STATISTICS

Marco Caserta
marco.caserta@ie.edu

IE University

1. Hypothesis testing
2. Sample size and power
3. Statistical vs. practical significance

Hypothesis testing framework

- We start with a NULL HYPOTHESIS (H_0) that represents the "status quo."

Hypothesis testing framework

- We start with a null hypothesis (H_0) that represents the "status quo."
- We also have an alternative hypothesis (H_A) that represents our research question, *i.e.* what we're testing for, our claim.

Hypothesis testing framework

- We start with a null hypothesis (H_0) that represents the "status quo."
- We also have an alternative hypothesis (H_A) that represents our research question, *i.e.* what we're testing for, our claim.
- We conduct a hypothesis test under the assumption that the null hypothesis is true, either via simulation or traditional methods based on the central limit theorem.

Hypothesis testing framework

- We start with a null hypothesis (H_0) that represents the "status quo."
- We also have an alternative hypothesis (H_A) that represents our research question, *i.e.* what we're testing for, our claim.
- We conduct a hypothesis test under the assumption that the null hypothesis is true, either via simulation or traditional methods based on the central limit theorem.
- If the test results suggest that the data do not provide convincing evidence for the alternative hypothesis, we stick with the null hypothesis. If they do, then we reject the null hypothesis in favor of the alternative.

Number of college applications

A survey conducted at a certain university asked how many colleges students applied to, and 206 students responded to this question. This sample yielded an average of 9.7 college applications with a standard deviation of 7. College Board website states that counselors recommend students apply to roughly 8 colleges. Do these data provide convincing evidence that the average number of colleges all students at that university apply to is higher than recommended?

http://www.collegeboard.com/student/apply/the-application/151680.html

Setting the hypotheses

- The parameter of interest is the average number of schools applied to by <u>all</u> students at a given university.

Setting the hypotheses

- The PARAMETER OF INTEREST is the average number of schools applied to by all students at a given university.
- There may be two explanations why our sample mean is higher than the recommended 8 schools.
  - The true population mean is different.
  - The true population mean is 8, and the difference between the true population mean and the sample mean is simply due to natural sampling variability.

Setting the hypotheses

- The PARAMETER OF INTEREST is the average number of schools applied to by all students at a given university.
- There may be two explanations why our sample mean is higher than the recommended 8 schools.
  - The true population mean is different.
  - The true population mean is 8, and the difference between the true population mean and the sample mean is simply due to natural sampling variability.
- We start with the assumption the average number of colleges students apply to is 8 (as recommended)

$$H_0 : \ \mu \leq 8$$

Setting the hypotheses

- The parameter of interest is the average number of schools applied to by all students at a given university.
- There may be two explanations why our sample mean is higher than the recommended 8 schools.
  - The true population mean is different.
  - The true population mean is 8, and the difference between the true population mean and the sample mean is simply due to natural sampling variability.
- We start with the assumption the average number of colleges students apply to is 8 (as recommended)

$$H_0 : \mu \leq 8$$

- We test the claim that the average number of colleges students apply to is greater than 8

$$H_A : \mu > 8$$

Number of college applications - conditions

Which of the following is not a condition that needs to be met to proceed with this hypothesis test?

a. Students in the sample should be independent of each other with respect to how many colleges they applied to.
b. Sampling should have been done randomly.
c. The sample size should be less than 10% of the population of all students at that university.
d. There should be at least 10 successes and 10 failures in the sample.
e. The distribution of the number of colleges students apply to should not be extremely skewed.

Number of college applications - conditions

Which of the following is <u>not</u> a condition that needs to be met to proceed with this hypothesis test?

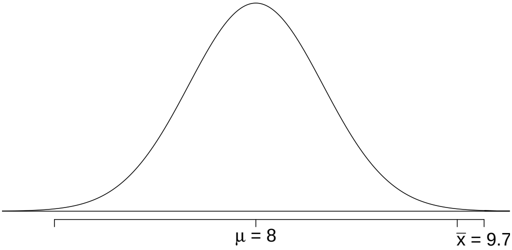
a. Students in the sample should be independent of each other with respect to how many colleges they applied to.
b. Sampling should have been done randomly.
c. The sample size should be less than 10% of the population of all students at that university.
d. *There should be at least 10 successes and 10 failures in the sample.*
e. The distribution of the number of colleges students apply to should not be extremely skewed.

Test statistic

In order to evaluate if the observed sample mean is unusual for the hypothesized sampling distribution, we determine how many standard errors away from the null it is, which is also called the TEST STATISTIC.

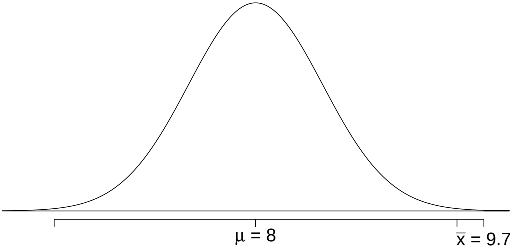
Test statistic

In order to evaluate if the observed sample mean is unusual for the hypothesized sampling distribution, we determine how many standard errors away from the null it is, which is also called the TEST STATISTIC.



Test statistic

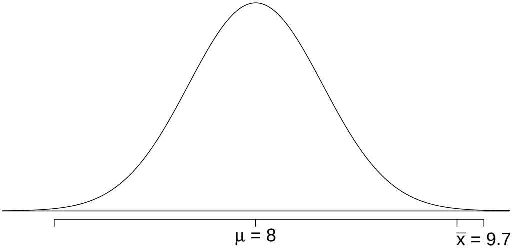
In order to evaluate if the observed sample mean is unusual for the hypothesized sampling distribution, we determine how many standard errors away from the null it is, which is also called the TEST STATISTIC.



$$\bar{x} \sim N(\mu = 8, SE = \frac{7}{\sqrt{206}} = 0.5)$$

Test statistic

In order to evaluate if the observed sample mean is unusual for the hypothesized sampling distribution, we determine how many standard errors away from the null it is, which is also called the TEST STATISTIC.

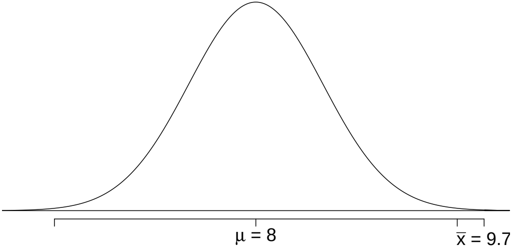


$$\bar{x} \sim N\left(\mu = 8, SE = \frac{7}{\sqrt{206}} = 0.5\right)$$

$$Z = \frac{9.7 - 8}{0.5} = 3.4$$

Test statistic

In order to evaluate if the observed sample mean is unusual for the hypothesized sampling distribution, we determine how many standard errors away from the null it is, which is also called the TEST STATISTIC.



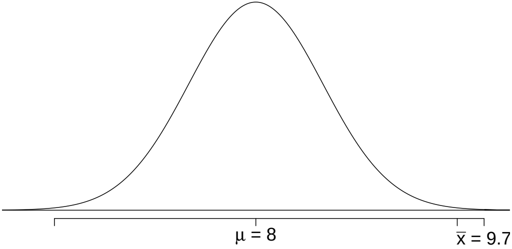
$$\bar{x} \sim N\left(\mu = 8, SE = \frac{7}{\sqrt{206}} = 0.5\right)$$

$$Z = \frac{9.7 - 8}{0.5} = 3.4$$

The sample mean is 3.4 standard errors away from the hypothesized value. Is this considered unusually high? That is, is the result STATISTICALLY SIGNIFICANT?

Test statistic

In order to evaluate if the observed sample mean is unusual for the hypothesized sampling distribution, we determine how many standard errors away from the null it is, which is also called the TEST STATISTIC.



$$\bar{x} \sim N\left(\mu = 8, SE = \frac{7}{\sqrt{206}} = 0.5\right)$$

$$Z = \frac{9.7 - 8}{0.5} = 3.4$$

The sample mean is 3.4 standard errors away from the hypothesized value. Is this considered unusually high? That is, is the result STATISTICALLY SIGNIFICANT?

Yes, and we can quantify how unusual it is using a p-value.

P-VALUES

- We then use this test statistic to calculate the P-VALUE, the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis were true.

P-VALUES

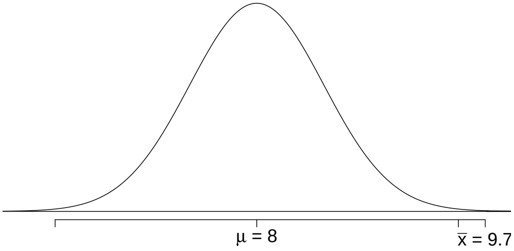
- We then use this test statistic to calculate the P-VALUE, the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis were true.
- If the p-value is LOW (lower than the significance level, α, which is usually 5%) we say that it would be very unlikely to observe the data if the null hypothesis were true, and hence REJECT H_0.

P-VALUES

- We then use this test statistic to calculate the P-VALUE, the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis were true.
- If the p-value is LOW (lower than the significance level, α, which is usually 5%) we say that it would be very unlikely to observe the data if the null hypothesis were true, and hence REJECT H_0.
- If the p-value is HIGH (higher than α) we say that it is likely to observe the data even if the null hypothesis were true, and hence DO NOT REJECT H_0.

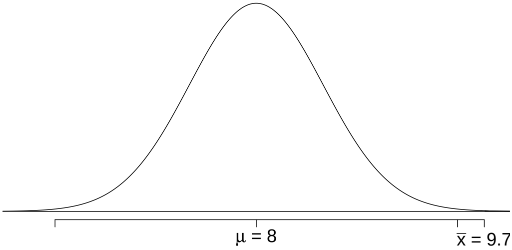
Number of college applications - p-value

p-value: probability of observing data at least as favorable to H_A as our current data set (a sample mean greater than 9.7), if in fact H_0 were true (the true population mean was 8).



Number of college applications - p-value

p-value: probability of observing data at least as favorable to H_A as our current data set (a sample mean greater than 9.7), if in fact H_0 were true (the true population mean was 8).



$$P(\bar{x} > 9.7 \mid \mu = 8) = P(Z > 3.4) = 0.0003$$

Number of college applications - Making a decision

- p-value = 0.0003

Number of college applications - Making a decision

- p-value $= 0.0003$
  - If the true average of the number of colleges students applied to is 8, there is only 0.03% chance of observing a random sample of 206 students who on average apply to 9.7 or more schools.

Number of college applications - Making a decision

- p-value = 0.0003
  - If the true average of the number of colleges students applied to is 8, there is only 0.03% chance of observing a random sample of 206 students who on average apply to 9.7 or more schools.
  - This is a pretty low probability for us to think that a sample mean of 9.7 or more schools is likely to happen simply by chance.

Number of college applications - Making a decision

- p-value = 0.0003
  - If the true average of the number of colleges students applied to is 8, there is only 0.03% chance of observing a random sample of 206 students who on average apply to 9.7 or more schools.
  - This is a pretty low probability for us to think that a sample mean of 9.7 or more schools is likely to happen simply by chance.
- Since p-value is low (lower than 5%) we reject H_0.

Number of college applications - Making a decision

- p-value = 0.0003
  - If the true average of the number of colleges students applied to is 8, there is only 0.03% chance of observing a random sample of 206 students who on average apply to 9.7 or more schools.
  - This is a pretty low probability for us to think that a sample mean of 9.7 or more schools is likely to happen simply by chance.
- Since p-value is low (lower than 5%) we reject H_0.
- The data provide convincing evidence that students apply to more than 8 schools on average.

Number of college applications - Making a decision

- p-value $= 0.0003$
    - If the true average of the number of colleges students applied to is 8, there is only 0.03% chance of observing a random sample of 206 students who on average apply to 9.7 or more schools.
    - This is a pretty low probability for us to think that a sample mean of 9.7 or more schools is likely to happen simply by chance.
- Since p-value is low (lower than 5%) we reject H_0.
- The data provide convincing evidence that students apply to more than 8 schools on average.
- The difference between the null value of 8 schools and observed sample mean of 9.7 schools is not due to chance or sampling variability.

A poll by the National Sleep Foundation found that college students average about 7 hours of sleep per night. A sample of 169 college students taking an introductory statistics class yielded an average of 6.88 hours, with a standard deviation of 0.94 hours. Assuming that this is a random sample representative of all college students *(bit of a leap of faith?)*, a hypothesis test was conducted to evaluate if college students on average sleep less than 7 hours per night. The p-value for this hypothesis test is 0.0485. Which of the following is correct?

a. Fail to reject H_0, the data provide convincing evidence that college students sleep less than 7 hours on average.
b. Reject H_0, the data provide convincing evidence that college students sleep less than 7 hours on average.
c. Reject H_0, the data prove that college students sleep more than 7 hours on average.
d. Fail to reject H_0, the data do not provide convincing evidence that college students sleep less than 7 hours on average.
e. Reject H_0, the data provide convincing evidence that college students in this sample sleep less than 7 hours on average.

A poll by the National Sleep Foundation found that college students average about 7 hours of sleep per night. A sample of 169 college students taking an introductory statistics class yielded an average of 6.88 hours, with a standard deviation of 0.94 hours. Assuming that this is a random sample representative of all college students *(bit of a leap of faith?)*, a hypothesis test was conducted to evaluate if college students on average sleep less than 7 hours per night. The p-value for this hypothesis test is 0.0485. Which of the following is correct?

a. Fail to reject H_0, the data provide convincing evidence that college students sleep less than 7 hours on average.
b. *Reject H_0, the data provide convincing evidence that college students sleep less than 7 hours on average.*
c. Reject H_0, the data prove that college students sleep more than 7 hours on average.
d. Fail to reject H_0, the data do not provide convincing evidence that college students sleep less than 7 hours on average.
e. Reject H_0, the data provide convincing evidence that college students in this sample sleep less than 7 hours on average.

Two-sided hypothesis testing with p-values

- If the research question was "Do the data provide convincing evidence that the average amount of sleep college students get per night is different than the national average?", the alternative hypothesis would be formulated as follows:

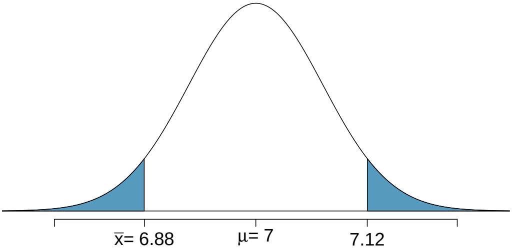
$$H_0 : \mu = 7$$
$$H_A : \mu \neq 7$$

Two-sided hypothesis testing with p-values

- If the research question was "Do the data provide convincing evidence that the average amount of sleep college students get per night is different than the national average?", the alternative hypothesis would be formulated as follows:

$$H_0 : \mu = 7$$
$$H_A : \mu \neq 7$$

- Hence the p-value would change as well:



p-value
$= 0.0485 \times 2$
$= 0.097$

In other words, we need to "split" $\alpha = 0.05$ over two sides (*i.e.*, $\alpha/2$)

Hypothesis Testing: Stating the Hypothesis

- A school publicizes that the proportion of alumni getting a job in the first three months after graduation is 85%.

Hypothesis Testing: Stating the Hypothesis

- A school publicizes that the proportion of alumni getting a job in the first three months after graduation is 85%.

$$\begin{cases} H_0: & p = 0.85 \quad \text{(claim)} \\ H_A: & p \neq 0.85 \end{cases}$$

Hypothesis Testing: Stating the Hypothesis

- A school publicizes that the proportion of alumni getting a job in the first three months after graduation is 85%.

$$\begin{cases} H_0: & p = 0.85 \quad \text{(claim)} \\ H_A: & p \neq 0.85 \end{cases}$$

- A company advertises that the mean life of its batteries is more than 6 months

Hypothesis Testing: Stating the Hypothesis

- A school publicizes that the proportion of alumni getting a job in the first three months after graduation is 85%.

$$\begin{cases} H_0: & p = 0.85 \quad \text{(claim)} \\ H_A: & p \neq 0.85 \end{cases}$$

- A company advertises that the mean life of its batteries is more than 6 months

$$\begin{cases} H_0: & \mu \leq 6 \\ H_A: & \mu > 6 \quad \text{(claim)} \end{cases}$$

Hypothesis Testing: Stating the Hypothesis

- A school publicizes that the proportion of alumni getting a job in the first three months after graduation is 85%.

$$\begin{cases} H_0: & p = 0.85 \quad \text{(claim)} \\ H_A: & p \neq 0.85 \end{cases}$$

- A company advertises that the mean life of its batteries is more than 6 months

$$\begin{cases} H_0: & \mu \leq 6 \\ H_A: & \mu > 6 \quad \text{(claim)} \end{cases}$$

- A car dealership announces that the mean time for an oil change is less than 15 minutes.

Hypothesis Testing: Stating the Hypothesis

- A school publicizes that the proportion of alumni getting a job in the first three months after graduation is 85%.

$$\begin{cases} H_0: & p = 0.85 \quad \text{(claim)} \\ H_A: & p \neq 0.85 \end{cases}$$

- A company advertises that the mean life of its batteries is more than 6 months

$$\begin{cases} H_0: & \mu \leq 6 \\ H_A: & \mu > 6 \quad \text{(claim)} \end{cases}$$

- A car dealership announces that the mean time for an oil change is less than 15 minutes.

$$\begin{cases} H_0: & \mu \geq 15 \\ H_A: & \mu < 15 \quad \text{(claim)} \end{cases}$$

Recap: Hypothesis testing framework

1. Set the hypotheses.
2. Check assumptions and conditions.
3. Calculate a test statistic and a p-value.
4. Make a decision, and interpret it in context of the research question.

Recap: Hypothesis testing for a population mean

1. Set the hypotheses
   - $H_0 : \mu = null\ value$
   - $H_A : \mu <$ or $>$ or $\neq null\ value$
2. Calculate the point estimate
3. Check assumptions and conditions
   - Independence: random sample/assignment, 10% condition when sampling without replacement
   - Normality: nearly normal population or $n \geq 30$, no extreme skew – or use the t distribution
4. Calculate a TEST STATISTIC and a p-value (draw a picture!)

$$z = \frac{\bar{x} - \mu}{SE}, \text{ where } SE = \frac{s}{\sqrt{n}}$$

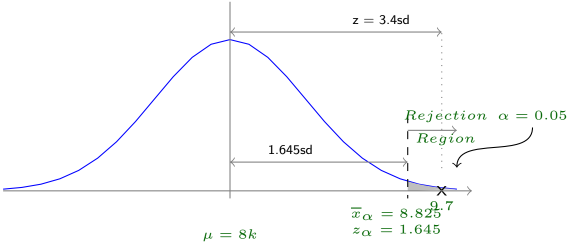
Approach based on the P-VALUE

- If p-value $< \alpha$, reject H_0, data provide evidence for H_A
- If p-value $> \alpha$, do not reject H_0, data do not provide evidence for H_A

Approach based on the REJECTION REGION

- If the z statistic fall within the rejection region, reject H_0, data provide evidence for H_A
- If the z statistic fall outside of the rejection region, do not reject H_0, data do not provide evidence for H_A

Hypothesis Testing: Rejection Region Approach

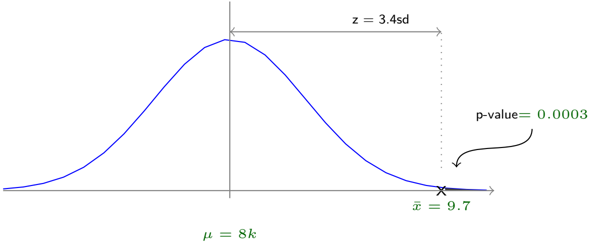


i. Assume H_0 is true
ii. Find threshold value for which the prob of falling above that value is α, *e.g.*, $\alpha = 0.05$ (`NORMSINV(0.95)` $= 1.645 \Rightarrow \bar{x}_\alpha = 8 + 1.645 \times 7/\sqrt{206} = 8.825$)
iii. We compute a test statistics from the sample: $z = \frac{\overline{x}-\mu}{SE}$
iv. Now we compare the two quantities:

$$z = \frac{\overline{x} - \mu}{SE} = 3.4 \quad \text{and} \quad z_\alpha = 1.645$$

- If $\bar{x}$ falls inside the rejection region (*i.e.*, the z score is beyond z_α), we *reject* H_0 (and thus we accept H_a)
- If $\bar{x}$ does not fall inside the rejection region, we *fail to reject* H_0 (does not mean that we accept H_0 as true)

Hypothesis Testing: p-value Approach



i. Assume H_0 is true

ii. We compute a test statistics from the sample: $z = \frac{\overline{x}-\mu}{SE} = \frac{9.7-8}{0.5} = 3.4$

iii. Now obtain the p-value:

$$p(\bar{x} > 3.4) = 0.0003$$

- If p-value is below α, we *reject* H_0 (and thus we accept H_a)
- If p-value is above α, we *fail to reject* H_0 (does not mean that we accept H_0 as true)

Hypothesis Testing for Small Samples

If the sample size is small (*i.e.*, below 30):

1. CLT is no longer valid $\Rightarrow$ We require *normality* of the underlying population
2. St.dev. σ can no longer be approximated using s within a z statistic $\Rightarrow$ We need to use a t-statistic with $n-1$ df:

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}}$$

Example

A car manufacturer wants to test emission level. The mean emission level μ must be less than 20 ppm of carbon. Ten engines are manufactured for testing. Can we conclude that this type of engine meets the pollution standards? Use $\alpha = 0.05$.

15.6	16.2	22.5	20.5	16.4	19.4	19.6	17.9	12.7	14.9

Hypothesis Testing for Population Proportion

A1. A random sample is selected from a binomial experiment

A2. The sample size n is large, *i.e.*, both $np_0 \geq 15$ and $nq_0 \geq 15$ hold

We thus assume CLT holds and we use the normal distribution as a reasonable approximation for the sampling distribution of $\hat{p}$:

$$z = \frac{\hat{p} - p_0}{\sigma_{\hat{p}}},$$

where $\sigma_{\hat{p}} = \sqrt{p_0 q_0 / n}$.

Pepsi Challenge

Coca-Cola drinkers participated in a blind taste test where they were asked to taste unmarked cups of Pepsi and Coke and select their favorite.
Pepsi claim: "More than half the Diet Coke drinkers surveyed said they preferred the taste of the Diet Pepsi."

- $n = 100$ Diet Coke drinkers
- $x = 56$ preferred taste of Diet Pepsi

What can we conclude based on the test?

Hypothesis Testing for Population Variance

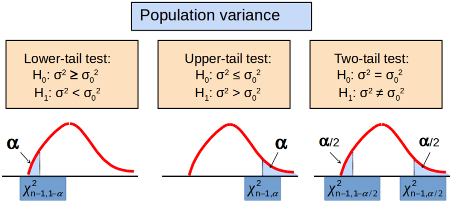
- If the population is normally distributed, then

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

follows a χ^2 distribution with $n-1$ degrees of freedom.

- The test statistic for hypothesis test about one population variance is:

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$



Decision errors

- Hypothesis tests are not flawless.
- In the court system innocent people are sometimes wrongly convicted and the guilty sometimes walk free.
- Similarly, we can make a wrong decision in statistical hypothesis tests as well.
- The difference is that we have the tools necessary to quantify how often we make errors in statistics.

Decision errors (cont.)

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

Decision errors (cont.)

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true		
	H_A true		

Decision errors (cont.)

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	
	H_A true		

Decision errors (cont.)

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	
	H_A true		✓

Decision errors (cont.)

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	Type 1 Error
	H_A true		✓

- A Type 1 Error is rejecting the null hypothesis when H_0 is true.

Decision errors (cont.)

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	Type 1 Error
	H_A true	Type 2 Error	✓

- A Type 1 Error is rejecting the null hypothesis when H_0 is true.
- A Type 2 Error is failing to reject the null hypothesis when H_A is true.

Decision errors (cont.)

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	Type 1 Error
	H_A true	Type 2 Error	✓

- A Type 1 Error is rejecting the null hypothesis when H_0 is true.
- A Type 2 Error is failing to reject the null hypothesis when H_A is true.
- We (almost) never know if H_0 or H_A is true, but we need to consider all possibilities.

Hypothesis Test as a trial

If we again think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

$$H_0: \text{ Defendant is not guilty}$$
$$H_A: \text{ Defendant is guilty}$$

Which type of error is being committed in the following cirumstances?

- Declaring the defendant not guilty when he is actually guilty
- Declaring the defendant guilty when he is actually innocent

Hypothesis Test as a trial

If we again think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

$$H_0: \text{ Defendant is not guilty}$$
$$H_A: \text{ Defendant is guilty}$$

Which type of error is being committed in the following cirumstances?

- Declaring the defendant not guilty when he is actually guilty

  Type 2 error

- Declaring the defendant guilty when he is actually innocent

Hypothesis Test as a trial

If we again think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

$$H_0: \text{ Defendant is not guilty}$$
$$H_A: \text{ Defendant is guilty}$$

Which type of error is being committed in the following cirumstances?

- Declaring the defendant not guilty when he is actually guilty

  Type 2 error

- Declaring the defendant guilty when he is actually innocent

  Type 1 error

Hypothesis Test as a trial

If we again think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

$$H_0: \text{ Defendant is not guilty}$$
$$H_A: \text{ Defendant is guilty}$$

Which type of error is being committed in the following cirumstances?

- Declaring the defendant not guilty when he is actually guilty

  Type 2 error

- Declaring the defendant guilty when he is actually innocent

  Type 1 error

Which error do you think is the worse error to make?

Hypothesis Test as a trial

If we again think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

$$H_0: \text{ Defendant is not guilty}$$
$$H_A: \text{ Defendant is guilty}$$

Which type of error is being committed in the following cirumstances?

- Declaring the defendant not guilty when he is actually guilty

  Type 2 error

- Declaring the defendant guilty when he is actually innocent

  Type 1 error

Which error do you think is the worse error to make?

"better that ten guilty persons escape than that one innocent suffer"

– William Blackstone

Type 1 error rate

- As a general rule we reject H_0 when the p-value is less than 0.05, i.e. we use a SIGNIFICANCE LEVEL of 0.05, $\alpha = 0.05$.

Type 1 error rate

- As a general rule we reject H_0 when the p-value is less than 0.05, i.e. we use a significance level of 0.05, $\alpha = 0.05$.
- This means that, for those cases where H_0 is actually true, we do not want to incorrectly reject it more than 5% of those times.

Type 1 error rate

- As a general rule we reject H_0 when the p-value is less than 0.05, i.e. we use a SIGNIFICANCE LEVEL of 0.05, $\alpha = 0.05$.
- This means that, for those cases where H_0 is actually true, we do not want to incorrectly reject it more than 5% of those times.
- In other words, when using a 5% significance level there is about 5% chance of making a Type 1 error if the null hypothesis is true.

$$P(\text{Type 1 error}) = \alpha$$

Type 1 error rate

- As a general rule we reject H_0 when the p-value is less than 0.05, i.e. we use a SIGNIFICANCE LEVEL of 0.05, $\alpha = 0.05$.
- This means that, for those cases where H_0 is actually true, we do not want to incorrectly reject it more than 5% of those times.
- In other words, when using a 5% significance level there is about 5% chance of making a Type 1 error if the null hypothesis is true.

$$P(\text{Type 1 error}) = \alpha$$

- This is why we prefer small values of α – increasing α increases the Type 1 error rate.

Choosing a significance level

- Choosing a significance level for a test is important in many contexts, and the traditional level is 0.05. However, it is often helpful to adjust the significance level based on the application.
- We may select a level that is smaller or larger than 0.05 depending on the consequences of any conclusions reached from the test.
- If making a Type 1 Error is dangerous or especially costly, we should choose a small significance level (e.g. 0.01). Under this scenario we want to be very cautious about rejecting the null hypothesis, so we demand very strong evidence favoring H_A before we would reject H_0.
- If a Type 2 Error is relatively more dangerous or much more costly than a Type 1 Error, then we should choose a higher significance level (e.g. 0.10). Here we want to be cautious about failing to reject H_0 when the null is actually false.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true		
	H_A true		

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true		Type 1 Error, α
	H_A true		

- Type 1 error is rejecting H_0 when you shouldn't have, and the probability of doing so is α (significance level)

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true		Type 1 Error, α
	H_A true	Type 2 Error, β	

- Type 1 error is rejecting H_0 when you shouldn't have, and the probability of doing so is α (significance level)
- Type 2 error is failing to reject H_0 when you should have, and the probability of doing so is β (a little more complicated to calculate)

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	$1 - \alpha$	Type 1 Error, α
	H_A true	Type 2 Error, β	

- Type 1 error is rejecting H_0 when you shouldn't have, and the probability of doing so is α (significance level)
- Type 2 error is failing to reject H_0 when you should have, and the probability of doing so is β (a little more complicated to calculate)
- POWER of a test is the probability of correctly rejecting H_0, and the probability of doing so is $1 - \beta$

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	$1 - \alpha$	Type 1 Error, α
	H_A true	Type 2 Error, β	Power, $1 - \beta$

- Type 1 error is rejecting H_0 when you shouldn't have, and the probability of doing so is α (significance level)
- Type 2 error is failing to reject H_0 when you should have, and the probability of doing so is β (a little more complicated to calculate)
- Power of a test is the probability of correctly rejecting H_0, and the probability of doing so is $1 - \beta$
- In hypothesis testing, we want to keep α and β low, but there are inherent trade-offs.

Type 2 error rate

If the alternative hypothesis is actually true, what is the chance that we make a Type 2 Error, i.e. we fail to reject the null hypothesis even when we should reject it?

- The answer is not obvious.
- If the true population average is very close to the null hypothesis value, it will be difficult to detect a difference (and reject H_0).
- If the true population average is very different from the null hypothesis value, it will be easier to detect a difference.
- Clearly, β depends on the effect size (δ)

Example - Blood Pressure

Blood pressure oscillates with the beating of the heart, and the systolic pressure is defined as the peak pressure when a person is at rest. The average systolic blood pressure for people in the U.S. is about 130 mmHg with a standard deviation of about 25 mmHg.

We are interested in finding out if the average blood pressure of employees at a certain company is greater than the national average, so we collect a random sample of 100 employees and measure their systolic blood pressure. What are the hypotheses?

Example - Blood Pressure

Blood pressure oscillates with the beating of the heart, and the systolic pressure is defined as the peak pressure when a person is at rest. The average systolic blood pressure for people in the U.S. is about 130 mmHg with a standard deviation of about 25 mmHg.

We are interested in finding out if the average blood pressure of employees at a certain company is greater than the national average, so we collect a random sample of 100 employees and measure their systolic blood pressure. What are the hypotheses?

$$H_0 : \mu = 130$$
$$H_A : \mu > 130$$

Example - Blood Pressure

Blood pressure oscillates with the beating of the heart, and the systolic pressure is defined as the peak pressure when a person is at rest. The average systolic blood pressure for people in the U.S. is about 130 mmHg with a standard deviation of about 25 mmHg.

We are interested in finding out if the average blood pressure of employees at a certain company is greater than the national average, so we collect a random sample of 100 employees and measure their systolic blood pressure. What are the hypotheses?

$$H_0 : \mu = 130$$
$$H_A : \mu > 130$$

We'll start with a very specific question – "What is the power of this hypothesis test to correctly detect an increase of 2 mmHg in average blood pressure?"

Calculating power

The preceding question can be rephrased as "How likely is it that this test will reject H_0 when the true average systolic blood pressure for employees at this company is 132 mmHg?"

Calculating power

The preceding question can be rephrased as "How likely is it that this test will reject H_0 when the true average systolic blood pressure for employees at this company is 132 mmHg?"

Hint: Break this down intro two simpler problems

Calculating power

The preceding question can be rephrased as "How likely is it that this test will reject H_0 when the true average systolic blood pressure for employees at this company is 132 mmHg?"

Hint: Break this down intro two simpler problems

1. Problem 1: Which values of $\bar{x}$ represent sufficient evidence to reject H_0?

Calculating power

The preceding question can be rephrased as "How likely is it that this test will reject H_0 when the true average systolic blood pressure for employees at this company is 132 mmHg?"

Hint: Break this down intro two simpler problems

1. Problem 1: Which values of $\bar{x}$ represent sufficient evidence to reject H_0?
2. Problem 2: What is the probability that we would reject H_0 if $\bar{x}$ had come from $N\left(mean = 132, SE = \frac{25}{\sqrt{100}} = 2.5\right)$, i.e. what is the probability that we can obtain such an $\bar{x}$ from this distribution?

Calculating power

The preceding question can be rephrased as "How likely is it that this test will reject H_0 when the true average systolic blood pressure for employees at this company is 132 mmHg?"

Hint: Break this down intro two simpler problems

1. Problem 1: Which values of $\bar{x}$ represent sufficient evidence to reject H_0?
2. Problem 2: What is the probability that we would reject H_0 if $\bar{x}$ had come from $N\left(mean = 132, SE = \frac{25}{\sqrt{100}} = 2.5\right)$, i.e. what is the probability that we can obtain such an $\bar{x}$ from this distribution?

Determine how power changes as sample size, standard deviation of the sample, α, and effect size increases.

Problem 1

Which values of $\bar{x}$ represent sufficient evidence to reject H_0?
(Remember $H_0 : \mu = 130$, $H_A : \mu > 130$)

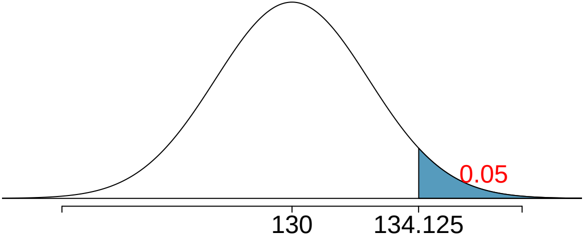
Problem 1

Which values of $\bar{x}$ represent sufficient evidence to reject H_0?
(Remember $H_0 : \mu = 130$, $H_A : \mu > 130$)

$$P(Z > z) < 0.05 \quad \Rightarrow \quad z > 1.65$$

$$\frac{\bar{x} - \mu}{s/\sqrt{n}} > 1.65$$

$$\bar{x} > 130 + 1.65 \times 2.5$$
$$\bar{x} > 134.125$$



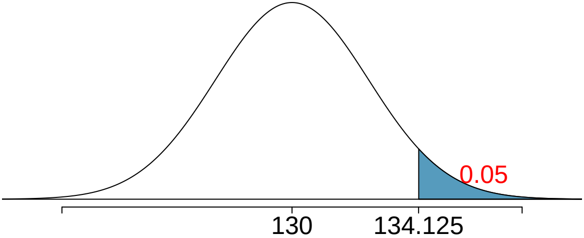
Problem 1

Which values of $\bar{x}$ represent sufficient evidence to reject H_0?
(Remember $H_0 : \mu = 130$, $H_A : \mu > 130$)

$$P(Z > z) < 0.05 \quad \Rightarrow \quad z > 1.65$$

$$\frac{\bar{x} - \mu}{s/\sqrt{n}} > 1.65$$

$$\bar{x} > 130 + 1.65 \times 2.5$$
$$\bar{x} > 134.125$$



Any $\bar{x} > 134.125$ would be sufficient to reject H_0 at the 5% significance level.

Problem 2

What is the probability that we would reject H_0 if $\bar{x}$ did come from $N(mean = 132, SE = 2.5)$.

Problem 2

What is the probability that we would reject H_0 if $\bar{x}$ did come from $N(mean = 132, SE = 2.5)$.

This is the same as finding the area above $\bar{x} = 134.125$ *if* $\bar{x}$ *came from* $N(132, 2.5)$.

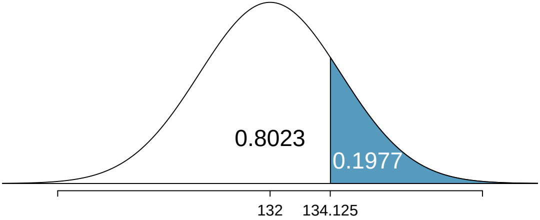
Problem 2

What is the probability that we would reject H_0 if $\bar{x}$ did come from $N(mean = 132, SE = 2.5)$.

This is the same as finding the area above $\bar{x} = 134.125$ *if* $\bar{x}$ *came from* $N(132, 2.5)$.

$$Z = \frac{134.125 - 132}{2.5} = 0.85$$

$$P(Z > 0.85) = 1 - 0.8023 = 0.1977$$



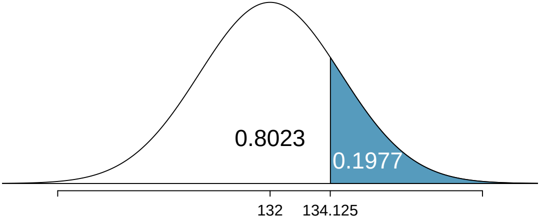
Problem 2

What is the probability that we would reject H_0 if $\bar{x}$ did come from $N(mean = 132, SE = 2.5)$.

This is the same as finding the area above $\bar{x} = 134.125$ if $\bar{x}$ came from $N(132, 2.5)$.

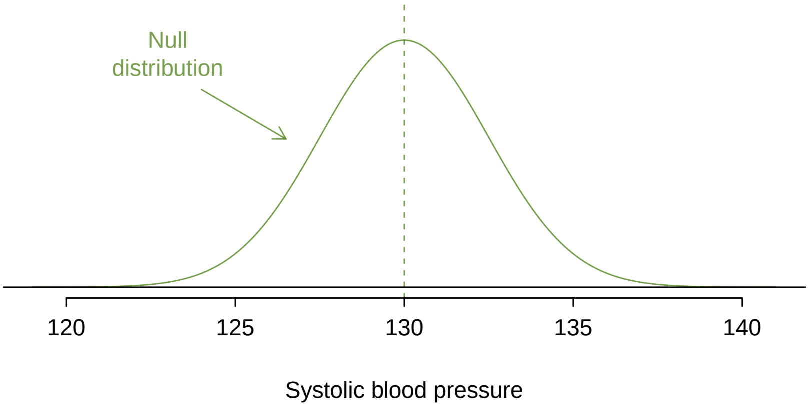
$$
\begin{aligned}
Z &= \frac{134.125 - 132}{2.5} \\
&= 0.85
\end{aligned}
$$

$$
\begin{aligned}
P(Z > 0.85) &= 1 - 0.8023 \\
&= 0.1977
\end{aligned}
$$

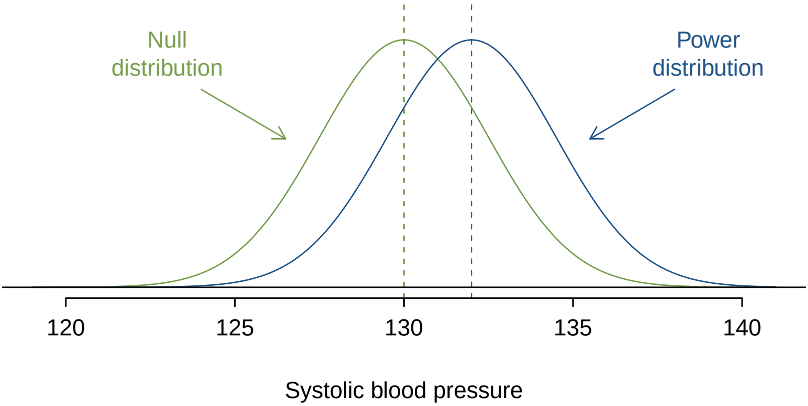


- *The probability of rejecting $H_0 : \mu = 130$, if the true average systolic blood pressure of employees at this company is 132 mmHg, is 0.1977 which is the power of this test ($1 - \beta$).*
- *Therefore, $\beta = 0.8023$ for this test.*

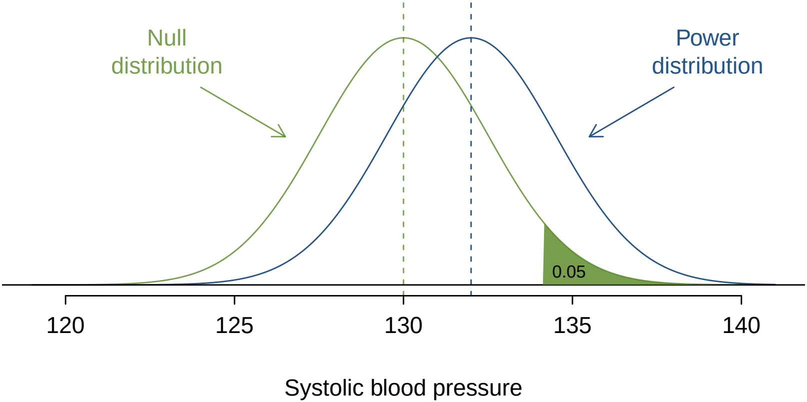
Putting it all together



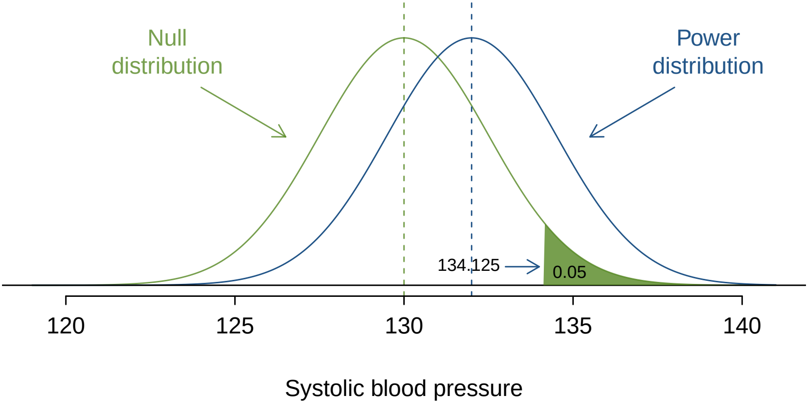
Putting it all together



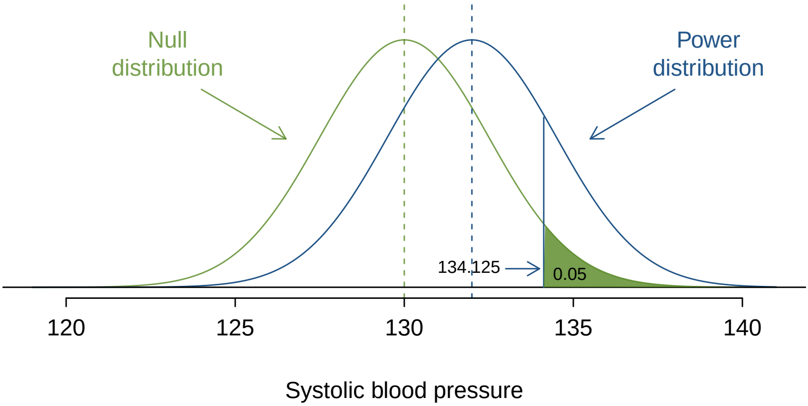
Putting it all together



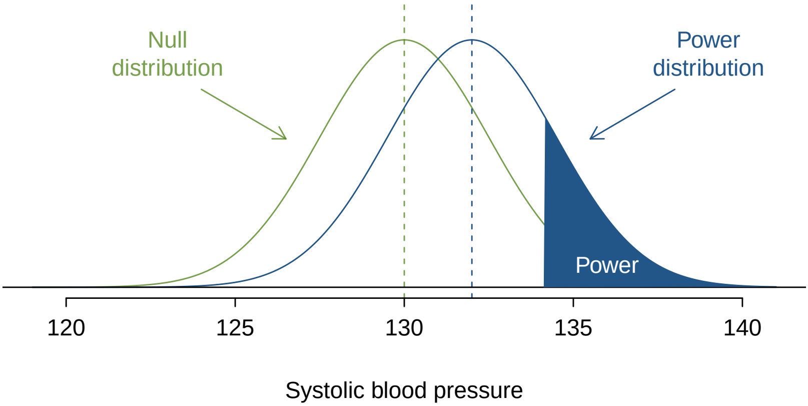
Putting it all together



Putting it all together



Putting it all together



Achieving desired power

There are several ways to increase power (and hence decrease type 2 error rate):

Achieving desired power

There are several ways to increase power (and hence decrease type 2 error rate):

1. Increase the sample size.

Achieving desired power

There are several ways to increase power (and hence decrease type 2 error rate):

1. Increase the sample size.
2. Decrease the standard deviation of the sample, which essentially has the same effect as increasing the sample size (it will decrease the standard error). With a smaller s we have a better chance of distinguishing the null value from the observed point estimate. This is difficult to ensure but cautious measurement process and limiting the population so that it is more homogenous may help.

Achieving desired power

There are several ways to increase power (and hence decrease type 2 error rate):

1. Increase the sample size.
2. Decrease the standard deviation of the sample, which essentially has the same effect as increasing the sample size (it will decrease the standard error). With a smaller s we have a better chance of distinguishing the null value from the observed point estimate. This is difficult to ensure but cautious measurement process and limiting the population so that it is more homogenous may help.
3. Increase α, which will make it more likely to reject H_0 (but note that this has the side effect of increasing the Type 1 error rate).

Achieving desired power

There are several ways to increase power (and hence decrease type 2 error rate):

1. Increase the sample size.
2. Decrease the standard deviation of the sample, which essentially has the same effect as increasing the sample size (it will decrease the standard error). With a smaller s we have a better chance of distinguishing the null value from the observed point estimate. This is difficult to ensure but cautious measurement process and limiting the population so that it is more homogenous may help.
3. Increase α, which will make it more likely to reject H_0 (but note that this has the side effect of increasing the Type 1 error rate).
4. Consider a larger effect size. If the true mean of the population is in the alternative hypothesis but close to the null value, it will be harder to detect a difference.

Recap - Calculating Power

- Begin by picking a meaningful effect size δ and a significance level α
- Calculate the range of values for the point estimate beyond which you would reject H_0 at the chosen α level.
- Calculate the probability of observing a value from preceding step if the sample was derived from a population where $\bar{x} \sim N(\mu_{H_0} + \delta, SE)$

Example - Using power to determine sample size

Going back to the blood pressure example, how large a sample would you need if you wanted 90% power to detect a 4 mmHg increase in average blood pressure for the hypothesis that the population average is greater than 130 mmHg at $\alpha = 0.05$?

Example - Using power to determine sample size

Going back to the blood pressure example, how large a sample would you need if you wanted 90% power to detect a 4 mmHg increase in average blood pressure for the hypothesis that the population average is greater than 130 mmHg at $\alpha = 0.05$?

$$Given: \quad H_0 : \mu = 130, \quad H_A : \mu > 130, \quad \alpha = 0.05, \quad \beta = 0.10, \quad \sigma = 25, \quad \delta = 4$$

Example - Using power to determine sample size

Going back to the blood pressure example, how large a sample would you need if you wanted 90% power to detect a 4 mmHg increase in average blood pressure for the hypothesis that the population average is greater than 130 mmHg at $\alpha = 0.05$?

$$Given: \quad H_0 : \mu = 130, \quad H_A : \mu > 130, \quad \alpha = 0.05, \quad \beta = 0.10, \quad \sigma = 25, \quad \delta = 4$$

Step 1: Determine the cutoff – in order to reject H_0 at $\alpha = 0.05$, we need a sample mean that will yield a Z score of at least 1.65.

$$\bar{x} > 130 + 1.65 \frac{25}{\sqrt{n}}$$

Example - Using power to determine sample size

Going back to the blood pressure example, how large a sample would you need if you wanted 90% power to detect a 4 mmHg increase in average blood pressure for the hypothesis that the population average is greater than 130 mmHg at $\alpha = 0.05$?

$$Given: \quad H_0 : \mu = 130, \quad H_A : \mu > 130, \quad \alpha = 0.05, \quad \beta = 0.10, \quad \sigma = 25, \quad \delta = 4$$

Step 1: Determine the cutoff – in order to reject H_0 at $\alpha = 0.05$, we need a sample mean that will yield a Z score of at least 1.65.

$$\bar{x} > 130 + 1.65 \frac{25}{\sqrt{n}}$$

Step 2: Set the probability of obtaining the above $\bar{x}$ if the true population is centered at 130 + 4 = 134 to the desired power, and solve for n.

$$p(\bar{x} > 130 + 1.65 \frac{25}{\sqrt{n}}) = 0.9$$

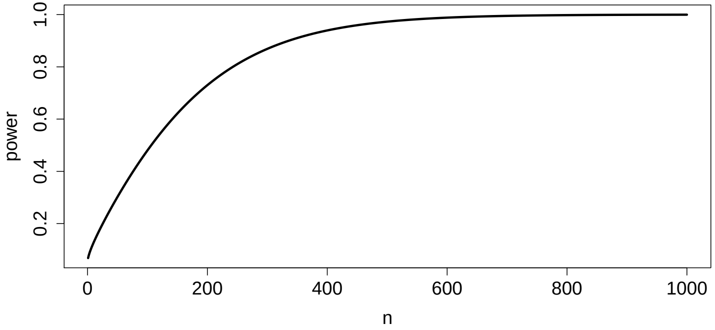
$$P\left(Z > \frac{\left(130 + 1.65 \frac{25}{\sqrt{n}}\right) - 134}{\frac{25}{\sqrt{n}}}\right) = P\left(Z > 1.65 - 4\frac{\sqrt{n}}{25}\right) = 0.9$$

Example - Using power to determine sample size (cont.)

You can either directly solve for n, or use computation to calculate power for various n and determine the sample size that yields the desired power:

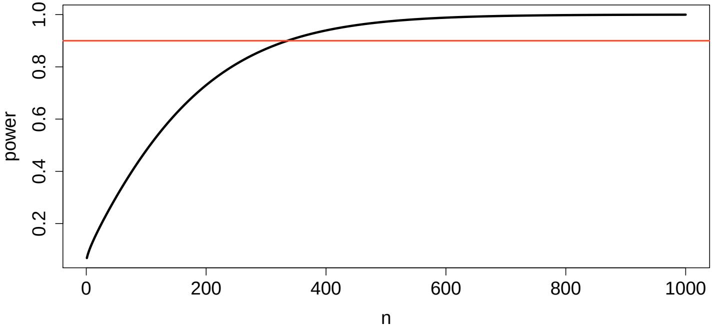
Example - Using power to determine sample size (cont.)

You can either directly solve for n, or use computation to calculate power for various n and determine the sample size that yields the desired power:



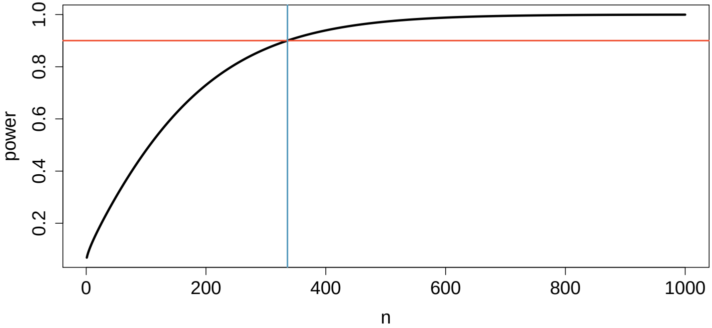
Example - Using power to determine sample size (cont.)

You can either directly solve for n, or use computation to calculate power for various n and determine the sample size that yields the desired power:



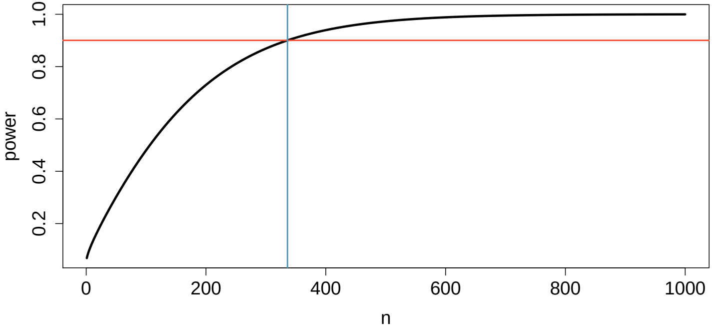
Example - Using power to determine sample size (cont.)

You can either directly solve for n, or use computation to calculate power for various n and determine the sample size that yields the desired power:



Example - Using power to determine sample size (cont.)

You can either directly solve for n, or use computation to calculate power for various n and determine the sample size that yields the desired power:



For $n = 336$, power $= 0.9002$, therefore we need 336 subjects in our sample to achieve the desired level of power for the given circumstance.

All else held equal, will the p-value be lower if $n = 100$ or $n = 10,000$?

a. $n = 100$
b. $n = 10,000$

All else held equal, will the p-value be lower if $n = 100$ or $n = 10,000$?

a. $n = 100$
b. $n = 10,000$

All else held equal, will the p-value be lower if $n = 100$ or $n = 10,000$?

(a) $n = 100$

(b) $n = 10,000$

Suppose $\bar{x} = 50$, $s = 2$, $H_0 : \mu = 49.5$, *and* $H_A : \mu \geq 49.5$.

All else held equal, will the p-value be lower if $n = 100$ or $n = 10,000$?

a. $n = 100$

b. $n = 10,000$

Suppose $\bar{x} = 50$, $s = 2$, $H_0 : \mu = 49.5$, *and* $H_A : \mu \geq 49.5$.

$$Z_{n=100} = \frac{50 - 49.5}{\frac{2}{\sqrt{100}}}$$

All else held equal, will the p-value be lower if $n = 100$ or $n = 10,000$?

(a) $n = 100$

(b) $n = 10,000$

Suppose $\bar{x} = 50$, $s = 2$, $H_0 : \mu = 49.5$, *and* $H_A : \mu \geq 49.5$.

$$Z_{n=100} \quad = \quad \frac{50 - 49.5}{\frac{2}{\sqrt{100}}} = \frac{50 - 49.5}{\frac{2}{10}} = \frac{0.5}{0.2} = 2.5, \quad \textit{p-value} = 0.0062$$

All else held equal, will the p-value be lower if $n = 100$ or $n = 10,000$?

(a) $n = 100$

(b) $n = 10,000$

Suppose $\bar{x} = 50$, $s = 2$, $H_0 : \mu = 49.5$, *and* $H_A : \mu \geq 49.5$.

$$Z_{n=100} = \frac{50 - 49.5}{\frac{2}{\sqrt{100}}} = \frac{50 - 49.5}{\frac{2}{10}} = \frac{0.5}{0.2} = 2.5, \quad \textit{p-value} = 0.0062$$

$$Z_{n=10000} = \frac{50 - 49.5}{\frac{2}{\sqrt{10000}}}$$

All else held equal, will the p-value be lower if $n = 100$ or $n = 10,000$?

(a) $n = 100$

(b) $n = 10,000$

Suppose $\bar{x} = 50$, $s = 2$, $H_0 : \mu = 49.5$, *and* $H_A : \mu \geq 49.5$.

$$
\begin{aligned}
Z_{n=100} &= \frac{50 - 49.5}{\frac{2}{\sqrt{100}}} = \frac{50 - 49.5}{\frac{2}{10}} = \frac{0.5}{0.2} = 2.5, \quad \textit{p-value} = 0.0062 \\
Z_{n=10000} &= \frac{50 - 49.5}{\frac{2}{\sqrt{10000}}} = \frac{50 - 49.5}{\frac{2}{100}} = \frac{0.5}{0.02} = 25, \quad \textit{p-value} \approx 0
\end{aligned}
$$

All else held equal, will the p-value be lower if $n = 100$ or $n = 10,000$?

(a) $n = 100$

(b) $n = 10,000$

Suppose $\bar{x} = 50$, $s = 2$, $H_0 : \mu = 49.5$, *and* $H_A : \mu \geq 49.5$.

$$\begin{aligned} Z_{n=100} &= \frac{50 - 49.5}{\frac{2}{\sqrt{100}}} = \frac{50 - 49.5}{\frac{2}{10}} = \frac{0.5}{0.2} = 2.5, \quad \textit{p-value} = 0.0062 \\ Z_{n=10000} &= \frac{50 - 49.5}{\frac{2}{\sqrt{10000}}} = \frac{50 - 49.5}{\frac{2}{100}} = \frac{0.5}{0.02} = 25, \quad \textit{p-value} \approx 0 \end{aligned}$$

As n *increases -* $SE \downarrow$, $Z \uparrow$, *p-value* $\downarrow$

Test the hypothesis $H_0 : \mu = 10$ vs. $H_A : \mu > 10$ for the following 8 samples. Assume $\sigma = 2$.

$\bar{x}$	10.05	10.1	10.2
$n = 30$	$p - value = 0.45$	$p - value = 0.39$	$p - value = 0.29$
$n = 5000$	$p - value = 0.39$	$p - value = 0.0002$	$p - value \approx 0$

Test the hypothesis $H_0 : \mu = 10$ vs. $H_A : \mu > 10$ for the following 8 samples. Assume $\sigma = 2$.

$\bar{x}$	10.05	10.1	10.2
$n = 30$	$p - value = 0.45$	$p - value = 0.39$	$p - value = 0.29$
$n = 5000$	$p - value = 0.39$	$p - value = 0.0002$	$p - value \approx 0$

Test the hypothesis $H_0 : \mu = 10$ vs. $H_A : \mu > 10$ for the following 8 samples. Assume $\sigma = 2$.

$\bar{x}$	10.05	10.1	10.2
$n = 30$	$p - value = 0.45$	$p - value = 0.39$	$p - value = 0.29$
$n = 5000$	$p - value = 0.04$	$p - value = 0.0002$	$p - value \approx 0$

Test the hypothesis $H_0 : \mu = 10$ vs. $H_A : \mu > 10$ for the following 8 samples. Assume $\sigma = 2$.

$\bar{x}$	10.05	10.1	10.2
$n = 30$	$p - value = 0.45$	$p - value = 0.39$	$p - value = 0.29$
$n = 5000$	$p - value = 0.04$	$p - value = 0.0002$	$p - value \approx 0$

Test the hypothesis $H_0 : \mu = 10$ vs. $H_A : \mu > 10$ for the following 8 samples. Assume $\sigma = 2$.

$\bar{x}$	10.05	10.1	10.2
$n = 30$	$p - value = 0.45$	$p - value = 0.39$	$p - value = 0.29$
$n = 5000$	$p - value = 0.04$	$p - value = 0.0002$	$p - value \approx 0$

Test the hypothesis $H_0 : \mu = 10$ vs. $H_A : \mu > 10$ for the following 8 samples. Assume $\sigma = 2$.

$\bar{x}$	10.05	10.1	10.2
$n = 30$	$p - value = 0.45$	$p - value = 0.39$	$p - value = 0.29$
$n = 5000$	$p - value = 0.04$	$p - value = 0.0002$	$p - value \approx 0$

Test the hypothesis $H_0 : \mu = 10$ vs. $H_A : \mu > 10$ for the following 8 samples. Assume $\sigma = 2$.

$\bar{x}$	10.05	10.1	10.2
$n = 30$	$p - value = 0.45$	$p - value = 0.39$	$p - value = 0.29$
$n = 5000$	$p - value = 0.04$	$p - value = 0.0002$	$p - value \approx 0$

Test the hypothesis $H_0 : \mu = 10$ vs. $H_A : \mu > 10$ for the following 8 samples. Assume $\sigma = 2$.

$\bar{x}$	10.05	10.1	10.2
$n = 30$	$p - value = 0.45$	$p - value = 0.39$	$p - value = 0.29$
$n = 5000$	$p - value = 0.04$	$p - value = 0.0002$	$p - value \approx 0$

- *When n is large, even small deviations from the null (small effect sizes), which may be considered practically insignificant, can yield statistically significant results.*

Test the hypothesis $H_0 : \mu = 10$ vs. $H_A : \mu > 10$ for the following 8 samples. Assume $\sigma = 2$.

$\bar{x}$	10.05	10.1	10.2
$n = 30$	$p - value = 0.45$	$p - value = 0.39$	$p - value = 0.29$
$n = 5000$	$p - value = 0.04$	$p - value = 0.0002$	$p - value \approx 0$

- *When n is large, even small deviations from the null (small effect sizes), which may be considered practically insignificant, can yield statistically significant results.*
- *Confidence intervals can give us a* better idea *of the effect size. E.g., we know that the average salary is $\mu > 100k$ but, how much higher?*

Statistical vs. practical significance

- Real differences between the point estimate and null value are easier to detect with larger samples.
- However, very large samples will result in statistical significance even for tiny differences between the sample mean and the null value (effect size), even when the difference is not practically significant.
- This is especially important to research: if we conduct a study, we want to focus on finding meaningful results (we want observed differences to be real, but also large enough to matter).
- The role of a statistician is not just in the analysis of data, but also in planning and design of a study.

"To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of." – R.A. Fisher