

Unit 7.
Hypothesis Testing

Topic	1. The Logic of Hypothesis Testing	2
	2. Beware the Statistical Hypothesis Test	15
	3. Introduction to Type I, II Error and Statistical Power	18
	4. Normal: Test for μ , σ^2 Known	24
	5. Normal: Test for μ , σ^2 Known – Critical Region Approach	27
	6. Normal: Test for μ , σ^2 Unknown	31
	7. Normal: Test for σ^2	34
	8. Normal Test for $\mu_{\text{DIFFERENCE}}$ – Paired Data Setting	36
	9. Normal: Test for $[\mu_1 - \mu_2]$ – Two Independent Groups	40
	10. Normal: Test for Equality of Two Variances (σ_1^2 / σ_2^2).....	46
	11. Single Binomial: Test for Proportion π	49
	12. Two Binomials: Test for $[\pi_1 - \pi_2]$ – Two Independent Groups.....	51
	Appendix	
URL's for the Computation of Probabilities	55	

1. The Logic of Hypothesis testing

What if we want to do comparisons? Test hypotheses? This is inferential statistics. Inferential statistics require probability models. The concept of a probability model was introduced previously.

- Recall that, loosely, a “probability” tells us the chances of observing something.
- We use “probabilities” to compare the reasonableness of competing hypotheses. Thus, they are tools in decision making.

Example Given a particular exposure (smoking), what is the probability of a particular disease? (“Tobacco companies on trial”)

Example What are the chances that a person without disease (no HIV infection) will obtain a positive result on an HIV antibody test? (“False positive”)

Inasmuch as we’re after an understanding of nature, we use the tool of “chance” only as long as we have to.

- Probability models, i.e.- “chance”, describe the unknown.
- “Noise” in the signal-to-noise concept is “chance”. Thus, what we do know is modeled (“signal”). The rest, representing what we cannot explain, is regarded as “due chance”.

As science progresses, increasingly, “due chance” variability is explained.

- Hypotheses are formulated, experiments are performed, and results are evaluated for their consistency (their non-consistency, actually) with a hypothesis.
- With the conclusion that a hypothesis is reasonable, the investigator has “explained” some of the current total pool of “due chance”. The pool of unknown, the “due chance”, is now smaller.
- Perhaps the next investigator, with his or her refined hypothesis, will reduce further the pool of “due chance”.

Inferential statistics proceeds similarly.

Consider the following scenario (hypothetical):

Interest is in investigating whether the *type* of access to clean injection paraphernalia will affect a person’s frequency of drug injection. A randomized trial investigating, among other things, frequency of injection is comparing two groups: 1) needle exchange and legal pharmacy sales versus 2) legal pharmacy sales only.

Comparison of 2 Groups

Analysis reveals no overall effect of randomization assignment on frequency of drug injection. Persons with access only to pharmacy sales appear to have similar frequencies of drug injection as persons with access to both pharmacy sales and needle exchange. However, the variability in the data is great. Another way of saying this is to say that the “noise” is great. *Perhaps in this “noise” there is another story to uncover.* This prompts a closer look. The mechanics of the subsequent closer looks might take the form of stratified analyses, regression modeling, etc.

Comparison of More Than 2 Groups

When the data are analyzed separately for men and women, it appears that access to needle exchange is beneficial among women and harmful among men, at least with respect to frequency of drug injection.

Thus, scientific inquiry, through the use of statistical modeling and hypothesis testing, treats deterministic events as stochastic until their nature is understood.

Statistical Hypothesis Testing is a Tool for the Investigation of Research Hypotheses.
Here are some examples of research hypotheses – also some study designs.

- Following counseling, access to needle exchange and pharmacies, compared to access to pharmacies alone, results in a lower 6-month sero-incidence of HIV infection.

Study design - Randomized controlled trial
Analysis Goal - Comparison of two groups

- The implementation of the policy of banning legal pharmacy sales of syringes will reduce the prevalence of drug injection in Anchorage.

Study design - Repeated cross-sectional survey
Analysis Goal - Comparison of two groups

- The delivery of an educational intervention to injection drug users in residential treatment will produce “safer” injection practices upon discharge.

Study design - Intervention study
Analysis Goal - Paired (Pre Test/Post Test) longitudinal comparison

- The cost to Anchorage, Alaska of screening 1000 injection drug users for Hepatitis C is \$X.

The logic of **proof by contradiction** is used to evaluate alternative explanations for observed phenomena in what is called **statistical hypothesis testing**. *As we will see, statistical inference is not biological inference.*

In evaluating competing explanations for observed phenomena, we draw upon concepts of null and alternative hypotheses.

Following are examples of null (H_0) and alternative (H_A) hypotheses.

- Following counseling, access to needle exchange and pharmacies, compared to access to pharmacies alone, results in a different 6-month sero-incidence of HIV infection.

Let μ represent the mean 6-month sero-incidence of HIV infection

Group 1: Pharmacy Sales Access only (mean = μ_1)

Group 2: Pharmacy Sales + Needle Exchange Access (mean = μ_2)

$$H_0: \mu_2 = \mu_1$$

$$H_A: \mu_2 \neq \mu_1 \quad (\text{two sided})$$

Note: For ethical reasons, many randomized trial involving human subjects cannot be justified without the belief that the alternative is two sided. The exception is equivalence trials.

- The implementation of the policy of banning legal pharmacy sales of syringes will reduce the prevalence of drug injection in Anchorage.

Let π represent the prevalence of drug injection

Group 1: 1998 Anchorage population of drug injectors (mean = π_1)

Group 2: 2000 Anchorage population of drug injectors (mean = π_2)

$$H_0: \pi_2 = \pi_1$$

$$H_A: \pi_2 < \pi_1 \quad (\text{one sided})$$

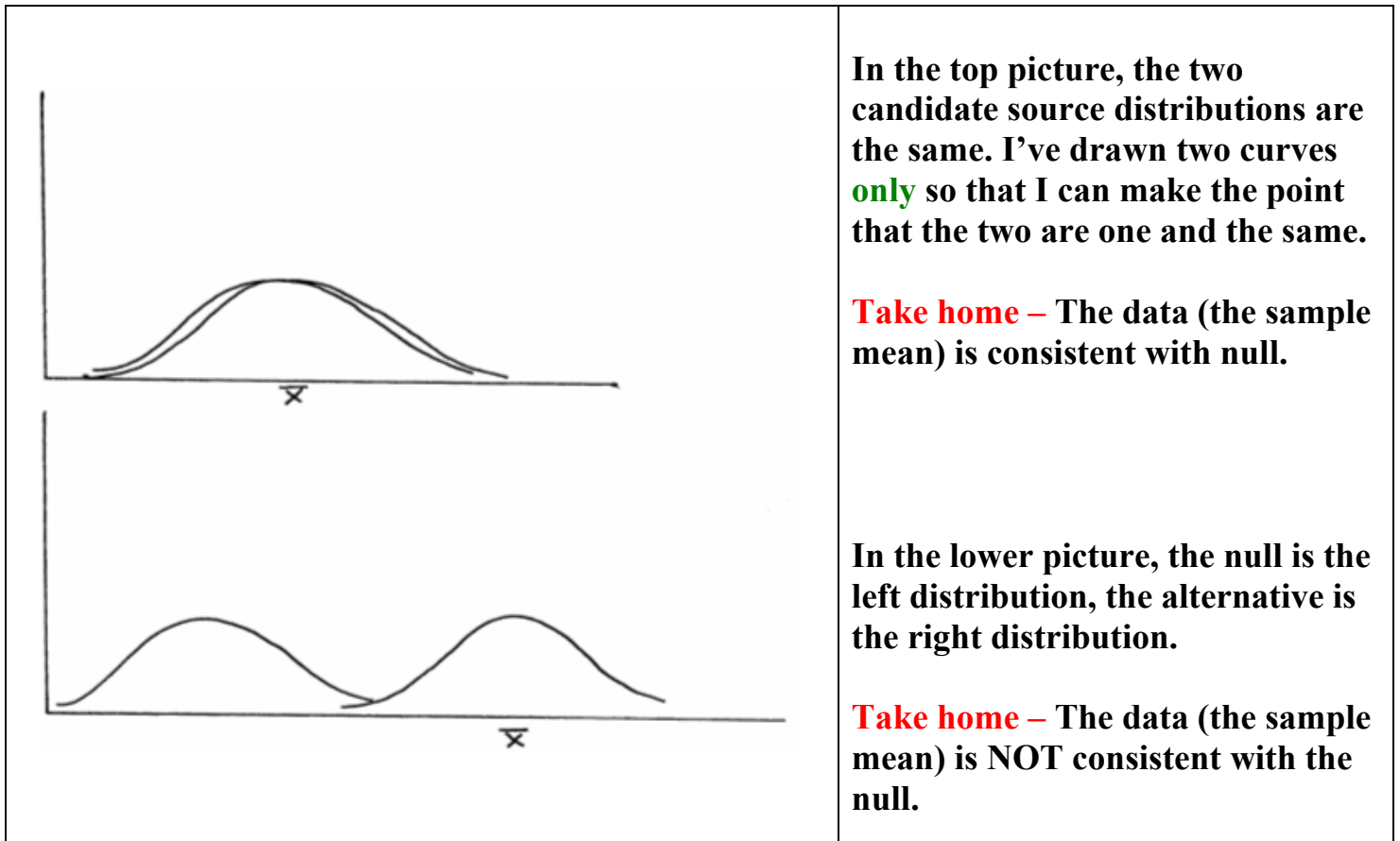
Lucky for us, it is possible to identify a reasonably consistent paradigm of steps in constructing a statistical hypothesis test, and it works for a variety of study design and analysis goal settings.

Here they are.

- 1. Identify the research question.**
- 2. State the 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)**

Following is a schematic of the thinking that underlies a statistical hypothesis test.

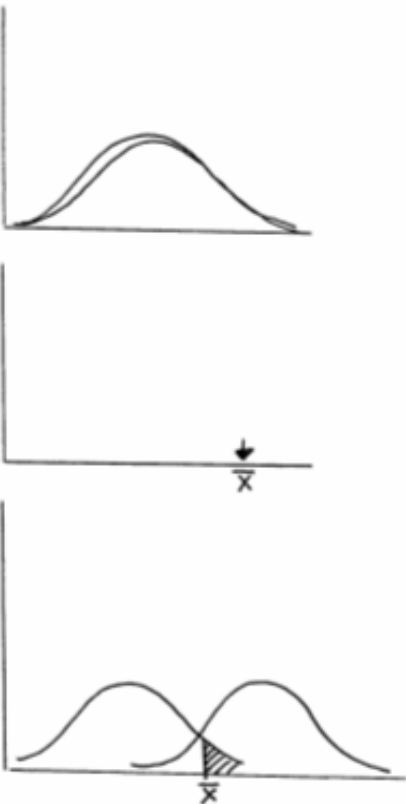
In each picture below, the scenario considered is that there are two candidate source probability distributions that might have given rise to the observed sample mean. These are **null** versus **alternative**.



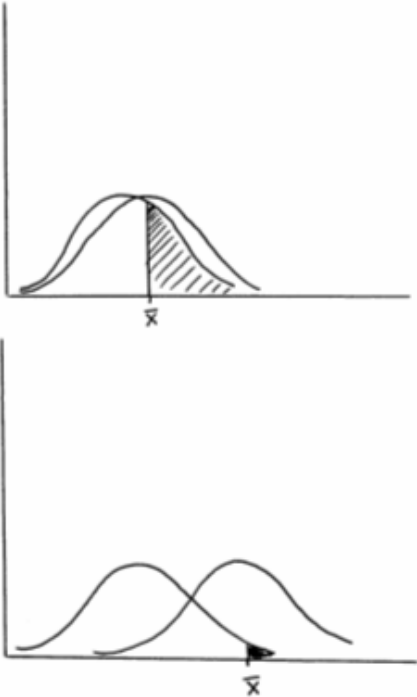
The next schematic is intended to show you, with pictures, how the logic of “proof by contradiction” works.

The setting is that the investigator wishes to assess, utilizing the tools of statistical hypothesis testing, the relative plausibility of two explanations for the observed data. As before, one explanation is the null and the other is the alternative.

In many (but not all) settings of the proof by contradiction argument is the strategy of designating as “null” the “there is nothing going on” explanation) and seeking to advance the “alternative” (there is a treatment benefit, or there is a change over time or there is a difference between groups explanation).

	<p>Step 1 – Grant the null ... The top picture represents the starting point for “proof by contradiction”. It is saying, schematically, “<i>assume the null hypothesis is true</i>”. Under this assumption, the true and the null curves are essentially the same. This is why the two curves are right on top of each other.</p> <p>Step 2 – Collect data ... The middle picture represents the starting point for the investigator. He or she collects data and might summarize it in the form of the sample mean. The absence of a graph of a distribution is a reminder that the investigator doesn’t actually know which distribution gave rise to the data.</p> <p>Step 3 – Argue “yes” or “no” does data contradict null. Represented in the lower picture is the sample mean again. Also shown is the distribution that gave rise to the sample mean if the null is true (left) and the distribution that gave rise to the sample mean if the alternative is true (right)</p> <p>The shaded area is a probability calculation under the assumption that the null is true. It answers the question “<i>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?</i>”</p> <p>Small chances contradict the null suggesting REJECT Large chances are consistent w null suggesting ACCEPT</p>
--	--

Let's look at the question “*what are the chances of a sample mean as extreme or more extreme*”, separately for two scenarios.

	<p>Scenario 1 - NULL is true</p> <ul style="list-style-type: none"> • Observed sample mean is close to null mean. • Likelihood of being “this far away”, when calculated pretending that the null is true, produces a large value. • Statistical decision - “do NOT reject”. <p>Scenario 2 - ALTERNATIVE is true</p> <ul style="list-style-type: none"> • Observed sample mean is now close to the alternative mean. • Likelihood of being “this far away” when calculated pretending that the null is true produces a small value. • Statistical decision – “reject”
--	---

- Do you notice in this logical framework the implicit assumption that the \bar{X} value that is available to us is be close to its true mean?
- In the next pages, you will learn that the calculation shown here to answer the question “If I pretend that the null hypothesis is true, then what were my chances of observing a sample mean as far away as the value obtained” is a **p-value** calculation.
- “p-value” goes by a variety of names: **p-value**, **significance level**, **achieved significance**.

Illustration of the logic of hypothesis testing.

You may recall this example from the course introduction. Consider a setting where, 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. 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 survival statistically significantly improved (relative to standard care) with receipt of the new therapy?**

This illustration follows the steps outlined on page 6.

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 $\mu_{\text{true}} > 38.3$?*

Assumptions are needed for computing probabilities.

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

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

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

Note - The null and alternative hypotheses must accommodate all possibilities. However, for computation purposes, we use the particular null hypothesis that is the closest to the alternative hypothesis. Here, it is $\mu_o = 38.3$. Rationale is to be conservative.

3. Reason “proof by contradiction”

IF: 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 “at least as far away” from 38.3 as the observed value of 46.9?

4. Specify a “proof by contradiction” rule.

Statistically, the data are inconsistent with the null (H_0) 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 value of such chances as

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

*Note - In this probability expression, notice the vertical bar after 46.9. This vertical bar is a shorthand for saying that we are doing this calculation **CONDITIONAL** on or under the assumption that the mean is 38.3*

5. Perform the calculation of such chances presuming H_0 true.

Recall what the null hypothesis says. It says

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

This allows us to say the following, too:

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

The concept of statistical hypothesis testing can be appreciated as an example of the idea of “signal-to-noise”.

Consider the observation that the observed average = 46.9 is 8.6 months different from the value of 38.3 months. Here is how “**signal-to-noise**” helps us in understanding the data:

<p>Signal -</p> <p>“46.9 is 8.6 months away from 38.3”</p>	$(46.9 - 38.3) = 8.6$
<p>Noise -</p> <p>Noise is the scatter of the average. We learned that this is the SE</p>	$SE(\bar{X}_{n=100}) = \frac{\sigma}{\sqrt{100}} = \frac{43.3}{10} = 4.33$
<p>Signal-to-Noise (Z-score)</p> <p>Thus, signal, expressed in units of noise</p> <p>“46.9 is 1.99 SE units away from 38.3”</p>	$\frac{(46.9-38.3)}{SE(\bar{X}_{n=100})} = \frac{8.6 \text{ months}}{4.33 \text{ months}} = 1.99$

*This signal-to-noise is an example of a z-score.
The distribution of a z-score is Normal($\mu = 0, \sigma^2 = 1$)*

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:

“Nature tends to give us the ‘likely’. Accordingly, if we see something that is ‘unlikely’, then perhaps the explanation is something other than what we presumed. This leads us to

Statistical rejection of the null hypothesis.

The Z-score is a Generic Statistic

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

Hint/Suggestion: The computation of many statistical hypothesis tests will result in the calculation of a z-score magnitude. As z-scores are distributed Standard Normal (0,1), a familiarity with this distribution is a helpful tool in gauging the extremeness of study data relative to the null hypothesis that is being challenged.

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.
- * Translation back to the data at hand – The chances of a study sample mean (distributed normal) having value that is more than approximately 2.5 SE distant from its expected value under some null hypothesis is a 1% likelihood.

2. Beware the Statistical Hypothesis Test

There are a variety of reasons for utilizing *only with caution* the tools of statistical hypothesis testing.

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.

2. An isolated p-value communicates limited information only.

Definition p-value

There are a variety of wordings of the meaning of a p-value. Here are some.

- **Source: Fisher and van Belle.** “The null hypothesis value of the parameter is used to calculate the probability of the observed value of the statistic or an observation more extreme.”
- **Source: Kleinbaum, Kupper and Muller.** “The p-value gives the probability of obtaining a value of the test statistic that is at least as unfavorable to H_0 as the observed value”
- **Source: Bailar and Mosteller.** “P-values are used to assess the degree of dissimilarity between two or more sets of measurement or between one set of measurements and a standard. A p-value is actually a probability, usually the probability of obtaining a result as extreme or more extreme than the one observed if the dissimilarity is entirely due to variation in measurements or in subject response – that is if it is the result of chance alone.”
- **Source: Freedman, Pisani, and Purves.** “The observed significance level is the chance of getting a test statistic value as extreme or more extreme than the observed one. The chance is computed on the basis that the null hypothesis is right. The smaller this chance is, the stronger the evidence against the null. ... At this point, the logic of the test can be seen more clearly. It is an argument by contradiction, designed to show that the null hypothesis will lead to an absurd conclusion and must therefore be rejected.”

Beware!

- The p-value is **NOT** the probability of the null hypothesis being correct.
- The p-value is **NOT** the probability of obtaining the observed data “by chance”.
- The p-value is **NOT** the probability of the observed data itself calculated under the assumption of the null hypothesis being correct.
- *Source: Rothman and Greenland.* A p-value is **NOT** “the probability that the data would show as strong an association as observed or stronger if the null hypothesis were correct”.

3. Other criteria are essential to biological inference.

- A conclusion of a treatment effect is *strengthened* by
 - A dose-response relationship
 - Existence in sub-groups as well as existence overall
 - Epidemiological evidence
 - Consistency with findings of independent trials.
 - Its observation in a large scale (meaning large sample size) trial
- A conclusion of a treatment effect is *weakened* by
 - Its unusualness; such a finding should be “checked” with new data
 - Its isolation; that is – it is observed in a selected subgroup only and nowhere else; such a finding is intriguing, however and should be explored further
 - Its emergence as a unique finding among many examinations of the data.

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

A statistical hypothesis test uses probabilities based only on the 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 do so.
 - 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”**

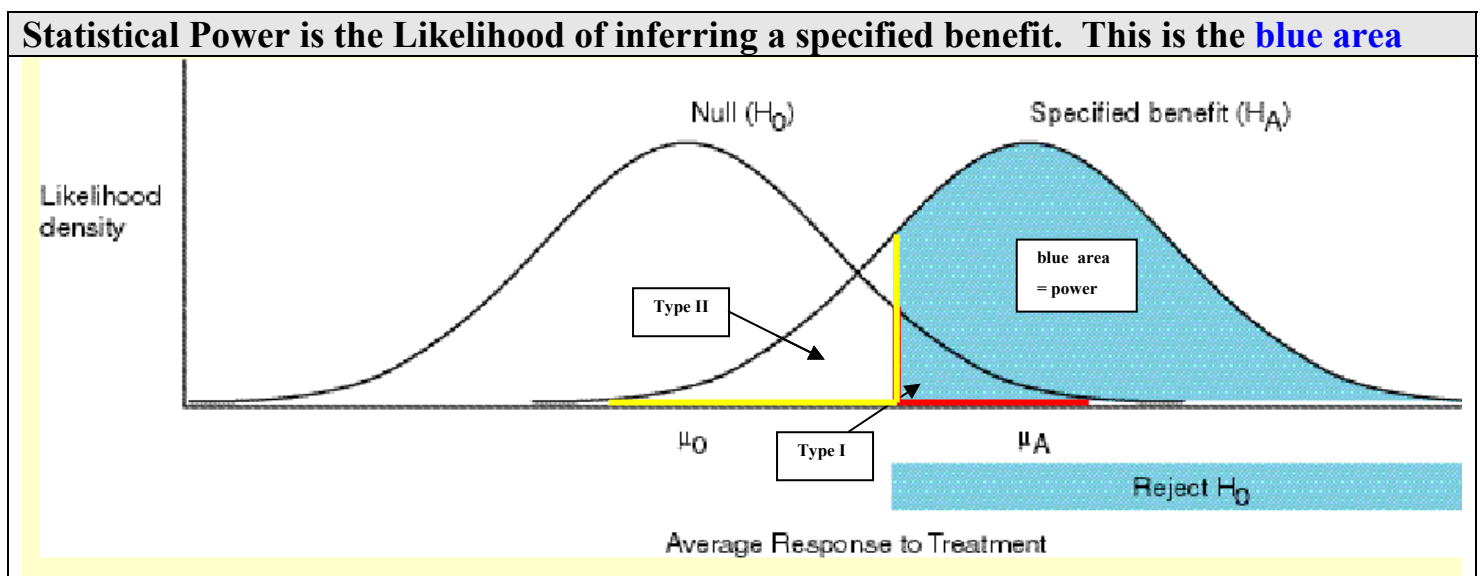
The goal is to get the right answer (power).

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

- This brings us into the realm of sample size calculations. You probably already have a sense for the wisdom that says larger sample size studies are more powerful.
- Know too, however, that there are other factors that will influence the power of a study.

Following are some pictures to illustrate the ideas of statistical power.

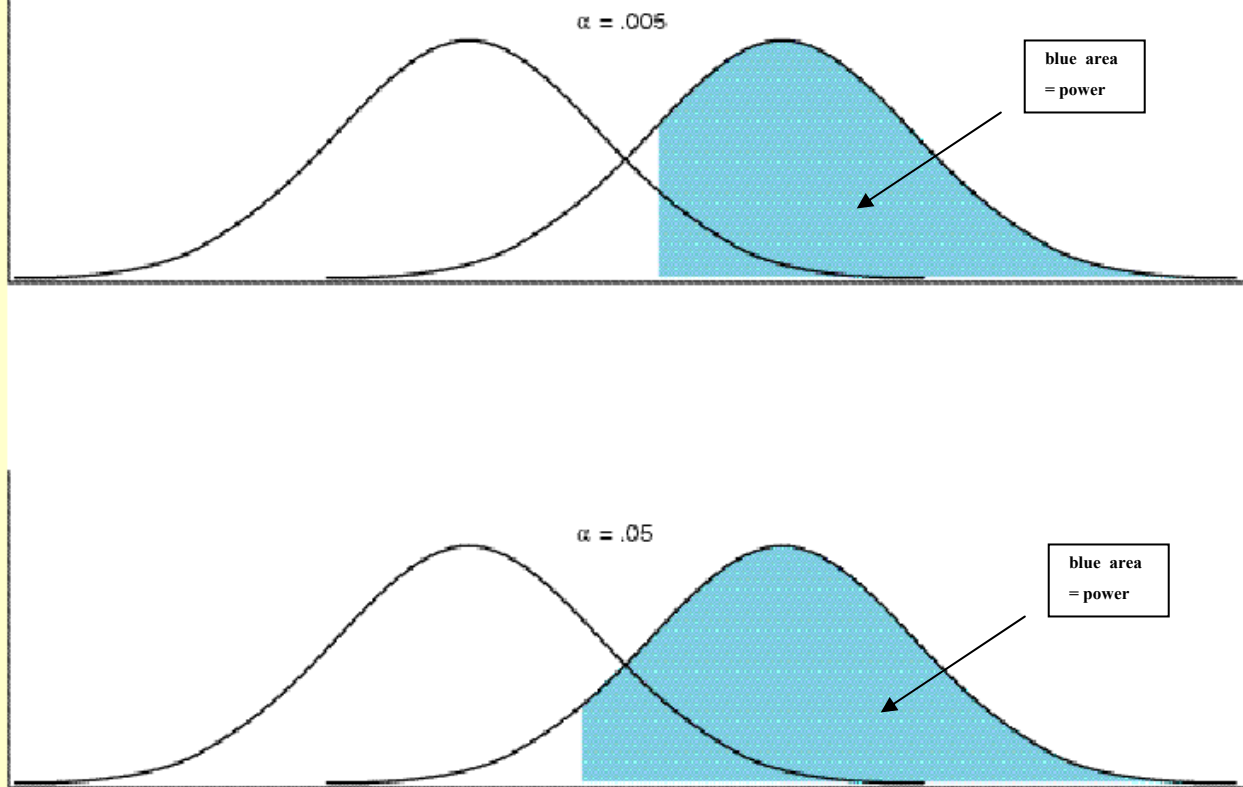
- The techniques of sample size and power calculations are not addressed in this course.



- **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.
- **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

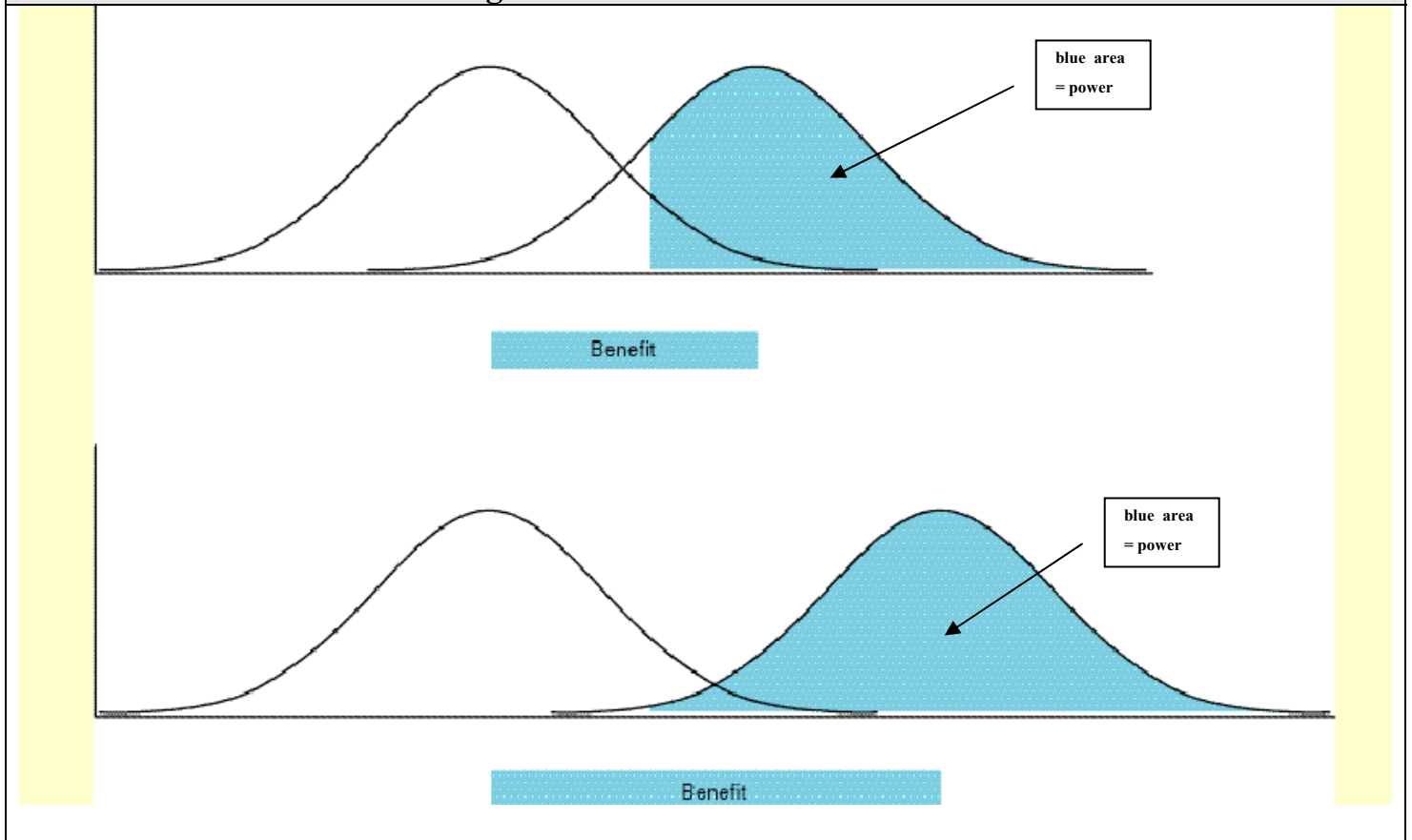
1. Type I Error



- 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.

The Power of a Study Depends on Four Parameters

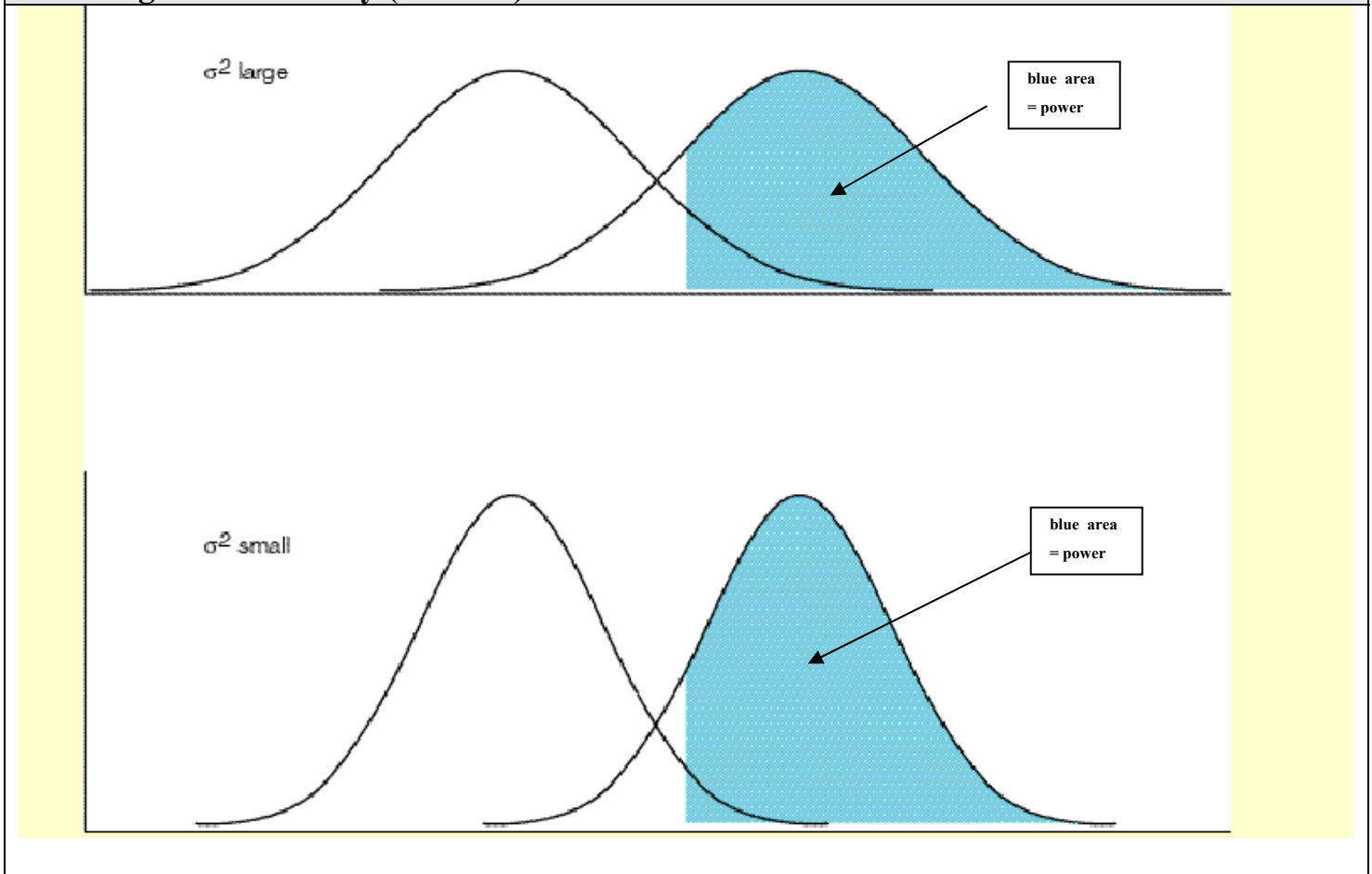
2. The Benefit Worth Detecting



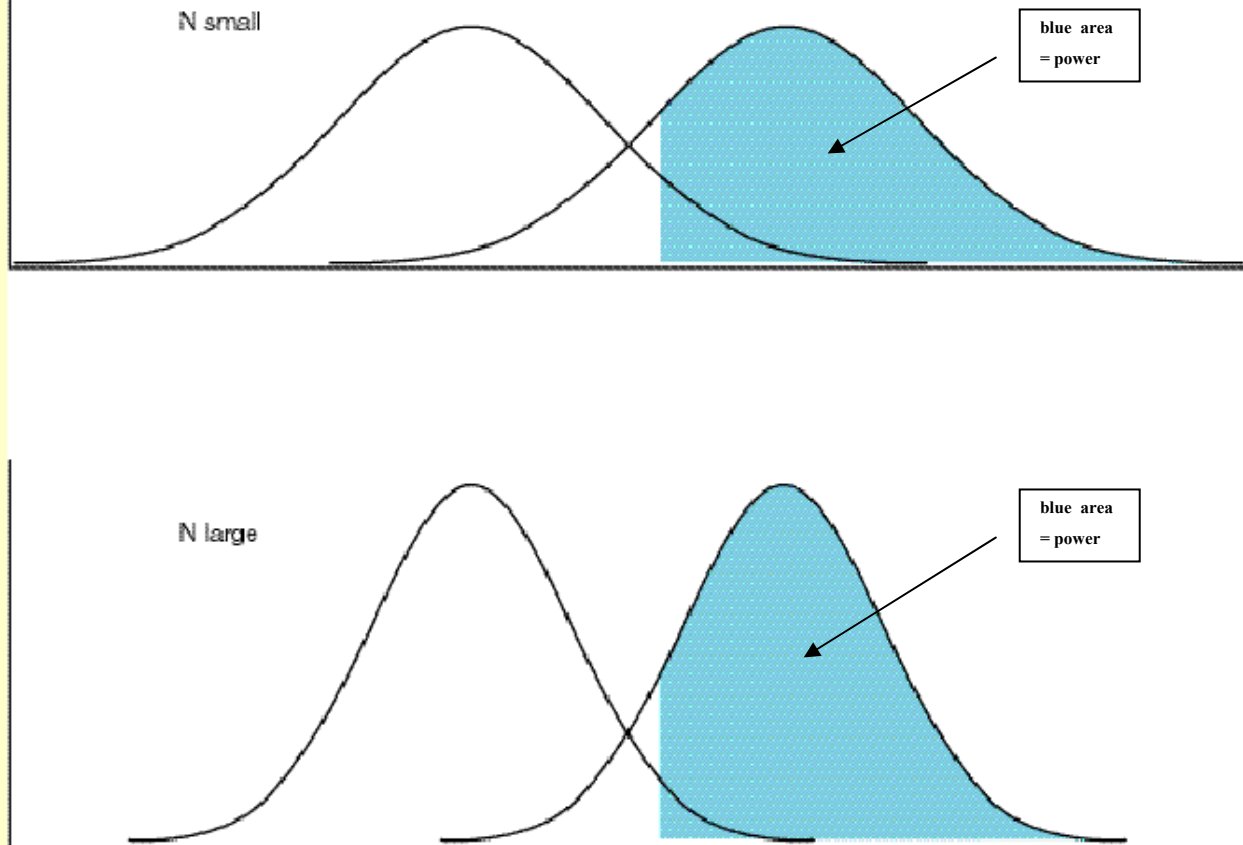
- In this picture, the null hypothesis is the same in the top and bottom panels.
- However, the alternative is closer to the null in the top panel and more distant from the null in the bottom panel.
- The “threshold” value of the sample mean that prompts rejection of the null hypothesis is the SAME in both top and bottom panels.
- What’s illustrated is that, 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.
- What’s illustrated is that, all other things being equal, selecting for measurement an outcome that is less noisy (**more precise**) will **increase** study power (the blue area under the curve).

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

- 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).
- What's illustrated is that, all other things being equal, utilization of a larger sample size will increase study power (the blue area under the curve).

4. Normal: Test for μ , σ^2 Known

The sections that follow in this reading parallel closely sections 5-15 of Topic 6, Estimation. Specifically, what are presented here are examples of how to perform statistical hypothesis tests in the settings of the normal and binomial distributions.

- It might be helpful to re-read the sections of topic 6 that are introductions to the student's t, chi square, and F distributions.
- Also presented is the idea of a **pivotal quantity** (choice of test statistic).
- Also presented is the idea of a **critical region** test.
- The discussion also includes some remarks on the choice between a one tailed versus a two tailed test.
- 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 10-14.

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 = 43.3² months squared. Is this statistically significant evidence of improved survival?

Probability Model Assumptions.

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

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

Reason “proof by contradiction”

IF: it 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 “at least as far away” from 38.3 as the observed value of 46.9?

Specify a “proof by contradiction” rule.

Statistically, the data are inconsistent with the null (H_0) 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 value of such chances as

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

The appropriate PIVOTAL QUANTITY is a Z-Score

The null hypothesis gives us the following:

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

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

Calculate Achieved Significance

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

“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 low likelihood of an average survival time being as great or greater than 46.9 months is **NOT** consistent with the null hypothesis expected mean survival time of 38.3 months.

Reject the null hypothesis.

5. Normal: Test for μ , σ^2 Known Critical Region Test Approach

The paradigm presented in section 4 leads to the calculation of the achieved significance of the data with respect to an assumed null hypothesis. This speaks to 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 paradigm** introduced here considers a slightly different, albeit related, framework.

- **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?

The idea is this. In evaluating our data, we tend naturally to regard as “unusual” an extreme value and are then inclined to regard this as evidence that the null hypothesis should be rejected. Sometimes, this will be a mistake and, if so, we will have made a type I error. The essence of developing a critical region test is to acknowledge this possibility up front and to determine, ahead of time, what extreme values (the name we give these is the critical region) will lead us to rejecting the null hypothesis.

Example is the same—

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 = 43.3² 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.

Probability Model Assumptions.

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

Specify the 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 PIVOTAL QUANTITY is a Z-Score

The null hypothesis gives us the following:

- 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 a pivotal quantity defined as the z-score standardization of $\bar{X}_{n=100}$, developed under the assumption that the null hypothesis is correct.

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

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

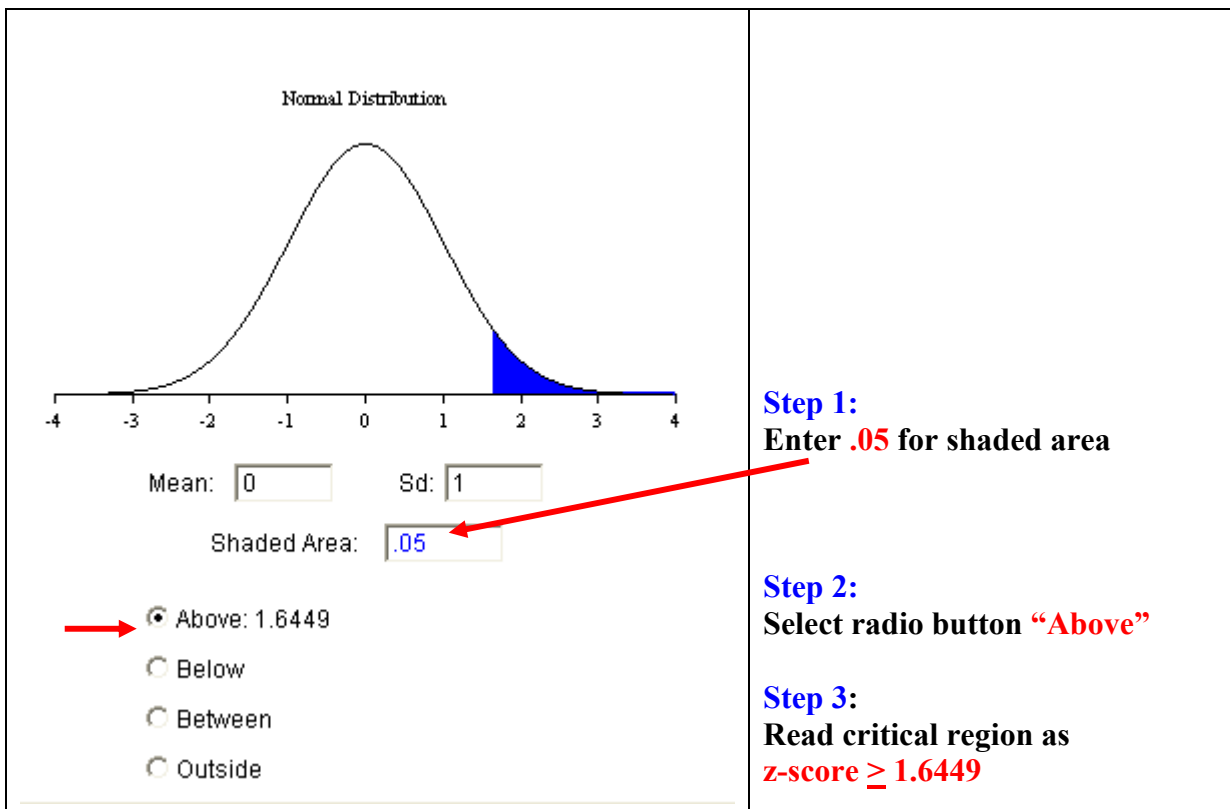
Step 1: Identify what is meant by “extreme” or “critical”:

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 is 0.05.

I used the link http://davidmlane.com/hyperstat/z_table.html
Here, you will find two calculators. Scroll down to the second.



Normal Distribution

Mean: 0 Sd: 1

Shaded Area: .05

Above: 1.6449
 Below
 Between
 Outside

Step 1:
Enter **.05** for shaded area

Step 2:
Select radio button **“Above”**

Step 3:
Read critical region as **$z\text{-score} \geq 1.6449$**

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

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

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

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 values of $\bar{X}_{n=100} \geq 45.42$.”

Compare the observed to 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.**

6. Normal: Test for μ , σ^2 UNKNOWN

The machinery of 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.

- We'll see that 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?

Probability Model Assumptions.

X_1, X_2, \dots, X_{100} is each distributed $\text{Normal}(\mu, \sigma^2)$

σ^2 is **NOT** known.

Specify the null and alternative hypotheses

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

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

Specify a “proof by contradiction” rule.

Statistically, the data are inconsistent with the null (H_0) 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 value of such chances as

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

The appropriate PIVOTAL QUANTITY is a T-Score

The null hypothesis gives us the following:

- X_1, X_2, \dots, X_{100} is each distributed Normal($\mu = 38.3, \sigma^2$).
- $\bar{X}_{n=100}$ is distributed Normal ($\mu = 38.3, \sigma_{\bar{X}}^2 = \sigma^2/100$)
- We’ll use as our test statistic a pivotal quantity defined as the t-score standardization of $\bar{X}_{n=100}$, developed under the assumption that the null hypothesis is correct.

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

- Recall that the denominator is a guess of SE, $\hat{SE}(\bar{X}_{n=100}) = \frac{S}{\sqrt{100}} = \frac{43.3}{10} = 4.33$

Calculate Achieved Significance

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

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

“Evaluate”.

IF 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 an **estimated 2.4% chance of observing an average survival time as great or greater than the observed average survival time of 46.9 months.**

Interpret.

The low likelihood of an average survival time being as great or greater than 46.9 months is NOT consistent with the null hypothesis expected mean survival time of 38.3 months. Reject the null hypothesis.

7. 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 $(.5)^2 = 0.25 \text{ mg}^2$?

Assumptions:

The data are a 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”.

Statistically, the data are inconsistent with the null (H_0) if there is at most a small chance of a sample SD among n=20 capsules being as large as 0.787 when it is correct that the laboratory $\sigma=0.5$. Thus, we'd like to know

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

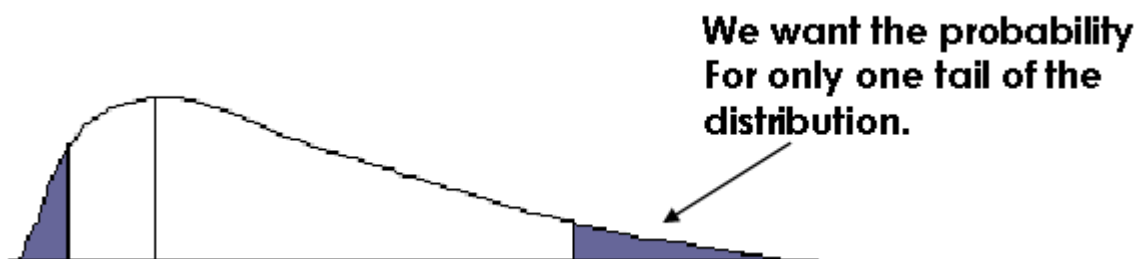
Test Statistic/Pivotal Quantity is a Chi Square:

We don't know how to calculate the above probability directly. Thus, the sample standard deviation S cannot be our pivotal quantity. However, recall from Topic 6 that, in constructing a confidence interval for σ^2 , we utilized a chi square random variable derivation. We can use it here, too, to arrive at an appropriate pivotal quantity for investigating the null and alternative hypotheses.

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

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

Following is a picture to give a feel for the p-value calculation we are after.



Calculations

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

$$\text{p-value} = \Pr[\text{Chi Square}_{\text{DF}=19} \geq 47.072] = 0.00035$$

“Evaluate”.

Under the null hypothesis of a laboratory variance $\sigma^2=0.25$, there is an **estimated 0.035%** chance of observing a sample variance as large as 0.787^2 . This is a very small likelihood!

Interpret.

Reject the null hypothesis and recommend that the company adjust its machines.

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

The following example is presented according to two scenarios

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

Example Scenario #1 – Variance Assumed Known:

(Note: These data are hypothetical.)

Twelve patients in the 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 at the 6 month visit. Interest is in whether participation in the trial alone effected a reduction in the 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 in an intervention study effect a reduction in 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:

$$d_1 = (Y_1 - X_1)$$

...

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

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

Assumptions.

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 .

$$H_0 : \mu_d = 0$$

$$H_A : \mu_d < 0$$

Test statistic/Pivotal Quantity is a Z-Score when the Variance is Known.

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

“Evaluation” rule.

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

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

Calculations.

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 ($\mu = 0, \sigma^2_d = \left[\frac{23.15^2}{12} \right]$)

p-value =

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

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

“Evaluate”.

IF participation in the needle exchange trial in the pharmacy sales condition has no effect on cocaine use, THEN there is a .1% chance of obtaining an observed mean change in hair content of -20.17 or greater among 12 participants.

Interpret.

This low likelihood of reduction in hair content is NOT consistent with the null hypothesis (no effect of study participation among persons in the pharmacy sales group). Possible explanations for the observed findings are

- Trial participation results in less actual use of cocaine.
- Trial participation results in less detection of use of cocaine.

Example Scenario #2 – Variance is UNKNOWN

- The paradigm of statistical hypothesis development now leads to a t-score
- Otherwise the thinking is the same.
- Suppose that $s=23.15$

 H_0 and H_A .

$$H_0 : \mu_d = 0$$

$$H_A : \mu_d < 0$$

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

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

Calculations.

p-value =

$$\begin{aligned} pr[\bar{d}_{n=12} \leq -20.17] &= pr\left[\left(\frac{\bar{d}_{12} - 0}{S_d/\sqrt{n}}\right) \leq \left(\frac{-20.17}{23.15/\sqrt{12}}\right)\right] \\ &= pr[\text{Student's } t_{DF=11} \leq -3.02] = \mathbf{0.00583} \text{ notice - bigger than the .00126 obtained previously!} \end{aligned}$$

Interpret.

The conclusion is the same.

9. 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. We may assume that the scores are normally distributed with distributions that have the same variance σ^2 . However, σ^2 is unknown. Data are:

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?

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 :

$$H_0 : \mu_1 = \mu_2$$

$$H_A : \mu_1 > \mu_2$$

Test statistic/Pivotal Quantity is a t-score.

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

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

We learned previously (see Topic 6) how to estimate the SE of the difference between two independent means, each of which is distributed Normal. In the setting where the two variances are assumed the same, recall how the solution went:

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

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

For these data:

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

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

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

“Evaluation” rule.

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 small likelihood.

Interpret.

The investigator infers a benefit of zidovudine on functional status.

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 following form:

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

What is our guess of the standard error of $(\bar{X}_1 - \bar{X}_2)$ now?

Answer:

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

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

We have to modify our calculation of the degrees of freedom, however – just as we did previously.

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

Thus,

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

Interpret.

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

10. 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 .

$$H_0 : \sigma_1 = \sigma_2$$

$$H_A : \sigma_1 \neq \sigma_2$$

Test statistic/Pivotal Quantity is an F-statistic.

Remark – Whereas the equality of continuous means was evaluated by looking at their difference, the equality of variances is evaluated by looking at their ratio. Thus, ratio values departing appreciably from 1 are evidence of non-equality of variances.

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

“Evaluation” rule.

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

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

Calculations.**p-value**

=

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

$$= (2) (0.0024)$$

$$= 0.0048$$

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"/>

“Evaluate”.

Under the null hypothesis H_0 , 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.

Interpret.

Reject the null hypothesis and conclude that the data are suggestive of a significant inequality in the variability of length of stay elective versus emergency patients.

11. Single Binomial: Test for Proportion π

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%?

Assumptions

- Data are a random sample of patients (over time), and the outcome of mortality, X =(# patients among the 200 who die in hospital) is Binomial ($N=200, \pi$).
- Observed is $X=40$
- As the parameter N (recall – this is the “number of trials”) is large, the central limit theorem (See Topic 5, The Normal Distribution) gives us the following very reasonable approximation:

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

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

H_0 and H_A :

$$H_0 : \pi = 0.25$$

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

Test statistic/Pivotal Quantity is a z-score.

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

“Evaluation” rule.

The likelihood of a mortality rate as different from the expected value of 25% as the observed 20% if H_0 is true, with respect to a two sided alternative is

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

Calculations.

$$\begin{aligned} \text{p-value} &= (2) \Pr \left[\text{Normal}(0,1) \leq \left(\frac{0.20 - 0.25}{\sqrt{\frac{0.25(0.75)}{200}}} \right) \right] = (2) \Pr[\text{Normal}(0,1) \leq -1.63] \\ &= (2) (0.051) \\ &= 0.102 \end{aligned}$$

“Evaluate”.

Under the null hypothesis H_0 , that the mortality rate at Baystate is 0.25, the likelihood of an observed mortality rate as far away (in either direction) as 20% approximately 10 chances in 100. This is a reasonable likelihood.

Interpret.

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

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

We will again use the idea of the z-score.

Example

Consider 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)?

Assumptions.

- In this example, gender is the outcome. In each group (pharmacy sales versus pharmacy sales + needle exchange), the number of women in the group is distributed Binomial.
- 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 ($n_2=53, \pi_2$) where

$\pi_1 =$ Proportion women in Pharmacy Sales

$\pi_2 =$ Proportion women in Pharmacy Sales + Needle Exchange

 H_0 and H_A .

$H_0 : \pi_1 = \pi_2$

$H_A : \pi_1 \neq \pi_2$ (Note that this is a 2 sided alternative)

Test statistic/Pivotal Quantity is a Z-score.

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

Two Independent Binomials – Calculation of $S\hat{E}[(\bar{X}_1 - \bar{X}_2)|H_o \text{ true}]$

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

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

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

For these data:

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

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

“Evaluation” rule.

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

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

Calculations.

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

Here: $z_{\text{score}}=1.23$ 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.

We will conclude that there is not a significant difference in the proportion of women in the two study conditions among the 101 available for interim analysis.

Appendix
URL's for the Computation of Probabilities

The Normal (0,1) Distribution

<http://www-stat.stanford.edu/%7Enaras/jsm/FindProbability.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://davidmlane.com/hyperstat/F_table.html