CHAPTER 1

BASIC STATISTICAL TOOLS

This Chapter provides some basic statistical concepts and tools. Pointwise estimation and confidence interval estimation are introduced; conservative estimation follows. Then, an explanation is given on what statistical tests are. The power function of the tests together with the errors of first and second type are defined. The p-value is presented, as an index for evaluating the outcome of the test. Some applications in the context of clinical trials are shown and numerical examples and figures are also provided. The probability of success in a trial (i.e. success probability) is illustrated, including how to estimate it. Superiority tests are adopted first to illustrate the above topics. Then, inequality tests are considered. Finally, there is a brief Section regarding how success probability estimation can be derived for tests of clinical superiority, of non-inferiority and for equality tests.

Success Probability Estimation with Applications to Clinical Trials, First Edition.
By Daniele De Martini Copyright © 2013 John Wiley & Sons, Inc.

1.1 Pointwise estimation

Two populations are in the study, representing tho one treated with a new drug and that treated with a control drug. In order to estimate the unknown parameters of interest of these populations, for example their averages (i.e. means), experimental samples are drawn from the populations. Then, experimental data are used to provide an estimated value, which is often the most likely one, of these parameters: this is *pointwise estimation.*

Often, the effect size of the new drug is the quantity of clinical interest and is expressed as the ratio of the difference between the two means and the common standard deviation of the distributions. So, the effect size can be estimated on the basis of point estimates of means and standard deviation derived from population samples. It is worth noting that the frequency of effect size estimates that are far from the true (and unknown) value of the effect size tends to be smaller as the size of the samples tends to be larger.

Nevertheless, although the point estimate is probably close to the unknown effect size, especially for large samples, it is almost surely different from the true effect size. What is most important is that pointwise estimation has its peculiar *random variation*, since it depends on random samples. Estimates, indeed, can vary from one sample to another and they vary randomly since samples are randomly drawn. Consequently, the variability of pointwise estimates must be taken into account during the estimation process. This is a very basic and essential statistical concept.

Consider two Gaussian distributions representing the distributions of the quantity of interest in the two patient populations being treated with the new drug (population 1) and with the control (population 2), defined by X_1 and X_2, respectively. These distributions have generic means μ' and μ'', respectively, and common standard deviation σ. This σ is a measure of the variability of population data around their means - it is a mean of the distances of data from the respective means.

The *effect size* of the experiment is considered here to be the standardized difference between the means, that is $\delta = (\mu' - \mu'')/\sigma$. The effect size can cover a wide range of values, depending on how the new and the control drugs perform. Nonetheless, there is only one "true" value of δ, namely δ_t and given by the true means μ_1 and μ_2. To investigate on δ_t is the aim of our research.

A sample of size m is randomly drawn from each patient population and the random variables representing the data provided by the patients are denoted by $X_{i,j}$, with $j = 1, \ldots, m$ and $i = 1, 2$. The sample averages $\sum_{j=1}^{m} X_{i,j}/m$, denoted by $\bar{X}_{i,m}$, aim to estimate the distribution means μ_i, $i = 1, 2$, and are *estimators* of the latter. Since estimators are functions of the random samples they are random variables too. In general, estimators are built in order to fall with high probability close to the parameters they are estimating.

Here, σ is assumed to be known. The pointwise estimator of δ_t is, then, based on sample averages: $d_m = (\bar{X}_{1,m} - \bar{X}_{2,m})/\sigma$.

The value that a random variable assumes, i.e. the observed value of the latter, is called *realization* of the random variable; the realizations of estimators are called *estimates*.

Remark 1.1. *Often, distinct notations are adopted to emphasize the difference between a random variable and its realization: uppercase and lowercase letters usually indicate the former and the latter object, respectively. For example, the realizations of the average of the first sample (i.e. the estimates of μ_1) might be denoted by $\bar{x}_{1,m}$. Throughout the book, this distinction will not be made and the reader will deduce from the context of the sentence if either a random variable or a realization of it is being considered.*

EXAMPLE 1.1

The variables X_1 and X_2 of the two populations under study have true mean $\mu_1 = 5$ and $\mu_2 = 1$, respectively, and standard deviation $\sigma = 8$. The shape of their Gaussian distributions is shown in Figure 1.1. The true effect size, therefore, is $\delta_t = (5-1)/8 = 0.5$. A sample of size $m = 85$ is drawn from each population, providing the estimates $\bar{X}_{1,85} = 3.816$ and $\bar{X}_{2,85} = 1.152$ (these are realizations of $\bar{X}_{1,85}$ and $\bar{X}_{1,85}$). It is not surprising that sample mean estimates do not coincide with population ones: this is due to random variation. The effect size estimate is $d_{85} = (3.816 - 1.152)/8 = 0.333$.

Without loss of generality, and in order to handle as few symbols as possible, σ is set equal to 1 throughout Part I of this book, whenever not explicitly claimed, the contrary. Then, $\delta_t = \mu_1 - \mu_2$. Consequently the pointwise estimator of the effect size, which is a function of the estimators of the unknown parameters, is:

$$d_m = \bar{X}_{1,m} - \bar{X}_{2,m} \tag{1.1}$$

The probability distribution of d_m, as a consequence of the Gaussian distribution shapes of the populations, is Gaussian too - this is due to a mathematical theorem. The mean of d_m is δ_t, and its standard deviation is $\sigma_{d_m} = \sigma\sqrt{2/m}$, which is also called the *standard error* of d_m; being $\sigma = 1$, we have $\sigma_{d_m} = \sqrt{2/m}$. This means that d_m is more dense around δ_t, and that the probability that d_m falls far from δ_t decreases, as the sample size m increases. In other words, the precision in estimating δ_t improves when m increases (see Figure 1.2). Formally speaking, d_m is a *consistent* estimator of δ_t. Recall that d_m is a random quantity, depending on random samples, and that its realizations vary from one couple of samples to another.

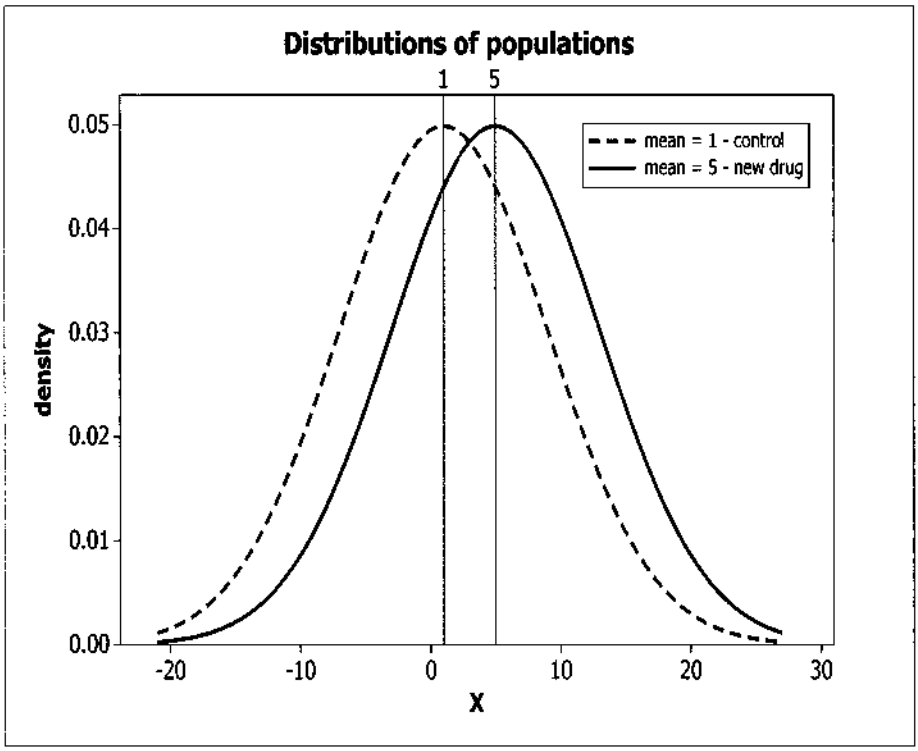

Figure 1.1 *Distributions of the populations in study - Example 1.1.*

1.2 Confidence interval estimation, conservative estimation

The variability of pointwise estimates is here taken into account. Indeed, beside the point estimate, an interval of admissible value for the unknown quantity of interest (in this case, the effect size) should be provided. This interval, too, is based on random samples, and so it may or may not include the effect size.

The key concept is that the probability that the so-called *Confidence Interval* contains the true effect size can be controlled. This probability, namely *confidence level* of the interval (viz. γ), is usually set high (e.g. $90 - 95\%$). Its complement to 1 (i.e. $1 - \gamma$) is the error probability, i.e. the probability that the Confidence Interval does not contain the true effect size.

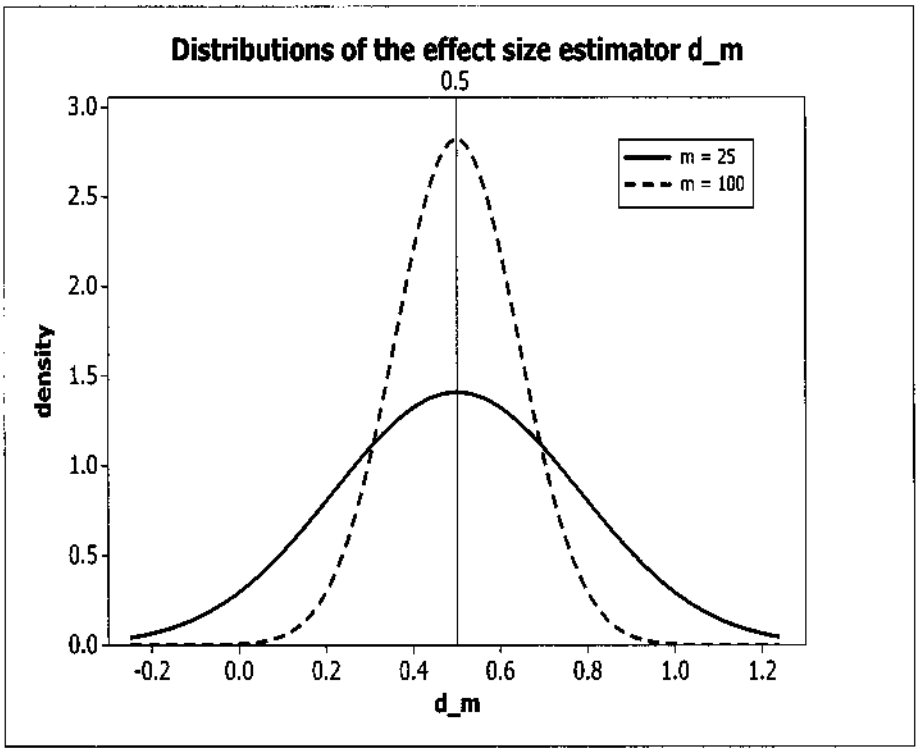

Figure 1.2 *Distributions of the effect size estimator d_m with $\delta_t = 0.5$, and with $m = 25$ and 100. It is of note that the distribution of d_{100} is more dense around $\delta_t = 0.5$ than that of d_{25}.*

Note that as the sample sizes increase the amplitude of the interval decreases, so that confidence interval estimation becomes more precise. On the contrary, the probability of error does not change with the sample sizes.

The lower bound of the confidence interval can be viewed as a conservative estimate of the effect size: it tends to the latter as the sample size increases, remaining below it with a given (high) probability.

In order to derive an interval of plausible values for δ_t (namely Confidence Interval) the standardized version of d_m is introduced, that is $(d_m - \delta_t)/\sqrt{2/m} = \sqrt{m/2}(d_m - \delta_t)$. The latter quantity has a Gaussian distribution with mean 0 and variance 1, namely a standard Gaussian distribution.

A random variable with the latter distribution is represented by the symbol Z. The terms Gaussian distribution and normal distribution will be used as synonyms.

Now, let Φ be the cumulative distribution function of the standard normal, that is the probability that Z falls below a certain value t: $\Phi(t) = P(Z \leq t)$). Moreover, let z_γ be the γ-th percentile of the standard normal (i.e. z_γ is such that $P(Z \leq z_\gamma) = \gamma$). Then, the γ-percentile is $z_\gamma = \Phi^{-1}(\gamma)$.

It follows that the central part of the distribution of Z lies, with probability γ, in the interval $[z_{(1-\gamma)/2}, z_{(1+\gamma)/2}]$. Since the standardized version of d_m is Z-distributed, we obtain:

$$P(z_{(1-\gamma)/2} \leq \sqrt{m/2}(d_m - \delta_t)/ \leq z_{(1+\gamma)/2}) = \gamma$$

Inverting the two inequalities above, and being $z_{1-\gamma} = -z_\gamma$ due to the symmetry of the Gaussian distributions, we finally have:

$$P(d_m - z_{(1+\gamma)/2}\sqrt{2/m} \leq \delta_t \leq d_m + z_{(1+\gamma)/2}\sqrt{2/m}) = \gamma \tag{1.2}$$

In other words, the (random) interval $[d_m - z_{(1+\gamma)/2}\sqrt{2/m}, d_m + z_{(1+\gamma)/2}\sqrt{2/m}]$ contains δ_t with probability γ, and so it is a γ-Confidence Interval for the effect size. This is a two-sided confidence interval, since it is both upward and downward bounded. The confidence level γ is usually set high, e.g. $90\%, 95\%, 99\%$.

EXAMPLE 1.2

Let us continue Example 1.1, where the effect size estimate was $d_{85} = 0.333$. *With a confidence level* $\gamma = 95\%$, $z_{(1+\gamma)/2}$ *becomes* $z_{97.5\%} = 1.96$ *(see Table A.1). The realizations of the bounds of the interval result:* $0.333 - 1.96\sqrt{2/85} = 0.032$ *and* $0.333 + 1.96\sqrt{2/85} = 0.634$. *In this case, the realization of the interval contains* $\delta_t = 0.5$.

Note that the confidence interval is defined by a random quantity (i.e. d_m) and so it varies from one couple of samples to another. γ is the frequency of sampled confidence intervals that contain δ_t, independently on m. The amplitude of the interval is given by the difference between its upper and lower bounds and it results $2z_{(1+\gamma)/2}\sqrt{2/m}$, which decreases as m increases. This means that confidence interval estimation improves its precision as the sample size increases. The confidence level γ does not change when m varies.

Confidence Intervals can also be one-sided. In fact, when a statistical lower bound for the effect size is of interest, the one-sided γ-Confidence Interval is:

$$[d_m - z_\gamma\sqrt{2/m}, +\infty)$$

which provides:

$$P(d_m - z_\gamma\sqrt{2/m} \leq \delta_t) = \gamma \tag{1.3}$$

The lower bound of the interval can be used for *conservative estimation*. Being:

$$d_m^{\gamma} = d_m - z_{\gamma}\sqrt{2/m} \tag{1.4}$$

the latter can be viewed as a conservative estimator of the effect size In other words, d_m^{γ} tends to be much closer to δ_t as the size of the sample m increases, with the condition of falling below δ_t with (high) probability γ (see Figure 1.3). Let us call d_m^{γ} the *γ-lower bound for δ_t*.

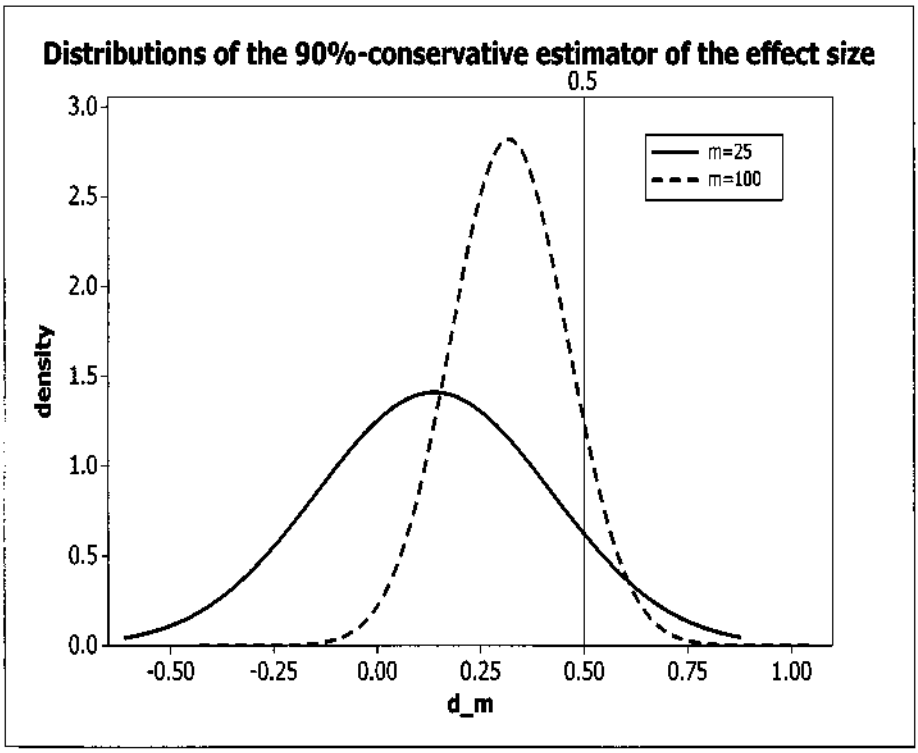

Figure 1.3 *Distributions of the effect size estimator $d_m^{90\%}$ with $\delta_t = 0.5$, and with $m = 25$ and 100. It is of note that $d_{100}^{90\%}$ is more dense around $\delta_t = 0.5$ than $d_{25}^{90\%}$, and that the area under the curves below 0.5 equals 90% (i.e. the amount of conservativeness of $d_m^{90\%}$) for each m.*

1.3 The statistical hypotheses, the statistical test and the type I error for one-tailed tests

The statistical test is the procedure that leads, through the statistical analysis of experimental data, to one of these two outcomes: "it is experimentally proved that the new drug is more effective than the control drug", or "it is not proved that the new drug is more effective". The possibility to prove that the new drug is less effective than the control treatment will be considered later.

In practice, the assumption that the mean of the effect of the new drug is lower than, or at least equal to, that of the control drug is made. In other words, it is assumed that the true effect size is lower than, or equal to, zero, viz. *null hypothesis*. If clinical trial data show, under the latter assumption, an unexpected result, that the patients respond considerably better under the new drug, then the assumption above (against the new drug) is rejected, the complementary assumption (i.e. the true effect size is greater than zero, viz. *alternative hypothesis*) is assumed to be true, and the effectiveness of the new drug is considered to be experimentally proved.

Specifically, a sample of patients is randomly drawn from each population and the respective sample means are computed. Hence, if a high difference is observed in favor of the new drug, so high that this observed event falls within the predefined set of events having globally, under the null hypothesis, a low probability (namely α), then the statement "it is experimentally proved that the new drug is more effective than the control drug" is the outcome of the test - this is called a *significant* outcome. Otherwise, the outcome is "it is not proved that the new drug is more effective", and nothing is proved.

In case the drug is considered to be effective where in actual fact it is not, an error is made: this is the type I error of the test. The probability that this error occurs is at most α. This α is set before recruiting patients, and so before analyzing data, it is often equal to 5% or to 2.5%.

The assumptions on the means are the *statistical hypotheses*, which formally result in:

$$H_0 : \mu_1 \leq \mu_2 \qquad \text{and} \qquad H_1 : \mu_1 > \mu_2$$

namely the *null* and the *alternative* hypotheses, respectively. The latter is the one-sided alternative of *superiority*. Further hypotheses (such as $H_1 : \mu_1 > \mu_2 + \delta_0$) will be considered later (see Section 1.8) . Note that the hypotheses can also be viewed in terms of effect size:

$$H_0 : \delta_t \leq 0 \qquad \text{and} \qquad H_1 : \delta_t > 0$$

The test statistic, namely T_m, has to reflect the behavior of the phenomenon of interest and is, therefore, a function of the samples of size m drawn from the two

populations. Here, the difference between the means is under study, and so T_m is built on the basis of sample averages, and in particular of d_m. Moreover, d_m is divided by its standard deviation σ_{d_m}, so that T_m has a unitary standard deviation:

$$T_m = (\bar{X}_{1,m} - \bar{X}_{2,m})/\sigma_{d_m} = d_m/\sqrt{2/m} = \sqrt{m/2}d_m \tag{1.5}$$

T_m has a Gaussian distribution with mean $\delta_t\sqrt{m/2}$ and variance 1 - see Figure 1.4. (When the null hypothesis is $H_0 : \mu_1 \leq \mu_2 + \delta_0$, i.e. $H_0 : \delta_t \leq \delta_0$, then $d_m - \delta_0$ is considered so that $T_m = \sqrt{m/2}(d_m - \delta_0)$ - see Section 1.8). Under the null hypothesis, T_m has mean at most 0.

It follows that large values of T_m could lead one to consider it true that $\delta_t > 0$ (or that $\delta_t > \delta_0$ when $H_1 : \mu_1 > \mu_2 + \delta_0$ is under testing) and induce H_0 rejection.

Then, the probability, namely α, permitted to the type I error, i.e. rejecting H_0 when it is true, is set. So, the rejection region, i.e. the set of values that if assumed by T_m induce H_0 rejection, is that on the right tail of the standard Gaussian distribution whose total probability is α, that is $(z_{1-\alpha}, +\infty)$. In other words, the null hypothesis is rejected when T_m results greater than $z_{1-\alpha}$, which is named the *critical value of the test*. Note that the probability to reject H_0 when it is true is actually, at most, α: $P_{\delta_t=0}(T_m > z_{1-\alpha}) = \alpha$.

The statistical test ψ_α, therefore, is:

$$\psi_\alpha(T_m) = \begin{cases} 1 & \text{if} \quad T_m > z_{1-\alpha} \\ 0 & \text{if} \quad T_m \leq z_{1-\alpha} \end{cases} \tag{1.6}$$

where "1" stands for "H_1 is experimentally proved" and "0" for "nothing is proved".

This ψ_α is also called Z-test. Here, the rejection region is defined on one tail of the distribution of the test statistic under the null, and so ψ_α is named *one-tailed* test.

1.4 The power function and the type II error

The power function of the test reports the probability to reject the null hypothesis, that is the probability to prove that the new drug is more effective than the control drug. This probability depends on the type I error, on the sample size and on the generic effect size δ.

When the null hypothesis is true, the power function, i.e. the probability to reject the null, provides the probability of an error: this is the type I error, whose probability assumes at most the value of α.

When the alternative hypothesis is true, that is when the new drug is effective, there is actually the possibility of not rejecting the null hypothesis. If this hap-

pens an error is made: this is the type II error, whose probability is named β and is given by 1 minus the power function.

Table 1.1 Errors and Right Decisions (RD), with their probabilities, in hypotheses testing.

	Hypotheses	
Decisions	H_0 true	H_1 true
Accept H_0	RD	type II error
	$1-\alpha$	β
Reject H_0	type I error	RD
	α	$\pi(\alpha, m, \delta)$

Consider now a generic value δ of the effect size, not the fixed and unique true effect size δ_t. Then, the generic test statistic T_m is normally distributed with mean $\delta\sqrt{m/2}$ (not $\delta_t\sqrt{m/2}$) and unitary variance. From (1.6) the probability to reject H_0 is $P_\delta(\psi_\alpha(T_m) = 1)$, i.e. $P_\delta(T_m > z_{1-\alpha})$. The latter quantity depends on α, m and δ and it is called the power function: $\pi(\alpha, m, \delta)$.

From the knowledge of the distribution of T_m we have:

$$\pi(\alpha, m, \delta) = P_\delta(T_m > z_{1-\alpha}) = P_\delta(T_m - \delta\sqrt{m/2} > z_{1-\alpha} - \delta\sqrt{m/2})$$

$$= P(Z > z_{1-\alpha} - \delta\sqrt{m/2}) = \Phi(\delta\sqrt{m/2} - z_{1-\alpha}) \tag{1.7}$$

Note that under H_0 (i.e. with values of $\delta \leq 0$) the power function is lower than, or equal to, the type I error:

$$\pi(\alpha, m, \delta) \leq \pi(\alpha, m, 0) = \Phi(-z_{1-\alpha}) = \Phi(z_\alpha) = \alpha$$

Given α and m, the power function (1.7) increases as δ increases, meaning that the probability to prove that $\delta > 0$ grows as the effect size becomes higher. The power function also increases, given α and $\delta > 0$, as the available information grows, that is as m increases. To complete, given m and δ, the power function is higher for larger αs.

Under the alternative hypothesis (i.e. with $\delta > 0$) a possible error could be to fail to reject H_0 - this is often called "to accept H_0". This is the type II error, whose probability is $\beta = 1 - \pi(\alpha, m, \delta)$. Under H_1, the power function (1.7) is higher than the type I error: $\pi(\alpha, m, \delta) > \alpha$ if $\delta > 0$. In Table 1.1 the possible decisions are summarized, together with their respective errors.

Remark 1.2. *It is noteworthy that to accept H_0 signifies that there is not enough information to reject it, and therefore to prove H_1. To accept H_0 does not mean,*

therefore, that the null hypothesis is proved: the null hypothesis is never proved. On the contrary, when H_0 is rejected H_1 is experimentally proved, unless the type I error is made, this probability is α.

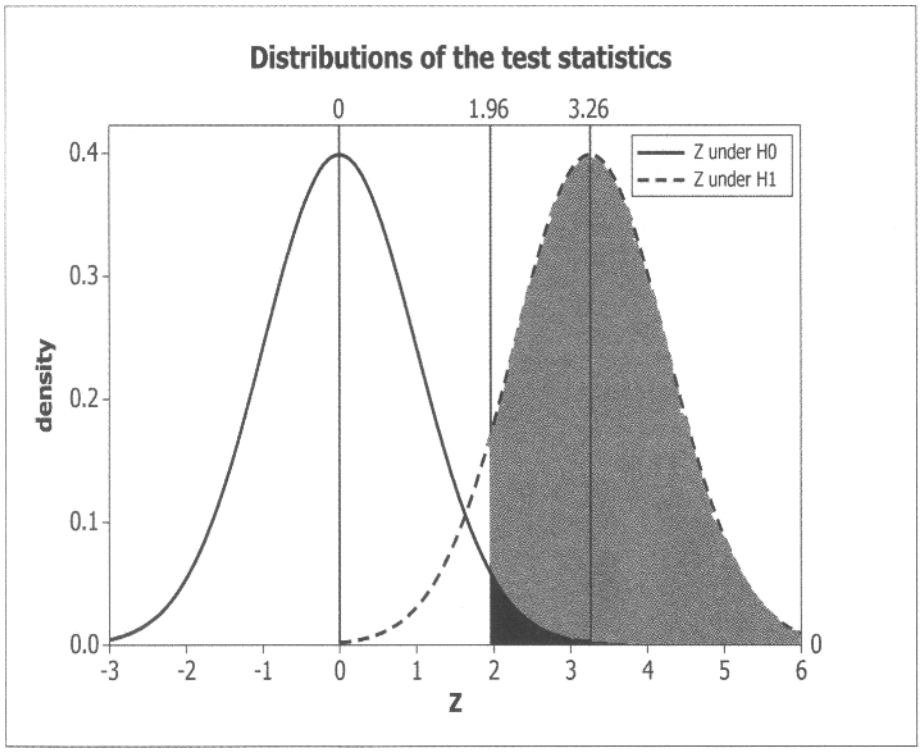

Figure 1.4 *Distributions of the test statistic under the null and under the alternative of Example 1.3 with $m = 85$. Note that the mean of T_{85} is $\delta_t\sqrt{m/2} = 0.5\sqrt{85/2} = 3.26$. The probabilities are represented by the area under the curves: the black area represents the type I error probability $\alpha = 2.5\%$ under the null; the gray-dashed area (which includes the black one) represents the probability, under the alternative, to fall in the rejection region with $m = 85$ and $\delta = 0.5$, that is $\approx 90\%$.*

EXAMPLE 1.3

When $\alpha = 2.5\%$ the critical value is $z_{97.5\%} = 1.96$ (use Table A.1 to obtain probabilities and deviates for the standard normal distribution, and so even for power computation). When $\delta = 0.5$ the power function with $m = 17$, 40 and 85 data per group provides 30.78%, 60.88% and 90.31%, respectively, according

to (1.7). This means that with $m = 85$ available data, there is approximately a 90% probability to prove that $\mu_1 > \mu_2$, allowing for a type I error of 2.5%. In Figure 1.4 this 90.31% power can be viewed as a probability (the label Z is adopted for values of T_m, as in the next Figures). The power functions are reported in Figure 1.5.

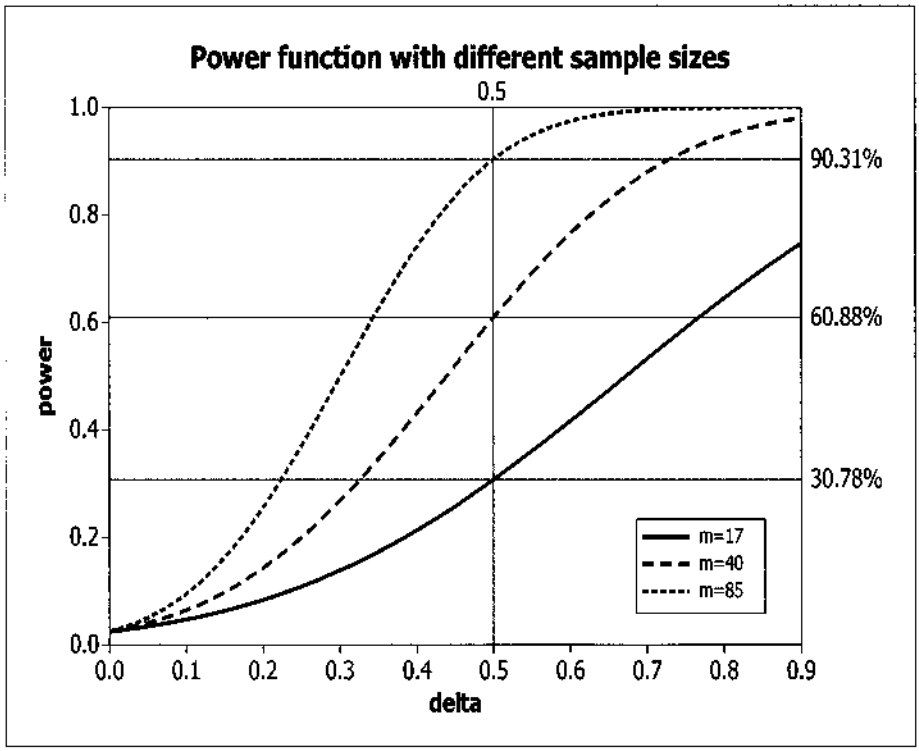

Figure 1.5 *Power functions for the Z-test with $\alpha = 2.5\%$, with $m = 17$, 40 and 85 (i.e. for scenarios in Example 1.3). The values assumed by the power functions with $\delta = 0.5$ are reported.*

1.5 The p-value

The p-value is the statistical index traditionally used for evaluating the outcome of statistical tests: given the data, the p-value is the maximum type I error for which a test statistic is not significant.

Another possible way to introduce the p-value is that it represents the probability, computed under the null hypothesis, of finding, in a new experiment completely analogous to that just performed but independent of it, a result even farther from the null hypothesis than the one just observed.

In other words, the p-value answers this question: if the null hypothesis is really true (in this case, if $\delta_t \leq 0$), what is the probability that random sampling would lead to a result better than the one observed in favor of the new drug (e.g. a difference between sample means larger than the one observed)?

The p-value is an index of the strength adopted to reject the null hypothesis: the lower than α the p-value is, the higher the strength.

Moreover, the p-value can also be used to compute the outcome of the test: the null hypothesis is, indeed, rejected only when the p-value results lower than the prefixed type I error probability α.

Let us define the p-value formally:

$$\text{p-value} \quad = \quad \max\{\, \alpha' \text{ s.t. } \psi_{\alpha'}(T_m) = 0 \,\} \tag{1.8}$$

From the test statistic (1.6) and the definition (1.8), the p-value is such that $T_m = z_{1-\text{p-value}}$. Then, recalling the definition of $z_x = \Phi^{-1}(x)$ and applying the monotone function Φ to both members of the latter equality we obtain:

$$\Phi(T_m) = \Phi(\Phi^{-1}(1 - \text{p-value})) = 1 - \text{p-value}$$

giving, finally:

$$\text{p-value} = 1 - \Phi(T_m) \tag{1.9}$$

Alternatively, the p-value can be defined by considering a new experiment, identical to the one just performed, which gives a new test statistic T_m^* independent of T_m. Hence, the p-value, being the probability under H_0 that the random statistic T_m^* would be larger than the observed T_m, can be defined as follows:

$$\text{p-value} = P_{\delta_t=0}(T_m^* > T_m \,|T_m) = 1 - \Phi(T_m)$$

where the last equality follows from the condition that the test new statistic T_m^* is Gaussian distributed, as in this introductive testing situation. (The notation $P(A\,|B)$ means "the probability of A once event B has been observed".)

EXAMPLE 1.4

Consider a one-tailed Z-test with type I error probability $\alpha = 2.5\%$*, so that the critical value is* $z_{97.5\%} = 1.96$*. A total of 170 patients are recruited, and they are randomized into two groups of size* 85*. The latter represent the two samples from the two populations under the new drug and under the control treatment, providing sample averages of* $\bar{X}_{1,85} = 0.477$ *and* $\bar{X}_{2,85} = 0.144$*, respectively. From 1.5, the observed test statistic is:* $T_{85} = \sqrt{85/2}(0.477 - 0.144) = 2.17$*. Since* T_{85} *is greater than* 1.96*, the test outcome is significant:* $\psi_{2.5\%}(2.17) = 1$*. According to (1.9) the p-value is* $1 - \Phi^{-1}(2.17) = 1.5\%$*, which seems quite a bit lower than* α *(use Table A.1 also to compute p-values). Consequently, this outcome appears quite a reassuring one. The p-value is reported as a black area in Figure 2.1.*

It is interesting to note that the p-value in (1.9) is lower than α only when T_m is over the critical value of the test, that is only when the outcome of the test is significant. Indeed, we have that:

$$\text{p-value} = 1 - \Phi(T_m) < \alpha \quad \text{iff} \quad \Phi(T_m) > 1 - \alpha \quad \text{iff} \quad T_m > \Phi^{-1}(1-\alpha) = z_{1-\alpha},$$

where iff means *If and Only if*. So, the p-value can also be employed to define the statistical test (1.6) itself:

$$\psi_\alpha(T_m) = \begin{cases} 1 & \text{if} \quad \text{p-value} < \alpha \\ 0 & \text{if} \quad \text{p-value} \geq \alpha \end{cases} \tag{1.10}$$

Equation (1.10) may be referred to as *p-value testing*.

As a consequence, the power function (1.7) (i.e. the probability of finding a significant outcome) can be viewed as the probability of finding a p-value lower than the type I error probability α:

$$\pi(\alpha, m, \delta) = P_\delta(T_m > z_{1-\alpha}) = P_\delta(\text{p-value} < \alpha)$$

Since the p-value depends on data, it is a random variable with a certain distribution. In Figure 1.6 the distributions of the p-value in scenarios of Example 1.3 are reported. Note that p-value distributions under the alternatives are much denser to the left for high power values, i.e. for high values of m. Moreover, when $\delta = 0$ (i.e. for the highest of the possible values under H_0) the p-value is uniformly distributed in $(0, 1)$, that is $P_{\delta_t=0}(\text{ p-value} \leq t) = t$, with $t \in (0, 1)$. In other words, in this case the density of the p-value is uniformly equal to 1 in the domain $(0, 1)$.

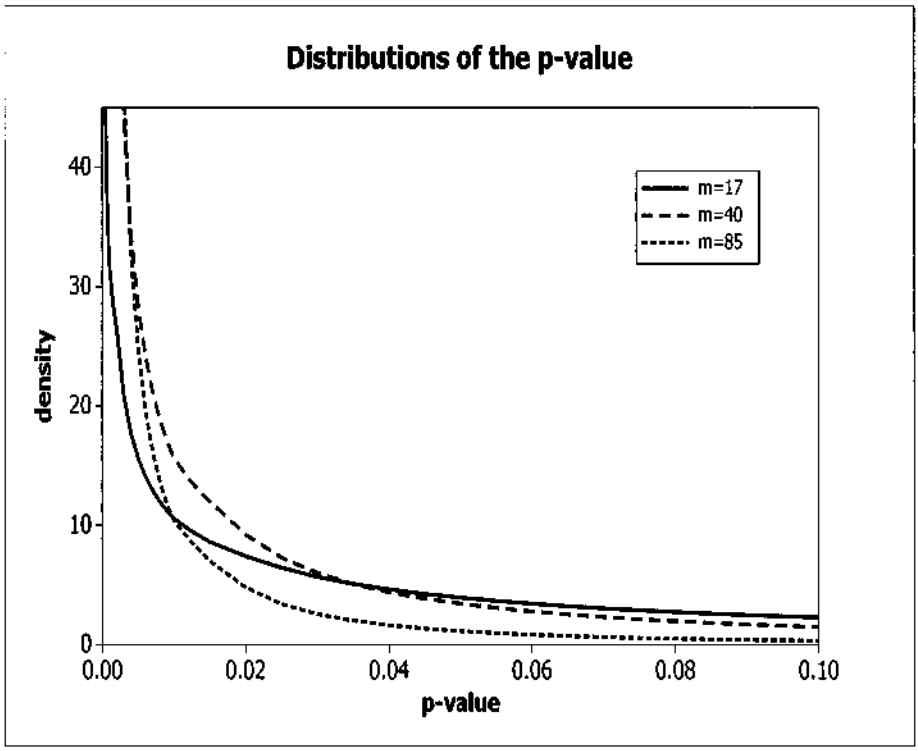

Figure 1.6 *Distributions of the p-value for the Z-test with $\alpha = 2.5\%$, under the alternative hypothesis with $\delta = 0.5$, $m = 17$, 40 and 85 (i.e. for scenarios in Example 1.3).*

1.6 The success probability and its estimation

The true probability of proving that the new drug is more effective than the control drug when it actually is, that is, the true probability of rejecting the null hypothesis under the alternative, is called the success probability, SP.

SP depends on the "true" value of the effect size, which is actually unknown. Of course, SP is related to the power function: in particular, SP *is* the power function evaluated at δ_t, when $\delta_t > 0$.

Since δ_t is unknown, the same holds true for SP. Nevertheless, it would be very useful to acquire information on SP, that can in fact be estimated.

SP estimation can be applied in solving the problems presented in Section I.4. In particular, it can be applied for estimating the sample size of a phase III trial on

the basis of phase II data, and for estimating, once a trial has been performed, the probability of finding a statistical significance in a new trial whose settings are identical to those of the one just performed (namely *reproducibility probability*), in order to evaluate the *stability* of the results of the trial.

The SP is the power function computed at δ_t, that is:

$$SP = P_{\delta_t}(T_m > z_{1-\alpha}) = \pi(\alpha, m, \delta_t) = \Phi(\delta_t \sqrt{m/2} - z_{1-\alpha}) \qquad (1.11)$$

In other words, SP is the true power of the test.

Now, let us assume that two samples of size n (not necessarily equal to m) are available from the same two populations, one from each. (Actually, in Chapter 3, Section 3.2, it will be explained that the condition of sampling from the same populations can be relaxed). These samples can be viewed as pilot ones and are independent from those of size m used to define the Z-test in (1.6). On the basis of these data an estimator of δ_t can be computed: let us call it $d_n^\bullet$, where $\bullet$ indicates that several statistical approaches can be adopted for estimating δ_t (the simplest one is $d_n^\bullet = d_n = \bar{X}_{1,n} - \bar{X}_{2,n}$). Hence, an estimator of the SP is obtained by putting the estimator of δ_t in the power function definition (1.7):

$$\hat{SP} = P_{d_n^\bullet}(T_m > z_{1-\alpha} \mid d_n^\bullet) = \pi(\alpha, m, d_n^\bullet) = \Phi(d_n^\bullet \sqrt{m/2} - z_{1-\alpha}) \qquad (1.12)$$

This statistical practice of substituting an estimator of a parameter into a certain function of the parameter itself is called the "plug-in principle".

EXAMPLE 1.5

Consider the same sample data of Example 1.4. The observed effect size is $d_{85} = 0.477 - 0.144 = 0.333$ *(as in Example 1.1) and it is considered the estimate of* δ_t. *Assume that it is of interest to estimate the SP of a one-tailed Z-test with* $\alpha = 2.5\%$ *enrolling* $m = 120$ *data per group (here* 85 *plays the role of* n*). From (1.12) the estimate of the latter quantity resulted:* $\hat{SP} = P_{0.333}(T_{120} > z_{97.5\%}) = \Phi(\sqrt{120/2}\,0.333 - 1.96) = 73.22\%$. *According to Example 1.1, where* $\delta_t = 0.5$, $SP = P_{0.5}(T_{120} > z_{97.5\%}) = 97.21\%$. *The power function with* $m = 120$ *is reported in Figure 1.7, together with* SP *and* $\hat{SP}$.

SP estimation can be applied to solve the problems presented in Situations I and II of Section I.4.

Actually, Situation I considered a particular case of SP estimation, where the size of the test sample (i.e. m) is equal to that of a sample actually available, n. The SP estimation assumes, therefore, the meaning of *reproducibility probability* estimation. This topic will be developed in Chapter 2.

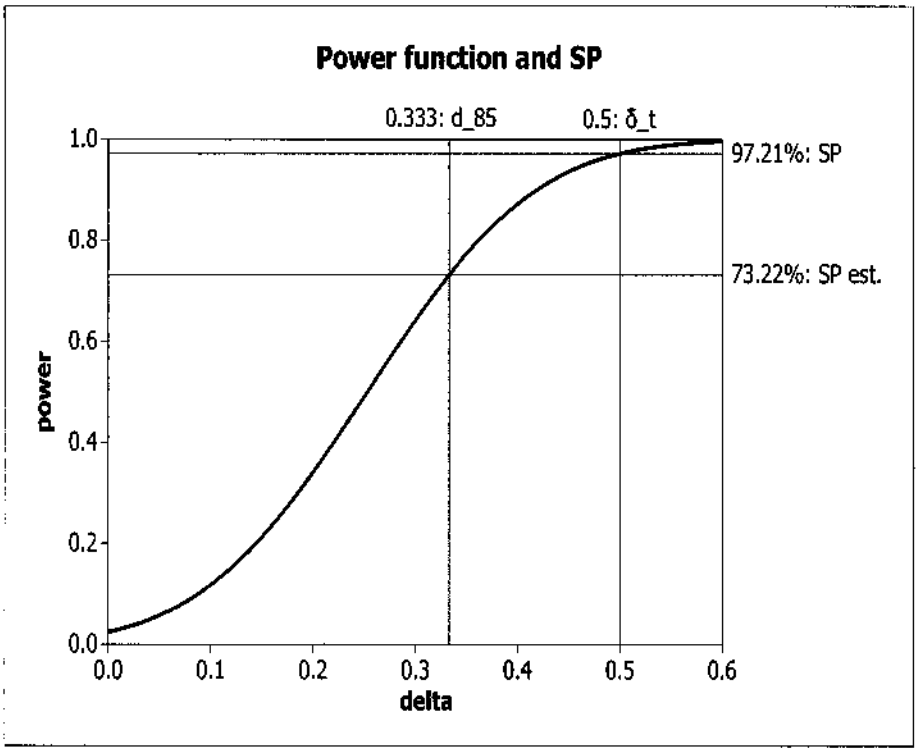

Figure 1.7 *Power function, SP and estimated SP for the Z-test with* $\alpha = 2.5\%$*, with* $m = 120$*, illustrating Example 1.5.*

Situation II concerned *sample size estimation*: in practice the sample size for a phase III can be computed on the basis of SP estimates given by phase II data. Different approaches to SP estimation are available, and they will be developed in Chapters 3 and 4.

1.7 Basic statistical tools for two-tailed tests

In this Section the basic statistical tools and SP definition will be extended to the two-tailed setting.

1.7.1 Two-sided hypotheses and two-tailed statistical test

When the statistical test procedure should provide one of these two statements: "it is experimentally proved that one of the drugs is more effective than the other one", or "it is not proved that neither drug is more effective", two-sided hypotheses are to be adopted.

The possibility that the new drug is less effective than the control treatment is included here. In other words, the assumption of *inequality* is being proved, instead of that of superiority.

In practice, the assumption that there is no difference between the means (i.e. the true effect size is zero, viz. *null hypothesis*) is made. If clinical trial data show a "strange" result, that is, a considerably better patient response to any of the drugs, then the assumption of no difference is rejected and the effectiveness of the drug which performed better is considered as experimentally proved.

Specifically, a sample of patients is randomly drawn from each population and the respective sample means are computed. If a high difference is observed, high enough that this observed event falls within the predefined set of events having globally, under the null hypothesis, a low probability α, then the first statement is the outcome of the test. Otherwise, the outcome is the second statement: nothing is proved.

The concept of type I error is analogous to that of Section 1.3: in case one of the drugs is considered to be more effective when actually it is not, an error is made, namely type I error of the test.

When two-sided alternatives are considered the statistical hypotheses under testing are:

$$H_0 : \mu_1 = \mu_2 \qquad \text{and} \qquad H_1 : \mu_1 \neq \mu_2$$

(one can also state $H_0 : \delta_t = 0$ and $H_1 : \delta_t \neq 0$). The latter H_1 is the two-sided alternative of *inequality*. The test statistic T_m is the same as the one defined in (1.5). Low values of T_m lead to infer that $\delta_t < 0$, and high values that $\delta_t > 0$, so that in both cases H_0 rejection is a probable outcome.

Under the null, T_m has a standard normal distribution. The probability allowed to the type I error, i.e. rejecting H_0 when it is true, is still α. So, the rejection region, i.e. the set of values that if assumed by T_m induce H_0 rejection, should consider both tails of the standard Gaussian distribution and α is, therefore, shared between the two tails, so that each part of the rejection region has $\alpha/2$ probability. Consequently, the null hypothesis is rejected when T_m results lower than $z_{\alpha/2}$ or greater than $z_{1-\alpha/2}$. The rejection region is, therefore, $(-\infty, z_{\alpha/2}) \cup (z_{1-\alpha/2}, +\infty)$. Note that the type I probability remains α: $P_{\delta_t=0}(T_m < z_{\alpha/2} \text{ or } T_m > z_{1-\alpha/2}) = P_{\delta_t=0}(T_m < z_{\alpha/2}) + P_{\delta_t=0}(T_m > z_{1-\alpha/2}) = \alpha/2 + \alpha/2 = \alpha$.

The statistical test ψ_α, therefore, is:

$$\psi_\alpha(T_m) = \begin{cases} 1 & \text{if} \quad T_m < z_{\alpha/2} \text{ or } T_m > z_{1-\alpha/2} \\ 0 & \text{if} \quad z_{\alpha/2} \leq T_m \leq z_{1-\alpha/2} \end{cases} \tag{1.13}$$

This test, where the rejection region is the union of two regions, one on each tail of the distribution of the test statistic under the null, is named *two-tailed* test.

1.7.2 Two-tailed power function, type II and III errors and SP

The power function reports the probability to reject the null hypothesis, that is the probability to prove that any of the drugs is effective. In this two-tailed context too, this probability depends on the type I error, on the sample size and on the generic effect size δ.

When the null hypothesis is true, the power function (i.e. the probability to reject H_0) coincides with the probability of the type I error, i.e. α.

When the alternative hypothesis is true, that is, when any of the drugs is more effective and null hypothesis is not rejected, an error is made: this is the type II error, whose probability (i.e. β) is 1 minus the power function (like in the one-tailed setting).

In case the alternative hypothesis is true and the null hypothesis is rejected but the worst drug is actually considered to be the best one, an error is committed: this is the type III error. The probability of the type III error is, in practice, very small and it can, therefore, be ignored.

In two-sided hypotheses testing, the SP is the probability to prove that any of the drugs is effective when one of the two is actually so (i.e. when $\delta_t \neq 0$), *avoiding the type III error*. This probability depends on the "true" value of the effect size δ_t, which is actually unknown. Even this two-tailed SP can be estimated.

In the two-tailed setting there are two possible significant outcomes (one for each tail of the null distribution) and so the power function of the test is the sum of the two following quantities:

$$\pi_2(\alpha, m, \delta) = P_\delta(T_m < z_{\alpha/2}) + P_\delta(T_m > z_{1-\alpha/2}) \tag{1.14}$$

We call these two summands $\pi_L(\alpha/2, m, \delta)$ and $\pi_R(\alpha/2, m, \delta)$ respectively, to indicate the probability to fall on the Left tail (i.e. below $z_{\alpha/2}$) and on the Right tail (over $z_{1-\alpha/2}$).

When $\delta \neq 0$ and the test fails to reject H_0 the type II error is made, whose probability is $\beta = 1 - \pi_2(\alpha, m, \delta)$. As for one-tailed tests, under the alternative the power function (1.14) is higher than the type I error: $\pi_2(\alpha, m, \delta) > \alpha$ if $\delta \neq 0$.

The type III error (Harter, 1957) consists in rejecting the null hypothesis and in deciding that the worst drug is the best one. Hence, under H_1, one of the two summands of the power function above (viz. $\pi_L(\alpha/2, m, \delta)$, $\pi_R(\alpha/2, m, \delta)$) is the probability of the type III error, and the other one can be viewed as the "good power function". Often, the type III error probability is very small and it always is lower than $\alpha/2$.

EXAMPLE 1.6

The values assumed by the type III error probability are often very small. For example, when $\alpha = 5\%$, *if* $\delta = 0.5 > 0$ *and* $m = 17$, *the type III error probability is* $\pi_L(2.5\%, 17, 0.5) = P_{0.5}(T_{17} < z_{2.5\%}) = 0.0316\%$. *If* δ *decreases to* 0.2 *then the probability of the type III error increases to* 0.55%. *With* $m = 40$ *these two latter probabilities are* 0.0014% *and* 0.22%, *respectively, and they become even lower when* m *grows.*

Now, let us define the SP on the basis of (1.14) and in accordance with its one-tailed definition (1.11). Note that the "good power function" is one of the two summands of (1.14), but which of the two is unknown. So, the SP for the two-tailed setting is the "good power function" evaluated at δ_t under the alternative hypothesis:

$$SP_2 = \begin{cases} P_{\delta_t}(T_m < z_{\alpha/2}) & \text{if} \quad \delta_t < 0 \\ P_{\delta_t}(T_m > z_{1-\alpha/2}) & \text{if} \quad \delta_t > 0 \end{cases} \tag{1.15}$$

SP estimation for two-tailed tests will be developed later (see Sections 2.9 and 3.10).

1.7.3 Two-tailed p-value

The concept of the p-value has not changed: it remains, therefore, the maximum type I error for which a test statistic is not significant.

Here, the p-value answers this question: if the null hypothesis is really true (in this case, if $\delta_t = 0$), what is the probability that random sampling in a new experiment completely analogous to that just performed would lead to a difference between sample means larger than that observed, *in favor of each of the drugs?*

Even in this case the null hypothesis is rejected only when the p-value is lower than α.

The formal definition of the p-value in (1.8) is still valid, and through (1.13) gives:

$$\text{p-value} = 2(1 - \Phi(|T_m|)) \tag{1.16}$$

In analogy with the one-tailed setting, the p-value can also be defined by considering the probability, under the null hypothesis, that a test statistic T_m^* given by a new experiment identical to the one just performed would be farther from 0, potentially in both directions/tails, than the observed T_m. This reflects in:

$$\text{p-value} = 2P_{\delta_t=0}(T_m^* > |T_m| \, | \, T_m) = 2(1 - \Phi(|T_m|))$$

In practice, the p-value is two times the probability mass, computed under the null, of the tail delimited by the observed test statistic. In other words, if, for example, $T_m > 0$ then the p-value is $2P_{\delta_t=0}(T_m^* > T_m \, | \, T_m) = 2(1 - \Phi(T_m))$.

Finally, the p-value based definition of the statistical test (1.10), i.e. p-value testing, remains valid.

1.8 Other statistical hypotheses and tests

In some circumstances, the statement "the new drug is more effective than the control" is related to a clinical minimum threshold of improvement. This concept introduces statistical tests of *clinical superiority*.

In other occasions different assumptions should be proved, such as "the new drug is as effective as a standard therapy". When this statement is related to a clinical maximum threshold of worsening of the new drug with respect to the standard therapy, this approach introduces statistical tests of *clinical non-inferiority*.

Further, *equivalence* statistical tests consider the simultaneous experimental proof of clinical non-superiority and clinical non-inferiority.

For all these tests, statistical analysis and SP estimation can be performed in analogy with those shown in previous Sections of this Chapter.

Let us consider a threshold of minimum clinical improvement $\delta_0 > 0$. It is interesting to prove the one-sided alternative of *superiority* $H_1 : \mu_1 > \mu_2 + \delta_0$, where the null hypothesis is $H_0 : \mu_1 \leq \mu_2 + \delta_0$. For these hypotheses the test statistic (1.5) becomes:

$$T_m = \sqrt{m/2}(\bar{X}_{1,m} - \bar{X}_{2,m} - \delta_0)$$

and the test of clinical superiority is defined as the test in (1.6).

Now, let $\delta_0 > 0$ be the threshold of maximum clinical worsening. Thus, the one-sided alternative of *non-inferiority* to be proved is $H_1 : \mu_1 > \mu_2 - \delta_0$, being $H_0 : \mu_1 \leq \mu_2 - \delta_0$. The test statistic, therefore, is:

$$T_m = \sqrt{m/2}(\bar{X}_{1,m} - \bar{X}_{2,m} + \delta_0)$$

and the test of clinical non-inferiority is once again equal to (1.6).

When $\delta_0 > 0$ represents the threshold of *clinical equivalence*, the alternative hypothesis of non-inferiority *and* non-superiority to be proved is $H_1 : |\mu_1 - \mu_2| < \delta_0$, that is $H_1 : \mu_2 - \delta_0 < \mu_1 < \mu_2 + \delta_0$. Complementarily, the null hypothesis of no equivalence is $H_0 : \mu_1 \leq \mu_2 - \delta_0 \cup \mu_1 \geq \mu_2 + \delta_0$. In this context, when the set representing H_0 is not convex, the statistical test is significant when $\sqrt{m/2}(\bar{X}_{1,m} - \bar{X}_{2,m} + \delta_0) > z_{1-\alpha}$ and $\sqrt{m/2}(\bar{X}_{1,m} - \bar{X}_{2,m} - \delta_0) < -z_{1-\alpha}$.

For clinical superiority and clinical non-inferiority tests, the SP is defined as in (1.11). In particular, for superiority tests the δ_t in (1.11) should be replaced by $\delta_t - \delta_0$, and for non-inferiority tests by $\delta_t + \delta_0$. The SP for equivalence tests can be obtained analogously, but this topic is not developed here.