2.

air pollution training institute

introduction to environmental statistics

self-instructional course

SI 473

MODULE II

HYPOTHESIS TESTING

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY

UNIT ONE

UNIT 1 - HYPOTHESIS TESTING

1

This unit introduces you to hypothesis testing. Hypothesis testing is the topic that makes the use of statistics so intriguing, useful, and necessary. In hypothesis testing, the statistician makes a statement (hypothesis) about an unknown population parameter, and then uses statistical methods to determine (test) whether the observed sample data support that statement.

The important notion to understand in this unit is the relationship of the null hypothesis to the alternative hypothesis (or hypotheses).

9a

Answer: null hypothesis is $H_o : \mu = 50$

alternative hypothesis is $H_1 : \mu \neq 50$

10

In all cases the null hypotheses and alternatives are written in terms of population parameters. You'll get to see several more types of hypotheses in later units.

2

As you recall, statistics are characteristics of samples and can be used as ____________________ of population characteristics.

11

Let's review:

A statistic can be used to ________________ a population ________________.

2a

Answer: estimates

3

Since these statistics are only estimates of a "true" population parameter based on a sample of observations, it is reasonable to expect that the value of the estimate will vary from the value of the "true" population parameter. To decide whether the value of the statistic is consistent with an hypothesized value for the population parameter the statistician uses a method called * ______________________.

(Hint: see title of this unit)

11a

Answer: estimate parameter (or characteristic)

12

When such an estimate is made, a(n) ______________ is made concerning an assumed population parameter.

3a

Answer: hypothesis testing

4

First, let's define hypothesis. A hypothesis is a conjectural statement about one or more parameters. Check which of these would be synonyms for "hypothesis":

____ Assumption

____ Claim

____ Guess

12a

Answer: hypothesis

This hypothesis must then be ______________ to verify its credibility. 13

4a

Answer: all three

x Assumption

x Claim

x Guess

The point is, you don't have to have much information to form a hypothesis. Any testable, conjectural statement could be a hypothesis.

5

In hypothesis testing, the statistician uses various techniques to decide whether to accept or reject hypotheses. If, for example, someone assumed a population mean was 50, and a sample from that population was selected that had a mean of 60, he might want to perform a test to see if the population mean of 50 could still be assumed. The statistician would use a statistical technique to ____________ or ____________ the hypothesis that the population mean equals 50.

13a

Answer: tested

14

The hypothesis being tested is the ____________ ____________.

5a

Answer: accept

reject

6

Statisticians have developed their own language to talk about hypothesis testing. For example, in our situation of a sample mean equal to 60 being used to test a hypothesized population mean of 50, the statistician would say it a little differently. Let's put the situation into symbols. Symbolically, the situation looks like this:

$\bar{x} = 60$

and the hypothesis is:

$\mu = 50$ (μ is a Greek letter, pronounced "mew")

Which statement refers to the sample? ____________

Which statement refers to the population? ____________

14a

Answer: null hypothesis

15

Each null hypothesis is accompanied by a(n) ____________ ____________.

6a

Answer: $\bar{x} = 60$ (Sample)

$\mu = 50$ (Population)

7

The hypothesis being tested in this instance can be written:

$$\mu = 50$$

This hypothesis has become known as the null hypothesis, "null" meaning there is no difference between the hypothesized value (50) and the true value (μ).

The null hypothesis is often written:

$$H_o : \mu = 50$$

15a

Answer: alternative hypothesis

16

When a null hypothesis is rejected, an alternative hypothesis is ____________________.

It is the null hypothesis that the statistician wishes to accept or reject. In rejecting a null hypothesis, such as $H_o : \mu = 50$, the statistician is at the same time accepting another hypothesis (in this case, $\mu \neq 50$). (The symbol $\neq$ means "does not equal.") This second hypothesis is often called the alternative hypothesis. The alternative hypothesis is usually written:

$$H_1 : \mu \neq 50$$

16a

Answer: accepted

17

For any null hypothesis, one of several alternative hypotheses could be chosen.

For example:

if $H_o : \mu = 60$

alternatives include:

$H_1 : \mu \neq 60$

or

$\mu > 60$ (μ is larger than 60)

or

$\mu < 60$ (μ is smaller than 60)

It is important when testing hypotheses to know both the ____________ ________________ and the ________________ ________________ because they will determine which type of statistical test will be used.

In our example, then, we have:

$$H_o : \mu = 50$$

$$H_1 : \mu \neq 50$$

Which is the null hypothesis? ____________

Which is the alternative hypothesis? ____________

Go to:

Frame 9a

Page II-1

17a

Answer: null hypothesis
alternative hypothesis

18

The concept of hypothesis testing is basic to the realm of statistics, for it is through the acceptance or rejection of the null hypothesis that the statistician can infer certain properties about the population from which a sample is taken. The null hypothesis is an assumption made about a population. It is the statistician's job to determine if that assumption should be retained, based on a statistic computed for a sample of that population. The remainder of this module is devoted to explaining:

1. How to determine what the null hypothesis is.
2. How to test the null hypothesis.
3. How to determine the probability of rejecting a true hypothesis and accepting a false hypothesis.

Go to Unit 2.

UNIT 2 - HYPOTHESIS TESTING: THE MECHANICS

Unit 1 of this module introduced the concept of hypothesis testing. This unit will continue with that concept by teaching you the mechanics of how the testing is accomplished. The more important aspects of hypothesis testing for which you should be prepared include:

1. The setting of significance levels.
2. The determination of critical values.
3. The determination of confidence intervals.
4. The way in which the above three can be used to increase the frequency of correct acceptance or rejection of the null hypothesis.

These four points will be presented using a statistical test called the "*z* test." The *z* test is a convenient example for teaching purposes. Toward the end of the unit, however, other tests will be briefly discussed (t, *ANOVA*, F, and χ^2).

Let's use this formula:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \qquad \bar{x} = 29$$

$$\mu = 25$$

$$= \frac{29-25}{12/\sqrt{36}} \qquad \sigma = 12$$

$$n = 36$$

$$= 2.00$$

By again consulting the z table, we find that # ______________ of the area under the standard normal distribution curve falls between $z = 2.00$ and the mean. [Some of you have probably noticed that we used the sample standard deviation (s) in the formula instead of a population standard deviation (σ). The reason for this is that when n is large (>30) the sample standard deviation is a good estimate of the population standard deviation. In this case, $n = 36$, so we can use our $s = 12$ in the formula.]

2

In order to continue with hypothesis testing, a brief review of the properties of the normal distribution is necessary since most of the statistical tests we will be dealing with require you to assume that the population from which you are drawing your sample is normally distributed.

A normal distribution is bell-shaped and (symmetrical/asymmetrical).

31a

Answer: .4772 (47.72%)

32

Since .4772 of the area under the curve falls between $z = 2.00$ and the mean, then # __________ (.5000 - .4772) falls to the right of $z = 2.00$.

2a

Answer: symmetrical

3

The mean of a population distribution is symbolized by ________. (Greek letter)

Its standard deviation is symbolized by ____. (Greek letter)

32a

Answer: .0228 (or 2.28%)

33

This means that if $\mu = 25$, we would expect to obtain $\bar{x} = 29$ or greater only 2.28% of the time. This percentage is well below the 5% level commonly set by statisticians as the level of significance. We therefore would reject:

$$H_o: \mu = 25$$

on the basis that we would expect to obtain sample data like those we collected less than 5% of the time.

We could say that we have rejected $H_o: \mu = 25$ at the 5% level of significance.

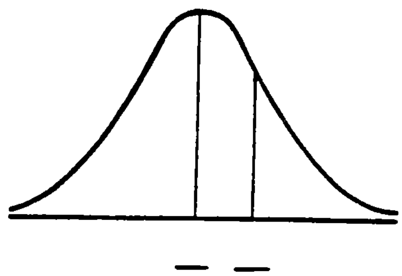
3a

Answer: μ

σ (sigma)

4

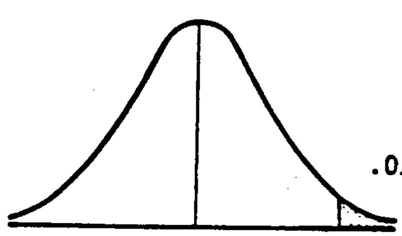
Place the symbols μ and +1σ in their proper positions on the figure below:



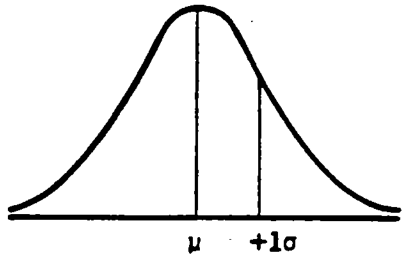
34

In normal practice, the statistician will obtain from the z table the value of z he would have to exceed to reject H_o. Using the 5% level, we want to determine what z value corresponds to .0500 of the curve falling to the right of the score.

.5000 - .0500 = # ____________

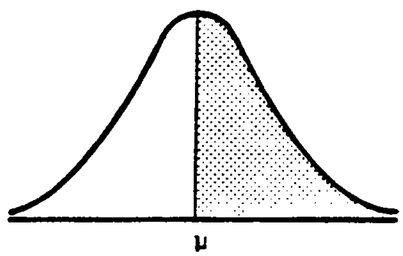


Answer:



5

Since the normal distribution is symmetrical, with the mean as its mid-point, what percent of the distribution is to the right of the mean? #_______%



34a

Answer: .4500

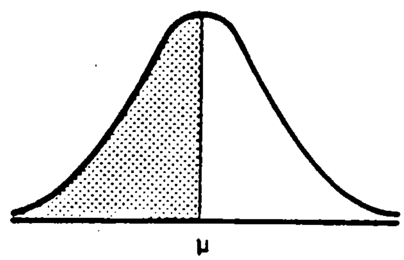
35

Go to the *z* table again. In the body of the table find the figure closest to .4500. The *z* value corresponding to that number is #________.

Answer: 50%

Okay, let's see if you are still awake. 6

What percent lies to the left of mean? #_______%



35a

Answer: $z = 1.64$. (Through interpolation the score is 1.645. A rule of rounding suggests that since 4 is an even digit followed by a 5, the accepted way to round is to the even digit. We, therefore, obtain 1.64. If the score was 1.655 we would again round to the even digit or 1.66. This is sometimes difficult to get used to after always hearing "five or more ? round up.")

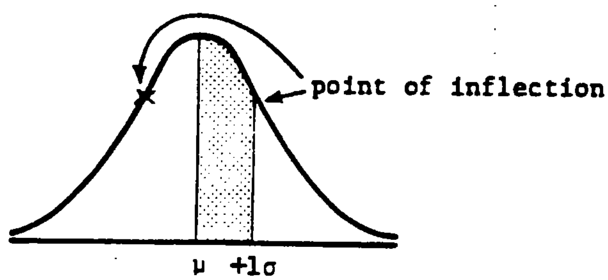
36

In our example, then, if the z value we obtain is greater than 1.64, we reject the null hypothesis (since a score of 1.64 cuts off .05 of the curve in that tail).

Answer: 50%

7

In Module I, Unit 3 you were told that 68% of the values fall between the two points of inflection. Remembering that the curve is symmetrical, what percent of the distribution is in the shaded area below? # ______%



37

In the example we just went through, we determined the:

1. null hypothesis.
2. alternative hypothesis.
3. significance level (α "alpha").
4. value that needed to be exceeded in order to reject the null hypothesis.
5. result leading to rejection.

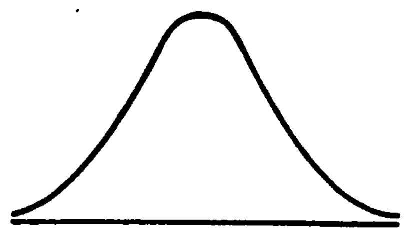
[Let's look for a moment at number 3 above. The significance level (often written α = .05 or α = .01) is the level selected by the statistician that determines the cut-off point for statistical significance. This will be discussed in greater detail in the following units.]

7a

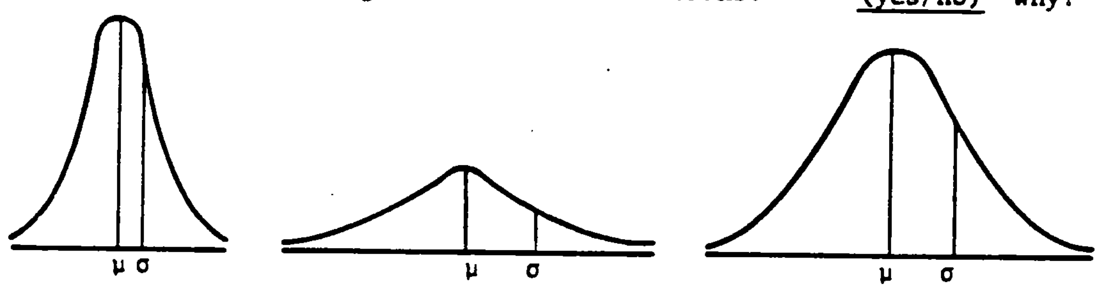
Answer: 34%

8

So far the normal distribution has taken the form of:



Can each of the following be normal distributions? (yes/no) Why?



†____

____

38

Now let's consider number 4 in the preceding frame--the value that needed to be exceeded in order to reject the null hypothesis. In deciding what value needed to be exceeded to reject the null hypothesis, given $\alpha = .05$, you (as the statisticians put it) determined the critical value.

Answer: yes (they are all bell-shaped and symmetrical)

9

As it would be physically impossible to construct tables for each μ and σ value, the statistician likes to talk in terms of "standard units." Putting values into standard units converts the distribution to one with a mean of 0 and standard deviation of 1. Conversion to standard units requires the use of the formula:

$$z = \frac{x_i - \mu}{\sigma}$$

where z = the standard value

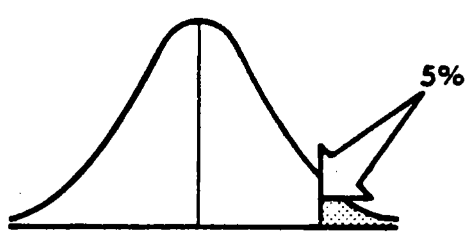
x_i = the ith data value

μ = population mean

σ = population standard deviation

39

The critical value is determined by the α selected, and the alternative hypothesis. In the example we had $\alpha = .05$ and $H_1: \mu > 25$.



The critical value was the z value above which point .05 of the values would fall:

$$z_{.05} = 1.64$$

10

Since this isn't a very complex formula, let's try a few quick ones.

Given: $\mu = 50$

$\sigma = 5$

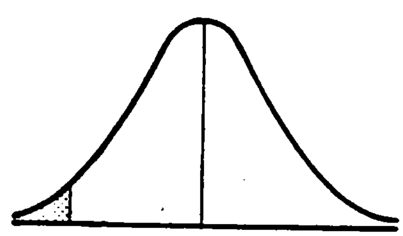
And the formula: $z = \frac{x_i - \mu}{\sigma}$

Calculate z (standard value) for each of the following:

40 45 55 60 62 50

40

If, however, $H_1: \mu < 25$ were used instead:



the critical value would be the z value below which .05 of the values would fall.

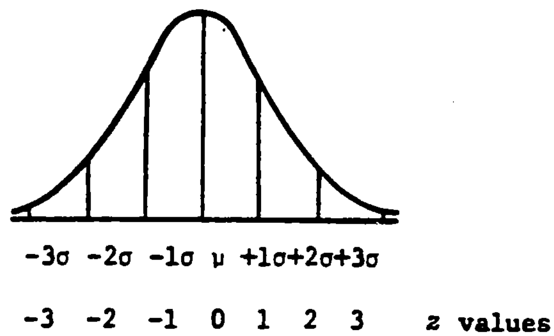
$$z_{.05} = -1.64$$

10a

Answer: -2, -1, +1, +2, +2.4, 0

11

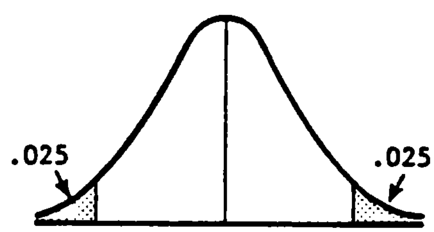
Changing scales on the normal distribution from standard deviation values to z values shows the relationships between them.



What would the z value for $+.5\sigma$ be? #______

41

If the $H_1: \mu \neq 25$, we would be concerned about extremes in both directions:



The critical value here would be the z value above or below which $\frac{.05}{2}$ of the values fall (since there are now two tails to equal the .05). In this case, if the z value obtained was greater than +1.96 or less than -1.96, the H_o would be rejected.

We found 1.96 the same way we found 1.64. Find the z value corresponding to .5000 - .0250 or .4750 in Table E: ________

11a

Answer: .5

12

By means of Table E in the Guide, we can find out what percent of the distribution falls between any z value and the mean.

What is the title of Table E? * ______________________________

41a

Answer: 1.96

42

Find the critical values given $\alpha = .01$ and

A. $H_o: \mu = 30$

$H_1: \mu < 30$ # __________

B. $H_o: \mu = 30$

$H_1: \mu > 30$ # __________

C. $H_o: \mu = 30$

$H_1: \mu \neq 30$ # __________

12a

Answer: The Standard Normal Distribution

13

Let's see how you go about finding this percentage in the table. Suppose we have:

$z = 1.96$

To find what percent of the values fall between $z = 1.96$ and the mean ($z = 0$) go down the left column to 1.9 and then across to the column headed by .06. At the intersection we find .4750. This means that .4750 (47.5%) of the values fall between $z = 1.96$ and the mean.

Use Table E to find the percentage between the mean and z values of .5, 2.54, 3.0.

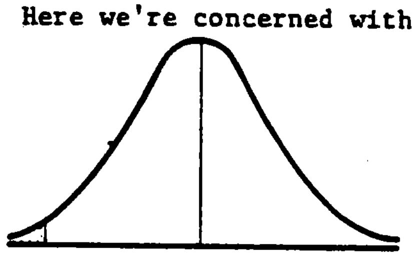
____________ # ____________ # ____________

42a

Answer:

Here we're concerned with values in the shaded area:

A:

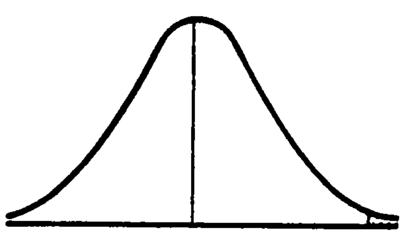


z = the point below which 1% of the values fall.

.5000 - .0100 = .4900

$z = -2.33$

B:

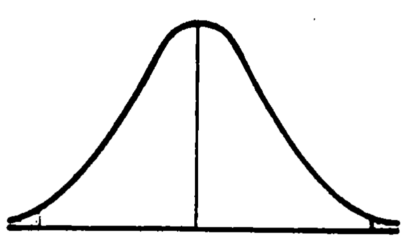


z = the point above which 1% of the values fall

.5000 - .0100 = .4900

$z = +2.33$

C:



z = the points above and below which .5% of the values fall

z corresponding to .5000 - .0050 = .4950

$z = \pm 2.58$

13a

Answer: .1915 (19.15%)

.4945 (49.45%)

.4987 (49.87%)

14

You may have noticed that there are no negative numbers in the table. What do you do if your z value is -1.96?

(Hint: remember the curve is symmetrical) † ______________________

____

43

For simplicity the statisticians have named the preceding tests.

Since the first two are concerned with only one tail of the distribution, they are called #____-tailed tests.

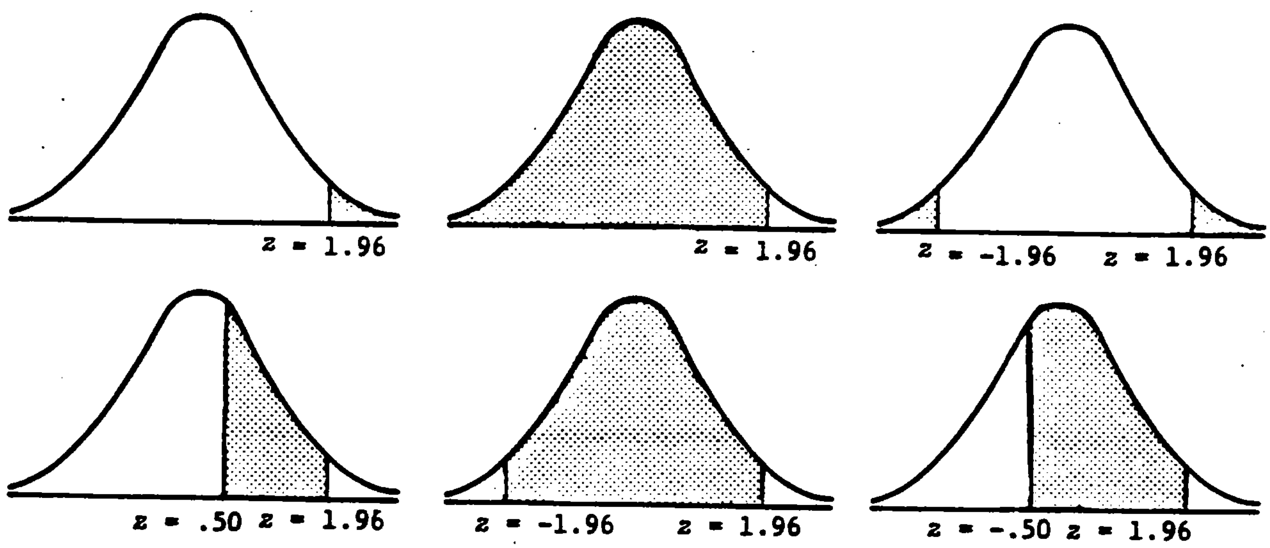
Since the third one is concerned with two tails of distribution, it is a #____-tailed test.

Answer: Look up the percentage corresponding to +1.96. 14a

The percentage for −1.96 is the same as for +1.96 so there was no need to include negative numbers in the table we already have. (It is important, however, to notate if the z value is positive or negative.)

15

In the following frames, we'll use some of the properties of the normal distribution and Table E to find the percent in the following shaded areas:



43a

Answer: one-tailed

two-tailed

44

Which of the following represent a:

one-tailed test? ____________

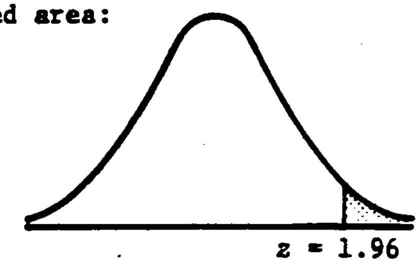
two-tailed test? ____________

A: $H_o: \mu = 100$
$H_1: \mu \neq 100$

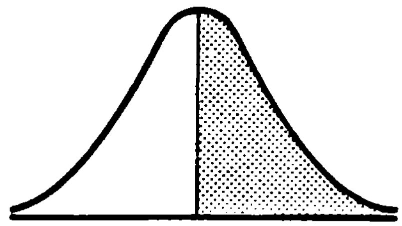
B: $H_o: \mu = 100$
$H_1: \mu > 100$

C: $H_o: \mu = 100$
$H_1: \mu < 100$

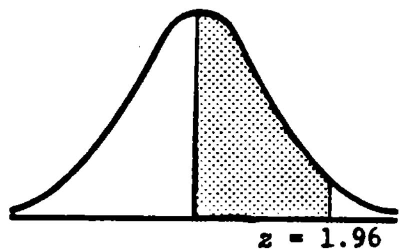
Find the % in the shaded area:



We know .50 of the values fall to the right of the mean:



From the table, .4750 falls between z = 1.96 and the mean



The remainder is equal to

$$\begin{array}{r} .5000 \\ -\ .4750 \\ \hline .0250 \end{array}$$

Therefore, .0250, or 2.5%, of the values fall to the right of $z = 1.96$.

Answer: one-tailed B & C 44a

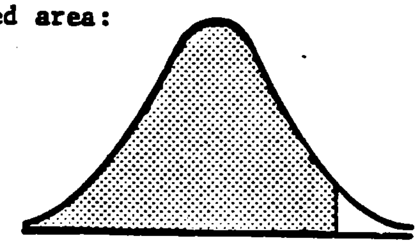
two-tailed A

45

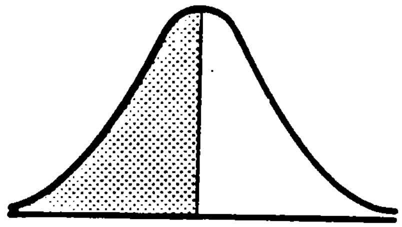
Once the critical values have been determined for a two-tailed test, the interval between them is known as the confidence interval or acceptance region. With $\alpha = .05$ the confidence interval is the interval between +1.96 and -1.96 (the critical values).

If $\alpha = .01$, what is the confidence interval? The interval between $z = +$#_____ and $z = -$#_____ (critical values).

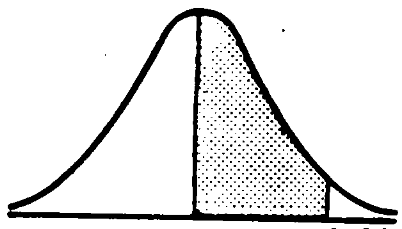
Find the % in the shaded area:



We know .50 of the values fall to the left of the mean.



From the table, .4750 falls between z = 1.96 and the mean.



Added together

$$\begin{array}{r} .5000 \\ +\ .4750 \\ \hline .9750 \end{array}$$

We find .9750 or 97.5%, to the left of 1.96.

45a

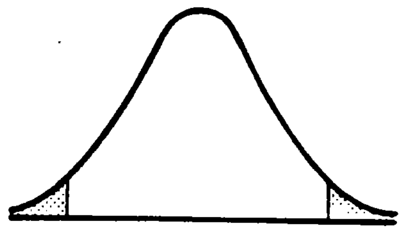
Answer: +2.58 -2.58

46

With $\alpha = .05$, if the z value we obtain falls between ±1.96, we can say (because we are accepting H_o if our z falls within the confidence interval) with 95% confidence (100% - 5% level of significance) that μ does not differ significantly from the assumed value. With $\alpha = .01$. if our obtained z falls between ±2.58, our confidence increases to ________%.

18

By now you should be familiar with the idea, so we'll let you try a few. Determine what portion of the area under the curve falls in the shaded area.



$z = -1.96$ $z = 1.96$

Answer: ____________

46a

Answer: 99%

47

One method used to write the confidence interval looks like this:

$$-1.96 < z_{OBTAINED} < +1.96$$

which means if the z value obtained is greater than -1.96 and less than +1.96, then we are 95% confident that the mean does not differ significantly from the assumed value (H_o). Write the confidence interval with an $\alpha = .01$.

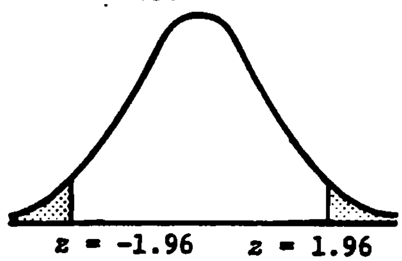
____

Answer: 5% of the area under the curve is in the shaded area.

Area between 1.96 and the mean = .4750

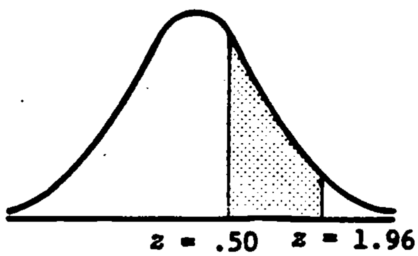
$$\begin{array}{r} .5000 \\ - .4750 \\ \hline .0250 \\ \times\ 2 \\ \hline .0500 \end{array}$$

.0250 = the area to right of 1.96 (since we have both ends)



19

Determine what portion of the curve falls in the shaded area:



Answer: #________

47a

Answer: $-2.58 < z_{OBTAINED} < +2.58$

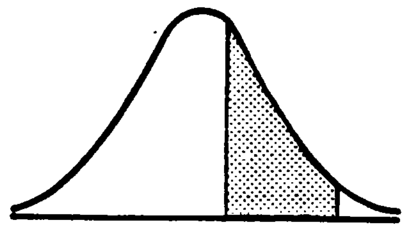
Usually the $_{OBTAINED}$ can be dropped, but you must be sure <u>you know</u> that's the z you have calculated.

Answer: 28.35%

Area between $z = 1.96$ and the mean equals .4750.

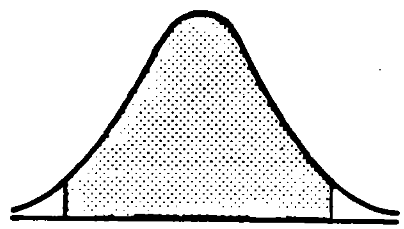
Area between $z = .50$ and the mean equals .1915.

$.4750 - .1915 = .2835$ is the area between $z = 1.96$ and $z = .50$.



20

Determine the portion of the area under the curve in the shaded area:



Answer: #________

48

In the following example, determine:

1. H_o : ____________
2. H_1 : ____________
3. α = #____________
4. critical values = #____________
5. confidence interval (-#______ < $z_{OBTAINED}$ < +#______)
6. decision to accept or reject H_o ____________

We have received data on 49 air samples yielding a mean SO_2 concentration of 60 $\mu g/m^3$ with a standard deviation of 14 $\mu g/m^3$.

Up to now we thought the mean equaled 55 $\mu g/m^3$, and we need to test this assumption against the hypothesis that the mean doesn't equal 55 at the 5% level of significance.

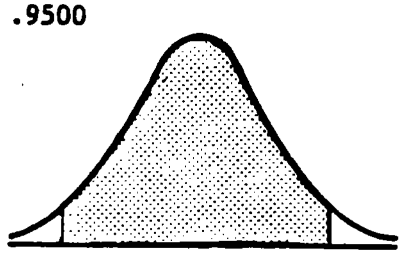
Use this formula: $z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

Answer: 95%

Area between $z = 1.96$ and the mean equals .4750

Since we want to know the area on both sides

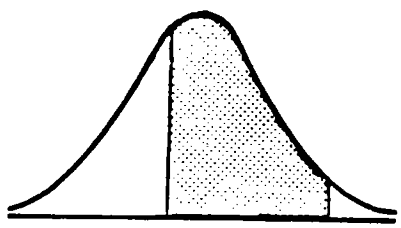
$.4750 \times 2 = .9500$



$z = -1.96$ $z = 1.96$

21

Determine the portion in the shaded area:



$z = -.50$ $z = 1.96$

Answer: #________

48a

Answer:

1. $H_o: \mu = 55$
2. $H_1: \mu \neq 55$
3. $\alpha = .05$
4. critical values = −1.96 and +1.96 (since this is a two-tailed test)
5. confidence interval = $-1.96 < z < +1.96$
6. reject H_o $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{60 - 55}{14/7} = \frac{5}{2} = +2.50$

since $2.50 > 1.96$ we reject H_o

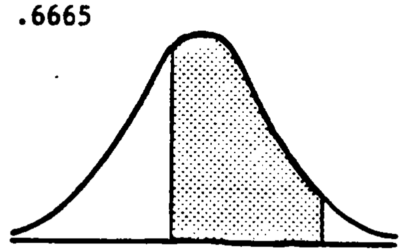
21a

Answer: 66.65%

Area between $z = 1.96$ and the mean equals .4750

Area between $z = -.50$ and mean equals .1915

.4750 + .1915 = .6665



22

Find the proportion of the area under the curve:

between $z = 2.54$ and the mean #________

between $z = 2.54$ and $z = -2.54$ #________

to the right of $z = 2.54$ #________

to the right of $z = 2.54$ and to the left of $z = -2.54$ #________

49

You have just been exposed to one type of statistical test--the z test, which is based on the standard normal distribution. There are, however, several others commonly used by statisticians that will be introduced in the following frames, and discussed in greater detail in later modules.

22a

Answer: .4945

.9890

.0055

.0110

23

Let's now see how this information can be applied. According to the table, the proportion of the area between the mean and:

z = 1 is #________

z = 2 is #________

z = 3 is #________

50

The first is the t test. Like the z test, the t test is used to test means, but is more often used than the z test for two reasons:

1. The z test requires a large sample size (over 30), while the t test can be used with <u>any</u> sample size.
2. The t test can be used not only to test a sample mean against a postulated population mean, but also to test the differences between two sample means.

23a

Answer: .3413

.4772

.4987

24

Using these proportions, we find that between $z = 1$ and $z = -1$, 2 x .3413 or roughly 68% of the area under the curve falls; between $z = 2$ and $z = -2$, 2 x #_____ or roughly #_____% falls; and between $z = 3$ and $z = -3$, 2 x #_____ or roughly #_____% falls.

This explains where those percentages came from that were mentioned in the slide/tape presentation.

51

The next test we will encounter is the analysis of variance (it's not as bad as it sounds). The ANalysis Of VAriance (ANOVA) is an extension of the t test in that it allows for testing the differences between more than two sample means, by partitioning the variance of the samples.

24a

Answer: .4772 95%

.4987 99%

25

We now know that in a normal distribution the proportions falling between ±1, ±2, and ±3 standard deviations are approximately .68, .95 and .99 respectively. In other words, if we randomly selected values from this distribution, we would expect 99 out of the 100 values to fall between #_____ standard deviations from the mean, 95 out of the 100 between #_____ standard deviations, and 68 out of the 100 between #_____ standard deviations.

52

Part of the ANOVA requires use of the *F* test, which consists simply of a ratio of variances.

25a

Answer: ±3

±2

±1

Recalling our previous exercise, if 95 out of 100 values will fall between ±2 standard deviations, only # ____ out of 100 will fall outside 2 standard deviations. 26

53

Finally, the *chi-square* (χ^2) test will be discussed. The chi-square test can be used to determine the likelihood of obtaining certain frequencies of results in comparison with the expected frequencies.

26a

Answer: 5

The meaning of this 5% will become more evident later in this unit and in units to follow.

27

Let's get back to hypothesis testing. Suppose we wanted to test the hypothesis that the mean distance required to stop a car at 20 miles an hour is 25 feet, against the alternative hypothesis that it takes more than 25 feet.

Write the null hypothesis and the alternative hypothesis.

H_o: ____________

H_1: ____________

54

In shorter notation:

z test	tests H_o: μ = any number, but requires a large sample size. ($n \geq 30$)
t test	tests H_o: μ = any number (any sample size) H_o: $\mu_1 = \mu_2$ (comparison of two means)
ANOVA	tests H_o: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_n$ (comparison of more than two means)
F test	tests H_o: $\sigma_1^2 = \sigma_2^2$ (comparison of variances)
chi-square (χ^2)	tests H_o: observed frequencies = expected frequencies

Answer: H_o: $\mu = 25$ feet

H_1: $\mu > 25$ feet

This type of situation can be handled by a slight variation of the *z* values we just talked about. Remember, we've been dealing with the percentage of scores that will fall into a certain area under the normal curve. Assuming that stopping distances are normally distributed, we can utilize *z* values to solve our problem.

Match the following null hypotheses with their appropriate test(s):

Test	Write name of test		H_o
z	______	A.	$H_o: \mu_1 = \mu_2 = \mu_3 = \mu_4$
t	______	B.	$H_o: \sigma_1^2 = \sigma_2^2$
ANOVA	______	C.	$H_o: \mu_1 = \mu_2$
F	______	D.	$H_o: \mu$ = any no. ($n = 15$)
χ^2	______	E.	$H_o: \mu$ = any no. ($n = 45$)
	______	F.	H_o: observed frequencies = expected frequencies

In our example, we want to know whether the mean distance to stop a car going 20 miles an hour is 25 feet, or if the mean distance is more than 25 feet. Let's add some sample data. 36 observations yield a mean of 29 feet necessary to stop a car at 20 mph, with a standard deviation of 12 feet. Based on this data, we want to test: 29

$$H_o: \mu = 25$$

$$H_1: \mu > 25$$

55a

Answer: A - *ANOVA*

B - *F*

C - *t*

D - *t*

E - *z*

F - χ^2

Let's review. 56

Statisticians have many ways to test many different kinds of hypotheses. This unit discussed the basis for all statistical tests:

1. Stating the null hypothesis.
2. Stating the alternative hypothesis.
3. Setting the level of significance.
4. Determining the critical values and confidence interval.

Once these have been accomplished, and the type of statistical test to be used is determined, all that remains is computation and the decision of acceptance or rejection, based on the result of the computation.

(Of course, after the statistical analysis is complete, there is still the interpretation of the results with which to contend.)

30

Just as before, using a z value we can determine what percent of the time we can expect to obtain this sample mean when the null hypothesis is true. The z value we use to do this is a bit different, however. Let's examine this difference briefly. If you draw a large number of samples from a normally distributed population, the means of those samples would also be normally distributed, with μ as its central value. The second formula below lets you obtain a z value for any sample mean within the distribution of sample means. The denominator of the formula is the standard deviation of the distribution of sample means (sometimes called "standard error of the mean") and defines the variability or dispersion of the distribution of sample means.

$z = \frac{x_i - \mu}{\sigma}$ is the formula we used before to determine the z values of raw data.

$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ is the formula we now use to obtain a z value of our sample mean.

Go to:

Frame 31

Page II-10

57

This unit also introduced the notion of standard units. To convert any data to standard units you use the following formula:

$$z = \frac{x_i - \mu}{\sigma}$$

This will always result in a mean of 0 and a standard deviation of 1.

z values can also be used to test the H_o: μ = any number (provided $n > 30$), through use of the formula:

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

where $\bar{x}$ = sample mean
μ = assumed or postulated mean of the population from which the sample was drawn
σ = standard deviation (estimated for the population because of the large *n*)
n = sample size

The obtained *z* is then compared with the value in Table E of the Guide and the appropriate conclusion drawn. Other tests briefly mentioned in this unit (*t*, *ANOVA*, *F*, X^2) will be explained in detail in the next three modules. It should be noted that all these statistical tests assume randomly selected observations from a normally distributed population.

Go to Unit 3

UNIT 3 - ERRORS

1

When testing the null hypothesis, there are four possible outcomes concerning its acceptance or rejection:

1. Accepting the null hypothesis when it is true.
2. Rejecting the null hypothesis when it is false.
3. Accepting the null hypothesis when it is false.
4. Rejecting the null hypothesis when it is true.

These possibilities can be diagrammed:

	H_o true	H_o false
Accept H_o	correct decision	error
Reject H_o	error	correct decision

As you can see, two of the possibilities are errors in the decision of acceptance or rejection. The first error is rejecting the null hypothesis when in fact it is true. This is termed "Type I" error [also "error of the first kind," or "alpha" (α)]. The second kind of error, conveniently termed "Type II" error ["error of the second kind," or "beta" (β)], is accepting the null hypothesis when it is false.

This unit explains these two kinds of errors, and shows how to determine the probabilities of committing them.

7

Fill in the blanks with "Type I" or "Type II":

	H_o true	H_o false
Accept H_o	correct decision	__________
Reject H_o	__________	correct decision

2

Since hypothesis testing calls for a decision to be made (acceptance or rejection of the null hypothesis) based on only a portion of the whole population, the possibility of making the WRONG decision exists. These wrong decisions are known as Type ______ and Type _____ errors.

7a

Answer:

	H_o true	H_o false
Accept H_o		Type II error
Reject H_o	Type I error	

Answer: Type I & Type II errors 2a

3

A Type I error occurs when we reject the null hypothesis when it is in fact true. When we (accept/reject) a (true/false) null hypothesis we have committed a Type I error.

8

The probability of making a Type I error (symbolized by α) can be controlled by the statistician fairly easily by increasing the level of significance to be used in testing the null hypothesis. We already know that convention suggests using a 5% or 1% level of significance. By changing the level from 5% to 1%, we are decreasing the probability of making a Type I error. Such change, however, does increase the probability of making a Type II error, and, in some cases, the probability of accepting a false hypothesis (Type II error) needs to be minimized. In such cases, the level of significance may be set higher than 5%.

$\alpha = .05$ means "the probability of making a Type I error is 5 chances in 100 (5%)."

$\alpha = .01$ means *______________________________

____

3a

Answer: reject

true

4

For example, if we have

H_o: $\mu = 50$

H_1: $\mu \neq 50$

and through testing of a sample mean of 60 we reject H_o when in fact μ does = 50, we have committed a Type #____ error.

8a

Answer: "the probability of making a Type I error is 1 chance in 100 (1%)".

9

By this time you may already have realized that

α = probability of making a Type I error = level of significance.

so when we set the level of significance we also then know the ____________ of making a ___________ ____ ____________

Answer: I 4a

5

A Type II error occurs when we accept a false null hypothesis. If a null hypothesis is (true/false) but we (accept/reject) it anyway, we have made a Type II error.

9a

Answer: probability of making a Type I error.

10

Given $\alpha = .05$, what is the probability of making a Type I error?

Given $\alpha = .01$, what is the probability of making a Type I error?

5a

Answer: false

accept

6

Given $H_o: \mu = 100$

$H_1: \mu \neq 100$

If on the basis of sample data we reject H_o, but μ does = 100, we have committed a Type_____ error.

If we accept H_o, but in fact $\mu \neq 100$, we have made a Type_____ error.

10a

Answer: 5 in 100 (5%)

1 in 100 (1%)

11

The probability of making a Type II error (represented by β) can also be determined, but a discussion of this will not be presented in this course since such a discussion is better suited for a more advanced course.

For more information on Type II error, see:

Glass & Stanley. Statistical methods in education and psychology. Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1970.

6a

Answer: I

II

Go to:

Frame 7

Page II-41

12

To review:

In hypothesis testing, there are two errors that can be made when the null hypothesis is accepted or rejected. These are:

	H_o True	H_o False
Accept H_o	Correct Decision	Type II Error
Reject H_o	Type I Error	Correct Decision

The probability of making a Type I error is given by α. α is also the level of significance of the statistical test, and is usually .05 or .01. This means that when we reject a null hypothesis, we are risking the possibility of rejecting a true null hypothesis only 5 times in 100, or once in 100, depending on the α level selected. Go to Unit 4.

UNIT 4 - THE STATISTICAL PROBLEM SOLVING PROCEDURE

1

The general problem of hypothesis testing is often difficult for beginners. This procedure has been developed to greatly reduce your difficulty in understanding hypothesis testing:

Step 1. Define the problem.

Step 2. Select the appropriate statistical technique and significance level.

Step 3. Organize and, if necessary, reduce the data.

Step 4. Compute the statistic.

Step 5. Consult the appropriate table for the critical values.

Step 6. Make the appropriate decision based upon the statistical results.

10

Now try this one on your own:

It is assumed from previous data that the mean TSP concentration for City A is 50 $\mu g/m^3$. Recently, 36 readings have shown a mean of 58 $\mu g/m^3$ with a standard deviation of 26 $\mu g/m^3$. Can we, using a 5% level of significance, conclude that the actual mean is greater than 50?

Go on to
Frame 11

<u>Step 1: Define the Problem</u>. First, define what the problem is. This includes defining the null hypothesis and the alternative hypothesis.

There are different null hypotheses, and they will be discussed in more detail in the later units on the various statistical techniques. To introduce you to the upcoming tests, some of these null hypotheses include:

H_o: $\mu = a$ (a = any number)
(the population mean equals a)

H_o: $\mu_1 = \mu_2$
(the mean of population 1 equals the mean of population 2)

H_o: $\sigma_1^2 = \sigma_2^2$
(the variance of population 1 equals the variance of population 2)

H_o: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
(the mean of population 1 equals the mean of population 2 equals the mean of population 3 equals the mean of population 4)

Each of these hypotheses calls for a different test. They will be discussed later.

11

Step 1: Define the problem.

H_o: ________

H_1: ________

3

Step 2: Select the Appropriate Statistical Technique and Significance Level. On the basis of the null hypothesis of the problem, the appropriate type of statistical technique must be selected. (The specific technique will also often rely on the nature of the data. This aspect of the selection process will be discussed in later units along with the specific tests.)

The significance level should also be set at this point. Careful thought must go into this selection, because once set it should not be changed. Basically, the selection of significance level boils down to asking yourself the question: "At what level of significance will I and other researchers be satisfied with the results obtained by the test?" Convention usually suggests the 5% or 1% level.

11a

Answer: $H_0: \mu = 50$

$H_1: \mu > 50$

12

Step 2: Select the appropriate statistical technique and significance level.

test ______________

α = # ______________

4

<u>Step 3: Organize and, If Necessary, Reduce the Data</u>. This involves accumulating the data and putting it into the form required by the test selected. Sometimes this will mean rank ordering, and other times it will mean categorizing the data. In any case, you should do whatever is necessary to facilitate the computation of the selected statistic.

12a

Answer: z test, since we are testing a sample mean against a postulated mean and since $n > 30$.

$\alpha = .05$

13

Step 3: Organize and, if necessary, reduce the data.

5

Step 4: Compute the Statistic. Now that the data are organized, compute the test you have selected.

13a

Answer: Not necessary

14

Step 4: Compute the statistic

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

6

<u>Step 5. Consult the Appropriate Table for the Critical Values</u>. Find the table corresponding to the test and determine the critical values.

14a

Answer:

$$z = \frac{58 - 50}{26/\sqrt{36}}$$

$$= \frac{8}{4.33}$$

$$= 1.85$$

15

Step 5: Consult the appropriate table for the critical values.

Step 6: Make the Appropriate Decision Based Upon the Statistical Results. Compare the results of the selected statistical tests with the critical values in the tables and make the appropriate decision to either accept or reject the null hypothesis.

15a

Answer: critical value is +1.64

Since this is a one-tailed test, we need the z corresponding to .5000 - .0500 or .4500.

16

Step 6: Make the appropriate decision based on the statistical results.

Here is an example of the application of the 6-step technique, using the z test:

It is stated in a report that the mean CO concentration for a location is 8 mg/m^3. To test this assumption, 49 readings of CO concentrations are taken, yielding a mean of 6.5 mg/m^3 and standard deviation of 3.5 mg/m^3. We want to know, with 95% confidence, if the mean is actually 8 mg/m^3, or if it differs significantly.

Step 1: Define the problem.

H_o: $\mu = 8$ mg/m^3

H_1: $\mu \neq 8$ mg/m^3

16a

Answer: Reject H_o: $\mu = 50$

You may have noticed in this example that if you had selected a critical value of +1.96 based on a two-tailed test, you would have accepted H_o: $\mu = 50$. You must be careful to select the correct critical value.

Step 2: Select the appropriate statistical technique and significance level.

Since the null hypothesis concerns testing a sample mean against a postulated mean, and $n > 30$, we select the z test. Since the specified confidence level is 95%, $\alpha = .05$ (100 % - 95%).

Step 3: Organize and, if necessary, reduce the data.

This is not necessary in the z test.

Step 4: Compute the statistic.

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$$= \frac{6.5 - 8.0}{3.5/\sqrt{49}}$$

$$= \frac{-1.5}{.5}$$

$$= -3.0$$

Step 5: Consult the appropriate table for the critical values.

In Table E we find the z value corresponding to .4750.

$$\left(.5000 - \frac{.0500}{2}\right)$$

Critical values = ±1.96

Step 6: Make the appropriate decision based upon the statistical results.

Since the obtained z (-3.00) beyond the critical values (±1.96), we reject the null hypothesis that $\mu = 8\ mg/m^3$.

Go to:

Frame 10

Page II-48

17

In future statistical tests, you will be expected to show your results in this 6-step procedure, but you need not memorize each step. All that you must be able to do is follow it. Anytime you must use the procedure you can refer to page ii in the Guide, which contains the six steps.