Chapter 10 HYPOTHESIS TESTING

Bibliography: Sheldon Ross

10.1 INTRODUCTION

We assume that we have a *random sample from a population distribution*, specified except for a vector of unknown parameters.

This time, we do not want to explicitly estimate the *unknown parameters*, now suppose that we are concerned with *using the resulting sample to test some particular hypothesis concerning them*. As an illustration, suppose that a minimarket has just purchased a large supply of bread that have been guaranteed to have an average weight of at least 700 g each. To verify this claim, the firm has decided to take a random sample of 10 bread loaves to determine their weight. They will then use the result of this experiment to ascertain whether or not they accept the baker's hypothesis that the population mean is at least 700 g per bread loaf.

A *statistical hypothesis* is usually a statement about a set of parameters of a population distribution. It is called a hypothesis because it is not known whether or not it is true. A primary problem is *to develop a procedure* for determining whether or not *the values of a random sample from this population are consistent with the hypothesis*. For instance, consider a particular normally distributed population having an unknown mean value θ and known variance 1. The statement "θ is less than 1" is a statistical hypothesis that we could try to test by observing a random sample from this population. If the random sample is deemed to be consistent with the hypothesis under consideration, we say that the hypothesis has been "accepted"; otherwise we say that it has been "rejected."

Note that in accepting a given hypothesis we are not actually claiming that it is true but rather we are saying that the resulting data appear to be consistent with it. For instance, in the case of a normal $(\theta,1)$ population, if a resulting sample of size 10 has an average value of 1.25, then although such a result cannot be regarded as being evidence in favor of the hypothesis "$\theta<1$," it is not inconsistent with this hypothesis, which would thus be accepted. On the other hand, if the sample of size 10 has an average value of 3, then even though a sample value that large is possible when $\theta<1$, it is so unlikely that it seems inconsistent with this hypothesis, which would thus be rejected.

10.2 SIGNIFICANCE LEVELS

Consider a population having distribution F_θ, where θ is unknown, and suppose we want to test a specific hypothesis about θ. We shall denote this hypothesis by H_0 and call it the *null hypothesis*. For example, if F_θ is a normal distribution function with mean θ and variance equal to 1, then two possible null hypotheses about θ are

$$a)\ H_0 : \theta = 1 \tag{10.1}$$

$$b)\ H_0 : \theta \le 1 \tag{10.2}$$

The first of these hypotheses states that the population is normal with mean 1 and variance 1, whereas the second states that it is normal with variance 1 and a mean less than or equal to 1. Note that the null hypothesis in (a), when true, completely specifies the population distribution, whereas the null hypothesis in (b) does not.

Suppose now that in order *to test a specific null hypothesis* H_0, a population sample of size *n* — say $X_1, X_2, \ldots, X_n$ — is to be observed. Based on these *n* values, we must decide whether or not to accept H_0. A test for H_0 can be specified by defining a region *C* in *n*-dimensional space with the proviso that *the hypothesis is to be rejected if the random sample* $X_1, X_2, \ldots, X_n$ *turns out to lie in C and accepted otherwise.* The region *C* is called the *critical region*. In other words, the statistical test determined by the critical region *C* is the one that

$$\text{accepts } H_0 \text{ if } (X_1, X_2, \ldots, X_n) \notin C \tag{10.3}$$

$$\text{and rejects } H_0 \text{ if } (X_1, X_2, \ldots, X_n) \in C \tag{10.4}$$

Example: A common test of the hypothesis that θ, the mean of a normal population with variance 1, is equal to 1 has a critical region given by

$$C = \left\{ (X_1, X_2, \ldots, X_n) : \left| \bar{X} - 1 \right| > 1.96 / \sqrt{n} \right\} \tag{10.5}$$

This test calls for rejection of the null hypothesis that θ = 1 when the sample average differs from 1 by more than 1.96 divided by the square root of the sample size.

When developing a procedure for testing a given null hypothesis H_0, two different types of errors can result.

- a *type I error*, results if the test incorrectly calls for rejecting H_0 when it is indeed correct.
- a *type II error*, results if the test calls for accepting H_0 when it is false.

The objective of a statistical test of H_0 is not to explicitly determine whether or not H_0 is true but rather to determine if its validity is consistent with the resultant data. Hence, with this objective it seems reasonable *that H_0 should only be rejected if the resultant data are very unlikely when H_0 is true*. The classical way of accomplishing this is to specify a value α and then require the test to have the property that whenever H_0 is true its probability of being rejected is never greater than α. The value α, called the *level of significance of the test*, is usually set in advance, with commonly chosen values being $\alpha = 0.1,\ 0.05,\ 0.005$. In other words, the classical approach to testing H_0 is to fix a significance level α and then require that the test have the property that the probability of a type I error occurring can never be greater than α.

Suppose now that we are interested in testing a certain hypothesis concerning θ, an unknown parameter of the population. Specifically, for a given set of parameter values *w*, suppose we are interested in testing

$$H_0 : \theta \in w \qquad (10.6)$$

A common approach to developing a test of H_0, say at level of significance α, is to start by determining a *point estimator of* θ — say $d(\mathbf{X})$. *The hypothesis is then rejected if* $d(\mathbf{X})$ *is "far away" from the region w*. However, to determine how "far away" it need be to justify rejection of H_0, we need to determine the probability distribution of the estimator $d(\mathbf{X})$ when H_0 is true since this will usually enable us to determine the appropriate *critical region* so as to make the test have the required significance level α.

Example: the test of the hypothesis that the mean of a normal $(\theta,1)$ population is equal to 1, given by Equation (10.5), calls for rejection when the point estimate of θ —that is, the sample average—is farther than $1.96/\sqrt{n}$ away from 1. The value $1.96/\sqrt{n}$ was chosen to meet a level of significance of $\alpha = 0.05$.

10.3 TESTS CONCERNING THE MEAN OF A NORMAL POPULATION

10.3.1 *Case of Known Variance*

Suppose that $X_1, X_2, \ldots, X_n$ is a sample of size *n* from a normal distribution having an unknown mean μ and a known variance σ^2 and suppose we are interested in testing the null hypothesis

$$H_0 : \mu = \mu_0 \tag{10.7}$$

against the alternative hypothesis

$$H_1 : \mu \neq \mu_0 \tag{10.8}$$

where μ_0 is some specified constant.

Since $\bar{X} = \sum X_i / n$ is a natural point estimator of μ, it seems reasonable to accept H_0 if $\bar{X}$ is not too far from μ_0. That is, the *critical region* of the test would be of the form

$$C = \left\{ (X_1, X_2, \ldots, X_n) : \left| \bar{X} - \mu_0 \right| > c \right\} \tag{10.9}$$

for some suitably chosen value *c*.

If we desire that the test has significance level α, then we must determine the critical value *c* in Equation (10.9) that will make the *type I error* probability equal to α. That is, *c* must be such that

$$P_{\mu_0} \left(\left| \bar{X} - \mu_0 \right| > c \right) = \alpha \tag{10.10}$$

where we write P_{μ_0} to mean that the preceding probability is to be computed under the assumption that $\mu = \mu_0$. When $\mu = \mu_0$, $\bar{X}$ will be normally distributed with mean μ_0 and variance σ^2 / n and so *Z*, defined by

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} = \frac{\sqrt{n}\left(\bar{X} - \mu_0 \right)}{\sigma} \tag{10.11}$$

will have a standard normal distribution. Now Equation (10.10) is equivalent to

$$P\left(|Z| > \frac{c\sqrt{n}}{\sigma} \right) = \alpha \tag{10.12}$$

$$2P\left(Z > \frac{c\sqrt{n}}{\sigma}\right) = \alpha \qquad P\left(Z > \frac{c\sqrt{n}}{\sigma}\right) = \frac{\alpha}{2} \tag{10.13}$$

where Z is a standard normal random variable. However, we know that

$$P\left(Z > z_{\alpha/2}\right) = \frac{\alpha}{2} \tag{10.14}$$

And so

$$\frac{c\sqrt{n}}{\sigma} = z_{\alpha/2}$$

$$c = \frac{z_{\alpha/2}\sigma}{\sqrt{n}} \tag{10.15}$$

Thus, *the significance level α test* is to reject H_0 if $\left|\bar{X} - \mu_0\right| > z_{\alpha/2}\sigma/\sqrt{n}$ and accept otherwise; or, equivalently, to

$$\text{reject } H_0 \text{ if } \frac{\sqrt{n}}{\sigma}\left|\bar{X} - \mu_0\right| > z_{\alpha/2} \tag{10.16}$$

$$\text{accept } H_0 \text{ if } \frac{\sqrt{n}}{\sigma}\left|\bar{X} - \mu_0\right| \le z_{\alpha/2} \tag{10.17}$$

This test can be pictorially represented as shown in Figure 10.1, where we have superimposed the standard normal density function [which is the density probability function of the test statistic $\sqrt{n}\left(\bar{X} - \mu_0\right)/\sigma$ when H_0 is true].

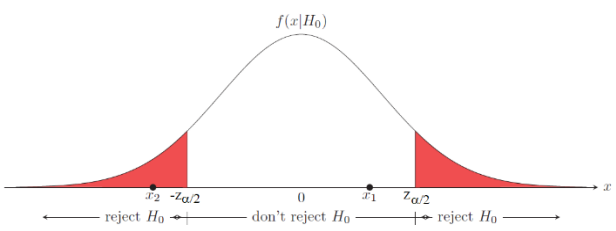

Figure 10.1 The significance level α test.

If our data produced the test statistic x_1 then we would not reject the null hypothesis H_0. On the other hand, if our data produced x_2 then we would reject the null hypothesis in favor of the alternative hypothesis.

Exercise 1. It is known that if a signal of value μ is sent from location *A*, then the value received at location *B* is normally distributed with mean μ and standard deviation 2. That is, the random noise added to the signal is a $N(0,4)$ random variable. There is reason for the people at location *B* to suspect that the signal value $\mu = 8$ will be sent today. Test this hypothesis if the same signal value is independently sent five times and the mean value received at location *B* is $\bar{X} = 9.5$.

Suppose we are testing at the 5 percent level of significance $\alpha = 0.05$. To begin, we compute the test statistic

$$\frac{\sqrt{n}}{\sigma}\left|\bar{X} - \mu_0\right| = \frac{\sqrt{5}}{2}\left|9.5 - 8\right| = 1.68$$

Since this value is less than $z_{0.025} = 1.96$, the hypothesis is *accepted*. The data are not inconsistent with the null hypothesis in the sense that a sample average as far from the value 8 as observed would be expected, when the true mean is 8.

If a less stringent significance level were chosen — say $\alpha = 0.1$ — then the null hypothesis would have been rejected. This follows since $z_{0.05} = 1.645$, which is less than 1.68. Hence, if we would have chosen a test that had a 10 percent chance of rejecting H_0 when H_0 was true, then the null hypothesis would have been rejected.

The "correct" level of significance to use in a given situation depends on the individual circumstances involved in that situation. For instance, if rejecting a null hypothesis H_0 would result in large costs that would thus be lost if H_0 were indeed true, then we might elect to be quite conservative and so choose a significance level of 0.05 or 0.01.

The test given by Equation (10.16)

$$\text{reject } H_0 \text{ if } \frac{\sqrt{n}}{\sigma}\left|\bar{X} - \mu_0\right| > z_{\alpha/2} \qquad (10.18)$$

can be described as follows: For any observed value of the test statistic $\sqrt{n}\left(\bar{X} - \mu_0\right)/\sigma$, call it *v*, the test calls for rejection of the null hypothesis if the probability that the test statistic would be as larger than *v* (exceed in absolut value) when H_0 is true is less than or equal to the significance level α.

From this, it follows that:

We can determine whether or not to accept the null hypothesis by computing, first, the value of the test statistic and, second, the probability that a standard normal would (in absolute value) exceed that quantity. This probability—called the *p*-value of the test—gives *the critical significance level* in the sense that H_0 will be accepted if the significance level α is less than the *p*-value and rejected if α is greater than or equal.

In practice, the significance level is often not set in advance but rather the data are looked at to determine the resultant *p*-value. Sometimes, this critical significance level is clearly much larger than any we would want to use, and so the null hypothesis can be readily accepted. At other times the *p*-value is so small that it is clear that the hypothesis should be rejected.

Exercise 2. In Exercise 1, suppose that the mean of the 5 values received is $\bar{X} = 8.5$. In this case,

$$\frac{\sqrt{n}}{\sigma}\left|\bar{X} - \mu_0\right| = \frac{\sqrt{5}}{2}\left|8.5 - 8\right| = \frac{\sqrt{5}}{4} = 0.559$$

Since

$$p - value = P\left(|Z| > 0.559\right) = 2P\left(Z > 0.559\right) = 2 \times 0.288 = 0.576$$

it follows that the *p*-value is 0.576 and thus the null hypothesis H_0 that the signal sent has value 8 would be accepted at any significance level $\alpha < 0.576$.

On the other hand, if the average of the data values were 11.5, then the *p*-value of the test that the mean is equal to 8 would be

$$\frac{\sqrt{n}}{\sigma}\left|\bar{X} - \mu_0\right| = \frac{\sqrt{5}}{2}\left|11.5 - 8\right| = \frac{\sqrt{5}}{2} \times 3.5 = 3.913$$

$$p - value = P\left(|Z| > 3.913\right) = 2P\left(Z > 3.913\right) = 2 \times 0.0000455 = 0.0000911$$

For such a small *p*-value, the hypothesis that the value 8 was sent is rejected.

We have not yet talked about the *probability of a type II error*—that is, the probability of accepting the null hypothesis when the true mean μ is unequal to μ_0. This probability will depend on the value of μ, and so let us define $\beta(\mu)$ by

$$\beta(\mu) = P_\mu\left(\text{acceptance of } H_0\right) \tag{10.19}$$

$$= P_\mu\left(\left|\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\right| \le z_{\alpha/2}\right)$$

$$= P_\mu\left(-z_{\alpha/2} \le \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right)$$

The function $\beta(\mu)$ is called the *operating characteristic* (or OC) *curve* and represents the probability that H_0 will be accepted when the true mean is μ.

To compute this probability, we use the fact that $\bar{X}$ is normal with mean μ and variance σ^2/n and so

$$Z \equiv \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) \tag{10.20}$$

Hence,

$$\beta(\mu) = P_\mu\left(-z_{\alpha/2} \le \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right)$$

$$= P_\mu\left(-z_{\alpha/2} - \frac{\mu}{\sigma/\sqrt{n}} \le \frac{\bar{X}-\mu_0-\mu}{\sigma/\sqrt{n}} \le z_{\alpha/2} - \frac{\mu}{\sigma/\sqrt{n}}\right)$$

$$= P_\mu\left(-z_{\alpha/2} - \frac{\mu}{\sigma/\sqrt{n}} \le Z - \frac{\mu_0}{\sigma/\sqrt{n}} \le z_{\alpha/2} - \frac{\mu}{\sigma/\sqrt{n}}\right)$$

$$= P_\mu\left(-z_{\alpha/2} + \frac{\mu_0-\mu}{\sigma/\sqrt{n}} \le Z \le z_{\alpha/2} + \frac{\mu_0-\mu}{\sigma/\sqrt{n}}\right)$$

$$= \phi\left(z_{\alpha/2} + \frac{\mu_0-\mu}{\sigma/\sqrt{n}}\right) - \phi\left(-z_{\alpha/2} + \frac{\mu_0-\mu}{\sigma/\sqrt{n}}\right) \tag{10.21}$$

where ϕ is the standard normal distribution function.

For a fixed significance level α, the OC curve given by Equation (10.21) is symmetric about μ_0 and will depend on μ only through $\sqrt{n}|\mu-\mu_0|/\sigma$. This curve with the abscissa changed from μ to $d = \sqrt{n}|\mu-\mu_0|/\sigma$ is presented in Figure 8.2 (Sheldon) when $\alpha = 0.05$.

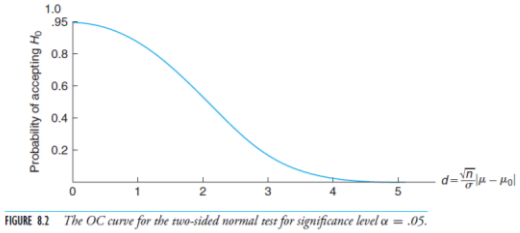

FIGURE 8.2 *The OC curve for the two-sided normal test for significance level* $\alpha = .05$.

Figure 10.2 Operating characteristic curve.

Exercise 3. For the problem presented in Exercise 1, let us determine the probability of accepting the null hypothesis that $\mu = 8$ when the actual value sent is 10. To do so, we compute

$$\frac{\sqrt{n}}{\sigma}(\mu_0 - \mu) = -\frac{\sqrt{5}}{2} \times 2 = -\sqrt{5}$$

As $z_{0.025} = 1.96$, the desired probability is, from Equation (10.21),

$$\beta = \phi\left(-\sqrt{5} + 1.96\right) - \phi\left(-\sqrt{5} - 1.96\right) = 0.392$$

```
pnorm(-sqrt(5)+1.96,0,1)-pnorm(-sqrt(5)-1.96,0,1)
[1] 0.3912343
```

REMARK

The function $1-\beta(\mu)$ is called the *power-function* of the test. Thus, for a given value μ, the power of the test is equal to the probability of rejection H_0 when μ is the true value.

Significance level and *power* are used to quantify the quality of the test. Ideally a significance test would not make errors. That is, it would not reject H_0 when H_0 was true and would reject H_0 in favor of H_1 when H_1 was true.

The two probabilities we focus on are:

Significance level = $P\left(\text{reject } H_0 \middle| H_0\right)$

= probability we incorrectly reject $H_0 = P(\text{type I error}) = \alpha$

Power = probability we correctly reject H_0

$$= P\left(\text{reject } H_0 \middle| H_1\right) = 1 - P(\text{type II error}) = 1 - \beta$$

Ideally, a hypothesis test should have a small significance level (near 0) and a large power (near 1).

Here are two analogies to help you remember the meanings of significance level and power:

1. Think of H_0 as the hypothesis 'nothing noteworthy is going on', i.e. 'the coin is fair', 'the treatment is no better than placebo' etc. And think of H_1 as the opposite: 'something interesting is happening'. Then power is the probability of detecting something interesting when it's present and significance level is the probability of mistakenly claiming some thing interesting has occured.

2. In the U.S. criminal defendents (inculpati penali) are presumed innocent until proven guilty beyond a reasonable doubt. We can phrase this in Null Hypothesis terms as

   H_0: the defendant is innocent (the default)

   H_1: the defendant is guilty.

Significance level is the probability of finding an innocent person guilty. *Power* is the probability of correctly finding a guilty party guilty. 'Beyond a reasonable doubt' means we should demand the significance level be very small.

The next two figures show high and low power tests.

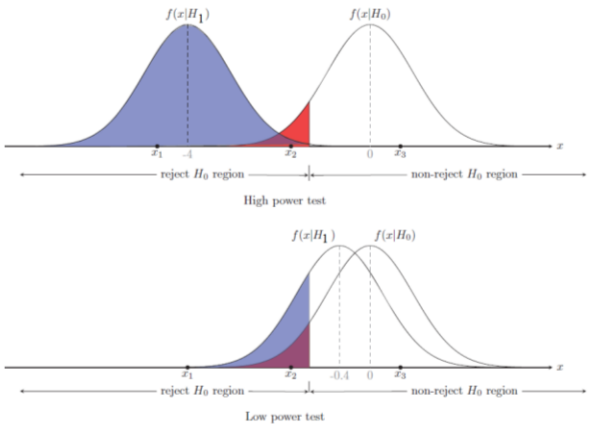

Figure 10.3 High and low power tests.

The shaded area under $f(x|H_0)$ represents the *significance level*. Remember the significance level is:

•The probability of falsely rejecting the null hypothesis when it is true.

•The probabilitiy the test statistic falls in the rejection region even though H_0 is true.

Likewise, the shaded area under $f(x|H_1)$ represents the *power*, i.e. the probability that the test statistic is in the rejection (of H_0) region when H_1 is true. Both tests have the same significance level, but if $f(x|H_1)$ has considerable overlap with $f(x|H_0)$ the power is much lower.

In both tests both distributions are standard normal. The null distribution, rejection region and significance level are all the same. (The significance level is the red/purple area under $f(x|H_0)$ and above the rejection region.)

In the top figure we see the means of the two distributions are 4 standard deviations apart. Since the areas under the densities have very little overlap the test has high power. That is if the data *x* is drawn from H_1 it will almost certainly be in the rejection region. For example x_3 would be a very surprising outcome for the H_1 distribution.

In the bottom figure we see the means of the two distributions are just 0.4 standard deviations apart. Since the areas under the densities have a lot of overlap the test has low power. That is if the data *x* is drawn from H_1 it is highly likely to be in the non-rejection region. For example x_3 would be not be a very surprising outcome for the H_1 distribution.

Typically we can increase the power of a test by increasing the amount of data and thereby decreasing the variance of the null and alternative distributions. In experimental design it is important to determine ahead of time the number of trials or subjects needed to achieve a desired power.

The operating characteristic function is useful in determining how large the random sample need be to meet certain specifications concerning type II errors. For instance, suppose that we desire to determine the sample size *n* necessary to ensure that the probability of accepting $H_0 : \mu = \mu_0$ when the true mean is actually μ_1 is approximately β.

That is, we want *n* to be such that

$$\beta(\mu_1) \approx \beta$$

$$\phi\left(z_{\alpha/2} + \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}}\right) - \phi\left(-z_{\alpha/2} + \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}}\right) \approx \beta \qquad (10.22)$$

A solution can be obtained by using the standard normal distribution table. In addition, an approximation for *n* can be derived from Equation (10.22). To start, suppose that $\mu_1 > \mu_0$. Then, because this implies that

$$\frac{\sqrt{n}(\mu_0 - \mu_1)}{\sigma} - z_{\alpha/2} \le -z_{\alpha/2} \qquad (10.23)$$

it follows, since ϕ is an increasing function, that

$$\phi\left(\frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} - z_{\alpha/2}\right) \le \phi(-z_{\alpha/2}) = P(Z \le -z_{\alpha/2}) = P(Z \ge z_{\alpha/2}) = \alpha/2 \qquad (10.24)$$

Hence, we can take

$$\phi\left(\frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} - z_{\alpha/2}\right) \approx 0 \qquad (10.25)$$

$$\beta \approx \phi\left(z_{\alpha/2} + \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}}\right) \qquad (10.26)$$

Or, since

$$\beta = P(Z > z_\beta) = P(Z < -z_\beta) = \phi(-z_\beta) \qquad (10.27)$$

we obtain from Equation (10.26) that

$$-z_\beta \approx (\mu_0 - \mu_1)\frac{\sqrt{n}}{\sigma} + z_{\alpha/2}$$

$$n \approx \frac{(z_{\alpha/2} + z_\beta)^2 \sigma^2}{(\mu_0 - \mu_1)^2} \qquad (10.28)$$

In fact, the same approximation would result when $\mu_1 < \mu_0$ and so Equation (10.28) is in all cases a reasonable approximation to the sample size necessary to ensure that the type II error at the value $\mu = \mu_1$ is approximately equal to β.

Exercise 4. For the problem of Exercise 1, how many signals need be sent so that the 0.05 level test of $H_0 : \mu_0 = 8$ has at least a 75 percent probability of rejection when $\mu = 9.2$?

Since $1-\beta = 0.75$, $\beta = 0.25$, $z_{0.025} = 1.96$, $z_{0.25} = 0.67$, the approximation (10.28) yields

$$n \approx \frac{\left(z_{\alpha/2} + z_{\beta}\right)^2 \sigma^2}{\left(\mu_0 - \mu_1\right)^2} = \frac{\left(1.96 + 0.67\right)^2 \times 4}{\left(1.2\right)^2} = 19.21$$

Hence a sample of size 20 is needed. From Equation (10.21), we see that with $n = 20$

$$\beta(9.2) = \phi\left(z_{\alpha/2} + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}\right) - \phi\left(-z_{\alpha/2} + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}\right) = \phi\left(1.96 - \frac{1.2\sqrt{20}}{2}\right) - \phi\left(-1.96 - \frac{1.2\sqrt{20}}{2}\right)$$

$$= \phi(-0.723) - \phi(-4.643) \approx 1 - \phi(0.723) \approx 0.235$$

Therefore, if the message is sent 20 times, then there is a 76.5 percent chance that the null hypothesis $\mu = 8$ will be rejected when the true mean is 9.2.