21DS636-Statistical Modelling

Estimation in Statistics

Dr. Deepa Gupta, professor
Human Language Technology Lab
Department of Computer Science & Engineering
ASE-Bangalore

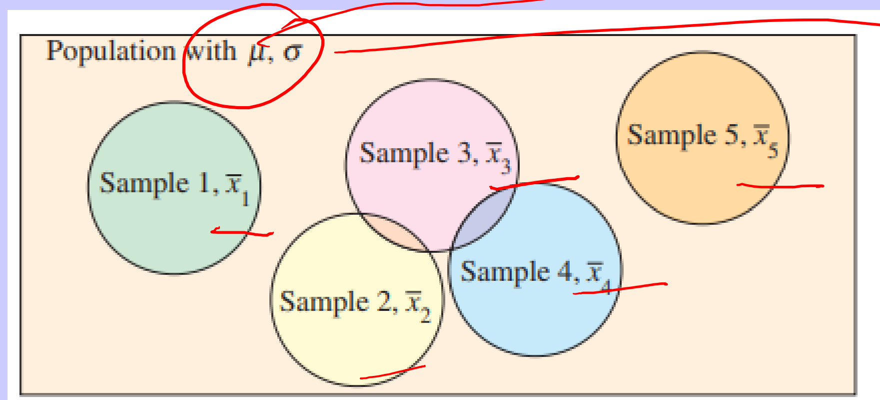
Slides source:
Professor J. Petry, multiple open web materials

Recap

A **sampling distribution** is the probability distribution of a sample statistic that is formed when samples of size n are repeatedly taken from a population. If the sample statistic is the sample mean, then the distribution is the sampling distribution of sample means. Every sample statistic has a sampling distribution.

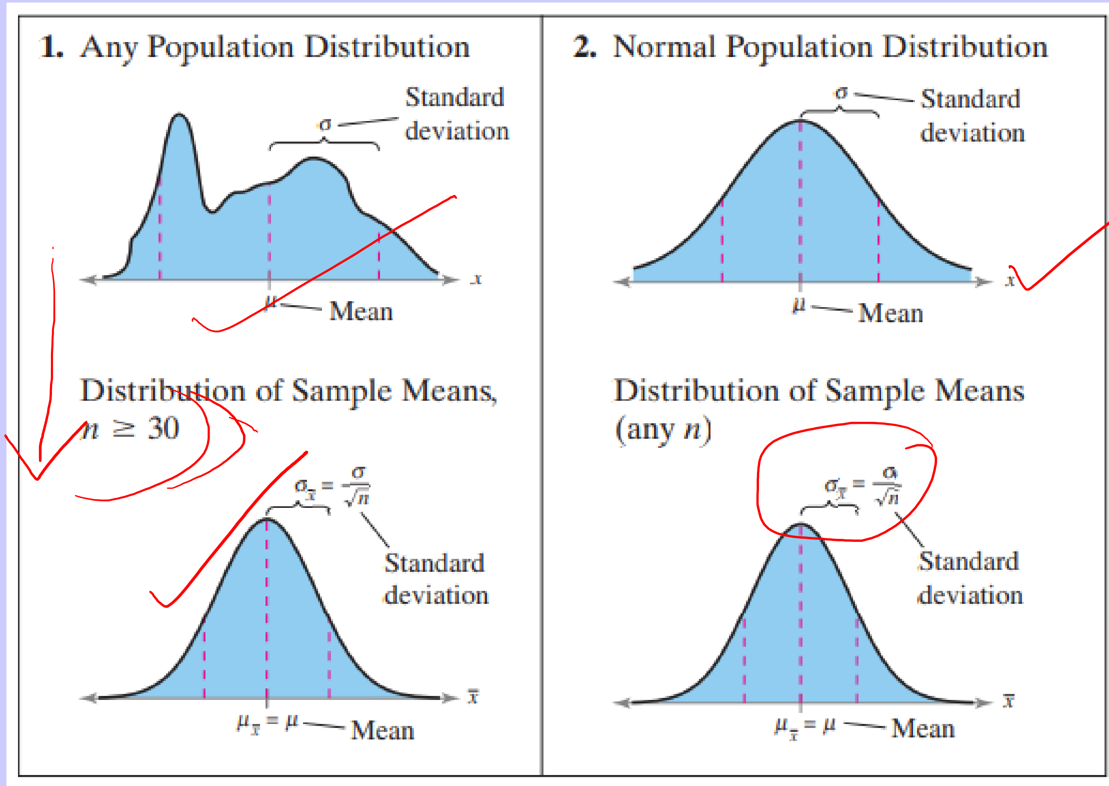
μ ?

unknown



INSIGHT: Sample means can vary from one another and can also vary from the population mean. This type of variation is to be expected and is called sampling error.

THE CENTRAL LIMIT THEOREM



$\bar{X}$

$\mu, \frac{\sigma}{\sqrt{n}}$

INSIGHT: The distribution of sample means has the same mean as the population. But its standard deviation is less than the standard deviation of the population. This tells you that the distribution of sample means has the same center as the population, but it is not as spread out. Moreover, the distribution of sample means becomes less and less spread out (tighter concentration about the mean) as the sample size n increases

Introduction

- Statistical inference is the process by which we acquire information about populations from samples.
- There are two procedures for making inferences:
  - Estimation.
  - Hypotheses testing.

Concepts of Estimation

- The objective of estimation is to determine the value of a population parameter based on a sample statistic.
- There are two types of estimators
  - Point Estimator
  - Interval estimator

STATISTICAL INFERENCE

μ

- Determining certain unknown properties of a probability distribution on the basis of a sample (usually, a r.s.) obtained from that distribution

μ

Point Estimation: $\hat{\mu} = 5$ $\qquad$ μ

Interval Estimation: $3 \le \mu \le 8$ $\qquad$ μ

Hypothesis Testing: $H_0 : \mu = 5$

$H_1 : \mu \ne 5$

STATISTICAL INFERENCE

$\Omega = \{\theta_1, \theta_2 - \theta_k\}$

$X \sim Bin(n, p)$
$p = \theta$

- ***Parameter Space (Ω or Θ):*** The set of all possible values of an unknown parameter, θ; $\theta \in \Omega$.

$X \sim Poi(\lambda)$
$\theta = \lambda$

- A pdf with unknown parameter: $f(x;\ \theta)$, $\theta \in \Omega$.
- **Estimation:** Where in Ω, θ is likely to be?

$\{\ f(x;\ \theta),\ \theta \in \Omega\ \} \longrightarrow$ The family of pdfs

STATISTICAL INFERENCE

- **Statistic:** A function of rvs (usually a sample rvs in an estimation) which does not contain any unknown parameters.

$$\bar{X}, S^2, etc$$

- **Estimator** of an unknown parameter θ: $\hat{\theta}$

A statistic used for estimating θ.

$$\hat{\theta} : estimator = U\left(X_1, X_2, \cdots, X_n\right)$$

$$\bar{X} : Estimator$$

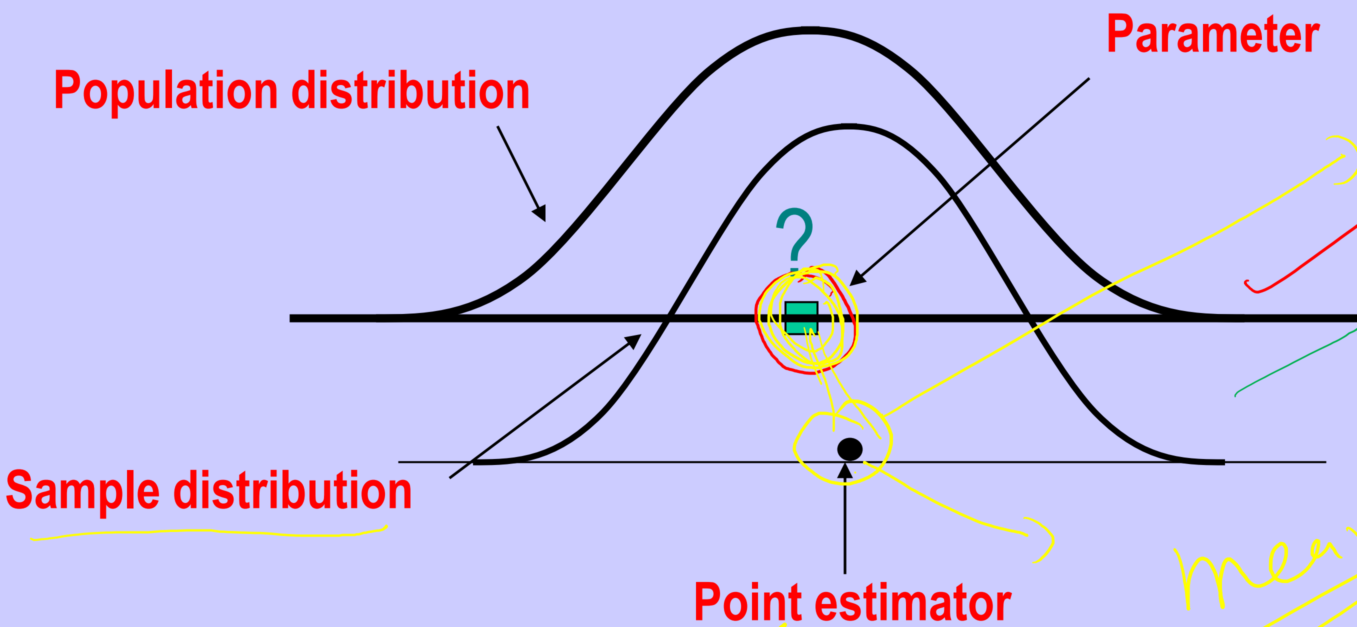
$$\downarrow$$

An observed value

$\bar{x}$: *Estimate* : A particular value of an estimator

Point Estimator

- point estimate is a single value estimate for a population parameter.
- A point estimator draws inference about a population by estimating the value of an unknown parameter using a single value or a point.



Point estimat

mean

θ: *a parameter of interest; unknown*

Goal: Find *good* estimator(s) for *θ or its function g(θ).*

AMRITA VISHWA VIDYAPEETHAM

Methods of Point Estimation

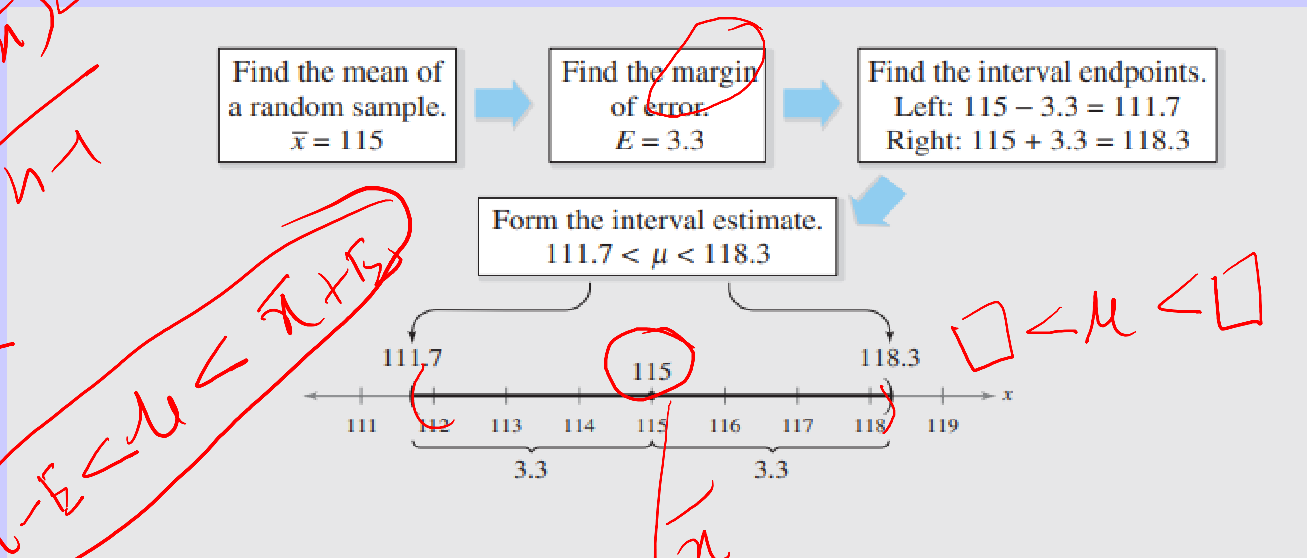
- Two methods for obtaining point estimators:
  - the method of moments
  - and the method of maximum likelihood.

- Although maximum likelihood estimators are generally preferable to moment estimators because of certain efficiency properties, they often require significantly more computation than do moment estimators.
- It is sometimes the case that these methods yield unbiased estimators.

Interval Estimation

For instance, a chess club wants to estimate the mean IQ of its members. The mean of a random sample of members is 115. Because this estimate consists of a single number represented by a point on a number line, it is called a point estimate. The problem with using a point estimate is that it is rarely equal to the exact parameter (mean, standard deviation, or proportion) of the population.

$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$



So, the club can be 90% confident that the mean IQ of its members is between 111.7 and 118.3.

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

$\boxed{} < \mu < \boxed{}$

x

$\mu \rightarrow \bar{x}$

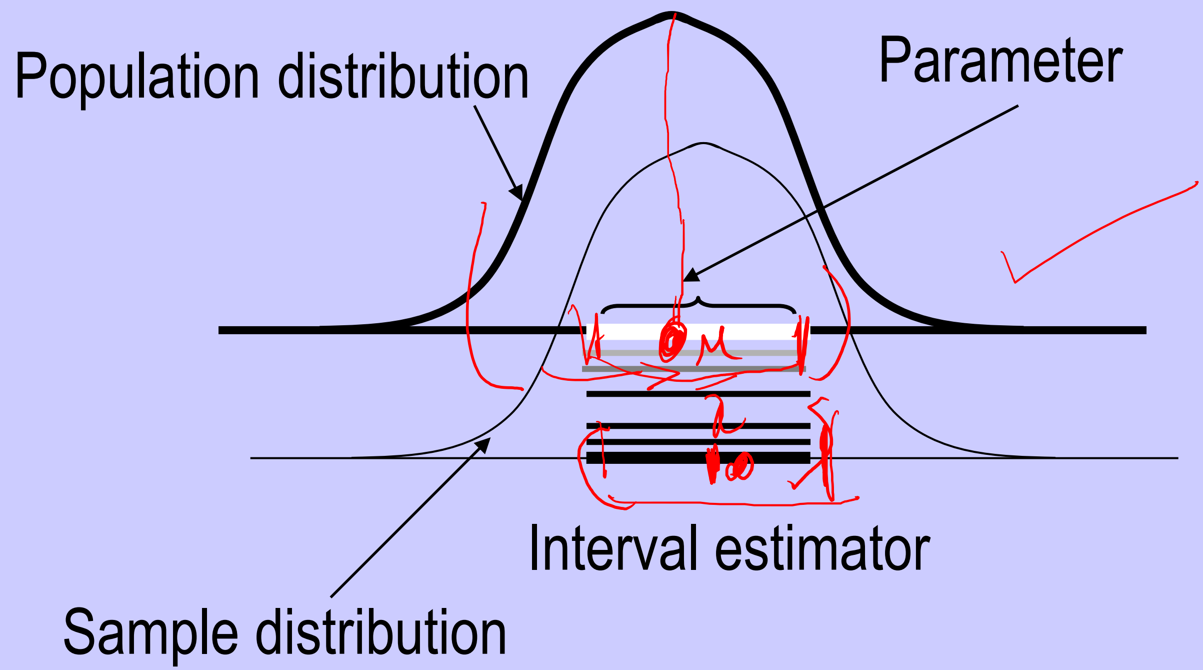
$\mu \rightarrow \bar{x}$

Definition: An interval estimate is an interval, or range of values, used to estimate a population parameter.

Definition: The difference between the point estimate and the actual parameter value is called the sampling error.

Interval Estimator

- An interval estimator draws inferences about a population by estimating the value of an unknown parameter using an interval.
- The interval estimator is affected by the sample size.

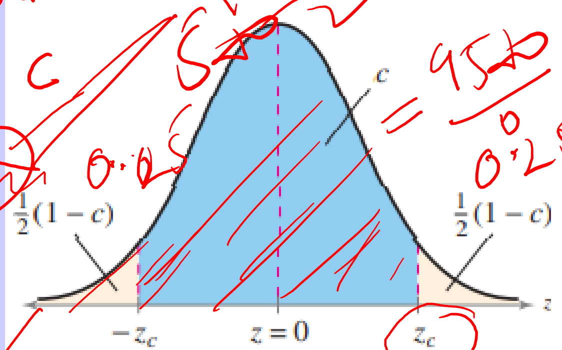


99%

Level of confidence and Critical Values

$c = (1-\alpha)\,100\%$

- **Definition:** The level of confidence c is the probability that the interval estimate contains the population parameter.
- You know from the Central Limit Theorem that when $n \geq 30$, the sampling distribution of sample means is a normal distribution.
- The level of confidence c is the area under the standard normal curve between the critical values, $-z_c$, and z_c
- **Definition:** Critical values are values that separate sample statistics that are probable from sample statistics that are improbable, or unusual.



If c = 90%:	
$c = 0.90$	Area in blue region
$1 - c = 0.10$	Area in yellow regions
$\frac{1}{2}(1-c) = 0.05$	Area in each tail
$-z_c = -1.645$	Critical value separating left tail
$z_c = 1.645$	Critical value separating right tail

$P(Z \leq 3) = \text{area}$

$1 - c$

prob = area

Margin of ERROR E

DEFINITION

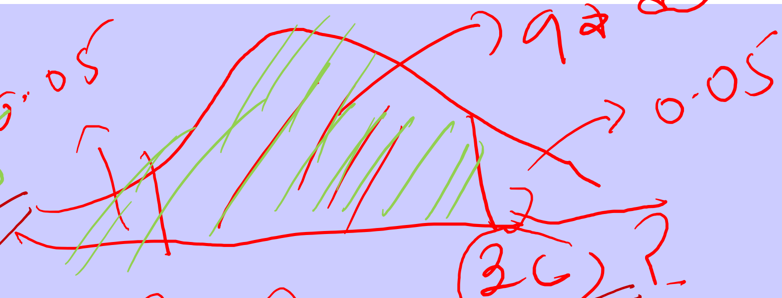
Given a level of confidence c, the **margin of error** $\boldsymbol{E}$ (sometimes also called the maximum error of estimate or error tolerance) is the greatest possible distance between the point estimate and the value of the parameter it is estimating.

$$E = z_c \sigma_{\bar{x}} = z_c \frac{\sigma}{\sqrt{n}}$$

In order to use this technique, it is assumed that the population standard deviation is known. This is rarely the case, but when $n \geq 30$, the sample standard deviation s can be used in place of σ.

$P(z \leq 3)$

$= 0.95$ 0.05



Remember that you can calculate the sample standard deviation s ~~using the~~ formula

$$s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$$

or the shortcut formula

$$s = \sqrt{\frac{\Sigma x^2 - (\Sigma x)^2/n}{n - 1}}$$

Confidence Intervals for the Population Mean

At μ

STUDY TIP

In this course, you will usually use 90%, 95%, and 99% levels of confidence. The following *z*-scores correspond to these levels of confidence.

Level of Confidence	z_c
90%	1.645
95%	1.96
99%	2.575

σ — s

this is unknown

DEFINITION

A ***c*-confidence interval for the population mean μ** is

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

Point estimator

The probability that the confidence interval contains μ is c.

GUIDELINES

Finding a Confidence Interval for a Population Mean ($n \geq 30$ or σ known with a normally distributed population)

IN WORDS	IN SYMBOLS
1. Find the sample statistics n and $\bar{x}$.	$\bar{x} = \dfrac{\sum x}{n}$
2. Specify σ, if known. Otherwise, if $n \geq 30$, find the sample standard deviation s and use it as an estimate for σ.	$s = \sqrt{\dfrac{\sum (x - \bar{x})^2}{n - 1}}$
3. Find the critical value z_c that corresponds to the given level of confidence.	Use the Standard Normal Table or technology.
4. Find the margin of error E.	$E = z_c \dfrac{\sigma}{\sqrt{n}}$
5. Find the left and right endpoints and form the confidence interval.	Left endpoint: $\bar{x} - E$ Right endpoint: $\bar{x} + E$ Interval: $\bar{x} - E < \mu < \bar{x} + E$

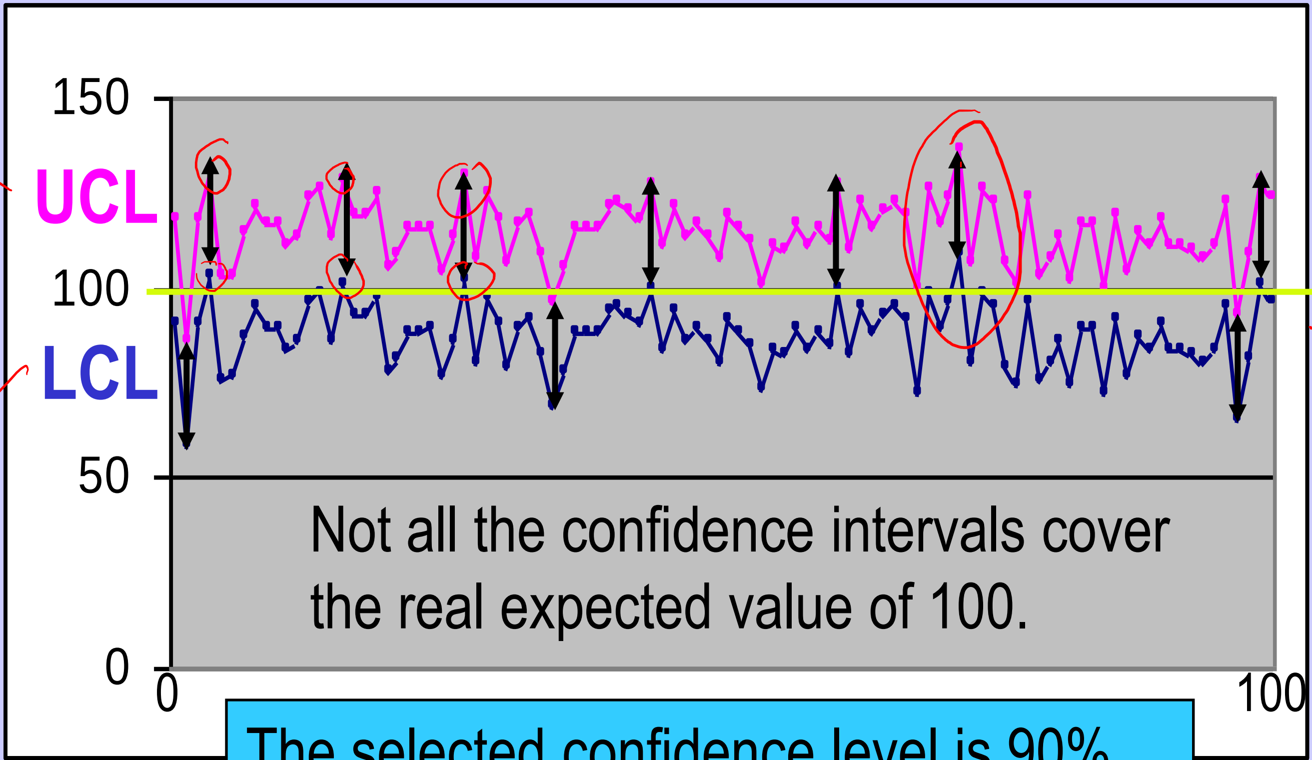
critical point

sample error

The confidence interval are correct most, but not all, of the time.

$\bar{x}+E$

$\bar{x}-E$



The selected confidence level is 90%, and 10 out of 100 intervals do not cover the real μ.

Example

A social networking website allows its users to add friends, send messages, and update their personal profiles. The following represents a random sample of the number of friends for 40 users of the website. Find a point estimate of the population mean μ. *(Adapted from Facebook)*

140	105	130	97	80	165	232	110	214	201	122
98	65	88	154	133	121	82	130	211	153	114
58	77	51	247	236	109	126	132	125	149	122
74	59	218	192	90	117	105				

$n = 40$

$\bar{x} = \dfrac{140 + 105 + \cdots + 105}{40} = 130.8$

Construct a 95% confidence interval for the mean number of friends for all users of the website.

$Z_c =$ $\quad n = 40 > 30, \quad \sigma = ?$

$s = \sqrt{\dfrac{\Sigma(x_i - \bar{x})^2}{n-1}}$

95% C.I for $\mu \Rightarrow \bar{x} - E < \mu < \bar{x} + E$

$E = z_c \dfrac{s}{\sqrt{n}} = 1.96 \cdot \dfrac{s}{\sqrt{40}} \approx 16.4$

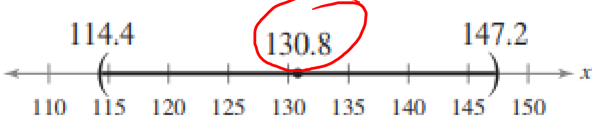
Left Endpoint

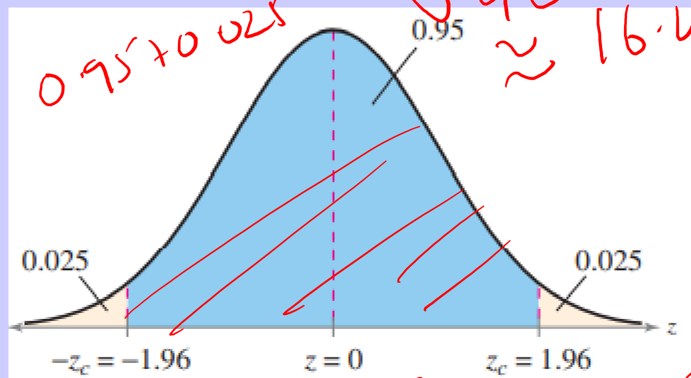
$\bar{x} - E \approx 130.8 - 16.4 = 114.4$

Right Endpoint

$\bar{x} + E \approx 130.8 + 16.4 = 147.2$

$114.4 < \mu < 147.2$





$0.95 + 0.025$

$P(Z \leq z_c) = 0.975$

Example: A college admissions director wishes to estimate the mean age of all students currently enrolled. In a random sample of 20 students, the mean age is found to be 22.9 years. From past studies, the standard deviation is known to be 1.5 years, and the population is normally distributed. Construct a 90% confidence interval of the population mean age.

Sol Given $n = 20$, $\bar{x} = 22.9$, $\sigma = 1.5$

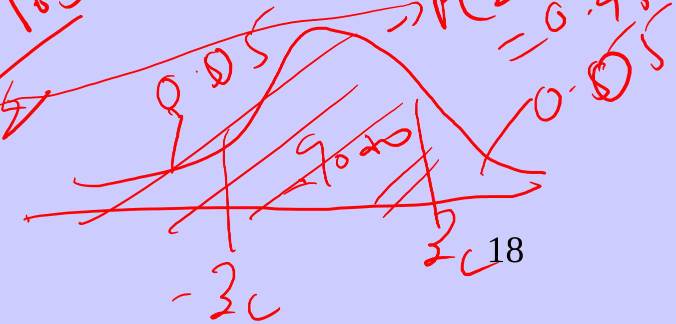
90% C.I $\Rightarrow$ $\bar{x} - E < \mu < \bar{x} + E$

$E = z_c \dfrac{\sigma}{\sqrt{n}} = 1.645 \cdot \dfrac{1.5}{\sqrt{20}}$

$22.9 - \quad < \mu < 22.9 +$

$1.64 + 1.65$

$P(Z \le z_c) = 0.95$



Example: Ten measurements of impact energy (*J*) on specimens of A238 steel cut at 60°C are as follows: 64.1, 64.7, 64.5, 64.6, 64.5, 64.3, 64.6, 64.8, 64.2, and 64.3. The impact energy is normally distributed with σ = 1*J*. Find a 95% CI for μ, the mean impact energy. Xbar=64.46

Sol//

$n = 10$, $\bar{x} = \frac{64.1 + 64.7 + \dots + 64.3}{10} = 64.46$, $\sigma = 1$

$C.I = 98\%$

$E = z_c \frac{\sigma}{\sqrt{n}} = 1.96 \cdot \frac{1}{\sqrt{10}}$

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

$64.46 - \frac{1.96}{\sqrt{10}} < \mu < 64.46 + \frac{1.96}{\sqrt{10}}$ //

Interpretation: Based on the sample data, a range of highly plausible values for mean impact energy for A238 steel at 60°C is between 63.84*J* and 65.08*J*

Confidence Intervals for the mean (Small Samples)

$n < 30$ and σ is unknown

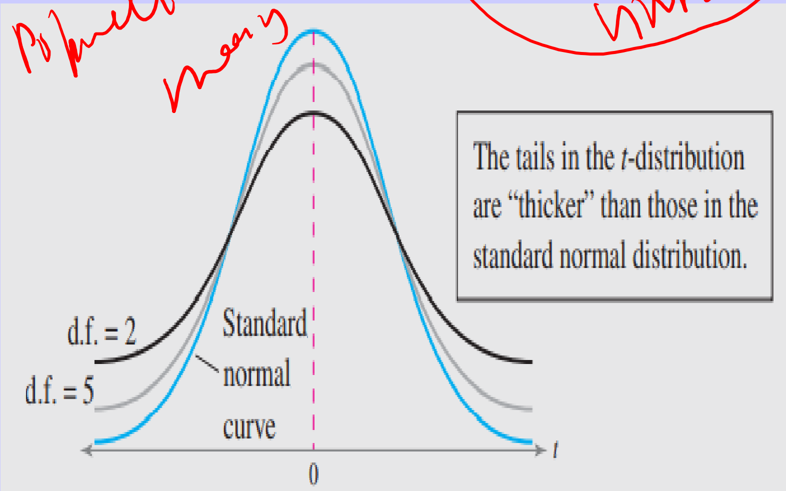
The *t* distribution

Let $X_1, X_2, \ldots, X_n$ be a random sample from a approximately normal distribution with unknown mean μ and unknown variance σ^2. The random variable

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

Population mean

has a *t* distribution with $n - 1$ degrees of freedom.



1. **The t-distribution is a family of curves, each determined by a parameter called the degrees of freedom. The degrees of freedom are the number of free choices left after a sample statistic such as is calculated. When you use a t-distribution to estimate a population mean, the degrees of freedom are equal to one less than the sample size.** The degrees of freedom (DF) in statistics **indicate the number of independent values that can vary in an analysis without breaking any constraints**.
2. **The mean, median, and mode of the t-distribution are equal to 0.**
3. **As the degrees of freedom increase, the t-distribution approaches the normal distribution. After 30 d.f., the t-distribution is very close to the standard normal z-distribution.**
4. **The t-distribution is bell-shaped and symmetric about the mean.**

Confidence Intervals for the Mean (Small Samples)

$n < 30$, σ is unknown

GUIDELINES

Constructing a Confidence Interval for the Mean: *t*-Distribution

IN WORDS	IN SYMBOLS
1. Find the sample statistics n, $\bar{x}$, and s.	$\bar{x} = \dfrac{\sum x}{n}, \quad s = \sqrt{\dfrac{\sum (x - \bar{x})^2}{n - 1}}$
2. Identify the degrees of freedom, the level of confidence c, and the critical value t_c.	d.f. $= n - 1$
	30
3. Find the margin of error E.	$E = t_c \dfrac{s}{\sqrt{n}}$
4. Find the left and right endpoints and form the confidence interval.	Left endpoint: $\bar{x} - E$ Right endpoint: $\bar{x} + E$ Interval: $\bar{x} - E < \mu < \bar{x} + E$

Example 2 Alloy Adhesion

df | 0.25 0.20
1
2

t

Construct a 95% CI on μ to the following data.

19.8	10.1	14.9	7.5	15.4	15.4
15.4	18.5	7.9	12.7	11.9	11.4
11.4	14.1	17.6	16.7	15.8	
19.5	8.8	13.6	11.9	11.4	

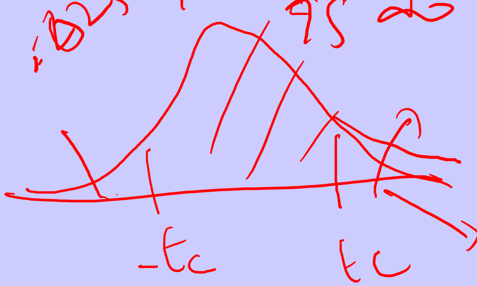
Given

$n = 22$, $\bar{x} =$, $t_c =$, $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$

< 30, σ is known, $d.f = n - 1 = 21$

.025, .95

95% C.I. is $\bar{x} - E < \mu < \bar{x} + E$



$E = t_c \cdot \frac{s}{\sqrt{n}} = t_c \cdot \frac{s}{\sqrt{22}}$ $t_c = 2.080$

Interpretation: The CI is fairly wide because there is a lot of variability in the measurements. A larger sample size would have led to a shorter interval.

$P(T \geq t_0) = P$ 0.25

Example: You randomly select 16 coffee shops and measure the temperature of the coffee sold at each. The sample mean temperature is 162.0^{0}F with a sample standard deviation of 10.0^{0}F. Construct a 95% confidence interval for the population mean temperature. Assume the temperatures are approximately normally distributed.

"σ is unknown"

Sol given: $n = 16 < 30$ $\bar{x} = 162.0$ $s = 10.0$

- 95% C.I for μ $\Rightarrow \bar{x} - E \le \mu \le \bar{x} + E$

$d.f = n-1 = 15$ $E = ? = t_c \frac{s}{\sqrt{n}}$

$= 2.131 \cdot \frac{10}{\sqrt{16}}$

Example: You randomly select 20 cars of the same model that were sold at a car dealership and determine the number of days each car sat on the dealership's lot before it was sold. The sample mean is 9.75 days, with a sample standard deviation of 2.39 days. Construct a 99% confidence interval for the population mean number of days the car model sits on the dealership's lot. Assume the days on the lot are normally distributed.

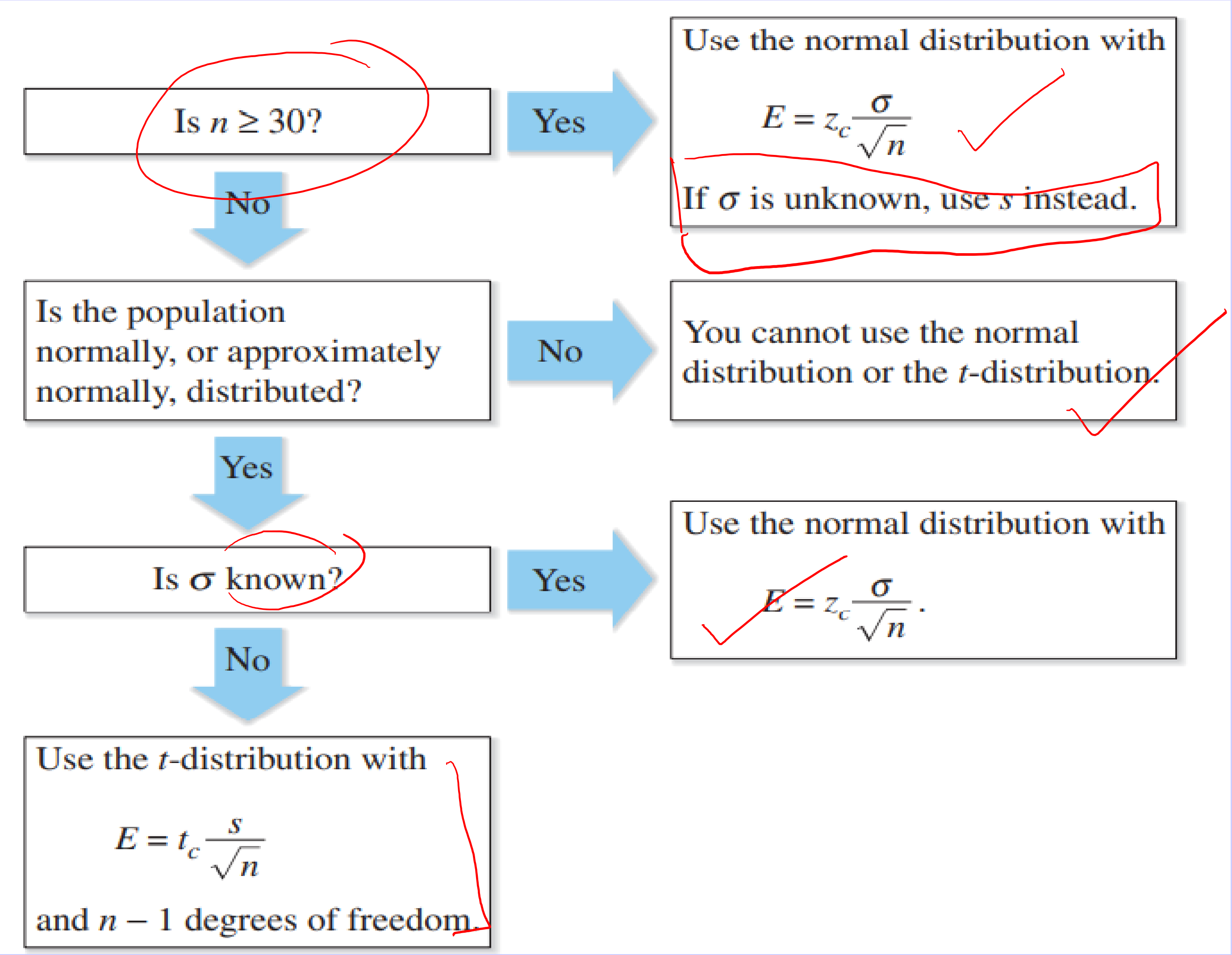
$9.75 - \cancel{3.883} \times \frac{2.39}{\sqrt{20}} < 9.75 + 3.883 \times \frac{2.39}{\sqrt{20}}$

$d.f = 19$

$p = 0.005$

2.861

The following flowchart describes when to use the normal distribution and when to use a t-distribution to construct a confidence interval for the population mean.



Confidence Interval for a population proportion

- We are going to learn:
  - How to find a point estimate for a population proportion
  - How to construct a confidence interval for a population proportion
  - How to determine the minimum sample size required when estimating a population proportion

Recap: Point Estimate for a Population Proportion

DEFINITION

The **point estimate for *p*,** the population proportion of successes, is given by the proportion of successes in a sample and is denoted by

$$\hat{p} = \frac{x}{n}$$ Sample proportion

where x is the number of successes in the sample and n is the sample size. The point estimate for the population proportion of failures is $\hat{q} = 1 - \hat{p}$. The symbols $\hat{p}$ and $\hat{q}$ are read as "p hat" and "q hat."

$\hat{\sigma}^2 = s^2$

Confidence Intervals For A Population Proportion

DEFINITION

A *c*-**confidence interval for a population proportion** p is

$$\hat{p} - E < p < \hat{p} + E$$

where

$$E = z_c \sqrt{\frac{\hat{p}\hat{q}}{n}}.$$

The probability that the confidence interval contains p is c.

c

$\hat{p} - z_c\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_c\sqrt{\frac{\hat{p}\hat{q}}{n}}$

GUIDELINES

Constructing a Confidence Interval for a Population Proportion

IN WORDS	IN SYMBOLS
1. Identify the sample statistics n and x.	
2. Find the point estimate $\hat{p}$.	$\hat{p} = \frac{x}{n}$
3. Verify that the sampling distribution of $\hat{p}$ can be approximated by a normal distribution.	$n\hat{p} \geq 5,\ n\hat{q} \geq 5$
4. Find the critical value z_c that corresponds to the given level of confidence c.	Use the Standard Normal Table or technology.
5. Find the margin of error E.	$E = z_c \sqrt{\frac{\hat{p}\hat{q}}{n}}$
6. Find the left and right endpoints and form the confidence interval.	Left endpoint: $\hat{p} - E$ Right endpoint: $\hat{p} + E$ Interval: $\hat{p} - E < p < \hat{p} + E$

Example 1: In a survey of 1000 U.S. adults, 662 said that it is acceptable to check personal e-mail while at work. Find a point estimate for the population proportion of U.S. adults who say it is acceptable to check personal e-mail while at work. Construct a 95% confidence interval for the population proportion of U.S. adults who say that it is acceptable to check personal e-mail while at work.

Sol://

Given: $n = 1000$ $\quad x = 662$ $\quad \hat{p} = \frac{x}{n} = \frac{662}{1000} = 0.662$ (P.E)

$c = 0.95$

$\hat{q} = 1 - \hat{p} = 1 - 0.662 = 0.338$

$n\hat{p}$

$n\hat{q} \geq 5$

$$\hat{p} - z_c\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_c\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$$\Rightarrow 0.662 - 1.96\sqrt{\frac{0.338 \times 0.662}{1000}} < p < 0.662 + 1.96\sqrt{\frac{0.338 \times 662}{1000}}$$

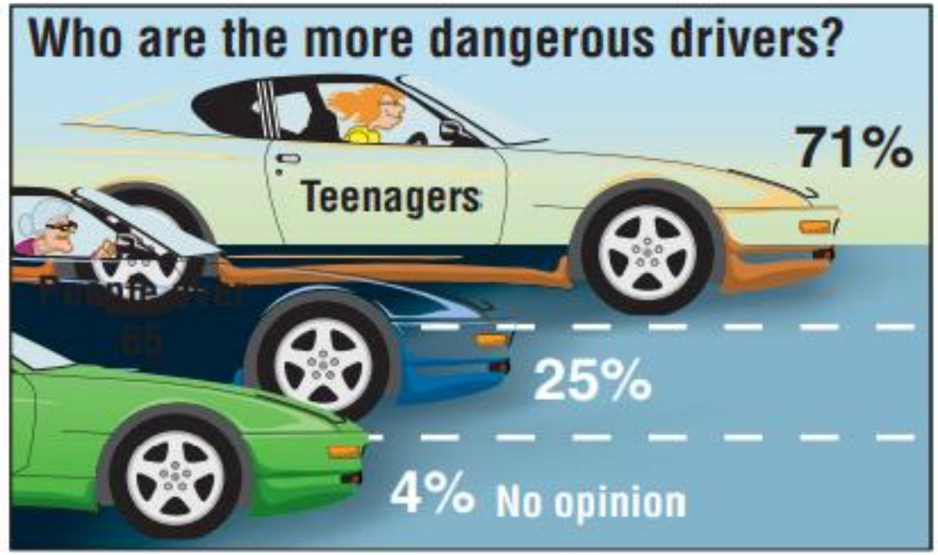
0.629

Interpretation: With 95% confidence, you can say that the population proportion of U.S. adults who say that it is acceptable to check personal e-mail while at work is between 63.3% and 69.1%

Example 2:

▸ Constructing a Confidence Interval for *p*

The graph shown at the right is from a survey of 498 U.S. adults. Construct a 99% confidence interval for the population proportion of U.S. adults who think that teenagers are the more dangerous drivers. *(Source: The Gallup Poll)*



Solution:

$0.005 \quad 0.99 \quad 0.005$

$0.995 = 0.995 \quad z_c$

$z_c = 2.575$

$n = 498 \qquad \hat{p} = 71\% \qquad \hat{q} = 1 - 0.71 = 0.29$

$n\hat{p} = 498 \times 0.71 > 5$

$n\hat{q} = 498 \times 0.29 > 5$

99% C.I for $p = \hat{p} - E < p < \hat{p} + E$

$E = z_c \sqrt{\dfrac{\hat{p}\hat{q}}{n}}$

Example 3: Use the data given in Example 2 to construct a 99% confidence interval for the population proportion of adults who think that people over 65 are the more dangerous drivers.

Interpretation: With 99% confidence, you can say that the proportion of U.S. adults who think that people over 65 are the more dangerous drivers is between 20% and 30%

Confidence Intervals for Variance and Standard Deviation

Learning Objective:

- How to interpret the chi-square distribution and use a chi-square distribution table

- How to use the chi-square distribution to construct a confidence interval for the variance and standard deviation

The Chi-square Distribution

- In manufacturing, it is necessary to control the amount that a process varies.
- For instance, an automobile part manufacturer must produce thousands of parts to be used in the manufacturing process. It is important that the parts vary little or not at all.
- How can you measure, and consequently control, the amount of variation in the parts? You can start with a point estimate.

DEFINITION

The **point estimate for σ^2** is s^2 and the **point estimate for σ** is s. The most unbiased estimate for σ^2 is s^2.

$\hat{\sigma}^2 = s^2$ $\hat{\sigma} = s$

Note: chi-square distribution to construct a confidence interval for the variance and standard deviation

χ^2 Distribution

DEFINITION

If a random variable x has a normal distribution, then the distribution of

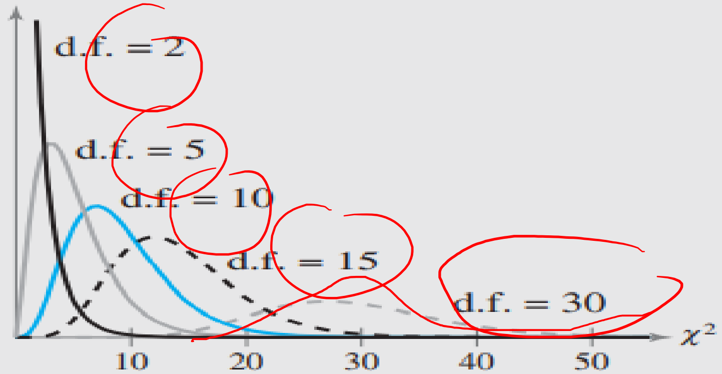
$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

forms a **chi-square distribution** for samples of any size $n > 1$. Four properties of the chi-square distribution are as follows.

1. All chi-square values χ^2 are greater than or equal to 0.
2. The chi-square distribution is a family of curves, each determined by the degrees of freedom. To form a confidence interval for σ^2, use the χ^2-distribution with degrees of freedom equal to one less than the sample size.

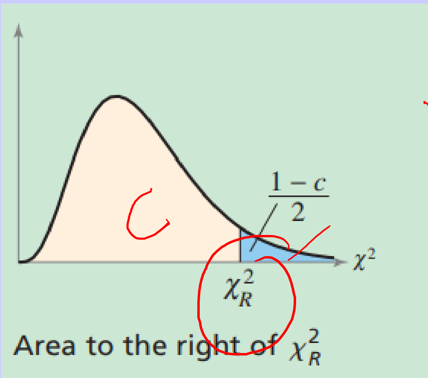
   d.f. $= n - 1$ Degrees of freedom

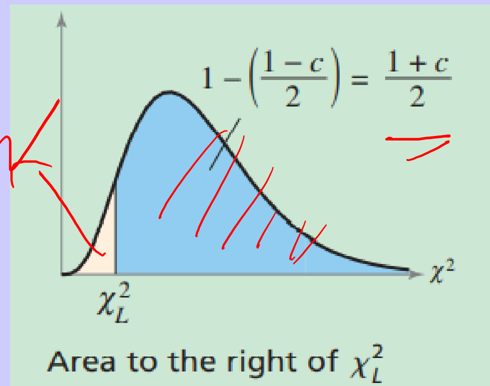
3. The area under each curve of the chi-square distribution equals 1.
4. Chi-square distributions are positively skewed.

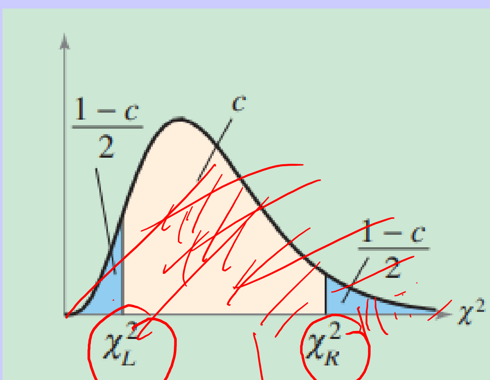


Chi-Square Distributions

Chi-square Critical Values with a C-confidence Level



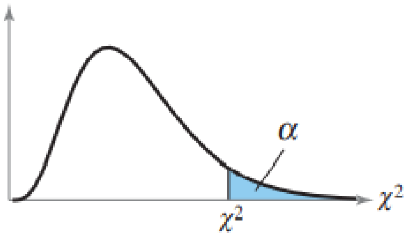




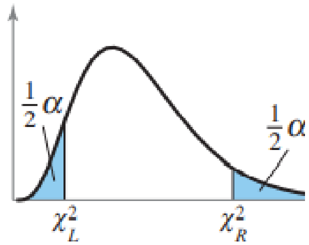
$c + \frac{1-c}{2} = \frac{2c+1-c}{2} = \frac{1+c}{2}$

- There are two critical values for each level of confidence.
- The value X^2_R represents the right-tail critical value and X^2_L represents the left-tail critical value.
-
- Chi-square Table lists critical values of for various degrees of freedom and areas.
- Each area in the table represents the region under the chi-square curve to the right of the critical value

Table 6 — Chi-Square Distribution



Right tail



Two tails

Because the sample size is 18, there are

$$\text{d.f.} = n - 1 = 18 - 1 = 17 \text{ degrees of freedom.}$$

The areas to the right of χ^2_R and χ^2_L are

$$\text{Area to right of } \chi^2_R = \frac{1-c}{2} = \frac{1-0.95}{2} = 0.025$$

and

$$\text{Area to right of } \chi^2_L = \frac{1+c}{2} = \frac{1+0.95}{2} = 0.975.$$

$\chi^2_L = 7.564$ $\quad$ $\chi^2_R = 30.191$

Degrees of freedom	α									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.299
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582

Question

Find the critical values χ^2_R and χ^2_L for a 90% confidence interval when the sample size is 30.

a. Identify the *degrees of freedom* and the *level of confidence*.
b. *Find the areas* to the right of χ^2_R and χ^2_L.

χ_R^2

find the critical values X^2_R and and X^2_L for the given confidence level c and sample size n

χ_L^2

Confidence Interval on the Variance and Standard Deviation

DEFINITION

The c-confidence intervals for the population variance and standard deviation are as follows.

Confidence Interval for σ^2:

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

Confidence Interval for σ:

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

The probability that the confidence intervals contain σ^2 or σ is c.

Example: You randomly select and weigh 30 samples of an allergy medicine. The sample standard deviation is 1.20 milligrams. Assuming the weights are normally distributed, construct 99% confidence intervals for the population variance and standard deviation.

Interpretation: With 99% confidence, you can say that the population variance is between 0.80 and 3.18, and the population standard deviation is between 0.89 and 1.78 milligrams

Example: Detergent Filling

An automatic filling machine is used to fill bottles with liquid detergent. A random sample of 20 bottles results in a sample variance of fill volume of $s^2 = 0.0153^2$. Assume that the fill volume is approximately normal. Compute a 95% confidence interval for the population variance and standard deviation.

Understanding Questions

- Does a population have to be normally distributed in order to use the chi-square distribution?

- What happens to the shape of the chi-square distribution as the degrees of freedom increase?

Factors Affecting Confidence Intervals

The width of the interval estimate is a function of:

- the confidence level
- the population standard deviation
- the sample size.

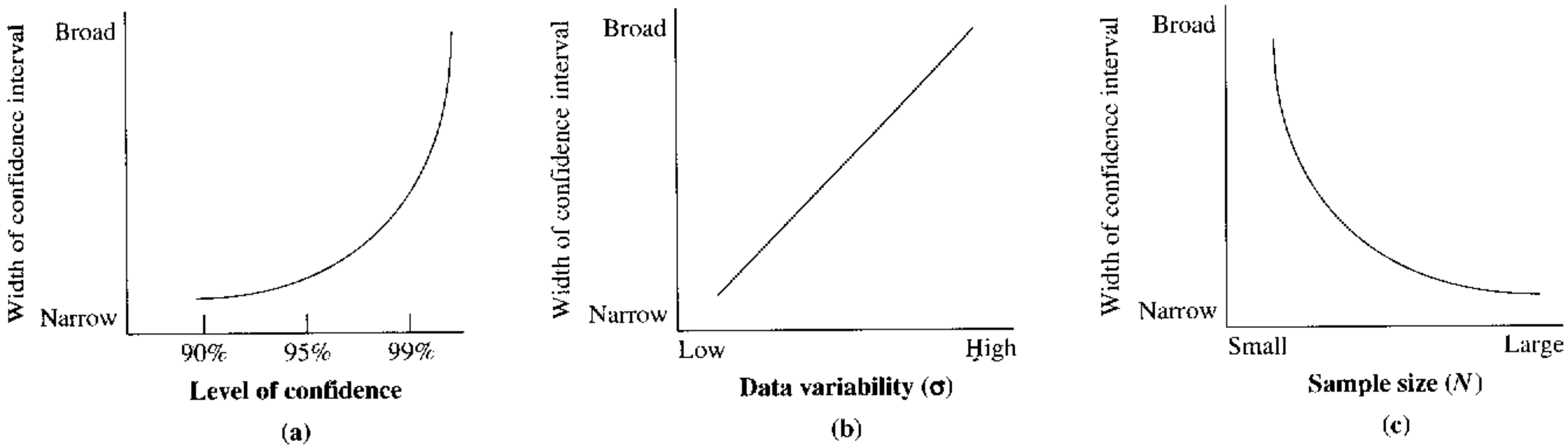


FIGURE 5.4

The width of a confidence interval varies as a function of level of confidence (a), data variability (b), and sample size (c).