Chapter 4

Significance Tests

Aim: Upon completing the chapter, the learner should be able to perform significance tests that are concerned with the expected values of continuous random variables.

4.1 Introduction

Classical statistical tests are performed to compare the expected values of a random variable, under the assumption that these values are constant parameters of the target population. The Bayesian approach assumes that these parameters are another random variable. In this chapter we will concentrate our analysis using classical statistical tests for comparing the expected values of a continuous random variable in two independent groups.

The classical statistical tests are based on the initial formulation of two complementary hypotheses that are related to the parameters of the target population; these hypotheses are the null and the alternative hypotheses. The null hypothesis, denoted by H_0, is the hypothesis that is to be tested. The alternative hypothesis, usually denoted by H_a, is the hypothesis that contradicts the null hypothesis (Rosner, 2010); usually, the alternative hypothesis will be related to a research hypothesis. To assess the null hypothesis, a sample of data is collected to compute a *test statistic* for supporting a decision in favor of or against the H_0; there are four possible outcomes:

1. Evidence in favor of H_0, with H_0 in fact being true
2. Evidence against H_0, though H_0 is in fact true (Type I error)
3. Evidence in favor of H_0, though H_0 is in fact false (Type II error)
4. Evidence against H_0, with H_0 in fact being false

In classical statistics, the probabilities of the occurrences of these outcomes are summarized in the following table:

Decision based on the sample data	H_0	
	True	*False*
Evidence in favor of H_0 (do not reject H_0)	$1 - \alpha = \Pr[\text{accept } H_0 \mid H_0 \text{ true}]$	$\beta = \Pr[\text{accept } H_0 \mid H_0 \text{ false}]$ *Probability of Type II error*
Evidence against H_0 (reject H_0)	$\alpha = \Pr[\text{reject } H_0 \mid H_0 \text{ true}]$ *Probability of Type I error* **(significance level)**	$1 - \beta = \Pr[\text{reject } H_0 \mid H_0 \text{ false}]$ **Statistical power**

The general aim in hypothesis testing is to use statistical tests that make α and β as small as possible. Typically, the evidence against H_0 is determined with a significance level less than or equal to 5%, while a statistical power of 80% or higher is considered adequate.

The significance level can be defined prior to performing the test; when this is done, two regions for the test statistics are defined: the *acceptance region* (evidence for accepting H_0) and the *rejection region* (evidence against the null hypothesis). However, the output of the statistical programs usually shows the probability (called *P*-value) for each statistical test. *P*-value is defined as the probability of obtaining a test statistic as extreme as or more extreme than the test statistic actually obtained, given that the null hypothesis is true. As a consequence, the *P*-value is interpreted as α level at which the given value of the test statistic is on the borderline between the acceptance and rejection regions (Rosner, 2010). In Stata, the *P*-value will be presented according to the test statistic used and the probability distribution assumed for this statistic; for example, assuming the Student's *t*-test statistic with t-probability distribution, the output for identifying the *P*-value will be expressed as $\Pr(T > t)$. To interpret *P*-values, we can use one of the following statements (Rosner, 2010):

If the *P*-value $\geq .05$, then the results are considered not statistically significant.
If $.01 <$ *P*-value $< .05$, then the results are significant.
If $.001 <$ *P*-value $\leq .01$, then the results are highly significant.
If the *P*-value $\leq .001$, then the results are very highly significant.

However, if $.05 \leq$ *P*-value $< .10$, then a trend toward statistical significance is sometimes noted.

4.2 Normality Test

When we want to estimate or compare the expected value of a continuous random variable, usually we assume that this variable follows a normal probability distribution. To assess whether the normality assumption is met, different statistical tests can be performed. A formal test to evaluate the normality of a continuous random variable is the *Shapiro–Wilk test,* whose null hypothesis states that a given random variable follows a normal distribution (Rosner, 2010). The *swilk* command can be used for this purpose. For example, to determine whether the continuous variables of the previous database follow a normal distribution, the command line below can be used:

```
swilk age weikg heimt bmi
```

Output

```
                Shapiro-Wilk W test for normal data

    Variable |       Obs        W           V          z       Prob>z
-------------+--------------------------------------------------------
         age |        10    0.82089      2.760      1.943    0.02598
       weikg |        10    0.94243      0.887     -0.203    0.58031
       heimt |        10    0.93312      1.031      0.052    0.47923
         bmi |        10    0.95562      0.684     -0.628    0.73506
```

The results above provide evidence in favor of the null hypothesis for all variables (P-value > .05) with the exception of the variable age (P-value = .0259).

4.3 Variance Homogeneity

An assumption that can be used in the performance of a parametric test to compare the expected values of a continuous random variable in two unmatched groups is that the variances are equal (variance homogeneity). This indicates that the location parameters (expected values) of the continuous random variables can be different in each group, but the dispersion parameter (variance) is equal in all groups. To perform this assessment in two groups, the *sdtest* command is available; for this assessment, it is assumed that the variance ratio of the two groups follows an F-Fisher probability distribution $\left(\sigma_1^2/\sigma_2^2 \sim F\right)$. The command lines are as follows for the variables *bmi, weight,* and *height* (all from the previous database):

```
sdtest  bmi, by(sex)
sdtest  weikg, by(sex)
sdtest  heimt, by(sex)
```

Output

```
. sdtest  bmi, by(sex)

Variance ratio test
------------------------------------------------------------------------------
   Group |  Obs      Mean   Std. Err. Std. Dev. [95% Conf. Interval]
---------+--------------------------------------------------------------------
    Male |    5  24.59281   2.133509  4.770671   18.66924   30.51638
  Female |    5  32.31535   3.094344  6.919164   23.72407   40.90663
---------+--------------------------------------------------------------------
combined |   10  28.45408   2.189954  6.925241   23.50006    33.4081
------------------------------------------------------------------------------
    ratio = sd(Male) / sd(Female)                              f = 0.4754
Ho: ratio = 1                          degrees of freedom =   4, 4

   Ha: ratio < 1        Ha: ratio != 1            Ha: ratio > 1
 Pr(F < f) = 0.2446  2*Pr(F < f) = 0.4891      Pr(F > f) = 0.7554

. sdtest  weik, by(sex)

Variance ratio test
------------------------------------------------------------------------------
   Group |  Obs    Mean   Std. Err.  Std. Dev.  [95% Conf. Interval]
---------+--------------------------------------------------------------------
    Male |    5    60.4   6.910861   15.45316   41.21237   79.58763
  Female |    5    72.2   5.580323   12.47798   56.70654   87.69346
---------+--------------------------------------------------------------------
combined |   10    66.3   4.626133   14.62912   55.83496   76.76504
------------------------------------------------------------------------------
    ratio = sd(Male) / sd(Female)                              f = 1.5337
Ho: ratio = 1                          degrees of freedom =   4, 4

   Ha: ratio < 1        Ha: ratio != 1            Ha: ratio > 1
 Pr(F < f) = 0.6556  2*Pr(F > f) = 0.6887      Pr(F > f) = 0.3444

. sdtest  heimt, by(sex)

Variance ratio test
------------------------------------------------------------------------------
   Group |  Obs   Mean   Std. Err.  Std. Dev.  [95% Conf. Interval]
---------+--------------------------------------------------------------------
    Male |    5   1.56   .0692098   .1547579   1.367843  1.752157
  Female |    5    1.5   .0219089   .0489897   1.439171  1.560829
---------+--------------------------------------------------------------------
combined |   10   1.53   .0356526   .1127435   1.449348  1.610652
------------------------------------------------------------------------------
    ratio = sd(Male) / sd(Female)                              f = 9.9792
Ho: ratio = 1                          degrees of freedom =   4, 4

   Ha: ratio < 1        Ha: ratio != 1            Ha: ratio > 1
 Pr(F < f) = 0.9766  2*Pr(F > f) = 0.0468      Pr(F > f) = 0.0234
```

For each variable (in the above case, *sex*), a table displays a description of the summary measures in each category of that variable: *Obs* (number of observations), *Mean, Std. Err.* (standard error), *Std. Dev.* (standard deviation), and 95% Conf. Interval (the 95% confidence interval is used to estimate the expected value of the random variables). In the above table, the user can see that the standard deviation of the variable bmi among males is 4.77 (variance = 22.75), while among females it is 6.92 (variance = 47.74); so the estimated ratio of the variances is 0.4754 (male/female). If the variances are equal in these two groups, the expected value of this ratio must be 1 (Rosner, 2010). Near the bottom of the table, the user can see that the null hypothesis is "H_0: ratio = 1." The alternative hypothesis is expressed in three ways: H_a: ratio < 1, H_a: ratio != 1 (different than 1), and H_a: ratio > 1; it is recommended that only the second alternative hypothesis *(ratio is different from 1)* be considered, if the purpose is assessing the variance homogeneity. Below each alternative hypothesis, the corresponding *P*-values are presented. Only for the variable *height* does the statistical evidence not support the assumption of variance homogeneity (*P*-value = .0468).

4.4 Student's *t*-Test for Independent Samples

Assuming that the assumptions of normality and variance homogeneity are met, the next step is to determine whether the expected value of the continuous variable changes in different groups. Suppose that the user wants to compare the expected value of the variable bmi by sex, assuming that the selection of a male is independent of the selection of a female in the study sample; for this, Student's *t*-test for independent samples (Rosner, 2010) can be used, as is demonstrated with the following expression:

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\mathrm{Var}\left(\bar{Y}_1 - \bar{Y}_2\right)}} \sim t_{k|H_0}$$

where:

$\bar{Y}_i$ indicates the sample mean of the variable bmi for the ith group
t_k is the *t-probability distribution* with k degrees of freedom
$\mathrm{Var}\left(\bar{Y}_1 - \bar{Y}_2\right)$ is the variance of $\left(\bar{Y}_1 - \bar{Y}_2\right)$

To compute the *P*-value, it is assumed that this expression follows the *t-probability distribution* under the null hypothesis assumption. To perform this kind of *t*-test, the user can utilize the *ttest* command. The specifications for this command can change depending on the structure of the database. For example, assuming that the

previous database is being used and that the aim of the user is to assess the variance homogeneity in the variable bmi by sex group, the command line for performing student's *t*-test is as follows:

```
ttest bmi, by(sex)
```

Output

```
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group| Obs      Mean Std. Err. Std. Dev.  [95% Conf. Interval]
---------+--------------------------------------------------------------------
  "Male"|   5  24.59281  2.133509  4.770671    18.66924  30.51638
"Female"|   5  32.31535  3.094344  6.919164    23.72407  40.90663
---------+--------------------------------------------------------------------
combined|  10  28.45408  2.189954  6.925241    23.50006   33.4081
---------+--------------------------------------------------------------------
    diff|     -7.722543  3.758567             -16.38981  .9447286
------------------------------------------------------------------------------
    diff = mean("Male") - mean("Female")                 t = -2.0547
Ho: diff = 0                     degrees of freedom =       8

   Ha: diff < 0           Ha: diff != 0                 Ha: diff > 0
Pr(T < t) = 0.0370  Pr(|T| > |t|) = 0.0740      Pr(T > t) =  0.9630
```

The above table is the same as the one described by the *sdtest* command. However, the null hypothesis formulated below in this table is different. The null hypothesis states that the expected bmi value is the same for both sexes ($\mu_{Male} = \mu_{Female}$). In Stata notation, this hypothesis is formulated as the following: diff = mean(Male) – mean(Fem) = 0. The alternative hypotheses that can be assessed are H_a: diff < 0, H_a: diff != 0 *(different than zero)*, and H_a: diff > 0. Assuming that the research hypothesis is that *males have a lower mean body mass index than females do,* the user has to assess the *P*-value below the first alternative hypothesis (one-tailed alternative hypothesis), with the result indicating that there is statistical evidence against the null hypothesis (*P*-value = .037); this finding suggests that the expected bmi in males is lower than the expected bmi in females. If the research hypothesis is that *males have different mean body mass index than females do*, then the user has to assess the *P*-value below the second alternative hypothesis (two-tailed alternative hypothesis), with this result indicating that there is statistical evidence in favor of the null hypothesis (*P*-value = .074); this finding suggests that the expected bmi in males is not different from the expected bmi in females.

4.5 Confidence Intervals for Testing the Null Hypothesis

Another way to assess the null hypothesis is to use a confidence interval. For example, to assess the statistical hypotheses, the following can be used: $H_0 : \mu_{\text{Male}} - \mu_{\text{Female}} = 0$ vs. $H_a : \mu_{\text{Male}} - \mu_{\text{Female}} \neq 0$. In this case, to test the null hypothesis (with a 5% significance level), the 95% confidence level for the mean difference provided in the *ttest* command can be used. The interval indicates that with a 95% confidence level, the difference of the expected bmi by sex $\left(\mu_{\text{Male}} - \mu_{\text{Female}}\right)$ is between −16.39 and 0.94. As zero is included in the interval, this finding provides evidence in favor of the null hypothesis (*P*-value > .05).

4.6 Nonparametric Tests for Unpaired Groups

There are several tests available to assess the distribution of the quantitative random variable in two groups when the basic assumptions of the *t*-test are violated. One of these tests is the *Mann–Whitney test,* also called the *Wilcoxon rank sum test,* which is a nonparametric test that compares two unpaired groups. To perform the Mann–Whitney test, all data are ranked, paying no attention to which group each value belongs; the smallest number gets a rank of 1 and the largest number gets a rank of *n*, where *n* is the total number of values in the two groups. Then, this test computes the average of the ranks in each group and reports the two averages. If the means of the ranks in the two groups are very different, the *P*-value will be small. The null hypothesis is that the distributions of both groups will be identical, so that there is a 50% probability that an observation from a value randomly selected from one population will exceed an observation randomly selected from the other population. In Stata, this test can be performed with the command *kwallis*. Using the data from the previous example, the syntax of the *kwallis* command is as follows:

```
kwallis bmi, by(sex)
```

Output

```
Kruskal-Wallis equality-of-populations rank test
  +---------------------------+
  |      sex | Obs | Rank Sum |
  |----------+-----+----------|
  |     Male |   5 |    20.00 |
  |   Female |   5 |    35.00 |
  +---------------------------+
```

```
chi-squared =     2.455 with 1 d.f.
probability =     0.1172

chi-squared with ties =     2.455 with 1 d.f.
probability =     0.1172
```

As can be seen above, the results of this test show that the frequency distributions of the bmi for both sexes are identical (P-value = .1172). This interpretation is consistent with that of Student's t-test for the two-tailed alternative hypothesis. An extensive review of the parametric and nonparametric statistical procedures can be found in the book of Sheskin (2007).

4.7 Sample Size and Statistical Power

The study design for comparing two means involves the minimum sample size and statistical power needed to reduce the probability of a false conclusion. Stata provides the *Power and sample size* option on the *Statistics* dropdown menu (Figure 4.1).

Once the *Statistics* menu is clicked, a new window will be displayed (Figure 4.2). If you choose the test comparing two independent means, a dialog window, in which the user enters the data according to the study problem, is displayed. For example, assuming that we are interested in determining the minimum sample size of the previous example for a two-sided test with a 5% significance level, an 80% statistical power, and an allocation ratio equal to 1 (equal sample size in each group), the window should be completed in the manner seen in Figure 4.3.

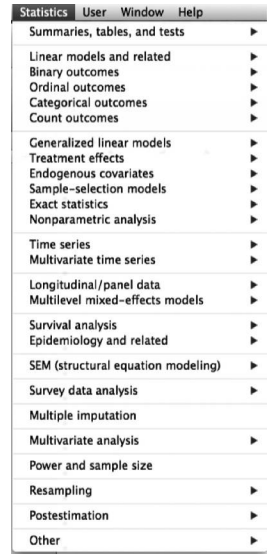

Figure 4.1 Statistics options.

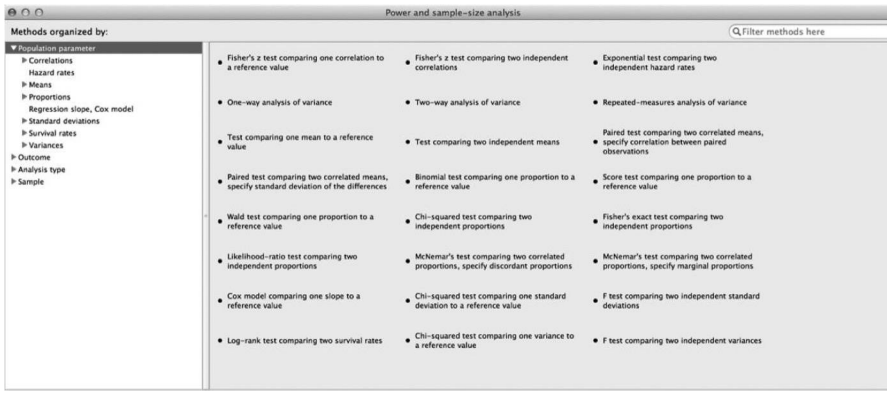

Figure 4.2 Power and sample size analysis options.

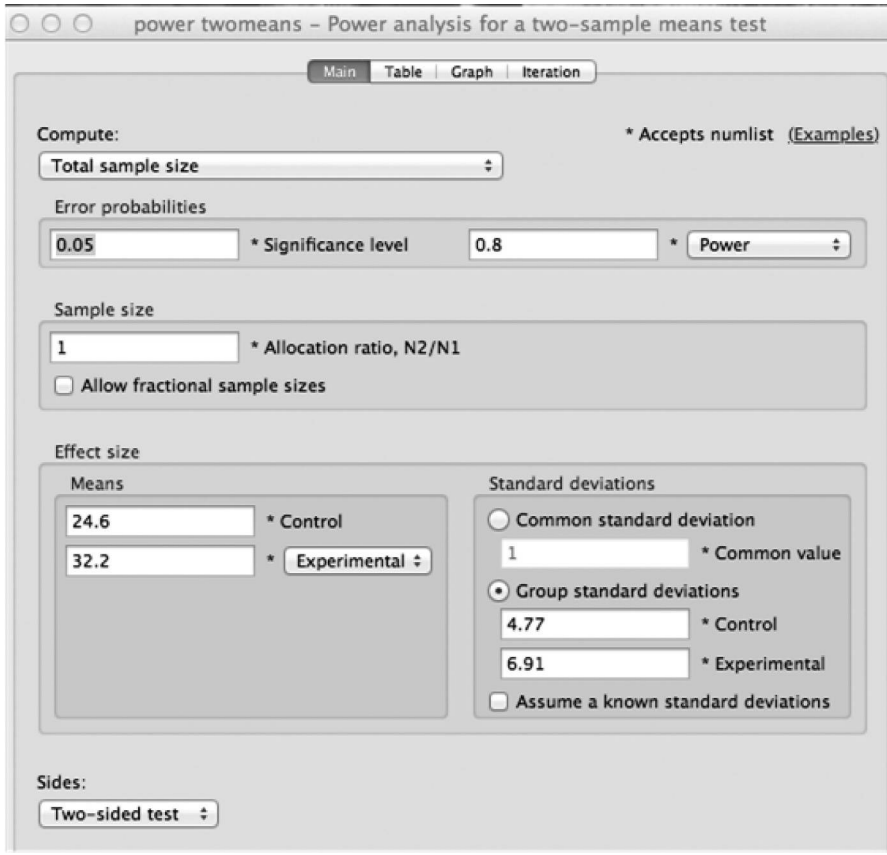

Figure 4.3 Power analyses for a two-sample means test.

After clicking the *Submit* option, the output will be as seen below:

```
. power twomeans 24.6 32.3, sd1(4.77) sd2(6.91)

Performing iteration ...

Estimated sample sizes for a two-sample means test
Satterthwaite's t test assuming unequal variances
Ho: m2 = m1  versus  Ha: m2 != m1

Study parameters:

       alpha =    0.0500
       power =    0.8000
       delta =    7.7000
          m1 =   24.6000
          m2 =   32.3000
         sd1 =    4.7700
         sd2 =    6.9100

Estimated sample sizes:

           N =        22
 N per group =        11
```

To compare the statistical hypotheses with a 5% significance level, an 80% statistical power, and an allocation ratio equal to 1, the minimum sample size needed (for this problem) is 11 per group. If the user is interested in displaying a graph for different options in the statistical power (0.8, 0.85, 0.9), using common standard deviation (5), and different allocation ratios (1, 2, 3), the window should be completed in the manner seen in Figure 4.4.

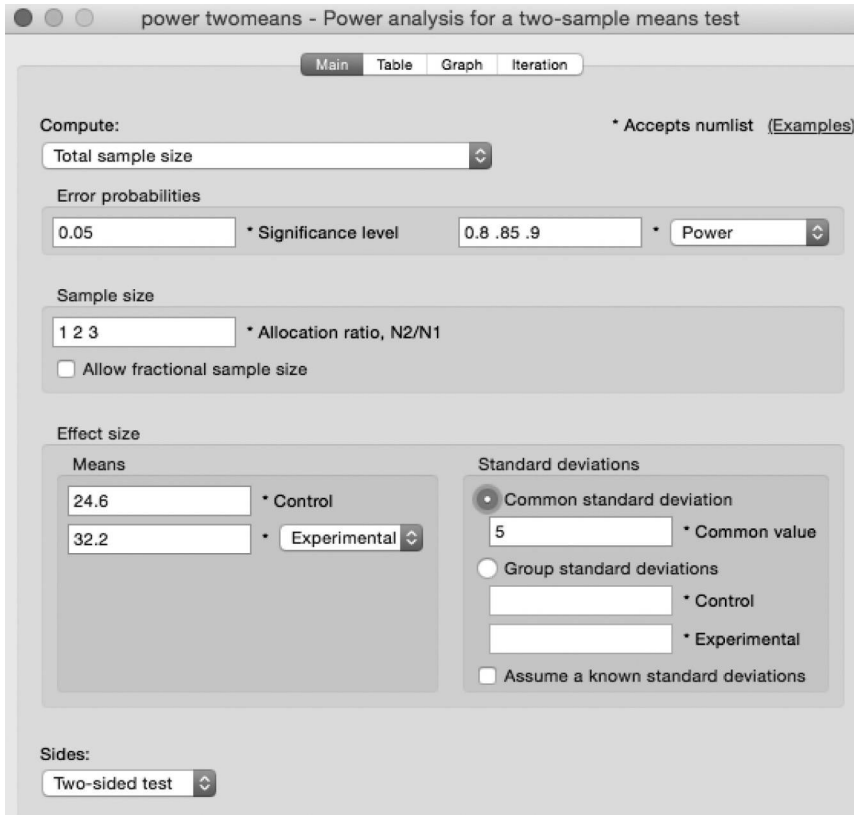

Figure 4.4 Sample size for two-sample means tests with several power levels and allocation ratios.

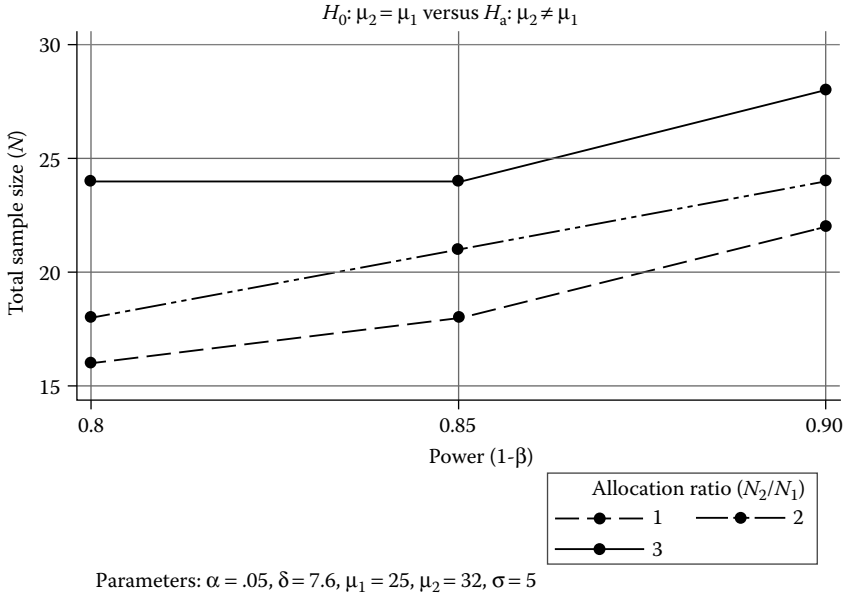

Figure 4.5 Sample size for two-sample means tests using graph option.

When the graph option is used, the output will be as seen in Figure 4.5. The total sample size requested will increase when the statistical power is increased; however, the changes in the sample size will depend on the allocation ratio.