One Sample

MATH 2441, BCIT

Technical Mathematics for Food Technology

March 8, 2017

Estimating a Population Proportion

When you take a sample, you can try to estimate population parameters such as mean, variance, and proportion on the basis of your sample data. Let's think about population proportions first.

point estimate The sample proportion (denoted by $\hat{p}$) is the best point estimate of the population proportion p.

confidence interval We can use a sample proportion to construct a confidence interval for the true value of a population proportion.

sample size We can find the sample size necessary to estimate a population proportion to a given degree of accuracy.

A point estimate is a single value (or point) used to approximate a population parameter.

Estimating a Population Proportion

Example 1: Yoghurt Expiry. You test 243 yoghurts three days after their expiry date and find that 19 of them are spoiled. Find the best point estimate of the percentage of this kind of yoghurt that spoils by day three after the expiry date. Answer: the best point estimate is 19/243 or approximately 7.82%.

confidence interval
: A confidence interval is a range of values used to estimate the true value of a population parameter.

confidence level
: The confidence level is the probability $1 - \alpha$ (such as 0.95 or 95%) that the confidence interval actually contains the population parameter. The confidence level is also called the degree of confidence or the confidence coefficent.

Estimating a Population Proportion

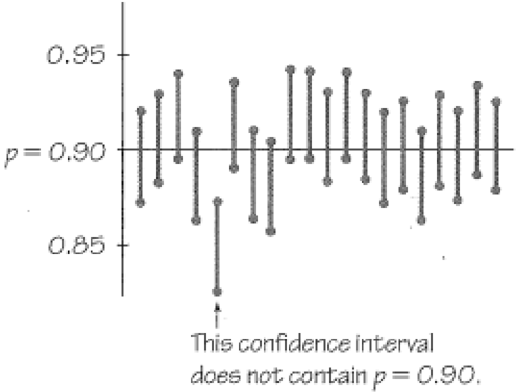


Figure 7-1 Confidence Intervals from 20 Different Samples

Estimating a Population Proportion

Interpreting a Confidence Interval

We must be careful to interpret confidence intervals correctly. There is a correct interpretation and many different and creative incorrect interpretations of the confidence interval $0.828 < p < 0.872$.

Correct: "We are 95% confident that the interval from 0.828 to 0.872 actually does contain the true value of the population proportion p." This means that if we were to select many different samples of size 1007 and construct the corresponding confidence intervals, 95% of them would actually contain the value of the population proportion p. (In this correct interpretation, the confidence level of 95% refers to the *success rate of the process* used to estimate the population proportion.)

Incorrect: "There is a 95% chance that the true value of p will fall between 0.828 and 0.872."

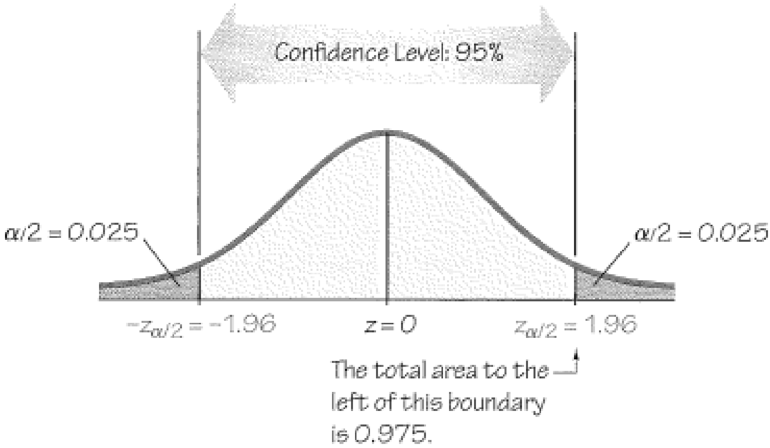
Incorrect: "95% of sample proportions will fall between 0.828 and 0.872."

Estimating a Population Proportion

critical value A critical value is the number on the borderline separating sample statistics that are likely to occur from those that are unlikely. The number $z_{\alpha/2}$ is a critical value that is a z-score with the property that it separates an area of $\alpha/2$ in the right tail of the standard normal distribution.

margin of error When data from a simple random sample are used to estimate a population proportion p, the margin of error, denoted by E, is the maximum likely difference (with probability $1-\alpha$, such as 0.95) between the observed sample probability $\hat{p}$ and the true value of the population proportion p.

Estimating a Population Proportion



Estimating a Population Proportion

Confidence Level	α	Critical Value, $z_{\alpha/2}$
90%	0.10	1.645
95%	0.05	1.96
99%	0.01	2.575

Estimating a Population Proportion

Formula for Margin of Error

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} \tag{1}$$

The confidence interval is $\hat{p} - E < p < \hat{p} + E$.

Sample Size to Estimate a Population Proportion

Objective: Determine how large the sample n should be in order to estimate the population proportion p

Notation: p is the population proportion; $\hat{p}$ is the sample proportion; n is the number of sample values; E is the desired margin of error; and $z_{\alpha/2}$ is the z-score separating an area of $\alpha/2$ in the right tail of the standard normal distribution

Required: The sample must be a simple random sample of independent sample units

Sample Size to Estimate a Population Proportion

estimate $\hat{p}$ is known

$$n = \frac{\left(z_{\alpha/2}\right)^2 \hat{p}\hat{q}}{E^2}$$

estimate $\hat{p}$ is not known

$$n = \frac{\left(z_{\alpha/2}\right)^2 \cdot 0.25}{E^2}$$

If the computed sample size n is not a whole number, round the value of n up to the next larger whole number.

Exercise 1: A company is interested in the percentage of adults who buy clothing online. How many adults must be surveyed in order to be 95% confident that the sample percentage is in error by no more than three percentage points?

1. Use this recent result from Statistics Canada: 66% of adults buy clothing online.
2. Assume that you have no prior information suggesting a possible value of the proportion.

Exercises

Exercise 2: A poll asks respondents if they felt vulnerable to identity theft. The results are as follows: $n = 1002, x = 531$, where x is the number of people responding with "yes." Construct the confidence interval, using a 95% confidence level. Constructing a confidence interval means:

1. find the point estimate for the population proportion
2. identify the value of the margin of error
3. identify the confidence interval, for example $0.51 < p < 0.63$

Exercise 3: From a poll in which respondents were asked to identify their favourite seat when they fly: $n = 806, x = 492$, where x is the number of people choosing the window seat. Construct the confidence interval, using a 99% confidence level.

Exercise 4: A company devises a method to increase the probability that a baby is a girl or a boy (we shall call this privileged sex a "birl"). 879 out of 945 babies born to parents using this method are birls.

1. What is the best point estimate of the population proportion of birls born to parents using this method?
2. Use the sample data to construct a 95% confidence interval estimate of the proportion of birls born to parents using this method.
3. Is the method effective?

Exercise 5: You plan to develop new softward and need to know how many people use Microsoft Windows. How many computers must be surveyed in order to be 99% confident that your estimate is in error by no more than one percentage point?

1. Assume that nothing is known about the percentage of computers with Windows operating systems.
2. Assume that a recent survey suggests that about 90% of computers use Windows operating systems.

Estimating a Population Mean

Point Estimate The sample mean $\bar{x}$ is the best point estimate of the population mean μ

Confidence Interval We can use a sample mean to construct a confidence interval estimate of the true value of a population mean

Sample Size We can find the sample size necessary to estimate a population mean within certain parameters

To construct the confidence interval, we need a normal distribution of the sample mean. According to the Central Limit Theorem, we are justified in assuming this if either the underlying distribution is normal or the sample size $n > 30$.

Unknown Population Standard Deviation

We wouldn't usually know the population standard deviation σ if were trying to find out the population mean μ. We use the sample mean s to approximate σ, but at a cost. To adjust for this assumption, we must use Gosset's Student t distribution instead of the standard normal distribution.

The Student t Distribution

If a population has a normal distribution, then the distribution of

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \tag{2}$$

is a Student t distribution for all samples of size n.

Degrees of Freedom

Finding a critical value $t/2$ requires a value for the degree of freedom, often labelled df. The df is the number of individuals in a sample that can vary freely if the mean is known, so

$$df = n - 1 \tag{3}$$

Confidence Interval for Mean with σ Not Known

Objective: Construct a confidence interval used to estimate a population mean

Notation: μ is the population mean; $\bar{x}$ is the sample mean; s is the sample standard deviation; n is the sample size; and E is the margin of error

Required: A simple random sample and either or both of these conditions fulfilled: the population is normally distributed or $n > 30$.

Confidence Interval for Mean with σ Not Known

confidence interval for mean, σ not known

$$\bar{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

with $df = n - 1$.

Sample Size Required to Estimate a Population Mean

Objective Determine the sample size n required to estimate the value of a population mean μ

Notation μ is the population mean; $\bar{x}$ is the sample mean; σ is the population standard deviation; and E is the desired margin of error; and $z_{\alpha/2}$ is the z-score separating an area of $\alpha/2$ in the right tail of the standard normal distribution

Required A simple random sample

The required sample size is found by using the formula

$$n = \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2 \tag{4}$$

Sample Size Required to Estimate a Population Mean

Here are ways to deal with the fact that σ is usually not known.

1. Use the range rule of thumb to estimate the standard deviation as follows: $\sigma \approx$range/4. With a sample of 87 or more values randomly selected from a normally distributed population, this approximation will yield a value that is greater than or equal to σ at least 95% of the time.
2. Start the sample collection process and use s instead of σ. You can improve s as you go along.
3. Estimate the value of σ using prior information.

Exercise 6: How many statistics students must be randomly selected for IQ tests if we want 95% confidence that the sample mean is within 3 IQ points of the population mean? Assume that $\sigma = 15$.

Confidence Interval for Mean with σ Known

Objective: Construct a confidence interval used to estimate a population mean

Notation: μ is the population mean; $\bar{x}$ is the sample mean; σ is the population standard deviation; n is the sample size; and E is the margin of error

Required: A simple random sample and either or both of these conditions fulfilled: the population is normally distributed or $n > 30$.

Confidence Interval for Mean with σ Known

confidence interval for mean, σ known

$$\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

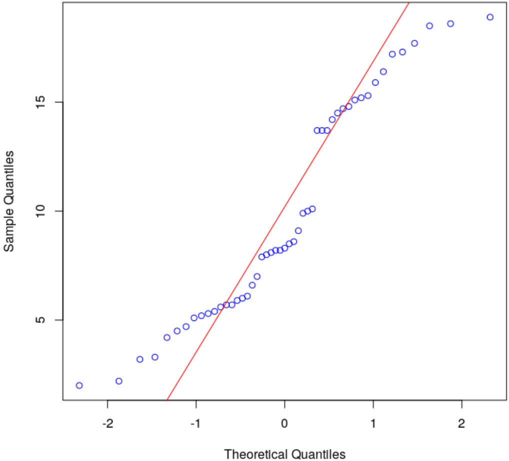
Exercises

Exercise 7: Consider the following data for the depth of earthquakes (in kilometres):

6.6	2.0	15.3	17.2	3.2	2.2	14.8
5.6	6.1	9.1	18.5	8.1	10.0	13.7
8.0	7.0	18.6	8.2	5.7	18.9	13.7
4.5	8.3	6.0	14.2	5.4	17.7	9.9
17.3	5.1	5.3	15.9	13.7	4.2	5.7
5.9	15.1	8.5	14.7	16.4	4.7	8.6
8.2	15.2	10.1	14.5	5.2	7.9	3.3

Exercises

The mean of this dataset is approximately 9.878, the standard deviation is approximately 5.0409. Construct a 98% confidence interval estimate of the mean depth. The data does not appear to be normally distributed.



Exercise 8: Listed below are the amounts of mercury (in parts per million, or ppm) found in tuna sushi sampled at different stores in New York City. The study was sponsored by the New York Times, and the stores (in order) are D'Agostino, Eli's Manhattan, Fairway, Food Emporium, Gourmet Garage, Grace's Marketplace, and Whole Foods. The sample mean is 0.719 ppm and the standard deviation is 0.366 ppm. Construct a 90% confidence interval estimate of the mean amount of mercury in the population.

```
0.50 0.75 0.10 0.95 1.25 0.54 0.88
```

Exercise 9: Here are the numbers of chocolate chips in a sample of 40 Chips Ahoy regular cookies. The mean is 23.95 chocolate chips and the standard deviation is 2.55 chocolate chips. Construct a 99% confidence interval estimate of the mean number of chocolate chips in all such cookies. How does the confidence interval not contradict the fact that most of the original values do not fall between the confidence interval limits?

Exercise 10: The following data describes a sample of 106 body temperatures having a mean of 98.20°F and a standard deviation of 0.62°F. Construct a 95% confidence interval estimate of the mean body temperature for the entire population. What does the result suggest about the common belief that 98.6°F is the mean body temperature?

Body Temperature Data

98.6	98.6	98.0	98.0	99.0	98.4	98.4	98.4	98.4	98.6
98.6	98.8	98.6	97.0	97.0	98.8	97.6	97.7	98.8	98.0
98.0	98.3	98.5	97.3	98.7	97.4	98.9	98.6	99.5	97.5
97.3	97.6	98.2	99.6	98.7	99.4	98.2	98.0	98.6	98.6
97.2	98.4	98.6	98.2	98.0	97.8	98.0	98.4	98.6	98.6
97.8	99.0	96.5	97.6	98.0	96.9	97.6	97.1	97.9	98.4
97.3	98.0	97.5	97.6	98.2	98.5	98.8	98.7	97.8	98.0
97.1	97.4	99.4	98.4	98.6	98.4	98.5	98.6	98.3	98.7
98.8	99.1	98.6	97.9	98.8	98.0	98.7	98.5	98.9	98.4
98.6	97.1	97.9	98.8	98.7	97.6	98.2	99.2	97.8	98.0
98.4	97.8	98.4	97.4	98.0	97.0				

Exercise 11: In a test of weight loss programs, 40 adults used the Atkins weight loss program. After 12 months, their mean weight loss was found to be 2.1 lb, with a standard deviation of 4.8 lb. Construct a 90% confidence interval estimate of the mean weight loss for all such subjects. Does the Atkins program appear to be effective? Does it appear to be practical?

Exercise 12: The Wechsler IQ test is designed so that the mean is 100 and the standard deviation is 15 for the population of normal adults. Find the sample size necessary to estimate the mean IQ score of professional pilots. We want to be 90% confident that our sample mean is within 3 IQ points of the true mean. The mean for this population is clearly greater than 100. The standard deviation for this population is probably less than 15 because it is a group with less variation than a group randomly selected from the general population; therefore, if we use $\sigma = 15$ we are being conservative by using a value that will make the sample size at least as large as necessary. Assume then that $\sigma = 15$ and determine the required sample size. Does the sample size appear to be reasonable?

Exercise 13: As part of a study of grade inflation, you want to estimate the mean grade point average of all current college students in the United States. All grade point averages are to be standardized for a scale between 0 and 4. How many grade point averages must be obtained so that the sample mean is within 0.2 of the population mean? Assume that a 99% confidence level is desired. Also assume that a pilot study showed that the population standard deviation is estimated to be 0.79.

Estimating a Population Variance

Point Estimate The sample variance s^2 is the best point estimate of the population variance σ^2. The sample standard deviation s is commonly used as a point estimate of σ even though it is a biased estimator.

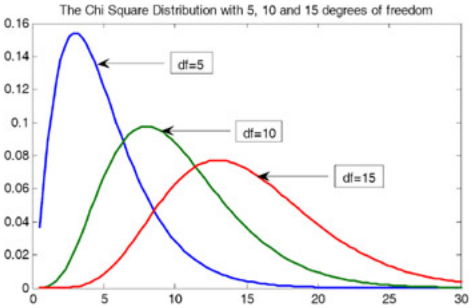
Confidence Interval When constructing a confidence interval estimate of a population standard deviation or population variance, we construct the confidence interval using the χ^2 (chi-square) distribution.

Chi-Square Distribution

The sample statistic

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \tag{5}$$

has a sampling distribution called the chi-square distribution. The chi-square distribution is skewed, so the right tail and the left tail are not symmetrical. We call the critical value for the right tail χ^2_R and the critical value for the left tail χ^2_L.



Confidence Interval for Population Variance

Objective	Construct a confidence interval used to estimate a population standard deviation or variance
Notation	σ is the population standard deviation; σ^2 is the population variance; s is the sample standard deviation; s^2 is the sample variance; n is the sample size; and E is the margin of error
Required	A simple random sample and a normally distributed population even if the sample is large!

Confidence Interval for Population Variance

confidence interval for population variance

$$\frac{(n-1)s^2}{\chi_R^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_L^2} \tag{6}$$

confidence interval for population standard deviation

$$\sqrt{\frac{(n-1)s^2}{\chi_R^2}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi_L^2}}$$

with $df = n - 1$.

Exercise 14: The following data describes a sample of 106 body temperatures having a mean of 98.20°F and a standard deviation of 0.62°F. Construct a 90% confidence interval estimate of the standard deviation of the body temperature for the entire population.

Body Temperature Data

98.6	98.6	98.0	98.0	99.0	98.4	98.4	98.4	98.4	98.6
98.6	98.8	98.6	97.0	97.0	98.8	97.6	97.7	98.8	98.0
98.0	98.3	98.5	97.3	98.7	97.4	98.9	98.6	99.5	97.5
97.3	97.6	98.2	99.6	98.7	99.4	98.2	98.0	98.6	98.6
97.2	98.4	98.6	98.2	98.0	97.8	98.0	98.4	98.6	98.6
97.8	99.0	96.5	97.6	98.0	96.9	97.6	97.1	97.9	98.4
97.3	98.0	97.5	97.6	98.2	98.5	98.8	98.7	97.8	98.0
97.1	97.4	99.4	98.4	98.6	98.4	98.5	98.6	98.3	98.7
98.8	99.1	98.6	97.9	98.8	98.0	98.7	98.5	98.9	98.4
98.6	97.1	97.9	98.8	98.7	97.6	98.2	99.2	97.8	98.0
98.4	97.8	98.4	97.4	98.0	97.0				

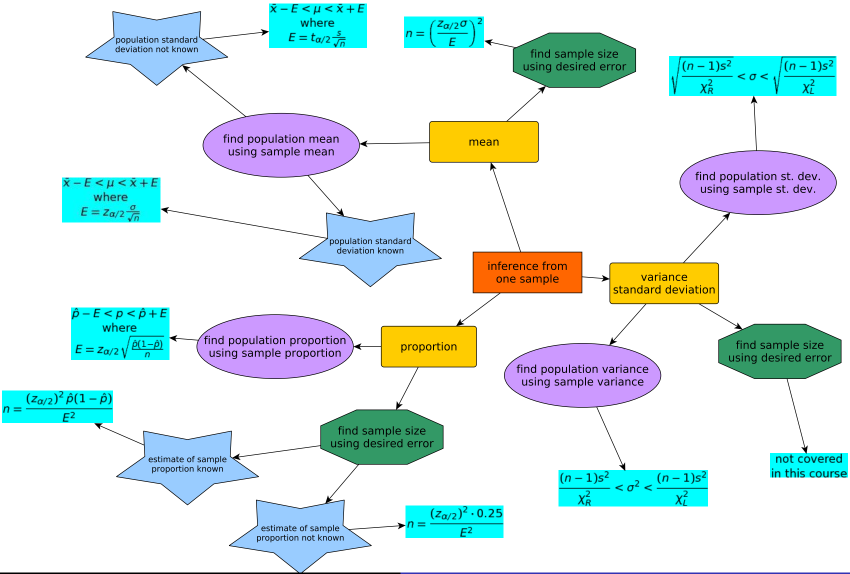
Exercise 15: A container of car antifreeze is supposed to hold 3785 mL of the liquid. Realizing that fluctuations are inevitable, the quality control manager wants to be sure that the standard deviation is less than 30 mL. Otherwise, some containers would overflow while others would not have enough of the coolant. She selects a simple random sample of 24 containers and finds that the mean is 3789 mL and the standard deviation is 42.8 mL. Use these sample results to construct the 99% confidence interval for the true value of σ. Does this confidence level suggest that the variation is at an acceptable level?

Exercise 16: Use the following data set (next slide) of the weight of post-1983 pennies to construct a 98% confidence interval estimate of the standard deviation of the weights of all post-1983 pennies. Assume that the weights are normally distributed. The sample mean is $\bar{x} = 2.4988$, the sample standard deviation is $s = 0.016636$.

Exercises

2.5113	2.4907	2.5024	2.5298	2.4950	2.5127
2.4998	2.4848	2.4823	2.5163	2.5222	2.5004
2.5248	2.5058	2.4900	2.5068	2.5016	2.4797
2.5067	2.5139	2.4762	2.5004	2.5170	2.4925
2.4876	2.4933	2.4806	2.4907	2.5017	2.4950
2.4973	2.5252	2.4978	2.5073	2.4658	2.4529

One Sample Flow Chart



One Sample Summary

	Proportion	Mean	Variance/StDev
Central Limit Theorem	yes	yes	no
Population	p	μ	σ^2 / σ
Sample	$\hat{p}$	$\bar{x}$	s^2 / s
Requirements	n>30	n>30 or x normally distributed	x normally distributed
Distribution	normal	normal	chi-squared distribution
Critical Values	$z_{\alpha/2}$	$t_{\alpha/2}$ (if σ unknown, otherwise $z_{\alpha/2}$)	χ^2_L and χ^2_R
Parameters of Distribution	$\mu_{\hat{p}}=p$	$\mu_{\bar{x}}=\mu$	
	$\sigma_{\hat{p}}=\sqrt{\frac{p(1-p)}{n}}$	$\sigma_{\bar{x}}=\frac{\sigma}{\sqrt{n}}$	
		$df=n-1$	$df=n-1$
Confidence Interval	$\hat{p}-E<p<\hat{p}+E$	$\bar{x}-E<\mu<\bar{x}+E$	$\frac{(n-1)s^2}{\chi^2_R}<\sigma^2<\frac{(n-1)s^2}{\chi^2_L}$
Margin of Error	$E=z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$	$E=t_{\alpha/2}\cdot\frac{s}{\sqrt{n}}$	n/a
Sample Size Determination	$\frac{(z_{\alpha/2})^2\hat{p}\hat{q}}{E^2}$ $\frac{(z_{\alpha/2})^2\cdot 0.25}{E^2}$	$\left(\frac{z_{\alpha/2}\sigma}{E}\right)^2$	n/a

End of Lesson

Next Lesson: Hypothesis Testing