Chapter 6

Hypothesis Testing

The second general type of statistical inference, after estimation, is hypothesis testing OR significance testing, which is used to determine whether an observed (sample) difference provides enough evidence for us to believe that there is a corresponding difference in the underlying population.

What is a hypothesis ?

A statistical hypothesis is a statement, assertion or claim about the nature of a population. It is a basic object of any experimental Inquiry.

Numerous examples can be cited where a person hypothesizes about a situation. Anyone who has watched commercial tv cannot fail to be aware of the constant barrage of claims. Brand X detergent will wash white clothes sparkling white; with a supergas, will get more miles to the gallon than before; a new radial tire will give more than 40,000 miles of trouble-free driving; and so on

In the framework of statistical investigation, very often a hypothesis takes the form of stipulating the values of the unknown parameters of the population being studied.

For example :

(a) A coin is unbiased . If $P(H) = p$. the hypothesis is that $p = \frac{1}{2}$.

(b) A tire manufacturer claims that the new radial produced by his company will give more than 40.000 miles of trouble-free driving. If μ represents the mean number of miles, the hypothesis that the manufacturer has in mind is $\mu > 40,000$

(c) A social scientist asserts that students with an urban background perform better than those with a rural background. Suppose the measure used is the percent of students who graduate from high school. If P_U and P_R represent, respectively, the population proportions of students with urban and rural background who graduate, then the hypothesis is that $P_U > P_R$

(d) An engineer claims that the machine manufactures bolts within specifications. If she uses the standard deviation of the diameter of the bolts as a measure the hypothesis may be that $\sigma < 0.1$ inch.

The only sure way of finding the truth or falsity of a hypothesis is by examining the entire population. Since this is <u>NOT</u> always feasible, instead we might examine a sample for the purpose of drawing conclusions

A hypothesis that is being tested for the purpose of possible rejection is Called a NULL hypothesis. It is a common convention to designate this hypothesis as Ho.

If, for example, the null hypothesis is that the Coin is a fair coin, letting P(head) = p, we will write

$$H_0 : p = \frac{1}{2} .$$

As another example, the null hypothesis that the mean life of bulbs produced by a machine is at least 700 hours will be written as

$$H_0 : \mu \geq 700 .$$

Also, in our discussion of the topic, we will always state the null hypothesis in such a way that it contains the equality sign. Thus we will never have a null hypothesis stating that the mean life of bulbs is greater than 700, that is, we will not state

$$H_0 : \mu > 700$$

The hypothesis against which the null hypothesis is tested is called the alternate hypothesis. We shall denote it as Ha.

The values of the parameter(s) stated under the alternative hypothesis are outside the region stated under the null hypothesis.

That is, there is NO overlap between the set of parameter values stipulated under the null hypothesis and those stipulated under the alternative hypothesis. For example, against the null hypothesis $H_0 : p = \frac{1}{2}$ stated above, we might have as an alternative hypothesis $H_A : p \neq \frac{1}{2}$, or $H_A : p = \frac{3}{4}$, or $H_A : p > \frac{2}{3}$. However, against $H_0 : p = \frac{1}{2}$ we will never have

$H_A : p > \frac{1}{4}$.

<u>Choice</u> of H_0 and H_A

In a given problem, which hypothesis constitutes the null hypothesis and which one the alternative hypothesis is an important question and one has to do with the logic of a statistical test procedure.
In formulating tests of hypotheses, the essential idea is that of proof by contradiction.
For examples, suppose a teacher strongly suspects that students from families at a higher economic level perform better than students at the lower level.

This is what the teacher believes she can establish & hence the onus of proof is on her. Under Ho we shall say that this is NOT the case and state

Ho: Students at a higher economic level do not perform better

Against this null hypothesis we shall give the alternative hypothesis as

HA: Students at a higher economic level do perform better

It is for this reason that the alternative hypothesis is sometimes refered to as the research hypothesis

e.g. A businessman wonders: "Are the sales this year going to be better than last year? Here we write

Ho: The sales are NOT going to be better than last year.

HA: The sales will be better than last yea

Hypothesis testing involve assessing to what extant the sample data are consistent with two competing hypothesis about the underlying population.
The Sample data are the known but incomplete evidence; the hypotheses are competing claims about the TRUE but UNKNOWN state of reality.

TWO Hypothesis

The two hypotheses in hypothesis testing are known as the null hypothesis and the alternate hypothesis. Both hypotheses are always about unknown population values, never about sample data.

Def^n: The NULL hypothesis (H_o) is an assumption concerning the value of the population parameter being studied.

Def^n: The Alternate hypothesis (H_A) specifies all possible values of the population parameters that are NOT specified in H_o

e.g. $H_o: \mu_{hyper} = \mu_{control}$ μ_i: mean ingestion of Calcium

We evaluate the conditional probability of obtaining a sample difference at least as large as the one actually obtained, given the null hypothesis (H_o). i.e. $P\{\text{such a large sample diff.} \mid H_o\}$.

If we find that the sample data are sufficiently unlikely to have occured in the situation proposed by H_o, we reject H_o. In this case, the sample difference is said to be statistically significant.

if we find that the sample data are <u>reasonably</u> likely to occur in the situation proposed by H_0, we do <u>NOT</u> reject H_0 & we attribute the sample difference to chance. This is the case in which the sample difference is said to be <u>NOT</u> statistically significant!

Notice that hypothesis testing is carried out like a game of "let's pretend": Let us suppose, that the H_0 is true; then we can see what the chances would be of getting a sample as unusual as the one that actually occured. Hypothesis testing does <u>NOT</u> tell us how likely it is that the null hypothesis is true, but rather how likely our sample data would be if the null hypothesis <u>were true</u>. That is, hypothesis testing does <u>NOT</u> evaluate $P(H_0)$, but rather $P(\text{such a large sample difference} \mid H_0)$. From that information, we make a judgement as to whether H_0 is sufficiently inconsistent with the sample data ~~to~~ for us to reject H_0 in favor of an H_A.

<u>Def^n</u>: <u>Test statistic</u> <u>T</u> is a random variable whose value is used to determine whether we reject H_0

<u>f^n</u>: <u>Decision rule</u>
The decision rule specifies the set of values of the test statistic for which H_0 is rejected in favor of H_A & the set of values for which H_0 is accepted (NOT rejected).

Defn: Rejection Region and acceptance Region
The rejection region of a test, also called the critical region consists of all values of the test statistic for which Ho is rejected. The acceptance region consists of all values of the test statistic for which Ho is accepted (NOT rejected).

Defn: Critical Value
A critical value of the test statistic is the value that separates the critical region from the acceptance region.

Four possible outcome of a hypothesis test:

	Actual Situation	
Decision	H_0 True	H_0 False
H_0 NOT rejected	correct decision	Type II error
H_0 rejected	Type I error	Correct decision

Defn: Type I error : {Reject H_0 when H_0 true}

Type II error : DO NOT Reject H_0 when H_0 false.

Defn: Level of significance.
The Level of sig. of a test is the probability that the test statistic falls in the critical region given that H_0 is true. It is denoted by α where

$$\alpha = P(\text{Type I error}) = P(H_0 \text{ rejected} \mid H_0 \text{ true}\}$$
$$= P(\text{test statistic in rejection region} \mid H_0 \text{ true}\}$$

4th Probability of Type II error

It is denoted by β. The probability β of making a Type II error is the probability that the test statistic falls in the acceptance when H_0 is false. i.e

$$\beta = P\{\text{test statistic in accep- regn} \mid H_0 \text{ false}\}$$

e.g. Example of Type I error: In 1987 USA Today reported that a secret government survey found that illegal gambling amounts to at least \$200 per year per adult in US. Suppose you feel that this figure is too high. You take a random sample of n individuals to estimate the mean amount of illegal gambling per person. You want to test:

$$H_0: \mu \geq 200$$
$$H_A: \mu < 200$$

Suppose in fact, H_0 is True but you obtain a sample $\bar{x}$ that is much lower than \$200. Based on the sample information. you rejected H_0

We may find the following analogy clarifies the nature of hypothesis testing. The analogy likens a significance to a criminal court trial:

ⓐ H_0: innocent
H_A: guilty

(b) the prosecutor typically seeks to demonstrate Ho should be rejected!

(c) The evidence is evaluated to see if it is consistant with the Ho (asscumption of innocence

(d) Statistical significance corresponds to verdict of guilty; lack of statistical significance, to a verdict of NOT guilty.

(e) Type I error – conviction of an innocent person

(f) Type II error – acquittal of a guilty person.

(g) Both the significance test and the Criminal trial involve a judgement about the TRUE state of affairs on the basis of less than complete information.

Def^n: Simple hypothesis and Composite Hypothesis

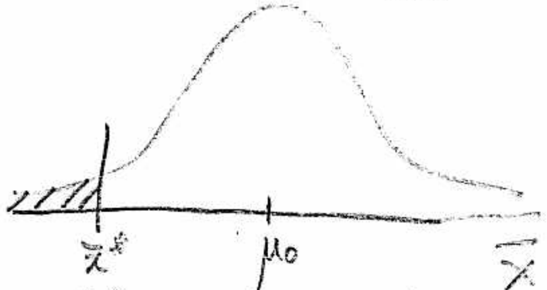
If a hypothesis states that a certain population parameter θ equals a single specific value, suchas $\theta = \theta_0$, then the hypothesis is said to be a simple hypothesis
This null hypothesis can be expressed as: $H_0: \mu = \mu_0$
When the hypothesis contains a range of possible values for the parameter of interest, then the hypothesis is said to be a composite hypothesis. e.g. $H_0: \theta \geq \theta_0$ or $H_0: \theta \leq \theta_0$

One-sided " test: : $H_0: \theta = \theta_0$, $H_A: \theta > \theta_0$; $H_0: \theta \geq \theta_0$, $H_A: \theta$
two-sided " test : $H_0: \theta = \theta_0$, $H_A: \theta \neq \theta_0$

ing hypothesis about the population mean when variance is known

(1) $H_0: \mu = \mu_0$
$H_A: \mu < \mu_0$

It is natural to base a test concerning the population mean on the sample mean $\bar{x}$. If the population is normal (or if sample size is large). Then when the null hypothesis H_0 is true, the random variable $\bar{X}$ has a normal distribution of mean μ_0 and variance $\frac{\sigma^2}{n}$ (by CLT)



We would reject H_0 in favor of H_A only if the observed mean $\bar{x}$ is substantially less than μ_0.

We would reject H_0 in favor of H_A only if the observed sample mean fell in the left-hand tail of the distribution to the left of a critical value, denoted by $\bar{x}^*$. For this test, the rejection region is the set of all values of the sample mean less than $\bar{x}^*$, and the acceptance region is the set of all values of the sample mean greater than or equal to $\bar{x}^*$.

When we test $H_0: \mu = \mu_0$,
$H_A: \mu < \mu_0$, we use the decision rule:

Reject H_0 in favor of H_A if and only if the observed sample mean $\bar{x}$ is less than the critical value $\bar{x}^*$

$X \sim N(\mu_0, \sigma^2)$ or n is large enough, we have $\bar{X} \sim N(\mu_0, \frac{\sigma^2}{n})$

$$\Rightarrow z = \frac{\bar{X}-\mu_0}{\left(\frac{\sigma}{\sqrt{n}}\right)} \sim N(0,1)$$

If the area in the left tail of the distribution is α, we have

$$P(Z < -z_\alpha) = \alpha$$

When H_0 is true, we obtain:

$$P\left(\frac{\bar{X}-\mu_0}{(\sigma/\sqrt{n})} < -z_\alpha\right) = \alpha$$

$$> P\left(\bar{X}-\mu_0 < -z_\alpha \frac{\sigma}{\sqrt{n}}\right) = \alpha$$

$$> P\left(\bar{X} < \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}}\right) = \alpha$$

Thus the critical value $\bar{x}^* = \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}}$

Whenever the observed mean $\bar{x}$ is less than $\bar{x}^*$, the observed Z score

$$z = \frac{\bar{x}-\mu_0}{(\sigma/\sqrt{n})}$$ is less than $-z_\alpha$, and vice versa.

Thus, the rule that we should reject H_0 if and only if the observed sample mean $\bar{x}$ is less than the critical value $\bar{x}^*$ is equivalent to the decision rule that we should reject H_0 if and only if the observed Z score is less than $-z_\alpha$

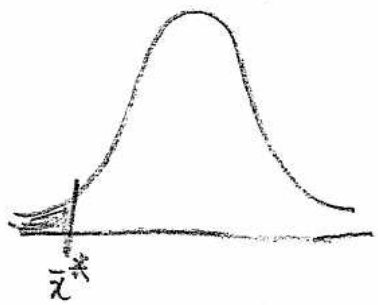
e.g. 1. Suppose a population is normally distributed with $\sigma^2 = 36$.

(a) Based on a sample of 16, describe a suitable test procedure at the 10% level of significance to test $H_0: \mu = 50$ against $H_A: \mu < 50$

(b) If a sample of 16 yields a mean 48.5, what decision would be appropriate at the 10% level of significance?

Soln :

(a) $H_0: \mu = 50$, $H_A: \mu < 50$, $\alpha = 0.10$



$$z_\alpha = z_{0.1} = 1.28$$

$$\sigma^2 = 36, \ \sigma = 6, \ n = 16$$

$$\frac{\bar{x} - \mu_0}{(\sigma/\sqrt{n})} = \frac{\bar{x} - 50}{(6/\sqrt{16})} = \frac{\bar{x} - 50}{1.5}$$

The critical value $\bar{x}^* = 50 - 1.28 \times 1.5 = 48.08$

reject H_0 if $\bar{x} < \bar{x}^* = 48.08$

or equivalently reject H_0 if $\frac{\bar{x} - 50}{1.5} < -1.28$

(b) The sample mean is $\bar{x} = 48.5$, the computed value of the test statistic is

$$\frac{\bar{x} - 50}{1.5} = \frac{48.5 - 50}{1.5} = -1$$

Since $\bar{x} = 48.5 > 48.08$, we do NOT reject H_0 or equivalently $\frac{\bar{x} - 50}{1.5} = -1 > -1.28$ we do NOT reject H_0 at 0.1 lev e.g. An auto dealer believes that his new model will give mean trouble-free service of at least 12000 miles. In a simulated test with 4 cars, the following numbers of trouble-free miles were obtained:

11000 12000 11800 11200

Test the null hypothesis that the mean mileage is at least 12000 miles against the alternative hypothesis that it is not. Use $\alpha = 0.05$ (Assume a normal distribution with known $\sigma = 476$.

Solution: $H_0: \mu = 12000$ (actually $\mu \geq 12000$)
$H_A: \mu < 12000$

$$z_{obs} = \frac{\bar{x} - 12000}{\sigma/\sqrt{n}} = \frac{11500 - 12000}{(476/2)} = -2.1008$$

$$-z_{0.05} = -1.645$$

$-2.353 < -1.645$, therefore reject H_0 at 0.05 level.
OR equivalently, the critical value $\bar{x}^* = 12000 - 1.645 \times \frac{47}{\sqrt{4}}$
$= 11608.49$
$\bar{x} = 11500 < 11608.49$. therefore reject H_0 at .05 level

(II) $H_0: \mu = \mu_0$
$H_A: \mu > \mu_0$

Reject H_0 in favour of H_A if $Z > Z_\alpha$ OR equivalently Reject H_0 in favour of H_A if $\bar{x} > \bar{x}^*$ where the critical value

$$\bar{x}^* = \mu_0 + Z_\alpha \frac{\sigma}{\sqrt{n}}$$

e.g. A car salesman claims that a particular model of car would give a mean mileage of greater than 20 miles per gallon. To test the claim, a field experiment was conducted where 10 cars were each run on a gallon of gasoline. The results (in miles) were:

23, 18, 22, 19, 19, 22, 18, 18, 24, 22

Is the salesman's claim justified? Use $\alpha = 0.05$ and assume normal distribution with $\sigma = 2.32$.

Solution: 1. $H_0: \mu = 20$
$H_A: \mu > 20$

$$Z_{obs} = \frac{\bar{x} - 20}{\sigma/\sqrt{n}} = \frac{20.5 - 20}{2.32/\sqrt{10}} = 0.681$$

$Z_{0.05} = 1.645$, $0.681 > 1.645$ DO NOT reject H_0

or equivalently reject H_0 in favour of H_A if $\bar{x} > \bar{x}^*$ where $\bar{x}^* = 20 + 1.645 \times \frac{2.32}{\sqrt{10}} = 21.206$.

$\bar{x} = 20.5 < \bar{x}^* = 21.206$, therefore do NOT reject H_0 at 0.05 level.

e.g. Suppose a population is normally distributed with $\sigma^2 = 36$

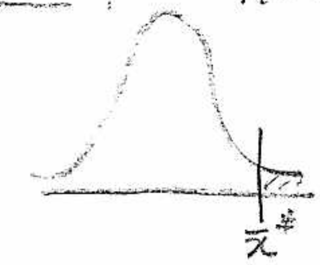
(a) Based on a sample size of 16, describe a suitable test procedure at 10% level of significance to

6.15

test $H_0 : \mu = 50$ against $H_A : \mu > 50$

(b) If a sample of 16 yields a mean of 53.0, what decision would be appropriate at 10% level of significance ?

Soln : $H_0 : \mu = 50$, $H_A : \mu > 50$



$\alpha = 0.10$, $\delta_\alpha = 1.28$, $\sigma^2 = 36$, $\sigma = 6$, n

$$\frac{\bar{x} - \mu_0}{(\sigma/\sqrt{n})} = \frac{\bar{x} - 50}{(6/4)} = \frac{\bar{x} - 50}{1.5}$$

The critical value $\bar{x}^* = 50 + 1.28 \times 1.5 = 51.92$

Reject H_0 if $\bar{x} > 51.92$ or equivalently reject H_0 if $\frac{\bar{x} - 50}{1.5} > 1.28$

(b) Sample mean is 53.0 which is great than 51.92 hence reject H_0 at 0.10 level

or

$$\frac{\bar{x} - 50}{1.5} = \frac{53 - 50}{1.5} = 2$$ which is greater than 1.28

$\therefore$ we reject H_0 at 0.10 level.

(III) $H_0 : \mu \geq \mu_0$, $H_A : \mu < \mu_0$

follow the same procedures as in (I)

(IV) $H_0 : \mu \leq \mu_0$, $H_A : \mu > \mu_0$

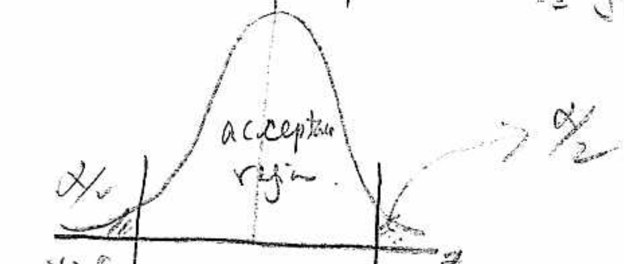
follow the same procedue as in (II)

1. $H_0: \mu = \mu_0$, $H_A: \mu \neq \mu_0$
for a selected α, find $-z_{\alpha/2}$, $z_{\alpha/2}$
Obtain the sample of size n,
calculate $z = \frac{\bar{x} - \mu_0}{\left(\sigma/\sqrt{n}\right)}$

Reject H_0 in favor of H_A if $z < -z_{\alpha/2}$ or if $z > z_{\alpha/2}$

An equivalent rule is
Reject H_0 if $\bar{x} < \bar{x}_1^*$, or if $\bar{x} > \bar{x}_2^*$

Where $\bar{x}_1^* = \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, $\bar{x}_2^* = \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$



e.g. A manufacturer of sports equipment has developed a new synthetic fishing line that he claims has a mean breaking strength of 8 kg with a s.d. of 0.5 kg. Test the hypothesis that $\mu = 8$ kg against the alternative that $\mu \neq 8$ if a r.s. of 50 lines is tested and found to have a mean breaking strength of 7.8 kg. Use 0.01 level of significance.

Soln: $H_0: \mu = 8$ kg $H_A: \mu \neq 8$ kg $\alpha = 0.01$
Critical region: $z < -2.575$ and $z > 2.575$ where
$$z = \frac{\bar{x} - \mu_0}{\left(\sigma/\sqrt{n}\right)}$$
$n = 50$, $\bar{x} = 7.8$ kg, $z = \frac{7.8 - 8}{\left(.5/\sqrt{50}\right)} = -2.83$

$-2.83 < -2.575$ ∴ reject H_0 & conclude that the average breaking strength is not equal to 8 but is, in fact less than 8 kg.

or equivalently,

$$\bar{x}_1^* = 8 - 2.575 \times \frac{0.5}{\sqrt{50}} = 7.818$$

$$\bar{x}_2^* = 8 + 2.575 \times \frac{0.5}{\sqrt{50}} = 8.182$$

rejection region: $\bar{x} < 7.818$, $\bar{x} > 8.182$

Since $7.8 < 7.818$ so reject H_0.

Testing hypothesis about the population mean when variance is unknown.

Same as in (I, p6.11), (II, p6.14), (III, p6.15), (IV, p6.15), (V, p6.16) except replace σ by s and z_α by $t_\alpha(n-1)$.

Def^n P-value

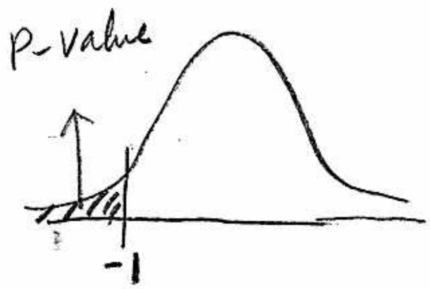
The p-value of a test is a probability of obtaining a value of the test statistic as extreme or more extreme than the observed sample value when the hypothesis H_0 is true.

A p-value answers the question, "If H_0 is true, what is the probability that a random sample will yield the observed test statistic or one whose value is further from the expected value?". If the probability is small, then we reject H_0 by concluding discrepancy is too large to be explained by chance alone. The observed difference is said to be statistically significant. If the probability is NOT small then we do NOT reject Ho, because the observed discrepancy may well be due to chance, that is sampling error.

The p-value is the smallest level of significance α at which Ho can be rejected. The smaller the p-value the more doubt is cast on the validity of Ho

e.g. Go back ~~(6.10)~~ (p10) (6.13) example

$$\text{the test statistic} = \frac{\bar{x}-50}{\sigma/\sqrt{n}} = \frac{48.5-50}{1.5} = -1$$



P-value is the prob. of obtaining a Z score more extreme than the observed Z score of -1.

We have $p\text{-value} = P(Z < -1) = 0.5 - 0.3413$
$= 0.1587$

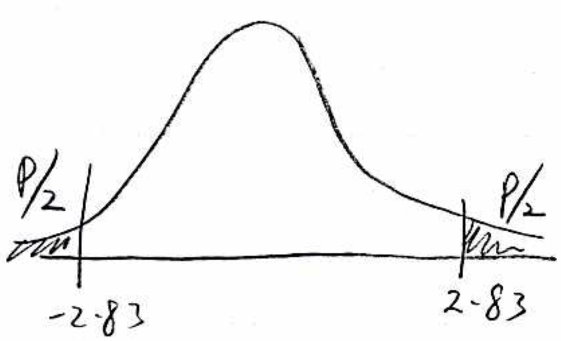
This p-value indicates that if Ho is true, then in repeated samples we should obtain a Z score as small or smaller than the one actually observed 158 ~ 159 times out of 1000 times. i.e. 0.1587.

In statistical decision making, a p-value provides more information than a report of whether the observed test is statistically significant at some preselected level of significance α. e.g. p-value = .03 which is significant at 0.05 level but not at 0.01 level

6.19

For 2-sided test if the sampling distribution of the test statistic under H_0 is symmetric then the p-value is simply 2 × p-value for 1-sided test.

e.g. use example in p. 6.16



$$p\text{-value} = P(|Z| > 2.83)$$
$$= 2 \times P(Z < -2.83)$$
$$= 0.0046$$

$p < \alpha = 0.01$ reject H_0.

j. Some theorists have suggested that the babies who are heavier at birth have a better chance of developing high intelligence than those with lower birth weights. If the theory is correct, then we might expect to find the individuals who are pursuing higher education had higher birth weight than the general population.
A sample of 16 female college students reported their birth weights. The recalled birth weights had a mean of 3.331 kg and a standard deviation of .333 kg. Official community health records show the mean female live birth weight to be 3.25 kg for the region from which the college draws its students. Let us test the null hypothesis that the $\mu_0 = 3.25$ kg against the alternative hypothesis that the population mean μ for female college students is not

6·20

equal to the female regional average of $\mu_0 = 3.25$ Kg.

$H_0 : \mu = 3.25$
$H_A : \mu \neq 3.25$

Soln :

If Ho is true, the observed value of the test statistic is

$$t_{obs} = \frac{3.331 - 3.25}{\left(\frac{.333}{\sqrt{16}}\right)} = 0.97$$

If Ho is true, $t_{obs} = 0.97$ may be regarded as a random selection from the t distribution with n-1= 16-1 = 15 degrees of freedom, whose .05-level critical values are $\pm t_{0.025}(15\ df) = \pm 2.131$.
Since $t_{obs} = 0.97$ falls between the critical values. we do NOT reject Ho at 0.05 significance level. p-value = 2 × 0.1737 = 0.3474

The data are consistent with the notion that these college students came from a population whose birth weight was 3.25 kg; the observed sample was NOT significantly different from the hypothesized value of 3.25 ($p = .3474$,). Of course, it could be that the students' recollections of their birth weights were inaccu But even if the recollections were perfectly accurate, we have not proved that the null hypothesis is true; perhaps, for example a larger sample would have given us a sufficiently precise estimate of college-student birth weights so that we have distinguished it from the hypothesized value of 3.25 kg.

A sample difference that is NOT statistically