Chapter 615

Multiple Testing for Two Means

Introduction

This chapter describes how to estimate power and sample size (e.g., number of arrays in a microarray experiment) for 2 group (two-sample) high-throughput studies using the Multiple Testing for Two Means procedure. False discovery rate and experiment-wise error rate control methods are available in this procedure. Values that can be varied in this procedure are power, false discovery rate and experiment-wise error rate, sample sizes (numbers of arrays) in each group, the minimum |mean difference| detected, the standard deviation(s), and in the case of false discovery rate control, the number of tests with minimum |mean difference| > δ.

Two-Sample Design

In a two-sample design, two groups are compared, which we will call Treatment 1 and Treatment 2. Several experimental units are randomly assigned to each of the two treatment groups. In the microarray scenario, a single mRNA or cDNA sample is obtained from each experimental unit of both groups. Each sample is exposed to a single microarray, resulting in a single expression value for each gene for each unit of each treatment group. The goal is to determine for each gene whether there is evidence that the expression is different between the two groups.

Null and Alternative Hypotheses

The two-sample null and alternative hypotheses are described here in terms of treatment groups: Treatment 1 and Treatment 2. These groups could equally be labeled Treatment A and Treatment B, Control and Treatment, etc. The two-sample null hypothesis for each test (e.g., gene) is $H_0: \mu_1 = \mu_2$, where μ_1 is the actual mean (expression for a particular gene) in the Treatment 1 environment, and μ_2 is the actual mean (expression for a particular gene) in the Treatment 2 environment. The alternative hypothesis may be any one of the following: $H_1: \mu_1 < \mu_2$, $H_1: \mu_1 > \mu_2$, or $H_1: \mu_1 \neq \mu_2$. The choice of the alternative hypothesis depends upon the goals of the research. For example, if the goal of a microarray experiment is only to determine which genes are up-regulated (increase in expression) over Treatment 1 when Treatment 2 is imposed, the alternative hypothesis would be $H_1: \mu_1 < \mu_2$. If the goal instead is to determine which genes are differentially expressed (up-regulated or down-regulated) when compared to the other treatment, the alternative hypothesis is $H_1: \mu_1 \neq \mu_2$.

Assumptions

The following assumptions are made when using the two-sample Z-test, T-test, or the Mann-Whitney *U* or Wilcoxon Rank-Sum test. One of the reasons for the popularity of the T-test is its robustness in the face of assumption violation. However, if an assumption is not met even approximately, the significance levels and the power of the T-test are unknown. You should take the appropriate steps to check the assumptions before you make important decisions based on these tests.

Two-Sample Z-Test Assumptions

The assumptions of the two-sample Z-test are:

1. The data are continuous (not discrete).
2. The data follow the normal probability distribution.
3. The variances of the two populations are equal. (If not, the Unequal-Variance test is used.)
4. The two samples are independent. There is no relationship between the individuals in one sample as compared to the other (as there is in the paired Z-test).
5. Both samples are simple random samples from their respective populations. Each individual in the population has an equal probability of being selected in the sample.
6. The standard deviation(s) is(are) known.

Two-Sample T-Test Assumptions

The assumptions of the two-sample T-test are:

1. The data are continuous (not discrete).
2. The data follow the normal probability distribution.
3. The variances of the two populations are equal. (If not, the Aspin-Welch Unequal-Variance test is used.)
4. The two samples are independent. There is no relationship between the individuals in one sample as compared to the other (as there is in the paired T-test).
5. Both samples are simple random samples from their respective populations. Each individual in the population has an equal probability of being selected in the sample.

Mann-Whitney U or Wilcoxon Rank-Sum Test Assumptions

The assumptions of the Mann-Whitney *U* or Wilcoxon Rank-Sum test for difference in means are:

1. The variable of interest is continuous (not discrete). The measurement scale is at least ordinal.
2. The probability distributions of the two populations are identical, except for location. That is, the variances are equal.
3. The two samples are independent.
4. Both samples are simple random samples from their respective populations. Each individual in the population has an equal probability of being selected in the sample.

Technical Details

Multiple Testing Adjustment

When the two-sample T-test is run for a replicated microarray experiment, the result is a list of P-values (Probability Levels) that reflect the evidence of difference in expression. When hundreds or thousands of genes are investigated at the same time, many 'small' P-values will occur by chance, due to the natural variability of the process. It is therefore requisite to make an appropriate adjustment to the P-value (Probability Level), such that the likelihood of a false conclusion is controlled.

Benjamini and Hochberg's (1995) False Discovery Rate Table

The following table (adapted to the subject of microarray data) is found in Benjamini and Hochberg's (1995) false discovery rate article. In the table, m is the total number of tests, m_0 is the number of tests for which there is no difference in expression, R is the number of tests for which a difference is declared, and U, V, T, and S are defined by the combination of the declaration of the test and whether or not a difference exists, in truth.

	Declared Not Different	Declared Different	Total
A true difference in expression does not exist	U	V	m_0
There exists a true difference in expression	T	S	$m - m_0$
Total	$m - R$	R	m

In the table, the m is the total number of hypotheses tested (or total number of genes) and is assumed to be known in advance. Of the m null hypotheses tested, m_0 is the number of tests for which there is no difference in expression, R is the number of tests for which a difference is declared, and U, V, T, and S are defined by the combination of the declaration of the test and whether or not a difference exists, in truth. The random variables U, V, T, and S are unobservable.

Need for Multiple Testing Adjustment

Following the calculation of a raw P-value (Probability Level) for each test, P-value adjustments need be made to account in some way for multiplicity of tests. It is desirable that these adjustments minimize the number of genes that are falsely declared different (V) while maximizing the number of genes that are correctly declared different (S). To address this issue the researcher must know the comparative value of finding a gene to the price of a false positive. If a false positive is very expensive, a method that focuses on minimizing V should be employed. If the value of finding a gene is much higher than the cost of additional false positives, a method that focuses on maximizing S should be used.

Error Rates – P-Value Adjustment Techniques

Below is a brief description of three common error rates that are used for control of false positive declarations. The commonly used P-value adjustment technique for controlling each error rate is also described.

Per-Comparison Error Rate (PCER) – No Multiple Testing Adjustment

The per-comparison error rate (PCER) is defined as

$$PCER = E(V)/m,$$

where *E*(*V*) is the expected number of genes that are falsely declared different, and *m* is the total number of tests. Preserving the PCER is tantamount to ignoring multiple testing altogether. If a method is used which controls a PCER of 0.05 for 1,000 tests, approximately 50 out of 1,000 tests will falsely be declared significant. Using a method that controls the PCER will produce a list of genes that includes most of the genes for which there exists a true difference in expression (i.e., maximizes *S*), but it will also include a very large number of genes which are falsely declared to have a true difference in expression (i.e., does not appropriately minimize *V*). Controlling the PCER should be viewed as overly weak control of Type I error.

To obtain P-values (Probability Levels) that control the PCER, no adjustment is made to the P-value. To determine significance, the P-value is simply compared to the designated alpha.

Experiment-Wise Error Rate (EWER)

The experiment-wise error rate (EWER) is defined as

$$EWER = \Pr(V > 0),$$

where *V* is the number of genes that are falsely declared different. Controlling EWER is controlling the probability that a single null hypothesis is falsely rejected. If a method is used which controls a EWER of 0.05 for 1,000 tests, the probability that any of the 1,000 tests (collectively) is falsely rejected is 0.05. Using a method that controls the EWER will produce a list of genes that includes a small (depending also on sample size) number of the genes for which there exists a true difference in expression (i.e., limits *S*, unless the sample size is very large). However, the list of genes will include very few or no genes that are falsely declared to have a true difference in expression (i.e., stringently minimizes *V*). Controlling the EWER should be considered very strong control of Type I error.

Assuming the tests are independent, the well-known Bonferroni P-value adjustment produces adjusted P-values (Probability Levels) for which the EWER is controlled. The Bonferroni adjustment is applied to all *m* unadjusted *P*-values (p_j) as

$$\tilde{p}_j = \min\left(mp_j, 1\right).$$

That is, each P-value (Probability Level) is multiplied by the number of tests, and if the result is greater than one, it is set to the maximum possible P-value of one.

False Discovery Rate (FDR)

The false discovery rate (FDR) (Benjamini and Hochberg, 1995) is defined as

$$FDR = E\left(\frac{V}{R}1_{\{R>0\}}\right) = E\left(\frac{V}{R}|R>0\right)\Pr(R>0),$$

where *R* is the number of genes that are declared significantly different, and *V* is the number of genes that are falsely declared different. Controlling FDR is controlling the expected *proportion* of falsely declared differences (false discoveries) to declared differences (true and false discoveries, together). If a method is used which controls a FDR of 0.05 for 1,000 tests, and 40 genes are declared different, it is expected that 40*0.05 = 2 of the 40 declarations are false declarations (false discoveries). Using a method that controls the FDR will produce a list of genes that includes an intermediate (depending also on sample size) number of genes for which there exists a true difference in expression (i.e., moderate to large S). However, the list of genes will include a small number of genes that are falsely declared to have a true difference in expression (i.e., moderately minimizes V). Controlling the FDR should be considered intermediate control of Type I error.

Assuming the tests are independent, the Benjamini and Hochberg P-value adjustment produces adjusted P-values (Probability Levels) for which the FDR is controlled. These adjusted *P*-values are found as

$$\tilde{p}_{r_i} = \min_{k=i,\ldots,m}\left\{\min(\frac{m}{k}p_{r_k}, 1)\right\},$$

where $p_{r_1} \le p_{r_2} \le \cdots \le p_{r_m}$ are the observed ordered unadjusted *P*-values. The procedure is defined in Benjamini and Hochberg (1995). The corresponding adjusted *P*-value definition given here is found in Dudoit, Shaffer, and Boldrick (2003).

Multiple Testing Adjustment Comparison

The following table gives a summary of the multiple testing adjustment procedures and error rate control. The power to detect differences also depends heavily on sample size.

Common Adjustment Technique	Error Rate Controlled	Control of Type I Error	Power to Detect Differences
None	PCER	Minimal	High
Bonferroni	EWER	Strict	Low
Benjamini and Hochberg	FDR	Moderate	Moderate/High

Type I Error: Rejection of a null hypothesis that is true.

Calculating Power

There are five separate test types, each requiring different formulas. Let the means of the two populations be represented by μ_1 and μ_2. The difference between these means will be represented by δ. Let the standard deviations of the two populations be represented as σ_1 and σ_2.

Equal-Variance Z-Test (Standard Deviations Known and Equal)

When $\sigma_1 = \sigma_2 = \sigma$ and σ is known, the power is calculated as follows for a directional alternative (one-tailed test) in which $\delta > 0$.

1. Find z_α such that $1 - \Phi(z_\alpha) = \alpha$, where $\Phi(x)$ is the area to the left of *x* under the standardized normal curve.
2. Calculate: $\sigma_{\bar{X}} = \sigma\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$.
3. Calculate: $z_1 = \frac{z_\alpha \sigma_{\bar{X}} - \delta}{\sigma_{\bar{X}}}$.
4. Power = $1 - \Phi(z_1)$.

Unequal-Variance Z-Test (Standard Deviations Known and Unequal)

When $\sigma_1 \neq \sigma_2$ and σ_1 and σ_2 are known, the power is calculated as follows for a directional alternative (one-tailed test) in which $\delta > 0$.

1. Find z_α such that $1 - \Phi(z_\alpha) = \alpha$, where $\Phi(x)$ is the area to the left of *x* under the standardized normal curve.
2. Calculate: $\sigma_{\bar{X}} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}$.
3. Calculate: $z_1 = \frac{z_\alpha \sigma_{\bar{X}} - \delta}{\sigma_{\bar{X}}}$.
4. Power = $1 - \Phi(z_1)$.

Equal-Variance T-Test (Standard Deviations Unknown and Equal)

When $\sigma_1 = \sigma_2 = \sigma$ and σ is unknown, the power is calculated as follows for a directional alternative (one-tailed test) in which $\delta > 0$.

1. Find t_α such that $1 - \mathrm{T}_{df}(t_\alpha) = \alpha$, where $\mathrm{T}_{df}(x)$ is the area to the left of *x* under a central-*t* curve with $df = N_1 + N_2 - 2$.
2. Calculate: $\sigma_{\bar{X}} = \sigma\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$.
3. Calculate the noncentrality parameter: $\lambda = \frac{\delta}{\sigma_{\bar{X}}}$.
4. Calculate: $t_1 = \frac{t_\alpha \sigma_{\bar{X}} - \delta}{\sigma_{\bar{X}}} + \lambda$.
5. Calculate: Power = $1 - T'_{df,\lambda}(t_1)$, where $T'_{df,\lambda}(x)$ is the area to the left of *x* under a noncentral-*t* curve with degrees of freedom df and noncentrality parameter λ.

Unequal-Variance T-Test (Standard Deviations Unknown and Unequal)

When $\sigma_1 \neq \sigma_2$ and σ_1 and σ_2 are unknown, the power is calculated as follows for a directional alternative (one-tailed test) in which $\delta > 0$. Note that in this case, an approximate T-Test is used.

1. Calculate: $\sigma_{\bar{X}} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}$.
2. Calculate: $df = \frac{\sigma_{\bar{X}}^4}{\frac{\sigma_1^4}{N_1^2(N_1-1)} + \frac{\sigma_2^4}{N_2^2(N_2-1)}}$,

   which is the adjusted degrees of freedom. Often, this is rounded to the next highest integer.
3. Find t_α such that $1 - \mathrm{T}_{df}(t_\alpha) = \alpha$, where $\mathrm{T}_{df}(x)$ is the area to the left of *x* under a central-*t* curve with df degrees of freedom.
4. Calculate the noncentrality parameter: $\lambda = \frac{\delta}{\sigma_{\bar{X}}}$.
5. Calculate: $t_1 = \frac{t_\alpha \sigma_{\bar{X}} - \delta}{\sigma_{\bar{X}}} + \lambda$.
6. Calculate: Power = $1 - T'_{df,\lambda}(t_1)$, where $T'_{df,\lambda}(x)$ is the area to the left of *x* under a noncentral-*t* curve with degrees of freedom df and noncentrality parameter λ.

Mann-Whitney U or Wilcoxon Rank-Sum Tests

The power calculation for the Mann-Whitney U or Wilcoxon Rank-Sum Test is the same as that for the two-sample equal-variance *t*-test except that an adjustment is made to the sample size based on an assumed data distribution as described in Al-Sunduqchi and Guenther (1990). The sample size n_i' used in power calculations is equal to

$$n_i' = n_i / W,$$

where W is the Wilcoxon adjustment factor based on the assumed data distribution.

The adjustments are as follows:

Distribution	***W***
Double Exponential	$2/3$
Logistic	$9/\pi^2$
Normal	$\pi/3$

This section describes the procedure for computing the power from n_1' and n_2', α, the assumed μ_1 and μ_2, and the assumed common standard deviation, $\sigma_1 = \sigma_2 = \sigma$. Two good references for these methods are Julious (2010) and Chow, Shao, Wang, and Lokhnygina (2018).

If we call the assumed difference between the means $\delta = \mu_1 - \mu_2$, the steps for calculating the power are as follows:

1. Find $t_{1-\alpha}$ based on the central-*t* distribution with degrees of freedom,

$$df = n_1' + n_2' - 2.$$

2. Calculate the non-centrality parameter:

$$\lambda = \frac{\delta}{\sigma\sqrt{\frac{1}{n_1'} + \frac{1}{n_2'}}}$$

3. Calculate the power as the probability that the test statistic *t* is greater than $t_{1-\alpha}$ under the non-central-*t* distribution with non-centrality parameter λ:

$$Power = \Pr_{Non-central-t}(t > t_{1-\alpha} | df = n_1' + n_2' - 2, \lambda).$$

The algorithms for calculating power for the opposite direction and the two-sided hypotheses are analogous to this method.

When solving for something other than power, **PASS** uses this same power calculation formulation, but performs a search to determine that parameter.

Adjusting Alpha

Experiment-wise Error Rate

When the Bonferroni method will be used to control the experiment-wise error rate, α_{EWER}, of all tests, the adjusted α, α_{adj}, for each test is given by

$$\alpha_{adj} = \frac{\alpha_{EWER}}{N_{tests}}$$

where N_{tests} is the total number of tests.

α_{adj} is the value that is used in the power and sample size calculations.

False Discovery Rate

When a false discovery rate controlling method will be used to control the false discovery rate for the experiment, fdr, the adjusted alpha, α_{adj}, for each test is given by Jung (2005) and Chow, Shao, Wang, and Lokhnygina (2018):

$$\alpha_{adj} = \frac{(K)(1-\beta)(fdr)}{(N_{tests} - K)(1 - fdr)}$$

where *K* is the number of genes with differential expression, β is the probability of a Type II error (not declaring a gene significant when it is), and N_{tests} is the total number of tests.

α_{adj} is the value that is used in the power and sample size calculations. Because α_{adj} depends on β, α_{adj} must be solved iteratively when the calculation of power is desired.

Example 1 – Finding Power

This example examines the power to detect differential expression for an experiment comparing a treatment group to a control group. There were 16 arrays used in each group. Each microarray produced intensity information for 5,000 genes. The 32 arrays were pre-processed by converting each expression value to the Log2 scale. In this example, the two-sample equal-variance T-Test was used to determine which genes were differentially expressed (upward or downward) when comparing the treatment group to the control group.

The researchers found very few differentially expressed genes and wish to examine the power of the experiment to detect two-fold differential expression (Log2-scale difference of 1). Typical standard deviations in each group ranged from 0.2 to 2.0.

The researchers guess the number of genes with at least 2-fold differential expression to be around 50 but will examine the effect of this estimate on power by trying 10 and 100 genes as well. A false discovery rate of 0.05 was used.

Setup

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 1** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Design Tab	
Solve For	**Power**
Test Type	**Equal-Variance T-Test**
Alternative Hypothesis	**Two-Sided**
False Discovery (Alpha) Method	**FDR (False Discovery Rate)**
FDR (False Discovery Rate)	**0.05**
Group Allocation	**Equal (N1 = N2)**
Sample Size Per Group	**16**
δ (Minimum |Mean Difference| Detected)	**1**
σ (Standard Deviation)	**0.2 to 2 by 0.2**
Number of Tests	**5000**
K (Number of Tests with |Mean Difference| > δ)	**10 50 100**

Output

Click the Calculate button to perform the calculations and generate the following output. The calculations should take a few moments.

Numeric Reports

Numeric Results

Solve For:	Power
Test Type:	Equal-Variance T-Test
Hypotheses:	H0: Diff = 0 vs. H1: Diff ≠ 0
False Discovery Method:	FDR (False Discovery Rate)
Number of Tests:	5000

Power for Each Test	Sample Size N1	N2	N	Minimum |Difference| Detected δ	Standard Deviation σ	Number of Tests with |Difference| > δ K	False Discovery Rate FDR	Single Test Alpha	Probability to Detect All K
1.00000	16	16	32	1	0.2	10	0.05	0.0001055	1.00000
1.00000	16	16	32	1	0.2	50	0.05	0.0005316	1.00000
1.00000	16	16	32	1	0.2	100	0.05	0.0010741	1.00000
0.98866	16	16	32	1	0.4	10	0.05	0.0001043	0.89217
0.99795	16	16	32	1	0.4	50	0.05	0.0005305	0.90250
0.99916	16	16	32	1	0.4	100	0.05	0.0010732	0.91949
0.52073	16	16	32	1	0.6	10	0.05	0.0000549	0.00147
0.75206	16	16	32	1	0.6	50	0.05	0.0003998	0.00000
0.83005	16	16	32	1	0.6	100	0.05	0.0008916	0.00000
0.06242	16	16	32	1	0.8	10	0.05	0.0000066	0.00000
0.23537	16	16	32	1	0.8	50	0.05	0.0001251	0.00000
0.34928	16	16	32	1	0.8	100	0.05	0.0003752	0.00000
0.00114	16	16	32	1	1.0	10	0.05	0.0000001	0.00000
0.02718	16	16	32	1	1.0	50	0.05	0.0000145	0.00000
0.06787	16	16	32	1	1.0	100	0.05	0.0000729	0.00000
0.00000	16	16	32	1	1.2	10	0.05	0.0000000	0.00000
0.00089	16	16	32	1	1.2	50	0.05	0.0000005	0.00000
0.00548	16	16	32	1	1.2	100	0.05	0.0000059	0.00000
0.00000	16	16	32	1	1.4	10	0.05	0.0000000	0.00000
0.00000	16	16	32	1	1.4	50	0.05	0.0000000	0.00000
0.00013	16	16	32	1	1.4	100	0.05	0.0000001	0.00000
0.00000	16	16	32	1	1.6	10	0.05	0.0000000	0.00000
0.00000	16	16	32	1	1.6	50	0.05	0.0000000	0.00000
0.00000	16	16	32	1	1.6	100	0.05	0.0000000	0.00000
0.00000	16	16	32	1	1.8	10	0.05	0.0000000	0.00000
0.00000	16	16	32	1	1.8	50	0.05	0.0000000	0.00000
0.00000	16	16	32	1	1.8	100	0.05	0.0000000	0.00000
0.00000	16	16	32	1	2.0	10	0.05	0.0000000	0.00000
0.00000	16	16	32	1	2.0	50	0.05	0.0000000	0.00000
0.00000	16	16	32	1	2.0	100	0.05	0.0000000	0.00000

Power	The individual probability of detecting a difference for each test with actual |mean difference| > δ.
N1, N2	The sample sizes (e.g., number of arrays for microarray studies) in groups 1 and 2, respectively, required to achieve the corresponding power.
N	The total sample size (e.g., number of arrays for microarray studies). N = N1 + N2.
δ	The smallest |mean difference| for which the power and sample size calculations are valid.
σ	The estimated standard deviation for both groups used in each test.
K	The number of tests for which the actual |mean difference| > δ.
FDR	The expected proportion of false declarations of significant difference (e.g., differential expression) to total declarations of significant difference.
Single Test Alpha	The probability of falsely declaring a significant difference for an individual test.
Detect All K	The probability of declaring significant difference for all K tests that have actual |mean difference| > δ.

Summary Statements

A parallel two-group design with 5000 individual tests will be used to test 5000 mean differences. Each comparison will be made using a two-sided, two-sample equal-variance t-test, with an individual test alpha of 0.0001055. The false discovery rate (FDR) for the experiment is 0.05. The common within-group standard deviation for both groups is assumed to be 0.2. To detect a |mean difference| of 1, with 10 of the 5000 individual tests having an actual |mean difference| greater than 1, with a sample size of 16 subjects in Group 1 and 16 subjects in Group 2, the power for each test is 1. Of the 10 tests with anticipated actual |mean difference| greater than 1, a significant difference is expected to be detected in 9 of them. The probability of detecting a difference in all 10 tests where the actual |mean difference| is greater than 1, is 1.

Dropout-Inflated Sample Size

	Sample Size			Dropout-Inflated Enrollment Sample Size			Expected Number of Dropouts		
Dropout Rate	**N1**	**N2**	**N**	**N1'**	**N2'**	**N'**	**D1**	**D2**	**D**
20%	16	16	32	20	20	40	4	4	8

Dropout Rate	The percentage of subjects (or items) that are expected to be lost at random during the course of the study and for whom no response data will be collected (i.e., will be treated as "missing"). Abbreviated as DR.
N1, N2, and N	The evaluable sample sizes at which power is computed (as entered by the user). If N1 and N2 subjects are evaluated out of the N1' and N2' subjects that are enrolled in the study, the design will achieve the stated power.
N1', N2', and N'	The number of subjects that should be enrolled in the study in order to obtain N1, N2, and N evaluable subjects, based on the assumed dropout rate. N1' and N2' are calculated by inflating N1 and N2 using the formulas N1' = N1 / (1 - DR) and N2' = N2 / (1 - DR), with N1' and N2' always rounded up. (See Julious, S.A. (2010) pages 52-53, or Chow, S.C., Shao, J., Wang, H., and Lokhnygina, Y. (2018) pages 32-33.)
D1, D2, and D	The expected number of dropouts. D1 = N1' - N1, D2 = N2' - N2, and D = D1 + D2.

Dropout Summary Statement

Anticipating a 20% dropout rate, 20 subjects should be enrolled in Group 1, and 20 in Group 2, to obtain final group sample sizes of 16 and 16, respectively.

References

Chow, S.C., Shao, J., Wang, H., and Lokhnygina, Y. 2018. Sample Size Calculations in Clinical Research, Third Edition. Taylor & Francis/CRC. Boca Raton, Florida.

Jung, S.-H. 2005. Sample size for FDR-control in microarray data analysis. Bioinformatics: Vol. 21 no. 14, pp. 3097-3104. Oxford University Press.

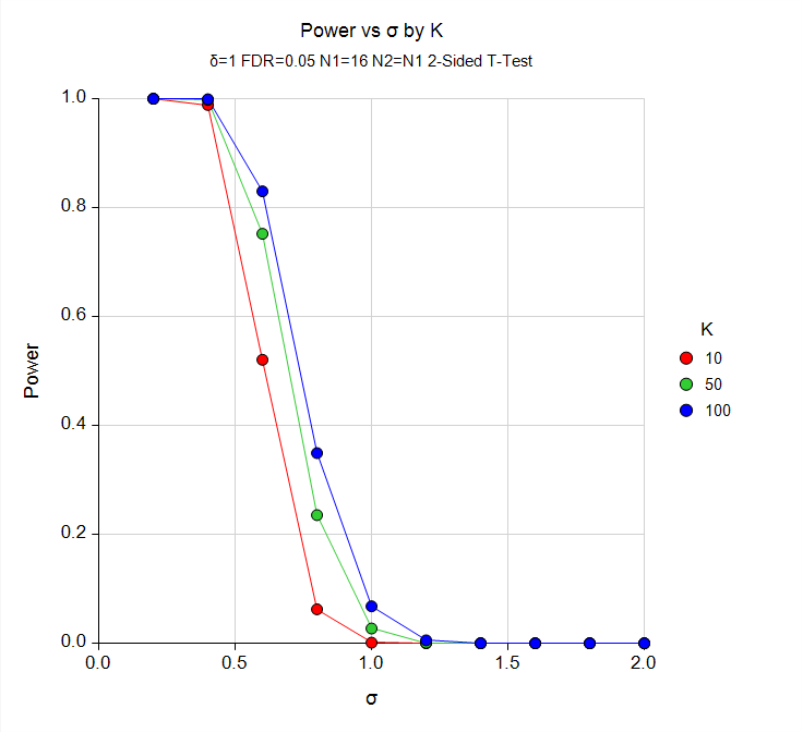
Machin, D., Campbell, M., Fayers, P., and Pinol, A. 1997. Sample Size Tables for Clinical Studies, 2nd Edition. Blackwell Science. Malden, MA.

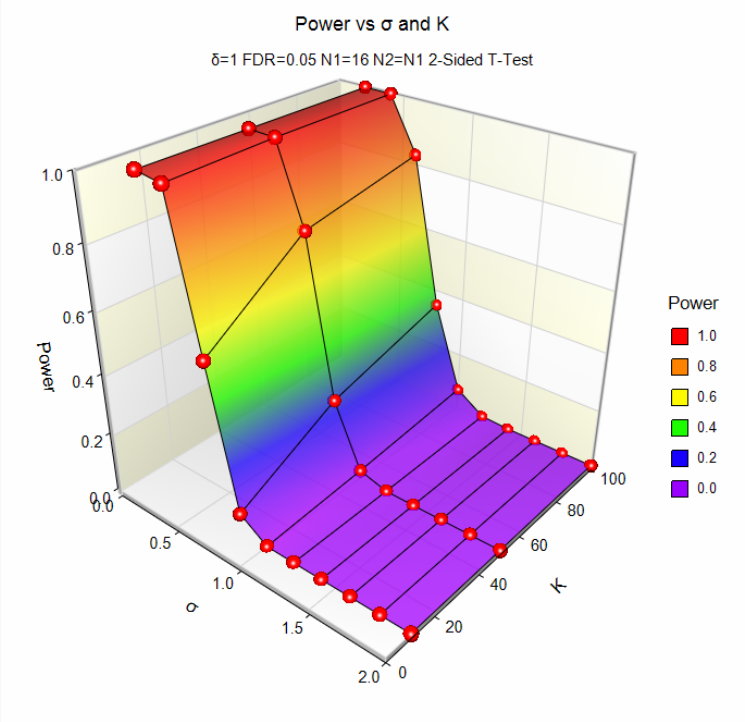
Zar, Jerrold H. 1984. Biostatistical Analysis (Second Edition). Prentice-Hall. Englewood Cliffs, New Jersey.

This report shows the values of each of the parameters, one scenario per row. The values of power were calculated from the other parameters. The definitions of each column are given in the Report Definitions section.

Plots Section

Plots





These plots show the relationship between power and the standard deviation of the differences for the three values of K. When the standard deviation within each group is greater than 1.0, the tests have very little power to detect 2-fold differences.

Example 2 – Finding the Sample Size

This example determines the number of arrays needed to achieve 80% power to detect differential expression for each gene. Each microarray will produce intensity information for 22,452 genes. The arrays will be pre-processed by converting each expression value to the Log2 scale. The two-sample equal-variance T-test will be used to determine which genes are differentially expressed (upward or downward) following exposure to the treatment.

The researchers wish to detect differential expression that is two-fold or greater (Log2-scale difference of 1). Typical standard deviations in each group are expected to range from 0.2 to 2.0.

The researchers guess the number of genes with at least 2-fold differential expression to be around 50, but will examine the effect of this estimate on sample size by trying 10 and 100 genes as well. A false discovery rate of 0.05 will be used.

Setup

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 2** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Design Tab	
Solve For	**Sample Size**
Test Type	**Equal-Variance T-Test**
Alternative Hypothesis	**Two-Sided**
Power for each Test	**0.8**
False Discovery (Alpha) Method	**FDR (False Discovery Rate)**
FDR (False Discovery Rate)	**0.05**
Group Allocation	**Equal (N1 = N2)**
δ (Minimum |Mean Difference| Detected)	**1**
σ (Standard Deviation)	**0.2 to 2 by 0.2**
Number of Tests	**22452**
K (Number of Tests with |Mean Difference| > δ)	**10 50 100**

Output

Click the Calculate button to perform the calculations and generate the following output. The calculations may take a few moments.

Numeric Reports

Numeric Results

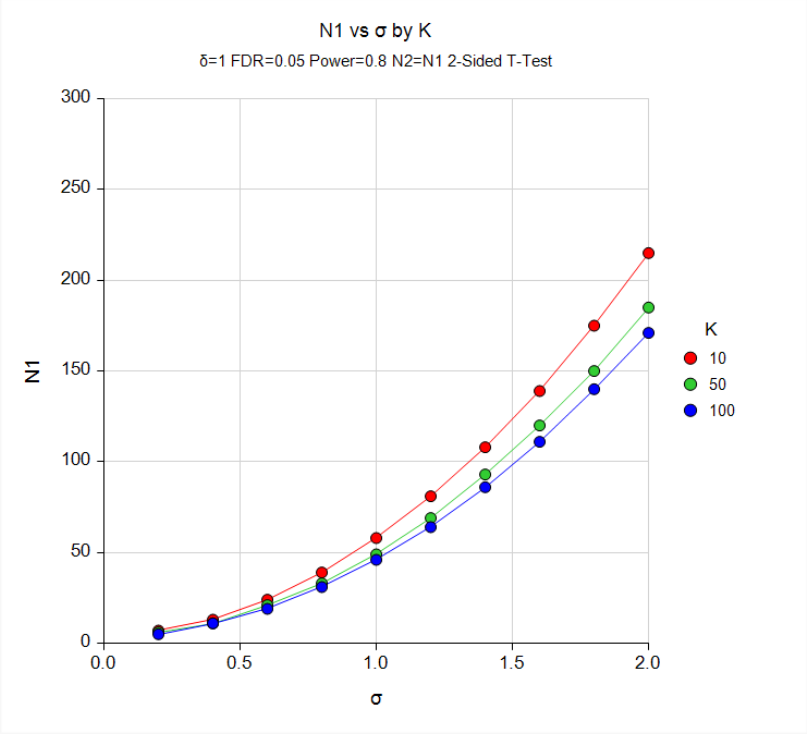
Solve For:	Sample Size
Test Type:	Equal-Variance T-Test
Hypotheses:	H0: Diff = 0 vs. H1: Diff ≠ 0
False Discovery Method:	FDR (False Discovery Rate)
Number of Tests:	22452

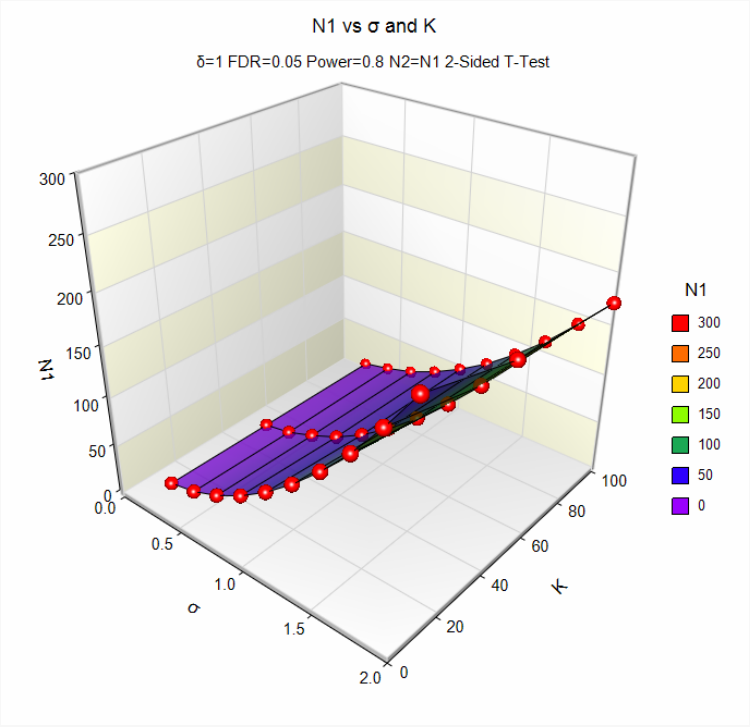
Power for Each Test		Sample Size			Minimum |Difference| Detected	Standard Deviation	Number of Tests with |Difference| > δ	False Discovery Rate	Single Test Alpha	Probability to Detect All K
Target	Actual	N1	N2	N	δ	σ	K	FDR		
0.8	0.93967	7	7	14	1	0.2	10	0.05	0.0000188	0.53673
0.8	0.92971	6	6	12	1	0.2	50	0.05	0.0000940	0.02615
0.8	0.80449	5	5	10	1	0.2	100	0.05	0.0001884	0.00000
0.8	0.81237	13	13	26	1	0.4	10	0.05	0.0000188	0.12518
0.8	0.80047	11	11	22	1	0.4	50	0.05	0.0000940	0.00001
0.8	0.86440	11	11	22	1	0.4	100	0.05	0.0001884	0.00000
0.8	0.82116	24	24	48	1	0.6	10	0.05	0.0000188	0.13940
0.8	0.83607	21	21	42	1	0.6	50	0.05	0.0000940	0.00013
0.8	0.81695	19	19	38	1	0.6	100	0.05	0.0001884	0.00000
0.8	0.81806	39	39	78	1	0.8	10	0.05	0.0000188	0.13424
0.8	0.80753	33	33	66	1	0.8	50	0.05	0.0000940	0.00002
0.8	0.81606	31	31	62	1	0.8	100	0.05	0.0001884	0.00000
0.8	0.81317	58	58	116	1	1.0	10	0.05	0.0000188	0.12643
0.8	0.80157	49	49	98	1	1.0	50	0.05	0.0000940	0.00002
0.8	0.80938	46	46	92	1	1.0	100	0.05	0.0001884	0.00000
0.8	0.80849	81	81	162	1	1.2	10	0.05	0.0000188	0.11934
0.8	0.80281	69	69	138	1	1.2	50	0.05	0.0000940	0.00002
0.8	0.80215	64	64	128	1	1.2	100	0.05	0.0001884	0.00000
0.8	0.80440	108	108	216	1	1.4	10	0.05	0.0000188	0.11343
0.8	0.80624	93	93	186	1	1.4	50	0.05	0.0000940	0.00002
0.8	0.80334	86	86	172	1	1.4	100	0.05	0.0001884	0.00000
0.8	0.80090	139	139	278	1	1.6	10	0.05	0.0000188	0.10859
0.8	0.80454	120	120	240	1	1.6	50	0.05	0.0000940	0.00002
0.8	0.80183	111	111	222	1	1.6	100	0.05	0.0001884	0.00000
0.8	0.80212	175	175	350	1	1.8	10	0.05	0.0000188	0.11026
0.8	0.80067	150	150	300	1	1.8	50	0.05	0.0000940	0.00001
0.8	0.80391	140	140	280	1	1.8	100	0.05	0.0001884	0.00000
0.8	0.80220	215	215	430	1	2.0	10	0.05	0.0000188	0.11037
0.8	0.80327	185	185	370	1	2.0	50	0.05	0.0000940	0.00002
0.8	0.80004	171	171	342	1	2.0	100	0.05	0.0001884	0.00000

This report shows the values of each of the parameters, one scenario per row. The sample size (number of arrays) estimates were calculated from the other parameters. The power is the actual power produced by the given sample size.

Plots Section

Plots





These plots show the relationship between sample size and the standard deviations within each group for three values of K.

Example 3 – Finding the Minimum Detectable Difference

This example finds the minimum difference in expression that can be detected with 90% power from a microarray experiment with two groups of 9 arrays in each group. The 9 arrays permit tests on 7,228 genes. The arrays will be pre-processed by converting each expression value to the Log2 scale. The two-sample equal-variance T-test will be used to determine which genes are differentially expressed (upward or downward) following exposure to the treatment. Typical standard deviations in each group for this experiment range from 0.2 to 1.8.

In this example we will examine a range for K (the number of genes with mean difference greater than the minimum detectable difference), since this should vary with the mean difference chosen. A false discovery rate of 0.05 will be used.

Setup

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 3** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Design Tab	
Solve For	**δ (Minimum |Mean Difference|)**
Test Type	**Equal-Variance T-Test**
Alternative Hypothesis	**Two-Sided**
Power for each Test	**0.9**
False Discovery (Alpha) Method	**FDR (False Discovery Rate)**
FDR (False Discovery Rate)	**0.05**
Group Allocation	**Equal (N1 = N2)**
Sample Size Per Group	**9**
σ (Standard Deviation)	**0.2 to 1.8 by 0.4**
Number of Tests	**7228**
K (Number of Tests with |Mean Difference| > δ)	**10 to 50 by 10**
Reports Tab	
δ Decimals	**4**

Output

Click the Calculate button to perform the calculations and generate the following output. The calculations may take a few moments.

Numeric Reports

Numeric Results

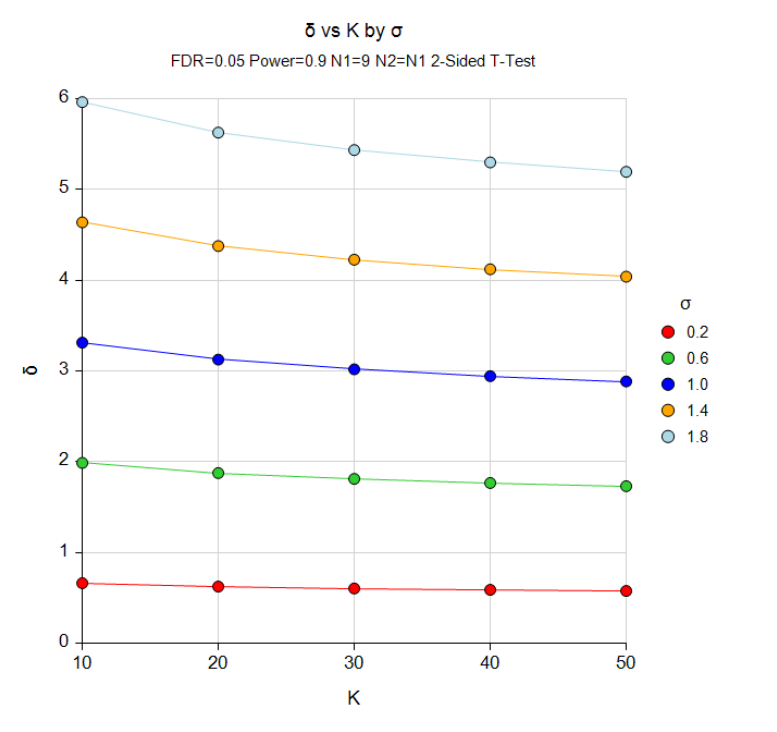
Solve For:	δ (Minimum |Mean Difference|)
Test Type:	Equal-Variance T-Test
Hypotheses:	H0: Diff = 0 vs. H1: Diff ≠ 0
False Discovery Method:	FDR (False Discovery Rate)
Number of Tests:	7228

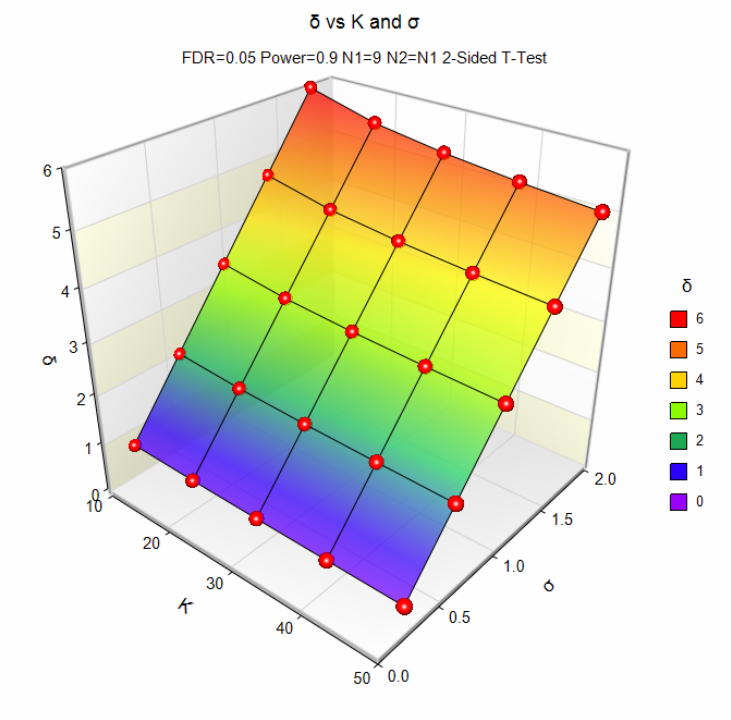
Power for Each Test	Sample Size N1	N2	N	Minimum |Difference| Detected δ	Standard Deviation σ	Number of Tests with |Difference| > δ K	False Discovery Rate FDR	Single Test Alpha	Probability to Detect All K
0.9	9	9	18	0.6626	0.2	10	0.05	0.0000656	0.34868
0.9	9	9	18	0.6253	0.2	20	0.05	0.0001314	0.12158
0.9	9	9	18	0.6038	0.2	30	0.05	0.0001974	0.04239
0.9	9	9	18	0.5888	0.2	40	0.05	0.0002636	0.01478
0.9	9	9	18	0.5772	0.2	50	0.05	0.0003300	0.00515
0.9	9	9	18	1.9879	0.6	10	0.05	0.0000656	0.34868
0.9	9	9	18	1.8759	0.6	20	0.05	0.0001314	0.12158
0.9	9	9	18	1.8115	0.6	30	0.05	0.0001974	0.04239
0.9	9	9	18	1.7663	0.6	40	0.05	0.0002636	0.01478
0.9	9	9	18	1.7315	0.6	50	0.05	0.0003300	0.00515
0.9	9	9	18	3.3132	1.0	10	0.05	0.0000656	0.34868
0.9	9	9	18	3.1265	1.0	20	0.05	0.0001314	0.12158
0.9	9	9	18	3.0192	1.0	30	0.05	0.0001974	0.04239
0.9	9	9	18	2.9439	1.0	40	0.05	0.0002636	0.01478
0.9	9	9	18	2.8858	1.0	50	0.05	0.0003300	0.00515
0.9	9	9	18	4.6385	1.4	10	0.05	0.0000656	0.34868
0.9	9	9	18	4.3770	1.4	20	0.05	0.0001314	0.12158
0.9	9	9	18	4.2269	1.4	30	0.05	0.0001974	0.04239
0.9	9	9	18	4.1214	1.4	40	0.05	0.0002636	0.01478
0.9	9	9	18	4.0402	1.4	50	0.05	0.0003300	0.00515
0.9	9	9	18	5.9638	1.8	10	0.05	0.0000656	0.34868
0.9	9	9	18	5.6276	1.8	20	0.05	0.0001314	0.12158
0.9	9	9	18	5.4346	1.8	30	0.05	0.0001974	0.04239
0.9	9	9	18	5.2990	1.8	40	0.05	0.0002636	0.01478
0.9	9	9	18	5.1945	1.8	50	0.05	0.0003300	0.00515

This report shows the values of each of the parameters, one scenario per row. The Minimum Mean Difference (δ) estimates were calculated from the other parameters.

Plots Section

Plots





These plots show the relationship between δ (the minimum detectable difference on the Log2 scale) and the standard deviations within each group for five values of K.

Example 4 – Validation (EWER) using Stekel (2003)

Stekel (2003), page 228, gives an example in which Power = 0.95, δ = 1, and σ1 = σ2 = 0.68 for a two-sided two-sample equal-variance T-Test. The number of genes tested is 10000. The control of false discoveries is "at most one false positive result." This corresponds to an EWER value of 1.0. The sample sizes obtained for this example are 33 per group.

Setup

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 4** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Design Tab

Solve For	**Sample Size**
Test Type	**Equal-Variance T-Test**
Alternative Hypothesis	**Two-Sided**
Power for each Test	**0.95**
False Discovery (Alpha) Method	**EWER (Experiment-wise Error Rate)**
EWER (Experiment-wise Error Rate)	**1**
Group Allocation	**Equal (N1 = N2)**
δ (Minimum |Mean Difference| Detected)	**1**
σ (Standard Deviation)	**0.68**
Number of Tests	**10000**

Reports Tab

σ, σ1, σ2 Decimals	**2**

Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results

Solve For:	Sample Size
Test Type:	Equal-Variance T-Test
Hypotheses:	H0: Diff = 0 vs. H1: Diff ≠ 0
False Discovery Method:	EWER (Experiment-Wise Error Rate)
Number of Tests:	10000

Power for Each Test		Sample Size			Minimum |Difference| Detected	Standard Deviation	Experiment-Wise Error Rate	Single Test Alpha
Target	Actual	N1	N2	N	δ	σ	EWER	
0.95	0.95785	33	33	66	1	0.68	1	0.0001

The sample sizes of 33 per group match Stekel's result.

Example 5 – Validation (EWER) using Lee (2004)

Lee (2004), pp. 218-220, gives an example in which Power = 0.90, δ = 1.0 1.5 2.0 2.5 and σ_paired = 1.0 for a two-sided Z-Test. The corresponding σ for a two-sample design is $1.0/\sqrt{2}$ = 0.707107. The number of genes tested is 1000. The control of false discoveries is 0.5. This corresponds to an EWER value of 0.5. This setup corresponds to the upper left corner of Table 14.3 on page 219. The sample sizes obtained for this setup are 23, 11, 6, and 4, respectively.

Setup

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 5** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Design Tab	
Solve For	**Sample Size**
Test Type	**Equal-Variance Z-Test**
Alternative Hypothesis	**Two-Sided**
Power for each Test	**0.9**
False Discovery (Alpha) Method	**EWER (Experiment-wise Error Rate)**
EWER (Experiment-wise Error Rate)	**0.5**
Group Allocation	**Equal (N1 = N2)**
δ (Minimum |Mean Difference| Detected)	**1 1.5 2 2.5**
σ (Standard Deviation)	**0.707107**
Number of Tests	**1000**

Reports Tab	
σ, σ1, σ2 Decimals	**3**

Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results

Solve For:	Sample Size
Test Type:	Equal-Variance Z-Test
Hypotheses:	H0: Diff = 0 vs. H1: Diff ≠ 0
False Discovery Method:	EWER (Experiment-Wise Error Rate)
Number of Tests:	1000

Power for Each Test		Sample Size			Minimum |Difference| Detected	Standard Deviation	Experiment-Wise Error Rate	Single Test
Target	Actual	N1	N2	N	δ	σ	EWER	Alpha
0.9	0.90576	23	23	46	1.0	0.707	0.5	0.0005
0.9	0.93244	11	11	22	1.5	0.707	0.5	0.0005
0.9	0.92194	6	6	12	2.0	0.707	0.5	0.0005
0.9	0.93565	4	4	8	2.5	0.707	0.5	0.0005

Group sample sizes of 23, 11, 6, and 4 per group match the results shown in Lee (2004).

Example 6 – Validation (FDR) using Jung (2005)

Jung (2005), page 3100, gives an example for the sample size needed to control FDR in a two-sample Z-Test. This example is repeated in Chow, Shao, Wang, and Lokhnygina (2018). In the example, Power = 0.60 (from 24/40), δ = 1.0, and σ = 1.0 for a one-sided two-sample equal-variance Z-Test. The number of genes tested is 4000. The FDR level is 1%. This setup corresponds to Example 1 on page 3100. The required sample size obtained in each group for this setup is 34.

Setup

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 6** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Design Tab

Solve For	**Sample Size**
Test Type	**Equal-Variance Z-Test**
Alternative Hypothesis	**One-Sided**
Power for each Test	**0.6**
False Discovery (Alpha) Method	**FDR (False Discovery Rate)**
FDR (False Discovery Rate)	**0.01**
Group Allocation	**Equal (N1 = N2)**
δ (Minimum |Mean Difference| Detected)	**1**
σ (Standard Deviation)	**1**
Number of Tests	**4000**
K (Number of Tests with |Mean Difference| > δ)	**40**

Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results

Solve For:	Sample Size
Test Type:	Equal-Variance Z-Test
Hypotheses:	H0: Diff ≤ 0 vs. H1: Diff > 0
False Discovery Method:	FDR (False Discovery Rate)
Number of Tests:	4000

Power for Each Test		Sample Size			Minimum |Difference| Detected	Standard Deviation	Number of Tests with |Difference| > δ	False Discovery Rate	Single Test Alpha	Probability to Detect All K
Target	Actual	N1	N2	N	δ	σ	K	FDR		
0.6	0.61099	34	34	68	1	1	40	0.01	0.0000612	0

A group sample size of 34 matches the result shown in Jung (2005). For Example 3 in Jung (2005), the alternative hypothesis is two-sided and results in a sample size of 73. This result may be validated in **PASS** by changing Alternative Hypothesis to "Two-Sided" in this example.