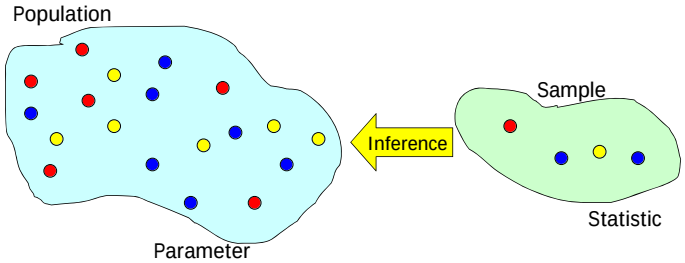
Chapter 12

Inference About A Population

Inference About A Population...



We will develop techniques to estimate and test three population parameters:

Population Mean μ

Population Variance σ^2

Population Proportion p

Inference With Variance Unknown...

Previously, we looked at estimating and testing the population mean when the population standard deviation (σ) was known or given:

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

But how often do we know the actual population variance?

Instead, we use the ***Student t-statistic***, given by:

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

Inference With Variance Unknown...

When σ is unknown, we use its point estimator $\boldsymbol{s}$

and the z-statistic is replaced by the the t-statistic, where the number of "degrees of freedom" $\boldsymbol{\nu}$, is n–1.

Testing μ when σ is unknown...

When the population standard deviation is unknown and the population is normal, the test statistic for testing hypotheses about μ is:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

which is Student ***t*** distributed with ν = n–1 degrees of freedom. The confidence interval estimator of μ is given by:

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

Example 12.1

It is likely that in the near future nations will have to do more to save the environment.

Possible actions include reducing energy use and recycling.

Currently (2007) most products manufactured from recycled material are considerably more expensive than those manufactured from material found in the earth.

Example 12.1

Newspapers are an exception.

It can be profitable to recycle newspaper.

A major expense is the collection from homes. In recent years a number of companies have gone into the business of collecting used newspapers from households and recycling them.

A financial analyst for one such company has recently computed that the firm would make a profit if the mean weekly newspaper collection from each household exceeded 2.0 pounds.

Example 12.1

In a study to determine the feasibility of a recycling plant, a random sample of 148 households was drawn from a large community, and the weekly weight of newspapers discarded for recycling for each household was recorded. Xm12-01*

Do these data provide sufficient evidence to allow the analyst to conclude that a recycling plant would be profitable?

Example 12.1

IDENTIFY

Our objective is to ***describe*** the population of the amount of newspaper discarded per household, which is an interval variable. Thus the parameter to be tested is the population mean μ.

We want to know if there is enough evidence to conclude that the mean is greater than 2. Thus,

$H_1: \mu > 2$

Therefore we set our usual null hypothesis to:

$H_0: \mu = 2$

Example 12.1

IDENTIFY

The test statistic is:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \quad \nu = n - 1$$

Because the alternative hypothesis is:

$H_1: \mu > 2$

the rejection region becomes:

$$t > t_{\alpha,\nu} = t_{.01,148} \approx t_{.01,150} = 2.351$$

Example 12.1

COMPUTE

From the data we determine :

$\sum x_i = 322.7 \quad \text{and} \quad \sum x_i^2 = 845.1$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{322.7}{148} = 2.18$$

$$s^2 = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1} = \frac{845.1 - \frac{(322.7)^2}{148}}{148-1} = .962$$

$$s = \sqrt{s^2} = \sqrt{.962} = .981$$

Example 12.1

COMPUTE

The Value of the test statistic :

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{2.18 - 2.0}{.981/\sqrt{148}} = 2.23$$

Example 12.1

COMPUTE

t-Test: Mean			
			Newspaper
Mean			2.1804
Standard Deviation			0.9812
Hypothesized Mean			2
df			147
t Stat			2.2369
P(T<=t) one-tail			0.0134
t Critical one-tail			2.352
P(T<=t) two-tail			0.0268
t Critical two-tail			2.6097

Example 12.1

INTERPRET

The value of the test statistic is t = 2.24 and its p-value is .0134.

There is not enough evidence to infer that the mean weight of discarded newspapers is greater than 2.0.

Note that there is some evidence; the p-value is .0134. However, because we wanted the Type I error to be small we insisted on a 1% significance level.

Thus, we cannot conclude that the recycling plant would be profitable.

Example 12.2

In 2004 (the latest year reported) 130,134,000 tax returns were files in the United States.

The Internal Revenue Service (IRS) examined 0.77% or 1,008,000 of them to determine if they were correctly done.

To determine how well the auditors are performing a random sample of these returns was drawn and the additional tax was reported. Xm12-02

Estimate with 95% confidence the mean additional income tax collected from the 1,008,000 files audited.

Example 12.2

IDENTIFY

The objective is to describe the population of additional tax collected.

The data are interval.

The parameter to be estimated is the mean additional tax.

The confidence interval estimator is

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$$

Example 12.2

COMPUTE

From the date we determine :

$\sum x_i = 1,254,240$ and $\sum x_i^2 = 9,232,718,166$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{1,254,240}{209} = 6,001$$

$$s^2 = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1} = \frac{9,232,718,166 - \frac{(1,254,240)^2}{209}}{209-1} = 8,201,144$$

$$s = \sqrt{s^2} = \sqrt{8,201,144} = 2,864$$

Example 12.2

COMPUTE

95% confidence interval estimate:

$1-\alpha = .95$

$\alpha = .05$

$\alpha/2 = .025$

$t_{\alpha/2,\nu} = t_{.025,208} \approx t_{.025,200} = 1.972$

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} = 6,001 \pm 1.972\frac{2,864}{\sqrt{209}} = 6,001 \pm 391$$

$LCL = \$5,610$ $UCL = \$6,392$

Example 12.2

COMPUTE

	A	B	C	D
1	**t-Estimate: Mean**			
2				
3				*Taxes*
4	Mean			6001
5	Standard Deviation			2864
6	LCL			5611
7	UCL			6392

Example 12.2...

INTERPRET

We estimate that the mean additional tax collected lies between $5,611 and $6,392 .

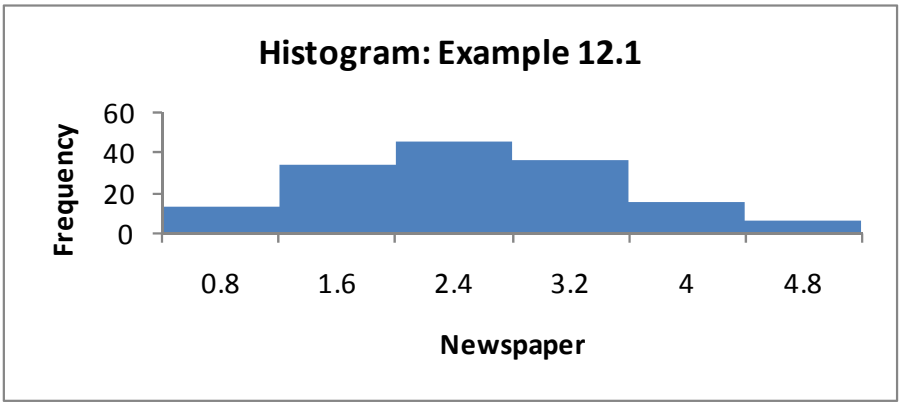
We can use this estimate to help decide whether the IRS is auditing the individuals who should be audited.

Check Required Conditions

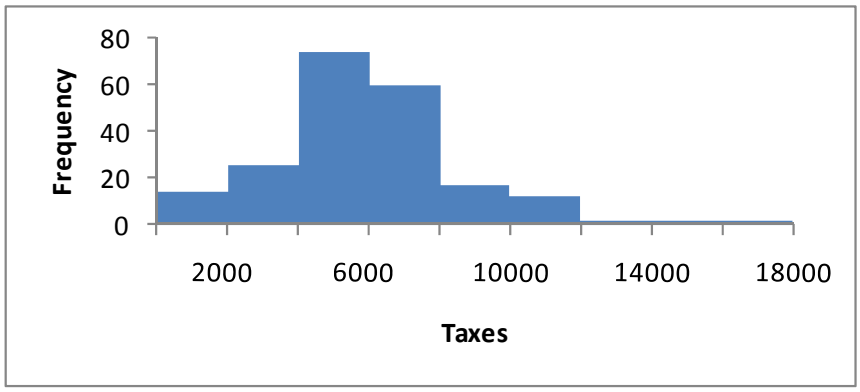
The Student t distribution is ***robust***, which means that if the population is nonnormal, the results of the t-test and confidence interval estimate are still valid provided that the population is "not ***extremely*** nonnormal".

To check this requirement, ***draw a histogram*** of the data and see how "bell shaped" the resulting figure is. If a histogram is extremely skewed (say in the case of an exponential distribution), that could be considered "extremely nonnormal" and hence t-statistics would be not be valid in this case.

Histogram for Example 12.1



Histogram for Example 12.2



Estimating Totals of Finite Populations

The inferential techniques introduced thus far were derived by assuming infinitely large populations. In practice however, most populations are finite.

When the population is small we must adjust the test statistic and interval estimator using the finite population correction factor introduced in Chapter 9.

However, in populations that are large relative to the sample size we can ignore the correction factor. Large populations are defined as populations that are at least 20 times the sample size.

Estimating Totals of Finite Populations

Finite populations allow us to use the confidence interval estimator of a mean to produce a confidence interval estimator of the population total.

To estimate the total we multiply the lower and upper confidence limits of the estimate of the mean by the population size.

Thus, the confidence interval estimator of the total is

$$N\left[\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}\right]$$

Estimating Totals of Finite Population

For example, a sample of 500 households (in a city of 1 million households) reveals a 95% confidence interval estimate that the ***household mean*** spent on Halloween candy lies between \$20 & \$30.

We can estimate the ***total*** amount spent in the city by multiplying these lower and upper confidence limits by the total population:

$$N\left[\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}\right] = 1{,}000{,}000[\$20 \ldots \$30]$$

Thus we estimate that the ***total*** amount spent on Halloween in the city lies between \$20 million and \$30 million.

Developing an Understanding of Statistical Concepts

The t-statistic like the z-statistic measures the difference between the sample mean and the hypothesized value of in terms of the number of standard errors.

However, when the population standard deviation is unknown we estimate the standard error as

$$s / \sqrt{n}$$

Developing an Understanding of Statistical Concepts

When we introduced the Student t distribution in Section 8.4 we pointed out that it is more widely spread out than the standard normal.

This circumstance is logical.

The only variable in the z-statistic is the sample mean, which will vary from sample to sample.

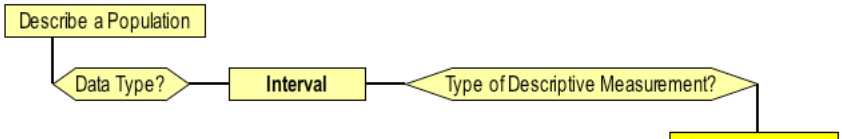
Developing an Understanding of Statistical Concepts

The t-statistic has two variables, the sample mean and the sample standard deviation s, both of which will vary from sample to sample.

Because of this feature the t-statistic will display greater variability.

Identifying Factors

Factors that identify the t-test and estimator of μ :



Parameter of interest: μ

Test Statistic:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

v=n-1 degrees of freedom

Interval Estimator:

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

Required Condition: Population is normal (or at least "not extremely nonnormal")

Inference About Population Variance

If we are interested in drawing inferences about a ***population's variability***, the parameter we need to investigate is the population variance: σ^2

The sample variance ($\mathbf{s^2}$) is an unbiased, consistent and efficient point estimator for σ^2. Moreover,

the statistic, $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$, has a chi-squared distribution,

with n–1 degrees of freedom.

Testing & Estimating Population Variance

The test statistic used to test hypotheses about σ^2 is:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

which is chi-squared with $\nu = n-1$ degrees of freedom.

Testing & Estimating Population Variance

Combining this statistic: $\chi^2 = \dfrac{(n-1)s^2}{\sigma^2}$

With the probability statement: $P\left(\chi^2_{1-\alpha/2} < \chi^2 < \chi^2_{\alpha/2}\right) = 1-\alpha$

Yields the ***confidence interval estimator for*** σ^2 :

$$LCL = \frac{(n-1)s^2}{\chi^2_{\alpha/2}} \qquad UCL = \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

lower confidence limit — upper confidence limit

Example 12.3...

Container-filling machines are used to package a variety of liquids; including milk, soft drinks, and paint.

Ideally, the amount of liquid should vary only slightly, since large variations will cause some containers to be underfilled (cheating the customer) and some to be overfilled (resulting in costly waste).

The president of a company that developed a new type of machine boasts that this machine can fill 1 liter (1,000 cubic centimeters) containers so consistently that the variance of the fills will be less than 1.

Example 12.3

To examine the veracity of the claim, a random sample of 25 l-liter fills was taken and the results (cubic centimeters) recorded. Xm12-03

Do these data allow the president to make this claim at the 5% significance level?

Example 12.3...

IDENTIFY

The problem objective is to describe the population of l-liter fills from this machine.

The data are interval, and we're interested in the variability of the fills.

It follows that the parameter of interest is the population variance σ^2.

Example 12.3...

IDENTIFY

Because we want to determine whether there is enough evidence to support the claim, the alternative hypothesis is

$$H_1 : \sigma^2 < 1$$

The null hypothesis is

$$H_0 : \sigma^2 = 1$$

and the test statistic we will use is

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

Example 12.3...

COMPUTE

Using a calculator, we find

$$\sum x_i = 24,992.0 \quad \text{and} \quad \sum x_i^2 = 24,984,017.76$$

$$s^2 = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1} = \frac{24,984,017.76 - \frac{(24,992.0)^2}{25}}{25-1} = .6333$$

The value of the test statistic :

$$x^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(25-1)(.6333)}{1} = 15.20$$

$$x^2 < X^2_{1-\alpha,n-1} = x^2_{1-.05,25-1} = x^2_{.95,24} = 13.85$$

Example 12.3

COMPUTE

	A	B	C	D
1	**Chi Squared Test: Variance**			
2				
3				*Fills*
4	Sample Variance			0.6333
5	Hypothesized Variance			1
6	df			24
7	chi-squared Stat			15.20
8	P (CHI<=chi) one-tail			0.0852
9	chi-squared Critical one tail	Left-tail		13.85
10		Right-tail		36.42
11	P (CHI<=chi) two-tail			0.1705
12	chi-squared Critical two tail	Left-tail		12.40
13		Right-tail		39.36

Example 12.3

INTERPRET

There is not enough evidence to infer that the claim is true.

As we discussed before, the result does not say that the variance is equal to 1; it merely states that we are unable to show that the variance is less than 1.

Example 12.4

$$x^2_{\alpha/2,n-1} = x^2_{.005,24} = 45.56$$

$$x^2_{1-\alpha/2,n-1} = x^2_{.995,24} = 9.89$$

$$LCL = \frac{(n-1)s^2}{x^2_{\alpha/2}} = \frac{15.20}{45.56} = .3336$$

$$UCL = \frac{(n-1)s^2}{x^2_{1-\alpha/2}} = \frac{15.20}{9.89} = 1.537$$

Example 12.4

Estimate with 99% confidence the variance of fills in Example 12.3. Xm12-03

	A	B
1	**Chi Squared Estimate: Variance**	
2		
3		*Fills*
4	Sample Variance	0.6333
5	df	24
6	LCL	0.3336
7	UCL	1.5375

Example 12.4...

INTERPRET

In Example 12.3, we saw that there was not sufficient evidence to infer that the population variance is less than 1.

Here we see that is estimated to lie between .3336 and 1.5375.

Part of this interval is above 1, which tells us that the variance may be larger than 1, confirming the conclusion we reached in Example 12.3.

Example 12.4...

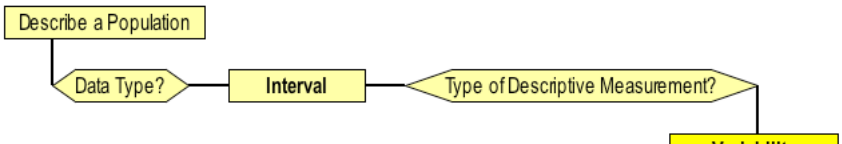
INTERPRET

We may be able to use the estimate to predict the percentage of overfilled and underfilled bottles.

This may allow us to choose among competing machines.

Identifying Factors...

Factors that identify the chi-squared test and estimator of σ^2:



Parameter of interest: σ^2

Test Statistic:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

Interval Estimator:

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

ν=n-1 degrees of freedom

Required Condition: Population is normal

Inference: Population Proportion...

When data are nominal, we count the number of occurrences of each value and calculate proportions. Thus, the parameter of interest in describing a population of nominal data is the population proportion ***p***.

This parameter was based on the binomial experiment.

Recall the use of this statistic: $\hat{p} = \frac{x}{n}$

where p-hat ($\hat{p}$) is the sample proportion: **x** successes in a sample size of **n** items.

Inference: Population Proportion...

When ***np*** and ***n(1–p)*** are both greater than 5, the sampling distribution of $\hat{p}$ is approximately normal with

mean: $\mu = p$

standard deviation: $\sigma = \sqrt{\dfrac{p(1-p)}{n}}$

Hence: $z = \dfrac{\hat{p} - p}{\sqrt{p(1-p)/n}}$

Inference: Population Proportion

Test statistic for ***p***:

$$z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$$

The confidence interval estimator for ***p*** is given by:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$$

(both of which require that ***np***>5 and ***n(1–p)***>5)

Example 12.5

After the polls close on election day networks compete to be the first to predict which candidate will win.

The predictions are based on counts in certain precincts and on exit polls.

Exit polls are conducted by asking random samples of voters who have just exited from the polling booth (hence the name) for which candidate they voted.

Example 12.5

In American presidential elections the candidate who receives the most votes in a state receives the state's entire Electoral College vote.

In practice, this means that either the Democrat or the Republican candidate will win.

Suppose that the results of an exit poll in one state were recorded where 1 = Democrat and 2 = Republican.

Xm12-05*

Example 12.5

The polls close at 8:00 P.M.

Can the networks conclude from these data that the Republican candidate will win the state?

Should the network announce at 8:01 P.M. that the Republican candidate will win?

Example 12.5

IDENTIFY

The problem objective is to describe the population of votes in the state. The data are nominal because the values are "Democrat" and "Republican." Thus the parameter to be tested is the proportion of votes in the entire state that are for the Republican candidate. Because we want to determine whether the network can declare the Republican to be the winner at 8:01 P.M., the alternative hypothesis is

$H_1: p > .50$

And hence our null hypothesis becomes:

$H_0: p = .50$

Example 12.5

IDENTIFY

The test statistic is

$$z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$$

Example 12.5

COMPUTE

The rejection region is

$z > z_\alpha = z_{.05} = 1.645$

$$\hat{p} = \frac{x}{n} = \frac{407}{765} = .532$$

$$z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} = \frac{.532 - .5}{\sqrt{.5(1-.5)/765}} = 1.77$$

$$p\text{-value} = p(z > 1.77) = 1 - p(z < 1.77) = 1 - .9616 = .0384$$

Example 12.5

COMPUTE

	A	B	C	D
1	**z-Test: Proportion**			
2				
3				*Votes*
4	Sample Proportion			0.532
5	Observations			765
6	Hypothesized Proportion			0.5
7	z Stat			1.77
8	P(Z<=z) one-tail			0.0382
9	z Critical one-tail			1.6449
10	P(Z<=z) two-tail			0.0764
11	z Critical two-tail			1.96

Example 12.5

INTERPRET

At the 5% significance level we reject the null hypothesis and conclude that there is enough evidence to infer that the Republican candidate will win the state.

However, is this the right decision?

One of the key issues to consider here is the cost of Type I and Type II errors.

A Type I error occurs if we conclude that the Republican will win when in fact he has lost.

Example 12.5

INTERPRET

Such an error would mean that a network would announce at 8:01 P.M. that the Republican has won and then later in the evening would have to admit to a mistake.

If a particular network were the only one that made this error it would cast doubt on their integrity and possibly affect the number of viewers.

Example 12.5

INTERPRET

This is exactly what happened on the evening of the U. S. presidential elections in November 2000.

Shortly after the polls closed at 8:00 P.M. all the networks declared that the Democratic candidate Albert Gore would win in the state of Florida.

A couple of hours later, the networks admitted that a mistake had been made and the Republican candidate George W. Bush had won.

Example 12.5

INTERPRET

Several hours later they again admitted a mistake and finally declared the race too close to call.

Fortunately for each network all the networks made the same mistake.

However, if one network had not done this it would have developed a better track record, which could have been used in future advertisements for news shows and likely drawn more viewers.

Considering the costs of Type I and II errors it would have been better to use a 1%significance level.

Estimating Totals for Large Populations

In much the same way as we saw earlier, when a population is ***large*** and ***finite*** we can estimate the total number of successes in the population by taking the product of the size of the population (**N**) and the confidence interval estimator:

$$N\left[\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p}}{n}}\right]$$

The Nielsen Ratings (used to measure TV audiences) uses this technique. Results from a small sample audience (5,000 viewers) is extrapolated to the total number of TV households (110 million).

Nielsen Ratings

Statistical techniques play a vital role in helping advertisers determine how many viewers watch the shows that they sponsor.

Although several companies sample television viewers to determine what shows they watch, the best known is the A. C. Nielsen firm.

The Nielsen ratings are based on a random sample of approximately 5,000 of the 110 million households in the United States with at least one television (in 2007).

Nielsen Ratings

A meter attached to the televisions in the selected households keeps track of when the televisions are turned on and what channels they are tuned to.

The data are sent to the Nielsen's computer every night from which Nielsen computes the rating and sponsors can determine the number of viewers and the potential value of any commercials.

Nielsen Ratings

The results from Sunday, April 1, 2007 for the time slot 9:00 to 9:30 P.M. have been recorded using the following codes:

Network	Show	Code
ABC	Desperate Housewives	1
CBS	The Amazing Race 11	2
NBC	Deal or No Deal	3
Fox	Family Guy	4
Television turned off or watched another channel		5

Nielsen Ratings

IDENTIFY

The problem objective is to describe the population of television shows watched by viewers across the country.

The data are nominal.

The combination of problem objective and data type make the parameter to be estimated the proportion of the entire population that watched the "Deal or No Deal."

The confidence interval estimator of the proportion is

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Nielsen Ratings Example

COMPUTE

	A	B
1	**z-Estimate: Proportion**	
2		*Program*
3	Sample Proportion	0.0836
4	Observations	5000
5	LCL	0.0759
6	UCL	0.0913

Nielsen Ratings

INTERPRET

We estimate that between 7.59% and 9.13% of all television sets had received "Deal or No Deal."

If we multiply these figures by the total number of televisions, 110 million, we produce an interval estimate of the number of televisions tuned to "Deal or No Deal".

Nielsen Ratings

INTERPRET

Thus, the 95% confidence interval estimate of the number of televisions tuned to "Deal or No Deal" lies between

LCL = .0759 (110 million) = 8.349 million

and

UCL = .0913 (110 million) = 10.043 million

Sponsoring companies can then determine the value of any commercials that appeared on the show.

Selecting the Sample Size

When we introduced the sample size selection method to estimate a mean in Section 10.3, we pointed out that the sample size depends on the confidence level and the bound on the error of estimation that the statistics practitioner is willing to tolerate.

When the parameter to be estimated is a proportion the bound on the error of estimation is

$$B = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Selecting the Sample Size

Solving for n we produce the required sample size to estimate p and where B is the bound on the error of Estimation

$$n = \left(\frac{z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})}}{B} \right)^2$$

Unfortunately we do not know the value of $\hat{p}$

Selecting the Sample Size

Two methods – in each case we choose a value for $\hat{p}$ then solve the equation for n.

Method 1 : no knowledge of even a rough value of $\hat{p}$ This is a 'worst case scenario' so we substitute $\hat{p} = .50$

Method 2 : we have some idea about the value of $\hat{p}$ This is a better scenario and we substitute in our estimated $\hat{p}$ value.

Selecting the Sample Size

Method 1 :: no knowledge of value of $\hat{p}$, use 50%:

$$n = \left(\frac{1.96\sqrt{.50(1-.50)}}{.03} \right)^2 = 1{,}068$$

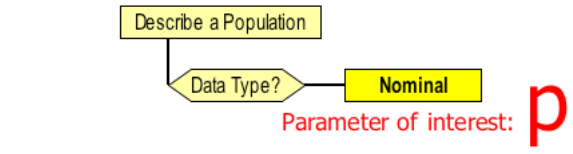
Method 2 :: some idea about a possible $\hat{p}$ value, say 20%:

$$n = \left(\frac{1.96\sqrt{.20(1-.20)}}{.03} \right)^2 = 683$$

Thus, we can sample fewer people if we already have a reasonable estimate of the population proportion before starting.

Identifying Factors

Factors that identify the z-test and interval estimator of ***p***:



Test Statistic:

$$z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$$

Interval Estimator:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Required Conditions: $np \geq 5$ *and* $n(1-p) \geq 5$ *(for test)*

$n\hat{p} \geq 5$ *and* $n(1-\hat{p}) \geq 5$ *(for estimate)*

APPLICATIONS IN MARKETING: Market Segmentation

Mass marketing refers to the mass production and marketing by a company of a single product for the entire market.

Mass marketing is especially effective for commodity goods such as gasoline, which are very difficult to differentiate from the competition,

It has given way to target marketing, which focuses on satisfying the demands of a particular segment of the entire market.

APPLICATIONS IN MARKETING: Market Segmentation

Because there is no single way to segment a market, managers must consider several different variables (or characteristics) that could be used to identify segments.

Surveys of customers are used to gather data about various aspects of the market, and statistical techniques are applied to define the segments.

Market segmentation separates consumers of a product into different groups in such a way that members of each group are similar to each other and there are differences between groups.

APPLICATIONS IN MARKETING: Market Segmentation

There are many ways to segment a market.

Table 12.1 lists several different segmentation variables and their market segments.

Table 12.1 Market Segmentation

Segmentation variable	Segments
Geographic	
Countries	Brazil, Canada, China, France, United States
Country regions	Midwest, Northeast, Southwest, Southeast
Demographic	
Age	Under 5, 5- 12, 13-19, 20-29, 30-50, over 50
Education	Some high school, high school graduate, some college, college or university graduate
Income	Under $30,000, $30,000-49,999, $50,000- 79,999, over $80,000
Marital status	Single, married, divorced, widowed

APPLICATIONS IN MARKETING: Market Segmentation

It is important for marketing managers to know the size of the segment because the size (among other parameters) determines its profitability.

Not all segments are worth pursuing. In some instances the size of the segment is too small or the costs of satisfying it may be too high.

APPLICATIONS IN MARKETING: Market Segmentation

The size can be determined in several ways.

The census provides useful information.

For example, we can determine the number of Americans in various age categories or the size of geographic residences.

For other segments we may need to survey members of a general population and use the inferential techniques introduced in the previous section where we showed how to estimate the total number of successes.

APPLICATIONS IN MARKETING: Market Segmentation

We can survey large populations to estimate the proportion of the population that fall into each segment.

From these estimates we can estimate the size of markets using the confidence interval estimator

$$N\left(\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p}}{n}}\right)$$

Example 12.6

In segmenting the breakfast cereal market a food manufacturer uses health and diet consciousness as the segmentation variable. Four segments are developed:

1. Concerned about eating healthy foods
2. Concerned primarily about weight
3. Concerned about health because of illness
4. Unconcerned

Example 12.6

To distinguish between groups surveys are conducted. On the basis of a questionnaire people are categorized as belonging to one of these groups.

A recent survey asked a random sample of 1250 American adults (20 and over) to complete the questionnaire. The categories were recorded using the codes. Xm12-06

The most recent census reveals that there are 207,347,000 Americans who are 20 and over.

Estimate with 95% confidence the number of American adults who are concerned about eating healthy foods.

Example 12.6

IDENTIFY

The problem objective is to describe the population of American adults' concerns about health and diet.

The data are nominal. Consequently, the parameter we wish to estimate is the proportion p of American adults who classify themselves as concerned about eating healthy (code = 1).

The confidence interval estimator we need to employ is

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

from which we will produce the estimate of the size of the market segment.

Example 12.6

COMPUTE

	A	B
1	**z-Estimate: Proportion**	
2		*Group*
3	Sample Proportion	0.2152
4	Observations	1250
5	LCL	0.1924
6	UCL	0.2380

Example 12.6

COMPUTE

We multiply the lower and upper limits by the population size.

$$\text{LCL} = N\left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] = 207{,}347{,}000\ (.1924)$$

$$= 39{,}893{,}563$$

and

$$\text{UCL} = N\left[\hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] = 207{,}347{,}000\ (.2380)$$

$$= 49{,}348{,}586$$

Example 12.6

INTERPRET

We estimate that the size of this market segment lies between 39,893,563 and 49,348,586.