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

$$\text{PCER} = E(V)/m\,,$$

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

$$\text{EWER} = \Pr(V > 0),$$

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

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

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

$$\text{FDR} = E(\frac{V}{R}1_{\{R>0\}}) = E(\frac{V}{R} \mid R > 0)\Pr(R > 0)\,,$$

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

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

where $p_{r_1} \leq p_{r_2} \leq \cdots \leq p_{r_m}$ 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.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Design and Options tabs. For more information about the options of other tabs, go to the Procedure Window chapter.

Design Tab

The Design tab contains most of the parameters and options involved in the power and sample size calculations.

Solve For

Solve For

This option specifies the parameter to be calculated from the values of the other parameters. Under most conditions, you would select either *Power* or *Sample Size*.

Select *Sample Size* when you want to determine the sample size needed to achieve a given power and false discovery rate (or alpha) error level.

Select *Power* when you want to calculate the power of an experiment.

Test

Test Type

Select the type of test that will be used when the analysis of the high-throughput data is carried out.

- **Equal-Variance T-Test**
  The Equal-Variance T-Test assumes the data values (e.g. expression values) come from normal distributions with an UNKNOWN common standard deviation (i.e., a standard deviation that will be estimated from the data).

- **Unequal-Variance T-Test**
  The Unequal-Variance T-Test assumes the data values (e.g. expression values) come from normal distributions with UNKNOWN and UNEQUAL standard deviations (i.e., standard deviations that will be estimated from the data).

- **Equal-Variance Z-Test**
  The Equal-Variance Z-Test assumes the data values (e.g. expression values) come from normal distributions with a KNOWN common standard deviation.

- **Unequal-Variance Z-Test**

  The Unequal-Variance Z-Test assumes the data values (e.g. expression values) come from normal distributions with KNOWN and UNEQUAL standard deviations.

- **Mann-Whitney U or Wilcoxon Rank-Sum Test**

  The Mann-Whitney U or Wilcoxon Rank-Sum Test is the nonparametric analog of the two-sample t-test.

Data Distribution

Displayed only if Test Type = Mann-Whitney U or Wilcoxon Rank-Sum Test

This option makes appropriate sample size adjustments for the Mann-Whitney U or Wilcoxon Rank-Sum test. Results by Al-Sunduqchi and Guenther (1990) indicate that power calculations for the Mann-Whitney U or Wilcoxon Rank-Sum test may be made using the standard t-test formulations with a simple adjustment to the sample size. The size of the adjustment depends upon the actual distribution of the data. They give sample size adjustment factors for three distributions. Select a distribution similar in shape to your data.

If Ni is the group sample size and W is the adjustment factor, then the distribution-adjusted group sample size is

$$Ni' = Ni/W$$

The options are:

- **Double Exponential**

  The sample size adjustment factor, W, is equal to “2/3”.

- **Logistic**

  The sample size adjustment factor, W, is equal to “$9/\pi^2$”.

- **Normal**

  The sample size adjustment factor, W, is equal to “$\pi/3$”.

Alternative Hypothesis

Specify whether the hypothesis tests are one-sided (directional) or two-sided (non-directional).

Recommendation for Microarray Studies

In most two-group experiments, differential expression in either direction (up-regulation or down-regulation) is of interest. Such experiments should have the Two-Sided alternative hypothesis.

For experiments for determining only whether expression has increased (or only decreased), a One-Sided alternative hypothesis is recommended. Often regulations dictate that the FDR or EWER level be divided by 2 for One-Sided alternative tests.

Error Rates

Power for each Test

Power is the probability of rejecting each null hypothesis when it is false. Power is equal to 1-Beta.

The POWER for each test represents that probability of detecting a difference (i.e. differential expression for microarray data) when it exists.

You can enter a single value such as *0.7* or a series of values such as *0.7 0.8 0.9* or *0.7 to 0.95 by 0.05*.

False Discovery (Alpha) Method

A type I error is declaring there to be a difference (e.g. a gene to be differentially expressed in microarray studies) when there is not. The two most common methods for controlling type I error are false discovery rate (FDR) control and Experiment-wise Error Rate (EWER) control.

- **FDR**

  Controlling the false discovery rate (FDR) controls the PROPORTION of tests for which the difference is falsely declared as significant (e.g. genes in microarray studies that are falsely declared as differentially expressed). For example, suppose that in a microarray study an FDR of 0.05 is used for 10000 tests (on 10000 genes). If differential expression is declared for 100 of the 10000 genes, 5 of the 100 genes are expected to be false discoveries.

- **EWER**

  Controlling the experiment-wise error rate (EWER) controls the PROBABILITY of ANY false significance declarations (e.g. of differential expression in microarray studies), across all tests. For example, suppose that in a microarray study an EWER of 0.05 is used for 10000 tests (on 10000 genes). If differential expression is declared for 100 of the 10000 genes, the probability that even one of the 100 declarations is false is 0.05.

Recommendation

For exploratory studies where a list of candidates (e.g. genes) for further study is the goal, FDR is the recommended Type I error control method, because of its higher power.

For confirmatory studies where final determination of the difference (e.g. differential expression) is the goal, EWER is the recommended Type I error control method, because of its strict control of false discoveries..

FDR (False Discovery Rate)

Specify the value for the False Discovery Rate. FDR controls the PROPORTION of tests for which the difference is falsely declared as significant (e.g. genes in microarray studies that are falsely declared as differentially expressed). For example, suppose that in a microarray study an FDR of 0.05 is used for 10000 tests (on 10000 genes). If differential expression is declared for 100 of the 10000 genes, 5 of the 100 genes are expected to be false discoveries.

Enter values in the range 0 < FDR ≤ 1. Commonly, 0.001 ≤ FDR ≤ 0.25. FDR is often set to 0.05 for two-sided tests and to 0.025 for one-sided tests. You can enter a single value such as *0.05* or a list of values such as *0.05 0.10 0.15* or *0.05 to 0.15 by 0.01*.

EWER (Experiment-wise Error Rate)

Specify the value for the Experiment-wise Error Rate. EWER controls the PROBABILITY of ANY false significance declarations (e.g. of differential expression in microarray studies), across all tests. For example, suppose that in a microarray study an EWER of 0.05 is used for 10000 tests (on 10000 genes). If differential expression is declared for 100 of the 10000 genes, the probability that even one of the 100 declarations is false is 0.05.

Enter values in the range 0 < EWER ≤ 1. Commonly, 0.001 ≤ EWER ≤ 0.25. EWER is often set to 0.05 for two-sided tests and to 0.025 for one-sided tests. You can enter a single value such as *0.05* or a list of values such as *0.05 0.10 0.15* or *0.05 to 0.15 by 0.01*.

Sample Size (When Solving for Sample Size)

Group Allocation

Select the option that describes the constraints on *N1* or *N2* or both.

The options are

- **Equal (N1 = N2)**
  This selection is used when you wish to have equal sample sizes in each group. Since you are solving for both sample sizes at once, no additional sample size parameters need to be entered.

- **Enter N1, solve for N2**
  Select this option when you wish to fix *N1* at some value (or values), and then solve only for *N2*. Please note that for some values of *N1*, there may not be a value of *N2* that is large enough to obtain the desired power.

- **Enter N2, solve for N1**
  Select this option when you wish to fix *N2* at some value (or values), and then solve only for *N1*. Please note that for some values of *N2*, there may not be a value of *N1* that is large enough to obtain the desired power.

- **Enter R = N2/N1, solve for N1 and N2**
  For this choice, you set a value for the ratio of *N2* to *N1*, and then PASS determines the needed *N1* and *N2*, with this ratio, to obtain the desired power. An equivalent representation of the ratio, *R*, is

  $$N2 = R * N1.$$

- **Enter percentage in Group 1, solve for N1 and N2**
  For this choice, you set a value for the percentage of the total sample size that is in Group 1, and then PASS determines the needed *N1* and *N2* with this percentage to obtain the desired power.

N1 (Number of Arrays, Group 1)

This option is displayed if Group Allocation = "Enter N1, solve for N2"

N1 is the number of items or individuals sampled from the Group 1 population.

N1 must be ≥ 2. You can enter a single value or a series of values.

N2 (Number of Arrays, Group 2)

This option is displayed if Group Allocation = "Enter N2, solve for N1"

N2 is the number of items or individuals sampled from the Group 2 population.

N2 must be ≥ 2. You can enter a single value or a series of values.

R (Group Sample Size Ratio)

This option is displayed only if Group Allocation = "Enter R = N2/N1, solve for N1 and N2."

R is the ratio of *N2* to *N1*. That is,

$$R = N2 / N1.$$

Use this value to fix the ratio of *N2* to *N1* while solving for *N1* and *N2*. Only sample size combinations with this ratio are considered.

N2 is related to *N1* by the formula:

$$N2 = [R \times N1],$$

where the value *[Y]* is the next integer ≥ *Y*.

For example, setting *R* = 2.0 results in a Group 2 sample size that is double the sample size in Group 1 (e.g., *N1* = 10 and *N2* = 20, or *N1* = 50 and *N2* = 100).

R must be greater than 0. If *R* < 1, then *N2* will be less than *N1*; if *R* > 1, then *N2* will be greater than *N1*. You can enter a single or a series of values.

Percent in Group 1

This option is displayed only if Group Allocation = "Enter percentage in Group 1, solve for N1 and N2."

Use this value to fix the percentage of the total sample size allocated to Group 1 while solving for *N1* and *N2*. Only sample size combinations with this Group 1 percentage are considered. Small variations from the specified percentage may occur due to the discrete nature of sample sizes.

The Percent in Group 1 must be greater than 0 and less than 100. You can enter a single or a series of values.

Sample Size (When Not Solving for Sample Size)

Group Allocation

Select the option that describes how individuals in the study will be allocated to Group 1 and to Group 2.

The options are

- **Equal (N1 = N2)**

  This selection is used when you wish to have equal sample sizes in each group. A single per group sample size will be entered.

- **Enter N1 and N2 individually**

  This choice permits you to enter different values for *N1* and *N2*.

- **Enter N1 and R, where N2 = R * N1**

  Choose this option to specify a value (or values) for *N1*, and obtain *N2* as a ratio (multiple) of *N1*.

- **Enter total sample size and percentage in Group 1**

  Choose this option to specify a value (or values) for the total sample size (*N*), obtain *N1* as a percentage of *N*, and then *N2* as *N* - *N1*.

Sample Size Per Group

This option is displayed only if Group Allocation = "Equal (N1 = N2)."

The Sample Size Per Group is the number of items or individuals sampled from each of the Group 1 and Group 2 populations. Since the sample sizes are the same in each group, this value is the value for *N1*, and also the value for *N2*.

The Sample Size Per Group must be ≥ 2. You can enter a single value or a series of values.

N1 (Number of Arrays, Group 1)

*This option is displayed if Group Allocation = "Enter N1 and N2 individually" or "Enter N1 and R, where N2 = R * N1."*

N1 is the number of items or individuals sampled from the Group 1 population.

N1 must be ≥ 2. You can enter a single value or a series of values.

N2 (Number of Arrays, Group 2)

This option is displayed only if Group Allocation = "Enter N1 and N2 individually."

N2 is the number of items or individuals sampled from the Group 2 population.

N2 must be ≥ 2. You can enter a single value or a series of values.

R (Group Sample Size Ratio)

*This option is displayed only if Group Allocation = "Enter N1 and R, where N2 = R * N1."*

R is the ratio of *N2* to *N1*. That is,

$$R = N2/N1$$

Use this value to obtain *N2* as a multiple (or proportion) of *N1*.

N2 is calculated from *N1* using the formula:

$$N2=[R \text{ x } N1],$$

where the value *[Y]* is the next integer ≥ *Y*.

For example, setting *R* = 2.0 results in a Group 2 sample size that is double the sample size in Group 1.

R must be greater than 0. If *R* < 1, then *N2* will be less than N1; if *R* > 1, then *N2* will be greater than *N1*. You can enter a single value or a series of values.

Total Sample Size (N)

This option is displayed only if Group Allocation = "Enter total sample size and percentage in Group 1."

This is the total sample size, or the sum of the two group sample sizes. This value, along with the percentage of the total sample size in Group 1, implicitly defines *N1* and *N2*.

The total sample size must be greater than one, but practically, must be greater than 3, since each group sample size needs to be at least 2.

You can enter a single value or a series of values.

Percent in Group 1

This option is displayed only if Group Allocation = "Enter total sample size and percentage in Group 1."

This value fixes the percentage of the total sample size allocated to Group 1. Small variations from the specified percentage may occur due to the discrete nature of sample sizes.

The Percent in Group 1 must be greater than 0 and less than 100. You can enter a single value or a series of values.

Effect Size

δ (Minimum |Mean Difference| Detected)

Specify the positive actual mean difference (e.g. in gene expression for microarray studies) such that tests with |mean difference| > δ will be detected with the given power at the corresponding sample size. This value must be entered as a positive number.

When δ is large, the resulting sample size will only detect the tests with extreme differences (e.g. differential expression). When δ is small, a larger sample size is required to have power sufficient to detect these small differences (e.g. in differential expression).

Microarray Expression Studies

In expression studies, it is very common that the expression is measured on the log scale. Values of δ should reflect the differences that will be used in testing. For example, if the log2 scale is used, δ = 1 implies a two-fold difference in expression, while δ = 2 implies a four-fold difference in expression.

Notes

You can enter a single value such as *1* or a series of values such as *1 2 3* or *0.2 to 2 by 0.1* in the range δ > 0.

σ (Standard Deviation)

Displayed only if Test Type = Equal-Variance T-Test, Equal-Variance Z-Test, or Mann-Whitney U or Wilcoxon Rank-Sum Test

The standard deviation entered here is the assumed common standard deviation for both Group 1 and Group 2. This standard deviation is assumed for all tests. σ should be on the same scale as δ.

To obtain the standard deviation from the standard deviation of paired differences, use σ = σ_paired/(√2).

Because the true variation will vary from test to test, it is recommended that a range of values be entered here.

Notes

You can enter a single value such as *1* or a series of values such as *1 2 3 4 5* or *0.2 to 2 by 0.1* in the range σ > 0.

σ1, σ2 (Group 1, 2 Standard Deviation)

Displayed only if Test Type = Unequal-Variance T-Test or Unequal-Variance Z-Test

Specify the standard deviation for group 1, 2. This standard deviation is assumed for all tests.

σ1, σ2 should be on the same scale as δ.

To obtain the standard deviation from the standard deviation of paired differences, use σ1, σ2 = σ_paired/(√2).

Because the true variation in expression values will vary from test to test, it is recommended that a range of values be entered here.

Notes

You can enter a single value such as *1* or a series of values such as *1 2 3 4 5* or *0.2 to 2 by 0.1* in the range σ1, σ2 > 0.

Number of Tests

Number of Tests

Specify the number of hypothesis tests that will be carried out. In microarray studies, this number will usually be the number of genes summarized on each array minus the number of housekeeping genes. Only one number may be entered in this box.

K (Number of Tests with |Mean Difference| > δ)

Displayed only if False Discovery (Alpha) Method = FDR

Specify the number of tests for which an actual absolute mean difference in expression greater than δ is expected.

The choice of K has a direct effect on the calculation of power or sample size when the False Discovery (Alpha) Method is set to FDR.

You can enter a single value such as *20* or a series of values such as *10 20 30 40 50* or *20 to 100 by 10*.

Options Tab

The Options tab contains convergence options that are rarely changed.

Convergence Options

FDR Power Convergence

When FDR is selected for False Discovery (Alpha) Method, and Find (Solve For) is set to Power, the corresponding search algorithm will converge when the search criteria is below this value.

This value will rarely be changed from the default value.

RECOMMENDED: 0.0000000001

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

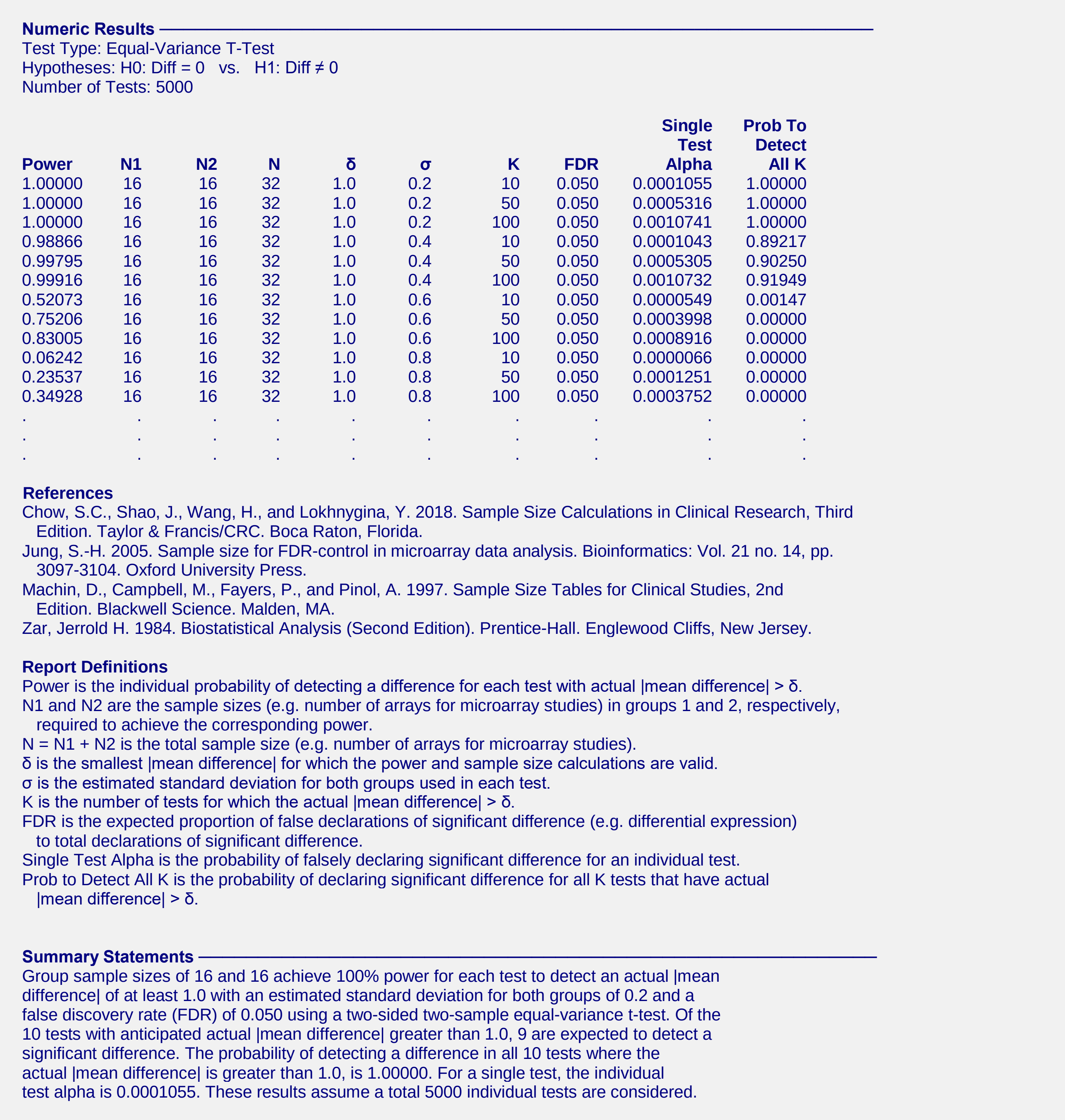
This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Multiple Testing for Two Means** procedure window by expanding **Means**, then **Two Independent Means**, then clicking on **Multiple Testing**, and then clicking on **Multiple Testing for Two Means**. You may then make the appropriate entries as listed below, or open **Example 1** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
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**

Annotated Output

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

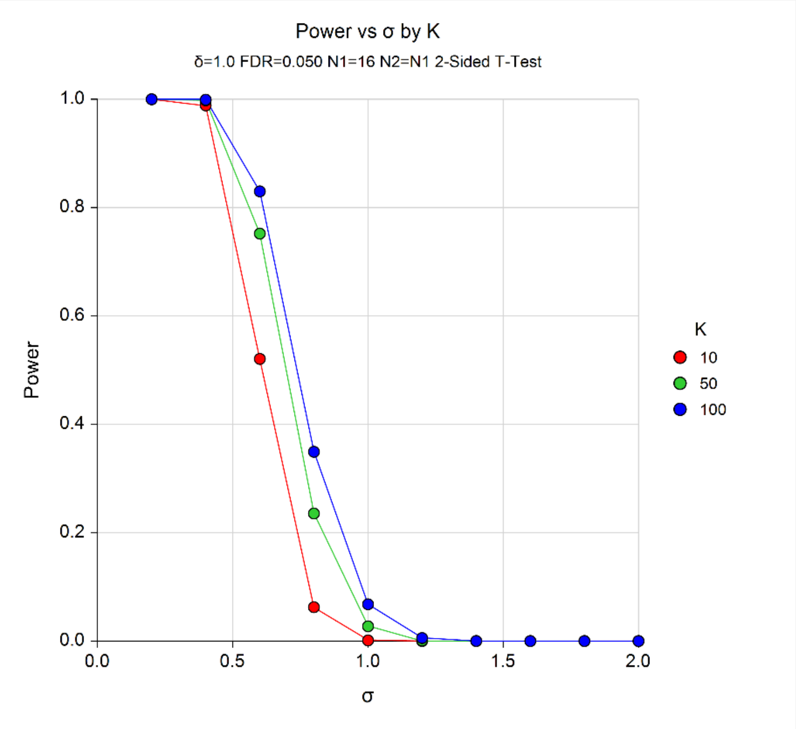
Numeric Results

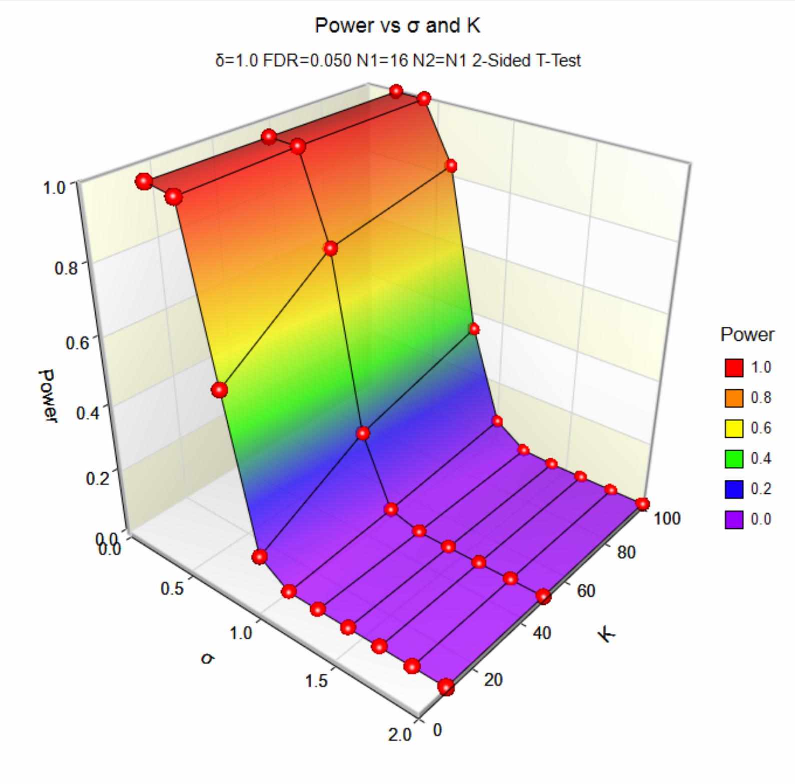


```
Numeric Results ──────────────────────────────────────────────────────────────────
Test Type: Equal-Variance T-Test
Hypotheses: H0: Diff = 0   vs.   H1: Diff ≠ 0
Number of Tests: 5000

                                                                    Single    Prob To
                                                                      Test     Detect
Power      N1      N2       N       δ       σ       K     FDR        Alpha      All K
1.00000    16      16      32     1.0     0.2      10   0.050    0.0001055    1.00000
1.00000    16      16      32     1.0     0.2      50   0.050    0.0005316    1.00000
1.00000    16      16      32     1.0     0.2     100   0.050    0.0010741    1.00000
0.98866    16      16      32     1.0     0.4      10   0.050    0.0001043    0.89217
0.99795    16      16      32     1.0     0.4      50   0.050    0.0005305    0.90250
0.99916    16      16      32     1.0     0.4     100   0.050    0.0010732    0.91949
0.52073    16      16      32     1.0     0.6      10   0.050    0.0000549    0.00147
0.75206    16      16      32     1.0     0.6      50   0.050    0.0003998    0.00000
0.83005    16      16      32     1.0     0.6     100   0.050    0.0008916    0.00000
0.06242    16      16      32     1.0     0.8      10   0.050    0.0000066    0.00000
0.23537    16      16      32     1.0     0.8      50   0.050    0.0001251    0.00000
0.34928    16      16      32     1.0     0.8     100   0.050    0.0003752    0.00000
.           .       .       .       .       .       .       .            .          .
.           .       .       .       .       .       .       .            .          .
.           .       .       .       .       .       .       .            .          .

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.

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


Summary Statements ───────────────────────────────────────────────────────────────
Group sample sizes of 16 and 16 achieve 100% power for each test to detect an actual |mean
difference| of at least 1.0 with an estimated standard deviation for both groups of 0.2 and a
false discovery rate (FDR) of 0.050 using a two-sided two-sample equal-variance t-test. Of the
10 tests with anticipated actual |mean difference| greater than 1.0, 9 are expected to detect a
significant difference. The probability of detecting a difference in all 10 tests where the
actual |mean difference| is greater than 1.0, is 1.00000. For a single test, the individual
test alpha is 0.0001055. These results assume a total 5000 individual tests are considered.
```

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





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

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Multiple Testing for Two Means** procedure window by expanding **Means**, then **Two Independent Means**, then clicking on **Multiple Testing**, and then clicking on **Multiple Testing for Two Means**. You may then make the appropriate entries as listed below, or open **Example 2** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
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

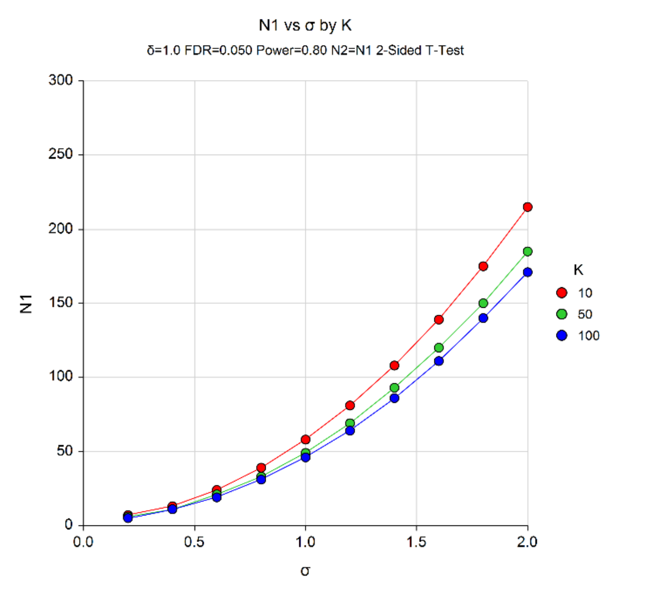
Click the Calculate button to perform the calculations and generate the following output. The calculations may take a few moments.

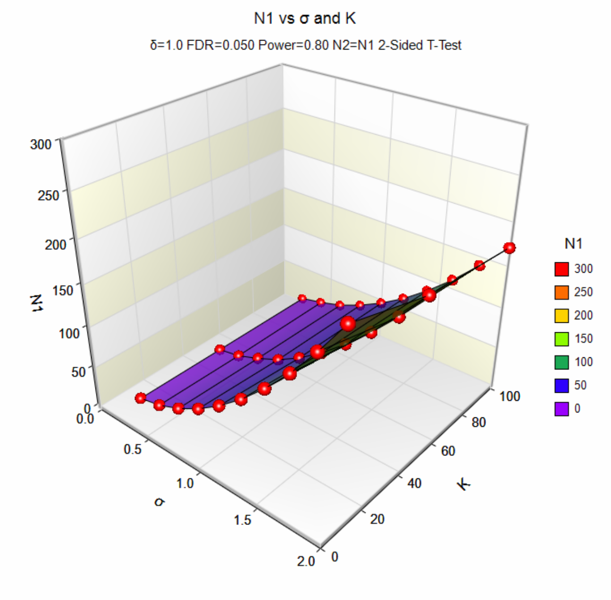
Numeric Results

```
Numeric Results ────────────────────────────────────────────────────────────
Test Type: Equal-Variance T-Test
Hypotheses: H0: Diff = 0   vs.   H1: Diff ≠ 0
Number of Tests: 22452

                                                                       Single    Prob To
Target    Actual                                                         Test     Detect
Power     Power     N1    N2     N     δ       σ      K     FDR         Alpha      All K
0.80    0.93967      7     7    14   1.0     0.2     10   0.050   0.0000188    0.53673
0.80    0.92971      6     6    12   1.0     0.2     50   0.050   0.0000940    0.02615
0.80    0.80449      5     5    10   1.0     0.2    100   0.050   0.0001884    0.00000
0.80    0.81237     13    13    26   1.0     0.4     10   0.050   0.0000188    0.12518
0.80    0.80047     11    11    22   1.0     0.4     50   0.050   0.0000940    0.00001
0.80    0.86440     11    11    22   1.0     0.4    100   0.050   0.0001884    0.00000
0.80    0.82116     24    24    48   1.0     0.6     10   0.050   0.0000188    0.13940
.             .      .     .     .     .       .      .       .           .          .
.             .      .     .     .     .       .      .       .           .          .
.             .      .     .     .     .       .      .       .           .          .
```

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





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

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Multiple Testing for Two Means** procedure window by expanding **Means**, then **Two Independent Means**, then clicking on **Multiple Testing**, and then clicking on **Multiple Testing for Two Means**. You may then make the appropriate entries as listed below, or open **Example 3** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
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

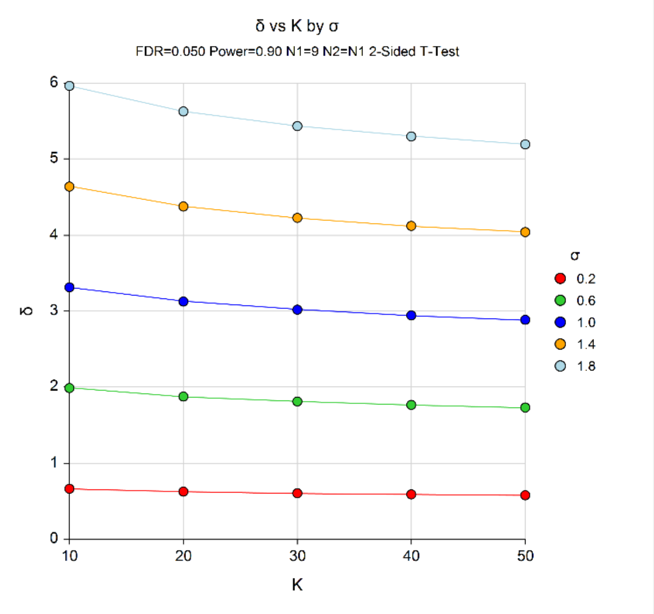
Click the Calculate button to perform the calculations and generate the following output. The calculations may take a few moments.

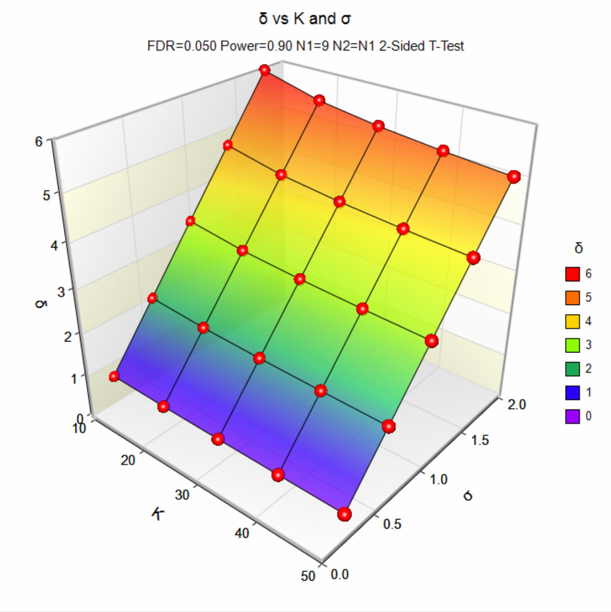
Numeric Results

```
Numeric Results ────────────────────────────────────────────────────────────
Test Type: Equal-Variance T-Test
Hypotheses: H0: Diff = 0   vs.   H1: Diff ≠ 0
Number of Tests: 7228

                                                                        Single    Prob To
                                                                          Test     Detect
Power        N1    N2     N        δ        σ        K     FDR          Alpha      All K
0.90000       9     9    18   0.6626      0.2       10   0.050      0.0000656    0.34868
0.90000       9     9    18   0.6253      0.2       20   0.050      0.0001314    0.12158
0.90000       9     9    18   0.6038      0.2       30   0.050      0.0001974    0.04239
0.90000       9     9    18   0.5888      0.2       40   0.050      0.0002636    0.01478
0.90000       9     9    18   0.5772      0.2       50   0.050      0.0003300    0.00515
0.90000       9     9    18   1.9879      0.6       10   0.050      0.0000656    0.34868
0.90000       9     9    18   1.8759      0.6       20   0.050      0.0001314    0.12158
0.90000       9     9    18   1.8115      0.6       30   0.050      0.0001974    0.04239
0.90000       9     9    18   1.7663      0.6       40   0.050      0.0002636    0.01478
0.90000       9     9    18   1.7315      0.6       50   0.050      0.0003300    0.00515
0.90000       9     9    18   3.3132      1.0       10   0.050      0.0000656    0.34868
0.90000       9     9    18   3.1265      1.0       20   0.050      0.0001314    0.12158
0.90000       9     9    18   3.0192      1.0       30   0.050      0.0001974    0.04239
.             .     .     .        .        .        .       .              .          .
.             .     .     .        .        .        .       .              .          .
.             .     .     .        .        .        .       .              .          .
```

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





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

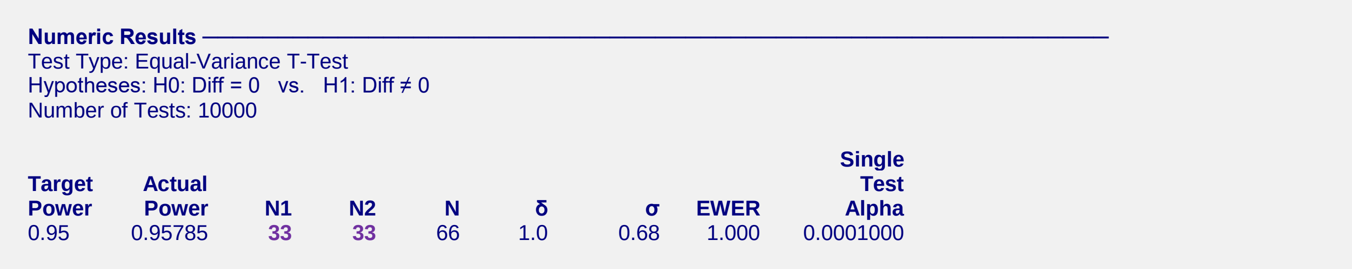
This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Multiple Testing for Two Means** procedure window by expanding **Means**, then **Two Independent Means**, then clicking on **Multiple Testing**, and then clicking on **Multiple Testing for Two Means**. You may then make the appropriate entries as listed below, or open **Example 4** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
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



```
Numeric Results
Test Type: Equal-Variance T-Test
Hypotheses: H0: Diff = 0   vs.   H1: Diff ≠ 0
Number of Tests: 10000

                                                                    Single
Target    Actual                                                      Test
Power     Power      N1     N2      N      δ        σ     EWER       Alpha
0.95      0.95785    33     33     66    1.0     0.68    1.000   0.0001000
```

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

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Multiple Testing for Two Means** procedure window by expanding **Means**, then **Two Independent Means**, then clicking on **Multiple Testing**, and then clicking on **Multiple Testing for Two Means**. You may then make the appropriate entries as listed below, or open **Example 5** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
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

```
Numeric Results
Test Type: Equal-Variance Z-Test
Hypotheses: H0: Diff = 0   vs.   H1: Diff ≠ 0
Number of Tests: 10000

                                                                  Single
Target    Actual                                                    Test
Power     Power     N1     N2      N      δ        σ     EWER      Alpha
0.90      0.90576   23     23     46    1.0    0.707    0.500  0.0005000
0.90      0.93244   11     11     22    1.5    0.707    0.500  0.0005000
0.90      0.92194    6      6     12    2.0    0.707    0.500  0.0005000
0.90      0.93565    4      4      8    2.5    0.707    0.500  0.0005000
```

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

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Multiple Testing for Two Means** procedure window by expanding **Means**, then **Two Independent Means**, then clicking on **Multiple Testing**, and then clicking on **Multiple Testing for Two Means**. You may then make the appropriate entries as listed below, or open **Example 6** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
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

```
Numeric Results
Test Type: Equal-Variance Z-Test
Hypotheses: H0: Diff ≤ 0   vs.   H1: Diff > 0
Number of Tests: 4000


                                                                       Single   Prob To
Target    Actual                                                         Test    Detect
Power      Power    N1     N2      N      δ       σ      K     FDR      Alpha     All K
0.60     0.61099    34     34     68    1.0     1.0     40   0.010  0.0000612   0.00000
```

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.