

Lab for Sampling Distribution of the Mean & Hypothesis Testing

Let's first review what we know about sampling distributions of the mean (Central Limit Theorem):

1. The mean of the sampling distribution will be equal to μ .
2. The standard error (standard deviation of the sampling distribution) typically will be less than σ . In fact, the standard error equals:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

3. As the sample size increases, the sampling distribution of the mean approaches a normal distribution, regardless of the shape of the parent population.

So, here's the good news about the sampling distribution of the mean. With all that we know about it, we can place some faith in characteristics of the sample means that we draw in research. In other words, if our sample is reasonably large, we can expect that its mean is actually drawn from a sampling distribution of the mean that is normal (regardless of the population's shape) and whose standard error is small enough that our sample mean is likely to be near the μ of the population from which it was drawn. Note that all of this is true in the abstract — when we don't even have a clue about what μ or σ might be. And, of course, in reality-land we'll rarely know the actual values for μ or σ .

Okay, how is all this information useful? Suppose that the population of IQ scores is normally distributed with $\mu = 100$ and $\sigma = 15$. Could you use this information to determine the likelihood that a sample of IQ scores has been randomly selected from the population of all IQ scores? In essence, you are testing the null hypothesis (H_0) that $\mu = 100$. What sample mean would be sufficiently different from 100 that you would be convinced that your sample was unlikely to have been drawn randomly from the population?

Hypothesis Testing

First of all, we need to define an unlikely event. An unlikely event is one that happens only rarely by chance. Try to think of some examples of unlikely events (hell freezing over, winning a lottery, tossing a fair coin and having it come up heads 20 times in a row, a human gestation period of 11 months, having an IQ of 160, getting a perfect score (1600) on the SATs, becoming president of the U.S. or a college or major business, etc.). What all such events have in common is that, all else being equal, they will occur only rarely.

Significance Level (α -level)

It should make sense to you that mean IQ scores that are near 100 would be consistent with H_0 . But how much larger or smaller than 100 would the sample mean need to be in order for you to conclude that H_0 seems unreasonable? Because we are typically interested in detecting differences in either direction, we will use two-tailed tests. But we still need to decide that probability level would determine extreme scores. Mean IQ scores that would occur as seldom as 1% of the time might strike you as fairly unusual. So might mean IQ scores that would occur 5% of the time...or 10% of the time. Mean IQ scores that would occur more than 10% of the time would probably not strike you as all that unusual.

OK, then, what would be the z -scores associated with these extreme scores? (Remember that you must take the % and divide it equally into the two tails of the population, so 1% means placing .5%, or .005, into each tail.)

	z-score
1% (.005 in each tail)	
5% (.025 in each tail)	
10% (.05 in each tail)	

In order to use statistical information to guide us in making decisions, we need to decide — in advance — what we’re going to define as a rare statistical event. The norm that has developed in psychology (and other disciplines) is to use a probability of .05 as determining a rare event. In other words, if an event is so unlikely as to occur only 5% of the time by chance, we are going to consider it to be rare. Because we want to detect events that are rare in both directions, we will typically consider the upper and lower 2.5% to determine the unlikely events to have occurred by chance. In this course, you will always use an α -level of .05.

Null and Alternative Hypotheses

As stated earlier, we will postulate something about the population parameter, such as $\mu = 100$. Such a hypothesis is called the null hypothesis, and it is this hypothesis that we set out to test. If we reject H_0 , which we are most often trying to do, we would then claim that the alternative hypothesis (H_1) is the better claim. So, if we were taking a sample of IQ scores, we would typically test a null hypothesis that the sample was drawn from the population with $\mu = 100$ against the alternative that $\mu \neq 100$. We would succinctly state the two hypotheses as:

$$H_0: \mu = 100$$

$$H_1: \mu \neq 100$$

Making Decisions

If your sample mean is so unusual that it would occur by chance 5% of the time or less when taking random samples of that size from the population, you would reject H_0 and conclude that the alternative hypothesis (H_1) is more reasonable. Suppose that you took a sample of $n = 25$ people and had each person take an IQ test. (Your first step would be to compute the standard error.)

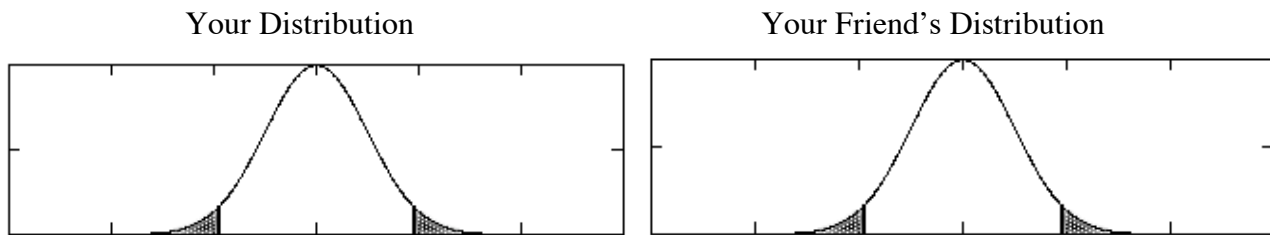
What sample means would be consistent with H_0 ?	
What sample means would lead you to think that the sample was likely to have been drawn from a population with $\mu > 100$?	
What sample means would lead you to think that the sample was likely to have been drawn from a population with $\mu < 100$?	

Impact of Sample Size

Suppose that you take a sample of $n = 9$ from the population to see if the proposed parameters are reasonable. Your friend takes a sample of $n = 25$ to test the same hypothesis. What's different between you and your friend is the sampling distribution from which your samples were drawn. What would the standard errors be for the two sampling distributions?

	Standard Error
$n = 9$	
$n = 25$	

To see what's different, calculate the scores that would determine the critical regions of the sampling distribution for both you and your friend, using $\alpha = .05$. Fill in the appropriate values on the figures below. Which of you requires larger scores to be able to reject H_0 ? Why is that so? Which of you is more likely to detect departures from H_0 ?



Errors in Decision Making

Suppose that you obtain a sample mean IQ of 109 for a sample of $n = 25$. Thus, the mean IQ of the sample would lead you to reject H_0 and conclude that the sample was drawn from a population with μ greater than 100. Keep in mind, however, that a mean IQ of 109 could represent one of two possibilities:

1. an unlikely event from the distribution of normal IQ scores, *or*
2. a more probable event from a different distribution with higher IQ scores.

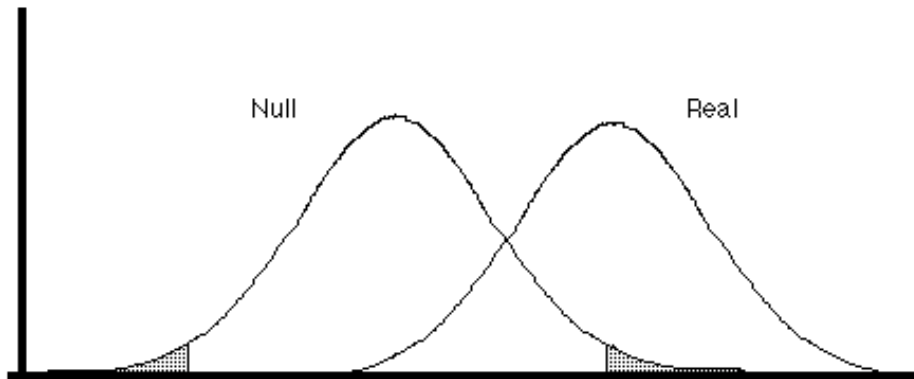
As statisticians, an outcome such as this would lead us to believe in the second alternative — that the sample was drawn from a population whose mean was greater than 100. Would we be right? Possibly, but you need to remember that we are never certain of our statistical decisions, because there is always the possibility — however slim — that the event we're observing is simply an unlikely event from the null distribution. In this case, that would mean that the sample was drawn from a population with $\mu = 100$.

We will characterize our options in the jargon displayed in the table below:

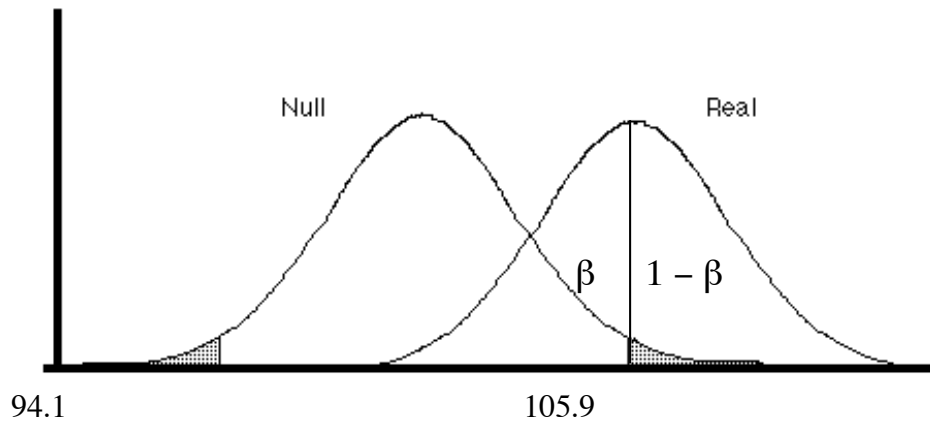
	H_0 is True	H_0 is False
Reject H_0	Type I Error	Correct Rejection
Retain H_0	Correct "Retention"	Type II Error

When we make a decision to reject H_0 , we could be correct or we could be making a Type I error. When we make a decision to retain H_0 , we could be correct or we could be making a Type II error. The probability of making a Type I error is α , and the probability of making a Type II error is β . We call the probability of a correct rejection *power*, and it has a probability of $1-\beta$.

In order to get a sense of the origin of Type I and Type II errors, you should understand the figure seen below:

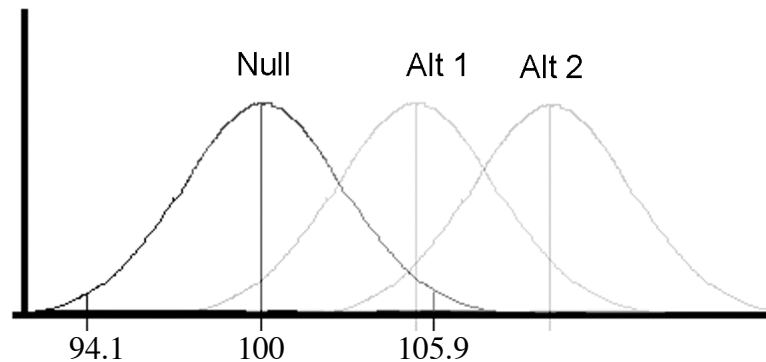


We first propose the null hypothesis, which is the curve seen on the left. We then identify the unlikely events in that curve (the two tails, each with 2.5% of the curve's area). Suppose that we could know that the score we're talking about came from the distribution on the right (the Real distribution). Most of the scores from that distribution would fall in one of the rejection regions (the right tail in this case). Thus, we could determine the probability of correctly rejecting H_0 (power), by finding the area under the Real curve marked off by scores that would lead us to reject H_0 . We could also determine the probability of making a Type II error by finding the area under the Real curve marked off by scores that would lead us to retain H_0 .



In this particular example, IQ scores between 94.1 and 105.9 would be consistent with H_0 . IQ scores less than 94.1 or greater than 105.9 would be inconsistent with H_0 .

Let's think of the situation like one in which you were making a bet. If you obtained a sample mean of 103, which distribution below would you bet it came from, the Null, Alternative 1, or Alternative 2? Can you see that the Null and Alternative 1 make more sense than Alternative 2? Can you also see that it could have actually come from any of these three distributions (as well as many other populations that I haven't illustrated)?



Given our hypothesis testing procedures, we would make the decision to retain H_0 if we obtained a sample mean of 103. However, we need to recognize that the sample could have been drawn from Alternative 1 (or some other alternative populations with means that fall between the Null and Alternative 1), in which case our decision would be a Type II Error.

Now, suppose that you obtained a sample mean of 107. Note that you would now make the decision to Reject H_0 , concluding that your sample was drawn from a population with $\mu > 100$. Thus, your sample could have been drawn from Alternative 1 or from Alternative 2. Can you see how either of those populations would make more sense than the Null (in terms of making a bet)? However, you should also see that it is possible to get a sample mean of 107 from the Null distribution, even though it is not a likely result.

Using the figure above, shade in the four options (Type I error, Type II error, Power, and Correct Retention of H_0). Can you see how Power and Type II probabilities differ depending on whether your sample mean came from Alternative 1 or Alternative 2?

IQ Testing

From Sternberg's intro text:

Two sophomores are hiking in the woods. One of them "aced" her freshman year courses, getting straight A's. Her college entrance test scores were phenomenal and she was admitted to college with a special scholarship reserved for the brightest entering students. The other student barely made it through her freshman year. Her college entrance test scores were marginal, and she just squeaked by in even getting into college. Nonetheless, people say of her that she is shrewd and clever her teachers call her "street-smart." As the friends are hiking, they encounter a huge, ferocious, obviously hungry grizzly bear. Its next meal has just come into sight, and they are it. The first student calculates that the grizzly bear will overtake them in 27.3 seconds. At that point, she panics, realizing there is no escape. She faces

her friend, the fear of death in her eyes. To her amazement, she observes that her friend is not scared at all. To the contrary, her friend is quickly but calmly taking off her hiking boots and putting on jogging shoes. “What do you think you’re doing?” the first hiker says to her companion: “You’ll never be able to outrun that grizzly bear” “That’s true,” says the companion, “but all I have to do is outrun you.”

Is there a measurable quantity called *intelligence*?

Is there only one kind of intelligence?

History

Sir Francis Galton (Darwin’s cousin) thought of intelligence as energy (the capacity for labor) and sensitivity to physical stimuli. Wissler (1901) showed that sensitivity to psychophysical stimuli was not related to intelligence.

Albert Binet and Theodisius Simon (in 1904) were charged with differentiating “mentally defective” children from those who were not succeeding in school for other reasons. The idea was to make sure that students were placed in classes from which they would profit. Binet and Simon (1916) thought that the core of intelligence is “judgment, otherwise called good sense, practical sense, initiative, the faculty of adapting one’s self to circumstances. To judge well, to comprehend well, to reason well, these are the essential activities of intelligence.”

For Binet, intelligent thought comprises three distinct elements: direction, adaptation, and criticism. *Direction* involves knowing what has to be done and how to do it. *Adaptation* refers to customizing a strategy for performing a task, then monitoring and adapting that strategy while implementing it. *Criticism* is the ability to criticize your own thoughts and actions.

$$IQ = [\text{Mental Age} / \text{Chronological age}] \times 100$$

Stanford-Binet Intelligence Scales (Lewis Terman @ Stanford)

Wechsler Scales (WAIS-R, WISC III, WPPSI)

Testing a Hypothesis About IQ

What is your sense about the IQ scores of Skidmore College students (or college students in general)? Do you think that they would be drawn randomly from a population with $\mu = 100$? Of course not! But how would you collect evidence to argue that the IQ scores of Skidmore College students were drawn from a population with $\mu > 100$? The trick is to test the null hypothesis that you think is unlikely, because if you can reject it, then you are in a position to argue that the alternative is more likely. In other words, you’d still test $H_0: \mu = 100$ against the alternative $H_1: \mu \neq 100$.

Although as a class you do not represent a random sample, we can use your IQ scores to test the null hypothesis that you, as a class, were randomly sampled from a population with $\mu = 100$. The IQ test that we will use is quite bogus, but might you have some criticism of all IQ tests? The use of IQ tests is certainly controversial in some areas, and you should become increasingly aware of the arguments swirling around this issue.