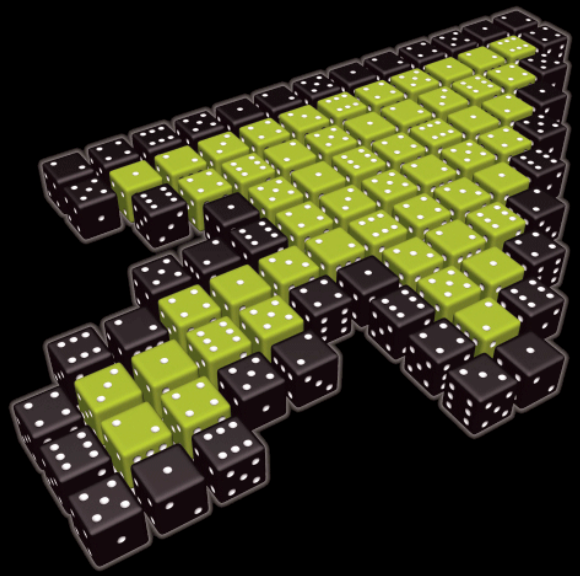
Chapter 8

Hypothesis testing



Testing as inference

- Along with estimation, **hypothesis testing** is one of the major fields of statistical inference
- In estimation, we:
  - don't know a population parameter
  - collect a sample and calculate a sample statistic
  - use that to provide a range of values for the parameter
- In testing, we start out with someone asserting a claim about the parameter
- We then collect a sample to test this claim

Claims that are tested

- In particular, we test a claim that the population parameter assumes some specific value

Examples

- The government's approval rating π is 52%
- The average life expectancy μ is 80 years

Null and alternative hypotheses

- The claim that is being tested is known as the **null hypothesis**, often denoted H_0

  Examples

  - H_0: $\pi = 52\%$
  - H_0: $\mu = 80$

- To the null hypothesis, there is always an **alternative hypothesis**, often denoted H_A

Different alternatives

- There are different kinds of alternative hypotheses

  Two-sided:

  - asserts that the value assigned in null hypothesis is wrong
  - will always be a simple “unequal” inequality
  - example: $H_A: \pi \neq 52\%$

  One-sided:

  - asserts the direction in which the null hypothesis is wrong
  - will either contain a “greater than” or “less than” sign
  - examples: $H_A: \pi > 52\%$ or $H_A: \pi < 52\%$

Reject or not reject

- The test is conducted by collecting a sample and comparing it to the claim made
  - Example: For claim $\pi = 52\%$, if a sample has a sample proportion $p = 37\%$, this suggests that the claim is wrong
- At the conclusion of a hypothesis test, you either:
  - have enough sample evidence to contradict the null hypothesis claim, so you reject it; or
  - do not have enough evidence to contradict it, so you do not reject it
- Note: the null hypothesis is never "proven"!

Conducting the test – assume H_0 is true

- Details in the methodology depend on the context
- But we always start by assuming the null is true!
- That is, we assume that the value assigned to the population parameter is the correct value

  Example

  H_0: $\pi = 52\%$

  H_A: $\pi \neq 52\%$
- We assume that the approval rating π is 52%

  (Note: we will return to this example throughout)

Where the assumption leads

- Why do we make this assumption?

- Not because we think it is true, but because we are trying to test the assumption

- In particular, we can:
  - see what the assumption implies for sampling distributions
  - then see how a sample stacks up against this

Example

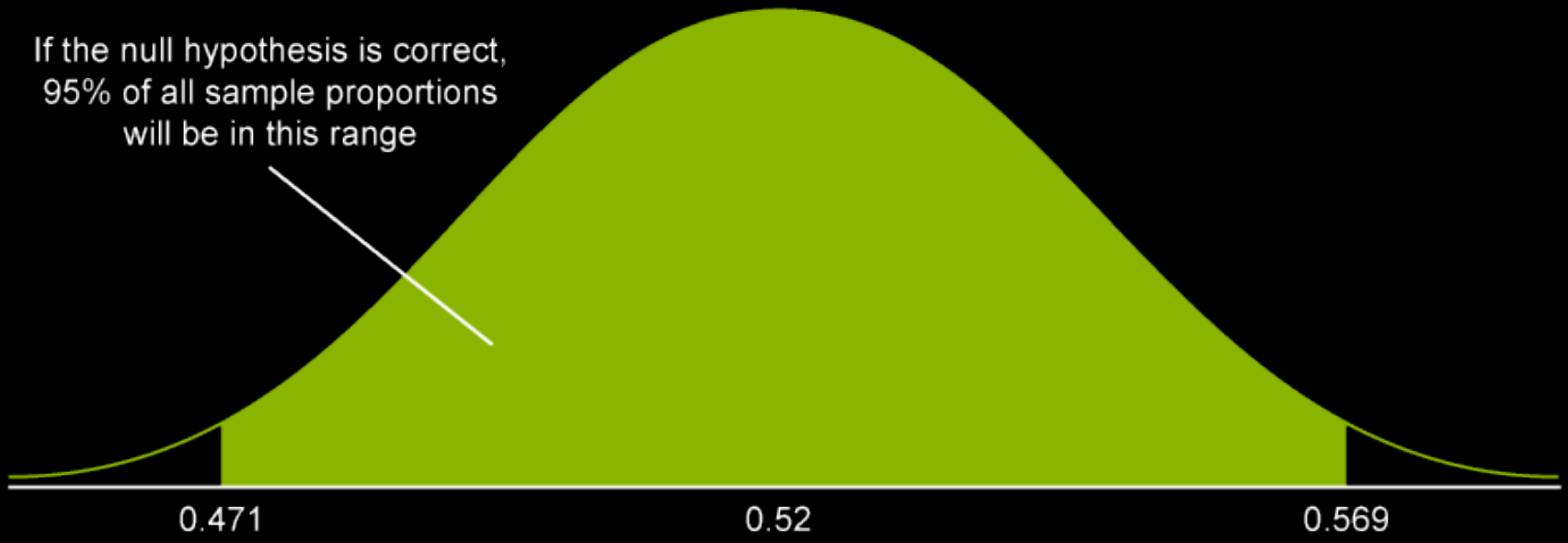
- Consider the government approval rating example
- Suppose we want a sample of $n = 400$ responses
- What does the assumption that π is 52% imply?
- Well, if that really is the population proportion, then:
  - the sampling distribution of the proportion is normal
  - it has mean $\pi = 0.52$
  - It has standard deviation $\sqrt{\frac{\pi(1-\pi)}{n}} = 0.025$

Example continued

- We can go further than this!
- By property of the normal distribution:
  - 95% of all values in sampling distribution lie within 1.96 standard deviations of the mean, so
  - 95% of sample proportions lie between 0.471 and 0.569
- So if we collect a sample and p is not in this range, we'll suspect the assumption is wrong!
- This provides us with a way of conducting, and concluding, our hypothesis test

Example continued

Sampling distribution of the proportion (if the null hypothesis is true)



- So once we have collected a sample, either p:
  - is outside 0.471 and 0.569, and we reject assumption
  - isn't outside the values, so we don't reject the assumption

Level of evidence

- Note: the boundary values depended on the fact that we used a '95% level' for the test
- This 'level' is really a measure of how much evidence we 'demand' before we reject H_0
- This will be made more precise soon, when we discuss the **level of significance** for the test
- To see how to conduct the test at any 'level', we must relate the sampling distribution to the standard normal distribution

The z-score of a sample statistic

- In the government survey of 400 people, say we got a sample proportion $p = 0.58$
- Under the assumption that H_0 is true (that $\pi = 0.52$), the z-score of this sample proportion is:

$$z = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{0.58 - 0.52}{\sqrt{\frac{0.52 \times 0.48}{400}}} = 2.4$$

- This is known as the **test statistic** for the test

Test statistic

- The test statistic (2.4) is the value of the sample proportion (0.58) in the standard normal distribution assuming that the null hypothesis is true ($\pi = 0.52$)
- It looks unlikely in that distribution – so we suspect the assumption was wrong!
- But can we be precise?

Level of significance

- We get more precise by defining a number at the beginning of every test, known as the level of significance α
- This is a number between 0 and 1 that determines how much ‘evidence’ you require before rejecting the null hypothesis
- The lower α is, the more evidence you require
- You actually choose the level of significance for your test at the beginning of the test

Level of significance (continued)

- It is analogous to the level of significance in a confidence interval estimate!
- In fact, α is related to the level of confidence, C%

$$C\% = 100 \times (1 - \alpha)\%$$

- Examples:
  - A '95% hypothesis test' means $\alpha = 0.05$
  - A '90% hypothesis test' means $\alpha = 0.1$
- This is what we were talking about before when we mentioned running a test at different 'levels'!

Level of significance (continued)

- Technically, α is defined to be the **probability of rejecting H_0 when it is true**
- But it also determines the **critical values** and **region of rejection** for your test, which determine the outcome of the test
- In particular, when you choose α this will determine z-scores in Z
- The conclusion of your test will depend on how the test statistic compares to these z-scores

Critical values and region of rejection

- There will be one or two **critical values** for the test
- They are z-scores in Z
- This critical value (or values) will define a **region of rejection**
- This will be an area of Z
- The conclusion of the test will depend on whether or not your test statistic is in the region of rejection
- Critical values and region of rejection will depend on whether the test is one-sided or two-sided

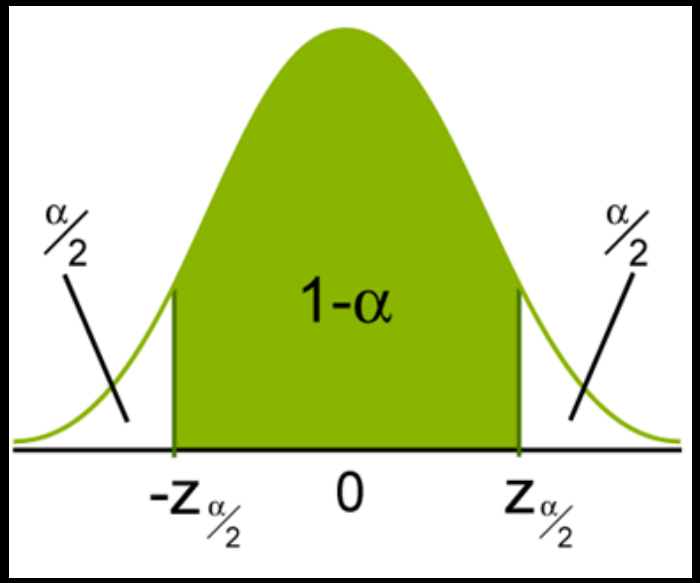
Two-sided tests: critical values

- The government approval survey was two-sided

  H_0: $\pi = 52\%$

  H_A: $\pi \neq 52\%$

- For a level of significance α, there are two **critical values** $z_{\alpha/2}$ and $-z_{\alpha/2}$
- As with a confidence interval, these are z-scores defined so that, as a proportion, α of Z lies outside the values



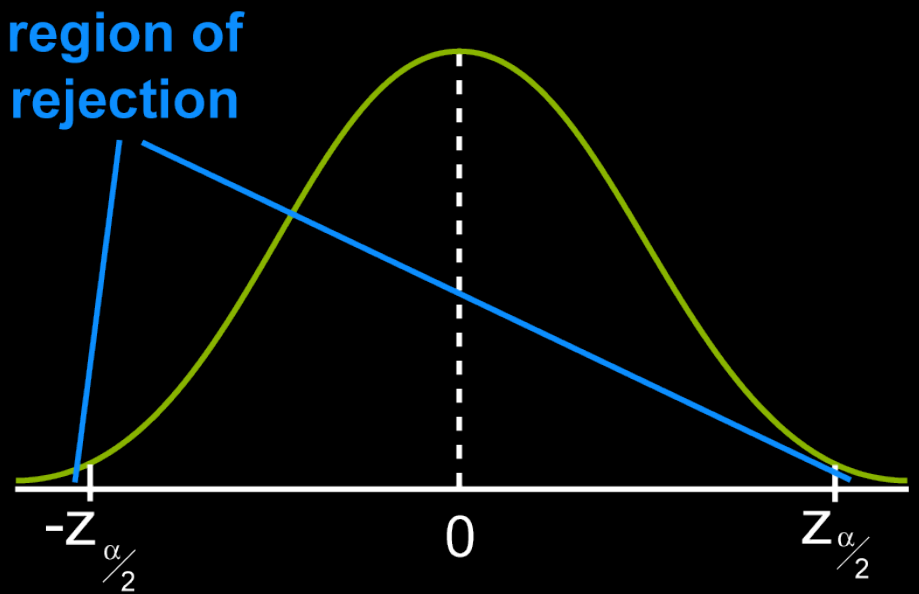
Two-sided tests: region of rejection

- For a two-sided test, the **region of rejection** is the area of Z outside of the critical values $z_{\alpha/2}$ and $-z_{\alpha/2}$
- Example:
- If $\alpha = 0.1$ is chosen,

  the critical values

  are $z_{0.05} = 1.645$

  and $-z_{0.05} = 1.645$



- The region of rejection is the set of values greater than 1.645 and values less than -1.645

Two-sided tests: conclusion

- Recall in the government approval survey, a sample proportion of $p = 58\%$ was calculated
- This sample proportion had a test statistic of $z = 2.4$
- If the test statistic is in the region of rejection, you reject H_0, if the test statistic isn't in the region of rejection, don't reject H_0
- The test statistic of 2.4 is greater than 1.645
- So it is in the region of rejection and H_0 is rejected
- That is, we conclude the approval rating isn't 52%

One-sided tests: critical values

- Now suppose the approval survey was one-sided

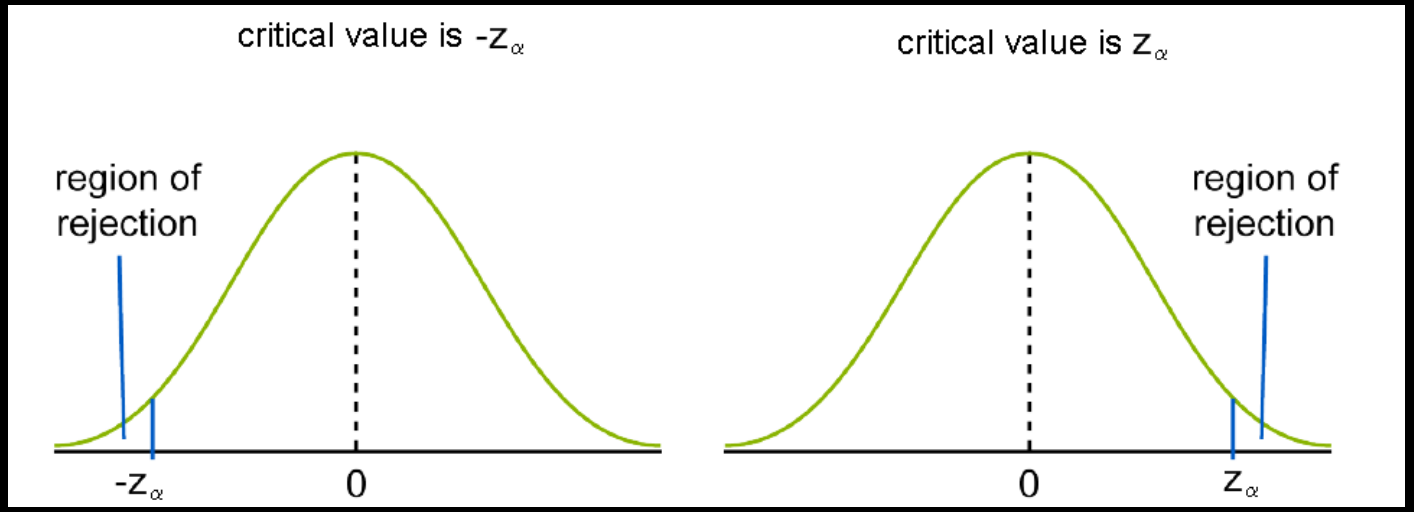
  H_0: $\pi = 52\%$

  H_A: $\pi > 52\%$

- For this test, there is only one critical value z_α
- Note that the critical value is z_α, not $z_{\alpha/2}$
- If the alternative hypothesis was the other way (that is, $\pi < 52\%$) then the critical value would be $-z_\alpha$
- Either way, the critical value is defined so that, as a proportion, α of Z lies to one side of the value

One-sided tests: region of rejection

- In a one-sided test, the region of rejection is the set of values to one side of the critical value, in the area that contains α of Z
  - if the critical value is z_α, it is the values greater than z_α
  - if the critical value is $-z_\alpha$, it is the values less than $-z_\alpha$



One-sided tests: conclusion

- In principle, the rule for concluding a one-sided test is the same as the rule for a two-sided test
- That is:
  - If the test statistic is in the region of rejection, reject H_0
  - If the test statistic isn't in the region of rejection, don't reject H_0
- So in either test, you compare the test statistic to the region of rejection to determine its conclusion

Hypothesis testing: step-by-step

- Step 1: State the hypotheses H_0 and H_A
  - this makes clear what claim is being tested
  - it also shows whether the test is one-sided or two-sided
- Step 2: Assume the null hypothesis H_0 is true
  - then the hypothesis test can test this assumption
- Step 3: Choose a level of significance α
  - this indicates the level of evidence you require
  - and it determines the critical values and region of rejection
  - some common levels are 0.1, 0.05 and 0.01

Hypothesis testing: step-by-step (cont'd)

- Step 4: Determine the critical value(s)
  - these set up the region of rejection
  - the number and nature of critical values depends on H_A
- Step 5: Determine the region of rejection
  - the region will depend on the critical values

Alternative hypothesis	$H_A: \pi \neq 52\%$	$H_A: \pi > 52\%$	$H_A: \pi < 52\%$
Critical value(s)	$z_{\alpha/2}$ & $-z_{\alpha/2}$	z_{α}	$-z_{\alpha}$
Region of rejection	$z > z_{\alpha/2}$ & $z < -z_{\alpha/2}$	$z > z_{\alpha}$	$z < -z_{\alpha}$

Hypothesis testing: step-by-step (cont'd)

- Step 6: Collect a sample, calculate a sample statistic
  - this is the evidence you use against the null hypothesis
- Step 7: Calculate the test statistic
  - this is a z-score that measures how different the sample statistic is to the null hypothesis
- Step 8: Conclusion
  - if the test statistic is in the region of rejection, reject the null hypothesis
  - if the test statistic is not in the region of rejection, do not reject the null hypothesis

Considerations in hypothesis testing

- The step-by-step guide shows how to conduct a test
- But there are some decisions that must be made!
- Two big examples are:
  1. What level of significance α should you choose?
  2. How large a sample n should you collect?
- The step-by-step guide doesn't tell you how to make these decisions!
- They're judgments that each statistician must make

Level of significance considerations

- As we've seen, α has a large impact on the test – it determines the critical values and region of rejection

- Broadly, it is a measure of how much evidence you need to reject the null hypothesis
- The lower you set α, the more evidence you need

- So how do you decide on a level?

Level of significance and error

- Your decision may be impacted by the fact that α is the probability of committing a type of error!
- α is the probability of rejecting the null hypothesis when it is true and shouldn't be rejected
- Note: this doesn't mean that you have made a miscalculation in your test!
- It only means that, while the null hypothesis is true, you happened to select a sample that suggested that it wasn't

Uncertainty in testing

- This highlights a vital fact: testing is not certain
- When you draw a conclusion from a test, you may be ‘wrong’
- But this doesn’t mean you’ve done anything wrong!
- It just means that the information you gathered from the sample didn’t match the population

Example

- Suppose you are testing whether a coin is fair
- You test whether heads turns up 50% of the time:

  $H_0: \pi = 0.5$

  $H_A: \pi \neq 0.5$
- Suppose you flip the coin 1,000 times and it turns up heads every time
- You'd probably conclude the coin wasn't fair!
- But it is possible that you were wrong, and you just got a really unlikely sample

Type I and Type II errors

- This is an example of a Type I error – you rejected the null hypothesis when it was true
- The probability of a Type I error, in general, is α

- There's another type of error – you commit an error if you do not reject the null hypothesis when it is false and should be rejected
- This is known as a Type II error
- The probability of a Type II error is denoted β

Type I and Type II errors (continued)

- Depending on how the conclusion from the test matches up (or doesn't match up) with what is actually happening, you will land in one of the four cells in this table:

	Null hypothesis true	**Null hypothesis false**
Null hypothesis rejected	Type I error	correct
Null hypothesis not rejected	correct	Type II error

The relationship between α and β

- At a fixed sample size, α and β are inversely proportional
- That is, decreasing one will increase the other:
  - As you decrease α, you decrease the region of rejection
  - This lowers the probability of a Type I error but increases the likelihood of a Type II error
- So at a fixed sample size, you can't make both errors completely unlikely!

Increasing the sample size

- So the answer lies in increasing the sample size
- Increasing the sample size will reduce the standard deviation in sampling distributions
- As a result, it is easier for us to tell the difference between a true null hypothesis and a false one
- If you increase the sample size, you can:
  - decrease α
  - decrease β; or
  - decrease both

Power

- The only positive conclusion you can draw from a hypothesis test is to reject the null hypothesis
- Remember – you can't conclude the null is true!
- So the 'power' of a test is a measure of your ability to correctly reject the null hypothesis
- In fact, the **power** of a test is defined to be the probability of rejecting the null when it is false, $1 - \beta$
- You can increase power by increasing sample size, decreasing the level of significance, or both

Testing the mean

- So far, the examples we've seen all relate to testing that a population proportion π assumes some value

- We can also do tests for a population mean μ
- The general step-by-step guide is the same!
- However, as with confidence interval estimation, if the population standard deviation σ is not known, you must use the t-distribution instead of Z
- This has a few effects on the running of the test

Testing the mean, σ unknown

- The critical value(s) will be t-scores from the correct t-distribution instead of z-scores from Z
  - For a two-sided test, they are $t_{\alpha/2}$ and $-t_{\alpha/2}$
  - For a one-sided test, it is t_{α} or $-t_{\alpha}$ (depending on H_A)
- Consequently, the region of rejection will be a region of the correct t-distribution, not Z
- The test statistic is found using the sample standard deviation s, not the population standard deviation σ
  - This test statistic will be a t-score, not a z-score
- But everything else about the test is the same!

Example

- Average life expectancy was 80 years, but we want to test if it is now bigger than this

  $H_0: \mu = 80$

  $H_A: \mu > 80$

- We will use a level of significance of $\alpha = 0.05$ and collect a sample of 100 life spans
- The critical value in this test is the t-score $t_{0.05}$ in the t-distribution with 99 degrees of freedom
- Statistical software tells us this value is $t_{0.05} = 1.66$

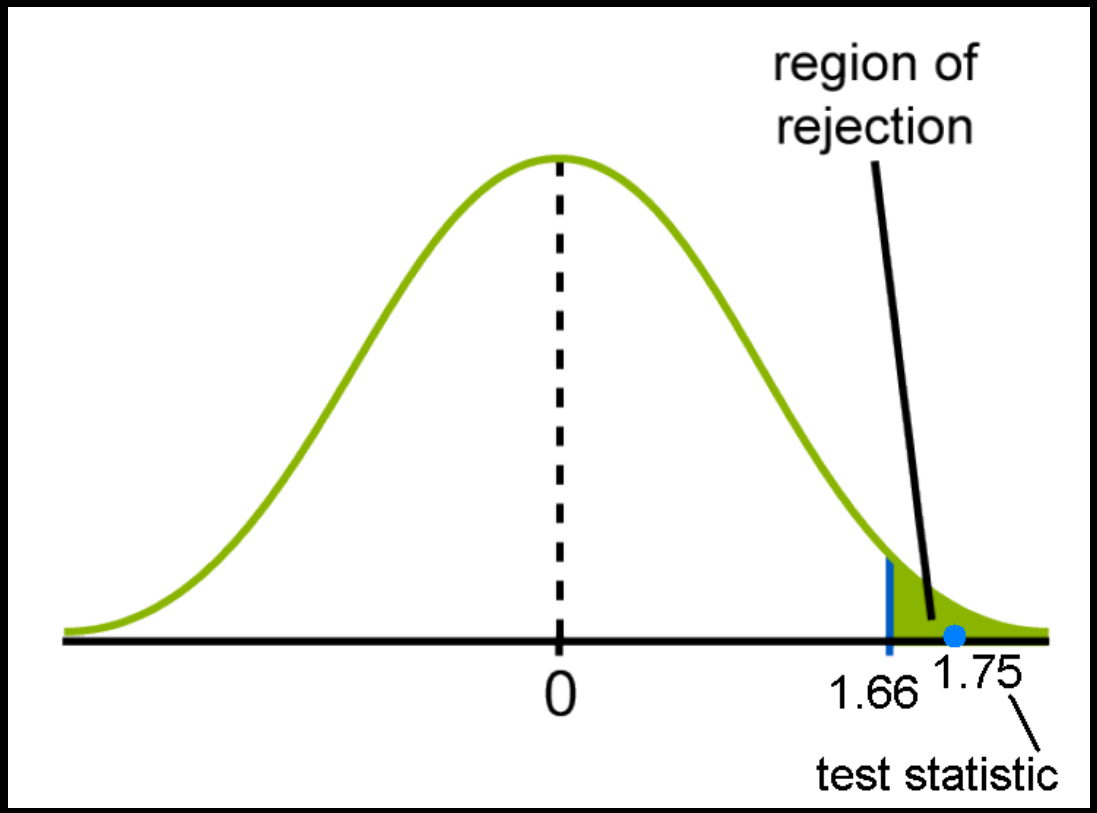
Example continued

- So the region of rejection is the set of values greater than 1.66
- Suppose the sample mean in the sample is $\bar{x} = 82.1$ and the sample standard deviation is $s = 12$ years
- Assuming H_0 is true, the population mean is 80 years and the test statistic is:

$$t = \frac{\bar{x} - 80}{s/\sqrt{n}} = \frac{82.1 - 80}{12/\sqrt{100}} = 1.75$$

Example continued

- The test statistic is in the region of rejection so we reject the null hypothesis



Test conclusions

- One of the big differences between estimation and testing is that tests result in one of two conclusions
- That is, a test is often conducted if a ‘black-and-white’ decision must be made
- The sample either constitutes ‘enough’ evidence to reject the null hypothesis, or it does not

- But is it always so black-and-white?

Likelihood of a sample

- The critical-value approach to testing requires us to answer the simple question:

  Is the test statistic in the region of rejection?

- But the value of the test statistic can tell us about how likely (or unlikely) our sample is if H_0 was true

If the sample is found to be very unlikely, we may think of this as evidence against the null hypothesis, even if the null hypothesis is not technically rejected

Example

- Suppose we have a one-sided test on the amount (in milligrams) of caffeine in a new brand of coffee

  H_0: $\mu = 300$

  H_A: $\mu > 300$

- The mean is $\bar{x} = 303.2$ in a sample of 25 coffees
- The standard deviation is assumed to be $\sigma = 10$
- This gives a test statistic of $z = 1.60$
- So how ‘extreme’ is this sample mean?

Example continued

- Well the chance of getting a sample mean that large is the same as the chance that Z will assume a z-score as large as $z = 1.60$
- The standard normal table tells us this is 5.48%
- That is, if the null hypothesis is true, there is only a 5.48% chance of obtaining a sample mean as large as 303.2 mg
- This is fairly small!
- We might even consider this evidence that the null hypothesis isn't true!

P-values

- This is known as the **P-value approach** to testing
- The probability 5.48% is referred to as the **P-value** for the test statistic (and for the sample)

- It is a measure of how likely a sample is, on the assumption that the null hypothesis is true
- The less likely, the stronger the evidence against the null hypothesis

One-sided vs two-sided P-values

- Just as with critical values, the P-value will depend on whether the test is one-sided or two-sided
- We just saw a one-sided test, and the P-value was the likelihood of obtaining a test statistic as large as the one obtained
- This is because the alternative hypothesis proposed that the mean was larger than 300 mg
- Depending on the nature of the alternative hypothesis, the P-value will change

One-sided vs two-sided P-values (cont'd)

- The P-value of a test statistic will depend on H_A

- If $H_A: \mu > 300$, P-value is the probability of getting a test statistic that large or larger (that is, positive)
- If $H_A: \mu < 300$, P-value is the probability of getting a test statistic that small or smaller (that is, negative)
- If $H_A: \mu \neq 300$, P-value is the probability of getting a test statistic that far away or further from 0