Unit 7
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"
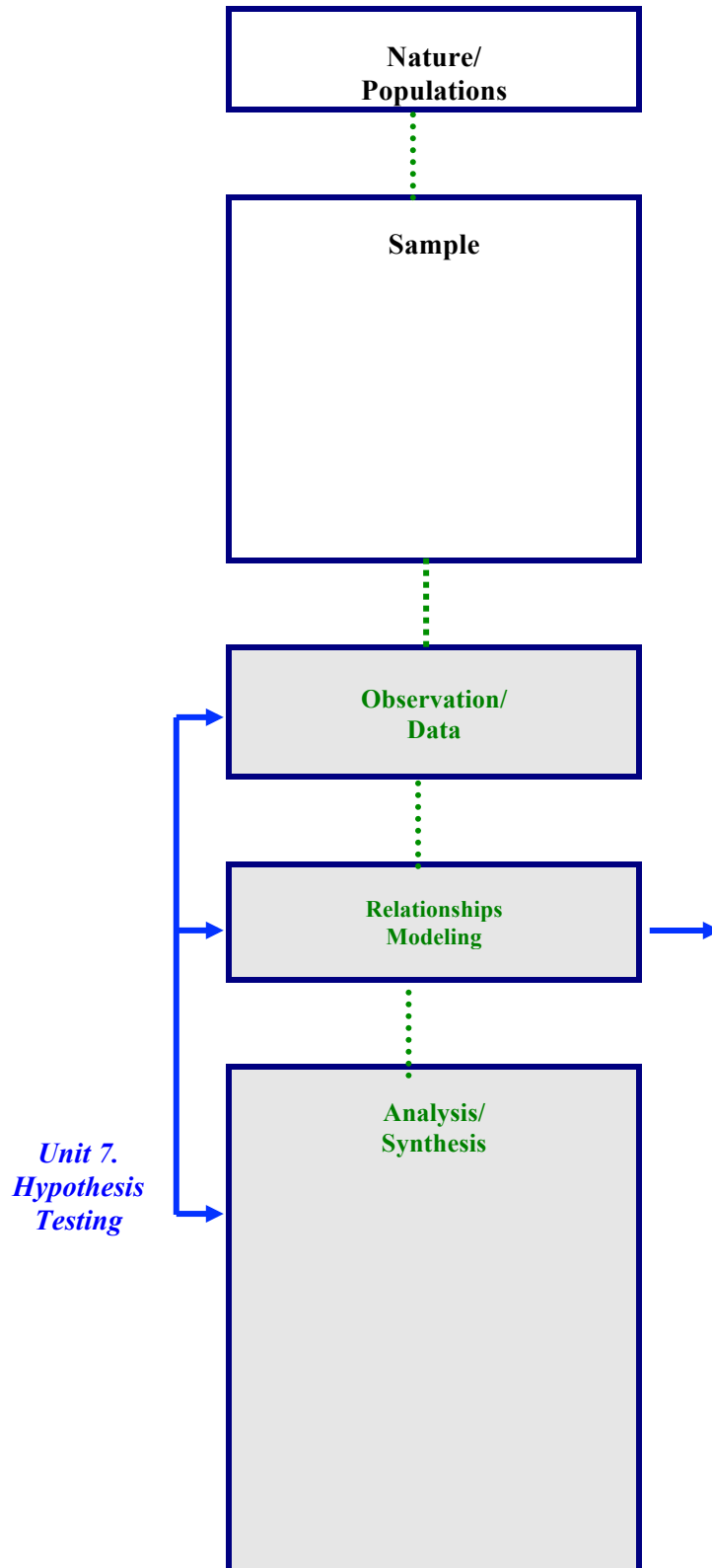
- Michel De Montaigne (1533-1592)

"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

Nature _____ **Population/** _____ **Observation/** _____ **Relationships/** _____ **Analysis/**
                              Sample   **Data**   **Modeling**   **Synthesis**

Table of Contents

Nature _____ **Population/** _____ **Observation/** _____ **Relationships/** _____ **Analysis/**
                        Sample   **Data**   **Modeling**   **Synthesis**

Unit Roadmap

Nature/ Populations	
Sample	
Observation/ Data	
Relationships Modeling	
Analysis/ Synthesis	

Unit 7. Hypothesis Testing

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!

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

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 (H_O) and alternative (H_A) 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 type II errors.

- Perform and interpret the statistical hypothesis tests described in the one and two sample settings described in these notes.

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

3. The Logic of Hypothesis testing

In 2010, Andrew Vickers of the Department of Epidemiology and Biostatistics at Memorial Sloan-Kettering Cancer Center published a wonderful (and very readable!) little book: ***What is a p-value anyway? 34 Stories to Help you Actually Understand Statistics*** (Addison Wesley, ISBN – 0-321-62930-2).

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.

*Source: Vickers, A. **What is a p-value Anyway? 34 Stories to Help You Actually Understand Statistics**. Addison Wesley, 2010. page 59*

Nature _____ **Population/** _____ **Observation/** _____ **Relationships/** _____ **Analysis/**
                Sample   **Data**   **Modeling**   **Synthesis**

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 $\overline{X}$=81 then a one sided p-value might be the calculation of** $Pr[\overline{X} \leq 81]$ **under the null hypothesis assumption that μ=100**

Nature _____ **Population/** _____ **Observation/** _____ **Relationships/** _____ **Analysis/**
            Sample   **Data**   **Modeling**   **Synthesis**

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**

$\Pr[\bar{X} \leq 81]=.02$ **under the null hypothesis assumption that μ=100. This is saying the following: The chances were 2 in 100 of obtaining an average IQ of 81 or lower if μ=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 tootbrush 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.

Nature _____   **Population/** _____ **Observation/** _____   **Relationships/** _____   **Analysis/**
                         Sample   **Data**   **Modeling**   **Synthesis**

Steps in Hypothesis Testing

1.   **Identify the research question.**

2.   **State the <u>null hypothesis</u> assumptions necessary for computing probabilities.**

3.   **Specify H_O 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.**

Nature _____ **Population/** _____ **Observation/** _____ **Relationships/** _____ **Analysis/**
                              Sample   **Data**   **Modeling**   **Synthesis**

Schematic of Statistical Hypothesis Testing

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

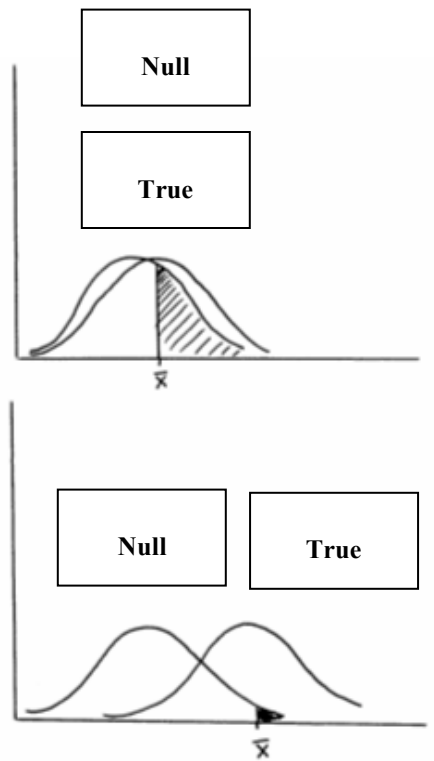|	
Null **True** $\overline{x}$	**Top Picture.** In the top picture, the two probability distributions (null & true) are essentially the same. The sample mean is close to its "true" expected value (a population mean). **Take home** – The null hypothesis *is* consistent with the data (the sample mean). We have no contradiction**.**
Null **True** $\overline{x}$	**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.

Nature _____ **Population/** _____ **Observation/** _____ **Relationships/** _____ **Analysis/**
                    Sample   **Data**   **Modeling**   **Synthesis**

A step-by-step schematic of how "proof by contradiction" and the rejection of the null works.

Null

True

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).

$\bar{X}$

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.

Null   True

$\bar{X}$

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:

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

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"

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

A closer look at $\Pr[\bar{X} \geq$ observed value | assuming null hypothesis model is true]

Scenario 1 - NULL is true

$\Pr[\bar{X} \geq$ observed value | assuming null is true]= **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 <u>large probability</u> (<u>large shaded area</u>). That is - assuming the null hypothesis model leads to a <u>likely outcome</u>.

- Statistical decision - "do NOT reject the null".

Scenario 2 - ALTERNATIVE is true

$\Pr[\bar{X} \geq$ observed value | assuming null is true]= **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 <u>small probability</u> (<u>small shaded area</u>). That is - assuming the null hypothesis model leads to an <u>unlikely outcome</u>.

- Statistical decision - "REJECT the null".

p-value =

$\Pr[$ Test statistic (eg $\bar{X}$) $=$ observed or more extreme | 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/ _____   Observation/ _____   Relationships/ _____   Analysis/
                              Sample   Data   Modeling   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 8.

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, $X_1, X_2, ..., X_{100}$ have average $\overline{X}_{n=100} = 46.9$ months. *Is this compelling evidence that* $\mu_{true} > 38.3$?

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$_O$: $\mu_{true} = \mu_O \leq 38.3$ **months**
H$_A$: $\mu_{true} = \mu_A > 38.3$ **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.

Nature _____ **Population/** _____ **Observation/** _____ **Relationships/** _____ **Analysis/**
     Sample   **Data**   **Modeling**   **Synthesis**

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 <u>small chance</u> 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_O) chances of "extremeness."   This value will be the p-value.*

Under the null hypothesis model::

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

This, in turn **(see again, course notes 6. Estimation, page 31)**, says that under the assumption that the null hypothesis is true:

$\bar{X}_{n=100}$ is distributed Normal ($\mu = 38.3, \sigma^2 = 43.3^2 / (n = 100)$)

Nature _____ **Population/** _____ **Observation/** _____ **Relationships/** _____ **Analysis/**
                         Sample   **Data**   **Modeling**   **Synthesis**

How extreme is "extreme" is an example of "signal-to-noise".

Signal - "46.9 is 8.6 months away from 38.3" Signal = 8.6 **Is 8.6 extreme or not?**	$$(46.9 - 38.3) = 8.6$$
Noise – Noise is the scatter/variability of the average. We measure this using the SE **How "noisy" is the mean typically? We measures this in units of SE.**	$$\text{SE}(\overline{X}_{n=100}) = \frac{\sigma}{\sqrt{100}} = \frac{43.3}{10} = 4.33$$
Signal-to-Noise (Z-score) Signal, in units of months, has been re-expressed in units of noise (SE units) **"46.9 is 1.99 SE units away from 38.3"**	$$\text{Z-score} = \frac{(\overline{X}_{n=100} - \mu_{\overline{X} \text{ under NULL}})}{\text{SE}(\overline{X}_{n=100})}$$ $$= \frac{(46.9\text{-}38.3)}{\text{SE}(\overline{X}_{n=100})}$$ $$= \frac{8.6 \text{ months}}{4.33 \text{ months}}$$ $$= \ 1.99 \text{ SE units}$$

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"

| **Nature** _____ | **Population/** _____ **Sample** | **Observation/** _____ **Data** | **Relationships/** _____ **Modeling** | **Analysis/** **Synthesis** |

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"

$$\Pr\left[\bar{X}_{n=100} \geq 46.9 \,|\, \mu_{true} = \mu_{null} = 38.3\right]$$
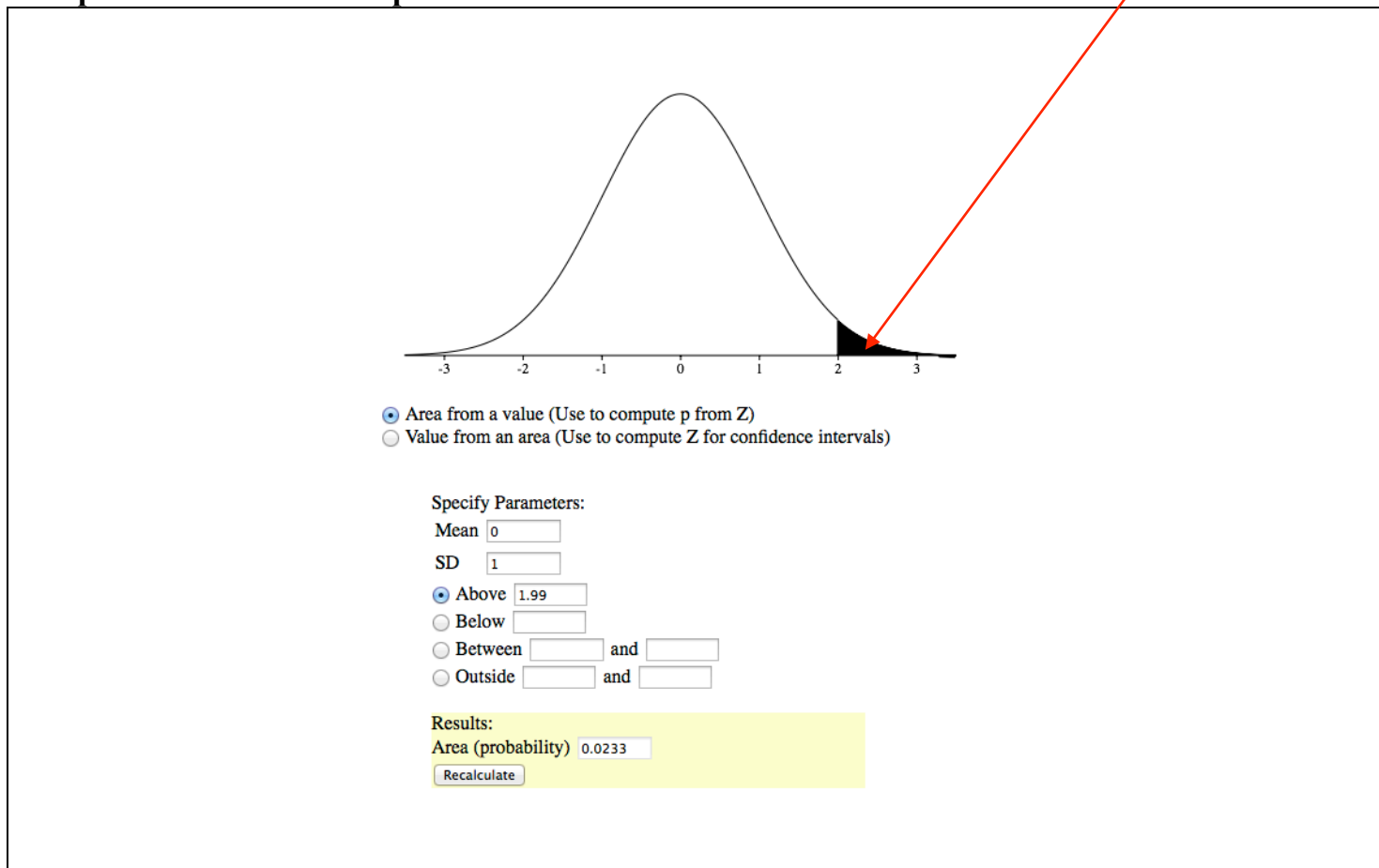
$$= \Pr[\text{Z-score} \geq 1.99] = .02$$

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

Area from a value (Use to compute p from Z)
Value from an area (Use to compute Z for confidence intervals)

Specify Parameters:
Mean 0
SD   1
Above 1.99
Below
Between   and
Outside   and

Results:
Area (probability) 0.0233
Recalculate

http://davidmlane.com/hyperstat/z_table.html

Nature _____ **Population/** _____ **Observation/** _____ **Relationships/** _____ **Analysis/**
Sample   **Data**   **Modeling**   **Synthesis**

The Z-score is a Signal-to-Noise Comparison

$$\text{Z-score} = \frac{\text{Signal}}{\text{Noise}} = \frac{\text{observed-expected}}{\text{SE(observed)}}$$ $$= \left[\frac{\overline{X}_{n=100} - \mu}{\left(\sigma \big/ \sqrt{n} \right)} \right]$$ **Example: z-score=1.99**	*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.*
p-value = Pr[Normal(0,1) ≥ z-score] **Example: pr[normal(0,1) ≥ 1.99] = .02**	p-value = The chances of obtaining a departure of this magnitude, or greater, calculated under the presumption that the null hypothesis is true.

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

TIP!

The Z-score is a very handy measure of extremeness
ps – and so is the T-score!

Normal Curve
Standard Deviation

Chart prepared by the NY State Education Department
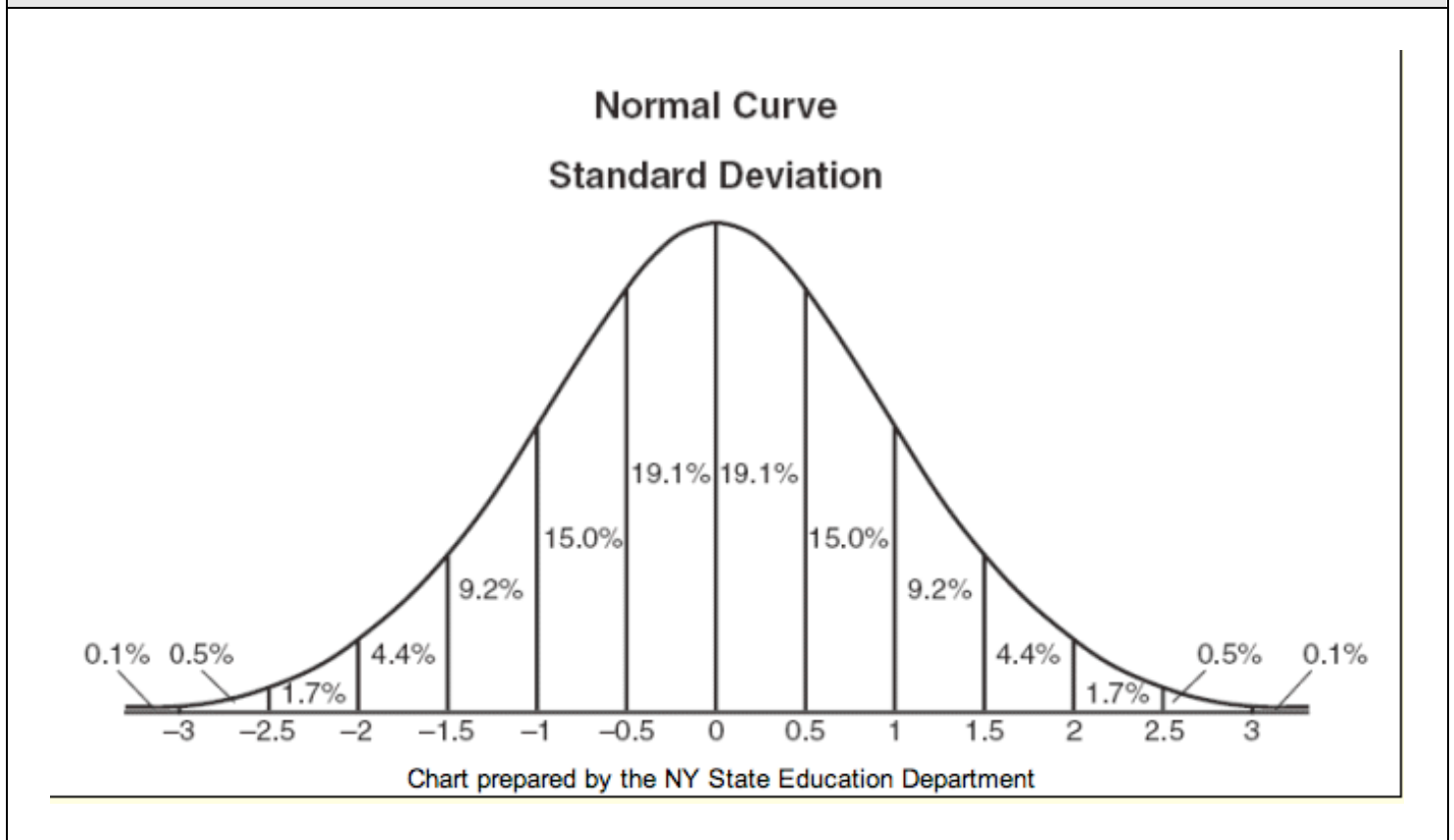
Example –

Imagine you are reading a manuscript and you see a sample mean (eg – 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%.

Nature _____ **Population/** _____ **Observation/** _____ **Relationships/** _____ **Analysis/**
                                Sample   **Data**   **Modeling**   **Synthesis**

3.1 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:

$$H_A: \quad \mu_A > 38.3 \text{ months}$$

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:

$$H_A: \quad \mu_A \neq 38.3 \text{ months}$$

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

    Step 1 – Obtain Z-score measure of "signal"

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

    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.

$$
\begin{aligned}
\text{p-value}_{TWO\,SIDED} &= \text{Prob[Normal(0,1)} \geq +1.99] + \text{Prob[Normal(0,1)} \leq -1.99] \\
&= (2)\,\text{Prob[Normal(0,1)>1.99]} \\
&= (2)(.02) \\
&= .04
\end{aligned}
$$

    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.

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

4. Beware the Statistical Hypothesis Test

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 ***statistically significantly*** better than results for patients receiving treatment "B".

There are actually multiple, different, explanations:

- *Explanation #1* - Treatment "A" is truly superior.

- *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.

- *Explanation #3* – An event of low probability has occurred. Treatment "B" is actually superior but sampling (*as it will occasionally do!*) yielded a rare outcome. **(Consider this - Events of low probability do occur sometimes, just not very often).**

Nature _____ **Population/** _____ **Observation/** _____ **Relationships/** _____ **Analysis/**
                            Sample   **Data**   **Modeling**   **Synthesis**

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_O 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."

Nature _____ **Population/** _____ **Observation/** _____ **Relationships/** _____ **Analysis/**
                          Sample   **Data**   **Modeling**   **Synthesis**

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.

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

5. Introduction to Type I and II Error and Statistical Power

A statistical hypothesis test uses probabilities based <u>only</u> on the null hypothesis (H$_O$) model!

- Our starting point is the assumption of the null hypothesis model; that is we presume that H$_O$ 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 (H$_O$) model, or we retain it (fail to reject).**

  - **We do <u>*not*</u> prove that the null hypothesis assumption is correct.**

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

<u>**Decision**</u>

		Retain the null	**Reject the null**
Truth	**Null true**	☺	α = type I error
	Alternative true	β = type II error	☺

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

Introduction to Type II Error
- **IF H$_a$ is true and we (incorrectly) fail to reject H$_O$**
  - We have made a type <u>II</u> error
  - We must have a specific H$_a$ model before we can calculate Pr [type II error]= β

Introduction to Power
- **IF H$_a$ is true and we (correctly) reject H$_O$**
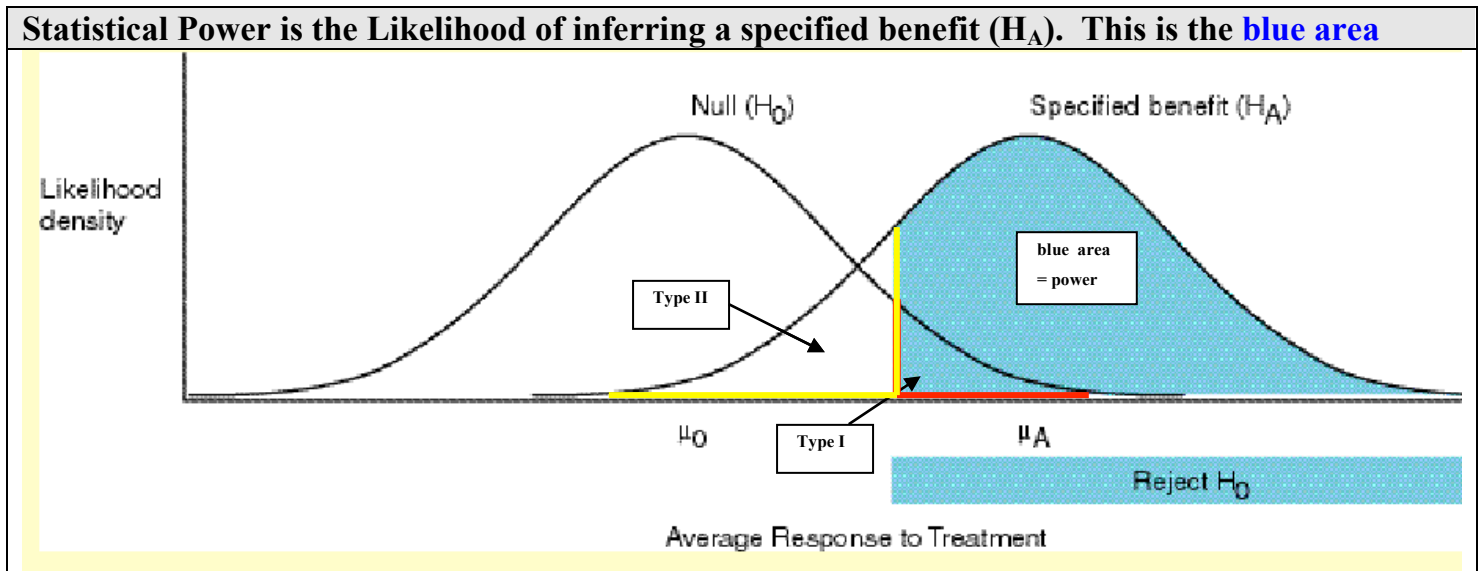 - **This occurs with probability = $\left(1-\beta\right)$ which we call the "POWER"**

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

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 = α) or a type II error (probability = β).

- **Sample size calculations!** Larger sample sizes will lower both probabilities: α and β

- 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.

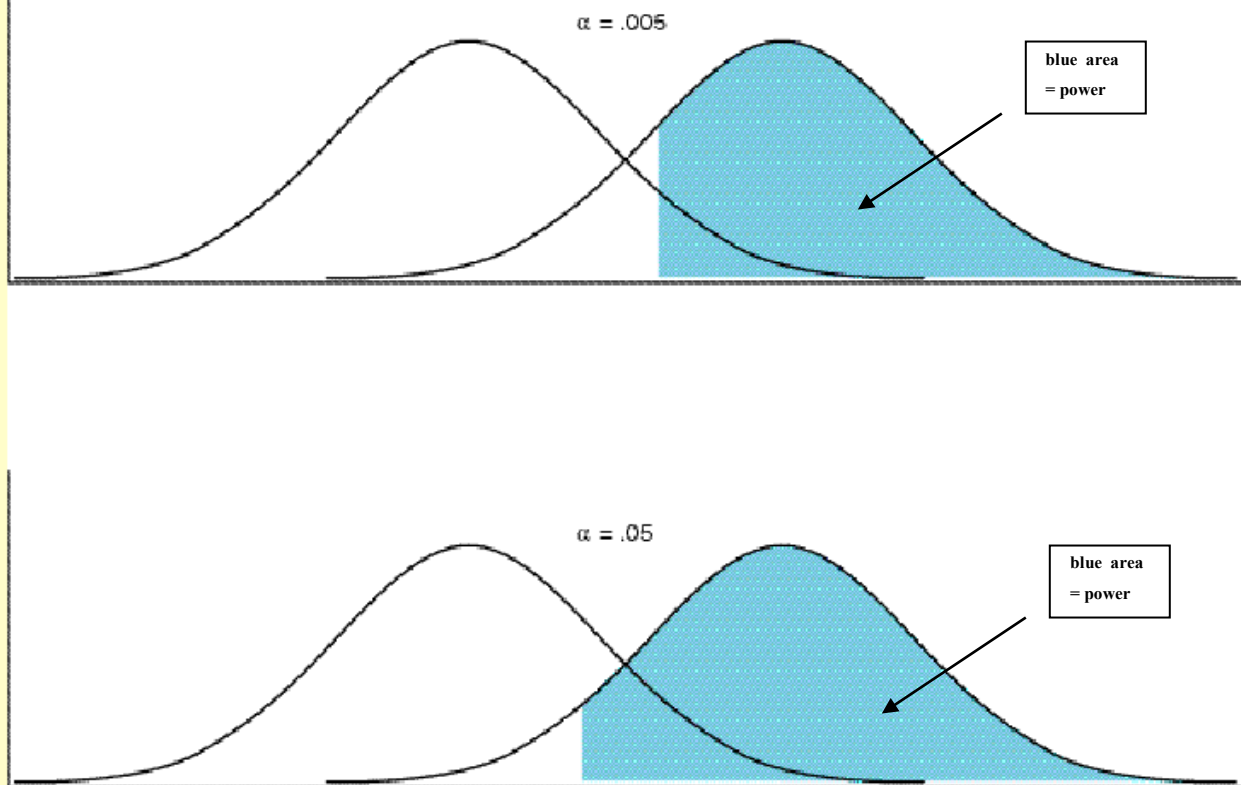**The techniques of sample size and power calculations are not addressed in this course.**

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$_O$" 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$_O$) curve: The type I error (probability = α).** This is the probability of mistakenly rejecting the null hypothesis; thus, it is calculated under the assumption that H$_O$ is true.

- **White area under the Alternative (H$_A$) curve: The type II error (probability = β).** This is the probability of mistakenly inferring the null; thus it is calculated under the assumption that H$_A$ is true**.**

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

**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, <u>then</u> 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).

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

The Power of a Study Depends on Four Parameters
**2. The Benefit Worth Detecting**
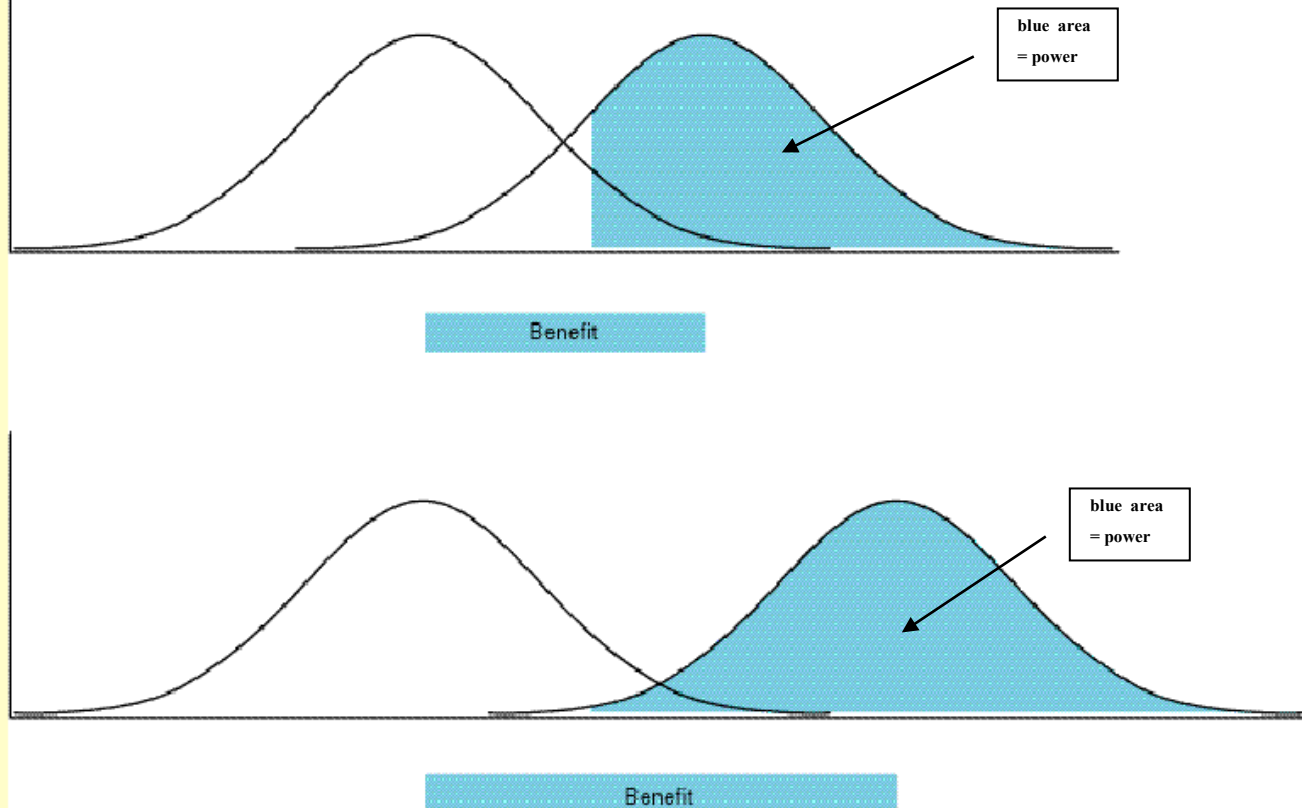
Take home message: The farther away the alternative (treatment benefit) is from the null hypothesis, the greater your chances of inferring it.   Conversely, the closer the alternative (treatment benefit) is to the null hypothesis, the lower your chances of inferring it.

- 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).

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

The Power of a Study Depends on Four Parameters
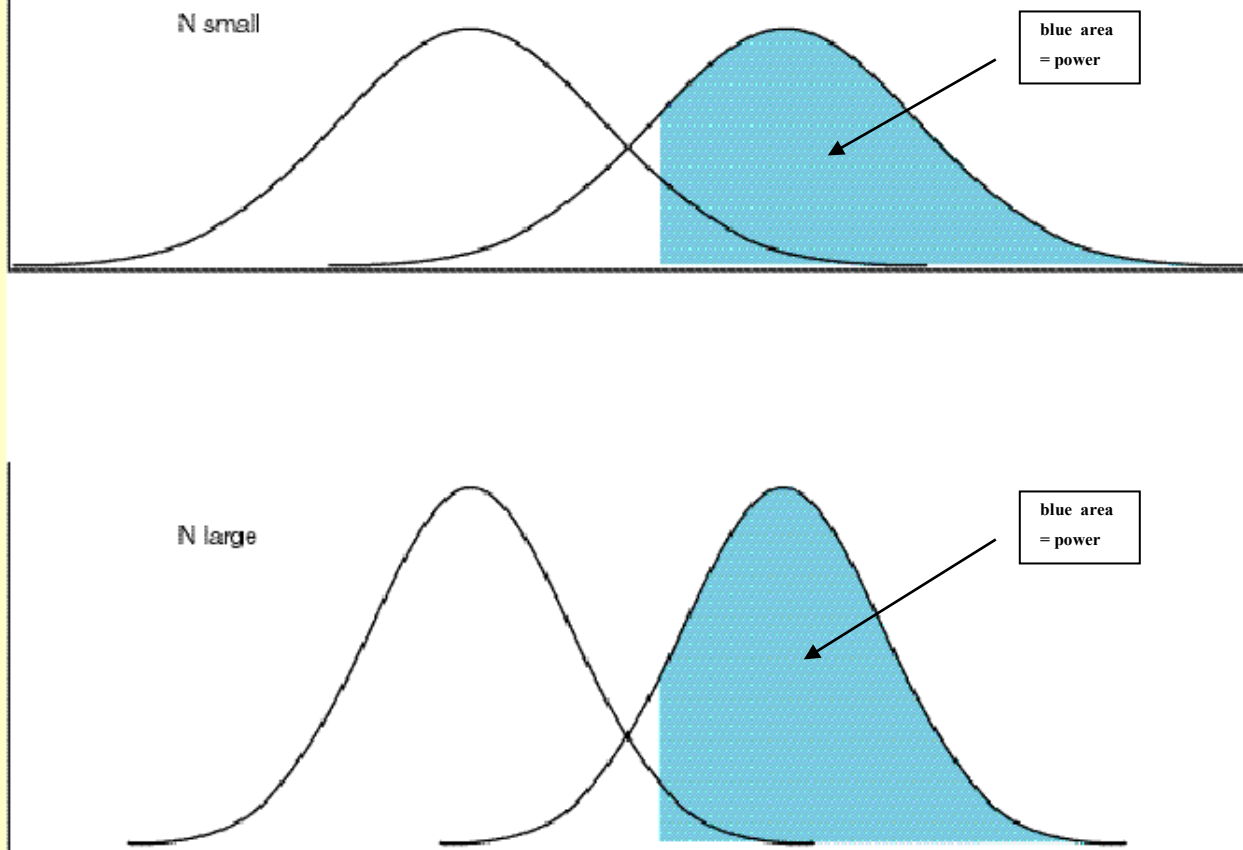**3. Biological Variability ("Noise")**
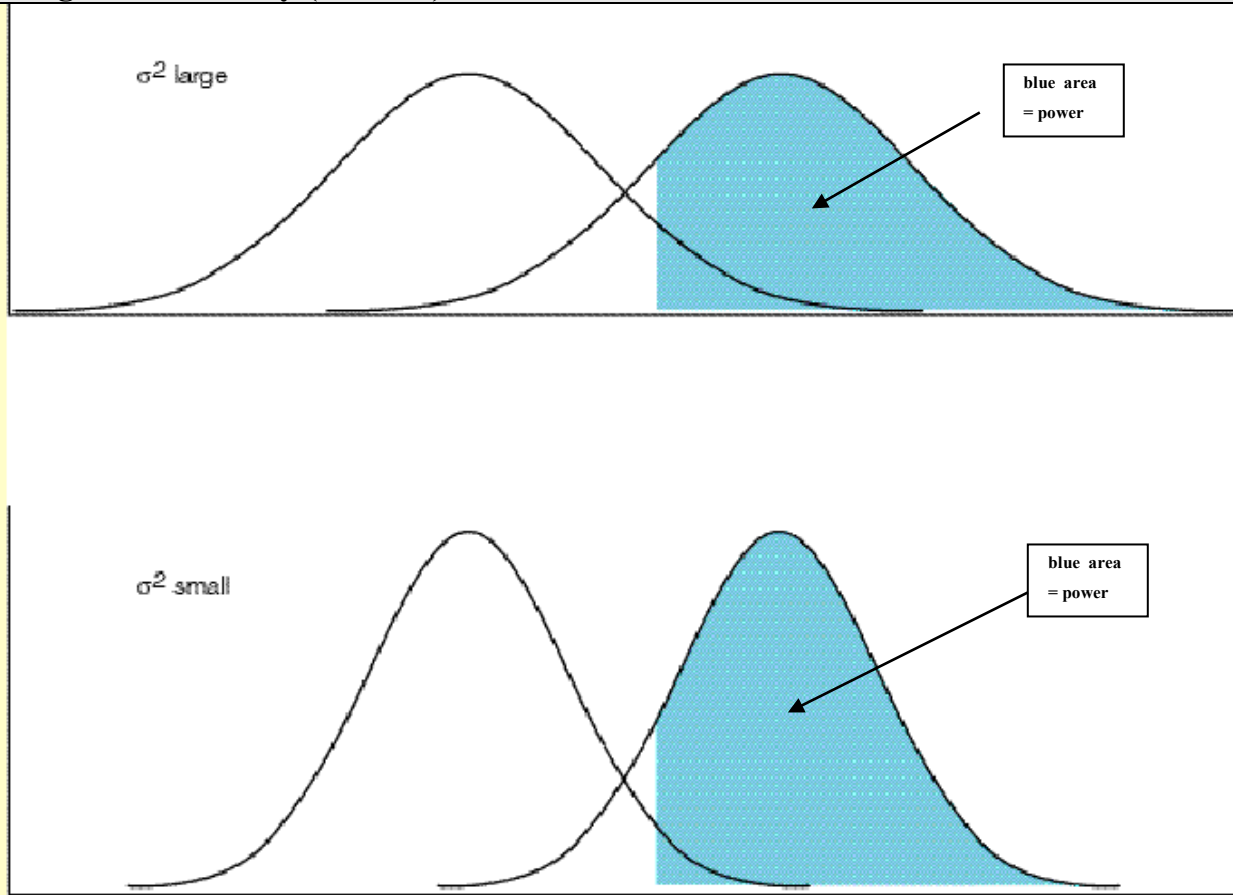
Take home message: The more precisely you can measure the outcomes of interest, the greater your chances of inferring a treatment benefit (the alternative); this is because the standard error is smaller.

- 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).

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

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).

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

1 Group, Outcome Continuous and Distributed Normal
6. Test for μ , σ^2 Known

The sections that follow in this reading parallel closely sections 5-9 of Unit 6, Estimation.

- *Tip* – Re-read course notes *6. Estimation*, pages 20-30, for review of the student's t, chi square, and F distributions.

- **Critical region** tests are also introduced.

- As we'll see, the steps in performing a statistical test are similar across the settings.

An example of a test for μ, when data are from a normal distribution with σ^2 known has been presented previously.

- Therefore, an abbreviated presentation is given here (so that these notes are easy to read!)
- For full details, see pp 12-16.

Example –

With standard care, cancer patients are expected to survive a mean duration of time equal to 38.3 months. Hypothesized is that a new therapy will improve survival. In this study, the new therapy is administered to 100 cancer patients. Their average survival time is 46.9 months. Suppose σ^2 known = 43.3^2 months squared. Is this statistically significant evidence of _improved_ survival?

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

Null Hypothesis Probability Model Assumptions.

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

Specify the null and alternative hypotheses

H$_O$: $\mu_{true} = \mu_O \leq 38.3$ **months**

H$_A$: $\mu_{true} = \mu_A > 38.3$ **months** **one sided.**   **This is because the investigator is researching "improvement"**

Reason "proof by contradiction"

   IF: we assume that the null hypothesis model is true, so that $\mu_{true} = \mu_O = 38.3$

THEN: what are the chances of observing a sample mean (average) survival time as great as 46.9 months, or greater, relative to the null hypothesis expected mean survival time of 38.3 months?

Specify the p-value calculation and hence the "proof by contradiction" reasoning.

Statistically, the null hypothesis applied to the observed data will lead to an <u>inconsistent conclusion</u> if there is at most a <u>small chance</u> of a mean of 100 survival times being 46.9 *or greater* when the expected value is 38.3. We calculate the p-value as

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

Nature _____ **Population/** _____ **Observation/** _____ **Relationships/** _____ **Analysis/**
                    Sample   **Data**   **Modeling**   **Synthesis**

The test statistic is a Z-Score

Under the assumption that the null hypothesis is true:

- $X_1, X_2, ..., X_{100}$ is each distributed Normal($\mu = 38.3, \sigma^2 = 43.3^2$).

- $\overline{X}_{n=100}$ is distributed Normal ($\mu = 38.3, \sigma^2 = 43.3^2 / 100$)

- We'll use as our test statistic a ***particular z-score standardization*** of $\overline{X}_{n=100}$. The standardization used is the one in which the population mean value subtracted is the one corresponding to the null hypothesis.

$$\text{Test Statistic} = \text{z-score} = \frac{\overline{X}_{n=100} - \mu_{null}}{SE(\overline{X}_{n=100})}$$

P-value calculation

$$\textbf{p-value} = \Pr\left[\overline{X}_{n=100} \geq 46.9 \mid \mu_{true} = \mu_{null} = 38.3\right]$$

$$= \Pr[\text{Z-score} \geq 1.99] = .02$$

"Evaluate".

IF the new therapy yields no improvement in survival so that the survival experience under the new therapy is identical to that experienced with receipt of standard care,

THEN there was a 2% chance of observing an average survival time as great or greater than the observed average survival time of 46.9 months.

Interpret.

The null hypothesis has led us to a very unlikely event (it had a 2% chance of occurrence) when considered in light of the observed data. This suggests abandoning the null hypothesis model. →. Reject the null hypothesis. Conclude that, compared to standard care, there is statistically significant evidence of longer survival times on the new treatment.

Nature _____ **Population/** _____ **Observation/** _____ **Relationships/** _____ **Analysis/**
               Sample   **Data**   **Modeling**   **Synthesis**

1 Group, Outcome Continuous and Distributed Normal
7. Test for μ, σ^2 Known
Critical Region Test Approach

The approach to hypothesis testing described in the previous pages is called the significance level approach. In this method, we were asking the question:

- Under the assumption that the null hypothesis is true, what were my chances of obtaining a test statistic as extreme or more extreme?

The critical region approach follows a slightly different (but related) thinking:

- **If** I assume that the null hypothesis is true,

- **And if** I agree that I will reject the null hypothesis under certain extreme conditions,

- **Then** what values of my test statistic will lead to rejection of the null hypothesis if I want my **type I error** to be a certain value?

How do Critical Regions Work?

- We agree *in advance (prior to collecting data)* that we will honor a *threshold test statistic value (this is what we mean by critical value)*, beyond which we will reject the null hypothesis.   We'll do this even though, theoretically under the null hypothesis, such extreme values are still possible.

- This means that whenever we then do get a test statistic value that is beyond the critical value, *if the null hypothesis is actually true*, then we will have *incorrectly reject a true null hypothesis.*   Under these circumstances, we will have made a *type I error*.

- *How does this actually work?* In developing a critical region test, we determine, *before actual data collection*, the *threshold statistic value* (the "*critical value*") and the *range of extreme values that lie beyond* (this is called the *critical region*) that will (in the future – after obtaining our data) prompt (rightly or wrongly!) rejection of the null hypothesis.

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

Example – again –

With standard care, cancer patients are expected to survive a mean duration of time equal to 38.3 months. Hypothesized is that a new therapy will improve survival. In this study, the new therapy is administered to 100 cancer patients. Their average survival time is 46.9 months. Suppose σ^2 known = 43.3^2 months squared. Is this statistically significant evidence of improved survival *at the 0.05 level*?

Notice the extra wording *at the 0.05 level* .
.05 represents a 5% chance. In the next pages, I will show you how to use this to
determine the threshold value of the test statistic and the associated 0.05 critical region.

Null Hypothesis Probability Model Assumptions.

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

Null and alternative hypotheses

H$_O$: $\mu_{true} = \mu_O \leq 38.3$ **months**
H$_A$: $\mu_{true} = \mu_A > 38.3$ **months**

The appropriate Test Statistic is a Z-Score

The null hypothesis gives us the following:

- $X_1, X_2,..., X_{100}$ is a simple random sample from a Normal($\mu = 38.3, \sigma^2 = 43.3^2$).
- $\overline{X}_{n=100}$ is distributed Normal ($\mu = 38.3, \sigma^2 = 43.3^2 / 100$)
- Again, we'll use as our test statistic the z-score standardization of $\overline{X}_{n=100}$, obtained under the assumption that the null hypothesis is correct.

$$\text{Test Statistic} = \text{z-score} = \frac{\overline{X}_{n=100} - \mu_{null}}{SE(\overline{X}_{n=100})}$$

Nature _____ **Population/** _____ **Observation/** _____ **Relationships/** _____ **Analysis/**
                          Sample   **Data**   **Modeling**   **Synthesis**

Using the direction of the alternative, obtain the 0.05 critical region

Step 1: **Consider the _direction_ of the alternative, relative to the null hypothesis**
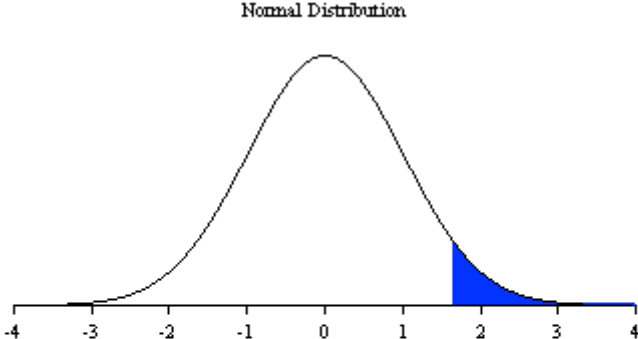**(eg – if the new treatment is better, survival will be longer):**
In this example, the alternative hypothesis model is one sided. Moreover, we are looking to reject the null hypothesis in favor of the alternative (improved survival) if the test statistic is extremely **large.**

Step 2: **Solve for the threshold and critical region of the test statistic**
Here is where our 5% chances (see previous page) get used. Work with the null hypothesis model. Specifically, in this example, the **Z-statistic** (see bottom of page 33) will be distributed Normal(0,1) when the null hypothesis is true. We have already agreed that we want to solve for a threshold value of our Z-statistic, beyond which all values will prompt rejection of the null hypothesis, even when the null might actually be true. **5% chances** and **"beyond which"** translates to **right tail area = .05**.

I used the link http://davidmlane.com/hyperstat/z_table.html
Be sure to choose **"value from an area"**.

Step 1:
Enter .05 for shaded area

Step 2:
Select radio button "Above"

Step 3:
Read threshold value as z-score = 1.6449 and critical region as z-score > 1.6449

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

Step 3: Using the critical region of the Z-statistic, solve for the critical region of $\overline{X}$:

> **How? We do this by setting the formula for the z-score equal to the value of the critical value for the z-score that was obtained in step 2, namely 1.6449.**

$$\text{z-score} \; \geq \; 1.6449 \; \rightarrow$$

$$\frac{\overline{X}_{n=100} - \mu_{null}}{SE(\overline{X}_{n=100})} \; \geq \; 1.6449 \; \rightarrow$$

$$\overline{X}_{n=100} - \mu_{null} \; \geq \; (1.6449)*SE(\overline{X}_{n=100}) \; \rightarrow$$

$$\overline{X}_{n=100} \; \geq \; \left[(1.6449)*SE(\overline{X}_{n=100})\right] + \mu_{null} \; \rightarrow$$

$$\overline{X}_{n=100} \; \geq \; \left[(1.6449)*(4.33)\right] + 38.3 \; \rightarrow$$

$$\overline{X}_{n=100} \; \geq \; 45.42$$

The critical region is $\overline{X}_{n=100} \geq 45.42$

Step 4: Interpret:

Going forward, after we have collected data and computed our test statistic, we will follow the following rule: If the observed $\overline{X}_{n=100}$ is **_at or beyond_** the threshold value of 45.42, we will infer the alternative hypothesis, even though such extreme values are theoretically possible when the null hypothesis is true. That is – "this **_critical region_** one sided .05 test of the null versus alternative hypotheses has been defined to reject the null hypothesis for any $\overline{X}_{n=100} \geq 45.42$.

Examine the observed to see if it is in the critical region

Now collect your data and compute your sample mean. In this example, the sample mean $\overline{X}_{n=100} = 46.9$. Because it is exceeds the threshold value of 45.42, it falls in the critical region.

Interpret.

Because the observed $\overline{X}_{n=100} = 46.9$ **_exceeds the value of the threshold 45.42_** and is, therefore, **_in the critical region_**, in critical region parlance we say **_"it is significant at the 0.05 level"_**. → **reject the null hypothesis.** The conclusion is the same: these data provide statistically significant evidence that, compared to standard care, survival times on the new treatment are longer.

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

1 Group, Outcome Continuous and Distributed Normal
8. Test for μ, σ^2 UNknown

Hypothesis testing in the setting of a sample from a single normal distribution with σ^2 not known is, not surprisingly, quite similar to that when the data are from a distribution with σ^2 known.

o   **The test statistic is a t-score instead of a z-score.**
     (Note – See course notes *6. Estimation* page 20 for a review of the Student t distribution)

Same example –
With standard care, cancer patients are expected to survive a mean duration of time equal to 38.3 months. Hypothesized is that a new therapy will improve survival. In this study, the new therapy is administered to 100 cancer patients. Their average survival time is 46.9 months. Suppose σ^2 is not known. Suppose instead that what is available is the sample variance of survival times $S^2 = 43.3^2$ months squared. Do these data provide statistically significant evidence of improved survival?

Null Hypothesis Probability Model.

$X_1, X_2, ..., X_{100}$ is a simple random sample from a Normal(μ, σ^2)
σ^2 is **NOT** known.

Null and alternative hypotheses

H$_O$: $\mu_{true} = \mu_O \le 38.3$ **months**
H$_A$: $\mu_{true} = \mu_A > 38.3$ **months** <u>one sided</u>

Reason "proof by contradiction" and use this to define the p-value calculation.

Statistically, the null hypothesis, when examined in light of the value of the test statistic, will have led to an unlikely outcome (an inconsistency or "contradiction") if the null hypothesis model probability of that test statistic value (a sample mean being as extreme as 46.9 months or larger) is small.   The chances of this (the p-value) is the following probability calculation:

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

Nature _____ **Population/** _____ **Observation/** _____ **Relationships/** _____ **Analysis/**
                              Sample   **Data**   **Modeling**   **Synthesis**

The appropriate Test Statistic is now a Student T-Test or T-Score

Under the assumption of the null hypothesis:

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

- $\overline{X}_{n=100}$ is distributed Normal ($\mu = 38.3, \sigma_{\overline{X}}^2 = \sigma^2/100$)

- Here, we'll use as our test statistic the ***t-score standardization*** of $\overline{X}_{n=100}$, obtained under the assumption that the null hypothesis is correct.

$$\text{Test Statistic = t-score=} \frac{\overline{X}_{n=100} - \mu_{null}}{\hat{SE}(\overline{X}_{n=100})}$$

- Our denominator has to be an estimate of the unknown SE, $\hat{SE}(\overline{X}_{n=100}) = \dfrac{S}{\sqrt{100}} = \dfrac{43.3}{10} = 4.33$

  because we do not know the value of the population σ.

P-value calculation

$$\textbf{p-value =} \Pr\left[\overline{X}_{n=100} \geq 46.9 \,|\, \mu_{true} = \mu_{null} = 38.3\right]$$

$$= \Pr\ [\text{t-score}_{\text{degrees of freedom=99}} \geq 1.99] = .02467 \quad \textit{quite close to .02 obtained on page 30}$$

"Evaluate".

IF we assume that the null hypothesis is true, meaning that the new therapy elicits ***no improvement*** in survival so that the survival experience under the new therapy is ***the same as*** that experienced with receipt of standard care,

THEN the probability was an **estimated 2.4%** (p-value = .024) chance that we would have obtained our observed average survival time of 46.9 months ***or greater***.

Interpret.

Nature _____   **Population/** _____   **Observation/** _____   **Relationships/** _____   **Analysis/**
                        Sample   **Data**   **Modeling**   **Synthesis**

The assumption that the null hypothesis is true, *when examined in light of the observed statistic's value*, has led to an unlikely conclusion. Abandon the null hypothesis. → **Reject the null hypothesis**. The conclusion is again the same: these data provide statistically significant evidence that, compared to standard care, survival times on the new treatment are longer.

Nature _____ **Population/** _____ **Observation/** _____ **Relationships/** _____ **Analysis/**
                              Sample   **Data**   **Modeling**   **Synthesis**

1 Group, Outcome Continuous and Distributed Normal
9. Test for σ^2
(Note – See course notes *6. Estimation* page 24 for a review of the Chi Square distribution)

Example -
Tip! – Studies of laboratory performance (reproducibility, precision) are studies of variance.

In drug manufacturing it is important, not only that the amount of drug in the capsules be a particular value on the average, but also that the <u>variation</u> around that value be very small.   The drug company will consider its machine accurate enough if the capsules are filled within 1 SD =.5 mg of the desired amount of the drug (2.5 mg). Data is collected for n=20 capsules. The observed sample standard deviation is S= 0.787. Is this variability statistically significantly greater than what the company will tolerate? Test whether the drug company should adjust its machines again. The company will only adjust the machine if the variance is too large.

Research Question:

Is the variance of drug in the capsules <u>greater</u> than $(.5)^2 = 0.25$ mg^2?

Null Hypothesis Assumptions:

The data are a simple random sample from a normal distribution.

Specify Hypotheses:

H$_0$: $\sigma^2 \leq 0.25$

H$_a$: $\sigma^2 > 0.25$ <u>one-sided</u>

Reason "proof by contradiction" and use it to define the p-value calculation.

Statistically, the null hypothesis, when examined *in light of the observed data*, lead to an inconsistency or "contradiction" if the null hypothesis probability is small that the observed sample SD among n=20 capsules is 0.787 or larger. Thus, the required p-value calculation is:

$$\text{P-value} = \Pr\left[S \geq 0.787 \,|\, \sigma_{true} = \sigma_O = 0.5\right]$$

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

The Test Statistic is a Chi Square:

S, as is, cannot be our test statistic. Its "behavior" isn't convenient. Thus, we do not calculate Pr [S > 0.787] directly.

Instead we work with a related random variable that does have convenient properties. It is called Y and is obtained under the assumption that the null hypothesis is true. Note – This is analogous working with a Z-score standardization of the sample mean.

In particular, under the assumption that the null hypothesis is true

$$Y=\frac{(n-1)S^2}{\sigma^2_{NULL}}$$ **is distributed chi square with degrees of freedom = (n-1)**

P-value = s the probability of that a chi square random variable is as extreme or more extreme (translation: larger) than the value that we obtained from our data

$$Y=\frac{(n-1)S^2}{\sigma^2_{NULL}}=\frac{(19)(0.787)^2}{0.25}=47.072$$

p-value = $\Pr\left[\text{Chi Square}_{DF=19}\geq 47.072\right]=0.0003$ **(the area under the curve is tiny!)**
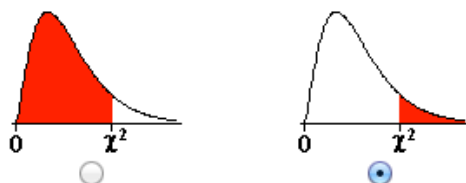
http://surfstat.anu.edu.au/surfstat-home/tables/chi.php

Nature _____ **Population/** _____ **Observation/** _____ **Relationships/** _____ **Analysis/**
              Sample   **Data**   **Modeling**   **Synthesis**

"Evaluate".

Assuming the null hypothesis to be true, and evaluating it in light of the data, leads to an inconsistency/contradiction. Under the null hypothesis model, the probability was an estimated 0.03% chance that our sample variance would be 0.787^2 or larger.

Interpret.

The null hypothesis is abandoned. → Reject the null hypothesis. Conclude that these data provide statistically significant evidence that the variability in capsule drug content is greater than .25 mg^2.

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

<div align="center">

Paired Data, Outcome Continuous and Distributed Normal
10. Test for $\mu_{DIFFERENCE}$

</div>

Two scenarios are presented.

- #1. Variance is assumed KNOWN
- #2. Variance is assumed NOT known

Example Scenario #1 – Variance Assumed Known:
(Note: These data are hypothetical.)

Twelve patients (n=12) in a needle exchange trial who were randomized to the pharmacy sales alone condition provided hair samples that were positive for cocaine at the baseline interview. Follow-up hair samples were obtained from these same n=12 at the 6 month visit, yielding paired data.

Is participation ***itself*** associated with reduced hair content of cocaine? **This is a reasonable concern. If participation itself is associated with reduced use of cocaine, then *at least* two issues arise: (1) there might be selection bias which makes generalization to a larger population difficult; and/or (2) it is difficult to attribute reduced cocaine use to the intervention and/or (3) ????**

 Research Question.

In the absence of an effect of study participation, it is expected that cocaine use would be stable over time. Accordingly, the hair content of cocaine would be expected to be the same at the baseline and follow-up visits. Does participation alone in an intervention study *reduce* cocaine use?

* Let the 12 pairs of cocaine measurements be denoted $(X_1, Y_1) \cdots (X_{12}, Y_{12})$.

* Focus is on the 12 differences because these represent change over 6 months:

$$d_1 = (Y_1 - X_1)$$
$$\cdots$$
$$d_{12} = (Y_{12} - X_{12})$$

   * Among n=12 participants, we observe $\overline{d}_{n=12} = -20.17$.

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

Null Hypothesis Probability Model.

The observed 12 differences in hair cocaine content is a sample, $d_1 \cdots d_{12}$, from a Normal population with unknown mean μ_d but known standard deviation $\sigma_d = 23.15$

H_O and H_A.

$$H_O : \quad \mu_d = \quad 0$$
$$H_A : \quad \mu_d < 0 \quad \underline{\text{one sided}} \quad \text{\textbf{The investigator is researching reduction in use}}$$

The test statistic is a Z-Score when the Variance is Known.

$$z_{score} = \left[\frac{\bar{d} - E(\bar{d} \mid H_O true)}{SE(\bar{d} \mid H_O true)} \right]$$

Proof by Contradiction and the definition of the p-value calculation .

The likelihood of these findings or ones more extreme if H_O is true is

$$\text{p-value} = \Pr\left[\bar{d}_{n=12} \leq -20.17 \mid \mu_d = 0 \right].$$

P-Value Calculation.

When the null hypothesis is true, the $d_1 \cdots d_{12}$ are a sample from a Normal ($\mu_d = 0$, $\sigma_d{}^2 = 23.15^2$) distribution.

Therefore, when the null is true, $\bar{d}_{n=12}$ is distributed Normal ($\mu = 0$, $\sigma_{\bar{d}}^2 = \left[\dfrac{23.15^2}{12} \right]$)

$$\text{p-value} =$$

$$pr\left[\bar{d}_{n=12} \leq -20.17 \right] = pr\left[\left(\frac{\bar{d}_{12} - 0}{\sigma_d / \sqrt{n}} \right) \leq \left(\frac{-20.17}{23.15 / \sqrt{12}} \right) \right]$$

$$= pr\left[\text{Normal}(0,1) \leq -3.02 \right] = 0.00126$$

Nature	_____	Population/	_____	Observation/	_____	Relationships/	_____	Analysis/
		Sample		**Data**		**Modeling**		**Synthesis**

"Evaluate".

 IF we assume that the null hypothesis model is true, then participation in the needle exchange trial in the pharmacy sales condition has no effect on cocaine use,

THEN there was a null hypothesis probability of .1% (0.00126 = 0.1% chance) of obtaining an observed sample mean change in hair content of –20.17 or greater reduction among 12 participants.

Interpret.

The null hypothesis, in light of the data, has led to an unlikely outcome, a 0.1% chance event. Abandon the null hypothesis. We conclude that there is statistically significant evidence of a "participation effect" on 6-month change over time in hair content of cocaine.

Now, having rejected the null hypothesis, consider that there is more than one possible explanation for what we have observed, including at least the following:

        - Trial participation results in less <u>use</u> of cocaine; ***and/or***

        - Trial participation results in less <u>detection</u> of use of cocaine.

Nature _____   **Population/** _____ **Observation/** _____   **Relationships/** _____   **Analysis/**
                  Sample   **Data**   **Modeling**   **Synthesis**

Example Scenario #2 – Variance is UNKNOWN

- Because the variance is unknown, the test statistic will be a *T-test*, a *t-score*
- Otherwise the thinking is the same.
- Suppose that the sample standard deviation of the differences is $S_d = 23.15$

H_O and H_A.

$$H_O :\ \mu_d =\ 0$$
$$H_A :\ \mu_d\ <\ 0$$

Test statistic is a *T-Score* when the Variance is UNKnown.

$$t_{score} = \left[\frac{\bar{d}\ -\ E(\bar{d}\,|\,H_O\,true)}{\hat{SE}(\bar{d}\,|\,H_O\,true)} \right]$$

P-Value Calculation.

p-value

$$=\quad pr\left[\bar{d}_{n=12} \leq -20.17 \right] =\ pr\left[\left(\frac{\bar{d}_{12}\text{-}0}{S_d\big/\sqrt{n}} \right) \leq \left(\frac{\text{-}20.17}{23.15\big/\sqrt{12}} \right) \right]$$

$$=\ pr\left[\text{Student's } t_{DF=11}\ \leq \text{-}3.02 \right] = \mathbf{0.00583}\quad \textit{notice – this is bigger than the .00126 on page 42}$$

Interpret.

The conclusion is the same.

Nature _____ Population/ _____ Observation/ _____ Relationships/ _____ Analysis/
                     Sample   **Data**   **Modeling**   **Synthesis**

2 Independent Groups, Outcome Continuous and Distributed Normal
11. Test of [μ₁ - μ₂]
(*Note – See course notes* 6. *Estimation* **page 31 for a review of sums and differences of Normals**)

In the examples presented here, it will be assumed that the variances are NOT known.   Two scenarios are considered:

- **#1. The two unknown variances are assumed equal**
- **#2. The two unknown variances are treated as unequal**

Example Scenario #1 - Equal Variances ($\sigma_1^2 = \sigma_2^2$):
(*Note: These data are hypothetical.*)

Functional status scores among patients receiving zidovudine for the treatment of AIDS were compared with those not receiving zidovudine to see if zidovudine is *beneficial*. We may assume that the scores are normally distributed with distributions that have the same variance σ^2. However, σ^2 is unknown. The data summaries are the following:

Zidovudine	Control
$n_1 = 15$	$n_2 = 22$
$\overline{X}_1 = 120$	$\overline{X}_2 = 96$
$S_1 = 40$	$S_2 = 35$

Research Question.

Do patients receiving zidovudine have higher functional status scores?

Null Hypothesis Assumptions.

$\overline{X}_1$ **is distributed Normal (μ_1 , σ^2/15)   and $\overline{X}_2$ is distributed Normal (μ_2 , σ^2/22)**

H$_O$ and H$_A$.

> $H_O : \mu_1 = \mu_2$
> $H_A : \mu_1 > \mu_2$ **one sided.**   The investigator is researching treatment benefit

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

The test statistic is a t-score.

$$t_{score} = \left[\frac{(\bar{X}_1 - \bar{X}_2) \ - \ E[(\bar{X}_1 - \bar{X}_2)|H_O \text{true}]}{\hat{SE}[(\bar{X}_1 - \bar{X}_2)|H_O \text{true}]} \right]$$

If σ^2 is unknown, what is our guess of the standard error of $\left(\bar{X}_1 - \bar{X}_2 \right)$?

Tip! See again Unit 6 (Estimation) page 51.

We learned previously (Unit 6) how to estimate the SE of the difference between two independent means, each of which is distributed Normal.

When the two variances are equal the estimate is:

$$\hat{SE}\left(\bar{X}_1 - \bar{X}_2 \right) = \sqrt{\frac{S^2_{pool}}{n_1} + \frac{S^2_{pool}}{n_2}} \qquad \textbf{where}$$

$$S^2_{pool} = \frac{(n_1 - 1)S^2_1 + (n_2 - 1)S^2_2}{(n_1 - 1) + (n_2 - 1)}$$

For these data:

$$\hat{\sigma}^2 = S^2_{pool} = \frac{\left(n_1 - 1\right)S^2_1 + \left(n_2 - 1\right)S^2_2}{\left(n_1 - 1\right) + \left(n_2 - 1\right)} = \frac{\left(15 - 1\right)40^2 + \left(22 - 1\right)35^2}{\left(15 - 1\right) + \left(22 - 1\right)} = 1375$$

$$\hat{SE}\left(\bar{X}_1 - \bar{X}_2 \right) = \sqrt{\frac{S^2_{pool}}{n_1} + \frac{S^2_{pool}}{n_2}} = \sqrt{\frac{1375}{15} + \frac{1375}{22}} = 12.42 \ ;$$

With degrees of freedom = $(n_1$-1) + $(n_2$-1) = (15-1) + (22-1) = 35.

Nature _____ **Population/** _____ **Observation/** _____ **Relationships/** _____ **Analysis/**
Sample   **Data**   **Modeling**   **Synthesis**

Proof by contradiction reasoning and the definition of the p-value calculation.

Under the null hypothesis model, the chances of this extremeness, or more so, is

$$\text{p-value} = \Pr\left[\left(\bar{X}_1 - \bar{X}_2\right) \geq (120 - 96) \mid H_O true\right].$$

Calculations.

$$\textbf{p-value} = \Pr\left[\left(\bar{X}_1 - \bar{X}_2\right) \geq (120 - 96)\right]$$

$$= \Pr\left[\frac{\left(\bar{X}_1 - \bar{X}_2\right) - (0)}{S\hat{E}\left(\bar{X}_1 - \bar{X}_2\right)} \geq \frac{(120 - 96) - (0)}{12.42}\right]$$

$$= \Pr\left[t_{score} \geq 1.93\right] \quad \textbf{where degrees of freedom} = \textbf{35}$$

$$= .03$$

Note: t_{score}=1.93 says "the observed difference in average functional status scores equal to (120-96) = 24 is 1.93 standard error units away from (greater than) the null hypothesis expected difference of 0."

"Evaluate".

Under the null hypothesis H_O, the chances that the 15 patients in the zidovudine treated group would have a mean score that is (120-96)=24 points higher than the average of the 22 scores among the control group is 3 in 100. This is a small chance, suggesting that we abandon the null hypothesis. → Reject the null hypothesis.

Interpret.

These data provide statistically significant evidence that treatment with zidovudine improves functional status.

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

Example Scenario #2 - UNequal Variances ($\sigma_1^2 \neq \sigma_2^2$):

Not surprisingly (we saw something similar in confidence interval development), the analysis is slightly different when the variances are unequal.

- The estimated SE should reflect the dissimilarity of the variances.
- With a larger # of unknowns, our degrees of freedom should be smaller.

Data are the same:

Zidovudine	Control
$n_1 = 15$	$n_2 = 22$
$\bar{X}_1 = 120$	$\bar{X}_2 = 96$
$S_1 = 40$	$S_2 = 35$

Our test statistic is still a t-score and has the same structure:

$$t_{score} = \left[\frac{(\bar{X}_1 - \bar{X}_2) - E[(\bar{X}_1 - \bar{X}_2) \mid H_o true]}{S\hat{E}[(\bar{X}_1 - \bar{X}_2) \mid H_o true]} \right]$$

But the estimate of the SE is now different. See again page 51 of the Unit 6 Notes

$$S\hat{E}(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}. \quad \textbf{For these data,}$$

$$= \sqrt{\frac{40^2}{15} + \frac{35^2}{22}}$$

$$= \mathbf{12.74}$$

Nature _____ **Population/** _____ **Observation/** _____ **Relationships/** _____ **Analysis/**
                    Sample   **Data**   **Modeling**   **Synthesis**

Eeew…. We have to use that horrible formula for the degrees of freedom that is on page 52 of the course notes 6. *Estimation*.

$$\text{Degrees of freedom} \quad = \quad \dfrac{\left(\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2} \right)^2}{\dfrac{\left(S_1^2/n_1\right)^2}{(n_1-1)} + \dfrac{\left(S_2^2/n_2\right)^2}{(n_2-1)}} . \qquad \text{In this example we get}$$

$$= \quad \dfrac{\left(\dfrac{40^2}{15} + \dfrac{35^2}{22} \right)^2}{\dfrac{\left(40^2/15\right)^2}{(14)} + \dfrac{\left(35^2/22\right)^2}{(21)}} \quad = 27.44 \approx 27 \quad \textit{Round DOWN}$$

Thus,

$$\textbf{p-value} \ = \Pr\left[\left(\bar{X}_1 - \bar{X}_2 \right) \geq (120 - 96) \right]$$

$$= \Pr\left[\dfrac{\left(\bar{X}_1 - \bar{X}_2 \right) - (0)}{S\hat{E}\left(\bar{X}_1 - \bar{X}_2 \right)} \geq \dfrac{(120 - 96) - (0)}{12.74} \right]$$

$$= \Pr\left[t_{score} \geq 1.88 \right] \quad \textbf{where degrees of freedom = 27}$$

$$= .035$$

Interpret.

The conclusion is the same – this is statistically significant evidence of a benefit of zidovudine on functional status.

<div align="center">

2 Independent Groups, Outcome Continuous and Distributed Normal
12. Test for Equality of Two Variances
(Note – See course notes *6. Estimation* page 28 for a review of the F distribution)

</div>

Example
Health services researchers are interested in patterns of length of stay (LOS) among patients entering the hospital through the emergency room as compared to those among elective hospitalizations.

Following are the data:

Group 1: Elective	Group 2: Emergency
$n_1 = 14$	$n_2 = 11$
$S_1 = 10.9$ days	$S_2 = 4.2$ days

Research Question.

Does the variability of LOS *differ* between emergency and elective patients?

Assumptions.

Two independent samples, each a simple random sample from a Normal distribution, $X_1 \ldots X_{n1}$ distributed Normal (μ_1, σ_1^2) and $Y_1 \ldots Y_{n2}$ distributed Normal (μ_2, σ_2^2)

H_O and H_A.

$$H_O : \sigma_1 = \sigma_2$$
$$H_A : \sigma_1 \neq \sigma_2 \ \textbf{two sided.}$$ The investigator is researching differences in variability, either way

Test statistic/Pivotal Quantity is an F-statistic.

Remark – When the means of continuous variables are compared, the analysis considers their difference. When two variances are compared, the analysis focuses on their quotient

Under the assumption of the null hypothesis model, the test statistic F defined as

$$F = \left[\frac{S_1^2}{S_2^2} \right] \quad \text{is distributed F with numerator df} = (n_1\text{-}1) \text{ and denominator df} = (n_2\text{-}1)$$

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

"Evaluation" rule.

Under the null hypothesis (H_O) model, the likelihood of these findings, or ones more extreme in either direction, is the following probability calculation:

$$\text{p-value} \; = \; (2) \; \Pr\left[F_{df=13,10} \geq \left(\frac{S_1^2}{S_2^2} \right) \mid H_O \, true \right].$$

Calculations.
p-value

$$= (2)\Pr\left[F_{df=13,10} \geq \left(\frac{S_1^2}{S_2^2} \right) \mid H_O \, true \right] = (2)\Pr\left[F_{df=13,10} \geq \left(\frac{10.9^2}{4.2^2} \right) \right] = (2)\Pr\left[F_{df=13,10} \geq 6.73 \right]$$

= (2) (0.0024)

= 0.0048

Dear class, The Daniel Soper F-distribution calculator doesn't provide right tail areas. So, here, I use the epitools.ausvet.com F distribution calculator. The only hitch is that, in order for it to do its thing, you have to supply a type I error value. So I input type I error, alpha = .05, even though this was not my goal – cb.

Get P and critical values for the F distribution **Results**

Input Values Calculate P values from the F distribution, level. . *Summary results*

Test (F) statistic:	6.73	Inputs are the test statistic, degrees of free
Degrees of freedom for the numerator:	13	The program outputs the P value correspor
Degrees of freedom for the denominator:	10	well as a summary and plot of the distributi
Alpha (significance) level :	.05	

Submit

	Value
P-value (F = 6.73)	0.0024
Critical value (alpha = 0.05)	2.89
P(F <= 6.73)	0.9976
P(F >= 6.73)	0.0024

http://epitools.ausvet.com.au/content.php?page=f_dist&t=6.73&df1=13&df2=10&Conf=.05

"Evaluate".

Under the null hypothesis (H_O) model that the variances are equal, the likelihood of an observed ratio of sample variances being as far away (in either direction) from 1 as the value 6.73 is approximately 4.8 chances in 1000. This is a small likelihood.

Nature _____ **Population/** _____ **Observation/** _____ **Relationships/** _____ **Analysis/**
       Sample   **Data**   **Modeling**   **Synthesis**

 Interpret.

Application of the null hypothesis model has led to an unlikely outcome. Reject the null hypothesis. Conclude that there is statistically significant evidence of a difference in the variability of length of stay for elective patients versus for emergency patients.

1 Group, Outcome Discrete and Distributed Binomial
13. Test for Proportion π
(Note – See course notes *4. Bernoulli and Binomial* pages 16-17 for a review of the Binomial distribution)

13.1 Exact Test – Use when the sample size is small

Tip! **-** The exact test described here is used in small sample size settings. Use the normal approximation test described in the next pages (Section 13.2, page 55) for larger sample sizes. See page 55 for a guideline to follow when choosing between the "exact" versus "normal approximation" tests.

Research Question:

In an ICU study, data were collected on 20 consecutive patients. Four (4) of the patients died in the hospital. Is there evidence that the mortality rate at Baystate Medical Center is different than 25%?

Null Hypothesis Probability Assumptions

- The data is the outcome of some Binomial random variable with number of trials n=20:
  Binomial (n=20, π).

- The number of events (mortality) observed in this sample of 20 trials is X=4

H_O and H_A.

$$H_O : \pi = 0.25$$
$$H_A :\ \pi \neq 0.25 \quad \textbf{two sided}$$

Tip – *Calculate the null hypothesis model value of the expected number of events of mortality. This is $E [X | H_O : \pi = 0.25]$. You will need this to calculate your p-value.*

$$E [X | H_O : \pi = 0.25]\ =\ [n] [0.25]\ =\ [20] [0.25] =\ 5$$

Nature _____ **Population/** _____ **Observation/** _____ **Relationships/** _____ **Analysis/**
                      Sample   **Data**   **Modeling**   **Synthesis**

The Test Statistic is X.

X = number of events of mortality

Proof by Contradiction Reasoning and the definition of the p-value calculation.

Because the alternative hypothesis is <u>two sided</u>, the p-value calculation we want to
do here answers the question: **"If H_O is true, what are the chances of obtaining
a number of events of death as far away from E [X | H_O : π = 0.25] = 5 in either direction?"**

The observed value of 4 deaths is 1 death different from the mean 5 in the left direction.
1 death different from the mean 5 in the other direction is 6 deaths. Thus,

p-value = Prob [Binomial(20, .25) $\geq$ 6]   + Prob [Binomial(20, .25) $\leq$ 4]

P-Value Calculation.

p-value = Prob [Binomial(20, .25) $\geq$ 6]   + Prob [Binomial(20, .25) $\leq$ 4]

=   0.4148   +   0.3828

=   0.7976

"Evaluate".
Under the null hypothesis H_O, that the mortality rate at Baystate is 0.25, the likelihood of an observed mortality
rate as small or smaller than 4/20 OR as large or larger than 6/20 is approximately 80 chances in 100.   ….
Aside: this is obviously too small a study to be useful!

Interpret.

The assumption of the null hypothesis model has not led to an unlikely outcome. Do NOT reject the null
hypothesis. Conclude that these data do not provide statistically significant evidence for its rejection.

Nature _____ **Population/** _____ **Observation/** _____ **Relationships/** _____ **Analysis/**
                        Sample   **Data**   **Modeling**   **Synthesis**

13.2 Normal Approximation (Z-Score) Test

Guideline and Definition of the Normal Approximation Test

Guidelines: Use a Z-score test approximation when the sample size is moderate or large. We can do this because the central limit theorem applies (proof not shown). As a rough rule of thumb, you can use the following normal approximation test when the following holds:

$$(n)\,(\pi_{null})\,(1 - \pi_{null})\ \geq\ 5$$

Definition: $\bar{X}$ is distributed Normal$(\pi, \dfrac{\pi(1-\pi)}{N})$ approximately

Research Question:

In an ICU study, data was collected on 200 consecutive patients. 40 of the patients died in the hospital. Is there evidence that the mortality rate at Baystate Medical Center is different (in either direction) than 25%?

Null Hypothesis Probability Assumptions

- Data are a random sample of patients (over time), and the outcome of mortality, X=(# patients among the 200 who die in hospital) has exact distribution that is Binomial (N=200, π).

- The observed number of events of mortality is X=40

- A quick check of the guidelines at the top of this page yields $(n)(\pi)(1 - \pi) = (200)(.25)(.75) = 37.5$. Thus, we have plenty of sample size for using a normal Z-score approximation test. In this setting, application of the central limit theorem to a binomial random variable outcome gives the following:

$$\bar{X} \text{ is distributed Normal}(\ \pi, \dfrac{\pi\,(1-\pi)}{N}) \text{ approximately}$$

- The observed proportion (the observed % who died) is $\bar{X}=40/200=0.20$

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

H_O and H_A.

$$H_O : \pi = 0.25$$
$$H_A : \pi \neq 0.25 \quad \text{two sided}$$

The Test statistic is a z-score.

$$\text{Z-score} = \frac{\overline{X} - \pi_0}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{N}}} = \frac{\overline{X} - 0.25}{\sqrt{\dfrac{0.25(0.75)}{200}}}$$

"Proof by Contradiction Reasoning and the definition of the p-value calculation.

Under the assumption that the null hypothesis is true, the probability of obtaining an observed sample mortality **as** *different* from the expected value of 25% as the observed 20% is a two sided calculation:

$$\text{p-value} = (2) \ \Pr\left[\text{Normal}(0,1) \leq \left(\frac{\overline{X} - 0.25}{\sqrt{\dfrac{0.25(0.75)}{200}}} \right) \right].$$

P-Value Calculation.

$$\textbf{p-value} = \textbf{(2)} \ \Pr\left[\text{Normal}(0,1) \leq \left(\frac{0.20 - 0.25}{\sqrt{\dfrac{0.25(0.75)}{200}}} \right) \right] = \textbf{(2)} \ \Pr\left[\text{Normal}(0,1) \leq -1.63 \right]$$

= (2) (0.051)

= 0.102

"Evaluate".

Under the null hypothesis H_O, that the mortality rate at Baystate is 0.25, the probability of an observed mortality rate as far away (in either direction) as 20% is .102, or approximately 10 chances in 100. The null hypothesis, when applied to the data, has not led to an unlikely outcome.

Interpret.

Do NOT reject the null hypothesis. Conclude that, in this sample, the observed mortality rate of 20% is consistent with the hypothesized rate of 25%.

Nature _____ **Population/** _____ **Observation/** _____ **Relationships/** _____ **Analysis/**
Sample   **Data**   **Modeling**   **Synthesis**

2 Independent Groups, Outcome Discrete and Distributed Binomial
14.   Test for Equality of Proportions via H$_O$: [$\pi_1 - \pi_2$]=0

A normal theory (Z-score) approximate test is described.

Example

Consider again the needle exchange trial introduced previously.   Among the preliminary aims was an analysis to identify variables that are associated with both randomization assignment and outcome. Such variables are potential confounders of response to intervention.

The literature suggests that women might respond differently to intervention than men. Therefore, an interim analysis sought to determine if there are gender differences in randomization assignment.

Among n=101 eligible and followed as of May 31, 1998:

Pharmacy Sales	Pharmacy Sales + Needle Exchange
$n_1 = 53$	$n_2 = 48$
# women = 9 = X$_1$	# women = 13 = X$_2$
% women = 17.0 = $\bar{X}_1$	% women = 27.1 = $\bar{X}_2$

Research Question.

Is the proportion of women in the pharmacy sales + needle exchange condition (27.1%) significantly different than the proportion of women in the pharmacy sales condition (17.0%), considering the limitations of sample size (53 and 48, respectively)?

Null Hypothesis Assumptions.

- The event of "female gender" is the outcome of interest. In each group (pharmacy sales versus pharmacy sales + needle exchange), the number (X) who are gender = female is a Binomial random variable.

- We will represent the proportions of women in the two groups as $\bar{X}_1$ and $\bar{X}_2$.
  $\bar{X}_1$ is distributed Binomial (n$_1$=53, π_1)   and $\bar{X}_2$ is distributed Binomial (n$_2$=48, π_2) where

$$\pi_1 = \text{ Proportion "female gender" in group = Pharmacy Sales}$$
$$\pi_2 = \text{ Proportion "female gender" in group = (Pharmacy Sales + Needle Exchange)}$$

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

 H$_O$ and H$_A$.

$$H_O : \pi_1 = \pi_2 \qquad\qquad\qquad \text{The representation of women in both groups is the same.}$$
$$H_A : \pi_1 \neq \pi_2 \qquad \underline{\text{two sided}} \qquad \text{The representation of women in both groups NOT the same}$$

The Test statistic is now a Z-score.

$$z_{score} = \left[\frac{(\bar{X}_1 - \bar{X}_2) \; - \; E[(\bar{X}_1 - \bar{X}_2)|H_O\text{true}]}{S\hat{E}[(\bar{X}_1 - \bar{X}_2)|H_O\text{true}]} \right]$$

Two Independent Binomials – Calculation of $S\hat{E}\left[\left(\bar{X}_1 - \bar{X}_2\right) \mid H_O true\right]$

$$S\hat{E}\left[\left(\bar{X}_1 - \bar{X}_2\right) \mid H_O\right] = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n_1} + \frac{\hat{\pi}(1-\hat{\pi})}{n_2}} \quad \text{where}$$

$\hat{\pi}$ **is our best guess of the common** π **and**

$$\hat{\pi} = \left[\frac{X_1 + X_2}{n_1 + n_2} \right]. \quad \textit{Notice that this is just the overall proportion}$$

For these data:

$$\hat{\pi} = \left[\frac{X_1 + X_2}{n_1 + n_2} \right] = \left[\frac{9+13}{53+48} \right] = .218$$

$$S\hat{E}\left(\bar{X}_1 - \bar{X}_2\right) = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n_1} + \frac{\hat{\pi}(1-\hat{\pi})}{n_2}} = \sqrt{\frac{.218(1-.218)}{53} + \frac{.218(1-.218)}{48}} = .0823$$

Nature _____ **Population/** _____ **Observation/** _____ **Relationships/** _____ **Analysis/**
                        Sample   **Data**   **Modeling**   **Synthesis**

"Proof by Contradiction Reasoning and the definition of the p-value calculation.

In the needle exchange trial, we want to know the null hypothesis model chances of obtaining a difference in the proportion of women in the two groups as great or greater than $|.271 - .170| = .1010$

The required p-value calculation is thus

$$\textbf{p-value} \; = \; 2\Pr\left[|(\bar{X}_2 - \bar{X}_1)| \; \geq \; |(.271 - .170)| \right].$$

P-Value calculation.

$$\textbf{p-value} \; = 2\Pr\left[(\bar{X}_2 - \bar{X}_1) \geq (.271 - .170) \right]$$

$$= 2\Pr\left[\frac{(\bar{X}_2 - \bar{X}_1) - E(\bar{X}_2 - \bar{X}_1)}{S\hat{E}(\bar{X}_2 - \bar{X}_1)} \geq \frac{(.271 - .170) - (0)}{.0823} \right]$$

$$= 2\Pr[z - score \geq 1.23] = 2[.10935]$$

$$= .22$$

$z_{score}=$**1.23 says** "the observed difference in % women in the two randomization groups equal to $(.271-.170) = .1010$ is 1.23 standard error units greater than the expected difference of 0 when the null hypothesis is true."

"Evaluate".

With sample sizes of 53 and 48, there was a reasonable chance, 22% chance, of obtaining a discrepancy in the % women in the two groups equal to 10 percentage points or more.

Interpret.

Application of the null hypothesis has NOT led to an unlikely outcome. Retain the null hypothesis and conclude that there is not a statistically significant difference in the proportion of women in the two study conditions among the 101 available for interim analysis.

Nature _____ **Population/** _____ **Observation/** _____ **Relationships/** _____ **Analysis/**
                          Sample   **Data**   **Modeling**   **Synthesis**

Appendix
URL's for the Computation of Probabilities

The Normal (0,1) Distribution
http://davidmlane.com/hyperstat/z_table.html

The Student's t Distribution
http://surfstat.anu.edu.au/surfstat-home/tables/t.php

The Chi Square Distribution
http://surfstat.anu.edu.au/surfstat-home/tables/chi.php

The F-Distribution (note – You'll need to scroll down to get to the F calculator)
http://www.socr.ucla.edu/Applets.dir/Normal_T_Chi2_F_Tables.htm

http://epitools.ausvet.com.au/content.php?page=f_dist&t=6.73&df1=13&df2=10&Conf=.05

Nature _____ **Population/** _____ **Observation/** _____ **Relationships/** _____ **Analysis/**
Sample   **Data**   **Modeling**   **Synthesis**