

Unit 6. Estimation

“Use at least twelve observations in constructing a confidence interval”

- Gerald van Belle

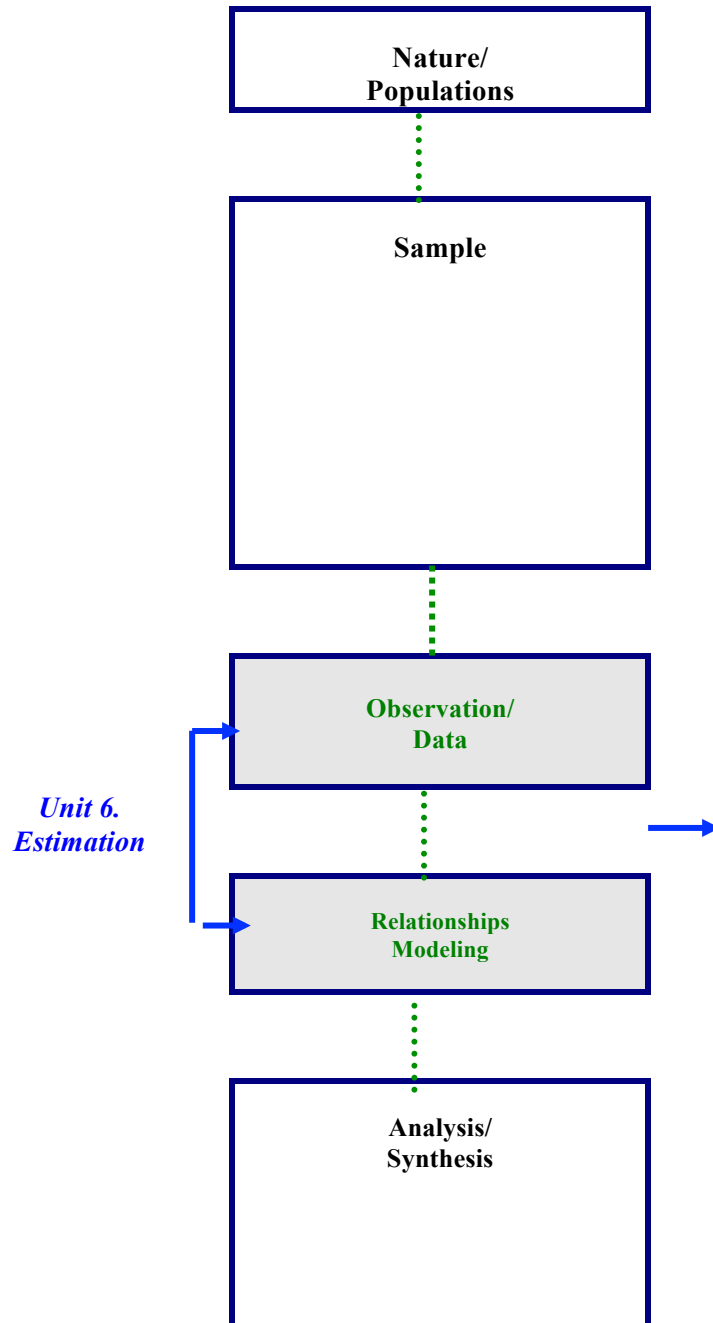
What is the mean of the blood pressures of all the students at the Amherst Regional High School? It's too much work to measure the blood pressure of every individual student in this population. So we will make a guess based on the blood pressures of a sample of students. This is estimation. Estimation involves using a statistic calculated for our sample as our estimate of the population quantity of interest. In this example, where the population mean is of interest, possible choices might be (1) the sample mean blood pressure; (2) the sample median blood pressure; (3) the average of the smallest and largest blood pressure; *and so on*. The point is, there is no one choice for estimation. What we mean by a **“good” choice of estimator** is one focus of this unit.

The other focus of this unit is **confidence interval estimation**. A confidence interval is a single estimate together with a “safety net”, which can also be thought of as a “margin of error”

Table of Contents

Topic		
	1. Unit Roadmap	3
	2. Learning Objectives	4
	3. Introduction	5
	a. Goals of Estimation	8
	b. Notation and Definitions	10
	c. How to Interpret a Confidence Interval	13
	4. Preliminaries: Some Useful Probability Distributions	20
	a. Introduction to the Student t- Distribution	20
	b. Introduction to the Chi Square Distribution	24
	c. Introduction to the F-Distribution	28
	d. Sums and Differences of Independent Normal Random Vars ..	31
	5. Normal Distribution: One Group	33
	a. Confidence Interval for μ , σ^2 Known	33
	b. Confidence Interval for μ , σ^2 Unknown	38
	c. Confidence Interval for σ^2	41
	6. Normal Distribution: Paired Data	44
	a. Confidence Interval for $\mu_{\text{DIFFERENCE}}$	45
	b. Confidence Interval for $\sigma^2_{\text{DIFFERENCE}}$	48
	7. Normal Distribution: Two Independent Groups:	49
	a. Confidence Interval for $[\mu_1 - \mu_2]$	49
	b. Confidence Interval for σ_1^2 / σ_2^2	57
	8. Binomial Distribution: One Group	60
	a. Confidence Interval for π	60
	9. Binomial Distribution: Two Independent Groups	64
	a. Confidence Interval for $[\pi_1 - \pi_2]$	64
	Appendices	
	i. Derivation of Confidence Interval for μ – Single Normal with σ^2 Known	67
	ii. Derivation of Confidence Interval for σ^2 – Single Normal	70
	iii. SE of a Binomial Proportion	72

1. Unit Roadmap



Recall that numbers that are calculated from the entirety of a population are called **population parameters**. They are represented by **Greek letters** such as μ (population mean) and σ^2 (population variance). Often, it is not feasible to calculate the value of a population parameter. Numbers that we calculate from a sample are called **statistics**. They are represented by **Roman letters** such as \bar{X} (sample mean) and S^2 (sample variance).

In our introduction to sampling distributions, we learned that a sample statistic such as \bar{X} is a random variable in its own right. This can be understood by imagining that there are infinitely many replications of our study so that there are infinitely many \bar{X} .

Putting this together, if a sample statistic such as \bar{X} is to be used as an estimate of a population parameter such as μ , the incorporation of a measure of its variability (in this case the standard error of \bar{X}) allows us to construct a “margin of error” about \bar{X} as our estimate of μ . The result is a **confidence interval estimate**.

2. Learning Objectives

When you have finished this unit, you should be able to:

- Explain that there is more than one way to estimate a population parameter based on data in a sample.
- Explain the criteria of unbiased and minimum variance in the selection of a “good” estimator.
- Define the Student t, chi square, and F probability distribution models.
- Explain that the sum and difference of independent random variables that are distributed normal are also distributed Normal.
- Interpret a confidence interval.
- Calculate point and confidence interval estimates of the mean and variance of a single Normal distribution.
- Calculate point and confidence interval estimates of the mean and variance of a single Normal distribution in the paired data setting.
- Calculate point and confidence interval estimates of the difference between the means of two independent Normal distributions.
- Calculate point and confidence interval estimates of the ratio of the variances of two independent Normal distributions.
- Calculate point and confidence interval estimates of the π (event probability) parameter of a single binomial distribution.
- Calculate point and confidence interval estimates of the difference between the π (event probability) parameters of two independent binomial distributions.

3. Introduction

Recall – Biostatistics is the application of probability models and associated tools to observed phenomena for the purposes of learning about a population and gauging the relative plausibility of alternative explanations.

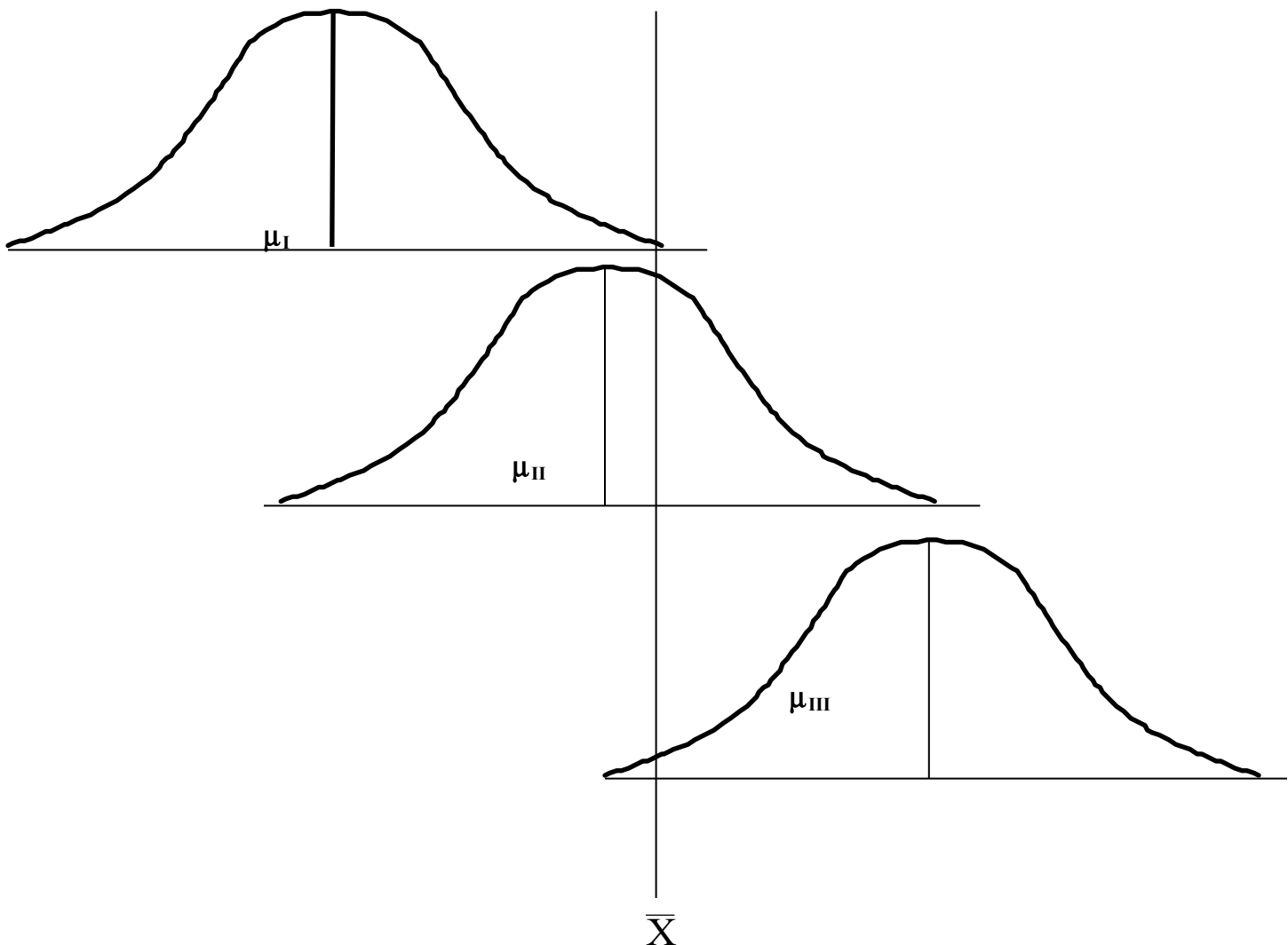
- ♣ **Description** - Information in a sample is used to summarize the sample itself. It is also used to make guesses of the characteristics of the source population.
- ♣ **Hypothesis Testing** – Information in a sample is used to help us compare different explanations for what we have observed.

Unit 6 is about using information in a sample to make estimates of the characteristics (parameters) of the population that gave rise to the sample. Recall the distinction between statistics and parameters:

Sample statistics are estimators	of population parameters
Sample mean \bar{X}	μ
Sample variance S^2	σ^2

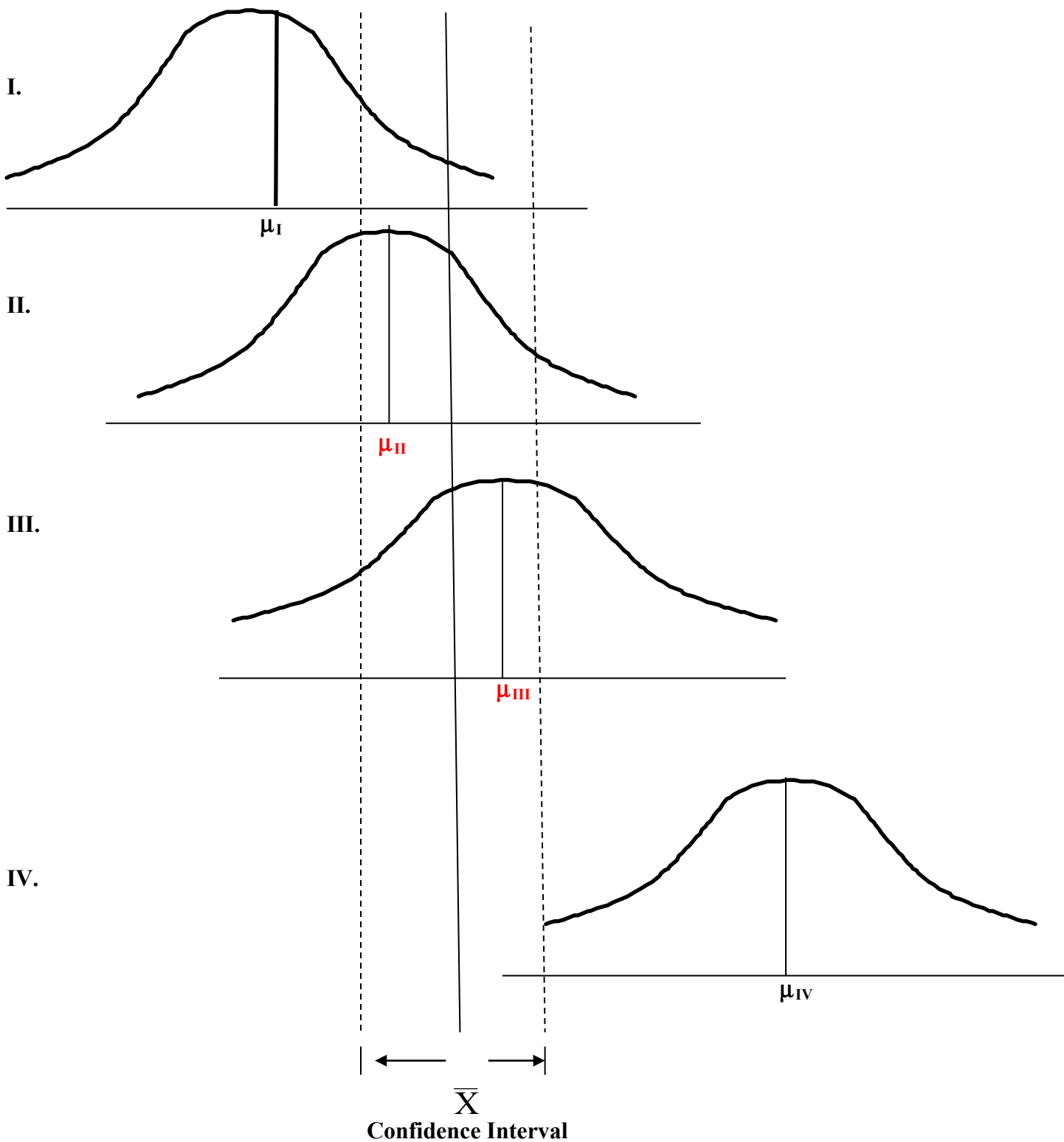
What does it mean to say we know \bar{X} from a sample but we don't know the population mean μ ?

Suppose we have a simple random sample of n observations $X_1 \dots X_n$ from some population. We have calculated the sample average \bar{X} . What population gave rise to our sample? In theory, there are infinitely many possible populations. For simplicity here, suppose there are just 3 possibilities, schematically shown below:



Suppose this is the location of our \bar{X} .

Okay, sorry. Here, I'm imagining four possibilities instead of three. Look at this page from the bottom up. Around our \bar{X} , I've constructed a "confidence" interval. Notice the dashed lines extending upwards into the 4 normal distributions. μ_I and μ_{IV} are **outside** the interval around \bar{X} . μ_{II} and μ_{III} are **inside** the interval.



We are “confident” that μ could be either μ_{II} or μ_{III} .

3a. Goals of Estimation

Whether an estimator is “good” or “not good” depends on what criteria we use to define “good”. There are potentially lots of criteria. Here, we’ll use one set of two criteria: unbiased and minimum variance.

Conventional Criteria for a Good Estimator –

1. “In the long run, correct” (unbiased)
2. “In the short run, in error by as little as possible” (minimum variance)

1. Unbiased - “In the Long Run Correct” -

Tip: Recall the introduction to statistical expectation and the meaning of unbiased (See Unit 4 – *Bernoulli & Binomial* pp 7-10).

“In the long run correct.” Imagine replicating the study over and over again, infinitely many times. Each time, calculate your statistic of interest so as to produce the sampling distribution of that statistic of interest. Now calculate the mean of the sampling distribution for your statistic of interest. Is it the same as the population parameter value that you are trying to estimate? If so, then that statistic is an unbiased estimate of the population parameter that is being estimated.

Example – Under normality and simple random sampling, S^2 as an unbiased estimate of σ^2 .

“In the long run correct” means that the statistical expectation of S^2 , computed over the sampling distribution of S^2 , is equal to its “target” σ^2 .

$$\sum_{\text{all possible samples "i"}} \left(\frac{S_i^2}{\# \text{ samples in sampling distn}} \right) = \sigma^2$$

Recall that we use the notation “E []” to refer to statistical expectation. Here it is $E [S^2] = \sigma^2$.

2. Minimum Variance “In Error by as Little as Possible” –

“In error by as little as possible.” We would like that our estimates not vary wildly from sample to sample; in fact, we’d like these to vary as little as possible. This is the idea of **precision**. When the estimates vary by as little as possible, we have **minimum variance**.

Putting together the two criteria (“long run correct” and “in error by as little as possible”)

Suppose we want to identify the minimum variance unbiased estimator of μ in the setting of a **simple random sample from a normal distribution**.

Candidate estimators might include the sample mean \bar{X} or the sample median \tilde{X} as estimators of the population mean μ . Which would be a better choice according to the criteria “in the long run correct” and “in the short run in error by as little as possible”?

Step 1 First, identify the unbiased estimators

Step 2 From among the pool of unbiased estimators, choose the one with minimum variance.

Illustration for data from a normal distribution

1. The unbiased estimators are the sample mean \bar{X} and median \tilde{X}

2. $\text{variance}[\bar{X}] < \text{variance}[\tilde{X}]$

Choose the sample mean \bar{X} . It is the minimum variance unbiased estimator.

For a random sample of data from a normal probability distribution, \bar{X} is the minimum variance unbiased estimator of the population mean μ .

Take home message:

Here, we will be using the criteria of “minimum variance unbiased”. However, other criteria are possible.

3b. Notation and Definitions

Estimation, Estimator, Estimate -

- ♣ **Estimation** is the computation of a statistic from sample data, often yielding a value that is an approximation (guess) of its target, an unknown true population parameter value.
- ♣ The statistic itself is called an **estimator** and can be of two types - point or interval.
- ♣ The value or values that the estimator assumes are called **estimates**.

Point versus Interval Estimators -

- ♣ An estimator that represents a "single best guess" is called a **point estimator**.
- ♣ When the estimate is of the form of a "range of plausible values", it is called an **interval estimator**.
Thus,

A **point estimate** is of the form:

[Value],

An **interval estimate** is of the form:

[lower limit, upper limit]

Example -

The sample mean \bar{X}_n , calculated using data in a sample of size n , is a point estimator of the population mean μ . If $\bar{X}_n = 10$, the value 10 is called a point estimate of the population mean μ .

Sampling Distribution

- ♣ **Recall the idea of a **sampling distribution**.** It is an theoretically obtained entity obtained by imagining that we repeat, over and over infinitely many times, the drawing of a simple random sample and the calculation of something from that sample, such as the sample mean \bar{X}_n based on a sample size draw of size equal to n . The resulting collection of “all possible” sample means is what we call the sampling distribution of \bar{X}_n .
- ♣ **Recall. The sampling distribution of \bar{X}_n plays a fundamental role in the central limit theorem.**

Unbiased Estimator

A statistic is said to be an **unbiased estimator** of the corresponding population parameter if its mean or expected value, taken over its sampling distribution, is equal to the population parameter value.

Intuitively, this is saying that the "long run" average of the statistic, taken over all the possibilities in the sampling distribution, has value equal to the value of its target population parameter.

Confidence Interval, Confidence Coefficient

- ♣ A **confidence interval** is a particular type of interval estimator.
- ♣ Interval estimates defined as confidence intervals provide not only several point estimates, but also a feeling for the precision of the estimates. This is because they are constructed using two ingredients:
 - 1) a point estimate, and
 - 2) the standard error of the point estimate.

Many Confidence Interval Estimators are of a Specific Form:

lower limit = (point estimate) - (confidence coefficient multiplier)(standard error)
upper limit = (point estimate) + (confidence coefficient multiplier)(standard error)

- ♣ The "multiple" in these expressions is related to the precision of the interval estimate; the multiple has a special name - **confidence coefficient**.
- ♣ A wide interval suggests imprecision of estimation. Narrow confidence interval widths reflects large sample size or low variability or both.
- ♣ Exceptions to this generic structure of a confidence interval are those for a variance parameter and those for a ratio of variance parameters

Take care when computing and interpreting a confidence interval!!

A common mistake is to calculate a confidence interval but then use it incorrectly by focusing only on its midpoint.

3c. How to Interpret a Confidence Interval

A confidence interval is a safety net.

Tip: In this section, the focus is on the **idea of a confidence interval**. For now, don't worry about the details.

Example

Suppose we want to estimate the average income from wages for a population of 5000 workers, X_1, \dots, X_{5000} . The average income that we want to estimate is the population mean μ .

$$\mu = \frac{\sum_{i=1}^{5000} X_i}{5000}$$

For purposes of this illustration, suppose we actually know the population $\sigma = \$12,573$. In real life, we wouldn't have such luxury!

Suppose the unknown $\mu = \$19,987$. Note – I'm only telling you this so that we can see how well this illustration performs!.

We'll construct two confidence interval estimates of μ to illustrate the **importance of sample size in confidence interval estimation**:

(1) from a sample size of $n=10$, versus

(2) from a sample size of $n=100$

(1) Carol uses a sample size n=10

Carol's data are X_1, \dots, X_{10}

$$\bar{X}_{n=10} = 19,887$$

$$\sigma = 12,573$$

$$SE_{\bar{X}_{n=10}} = \frac{\sigma}{\sqrt{10}} = 3,976$$

(2) Ed uses a sample size n=100

Ed's data are X_1, \dots, X_{100}

$$\bar{X}_{n=100} = 19,813$$

$$\sigma = 12,573$$

$$SE_{\bar{X}_{n=100}} = \frac{\sigma}{\sqrt{100}} = 1,257$$

Compare the two SE, one based on n=10 and the other based on n=100 ...

- The variability of an average of 100 is less than the variability of an average of 10.
- It seems reasonable that, all other things being equal, we should have **more confidence (smaller safety net)** in our sample mean as a guess of the population mean when it is based on a larger sample size (100 versus 10).
- Taking this one step further ... we ought to have complete (100%) confidence (no safety net required at all) if we interviewed the entire population!. This makes sense since we would obtain the correct answer of \$19,987 every time.

Definition Confidence Interval (Informal):

A confidence interval is a guess (point estimate) together with a “safety net” (interval) of guesses of a population characteristic. In most instances, it is easy to see the 3 components of a confidence interval:

- 1) A point estimate (e.g. the sample mean \bar{X})
- 2) The standard error of the point estimate (e.g. $SE_{\bar{X}} = \sigma / \sqrt{n}$)
- 3) A confidence coefficient (conf. coeff)

In most instances (**means, differences of means, regression parameters**, etc), the structure of a confidence interval is calculated as follows:

$$\begin{aligned} \text{Lower limit} &= (\text{point estimate}) - (\text{confidence coefficient})(SE) \\ \text{Upper limit} &= (\text{point estimate}) + (\text{confidence coefficient})(SE) \end{aligned}$$

In other instances (as you’ll see in the next pages), the structure of a confidence interval looks different, as for confidence intervals for

Population variance
Population standard deviation
Ratio of two population variances
relative risk
Odds ratio

Nature — Population/ Sample — Observation/ Data — Relationships/ Modeling — Analysis/ Synthesis

Example: Carol samples n = 10 workers.

Sample mean $\bar{X} = \$19,887$

Standard error of sample mean, $SE_{\bar{X}} = \sigma/\sqrt{n} = \$3,976$ for n=10

Confidence coefficient for 95% confidence interval = 1.96

Lower limit = (point estimate) – (confidence coefficient)(SE) = $\$19,887 - (1.96)(\$3,976) = \$12,094$

Upper limit = (point estimate) + (confidence coefficient)(SE) = $\$19,887 + (1.96)(\$3,976) = \$27,680$

Width = $(\$27,680 - \$12,094) = \$15,586$

Example: Ed samples n = 100 workers.

Sample mean $\bar{X} = \$19,813$

Standard error of sample mean, $SE_{\bar{X}} = \sigma/\sqrt{n} = \$1,257$ for n=100

Confidence coefficient for 95% confidence interval = 1.96

Lower limit = (point estimate) – (confidence coefficient)(SE) = $\$19,813 - (1.96)(\$1,257) = \$17,349$

Upper limit = (point estimate) + (confidence coefficient)(SE) = $\$19,813 + (1.96)(\$1,257) = \$22,277$

Width = $(\$22,277 - \$17,349) = \$4,928$

	n	Estimate	95% Confidence Interval	
Carol	10	\$19,887	(\$12,094, \$27,680)	Wide
Ed	100	\$19,813	(\$17,349, \$22,277)	Narrow
Truth	5000	\$19,987	\$19,987	No safety net

Definition 95% Confidence Interval

If all possible random samples (an infinite number) of a given sample size (e.g. 10 or 100) were obtained and if each were used to obtain its own confidence interval,

Then 95% of all such confidence intervals would contain the unknown; the remaining 5% would not.

But Carol and Ed Each Have Only ONE Interval:

*So now what?! The definition above doesn't seem to help us. What **can** we say?*

Carol says: “With 95% confidence, the interval \$12,094 to \$27,680 contains the unknown true mean μ .

Ed says: “With 95% confidence, the interval \$17,349 to \$22,277 contains the unknown true mean μ .

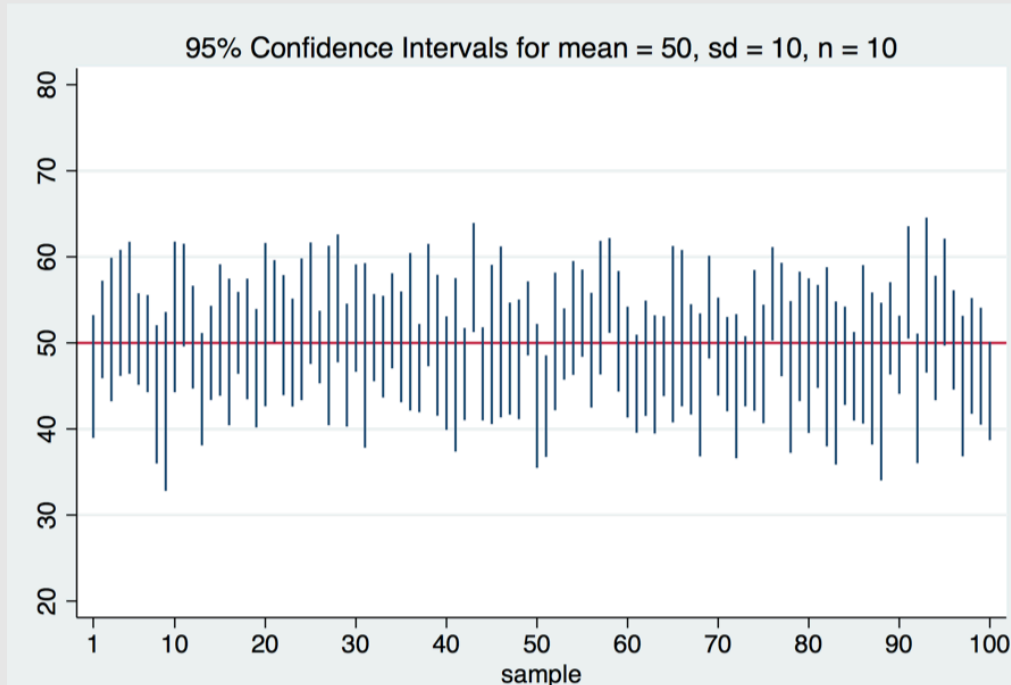
Caution on the use of Confidence Intervals:

1) It is **incorrect** to say – “*The probability that a given 95% confidence interval contains μ is 95%*”

A given interval either contains μ or it does not.

2) The **confidence coefficient** (recall – this is the multiplier we attach to the SE) for a 95% confidence interval is the number needed to ensure 95% coverage in the long run (in probability).

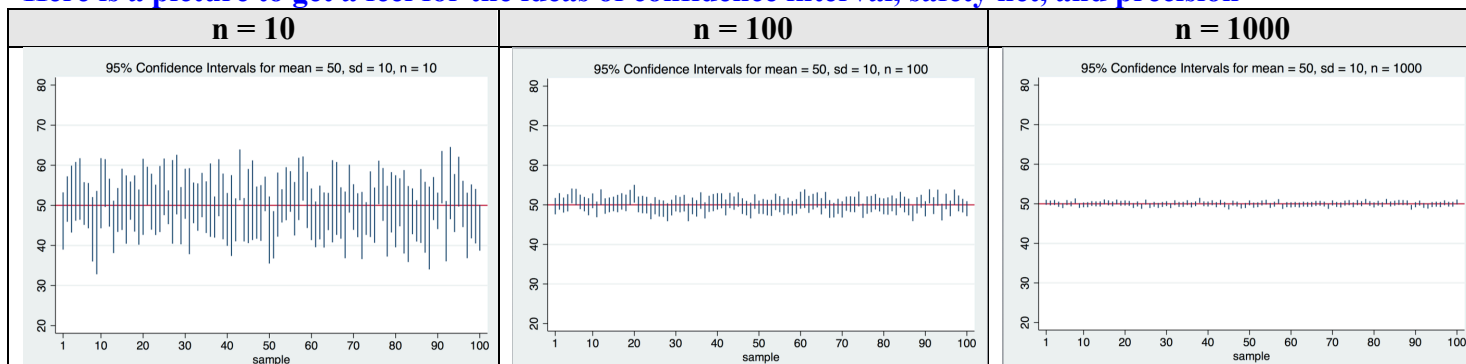
Here is a picture of a lot of confidence intervals, each based on a sample of size $n=10$



Notice ...

- (1) Any one confidence interval either contains μ or it does not. This illustrates that it is incorrect to say “There is a 95% probability that the confidence interval contains μ ”
- (2) For a given sample size (here, $n=10$), the width of all the confidence intervals is the same.

Here is a picture to get a feel for the ideas of confidence interval, safety net, and precision



Now you can also see ...

- (3) As the sample size increases, the confidence intervals are more narrow (*more precise*)
- (4) As $n \rightarrow \text{infinity}$, μ is in the interval every time.

Nature ——— Population/ Sample ——— Observation/ Data ——— Relationships/ Modeling ——— Analysis/ Synthesis

Some additional remarks on the interpretation of a confidence interval might be helpful

- Each sample gives rise to its own point estimate and confidence interval estimate built around the point estimate. The idea is to construct our intervals so that:

“IF all possible samples of a given sample size (an infinite #!) were drawn from the underlying distribution and each sample gave rise to its own interval estimate,

THEN 95% of all such confidence intervals would include the unknown μ while 5% would not”

- Another Illustration of - It is NOT CORRECT to say: “The probability that the interval (1.3, 9.5) contains μ is 0.95”.** Why? Because either μ is in (1.3, 9.5) or it is not. For example, if $\mu=5.3$ then μ is in (1.3, 9.5) with probability = 1. If $\mu=1.0$ then μ is in (1.3, 9.5) with probability=0.
- I toss a fair coin, but don’t look at the result. The probability of heads is 1/2. I am “50% confident” that the result of the toss is heads. In other words, I will guess “heads” with 50% confidence. Either the coin shows heads or it shows tails. I am either right or wrong on this particular toss. In the long run, if I were to do this, I should be right about 50% of the time – hence “50% confidence”. But for this particular toss, I’m either right or wrong.
- In most experiments or research studies we can’t look to see if we are right or wrong – but we define a confidence interval in a way that we know “in the long run” 95% of such intervals will get it right.

4. Preliminaries: Some Useful Probability Distributions

4a. Introduction to the Student t-Distribution

Looking ahead

Percentiles of the student t-distribution are used in confidence intervals for means when the population variance is NOT known.

There are a variety of definitions of a student t random variable. A particularly useful one for us here is the following. It appeals to our understanding of the z-score.

A Definition of a Student's t Random Variable

Consider a simple random sample $X_1 \dots X_n$ from a $\text{Normal}(\mu, \sigma^2)$ distribution. Calculate \bar{X} and S^2 in the usual way:

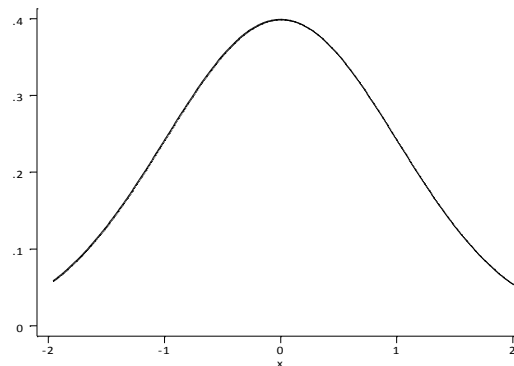
$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} \quad \text{and} \quad S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$

A student's t distributed random variable results if we construct a t-score instead of a z-score.

$$t\text{-score} = t_{DF=n-1} = \frac{\bar{X} - \mu}{s / \sqrt{n}} \text{ is distributed Student's t with degrees of freedom } = (n-1)$$

Note – The abbreviation “df” is often used to refer to “degrees of freedom”

The features of the Student's t-Distribution are similar, but not identical, to those of a Normal Distribution

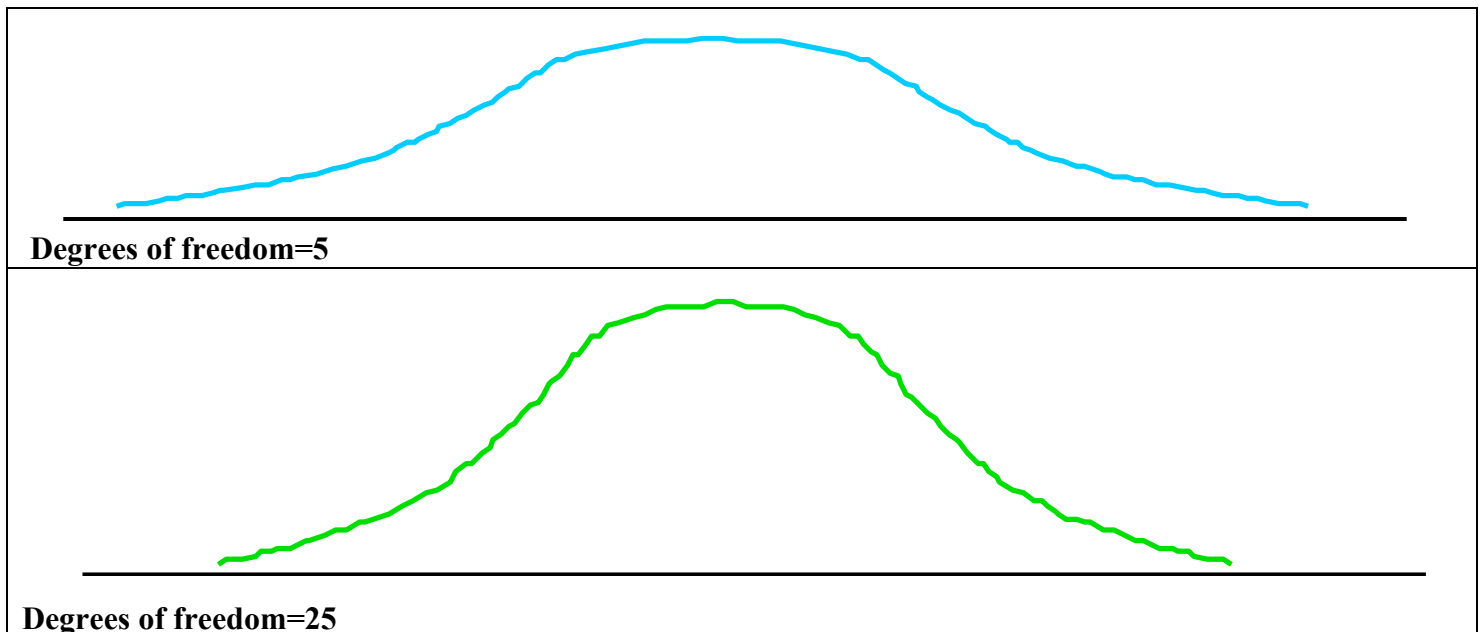
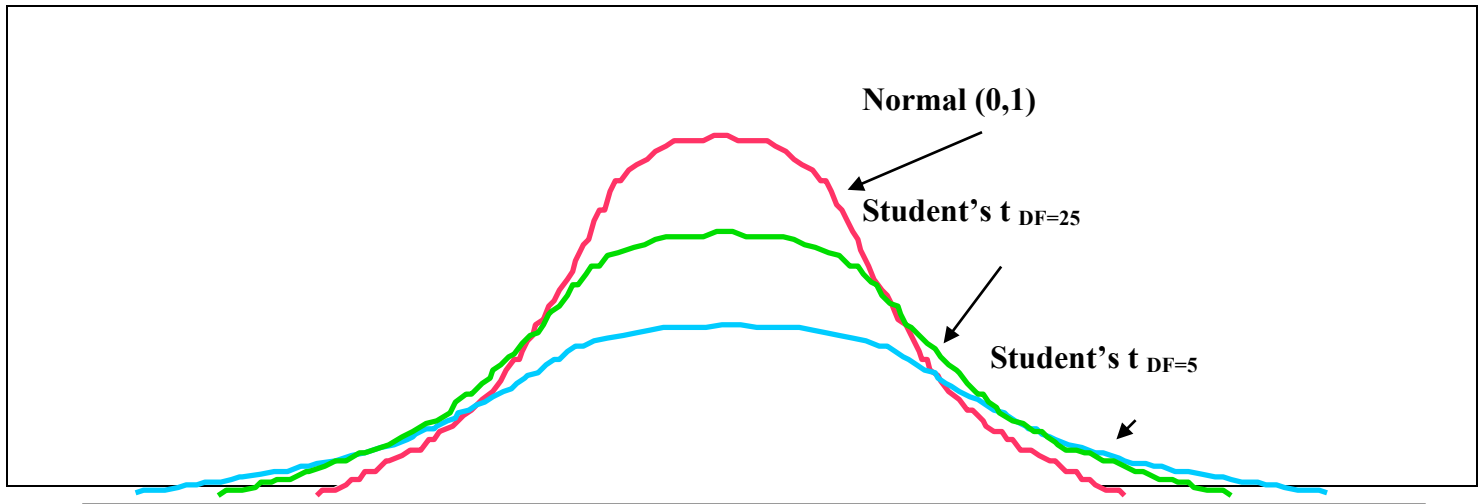


- **Bell Shaped**
- **Symmetric about zero**
- **Flatter than the Normal (0,1). This means**
 - (i) **The variability of a student t variable greater than that of a standard normal (0,1)**
 - (ii) **Thus, there is more area under the tails and less at center**
 - (iii) **Because variability is greater, resulting confidence intervals will be wider.**

The relative greater variability of a Student's t- distribution (compared to a Normal) makes sense.

We have added uncertainty in our confidence interval because we are using an estimate of the standard error rather than the actual value of the standard error.

Each degree of freedom (df) defines a separate student's t-distribution. As the degrees of freedom gets larger, the student's t-distribution looks more and more like the standard normal distribution with mean=0 and variance=1.



How to Use the Student t Distribution Calculator Provided by SurfStat

Source: <http://surfstat.anu.edu.au/surfstat-home/tables/t.php>

- From the pictures, choose between: *left tail*, *right tail*, *between*, or *two tailed*
- In the box **d.f.**, enter degrees of freedom
- To obtain a probability, enter your t-statistics in the box **t value**, enter the value
- To obtain a percentile, enter your cumulative probability in the box **probability**

Example – Solution for a probability: Probability [Student $t_{DF=1} < 3.078$] = .90

SurfStat t-distribution calculator

d.f. t value probability

SurfStat t-distribution calculator

d.f. t value probability

<http://surfstat.anu.edu.au/surfstat-home/tables/t.php>

Example – Solution for a percentile value: The 97.5th Percentile of a Student $t_{DF=9} = .90$

SurfStat t-distribution calculator

d.f. t value probability

SurfStat t-distribution calculator

d.f. t value probability

<http://surfstat.anu.edu.au/surfstat-home/tables/t.php>

4b. Introduction to the Chi Square Distribution

Looking ahead

Percentiles of the chi square distribution are used in confidence intervals for a single population variance or single population standard deviation.

Suppose we have a simple random sample from a Normal distribution. We want to calculate a confidence interval estimate of the normal distribution variance parameter, σ^2 . To do this, we work with a new random variable Y that is defined as follows:

$$Y = \frac{(n-1)S^2}{\sigma^2},$$

In this formula, S^2 is the sample variance. Under simple random sampling from a $\text{Normal}(\mu, \sigma^2)$

$$Y = \frac{(n-1)S^2}{\sigma^2} \text{ is distributed Chi Square with degrees of freedom } = (n-1)$$

Mathematical Definition Chi Square Distribution

The above can be stated more formally.

- (1) **If** the random variable X follows a normal probability distribution with mean μ and variance σ^2 ,

Then the random variable V defined:

$$V = \frac{(X - \mu)^2}{\sigma^2} \text{ is distributed chi square distribution with degree of freedom } = 1.$$

- (2) **If** each of the random variables V_1, \dots, V_k is distributed chi square with degree of freedom = 1, **and if** these are independent,

Then their sum, defined:

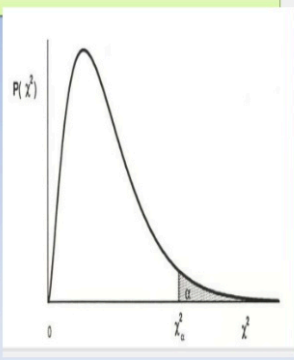
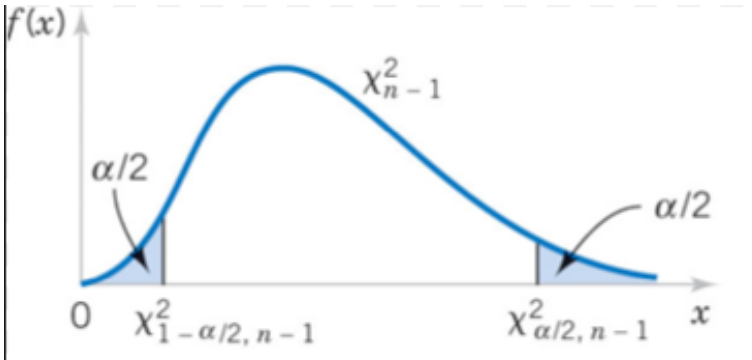
$$V_1 + \dots + V_k \text{ is distributed chi square distribution with degrees of freedom } = k$$

NOTE: For this course, it is not necessary to know the probability density function for the chi square distribution.

Features of the Chi Square Distribution:

- (1) When data are a random sample of independent observations from a normal probability distribution and interest is in the behavior of the random variable defined as the sample variance S^2 , the assumptions of the chi square probability distribution hold.
- (2) The first mathematical definition of the chi square distribution says that it is defined as the square of a standard normal random variable.
- (3) **A chi square random variable cannot be negative.** Because the chi square distribution is obtained by the squaring of a random variable, this means that a chi square random variable can assume **only non-negative** values. That is, the probability density function has domain $[0, \infty)$ and is not defined for outcome values less than zero. **Thus, the chi square distribution is NOT symmetric.** Here is a picture.

Two Pictures of the Chi Square Distribution:

<div data-bbox="121 997 730 1459"> <h4>The Chi-Square Distribution</h4> <ul style="list-style-type: none"> No negative values Mean is equal to the degrees of freedom The standard deviation increases as degrees of freedom increase, so the chi-square curve spreads out more as the degrees of freedom increase. As the degrees of freedom become very large, the shape becomes more like the normal distribution.  </div>	
<ul style="list-style-type: none"> Often, online calculators for the chi square distributions are “right tail” only. Tip! 1 = “left tail area” + “right tail area” 	<ul style="list-style-type: none"> Because this distribution is NOT symmetric about 0, Remember - You will need to solve for 2 percentile values when using the Chi square distribution in confidence intervals

Source: www.slideshare.net

Source: cmaps.cmapppers.net

Features of the Chi Square Distribution - continued:

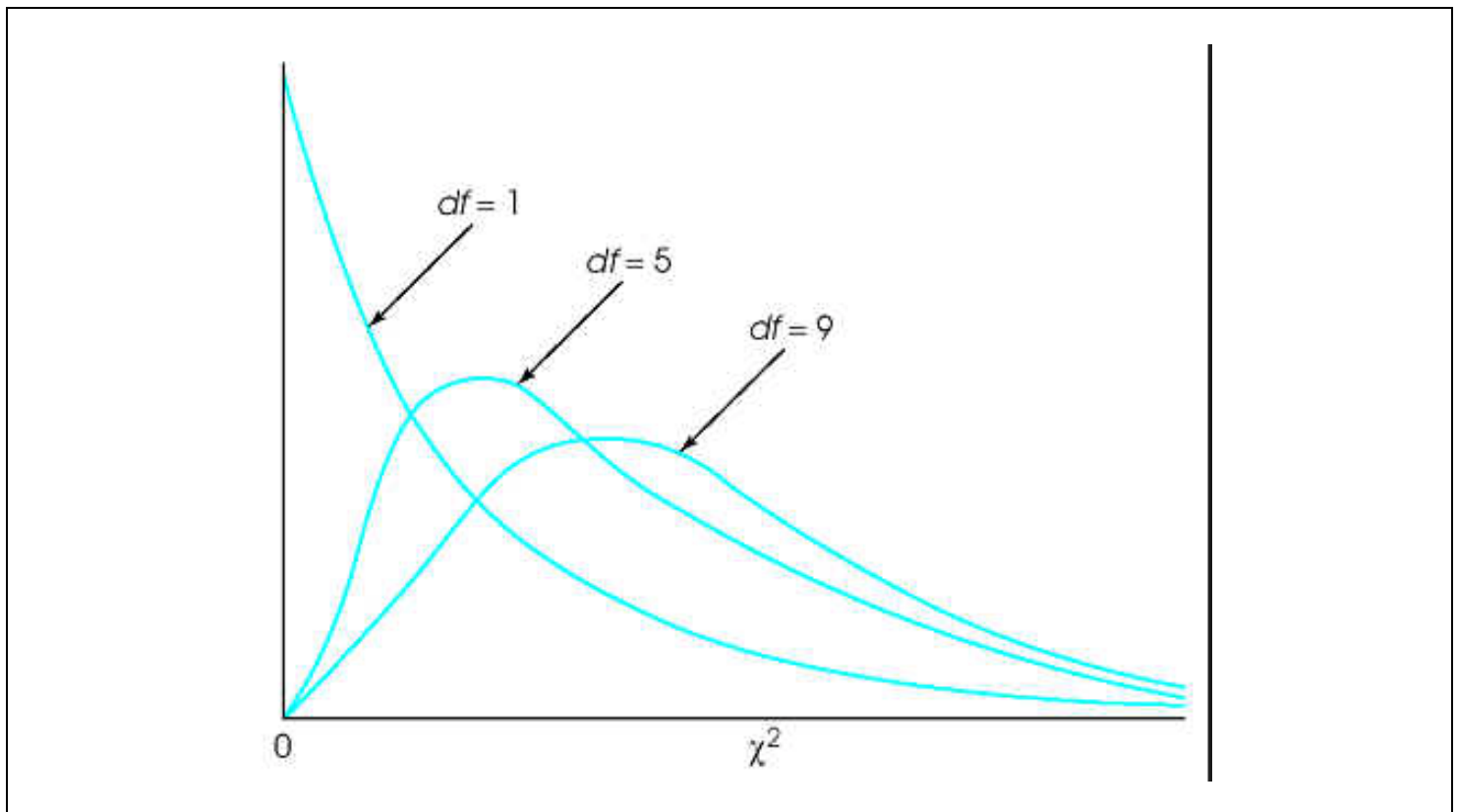
- (4) The fact that the chi square distribution is NOT symmetric about zero means that for $Y=y$ where $y>0$:

$$\Pr[Y > y] \text{ is NOT EQUAL to } \Pr[Y < -y]$$

However, because the total area under a probability distribution is 1, it is still true that

$$1 = \Pr[Y < y] + \Pr[Y > y]$$

- (5) The chi square distribution is less skewed as the number of degrees of freedom increases. See below.



Source: web.mnstate.edu

- (6) Like the degrees of freedom for the Student's t-Distribution, the degrees of freedom associated with a chi square distribution is an index of the extent of independent information available for estimating population parameter values. Thus, the chi square distributions with small associated degrees of freedom are relatively flat to reflect the imprecision of estimates based on small sample sizes. Similarly, chi square distributions with relatively large degrees of freedom are more concentrated near their expected value.

Nature — Population/ Sample — Observation/ Data — Relationships/ Modeling — Analysis/ Synthesis

How to Use the Chi Square Distribution Calculator Provided by SurfStat

Source: <http://surfstat.anu.edu.au/surfstat-home/tables/chi.php>

- From the pictures, choose between: *left tail*, *right tail*, *between*, or *two tailed*
- In the box **d.f.**, enter degrees of freedom
- To obtain a probability, enter your t-statistics in the box **t value**, enter the value
- To obtain a percentile, enter your cumulative probability in the box **probability**

Example – Solution for a probability: Probability [Chi-Square $t_{DF=4} > 6.2$] = .1847

SurfStat chi-squared calculator

d.f.
4

c² value
6.2

probability

SurfStat chi-squared calculator

d.f.
4

c² value
6.2

probability
0.1847

<http://surfstat.anu.edu.au/surfstat-home/tables/chi.php>

Example – Solution for a percentile value: The 97.5th Percentile of a Chi-Square_{DF=9} = 19.02

SurfStat chi-squared calculator

d.f.
9

c² value

probability
.975

SurfStat chi-squared calculator

d.f.
9

c² value
19.02

probability
.975

<http://surfstat.anu.edu.au/surfstat-home/tables/chi.php>

4c. Introduction to the F Distribution

Looking ahead

Percentiles of the F distribution are used in confidence intervals for the ratio of two independent variances.

Suppose we are Interested in Comparing Two Independent Variances

- Unlike the approach used to compare two means in the continuous variable setting (where we will look at their difference), the comparison of two variances is accomplished by looking at their ratio. Ratio values close to one are evidence of similarity.
- Of interest will be a confidence interval estimate of the ratio of two variances in the setting where data are comprised of two independent samples of data, each from a separate Normal distribution.

Examples -

- I have a new measurement procedure. Are the results more variable than those obtained using the standard procedure?
- I am doing a preliminary analysis to determine whether or not it is appropriate to compute a pooled variance estimate or not, when the goal is comparing the mean levels of two groups.

When comparing two independent variances, we will use a RATIO rather than a difference.

- Specifically, we will look at the ratios of variances of the form: s_x^2/s_y^2
- If the value of the ratio is close to 1, this suggests that the population variances are similar. If the value of the ratio is very different from 1, this suggests that the population variances are not the same.
- We use percentiles from the F distribution to construct a confidence interval for σ_x^2/σ_y^2

A Definition of the F-Distribution

Suppose X_1, \dots, X_{n_x} are independent and a simple random sample from a normal distribution with mean μ_X and variance σ_X^2 . Suppose further that Y_1, \dots, Y_{n_y} are independent and a simple random sample from a normal distribution with mean μ_Y and variance σ_Y^2 .

If the two sample variances are calculated in the usual way

$$S_X^2 = \frac{\sum_{i=1}^{n_x} (X_i - \bar{X})^2}{n_x - 1} \quad \text{and} \quad S_Y^2 = \frac{\sum_{i=1}^{n_y} (Y_i - \bar{Y})^2}{n_y - 1}$$

Then

$$F_{n_x-1, n_y-1} = \frac{S_X^2 / \sigma_X^2}{S_Y^2 / \sigma_Y^2} \quad \text{is distributed F with two degree of freedom specifications}$$

Numerator degrees of freedom = $n_x - 1$

Denominator degrees of freedom = $n_y - 1$

For the advanced reader

This can be skipped if you are using an online calculator

There is a relationship between the values of percentiles for pairs of F Distributions that is defined as follows:

$$F_{d_1, d_2; \alpha/2} = \frac{1}{F_{d_2, d_1; (1-\alpha)/2}}$$

Notice that (1) the degrees of freedom are in opposite order, and (2) the solution for a left tail percentile is expressed in terms of a right tail percentile.

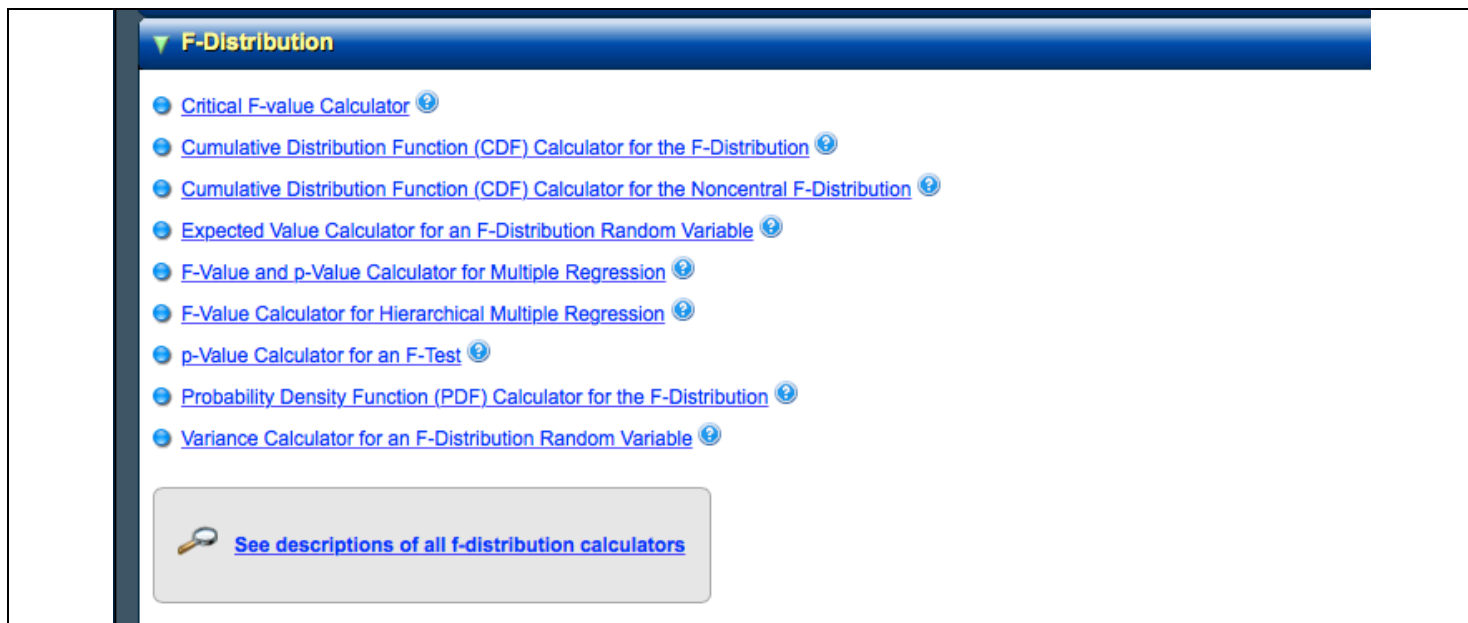
This is useful when the published table does not list the required percentile value; usually the missing percentiles are the ones in the left tail.

How to Use the F Distribution Calculator Provided by Danielsoper.com

Tip – This is a **right tail** calculator ONLY!!

Source: <http://www.danielsoper.com/statcalc3/default.aspx>

You will need to scroll down to get to the F distribution calculator. The drop down menu gives you choices:



Example – Solution for a 2.5th and 97.5th percentile value: The 2.5th and 97.5th Percentiles of an F-distribution with numerator df=4 and denominator df=23 are 0.12 and 3.41, respectively.

For the 2.5 th percentile, right tail area = .975	For the 97.5 th percentile, right tail area = .025
<p>Critical F-value Calculator</p> <p>This calculator will tell you the critical value of the F-distribution, given the probability level, denominator degrees of freedom.</p> <p>Please supply the necessary parameter values, and then click 'Calculate'.</p> <p>Degrees of freedom 1: <input type="text" value="4"/></p> <p>Degrees of freedom 2: <input type="text" value="23"/></p> <p>Probability level: <input type="text" value=".975"/></p> <p><input type="button" value="Calculate!"/></p> <p>Critical F-value: 0.11734906</p>	<p>Critical F-value Calculator</p> <p>This calculator will tell you the critical value of the F-distribution, given the probability level, denominator degrees of freedom.</p> <p>Please supply the necessary parameter values, and then click 'Calculate'.</p> <p>Degrees of freedom 1: <input type="text" value="4"/></p> <p>Degrees of freedom 2: <input type="text" value="23"/></p> <p>Probability level: <input type="text" value=".025"/></p> <p><input type="button" value="Calculate!"/></p> <p>Critical F-value: 3.40826783</p>

<http://www.danielsoper.com/statcalc3/default.aspx>

4d. Sums and Differences of Independent Normal Random Variables

This is review. See again course notes, 5. The Normal Distribution, pp 23-24

Looking ahead

We will be calculating confidence intervals of such things as the difference between two independent means (eg control versus intervention in a randomized controlled trial)

Suppose we have two independent random samples, from two independent normal distributions. eg – randomized controlled trial of placebo versus treatment groups). We suppose we want to compute a confidence interval estimates of the difference of the means.

Point Estimator: How do we obtain a point estimate of the difference $[\mu_{\text{Group 1}} - \mu_{\text{Group 2}}]$?

- A good point estimator of the difference between population means is the difference between sample means, $[\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}}]$

Standard Error of the Point Estimator: We need the standard error of $[\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}}]$

Definitions

IF

- (for group 1): $X_{11}, X_{12}, \dots, X_{1n_1}$ is a simple random sample from a Normal (μ_1, σ_1^2)
- (for group 2): $X_{21}, X_{22}, \dots, X_{2n_2}$ is a simple random sample from a Normal (μ_2, σ_2^2)
- This is great!** We already know the sampling distribution of each sample mean
 $\bar{X}_{\text{Group 1}}$ is distributed Normal $(\mu_1, \sigma_1^2 / n_1)$
 $\bar{X}_{\text{Group 2}}$ is distributed Normal $(\mu_2, \sigma_2^2 / n_2)$

THEN

$[\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}}]$ is also distributed Normal with

$$\text{Mean} = [\mu_{\text{Group 1}} - \mu_{\text{Group 2}}]$$

$$\text{Variance} = \left[\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \right]$$

Be careful!! The standard error of the difference is NOT the sum of the two separate standard errors.

Notice – You must first sum the variance and then take the square root of the sum.

$$SE[\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}}] = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

**A General Result
Handy!**

If random variables X and Y are **independent** with

$$E[X] = \mu_X \text{ and } \text{Var}[X] = \sigma_X^2$$

$$E[Y] = \mu_Y \text{ and } \text{Var}[Y] = \sigma_Y^2$$

Then

$$E[aX + bY] = a\mu_X + b\mu_Y$$

$$\text{Var}[aX + bY] = a^2 \sigma_X^2 + b^2 \sigma_Y^2 \text{ and}$$

$$\text{Var}[aX - bY] = a^2 \sigma_X^2 + b^2 \sigma_Y^2$$

Tip on variances: This result ALSO says that, when X and Y are independent, the variance of their difference is equal to the variance of their sum. This makes sense if it is recalled that variance is defined using squared deviations which are always positive.

5. Normal Distribution: One Group

5a. Confidence Interval for μ (σ^2 Known)

Introduction and “where we are going” ...

Hopefully, you will see that the logic and mechanics of confidence interval construction are very similar across a variety of settings.

In this unit, we consider the setting of data from a **normal** distribution (or two normal distributions) and the setting of data from a **binomial** distribution (or two binomial distributions).

We want to compute a confidence interval estimate of μ for a population distribution that is normal with σ known. Available are data from a random sample of size= n .

- These pages show you how to construct a confidence interval.
- Appendix 1 gives the [statistical theory](#) underlying this calculation

1. The Point Estimate of μ is the Sample Mean \bar{X}

Recall that, for a sample of size= n , the sample mean is calculated as

$$\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}$$

Features:

1. Under simple random sampling, the sample mean (\bar{X}) is an unbiased estimator of the population mean parameter μ , regardless of the underlying probability distribution.
2. When the underlying probability distribution is normal, the sample mean \bar{X} also satisfies the criterion of being minimum variance unbiased (See page 5).

2. The Standard Error of \bar{X}_n is σ/\sqrt{n}

The precision of \bar{X}_n as an estimate of the unknown population mean parameter μ is reflected in its standard error. Recall:

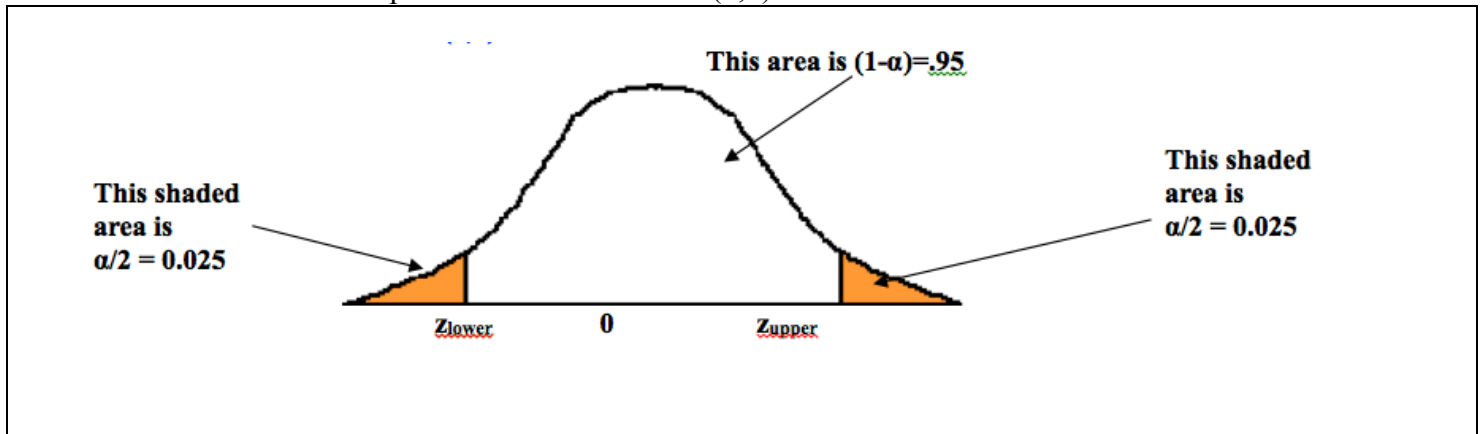
$$SE(\bar{X}_n) = \sqrt{\text{variance}(\bar{X}_n)} = \frac{\sigma}{\sqrt{n}}$$

- ♣ SE is smaller for smaller σ (measurement error)
- ♣ SE is smaller for larger n (study design)

3. The Confidence Coefficient

The **confidence coefficient** for a 95% confidence interval is the number needed to insure 95% coverage “in the long run” (in probability). See again, page 18.

- ♣ 95% coverage **leaves 5% in the tails**. This is split evenly in the 2 tails. Thus, for a 95% confidence interval (5% in tails total/ 2 tails) = **2.5% in each tail**. For a 95% confidence interval, this number will be the 97.5th percentile of the Normal (0,1) distribution = 1.96.



- ♣ For a $(1-\alpha)100\%$ confidence interval, this number will be the $(1-\alpha/2)100^{\text{th}}$ percentile of the Normal (0,1) distribution.
- ♣ The table below gives some of these values in the setting of constructing a confidence interval estimate of μ when data are from a Normal distribution with σ^2 known.

Confidence Level	Confidence Coefficient = Percentile Value from Normal (0,1)	
	Percentile	
.50	75 th	0.674
.75	87.5 th	1.15
.80	90 th	1.282
.90	95 th	1.645
.95	97.5 th	1.96
.99	99.5 th	2.576
$(1-\alpha)$	$(1-\alpha/2)100^{\text{th}}$	-

Example - For a 50% CI, $.50 = (1-\alpha)$ says $\alpha=.50$ and says $(1-\alpha/2)=.75$. Thus, use 75th percentile of $N(0,1)=0.674$

Example -

The following data are the weights (micrograms) of drug inside each of 30 capsules, after subtracting the capsule weight. Suppose it is known that $\sigma^2 = 0.25$. Under the assumption of normality, calculate a 95% confidence interval estimate of μ .

0.6	0.3	0.1	0.3	0.3
0.2	0.6	1.4	0.1	0.0
0.4	0.5	0.6	0.7	0.6
0.0	0.0	0.2	1.6	-0.2
1.6	0.0	0.7	0.2	1.4
1.0	0.2	0.6	1.0	0.3

- ♣ The data are simple random sample of size $n=30$ from a Normal distribution with mean = μ and variance = σ^2 .
- ♣ The population variance is known and has value $\sigma^2 = 0.25$
- ♣ **Remark – In real life, we will rarely know σ^2 !** This example is for illustration only.

The solution for the confidence interval is point estimate \pm safety net:

$$\begin{aligned}\text{Lower limit} &= (\text{point estimate}) - (\text{multiple}) (\text{SE of point estimate}) \\ \text{Upper limit} &= (\text{point estimate}) + (\text{multiple}) (\text{SE of point estimate})\end{aligned}$$

Point Estimate of μ is the Sample Mean $\bar{X}_{n=30}$

$$\bar{X}_{n=30} = \frac{\sum_{i=1}^n X_i}{n=30} = 0.51$$

The Standard Error of \bar{X}_n is σ/\sqrt{n}

$$SE(\bar{X}_{n=30}) = \sqrt{\text{variance}(\bar{X}_{n=30})} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{0.25}}{\sqrt{30}} = 0.0913$$

The Confidence Coefficient

For a 95% confidence interval, this number will be the 97.5th percentile of the Normal (0,1) distribution. From the table on page 34 (or a Normal(0,1) calculator on the web), obtain the value 1.96.

Desired Confidence Level	Value of Confidence Coefficient
.95	1.96

For a 95% CI, .95 = (1- α) says $\alpha=.05$ and says (1- $\alpha/2$)=.975. Thus, use 97.5th percentile of N(0,1)=1.96

Putting this all together –

$$\begin{aligned}\text{Lower limit} &= (\text{point estimate}) - (\text{multiple}) (\text{SE of point estimate}) \\ &= 0.51 - (1.96) (0.0913) \\ &= 0.33\end{aligned}$$

$$\begin{aligned}\text{Upper limit} &= (\text{point estimate}) + (\text{multiple}) (\text{SE of point estimate}) \\ &= 0.51 + (1.96) (0.0913) \\ &= 0.69\end{aligned}$$

Thus, we have the following general formula for a (1 - α)100% confidence interval -

$$\bar{X}_n \pm [(1-\alpha/2)100^{\text{th}} \text{ percentile of Normal}(0,1)] \text{SE}(\bar{X}_n)$$

How to Calculate the Proportion of Sample Means in a Given Interval (Use the idea of standardization)

We learned how to do this in Unit 5. *The Normal Distribution*.

Example

A sample of size $n=100$ from a normal distribution with unknown mean yields a sample mean $\bar{X}_{n=100} = 267.43$. The population variance of the normal distribution is known to be equal to $\sigma^2 = 36,764.23$. What proportion of means of size $n=100$ will lie in the interval $[200,300]$ if it is known that $\mu = 250$

Answer = 99.1%

Solution:

Step 1 -

The random variable that we need to “standardize” is $\bar{X}_{n=100}$.

$$\clubsuit \text{ Mean} = 250$$

$$\clubsuit \text{ SE} = \sigma / \sqrt{100} = \sqrt{36,764.23} / \sqrt{100} = 19.174$$

Step 2 -

Probability $[200 < \bar{X}_{n=100} < 300]$ by the standardization formula is

$$= \Pr \left[\frac{200-250}{19.174} < \frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}} < \frac{300-250}{19.174} \right] = \Pr [-2.608 < Z\text{-score} < +2.608]$$

$$= 0.9910.$$

5b. Confidence Interval for μ (σ^2 NOT known)

- In section 5a, we assumed that σ^2 is known and obtained a confidence interval for μ of the form

$$\text{lower limit} = \bar{X} - z_{(1-\alpha/2)100} \left(\sigma / \sqrt{n} \right)$$

$$\text{upper limit} = \bar{X} + z_{(1-\alpha/2)100} \left(\sigma / \sqrt{n} \right)$$

- The required confidence coefficient ($z_{1-\alpha/2}$) was obtained as a percentile from the standard normal, $N(0,1)$, distribution. (e.g. for a 95% CI, we used the 97.5th percentile)
- **More realistically, however, σ^2 will not be known. Now what?** Reasonably, we might replace σ with “s”. Recall that s is the sample standard deviation and we get it as follows:

$$s = \sqrt{s^2} \text{ where } s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

- So far so good. But there is a problem.

Whereas $\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$ IS distributed Normal (0,1)

$\frac{\bar{X} - \mu}{s / \sqrt{n}}$ is **NOT** distributed Normal (0,1).

- We have to modify our “machinery” (specifically the SE piece of our machinery) to accommodate the unknown-ness of σ^2 .

Whereas we previously used when σ^2 was known	With σ^2 unknown we now use
z-score	t-score
Percentile from Normal(0,1)	Percentile from Student's t

Under simple random sampling from a normal distribution, the confidence interval for an unknown mean μ , the confidence interval will be of the following form

$$\text{lower limit} = \bar{X} - t_{DF; (1-\alpha/2)100} \left(s / \sqrt{n} \right)$$

$$\text{upper limit} = \bar{X} + t_{DF; (1-\alpha/2)100} \left(s / \sqrt{n} \right)$$

When σ^2 is not known, the computation of a confidence interval for the mean μ is not altered much.

- We simply replace the confidence coefficient from the $N(0,1)$ with one from the appropriate Student's t-Distribution (the one with $df = n-1$)
- We replace the (now unknown) standard error with its estimate. The latter looks nearly identical except that it utilizes “s” in place of “ σ ”
- Recall

$$s = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{(n-1)}}$$

- Thus,

Confidence Interval for μ in two settings of a sample from a Normal Distribution	
σ^2 is KNOWN	σ^2 is NOT Known
$\bar{X} \pm (z_{1-\alpha/2})(\sigma/\sqrt{n})$	$\bar{X} \pm (t_{n-1; 1-\alpha/2})(s/\sqrt{n})$

Example

A random sample of size $n=20$ durations (minutes) of cardiac bypass surgeries has a mean duration of $\bar{X} = 267$ minutes, and variance $s^2 = 36,700$ minutes². Assuming the underlying distribution is normal with unknown variance, construct a 90% CI estimate of the unknown true mean, μ .

Answer: (193.2, 340.8) minutes

Solution:

Step 1 - Point Estimate of μ is the Sample Mean \bar{X}

$$\bar{X}_{n=20} = \frac{\sum_{i=1}^n X_i}{n=20} = 267 \text{ minutes.}$$

Step 2 – The Estimated Standard Error of \bar{X}_n is s/\sqrt{n}

$$\hat{SE}(\bar{X}_{n=20}) = \sqrt{\text{variance}(\bar{X}_{n=20})} = \frac{S}{\sqrt{n}} = \frac{\sqrt{36,700}}{\sqrt{20}} = 42.7 \text{ minutes}$$

Step 3 - The Confidence Coefficient

For a 90% confidence interval, this number will be the 95th percentile of the Student's t-Distribution that has degrees of freedom = $(n-1) = 19$. This value is 1.729.

Putting this all together –

$$\begin{aligned} \text{Lower limit} &= (\text{point estimate}) - (\text{confidence coefficient.}) (\text{SE of point estimate}) \\ &= 267 - (1.729)(42.7) \\ &= 193.17 \end{aligned}$$

$$\begin{aligned} \text{Upper limit} &= (\text{point estimate}) + (\text{confidence coefficient}) (\text{SE of point estimate}) \\ &= 267 + (1.729)(42.7) \\ &= 340.83 \end{aligned}$$

Thus, a 90% confidence interval for the true mean duration of surgery is (193.2, 340.8) minutes.

5c. Confidence Interval for σ^2

A confidence interval for σ^2 is calculated using percentiles from the chi square distribution.

- The following are some settings where our interest lies in estimation of the variance, σ^2
 - Standardization of equipment – repeated measurement of a standard should have small variability
 - Evaluation of technicians – are the results from a particular technician “too variable”
 - Comparison of measurement techniques – are the measurements obtained using a new technique too variable compared to the precision of the old technique?
- We have a point estimator of σ^2 . It is S^2 .
- How do we get a confidence interval? The answer will utilize a new standardized variable, based on the way in which S^2 is computed. It is a **chi square** random variable.

The definition of the chi square distribution gives us what we need to construct a confidence interval estimate of σ^2 when data are a simple random sample from a normal probability distribution. The approach here is similar to that for estimating the mean μ .

- The table below shows how to construct a confidence interval.
- For the interested reader, Appendix 2 is the **derivation** behind the calculation.

(1- α)100% Confidence Interval for σ^2 Setting – Normal Distribution	
Lower limit =	$\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}$
Upper limit =	$\frac{(n-1)S^2}{\chi^2_{\alpha/2}}$

Example

A precision instrument is guaranteed to read accurately to within ± 2 units. A sample of 4 readings on the same object yield 353, 351, 351, and 355. Find a 95% confidence interval estimate of the population variance σ^2 and also for the population standard deviation σ .

Answer: (1.18, 51.0) units squared

Solution:

1. Obtain the point estimate of σ^2 . It is the sample variance S^2

To get the sample variance S^2 , we will need to compute the sample mean first.

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = 352.5 \quad \text{and} \quad S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} = 3.67$$

2. Determine the correct chi square distribution to use.

It has degrees of freedom, $df = (4-1) = 3$.

3. Obtain the correct multipliers.

Because the desired confidence level is 0.95, we set $0.95 = (1-\alpha)$. Thus $\alpha = .05$

For a 95% confidence level, the percentiles we want are

- (i) $(\alpha/2)100^{\text{th}} = 2.5^{\text{th}}$ percentile
- (ii) $(1 - \alpha/2)100^{\text{th}} = 97.5^{\text{th}}$ percentile

Obtain percentiles for chi square distribution with degrees of freedom = 3

<http://www.stat.tamu.edu/~west/applets/chisqdemo.html>

- (i) $\chi_{df=3,025}^2 = 0.2158$
- (ii) $\chi_{df=3,975}^2 = 9.348$

4. Thus,

$$(i) \text{ Lower limit} = \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}} = \frac{(3)(3.67)}{9.348} = 1.178$$

$$(ii) \text{ Upper limit} = \frac{(n-1)S^2}{\chi^2_{\alpha/2}} = \frac{(3)(3.67)}{0.2158} = 51.02$$

Obtain a Confidence Interval for the Population Standard Deviation σ

Answer: (1.09, 7.14) units

Solution:

Step 1 – Obtain a confidence interval for σ^2
(1.178, 51.02)

Step 2 – The associated confidence interval for σ is obtained by taking the square root of each of the lower and upper limits

- 95% Confidence Interval = $(\sqrt{1.178}, \sqrt{51.02}) = (1.09, 7.14)$
- Point estimate = $\sqrt{3.67} = 1.92$

Remarks on the Confidence Interval for σ^2

- It is **NOT** symmetric about the point estimate; the “safety net” on each side of the point estimate is of different lengths.
- These intervals tend to be wide. Thus, large sample sizes are required to obtain reasonably narrow confidence interval estimates for the variance and standard deviation parameters.

6. Normal Distribution: Paired Data

Introduction to Paired Data

- Paired data occur when each individual (more specifically, each unit of measurement) in a sample is measured twice.
- Paired data are familiar: "pre/post", "before/after", "right/left", "parent/child", etc.
- **Examples -**
 - 1) Blood pressure prior to and following treatment,
 - 2) Number of cigarettes smoked per week measured prior to and following participation in a smoking cessation program,
 - 3) Number of sex partners in the month prior to and in the month following an HIV education campaign.
- In each of these examples that the two occasions of measurement are linked by virtue of the two measurements being made on the same individual.

We are interested in comparing the two paired outcomes.

When the paired data are **continuous**, the comparison focuses on the **difference** between the two paired measurements.

Note – We'll see later that when the data are discrete, an analysis of paired data might focus on the ratio (eg. relative risk) of the two measurements rather than on the difference.

Examples:

- 1) Blood pressure prior to and following treatment. Interest is $d = \text{pre-post}$. Large differences are evidence of blood pressure lowering associated w treatment.
- 2) Number of cigarettes smoked per week measured prior to and following participation in a smoking cessation program. Interest is $d = \text{pre-post}$. Large differences "d" are evidence of smoking reduction.
- 3) Number of sex partners in the month prior to and in the month following an HIV education campaign. Interest is $d = \text{pre} - \text{post}$. Large differences are evidence of safer sex behaviors.

Nature — Population/ Sample — Observation/ Data — Relationships/ Modeling — Analysis/ Synthesis

6a. Confidence Interval for $\mu_{\text{DIFFERENCE}}$

- Suppose two paired measurements are made of the same phenomenon (eg. blood pressure, # cigarettes/week, etc) on each individual in a sample. Call them X and Y. If X and Y are each normally distributed, then their difference is also distributed normal. (see again, section 4d)
- Thus, the setting is our focus on the difference D and the following assumptions
 - $D = (X - Y)$ is distributed Normal with
 - Mean of $D = \mu_{\text{difference}}$. Let's write this as μ_d
 - Variance of $D = \sigma_{\text{DIFFERENCE}}^2$. Let's write this as σ_d^2
- In this setting, estimation for paired data is a special case of selected methods already presented. Attention is restricted to the single random variable defined as the difference between the two measurements. The methods already presented that we can use here are

- Confidence Interval for μ_d - Normal Distribution σ_d^2 unknown
- Confidence Interval for σ_d^2 - Normal Distribution

Example

source: Anderson TW and Sclove SL. *Introductory Statistical Analysis*. Boston: Houghton Mifflin, 1974. page 339

A researcher is interested assessing the improvement in reading skills upon completion of the second grade (Y) in comparison to those prior to the second grade (X). For each child, his or her improvement is measured using the difference “d” which is defined $d = Y - X$. A sample of $n=30$ children are studied. The data are shown on the next page.

ID	PRE(X)	POST(Y)	d=(Y-X)
1	1.1	1.7	0.6
2	1.5	1.7	0.2
3	1.5	1.9	0.4
4	2.0	2.0	0.0
5	1.9	3.5	1.6
6	1.4	2.4	1.0
7	1.5	1.8	0.3
8	1.4	2.0	0.6
9	1.8	2.3	0.5
10	1.7	1.7	0.0
11	1.2	1.2	0.0
12	1.5	1.7	0.2
13	1.6	1.7	0.1
14	1.7	3.1	1.4
15	1.2	1.8	0.6
16	1.5	1.7	0.2
17	1.0	1.7	0.7
18	2.3	2.9	0.6
19	1.3	1.6	0.3
20	1.5	1.6	0.1
21	1.8	2.5	0.7
22	1.4	3.0	1.6
23	1.6	1.8	0.2
24	1.6	2.6	1.0
25	1.1	1.4	0.3
26	1.4	1.4	0.0
27	1.4	2.0	0.6
28	1.5	1.3	-0.2
29	1.7	3.1	1.4
30	1.6	1.9	0.3

Calculate

- (1) A 99% confidence interval for μ_d
- (2) An 80% confidence Interval for σ_d^2

Solution for a 99% Confidence Interval for μ_d

Step 1 – Point Estimate of μ_d is the Sample Mean $\bar{d}_{n=30}$

$$\bar{d}_{n=30} = \frac{\sum_{i=1}^n d_i}{n=30} = 0.51$$

Step 2 – The Estimated Standard Error of \bar{d}_n is S_d / \sqrt{n}

Calculate the sample variance of the individual differences:

$$S_d^2 = \frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n-1} = 0.2416$$

The estimated variance of the sample mean of the differences is therefore:

$$\text{variance}(\bar{d}_{n=30}) = \frac{S_d^2}{n} = \frac{0.2416}{30}$$

Thus,

$$SE(\bar{d}_{n=30}) = \sqrt{\text{variance}(\bar{d}_{n=30})} = \frac{S_d}{\sqrt{n}} = \frac{\sqrt{0.2416}}{\sqrt{30}} = 0.0897$$

Step 3 – The Confidence Coefficient

For a 99% confidence interval, this number will be the 99.5th percentile of the Student's t-Distribution that has degrees of freedom = (n-1) = 29. This value is 2.756.

Step 4 – Substitute into the formula for a confidence interval

$$\begin{aligned} \text{Lower limit} &= (\text{point estimate}) - (\text{confidence coefficient}) (\text{SE of point estimate}) \\ &= 0.51 - (2.756)(0.0897) \\ &= 0.2627 \end{aligned}$$

$$\begin{aligned} \text{Upper limit} &= (\text{point estimate}) + (\text{confidence coefficient.}) (\text{SE of point estimate}) \\ &= 0.51 + (2.756)(0.0897) \end{aligned}$$

$$= 0.7573$$

6b. Confidence Interval for σ^2 DIFFERENCE

Solution for an 80% Confidence Interval for σ_d^2 .

Step 1 - Obtain the point estimate of σ_d^2 .

$$S_d^2 = \frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n-1} = 0.2416$$

Step 2 - Determine the correct chi square distribution to use.

$$\text{It has } df = (30-1) = 29.$$

Step 3 - Obtain the correct multipliers.

Because the desired confidence level is 0.80, set $0.80 = (1-\alpha)$. Thus $\alpha = .20$

For a 80% confidence level, $\alpha = .20$ and $\alpha/2 = .10$ so we want:

- (i) $(\alpha/2)100^{\text{th}} = 10^{\text{th}}$ percentile
- (ii) $(1 - \alpha/2)100^{\text{th}} = 90^{\text{th}}$ percentile

Using a chi square distribution calculator, set degrees of freedom, $df = 29$

$$(i) \chi_{df=29, .10}^2 = 19.77$$

$$(ii) \chi_{df=29, .90}^2 = 39.09$$

Step 4 – Substitute into the formula for the confidence interval

$$(i) \text{ Lower limit} = \frac{(n-1)S_d^2}{\chi_{1-\alpha/2}^2} = \frac{(29)(0.2416)}{39.09} = 0.1792$$

$$(ii) \text{ Upper limit} = \frac{(n-1)S_d^2}{\chi_{\alpha/2}^2} = \frac{(29)(0.2416)}{19.77} = 0.3544$$

7. Normal Distribution: Two Independent Groups

Illustration of the Setting of Two Independent Groups

Example - A researcher performs a drug trial involving two independent groups.

- A **control** group is treated with a placebo while, separately;
- An independent **intervention** group is treated with an active agent.
- Interest is in a comparison of the mean control response with the mean intervention response under the assumption that the responses are independent.
- The tools of confidence interval construction described for paired data are **not** appropriate.

7a. Confidence Interval for $[\mu_{\text{GROUP1}} - \mu_{\text{GROUP2}}]$

Suppose we want to compare the mean response in one group with the mean response in a separate group. Suppose further that two groups are independent.

Examples -

- 1) Is mean blood pressure the same for males and females?
- 2) Is body mass index (BMI) similar for breast cancer cases versus non-cancer patients?
- 3) Is length of stay (LOS) for patients in hospital “A” the same as that for similar patients in hospital “B”?

For continuous data, the comparison of two independent groups focuses on the difference between the means of the two groups.

- Similarity of the two groups is reflected in a difference between means that is “near” zero.
- Focus is on $[\mu_{\text{Group 1}} - \mu_{\text{Group 2}}]$

Nature — Population/ Sample — Observation/ Data — Relationships/ Modeling — Analysis/ Synthesis

Point Estimator: We want a point estimate of the difference [$\mu_{\text{Group 1}} - \mu_{\text{Group 2}}$]

- Our point estimator will be the difference between sample means, [$\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}}$]

Standard Error of the Point Estimator: We need the value (or estimate of) the standard error of [$\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}}$]. We can use what we learned in section 4d (see pp 31-32)

- Since $X_{11}, X_{12}, \dots, X_{1n_1}$ is a simple random sample from a Normal (μ_1, σ_1^2)
- And since $X_{21}, X_{22}, \dots, X_{2n_2}$ is a simple random sample from a Normal (μ_2, σ_2^2)

- **We have**

$\bar{X}_{\text{Group 1}}$ is distributed Normal ($\mu_1, \sigma_1^2 / n_1$)

$\bar{X}_{\text{Group 2}}$ is distributed Normal ($\mu_2, \sigma_2^2 / n_2$)

- **And thus,**

[$\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}}$] is also distributed Normal with

Mean = [$\mu_{\text{Group 1}} - \mu_{\text{Group 2}}$]

Variance = $\left[\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \right]$

How to Estimate the Standard Error

The correct solution depends on σ_1^2 and σ_2^2 .

Solution 1 - σ_1^2 and σ_2^2 are both known

$$SE[\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}}] = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Solution 2 - σ_1^2 and σ_2^2 are both NOT known but are assumed EQUAL

$$SE[\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}}] = \sqrt{\frac{S_{\text{pool}}^2}{n_1} + \frac{S_{\text{pool}}^2}{n_2}}$$

S_{pool}^2 is a weighted average of the two separate sample variances, with weights equal to the associated degrees of freedom contributions.

$$S_{\text{pool}}^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}$$

Solution 3 - σ_1^2 and σ_2^2 are both NOT known and NOT EQUAL

$$SE[\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}}] = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

Confidence Coefficient (“Multiplier”) –

Again, the correct solution depends on σ_1^2 and σ_2^2 .

Solution 1 - σ_1^2 and σ_2^2 are both known

Use percentile of Normal(0,1)

Solution 2 - σ_1^2 and σ_2^2 are both NOT known but are assumed EQUAL

Use percentile of Student’s t
Degrees of freedom = $(n_1 - 1) + (n_2 - 1)$

Solution 3 - σ_1^2 and σ_2^2 are both NOT known and NOT EQUAL

Use percentile of Student’s t
Degrees of freedom = f where “f” is given by formula (Satterthwaite)

$$f = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right)^2}{\left(\frac{\left[\frac{S_1^2}{n_1} \right]^2}{n_1 - 1} + \frac{\left[\frac{S_2^2}{n_2} \right]^2}{n_2 - 1} \right)}$$

Horrible, isn’t it!

Summary ...

Normal Distribution: Confidence Interval for [$\mu_1 - \mu_2$] (Two Independent Groups) CI = [point estimate] \pm (conf.coeff)SE[point estimate]			
Scenario →	σ_1^2 and σ_2^2 are both known	σ_1^2 and σ_2^2 are both NOT known but are assumed EQUAL	σ_1^2 and σ_2^2 are both NOT known and NOT Equal
Estimate	$\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}}$	$\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}}$	$\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}}$
SE to use	$SE[\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}}] = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	$SE[\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}}] = \sqrt{\frac{S_{\text{pool}}^2}{n_1} + \frac{S_{\text{pool}}^2}{n_2}}$ where you already have obtained: $S_{\text{pool}}^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}$	$SE[\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}}] = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$
Confidence Coefficient Use Percentiles from	Normal	Student's t	Student's t
Degrees freedom	Not applicable	$(n_1 - 1) + (n_2 - 1)$	$f = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right)^2}{\left(\frac{\left[\frac{S_1^2}{n_1} \right]^2}{n_1 - 1} + \frac{\left[\frac{S_2^2}{n_2} \right]^2}{n_2 - 1} \right)}$

Example

Data are available on the weight gain of weanling rats fed either of two diets. The weight gain in grams was recorded for each rat, and the mean for each group computed:

Diet #1 Group **$n_1 = 12$ rats** **$\bar{X}_1 = 120$ grams****Diet #2 Group** **$n_2 = 7$ rats** **$\bar{X}_2 = 101$ grams**

On the basis of a 99% confidence interval, is there a difference in mean weight gain among rats fed on the 2 diets? For illustration purposes, we'll consider all three scenarios, depending on σ_1^2 and σ_2^2 .

Solution 1

σ_1^2 and σ_2^2 are both known = 400 grams²

Step 1 – Point Estimate of $[\mu_1 - \mu_2]$

$$\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}} = 19g$$

Step 2 – Standard Error of Point Estimate

$$SE[\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}}] = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{400}{12} + \frac{400}{7}} = 9.51g$$

Step 3 – The Confidence Coefficient

Since $\sigma_1^2 = \sigma_2^2 = \sigma^2$ is known, the multiplier is a percentile from the Normal (0,1). For a 99% confidence interval, the required percentile is the 99.5th. This has value 2.575

Step 4 – Substitute into the formula for a confidence interval

$$\begin{aligned} CI &= (\text{point estimate}) \pm (\text{confidence coefficient.}) (SE \text{ of point estimate}) \\ &= 19 \pm (2.575)(9.51) \\ &= (-5.5g, 43.5g) \end{aligned}$$

Solution 2

σ_1^2 and σ_2^2 are both NOT known but are assumed EQUAL and we have
 $S_1^2 = 457.25$, $S_2^2 = 425.33$

Step 1 – Point Estimate of $[\mu_1 - \mu_2]$

$$\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}} = 19g$$

Step 2 – Estimated Standard Error of Point Estimate is in two steps

$$S_{\text{pool}}^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)} = \frac{(11)(457.25) + (6)(425.33)}{(11) + (6)} = 445.98 \text{ grams}^2$$

$$\hat{SE}[\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}}] = \sqrt{\frac{S_{\text{pool}}^2}{n_1} + \frac{S_{\text{pool}}^2}{n_2}} = \sqrt{\frac{445.98}{12} + \frac{445.98}{7}} = 10.0437g$$

Step 3 – The Confidence Coefficient

Since $\sigma_1^2 = \sigma_2^2 = \sigma^2$ but UNknown, the multiplier is a percentile from the Student's t with degrees of freedom = $(12-1) + (7-1) = 17$. For a 99% confidence interval, the required percentile is the 99.5th. This has value 2.8982.

Step 4 – Substitute into the formula for a confidence interval

$$\begin{aligned} \text{CI} &= (\text{point estimate}) \pm (\text{confidence coefficient.}) (\text{SE of point estimate}) \\ &= 19 \pm (2.8982)(10.0437) \\ &= (-10.1 \text{ g}, 48.1 \text{ g}) \text{ considerably wider than for scenario 1!} \end{aligned}$$

Solution 3

σ_1^2 and σ_2^2 are both NOT known and are UNEQUAL and we have
 $S_1^2 = 457.25$, $S_2^2 = 425.33$

Step 1 – Point Estimate of $[\mu_1 - \mu_2]$

$$\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}} = 19g$$

Step 2 – Estimated Standard Error of Point Estimate is in just one step now

$$\hat{SE}[\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}}] = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} = \sqrt{\frac{457.25}{12} + \frac{425.33}{7}} = 9.943g$$

Step 3 – The Confidence Coefficient

With $\sigma_1^2 \neq \sigma_2^2$ and both UNknown, the multiplier is a percentile from the Student's t with degrees of freedom given by that horrible formula (see page 52).

$$f = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right)^2}{\left(\frac{\left[\frac{S_1^2}{n_1} \right]^2}{n_1 - 1} + \frac{\left[\frac{S_2^2}{n_2} \right]^2}{n_2 - 1} \right)} = \frac{\left(\frac{457.25}{12} + \frac{425.33}{7} \right)^2}{\left(\frac{\left[\frac{457.25}{12} \right]^2}{11} + \frac{\left[\frac{425.33}{7} \right]^2}{6} \right)} = 13.0793$$

Round down so as to obtain an appropriately conservative (wide) interval.

So we'll use $f=13$. The 99.5th percentile of the Student's t with $df=13$ has value 3.0123

Step 4 – Substitute into the formula for a confidence interval

$$\begin{aligned} \text{CI} &= (\text{point estimate}) \pm (\text{confidence coefficient.}) (\text{SE of point estimate}) \\ &= 19 \pm (3.0123)(9.943) \\ &= (-11.0g, 49.0g) \quad \text{Note – This is the widest of the 3 solutions!} \end{aligned}$$

7b. Confidence Interval for σ_1^2 / σ_2^2

Suppose we want to compare two independent variances; eg - .

Examples

- Are the reproducibilities of two laboratory assays similar?
- We might want to assess the similarity of two independent normal population distributions. This would include a comparison of their two variance parameters.
- We might want to assess the similarity of two variance parameters before doing an analysis that requires assuming them to be equal.

(1- α)100% Confidence Interval for σ_1^2 / σ_2^2 Setting – Two Independent Normal Distributions	
Lower limit =	$\left(\frac{1}{F_{n_1-1; n_2-1; (1-\alpha/2)}} \right) \left[\frac{S_1^2}{S_2^2} \right]$
Upper limit =	$\left(\frac{1}{F_{n_1-1; n_2-1; (\alpha/2)}} \right) \left[\frac{S_1^2}{S_2^2} \right]$

Example

(Source: Daniel WW. *Biostatistics: A Foundation for Analysis in the Health Sciences, Fourth Edition. 1987. Page 163*)

Reaction time to a stimulus was examined in two independent groups, each a simple random sample from a Normal population distribution. One group ($X_1 \dots X_{n_1}$) is comprised of $n_1=21$ healthy adults. The other group ($Y_1 \dots Y_{n_2}$) includes $n_2 = 16$ Parkinson's disease patients. Calculate a 95% confidence interval estimate of σ_1^2 / σ_2^2 .

Preliminary calculations yield the following:

$$S_1^2 = \frac{\sum_{i=1}^{n_1} (x_i - \bar{x})^2}{n_1 - 1} = 1600 \quad \text{and} \quad S_2^2 = \frac{\sum_{i=1}^{n_2} (y_i - \bar{y})^2}{n_2 - 1} = 1225$$

Numerator degrees of freedom = $n_1 - 1 = 20$

Denominator degrees of freedom = $n_2 - 1 = 15$

Step 1 – Solution for Point Estimator S_1^2 / S_2^2

$$S_1^2 / S_2^2 = 1600 / 1225 = 1.306$$

Step 2 – Solution for Confidence Coefficient Multipliers

$$\left(\frac{1}{F_{n_1-1; n_2-1; (1-\alpha/2)}} \right) = \left(\frac{1}{F_{20; 15; .975}} \right) = \left(\frac{1}{2.76} \right)$$

$$\left(\frac{1}{F_{n_1-1; n_2-1; (\alpha/2)}} \right) = \left(\frac{1}{F_{20; 15; .025}} \right) = \left(\frac{1}{0.3886} \right)$$

Step 3 – Solution for Lower and Upper Confidence Interval Limit Values

Lower Limit of confidence interval =

$$\left(\frac{1}{F_{n_1-1; n_2-1; (1-\alpha/2)}}\right) \left[\frac{S_1^2}{S_2^2}\right] = \left(\frac{1}{2.76}\right) [1.306] = 0.47$$

Upper Limit of confidence interval =

$$\left(\frac{1}{F_{n_1-1;n_2-1;(\alpha/2)}}\right)\left[\frac{S_1^2}{S_2^2}\right]=\left(\frac{1}{0.3886}\right)[1.306]=3.36$$

8. Binomial Distribution: One Group

8a. Confidence Interval for π

Recall – The **Binomial** Distribution was introduced in Unit 4, *Bernoulli and Binomial Distributions*

- We have **n independent Bernoulli trials**. To be general, let's call the two possible outcomes “event” and “non-event”.
- “**Event/non-event**” might refer to: “alive/dead”, “tumor/remission”, “success/failure”, “heads/tails”, etc.
- In each Bernoulli trial, the outcome of “event” occurs with the same probability = π
- Suppose that “event” occurs in x of the n trials. An estimate of π is the proportion of trials in which event occurred. This is equal to x/n .
- The number of occurrences of event, X, in n independent Bernoulli trials is distributed **Binomial(n, π)**.

(With apology) There are a variety of notations for representing an estimate of π

The most clear is $\hat{\pi}$. The **caret** on the top is an indication that this is a guess.

- Another notation for $\hat{\pi}$ is p for “proportion”. This is awkward because some texts use the notation “p” is used for the population parameter π itself. Therefore, I recommend against using this to represent the estimate of π .
- A better choice for a notation for $\hat{\pi}$ is to write it as \hat{p} because it has the caret on top.
- Still another notation for $\hat{\pi}$ is X/n . This is nice because you can recognize it as the observed proportion.
- Still another is \bar{X} . This also makes sense since it is the sum n Bernoulli outcomes coded as 0's (non-events) and 1's (events), divided by n, the number of trials. *Putting these all together ...*
- **Summarizing: $\hat{\pi}$ is equivalently written: $\hat{\pi} = \hat{p} = \bar{X} = X/n$** Notice I left off the notation “p”.

In constructing a confidence interval for π of a Binomial distribution - just as we did for the mean parameter μ of a Normal distribution – we need:

1. Point estimate
2. SE of the point estimate
3. Confidence coefficient

Example –

(Source: Daniel WW. *Biostatistics: A Foundation for Analysis in the Health Sciences*)

Suppose we are interested in estimating the proportion of individuals who obtain a dental check up twice a year in a certain urban population. In a simple random sample of $n=300$ persons, $X=123$ reported having had 2 dental check ups in the last year. Construct a 95% confidence interval for π , the unknown true proportion.

1. The Point Estimate of π is the Sample Mean $\hat{\pi} = \bar{X}$

$$\bar{X} = \frac{X}{n} = \frac{123}{300} = 0.41$$

2. The Standard Error of $\hat{\pi} = \bar{X}$ is estimated using $\hat{SE}(\hat{\pi}) = \sqrt{\frac{\bar{X}(1-\bar{X})}{n}}$

This formula makes sense for two reasons:

- If X is distributed Binomial(n, π) Then $\text{Variance}(X) = n \pi (1-\pi)$
- $\text{Variance}[(\text{constant})X] = (\text{constant})^2 \text{Variance}(X)$
- For the interested: Appendix 3 is the solution for this SE formula.

$$\hat{SE}(\hat{\pi}) = \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} = \sqrt{\frac{0.41(0.59)}{300}} = 0.028$$

3. The Confidence Coefficient is a Percentile from the Normal(0,1) Distribution

Do NOT use percentiles from the Student t-distribution here! The correct percentile is one from the Normal(0,1) for reasons having to do with the central limit theorem.

- ♣ As we saw before - For a 95% confidence interval, this number will be the 97.5th percentile of the Normal (0,1) distribution.
- ♣ And in general - For a $(1-\alpha)100\%$ confidence interval, this number will be the $(1-\alpha/2)100^{\text{th}}$ percentile of the Normal (0,1) distribution.

$$z_{.975} = 1.96$$

4. Putting it all together.

Lower limit of confidence interval

$$\begin{aligned} &= (\text{point estimate}) - (\text{multiple}) (\text{SE of estimate}) \\ &= 0.41 - (1.96)(0.028) \\ &= 0.36 \end{aligned}$$

Upper limit of confidence interval

$$\begin{aligned} &= (\text{point estimate}) + (\text{multiple}) (\text{SE of estimate}) \\ &= 0.41 + (1.96)(0.028) \\ &= 0.46 \end{aligned}$$

**Confidence Interval for a proportion π
Single sample from a Binomial(n, π) Distribution**

$$\hat{\pi} \pm (z_{1-\alpha/2}) \hat{SE}(\hat{\pi})$$

where the required calculations are

(1) $\bar{X} = \frac{X}{n}$ the observed proportion of events in the n trials

(2) $\hat{\pi} = \bar{X}$

(3) $\hat{SE} = \sqrt{\frac{\bar{X}(1-\bar{X})}{n}}$

(4) **For small number of trials ($n \leq 30$ or so) use $\hat{SE} = \sqrt{\frac{0.5(0.5)}{n}}$**

Why? For small number of trials n (say $n \leq 30$), it may be desirable to compute a more conservative (wider) confidence interval by using a slightly different SE calculation.

- A closer look at the SE calculation $\hat{SE} = \sqrt{\frac{\bar{X}(1-\bar{X})}{n}}$ reveals that it attains its largest value when $\bar{X} = 0.50$

9. Binomial Distribution: Two Independent Groups

9a. Confidence Interval for $[\pi_1 - \pi_2]$

Suppose we want to compare 2 independent “event” occurrence probabilities, π_1 versus π_2 :

- Are the probabilities of **disease occurrence** the same in two populations?
- Patients are treated with either drug D, or with placebo. Is the probability of the **event of “improvement”** the same in both groups?

Suppose that, available to us, are the results of two independent Binomial random variables:

- X distributed Binomial(n_1, π_1)
- Y distributed Binomial(n_2, π_2)

We have therefore the following

$\hat{\pi}_1 = \bar{X} = \frac{X}{n_1}$ $SE(\hat{\pi}_1) = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1}}$	$\hat{\pi}_2 = \bar{Y} = \frac{Y}{n_2}$ $SE(\hat{\pi}_2) = \sqrt{\frac{\pi_2(1-\pi_2)}{n_2}}$
---	---

We have what we need for developing a confidence interval for the difference $[\pi_1 - \pi_2]$

Example

In a clinical trial for a new drug to treat hypertension, $n_1 = 50$ patients were randomly assigned to receive the new drug, and $n_2 = 50$ patients to receive a placebo. $X = 34$ of the patients receiving the drug showed improvement, while $Y = 15$ of those receiving placebo showed improvement. Calculate a 95% confidence interval estimate for the difference between proportions improved.

1. The Point Estimate of $[\pi_1 - \pi_2]$ is difference between the sample means

$$\hat{\pi}_1 = \bar{X} = X/n_1 = 34/50 = 0.68$$

$$\hat{\pi}_2 = \bar{Y} = Y/n_2 = 15/50 = 0.30$$

$$[\hat{\pi}_1 - \hat{\pi}_2] = [\bar{X} - \bar{Y}] = [0.68 - 0.30] = 0.38$$

2. The Standard Error of $[\hat{\pi}_1 - \hat{\pi}_2]$ is estimated using $\hat{SE}(\hat{\pi}_1 - \hat{\pi}_2) = \sqrt{\frac{\bar{X}(1-\bar{X})}{n_1} + \frac{\bar{Y}(1-\bar{Y})}{n_2}}$

This formula is reasonable because both sample sizes are larger than 30.

$$\hat{SE}(\hat{\pi}_1 - \hat{\pi}_2) = \sqrt{\frac{\bar{X}(1-\bar{X})}{n_1} + \frac{\bar{Y}(1-\bar{Y})}{n_2}} = \sqrt{\frac{.68(.32)}{50} + \frac{.30(.70)}{50}} = 0.0925$$

3. The Confidence Coefficient is again a Percentile from the Normal(0,1) Distribution

$$z_{.975} = 1.96$$

4. Putting it all together.

$$\text{Lower} = (\text{point estimate}) - (\text{multiple}) (\text{SE of estimate}) = 0.38 - (1.96)(0.0925) = 0.20$$

$$\text{Upper} = (\text{point estimate}) + (\text{multiple}) (\text{SE of estimate}) = 0.38 + (1.96)(0.0925) = 0.56$$

**Confidence Interval for a difference between two independent proportions $[\pi_1 - \pi_2]$
Two Independent Binomial Distributions**

$$[\hat{\pi}_1 - \hat{\pi}_2] \pm (z_{1-\alpha/2}) \hat{SE}(\hat{\pi}_1 - \hat{\pi}_2)$$

where the required calculations are

$$(1) \quad \bar{X} = \frac{X}{n_1} \quad \text{and} \quad \bar{Y} = \frac{Y}{n_2}$$

$$(2) \quad \hat{\pi}_1 = \bar{X} \quad \text{and} \quad \hat{\pi}_2 = \bar{Y}$$

$$(3) \quad \hat{SE} = \sqrt{\frac{\bar{X}(1-\bar{X})}{n_1} + \frac{\bar{Y}(1-\bar{Y})}{n_2}}$$

(4) **For small number of trials ($n \leq 30$ or so) in either group, use**

$$\hat{SE} = \sqrt{\frac{0.5(0.5)}{n_1} + \frac{0.5(0.5)}{n_2}}$$

Appendices

i. Derivation of Confidence Interval for μ : Single Sample From Normal, σ^2 known

The setting is the example in Section 5a (Confidence Interval for μ , σ^2 known).

Recall that we were given the weight in micrograms of drug inside each of 30 capsules, after subtracting the capsule weight.

0.6	0.3	0.1	0.3	0.3
0.2	0.6	1.4	0.1	0.0
0.4	0.5	0.6	0.7	0.6
0.0	0.0	0.2	1.6	-0.2
1.6	0.0	0.7	0.2	1.4
1.0	0.2	0.6	1.0	0.3

We're told that $\sigma^2 = 0.25$

Step 1 – Obtain a point estimate \bar{X}

$$\bar{X}=0.51$$

$$n = 30$$

Step 2 – Obtain the SE of the point estimate \bar{X} by recalling that $SE(\bar{X})=\sigma/\sqrt{n}$

$$SE(\bar{X})=\sigma/\sqrt{n} = 0.5/\sqrt{30}$$

$$= 0.0913$$

Step 3 – Select desired confidence = $(1 - \alpha)$

Suppose we want a 95% confidence interval.

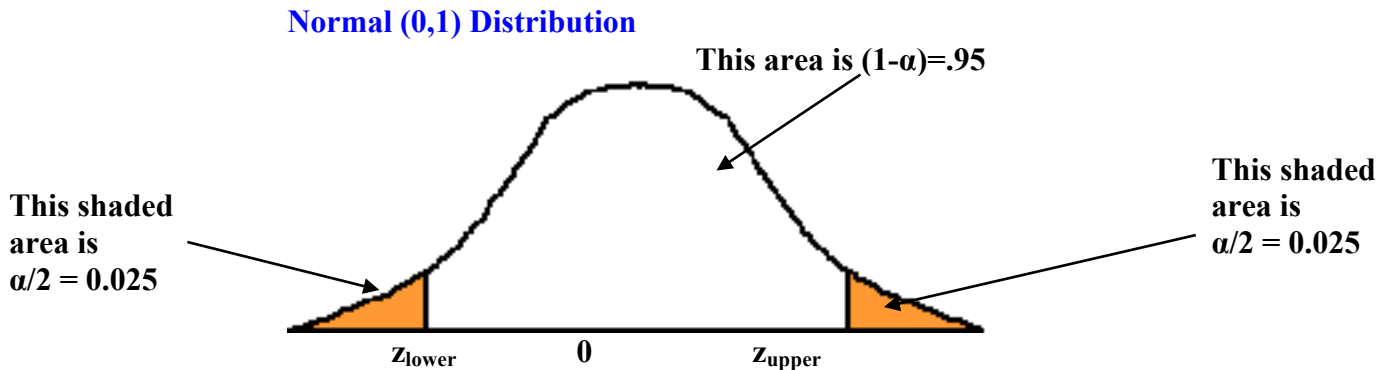
Then $(1 - \alpha) = 0.95$.

This means that $\alpha = 0.05$

The $\alpha = 0.05$ is the probability of error

Step 4 – Using a calculator for the Normal (0,1) distribution, obtain symmetric values of a standard normal deviate Z (call these z_{lower} and z_{upper}) such that

$$\text{Probability} [z_{\text{lower}} \leq Z \leq z_{\text{upper}}] = 0.95$$



$$\text{Probability} [-1.96 \leq Z \leq +1.96] = 0.95 \text{ so that}$$

$$z_{\text{lower}} = -1.96$$

$$z_{\text{upper}} = +1.96$$

This expression, $\text{Probability} [-1.96 \leq Z \leq +1.96] = 0.95$ in this example.

More generally, it is $\text{Probability} [z_{\text{lower}} \leq Z \leq z_{\text{upper}}] = (1 - \alpha)$

To get to this expression, standardize \bar{X}

$\text{Probability} [-1.96 \leq Z \leq +1.96] = 0.95$ in this example is actually

$$\text{Probability} [z_{\text{lower}} \leq Z \leq z_{\text{upper}}] = (1 - \alpha) \rightarrow$$

note #1 - Because the Normal(0,1) distribution is symmetric about the value 0

$$z_{\text{lower}} = (-1) z_{\text{upper}}$$

So let's call z_{upper} simply z

This allows us to simplify the above expression with two convenient substitutions

$$z_{\text{upper}} = z$$

$$z_{\text{lower}} = -z$$

Probability $[-z \leq Z \leq z] = (1 - \alpha) \rightarrow$

note #2 - Now we'll insert another convenient substitution

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

Probability $[-z \leq \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \leq z] = (1 - \alpha) \rightarrow$

note #3 - All that remains is to do the algebra necessary to “isolate” μ

Probability $\left[\left(\frac{\sigma}{\sqrt{n}} \right) -z \leq \bar{X} - \mu \leq \left(\frac{\sigma}{\sqrt{n}} \right) z \right] = (1 - \alpha) \rightarrow$

With confidence $(1 - \alpha)100\%$, $\left[\bar{X} - \left(\frac{\sigma}{\sqrt{n}} \right) z \leq \mu \leq \bar{X} + \left(\frac{\sigma}{\sqrt{n}} \right) z \right]$ **which matches.**

ii. Derivation of Confidence Interval for σ^2 : Single Sample from a Normal

The setting here is the example in Section 5c.

A precision instrument is guaranteed to read accurately to within ± 2 units. A sample of 4 readings on the same object yield 353, 351, 351, and 355. Calculate a 95% confidence interval estimate of the population variance σ^2 .

Step 1 – Obtain a point estimate S^2 and its associated degrees of freedom

$$S^2 = 3.67$$

$$df = 3$$

Step 2 – Recalling the material from section 2b, define the appropriate chi square random variable

$$Y = \frac{(n-1)S^2}{\sigma^2} \text{ is distributed Chi Square with degrees of freedom } = (n-1)$$

Step 3 – Select desired confidence = $(1 - \alpha)$

For desired confidence = .95, $(1 - \alpha) = 0.95$.

Step 4 – Substitute for χ^2 in the middle of the “area under the curve” calculation for a chi square random variable as follows.

$$\text{Probability} [\chi_{df, \alpha/2}^2 \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_{df, (1-\alpha/2)}^2] = (1 - \alpha)$$

Step 5 – Do the algebra to obtain an expression that is the confidence interval for σ^2 .

$$\text{Probability} \left[\chi_{\text{df}, \alpha/2}^2 \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_{\text{df}, (1-\alpha/2)}^2 \right] = (1-\alpha) \rightarrow$$

$$\text{Probability} \left[\frac{1}{\chi_{\text{df}, (1-\alpha/2)}^2} \leq \frac{\sigma^2}{(n-1)S^2} \leq \frac{1}{\chi_{\text{df}, \alpha/2}^2} \right] = (1-\alpha) \rightarrow$$

$$\text{With confidence } (1-\alpha)100\%, \left[\frac{(n-1)S^2}{\chi_{\text{df}, (1-\alpha/2)}^2} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi_{\text{df}, \alpha/2}^2} \right] \text{ which matches.}$$

iii. The Standard Error of $\hat{\pi} = \bar{X}$ is estimated using $\hat{SE}(\hat{\pi}) = \sqrt{\frac{\bar{X}(1-\bar{X})}{n}}$

We can take advantage of two results -

- If X is distributed Binomial(n,π) Then Variance(X) = n π (1-π)
- Variance[(constant)X] = (constant)² Variance (X)

Proof

$$SE(\bar{X}) = \sqrt{\text{Variance}(\bar{X})}$$

$$= \sqrt{\text{Variance}\left(\frac{X}{n}\right)}$$

$$= \sqrt{\left(\frac{1}{n^2}\right) (\text{Variance}[X])}$$

$$= \sqrt{\left(\frac{1}{n^2}\right) (n\pi[1-\pi])}$$

$$= \sqrt{\frac{\pi[1-\pi]}{n}}$$

The problem now is that π is not known. So it is replaced by its estimate

$$\approx \sqrt{\frac{\bar{X}[1-\bar{X}]}{n}}$$

which matches.