QM353
Business Statistics

Chapter 2
Sampling Distributions
and Estimating Population
Parameters

Chapter Goals

After completing this chapter, you should be able to:

- Define the concept of sampling error
- Determine the mean and standard deviation for the sampling distribution of the sample mean, $\bar{x}$
- Determine the mean and standard deviation for the sampling distribution of the sample proportion, $\bar{p}$
- Describe the Central Limit Theorem and its importance
- Apply sampling distributions for both $\bar{x}$ and $\bar{p}$

Chapter Goals

(continued)

- Distinguish between a point estimate and a confidence interval estimate
- Construct and interpret a confidence interval estimate for a single population mean using both the z and t distributions
- Form and interpret a confidence interval estimate for a single population proportion

Sampling Error

- **Sample Statistics are used to estimate Population Parameters**

  ex: $\bar{X}$ is an estimate of the population mean, μ

Problems:

- **Different samples provide different estimates of the population parameter**
- **Sample results have potential variability, thus sampling error exits**

Recall: With a random sample the goal is to gather a *representative* group from the population

Calculating Sampling Error

- Sampling Error:

  The difference between a value (a statistic) computed from a sample and the corresponding value (a parameter) computed from a population

Example: (for the mean)

$$\text{Sampling Error} = \bar{x} - \mu$$

where:

$\bar{x}$ = sample mean
μ = population mean

Always present just because you sample!

Review

- Population mean: Sample Mean:

$$\mu = \frac{\sum x_i}{N} \qquad \bar{x} = \frac{\sum x_i}{n}$$

See Chpt. 3

Population mean does NOT vary

Sample mean can vary when different samples are collected from the population

where:

μ = Population mean
$\bar{x}$ = sample mean
x_i = Values in the population or sample
N = Population size
n = sample size

Example

If the population mean is $\mu = 98.6$ degrees and a sample of n = 5 temperatures yields a sample mean of $\overline{x} = 99.2$ degrees, then the sampling error is

$$\boxed{\bar{x} - \mu = 99.2 - 98.6 = 0.6 \text{ degrees}}$$

Sampling Errors

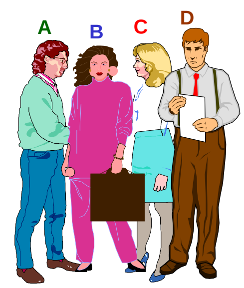
- Different samples will yield different sampling errors
- The sampling error may be positive or negative ($\overline{x}$ may be greater than or less than μ)
- The size of the error depends on the sample selected
  - i.e., a larger sample does not necessarily produce a smaller error if it is not a representative sample

Sampling Distribution

A sampling distribution is a distribution of the probability of possible values of a statistic for a given size sample selected from a population

Developing a Sampling Distribution

- **Assume there is a population …**
- Population size N=4
- Random variable, x, is age of individuals
- Values of x: 18, 20, 22, 24 (years)



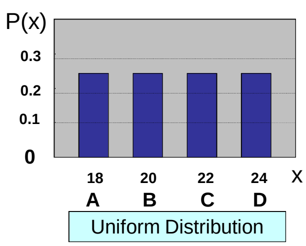
Developing a Sampling Distribution

(continued)

Summary Measures for the Population Distribution:

$$\mu = \frac{\sum x_i}{N} = \frac{18 + 20 + 22 + 24}{4} = 21$$

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}} = 2.236$$



Developing a Sampling Distribution

(continued)

Now consider all possible samples of size n=2

1st Obs	2nd Observation 18	20	22	24
18	18,18	18,20	18,22	18,24
20	20,18	20,20	20,22	20,24
22	22,18	22,20	22,22	22,24
24	24,18	24,20	24,22	24,24

16 possible samples (sampling with replacement)

16 Sample Means

1st Obs	2nd Observation 18	20	22	24
18	18	19	20	21
20	19	20	21	22
22	20	21	22	23
24	21	22	23	24

Developing a Sampling Distribution

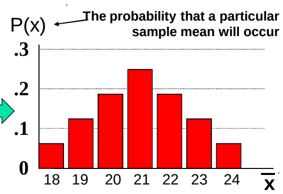
(continued)

Sampling Distribution of All Sample Means

16 Sample Means

1st Obs	2nd Observation 18	20	22	24
18	18	19	20	21
20	19	20	21	22
22	20	21	22	23
24	21	22	23	24

Sample Means Distribution



(no longer uniform)

Developing a Sampling Distribution

(continued)

Summary Measures of this Sampling Distribution:

$$\mu_{\bar{x}} = \frac{\sum \bar{X}_i}{N} = \frac{18+19+21+\cdots+24}{16} = 21$$

Average of the averages

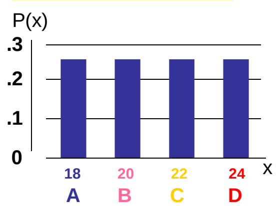
$$\sigma_{\bar{x}} = \sqrt{\frac{\sum(\bar{X}_i - \mu_{\bar{x}})^2}{N}}$$

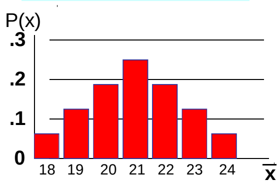
$$= \sqrt{\frac{(18-21)^2+(19-21)^2+\cdots+(24-21)^2}{16}} = 1.58$$

Comparing the Population with its Sampling Distribution

Population
N = 4
$\mu = 21 \quad \sigma = 2.236$

Sample Means Distribution
n = 2
$\mu_{\bar{x}} = 21 \quad \sigma_{\bar{x}} = 1.58$





Properties of a Sampling Distribution

- For any population,
  - the average value of all possible sample means computed from all possible random samples of a given size from the population is equal to the population mean:

Considered an "unbiased" estimator $\longrightarrow$ $\mu_{\bar{x}} = \mu$ Theorem 1

  - The standard deviation of the possible sample means computed from all random samples of size n is equal to the population standard deviation divided by the square root of the sample size:

Also called the standard error $\longrightarrow$ $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ Theorem 2

If the Population is Normal

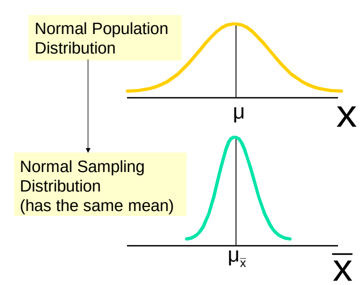
If a population is normal with mean μ and standard deviation σ, the sampling distribution of $\bar{X}$ is also normally distributed with

$\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ Theorem 3

As n increases the data *behaves* more like a normal distribution

Sampling Distribution Properties

- The sample mean is an unbiased estimator



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

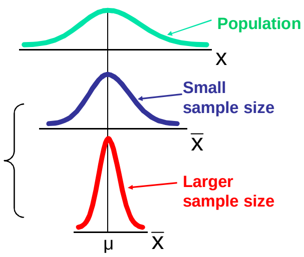
Sampling Distribution Properties

(continued)

- The sample mean is a consistent estimator

(the value of $\bar{x}$ becomes closer to μ as n increases):

As n increases, $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ decreases



z-value for Sampling Distribution of $\bar{x}$

- z-value for the sampling distribution of $\bar{x}$:

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

where:
$\bar{x}$ = sample mean
μ = population mean
σ = population standard deviation
n = sample size

Finite Population Correction

- Apply the **Finite Population Correction** if:
  - The sample is large relative to the population

    (n is greater than 5% of N)

  and…
  - Sampling is without replacement

Then

$$z = \frac{(\bar{x} - \mu)}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}}$$

Using the Sampling Distribution For Means

1. Compute the sample mean
2. Define the sampling distribution
3. Define the probability statement of interest
4. Convert sample mean to a z-value
5. Find the probability from the standard normal table (Appendix D)

If the Population is **not** Normal

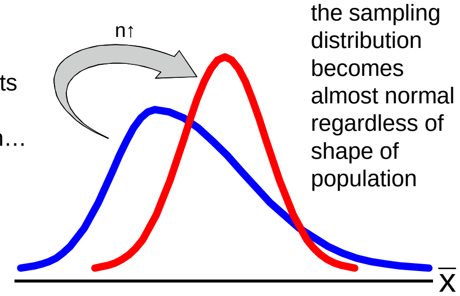
- We can apply the **Central Limit Theorem**:
  - Even if the population is not normal,
  - …sample means from the population will be approximately normal as long as the sample size is large enough
  - …and the sampling distribution will have

$$\mu_{\bar{x}} = \mu \quad \text{and} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Theorem 4

Central Limit Theorem

As the sample size gets large enough…



the sampling distribution becomes almost normal regardless of shape of population

How Large is Large Enough?

- For most distributions, $n > 30$ will give a sampling distribution that is nearly normal
- For fairly symmetric distributions, $n > 15$ is sufficient
- For normal population distributions, the sampling distribution of the mean is always normally distributed

Example

- Suppose a population has mean $\mu = 8$ and standard deviation $\sigma = 3$ and random sample of size $n = 36$ is selected.
- What is the probability that the sample mean is between 7.8 and 8.2?

Example

(continued)

Solution:

- Even if the population is not normally distributed, the central limit theorem can be used ($n > 30$)
- … so the sampling distribution of $\bar{X}$ is approximately normal
- … with mean $\mu_{\bar{x}} = \mu = 8$
- …and standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{36}} = 0.5$

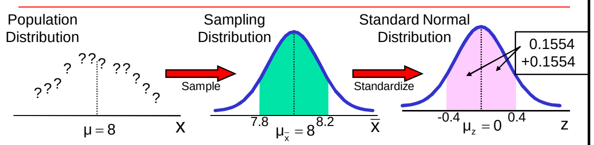
Example

(continued)

Solution (continued) -- find z-scores:

$$P(7.8 < \mu_{\bar{x}} < 8.2) = P\left(\frac{7.8 - 8}{3/\sqrt{36}} < \frac{\mu_{\bar{x}} - \mu}{\sigma/\sqrt{n}} < \frac{8.2 - 8}{3/\sqrt{36}} \right)$$

$$= P(-0.4 < z < 0.4) = \boxed{0.3108}$$



Population Proportions, π

π = the proportion of the population having some characteristic

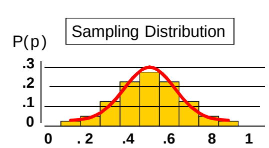
- Sample proportion (p) provides an estimate of π :

$$p = \frac{x}{n} = \frac{\text{number of successes in the sample}}{\text{sample size}}$$

- If two outcomes, p is a binomial distribution

Sampling Distribution of p

- Approximated by a normal distribution if:
  - $n\pi \geq 5$
    $n(1-\pi) \geq 5$



where

$\mu_p = \pi$ and $\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$ Theorem 5

(where π = population proportion)

z-Value for Proportions

Standardize p to a z value with the formula:

$$z = \frac{p - \pi}{\sigma_p} = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$

- If sampling is **without replacement** and n is greater than 5% of the population size, then σ_p must use the **finite population correction factor**:

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}\sqrt{\frac{N-n}{N-1}}$$

Using the Sample Distribution for Proportions

1. Determine the population proportion, π
2. Calculate the sample proportion, p
3. Derive the mean and standard deviation of the sampling distribution
4. Define the event of interest
5. If $n\pi$ and $n(1-\pi)$ are both > 5, then covert p to z-value
6. Use standard normal table (Appendix D) to determine the probability

Example

- If the true proportion of voters who support Proposition A is $\pi = 0.4$, what is the probability that a sample of size 200 yields a sample proportion between 0.40 and 0.45?
- i.e.: **if $\pi = 0.4$ and $n = 200$, what is $P(0.40 \le p \le 0.45)$?**

Example

(continued)

if $\pi = .4$ and $n = 200$, what is $P(0.40 \le p \le 0.45)$?

Find σ_p: $\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{.4(1-0.4)}{200}} = 0.03464$

Convert to standard normal (z-values): $P(0.40 \le p \le 0.45) = P\left(\frac{0.40-0.40}{.03464} \le z \le \frac{0.45-0.40}{.03464}\right)$

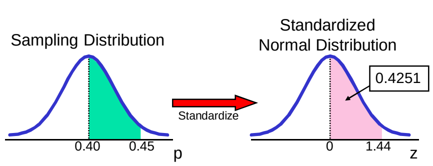
$= P(0 \le z \le 1.44)$

Example

(continued)

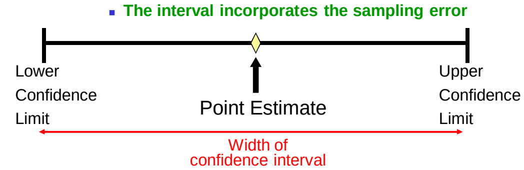
if $\pi = 0.4$ and $n = 200$, what is $P(0.40 \le p \le 0.45)$?

Use standard normal table: $P(0 \le z \le 1.44) = \boxed{0.4251}$



Point and Interval Estimates

- A point estimate is a single number, used to estimate an unknown population parameter
  - **The point estimate is not likely to exactly equal the population parameter**
- So a confidence interval provides additional information about variability within a range of z-values
  - **The interval incorporates the sampling error**



Point Estimates

We can estimate a Population Parameter …		with a Sample Statistic (a Point Estimate)
Mean	μ	$\bar{x}$
Proportion	π	p

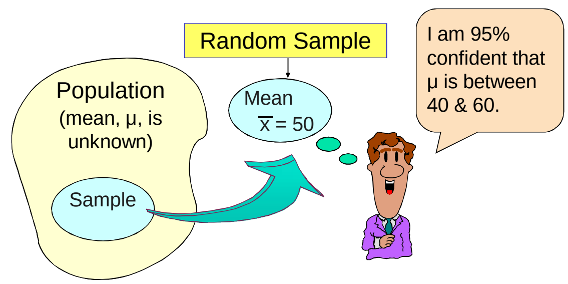
Confidence Intervals

- How much uncertainty is associated with a point estimate of a population parameter?
- An interval estimate provides more information about a population characteristic than does a point estimate
- Such interval estimates are called **confidence intervals**

Confidence Interval Estimate

- An interval gives a range of values:
  - Takes into consideration variation in sample statistics from sample to sample
  - Based on observation from 1 sample
  - Gives information about closeness to unknown population parameters
  - Stated in terms of level of confidence
    - Never 100% sure

Estimation Process



General Formula

- The general formula for all confidence intervals is:

Point Estimate ± (Critical Value)(Standard Error)

z-value based on the level of confidence desired

Confidence Level

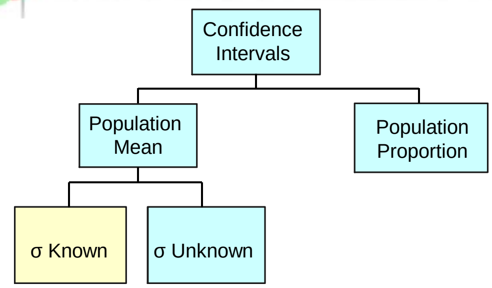
- Confidence Level
  - Confidence in which the interval will contain the unknown population parameter
- A percentage (less than 100%)
  - Most common: 90%, 95%, 99%

Confidence Level, (1-α)

(continued)

- Suppose confidence level = 95%
- A relative frequency interpretation:
  - In the long run, 95% of all the confidence intervals that can be constructed will contain the unknown true parameter
- A specific interval either will contain or will not contain the true parameter
  - No probability involved in a specific interval

Confidence Intervals



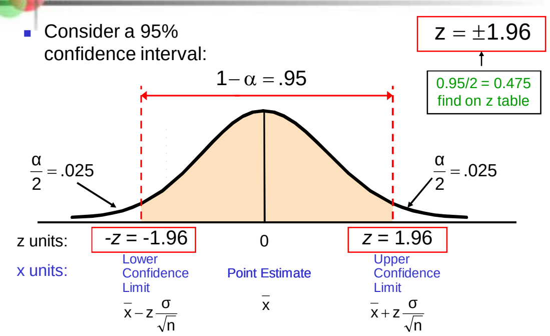
Confidence Interval for μ (σ Known)

- Assumptions
  - Population standard deviation σ is known
  - Population is normally distributed
  - If population is not normal, use large sample $n \geq 30$
- Confidence interval estimate

$$\bar{x} \pm z \frac{\sigma}{\sqrt{n}}$$

Finding the Critical Value

- Consider a 95% confidence interval:



Common Levels of Confidence

- Commonly used confidence levels are 90%, 95%, and 99%

Confidence Level	*Critical value, z*
80%	1.28
90%	1.645
95%	1.96
98%	2.33
99%	2.58
99.8%	3.08
99.9%	3.27

Computing a Confidence Interval Estimate for the Mean (σ known)

1. Select a random sample of size n
2. Specify the confidence level
3. Compute the sample mean
4. Determine the standard error
5. Determine the critical value (z) from the normal table
6. Compute the confidence interval estimate

Margin of Error

- Margin of Error (e): the amount added and subtracted to the point estimate to form the confidence interval
  - **Defines the relationship between the population parameter and the sample statistic,**

Example: Margin of error for estimating μ, σ known:

$$\bar{x} \pm z\frac{\sigma}{\sqrt{n}} \qquad e = z\frac{\sigma}{\sqrt{n}}$$

Factors Affecting Margin of Error

$$\boxed{e = z\frac{\sigma}{\sqrt{n}}}$$

- Data variation, σ : e ⬇ as σ ⬇
- Sample size, n : e ⬇ as n ⇧
- Level of confidence, 1 - α : e ⬇ if 1 - α ⬇

Example

- A sample of 11 circuits from a large normal population has a mean resistance of 2.20 ohms. We know from past testing that the population standard deviation is 0.35 ohms.
- Determine a 95% confidence interval for the true mean resistance of the population.

Example_Solution

(continued)

Solution:

$$\bar{x} \pm z\frac{\sigma}{\sqrt{n}}$$

$$= 2.20 \pm 1.96\,(0.35/\sqrt{11})$$

$$= 2.20 \pm .2068$$

$$\boxed{1.9932 \longrightarrow 2.4068}$$

95% confident that the mean resistance is between approximately 2 and 2.41 ohms

Interpretation

- We are 95% confident that the true mean resistance is between 1.9932 and 2.4068 ohms
- Although the true mean may or may not be in this interval, 95% of intervals formed in this manner will contain the true mean
- An **incorrect** interpretation is that there is 95% probability that this interval contains the true population mean.

  (This interval either does or does not contain the true mean, there is no probability for a single interval)

Confidence Interval for μ (σ Unknown)

- If the population standard deviation σ is unknown, we can substitute the sample standard deviation, s
- This introduces extra uncertainty, since s is variable from sample to sample
- So we use the t distribution instead of the normal distribution

Confidence Interval for μ (σ Unknown)

(continued)

- Assumptions
  - Population standard deviation is unknown
  - Population is normally distributed
  - If population is not normal, use large sample $n \geq 30$
- Use Student's *t* Distribution
- Confidence Interval Estimate

$$\bar{X} \pm t \frac{S}{\sqrt{n}}$$

Student's t Distribution

- The t is a family of distributions
- The t value depends on degrees of freedom (d.f.)
  - Number of observations that are free to vary after sample mean has been calculated

$$\text{d.f.} = n - 1$$

  - Only n-1 independent pieces of data information left in the sample because the sample mean has already been obtained

Degrees of Freedom (df)

Idea: Number of observations that are free to vary after sample mean has been calculated

Example: Suppose the mean of 3 numbers is 8.0

Let $x_1 = 7$
Let $x_2 = 8$
What is x_3?

If the mean of these three values is 8.0,
then x_3 must be 9
(i.e., x_3 is not free to vary)

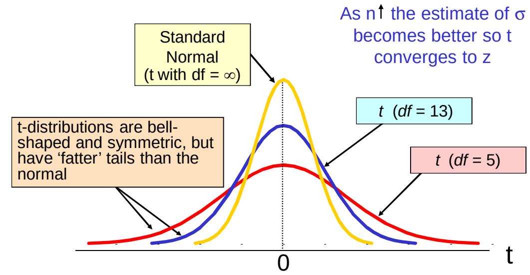
Here, n = 3, so degrees of freedom $= n - 1 = 3 - 1 = 2$

(2 values can be any numbers, but the third is not free to vary for a given mean)

Student's t Distribution

Note: t compared to z as n increases

As n↑ the estimate of σ becomes better so t converges to z

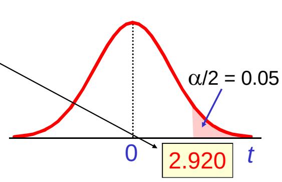


Student's t Table

df	Confidence Level 0.50	0.80	0.90
1	1.000	3.078	6.314
2	0.817	1.886	2.920
3	0.765	1.638	2.353

The body of the table contains t values, not probabilities

Let: n = 3
df = n - 1 = 2
confidence level: 90%



t Distribution Values

With comparison to the z value

Confidence Level	t (10 d.f.)	t (20 d.f.)	t (30 d.f.)	z
0.80	1.372	1.325	1.310	1.28
0.90	1.812	1.725	1.697	1.64
0.95	2.228	2.086	2.042	1.96
0.99	3.169	2.845	2.750	2.58

Note: t compared to z as n increases

Example

A random sample of n = 25 has $\bar{x}$ = 50 and s = 8. Form a 95% confidence interval for μ

- d.f. = n – 1 = 24, so $t_{\alpha/2,\,n-1} = t_{0.025,24} = 2.0639$

The confidence interval is

$$\bar{x} \pm t\frac{s}{\sqrt{n}} = 50 \pm (2.0639)\frac{8}{\sqrt{25}}$$

46.698 ⟶ 53.302

Approximation for Large Samples

- Since t approaches z as the sample size increases, an approximation is sometimes used when n is very large
- The text t-table provides t values up to 500 degrees of freedom
- Computer software will provide the correct t-value for any degrees of freedom

Correct formula, σ unknown

$$\bar{x} \pm t\frac{s}{\sqrt{n}}$$

Approximation for very large n

$$\bar{x} \pm z\frac{s}{\sqrt{n}}$$

Confidence Intervals for the Population Proportion, π

- An interval estimate for the population proportion (π) can be calculated by adding an allowance for uncertainty to the sample proportion (p)

Confidence Intervals for the Population Proportion, π

(continued)

- Recall that the distribution of the sample proportion is approximately normal if the sample size is large, with standard deviation

$$\sigma_\pi = \sqrt{\frac{\pi(1-\pi)}{n}}$$

See Chpt. 7!!

- We will estimate this with sample data:

$$s_p = \sqrt{\frac{p(1-p)}{n}}$$

Confidence Interval Endpoints

- Upper and lower confidence limits for the population proportion are calculated with the formula

$$p \pm z\sqrt{\frac{p(1-p)}{n}}$$

- where
  - z is the standard normal value for the level of confidence desired
  - p is the sample proportion
  - n is the sample size

Example

- A random sample of 100 people shows that 25 are left-handed.
- Form a 95% confidence interval for the true proportion of left-handers

Example

(continued)

- A random sample of 100 people shows that 25 are left-handed. Form a 95% confidence interval for the true proportion of left-handers.

1. $p = 25/100 = 0.25$
2. $S_p = \sqrt{p(1-p)/n} = \sqrt{0.25(0.75)/100} = 0.0433$
3. $0.25 \pm 1.96\,(0.0433)$

$$\boxed{0.1651 \longrightarrow 0.3349}$$

Interpretation

- We are 95% confident that the true percentage of left-handers in the population is between

  16.51% and 33.49%

- Although this range may or may not contain the true proportion, 95% of intervals formed from samples of size 100 in this manner will contain the true proportion.

Changing the sample size

- Increases in the sample size reduce the width of the confidence interval.

Example:

- If the sample size in the above example is doubled to 200, and if 50 are left-handed in the sample, then the interval is still centered at 0.25, but the width shrinks to

  $0.19 \longrightarrow 0.31$

Finding the Required Sample Size for Proportion Problems

Define the margin of error:

$$e = z\sqrt{\frac{\pi(1-\pi)}{n}}$$

Solve for n:

$$n = \frac{z^2\,\pi(1-\pi)}{e^2}$$

Will be in % units

π can be estimated with a pilot sample, if necessary (or conservatively use $\pi = 0.50$ – worst possible variation thus the largest sample size)

What sample size...?

- How large a sample would be necessary to estimate the true proportion defective in a large population within 3%, with 95% confidence?

(Assume a pilot sample yields $p = 0.12$)

What sample size...?

(continued)

Solution:

For 95% confidence, use $Z = 1.96$

$e = 0.03$

$p = 0.12$, so use this to estimate π

$$n = \frac{z^2\,\pi(1-\pi)}{e^2} = \frac{(1.96)^2(0.12)(1-0.12)}{(0.03)^2} = 450.74$$

So use $n = 451$

Chapter 2 Summary

- Discussed sampling error
- Introduced sampling distributions
- Described the sampling distribution of the mean
  - For normal populations
  - Using the Central Limit Theorem (normality unknown)
- Described the sampling distribution of a proportion
- Calculated probabilities using sampling distributions
- Discussed sampling from finite populations

Chapter 2 Summary

(continued)

- Discussed point estimates
- Introduced interval estimates
- Discussed confidence interval estimation for the mean [σ known]
- Discussed confidence interval estimation for the mean [σ unknown]
- Discussed confidence interval estimation for the proportion
- Addressed determining sample size for proportion problems