Chapter 7 - Estimates and Sample Sizes(Confidence Intervals)

Estimating a Population Proportion:

What we need:

1. **Random Sample size (Without it we could not estimate)**
2. **Conditions for Binomial:**
   a. **Fixed number of Trials**
   b. **Trials are Independent**
   c. **Two outcomes - Success & Failure**
   d. $n * p \geq 5, \ n * q \geq 5$
      n = Sample size
      p = Probability of Success

Point Estimate: **A point estimate is a single value (or point from a sample) used to approximate a population parameter.**

p = Population proportion of Success
$\hat{p}$ **= Sample proportion of Success**

$\hat{p} = \frac{x}{n}$ **(x = Number of Success, n = Number of Trials)**

$\hat{q}$ **= Sample proportion of Failure**

$\hat{q} = 1 - \hat{p}$

IMP: $\hat{p}$ **is a point estimate for P**

Eg: Touch therapist 280 trials, Touch therapist choose correct hand 123 times.

$\hat{p} = \frac{x}{n} = \frac{123}{280} = 0.44$

44% of the times she was correct or 44% chances that she will choose the right hand.

In Example, we saw that 0.44 was our best point estimate of the population proportion p, but we have no indication of just how good our best estimate is. Because a point estimate has the serious flaw of not revealing anything about how good it is, statisticians have cleverly developed another type of estimate. This estimate called a confidence interval or interval estimate, consists of a range (or an interval) of values instead of just a single value.

Confidence Intervals:

A confidence interval (or interval estimate) is a range (or an interval) of values used to estimate the true value of a population parameter. A confidence interval is sometimes abbreviated as CI.

They Have:

1. **Confidence level - This tells you how confident you are that the actual population parameter will be in the range(interval) you are giving me.**

   1 - α, α denotes the complement of a confidence level.

   Most common are,

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

Eg: The 95% CI for P is 0.381 < p < 0.497. Interpretation - " I don't know what p actually is but I'm 95% sure that p falls in the 95% range"

2. **Critical Value: A Z score that separates the 'Likely' region from the 'Unlikely' region.**

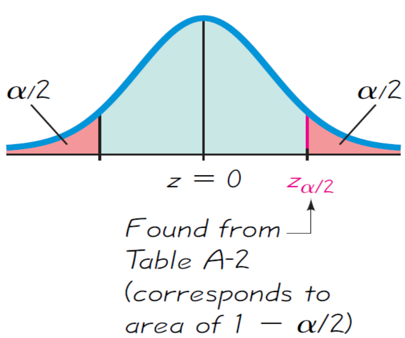

Figure 7-2 Critical Value $z_{\alpha/2}$ in the Standard Normal Distribution

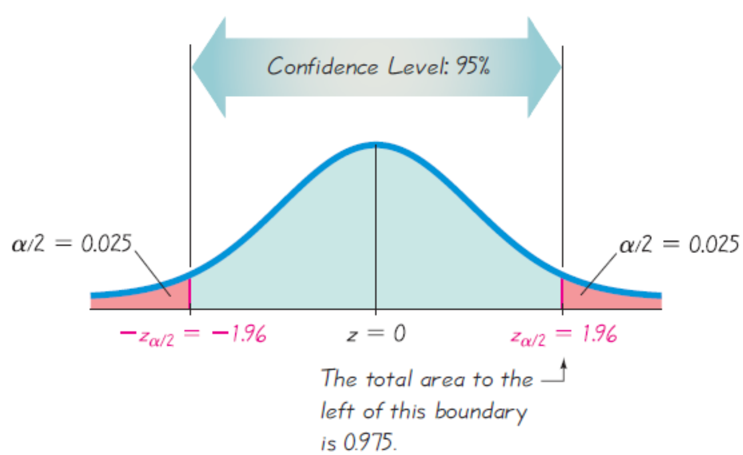

Figure 7-3 Finding $z_{\alpha/2}$ for a 95% Confidence Level

3. **Margin of Error: The maximum difference between $\hat{p}$ and p**

$$E = z_{\alpha/2}\sqrt{\frac{\widehat{p} * \widehat{q}}{n}}$$

Steps to find E:

1. **Find** $\widehat{p}$, $\widehat{q}$, n
2. **Use confidence level to find** $z_{\alpha/2}$ **(critical value)**
3. **Find E.**
4. **Confidence interval -** $\widehat{p}$ **- E < p <** $\widehat{p}$ **+ E, Also witten as - p =** $\widehat{p} \pm E$

Eg: 1. Touch therapist with 280 trials, 123 correct identifications construct a 95% CI for p.

$$\widehat{p} = \frac{123}{280} = 0.44$$

$$\widehat{q} = 1 - \widehat{p} = 1 - 0.44 = 0.56$$

n = 280

For 95% $\alpha = 0.05,\ z_{\alpha/2} = 1.96$ **(From above table)**

$$z_{\alpha/2} = 1.96$$

$$E = z_{\alpha/2}\sqrt{\frac{\widehat{p} * \widehat{q}}{n}}$$

$$E = 1.96\sqrt{\frac{\widehat{0.44} * \widehat{0.56}}{280}}$$

E = 0.0581

p = $\widehat{0.44} \pm 0.0581$

p = 0.4981 or 0.3819

p = 0.44 - 0.581 < p < 0.4393 + 0.0581 = 0.3819 < p < 0.4981
I don’t know what the exact value Population Proportion but I’m 95% sure that will follow 0.3819 < p < 0.4981 range.

IMP: As n increases E decreases

Determining Sample Size:

Given an 'E', you can find the sample size needed to get that 'E'

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

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

$$\frac{E^2}{z^2_{\alpha/2}} = \frac{\hat{p} * \hat{q}}{n}$$

$$\frac{n}{z^2_{\alpha/2}} = \frac{\hat{p} * \hat{q}}{E^2}$$

$$n = \frac{z^2_{\alpha/2} * \hat{p} * \hat{q}}{E^2}$$

$n = \frac{z^2_{\alpha/2} * 0.25}{E^2}$ **for Worst case scenario where there is 50% chance that** $\hat{p} = 0.5,\ \hat{q} = 0.5$

Eg: We want to determine the % of US people who use E mail at 95% Confidence level , What does the sample size needed to be to ensure a margin of error of 4%.

a. **16.9% of people used E mail in 1997.**

   n = ?, $\hat{p}$ **= 0.169,** $\hat{q} = 1 - \hat{p} = 0.831$**,** $z_{\alpha/2}$**= 1.96, E = 4%**

   $$n = \frac{z^2_{\alpha/2} * \hat{p} * \hat{q}}{E^2} = \frac{1.96^2 * 0.169 * 0.831}{0.04^2} = \frac{3.8416 * 0.169 * 0.831}{0.0016} = 337.194$$

b. **Assume we know nothing about previous E mail usage (So no** $\hat{p}$ **and so no** $\hat{q}$**)**

   n = ?, $\hat{p}$ **= 0.25 (Assuming),** $\hat{q} = 0.25$ **(Assuming),** $z_{\alpha/2}$**= 1.96, E = 4%**

   $$n = \frac{z^2_{\alpha/2} * \hat{p} * \hat{q}}{E^2} = \frac{1.96^2 * 0.25}{0.04^2} = 600.25$$

Given a C.I find $\widehat{p}$ & E

$\widehat{p}$ **- E < p <** $\widehat{p}$ **+ E**

$$\widehat{p} = \frac{(\widehat{p} - E) + (\widehat{p} + E)}{2} \ or \ \frac{upper + lower}{2}$$

$$E = \frac{(\widehat{p} + E) - (\widehat{p} - E)}{2} \ or \ \frac{upper - lower}{2}$$

Eg:

1. 95% CI and got 0.58 < p < 0.81

Upper = 0.81
Lower = 0.58

$$\widehat{p} = \frac{upper + lower}{2} = \frac{0.81 + 0.58}{2} = 0.695$$

$$E = \frac{upper - lower}{2} = \frac{0.81 - 0.58}{2} = 0.115$$

2. In a study of 1300 randomly selected Medical lawsuits, 900 of them were dropped.

a. Find the best Point Estimate for Population Proportion

$$\widehat{p} = \frac{900}{1300} = 0.6923$$

b. Construct a 99% Confidence interval

$\widehat{p} = 0.6923, \widehat{q} = 1 - 0.6923 = 0.3077, n = 1300$
$z_{\alpha/2} = 2.575$ **from table**

$$E = z_{\alpha/2}\sqrt{\frac{\widehat{p} * \widehat{q}}{n}} = 2.575\sqrt{\frac{\widehat{0.6923} * \widehat{0.3077}}{1300}} = 0.0329$$

p = $\widehat{p} \pm E = 0.6923 \pm 0.0329$

p = 0.6594
p = 0.7252

c. Are most cases dropped?

Yes because 65% to 72% of the cases were dropped.

Estimating a Population Mean(μ): Std Dev (σ) Known

What you need:

1. **Random Sample**
2. **Population std dev (σ) is known**
3. **n > 30 or Population is normally distributed.**

Point Estimate: $\bar{x}$ **(Sample mean) is the point estimate for** μ **(Population mean)**

Margin of error:

Confidence Interval for Estimating a Population Mean (with σ Known)

Objective

Construct a confidence interval used to estimate a population mean.

Notation

μ = population mean

σ = population standard deviation

$\bar{x}$ = sample mean

n = number of sample values

E = margin of error

$z_{\alpha/2}$ = z score separating an area of $\alpha/2$ in the right tail of the standard normal distribution

Requirements

1. The sample is a simple random sample.
2. The value of the population standard deviation σ is known.
3. Either or both of these conditions is satisfied: The population is normally distributed or $n > 30$.

Confidence Interval

$$\bar{x} - E < \mu < \bar{x} + E \quad \text{where} \quad E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

or

$$\bar{x} \pm E$$

or

$$(\bar{x} - E, \bar{x} + E)$$

Steps:

1. **Check Requirements**
2. **Find Critical Value** $z_{\alpha/2}$ **for given confidence value**
3. **Find** $E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
4. **Make confidence interval.** $\overline{x} - E < \mu < \overline{x} + E$

Eg:

1. Random sample of 40 students, the average resting heart rate for the sample was 76.3 bpm. Assume the population std dev is 12.5 bpm. Construct a 99% confidence interval of the resting heart rate.

$z_{\alpha/2}$ **for 99% CI is 2.575 (From table)**

$$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 2.575\frac{12.5}{\sqrt{40}} = 2.575\frac{12.5}{6.32456} = 5.0892$$

$$\overline{x} - E < \mu < \overline{x} + E$$

76.3 - 5.0892 < μ < 76.3 + 5.0892

71.2108 < μ < 81.3892

Interpretation : " Don't know what exact alue of population mean is but i'm 99% confident that it will fall in 71.2108 to 81.3892 range. "

Finding required sample size for a given margin of error:

$$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$\frac{E}{z_{\alpha/2}} = \frac{\sigma}{\sqrt{n}}$$

$$\frac{z_{\alpha/2}}{E} = \frac{\sqrt{n}}{\sigma}$$

$$\frac{z_{\alpha/2}{}^{*}\sigma}{E} = \sqrt{n}$$

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

Eg:

1. We want to construct a 95% CI for mean of IQ scores, assume population std dev is 15. I want the sample mean to be within 2 points of the pop mean. How big does the sample need to be?

$z_{\alpha/2}$ **= 1.96 ,** $\sigma = 15$ **, E = 2**

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

$$n = \left(\frac{1.96 * 15}{2}\right)^2$$

n = 216.09 ≅ 217

Estimating a Population Mean μ: σ Not Known

If you don't know σ, you can't use a Z score. Instead we use a T Score.

T - Score:

1. **We need to have a random sample size**
2. **n > 30 or Sample is from a normally distributed population**

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

s = std dev of a sample

Critical Values are given by $t_{\alpha/2}$

Confidence Interval for Estimating a Population Mean (with σ Not Known)

Objective

Construct a confidence interval used to estimate a population mean.

Notation

μ = population mean

$\bar{x}$ = sample mean

s = sample standard deviation

n = number of sample values

E = margin of error

$t_{\alpha/2}$ = critical t value separating an area of $\alpha/2$ in the right tail of the t distribution

Requirements

1. The sample is a simple random sample.
2. Either the sample is from a normally distributed population or $n > 30$.

Confidence Interval

$$\bar{x} - E < \mu < \bar{x} + E \quad \text{where} \quad E = t_{\alpha/2}\frac{s}{\sqrt{n}} \quad (\text{df} = n - 1)$$

or

$$\bar{x} \pm E$$

or

$$(\bar{x} - E, \bar{x} + E)$$

Degrees of Freedom: The number of degrees of freedom for a collection of sample data is the number of sample values that can vary after certain restrictions have been imposed on all data values. The number of degrees of freedom is often abbreviated as df.

df = n - 1

df = sample size - 1

Eg:

1. Sample 23 from a normal distributed population. Find critical value $t_{\alpha/2}$ for CI of 95%

α = 0.95, n = 23 , df = 23 - 1 = 22,

$t_{\alpha/2}$ = 2.074

Margin of error: E

$$E = t_{\alpha/2} * \frac{s}{\sqrt{n}}$$

Confidence Intervals: $\overline{x} - E < \mu < \overline{x} + E$

Eg:

1. Construct a 95% CI for the average age of people denied a promotion: In a random sample of 23 People, the average age was 47.0 with a std dev of 7.2. Assume this sample comes froma population that is normally distributed.

$n = 23,\ df = 22,\ \overline{x} = 47,\ \sigma = 7.2$

$E = t_{\alpha/2} * \frac{s}{\sqrt{n}}$

$E = 2.074 * \frac{7.2}{\sqrt{23}} = 2.074 * \frac{7.2}{4.7958} = 3.1137$

$\overline{x} - E < \mu < \overline{x} + E$

$\overline{47} - 3.1137 < \mu < \overline{47} + 3.1137$

43.8863 < μ < 50.1137

2. $\overline{x} - E < \mu < \overline{x} + E$

24.065 < μ < 27.218

To find $\overline{x}$ **:** $\frac{upper + lower}{2} = \frac{27.218 + 24.065}{2} = 25.6415$

To find E: $\frac{upper - lower}{2} = \frac{27.218 - 24.065}{2} = 1.5765$

3. In a random sample of 190 babies born to cocaine using mothers. The average weight was found to be 2700g with a std dev of 645g. Construct a 99% CI for the population average birth weight.

$n = 190,\ \bar{x} = 2700,\ s = 645,\ \alpha = 0.01,\ df = 189$

$t_{\alpha/2}$ **= 2.601**

$E = t_{\alpha/2} * \frac{s}{\sqrt{n}} = 2.601 * \frac{645}{\sqrt{190}} = 2.601 * \frac{645}{13.7840} = 121.7095$

$\bar{x} - E < \mu < \bar{x} + E$

$2700 - 121.7095 < \mu < 2700 + 121.7095$

2578.2905 < μ **< 2821.7095**

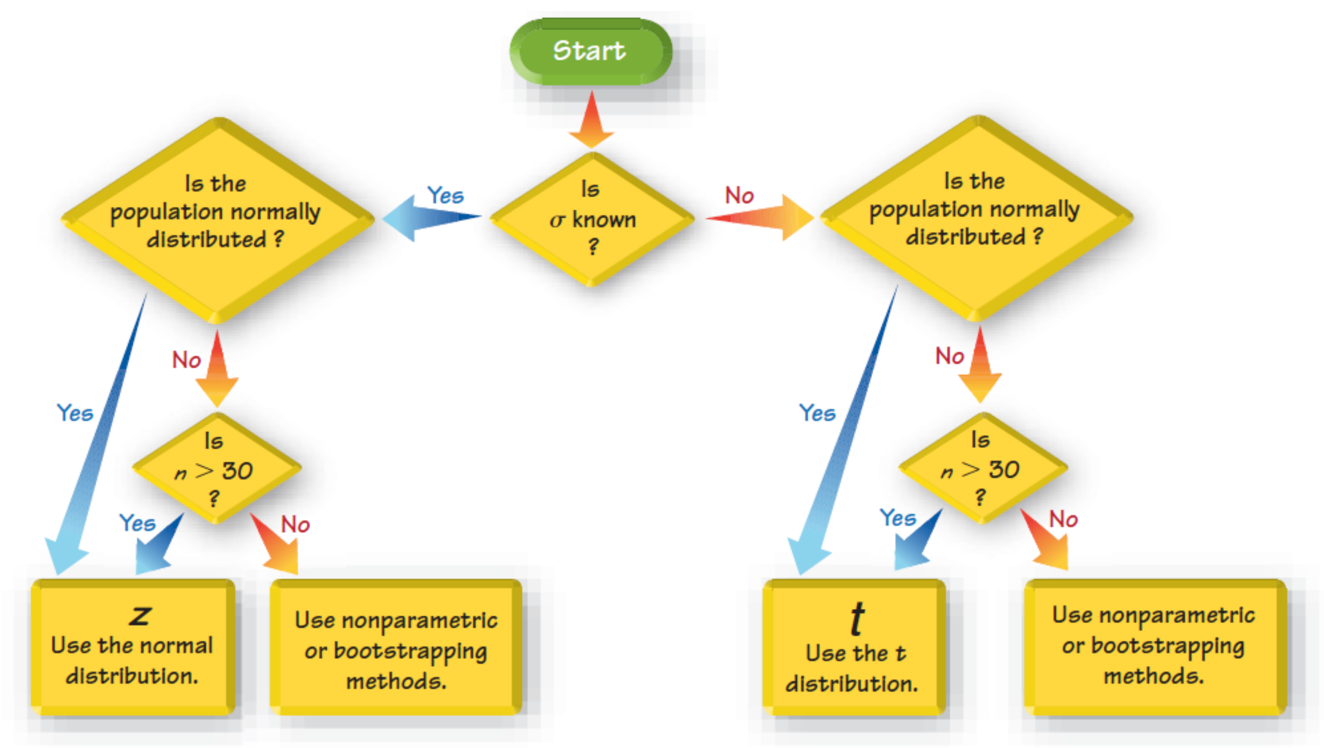

Figure 7-6 **Choosing Between *z* and *t***

Method	Conditions
Use normal (z) distribution.	σ known and normally distributed population *or* σ known and $n > 30$
Use t distribution.	σ not known and normally distributed population *or* σ not known and $n > 30$
Use a nonparametric method or bootstrapping.	Population is not normally distributed and $n \le 30$.

Notes:

1. **Criteria for deciding whether the population is normally distributed:** The population need not be exactly normal, but it should appear to be somewhat symmetric with one mode and no outliers.
2. **Sample size $n > 30$:** This is a common guideline, but sample sizes of 15 to 30 are adequate if the population appears to have a distribution that is not far from being normal and there are no outliers. For some population distributions that are extremely far from normal, the sample size might need to be much larger than 30.

Estimating a Population Variance

Chi-Square(χ) Distribution:

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

n - sample value
s^2 - Sample Variance
σ^2 - population variance

Properties of the Chi-Square Distribution:

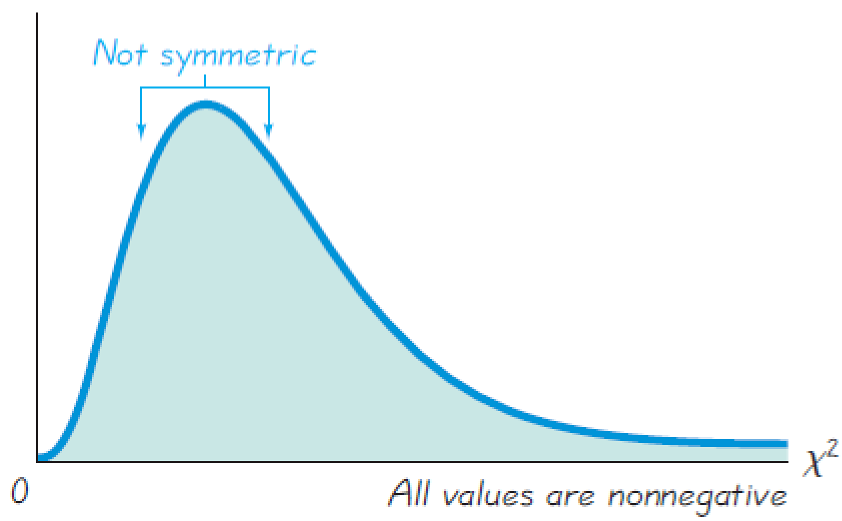

Figure 7-8 Chi-Square Distribution

1. **Not Symmetrical**
2. **Values are not negative**
3. **As df goes up distribution becomes more symmetrical**
4. **Gives critical values to the area to the Right.**

Eg:

n = 12, 95% Confidence level. Find the critical value.
α = 0.05, df = 12 - 1 = 11

Critical value = 21.920 (From table we found Right Chi squared distribution critical value $\left(\chi_R^2\right)$)

For left: 1 - 0.025 = 0.975
Critical value = 3.816 (From table $\left(\chi_L^2\right)$)

Confidence Interval: $\frac{(n-1)\,s^2}{(\chi^2_R)} < \sigma^2 < \frac{(n-1)s^2}{(\chi^2_L)}$

Eg:

1. Sample of 10 appliances, std dev of 0.15, construct a 95% CI for σ^2 & σ

$s = 0.15,\ n = 10,\ df = 10 - 1 = 9$

(χ^2_R) **= 19.023** (χ^2_L) **= 2.7**

$\frac{(n-1)\,s^2}{(\chi^2_R)} < \sigma^2 < \frac{(n-1)s^2}{(\chi^2_L)}$

$\frac{9 * 0.15^2}{19.023} < \sigma^2 < \frac{9 * 0.15^2}{2.7}$

0.0106 < σ^2 < 0.0833 (Variance)

0.1029 < σ^2 <0.2886 (S.D)

Ch. 7: Sample Size Determination

$$n = \frac{[z\alpha/2]^2 \cdot 0.25}{E^2} \quad \text{Proportion}$$

$$n = \frac{[z\alpha/2]^2 \hat{p}\hat{q}}{E^2} \quad \text{Proportion } (\hat{p} \text{ and } \hat{q} \text{ are known})$$

$$n = \left[\frac{z\alpha/2\sigma}{E}\right]^2 \quad \text{Mean}$$