

TESTS OF HYPOTHESES AND SIGNIFICANCE, OR DECISION RULES

If we suppose that a particular hypothesis is true but find that the results observed in a random sample differ markedly from the results expected under the hypothesis (i.e., expected on the basis of pure chance, using sampling theory), then we would say that the observed differences are *significant* and would thus be inclined to reject the hypothesis (or at least not accept it on the basis of the evidence obtained). For example, if 20 tosses of a coin yield 16 heads, we would be inclined to reject the hypothesis that the coin is fair, although it is conceivable that we might be wrong.

Procedures that enable us to determine whether observed samples differ significantly from the results expected, and thus help us decide whether to accept or reject hypotheses, are called *tests of hypotheses*, *tests of significance*, *rules of decision*, or simply *decision rules*.

TYPE I AND TYPE II ERRORS

If we reject a hypothesis when it should be accepted, we say that a *Type I error* has been made. If, on the other hand, we accept a hypothesis when it should be rejected, we say that a *Type II error* has been made. In either case, a wrong decision or error in judgment has occurred.

In order for decision rules (or tests of hypotheses) to be good, they must be designed so as to minimize errors of decision. This is not a simple matter, because for any given sample size, an attempt to decrease one type of error is generally accompanied by an increase in the other type of error. In practice, one type of error may be more serious than the other, and so a compromise should be reached in favor of limiting the more serious error. The only way to reduce both types of error is to increase the sample size, which may or may not be possible.

LEVEL OF SIGNIFICANCE

In testing a given hypothesis, the maximum probability with which we would be willing to risk a Type I error is called the *level of significance*, or *significance level*, of the test. This probability, often denoted by α , is generally specified before any samples are drawn so that the results obtained will not influence our choice.

In practice, a significance level of 0.05 or 0.01 is customary, although other values are used. If, for example, the 0.05 (or 5%) significance level is chosen in designing a decision rule, then there are about 5 chances in 100 that we would reject the hypothesis when it should be accepted; that is, we are about 95% *confident* that we have made the right decision. In such case we say that the hypothesis has been rejected at the 0.05 significance level, which means that the hypothesis has a 0.05 probability of being wrong.

TESTS INVOLVING NORMAL DISTRIBUTIONS

To illustrate the ideas presented above, suppose that under a given hypothesis the sampling distribution of a statistic S is a normal distribution with mean μ_S and standard deviation σ_S . Thus the distribution of the standardized variable (or z score), given by $z = (S - \mu_S)/\sigma_S$, is the standardized normal distribution (mean 0, variance 1), as shown in Fig. 10-1.

As indicated in Fig. 10-1, we can be 95% confident that if the hypothesis is true, then the z score of an actual sample statistic S will lie between -1.96 and 1.96 (since the area under the normal curve between these values is 0.95). However, if on choosing a single sample at random we find that the z score of its statistic lies *outside* the range -1.96 to 1.96 , we would conclude that such an event could happen with a probability of only 0.05 (the total shaded area in the figure) if the given hypothesis were true. We would then say that this z score differed *significantly* from what would be expected under the hypothesis, and we would then be inclined to reject the hypothesis.

The total shaded area 0.05 is the significance level of the test. It represents the probability of our being wrong in rejecting the hypothesis (i.e., the probability of making a Type I error). Thus we say that

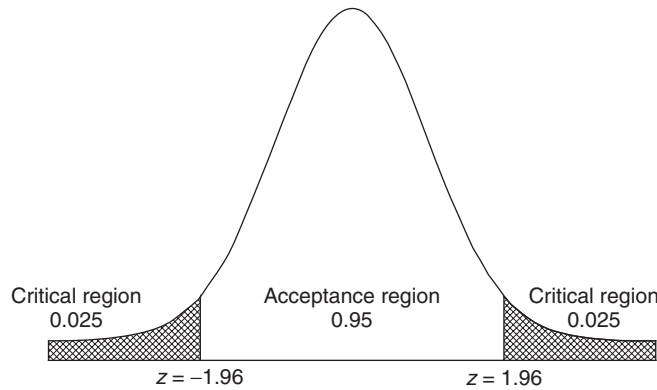


Fig. 10-1 Standard normal curve with the critical region (0.05) and acceptance region (0.95).

the hypothesis is *rejected at 0.05 the significance level* or that the z score of the given sample statistic is *significant at the 0.05 level*.

The set of z scores outside the range -1.96 to 1.96 constitutes what is called the *critical region of the hypothesis, the region of rejection of the hypothesis, or the region of significance*. The set of z scores inside the range -1.96 to 1.96 is thus called the *region of acceptance of the hypothesis, or the region of non-significance*.

On the basis of the above remarks, we can formulate the following decision rule (or test of hypothesis or significance):

Reject the hypothesis at the 0.05 significance level if the z score of the statistic S lies outside the range -1.96 to 1.96 (i.e., either $z > 1.96$ or $z < -1.96$). This is equivalent to saying that the observed sample statistic is significant at the 0.05 level.

Accept the hypothesis otherwise (or, if desired, make no decision at all).

Because the z score plays such an important part in tests of hypotheses, it is also called a *test statistic*.

It should be noted that other significance levels could have been used. For example, if the 0.01 level were used, we would replace 1.96 everywhere above with 2.58 (see Table 10.1). Table 9.1 can also be used, since the sum of the significance and confidence levels is 100%.

Table 10.1

Level of significance, α	0.10	0.05	0.01	0.005	0.002
Critical values of z for one-tailed tests	-1.28 or 1.28	-1.645 or 1.645	-2.33 or 2.33	-2.58 or 2.58	-2.88 or 2.88
Critical values of z for two-tailed tests	-1.645 and 1.645	-1.96 and 1.96	-2.58 and 2.58	-2.81 and 2.81	-3.08 and 3.08

TWO-TAILED AND ONE-TAILED TESTS

In the above test we were interested in extreme values of the statistic S or its corresponding z score on *both* sides of the mean (i.e., in both tails of the distribution). Such tests are thus called *two-sided tests, or two-tailed tests*.

Often, however, we may be interested only in extreme values to one side of the mean (i.e., in one tail of the distribution), such as when we are testing the hypothesis that one process is better than another (which is different from testing whether one process is better or worse than the other). Such tests are

called *one-sided tests*, or *one-tailed tests*. In such cases the critical region is a region to one side of the distribution, with area equal to the level of significance.

Table 10.1, which gives critical values of z for both one-tailed and two-tailed tests at various levels of significance, will be found useful for reference purposes. Critical values of z for other levels of significance are found from the table of normal-curve areas (Appendix II).

SPECIAL TESTS

For large samples, the sampling distributions of many statistics are normal distributions (or at least nearly normal), and the above tests can be applied to the corresponding z scores. The following special cases, taken from Table 8.1, are just a few of the statistics of practical interest. In each case the results hold for infinite populations or for sampling with replacement. For sampling without replacement from finite populations, the results must be modified. See page 182.

1. **Means.** Here $S = \bar{X}$, the sample mean; $\mu_S = \mu_{\bar{X}} = \mu$, the population mean; and $\sigma_S = \sigma_{\bar{X}} = \sigma/\sqrt{N}$, where σ is the population standard deviation and N is the sample size. The z score is given by

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{N}}$$

When necessary, the sample deviation s or \hat{s} is used to estimate σ .

2. **Proportions.** Here $S = P$, the proportion of “successes” in a sample; $\mu_S = \mu_P = p$, where p is the population proportion of successes and N is the sample size; and $\sigma_S = \sigma_P = \sqrt{pq/N}$, where $q = 1 - p$.

The z score is given by

$$z = \frac{P - p}{\sqrt{pq/N}}$$

In case $P = X/N$, where X is the actual number of successes in a sample, the z score becomes

$$z = \frac{X - Np}{\sqrt{Npq}}$$

That is, $\mu_X = \mu = Np$, $\sigma_X = \sigma = \sqrt{Npq}$, and $S = X$.

The results for other statistics can be obtained similarly.

OPERATING-CHARACTERISTIC CURVES; THE POWER OF A TEST

We have seen how the Type I error can be limited by choosing the significance level properly. It is possible to avoid risking Type II errors altogether simply by not making them, which amounts to never accepting hypotheses. In many practical cases, however, this cannot be done. In such cases, use is often made of *operating-characteristic curves*, or *OC curves*, which are graphs showing the probabilities of Type II errors under various hypotheses. These provide indications of how well a given test will enable us to minimize Type II errors; that is, they indicate the *power of a test* to prevent us from making wrong decisions. They are useful in designing experiments because they show such things as what sample sizes to use.

p -VALUES FOR HYPOTHESES TESTS

The p -value is the probability of observing a sample statistic as extreme or more extreme than the one observed under the assumption that the null hypothesis is true. When testing a hypothesis, state the

value of α . Calculate your p -value and if the p -value $\leq \alpha$, then reject H_0 . Otherwise, do not reject H_0 . For testing means, using large samples ($n > 30$), calculate the p -value as follows:

1. For $H_0: \mu = \mu_0$ and $H_1: \mu < \mu_0$, p -value = $P(Z < \text{computed test statistic})$,
2. For $H_0: \mu = \mu_0$ and $H_1: \mu > \mu_0$, p -value = $P(Z > \text{computed test statistic})$, and
3. For $H_0: \mu = \mu_0$ and $H_1: \mu \neq \mu_0$, p -value = $P(Z < -|\text{computed test statistic}|) + P(Z > |\text{computed test statistic}|)$.

The computed test statistic is $\frac{\bar{x} - \mu_0}{(s/\sqrt{n})}$, where \bar{x} is the mean of the sample, s is the standard deviation of the sample, and μ_0 is the value specified for μ in the null hypothesis. Note that if σ is unknown, it is estimated from the sample by using s . This method of testing hypothesis is equivalent to the method of finding a critical value or values and if the computed test statistic falls in the rejection region, reject the null hypothesis. The same decision will be reached using either method.

CONTROL CHARTS

It is often important in practice to know when a process has changed sufficiently that steps should be taken to remedy the situation. Such problems arise, for example, in quality control. Quality control supervisors must often decide whether observed changes are due simply to chance fluctuations or are due to actual changes in a manufacturing process because of deteriorating machine parts, employees' mistakes, etc. *Control charts* provide a useful and simple method for dealing with such problems (see Problem 10.16).

TESTS INVOLVING SAMPLE DIFFERENCES

Differences of Means

Let \bar{X}_1 and \bar{X}_2 be the sample means obtained in large samples of sizes N_1 and N_2 drawn from respective populations having means μ_1 and μ_2 and standard deviations σ_1 and σ_2 . Consider the null hypothesis that there is *no difference* between the population means (i.e., $\mu_1 = \mu_2$), which is to say that the samples are drawn from two populations having the same mean.

Placing $\mu_1 = \mu_2$ in equation (5) of Chapter 8, we see that the sampling distribution of differences in means is approximately normally distributed, with its mean and standard deviation given by

$$\mu_{\bar{X}_1 - \bar{X}_2} = 0 \quad \text{and} \quad \sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}} \tag{1}$$

where we can, if necessary, use the sample standard deviations s_1 and s_2 (or \hat{s}_1 and \hat{s}_2) as estimates of σ_1 and σ_2 .

By using the standardized variable, or z score, given by

$$z = \frac{\bar{X}_1 - \bar{X}_2 - 0}{\sigma_{\bar{X}_1 - \bar{X}_2}} = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}_1 - \bar{X}_2}} \tag{2}$$

we can test the null hypothesis against alternative hypotheses (or the significance of an observed difference) at an appropriate level of significance.

Differences of Proportions

Let P_1 and P_2 be the sample proportions obtained in large samples of sizes N_1 and N_2 drawn from respective populations having proportions p_1 and p_2 . Consider the null hypothesis that there is *no difference* between the population parameters (i.e., $p_1 = p_2$) and thus that the samples are really drawn from the same population.

Placing $p_1 = p_2 = p$ in equation (6) of Chapter 8, we see that the sampling distribution of differences in proportions is approximately normally distributed, with its mean and standard deviation given by

$$\mu_{P_1 - P_2} = 0 \quad \text{and} \quad \sigma_{P_1 - P_2} = \sqrt{pq \left(\frac{1}{N_1} + \frac{1}{N_2} \right)} \quad (3)$$

where

$$p = \frac{N_1 P_1 + N_2 P_2}{N_1 + N_2}$$

is used as an estimate of the population proportion and where $q = 1 - p$.

By using the standardized variable

$$z = \frac{P_1 - P_2 - 0}{\sigma_{P_1 - P_2}} = \frac{P_1 - P_2}{\sigma_{P_1 - P_2}} \quad (4)$$

we can test observed differences at an appropriate level of significance and thereby test the null hypothesis.

Tests involving other statistics can be designed similarly.

TESTS INVOLVING BINOMIAL DISTRIBUTIONS

Tests involving binomial distributions (as well as other distributions) can be designed in a manner analogous to those using normal distributions; the basic principles are essentially the same. See Problems 10.23 to 10.28.

Solved Problems

TESTS OF MEANS AND PROPORTIONS, USING NORMAL DISTRIBUTIONS

10.1 Find the probability of getting between 40 and 60 heads inclusive in 100 tosses of a fair coin.

SOLUTION

According to the binomial distribution, the required probability is

$$\binom{100}{40} \left(\frac{1}{2}\right)^{40} \left(\frac{1}{2}\right)^{60} + \binom{100}{41} \left(\frac{1}{2}\right)^{41} \left(\frac{1}{2}\right)^{59} + \cdots + \binom{100}{60} \left(\frac{1}{2}\right)^{60} \left(\frac{1}{2}\right)^{40}$$

Since $Np = 100\left(\frac{1}{2}\right)$ and $Nq = 100\left(\frac{1}{2}\right)$ are both greater than 5, the normal approximation to the binomial distribution can be used in evaluating this sum. The mean and standard deviation of the number of heads in 100 tosses are given by

$$\mu = Np = 100\left(\frac{1}{2}\right) = 50 \quad \text{and} \quad \sigma = \sqrt{Npq} = \sqrt{(100)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)} = 5$$

On a continuous scale, between 40 and 60 heads inclusive is the same as between 39.5 and 60.5 heads. We thus have

$$39.5 \text{ in standard units} = \frac{39.5 - 50}{5} = -2.10 \quad 60.5 \text{ in standard units} = \frac{60.5 - 50}{5} = 2.10$$

$$\begin{aligned} \text{Required probability} &= \text{area under normal curve between } z = -2.10 \text{ and } z = 2.10 \\ &= 2(\text{area between } z = 0 \text{ and } z = 2.10) = 2(0.4821) = 0.9642 \end{aligned}$$

10.2 To test the hypothesis that a coin is fair, adopt the following decision rule:

Accept the hypothesis if the number of heads in a single sample of 100 tosses is between 40 and 60 inclusive.

Reject the hypothesis otherwise.

- Find the probability of rejecting the hypothesis when it is actually correct.
- Graph the decision rule and the result of part (a).
- What conclusions would you draw if the sample of 100 tosses yielded 53 heads? And if it yielded 60 heads?
- Could you be wrong in your conclusions about part (c)? Explain.

SOLUTION

(a) From Problem 10.1, the probability of not getting between 40 and 60 heads inclusive if the coin is fair is $1 - 0.9642 = 0.0358$. Thus the probability of rejecting the hypothesis when it is correct is 0.0358.

(b) The decision rule is illustrated in Fig. 10-2, which shows the probability distribution of heads in 100 tosses of a fair coin. If a single sample of 100 tosses yields a z score between -2.10 and 2.10 , we accept the hypothesis; otherwise, we reject the hypothesis and decide that the coin is not fair.

The error made in rejecting the hypothesis when it should be accepted is the *Type I error* of the decision rule; and the probability of making this error, equal to 0.0358 from part (a), is represented by the total shaded area of the figure. If a single sample of 100 tosses yields a number of heads whose z score (or z statistic) lies in the shaded regions, we would say that this z score differed *significantly* from what would be expected if the hypothesis were true. For this reason, the total shaded area (i.e., the probability of a Type I error) is called the *significance level* of the decision rule and equals 0.0358 in this case. Thus we speak of rejecting the hypothesis at the 0.0358 (or 3.58%) significance level.

(c) According to the decision rule, we would have to accept the hypothesis that the coin is fair in both cases. One might argue that if only one more head had been obtained, we would have rejected the hypothesis. This is what one must face when any sharp line of division is used in making decisions.

(d) Yes. We could accept the hypothesis when it actually should be rejected—as would be the case, for example, when the probability of heads is really 0.7 instead of 0.5. The error made in accepting the hypothesis when it should be rejected is the *Type II error* of the decision.

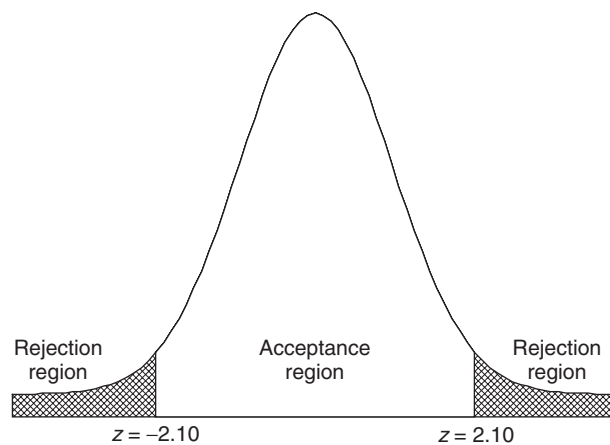


Fig. 10-2 Standard normal curve showing the acceptance and rejection regions for testing that the coin is fair.

- 10.3** Using the binomial distribution and not the normal approximation to the binomial distribution, design a decision rule to test the hypothesis that a coin is fair if a sample of 64 tosses of the coin is taken and a significance level of 0.05 is used. Use MINITAB to assist with the solution.

SOLUTION

The binomial plot of probabilities when a fair coin is tossed 64 times is given in Fig. 10.3. Partial cumulative probabilities generated by MINITAB are shown below the Fig. 10-3.

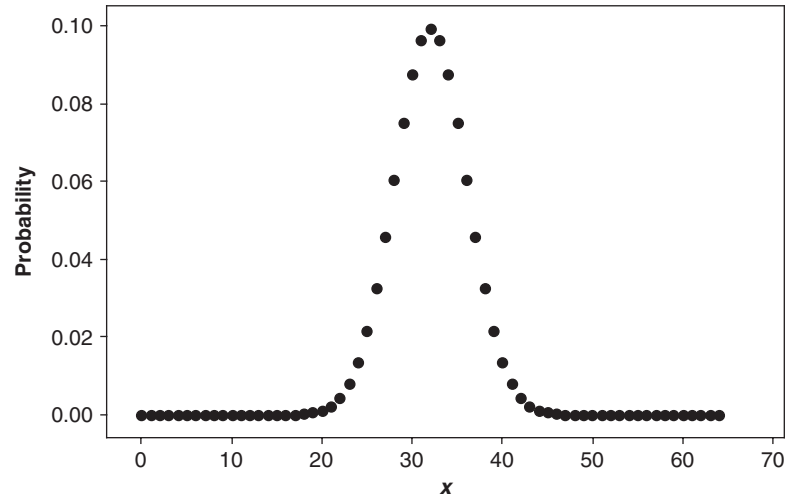


Fig. 10-3 MINITAB plot of the binomial distribution for $n=64$ and $p=0.5$.

x	Probability	Cumulative	x	Probability	Cumulative
0	0.0000000	0.0000000	13	0.0000007	0.0000009
1	0.0000000	0.0000000	14	0.0000026	0.0000035
2	0.0000000	0.0000000	15	0.0000086	0.0000122
3	0.0000000	0.0000000	16	0.0000265	0.0000387
4	0.0000000	0.0000000	17	0.0000748	0.0001134
5	0.0000000	0.0000000	18	0.0001952	0.0003087
6	0.0000000	0.0000000	19	0.0004727	0.0007814
7	0.0000000	0.0000000	20	0.0010636	0.0018450
8	0.0000000	0.0000000	21	0.0022285	0.0040735
9	0.0000000	0.0000000	22	0.0043556	0.0084291
10	0.0000000	0.0000000	23	0.0079538	0.0163829
11	0.0000000	0.0000001	24	0.0135877	0.0299706
12	0.0000002	0.0000002	25	0.0217403	0.0517109

We see that $P(X \leq 23) = 0.01638$. Because the distribution is symmetrical, we also know $P(X \geq 41) = 0.01638$. The rejection region $\{X \leq 23 \text{ and } X \geq 41\}$ has probability $2(0.01638) = 0.03276$. The rejection region $\{X \leq 24 \text{ and } X \geq 40\}$ would exceed 0.05. When the binomial distribution is used we cannot have a rejection region equal to exactly 0.05. The closest we can get to 0.05 without exceeding it is 0.03276.

Summarizing, the coin will be flipped 64 times. It will be declared unfair or not balanced if 23 or fewer heads or 41 or more heads are obtained. The chance of making a Type 1 error is 0.03276 which is as close to 0.05 as you can get without exceeding it.

- 10.4** Refer to Problem 10.3. Using the binomial distribution and not the normal approximation to the binomial distribution, design a decision rule to test the hypothesis that a coin is fair if a sample

of 64 tosses of the coin is taken and a significance level of 0.05 is used. Use EXCEL to assist with the solution.

SOLUTION

The outcomes 0 through 64 are entered into column A of the EXCEL worksheet. The expressions =BINOMDIST(A1,64,0.5,0) and =BINOMDIST(A1,64,0.5,1) are used to obtain the binomial and cumulative binomial distributions. The 0 for the fourth parameter requests individual probabilities and the 1 requests cumulative probabilities. A click-and-drag in column B gives the individual probabilities, and in column C a click-and-drag gives the cumulative probabilities.

A	B	C	A	B	C
X	Probability	Cumulative	x	Probability	Cumulative
0	5.42101E-20	5.42101E-20	13	7.12151E-07	9.40481E-07
1	3.46945E-18	3.52366E-18	14	2.59426E-06	3.53474E-06
2	1.09288E-16	1.12811E-16	15	8.64754E-06	1.21823E-05
3	2.25861E-15	2.37142E-15	16	2.64831E-05	3.86654E-05
4	3.44438E-14	3.68152E-14	17	7.47758E-05	0.000113441
5	4.13326E-13	4.50141E-13	18	0.000195248	0.000308689
6	4.06437E-12	4.51451E-12	19	0.000472706	0.000781395
7	3.36762E-11	3.81907E-11	20	0.001063587	0.001844982
8	2.39943E-10	2.78134E-10	21	0.002228469	0.004073451
9	1.49298E-09	1.77111E-09	22	0.004355644	0.008429095
10	8.21138E-09	9.98249E-09	23	0.007953785	0.01638288
11	4.03104E-08	5.02929E-08	24	0.013587715	0.029970595
12	1.78038E-07	2.28331E-07	25	0.021740344	0.051710939

It is found, as in Problem 10.3, that $P(X \leq 23) = 0.01638$ and because of symmetry, $P(X \geq 41) = 0.01638$ and that the rejection region is $\{X \leq 23 \text{ or } X \geq 41\}$ and the significance level is $0.01638 + 0.01638$ or 0.03276.

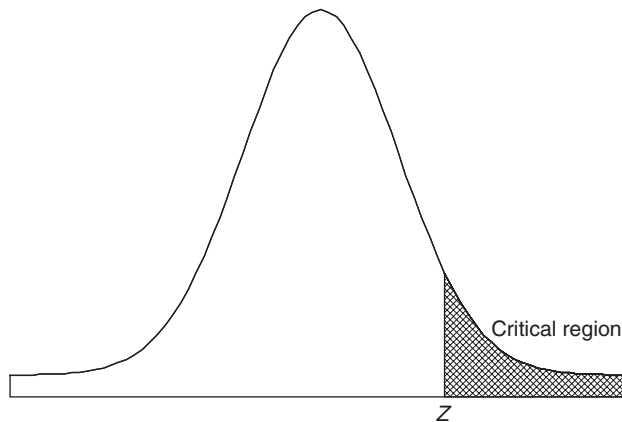


Fig. 10-4 Determining the Z value that will give a critical region equal to 0.05.

- 10.5** In an experiment on extrasensory perception (ESP), an individual (subject) in one room is asked to state the color (red or blue) of a card chosen from a deck of 50 well-shuffled cards by an individual in another room. It is unknown to the subject how many red or blue cards are in the deck. If the subject identifies 32 cards correctly, determine whether the results are significant at the (a) 0.05 and (b) 0.01 levels.

SOLUTION

If p is the probability of the subject choosing the color of a card correctly, then we have to decide between two hypotheses:

$H_0 : p = 0.5$, and the subject is simply guessing (i.e., the results are due to chance).

$H_1 : p > 0.5$, and the subject has powers of ESP.

Since we are not interested in the subject's ability to obtain extremely low scores, but only in the ability to obtain high scores, we choose a one-tailed test. If hypothesis H_0 is true, then the mean and standard deviation of the number of cards identified correctly are given by

$$\mu = Np = 50(0.5) = 25 \quad \text{and} \quad \sigma = \sqrt{Npq} = \sqrt{50(0.5)(0.5)} = \sqrt{12.5} = 3.54$$

- (a) For a one-tailed test at the 0.05 significance level, we must choose z in Fig. 10-4 so that the shaded area in the critical region of high scores is 0.05. The area between 0 and z is 0.4500, and $z = 1.645$; this can also be read from Table 10.1. Thus our decision rule (or test of significance) is:

If the z score observed is greater than 1.645, the results are significant at the 0.05 level and the individual has powers of ESP.

If the z score is less than 1.645, the results are due to chance (i.e., not significant at the 0.05 level).

Since 32 in standard units is $(32 - 25)/3.54 = 1.98$, which is greater than 1.645, we conclude at the 0.05 level that the individual has powers of ESP.

Note that we should really apply a continuity correction, since 32 on a continuous scale is between 31.5 and 32.5. However, 31.5 has a standard score of $(31.5 - 25)/3.54 = 1.84$, and so the same conclusion is reached.

- (b) If the significance level is 0.01, then the area between 0 and z is 0.4900, from which we conclude that $z = 2.33$.

Since 32 (or 31.5) in standard units is 1.98 (or 1.84), which is less than 2.33, we conclude that the results are *not significant* at the 0.01 level.

Some statisticians adopt the terminology that results significant at the 0.01 level are *highly significant*, that results significant at the 0.05 level but not at the 0.01 level are *probably significant*, and that results significant at levels larger than 0.05 are *not significant*. According to this terminology, we would conclude that the above experimental results are *probably significant*, so that further investigations of the phenomena are probably warranted.

Since significance levels serve as guides in making decisions, some statisticians quote the actual probabilities involved. For instance, since $pr\{z \geq 1.84\} = 0.0322$, in this problem, the statistician could say that on the basis of the experiment the chances of being wrong in concluding that the individual has powers of ESP are about 3 in 100. The quoted probability (0.0322 in this case) is called the p -value for the test.

- 10.6** The claim is made that 40% of tax filers use computer software to file their taxes. In a sample of 50, 14 used computer software to file their taxes. Test $H_0 : p = 0.4$ versus $H_a : p < 0.4$ at $\alpha = 0.05$ where p is the population proportion who use computer software to file their taxes. Test using the binomial distribution and test using the normal approximation to the binomial distribution.

SOLUTION

If the exact test of $H_0 : p = 0.4$ versus $H_a : p < 0.4$ at $\alpha = 0.05$ is used, the null is rejected if $X \leq 15$. This is called the rejection region. If the test based on the normal approximation to the binomial is used, the null is rejected if $Z < -1.645$ and this is called the rejection region. $X = 14$ is called the test statistic. The binomial test statistic is in the rejection region and the null is rejected. Using the normal approximation, the test statistic is $z = \frac{14 - 20}{3.46} = -1.73$. The actual value of α is 0.054 and the rejection region is $X \leq 15$ and the cumulative binomial probability $P(X \leq 15)$ is used. If the normal approximation is used, you would also reject since $z = -1.73$ is in the rejection region which is $Z < -1.645$. Note that if the binomial distribution is used to perform the test, the test statistic has a binomial distribution. If the normal distribution is used to test the hypothesis, the test statistic, Z , has a standard normal distribution.

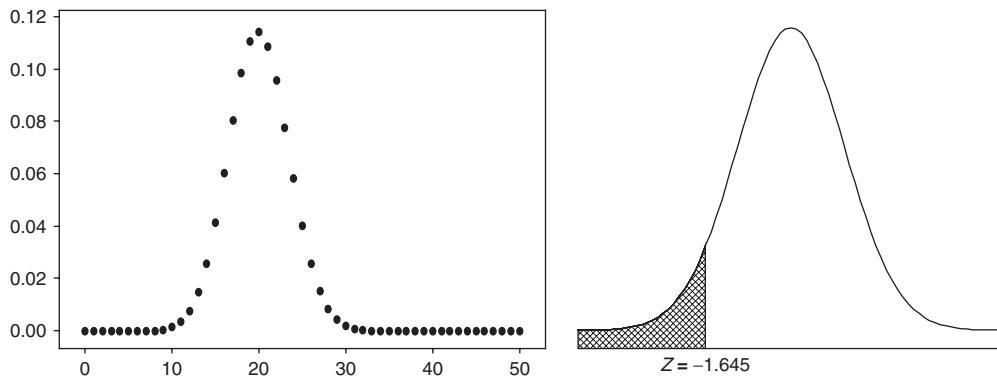


Fig. 10-5 Comparison of the exact test on the left (Binomial) and the approximate test on the right (standard normal).

10.7 The p -value for a test of hypothesis is defined to be the smallest level of significance at which the null hypothesis is rejected. This problem illustrates the computation of the p -value for a statistical test. Use the data in Problem 9.6 to test the null hypothesis that the mean height of all the trees on the farm equals 5 feet (ft) versus the alternative hypothesis that the mean height is less than 5 ft. Find the p -value for this test.

SOLUTION

The computed value for z is $z = (59.22 - 60)/1.01 = -0.77$. The smallest level of significance at which the null hypothesis would be rejected is $p\text{-value} = P(z < -0.77) = 0.5 - 0.2794 = 0.2206$. The null hypothesis is rejected if the p -value is less than the pre-set level of significance. In this problem, if the level of significance is pre-set at 0.05, then the null hypothesis is not rejected. The MINITAB solution is as follows where the subcommand `Alternative-1` indicates a lower-tail test.

```
MTB > ZTest mean = 60 sd = 10.111 data in c1 ;
SUBC> Alternative -1.
```

Z-Test

Test of $\mu = 60.00$ vs $\mu < 60.00$
 The assumed sigma = 10.1

Variable	N	Mean	StDev	SE Mean	Z	P
height	100	59.22	10.11	1.01	-0.77	0.22

10.8 A random sample of 33 individuals who listen to talk radio was selected and the hours per week that each listens to talk radio was determined. The data are as follows.

- 9 8 7 4 8 6 8 8 7 10 8 10 6 7 7 8 9
- 6 5 8 5 6 8 7 8 5 5 8 7 6 6 4 5

Test the null hypothesis that $\mu = 5$ hours (h) versus the alternative hypothesis that $\mu \neq 5$ at level of significance $\alpha = 0.05$ in the following three equivalent ways:

- (a) Compute the value of the test statistic and compare it with the critical value for $\alpha = 0.05$.
- (b) Compute the p -value corresponding to the computed test statistic and compare the p -value with $\alpha = 0.05$.
- (c) Compute the $1 - \alpha = 0.95$ confidence interval for μ and determine whether 5 falls in this interval.

SOLUTION

In the following MINITAB output, the standard deviation is found first, and then specified in the Ztest statement and the Zinterval statement.

```
MTB > standard deviation c1
Standard deviation of hours = 1.6005

MTB > ZTest 5.0 1.6005 'hours';
SUBC> Alternative 0.
```

Z-Test

```
Test of mu = 5.000 vs mu not = 5.000
The assumed sigma = 1.60
```

Variable	N	Mean	StDev	SE Mean	Z	P
hours	33	6.897	1.600	0.279	6.81	0.0000

```
MTB > ZInterval 95.0 1.6005 'hours'.
```

Variable	N	Mean	StDev	SE Mean	95.0 % CI
hours	33	6.897	1.600	0.279	(6.351, 7.443)

- (a) The computed value of the test statistic is $Z = \frac{6.897 - 5}{0.279} = 6.81$, the critical values are ± 1.96 , and the null hypothesis is rejected. Note that this is the computed value shown in the MINITAB output.
- (b) The computed p -value from the MINITAB output is 0.0000 and since the p -value $< \alpha = 0.05$, the null hypothesis is rejected.
- (c) Since the value specified by the null hypothesis, 5, is not contained in the 95% confidence interval for μ , the null hypothesis is rejected.

These three procedures for testing a null hypothesis against a two-tailed alternative are equivalent.

- 10.9** The breaking strengths of cables produced by a manufacturer have a mean of 1800 pounds (lb) and a standard deviation of 100 lb. By a new technique in the manufacturing process, it is claimed that the breaking strength can be increased. To test this claim, a sample of 50 cables is tested and it is found that the mean breaking strength is 1850 lb. Can we support the claim at the 0.01 significance level?

SOLUTION

We have to decide between the two hypotheses:

$H_0 : \mu = 1800$ lb, and there is really no change in breaking strength.

$H_1 : \mu > 1800$ lb, and there is a change in breaking strength.

A one-tailed test should be used here; the diagram associated with this test is identical with Fig. 10-4 of Problem 10.5(a). At the 0.01 significance level, the decision rule is:

If the z score observed is greater than 2.33, the results are significant at the 0.01 level and H_0 is rejected.

Otherwise, H_0 is accepted (or the decision is withheld).

Under the hypothesis that H_0 is true, we find that

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{N}} = \frac{1850 - 1800}{100/\sqrt{50}} = 3.55$$

which is greater than 2.33. Hence we conclude that the results are *highly significant* and that the claim should thus be supported.