less than the value of α. When *multiple inference procedures* (MCP) are employed simultaneously, a single inference approach increases the false positive significance rate (type I error rate). John Tukey (1953) developed the concept of a familywise error rate (FWER) to control type I errors (Hochberg and Tamhane, 1987; Hochberg and Benjamini, 1990). Benjamini and Hochberg (1995) proposed an alternative technique to FWER, the false discovery rate (FDR), which addresses the expected proportion of incorrectly rejected null hypotheses.

Should the FDR be used in a *single inference* test if the testing strategy is incorrect? The two-sample independent t-test examines the mean differences between two independent populations. When two samples are paired,

the paired t-test is proposed by Snedecor and Cochran (1989, p. 83). This study examines the performance of the FDR and p-value using simulation when a two-sample independent t-test is employed instead of a paired t-test.

Method

Two-Sample t-Test with Equal Variances

The two-sample independent t-test assumes that two samples are selected from normal distributions using simple random sampling (i.e., observation independence) and that both populations are independent with equal variances. A two-sided (two-tailed) hypothesis is defined as

$$\mathrm{H}_0 : \mu_1 - \mu_2 = 0$$

$$\mathrm{H}_\mathrm{a} : \mu_1 - \mu_2 \neq 0,$$

where μ_1 and μ_2 are the means of the populations 1 and 2, respectively. The test statistics is

$$t^* = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

with pooled sample variance

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

where $n_1, \bar{x}_1$, and s_1^2 are the sample size, mean, and variance for sample 1, respectively. Similarly, $n_2, \bar{x}_2$, and s_2^2 are defined for sample 2. Under the null hypothesis, t^*follows t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

Paired t-Test

A paired t-test, also known as a correlated or dependent test, is a statistical test designed to assess whether the mean difference between two groups is equal. The dependent variable is defined as the difference between two sets of observations. Thus, the assumptions are based on the dependent variable rather than the original data. The assumptions are that the observations within each group are independent and that the dependent variable has an approximately normal distribution. A two-sided hypothesis is defined as

$$H_0 : \mu_d = 0$$

$$H_a : \mu_d \neq 0,$$

where $\mu_d$is the mean difference. The paired t-test is defined as

$$t^* = \frac{\bar{x}_d}{s_d/\sqrt{n}}$$

where $n, \bar{x}_d$, and s_d are the sample size, mean, and standard deviation of the dependent variable, respectively. Under the null hypothesis, t^*follows t-distribution with $n-1$ degrees of freedom.

Types of Errors

Table 1 shows possible errors that could occur in m hypothesis tests. Any statistical testing procedure can result in two sorts of errors. Type I errors occur when the null hypothesis is rejected given that it is true. This error is also known as a False Positive (FP). The type II error fails to reject the null hypothesis when the alternative hypothesis is true. This kind of error is commonly known as a False Negative (FN).

In Table 1, let U denote the number of True Negatives (TN), V the number of False Positives (FP), T the number of False Negatives (FN), and R the number of True Negatives (TN). The False Positive Rate (FPR) is the percentage of positive cases that are incorrectly classified as positive. Alternatively, it is the percentage of people who do not have the disease but are diagnosed as having it. The FPR is defined

$$FPR = \frac{V}{V+U} = \frac{V}{m_0}$$

FPR and α are often used interchangeably in the literature, but there is a slight difference. The α is a predefined value before the experiment, while FPR is an observed value that takes into account α. Are the p-value and FPR the same? The two terms should not be used interchangeably since the p-value is a probability calculated from data for a specific test, whereas the FPR is a proportion derived in the context of a classification problem.

The FDR measures the proportion of false discoveries among significant findings. That is the proportion of people identified as having the disease who do not have the disease. The FDR is defined as

$$FDR = \frac{V}{V+S} = \frac{V}{R}$$

The FDR is not the same as the p-value. The FDR measures the proportion of false positive results among overall positive test results, whereas the p-value calculates false positives across all tests. For example, a p-value of 0.01 indicates that one percent of all tests will produce FPs. An FDR of 0.01 states that one percent of significant discoveries will produce FPs.

Table 1. Possible types of errors in the statistical testing procedure.

	Declared non-significant (identified as not having disease)	**Declared significant (Identified as having disease)**	**TOTAL**
H_o True (doesn't have disease)	Correct Decision True negative (TN) U	Type I Error False Positive (FP) V	m_0
H_a True (has disease)	Type II Error False Negative (FN)	Correct Decision True Positive (TP)	$m - m_0$

	T	S	
TOTAL	$m - R$	R	m

Results

The simulation program was written using R, a free software environment for statistical computing and graphing. We used 100,000 data sets to test the null and alternative hypotheses. Using the built-in function "rnorm_multi", we generated two correlated samples from normal distributions, with means μ_1 and μ_2, and variances σ_1^2 and σ_2^2. For purposes of simplicity, we assumed $n_1 = n_2 = n$ and $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Sample sizes of $n = 5,\ 10,\ 15,\ 20,\ 30$ and $n = 50$ were considered. The correlation coefficient, ρ, ranged from -0.9 to 0.9 with a step of 0.1 Under the null hypothesis, samples 1 and 2 had normal distributions with $\mu_1 = \mu_2 = \mu = 0$ with $\sigma_1^2 = \sigma_2^2 = \sigma^2 = 1$ with various correlations. Under the alternative hypothesis, sample 1 had a normal distribution with $\mu_1 = 0$ with $\sigma_1^2 = 1,$ while sample 2 had a normal distribution with $\mu_2 = 1$ with $\sigma_1^2 = 1$ with different correlations. Hence, the hypotheses are

$$\mathrm{H}_0 : \mu_1 - \mu_2 = 0$$

$$\mathrm{H}_a : \mu_1 - \mu_2 = 1$$

The goal of this study is to investigate the p-value and FDR if the classical t-test is incorrectly utilized while the proper approach is the paired t-test. The p-values were obtained from the built-in function "t.test".

Table 2 shows the results of the methods when the two samples are negatively correlated for small to large sample sizes. The values in the table are interpreted in the following way: For $n = 5$ and $\rho = -0.9$, the p-value of 0.161 indicates that 16.1% of all tests produce false positives, while the FDR of 0.31 implies that 31% of significant discoveries result in false positives. When the correlation coefficient increases in absolute value, the FDR and p-values increase for all sample sizes. While the sample size appears to have almost no effect on the p-value for a given correlation, that is, the p-value produces similar results for a given correlation coefficient as the sample size increases, the FDR produces smaller values. For example, when $\rho = -0.5,\ n = 5,\ 10,\ 15,\ 20, 30$ and $n = 50$, the p-values are $0.104, 0.108, 0.108, 0.111, 0.109, 0.108$, whereas the FDR values are $0.243, 0.163, 0.131, 0.120, 0.104, 0.098$. For small to moderate sample sizes ($n = 5, 10, 15,\ 20$), the FDR provides a larger value than the p-value for any given ρ. However, for large sample sizes, the opposite is a fact.

Table 3 displays the outcomes of the methods when the two samples are positively correlated. When the correlation coefficient increases, the FDR and p-values decrease for all sample sizes, as opposed to the results in negative correlation case. While the sample size does not appear to affect the p-value for a given correlation, the FDR decreases as the sample size increases. For small to moderate sample sizes, the FDR provides a larger value than the p-value for any given ρ. However, for large sample sizes, both produce comparable results. For small to moderate values of ρ (between 0.1 and 0.4) and small to moderate sample sizes, the FDR is slightly bigger than the p-value. However, when the sample size and correlation coefficient increase, both approaches produce the same results. Both methods are compatible with large sample sizes and any magnitude of ρ.

Table 2. Comparison of p-value and FDR when two samples are negatively correlated.

n = 5 ρ									
METHOD	-0.9	-0.8	-0.7	-0.6	-0.5	-0.4	-0.3	-0.2	-0.1
P-VAL	0.161	0.145	0.129	0.117	0.104	0.091	0.079	0.069	0.060
FDR	0.310	0.291	0.275	0.260	0.243	0.223	0.203	0.186	0.169
n = 10 ρ									
METHOD	-0.9	-0.8	-0.7	-0.6	-0.5	-0.4	-0.3	-0.2	-0.1
P-VAL	0.160	0.148	0.134	0.119	0.108	0.093	0.083	0.072	0.060
FDR	0.225	0.212	0.195	0.177	0.163	0.144	0.130	0.115	0.098
n = 15 ρ									
METHOD	-0.9	-0.8	-0.7	-0.6	-0.5	-0.4	-0.3	-0.2	-0.1
P-VAL	0.157	0.145	0.134	0.121	0.108	0.096	0.084	0.073	0.060
FDR	0.185	0.172	0.159	0.146	0.131	0.118	0.104	0.090	0.075
n = 20 ρ									
METHOD	-0.9	-0.8	-0.7	-0.6	-0.5	-0.4	-0.3	-0.2	-0.1
P-VAL	0.156	0.147	0.133	0.121	0.111	0.097	0.084	0.073	0.061
FDR	0.164	0.155	0.141	0.129	0.120	0.104	0.091	0.079	0.066
n = 30 ρ									
METHOD	-0.9	-0.8	-0.7	-0.6	-0.5	-0.4	-0.3	-0.2	-0.1
P-VAL	0.156	0.144	0.132	0.121	0.109	0.097	0.085	0.072	0.061
FDR	0.146	0.136	0.125	0.115	0.104	0.093	0.082	0.070	0.060
n = 50 ρ									
METHOD	-0.9	-0.8	-0.7	-0.6	-0.5	-0.4	-0.3	-0.2	-0.1

P-VAL	0.158	0.145	0.133	0.121	0.108	0.097	0.085	0.072	0.061
FDR	0.138	0.128	0.119	0.109	0.098	0.089	0.078	0.068	0.058

Table 3. Comparison of p-value and FDR when two samples are positively correlated.

$n = 5$ ρ										
METHOD	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
P-VAL	0.050	0.041	0.034	0.026	0.020	0.014	0.010	0.005	0.002	0.001
FDR	0.148	0.129	0.110	0.089	0.072	0.053	0.037	0.020	0.010	0.002
$n = 10$ ρ										
METHOD	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
P-VAL	0.050	0.039	0.031	0.022	0.015	0.010	0.005	0.002	0.000	0.000
FDR	0.082	0.065	0.052	0.037	0.025	0.017	0.008	0.003	0.001	0.000
$n = 15$ ρ										
METHOD	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
P-VAL	0.051	0.040	0.029	0.022	0.014	0.009	0.004	0.001	0.000	0.000
FDR	0.063	0.050	0.037	0.027	0.017	0.010	0.005	0.002	0.000	0.000
$n = 20$ ρ										
METHOD	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
P-VAL	0.050	0.040	0.029	0.021	0.014	0.007	0.003	0.001	0.000	0.000
FDR	0.054	0.043	0.032	0.023	0.015	0.008	0.003	0.001	0.000	0.000
$n = 30$ ρ										
METHOD	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
P-VAL	0.049	0.039	0.028	0.021	0.012	0.007	0.003	0.001	0.000	0.000
FDR	0.048	0.038	0.028	0.021	0.012	0.007	0.003	0.001	0.000	0.000
$n = 50$ ρ										
METHOD	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
P-VAL	0.050	0.037	0.029	0.020	0.012	0.006	0.002	0.001	0.000	0.000
FDR	0.048	0.036	0.028	0.020	0.012	0.006	0.002	0.001	0.000	0.000

Conclusion

In multiple comparison tests, the FWER is commonly employed to control type I errors. Benjamini and Hochberg (1995) proposed the FDR as an alternative to the FWER for reducing erroneous rejections of null hypotheses while increasing the likelihood of identifying real effects. For example, while a p-value of 0.05 or lower can be utilized to detect significant results, the FDR will be set to greater than 0.05. The FDR has gained popularity over the FWER in genomic research for evaluating hundreds of genes at the same time to identify significant genes that cause disease. However, when an improper method is employed, the FDR has not been considered yet in literature, as far as we know.

This paper investigates the p-value and FDR for various sample sizes, as well as the correlation between two normally distributed samples using simulation when the t-test is used instead of the paired t-test. The simulation findings show that sample size, strength, and correlation signs have an impact on methods. As sample size increases, the values between FDR and p-value decrease. For some correlations and sample sizes, both techniques provide similar results. In conclusion, the FDR is recommended for decision-making when a paired t-test was expected to be used but the t-test was used in analyzing data.

Acknowledgments

I would like to thank Dr. Cyril Rakovski for guiding me in this research.

References

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)*, 57(1), 289 - 300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Encyclopedia Britannica. Student's t-test. https://www.britannica.com/science/Students-t-test

Fisher RA (1925). Application of student's distribution new tables for testing the significance of observations expansion of student's integral in powers of n-1. *Metron,* 5, 90 - 104

Hochberg, Y. and Tamhane, A. (1987). Multiple Comparison Procedures. New York: Wiley.

Hochberg, Y. & Benjamini, Y. (1990). More powerful procedures for multiple significance testing. *Statist. Med.*, 9, 811 - 818

Snedecor , G. W. & Cochran , W. G. (1989). Statistical Methods, 8th ed., Ames , IA: Iowa State University Press.

Student (1908). The Probable Error of a Mean. *Biometrika*. 6 (1), 1- 25. https://doi.org/10.2307/2331554

Tukey, J. W. (1953). The problem of multiple comparisons. *In Mimeographed Notes*, Princeton, NJ: Princeton University