

Unit 6
Estimation

Topic	<ol style="list-style-type: none"> 1. Introduction 2 <ol style="list-style-type: none"> a. Goals of Estimation 5 b. Notation and Definitions 7 c. How to Interpret a Confidence Interval 10 2. Preliminaries: Some Useful Probability Distributions 17 <ol style="list-style-type: none"> a. Introduction to the Student t- Distribution 17 b. Introduction to the Chi Square Distribution 21 c. Introduction to the F-Distribution 25 d. Sums and Differences of Independent Normal Random Vars .. 28 3. Normal Distribution: One Group 30 <ol style="list-style-type: none"> a. Confidence Interval for μ, σ^2 Known 30 b. Confidence Interval for μ, σ^2 Unknown 35 c. Confidence Interval for σ^2 38 4. Normal Distribution: Paired Data 41 <ol style="list-style-type: none"> a. Confidence Interval for $\mu_{\text{DIFFERENCE}}$ 42 b. Confidence Interval for $\sigma^2_{\text{DIFFERENCE}}$ 45 5. Normal Distribution: Two Independent Groups: 46 <ol style="list-style-type: none"> a. Confidence Interval for $[\mu_1 - \mu_2]$ 46 b. Confidence Interval for σ_1^2 / σ_2^2 54 6. Binomial Distribution: One Group 57 <ol style="list-style-type: none"> a. Confidence Interval for π 57 7. Binomial Distribution: Two Independent Groups 61 <ol style="list-style-type: none"> a. Confidence Interval for $[\pi_1 - \pi_2]$ 61 	
	<p>Appendices</p> <ol style="list-style-type: none"> i. Derivation of Confidence Interval for μ – Single Normal with σ^2 Known 64 ii. Derivation of Confidence Interval for σ^2 – Single Normal 67 iii. SE of a Binomial Proportion 69 	

1. Introduction

Recall our introduction to biostatistics. It 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.

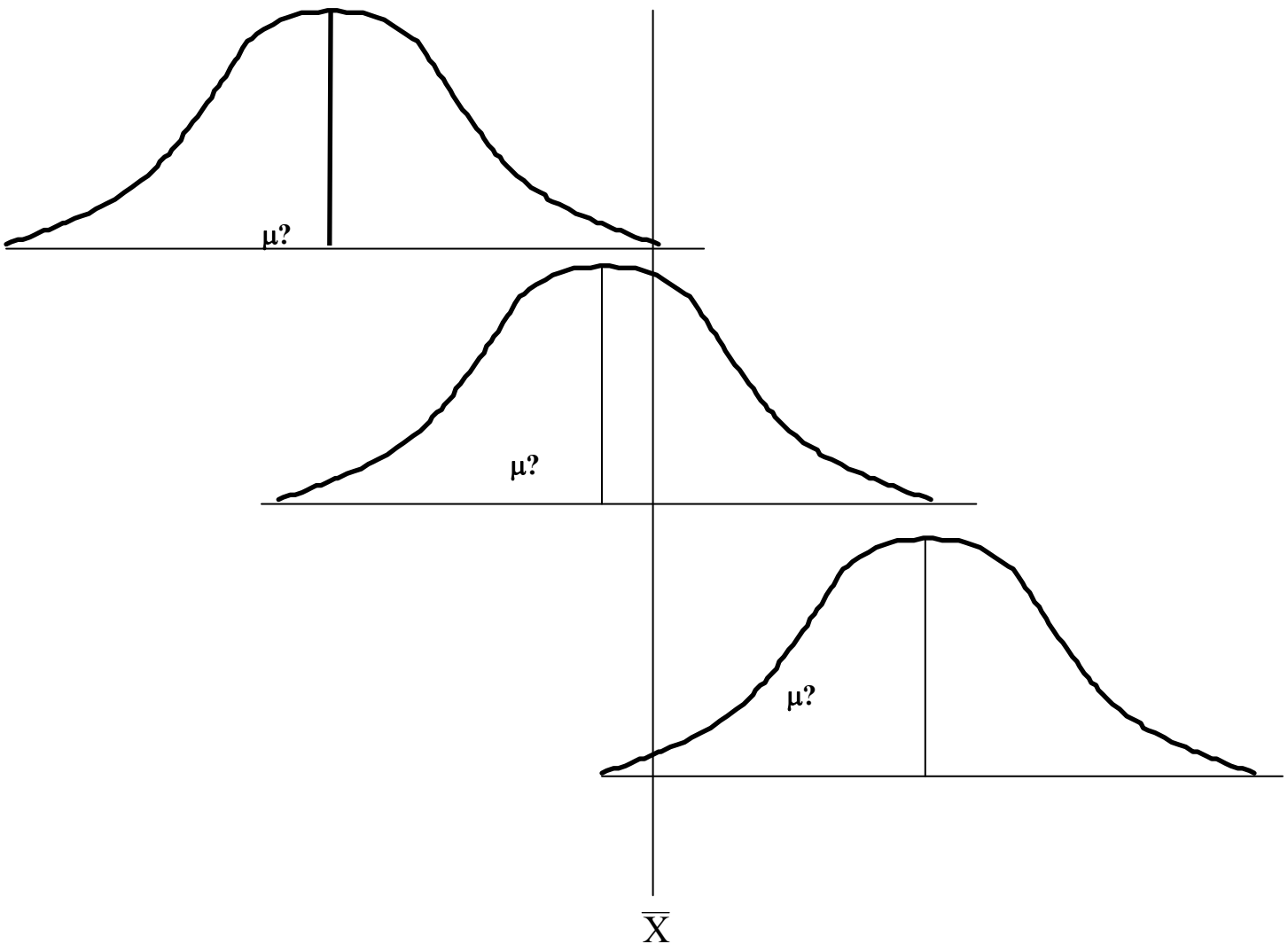
- ♣ **Description** - Information in a sample is used to summarize the sample itself. It is also used to make guesses of the characteristics of the source population.
- ♣ **Hypothesis Testing** – Information in a sample is used to gauge the plausibility of competing explanations for the phenomena observed.

Unit 6 is about using information in a sample to make estimates of the characteristics (parameters) of the source population. We already have a feel for the distinction between statistics and parameters:

<u>Sample Statistics are Estimators</u>	<u>of Population Parameters</u>
Sample mean \bar{X}	μ
Sample variance S^2	σ^2

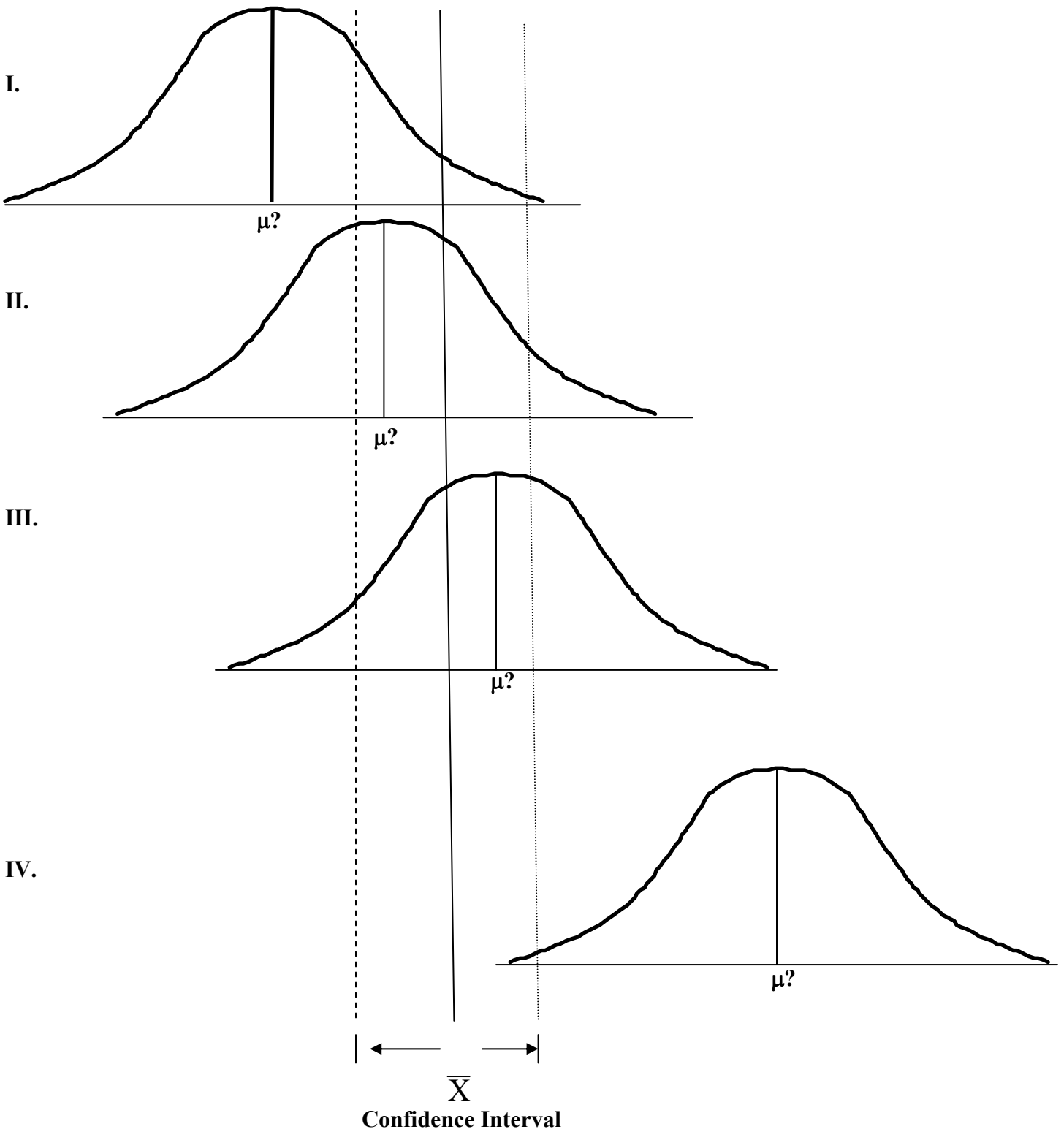
What does it mean to say we know \bar{X} from a sample but we don't know the population mean μ ?

We have a sample of n observations $X_1 \dots X_n$, and from these we have calculated their average, \bar{X} . Interest is in learning about the population from which these observations came. Any one of a number of possible populations might have been the source of our sample of data. Which population appears most likely to have given rise to the sample mean \bar{X} that we have observed? For simplicity here, suppose there are 3 possibilities.



This is the \bar{X} we see.

We cannot ask: What is the correct μ ? Any of these distributions could have been the source population distribution. So we construct an interval of plausible μ . In the following picture, I'm imagining four possibilities instead of three.



We are “confident” that μ could be the mean parameter in populations “II” or “III”.

1a. Goals of Estimation

What we regard as a good estimator depends on the criteria we use to define “good”. There are actually multiple criteria. Here we consider one set of criteria.

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**)

“In the Long Run Correct” -

Tip: Recall the introduction to statistical expectation and the meaning of unbiased (See Unit 4 – *Bernoulli & Binomial* pp 4-6.

“In the long run correct” says - If we imagine replicating the study over and over again and each time calculating a statistic of interest (so as to produce the sampling distribution of that statistic of interest), the mean of the sampling distribution for that statistic of interest is actually equal to the target parameter value being estimated.

E.g. – Consider S^2 as an 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 , has the value that is equal to its “target” σ^2 .

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

Mathematically, this is actually saying that the statistical expectation of S^2 is equal to its target σ^2 .

Recall – we write this as $E[S^2] = \sigma^2$. For the “mathematically inclined” among you:

$$\int S^2 f_{S^2}(S^2) dS^2 = \sigma^2$$

“In Error by as Little as Possible” –

“In error by as little as possible” is about the following - 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. Their smallest variability from sample to sample is the idea of 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 **data 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: In this course, we will be using the criteria of “minimum variance unbiased”. However, other criteria are possible.

1b. 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],

Whereas, an **interval estimate** is of the form:

[lower limit, upper limit]

Example -

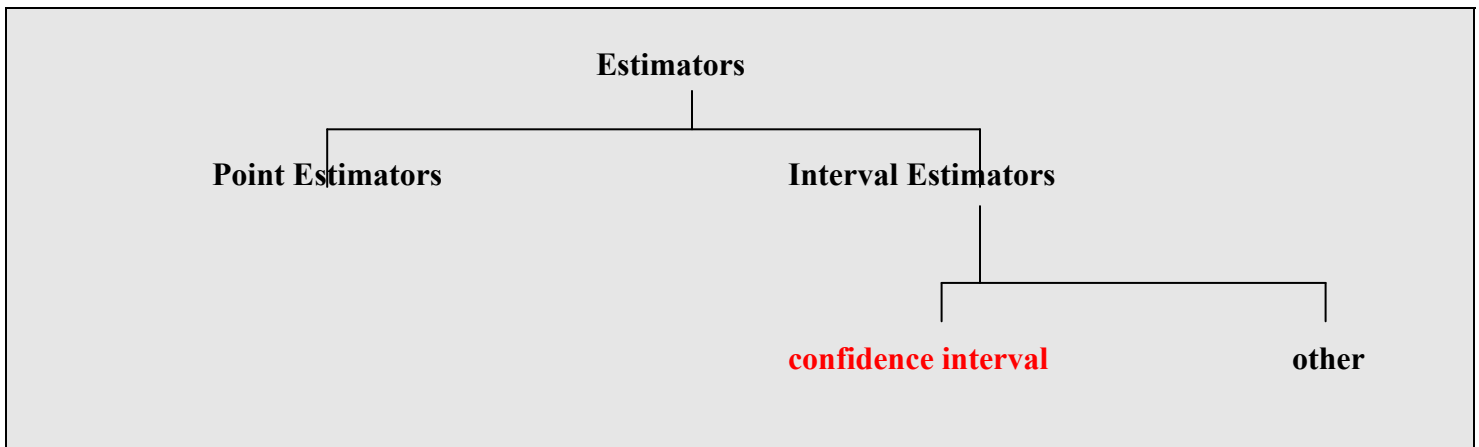
The sample mean \bar{X}_n , calculated using data in a sample of size n , is a point estimator of the population mean μ . If $\bar{X}_n = 10$, the value 10 is called a point estimate of the population mean μ .

Sampling Distribution

- ♣ It's helpful to recall the idea of a **sampling distribution** again. One can produce a "population" of all possible sample means \bar{X}_n by replicating simple random sampling over and over again, each time with some same sample size = "n" and compiling the resulting collection of sample means \bar{X}_n into a kind of "population" that we now call a sampling distribution.
- ♣ It is also helpful to recall that **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 the mean or expected value of the statistic, taken over its sampling distribution, is equal to the population parameter value. Intuitively, this is saying that the "long run" average of the statistic is equal to the population parameter value.



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) - (multiple)(standard error)

upper limit = (point estimate) + (multiple)(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!!

Many users of the confidence interval idea produce an interval of estimates but then err in focusing on its midpoint.

1c. 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 just yet. We'll come to these later.

Example

Interest is in estimating the average income from wages for a population of 5000 workers, X_1, \dots, X_{5000}

The average income to be estimated is the population mean μ .

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

Suppose $\mu = \$19,987$ and suppose we do not know this value.

We wish to estimate μ from a sample of wages.

The population standard deviation σ is a population parameter describing the variability among the 5000 individual wages.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{5000} (X_i - \mu)^2}{5000}} = \sqrt{\frac{\sum_{i=1}^{5000} (X_i - 19987)^2}{5000}}$$

Reminder – Notice in the calculation of σ that (1) squared deviations are computed about the reference value equal to the actual population mean μ and (2) division is by the actual population size $N=5000$; division is **not** by $N-1$. Do you remember why this is the correct calculation? **Answer** – this is a population parameter value calculation, not a statistic calculated from a sample.

Suppose we know that $\sigma = \$12,573$.

We will use the standard error to describe a typical departure of a sample mean away from the population mean μ

We illustrate the meaning of a confidence interval using two samples of different sizes.

Carol wants to estimate μ

She has a sample of data from interviews of $n=10$ workers

Data are X_1, \dots, X_{10}

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

$$\sigma = 12,573$$

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

Ed wants to estimate μ

He has a sample of data from interviews of $n=100$ workers

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 $n=100$...

- Notice that the variability of an average of 100 is less than the variability of an average of 10.
- It seems reasonable that we should have more confidence (smaller safety net) in our sample mean as a guess of the population mean when it is based on 100 observations than when it is based on 10.
- By the same token, we ought to have complete (100%) confidence (no safety net required at all) if we could afford to interview all 5000. This is because 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. It has 3 components:

- 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)

The “safety net” (confidence interval) that we construct has “lower” and “upper” limits defined

Lower limit = (point estimate) – (confidence coefficient)(SE)

Upper limit = (point estimate) + (confidence coefficient)(SE)

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 .”

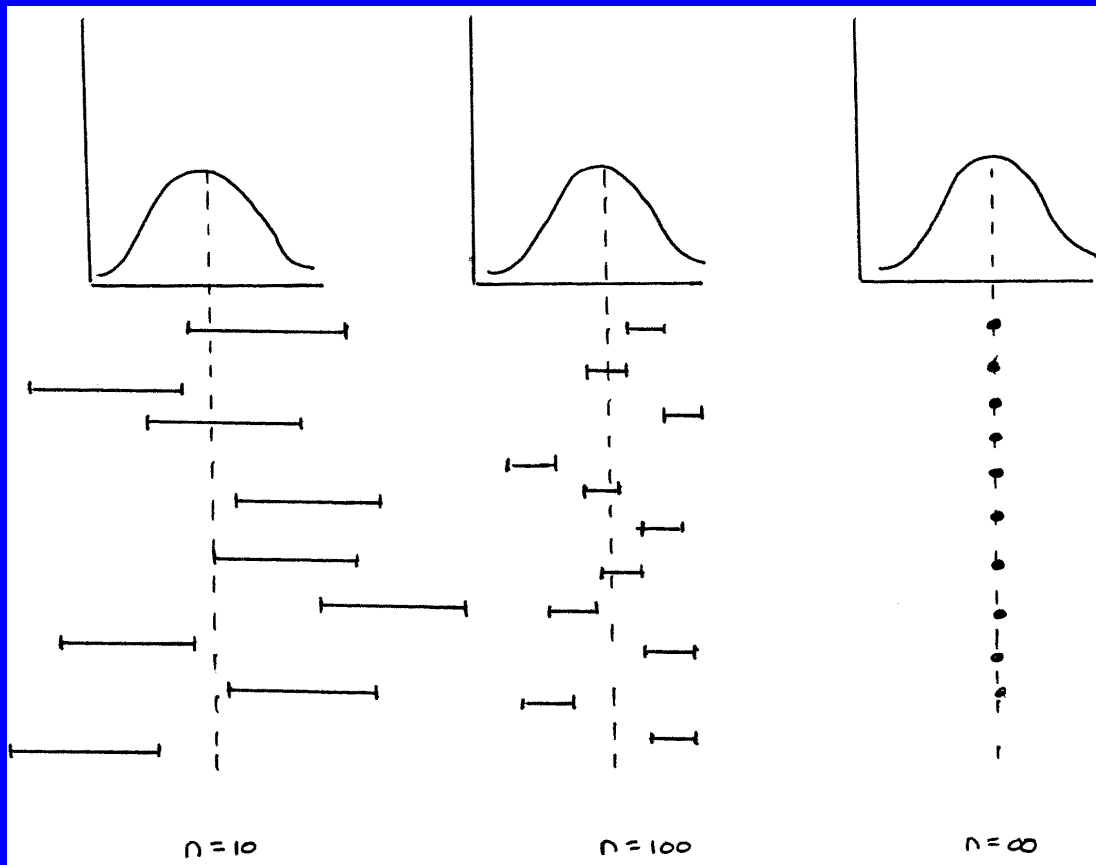
Caution on the use of Confidence Intervals:

- 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).

A picture helps in getting a feel for the ideas of confidence interval, safety net, and precision:



For each sampling plan ($n=10$, or $n=100$, or $n=\infty$), the figure (admittedly not the fanciest of art) gives a feel for the collection of all possible confidence intervals.

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 (10 or 100 or ∞), the width of all the confidence intervals is the same.
- (3) Confidence intervals based on larger sample sizes are more narrow (*more precise*)
- (4) When n is equal to the size of the population, μ is in the interval every time.

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.

2. Preliminaries: Some Useful Probability Distributions

2a. Introduction to the Student t-Distribution

There are a variety of random variables (or transformations of random variables that are distributed student's t. A particularly advantageous definition is one that appeals to our understanding of the z-score.

A Definition of a Student's t Random Variable

In the setting of a random sample $X_1 \dots X_n$ of independent, identically distributed outcomes of a $\text{Normal}(\mu, \sigma^2)$ distribution, where we calculate \bar{X} and S^2 in the usual way:

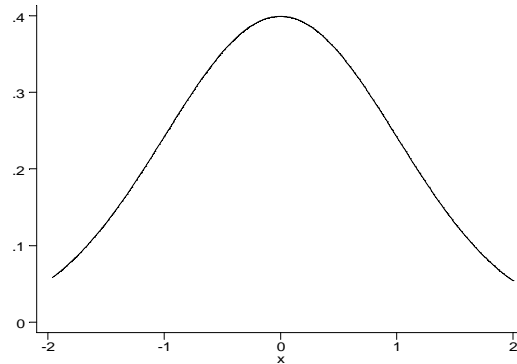
$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} \quad \text{and} \quad S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$

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

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

Note – We often use the abbreviation “df” to refer to “degrees of freedom”

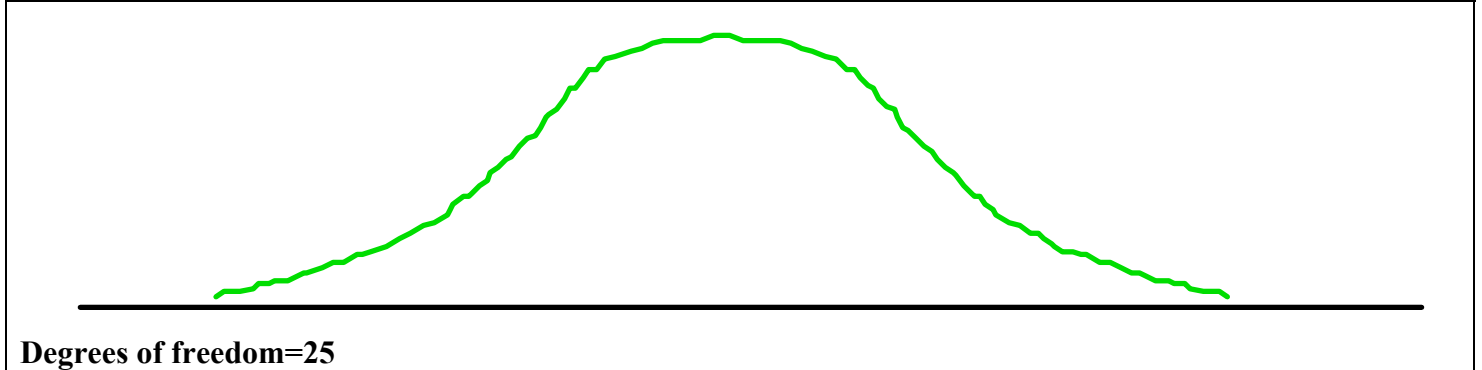
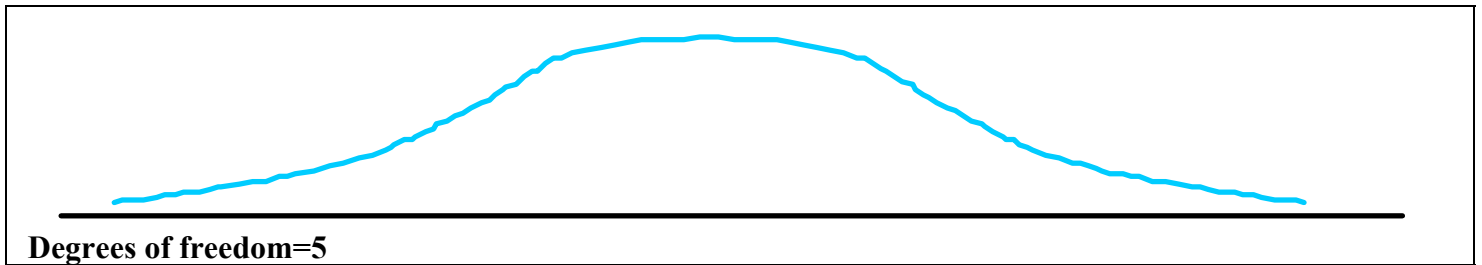
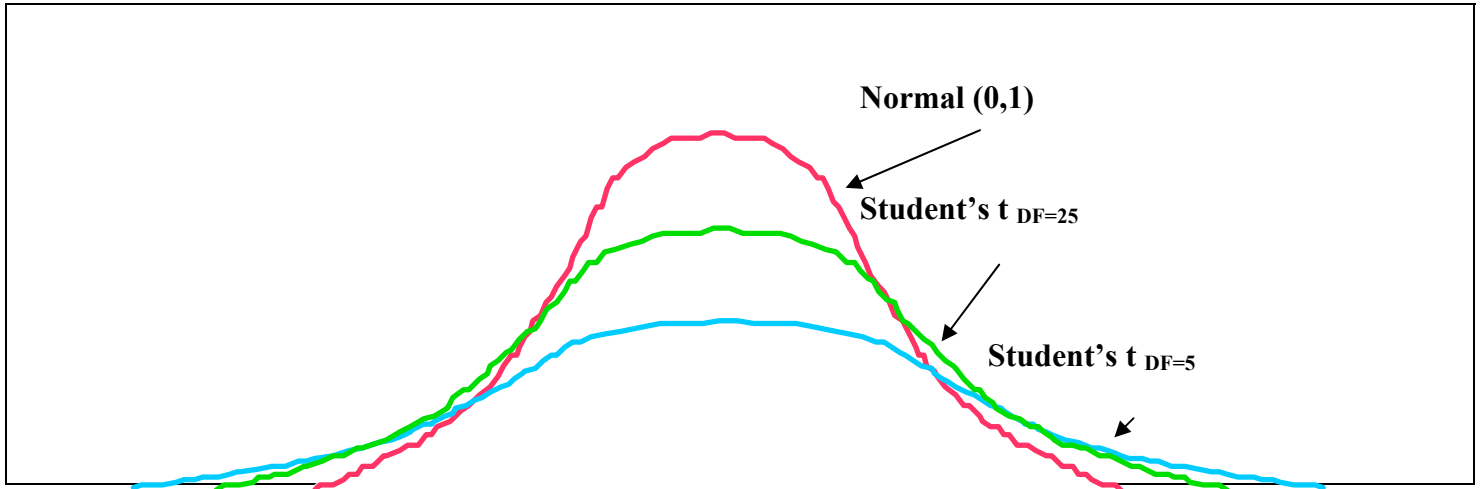
The features of the Student's t-Distribution are similar, but not identical, to those of a Normal Distribution



- **Bell Shaped**
- **Symmetric about zero**
- **Flatter than the Normal (0,1). This means**
 - (i) **The variability of a t is greater than that of a Z that is 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) should make intuitive 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's t-Table in the course text \(Table A3 of Kirkwood and Sterne, 2nd Edition\)](#)

Source: page 473.

Each row gives information for a separate t -distribution defined by the degree of freedom

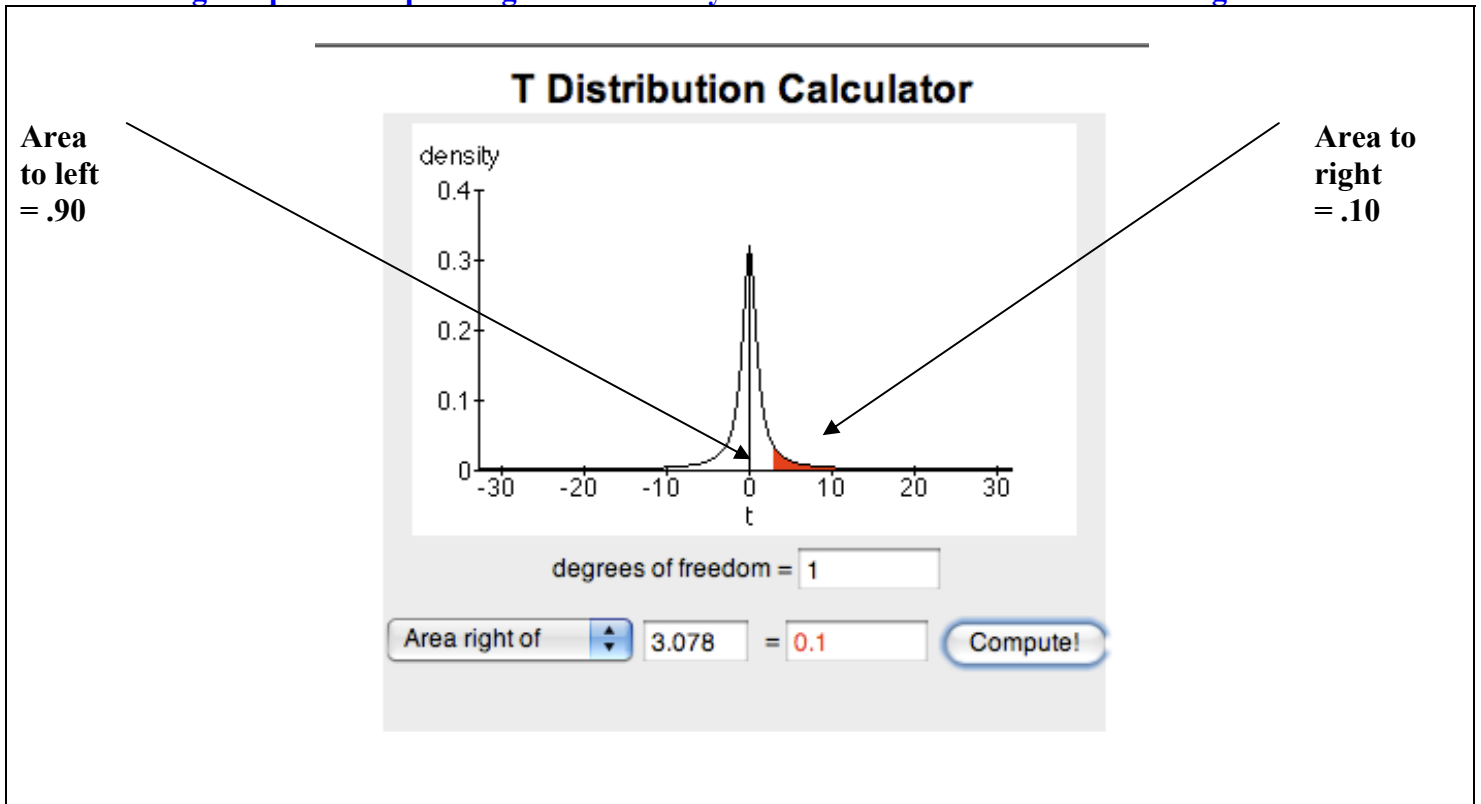
- The notation “df” to denote degrees of freedom

The column heading tells you the a right tail area (“one sided p-value) or two tailed area (“two sided p-value)

- eg. the column heading 0.25 just below “one sided p-value” separates the upper 25% of the distribution from the lower 75%

The body of the table is comprised of values of the student t random variable

The following is a picture explaining the table entry of 3.08 for df=1 and column heading=.10



Calculator used: <http://www.stat.tamu.edu/~west/applets/tdemo.html>

2b. Introduction to the Chi Square Distribution

Heuristic Definition of a Chi Square Random Variable:

We will be interested in calculating, on the basis of a random sample from a Normal distribution, a confidence interval estimate of the normal distribution variance parameter, σ^2 . The required standardizing transformation will be seen to be given by:

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

where S^2 is the sample variance. This new random variable (Y) that is a function of information in a random sample from a Normal(μ, σ^2) distribution – through the calculation of S^2 - is said to follow a chi square distribution with (n-1) degrees of freedom.

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

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:

$$V = \frac{(X - \mu)^2}{\sigma^2}$$

follows a 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:

$$V_1 + \dots + V_k$$

is distributed chi square with degrees of freedom = k.

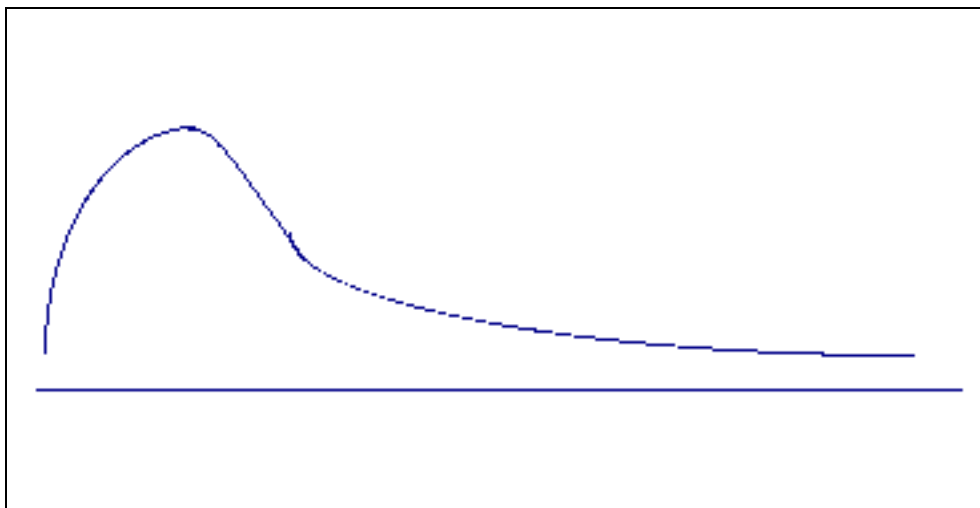
Now we need to Reconcile the Two Definitions of a Chi Square Distribution

The two definitions on the previous page are consistent because it is possible (with a little algebra) to re-write $Y = \frac{(n-1)S^2}{\sigma^2}$ as the sum of $(n-1)$ independent chi square random variables V , each with degrees of freedom = 1.

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) 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.



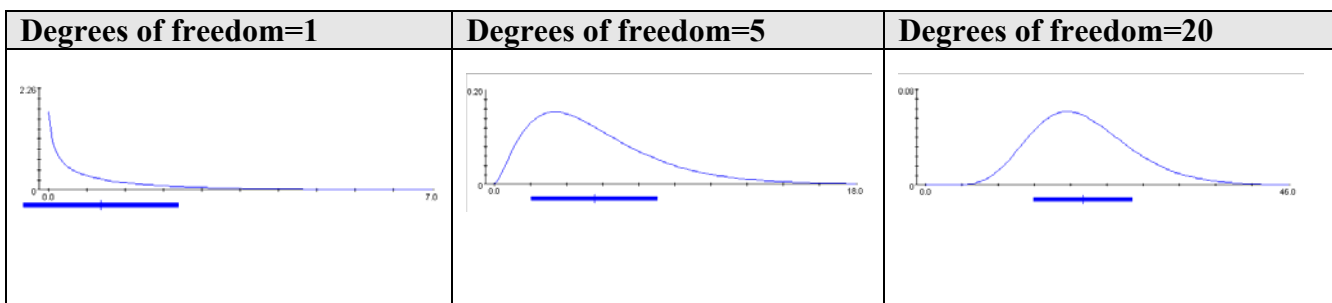
- (4) The fact that the chi square distribution is NOT symmetric about zero means that for $Y=y$ where $y>0$:

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

However, because the total areas 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. Following is an illustration of this point.



Source: http://www.ds.unifi.it/VL/VL_EN/special/special4.html

- (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.

[How to Use the Chi Square-Table in the course text \(Kirkwood and Sterne, 2nd Edition\)](#)

Source: Table A5 page 476

The format of this table is similar to that of Table A3 (Student's t) on page 473 of Kirkwood and Sterne.

Each row gives information for a separate chi square distribution defined by the degree of freedom

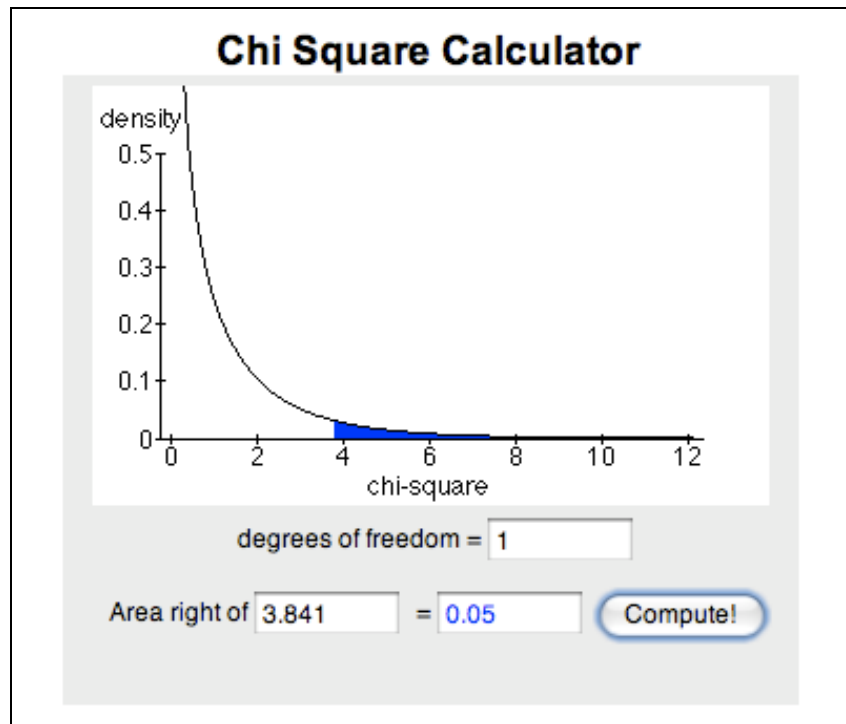
- The notation "df" tells you the degrees of freedom

The column heading tells you the right tail probability (or area under the curve)

- eg. The column heading 0.5 separates the lower 50% of the distribution from the upper 50%

The body of the table is comprised of values of the chi square random variable.

The following is a picture explaining the table entry of 3.84 for df=1 and column heading=.05



Calculator used:

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

2c. Introduction to the F Distribution

A Rationale for Introducing the F Distribution is Interest in Comparing Two 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.

We are often interested in comparing the variances of 2 groups.

- This may be the primary question of interest: I have a new measurement procedure – are the results more variable than those obtained using the standard procedure?
- Comparison of variances is sometimes 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.

Thus, for comparing 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 this ratio is 1, then the variances are the same. If it is far from 1, then the variances differ.
- It is because we wish to make probability statements about ratios of variances, and to compute confidence intervals that we introduce the F distribution.

A Definition of the F-Distribution

Suppose X_1, \dots, X_{n_x} is 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} is 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 the following is said to be distributed F

$$F_{n_x-1, n_y-1} = \frac{S_X^2 / \sigma_x^2}{S_Y^2 / \sigma_y^2} \quad \text{with two degree of freedom specifications}$$

Numerator degrees of freedom = $n_x - 1$

Denominator degrees of freedom = $n_y - 1$

A Wonderful Result

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

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

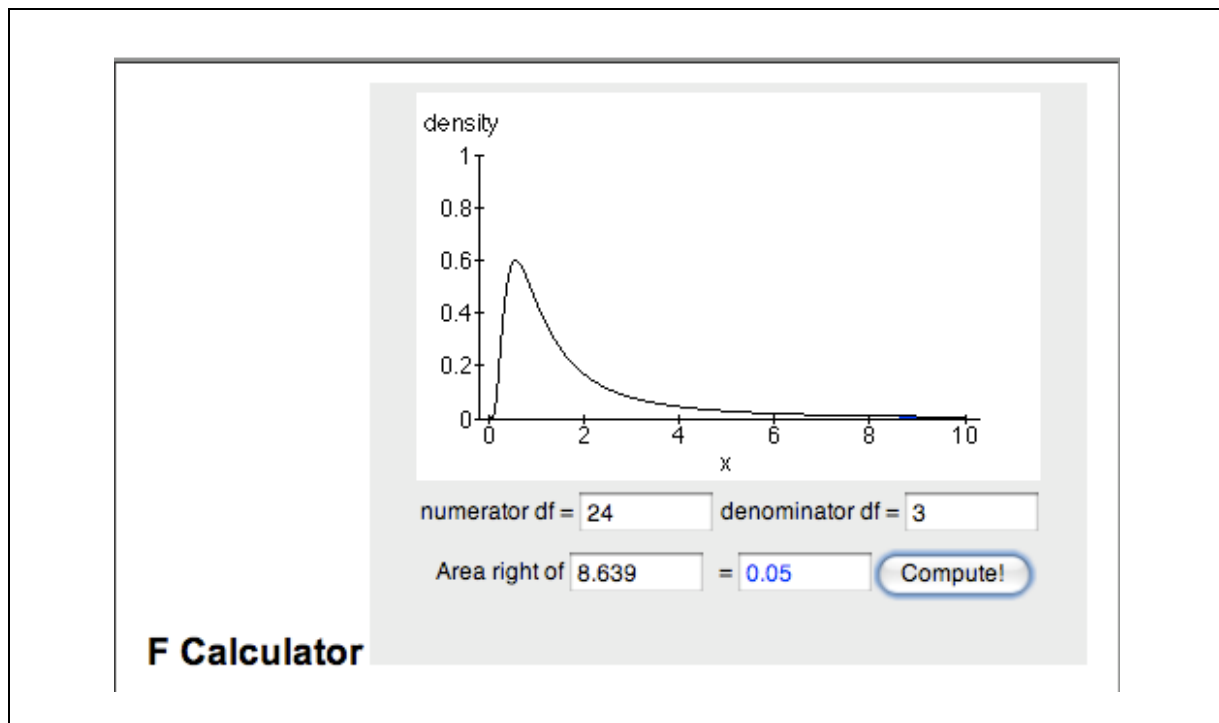
Notice that (1) the degrees of freedom are in opposite order, and (2) the solution for a left tail percentile is expressed in terms of a right tail percentile.

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 Texas A & M

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

- “df” stands for “degrees of freedom”. Enter numerator df and denominator df.
 - If you want a right tail probability, type in F-value in box “Area right of”.
 - If you want the value of a percentile, type in the right tail area in the box “=”
 - Click “Compute!”.
-
- **Example** - What is the 95th percentile value of an F distribution random variable with numerator degrees of freedom equal to 24 and denominator degrees of freedom equal to 3?



Answer: The 95th percentile of an F distribution with df=24, 3 is equal to 8.639

2d. Sums and Differences of Independent Normal Random Variables

We will be interested in calculating, on the basis of two independent random samples, one from each of two groups of interest (eg – randomized controlled trial of placebo versus treatment groups), confidence interval estimates of the difference of the means.

:

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

- We will see that 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 will need to know the standard error of

$$[\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}}]$$

Definitions

IF

- Each X_i in the sample of size= n_1 from group #1 is Normal (μ_1, σ_1^2)
- Each X_i in the sample of size= n_2 from group #2 is Normal (μ_2, σ_2^2)
- This is great!** We *already* know the sampling distribution of each sample mean

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

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

THEN

- Now, without worrying about the details,** we now also have the following tool for the sampling distribution of the difference between two independent sample means, each of which is distributed normal.

$[\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}}]$ 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. Following is the correct formula.** *Notice – You must first sum the variance and then take the square root of the sum.*

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

A General Result that is Sometimes Useful
If random variables X and Y are independent with

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

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

Then

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

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

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

NOTE: 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.

3. Normal Distribution: One Group

3a. Confidence Interval for μ (σ^2 Known)

Introduction and “where we are going” ...

In this and in subsequent sections, the “idea” of the confidence interval introduced in section 1 is operationalized. Hopefully you will see that the logic and mechanics of confidence interval construction are very similar across a variety of settings.

In this lecture in particular, 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 have seen that there are 3 elements to a confidence interval:

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

Consider the task of computing 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 .

- Presented in this and the next pages is **instruction** in how to construct a confidence interval.
- Presented in Appendix 1 is the **statistical theory** underlying this methodology. I encourage you strongly to have a look at this, too!

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 section 2, 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 the picture on page 15 to get a feel for this.*

- ♣ For a 95% confidence interval, this number will be the 97.5th percentile of the Normal (0,1) distribution.
- ♣ 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.
- ♣ On the next page are 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.

<u>Confidence Level</u>	<u>Percentile</u>	<u>Confidence Coefficient = Percentile Value from Normal (0,1)</u>
.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- α)	(1- $\alpha/2$)100 th	-

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

Example -

We are given the weight in micrograms of drug inside each of 30 capsules, after subtracting the capsule weight. Requested is 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

We're told

- ♣ 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 !!** Thus, the solution in real life is actually slightly different (it will involve using a new distribution, the Student's t-distribution). Here however, it is considered known so that the ideas can be introduced more simply.

Recall that the basic structure of the required confidence interval is point estimate \pm safety net:

Lower limit = (point estimate) - (multiple) (SE of point estimate)

Upper limit = (point estimate) + (multiple) (SE of point estimate)

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. See the table on page 19 and locate that value is 1.96.

<u>Desired Confidence Level</u>	<u>Value of Confidence Coefficient</u>
.95	1.96

Here 1.96 = (1-.05/2)100th = 97.5th percentile of the Normal(0,1) distribution

Putting this all together –

Lower limit = (point estimate) - (multiple) (SE of point estimate)

$$= 0.51 - (1.96) (0.0913)$$

$$= 0.33$$

Upper limit = (point estimate) + (multiple) (SE of point estimate)

$$= 0.51 + (1.96) (0.0913)$$

$$= 0.69$$

Thus, we have the following general formula for a $(1 - \alpha)100\%$ 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)

This is an exercise in computing a probability and draws on the ideas of topic 5 (*The Normal Distribution*). We consider an example. This question is addressed using the transformation formula for obtaining a z-score.

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 $\mu = 250$

Solution:

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

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

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

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

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

$$= 0.9910.$$

3b. Confidence Interval for μ (σ^2 NOT known)

- In section 3a, we permitted ourselves the luxury of pretending that σ^2 is known and obtained a confidence interval for μ of the form

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

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

- 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).

- Thus, we have to modify our “machinery” (specifically the SE piece of our machinery) to accommodate the unknown-ness of σ^2 . Fortunately, this is conceptually not difficult.

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

- Thus, for the setting of seeking a confidence interval for an unknown mean μ , the confidence interval will be of the following form

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

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

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 .

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}

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

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 –

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

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

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

3c. Confidence Interval for σ^2

Where we are going: We want a confidence interval for σ^2 . Its solution involves 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 person *I* “too variable”
 - Comparison of measurement techniques – is a new method more variable than a standard method?
- We have an obvious point estimator of σ^2 . It is S^2 , which we have shown earlier is an unbiased estimator.
- 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 .

- Presented here is **instruction** in how to construct a confidence interval.
- Presented in Appendix 2 is the **derivation** of the formula that you will be using.

Formula for a $(1-\alpha)100\%$ 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 to Illustrate the Calculation

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 .

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

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

Note – I used the following URL.

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

4. Put it all together, obtain

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

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

Obtain a Confidence Interval for the Population Standard Deviation σ

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.

4. Normal Distribution: Paired Data

Introduction to Paired Data

- Paired data arises when each individual (more specifically, each unit of measurement) in a sample is measured twice.
- Paired data are familiar. The two occasions of measurement might be "pre/post", "before/after", "right/left", "parent/child", etc.
- Here are some examples of paired data:
 - 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.
- Notice 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.
- One focus in an analysis of paired data is the **comparison of the two outcomes**. For **continuous** data, especially, this comparison is often formulated using the **difference** between the two 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.*

For example:

- 1) Blood pressure prior to and following treatment. Interest is $d = \text{pre} - \text{post}$. 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 $d = \text{pre} - \text{post}$. 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.

4a. Confidence Interval for $\mu_{\text{DIFFERENCE}}$

- Without worrying (for the present) about the details, consider the following: if two measurements of the same phenomenon (eg. blood pressure, # cigarettes/week, etc) X and Y are measured on an individual and if each is normally distributed, then their difference is also distributed normal.
- Thus, the setting is our focus on the difference D and the following assumptions
 - (1) $D = (X - Y)$ is distributed Normal with
 - (2) Mean of $D = \mu_{\text{difference}}$. Let's write this as μ_d
 - (3) Variance of $D = \sigma_{\text{DIFFERENCE}}^2$ Let's write this as σ_d^2
- Thus, 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
 - (1) Confidence Interval for μ_d - Normal Distribution σ_d^2 unknown
 - (2) 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). The comparison is made by calculating the difference "d" in the scores on a standard reading test. A total of n=30 children are studied. Following are the data.

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

- Of interest are

- (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}

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

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

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

$$\begin{aligned} \text{Upper limit} &= (\text{point estimate}) + (\text{conf coeff.}) (\text{SE of point estimate}) \\ &= 0.51 + (2.756)(0.0897) \\ &= 0.7573 \end{aligned}$$

4b. 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.

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, the percentiles we want are

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

From either your text or the url online, use the row for $df = 29$

- (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

$$\begin{aligned} \text{(i) Lower limit} &= \frac{(n-1)S_d^2}{\chi_{1-\alpha/2}^2} = \frac{(29)(0.2416)}{39.09} = 0.1792 \\ \text{(ii) Upper limit} &= \frac{(n-1)S_d^2}{\chi_{\alpha/2}^2} = \frac{(29)(0.2416)}{19.77} = 0.3544 \end{aligned}$$

5. Normal Distribution: Two Independent Groups

Illustration of the Setting of Two Independent Groups

A researcher performs a drug trial involving two independent groups.

- A **control** group is treated with a placebo while, separately;
- The **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.

5a. Confidence Interval for $[\mu_{\text{GROUP1}} - \mu_{\text{GROUP2}}]$

Interest is in a comparison of the mean response in one group with the mean response in a separate group under the assumption that the responses are independent.

Here are some examples. In every example, we are interested in the similarity of the two groups.

- 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 is often formulated using the **difference** between the means of the two groups.

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

Recall again the idea that there are 3 components to a confidence interval

- (1) A point estimator
- (2) The standard error of the point estimator
- (3) Confidence coefficient

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

- An obvious 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: What “noise” is associated with this point estimator? Here is where we need to know the standard error of [$\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}}$] that we learned in section 2d.

- Each X_i in the sample of size= n_1 from group #1 is Normal (μ_1, σ_1^2)
- Each X_i in the sample of size= n_2 from group #2 is Normal (μ_2, σ_2^2)
- Recall - we 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 ($\mu_2, \sigma_2^2 / n_2$)

- Now we make use of the following.

[$\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}}$] 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 what is known in your situation. There are 3 possible formulae. Use the one that is appropriate to your situation.

Solution 1 - Use this when σ_X^2 and σ_Y^2 are both known

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

Solution 2 - Use when σ_X^2 and σ_Y^2 are both NOT known but are assumed EQUAL

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

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 - Use when σ_X^2 and σ_Y^2 are both NOT known and NOT EQUAL

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

Confidence Coefficient (“Multiplier”) –

There are 3 solutions, depending on the standard error calculation (as per above).

Solution 1 - Use this when σ_X^2 and σ_Y^2 are both known

Use percentile of Normal(0,1)

Solution 2 - Use when σ_X^2 and σ_Y^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 - Use when σ_X^2 and σ_Y^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!

Here's a summary table ...

Normal Distribution: Confidence Interval for [$\mu_1 - \mu_2$] (Two Independent Groups) CI = [point estimate] \pm (conf.coeff)SE[point estimate]			
	σ_X^2 and σ_Y^2 are both known	σ_X^2 and σ_Y^2 are both NOT known but are assumed EQUAL	σ_X^2 and σ_Y^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 (according to what is known or can or cannot be assumed about the separate population variances).

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

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

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

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

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

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

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

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

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

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.

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

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

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

5b. Confidence Interval for σ_1^2 / σ_2^2

Of interest now is a comparison of the **two variances**; eg - .

- We might want to know if the reproducibilities of two laboratory assays are similar.
- More simply, we might be interested in an exploration of two independent normal population distributions; this would include a comparison of their two variance parameters.
- Sometimes, we are interested in a comparison of two variance parameters as a preliminary step in a larger analysis plan for the reason that some statistical analysis techniques make assumptions about equality of variances.

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

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 ($X_1 \dots X_{n_x}$) is comprised of $n_x=21$ healthy adults. The other group ($Y_1 \dots Y_{n_y}$) includes $n_y = 16$ Parkinson's disease patients. Interest is in a 95% confidence interval estimate of σ_x^2 / σ_y^2 . Preliminary calculations yield the following statistics:

$$S_X^2 = \frac{\sum_{i=1}^{n_x} (X_i - \bar{X})^2}{n_x - 1} = 1600 \quad \text{and} \quad S_Y^2 = \frac{\sum_{i=1}^{n_y} (y_i - \bar{y})^2}{n_y - 1} = 1225$$

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

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

Step 1 – Solution for Point Estimator S_x^2 / S_y^2

$$S_x^2 / S_y^2 = 1600 / 1225 = 1.306$$

Step 2 – Solution for Confidence Coefficient Multipliers

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

$$\left(\frac{1}{F_{n_1-1; n_2-1; (\alpha/2)}} \right) = \left(\frac{1}{F_{20; 15; .025}} \right) = 2.57$$

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

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

Upper Limit Value =

$$\left(\frac{1}{F_{n_1-1; n_2-1; (\alpha/2)}} \right) \left[\frac{S_1^2}{S_2^2} \right] = \left(F_{n_2-1; n_1-1; (1-\alpha/2)} \right) \left[\frac{S_1^2}{S_2^2} \right] = (2.57) [1.306] = 3.36$$

6. Binomial Distribution: One Group

6a. Confidence Interval for π

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

- The setting is results of N independent trials, each of which produces two possible outcomes; we called these “event” and “non-event”.
- “Event/non-event” might refer to: “alive/dead”, “tumor/remission”, “success/failure”, “heads/tails”, etc.
- Associated with each trial is the same probability of event occurrence, π
- An estimate of the probability of event occurrence π is given by the observed proportion of the N trials that yielded event occurrence.
- The binomial distribution is the probability model used to describe the outcome of $X=x$ “events” among the N independent trials.

(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 is p for “proportion”. This is awkward because sometimes the notation “ p ” is used for the population parameter π itself. Therefore, I recommend against using this to represent the estimate of π .
- Better is the use of \hat{p} because it has the caret on top. This is most likely to be used when the writers of the text you are reading refer to the population parameter π as p .
- Another is X/N . This makes sense since you can discern from this that it is referring to an observed proportion.
- Still another is \bar{X} . This also makes sense since it is the sum of 0’s and 1’s, divided by N , the number of trials. *Putting these all together ...*
- **Estimate of π notations:** $\hat{\pi} = \hat{p} = \bar{X} = X/N$ 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, Fourth Edition*, 1987 page 149)

Interest is in estimating the proportion of individuals who obtain a dental check up twice a year in a certain urban population. Of $N=300$ persons identified by simple random sampling and interviewed, $X=123$ 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 Variance(X)= $N\pi(1-\pi)$
- Variance[(constant) X] = (constant)² 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**

This may seem counterintuitive but it is not. It is not correct that, because the SE has to be estimated, that the percentile is Student's t. 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 = (point estimate) - (multiple) (SE of estimate) = $0.41 - (1.96)(0.028) = 0.36$

Upper = (point estimate) + (multiple) (SE of estimate) = $0.41 + (1.96)(0.028) = 0.46$

Confidence Interval for a proportion π - a 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 the largest value it can have is the one for which $\bar{X} = 0.50$ in the SE calculation

7. Binomial Distribution: Two Independent Groups

7a. Confidence Interval for $[\pi_1 - \pi_2]$

We are often interested in comparing proportions from 2 populations:

- Is the incidence of disease A the same in two populations?
- Patients are treated with either drug D, or with placebo. Is the proportion “improved” 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. Compute 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 $S\hat{E}(\hat{\pi}_1 - \hat{\pi}_2) = \sqrt{\frac{\bar{X}(1-\bar{X})}{N_1} + \frac{\bar{Y}(1-\bar{Y})}{N_2}}$

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

$$S\hat{E}(\hat{\pi}_1 - \hat{\pi}_2) = \sqrt{\frac{\bar{X}(1-\bar{X})}{N_1} + \frac{\bar{Y}(1-\bar{Y})}{N_2}} = \sqrt{\frac{.68(.32)}{50} + \frac{.30(.70)}{50}} = 0.0925$$

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 Normal σ^2 known

The setting is the example in Section 3a (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}

$$\begin{aligned}\bar{X} &= 0.51 \\ n &= 30\end{aligned}$$

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

$$\begin{aligned}SE(\bar{X}) &= \sigma / \sqrt{n} = 0.5 / \sqrt{30} \\ &= 0.0913\end{aligned}$$

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

Suppose we want a 95% confidence interval.

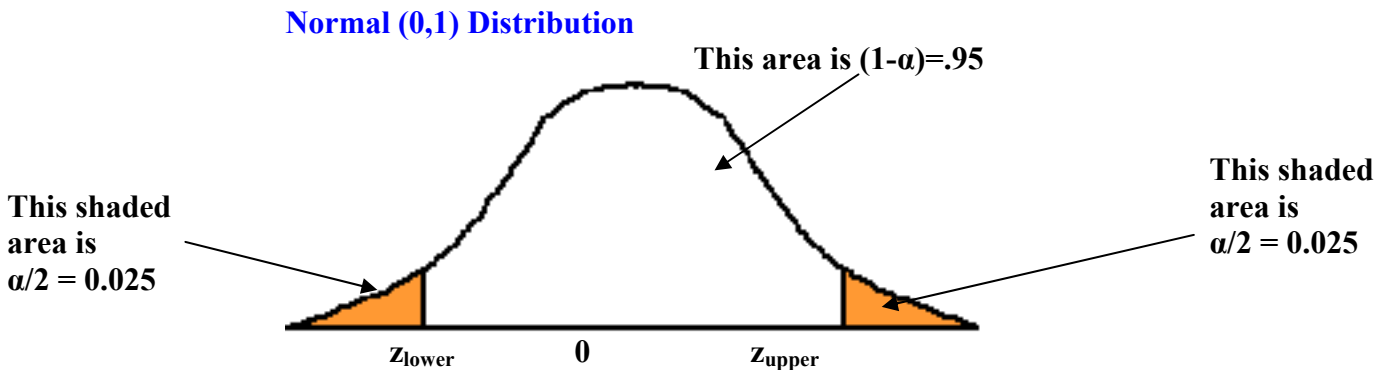
Then $(1 - \alpha) = 0.95$.

This means that $\alpha = 0.05$

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

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

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



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

$$z_{lower} = -1.96$$

$$z_{upper} = +1.96$$

This expression, $\text{Probability} [-1.96 \leq Z \leq +1.96] = 0.95$ in this example and $\text{Probability} [z_{lower} \leq Z \leq z_{upper}] = (1 - \alpha)$ more generally is the origin of the formula for a confidence interval. To arrive at the latter involves insertion of the standardization of \bar{X}

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

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

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

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

So let's call z_{upper} simply z

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

$$z_{upper} = z$$

$$z_{lower} = -z$$

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

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

ii. Derivation of Confidence Interval for σ^2 Single Normal

The setting here is the example in Section 3c.

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 .

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 = $(1 - \alpha)$

As we want a 95% confidence interval, $(1 - \alpha) = 0.95$.

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

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

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

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

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

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

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

We take advantage of two statistical results -

- If X is distributed Binomial(N,π) Then Variance(X)=Nπ(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

$$\approx \sqrt{\frac{\bar{X}[1-\bar{X}]}{N}} \text{ which matches.}$$