Chapter 610

Multiple Testing for One Mean (One-Sample or Paired Data)

Introduction

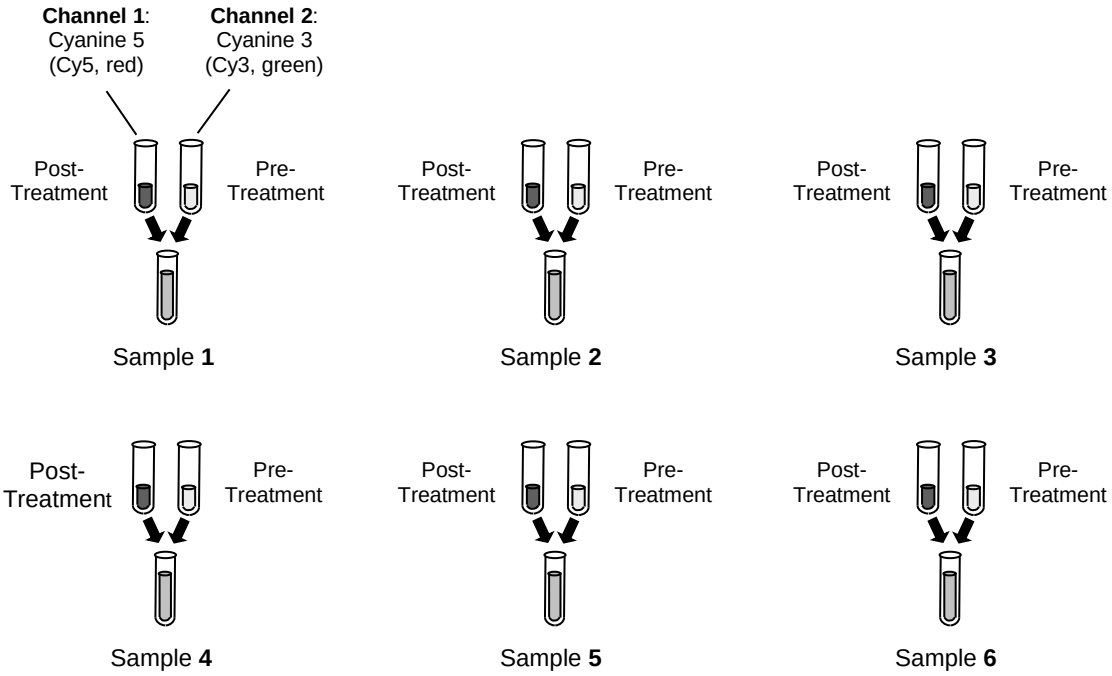
This chapter describes how to estimate power and sample size (e. g. number of arrays in a microarray experiment) for paired and one sample high-throughput studies using the Multiple Testing for One Mean (One-Sample or Paired Data) procedure. False discovery rate and experiment-wise error rate control 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 size (number of arrays), the minimum |mean difference| detected, the standard deviation, and in the case of false discovery rate control, the number of tests with minimum |mean difference| > δ.

Paired Design (Two-Channel Arrays)

The paired design is often used in two-channel microarray experiments when the gene expression comparison to be made involves a natural pairing of experimental units.

As an example, suppose 6 cell samples will be available for comparison. A portion of each of the 6 cell samples (before treatment) is to be reserved as a control. The same treatment will then be given to each of the 6 remaining portions of the samples. It is of interest to determine the genes that are differentially expressed when the treatment is given. In this scenario there is a natural before/after treatment pairing for each sample. The reserved control portions of each sample will be labeled with Cyanine 3 (Cy3, green) dye, while the treatment portions are to be labeled with Cyanine 5 (Cy5, red) dye. From each sample, the labeled control and the labeled treatment portions will be mixed and exposed to an array. The control and treatment portions compete to bind at each spot. The expression of treatment and control samples for each gene will be measured with laser scanning. A pre-processing procedure is then used to obtain expression difference values for each gene. In this example, the result will be 6 relative expression values (e.g., $\text{Log}_2(Post / Pre)$) for each gene represented on the arrays.

Paired Design, Six Arrays



Null and Alternative Hypotheses

The paired test null hypothesis for each gene is $H_0: \mu_{\text{pair}} = \mu_0$, where μ_{pair} is the actual mean paired difference (in expression for a particular gene), and μ_0 is the null-hypothesized paired difference (in expression). A common value for μ_0 in a paired sample experiment is 0. The alternative hypothesis may be any one of the following: $H_1: \mu_{\text{pair}} < \mu_0$, $H_1: \mu_{\text{pair}} > \mu_0$, or $H_1: \mu_{\text{pair}} \neq \mu_0$. The choice of the alternative hypothesis depends upon the goals of the research. For example, if the goal of a microarray experiment is to determine which genes are differentially expressed without regard to direction, the alternative hypothesis would be $H_1: \mu_{\text{pair}} \neq \mu_0$. If, however, the goal is to identify only genes which have increased expression after the treatment is applied, the alternative hypothesis would be $H_1: \mu_{\text{pair}} > \mu_0$.

Paired T-Test Formula

For testing the hypothesis $H_0: \mu_{\text{pair}} = \mu_0$, the formula for the paired T-statistic is:

$$T_{\text{pair}} = \frac{\bar{x}_{\text{pair}} - \mu_0}{\frac{s_{\text{pair}}}{\sqrt{n}}}$$

where $\bar{x}_{\text{pair}}$ is mean paired difference (in expression) of n replicates (for a given gene), μ_0 is the hypothesized mean paired difference, and s_{pair} is standard deviation of the paired differences of the n replicates. If the assumptions (described below) of the test are met, the distribution of T_{pair} is the standard t distribution with $n - 1$ degrees of freedom. P-values are obtained from T_{pair} by finding the proportion of the t distribution that is more extreme than T_{pair}.

Assumptions

Paired Z-Test Assumptions

The assumptions of the paired *z*-test are:

1. The paired (expression) differences are continuous (not discrete). Because of the large range of possible intensities, microarray expression values can be considered continuous.
2. The paired (expression) differences follow a normal probability distribution. This assumption can be examined in microarray data only if the number of arrays in the experiment is reasonably large (>100).
3. The sample of pairs is a simple random sample from its population. Each individual in the population has an equal probability of being selected in the sample. If the samples used in the microarray experiment are not random, bias may easily be introduced into the results.
4. The population standard deviation is known.

Paired T-Test Assumptions

The assumptions of the one-sample or paired *t*-test are:

1. The paired (expression) differences are continuous (not discrete). Because of the large range of possible intensities, microarray expression values can be considered continuous.
2. The paired (expression) differences follow a normal probability distribution. This assumption can be examined in microarray data only if the number of arrays in the experiment is reasonably large (>100).
3. The sample of pairs is a simple random sample from its population. Each individual in the population has an equal probability of being selected in the sample. If the samples used in the microarray experiment are not random, bias may easily be introduced into the results.

Paired Wilcoxon Signed-Rank Test Assumptions

The assumptions of the Wilcoxon signed-rank test are as follows (note that the difference is between a data value and the hypothesized median or between the two data values of a pair):

1. The differences are continuous (not discrete).
2. The distribution of each difference is symmetric.
3. The differences are mutually independent.
4. The differences all have the same median.
5. The measurement scale is at least interval.

One-Sample Design

The one-sample design is the simplest of all designs. A single mRNA or cDNA sample is obtained from each experimental unit of a single group. In the microarray scenario, each sample is exposed to a single microarray, resulting in a single expression value for each gene for each unit of the group. The goal is to determine for each gene whether there is evidence that the expression is different from some null value. This design may be useful for determining whether or not each gene is expressed at all, or for comparing expression of each gene to a hypothesized expression level.

Null and Alternative Hypotheses

The one-sample null hypothesis for each gene is $H_0: \mu = \mu_0$ where μ is the actual mean (expression for a particular gene), and μ_0 is the null-hypothesized mean, or the mean to be compared against. The alternative hypothesis may be any one of the following: $H_1: \mu < \mu_0$, $H_1: \mu > \mu_0$, or $H_1: \mu \neq \mu_0$. The choice of the alternative hypothesis depends upon the goals of the research. For example, if the goal of the experiment is to determine which genes are expressed above a certain level, the alternative hypothesis would be $H_1: \mu > \mu_0$.

T-Test Formula

For testing the hypothesis $H_0: \mu = \mu_0$, the formula for the one-sample T-statistic is:

$$T = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

where $\bar{x}$ is mean (expression) of n replicates (for a given gene), μ_0 is the null-hypothesized mean, and s is standard deviation of the n replicates. If the assumptions (described below) of the test are met, the distribution of T is the standard t distribution with $n - 1$ degrees of freedom. P-values are obtained from T by finding the proportion of the t distribution that is more extreme than T.

Wilcoxon Signed-Rank Test Statistic

The Wilcoxon signed-rank test is a popular, nonparametric substitute for the *t*-test. It assumes that the data follow a symmetric distribution. The test is computed using the following steps.

1. Subtract the hypothesized mean, μ_0, from each data value. Rank the values according to their absolute values.
2. Compute the sum of the positive ranks *Sp* and the sum of the negative ranks *Sn*. The test statistic, W_R, is the minimum of *Sp* and *Sn*.
3. Compute the mean and standard deviation of W_R using the formulas

$$\mu_{W_R} = \frac{n(n+1)}{4}$$

$$\sigma_{W_R} = \sqrt{\frac{n(n+1)(2n+1)}{24} - \frac{\sum t^3 - \sum t}{48}}$$

   where t represents the number of times the i^{th} value occurs.
4. Compute the *z*-value using

$$z_W = \frac{W_R - \mu_{W_R}}{\sigma_{W_R}}$$

The significance of the test statistic is determined by computing the p-value using the standard normal distribution. If this p-value is less than a specified level (usually 0.05), the null hypothesis is rejected in favor of the alternative hypothesis. Otherwise, no conclusion can be reached.

Assumptions

One-Sample Z-Test Assumptions

The assumptions of the one-sample or paired *z*-test are:

1. The data are continuous (not discrete). Because of the large range of possible intensities, microarray expression values can be considered continuous.
2. The data (e.g. the expression values) follow a normal probability distribution. This assumption can be examined in microarray data only if the number of arrays in the experiment is reasonably large (>300).
3. The sample is a simple random sample from its population. Each individual in the population has an equal probability of being selected in the sample. If the samples used in the microarray experiment are not random, bias may easily be introduced into the results.
4. The population standard deviation is known.

One-Sample T-Test Assumptions

The assumptions of the one-sample or paired *t*-test are:

1. The data are continuous (not discrete). Because of the large range of possible intensities, microarray expression values can be considered continuous.
2. The data (e.g. the expression values) follow a normal probability distribution. This assumption can be examined in microarray data only if the number of arrays in the experiment is reasonably large (>300).
3. The sample is a simple random sample from its population. Each individual in the population has an equal probability of being selected in the sample. If the samples used in the microarray experiment are not random, bias may easily be introduced into the results.

One-Sample Wilcoxon Signed-Rank Test Assumptions

The assumptions of the Wilcoxon signed-rank test are as follows:

1. The data are continuous (not discrete).
2. The distribution is symmetric.
3. The data are mutually independent.
4. The data have the same median.
5. The measurement scale is at least interval.

Technical Details

Multiple Testing Adjustment

When a one-sample/paired T-test is run for a replicated microarray experiment, the result is a list of P-values (Probability Level) 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

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

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

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

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

One-Sample Z-Test

When the standard deviation is known, the power is calculated as follows for a directional alternative (one-tailed test) in which $\mu_1 > \mu_0$.

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: $X_1 = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$.
3. Calculate: $z_1 = \frac{X_1 - \mu_1}{\frac{\sigma}{\sqrt{n}}}$.
4. Power = $1 - \Phi(z_1)$.

One-Sample T-Test

When the standard deviation is unknown, the power is calculated as follows for a directional alternative (one-tailed test) in which $\mu_1 > \mu_0$.

1. Find t_α such that $1 - T_{df}(t_\alpha) = \alpha$, where $T_{df}(t_\alpha)$ is the area under a central-*t* curve to the left of *x* and $df = n - 1$.
2. Calculate: $X_1 = \mu_0 + t_\alpha \frac{\sigma}{\sqrt{n}}$.
3. Calculate the noncentrality parameter: $\lambda = \frac{\mu_1 - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{\delta_1}{\frac{\sigma}{\sqrt{n}}}$.
4. Calculate: $t_1 = \frac{X_1 - \mu_1}{\frac{\sigma}{\sqrt{n}}} + \lambda$.
5. 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.

Wilcoxon Signed-Rank Test

The power calculation for the Wilcoxon signed-rank test is the same as that for the one-sample *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' used in power calculations is equal to

$$n' = n/W,$$

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

The adjustments are as follows:

Distribution	W
Uniform	1
Double Exponential	$2/3$
Logistic	$9/\pi^2$
Normal	$\pi/3$

The power is calculated as follows for a directional alternative (one-tailed test) in which $\mu_1 > \mu_0$.

1. Find t_α such that $1 - T_{df}(t_\alpha) = \alpha$, where $T_{df}(t_\alpha)$ is the area under a central-*t* curve to the left of *x* and $df = n' - 1$.
2. Calculate: $X_1 = \mu_0 + t_\alpha \frac{\sigma}{\sqrt{n'}}$.
3. Calculate the noncentrality parameter: $\lambda = \frac{\mu_1 - \mu_0}{\frac{\sigma}{\sqrt{n'}}} = \frac{\delta_1}{\frac{\sigma}{\sqrt{n'}}}$.
4. Calculate: $t_1 = \frac{X_1 - \mu_1}{\frac{\sigma}{\sqrt{n'}}} + \lambda$.
5. 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.

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.

- **T-Test**

  The T-Test assumes the differences come from a normal distribution with UNKNOWN standard deviation (i.e., a standard deviation that will be estimated from the data).

- **Z-Test**

  The Z-Test assumes the differences come from a normal distribution with KNOWN standard deviation.

- **Wilcoxon Signed-Rank Test**

  The Wilcoxon Signed-Rank Test is the nonparametric analog of the one-sample or paired T-Test.

Data Distribution

Displayed only if Test Type = Wilcoxon Signed-Rank Test

This option makes appropriate sample size adjustments for the Wilcoxon Signed-Rank test. Results by Al-Sunduqchi and Guenther (1990) indicate that power calculations for the Wilcoxon Signed-Rank 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 four distributions.

The options are as follows:

- **Uniform**

  The sample size adjustment factor, W, is equal to “1”. This selection gives the same result as the one-sample t-test.

- **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 paired 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 whether there is expression above some threshold, 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 $0 < \text{FDR} \leq 1$. Commonly, $0.001 \leq \text{FDR} \leq 0.25$. 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 $0 < \text{EWER} \leq 1$. Commonly, $0.001 \leq \text{EWER} \leq 0.25$. 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

N (Sample Size)

Enter a value for the sample size (N).

Microarray Data

For a microarray experiment, this is the number of arrays. For two-channel paired microarray experiments, this is the number of arrays, not the number of samples.

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

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).

Paired Data

For paired data, δ is the minimum paired difference.

One-Sample Data

For one-sample data, δ = μ - μ0, where μ is the minimum mean to detect and μ0 is the value of the mean for comparison under the null hypothesis.

Microarray Expression Studies

In paired expression studies, it is very common that the difference in expression is measured on the log scale (e.g., log2(A) - log2(B)). Values of δ should reflect the differences that will be used in testing. For example, if log2 differences are 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 of Differences)

Specify the standard deviation of the differences. This standard deviation is assumed for all tests. σ should be on the same scale as δ. Because the true variation will vary from test to test, it is recommended that a range of values be entered here.

Paired Data

For paired data, this is the standard deviation of paired differences. To obtain the standard deviation of paired differences from the standard deviation of the data, use σ_paired = (√2)*(σ_data).

One-Sample Data

For one-sample data, this is the standard deviation of the data.

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.

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 that involved 22 two-channel arrays. Two samples were obtained from each of 22 individuals. One of the two samples was randomly assigned the treatment and the other remained as the control. Following treatment, the two samples were exposed to a single microarray. Each microarray produced intensity information for 10,000 genes. The 22 arrays were pre-processed by subtracting the control intensity (Log2) from the treatment intensity for each gene on each array. Thus, a positive value implies upward expression in the treatment, while a negative value implies down-regulation in the treatment. In this example, the paired T-test was used to determine which genes were differentially expressed (upward or downward) following exposure to the treatment.

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 of the Log2 paired differences 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 One Mean (One-Sample or Paired Data)** procedure window by expanding **Means**, then **One Mean**, then clicking on **Multiple Testing**, and then clicking on **Multiple Testing for One Mean (One-Sample or Paired Data)**. 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	**T-Test**
Alternative Hypothesis	**Two-Sided**
False Discovery (Alpha) Method	**FDR (False Discovery Rate)**
FDR (False Discovery Rate)	**0.05**
N (Sample Size)	**22**
δ (Minimum |Mean Difference| Detected)	**1**
σ (Standard Deviation of Differences)	**0.2 to 2 by 0.2**
Number of Tests	**10000**
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: T-Test
Hypotheses: H0: Diff = 0 vs. H1: Diff ≠ 0
Number of Tests: 10000

Power	Sample Size N	Min Mean Diff δ	Std Dev of Diffs σ	Effect Size ES	Number of Tests with |Diff| > δ K	False Disc Rate FDR	Single Test Alpha	Prob To Detect All K	Beta
1.00000	22	1.0	0.2	5.0	10	0.050	0.0000527	1.00000	0.00000
1.00000	22	1.0	0.2	5.0	50	0.050	0.0002645	1.00000	0.00000
1.00000	22	1.0	0.2	5.0	100	0.050	0.0005316	1.00000	0.00000
1.00000	22	1.0	0.4	2.5	10	0.050	0.0000527	1.00000	0.00000
1.00000	22	1.0	0.4	2.5	50	0.050	0.0002645	1.00000	0.00000
1.00000	22	1.0	0.4	2.5	100	0.050	0.0005316	1.00000	0.00000
0.98617	22	1.0	0.6	1.7	10	0.050	0.0000520	0.86996	0.01383
0.99793	22	1.0	0.6	1.7	50	0.050	0.0002639	0.90158	0.00207
0.99924	22	1.0	0.6	1.7	100	0.050	0.0005312	0.92649	0.00076
0.71696	22	1.0	0.8	1.3	10	0.050	0.0000378	0.03589	0.28304
0.89092	22	1.0	0.8	1.3	50	0.050	0.0002356	0.00310	0.10908
0.93538	22	1.0	0.8	1.3	100	0.050	0.0004973	0.00126	0.06462
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.

Report Definitions

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

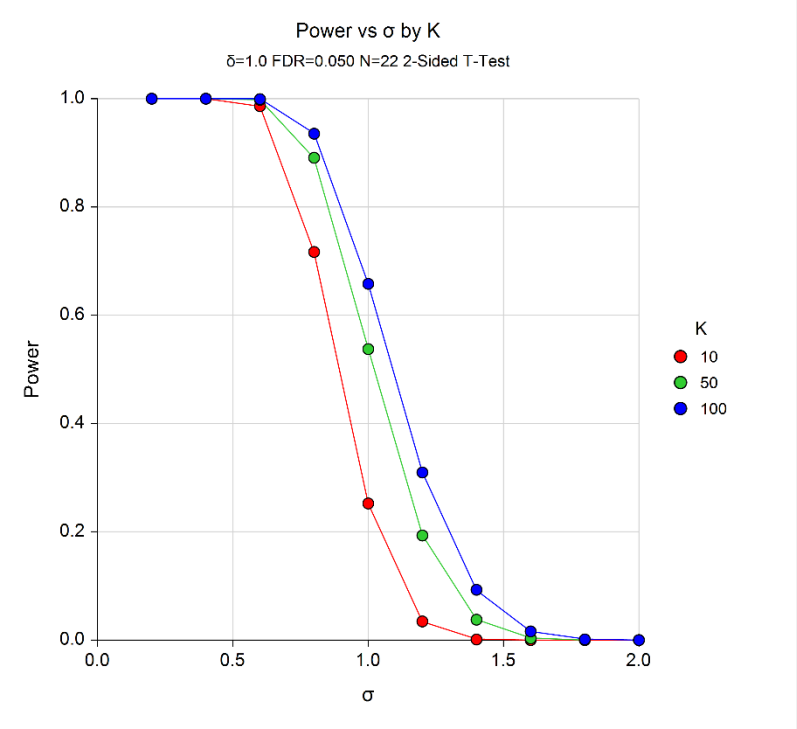
Power is the individual probability of detecting a difference for each test with actual |mean difference| > δ.
N is the sample size (e.g. number of arrays for microarray studies) required to achieve the corresponding power.
δ is the smallest |mean difference| for which the power and sample size calculations are valid.
σ is the standard deviation of differences used in each test.
ES = δ/σ, is the relative magnitude of the mean expression difference for the tests where |mean difference| > δ.
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.
Beta is the individual probability of failing to detect a difference for each test with |mean difference| > δ.

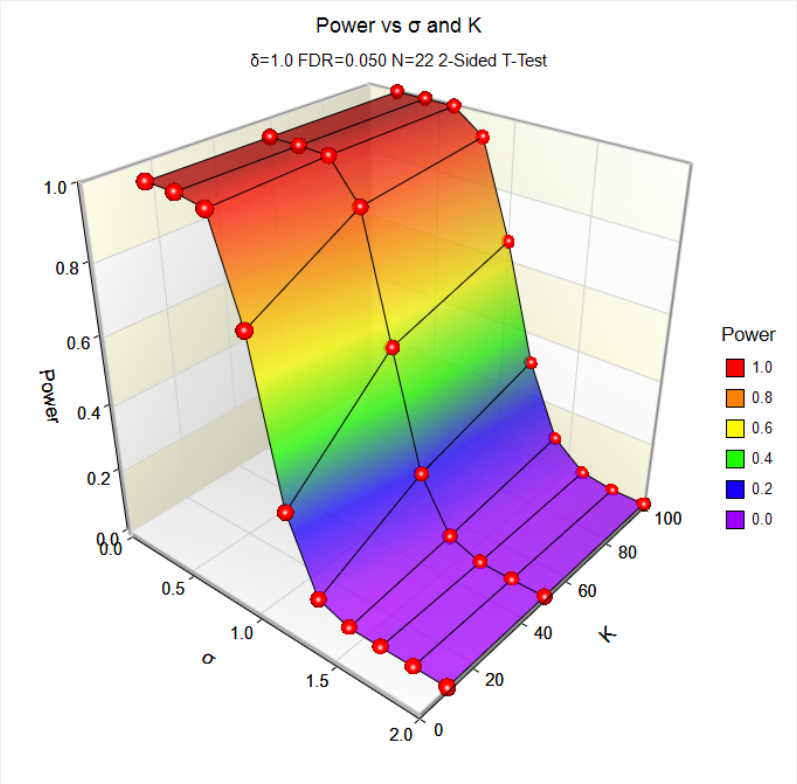
Summary Statements

A sample size of 22 achieves 100% power for each test to detect an actual |mean difference| of at least 1.0 with an estimated standard deviation of 0.2 and a false discovery rate (FDR) of 0.050 using a two-sided one-sample 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 0.0000527. These results assume a total 10000 individual tests are considered.

This report shows the values of each of the parameters, one scenario per row. The values of power and beta 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.

Example 2 – Finding the Sample Size

This example determines the number of two-channel arrays needed to achieve 80% power to detect differential expression for each gene. Two samples will be obtained from each of the sampled individuals. One of the two samples will be randomly assigned the treatment and the other will remain as the control. Following treatment, the two samples will be exposed to a single microarray. Each microarray will produce intensity information for 12,682 genes. The arrays will be pre-processed by subtracting the control intensity (Log2) from the treatment intensity for each gene on each array. Thus, a positive value implies upward expression in the treatment, while a negative value implies down-regulation in the treatment. The paired 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 of the Log2 paired differences for this experiment 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 One Mean (One-Sample or Paired Data)** procedure window by expanding **Means**, then **One Mean**, then clicking on **Multiple Testing**, and then clicking on **Multiple Testing for One Mean (One-Sample or Paired Data)**. 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	**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**
δ (Minimum |Mean Difference| Detected)	**1**
σ (Standard Deviation of Differences)	**0.2 to 2 by 0.2**
Number of Tests	**12682**
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.

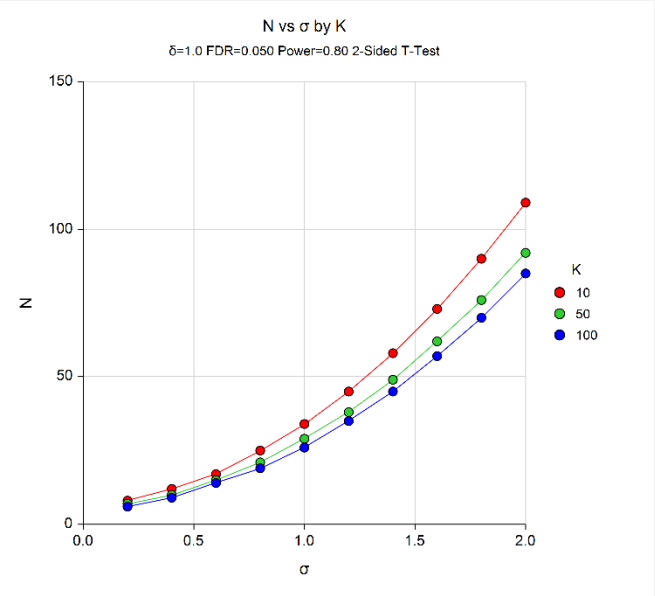
Numeric Results

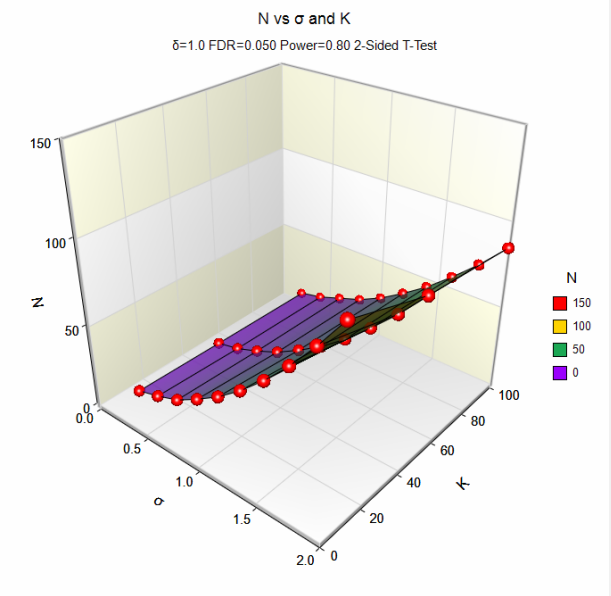
Numeric Results
Test Type: T-Test
Hypotheses: H0: Diff = 0 vs. H1: Diff ≠ 0
Number of Tests: 12682

Power	Sample Size N	Min Mean Diff δ	Std Dev of Diffs σ	Effect Size ES	Number of Tests with |Diff| > δ K	False Disc Rate FDR	Single Test Alpha	Prob To Detect All K	Beta
0.96741	8	1.0	0.2	5.0	10	0.050	0.0000332	0.71799	0.03259
0.97509	7	1.0	0.2	5.0	50	0.050	0.0001667	0.28324	0.02491
0.91190	6	1.0	0.2	5.0	100	0.050	0.0003346	0.00010	0.08810
0.88530	12	1.0	0.4	2.5	10	0.050	0.0000332	0.29573	0.11470
0.86231	10	1.0	0.4	2.5	50	0.050	0.0001667	0.00061	0.13769
0.83278	9	1.0	0.4	2.5	100	0.050	0.0003346	0.00000	0.16722
0.81531	17	1.0	0.6	1.7	10	0.050	0.0000332	0.12979	0.18469
0.85472	15	1.0	0.6	1.7	50	0.050	0.0001667	0.00039	0.14528
0.86398	14	1.0	0.6	1.7	100	0.050	0.0003346	0.00000	0.13602
0.83661	25	1.0	0.8	1.3	10	0.050	0.0000332	0.16797	0.16339
0.82805	21	1.0	0.8	1.3	50	0.050	0.0001667	0.00008	0.17195
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.

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 deviation of the differences 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 14 two-channel arrays. The 14 arrays permit tests on 5,438 genes. The arrays will be pre-processed by subtracting the control intensity (Log2) from the treatment intensity for each gene on each array. Thus, a positive value implies upward expression in the treatment, while a negative value implies down-regulation in the treatment. The paired T-test will be used to determine which genes are differentially expressed (upward or downward) following exposure to the treatment. Standard deviations of the Log2 paired differences 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 One Mean (One-Sample or Paired Data)** procedure window by expanding **Means**, then **One Mean**, then clicking on **Multiple Testing**, and then clicking on **Multiple Testing for One Mean (One-Sample or Paired Data)**. 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	**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**
N (Sample Size)	**14**
σ (Standard Deviation of Differences)	**0.2 to 1.8 by 0.4**
Number of Tests	**5438**
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.

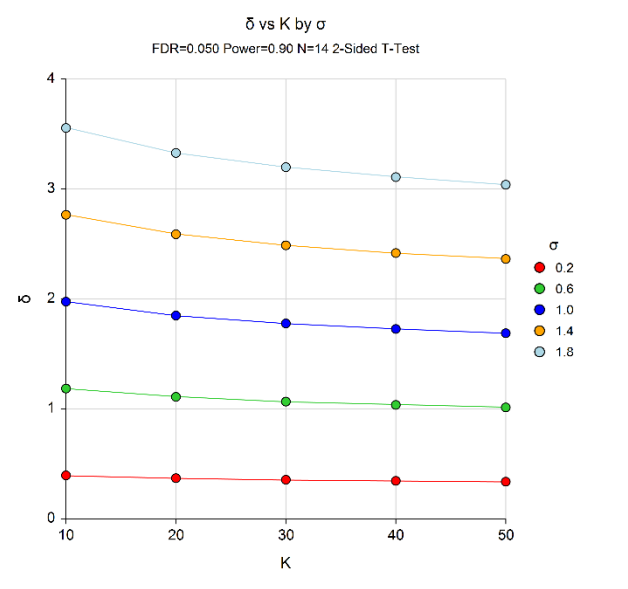
Numeric Results

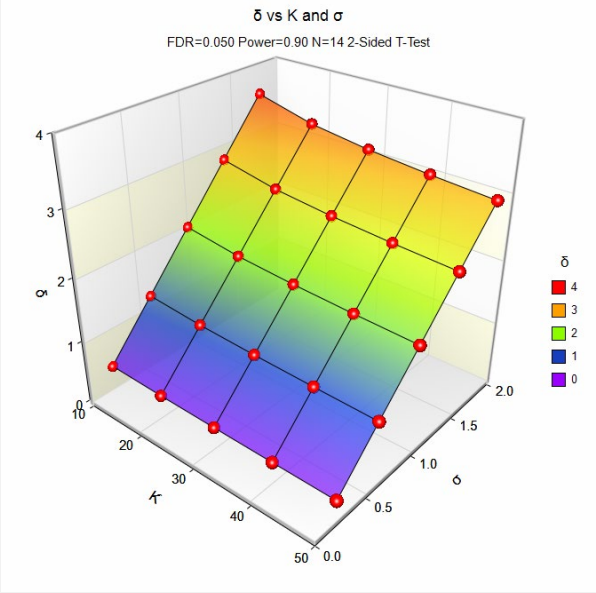
Numeric Results
Test Type: T-Test
Hypotheses: H0: Diff = 0 vs. H1: Diff ≠ 0
Number of Tests: 5438

Power	Sample Size N	Min Mean Diff δ	Std Dev of Diffs σ	Effect Size ES	Number of Tests with |Diff| > δ K	False Disc Rate FDR	Single Test Alpha	Prob To Detect All K	Beta
0.90000	14	0.3951	0.2	2.0	10	0.050	0.0000873	0.34868	0.10000
0.90000	14	0.3699	0.2	1.8	20	0.050	0.0001749	0.12158	0.10000
0.90000	14	0.3555	0.2	1.8	30	0.050	0.0002628	0.04239	0.10000
0.90000	14	0.3454	0.2	1.7	40	0.050	0.0003510	0.01478	0.10000
0.90000	14	0.3377	0.2	1.7	50	0.050	0.0004396	0.00515	0.10000
0.90000	14	1.1854	0.6	2.0	10	0.050	0.0000873	0.34868	0.10000
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.

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 deviation of the differences for five values of K.

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

Stekel (2003), pp. 226-228, gives an example in which N = 20, δ = 1, and σ = 0.68 for a two-sided paired T-Test. The number of genes tested is 6500. The control of false discoveries is "no more than one false positive." This corresponds to an EWER value of 0.975. The power obtained for this example is 0.94.

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 One Mean (One-Sample or Paired Data)** procedure window by expanding **Means**, then **One Mean**, then clicking on **Multiple Testing**, and then clicking on **Multiple Testing for One Mean (One-Sample or Paired Data)**. 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	**Power**
Test Type	**T-Test**
Alternative Hypothesis	**Two-Sided**
False Discovery (Alpha) Method	**EWER (Experiment-wise Error Rate)**
EWER (Experiment-wise Error Rate)	**0.975**
N (Sample Size)	**20**
δ (Minimum |Mean Difference| Detected)	**1**
σ (Standard Deviation of Differences)	**0.68**
Number of Tests	**6500**
Reports Tab	
σ Decimals	**2**

Output

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

Numeric Results

Numeric Results
Test Type: T-Test
Hypotheses: H0: Diff = 0 vs. H1: Diff ≠ 0
Number of Tests: 6500

Power	Sample Size N	Min Mean Diff δ	Std Dev of Diffs σ	Effect Size ES	Experiment -Wise Error Rate EWER	Single Test Alpha	Beta
0.93591	20	1.0	0.68	1.5	0.975	0.0001500	0.06409

The power of 0.93591 matches 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 = 1.0 for a two-sided paired Z-Test. 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 One Mean (One-Sample or Paired Data)** procedure window by expanding **Means**, then **One Mean**, then clicking on **Multiple Testing**, and then clicking on **Multiple Testing for One Mean (One-Sample or Paired Data)**. 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	**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**
δ (Minimum |Mean Difference| Detected)	**1 1.5 2 2.5**
σ (Standard Deviation of Differences)	**1**
Number of Tests	**1000**

Output

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

Numeric Results

Numeric Results
Test Type: Z-Test
Hypotheses: H0: Diff = 0 vs. H1: Diff ≠ 0
Number of Tests: 1000

Power	Sample Size N	Min Mean Diff δ	Std Dev of Diffs σ	Effect Size ES	Experiment-Wise Error Rate EWER	Single Test Alpha	Beta
0.90576	23	1.0	1.0	1.0	0.500	0.0005000	0.09424
0.93244	11	1.5	1.0	1.5	0.500	0.0005000	0.06756
0.92194	6	2.0	1.0	2.0	0.500	0.0005000	0.07806
0.93565	4	2.5	1.0	2.5	0.500	0.0005000	0.06435

Sample sizes of 23, 11, 6, and 4 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). We adapt the effect size in this validation to correspond to a one-sample test. Namely, the effect size is reduced by one half. In the example, Power = 0.60 (from 24/40), δ = 1.0, and σ = 1.0 for a one-sided two-sample Z-Test. We use σ = 2.0 to correspond to the equivalent in the one-sample 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 for this setup is 68.

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 One Mean (One-Sample or Paired Data)** procedure window by expanding **Means**, then **One Mean**, then clicking on **Multiple Testing**, and then clicking on **Multiple Testing for One Mean (One-Sample or Paired Data)**. 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	**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.

Numeric Results

Numeric Results

Test Type: Z-Test
Hypotheses: H0: Diff ≤ 0 vs. H1: Diff > 0
Number of Tests: 4000

Power	Sample Size N	Min Mean Diff δ	Std Dev of Diffs σ	Effect Size ES	Number of Tests with |Diff| > δ K	False Disc Rate FDR	Single Test Alpha	Prob To Detect All K	Beta
0.61099	**68**	1.0	2.0	0.5	40	0.010	0.0000612	0.00000	0.38901

A sample size of 68 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.