

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

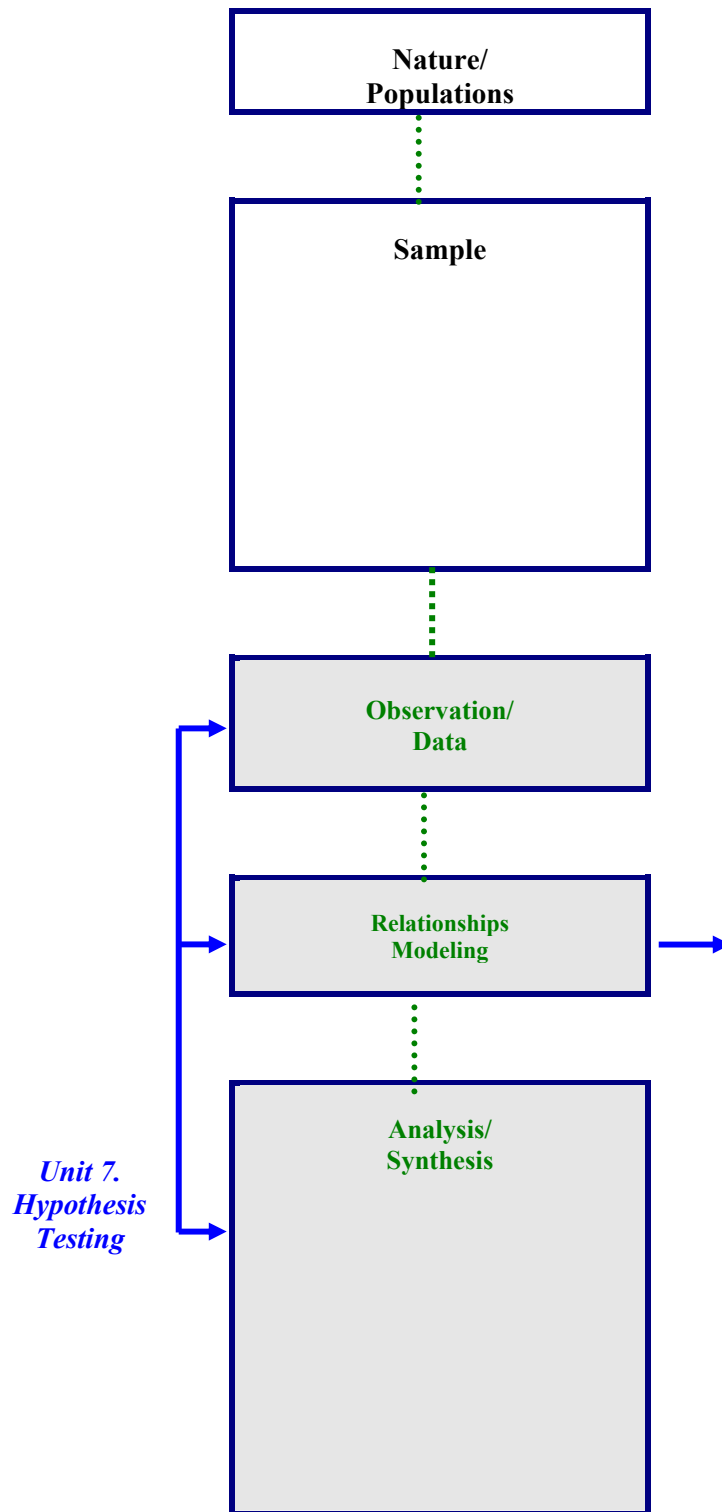
Nature

Population/
SampleObservation/
DataRelationships/
ModelingAnalysis/
Synthesis

Table of Contents

Topic	1. Unit Roadmap.....	3
	2. Learning Objectives	4
	3. The Logic of Hypothesis Testing	18
	3.1. One Sided versus Two Sided Tests	18
	4. Beware the Statistical Hypothesis Test	19
	5. Introduction to Type I, II Error and Statistical Power	22
	6. Normal: Test for μ , σ^2 Known	28
	7. Normal: Test for μ , σ^2 Known – Critical Region Approach	31
	8. Normal: Test for μ , σ^2 Unknown	35
	9. Normal: Test for σ^2	38
	10. Normal Test for $\mu_{\text{DIFFERENCE}}$ – Paired Data Setting	41
	11. Normal: Test for $[\mu_1 - \mu_2]$ – Two Independent Groups	45
	12. Normal: Test for Equality of Two Variances (σ_1^2/σ_2^2).....	50
	13. Single Binomial: Test for Proportion π	53
	13.1 Exact Test	53
	13.2 Normal Approximation Test	55
	14. Two Binomials: Test for $[\pi_1 - \pi_2]$ – Two Independent Groups.....	57
	Appendix	
	URL’s for the Computation of Probabilities	60

1. Unit Roadmap



Statistical significance assessment is a tool that informs our understanding of nature but ***does not determine 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. Call this 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 is rejected.

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

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

Nature
Population/
Sample
Observation/
Data
Relationships/
Modeling
Analysis/
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_0) 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.

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 so nicely. So here it is, in its entirety. Hopefully, I have given the proper attribution:

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/
Sample

Observation/
Data

Relationships/
Modeling

Analysis/
Synthesis

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

1. “Inference involves testing a hypothesis..”

An important reminder here (and one that was already mentioned on page 3) is one of perspective. It is the hypotheses that are considered and then abandoned or retained, depending on their consistency with the data. The data are the “fixed.” Not the other way around. Be careful not to make statements such as, “the data are inconsistent with a hypothesis”. Instead, you should say, “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** probability distribution. **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 $\mu=100$.**

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 $\mu=100$ and your observed sample mean is $\bar{X}=81$ then a one sided p-value might be the calculation of $\Pr[\bar{X} \leq 81]$ under the null hypothesis assumption that $\mu=100$**

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

Steps in Hypothesis Testing

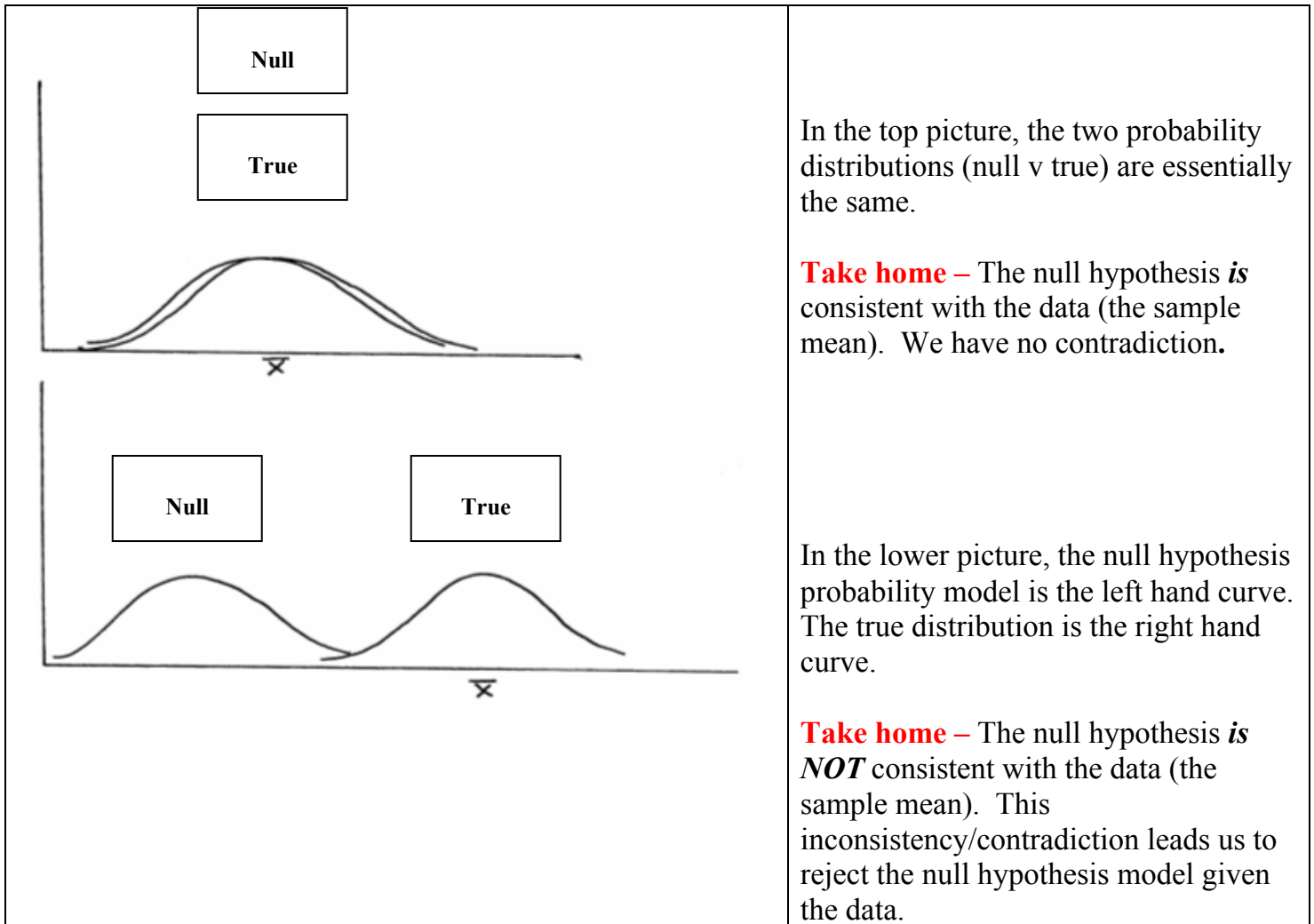
1. Identify the research question.
2. State the null hypothesis assumptions necessary for computing probabilities.
3. Specify H_0 and H_A .
4. “Reason” an appropriate test statistic.
5. Specify an “evaluation” rule.
6. Perform the calculations.
7. “Evaluate” findings and report.
8. Interpret in the context of biological relevance.
9. (Accompany the procedure with an appropriate confidence interval)

Nature

Population/
SampleObservation/
DataRelationships/
ModelingAnalysis/
Synthesis

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.



In the top picture, the two probability distributions (null v true) are essentially the same.

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

In the lower picture, the null hypothesis probability model is the left hand curve. The true distribution is the right hand curve.

Take home – The null hypothesis *is NOT* consistent with the data (the sample mean). This inconsistency/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.

	<p>Step 1 – Begin by assuming that the null hypothesis is true Under this assumption, 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. <i>Note</i> – We have not yet taken into consideration the given observed data. This comes next.</p> <p>Step 2 – Consider the observed data ... This picture represents the given data. Notice that there is no probability distribution shown. This is a reminder that, in actuality, we don’t actually know which distribution gave rise to the data.</p> <p>Step 3 – Argue “yes” or “no” does data contradict null. In this picture, the true distribution that gave rise to the data is on the right. The null hypothesis model is on the left. The shaded area is a probability calculation under the assumption that the null is true:</p> $\Pr [\bar{X} \geq \text{observed}].$ <p>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?”</p> <p>Small probability says “Assuming the null leads to an unlikely event” Large probability says “Assuming the null leads to plausible event”</p>
--	--

$\Pr [\bar{X} \geq \text{observed value}]$ and rejection or non-rejection of the null hypothesis involves the question: “what are the chances of a sample mean as extreme or more extreme”.

	<p>Scenario 1 - NULL is true $\Pr [\bar{X} \geq \text{observed value}] = \text{large}$</p> <ul style="list-style-type: none"> Observed sample mean <i>is</i> close to null mean. Assuming the null hypothesis model leads to a relatively large probability that \bar{X} is the value actually observed or more extreme. (Note – extreme is always in the direction of the alternative) Statistical decision - “do NOT reject the null”. <p>Scenario 2 - ALTERNATIVE is true $\Pr [\bar{X} \geq \text{observed value}] = \text{small}$</p> <ul style="list-style-type: none"> Observed sample mean is <i>not</i> close to null mean. Assuming the null hypothesis model leads to a small probability that \bar{X} is the value actually observed or more extreme. Statistical decision - “REJECT the null”.
--	---

$$\text{p-value} = \Pr [\text{Test statistic (eg } \bar{X}) = \text{observed or more extreme}]$$

EG - “If I assume that the null hypothesis is true and use this model, what was my probability of obtaining \bar{X} as large or larger than the value that I observed?”.

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



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. Next, suppose that the new therapy is administered to 100 cancer patients. It is observed that they experience instead an average survival time of 46.9 months. **Is the observed survival under the new treatment statistically significantly improved (relative to standard care)?**

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

Begin by assuming the null hypothesis is true and state the corresponding null hypothesis probability model that you will use for computing the p-value probability value

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 $\sigma^2 = 43.3^2$ months². *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

$$H_0: \mu_{\text{true}} = \mu_0 \leq 38.3 \text{ months}$$

$$H_A: \mu_{\text{true}} = \mu_A > 38.3 \text{ months}$$

Note – Strictly speaking, the null and alternative hypotheses must span all possibilities. That's why they are written as you see them here. However only one value for the null hypothesis mean can be used to calculate a p-value probability. We choose the value that is closest to the alternative hypothesis. Here, it is $\mu_0 = 38.3$. Rationale is to be conservative.

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

3. Reason “proof by contradiction”

IF: the null hypothesis is true, so that $\mu_{true} = \mu_o = 38.3$

THEN: what are the chances that a mean of 100 survival times will be “as extreme or more extreme than the value 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

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

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. Perform the calculation of such chances presuming H_o true.

Under the assumption that the null hypothesis is true:

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

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”.



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

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 mean as far away from 38.3 as the value of 46.9”

$$\Pr[\bar{X}_{n=100} \geq 46.9 | \mu_{true} = \mu_{null} = 38.3]$$

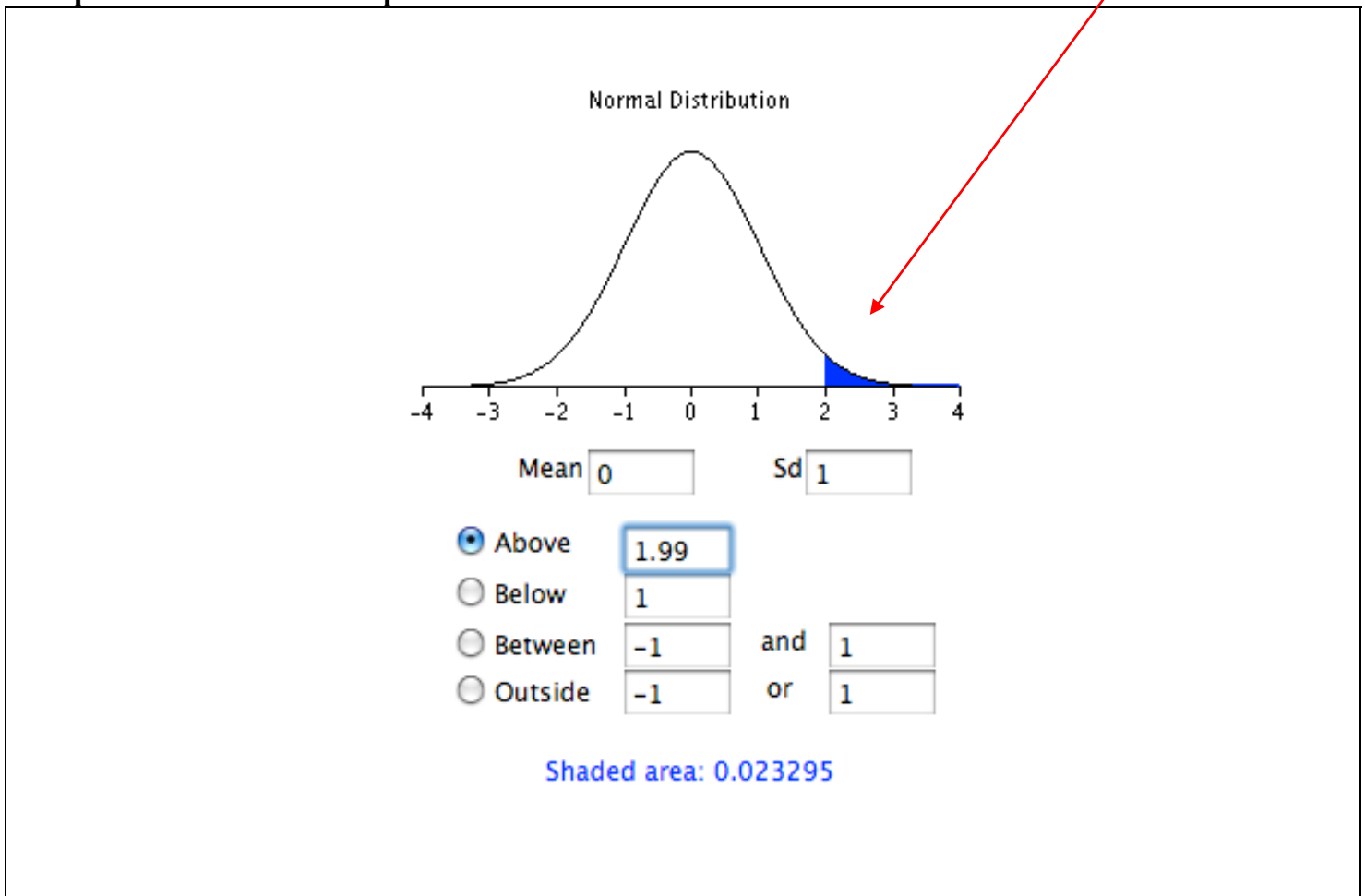
$$= \Pr[Z\text{-score} \geq 1.99] = .02$$

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, if not contradicted. →

Statistical rejection of the null hypothesis.

Graphical illustration of a p-value calculation

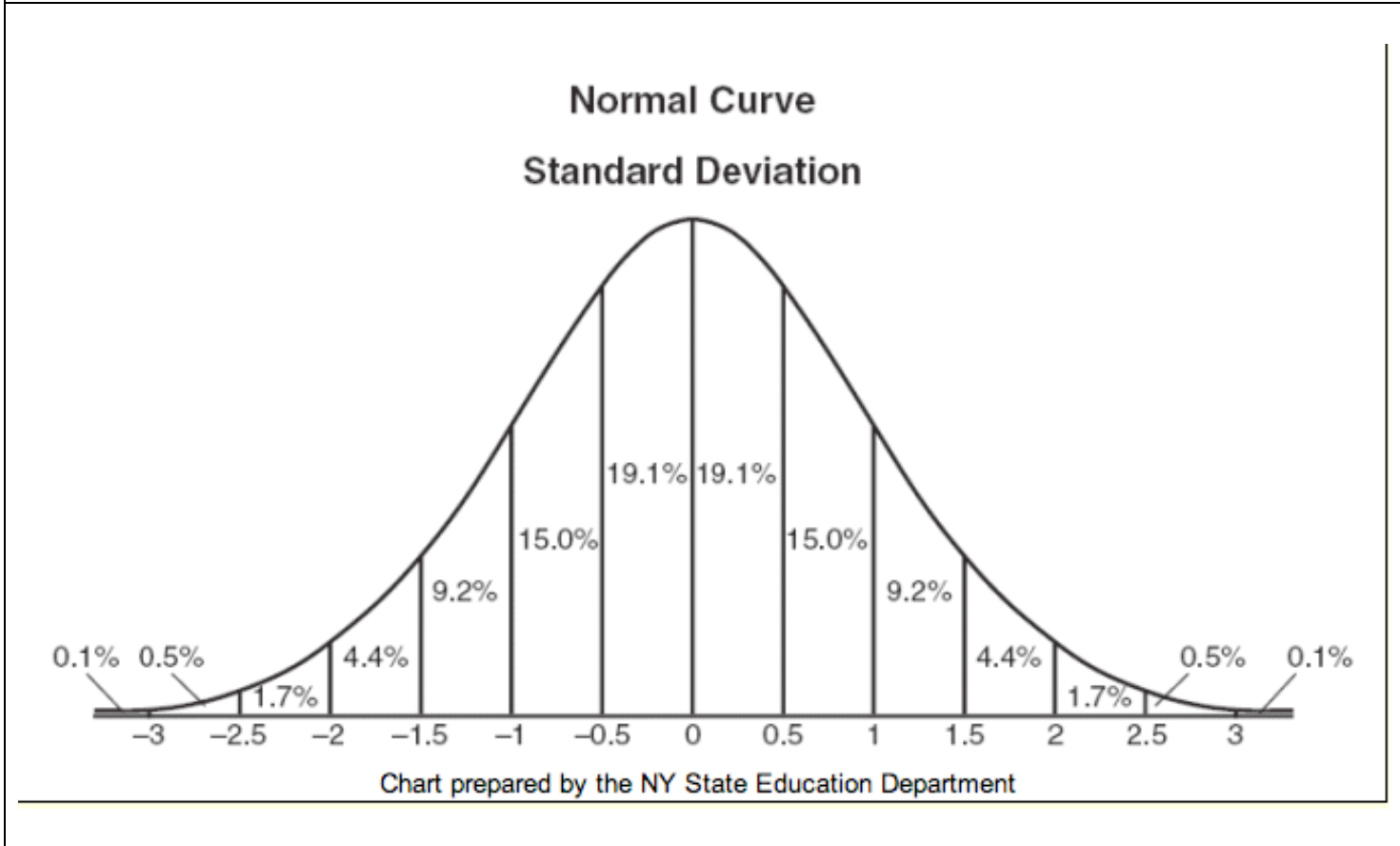


The Z-score is a Signal-to-Noise Comparison

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

TIP!

Often, Test Statistic = Z-score
Use the Normal(0,1) distribution curve to assess extremeness



Example –

You are reading a manuscript and you see a sample mean and its SE. Of interest to you, as you are reading, is a rough sense of the extent to which the data are consistent with some hypothesis. Using the hypothesis, 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 1% likelihood (0.5% + 0.5%).
- * Translation: – The probability that an observed sample mean (distributed normal) is as extreme or more extreme than a distance of 2.5 SE units away from its expected value is 0.5% + 0.5% = 1%.

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:

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

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.

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

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

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

Step 2 – Calculate p-value = Probability of “extremeness” in either of two directions

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

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.

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, rendering the apparent finding of superiority of “A” 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 sample data that are quite distant from its expected value.

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/
SampleObservation/
DataRelationships/
ModelingAnalysis/
Synthesis

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

A statistical hypothesis test uses probabilities based only on the null hypothesis (H_0) model!

- The proof by contradiction thinking asks us to presume that H_0 is true and to then examine the plausibility of our data in light of this assumption.
 - We either reject it, or we fail to reject.
 - We do not prove that H_0 is correct.

We can summarize the results of statistical hypothesis testing as follows:

		<u>Truth</u>	
		Null True	Alternative True
<u>Decision</u>	Retain null	☺	β or type II error
	Reject null	α or type I error	☺

Introduction to Type I Error

- IF H_0 is true and we (incorrectly) reject H_0
 - We have made a type I error
 - We can calculate its probability as $\text{Pr}[\text{type I error}] = \alpha$

Introduction to Type II Error

- IF H_a is true and we (incorrectly) fail to reject H_0
 - We have made a type II error
 - We must have a specific H_a model before we can calculate $\text{Pr}[\text{type II error}] = \beta$

Introduction to Power

- IF H_a is true and we (correctly) reject H_0
 - This occurs with probability = $(1 - \beta)$ which we call the **“POWER”**

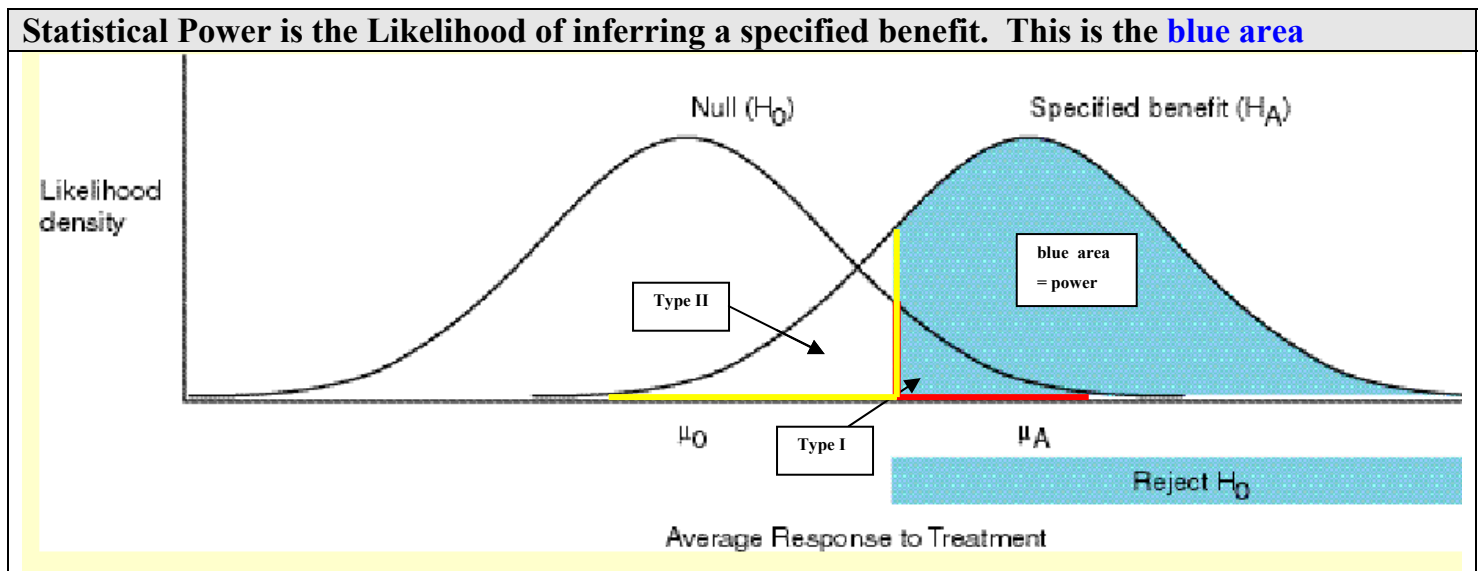


The goal is to get the right answer (power).

Either type of error is undesirable. We'd like both α and β to be small.

- Sample size calculations! Larger sample sizes will lower both 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.

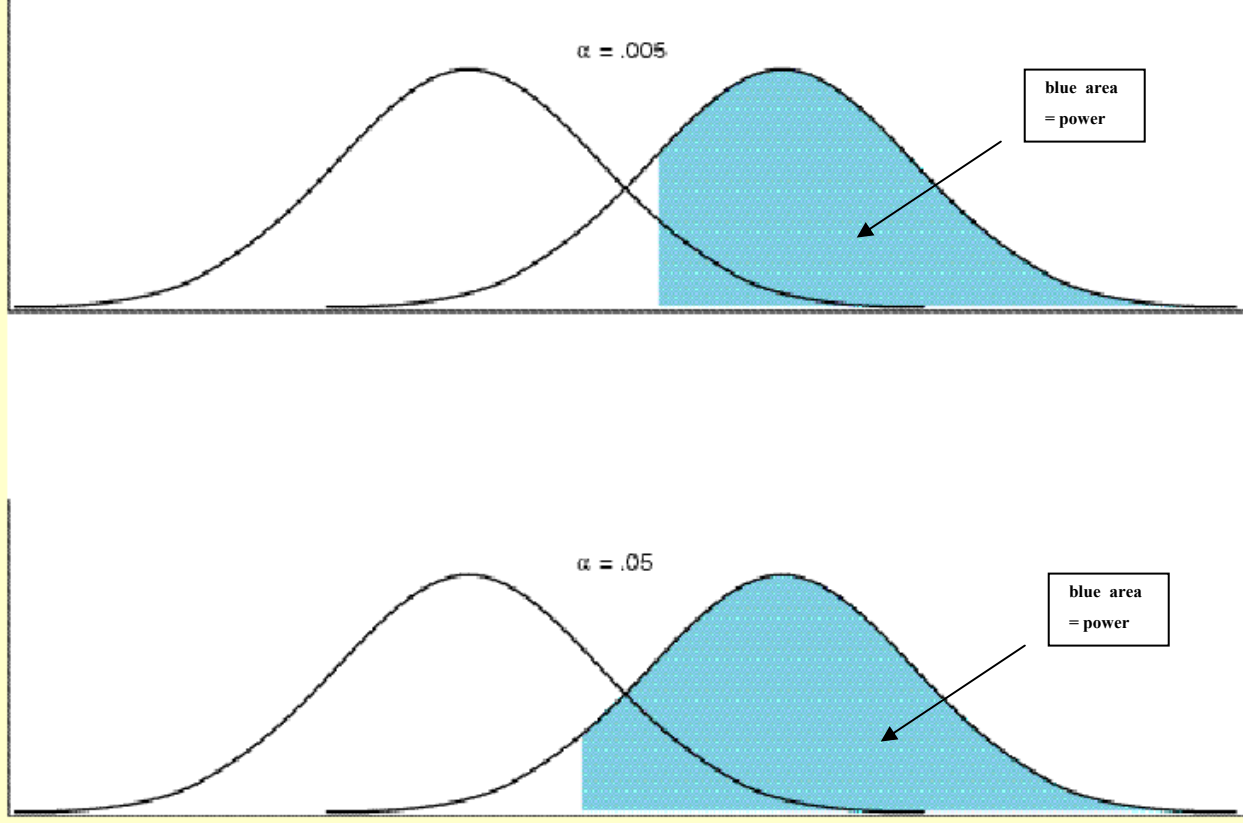


Key:

- **Blue ribbon along the horizontal axis with “reject H_0 ” typed inside:** The values of the sample average that will prompt rejection of the null hypothesis, also called the **critical region**.
- **Blue area under the Null (H_0) curve: The type I error.** 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.** 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



- 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.

Nature

Population/
Sample

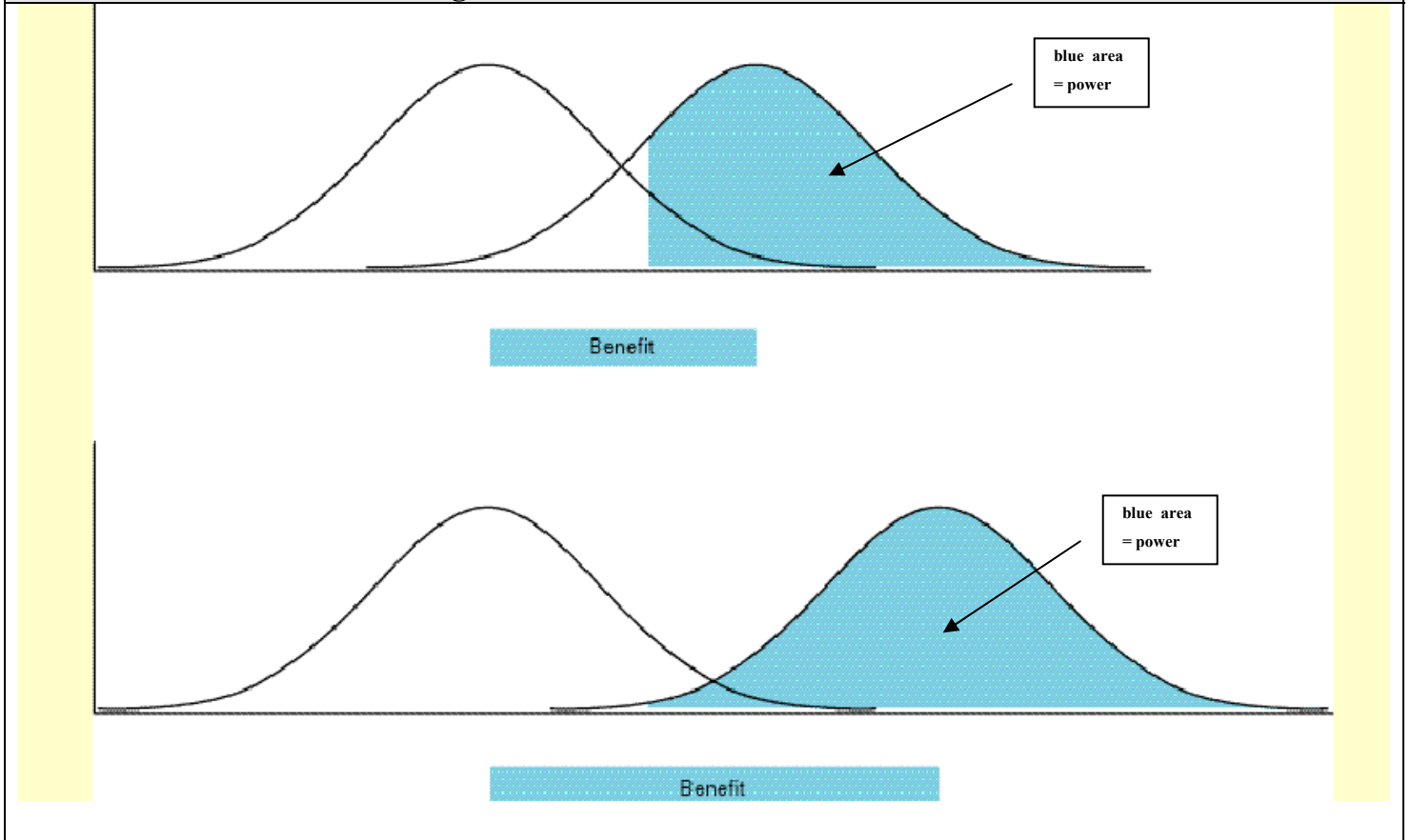
Observation/
Data

Relationships/
Modeling

Analysis/
Synthesis

The Power of a Study Depends on Four Parameters

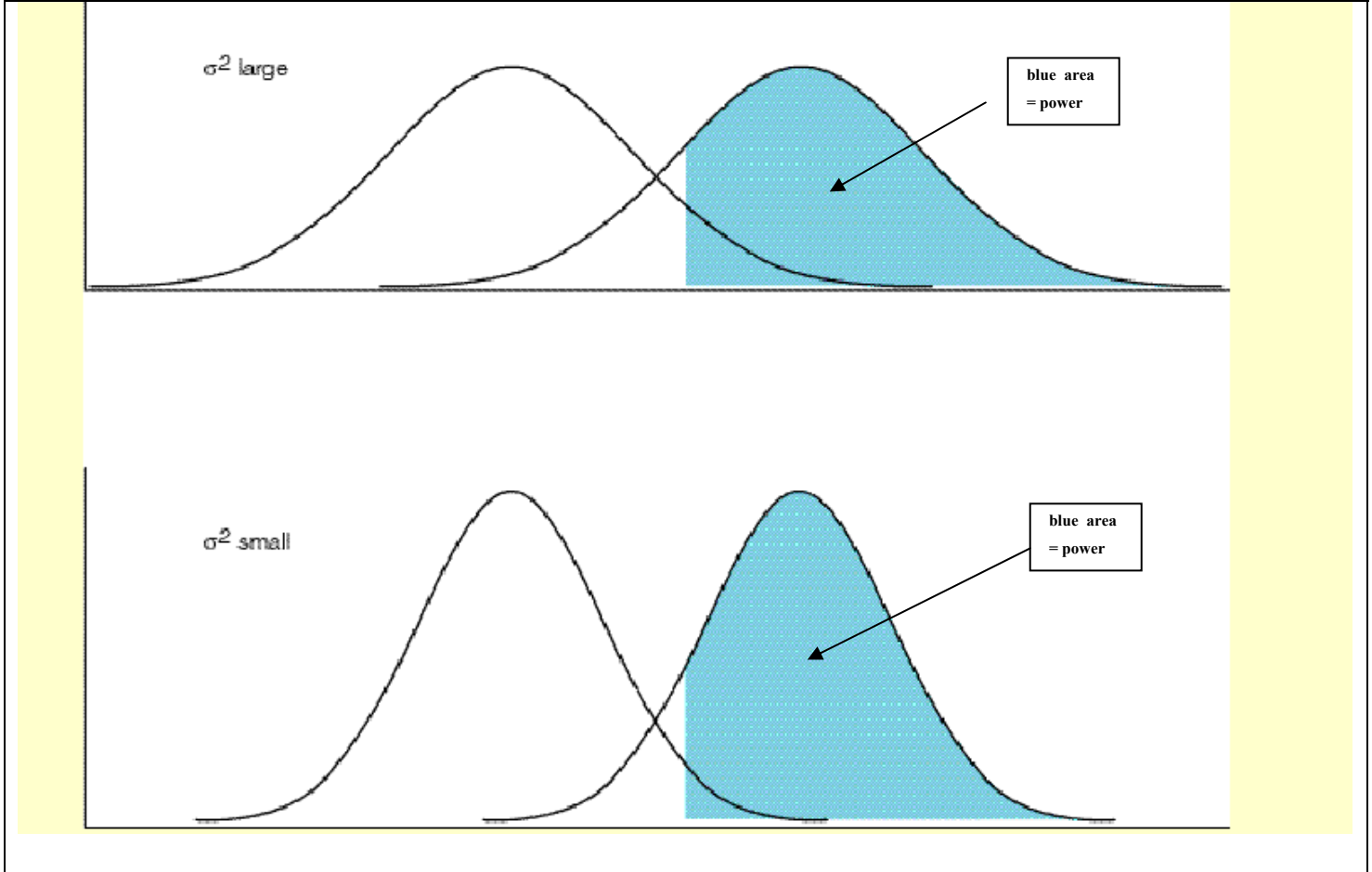
2. The Benefit Worth Detecting



- 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

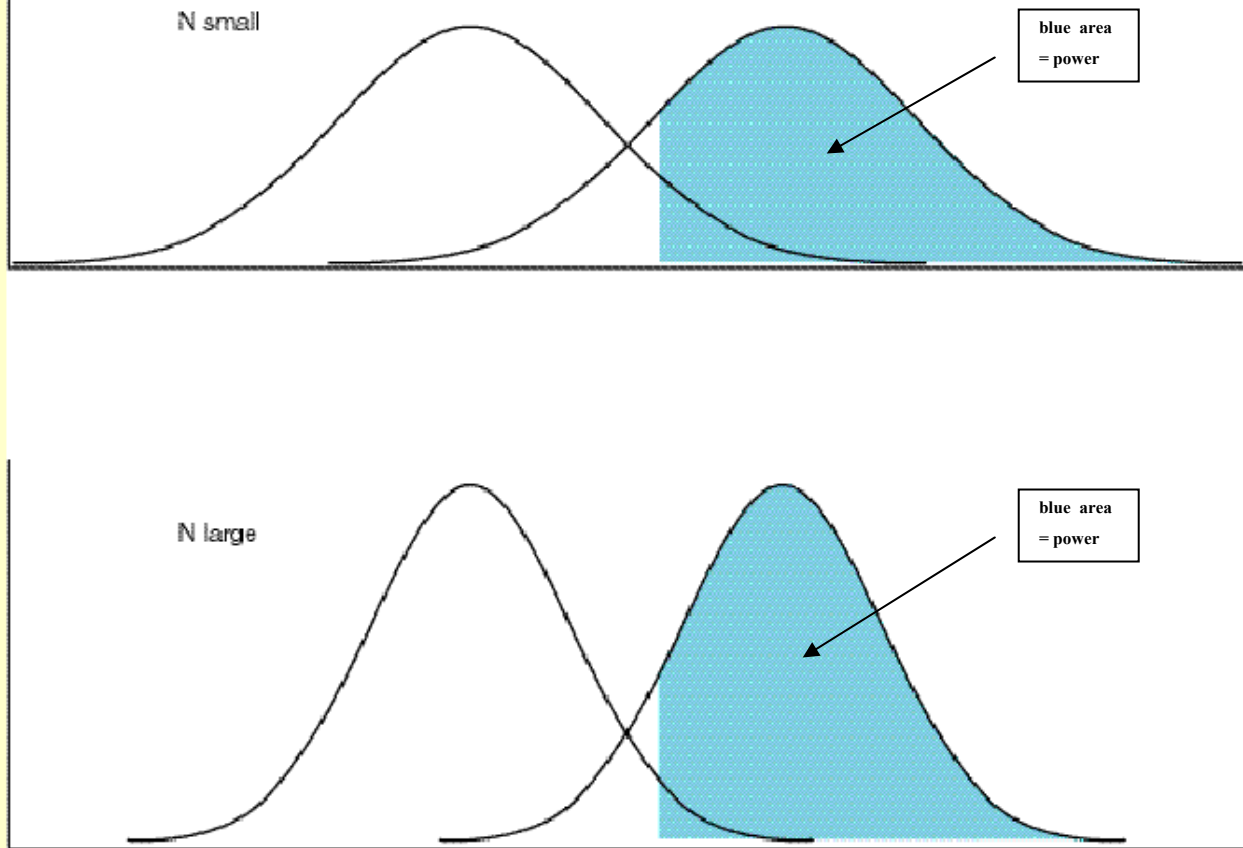
3. Biological Variability (“Noise”)



- 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/
Sample ——— Observation/
Data ——— Relationships/
Modeling ——— Analysis/
Synthesis

The Power of a Study Depends on Four Parameters
4. Sample Size (“Design”)



- 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).

6. Normal: Test for μ , σ^2 Known

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

- **Tip** - Reread Unit 6 notes pp 20-32 for review of the student's t, chi square, and F distributions.
- This section includes an introduction to the idea of a **pivotal quantity** (choice of test statistic).
- **Critical region** tests are also introduced.
- As previously mentioned, the steps are very 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 $\sigma^2_{\text{known}} = 43.3^2$ months squared. Is this statistically significant evidence of improved survival?

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

Null Hypothesis Probability Model Assumptions.

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

Specify the null and alternative hypotheses

$H_0: \mu_{true} = \mu_o \leq 38.3$ 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 is true, so that $\mu_{true} = \mu_o = 38.3$

THEN: what is the probability (this will be our p-value) that we observed a mean of 100 survival times that is as extreme as 46.9 months (meaning 46.9 or greater) relative to the null hypothesis expected mean 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 inconsistent conclusion if there is at most a small chance 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

$$\Pr[\bar{X}_{n=100} \geq 46.9 | \mu_{true} = \mu_o = 38.3]$$

Nature

Population/
SampleObservation/
DataRelationships/
ModelingAnalysis/
Synthesis

The test statistic is a Z-Score

Under the assumption that the null hypothesis is true:

- X_1, X_2, \dots, X_{100} is each distributed $\text{Normal}(\mu = 38.3, \sigma^2 = 43.3^2)$.
- $\bar{X}_{n=100}$ is distributed $\text{Normal}(\mu = 38.3, \sigma^2 = 43.3^2/100)$
- We'll use as our test statistic the z-score standardization of $\bar{X}_{n=100}$, obtained under the assumption that the null hypothesis is correct.

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

P-value calculation

$$\begin{aligned} \text{p-value} &= \Pr[\bar{X}_{n=100} \geq 46.9 | \mu_{\text{true}} = \mu_{\text{null}} = 38.3] \\ &= \Pr[Z\text{-score} \geq 1.99] = .02 \end{aligned}$$

“Evaluate”.

IF the new therapy elicits 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 is 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 when considered in light of the observed data. **Reject the null hypothesis.**



7. Normal: 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?

Critical Region Reasoning.

- We agree *in advance (prior to collecting data)* that we will judge “extreme” values as inconsistent with the null hypothesis and then reject the null hypothesis, even though null hypothesis extreme values are theoretically possible.
- Should such “extreme” occur and we *incorrectly reject a true null hypothesis* we will have made a *type I error*.
- In developing a critical region test, we determine, ahead of time, the set of extreme values (this is called the *critical region*) that will prompt (rightly or wrongly!) rejection of the null hypothesis.

Nature Population/
Sample Observation/
Data Relationships/
Modeling Analysis/
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 $\sigma^2_{\text{known}} = 43.3^2$ months squared. Is this statistically significant evidence of improved survival *at the 0.05 level*?

Notice the extra wording *at the 0.05 level*. We will use this to develop a 0.05 critical region.

Null Hypothesis Probability Model Assumptions.

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

Null and alternative hypotheses

$H_0: \mu_{\text{true}} = \mu_0 \leq 38.3$ months

$H_A: \mu_{\text{true}} = \mu_A > 38.3$ months

The appropriate Test Statistic is a Z-Score

The null hypothesis gives us the following:

- X_1, X_2, \dots, X_{100} is a simple random sample from a Normal($\mu = 38.3, \sigma^2 = 43.3^2$).
- $\bar{X}_{n=100}$ 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 $\bar{X}_{n=100}$, obtained under the assumption that the null hypothesis is correct.

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

Nature

Population/
SampleObservation/
DataRelationships/
ModelingAnalysis/
Synthesis

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

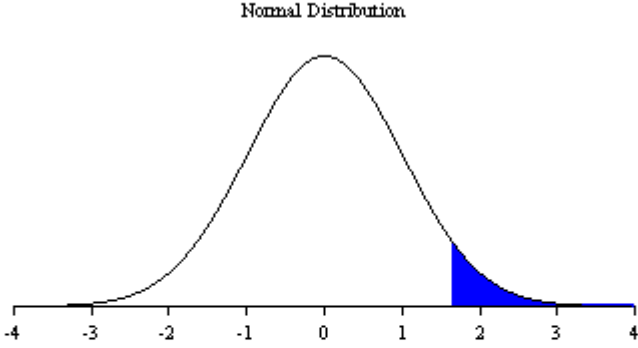
Step 1: Identify what is meant by “extreme” or “critical”:

In this example, the alternative is one sided and extreme values in the direction of the alternative are large positive values of the pivotal quantity.

Step 2: Solve for the critical region of the pivotal quantity:

In this example, solve for the range of extreme values of a Z-score random variable distributed Normal(0,1) such that the area under the null hypothesis curve in the direction of large positive (right tail) is 0.05.

I used the link http://davidmlane.com/hyperstat/z_table.html
Be sure to scroll down to the second calculator that is provided.

 <p>Normal Distribution</p> <p>Mean: <input type="text" value="0"/> Sd: <input type="text" value="1"/></p> <p>Shaded Area: <input type="text" value=".05"/></p> <p><input checked="" type="radio"/> Above: 1.6449 <input type="radio"/> Below <input type="radio"/> Between <input type="radio"/> Outside</p>	<p>Step 1: Enter .05 for shaded area</p> <p>Step 2: Select radio button “Above”</p> <p>Step 3: Read critical region as z-score \geq 1.6449</p>
--	---

Step 3: Solve for the critical region of \bar{X} :

$$\text{Pivotal Quantity} = \text{z-score} = \frac{\bar{X}_{n=100} - \mu_{\text{null}}}{\text{SE}(\bar{X}_{n=100})} \geq 1.6449 \rightarrow$$

$$\frac{\bar{X}_{n=100} - 38.3}{4.33} \geq 1.6449 \rightarrow$$

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

Step 4: Interpret:

In words, “this one sided .05 test of the null versus alternative hypotheses rejects the null hypothesis for critical region values of $\bar{X}_{n=100} \geq 45.42$.

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

$\bar{X}_{n=100} = 46.9$ is in the critical region because it is greater than 45.42.

Interpret.

Because $\bar{X}_{n=100} = 46.9$ and is in the critical region, it is significant at the 0.05 level. According to the critical region approach with type I error = 0.05, **reject the null hypothesis.**

8. Normal: 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.

- The pivotal quantity is a **t-score** instead of a z-score.

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

Null and alternative hypotheses

$$H_0: \mu_{true} = \mu_0 \leq 38.3 \text{ months}$$

$$H_A: \mu_{true} = \mu_A > 38.3 \text{ months} \quad \underline{\text{one sided}}$$

Nature

Population/
SampleObservation/
DataRelationships/
ModelingAnalysis/
Synthesis

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

Statistically, the null hypothesis, when examined in light of the data, lead to an inconsistency or “contradiction” if the null hypothesis model probability of a sample mean being as extreme as 46.9 months (or larger) is small. We calculate the value of such chances as

$$\Pr[\bar{X}_{n=100} \geq 46.9 | \mu_{true} = \mu_o = 38.3]$$

The appropriate Test Statistic is now a T-Score

Under the assumption of the null hypothesis:

- X_1, X_2, \dots, X_{100} is a simple random sample from a Normal($\mu = 38.3, \sigma^2$).
- $\bar{X}_{n=100}$ is distributed Normal ($\mu = 38.3, \sigma_{\bar{X}}^2 = \sigma^2/100$)
- Here, we’ll use as our test statistic the *t-score standardization* of $\bar{X}_{n=100}$, obtained under the assumption that the null hypothesis is correct.

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

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

P-value calculation

$$\text{p-value} = \Pr[\bar{X}_{n=100} \geq 46.9 | \mu_{true} = \mu_{null} = 38.3]$$

$$= \Pr[\text{t-score}_{\text{degrees of freedom}=99} \geq 1.99] = .02467 \text{ quite close to .02 obtained previously!}$$

“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%** chance that we would have obtained our observed average survival time of 46.9 months *or greater*.

Interpret.

The assumption that the null hypothesis is true, when examined in light of the observed data, has led to an unlikely conclusion. Reject the null hypothesis.

9. Normal: Test for σ^2

Example -

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 variation 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 greater than $(.5)^2 = 0.25 \text{ mg}^2$?

Null Hypothesis Assumptions:

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

Specify Hypotheses:

$$H_0: \sigma^2 \leq 0.25$$

$$H_a: \sigma^2 > 0.25 \text{ one-sided}$$

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:

$$\text{P-value} = \Pr[S \geq 0.787 \mid \sigma_{true} = \sigma_0 = 0.5]$$

Nature

Population/
SampleObservation/
DataRelationships/
ModelingAnalysis/
Synthesis

The Test Statistic is a Chi Square:

S, as is, cannot be our test statistic. That is, we do not calculate $\Pr [S > 0.787]$ directly.

Instead we work with a related random variable, obtained under the assumption that the null hypothesis is true. Note – This is analogous working with a Z-score standardization of the sample mean.

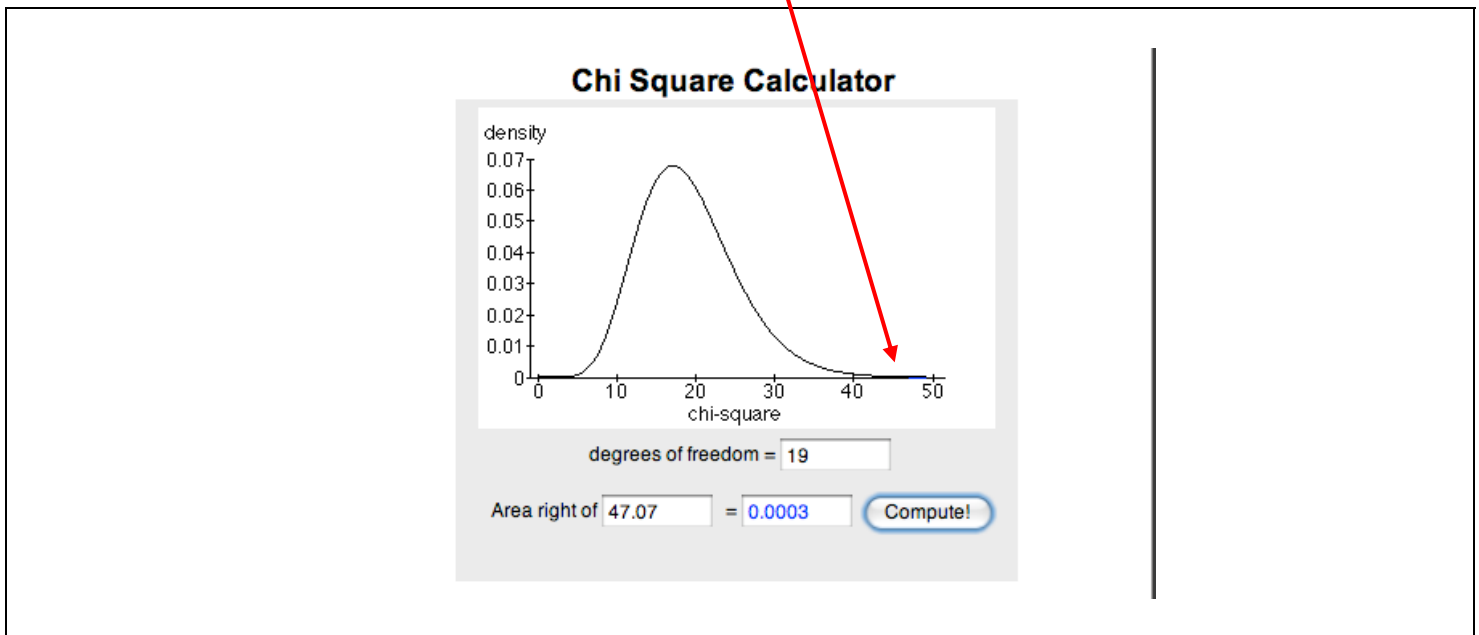
In particular, under the assumption that the null hypothesis is true

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

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

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

p-value = $\Pr [\text{Chi Square}_{\text{DF}=19} \geq 47.072] = 0.00035$ (the area under the curve is tiny!)



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

10. Normal: Test for $\mu_{\text{DIFFERENCE}}$ - Paired Data Setting

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?

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 $(X_1, Y_1) \dots (X_{12}, Y_{12})$.
- * Focus is on the 12 differences because these represent change over 6 months:

$$d_1 = (Y_1 - X_1)$$

$$\dots$$

$$d_{12} = (Y_{12} - X_{12})$$

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

Nature

Population/
Sample

Observation/
Data

Relationships/
Modeling

Analysis/
Synthesis

Null Hypothesis Probability Model.

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

 H_0 and H_A .

$$H_0 : \mu_d = 0$$

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

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

$$z_{score} = \left[\frac{\bar{d} - E(\bar{d} | H_0 \text{ true})}{SE(\bar{d} | H_0 \text{ true})} \right]$$

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

The likelihood of these findings or ones more extreme if H_0 is true is

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

P-Value Calculation.

When the null hypothesis is true, the $d_1 \dots d_{12}$ 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 ($\mu = 0, \sigma^2 = \left[\frac{23.15^2}{12} \right]$)

p-value =

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

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

Nature

Population/
SampleObservation/
DataRelationships/
ModelingAnalysis/
Synthesis

Example Scenario #2 – Variance is UNKNOWN

- Because the variance is unknown, the test statistic will be a ***t-score***
- Otherwise the thinking is the same.
- Suppose that $s=23.15$

H₀ and H_A

$$H_0 : \mu_d = 0$$

$$H_A : \mu_d < 0$$

Test statist is a T-Score when the Variance is UNKnown.

$$t_{score} = \left[\frac{\bar{d} - E(\bar{d} | H_0 \text{ true})}{\widehat{SE}(\bar{d} | H_0 \text{ true})} \right]$$

P-Value Calculation.

p-value

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

$$= pr[\text{Student's } t_{DF=11} \leq -3.02] = \mathbf{0.00583} \quad \text{notice – this is bigger than the .00126 obtained previously!}$$

Interpret.

The conclusion is the same.

11. Normal: Test of $[\mu_1 - \mu_2]$ - Two Independent Groups

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

Research Question.

Do patients receiving zidovudine have higher functional status scores?

Null Hypothesis Assumptions.

\bar{X}_1 is distributed Normal ($\mu_1, \sigma^2/15$) and \bar{X}_2 is distributed Normal ($\mu_2, \sigma^2/22$)

H_0 and H_A .

$H_0 : \mu_1 = \mu_2$

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

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

The test statistic is a t-score.

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

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

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:

$$SE\hat{E}(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{S_{pool}^2}{n_1} + \frac{S_{pool}^2}{n_2}} \quad \text{where}$$

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

For these data:

$$\hat{\sigma}^2 = S_{pool}^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)} = \frac{(15 - 1)40^2 + (22 - 1)35^2}{(15 - 1) + (22 - 1)} = 1375$$

$$SE\hat{E}(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{S_{pool}^2}{n_1} + \frac{S_{pool}^2}{n_2}} = \sqrt{\frac{1375}{15} + \frac{1375}{22}} = 12.42$$

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

Nature

Population/
Sample

Observation/
Data

Relationships/
Modeling

Analysis/
Synthesis

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

The likelihood of these findings or ones more extreme if H_0 is true is

$$\text{p-value} = \Pr[(\bar{X}_1 - \bar{X}_2) \geq (120 - 96) | H_0 \text{ true}].$$

Calculations.

$$\begin{aligned} \text{p-value} &= \Pr[(\bar{X}_1 - \bar{X}_2) \geq (120 - 96)] \\ &= \Pr\left[\frac{(\bar{X}_1 - \bar{X}_2) - (0)}{\hat{SE}(\bar{X}_1 - \bar{X}_2)} \geq \frac{(120 - 96) - (0)}{12.42}\right] \\ &= \Pr[t_{\text{score}} \geq 1.93] \quad \text{where degrees of freedom} = 35 \\ &= .03 \end{aligned}$$

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

“Evaluate”.

Under the null hypothesis H_0 , 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 probability. Reject the null hypothesis.

Interpret.

The investigator infers a benefit of zidovudine on functional status.

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

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

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

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

$$= \sqrt{\frac{40^2}{15} + \frac{35^2}{22}} = 12.74$$

Nature

Population/
SampleObservation/
DataRelationships/
ModelingAnalysis/
Synthesis

We have to use that horrible formula for the degrees of freedom that is on page 52 of the Unit 6 (Estimation) notes.

$$\begin{aligned} \text{Degrees of freedom} &= \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{(n_1-1)} + \frac{(S_2^2/n_2)^2}{(n_2-1)}}. && \text{In this example we get} \\ &= \frac{\left(\frac{40^2}{15} + \frac{35^2}{22}\right)^2}{\frac{(40^2/15)^2}{(14)} + \frac{(35^2/22)^2}{(21)}} = 27.44 \approx 27 \text{ Round DOWN} \end{aligned}$$

Thus,

$$\begin{aligned} \text{p-value} &= \Pr[(\bar{X}_1 - \bar{X}_2) \geq (120 - 96)] \\ &= \Pr\left[\frac{(\bar{X}_1 - \bar{X}_2) - (0)}{\widehat{SE}(\bar{X}_1 - \bar{X}_2)} \geq \frac{(120 - 96) - (0)}{12.74}\right] \\ &= \Pr[t_{\text{score}} \geq 1.88] \quad \text{where degrees of freedom} = 27 \\ &= .035 \end{aligned}$$

Interpret.

The conclusion is the same - infer a benefit of zidovudine on functional status.

12. Normal: Test for Equality of Two Variances

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, $X_1 \dots X_{n_1}$ distributed Normal (μ_1, σ_1^2) and $Y_1 \dots Y_{n_2}$ distributed Normal (μ_2, σ_2^2)

H_0 and H_A .

$$H_0 : \sigma_1 = \sigma_2$$

$$H_A : \sigma_1 \neq \sigma_2 \text{ two sided. } \quad \text{The investigator is researching inequality}$$

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

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



“Evaluation” rule.

The likelihood of these findings or ones more extreme if H_0 is true, with respect to a two sided alternative is

$$\text{p-value} = (2) \Pr \left[F_{df=13,10} \geq \left(\frac{S_1^2}{S_2^2} \right) \mid H_0 \text{ true} \right].$$

Calculations.**p-value**

=

$$(2) \Pr \left[F_{df=13,10} \geq \left(\frac{S_1^2}{S_2^2} \right) \mid H_0 \text{ true} \right] = (2) \Pr \left[F_{df=13,10} \geq \left(\frac{10.9^2}{4.2^2} \right) \right] = (2) \Pr [F_{df=13,10} \geq 6.73]$$

= (2) (0.0024)

= 0.0048

URL for obtaining F-Distribution Probabilities (Provided are RIGHT tail areas)

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

df numerator =	<input type="text" value="13"/>
df denominator =	<input type="text" value="10"/>
F =	<input type="text" value="6.73"/>
p =	<input type="text" value="0.00240"/>

Nature

Population/
SampleObservation/
DataRelationships/
ModelingAnalysis/
Synthesis

13. Single Binomial: Test for Proportion π

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 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

- Data are the outcomes of Binomial random variable with number of trials $n=20$. Binomial ($n=20, \pi$).
- Observed is $X=4$

H_0 and H_A :

$$H_0 : \pi = 0.25$$

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

Tip - Calculate $E [X | H_0 : \pi = 0.25]$. You will need this to calculate your p-value.

$$E [X | H_0 : \pi = 0.25] = [n] [0.25] = [20] [0.25] = 5$$

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 two sided, the p-value calculation we want to do here answers the question: **“If H₀ is true, what are the chances of obtaining a number of events of death as far away from E [X | H₀ : 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,

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

P-Value Calculation.

$$\begin{aligned} p\text{-value} &= \text{Prob} [\text{Binomial}(20, .25) \geq 6] + \text{Prob} [\text{Binomial}(20, .25) \leq 4] \\ &= 0.4148 + 0.3828 \\ &= 0.7976 \end{aligned}$$

“Evaluate”.

Under the null hypothesis H₀, 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.

Do NOT reject the null hypothesis. Conclude that these data do not provide statistically significant evidence for its rejection.



13.2 Normal Approximation Test

Guideline for Use of the Normal Approximation Test

Use this test for moderate to large size data settings. Its appropriateness is an application of the central limit theorem.

As a rough rule of thumb, you can use the following normal approximation test when the following holds:

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

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 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 ($N=200, \pi$).
- The observed $X=40$
- A quick check of $(n)(\pi)(1 - \pi) = (200)(.25)(.75) = 37.5$ tells us that we have plenty of sample size for using a normal approximation test approach. In this setting, application of the central limit theorem to a binomial random variable outcome gives the following:

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

- The observed proportion is $\bar{X}=40/200=0.20$

Nature

Population/
SampleObservation/
DataRelationships/
ModelingAnalysis/
Synthesis

H₀ and H_A

$$H_0 : \pi = 0.25$$

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

The Test statistic is a z-score.

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

“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 mortality rate **as different** from the expected value of 25% as the observed 20% is a two sided calculation:

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

P-Value Calculation.

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

$$= (2) (0.051)$$

$$= 0.102$$

“Evaluate”.

Under the null hypothesis H_0 , 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 a contradiction.

Interpret.

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

Nature

Population/
SampleObservation/
DataRelationships/
ModelingAnalysis/
Synthesis

14. Two Binomials: Test for Equality of Proportions [$\pi_1 - \pi_2$]

A normal theory approximate test is described.

Example

Consider again the needle exchange trial introduced previously. Among the preliminary aims is 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 greater 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.

- Gender is the random variable of interest. In each group (pharmacy sales versus pharmacy sales + needle exchange), the number 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 ($n_1=53, \pi_1$) and \bar{X}_2 is distributed Binomial ($n_2=48, \pi_2$) where

π_1 = Proportion women in Pharmacy Sales

π_2 = Proportion women in Pharmacy Sales + Needle Exchange

H_0 and H_A :

$$H_0 : \pi_1 = \pi_2$$

$$H_A : \pi_1 \neq \pi_2 \quad \text{two sided}$$

The Test statistic is now a Z-score.

$$z_{score} = \left[\frac{(\bar{X}_1 - \bar{X}_2) - E[(\bar{X}_1 - \bar{X}_2) | H_0 true]}{SE[(\bar{X}_1 - \bar{X}_2) | H_0 true]} \right]$$

Two Independent Binomials – Calculation of $SE[(\bar{X}_1 - \bar{X}_2) | H_0 true]$

$$SE[(\bar{X}_1 - \bar{X}_2) | H_0] = \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n_1} + \frac{\hat{\pi}(1 - \hat{\pi})}{n_2}} \quad \text{where}$$

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

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

For these data:

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

Nature

Population/
SampleObservation/
DataRelationships/
ModelingAnalysis/
Synthesis

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

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

In the needle exchange trial, interest is in the likelihood of obtaining a magnitude of difference as great or greater than $|.271 - .170| = .1010$

The required p-value calculation is thus

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

P-Value calculation.

$$\begin{aligned} \text{p-value} &= 2 \Pr[(\bar{X}_2 - \bar{X}_1) \geq (.271 - .170)] \\ &= 2 \Pr\left[\frac{(\bar{X}_2 - \bar{X}_1) - E(\bar{X}_2 - \bar{X}_1)}{SE(\bar{X}_2 - \bar{X}_1)} \geq \frac{(.271 - .170) - (0)}{.0823}\right] \\ &= 2 \Pr[z\text{-score} \geq 1.23] = 2[.10935] \\ &= .22 \end{aligned}$$

$z_{\text{score}} = 1.23$ 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 is a 22% chance of obtaining a discrepancy in the % women in the two groups equal to 10 percentage points or more.

Interpret.

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/
Sample _____ Observation/
Data _____ Relationships/
Modeling _____ Analysis/
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://www.stat.tamu.edu/~west/applets/tdemo.html>

The Chi Square Distribution

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

The F-Distribution

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

Nature

Population/
SampleObservation/
DataRelationships/
ModelingAnalysis/
Synthesis