Unit 8
Statistical Literacy – Estimation and Hypothesis Testing

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

- Michel De Montaigne (1533-1592)

"A hypothesis is a contention that may or may not be true, but is provisionally assumed to be true until new evidence suggests otherwise.

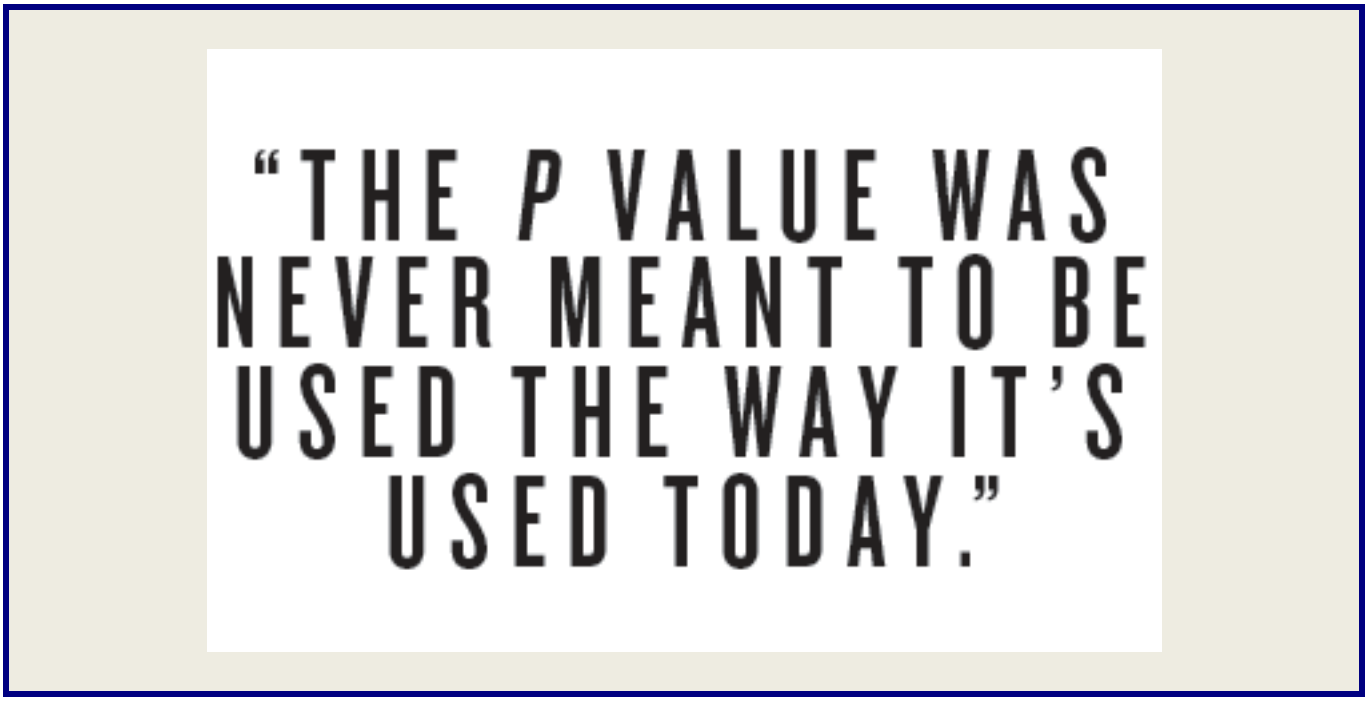
A hypothesis may be proposed from a hunch, from a guess, or on the basis of preliminary observations. A statistical hypothesis is a contention about a population, and we investigate it by performing a study on a sample collected from that population.

We then examine the sample information to see how consistent the data are with the hypothesis under question; if there are discrepancies, we tend to disbelieve the hypothesis and reject it.

So, the question arises, how inconsistent with the hypothesis do the sample data have to be before we are prepared to reject the statistical hypothesis? It is to answer questions such as this that we use statistical tests of hypotheses, or significance tests."

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

Cheers!

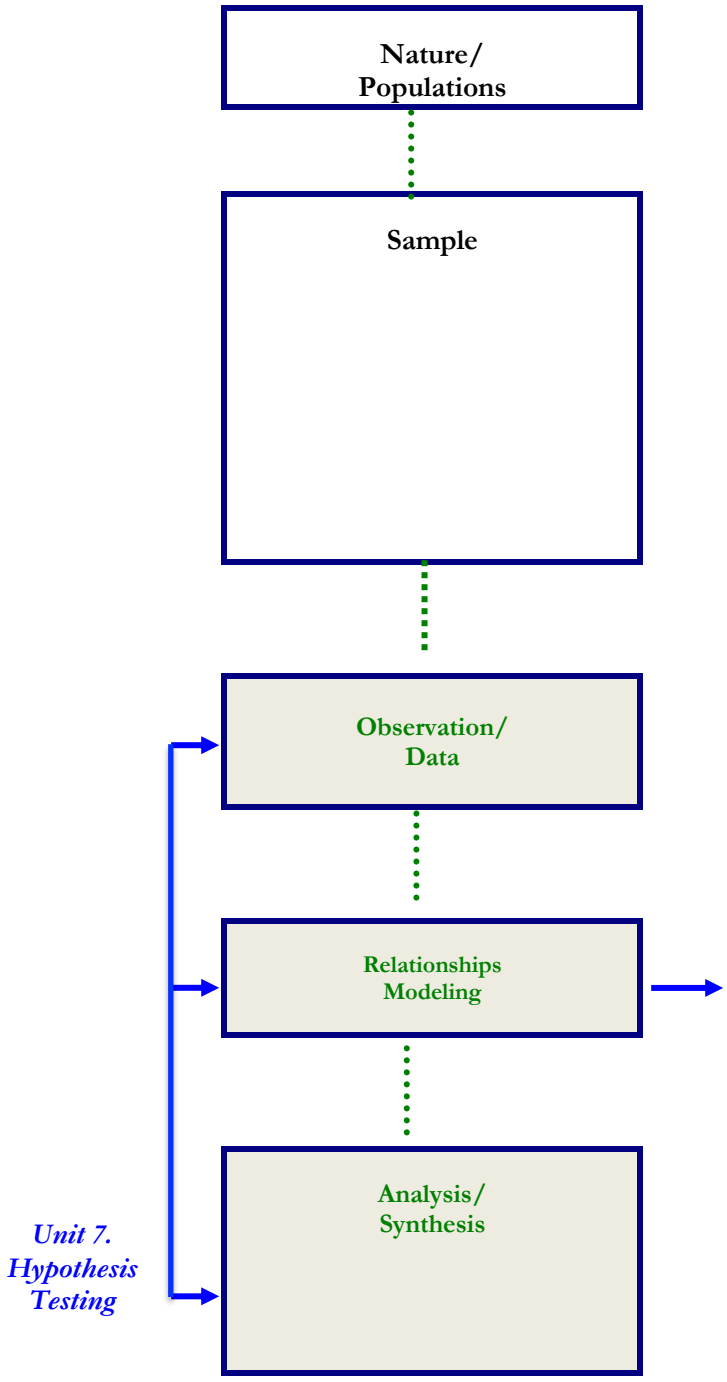


Source: Nuzzo, R. (2014) Statistical Errors. Nature Vol. 506, 13 February 2014

Table of Contents

Topic	**1. Unit Roadmap**…….....……..…	4
	2. Learning Objectives …….....……	5
	3. Introduction to Hypothesis Testing	6
	3.1. The Logic of Hypothesis Testing …..…….…….....……	6
	3.2. P-value Approach …….....	9
	3.3 Introduction to Type I and II Error and Statistical Power …….....	19
	3.4. Critical Region Approach …….....……..	25
	3.5. One Sided versus Two Sided Tests …….…….....…	30
	3.6. Beware the Statistical Hypothesis Test …….....……	32
	4. Introduction to Confidence Interval Estimation	35
	4.1. Goals of Estimation ……..……...……..….....…	35
	4.2. Notation and Definitions ……..…….....…	39
	4.3. How to Interpret a Confidence Interval ……..…….....	42

Unit Roadmap



Statistical significance testing is a tool that informs our understanding of nature but ***does not establish biological significance*** one way or the other.

The logic of statistical hypothesis testing is a "proof by contradiction" argument:

Step 1 –Begin with the "skeptic's" perspective. Define a "chance" model. This is the **null hypothesis**.

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

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

Step 4 – State the *statistical* inference:

If the outcome is unlikely,
The null hypothesis model is rejected.

If the outcome is plausible,
The null hypothesis model is NOT rejected.

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

2. Learning Objectives

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

- Explain the logic of statistical hypothesis testing;
- Translate the statement of a research question into a testable hypothesis; and specifically,
- Translate the statement of a research question into its associated null (H_O) and alternative (H_A) hypotheses.
- For a given data situation, define and explain the null hypothesis model.
- Explain the steps in performing a statistical hypothesis test.
- Explain the meaning of a p-value.
- Interpret the value of a p-value with respect to rejection or non-rejection of a null hypothesis.
- Interpret p-values in publications.
- Explain the utility of accompanying statistical hypothesis tests with confidence intervals.
- Explain Type I and II error and the meaning of statistical power
- Explain that there is more than one way to estimate a population parameter based on data in a sample.
- Explain the criteria of unbiased and minimum variance in the selection of a "good" estimator.
- Define the Student t, chi square, and F probability distribution models.
- Interpret a confidence interval.

3. Introduction to Hypothesis testing

3.1 The Logic of Hypothesis Testing

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

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

Quoting …

"Things to Remember"

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

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

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

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

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

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

A little more detail on Andrew Vickers' "Things to Remember" reveals the logic of hypothesis testing.

1. "Inference involves testing a hypothesis.."

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

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

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

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

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

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

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

A p-value number (such as .05) is a probability calculation. It's not really a calculation that the "***data*** *would be at least as extreme as those observed"*…It is the calculation that "***some statistic*** *would be at least as extreme as that observed"*. The statistic might be the sample mean. **Example, continued. If your null hypothesis is that your sample of n=25 IQ test scores are a simple random sample from a normal distribution with mean μ=100 and your observed sample mean is $\bar{X}=81$ then a one sided p-value might be the calculation of $\Pr[\bar{X} \leq 81]$ under the null hypothesis assumption that μ=100**

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

The logic of statistical hypothesis testing is a **proof by contradiction argument.** Having first assumed that the null hypothesis true, you then examine where this takes you in light of the observed data. Does the null hypothesis, in light of the data, take you to an unlikely outcome (a small p-value)? If so, the null hypothesis is inconsistent with the data. Since the data can't be wrong, the problem must be with the null hypothesis. The null hypothesis is then rejected.. **Example, continued. Suppose that** $\Pr[\bar{X} \leq 81]=.02$ **under the null hypothesis assumption that** $\mu=100$**. This is saying the following: The chances were 2 in 100 of obtaining an average IQ of 81 or lower if** $\mu=100$**. These are small chances! So small that we will reject the null hypothesis and conclude that the mean IQ for the population that gave rise to this sample is some figure that is lower than the null hypothesis model values of 100.**

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

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

Hence the quick rinse under the tap….

As we will see, statistical inference is not biological inference.

3.2 P-value Approach

There are two approaches to performing a statistical hypothesis test: 1) p-value; and 2) critical region. Here, we consider the p-value approach. The critical region approach is introduced in Section 3.4.

Steps in Hypothesis Testing
P-Value Approach

1. Identify the research question.

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

3. Specify H_O and H_A.

4. "Reason" an appropriate test statistic.

5. Specify an "evaluation" rule.

6. Perform the calculations.

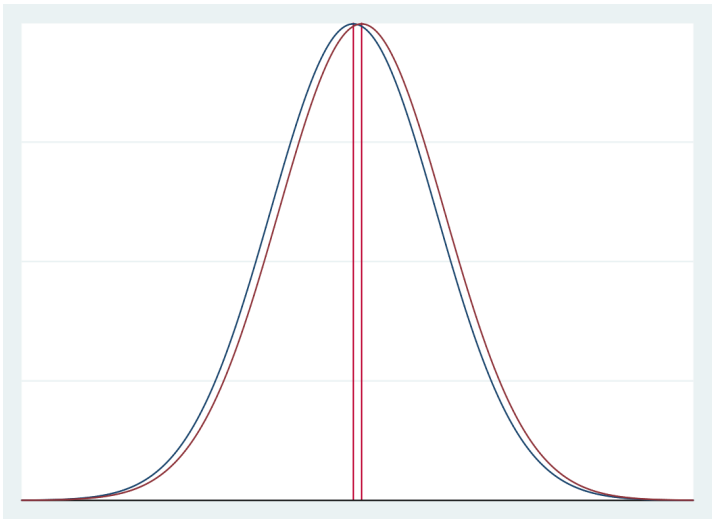
7. "Evaluate" findings and report.

8. Interpret in the context of biological relevance.

9. Compute an appropriate confidence interval estimate.

Schematic of Statistical Hypothesis Testing Using P-Value Approach

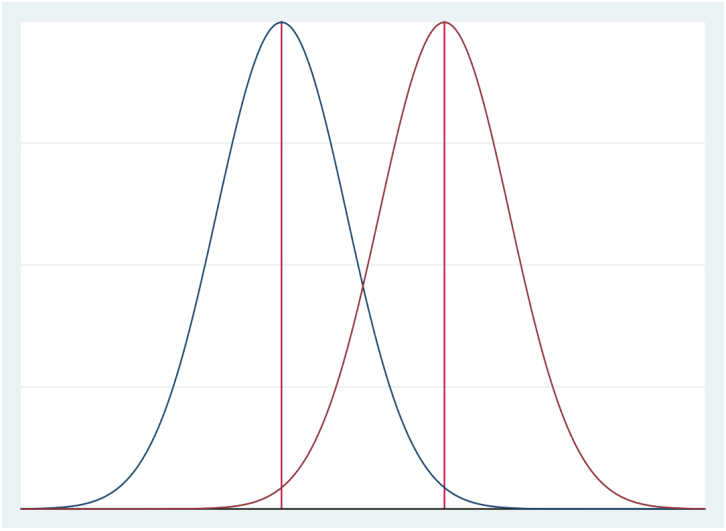
In each picture below, the data is summarized using $\bar{X}$. Above it are two probability models that might have given rise to the data. One is a null hypothesis probability model (in blue). The other is the true (in red). Notice that, in both scenarios, the value of $\bar{X}$ tends to be close to its mean.



$\bar{X}$

In this picture, the two probability distributions (null & true) are essentially the same. The sample mean is close to its "true" expected value (a population mean).

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



$\bar{X}$

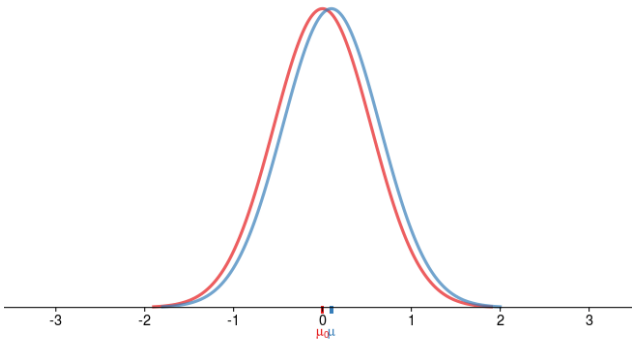
In this picture picture, the null hypothesis probability model is the left hand curve. The true distribution is the right hand curve. Here, too, the sample mean is close to its "true" expected value (a population mean).

Take home – The null hypothesis ***is NOT*** consistent with the data. We say this because the sample mean is far away from the null hypothesis model mean. This inconsistency or contradiction leads us to reject the null hypothesis model given the data.

Stata command used: zdemo2 mean1 sd1 mean2 sd2

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

Sorry!!! The colors are now reversed (Null is red, True is blue)



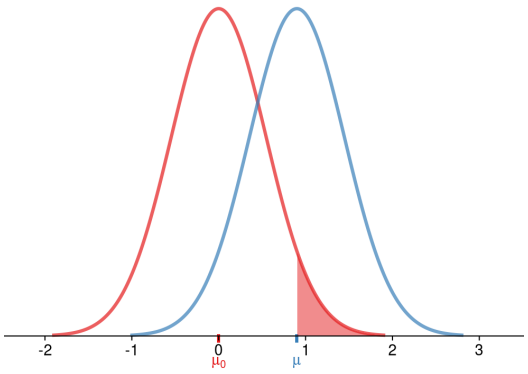
Step 1 –Begin by assuming that the null hypothesis is true

Under the null hypothesis assumption model, the two probability distributions ("true" and "null hypothesis") are identical or very nearly the same. This is why the two curves are right on top of each other. ***Note –*** So far, no data is being considered or incorporated into the picture in any way.

$\bar{X}$

Step 2 – Consider the observed sample mean.

In this picture, ONLY the data is being considered. Any models that might be considered are not incorporated into the picture in any way. Thus, no probability distribution models are shown. ***Remember -*** we don't actually know which distribution gave rise to the data.



$\bar{X}$

Step 3 – Argue "yes" or "no" consistency of the null hypothesis model assumption with the data.

Now we put the two together, namely the data together with an overlay of two models. In this picture, the "true" distribution that gave rise to the data is on the right. The null hypothesis assumption model is on the left. The shaded area is a probability calculation under the assumption that the null is true:

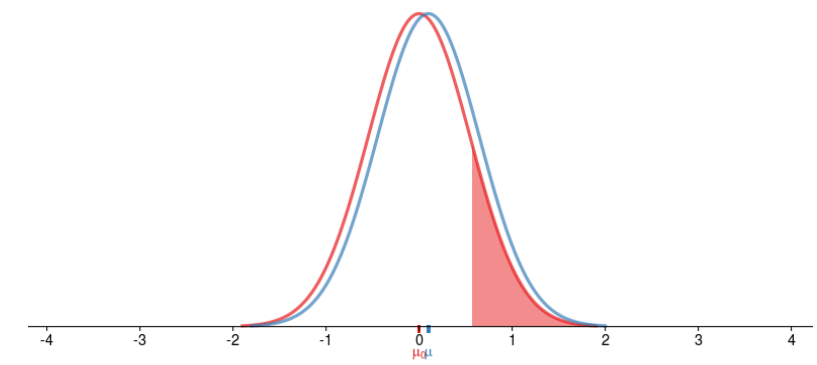
$$\Pr[\bar{X} \geq \text{observed} \mid \text{assuming null model}].$$

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

Small probability says "Assuming the null led to an unlikely event."
Large probability says "Assuming the null led to a likely event."

Source: https://istats.shinyapps.io/power/

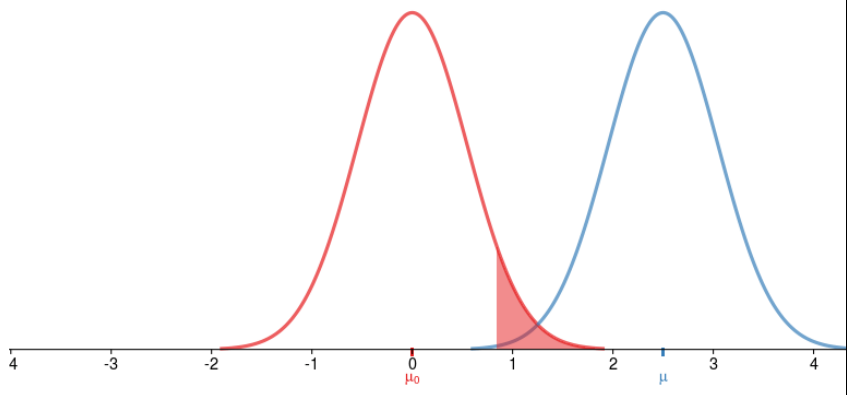
A closer look at Pr [$\bar{X}$ ≥ observed value | assuming null hypothesis model is true]



Scenario 1 - NULL is true

Pr [$\bar{X}$ ≥ observed value | assuming null is true] **= large**

- The sample mean tends to be close to its "true" expected value, and this is also close to the null hypothesis model expected value. .
- The null hypothesis model probability that the sample mean $\bar{X}$ is as far away from the null hypothesis mean, or more extreme, is a large probability (large shaded area). That is - assuming the null hypothesis model leads to a likely outcome.
- Statistical decision - "do NOT reject the null".



Scenario 2 - ALTERNATIVE is true

Pr [$\bar{X}$ ≥ observed value | assuming null is true] **= small**

- Observed sample mean is ***not*** close to null mean. Rather it tends to be close to its true mean which is now the alternative
- The null hypothesis model probability that the sample mean $\bar{X}$ is as far away from the null hypothesis mean, or more extreme, is a small probability (small shaded area). That is - assuming the null hypothesis model leads to an unlikely outcome.
- Statistical decision - "REJECT the null".

Source: https://istats.shinyapps.io/power/

p-value =

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

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

Three ways of saying 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.

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

This illustration follows the steps outlined on page 9.

1. Identify the research question

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

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

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

2. Specify the null and alternative hypotheses

H_O: $\mu_{true} = \mu_O \leq 38.3$ **months**
H_A: $\mu_{true} = \mu_A > 38.3$ **months**

Note – Strictly speaking, the null and alternative hypotheses must cover all possibilities. That's why they are written as you see them here. In calculating the p-value, however, we must choose a single null hypothesis value of the mean. We choose the value that is at the "border" between the null and alternative possibilities. This gives us the most "conservative" calculation of the p-value. Thus, we will use $\mu_O = 38.3$ ***in step #5.***

3. Reason "proof by contradiction"

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

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

4. Specify a "proof by contradiction" rule.

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

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

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

5. Calculate the null hypothesis model (H_O) chances of "extremeness." This value will be the p-value.

Under the null hypothesis model::

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

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

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

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

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

Z-score=1.99 says:

"The observed mean of 46.9 is 1.99 SE units away from the null hypothesis expected value of 38.3"

Logic of Proof-by-Contradiction says:
"Under the assumption that the null hypothesis is true, there are 2 in 100 chances of obtaining a sample mean as "extremely" far away (or more extremely far away) from 38.3 as the value of 46.9"

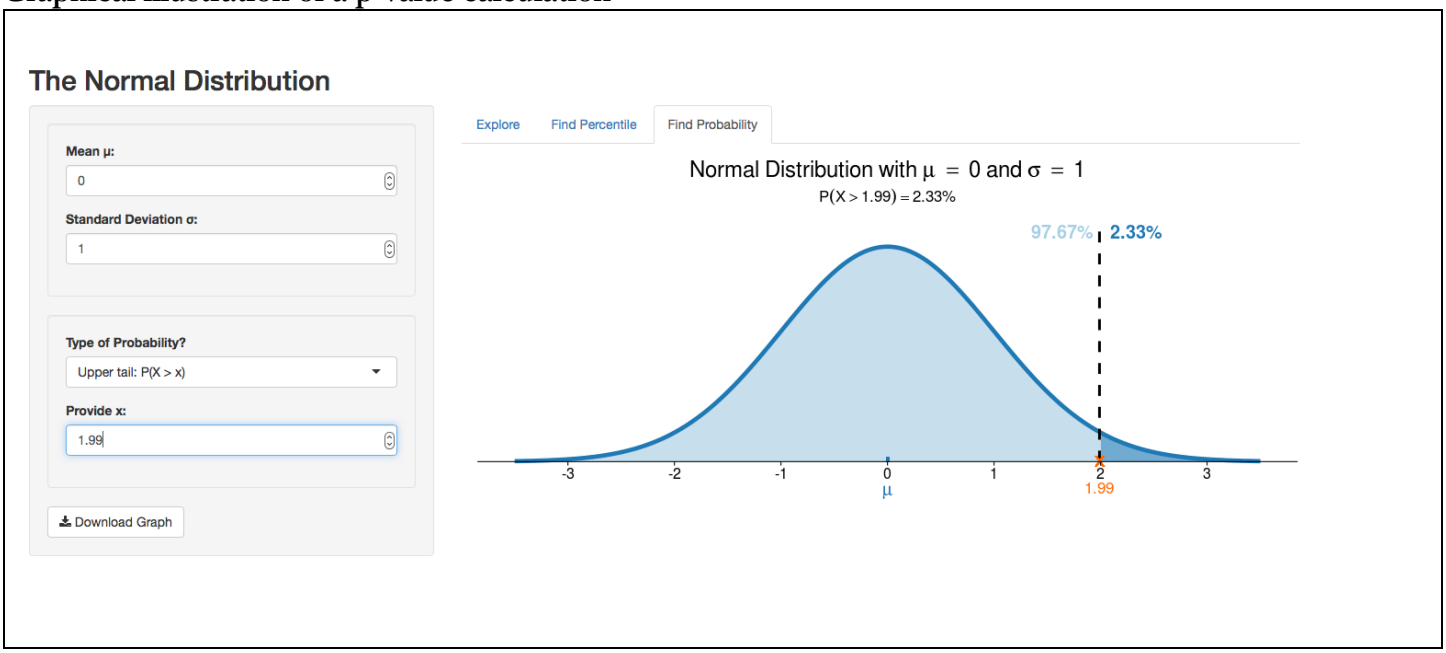
$$\Pr\left[\bar{X}_{n=100} \geq 46.9 \mid \mu_{true} = \mu_{null} = 38.3\right]$$

$$=\Pr[\text{Z-score} \geq 1.99]=.0233$$

Statistical Reasoning of "likely" says:
"If the null hypothesis, when examined in light of the data, leads us to something that is 'unlikely', namely a small p-value (shaded area in blue below), then the null hypothesis is severely challenged and possibly contradicted. →

Statistical rejection of the null hypothesis.

Graphical illustration of a p-value calculation



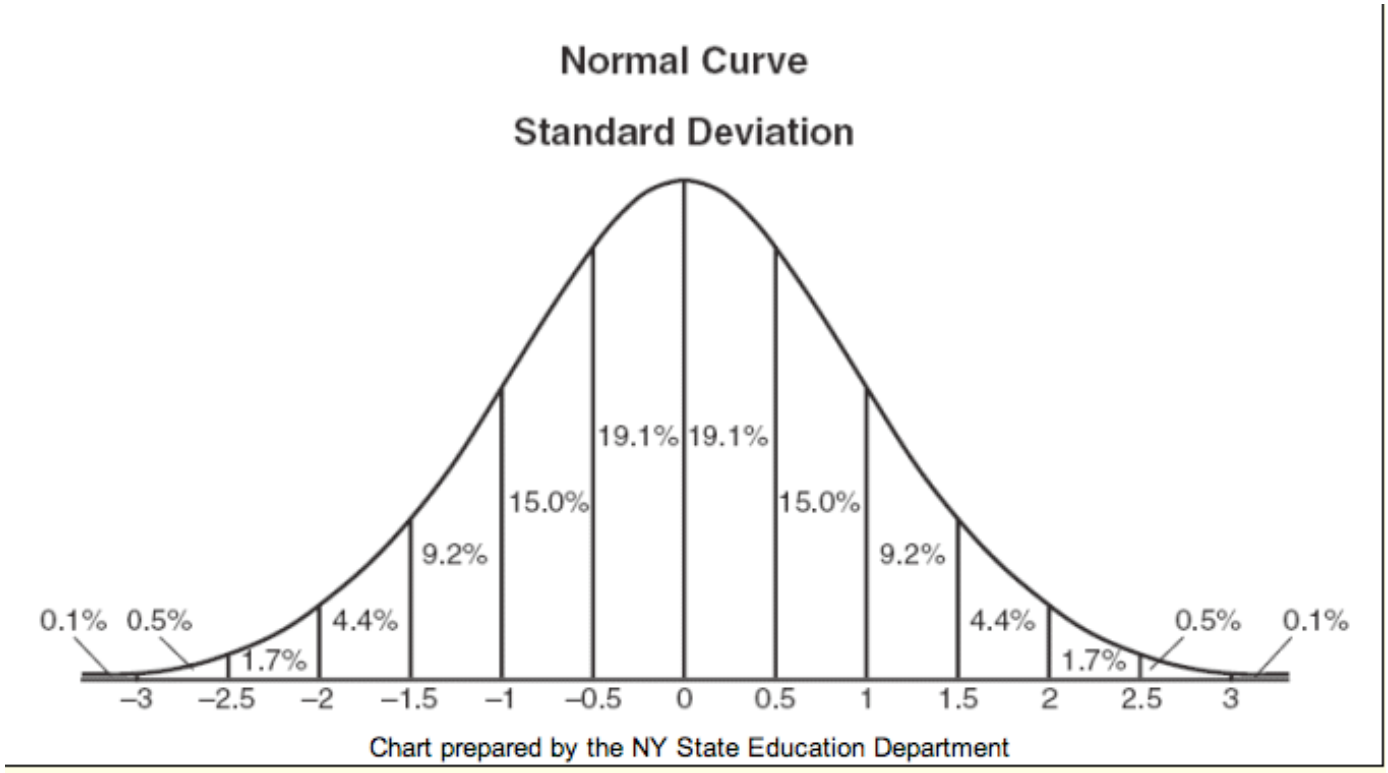
https://istats.shinyapps.io/NormalDist/

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

$$\text{Z-score}=\frac{\text{Signal}}{\text{Noise}}=\frac{\text{observed-expected}}{\text{SE(observed)}}$$ $$=\left[\frac{\bar{X}_{n=100}-\mu}{\left(\sigma/\sqrt{n}\right)}\right]$$ **Example: z-score=1.99**	*Z-Score* = *The magnitude of the departure, from the null hypothesis expectation, of the observed sample statistic (in this case the sample mean), expressed on the scale of SE units.*
p-value = Pr[Normal(0,1) ≥ z-score] **Example: pr[normal(0,1) ≥ 1.99] = .0233**	p-value = The chances of obtaining a departure of this magnitude, or greater, calculated under the presumption that the null hypothesis is true.

TIP!

The Z-score is a very handy measure of extremeness
spoiler … and so is the T-score (coming soon)!



Example –

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

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

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

3.3 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!

- Our starting point is the assumption of the null hypothesis model; that is we presume that H_0 is true. We then used this model to estimate the likelihood of what we actually observed, namely our test statistic value, or something more extreme. Depending on how big this likelihood is:
  - **It's all about the presumed null hypothesis model. We either abandon (reject) the null hypothesis (H_0) model, or we retain it (fail to reject).**
  - **We do *not* prove that the null hypothesis assumption is correct.**

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

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

Introduction to Type I Error

- **IF H_0 is true and we (incorrectly) reject H_0**
  - We have made a type I error
  - We can calculate its probability as Pr [type I error]= α

Introduction to Type II Error

- **IF H_a is true and we (incorrectly) fail to reject H_0**
  - We have made a type II error
  - We must have a specific H_a model before we can calculate Pr [type II error]= β

Introduction to Power

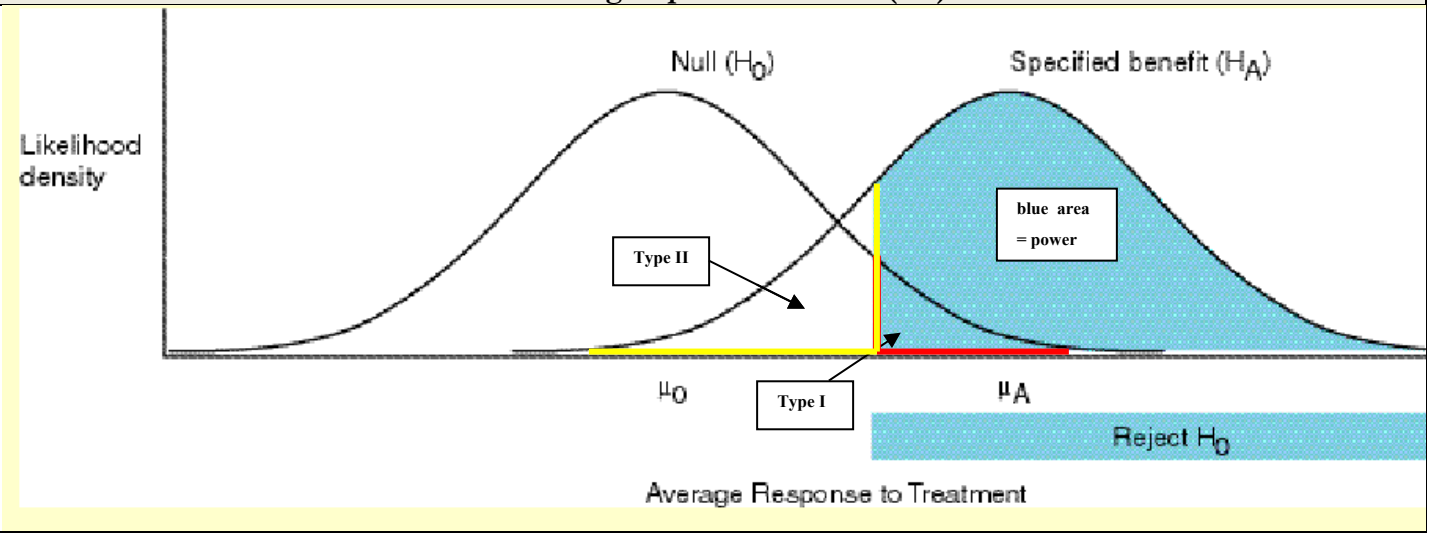
- **IF H_a is true and we (correctly) reject H_0**
  - This occurs with probability = $(1-\beta)$ which we call the "POWER"

The goal is to get the right answer (power). Either type of error is undesirable. We'd like to minimize the chances of either a type I error (probability = α) or a type II error (probability= β).

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

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

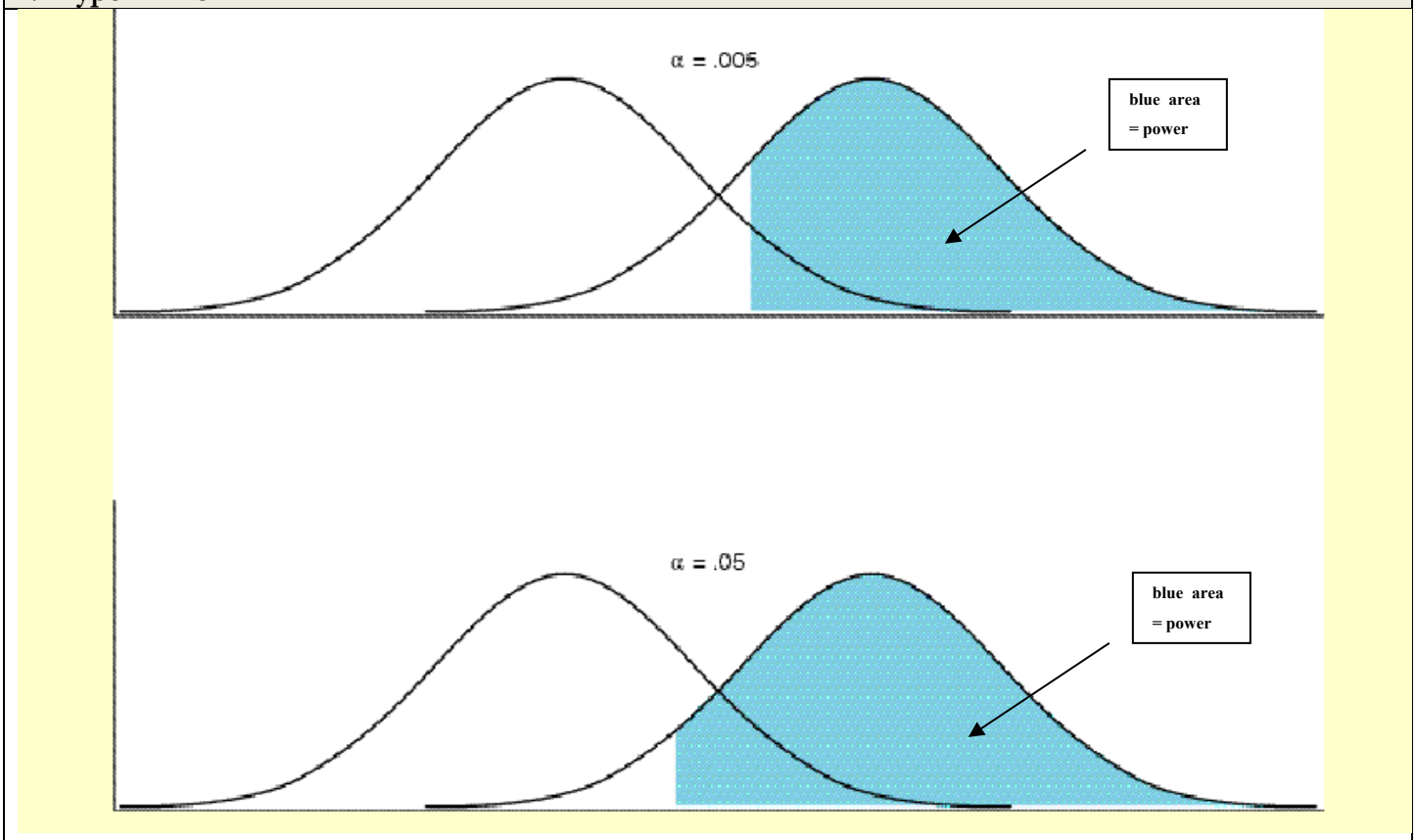
Statistical Power is the Likelihood of inferring a specified benefit (H_A). This is the blue area



Key:

- **Blue ribbon along the horizontal axis with "reject H_0" typed inside:** These are the values of the sample average that will prompt rejection of the null hypothesis, also called the **critical region**. **Note – "Critical regions" and "critical region tests" are introduced and explained beginning on page 31**
- **Blue area under the Null (H_0) curve: The type I error (probability = α).** This is the probability of mistakenly rejecting the null hypothesis; thus, it is calculated under the assumption that H_0 is true.
- **White area under the Alternative (H_A) curve: The type II error (probability = β).** This is the probability of mistakenly inferring the null; thus, it is calculated under the assumption that H_A is true**.**

The Power of a Study Depends on Four Parameters
1. Type I Error



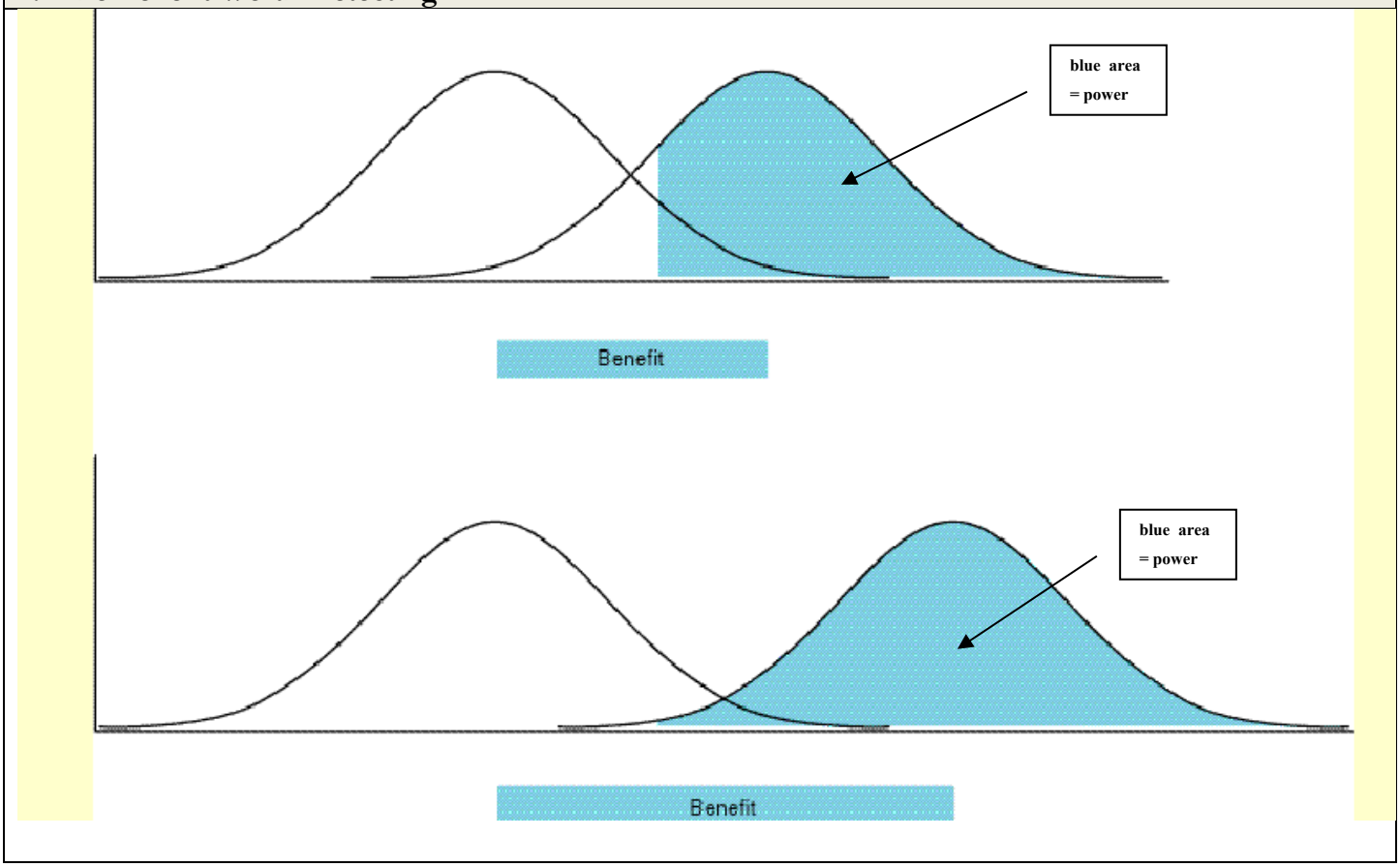
Take home message: ***If you're willing to live with more type I error, then you can increase your chances of inferring a treatment benefit (the alternative).***

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

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

- Thus, all other things being equal, use of a smaller p-value criterion (e.g. 0.005 versus 0.05) reduces the power to detect a true alternative explanation (the blue area in the top panel is smaller than the blue area in the bottom panel).

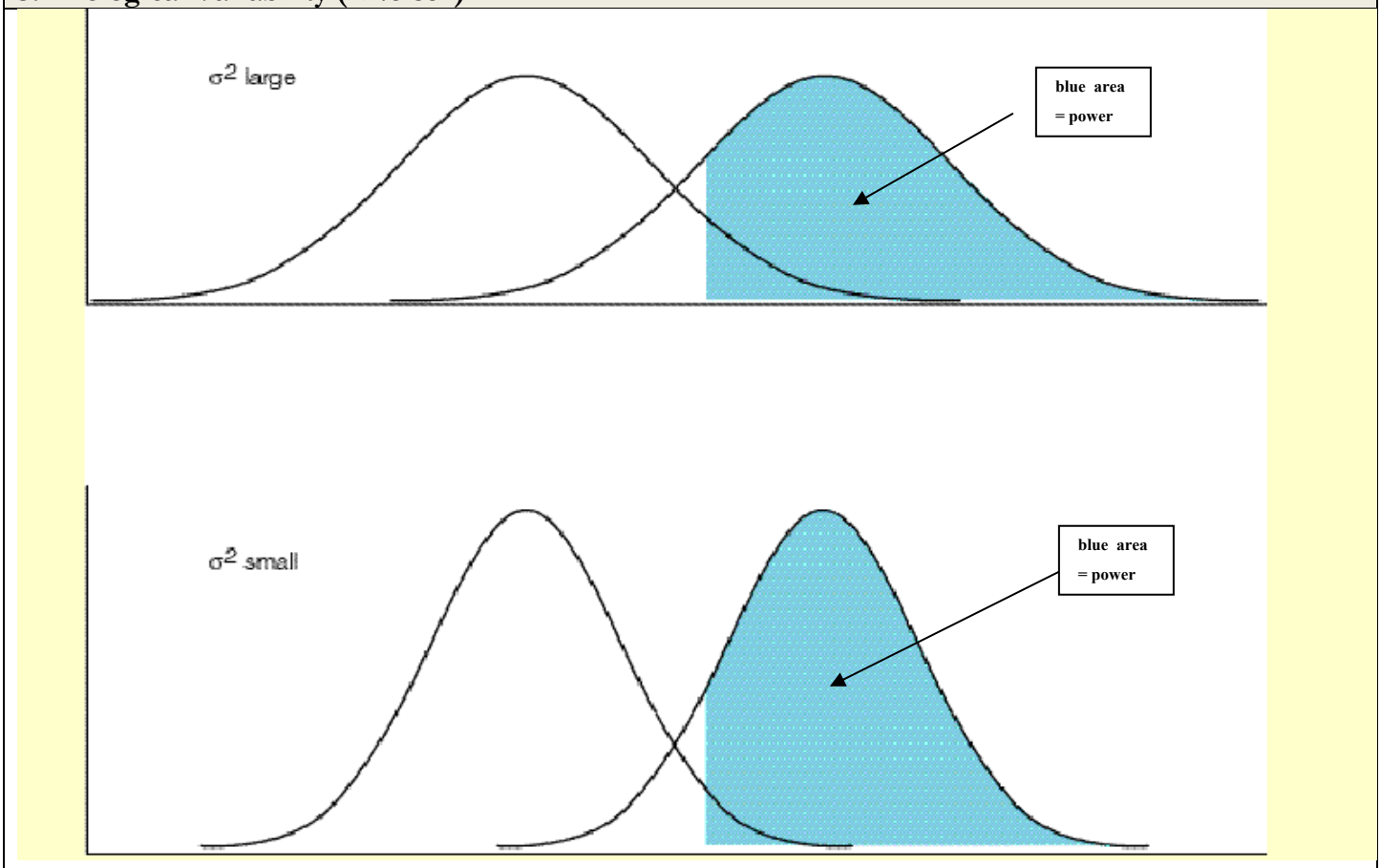
The Power of a Study Depends on Four Parameters
2. The Benefit Worth Detecting



***Take home message:** The farther away the alternative (treatment benefit) is from the null hypothesis, the greater your chances of inferring it. Conversely, the closer the alternative (treatment benefit) is to the null hypothesis, the lower your chances of inferring it.*

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

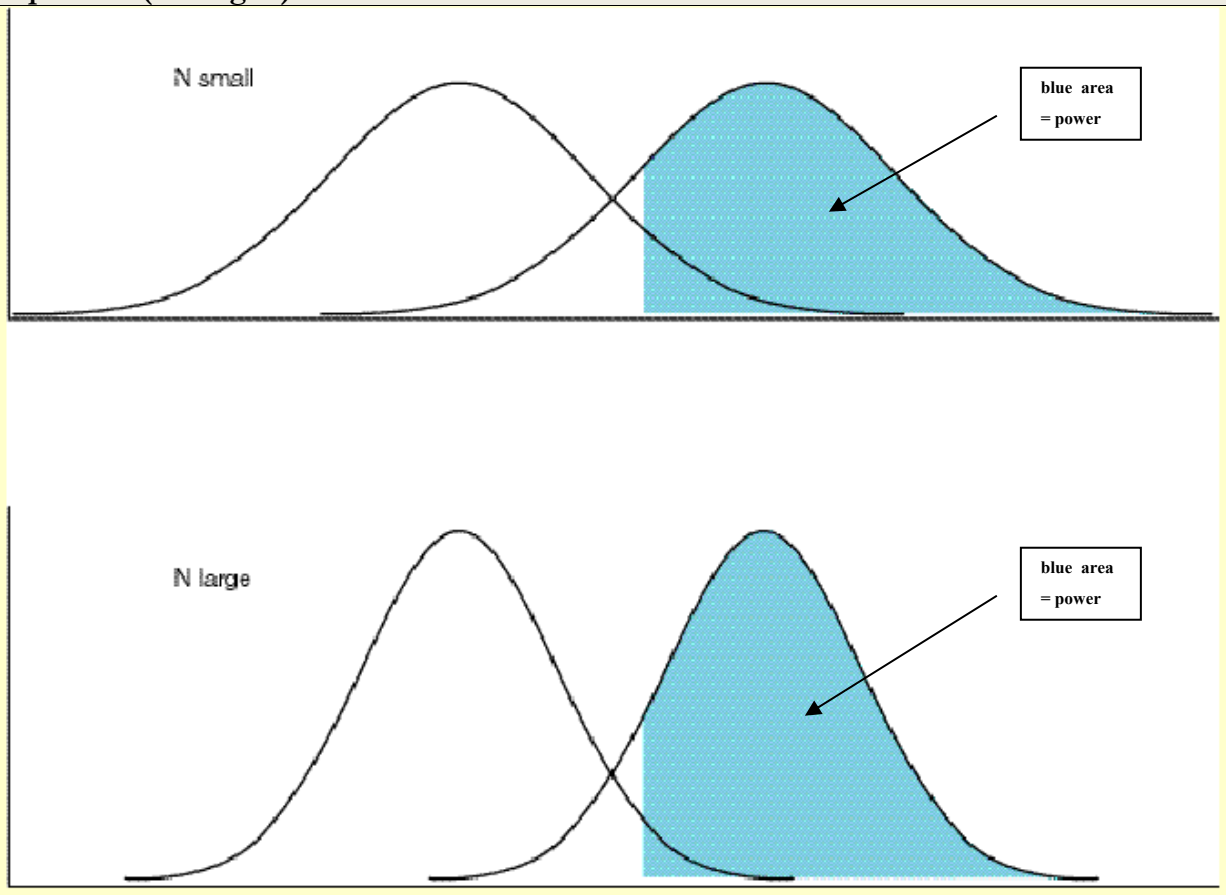
The Power of a Study Depends on Four Parameters
3. Biological Variability ("Noise")



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

- In this picture, the null hypothesis is the same in the top and bottom panels. As well, the alternative hypothesis is the same in the top and bottom panels.
- The distinction is that the underlying variability of the outcomes (a combination of naturally occurring biological variability and measurement error) is smaller in the bottom panel.
- The "threshold" value of the sample mean that prompts rejection of the null hypothesis is the SAME in both top and bottom panels.
- Here, all other things being equal, using a measurement tool that is less noisy (more precise) will increase study power (the blue area under the curve).

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



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

- In this picture, the null hypothesis is the same in the top and bottom panels. As well, the alternative hypothesis is the same in the top and bottom panels.
- In this picture, too, the underlying variability of the outcomes (a combination of naturally occurring biological variability and measurement error) is the same in the two panels.
- However, the sample size N is larger in the bottom panel. The result is that the SE of the sample mean ($SE(\bar{X})=\sigma/\sqrt{n}$) has a smaller value (by virtue of division in the denominator by a larger square root of n).
- Here, all other things being equal, using a larger sample size will increase study power (the blue area under the curve).

3.4 Critical Region Approach

Steps in Hypothesis Testing
Critical Region Approach

1. Identify the research question.

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

3. Specify H_O and H_A.

4. "Reason" an appropriate test statistic.

5. Set the value of the type I error, defined as the (null hypothesis) probability of rejecting the null hypothesis when it is in fact true.

6. Using the type I error, determine the critical region, defined by "critical" value(s) of the test statistic. Observed test statistic values falling in the associated critical region will prompt statistical rejection of the null hypothesis.

7. Now obtain study data and calculate the observed value of the test statistic.

8. "Evaluate" findings. Is the observed test statistic in the critical region (reject the null) or not in the critical region (do not reject the null)?

9. Interpret in the context of biological relevance.

10. Compute an appropriate confidence interval estimate.

Schematic of Statistical Hypothesis Testing Using Critical Region Approach

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 true null hypothesis (in error) under certain extreme conditions,
- **Then** what values of my test statistic will lead to rejection of the true null hypothesis if I want my **type I error** to be a certain value?

How do Critical Regions Work?

- We agree ***in advance (prior to collecting data)*** that we will honor a ***threshold test statistic value (this is what we mean by critical value)***, beyond which we will reject the null hypothesis. We'll do this even though, theoretically under the null hypothesis, such extreme values are still possible.
- This means that whenever we then do get a test statistic value that is beyond the critical value, ***if the null hypothesis is actually true***, then we will have ***incorrectly reject a true null hypothesis.*** Under these circumstances, we will have made a ***type I error***.
- ***How does this actually work?*** In developing a critical region test, we determine, ***before actual data collection***, the ***threshold statistic value*** (the "***critical value***") and the ***range of extreme values that lie beyond*** (this is called the ***critical region***) that will (in the future – after obtaining our data) prompt (rightly or wrongly!) rejection of the null hypothesis.

Example, continued

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

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

Null Hypothesis Probability Model Assumptions.

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

Null and alternative hypotheses

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

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

The appropriate Test Statistic is a Z-Score

The null hypothesis gives us the following:

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

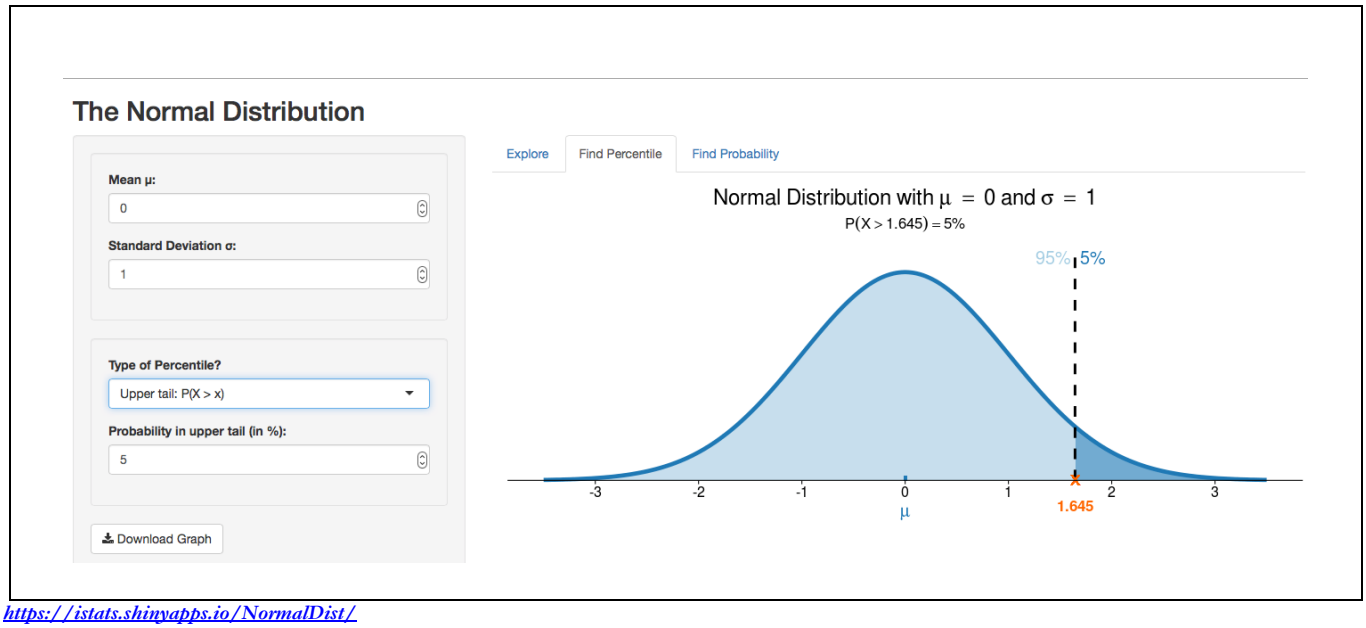
$$\text{Test Statistic} = \text{z-score} = \frac{\bar{X}_{n=100} - \mu_{null}}{SE(\bar{X}_{n=100})}$$

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

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

Step 2: Solve for the threshold and critical region of the Z-Score test statistic
Here is where our 5% chances (see previous page) get used.
- Assume the null hypothesis is true
- Under the null hypothesis assumption our **Z-statistic** will be distributed Normal(0,1)
- What cut off value of Z-statistic will prompt an incorrect rejection of the null?
- Answer: That value for which area under the curve in the right tail = .05
- That is: ***5% chances*** and ***"beyond which"*** translates to ***right tail area = .05***.

Now we go to our "art of stat" link *https://istats.shinyapps.io/NormalDist/*
- To get cut-off value of Z-statistic, click at top on the tab **Find Percentile**
- At left, from the dropdown **Type of** Percentile, choose **"Upper tail: P(X > x)"**
- Art of Stat now returns the Z-socre critical value = **1.645.**



https://istats.shinyapps.io/NormalDist/

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

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

$$\text{z-score} \geq 1.645 \rightarrow$$

$$\frac{\bar{X}_{n=100} - \mu_{null}}{SE(\bar{X}_{n=100})} \geq 1.645 \rightarrow$$

$$\bar{X}_{n=100} - \mu_{null} \geq (1.645)*SE(\bar{X}_{n=100}) \rightarrow$$

$$\bar{X}_{n=100} \geq \left[(1.645)*SE(\bar{X}_{n=100})\right] + \mu_{null} \rightarrow$$

$$\bar{X}_{n=100} \geq \left[(1.645)*(4.33)\right] + 38.3 \rightarrow$$

$$\bar{X}_{n=100} \geq 45.42$$

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

Step 4: Interpret:

So where are we at this point? Answer: we've merely "set the rules". Going forward, after we have collected data and computed our test statistic, we will follow the following rule: If the future observed $\bar{X}_{n=100}$ is ***at or beyond*** the threshold value of 45.42, we will infer the alternative hypothesis, even though such extreme values are theoretically possible when the null hypothesis is true. That is – "this ***critical region*** one sided .05 test of the null versus alternative hypotheses has been defined to reject the null hypothesis for any $\bar{X}_{n=100} \geq 45.42$.

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

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

Interpret.

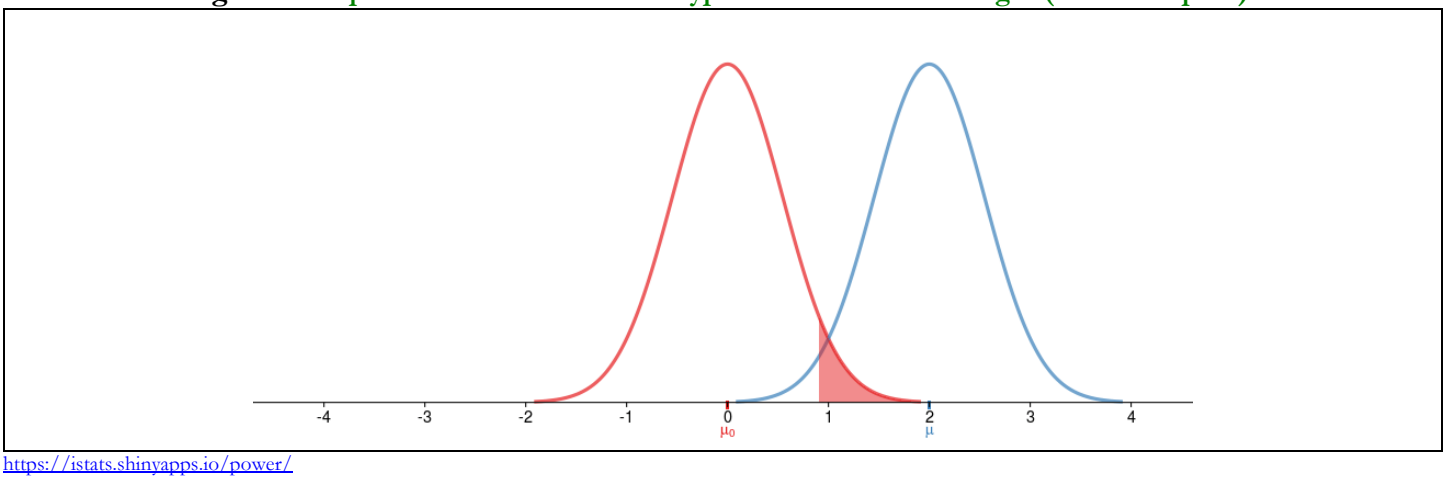
Because the observed $\bar{X}_{n=100} = 46.9$ ***exceeds the value of the threshold 45.42*** and is, therefore, ***in the critical region***, in critical region parlance we say *"it is significant at the 0.05 level".* → **reject the null hypothesis.** The conclusion is the same: these data provide statistically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

3.5 One Sided versus Two Sided tests **(p-values are areas under the NULL distribution in the direction of the alternative)**

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 <u>one sided test</u> because the alternative hypothesis probability models are to one side of the null hypothesis model:

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

Rejection of the null hypothesis occurs for extreme values of the test statistic ***in the direction of the alternative***, in this case ***to the right***: **Here p-value = Area under null hypothesis model to the right (shaded in pink)**

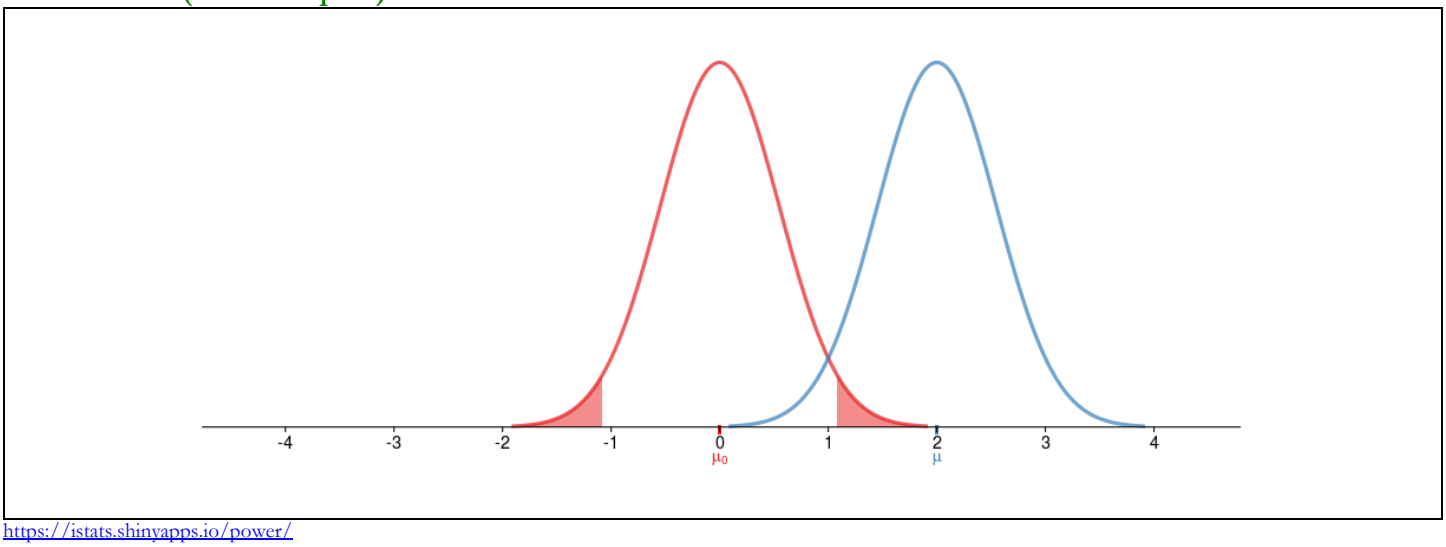


https://istats.shinyapps.io/power/

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 <u>two sided test</u> because the alternative hypothesis probability models are on either side of the null hypothesis model:

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

Again, rejection of the null hypothesis occurs for extreme values of the test statistic ***in the direction of the alternative***, in this case ***to the right AND to the left***: **Here p-value = Area under null hypothesis model to the right and to the left (shaded in pink)**



https://istats.shinyapps.io/power/

How to Calculate a TWO sided p-value

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

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

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

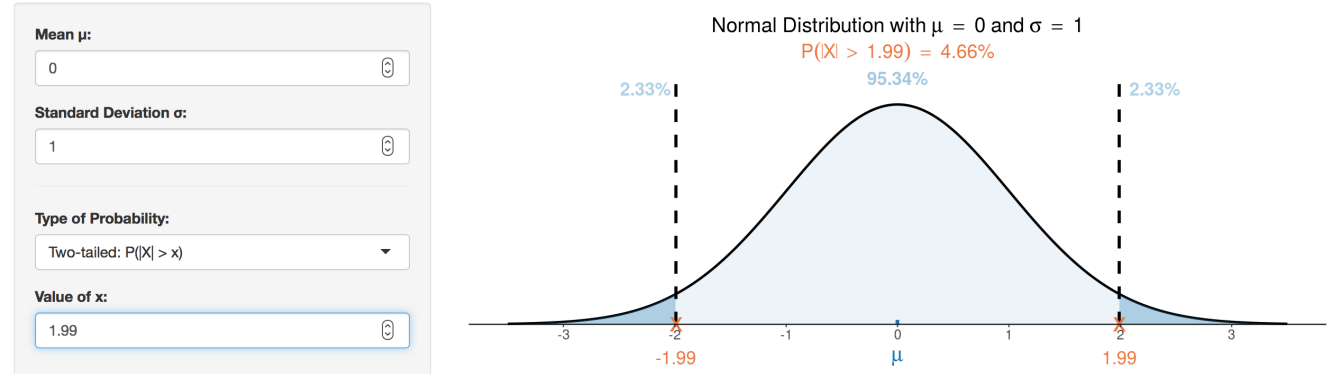
Interpretation: The observed mean is 1.99 SE units away (to the right) of the null

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

$$\text{p-value}_{\text{TWO SIDED}} = \text{Prob}[\text{Normal}(0,1) \geq +1.99] + \text{Prob}[\text{Normal}(0,1) \leq -1.99]$$
$$= (2)\,\text{Prob}[\text{Normal}(0,1) > 1.99]$$
$$= (2)(.0233)$$
$$= .0466$$

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 4.6 chances in 100.

artofstat.com > WebApps > Normal Distribution > FIND PROBABILITY > Type of probability: "Two –tailed"
The app returns Pr[Normal(0,1) > | 1.99|] = .0466. Thus, the 2-sided p-value = .0466



3.6 Beware the Statistical Hypothesis Test

Statistical significance is not biological inference. P-values, *by themselves*, don't help us much. Other criteria are essential to biological inference.

1. Statistical Significance is NOT Biological Inference.

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

> Results for patients receiving treatment "A" are ***statistically significantly*** better than results for patients receiving treatment "B".

There are actually multiple, different, explanations:

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

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

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

2. P-values, *by themselves*, don't help us much.

Definition p-value

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

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

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

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

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

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

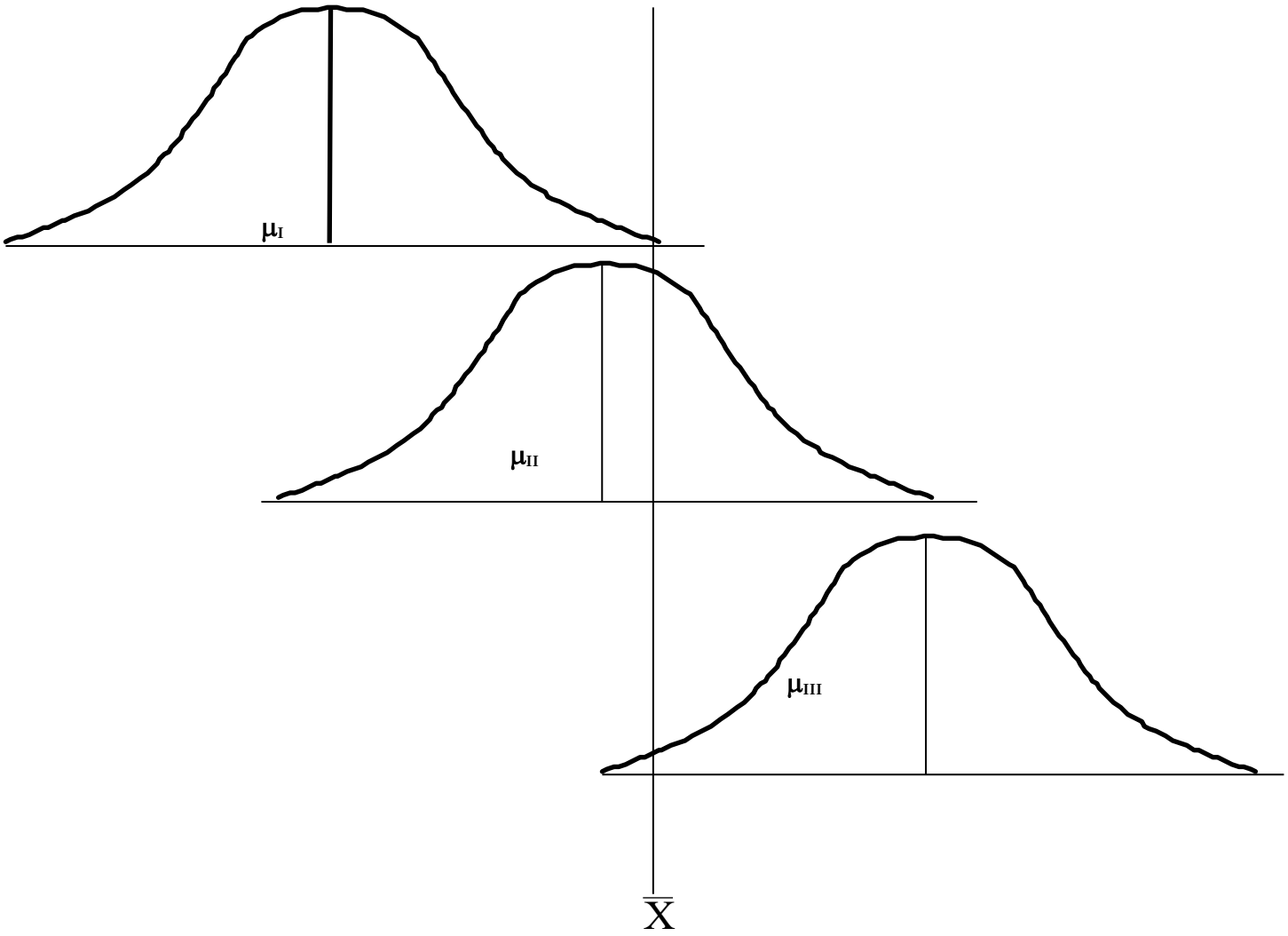
- **A conclusion of a treatment effect is *weakened* by**
    - Its unusualness; such a finding should be "checked" with new data
    - Its isolation; that is – it is observed in a selected subgroup only and nowhere else; such a finding is intriguing, however and should be explored further
    - Its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

4. Introduction to Confidence Interval Estimation

4.1 Goals of Estimation

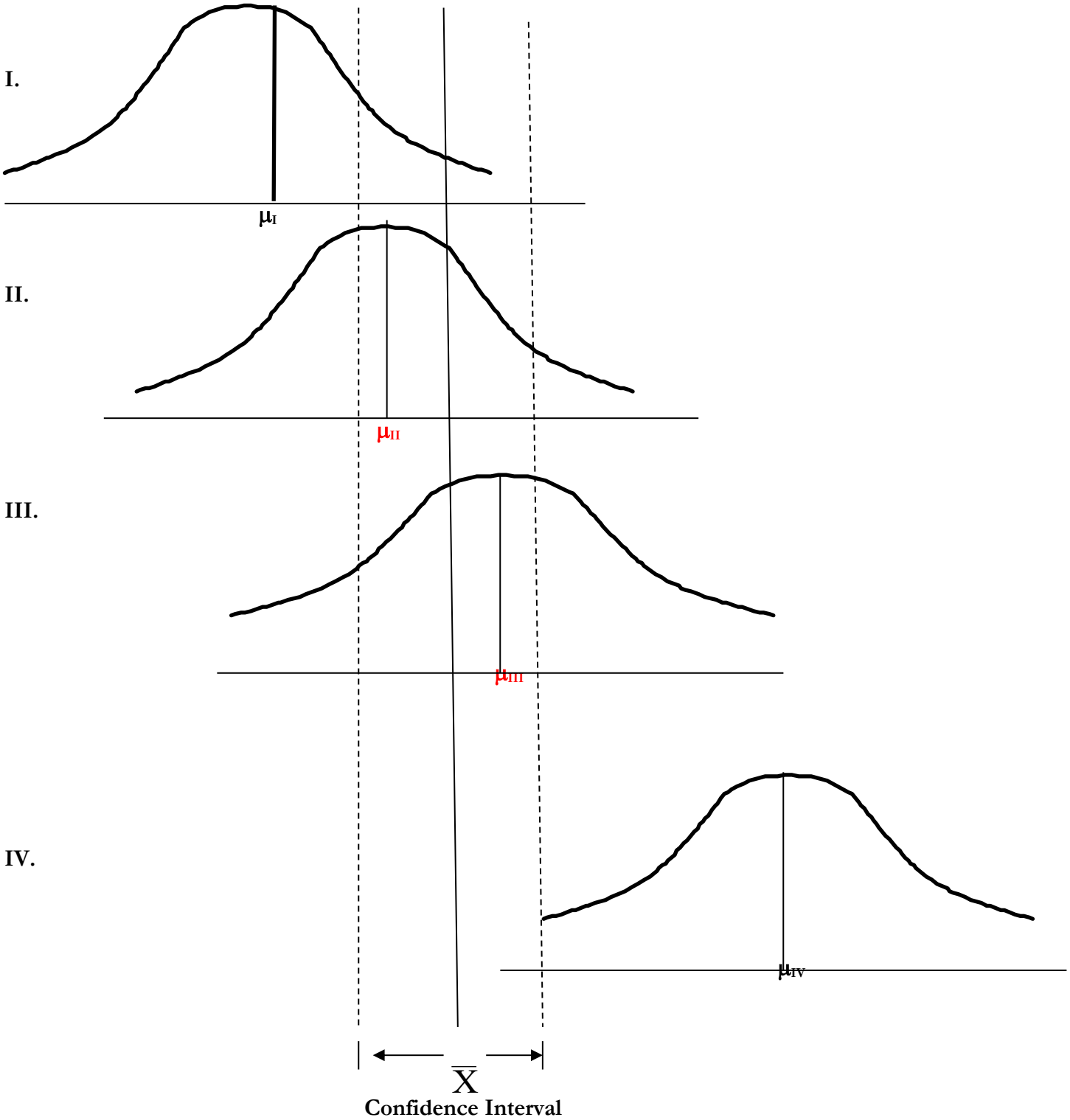
What does it mean to say we know $\overline{X}$ from a sample but we don't know the population mean μ?

Suppose we have a simple random sample of n observations $X_1 \ldots X_n$ from some population. We have calculated the sample average $\overline{X}$. What population gave rise to our sample? In theory, there are infinitely many possible populations. For simplicity here, suppose there are just 3 possibilities, schematically shown below:



Suppose this is the location of our $\overline{X}$.

Okay, sorry. Here, I'm imagining four possibilities instead of three. Look at this page from the bottom up. Around our $\bar{X}$, I've constructed a "confidence" interval. Notice the dashed lines extending upwards into the 4 normal distributions. μ_I and μ_{IV} are **outside** the interval around $\bar{X}$. μ_{II} and μ_{III} are **inside** the interval.



Confidence Interval

We are "confident" that μ could be either μ_{II} or μ_{III}.

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

Whether an estimator is "good" or "not good" depends on what criteria we use to define "good". There are potentially lots of criteria. Here, we'll use one set of two criteria: unbiased and minimum variance.

Conventional Criteria for a Good Estimator –

1. **"In the long run, correct" (unbiased)**
2. **"In the short run, in error by as little as possible" (minimum variance)**

1. Unbiased - "In the Long Run Correct" -

Tip: Recall the introduction to statistical expectation and the meaning of unbiased (See Unit 6 – *Bernoulli & Binomial* pp 8-11.

"In the long run correct." Imagine replicating the study over and over again, infinitely many times. Each time, calculate your statistic of interest so as to produce the sampling distribution of that statistic of interest. Now calculate the mean of the sampling distribution for your statistic of interest. Is it the same as the population parameter value that you are trying to estimate? If so, then that statistic is an unbiased estimate of the population parameter that is being estimated.

Example – Under normality and simple random sampling, S^2 as an unbiased estimate of σ^2.
"In the long run correct" means that the statistical expectation of S^2, computed over the sampling distribution of S^2, is equal to its "target" σ^2.

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

Recall that we use the notation "E [ ]" to refer to statistical expectation. Here it is E [S^2] = σ^2.

2. Minimum Variance "In Error by as Little as Possible" –

"In error by as little as possible." We would like that our estimates not vary wildly from sample to sample; in fact, we'd like these to vary as little as possible. This is the idea of **precision**. When the estimates vary by as little as possible, we have **minimum variance.**

Putting together the two criteria ("long run correct" and "in error by as little as possible")

Suppose we want to identify the "minimum variance unbiased" estimator of μ in the setting of a **simple random sample from a normal distribution**.

Candidate estimators might include the sample mean $\bar{X}$ or the sample median $\tilde{X}$ as estimators of the population mean μ. Which would be a better choice according to the criteria "in the long run correct" and "in the short run in error by as little as possible"?

Step 1 **First, identify the unbiased estimators**

Step 2 **From among the pool of unbiased estimators, choose the one with minimum variance.**

Illustration for data from a normal distribution

1. The unbiased estimators are the sample mean $\bar{X}$ and median $\tilde{X}$

2. variance [$\bar{X}$] < variance [$\tilde{X}$]

Choose the sample mean $\bar{X}$. It is the minimum variance unbiased estimator.

For a random sample of data from a normal probability distribution, $\bar{X}$ is the minimum variance unbiased estimator of the population mean μ.

Take home message:

Here, we will be using the criteria of "minimum variance unbiased".
However, other criteria are possible.

4.2 Notation and Definitions

Estimation, Estimator, Estimate -

- **Estimation** is the computation of a statistic from sample data, often yielding a value that is an approximation (guess) of its target, an unknown true population parameter value.
- The statistic itself is called an **estimator** and can be of two types - point or interval.
- The value or values that the estimator assumes are called **estimates.**

Point versus Interval Estimators -

- An estimator that represents a "single best guess" is called a **point estimator.**
- When the estimate is of the form of a "range of plausible values", it is called an **interval estimator.** Thus,

A point estimate is of the form:

[Value],

An interval estimate is of the form:

[lower limit, upper limit]

Example -

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

Sampling Distribution

- **Recall the idea of a sampling distribution.** It is an theoretically obtained entity obtained by imagining that we repeat, over and over infinitely many times, the drawing of a simple random sample and the calculation of something from that sample, such as the sample mean $\bar{X}_n$ based on a sample size draw of size equal to n. The resulting collection of "all possible" sample means is what we call the sampling distribution of $\bar{X}_n$.

- **Recall. The sampling distribution of $\bar{X}_n$ plays a fundamental role in the central limit theorem.**

Unbiased Estimator

A statistic is said to be an **unbiased estimator** of the corresponding population parameter if its mean or expected value, taken over its sampling distribution, is equal to the population parameter value.

Intuitively, this is saying that the "long run" average of the statistic, taken over all the possibilities in the sampling distribution, has value equal to the value of its target population parameter.

Confidence Interval, Confidence Coefficient

- A **confidence interval** is a particular type of interval estimator.
- Interval estimates defined as confidence intervals provide not only several point estimates, but also a feeling for the precision of the estimates. This is because they are constructed using two ingredients:

  1) a point estimate, and

  2) the standard error of the point estimate.

Many Confidence Interval Estimators are of a Specific Form:

lower limit = (point estimate) - (confidence coefficient multiplier)(standard error)
upper limit = (point estimate) + (confidence coefficient multiplier)(standard error)

- The "multiple" in these expressions is related to the precision of the interval estimate; the multiple has a special name - **confidence coefficient**.
- A wide interval suggests imprecision of estimation. Narrow confidence interval widths reflect large sample size or low variability or both.
- Exceptions to this generic structure of a confidence interval are those for a variance parameter and those for a ratio of variance parameters

Take care when computing and interpreting a confidence interval!!

A common mistake is to calculate a confidence interval but then use it incorrectly by focusing only on its midpoint.

4.3 How to Interpret a Confidence Interval

A confidence interval is a <u>safety net</u>.

Tip: **In this section, the focus is on the idea of a confidence interval. For now, don't worry about the details.**

Example

Suppose we want to estimate the average income from wages for a population of 5000 workers, $X_1,...,X_{5000}$

The average income that we want to estimate is the population mean μ.

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

For purposes of this illustration, suppose we actually know the population σ = \$12,573. In real life, we wouldn't have such luxury!

Suppose the unknown μ = \$19,987. Note – I'm only telling you this so that we can see how well this illustration performs!

We'll construct two confidence interval estimates of μ to illustrate the **importance of sample size in confidence interval estimation:**

(1) from a sample size of n=10, versus

(2) from a sample size of n=100

(1) Carol uses a sample size n=10
Carol's data are X_1, …, X_{10}

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

(2) Ed uses a sample size n=100
Ed's data are X_1, …, X_{100}

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

Compare the two SE, one based on n=10 and the other based on n=100 …

- The variability of an average of 100 is less than the variability of an average of 10.
- It seems reasonable that, all other things being equal, we should have ***more confidence (smaller safety net)*** in our sample mean as a guess of the population mean when it is based on a larger sample size (100 versus 10).
- …. Taking this one step further … we ought to have complete (100%) confidence (no safety net required at all) if we interviewed the entire population! This makes sense since we would obtain the correct answer of $19,987 every time.

Definition Confidence Interval (Informal):

A confidence interval is a guess (point estimate) together with a "safety net" (interval) of guesses of a population characteristic. In most instances, it is easy to see the 3 components of a confidence interval:

1) **A point estimate (e.g. the sample mean $\bar{X}$)**

2) **The standard error of the point estimate (e.g. $SE_{\bar{X}} = \sigma/\sqrt{n}$)**

3) **A confidence coefficient (conf. coeff)**

In most instances (**means, differences of means, regression parameters,** etc), the structure of a confidence interval is calculated as follows:

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

In other instances (as you'll see in the next pages), the structure of a confidence interval looks different, as for confidence intervals for

Population variance
Population standard deviation
Ratio of two population variances
relative risk
Odds ratio

Example: Carol samples n = 10 workers.

Sample mean $\bar{X}$ = \$19,887

Standard error of sample mean, $SE_{\bar{X}} = \sigma/\sqrt{n}$ = \$3,976 for n=10

Confidence coefficient for 95% confidence interval = 1.96

Lower limit = (point estimate) – (confidence coefficient)(SE) = \$19,887 – (1.96)(\$3976) = \$12,094

Upper limit = (point estimate) + (confidence coefficient)(SE) = \$19,887 + (1.96)(\$3976) = \$27,680

Width = (\$27,680 - \$12,094) = \$15,586

Example: Ed samples n = 100 workers.

Sample mean $\bar{X}$ = \$19,813

Standard error of sample mean, $SE_{\bar{X}} = \sigma/\sqrt{n}$ = \$1,257 for n=100

Confidence coefficient for 95% confidence interval = 1.96

Lower limit = (point estimate) – (confidence coefficient)(SE) = \$19,813 – (1.96)(\$1257) = \$17,349

Upper limit = (point estimate) + (confidence coefficient)(SE) =\$19,813 + (1.96)(\$1257) = \$22,277

Width = (\$22,277 - \$17,349) = \$4,928

	n	Estimate	95% Confidence Interval	
Carol	**10**	**\$19,887**	**(\$12,094, \$27,680)**	**Wide**
Ed	**100**	**\$19,813**	**(\$17,349, \$22,277)**	**Narrow**
Truth	**5000**	**\$19,987**	**\$19,987**	**No safety net**

Definition 95% Confidence Interval

If all possible random samples (an infinite number) of a given sample size (e.g. 10 or 100) were obtained and if each were used to obtain its own confidence interval,

Then 95% of all such confidence intervals would contain the unknown; the remaining 5% would not.

But Carol and Ed Each Have Only ONE Interval:

So now what? The definition above doesn't seem to help us. What can we say?

Carol says: "With 95% confidence, the interval \$12,094 to \$27,680 contains the unknown true mean μ.

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

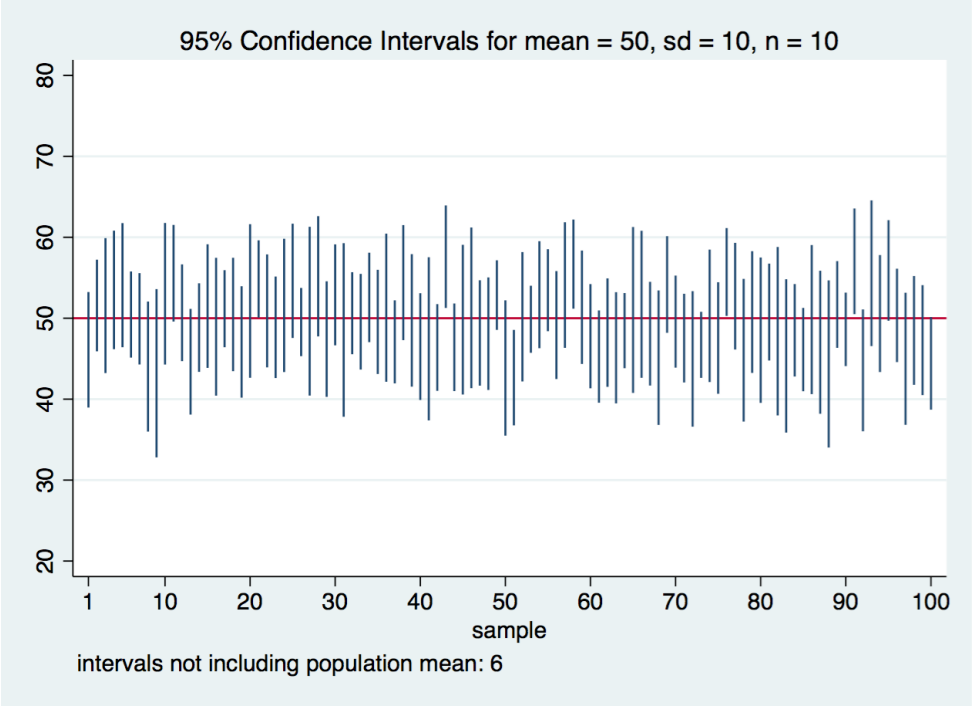
Caution on the use of Confidence Intervals:

1) It is incorrect to say – *"The probability that a given 95% confidence interval contains μ is 95%"*

A given interval either contains μ or it does not.

2) **The confidence coefficient (recall – this is the multiplier we attach to the SE) for a 95% confidence interval is the number needed to ensure 95% coverage in the long run (in probability).**

Here is a picture of 100 95% confidence intervals, each based on a sample size of n=10

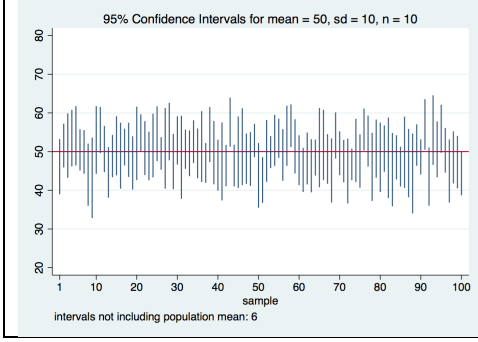
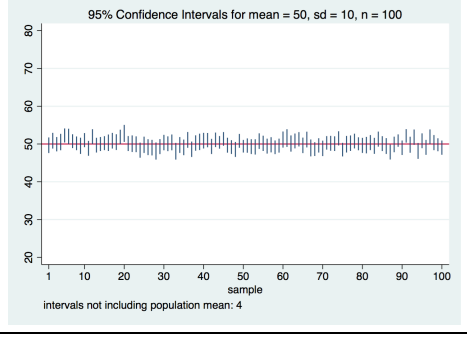
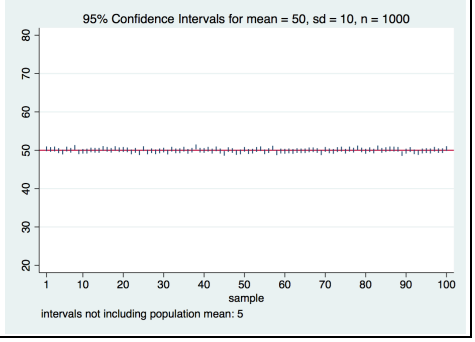


Stata command used: cidemo2 10, level(95) samples(100)

Notice …

(1) Any one confidence interval either contains μ or it does not. In fact, 6 of the intervals do not.
(2) For a given samples size (here, n=10 every time), the width of the confidence interval is the same.

Here is a picture to get a feel for the role of sample size in confidence interval width and precision of estimation

n=10	n=100	n=1000
		

Now notice …

(1) As the sample size increases from 10 to 100 to 1000, the confidence intervals are narrower (*more precise*)

(2) As n → infinity, μ is in the interval every time.

Some additional remarks on the interpretation of a confidence interval might be helpful

- Each sample gives rise to its own point estimate and confidence interval estimate built around the point estimate. The idea is to construct our intervals so that:

"IF all possible samples of a given sample size (an infinite #!) were drawn from the underlying distribution and each sample gave rise to its own interval estimate,

THEN 95% of all such confidence intervals would include the unknown μ while 5% would not"

- **Another Illustration of - It is NOT CORRECT to say:** ***"The probability that the interval (1.3, 9.5) contains μ is 0.95".*** Why? Because either μ is in (1.3, 9.5) or it is not. For example, if μ=5.3 then μ is in (1.3, 9.5) with probability = 1. If μ=1.0 then μ is in (1.3, 9.5) with probability=0.

- I toss a fair coin, but don't look at the result. The probability of heads is 1/2.
  I am "50% confident" that the result of the toss is heads. In other words, I will guess "heads" with 50% confidence. Either the coin shows heads or it shows tails. I am either right or wrong on this particular toss. In the long run, if I were to do this, I should be right about 50% of the time – hence "50% confidence". But for this particular toss, I'm either right or wrong.

- In most experiments or research studies we can't look to see if we are right or wrong – but we define a confidence interval in a way that we know "in the long run" 95% of such intervals will get it right.