Unit 8
Statistical Literacy – Estimation and Hypothesis Testing

“Who would not say that the glosses (commentaries on the law) increase doubt and ignorance? It is more of a business to interpret the interpretations than to interpret the things”

- Michel De Montaigne (1533-1592)

“A hypothesis is a contention that may or may not be true, but is provisionally assumed to be true until new evidence suggests otherwise. A hypothesis may be proposed from a hunch, from a guess, or on the basis of preliminary observations. A statistical hypothesis is a contention about a population, and we investigate it by performing a study on a sample collected from that population. We then examine the sample information to see how consistent the data are with the hypothesis under question; if there are discrepancies, we tend to disbelieve the hypothesis and reject it. So the question arises, how inconsistent with the hypothesis do the sample data have to be before we are prepared to reject the statistical hypothesis? It is to answer questions such as this that we use statistical tests of hypotheses, or significance tests.”

Source: Elston RC and Johnson WD. Essentials of Biostatistics. FA Davis Company. 1987. page 126
Cheers!

“THE P VALUE WAS NEVER MEANT TO BE USED THE WAY IT’S USED TODAY.”

## Table of Contents

<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Unit Roadmap</td>
<td>4</td>
</tr>
<tr>
<td>2. Learning Objectives</td>
<td>5</td>
</tr>
<tr>
<td>3. Introduction to Hypothesis Testing</td>
<td>6</td>
</tr>
<tr>
<td>3.1 The Basics</td>
<td>6</td>
</tr>
<tr>
<td>3.2 One Sided versus Two Sided Tests</td>
<td>19</td>
</tr>
<tr>
<td>3.3. Beware!!</td>
<td>20</td>
</tr>
<tr>
<td>3.4. Introduction to Type I and II Error and Statistical Power</td>
<td>23</td>
</tr>
<tr>
<td>4. Introduction to Confidence Interval Estimation</td>
<td>29</td>
</tr>
<tr>
<td>4.1 Goals of Estimation</td>
<td>29</td>
</tr>
<tr>
<td>4.2 Notation and Definitions</td>
<td>33</td>
</tr>
<tr>
<td>4.3. How to Interpret a Confidence Interval</td>
<td>36</td>
</tr>
<tr>
<td>5. Preliminaries: Some Useful Probability Distributions</td>
<td>43</td>
</tr>
<tr>
<td>5.1 Student t-Distribution</td>
<td>43</td>
</tr>
<tr>
<td>5.2 Chi Square Distribution</td>
<td>47</td>
</tr>
<tr>
<td>5.3. F-Distribution</td>
<td>51</td>
</tr>
<tr>
<td>5.4. Sums and Differences of Independent Normal Random Variables</td>
<td>54</td>
</tr>
</tbody>
</table>
Statistical significance assessment is a tool that informs our understanding of nature but does not establish biological significance one way or the other.

The logic of statistical hypothesis testing is a “proof by contradiction” argument that proceeds as follows:

**Step 1** – Begin with the “skeptic’s” perspective. Define a “chance” model. This is the null hypothesis.

**Step 2** – Assume that the null hypothesis model is true.

**Step 3** – Apply the null hypothesis model to the data. Show (or not show) that the null hypothesis model, when applied to the given data, leads to an unlikely conclusion. Important point – the hypotheses are up for debate, but the data are not. The data are “givens”.

**Step 4** – State the statistical inference:

- If the conclusion is unlikely, The null hypothesis model is rejected.
- If the conclusion is not unlikely, The null hypothesis model is NOT rejected.

**Step 5** – Proceed onward to the next step in inference making. You’re not done yet!
2. Learning Objectives

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

- Explain the logic of statistical hypothesis testing;
- Translate the statement of a research question into a testable hypothesis; and specifically,
  - Translate the statement of a research question into its associated null (H₀) and alternative (Hₐ) hypotheses.
- For a given data situation, define and explain the null hypothesis model.
- Explain the steps in performing a statistical hypothesis test.
- Explain the meaning of a p-value.
- Interpret the value of a p-value with respect to rejection or non-rejection of a null hypothesis.
- Interpret p-values in publications.
- Explain the utility of accompanying statistical hypothesis tests with confidence intervals.
- Explain Type I and II error and the meaning of statistical power
- 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.
- Interpret a confidence interval.
3. The Logic of Hypothesis testing

3.1 The Basics


Chapter 14 is titled “The probability of a dry toothbrush: what is a p-value anyway?” In it, on page 59, he provides a little box, “Things to Remember”. It captures the logic of hypothesis testing very nicely. So here it is, in its entirety.

Quoting …

```
“Things to Remember”

1. Inference statistics involves testing a hypothesis, specifically, a null hypothesis.

2. A null hypothesis is a statement suggesting that nothing interesting is going on, for example, that there is no difference between the observed data and what was expected, or no difference between two groups.

3. The p-value is the probability that the data would be at least as extreme as those observed if the null hypothesis were true.

4. If the data would be unlikely if the null hypothesis were true, we conclude that the null hypothesis is not true.

5. My son has now worked out my trick and has taken to running his toothbrush under the tap for a second or two before heading to bed.

```
A little more detail on Andrew Vickers’ “Things to Remember” reveals the logic of hypothesis testing.

1. “Inference involves testing a hypothesis..”

It’s all about perspective (already mentioned on page 3). It is the hypotheses, not the data, that are abandoned or retained. The data themselves are “givens”, “non-negotiable.” Take care not to make statements such as the following: “the data are inconsistent with a hypothesis”. Instead, make statements such as the following: “the hypothesis is not consistent with the data” or “the hypothesis is consistent with the data”.

2. “A null hypothesis is a statement suggesting that nothing interesting is going on…”

As you’ll see in the pages that follow, statistical hypothesis testing makes use of two kinds of hypotheses: null and alternative.

With some important exceptions (described later), often, it is the alternative hypothesis that the investigator hopes to advance. The interesting hypothesis! Examples of alternative hypotheses are the following: (1) the new drug is beneficial and significantly more so than the old drug; (2) the observations of ill health are associated with some exposure; (3) the prevalence of injection drug use has declined in the past 5 years. And so on.

And, as Andrew Vickers expressed it, the null hypothesis is the “nothing is going on” hypothesis; eg (1) the benefits accompanying administration of the new drug are no different than what occur with the old drug; (2) the observations of ill health are unrelated to the suspected exposure; (3) the prevalence of injection drug use is the same as what it was 5 years ago. And so on.

3. “The p-value is the probability that the data would be at least as extreme as those observed if the null hypothesis were true”

An important point to remember is this. We start by assuming that the null hypothesis is true. More specifically, we start by assuming that the given data are a sample from some null hypothesis model probability distribution. This is the chance model! For example, you might assume that your observed set of n=25 IQ test scores are a simple random sample from a normal distribution with mean µ=100.

A p-value number (such as .05) is probability calculation. It’s not really a calculation that the “data would be at least as extreme as those observed”...It is the calculation that “some statistic would be at least as extreme as that observed”. The statistic might be the sample mean. Example, continued. If your null hypothesis is that your sample of n=25 IQ test scores are a simple random sample from a normal distribution with mean µ=100 and your observed sample mean is X̄=81 then a one sided p-value might be the calculation of Pr[X ≤ 81] under the null hypothesis assumption that µ=100.
4. “If the data would be unlikely if the null hypothesis were true, we conclude that the null hypothesis is not true.”

The logic of statistical hypothesis testing is a proof by contradiction argument! Having first assumed that the null hypothesis true, you then examine where this takes you in light of the observed data. Does the null hypothesis, in light of the data, take you to an unlikely conclusion (a small p-value)? If so, the data and the null hypothesis are inconsistent. Since the data can’t be wrong, the problem must be with the null hypothesis. The null hypothesis is then rejected.

Example, continued. Suppose that \( \Pr[\bar{X} \leq 81] = 0.02 \), under the null hypothesis assumption that \( \mu = 100 \). This is saying the following: The chances were 2 in 100 of obtaining an average IQ of 81 or lower if \( \mu = 100 \). These are small chances! So small that we will reject the null hypothesis and conclude that the mean IQ for the population that gave rise to this sample is some figure that is lower than the null hypothesis model values of 100.

5. “My son has now worked out my trick and has taken to running his toothbrush under the tap for a second or two before heading to bed.”

Andrew Vickers’ son has figured out an “end run” around his statistician dad’s reasoning: “the toothbrush is dry; it is unlikely that the toothbrush would be dry if my son had cleaned his teeth; therefore, he hasn’t cleaned his teeth” Source: Vickers, A. What is a p-value Anyway? 34 Stories to Help You Actually Understand Statistics. Addison Wesley, 2010. page 58

Hence the quick rinse under the tap….

As we will see, statistical inference is not biological inference.
Steps in Hypothesis Testing

1. Identify the research question.

2. State the null hypothesis assumptions necessary for computing probabilities.

3. Specify $H_0$ and $H_A$.

4. “Reason” an appropriate test statistic.

5. Specify an “evaluation” rule.

6. Perform the calculations.

7. “Evaluate” findings and report.

8. Interpret in the context of biological relevance.

9. Compute an appropriate confidence interval estimate.
Schematic of Statistical Hypothesis Testing

In each picture below, the data is summarized using $\bar{X}$. Above it are two probability models that might have given rise to the data. One is the true probability distribution and $\bar{X}$ is located at a value near the center. The other is the null hypothesis probability model.

**Top Picture.**
In the top picture, the two probability distributions (null & true) are essentially the same. The sample mean is close to its “true” expected value (a population mean).

**Take home** – The null hypothesis is consistent with the data (the sample mean). We have no contradiction.

**Lower Picture.**
In the lower picture, the null hypothesis probability model is the left hand curve. The true distribution is the right hand curve. Here, too, the sample mean is close to its “true” expected value (a population mean).

**Take home** – The null hypothesis is *NOT* consistent with the data. We say this because the sample mean is far away from the null hypothesis model mean. This inconsistency or contradiction leads us to reject the null hypothesis model given the data.
A step-by-step schematic of how “proof by contradiction” and the rejection of the null works.

**Top Picture.**
**Step 1 – Begin by assuming that the null hypothesis is true**
Under the null hypothesis assumption model, the two probability distributions (“true” and “null hypothesis”) are identical or very nearly the same. This is why the two curves are right on top of each other. **Note** – We have not yet taken into consideration the given observed data. This comes next (middle picture).

**Middle Picture.**
**Step 2 – Consider the observed sample mean.**
This picture represents the given data. Notice that there is no probability distribution shown. **Remember** - we don’t actually know which distribution gave rise to the data.

**Bottom Picture.**
**Step 3 – Argue “yes” or “no” consistency of the null hypothesis model assumption with the data.**
In this picture, the “true” distribution that gave rise to the data is on the right. The null hypothesis assumption model is on the left. The shaded area is a probability calculation under the assumption that the null is true:

\[
\Pr \left[ \bar{X} \geq \text{observed} \mid \text{assuming null model} \right].
\]

It answers the question “*Under the assumption of the null hypothesis, what are the chances of a value of the sample mean as extreme, or more, than was observed?”*

Small probability says “Assuming the null led to an unlikely event”
Large probability says “Assuming the null led to a likely event”
A closer look at \( \Pr [ \bar{X} \geq \text{observed value} | \text{assumed null hypothesis model is true}] \)

### Scenario 1 - NULL is true
\[
\Pr [ \bar{X} \geq \text{observed value} | \text{assuming null is true}] = \text{large}
\]
- The sample mean is close to its “true” expected value, and this is also close to the null hypothesis model expected value.
- The null hypothesis model probability that the sample mean \( \bar{X} \) is as far away from the null hypothesis mean, or more extreme, is a large probability (large shaded area). That is - assuming the null hypothesis model leads to a likely outcome.
- Statistical decision - “do NOT reject the null”.

### Scenario 2 - ALTERNATIVE is true
\[
\Pr [ \bar{X} \geq \text{observed value} | \text{assuming null is true}] = \text{small}
\]
- Observed sample mean is not close to null mean.
- The null hypothesis model probability that the sample mean \( \bar{X} \) is as far away from the null hypothesis mean, or more extreme, is a small probability (small shaded area). That is - assuming the null hypothesis model leads to an unlikely outcome.
- Statistical decision - “REJECT the null”.

\[ \text{p-value} = \Pr [ \text{Test statistic (eg } \bar{X} \text{)} = \text{observed or more extreme} | \text{assuming null is true}] \]

EG - “If I assume that the null hypothesis is true and use this model, what was my probability of obtaining \( \bar{X} \) as far away from the null hypothesis expectation, or more so, than the value that I observed?”

The same thing: “p-value” “significance level” “achieved significance”.

| Nature | Population/ Sample | Observation/ Data | Relationships/ Modeling | Analysis/ Synthesis |
Illustration.

Suppose that, with standard care, cancer patients are expected to survive a mean duration of time equal to 38.3 months. Investigators are hopeful that a new therapy will improve survival.

A new therapy is administered to a sample of 100 cancer patients. The sample average survival time is obtained and it is equal to 46.9 months. Is the sample average of 46.9 months sufficiently unlikely, relative to the null hypothesis expected value of 38.3 months, to warrant abandoning the null hypothesis in favor of the alternative conclusion, namely: improved survival?

This illustration follows the steps outlined on page 9.

1. Identify the research question

With standard care, the expected survival time is \( \mu = 38.3 \) months. With the new therapy, the observed 100 survival times, \( X_1, X_2, \ldots, X_{100} \) have average \( \bar{X}_{100} = 46.9 \) months. Is this compelling evidence that \( \mu_{\text{true}} > 38.3 \) ?

Invoke the null hypothesis model assumption. In particular, state the corresponding null hypothesis probability model value of the mean. This will be used when computing the p-value probability value (chances of extremeness)

For now, we’ll assume that the 100 survival times follow a distribution that is Normal (Gaussian). We’ll suppose further that it is known that \( \sigma^2 = 43.3^2 \) months\(^2\). Note – In real life, this would not be a very reasonable assumption as survival distributions tend to be quite skewed. Normality is assumed here, and only for illustration purposes, so as to keep the example simple.

2. Specify the null and alternative hypotheses

\[ H_0: \mu_{\text{true}} = \mu_O \leq 38.3 \text{ months} \]
\[ H_A: \mu_{\text{true}} = \mu_A > 38.3 \text{ months} \]

Note – Strictly speaking, the null and alternative hypotheses must cover all possibilities. That’s why they are written as you see them here. In calculating the p-value, however, we must choose a single null hypothesis value of the mean. We choose the value that is at the “border” between the null and alternative possibilities. This gives us the most “conservative” calculation of the p-value. Thus, we will use \( \mu_O = 38.3 \) in step #5 on the next page.
3. **Reason “proof by contradiction”**

IF: the null hypothesis model is true so that the expected average survival time was $\mu_{true} = \mu_O = 38.3$

THEN: what were the chances of obtaining a sample average survival time as far away from 38.3, or more so in the direction of longer survival, than what was actually observed, namely 46.9?

4. **Specify a “proof by contradiction” rule.**

Statistically, assuming the null hypothesis in light of the observed data leads to an unlikely conclusion (translation: small p-value) if there is at most a small chance that the mean of 100 survival times is 46.9 or greater when its expected value is 38.3. We calculate the value of such chances as

$$\Pr\left[\bar{X}_{n=100} \geq 46.9 \mid \mu_{true} = \mu_O = 38.3\right]$$

Reminder - The vertical bar is a shorthand for saying that we are doing this calculation under the assumption that the mean is 38.3

5. **Calculate the null hypothesis model ($H_0$) chances of “extremeness.” This value will be the p-value.**

Under the null hypothesis model:

$$X_1, X_2, \ldots, X_{100} \text{ is a simple random sample from a Normal}(\mu = 38.3, \sigma^2 = 43.3^2).$$

This, in turn says that under the assumption that the null hypothesis is true:

$$\bar{X}_{n=100} \text{ is distributed Normal}(\mu = 38.3, \sigma^2 = 43.3^2/(n = 100))$$
How extreme is “extreme” is an example of “signal-to-noise”.

<table>
<thead>
<tr>
<th>Signal</th>
<th>“46.9 is 8.6 months away from 38.3”</th>
<th>(46.9 – 38.3) = 8.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is 8.6 extreme or not?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Noise</th>
<th>Noise is the scatter/variability of the average. We measure this using the SE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SE((\bar{X}_{n=100})) = (\frac{\sigma}{\sqrt{100}}) = (\frac{43.3}{10}) = 4.33</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Signal-to-Noise (Z-score)</th>
<th>Signal, in units of months, has been re-expressed in units of noise (SE units)</th>
</tr>
</thead>
<tbody>
<tr>
<td>“46.9 is 1.99 SE units away from 38.3”</td>
<td></td>
</tr>
</tbody>
</table>

\[
Z\text{-score} = \frac{(\bar{X}_{n=100} - \mu_{\bar{X} \text{ under NULL}})}{SE(\bar{X}_{n=100})} = \frac{(46.9-38.3)}{SE(\bar{X}_{n=100})} = \frac{8.6 \text{ months}}{4.33 \text{ months}} = 1.99 \text{ SE units}
\]

Z-score=1.99 says: “The observed mean of 46.9 is 1.99 SE units away from the null hypothesis expected value of 38.3”
**Logic of Proof-by-Contradiction says:**

“Under the assumption that the null hypothesis is true, there are 2 in 100 chances of obtaining a sample mean as "extremely" far away from 38.3 as the value of 46.9”

\[
\Pr\left[ \bar{X}_{n=100} \geq 46.9 \mid \mu_{true} = \mu_{null} = 38.3 \right]
\]

=Pr[Z-score ≥ 1.99]=.02

**Statistical Reasoning of “likely” says:**

“If the null hypothesis, when examined in light of the data, leads us to something that is ‘unlikely’, namely a small p-value (shaded area in blue below), then the null hypothesis is severely challenged and possibly contradicted. ➔

Statistical rejection of the null hypothesis.

**Graphical illustration of a p-value calculation**

http://davidmlane.com/hyperstat/z_table.html
The Z-score is a Signal-to-Noise Comparison

\[
Z\text{-score} = \frac{\text{Signal}}{\text{Noise}} = \frac{\text{observed-expected}}{\text{SE(observed)}}
\]

\[
= \frac{\bar{X}_{n=100} - \mu}{\sigma \sqrt{n}}
\]

Z-Score =

The magnitude of the departure, from the null hypothesis expectation, of the observed sample statistic (in this case the sample mean), expressed on the scale of SE units.

Example: \( z\text{-score}=1.99 \)

\[
p\text{-value} = \Pr[\text{Normal}(0,1) \geq z\text{-score}]
\]

p-value =

The chances of obtaining a departure of this magnitude, or greater, calculated under the presumption that the null hypothesis is true.

Example: \( \Pr[\text{normal}(0,1) \geq 1.99] = .02 \)
**TIP!**

The Z-score is a very handy measure of extremeness

*ps – and so is the T-score, as we shall see later in these notes!*

---

**Example –**

Imagine you are reading a manuscript and you see a sample mean (e.g., average treatment response) and its SE. You wonder if this average treatment response is “compellingly” different from a null hypothesis of “no benefit”. You can get an idea of this! To do so, you re-express the reported sample mean as a z-score.

* The chances of a z-score having value greater than 2.5 SE units away from its expected value of 0 *in either direction* is a small likelihood, namely a 1% likelihood in both tails or a 0.5% likelihood in each of the left and right tails (0.5% + 0.5%).

* Translation: – The assumption of the null hypothesis (“no treatment benefit”) has led to an unlikely outcome, because the chances of being 2.5 SE units distant from “not benefit” were 1% 0.5% + 0.5% = 1%.
3. The Logic of Hypothesis testing

3.2 One Sided versus Two Sided Tests

**One Sided** - In the example above, the investigators sought to assess whether the new treatment might be associated with an *improvement* in survival. The key word here is “*improvement*”. This is an example of a one sided test because the alternative hypothesis probability models are to one side of the null hypothesis model:

\[ H_A : \mu_A > 38.3 \text{ months} \]

**Two Sided** – What if, instead, the investigators had wished only to assess whether the new treatment is associated with a *different* survival? Here the key word is “*different*”. This would have been an example of a two sided test because the alternative hypothesis probability models are on either side of the null hypothesis model:

\[ H_A : \mu_A \neq 38.3 \text{ months} \]

**P-value calculations in two sided tests** consider “extremeness” in two directions (both “tails”).

**Step 1** – Obtain Z-score measure of “signal”

\[
Z - \text{score} = \frac{{(\bar{X}_{n=100} - \mu_{\text{NULL}})}}{{SE(\bar{X}_{n=100})}} = \frac{{(46.9 - 38.3)}}{{SE(\bar{X}_{n=100})}} = \frac{{8.6 \text{ months}}}{{4.33 \text{ months}}} = 1.99
\]

**Interpretation:** The observed mean is 1.99 SE units away (to the right) of the null

**Step 2** – Calculate p-value = Probability of “extremeness” to the right and to the left.

\[
p-value_{\text{TWO SIDED}} = \text{Prob}[\text{Normal}(0,1) \geq +1.99] + \text{Prob}[\text{Normal}(0,1) \leq -1.99]
\]

\[
= 2 \times \text{Prob}[\text{Normal}(0,1) > 1.99]
\]

\[
= 2 \times (0.02)
\]

\[
= 0.04
\]

**Interpretation:** Under the assumption that the mean survival is the null value of 38.3 months, the probability of an average survival being different by 8.6 months in *either direction* (more or less) is 4 chances in 100.
3. The Logic of Hypothesis testing

3.3 Beware

Yup. Beware:

1. Statistical significance is not biological inference
2. An isolated p-value communicates limited information only
3. Other criteria are essential to biological inference.

1. Statistical Significance is NOT Biological Inference.

To appreciate this, suppose that, upon completion of a statistical hypothesis test, you find that:

Results for patients receiving treatment “A” are \textit{statistically significantly} better than results for patients receiving treatment “B”.

There are actually multiple, different, explanations:

- \textit{Explanation #1} - Treatment “A” is truly superior.

- \textit{Explanation #2} - Groups “A” and “B” were not comparable to begin with. The apparent finding of superiority of “A” is an artifact. The nature of the “artifact” has to do with concepts of confounding that you are learning in your epidemiology courses.

- \textit{Explanation #3} – An event of low probability has occurred. Treatment “B” is actually superior but sampling (\textit{as it will occasionally do!}) yielded a rare outcome. (\textit{Consider this - Events of low probability do occur sometimes, just not very often}).
2. **P-values, by themselves, don’t help us much.**

**Definition p-value**

There are a variety of wordings of the meaning of a p-value; e.g -

- **Source: Fisher and van Belle.** “The null hypothesis value of the parameter is used to calculate the probability of the observed value of the statistic or an observation more extreme.”

- **Source: Kleinbaum, Kupper and Muller.** “The p-value gives the probability of obtaining a value of the test statistic that is at least as unfavorable to $H_0$ as the observed value”

- **Source: Bailar and Mosteller.** “P-values are used to assess the degree of dissimilarity between two or more sets of measurement or between one set of measurements and a standard. A p-value is actually a probability, usually the probability of obtaining a result as extreme or more extreme than the one observed if the dissimilarity is entirely due to variation in measurements or in subject response – that is if it is the result of chance alone.”

- **Source: Freedman, Pisani, and Purves.** “The observed significance level is the chance of getting a test statistic value as extreme or more extreme than the observed one. The chance is computed on the basis that the null hypothesis is right. The smaller this chance is, the stronger the evidence against the null. … At this point, the logic of the test can be seen more clearly. It is an argument by contradiction, designed to show that the null hypothesis will lead to an absurd conclusion and must therefore be rejected.”
Beware!

- The p-value is **NOT** the probability of the null hypothesis being correct.
- The p-value is **NOT** the probability of obtaining the observed data “by chance”.
- The p-value is **NOT** the probability of the observed data itself calculated under the assumption of the null hypothesis being correct.

*Source: Rothman and Greenland.* A p-value is **NOT** “the probability that the data would show as strong an association as observed or stronger if the null hypothesis were correct”.

3. **Other criteria are essential to biological inference.**

- A conclusion of a treatment effect is *strengthened* by
  - A dose-response relationship
  - Existence in sub-groups as well as existence overall
  - Epidemiological evidence
  - Consistency with findings of independent trials.
  - Its observation in a large scale (meaning large sample size) trial

- A conclusion of a treatment effect is *weakened* by
  - Its unusualness; such a finding should be “checked” with new data
  - Its isolation; that is – it is observed in a selected subgroup only and nowhere else; such a finding is intriguing, however and should be explored further
  - Its emergence as a unique finding among many examinations of the data.
3. The Logic of Hypothesis testing

3.4 Introduction to Type I and II Error and Statistical Power

A statistical hypothesis test uses probabilities based only on the null hypothesis (H₀) model!

- Our starting point is the assumption of the null hypothesis model; that is we presume that H₀ is true. We then used this model to estimate the likelihood of what we actually observed, namely our test statistic value, or something more extreme. Depending on how big this likelihood is:
  - It’s all about the presumed null hypothesis model. We either abandon (reject) the null hypothesis (H₀) model, or we retain it (fail to reject).
  - **We do not** prove that the null hypothesis assumption is correct.

So ….. how did we do? Either we draw the correct inference or the wrong inference:

<table>
<thead>
<tr>
<th>Decision</th>
<th>Retain the null</th>
<th>Reject the null</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null true</td>
<td>☺</td>
<td>α = type I error</td>
</tr>
<tr>
<td>Alternative true</td>
<td>β = type II error</td>
<td>☺</td>
</tr>
</tbody>
</table>

**Introduction to Type I Error**
- IF H₀ is true and we (incorrectly) reject H₀
  - We have made a type I error
  - We can calculate its probability as Pr [type I error]= α

**Introduction to Type II Error**
- IF Hₐ is true and we (incorrectly) fail to reject H₀
  - We have made a type II error
  - We must have a specific Hₐ model before we can calculate
    Pr [type II error]= β

**Introduction to Power**
- IF Hₐ is true and we (correctly) reject H₀
  - This occurs with probability = \(1 - β\) which we call the “POWER”
The goal is to get the right answer (power). Either type of error is undesirable. We’d like to minimize the chances of either a type I error (probability = $\alpha$) or a type II error (probability = $\beta$).

- **Sample size calculations!** Larger sample sizes will lower both probabilities: $\alpha$ and $\beta$
- All other things being equal, a larger sample size increases power (the probability of drawing the correct inference)
- There are other factors that influence the power of a study, too.

The techniques of sample size and power calculations are not addressed in this course.

### Statistical Power is the Likelihood of inferring a specified benefit ($H_A$). This is the blue area

#### Key:

- **Blue ribbon along the horizontal axis with “reject $H_0$” typed inside:** These are the values of the sample average that will prompt rejection of the null hypothesis, also called the **critical region**. Note – “Critical regions” and “critical region tests” are introduced and explained beginning on page 31

- **Blue area under the Null ($H_0$) curve:** The type I error (probability = $\alpha$). This is the probability of mistakenly rejecting the null hypothesis; thus, it is calculated under the assumption that $H_0$ is true.

- **White area under the Alternative ($H_A$) curve:** The type II error (probability = $\beta$). This is the probability of mistakenly inferring the null; thus it is calculated under the assumption that $H_A$ is true.
The Power of a Study Depends on Four Parameters
1. Type I Error

Take home message: If you’re willing to live with more type I error, then you can increase your chances of inferring a treatment benefit (the alternative).

- In this picture, the null and alternative distributions in the top panel are the same as the null and alternative distributions in the bottom panel.

- In the top panel, rejection of the null hypothesis occurs when the p-value calculation is any value smaller than or equal to 0.005. Whereas, in the bottom panel, rejection of the null hypothesis occurs when the p-value calculation is any value smaller than or equal to 0.05.

- Thus, all other things being equal, use of a smaller p-value criterion (e.g. 0.005 versus 0.05) reduces the power to detect a true alternative explanation (the blue area in the top panel is smaller than the blue area in the bottom panel).
Take home message: The farther away the alternative (treatment benefit) is from the null hypothesis, the greater your chances of inferring it. Conversely, the closer the alternative (treatment benefit) is to the null hypothesis, the lower your chances of inferring it.

- In this picture, the null hypothesis is the same in the top and bottom panels.
- However, the alternative is closer to the null in the top panel and more distant from the null in the bottom panel.
- The “threshold” value of the sample mean that prompts rejection of the null hypothesis is the SAME in both top and bottom panels.
- Here, all other things being equal, alternative hypotheses that are farther away from the null are easier (power is greater) to detect (larger blue area under the curve in the bottom panel) than are alternative hypotheses that are closer to the null (smaller blue area under the curve in the top panel).
The Power of a Study Depends on Four Parameters

3. Biological Variability ("Noise")

Take home message: The more precisely you can measure the outcomes of interest, the greater your chances of inferring a treatment benefit (the alternative); this is because the standard error is smaller.

- In this picture, the null hypothesis is the same in the top and bottom panels. As well, the alternative hypothesis is the same in the top and bottom panels.

- The distinction is that the underlying variability of the outcomes (a combination of naturally occurring biological variability and measurement error) is smaller in the bottom panel.

- The “threshold” value of the sample mean that prompts rejection of the null hypothesis is the SAME in both top and bottom panels.

- Here, all other things being equal, using a measurement tool that is less noisy (more precise) will increase study power (the blue area under the curve).
The Power of a Study Depends on Four Parameters

4. Sample Size (“Design”)

Take home message: The larger the sample size you use, the greater your chances of inferring a treatment benefit (the alternative); this is again because the standard error is smaller.

- In this picture, the null hypothesis is the same in the top and bottom panels. As well, the alternative hypothesis is the same in the top and bottom panels.

- In this picture, too, the underlying variability of the outcomes (a combination of naturally occurring biological variability and measurement error) is the same in the two panels.

- However, the sample size N is larger in the bottom panel. The result is that the SE of the sample mean (SE(\(\bar{X}\))=\(\sigma/\sqrt{n}\)) has a smaller value (by virtue of division in the denominator by a larger square root of n).

- Here, all other things being equal, using a larger sample size will increase study power (the blue area under the curve).
4. Introduction to Confidence Interval Estimation

4.1 Goals of Estimation

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

Suppose we have a simple random sample of $n$ observations $X_1 \ldots X_n$ 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. \( \mu_{I} \) and \( \mu_{IV} \) are outside the interval around \( \bar{X} \). \( \mu_{II} \) and \( \mu_{III} \) are inside the interval.

I.

II.

III.

IV.

\( \bar{X} \)  
Confidence Interval

Nature  Populations/Samples  Observation/Data  Relationships/Modeling  Analysis/Synthesis
We are “confident” that $\mu$ could be either $\mu_{II}$ or $\mu_{III}$.

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 6 – 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 $\sigma^2$.

“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” $\sigma^2$.

\[
\sum_{\text{all possible samples}^\prime} \left( \frac{S^2_i}{\#\text{samples in sampling distn}} \right) = \sigma^2
\]

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

| Nature         | Population/Sample | Observation/Data | Relationships/Modeling | Analysis/Synthesis |
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 \( \mu \) 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 \( \mu \). 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 \( \mu \).

**Take home message:**

Here, we will be using the criteria of “minimum variance unbiased”. However, other criteria are possible.
### 4. Introduction to Confidence Interval Estimation

#### 4.2 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:

  \[ \text{[ Value ]}, \]

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

  \[ \text{[ lower limit, upper limit ]} \]

**Example** -

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

ค่♣♣ A confidence interval is a particular type of interval estimator.(161,250),(414,272)

♣ 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.
4. Introduction to Confidence Interval Estimation

4.3 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, $X_1, \ldots, X_{5000}$. The average income that we want to estimate is the population mean $\mu$.

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

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 $\mu$ 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 $X_1, \ldots, X_{10}$

\[
\bar{X}_{n=10} = 19,887 \\
\sigma = 12,573 \\
SE_{\bar{X}_{n=10}} = \frac{\sigma}{\sqrt{10}} = 3,976
\]

(2) **Ed uses a sample size $n=100$**  
Ed’s data are $X_1, \ldots, X_{100}$

\[
\bar{X}_{n=100} = 19,813 \\
\sigma = 12,573 \\
SE_{\bar{X}_{n=100}} = \frac{\sigma}{\sqrt{100}} = 1,257
\]
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:

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

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

<table>
<thead>
<tr>
<th>Nature</th>
<th>Population/Sample</th>
<th>Observation/Data</th>
<th>Relationships/Modeling</th>
<th>Analysis/Synthesis</th>
</tr>
</thead>
</table>
Example: Carol samples \( n = 10 \) workers.

Sample mean \( \bar{X} = 19,887 \)

Standard error of sample mean, \( SE_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = 3,976 \) 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, \( SE_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = 1,257 \) 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 = \( 22,277 - 17,349 = 4,928 \)

<table>
<thead>
<tr>
<th></th>
<th>( n )</th>
<th>Estimate</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carol</td>
<td>10</td>
<td>$19,887</td>
<td>( $12,094, $27,680 )</td>
</tr>
<tr>
<td>Ed</td>
<td>100</td>
<td>$19,813</td>
<td>( $17,349, $22,277 )</td>
</tr>
<tr>
<td>Truth</td>
<td>5000</td>
<td>$19,987</td>
<td>$19,987</td>
</tr>
</tbody>
</table>

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 $\mu$.

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

Caution on the use of Confidence Intervals:

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

   A given interval either contains $\mu$ 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 \( n=10 \)

\[
\text{95\% Confidence Intervals for mean = 50, sd = 10, n = 10}
\]

Notice …

1. Any one confidence interval either contains \( \mu \) or it does not. This illustrates that it is incorrect to say “There is a 95\% probability that the confidence interval contains \( \mu \)”

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

<table>
<thead>
<tr>
<th>( n = 10 )</th>
<th>( n = 100 )</th>
<th>( n = 1000 )</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Confidence Intervals n=10" /></td>
<td><img src="image2.png" alt="Confidence Intervals n=100" /></td>
<td><img src="image3.png" alt="Confidence Intervals n=1000" /></td>
</tr>
</tbody>
</table>

Now you can also see …

3. As the sample size increases, the confidence intervals are more narrow (more precise)

4. As \( n \to \infty \), \( \mu \) 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 μ=5.3 then μ is in (1.3, 9.5) with probability = 1. If μ=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.
5. Preliminaries: Some Useful Probability Distributions

5.1 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 $X_1, \ldots, X_n$ from a 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.

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

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.

![Graph showing normal distribution and student's t-distributions with different degrees of freedom](image-url)
How to Use the Student t Distribution Calculator Provided by SurfStat

- 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 \( \text{Student t}_{DF=1} < 3.078 \) = .90

Example – Solution for a percentile value: The 97.5\(^{th}\) Percentile of a Student t\(_{DF=9}\) = .90
5. Preliminaries: Some Useful Probability Distributions

5.2 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, \( \sigma^2 \). 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 Normal(\( \mu, \sigma^2 \))

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

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 \( \mu \) and variance \( \sigma^2 \),

   Then the random variable \( V \) defined:

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

   is distributed chi square distribution with degree of freedom = 1.

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

   Then their sum, defined:

   \[
   V_1 + ... + V_k \quad \text{is distributed chi square distribution with degrees of freedom = k}
   \]
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:

- Often, online calculators for the chi square distributions are “right tail” only.
- Tip! $1 = \text{“left tail area” + “right tail area”}$
- 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.cmappers.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.

![Chi Square Distribution Graph](web.mnstate.edu)

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.
How to Use the Chi Square Distribution Calculator Provided by SurfStat

- 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 \( \chi^2 \text{DF}=4 > 6.2 \) = .1847

![Chi-Square Distribution Calculator](http://surfstat.anu.edu.au/surfstat-home/tables/chi.php)

**Example – Solution for a percentile value:** The 97.5\(^{th}\) Percentile of a Chi-Square \( \text{DF}=9 \) = 19.02

![Chi-Square Distribution Calculator](http://surfstat.anu.edu.au/surfstat-home/tables/chi.php)
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 $\sigma_x^2 / \sigma_y^2$
A Definition of the F-Distribution

Suppose \( X_1, \ldots, X_{nx} \) are independent and a simple random sample from a normal distribution with mean \( \mu_X \) and variance \( \sigma_X^2 \). Suppose further that \( Y_1, \ldots, Y_{ny} \) are independent and a simple random sample from a normal distribution with mean \( \mu_Y \) and variance \( \sigma_Y^2 \).

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!!


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 2.5th percentile, right tail area = 0.975

![Critical F-value Calculator](http://www.danielsoper.com/statcalc3/default.aspx)

**Critical F-value: 0.11734906**

### For the 97.5th percentile, right tail area = 0.025

![Critical F-value Calculator](http://www.danielsoper.com/statcalc3/default.aspx)

**Critical F-value: 3.40826783**

---
5. Preliminaries: Some Useful Probability Distributions

5.4 Sums and Differences of Independent Normal Distributions

This is review. See again course notes, 7. 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 to 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): \( X_{11}, X_{12}, \ldots, X_{1n_1} \) is a simple random sample from a Normal \( (\mu_1, \sigma_1^2) \)
- (for group 2): \( X_{21}, X_{22}, \ldots, X_{2n_2} \) is a simple random sample from a Normal \( (\mu_2, \sigma_2^2) \)

- 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

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

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

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

A General Result Handy!

If random variables \( X \) and \( Y \) are independent with

\[
\begin{align*}
E [X] &= \mu_X \text{ and } Var [X] = \sigma_X^2 \\
E [Y] &= \mu_Y \text{ and } Var [Y] = \sigma_Y^2
\end{align*}
\]

Then

\[
\begin{align*}
E [aX + bY] &= a\mu_X + b\mu_Y \\
Var [aX + bY] &= a^2\sigma_X^2 + b^2\sigma_Y^2 \text{ and} \\
Var [aX - bY] &= a^2\sigma_X^2 + b^2\sigma_Y^2
\end{align*}
\]

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.