VI.
Why do samples allow inference? How sure do we have to be? How many do I need to be that sure?

Sampling Distributions, Confidence Intervals, & Sample Size

PBAF 527 Winter 2005

1

Today

Theory

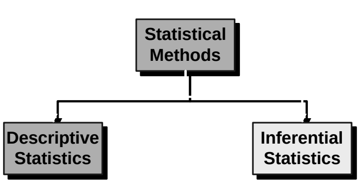
Sampling Distributions
- Describe the Properties of Estimators
- Explain Sampling Distribution
- Describe the Relationship between Populations & Sampling Distributions
- State the Central Limit Theorem
- Solve Probability Problems Involving Sampling Distributions

Practice

Tools for Samples
- Making Inference
- Confidence Intervals
  - State What is Estimated
  - Distinguish Point & Interval Estimates
  - Explain Interval Estimates
  - Compute Confidence Interval Estimates for Population Mean & Proportion
- Compute Sample Size

2

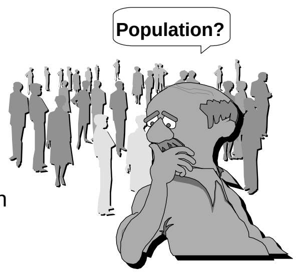
Statistical Methods



3

Inferential Statistics

1. Involves:
   - Estimation
   - Hypothesis Testing
2. Purpose
   - Make Decisions about Population Characteristics

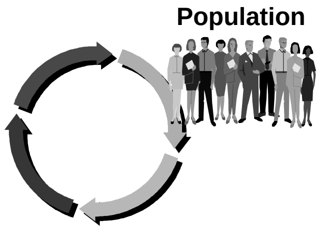


4

Inference Process

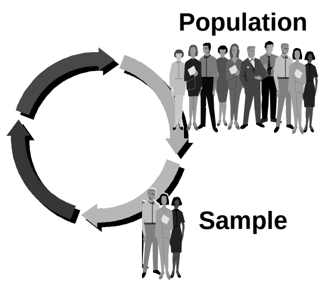
5

Inference Process



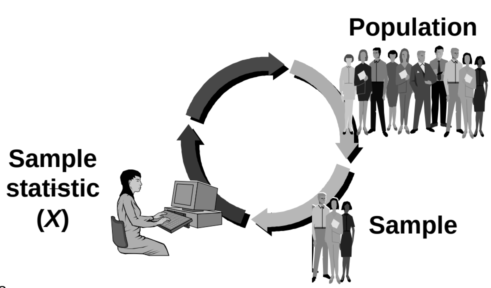
6

Inference Process



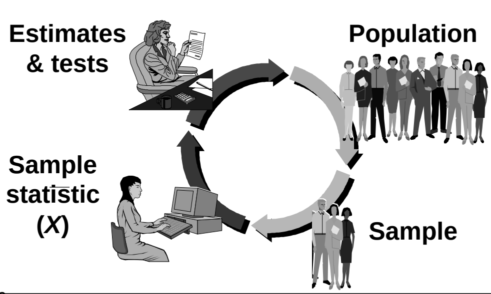
7

Inference Process



8

Inference Process



9

Non-Probability Sampling

- Cannot tell the probability of choosing each member of the population.
  - Quota Sampling
  - Volunteer Sampling
  - Snowball Sampling

10

Probability Sampling

Each population member sampled with known probability.
- ESSENTIAL for a representative sample
- Simple Random Sample (SRS)
- Stratified Random Sample
- Cluster Sample
- Systematic Sample

11

Estimators

- Random Variables Used to Estimate a Population Parameter
  - Sample Proportion $Var(\hat{P}) = \frac{P(1-P)}{n}$
    - The mean of the sample p= $\hat{P}$
  - Sample Mean

$$E(\overline{X}) = m \qquad Var(\overline{X}) = \frac{s^2}{n}$$

- Theoretical Basis Is Sampling Distribution

12

Sampling Distribution

1. **Theoretical** Probability Distribution
2. Random Variable is **Sample Statistic**
   - Sample Mean, Sample Proportion etc.
3. Results from Drawing **All** Possible Samples of a **Fixed** Size
4. List of All Possible [$\bar{X}$, P($\bar{X}$)] Pairs
   - Sampling Distribution of Mean

13

Standard Error of Mean

1. Standard Deviation of All Possible Sample Means, $\bar{x}$
   - How far a typical x-bar is from mu.
2. Less Than Pop. Standard Deviation
3. Formula (Sampling With Replacement)

For a sample proportion

$$SE \equiv SD(\hat{P}) = \sqrt{\frac{p(1-p)}{n}}$$

For a sample mean

$$SE \equiv SD(\bar{X}) = \frac{s}{\sqrt{n}}$$

14

Creating a Sampling Distribution

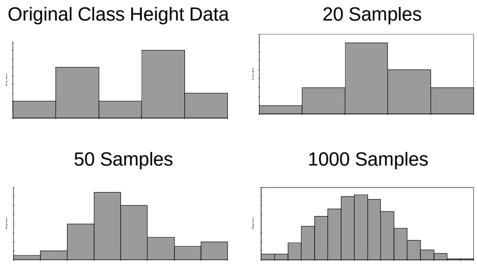
Remember, from a class of 27, there are over 33,000 possible samples of 5.

So, we can take multiple samples from the class

For each sample, we can calculate a mean and a standard deviation.

15

Sampling Distributions



16

Sampling Distributions

	Mean	SD($\bar{x}$)
10 Samples	66.6	0.55
20 Samples	67.3	0.41
50 Samples	67.2	0.21
250 Samples	67.0	0.10
1000 Samples	67.0	0.05

17

Properties of Sampling Distribution of Mean

1. Unbiasedness
   - Mean of Sampling Distribution Equals Population Mean
2. Efficiency
   - Sample Mean Comes Closer to Population Mean Than Any Other Unbiased Estimator
3. Consistency
   - As Sample Size Increases, Variation of Sample Mean from Population Mean Decreases

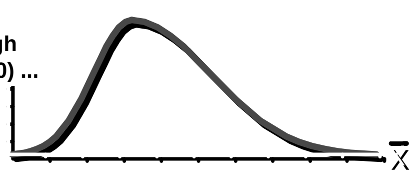
18

Central Limit Theorem

19

Central Limit Theorem

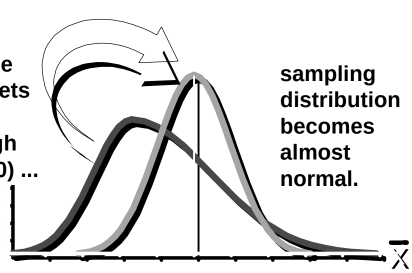
As sample size gets large enough (n ^{3} 30) ...



20

Central Limit Theorem

As sample size gets large enough (n ^{3} 30) ...

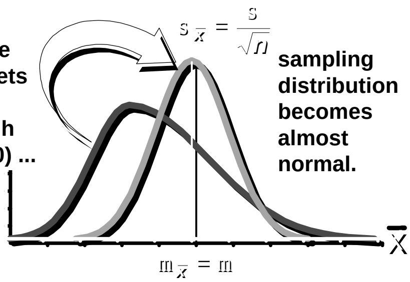


sampling distribution becomes almost normal.

21

Central Limit Theorem

As sample size gets large enough ($n \geq 30$) ...



sampling distribution becomes almost normal.

22

Problem 1

I know from the U.S. Current Population Survey that 1999 household income had a mean of $44,000 and a standard deviation of $62,000. If I were to pick a random sample of 100 U.S. households, how likely am I to get a sample mean of household income more than $1000 above the population mean?

1. Write down the target and draw a picture.
2. Find the standard error.
3. Find the standard score for our sample mean target value.

   SE … 6200

4. Look up the score in the z-table.

   there is a 44% probability of seeing a value of over $45K

23

Problem 1 (Part 2)

Suppose instead that we had a random sample of 10,000 U.S. household. What is the chance of getting a sample mean for income more than 1000 above the population mean of $44,000?

1. Find the target and draw picture.

   We want $P(\bar{X} > \$45K)$

2. Find the standard error.

   $\sqrt{n}$ $\sqrt{10,000}$

3. Find the stand…

   $P(\bar{X} > \$45K) = P\left(Z > \frac{\ldots}{SE}\right) = P\left(Z > \frac{\ldots}{620}\right) = P(Z > 1.61)$

4. Look up score in Z tab…

   percent of the time.

24

Proportions

Proportions are means of binary variables, so the sampling distribution is approximately normal.

Remember:

- Population Mean≡ P
- Population Variance≡ P(1-P)
- s =√P(1-P)

25

Problem 2

In 1999, about 12 % of persons in the U.S. lived in households with income under the poverty level. What's the probability that the poverty rate in a sample of 1000 people will be within 1 percentage point of the "true" proportion?

To find the probability for a sample proportion:

1. Find target and draw picture.
2. Find SE (Need pop SD first).
3. Find Z scores.

26

Today

Theory

Sampling Distributions

- Describe the Properties of Estimators
- Explain Sampling Distribution
- Describe the Relationship between Populations & Sampling Distributions
- State the Central Limit Theorem
- Solve Probability Problems Involving Sampling Distributions

Practice

Tools for Samples

- Making Inference
- Confidence Intervals
  - State What is Estimated
  - Distinguish Point & Interval Estimates
  - Explain Interval Estimates
  - Compute Confidence Interval Estimates for Population Mean & Proportion
- Compute Sample Size

27

Tools for Samples

- Point Estimate
  - Best guess of a population parameter based upon a sample
- Confidence Interval
  - Range estimate around point estimate
- Hypothesis Test
  - Decision rule for rejecting hypothesized population values (null hypotheses)
- p-value
  - Continuous measure of support for null hypothesis (a probability, α)

28

Making Inference

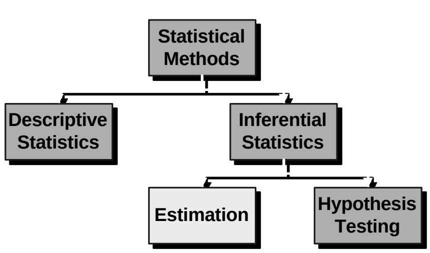
How can we make assertions about the unknown?

Scientific Method:

Theory ⇒	Hypothesis ⇒	Empirical Evidence
(model of how the world works)	(specific statement about the model)	(test of the hypothesis)

29

Statistical Methods



30

Example: the Polls

In the 1992 Presidential Election, George Bush was expected to receive 38% of the vote, and Bill Clinton 40%. Yet, Bush was hopeful of winning. Why?

THE MARGIN OF ERROR

38% ± 3% = [35%, 41%]

31

Example: the Polls

In the 1992 Presidential Election, George Bush was expected to receive 38% of the vote, and Bill Clinton 40%. Yet, Bush was hopeful of winning. Why?

THE MARGIN OF ERROR

38% ± 3% = [35%, 41%]

32

Example: the Polls (2)

In the 2000 Presidential Election, we had a very different situation. By Nov 1, 2000, the polls indicated that George W. Bush could expect to receive 47% and Al Gore 43%. The pundits commented on Bush's slim lead. Why?

THE MARGIN OF ERROR

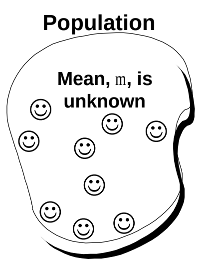
47% ± 2% = [45%, 49%]

33

Estimation Process

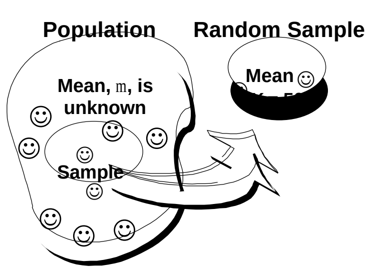
34

Estimation Process



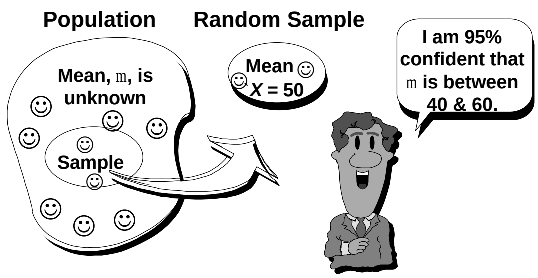
35

Estimation Process



36

Estimation Process



37

Unknown Population Parameters Are Estimated

Estimate Population Parameter...		with Sample Statistic
Mean	m	$\bar{x}$
Proportion	p	$\hat{p}$
Differences	$m_1 - m_2$	$\bar{x}_1 - \bar{x}_2$

38

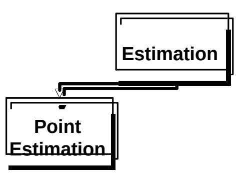
Estimation Methods

39

Estimation Methods

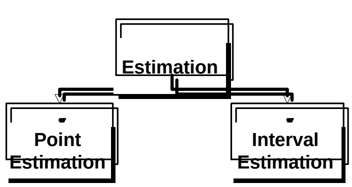
40

Estimation Methods



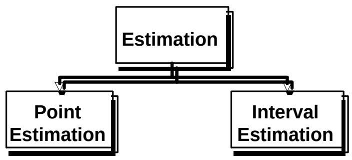
41

Estimation Methods



42

Estimation Methods



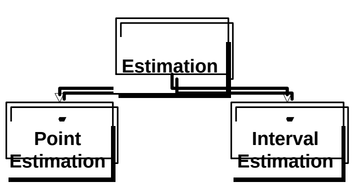
43

Point Estimation

1. Provides Single Value
   - Based on Observations from 1 Sample
2. Gives No Information about How Close Value Is to the Unknown Population Parameter
3. Example: Sample Mean $\bar{X}$ = 3 Is Point Estimate of Unknown Population Mean

44

Estimation Methods



45

Interval Estimation

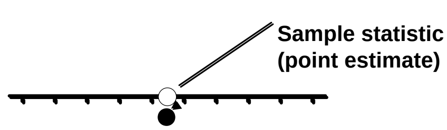
1. Provides Range of Values
   - Based on Observations from 1 Sample
2. Gives Information about Closeness to Unknown Population Parameter
   - Stated in terms of Probability
     - Knowing Exact Closeness Requires Knowing Unknown Population Parameter
3. Example: Unknown Population Mean Lies Between 50 & 70 with 95% Confidence

46

Key Elements of Interval Estimation

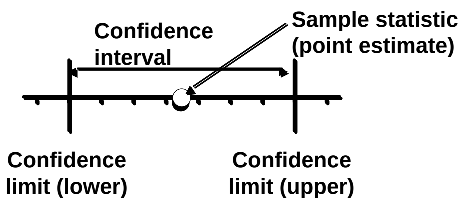
47

Key Elements of Interval Estimation



48

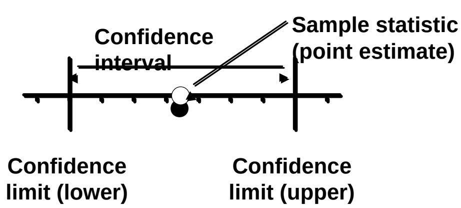
Key Elements of Interval Estimation



49

Key Elements of Interval Estimation

A probability that the population parameter falls somewhere within the interval.



50

Confidence Limits for Population Mean

(1) $m = \bar{X} \pm Error$

(2) $Error = \bar{X} - m$ or $\bar{X} + m$

(3) $Z = \frac{\bar{X} - m}{s_{\bar{x}}} = \frac{Error}{s_{\bar{x}}}$

(4) $Error = Zs_{\bar{x}}$

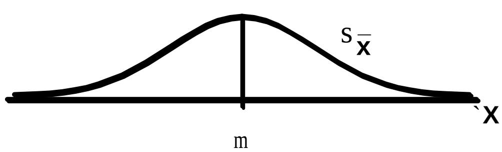
(5) $m = \bar{X} \pm Zs_{\bar{x}}$

51

Many Samples Have Same Interval

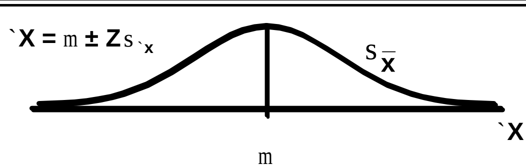
52

Many Samples Have Same Interval



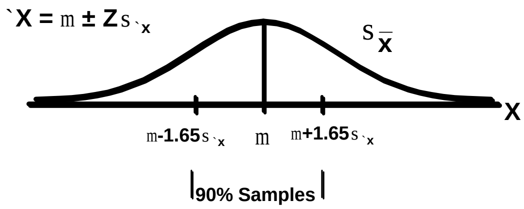
53

Many Samples Have Same Interval



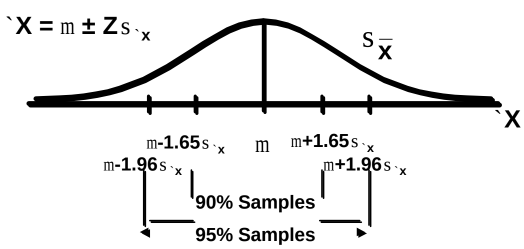
54

Many Samples Have Same Interval



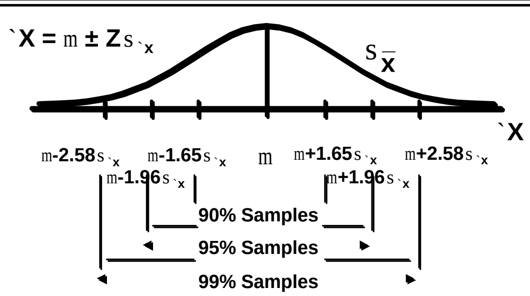
55

Many Samples Have Same Interval



56

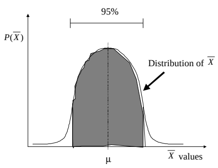
Many Samples Have Same Interval



57

Sampling Distribution Gives the Probability of Sample Mean Close to μ.

What's the probability of getting a sample mean within 1.96 standard errors of the population mean?



$P(\mu-1.96SE) < \overline{X} < P(\mu + 1.96SE)$

$$= P\left[\frac{(m-1.96SE)-m}{SE} < Z < \frac{(m+1.96SE)-m}{SE}\right]$$

=1-.025-.025=.4750+.4750=.95

58

Today

Theory | **Practice**

Sampling Distributions
- Describe the Properties of Estimators
- Explain Sampling Distribution
- Describe the Relationship between Populations & Sampling Distributions
- State the Central Limit Theorem
- Solve Probability Problems Involving Sampling Distributions

Tools for Samples
- Making Inference
- Confidence Intervals
  - State What is Estimated
  - Distinguish Point & Interval Estimates
  - Explain Interval Estimates
  - Compute Confidence Interval Estimates for Population Mean & Proportion
- Compute Sample Size

59

Confidence Intervals Explained: 95% CI

The interval from xbar-1.96*SE* to xbar+1.96*SE* should contain the population mean, μ, 95% of the time. Xbar is the mean of one sample.

In other words, 95% of the time when we take a sample of size n and take its mean, the population mean, μ, will fall within ±1.96*SE*'s around xbar. That is, in 95 of 100 samples.

We call the interval xbar ± 1.96*SE* a 95% confidence interval for the unknown population mean, μ.

μ either lies inside the confidence interval or it lies outside the confidence interval.

What we know is that 95% of all possible intervals constructed around all of the xbar (that we calculate from random samples of size n) will contain μ.

60

Confidence Interval Example

U.S. household income has a standard deviation of $62,000. Suppose you took a random sample of 100 households and got a sample mean of $40,000. What range estimate could you construct that will have a 95% chance of including the unknown population mean?

Want 95% confidence interval (with known s):

P(xbar- 1.96SE < μ < xbar+ 1.96SE)= .95

1.Get standard error $SE = \frac{s}{\sqrt{n}} = \frac{\$62,000}{\sqrt{100}} = 6200$

2. Plug in SE and sample mean:

P(xbar- 1.96SE < μ < xbar + 1.96SE)= .95

=P[40,000 -(1.96)(6200) < μ < 40,000 + (1.96)(6200)]

=P[27,848 < μ < 52,152]=.95 $\overline{X} \pm 1.96\frac{s}{\sqrt{n}}$=[27848,52152]

▪So, 95 percent of the time, the household mean for U.S. households is within the range $27,848 and $52,152.

61

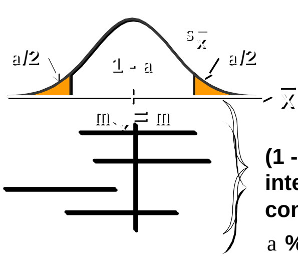
Confidence Level

1. Probability that the Unknown Population Parameter Falls Within Interval
2. Denoted (1 - α) %
   - a Is Probability That Parameter Is **Not** Within Interval (the error)
3. Typical Values Are 99%, 95%, 90%

62

Intervals & Confidence Level

Sampling Distribution of Mean



Intervals extend from `X - Zs`x to `X + Zs`x

(1 - a) % of intervals contain m.

a % do not.

Large number of intervals

63

Steps to Confidence Intervals

1. Decide what percentage of the probability you want inside the interval.
2. Take the percentage outside the interval and label it α
3. Calculate $\alpha/2$, the percentage of the probability in one of the tails outside the interval.
4. Look up the value of $z_{\alpha/2}$, the z value that cuts off a right tail area of $\alpha/2$, under the standard normal curve.

64

Common Values of $z_{?/2}$

Confidence Level	α=error probability	$\alpha/2$=one tail area	$z_{\alpha/2}$=critical value
90%	0.10	0.05	1.645
95%	0.05	0.025	1.96
98%	0.02	0.01	2.326
99%	0.01	0.005	2.576

65

An aside about the Empirical Rule and Common Confidence Intervals

SD on either side of the mean	Empirical Rule for Mound-Shaped Distributions	Probabilities Associated with ANY Normal Variable
1	Approximately 68%	.6826
2	Approximately 95%	.9544
3	Almost all	.9974

Estimated Probabilities

SD on either side of the mean	Probabilities Associated with ANY Normal Variable
1.645	.90
1.96	.95
2.575	.99

Exact Probabilities

66

Factors Affecting Interval Width

1. Data Dispersion
   - Measured by σ
2. Sample Size
   - $\sigma_{\bar{X}} = \sigma / \sqrt{n}$
3. Level of Confidence $(1 - \alpha)$
   - Affects Z

Intervals Extend from $\bar{X} - Z\sigma_{\bar{X}}$ to $\bar{X} + Z\sigma_{\bar{X}}$

67

Confidence Interval Estimates

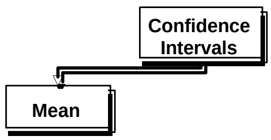
68

Confidence Interval Estimates

Confidence Intervals

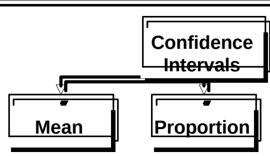
69

Confidence Interval Estimates



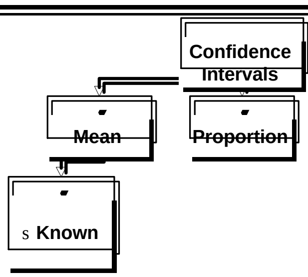
70

Confidence Interval Estimates



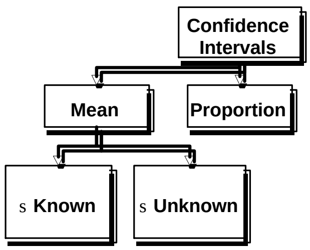
71

Confidence Interval Estimates



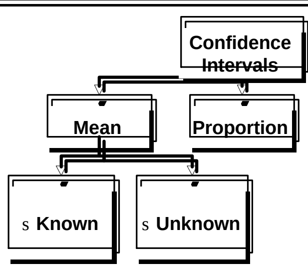
72

Confidence Interval Estimates



73

Confidence Interval Estimates



74

Confidence Interval Mean (s Known)

1. Assumptions
   - Population Standard Deviation Is Known
   - Population Is Normally Distributed
   - If Not Normal, Can Be Approximated by Normal Distribution ($n \geq 30$)
2. Confidence Interval Estimate

$$\bar{X} - Z_{a/2} \cdot \frac{s}{\sqrt{n}} \leq m \leq \bar{X} + Z_{a/2} \cdot \frac{s}{\sqrt{n}}$$

75

Estimation Example Mean (σ Known)

The mean of a random sample of *n* = 25 is $\bar{X}$ = 50. Set up a 95% confidence interval estimate for μ if σ = 10.

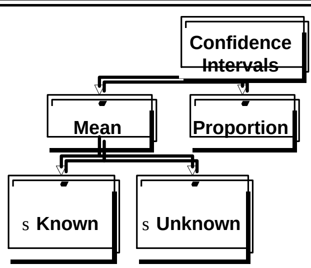
$$\bar{X} - Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

$$50 - 1.96 \cdot \frac{10}{\sqrt{25}} \le \mu \le 50 + 1.96 \cdot \frac{10}{\sqrt{25}}$$

$$46.08 \le \mu \le 53.92$$

76

Confidence Interval Estimates



77

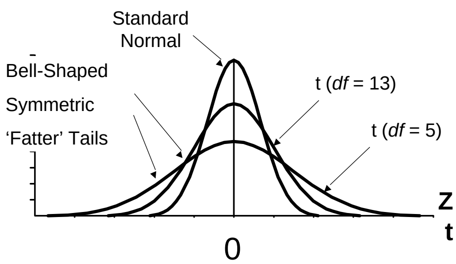
Confidence Interval Mean (σ Unknown)

1. Assumptions
   - Population Standard Deviation Is Unknown
   - Population Must Be **Normally Distributed**
2. Estimate SE $\frac{S}{\sqrt{n}}$
3. Use Student's t Distribution
4. Confidence Interval Estimate

$$\bar{X} - t_{\alpha/2, n-1} \cdot \frac{S}{\sqrt{n}} \le \mu \le \bar{X} + t_{\alpha/2, n-1} \cdot \frac{S}{\sqrt{n}}$$

78

Student's t Distribution



79

Student's t Table

80

Student's t Table

v	$t_{.10}$	$t_{.05}$	$t_{.025}$
1	3.078	6.314	12.706
2	1.886	2.920	4.303
3	1.638	2.353	3.182

81

Student's t Table

v	$t_{.10}$	$t_{.05}$	$t_{.025}$
1	3.078	6.314	12.706
2	1.886	2.920	4.303
3	1.638	2.353	3.182

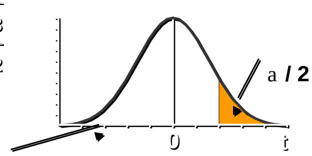
t values

82

Student's t Table

a / 2

v	$t_{.10}$	$t_{.05}$	$t_{.025}$
1	3.078	6.314	12.706
2	1.886	2.920	4.303
3	1.638	2.353	3.182



t values

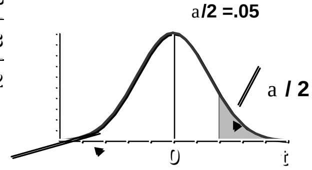
83

Student's t Table

a / 2

Assume:
n = 3
df = *n* - 1 = 2
a = .10
a/2 =.05

v	$t_{.10}$	$t_{.05}$	$t_{.025}$
1	3.078	6.314	12.706
2	1.886	2.920	4.303
3	1.638	2.353	3.182



t values

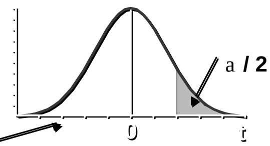
84

Student's t Table

v	$t_{.10}$	$t_{.05}$	$t_{.025}$
1	3.078	6.314	12.706
2	1.886	2.920	4.303
3	1.638	2.353	3.182

a / 2

Assume:
$n = 3$
$df = n - 1 = 2$
a = .10
a/2 = .05



t values

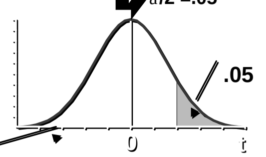
85

Student's t Table

v	$t_{.10}$	$t_{.05}$	$t_{.025}$
1	3.078	6.314	12.706
2	1.886	2.920	4.303
3	1.638	2.353	3.182

a / 2

Assume:
$n = 3$
$df = n - 1 = 2$
a = .10
a/2 = .05



t values

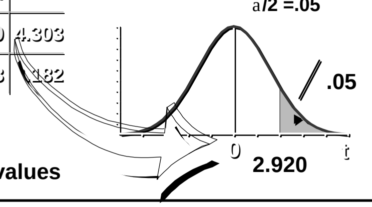
86

Student's t Table

v	$t_{.10}$	$t_{.05}$	$t_{.025}$
1	3.078	6.314	12.706
2	1.886	2.920	4.303
3	1.638	2.353	3.182

a / 2

Assume:
$n = 3$
$df = n - 1 = 2$
a = .10
a/2 = .05



t values

87

Degrees of Freedom (*df*)

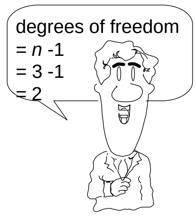
1. Number of Observations that Are Free to Vary After Sample Statistic Has Been Calculated
2. Example
   - Sum of 3 Numbers Is 6

     X_1 = 1 (or Any Number)

     X_2 = 2 (or Any Number)

     X_3 = **3** (Cannot Vary)

     Sum = 6



88

Estimation Example Mean (s Unknown)

Suppose my sample of 100 households has a mean income of $40,000 and a standard deviation (s) of $59,000. What is the 95% confidence interval?

Want 95% confidence interval (with σ unknown):

$$P\left(\overline{X} - t_{.025}\hat{S}E < m < \overline{X} + t_{.025}\hat{S}E\right) = .95$$

(α=.05, so α/2=.025, or 2.5% in each tail)

89

Estimation Example Mean (s Unknown) (2)

1. Find SE estimate. $\hat{S}E = \frac{s}{\sqrt{n}} = \frac{59{,}000}{\sqrt{100}} = 5900$

2. Find t-score for 95% CI with 2.5% in each tail.

degrees of freedom = n-1 = 100-1= 99

⇒$t_{.025}$= 2.00 2t rule of thumb

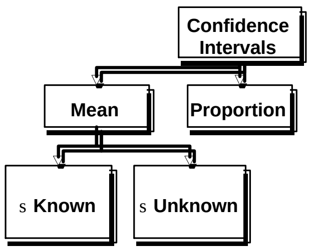
3. Plug in t-score, sample mean, and estimated sample SE:

$$P\left(\overline{X} - 2\hat{S}E < m < \overline{X} + 2\hat{S}E\right) = .95$$
$$= P\left[(40{,}000 - 2.00(5900)) < m < (40{,}000 + 2.00(5900))\right] = .95$$
$$= P\left[28{,}200 < m < 51{,}800\right] = .95$$

So, 95 percent of the time, the household mean for U.S. households is within the range $28,200 and $51,800.

90

Confidence Interval Estimates



91

Confidence Interval Proportion

1. Assumptions
   - Two Categorical Outcomes
   - Large sample
     - np and n(1-p) are greater than about 5 or 10.
   - Normal Approximation Can Be Used
2. Formula: a large sample (1-?)100% confidence interval for the population proportion, p:

$$\hat{p} \pm z_{a/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} \qquad \hat{q} = 1 - \hat{p}$$

92

Estimation Example Proportion

Suppose that in my random sample of 100 households I find that 58% would be willing to pay more than $100 per year in taxes to offset the national deficit. What is the 99% confidence interval for that proportion?

$$P\left(\hat{p} - z_{.005}\hat{S}E < p < \hat{p} + z_{.005}\hat{S}E\right) = .99$$

1. Find SE Estimate. For proportion $s = \sqrt{\hat{p}\hat{q}} = \sqrt{.58(1-.58)} = .49$

2. Find z-score of 99% CI. ?=.01, so ?/2=.005, or .5% in each tail
   $z_{.005}$=2.58

3. Plug in $\hat{S}E$, $\hat{p}$, z.

$$P\left(\hat{p} - 2.58\hat{S}E < p < \hat{p} + 2.58\hat{S}E\right) = .99$$
$$= P(.58 - 2.58(.049) < p < .58 + 2.58(.049)) = .99$$
$$= P(.45 < p < .71) = .99$$

93 99% of the time, the proportion will fall between .45 and .71.

A final word on confidence intervals

So, **confidence intervals** give range estimates of all values close enough for us to consider as possible population means or proportions.

94

How many do I need to be that sure?

Determining Sample Size

95

To determine sample size…

...You need to know something about your population!

You also need to think about the quality of the estimate:

- Close do you want to be to the unknown parameter? (B)
- Confidence level? (z)
- Estimate of Variance? (SD)

96

Sample Size Formulas

Estimating a population mean $n = \frac{z_{a/2}^2 s^2}{B^2}$

Estimating a population proportion $n = \frac{z_{a/2}^2 pq}{B^2}$

Need some prior experience to estimate sample size—or conservatism. p=.05 will yield the largest sample size for a proportion

These are the minimum sample sizes for SRS (with replacement) where you only want to estimate a mean or proportion! They do not take **population size** into account.

97

Finding Sample Sizes for Estimating m

(1) $Z = \frac{\overline{X} - m}{s_{\overline{x}}} = \frac{Error}{s_{\overline{x}}}$

(2) $Error = Zs_{\overline{x}} = Z\frac{s}{\sqrt{n}}$

(3) $n = \frac{Z^2 s^2}{Error^2}$

Error Is Also Called Bound, *B*

98

Sample Size Example

What if I was going to draw my sample of households from the population? I know that the standard deviation of household income in the U.S. is $62,000. What sample size do I need to estimate mean income with 95% confidence to within plus or minus $500?

1. Find B: B is the half-width bound, so B=500
2. What z? At 95% confidence, z=1.96
3. Variance estimate? σ=62000, so σ^2=3,844,000,000
4. Plug in B, z, and σ^2 (or its estimate) into the equation

$$n = \frac{z_{a/2}^2 s^2}{B^2} = \frac{(1.96)^2(62000)^2}{500^2} = 59{,}068$$

99

Sample Size Example (2)

What if I only wanted to come within $1000 of the mean?

$$n = \frac{z_{a/2}^2 s^2}{B^2} = \frac{(1.96)^2(62000)^2}{1000^2} = 14{,}767$$

How about $5000 of the mean?

$$n = \frac{z_{a/2}^2 s^2}{B^2} = \frac{(1.96)^2(62000)^2}{5000^2} = 591$$

At a larger distance, a smaller sample size is needed.

100

Correcting for Without Replacement Sampling

- SRS_{WOR} in reality!
- Also correct for population size
  - Important for small populations
- Formula

$$n_{WOR} = \frac{n}{1 + \frac{n}{N}}$$

Where
  - n_{WOR} is the sample size needed for without replacement sampling
  - n is the sample size needed for with replacement sampling (equation in book)
  - N is the population size.

101

Today

Theory	Practice

Sampling Distributions
- Describe the Properties of Estimators
- Explain Sampling Distribution
- Describe the Relationship between Populations & Sampling Distributions
- State the Central Limit Theorem
- Solve Probability Problems Involving Sampling Distributions

Tools for Samples
- Making Inference
- Confidence Intervals
  - State What is Estimated
  - Distinguish Point & Interval Estimates
  - Explain Interval Estimates
  - Compute Confidence Interval Estimates for Population Mean & Proportion
- Compute Sample Size

102 Next Time: Hypothesis Testing

End of Chapter

Any blank slides that follow are
blank intentionally.