Unit 25: Tests of Significance

SUMMARY OF VIDEO

Sometimes, when you look at the outcome of a particular study, it can be hard to tell just how noteworthy the results are. For example, if the severe injury and death rates due to car crashes on one state's roads have dropped from 4.7% down to 3.8% after enacting a seat belt law, how would we know whether this result was due to the seat belt law or simply due to chance variation?

To sort out whether results are due to chance or there is something else at work (such as the enactment of the seat belt law), statisticians turn to a tool of inference called tests of significance. Significance testing can be applied in a variety of situations. We next explore how researchers used it to help solve a controversy in classic literature.

In 1985, scholar Gary Taylor made a surprising find while conducting research for a new edition of the complete works of William Shakespeare. While going through a 17th century anthology at the Bodleian Library at Oxford University, he came upon a sonnet he had never seen before and it was attributed to William Shakespeare. Obviously, Taylor was excited about his new find and wanted to include it in his new edition of *The Complete Works*.

This discovery caused quite a controversy – some scholars were thrilled by the discovery but others didn't think the poem was good enough to be one of Shakespeare's. Statistics to the rescue! A decade earlier, statistician Ron Thisted had done a statistical analysis of Shakespeare's vocabulary. Thisted's program provided a detailed, numeric description of Shakespeare's vocabulary. For every work, Thisted could tell how many new words there were that Shakespeare didn't use anywhere else. Using this model, Thisted predicted that if Shakespeare had written the poem in question, it would have 7 unique words in it. When they ran the poem through the program, however, they found that there were 10 unique words. Did this difference reflect random variation within Shakespeare's writing? Or did it indicate that Shakespeare was not the author? This is where significance testing (or tests of hypotheses) can be helpful.

Thisted set up two opposing hypotheses: the null hypothesis, written as H_0, that basically means nothing unusual is happening; and the alternative hypothesis, the researchers' point of

view, written as H_a. Researchers aim to reject the null hypothesis with evidence that suggests something more is going on than random variation. In this case, the hypotheses are:

H_0: Shakespeare wrote the poem.

H_a: Someone other than Shakespeare wrote the poem.

The question was whether the discrepancy between the observed number of unique words, 10, and the predicted number of unique words, 7, was due to another author writing the poem rather than to chance variation. Is that three-word difference a big difference? To answer this question, Thisted assumed (based on his data) that the number of unique words in Shakespeare's poems had the approximately normal distribution with mean μ = 7 and standard deviation $\sigma = 2.6$ shown in Figure 25.1.

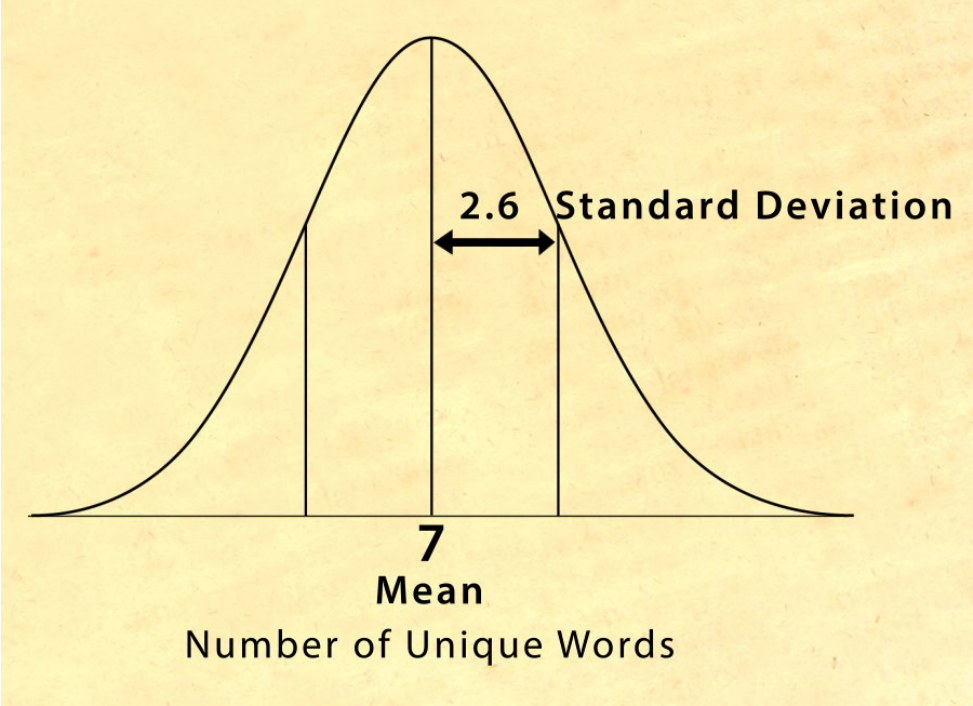


Figure 25.1. Distribution of the number of unique words in Shakespeare's poems.

The shaded area under the density curve in Figure 25.2 corresponds to the probability of a number of unique words at least as extreme as 10 (in other words, a difference from 7 of 3 or more words).

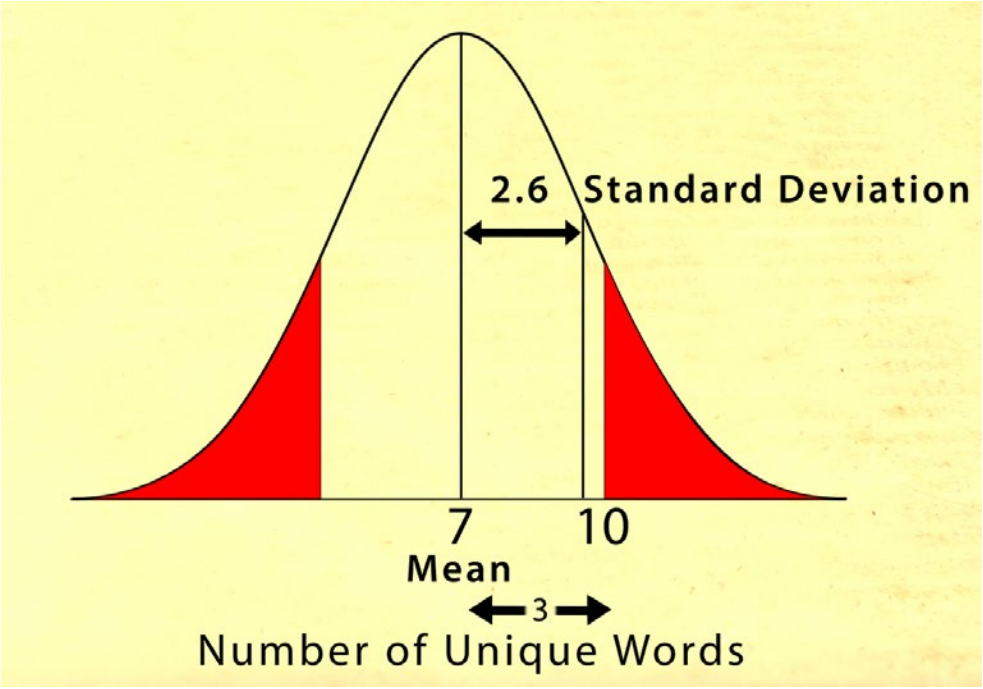


Figure 25.2. Finding the p-value.

Using technology, we find that the shaded area is 2(0.1243) = 0.2483. Thus, Thisted could expect to find a value at least as extreme as 10 unique words roughly 25% of the time. Therefore, Thisted failed to find significant evidence against the null hypothesis that Shakespeare wrote the poem. He could not reject H_0. In the absence of literary or statistical evidence against Shakespeare's authorship, the poem was published in Taylor's edition of *The Complete Works*.

Since we want to work with sample means, let's suppose researchers found a folio of five new poems that were attributed to Shakespeare. Suppose that our sample mean from the five poems in the folio is $\bar{x} = 8.2$. We want to know if, based on this evidence, we can conclude that Shakespeare did not write these poems. We set up our null and alternative hypotheses:

$H_0 : \mu = 7$

Shakespeare wrote the poems.

$H_a : \mu \neq 7$

Someone else wrote the poems.

One thing to decide, when setting up a significance test, is whether to use a one-sided or two-sided alternative hypothesis. In our Shakespeare example, we are using a two-sided alternative hypothesis because a different author might consistently use either more or fewer unique words than Shakespeare. But suppose we suspected the poem was written by a particular author who was known to consistently use more unique words than Shakespeare?

Then the alternative hypothesis would be one-sided:

$H_a : \mu > 7$

We begin by assuming the null hypothesis is true. Then we find the probability of getting a result at least as extreme as ours if the null hypothesis really is true. If these poems were written by Shakespeare, then the distribution of $\bar{x}$, the mean number of unique words per poem in five poems, would have a normal distribution with the following mean and standard deviation:

$\mu_{\bar{x}} = \mu$

$\sigma_{\bar{x}} = \frac{2.6}{\sqrt{5}} \approx 1.163$

Next, we need to find the probability that any sample of five of Shakespeare poems would have an $\bar{x}$ at least as far from 7 as what we observed from our sample, $\bar{x} = 8.2$. Figure 25.3 illustrates this probability. Notice that two areas are shaded because our alternative is two-sided.

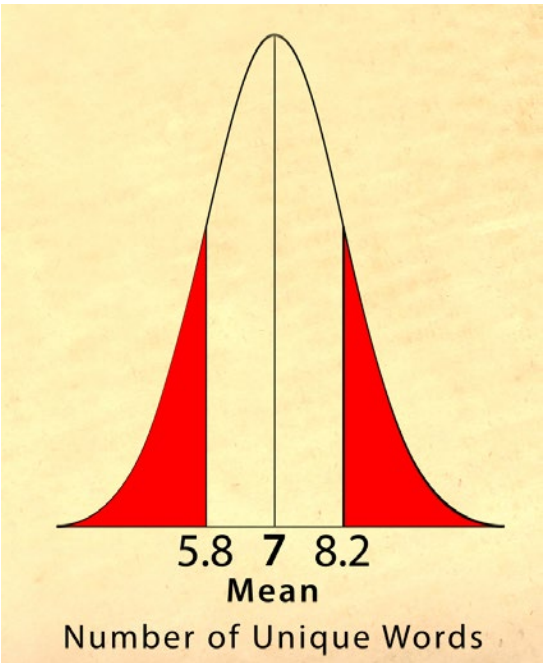


Figure 25.3. Sampling distribution of $\bar{x}$.

To calculate this probability from a standard normal table, we find the *z*-score for our observed sample mean. This is called a *z*-test statistic:

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

$$z = \frac{8.2 - 7}{2.6 / \sqrt{5}} \approx 1.03$$

So, the observed value of our test statistic *z* is 1.03, a little more than one standard deviation away from the mean, 0, on the standard normal curve. The final step in our test of significance is to find the probability of observing a value from a standard normal distribution that is at least this extreme. This probability is called the *p*-value. To find this *p*-value, we use $z = 1.03$ and look in the standard normal table (*z*-table). From Figure 25.4, we find that the area under the standard normal curve to the left of 1.03 is 0.8485.

z	.00	.01	.02	.03	.04	.05
0.7	.7580	.7611	.7642	.7673	.7704	.7734
0.8	.7881	.7910	.7939	.7967	.7995	.8023
0.9	.8159	.8186	.8212	.8238	.8264	.8289
1.0	.8413	.8438	.8461	.8485	.8505	.8531
1.1	.8643	.8665	.8686	.8708	.8729	.8749

Figure 25.4. Portion of standard normal table (z-table).

That means that 1 – 0.8485 or 0.1515 is the area in the right tail (the shaded region in Figure 25.5). Since we choose a two-sided alternative, we double this value because we are interested in the area under BOTH tails (the area to the right of 1.03 and the area to the left of -1.03). Our final result gives a *p*-value of 0.303.

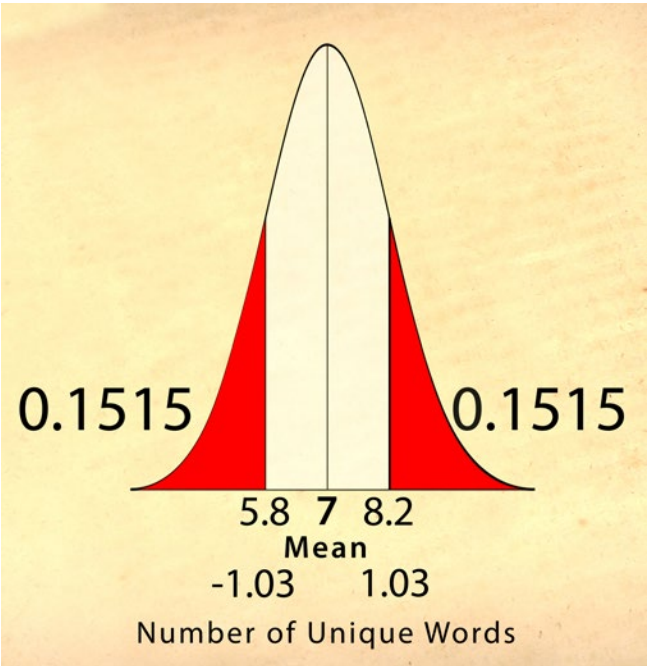


Figure 25.5. Finding the p-value from a standard normal distribution.

From the *p*-value, we know that there is a 30.3% chance that random variation would produce a mean unique word count as far from 7 in either direction as 8.2. Since a 30.3% chance is a pretty good chance, we have failed to disprove the null hypothesis. We have not found good evidence against Shakespeare's authorship of these new poems.

This example helps illustrate the general rule about *p*-values: Small *p*-values give evidence against the null hypothesis; large *p*-values fail to reject the null hypothesis. Since *p*-values can range from the very small – close to zero – to the very large – close to one, researchers need to decide when a *p*-value is small enough for them to reject the null hypothesis. One of the most common levels is 0.05 or 5%. If something is statistically significant at the 5% level, it means that the results produced a *p*-value less than 0.05. Another widely used level is 0.01 or the 1% level.

STUDENT LEARNING OBJECTIVES

A. Understand that a significance test answers the question “Is this sample outcome good evidence that an effect is present in the population, or could it easily occur just by chance?”

B. Be able to formulate the null hypothesis and alternative hypothesis for tests about the mean of a population. Understand that the alternative hypothesis is the researcher’s point of view.

C. Understand the concept of a *p*-value. Know that smaller *p*-values indicate stronger evidence against the null hypothesis.

D. Be able to calculate *p*-values as areas under a normal curve in the setting of tests about the mean of a normal population with known standard deviation.

E. Be able to test a population mean with a *z*-test.

CONTENT OVERVIEW

A **significance test** (also called a **test of hypotheses**) answers the question "Is this sample outcome good evidence that an effect is present in the population, or could it easily occur just by chance?" The reasoning is as follows: Suppose, for the moment, that we assume the effect is not present in the population. If the observed result is very unlikely to occur given this assumption, that's evidence that the supposition of "no population effect" is false.

The statement being tested in a test of significance is called the **null hypothesis**, written H_0. For example, H_0 might state that a population parameter, such as the mean μ, takes a specific value. Usually the null hypothesis is a statement of "no effect" or "no difference" or "status quo." The test of significance is designed to assess the strength of the evidence against the null hypothesis and in favor of an alternative hypothesis H_a that represents the effect we hope or suspect is true. (H_a is generally the researcher's point of view.) The alternative hypothesis might be that the parameter differs from its null value, in a specific direction (one-sided alternative) or in either direction (two-sided alternative).

Suppose that we want to conduct a test about the mean of a population. More specifically, suppose that we want to test that the mean has a specific value, which we'll call μ_0, or that it doesn't have that value, or is smaller than that value, or larger than that value. We form two opposing hypotheses – the null and alternative hypotheses – which we express symbolically as follows (select one of the possible alternatives):

$H_0 : \mu = \mu_0$
$H_a : \mu \neq \mu_0$ or $H_a : \mu > \mu_0$ or $H_a : \mu < \mu_0$

To test the hypothesis $H_0 : \mu = \mu_0$ based on a random sample of size n from a population with unknown mean μ and known standard deviation σ, we compute the sample mean $\bar{x}$. Here's a recap of what we know about $\bar{x}$:

- If H_0 is true and the population is normal, then $\bar{x}$ has the normal distribution with mean μ_0 and standard deviation $\sigma/\sqrt{n}$.
- Suppose instead that the population does not follow a normal distribution. If the sample size *n* is large, we can apply the Central Limit Theorem and conclude that $\bar{x}$ is approximately normally distributed with mean μ_0 and standard deviation $\sigma/\sqrt{n}$.

- Next, still assuming H_0 is true, we convert $\bar{x}$ into a *z*-score. The result is the z-test statistic given below:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

If H_0 is true, *z* has the standard normal distribution (at least approximately).

Now, we work through an example. Researchers studying the effects of smoking on sleep believe that men who smoke need more sleep than what is average for men, which is 7.5 hours per night. Let μ be the mean number of hours of sleep for men who smoke. Assume that the standard deviation is $\sigma = 0.5$ hours. The null and alternative hypotheses are:

$H_0 : \mathrm{m} = 7.5$
$H_a : \mathrm{m} > 7.5$

A random sample of 50 smokers completed a questionnaire in which they were asked to record the number of hours they sleep each night. The sample mean is $\bar{x} = 7.7$ hours. We compute the *z*-test statistic as follows:

$$z = \frac{7.7 - 7.5}{0.5 / \sqrt{50}} \approx 2.83$$

From the *z*-test statistic, we learn that the observed value of $\bar{x} = 7.7$ is 2.83 standard deviations from the hypothesized mean from H_0, $\mu = 7.5$. If H_0 is true, then *z* has the standard normal distribution. Now, we are ready to evaluate the evidence against H_0 – How likely would it be to observe a value from the standard normal distribution that is at least as extreme as 2.83? The answer, around 0.2%, is illustrated in Figure 25.6. Around 0.2% is pretty unlikely. So, in this case, we reject the null hypothesis and accept the alternative: Male smokers, on average, need more sleep than men in general.

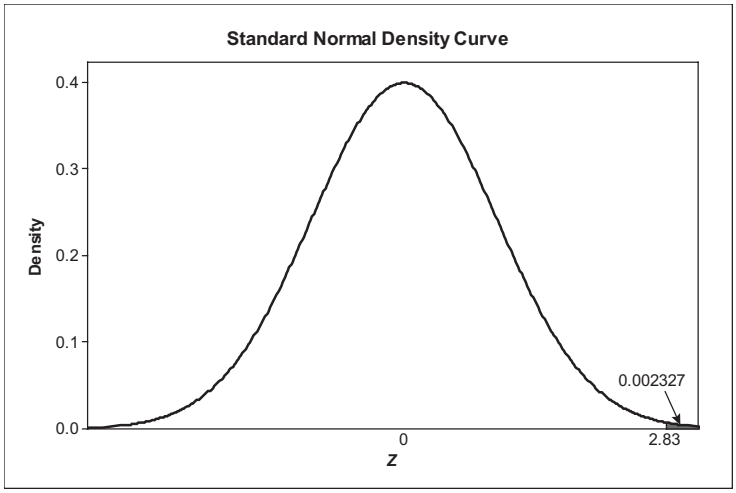


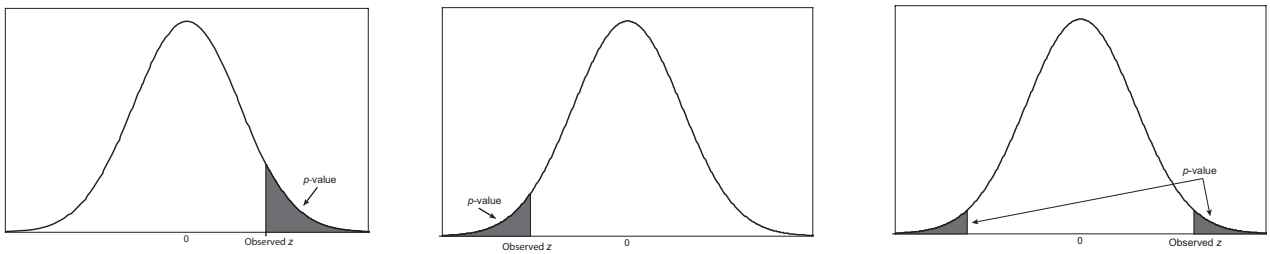
Figure 25.6. The evidence against H_0.

As we saw in the previous example, the distribution of the *z*-test statistic, under the assumption that H_0 is true, allows us to use the observed *z*-value to assess the evidence against H_0. We calculate the probability, assuming H_0 is true, of observing a value from the standard normal distribution as extreme or more extreme than the *z*-value we calculated – this probability is called the *p*-value. Because there are three possible alternatives, there are three possibilities for computing the *p*-value:

1. The *p*-value for a test of H_0 against $H_a : \mu > \mu_0$ is the probability of observing a value from the standard normal distribution that is at least as large as the observed *z*-test statistic. (See Figure 25.7 (1).)

2. The *p*-value for a test of H_0 against $H_a : \mu < \mu_0$ is the probability of observing a value from the standard normal distribution that is at least as small as the observed *z*-test statistic. (See Figure 25.7 (2).)

3. The *p*-value for a test of H_0 against $H_a : \mu \neq \mu_0$ is the probability of observing a value from the standard normal distribution that is at least as far from 0 (on either side of 0) as the observed *z*-test statistic. (See Figure 25.7 (3).)



(1) (2) (3)

Figure 25.7. Calculating p-values corresponding to alternative hypotheses (1 – 3).

Small *p*-values mean that the probability of observing standard normal values at least as extreme as the observed *z*-test statistic are very unlikely to occur assuming the null hypothesis is true. Hence, small *p*-values provide evidence against the null hypothesis in support of the alternative.

Sometimes we set certain cutoffs for the *p*-value called the **significance level**. For example, if the *p*-value is below 0.05 ($p < 0.05$), we say the results are significant at the 0.05 level, or the 5% level.

KEY TERMS

A **significance test** or **test of hypotheses** is a method that uses sample data to decide between two competing claims.

The claim tested by a significance test is called the **null hypotheses**. Usually the null hypothesis is a statement about "no effect" or "no change." The claim that we are trying to gather evidence for – the researcher's point of view – is called the **alternative hypothesis**. The alternative hypothesis is **two-sided** if it states that a parameter is different from the null hypothesis value. The alternative hypothesis is **one-sided** if it states that either a parameter is greater than or a parameter is less than the null hypothesis value.

A **test statistic** is a quantity computed from the sample data that measures the gap between the null hypotheses and the sample data. A test statistic is used to make a decision between the null and alternative hypotheses.

The ***p*-value** is the probability, computed under the assumption that the null hypothesis is true, of observing a value from the test statistic at least as extreme as the one that was actually observed.

The **significance level** of a test of hypotheses is the highest *p*-value for which we will reject the null hypothesis.

A ***z*-test statistic** for testing $H_0 : \mu = \mu_0$, where μ is the population mean, is given by:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

The *z*-test is used in situations where the population standard deviation σ is known and either the population has a normal distribution or the sample size *n* is large.

THE VIDEO

Take out a piece of paper and be ready to write down answers to these questions as you watch the video.

1. In the 1970s, statistician Ron Thisted did a statistical analysis of Shakespeare's vocabulary. Based on his analysis he created a computer program. What could his program tell you about a Shakespearean poem?

2. In analyzing a poem to see whether or not it was authored by Shakespeare, Thisted set up a null hypothesis and an alternative hypothesis. State those hypotheses in words.

3.What was the approximate distribution of the number of unique words per poem in Shakespeare's poems?

4. Thisted observed 10 unique words in the newly discovered poem. Was that sufficient evidence to conclude that Shakespeare did not write the poem?

5. Which is better evidence *against* the null hypothesis, a large *p*-value or a small *p*-value?

UNIT ACTIVITY:
CHIPS AHOY!

Nabisco Chips Ahoy is a popular brand of chocolate chip cookie. In the 1980s, Nabisco ran television ads claiming that their cookies had, on average, 16 chips per cookie. Since the 1980s many more brands of chocolate chip cookies have appeared on supermarket shelves, which could have put pressure on Nabisco to improve its product perhaps by increasing the amount of chips. On the other hand, the price of chocolate has increased, which could have had the opposite effect. In this activity, you will test whether or not Nabisco could run the same ad today.

1. Collect the data. Your instructor will provide directions and, after the data collection is complete, distribute the data. (Save the data for use in Unit 27's activity.)

2. Compute the mean and standard deviation of the number of chips per cookie.

3. a. State the null and alternative hypotheses.

b. Calculate the value of the *z*-test statistic. (Since the sample size is large, use *s* in place of σ.)

c. Calculate the *p*-value and state your conclusion.

4. Calculate a 95% confidence interval for μ. Does your confidence interval indicate that μ has increased, decreased, or remained the same from its value in the 1980s?

EXERCISES

1. Each of the following situations requires a significance test about a population mean μ. State the appropriate null hypothesis, H_0, and alternative hypothesis, H_a, in each case.

a. Larry's car averages 32 miles per gallon on the highway. He switches to a new motor oil that is advertised as increasing gas mileage. After driving 3000 highway miles with the new oil, he wants to determine if his gas mileage actually has increased.

b. A university gives credit in a French language course to students who pass a placement test. The language department wants to know if students who get credit in this way differ in their understanding of spoken French from students who actually take the French course. Some faculty think the students who test out of the course are better, but others argue that they are weaker in oral comprehension. Experience has shown that the mean score of students in the course on a standard listening test is 24. The language department gives the same listening test to a sample of 40 students who passed the placement test to see if their performance is different.

c. Experiments on learning in animals sometimes measure how long it takes a mouse to find its way through a maze. The mean time is 18 seconds for one particular maze. A student thinks that a loud noise will cause the mice to complete the maze faster. She measures how long each of 10 mice takes with a noise as stimulus.

2. The Survey of Study Habits and Attitudes (SSHA) is a psychological test that measures the motivation, attitude toward school, and study habits of students. Scores range from 0 to 200. The mean score for U.S. college students is about 115, and the standard deviation is about 30. A teacher who suspects that older students have better attitudes toward school gives the SSHA to 25 students who are at least 30 years of age. Their mean score is $\bar{x} = 125.2$.

Assume that σ = 30 for the population of older students, and that the students tested are a random sample from the population of older college students. Carry out a significance test of

$H_0 : \mu = 115$
$H_a : \mu > 115$

Report the value of the test statistic, the *p*-value of your test, and state your conclusion clearly.

3. Radon is a colorless, odorless gas that is naturally released by rocks and soils and may concentrate in tightly closed houses. Because radon is slightly radioactive, there is some concern that it may be a health hazard. Radon detectors are sold to homeowners worried about this risk, but the detectors may be inaccurate. Tricia wants to study the accuracy of radon detectors for a science fair project. At a nearby university, she places 12 detectors in a chamber where they are exposed to 105 picocuries per liter (pci/l) of radon over 3 days. Here are the readings given by the detectors.

91.9	97.8	111.4	122.3	105.4	95.0
99.6	96.6	119.3	104.8	101.7	03.8

a. In this case, the sample size $n = 12$ is relatively small. Check to see if it is reasonable to assume these data come from an approximately normal population.

b. Do these observations provide good evidence that the average detector reading differs from the true value of 105? Assume that you know that the standard deviation of readings for all detectors of this type is $\sigma = 9$.

4. The CDC publishes charts on Body Mass Index (BMI) percentiles for boys and girls of different ages. Based on the chart for girls, the mean BMI for 6-year-old girls is listed as 15.2 kg/m2. The data from which the CDC charts were developed is old and there is concern that the mean BMI for 6-year old girls has increased. The BMIs of a random sample of 30 6-year-old girls are given below.

24.5	16.3	15.7	20.6	15.3	14.5	14.7	15.7	14.4	13.2
16.3	15.9	16.3	13.5	15.5	14.3	13.7	14.3	13.7	16.0
14.2	17.3	19.5	22.8	16.4	15.4	18.2	13.9	17.6	15.5

a. State null and alternative hypotheses relevant to this situation.

b. Calculate the sample mean and standard deviation.

c. Since the sample size is relatively large, use *s* in place of σ and calculate the value of the *z*-test statistic. Then calculate the *p*-value.

d. Based on your answer to (c), do the sample data provide sufficient evidence that the mean BMI for 6-year-old girls has increased? Explain.

REVIEW QUESTIONS

1. Small amounts of sulfur compounds are often present in wine. Because these compounds have unpleasant odors, wine experts have determined the odor threshold, the lowest concentration of a compound that a trained human nose can detect. For example, the odor threshold for dimethyl sulfide (DMS) is 25 micrograms per liter of wine (μg/l). Untrained noses may be less sensitive, however. A wine researcher found the DMS odor thresholds for 10 students in his restaurant management class. Here are the data.

31	31	43	36	23	34	32	30	20	24

Assume that the standard deviation of the odor threshold for untrained noses is known to be σ = 7 μg/l.

a. Is it reasonable to assume the data are from an approximately normal population? Explain.

b. The researcher believes that the mean odor threshold for beginning students is higher than the published threshold, 25 μg/l, and decides to conduct a significance test. What are the null and alternative hypotheses?

c. Carry out a significance test. Report the value of the test statistic, the *p*-value, and your conclusion.

2. In 2010/2011 the national mean SAT Math score was 514. Faculty at a state university had disagreements over their students' mathematics preparation for college. Some felt that their students had fallen below the national average, and others felt that their students had made some advances. To help answer this question, math faculty took a random sample of 50 students who entered the university fall semester 2011. The SAT Math scores from those students are given below.

580	540	520	490	430	570	520	540	440	610
430	390	470	550	390	500	550	440	550	660
560	550	450	560	680	630	400	450	500	460
460	530	590	380	660	570	520	530	500	680
450	590	660	420	370	550	450	510	480	500

a. Calculate the sample mean and standard deviation.

b. Do these data provide sufficient evidence that the mean SAT Math scores of students entering the university in fall 2011 differed from the national mean? State the hypothesis you are testing, the value of the test statistic, the *p*-value and your conclusion. (Replace σ in the test statistic by *s* since the sample size is large.)

c. Construct a 95% confidence interval for μ, the mean Math SAT for students entering this university in fall 2011. (Refer to Unit 24, Confidence Intervals.) Does your confidence interval indicate that the true mean SAT Math score for students entering the university in fall 2011 is less than 514, could be 514, or is greater than 514? Explain.

3. The average length of calls coming into a municipal call center had been around 90 seconds. Lately, there has been some concern that more complicated calls are coming into the center causing the mean length of the calls to increase. In order to test this assumption, the city draws a random sample of 100 calls. The sample mean and standard deviation are $\bar{x} = 118.4$ seconds and *s* = 186.5 seconds, respectively.

a. State the hypotheses being tested.

b. Do these data provide good evidence that the average call length has increased from 90 seconds? (Since the sample size is large, use *s* in place of μ) Show the work needed to support your answer. Conduct the significance test at the 0.05 level.

c. Suppose city planners are willing to run the test at the 0.10 level. (They will reject the null hypothesis if the *p*-value is below 0.10.) Would this change the conclusion reached in (b)? Explain.

4. Eating fish contaminated with mercury can cause serious health problems. Mercury contamination from historic gold mining operations is fairly common in sediments of rivers, lakes and reservoirs today. A study was conducted on Lake Natoma in California to determine if the mercury concentration in fish in the lake exceeded guidelines for safe human consumption. Suppose that you are an inspector for the Fish and Game Department and that you are given the task of determining whether to prohibit fishing in Lake Natoma. You will close the lake to fishing if it is determined that fish from the lake have unacceptably high mercury content.

a. Assuming that mercury concentration of 5 ppm is considered the maximum safe concentration, which of the pairs of hypotheses below would you test? Justify your choice.

$H_0 : \mu = 5$ versus $H_a : \mu > 5$

or

$H_0 : \mu = 5$ versus $H_a : \mu < 5$

b. Would you prefer a significance level of 0.1 or 0.01 for your test? Explain your choice.