

# STATISTICS FOR CLINICIANS

---

## Power

Barry K. Moser, PhD  
Associate Director  
CALGB Statistical Center  
Duke University Medical Center  
Durham, NC

### H&O How do you define power?

**BM** Power is used in tests of statistical hypotheses. The power of a test procedure is the probability of observing a test statistic that is extreme enough to reject the null hypothesis of an  $\alpha$  level test, under the assumption that the alternative hypothesis is true. In other words, the power of the test is the probability of correctly deciding to reject the null hypothesis when the alternative hypothesis is true. This definition has several qualities. First, the power is the probability of rejecting the null hypothesis of an  $\alpha$  level test—ie, the power is a function of the  $\alpha$  level chosen for the test. Second, the power is calculated from the distribution of the test statistic under the alternative hypothesis.

### H&O How does one construct an $\alpha$ level test?

**BM** The null hypothesis usually states that the factor being tested will have a specified influence on the outcome. For example, in a study trying to determine the effect of a certain cancer treatment, the null hypothesis associated with a one-sided test states that the average outcome of a population of patients has a value less than or equal to some quantity. The  $\alpha$  level is the probability of rejecting the null hypothesis when the null hypothesis is true. In this example, small, average population outcomes are consistent with the null hypothesis, and thus the null hypothesis is rejected when the average sample outcome (and therefore the test statistic) is large. The  $\alpha$  level is therefore calculated as the probability that the test statistic has a large value when the null hypothesis is true. For this one-sided test, the region of large statistic values where the null hypothesis is rejected is called the  $\alpha$  level rejection region. The  $\alpha$  level rejection region of the test is

calculated from the distribution of the test statistic under the null hypothesis for a specified sample size. That is, if the observed test statistic lies in the rejection region then the null hypothesis is rejected. If the test statistic does not lie in the rejection region then the null hypothesis is not rejected. Note that an  $\alpha$  level test for a certain sample size can be constructed without any consideration of the power.

### H&O Once the rejection region is determined for a specific $\alpha$ level test, how is the power calculated?

**BM** The power is the probability that the observed test statistic lies in the rejection region when the alternative hypothesis is true. That is, the power of the test is the probability of correctly rejecting the null hypothesis. For example, consider the case where it is of interest to test hypotheses regarding the proportion of patients who achieved complete remission after the induction phase of a chemotherapy treatment on adult acute myeloid leukemia (AML) patients. Formally, let  $p$  represent the proportion of the population of adult AML patients who achieve a complete remission from the treatment. In general, a value such as  $p$  is known as the parameter of the test. It is of interest to construct a one-sided  $\alpha$  level test for the null hypothesis  $H_0 : p \leq p_0$  versus the alternative hypothesis  $H_a : p > p_0$ . Let the sample size be denoted by  $n$  and let the observed number of patients out of  $n$  who achieve a complete remission be designated by  $X$ . Then  $X$  follows a Binomial  $(n, p)$  distribution. The rejection region is calculated from this binomial distribution with  $p = p_0$ , the value of  $p$  under the null hypothesis and closest to the alternative hypothesis. In this example, large values of  $X$  are consistent with large values of  $p$  under the alternative hypothesis, therefore the rejection region takes the form: if the observed number of complete remissions  $X$  is greater than or equal to a value  $x$  then reject the null hypothesis  $H_0 : p \leq p_0$ . The value of  $x$  defining this level rejection region is calculated from the preceding binomial distribution such that  $\alpha$  equals the probability that the number of complete remissions  $X$  out of a sample size  $n$  is greater than or equal to  $x$  when  $p = p_0$ . Formally:

$$\alpha = P(X \geq x \text{ given } p = p_0, n).$$

As a numerical example, let  $p_0 = 0.7$  and  $n = 20$ . Then for  $x = 17$

$$0.10 = P(X \geq 17 \text{ given } p = 0.7, n = 20),$$

indicating for this example that an  $\alpha = 0.10$  one-sided rejection region takes the form: reject  $H_0 : p \leq 0.7$  in favor of  $H_a : p > 0.7$  if the number of complete remissions  $X$  out of a sample of size  $n = 20$  is greater than or equal to 17.

The power is then the probability that the number of complete remissions  $X$  out of a sample of size  $n = 20$  is greater than or equal to 17 when  $p > 0.7$ . Formally,

$$\text{Power} = P(X \geq 17 \text{ given } p > 0.7, n = 20).$$

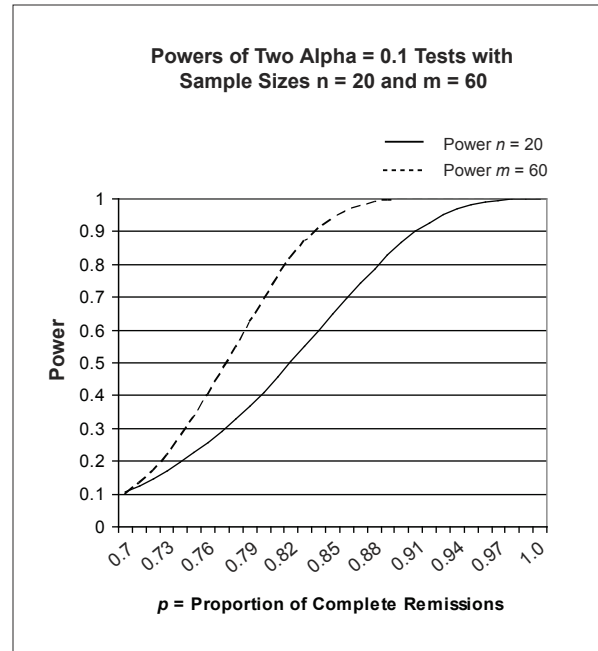
That is, the power of an  $\alpha$  level test is a function of the sample size  $n$  and the values of the proportion  $p$  under the alternative hypothesis. Figure 1 provides a graphic representation of the power as a function of  $p$ . The graph represented by the “Power  $n = 20$ ” curve in Figure 1 depicts the preceding numeric example. Note that for values of  $p$  near 0.7, the largest value of  $p$  under the null hypothesis, the power is near the  $\alpha$  level 0.10 and then the power increases to 1 as  $p$  approaches 1. In other words, it is harder to correctly decide that the true value of  $p$  lies in the alternative hypothesis region when  $p$  is near, but slightly larger than,  $p_0$  and easier as  $p$  increases away from  $p_0$  towards 1. Therefore, in this example, the power is an increasing function of  $p$  for  $p > 0.7$ .

### H&O Why should researchers be concerned with the power of a test?

**BM** For a given level  $\alpha$  test with a sample size  $n$ , the power function provides the probability of correctly rejecting the null hypothesis for values of the parameter under the alternative hypothesis. Therefore, the power function is important because it supplies information concerning the chances that the test will produce the correct conclusion when the alternative hypothesis is true.

As illustrated in the next example, the information provided by the power function can be used: a) to evaluate the “acceptability” of a specific test procedure, and b) to replace unacceptable test procedures with acceptable ones, when acceptable means “correctly concluding the alternative hypothesis is true with high probability.”

When designing a study, the researcher will generally determine a value of the parameter (under the alternative hypothesis) that dictates a clinically significant change from the null hypothesis parameter value. For instance, as in the last example, when testing the proportion of complete remissions  $H_0 : p \leq 0.7$  versus  $H_a : p > 0.7$ , suppose the researcher indicates that a proportion of 0.82



**Figure 1.** Power functions associated with binomial tests on proportions.

or higher provides a clinically significant improvement in the complete remission rate over the null hypothesis rate 0.7. The researcher’s goal therefore is to design an  $\alpha$  level test procedure that provides reasonable power to correctly conclude  $H_a : p > 0.7$  when  $p = 0.82$ . From Figure 1 when the sample size is  $n = 20$  and  $p = 0.82$  the power is approximately 0.50. That is, if the population proportion of complete remissions  $p$  equals 0.82 then the  $\alpha = 0.10$  level test with a sample size of  $n = 20$  (reject  $H_0 : p \leq 0.7$  in favor of  $H_a : p > 0.7$  when  $X \geq 17$ ) has only a 50/50 chance of correctly rejecting  $H_0$ . This test procedure is therefore underpowered, as the same probability of correctly rejecting  $H_0$  can be attained by foregoing any data collection and simply flipping a fair coin to make the decision between the hypotheses. The solution to this underpowered problem lies in the sample size of the test. The fact is: for two  $\alpha$  level procedures with different sample sizes (say,  $n$  and  $m$  with  $m > n$ ) whose statistics ( $X$  and  $Y$ ) have the same form (eg,  $X$  and  $Y$  are the number of complete remissions out of  $n$  and  $m$  patients, respectively) and whose rejection regions are constructed in the same manner, the test procedure with the larger sample size will have higher power for all values of the parameter under the alternative hypothesis. Therefore, in the example, the solution to the previously underpowered test is to construct an  $\alpha$  level test with a large enough sample size to provide an acceptable power when  $p = 0.82$ . For exam-

ple, by increasing the sample size from  $n = 20$  to  $m = 60$  we obtain:

$$0.10 = P(Y \geq 47 \text{ given } p = 0.7, m = 60).$$

That is, with a sample size of  $m = 60$  an  $\alpha = 0.10$  one-sided rejection region takes the form: reject  $H_0 : p \leq 0.7$  in favor of  $H_a : p > 0.7$  when the number of complete remissions  $Y$  is greater than or equal to 47.

The power is then the probability that the number of complete remissions  $Y$  out of a sample size of  $m = 60$  is greater than or equal to 47 when  $p > 0.7$ . Formally,

$$\text{Power} = P(Y \geq 47 \text{ given } p > 0.7, m = 60).$$

The specific power values for this case are represented by the “Power  $m = 60$ ” curve in Figure 1. From the figure, when  $m = 60$  and  $p = 0.82$ , the power is approximately 0.80. That is, if the true proportion of complete remissions  $p$  equals 0.82 then with a sample size of  $m = 60$ , the  $\alpha = 0.10$  level test (reject  $H_0 : p \leq 0.7$  in favor of  $H_a : p > 0.7$  when  $Y \geq 47$ ) has an 80% chance of correctly rejecting  $H_0$ . If 80% is an acceptable power for the  $\alpha = 0.10$  level test then the sample size and associated rejection region are also acceptable. If a higher power is sought for an  $\alpha = 0.10$  level test of  $H_0 : p \leq 0.7$  versus  $H_a : p > 0.7$ , then a larger sample must be utilized, a new rejection region calculated, and a new power curve generated.

In the previous examples, an  $\alpha = 0.10$  level one-sided test for the hypotheses  $H_0 : p \leq p_0$  versus  $H_a : p > p_0$  was examined using the binomial distribution. However, the same general concepts apply to any  $\alpha$  level test on hypotheses with any parameter using statistics from any known distribution.

### H&O In the design of a study, can the power be used in a misleading manner?

**BM** Yes, the power can be used in a misleading manner. For example, suppose a study design is described in the following manner. To test the hypotheses that the proportion of complete remissions is  $H_0 : p \leq 0.7$  versus  $H_a : p > 0.7$ , an  $\alpha = 0.10$  level test with 90% power is provided by the rejection region: reject  $H_0 : p \leq 0.7$  in favor of  $H_a : p > 0.7$  when the number of complete remissions out of 20 patients is greater than or equal to 17. On the surface the test appears to afford a powerful procedure for testing  $H_0$  versus  $H_a$ . Strictly speaking, the preceding statement is not incorrect, but it is misleading as it does not supply the complete remission rate used to power

the study, the rate that represents the minimal clinically significant improvement above 0.70. In this case if the true proportion of complete remissions  $p$  equals 0.92 then from the “Power  $n = 20$ ” curve in Figure 1 the power is approximately 0.90, verifying that the statement is not incorrect. However, for a clinically significant complete remission proportion of 0.82 the test procedure with a sample of 20 patients has only a 50% chance of correctly rejecting  $H_0 : p \leq 0.7$ , as noted earlier. Therefore, when evaluating the power of the test, all relevant aspects of the procedure should be understood and examined.

### H&O What relevant aspects should be examined when evaluating the power of a study?

**BM** The power for a specific test procedure is a function of the  $\alpha$  level of the test, the sample size, and the alternative hypothesis value of the parameter used to power the study. The alpha level of the test chosen to insure that the probability of rejecting the null hypothesis (when the null hypothesis is true) is small. The alternative hypothesis value of the parameter generally represents the smallest clinically significant change from the null hypothesis values of the parameter. For an  $\alpha$  level test, the sample size is then chosen to produce an acceptable power calculated for the specified alternative value of the parameter. The following relationships hold between the power, the  $\alpha$  level of the test, the alternative hypothesis value of the parameter, and the sample size.

- For a fixed sample size and alternative hypothesis value of the parameter, as the  $\alpha$  level of the test decreases, the power of the test decreases.
- For a fixed  $\alpha$  level of the test and alternative hypothesis value of the parameter, as the sample size increases, the power of the test increases.
- For a fixed  $\alpha$  level of the test and sample size, as the alternative hypothesis value of the parameter moves away from the null hypothesis parameter region, the power increases.

### H&O Should studies always be designed using power?

**BM** The power of a hypothesis test provides essential information concerning the probability of the test to correctly reject the null hypothesis when the alternative hypothesis is true. Therefore, any study that involves a hypothesis test should indicate the power of the test as a function of the sample size, the  $\alpha$  level, and the alternative hypothesis value of the parameter that is associated with smallest clinically significant change from the null hypothesis.