TESTING OF HYPOTHESIS

Dr. Krishan K. Pandey

**Professor & Dean, Office of Doctoral Studies (ODS),
O.P. Jindal Global University, Sonipat**

2011

POPULATION: A population consists of an entire set of objects, observations, or scores that have something in common. For example, a population might be defined as all males between the ages of 15 and 18. Some populations are only hypothetical. Consider an experimenter interested in the possible effectiveness of a new method of teaching reading. He or she might define a population as the reading achievement scores that would result if all six year olds in the US were taught with this new method. The population is hypothetical in the sense that there does not exist, a group of students who have been taught using the new method; the population consists of the scores that would be obtained if they were taught with this method.

The distribution of a population can be described by several parameters such as the mean and standard deviation. Estimates of these parameters taken from a sample are called statistics.

SAMPLE: A sample is a subset of a population. Since it is usually impractical to test every member of a population, a sample from the population is typically the best approach available. Inferential statistics generally require that sampling be random although some types of sampling (such as those used in voter polling) seek to make the sample as representative of the population as possible by choosing the sample to resemble the population on the most important characteristics.

STATISTICS: The word "statistics" is used in several different senses. In the broadest sense, "statistics" refers to a range of techniques and procedures for analyzing data, interpreting data, displaying data, and making decisions based on data. This is what courses in "statistics" generally cover.

In a second usage, a "statistic" is defined as a numerical quantity (such as the mean)   calculated in a sample. Such statistics are used to estimate parameters. The term "statistics" sometimes refers to calculated quantities regardless of whether or not they are from a sample. For example, one might ask about a baseball player's statistics and be referring to his or her batting average, runs batted in, number of home runs, etc. Or, "government statistics" can refer to any numerical indexes calculated by a governmental agency.

Although the different meanings of "statistics" have the potential for confusion, a careful consideration of the context in which the word is used should make its intended meaning clear.

PARAMETER: A parameter is a numerical quantity measuring some aspect of a population of scores. For example, the mean is a measure of central tendency. Greek letters are used to designate parameters. At the bottom of this page are shown several parameters of great importance in statistical analyses and the Greek symbol that represents each one. Parameters are rarely known and are usually estimated by statistics computed in samples. To the right of each Greek symbol is the symbol for the associated statistic used to estimate it from a sample.

Quantity	Parameter	Statistic
Mean	μ	M
Standard deviation	σ	s
Proportion	π	p
Correlation	ρ	r

SIGNIFICANCE TEST: A significance test is performed to determine if an observed value of a statistic differs enough from a hypothesized value of a parameter to draw the inference that the hypothesized value of the parameter is not the true value. The hypothesized value of the parameter is called the "null hypothesis." A significance test consists of calculating the probability of obtaining a statistic as different or more different from the null hypothesis (given that the null hypothesis is correct) than the statistic obtained in the sample. If this probability is sufficiently low, then the difference between the parameter and the statistic is said to be "statistically significant."

Just how low is sufficiently low? The choice is somewhat arbitrary but by convention levels of 0.05 and 0.01 are most commonly used. For instance, an experimenter may hypothesize that the size of a food reward does not affect the speed a rat runs down an alley. One group of rats receives a large reward and another receives a small reward for running the alley. Suppose the mean running time for the large reward were 1.5 seconds and the mean running time for the small reward were 2.1 seconds.

The difference between means is thus 2.1 - 1.5 = 0.6 seconds. The test of whether this difference is significant consists of determining the probability of obtaining a difference as large or larger than 0.6 seconds given there is really no effect of magnitude of reward. If the probability is low (below the significance level) then the null hypothesis that magnitude of reward makes no difference is rejected in favor of the alternate hypothesis that it does make a difference. The null hypothesis is not accepted just because it is not rejected.

NULL HYPOTHESIS: The null hypothesis is an hypothesis about a population parameter. The purpose of hypothesis testing is to test the viability of the null hypothesis in the light of experimental data. Depending on the data, the null hypothesis either will or will not be rejected as a viable possibility. Consider a researcher interested in whether the time to respond to a tone is affected by the consumption of alcohol. The null hypothesis is that $\mu_1 - \mu_2 = 0$ where μ_1 is the mean time to respond after consuming alcohol and μ_2 is the mean time to respond otherwise. Thus, the null hypothesis concerns the parameter $\mu_1 - \mu_2$ and the null hypothesis is that the parameter equals zero.

The null hypothesis is often the reverse of what the experimenter actually believes; it is put forward to allow the data to contradict it. In the experiment on the effect of alcohol, the experimenter probably expects alcohol to have a harmful effect. If the experimental data show a sufficiently large effect of alcohol, then the null hypothesis that alcohol has no effect can be rejected.

It should be stressed that researchers very frequently put forward a null hypothesis in the hope that they can discredit it. For a second example, consider an educational researcher who designed a new way to teach a particular concept in science, and wanted to test experimentally whether this new method worked better than the existing method. The researcher would design an experiment comparing the two methods. Since the null hypothesis would be that there is no difference between the two methods, the researcher would be hoping to reject the null hypothesis and conclude that the method he or she developed is the better of the two. The symbol H_0 is used to indicate the null hypothesis. For the example just given, the null hypothesis would be designated by the following symbols: H_0: $\mu_1 - \mu_2 = 0$ or by $H0$: $\mu_1 = \mu_2$.

The null hypothesis is typically a hypothesis of no difference as in this example where it is the hypothesis of no difference between population means. That is why the word "null" in "null hypothesis" is used -- it is the hypothesis of no difference.

Despite the "null" in "null hypothesis," there are occasions when the parameter is not hypothesized to be 0. For instance, it is possible for the null hypothesis to be that the difference between population means is a particular value. Or, the null hypothesis could be that the mean SAT score in some population is 600. The null hypothesis would then be stated as: $H_0: \mu = 600$. Although the null hypotheses discussed so far have all involved the testing of hypotheses about one or more population means, null hypotheses can involve any parameter. An experiment investigating the correlation between job satisfaction and performance on the job would test the null hypothesis that the population correlation (ρ) is 0. Symbolically, $H_0: \rho = 0$. Some possible null hypotheses are given below:

$H_0: \mu=0$; $H_0: \mu=10$; $H_0: \mu_1 - \mu_2 = 0$; $H_0: \pi = 0.5$; $H_0: \pi_1 - \pi_2 = 0$; $H_0: \mu_1 = \mu_2 = \mu_3$; $H_0: \rho_1 - \rho_2 = 0$

When a one-tailed test is conducted, the null hypothesis includes the direction of the effect. A one-tailed test of the differences between means might test the null hypothesis that $\mu_1 - \mu_2 \geq 0$. If M1 - M2 were much less than 0 then the null hypothesis would be rejected in favor of the alternative hypothesis: $\mu_1 - \mu_2 < 0$.

ALTERNATIVE HYPOTHESIS: In hypothesis testing, the null hypothesis and an alternative hypothesis are put forward. If the data are sufficiently strong to reject the null hypothesis, then the null hypothesis is rejected in favor of an alternative hypothesis. For instance, if the null hypothesis were that $\mu_1 = \mu_2$ then the alternative hypothesis (for a two-tailed test) would be $\mu_1 \neq \mu_2$.

ONE- AND TWO-TAILED TESTS: In the section on "Steps in hypothesis testing" the fourth step involves calculating the probability that a statistic would differ as much or more from parameter specified in the null hypothesis as does the statistic obtained in the experiment. This statement implies that a difference in either direction would be counted. That is, if the null hypothesis were: $H_0: \mu - \mu = 0$;

and the value of the statistic $M_1 - M_2$ were +5, then the probability of $M_1 - M_2$ differing from zero by five or more (in either direction) would be computed. In other words, probability value would be the probability that either $M_1 - M_2 \geq 5$ or $M_1 - M_2 \leq -5$. Assume that the figure shown below is the sampling distribution of $M_1 - M_2$.
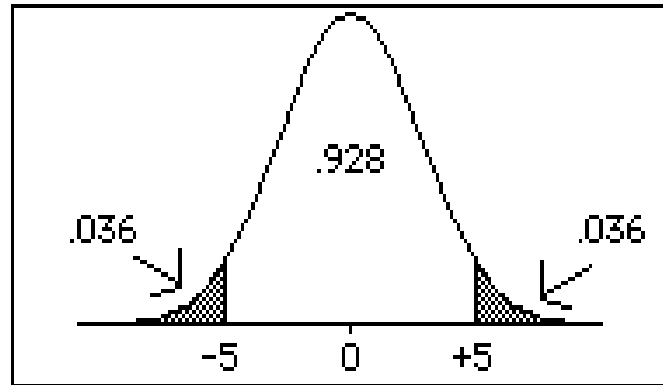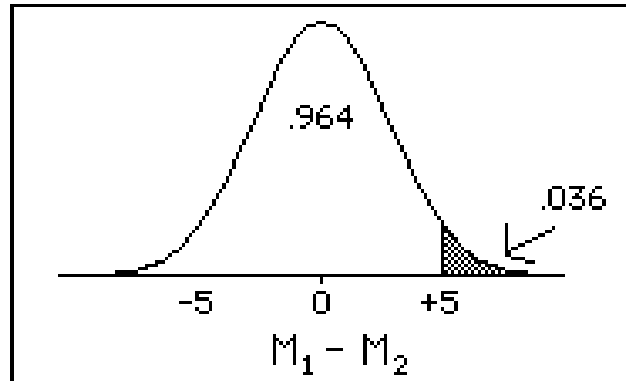

The figure shows that the probability of a value of +5 or more is 0.036 and that the probability of a value of -5 or less is .036. Therefore the probability of a value either greater than or equal to +5 or less than or equal to -5 is $0.036 + 0.036 = 0.072$.

A probability computed considering differences in both directions is called a "two-tailed" probability. The name makes sense since both tails of the sampling distribution are considered.

There are situations in which an experimenter is concerned only with differences in one direction. For example, an experimenter may be concerned with whether or not $\mu_1 - \mu_2$ is greater than zero. However, if $\mu_1 - \mu_2$ is not greater than zero, the experimenter may not care whether it equals zero or is less than zero. For instance, if a new drug treatment is developed, the main issue is whether or not it is better than a placebo. If the treatment is not better than a placebo, then it will not be used. It does not really matter whether or not it is worse than the placebo. When only one direction is of concern to an experimenter, then a "one-tailed" test can be performed. If an experimenter were only concerned with whether or not $\mu_1 - \mu_2$ is greater than zero,

then the one-tailed test would involve calculating the probability of obtaining a statistic as great or greater than the one obtained in the experiment.

In the example, the one-tailed probability would be the probability of obtaining a value of $M_1 - M_2$ greater than or equal to five given that the difference between population means is zero.

The shaded area in the figure is greater than five. The figure shows that the one-tailed probability is 0.036. It is easier to reject the null hypothesis with a one-tailed than with a two-tailed test as long as the effect is in the specified direction. Therefore, one-tailed tests have lower Type II error rates and more power than do two-tailed tests. In this example, the one-tailed probability (0.036) is below the conventional significance level of 0.05 whereas the two-tailed probability (0.072) is not. Probability values for one-tailed tests are one half the value for two-tailed tests as long as the effect is in the specified direction.

One-tailed and two-tailed tests have the same Type I error rate. One-tailed tests are sometimes used when the experimenter predicts the direction of the effect in advance. This use of one-tailed tests is questionable because the experimenter can only reject the null hypothesis if the effect is in the predicted direction. If the effect is in the other direction, then the null hypothesis cannot be rejected no matter how strong the effect is. A skeptic might question whether the experimenter would really fail to reject the null hypothesis if the effect were strong enough in the wrong direction. Frequently the most interesting aspect of an effect is that it runs counter to expectations. Therefore, an experimenter who committed himself or herself to ignoring effects in one direction may be forced to choose between ignoring a potentially important finding and using the techniques of statistical inference dishonestly. One-tailed tests are not used frequently. Unless otherwise indicated, a test should be assumed to be two-tailed.

TYPE I AND II ERRORS: There are two kinds of errors that can be made in significance testing: (1) a true null hypothesis can be incorrectly rejected and (2) a false null hypothesis can fail to be rejected. The former error is called a Type I error and the latter error is called a Type II error. These two types of errors are defined in the table.

Statistical Decision	True State of the Null Hypothesis	
	H_0 True	H_0 False
Reject H_0	Type I error	Correct
Do not Reject H_0	Correct	Type II error

The probability of a Type I error is designated by the Greek letter alpha (α) and is called the Type I error rate; the probability of a Type II error (the Type II error rate) is designated by the Greek letter beta (ß). A Type II error is only an error in the sense that an opportunity to reject the null hypothesis correctly was lost. It is not an error in the sense that an incorrect conclusion was drawn since no conclusion is drawn when the null hypothesis is not rejected.

A Type I error, on the other hand, is an error in every sense of the word. A conclusion is drawn that the null hypothesis is false when, in fact, it is true. Therefore, Type I errors are generally considered more serious than Type II errors. The probability of a Type I error (α) is called the significance level and is set by the experimenter. There is a tradeoff between Type I and Type II errors. The more an experimenter protects himself or herself against Type I errors by choosing a low level, the greater the chance of a Type II error. Requiring very strong evidence to reject the null hypothesis makes it very unlikely that a true null hypothesis will be rejected. However, it increases the chance that a false null hypothesis will not be rejected, thus lowering power. The Type I error rate is almost always set at .05 or at .01, the latter being more conservative since it requires stronger evidence to reject the null hypothesis at the .01 level then at the .05 level.

SIGNIFICANCE LEVEL: In hypothesis testing, the significance level is the criterion used for rejecting the null hypothesis. The significance level is used in hypothesis testing as follows: First,

the difference between the results of the experiment and the null hypothesis is determined. Then, assuming the null hypothesis is true; the probability of a difference that large or larger is computed. Finally, this probability is compared to the significance level. If the probability is less than or equal to the significance level, then the null hypothesis is rejected and the outcome is said to be statistically significant. Traditionally, experimenters have used either the 0.05 level (sometimes called the 5% level) or the 0.01 level (1% level), although the choice of levels is largely subjective. The lower the significance level, the more the data must diverge from the null hypothesis to be significant. Therefore, the 0.01 level is more conservative than the 0.05 level. The Greek letter alpha (α) is sometimes used to indicate the significance level. See also: Type I error and significance test.

POWER: Power is the probability of correctly rejecting a false null hypothesis. Power is therefore defined as: $1 - \beta$ where β is the Type II error probability. If the power of an experiment is low, then there is a good chance that the experiment will be inconclusive. That is why it is so important to consider power in the design of experiments. There are methods for estimating the power of an experiment before the experiment is conducted. If the power is too low, then the experiment can be redesigned by changing one of the factors that determine power. Consider a hypothetical experiment designed to test whether rats brought up in an enriched environment can learn mazes faster than rats brought up in the typical laboratory environment (the control condition). Two groups of 12 rats each are tested. Although the experimenter does not know it, the population mean number of trials it takes to learn the maze is 20 for the enriched condition and 32 for the control condition. The null hypothesis that the enriched environment makes no difference is therefore false.

The question is, "What is the probability that the experimenter is going to be able to demonstrate that the null hypothesis is false by rejecting it at the .05 level?" This is the same thing as asking "What is the power of the test?" Before the power of the test can be determined, the standard deviation (σ) must be known. If $\sigma = 10$ then the power of the significance test is 0.80. This means that there is a 0.80 probability that the experimenter will be able to reject the null hypothesis. Since power = 0.80, $\beta = 1-.80 = .20$.

It is important to keep in mind that power is not about whether or not the null hypothesis is true (It is assumed to be false). It is the probability the data gathered in an experiment will be sufficient to reject the null hypothesis. The experimenter does not know that the null hypothesis is false. The experimenter asks the question: If the null hypothesis is false with specified population means and standard deviation, what is the probability that the data from the experiment will be sufficient to reject the null hypothesis?

If the experimenter discovers that the probability of rejecting the null hypothesis is low (power is low) even if the null hypothesis is false to the degree expected (or hoped for), then it is likely that the experiment should be redesigned. Otherwise, considerable time and expense will go into a project that has a small chance of being conclusive even if the theoretical ideas behind it are correct.

STATISTICAL SIGNIFICANCE: Significance tests are performed to see if the null hypothesis can be rejected. If the null hypothesis is rejected, then the effect found in a sample is said to be statistically significant. If the null hypothesis is not rejected, then the effect is not significant. The experimenter chooses a significance level before conducting the statistical analysis. The significance level chosen determines the probability of a Type I error. A statistically significant effect is not necessarily practically significant.

STEPS IN HYPOTHESIS TESTING: The basic logic of hypothesis testing has been presented somewhat informally in the sections on "Ruling out chance as an explanation" and the "Null hypothesis." In this section the logic will be presented in more detail and more formally.

1. The first step in hypothesis testing is to specify the null hypothesis (H_0) and the alternative hypothesis (H_1). If the research concerns whether one method of presenting pictorial stimuli leads to better recognition than another, the null hypothesis would most likely be that there is no difference between methods (H_0: $\mu_1 - \mu_2 = 0$). The alternative hypothesis would be H_1: $\mu_1 \neq \mu_2$. If the research concerned the correlation between grades and SAT scores, the null hypothesis would most likely be that there is no correlation (H_0: $\rho = 0$). The alternative hypothesis would be H_1: $\rho \neq 0$.

2. The next step is to select a significance level. Typically the 0.05 or the 0.01 level is used.

3. The third step is to calculate a statistic analogous to the parameter specified by the null hypothesis. If the null hypothesis were defined by the parameter $\mu_1 - \mu_2$, then the statistic $M_1 - M_2$ would be computed.

4. The fourth step is to calculate the probability value (often called the p value). The p value is the probability of obtaining a statistic as different or more different from the parameter specified in the null hypothesis as the statistic computed from the data. The calculations are made assuming that the null hypothesis is true.

5. The probability value computed in Step 4 is compared with the significance level chosen in Step 2. If the probability is less than or equal to the significance level, then the null hypothesis is rejected; if the probability is greater than the significance level then the null hypothesis is not rejected. When the null hypothesis is rejected, the outcome is said to be "statistically significant" when the null hypothesis is not rejected then the outcome is said be "not statistically significant."

6. If the outcome is statistically significant, then the null hypothesis is rejected in favor of the alternative hypothesis. If the rejected null hypothesis were that $\mu_1 - \mu_2 = 0$, then the alternative hypothesis would be that $\mu_1 \neq \mu_2$. If M_1 were greater than M_2 then the researcher would naturally conclude that $\mu_1 \geq \mu_2$.

7. The final step is to describe the result and the statistical conclusion in an understandable way. Be sure to present the descriptive statistics as well as whether the effect was significant or not. For example, a significant difference between a group that received a drug and a control group might be described as follow:

Subjects in the drug group scored significantly higher ($M = 23$) than did subjects in the control group ($M = 17$), $t(18) = 2.4$, $p = 0.027$.

The statement that "$t(18) = 2.4$" has to do with how the probability value (p) was calculated. A small minority of researchers might object to two aspects of this wording. First, some believe that the significance level rather than the probability level should be reported. Second, since the alternative hypothesis was stated as $\mu_1 \neq \mu_2$, some might argue that it can

only be concluded that the population means differ and not that the population mean for the drug group is higher than the population mean for the control group.

This argument is misguided. Intuitively, there are strong reasons for inferring that the direction of the difference in the population is the same as the difference in the sample. There is also a more formal argument. A non significant effect might be described as follows:

Although subjects in the drug group scored higher (M = 23) than did subjects in the control group, (M = 20), the difference between means was not significant, t(18) = 1.4, p = 0.179.

It would not have been correct to say that there was no difference between the performances of the two groups. There was a difference. It is just that the difference was not large enough to rule out chance as an explanation of the difference. It would also have been incorrect to imply that there is no difference in the population. Be sure not to accept the null hypothesis.

<u>WHY THE NULL HYPOTHESIS IS NOT ACCEPTED:</u> A null hypothesis is not accepted just because it is not rejected. Data not sufficient to show convincingly that a difference between means is not zero do not prove that the difference is zero. Such data may even suggest that the null hypothesis is false but not be strong enough to make a convincing case that the null hypothesis is false. For example, if the probability value were 0.15, then one would not be ready to present one's case that the null hypothesis is false to the (properly) skeptical scientific community. More convincing data would be needed to do that. However, there would be no basis to conclude that the null hypothesis is true. It may or may not be true; there just is not strong enough evidence to reject it. Not even in cases where there is no evidence that the null hypothesis is false is it valid to conclude the null hypothesis is true. If the null hypothesis is that $\mu_1 - \mu_2$ is zero then the hypothesis is that the difference is exactly zero. No experiment can distinguish between the case of no difference between means and an extremely small difference between means. If data are consistent with the null hypothesis, they are also consistent with other similar hypotheses.

Thus, if the data do not provide a basis for rejecting the null hypothesis that $\mu_1 - \mu_2 = 0$ then they almost certainly will not provide a basis for rejecting the hypothesis that $\mu_1 - \mu_2 = 0.001$. The data are consistent with both hypotheses. When the null hypothesis is not rejected then it is legitimate to conclude that the data are consistent with the null hypothesis. It is not legitimate to conclude

that the data support the acceptance of the null hypothesis since the data are consistent with other hypotheses as well. In some respects, rejecting the null hypothesis is comparable to a jury finding a defendant guilty. In both cases, the evidence is convincing beyond a reasonable doubt. Failing to reject the null hypothesis is comparable to a finding of not guilty. The defendant is not declared innocent. There is just not enough evidence to be convincing beyond a reasonable doubt. In the judicial system, a decision has to be made and the defendant is set free. In science, no decision has to be made immediately. More experiments are conducted.

One experiment might provide data sufficient to reject the null hypothesis, although no experiment can demonstrate that the null hypothesis is true. Where does this leave the researcher who wishes to argue that a variable does not have an effect? If the null hypothesis cannot be accepted, even in principle, then what type of statistical evidence can be used to support the hypothesis that a variable does not have an effect. The answer lies in relaxing the claim a little and arguing not that a variable has no effect whatsoever but that it has, at most, a negligible effect. This can be done by constructing a confidence interval around the parameter value. Consider a researcher interested in the possible effectiveness of a new psychotherapeutic drug. The researcher conducted an experiment comparing a drug-treatment group to a control group and found no significant difference between them. Although the experimenter cannot claim the drug has no effect, he or she can estimate the size of the effect using a confidence interval. If μ_1 were the population mean for the drug group and μ_2 were the population mean for the control group, then the confidence interval would be on the parameter $\mu_1 - \mu_2$.

Assume the experiment measured "well being" on a 50 point scale (with higher scores representing more well being) that has a standard deviation of 10. Further assume the 99% confidence interval computed from the experimental data was: $-0.5 \leq \mu_1 - \mu_2 \leq 1$. This says that one can be confident that the mean "true" drug treatment effect is somewhere between -0.5 and 1. If it were -0.5 then the drug would, on average, be slightly detrimental; if it were 1 then the drug would, on average, be slightly beneficial. But, how much benefit is an average improvement of 1? Naturally that is a question that involves characteristics of the measurement scale. But, since 1 is only 0.10 standard deviations, it can be presumed to be a small effect. The overlap between two distributions whose means differ by 0.10 standard

deviations is shown below. Although the blue distribution is slightly to the right of the red distribution, the overlap is almost complete.
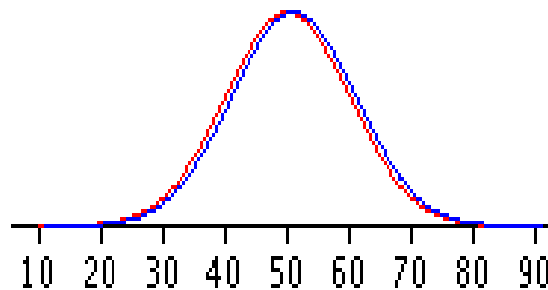
So, the finding that the maximum difference that can be expected (based on a 99% confidence interval) is itself a very small difference would allow the experimenter to conclude that the drug is not effective. The claim would not be that it is totally ineffective, but, at most, its effectiveness is very limited.

THE PRECISE MEANING OF THE PROBABILITY VALUE: There is often confusion about the precise meaning of the probability computed in a significance test. As stated in Step 4 of the steps in hypothesis testing, the null hypothesis (H_0) is assumed to be true. The difference between the statistic computed in the sample and the parameter specified by H_0 is computed and the probability of obtaining a difference this large or large is calculated. This probability value is the probability of obtaining data as extreme or more extreme than the current data (assuming H_0 is true). It is not the probability of the null hypothesis itself. Thus, if the probability value is 0.005, this does not mean that the probability that the null hypothesis is true is .005. It means that the probability of obtaining data as different or more different from the null hypothesis as those obtained in the experiment is 0.005.

The inferential step to conclude that the null hypothesis is false goes as follows: The data (or data more extreme) are very unlikely given that the null hypothesis is true. This means that: (1) a very unlikely event occurred or (2) the null hypothesis is false. The inference usually made is that the null hypothesis is false.To illustrate that the probability is not the probability of the hypothesis, consider a test of a person who claims to be able to predict whether a coin will come up heads or

tails. One should take a rather skeptical attitude toward this claim and require strong evidence to believe in its validity. The null hypothesis is that the person can predict correctly half the time (H_0: $\pi = 0.5$). In the test, a coin is flipped 20 times and the person is correct 11 times. If the person has no special ability (H_0 is true), then the probability of being correct 11 or more times out of 20 is 0.41. Would someone who was originally skeptical now believe that there is only a 0.41 chance that the null hypothesis is true? They almost certainly would not since they probably originally thought H_0 had a very high probability of being true (perhaps as high as 0.9999). There is no logical reason for them to decrease their belief in the validity of the null hypothesis since the outcome was perfectly consistent with the null hypothesis.

The proper interpretation of the test is as follows: A person made a rather extraordinary claim and should be able to provide strong evidence in support of the claim if the claim is to believed. The test provided data consistent with the null hypothesis that the person has no special ability since a person with no special ability would be able to predict as well or better more than 40% of the time. Therefore, there is no compelling reason to believe the extraordinary claim. However, the test does not prove the person cannot predict better than chance; it simply fails to provide evidence that he or she can. The probability that the null hypothesis is true is not determined by the statistical analysis conducted as part of hypothesis testing. Rather, the probability computed is the probability of obtaining data as different or more different from the null hypothesis (given that the null hypothesis is true) as the data actually obtained.

EXAMPLE CALCULATION: For this example assume a normal distribution with a population standard deviation (σ) known to be 100. The null hypothesis is that $\mu = 500$. A random sample of 50 subjects is obtained and the mean of their scores (M) is calculated to be 530. The statistic (M) differs from the parameter specified in the null hypothesis (μ) by 30 points. What is the probability of M differing from μ by 30 or more points? The sampling distribution of M is known to have a mean of μ and a standard error of

$$\sigma_M = \frac{\sigma}{\sqrt{n}}$$

For the present example, the mean of the sampling distribution is 500 and the standard error is: 100/7.07 = 14.14. The sampling distribution, then, looks like this:
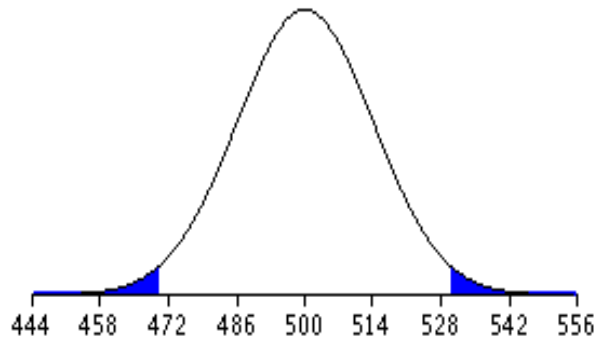
The mean is 500 and each tick mark represents 1 standard deviation (14 points). The shaded area is the portion of the distribution 30 or more points from μ. Therefore, the proportion of the area that is shaded is the probability of M differing from μ by 30 or more points? This area can be calculated using the methods described in the section "Area under portions of the normal curve." The area below 470 is 30/14.14 = 2.12 standard deviations below the mean. Therefore, the z score is -2.12.

A z table can be used to find that the probability of a z less than or equal to -2.12 is 0.017. Since the normal distribution is symmetric, the area greater than or equal to 530 is also 0.017. Therefore the probability of obtaining a sample mean 30 or more points from μ (M either ≤ 470 or ≥ 530) is 0.017 + 0.017 = .034. So, the probability that M would be as different or more different from 500 than the difference of 30 obtained in the experiment is 0.034. Or, more compactly: p = .034.