L20: Hypothesis testing

Hypothesis testing

Hypothesis testing

Hypothesis testing

We have covered how to draw confidence intervals around a sample mean and how to interpret this interval.

Today we will look at how we set us a statistical test against a "null hypothesis" and report a result with a "p-value"

There are differing opinions about the value of reporting p-values vs confidence intervals (CIs).

Often both are reported.

In public health we tend to favor CIs

Learning objectvies for today

Learning objectvies for today

- Define a hypothesis test
- Set up the steps for conducting a hypothesis test when the population standard deviation is known.

Note that we are still assuming we know the true SD - which is not realistic, however it is useful for the purposes of teaching how to set up a hypothesis test, and we will generalize this to a more realistic situation later.

Hypothesis testing

Hypothesis testing is the process we use to combine our assumptions with statistical inference to say how likely something is to be true given those assumptions and methods.

We start with a hypothesis. This is some specific question about a specific population. that we set up in terms of a parameter of interest

We then state this formally as a null and alternative hypothesis of that parameter.

We set a "significance level" or α

We calculate our test statistic and p-value based on the null distribution (implied by our null hypothesis) in other words, we assess how likely it is to see the data we saw under assumed conditions.

Then we write a conclusion for our tested hypothesis

Define the Hypothesis

A Null Hypothesis (H_0) is the hypothesis that is assumed to be true and the start of a test. This is often expressed as a statement of equality (ie. mean equal to a certain value or no difference between groups)

An Alternative Hypothesis (H_A) is usually the inverse of the null hypothesis and is expressed as a statement of difference.

- H_A: The mean is greater than the Null (one tailed)
- H_A: The mean is less than the null (one tailed)
- H_A: The mean is not equal to (greater or less than) the null (two tailed)

When we test a hypothesis, we are not trying to prove H_A, we are trying to disprove H_0

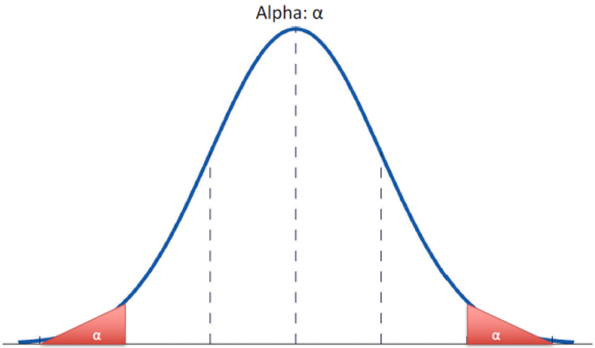
Decide on a threshold for rejecting the null

We choose a probability that we decide is small enough that we are unlikely to have observed it by chance if H_0 is true.

This threshold is our α.

We must decide if our hypothesis is one-tailed or two-tailed

What is my alpha?



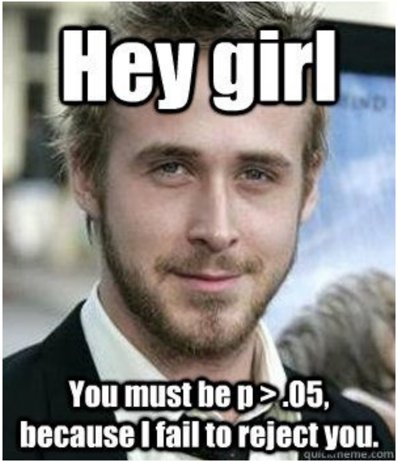
What is my alpha?

If we have a 95% confidence interval, we are saying that 95% of the time our true value will be in the given range, so 5% of the time my range does not contain the true value

5% here is the α

α represents the probability that you reject the null hypothesis when the null hypothesis is true

Decide on a threshold for rejecting the null



One sample questions

In our lecture on distributions we looked at using the normal distribution to ask how probable it was to observe a value in a population (ie. what proportion of the population would be as tall or taller than a value)

Now we are asking: "How likely is it that we observed a sample mean we did if the true population mean is μ_0"

One sample questions

Our H_0: There is no difference between the sample mean and the hypothesized null

Our H_A: There is a difference between our sample mean and the hypothesized null

Our α : The level at which we decide an outcome is not probable

Our tails: One tailed, or two tailed alternative?

- Let's consider an example.

Example: one tailed hypothesis

Inorganic phosphorus levels in the elderly

Levels of inorganic phosphorus in the blood are known to vary among adults Normally with $\mu = 1.2$ millimoles per liter and standard deviation 0.1 mmol/l.

A study examined inorganic phosphorus in older individuals to see if it decreases with age. Here are the data from 12 men and women between the ages of 75 and 79:

```
phos <- c(1.26, 1.39, 1, 1, 1, 1,
          0.87, 1.23, 1.19, 1.29, 1.03, 1.18)

known_sigma <- 0.1
```

*note that sigma here is given to you

Descriptives

The sample mean $\bar{x}$ is:

```
## [1] 1.12
```

Does the sample mean really differ from the null hypothesis mean?

The null hypothesis H_0 and alternative hypothesis H_a

The null hypothesis is a hypothesis of no effect or no difference. For our example:

$$H_0 : \mu = 1.2$$

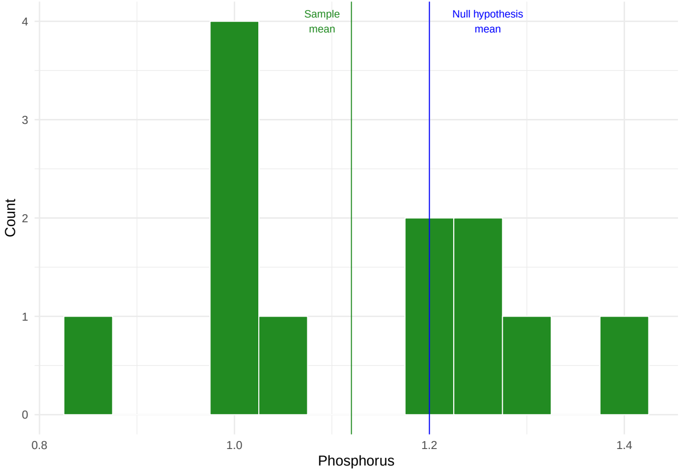
The alternative hypothesis (H_a) is the competing hypothesis. In this question, the competing hypothesis is that older individuals have an average phosphorus level lower than that seen in the overall adult population:

$$H_a : \mu < 1.2$$

This is a one-sided hypothesis

Descriptives

We can draw the data, the sample mean, and the hypothesized mean under H_0 all on one figure:



Performance of the hypothesis test

In our hypothesis test, we assume that the null hypothesis is true and calculate how likely we would observe a sample mean at least as extreme as the mean observed in the sampled data.

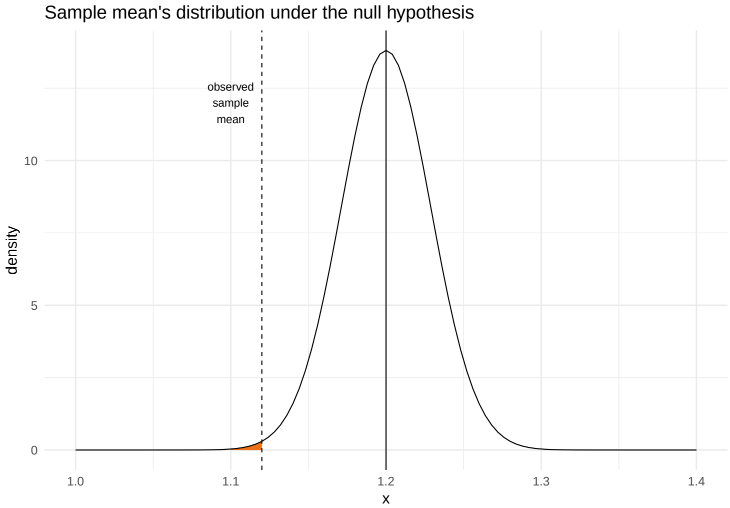
We need to evaluate the evidence *against* the null hypothesis that $\mu = 1.2$ assuming that the null hypothesis is true.

If it is, then the sampling distribution of $\bar{x}$ from n=12 individuals has:

- $\mu = 1.2$ and
- $sd(\bar{x}) = \frac{\sigma}{\sqrt{n}} = 0.1/\sqrt{12} = 0.0289$.

Draw the sampling distribution under the null hypothesis:

What is the probability of observing the sample mean (or lower) we observed if the null hypothesis were true?



Draw the sampling distribution under the null hypothesis:

What is the probability of observing the sample mean we observed *or lower* if the null hypothesis were true?

```
pnorm(q = 1.12, mean = 1.2, sd = 0.0289)
```

```
## [1] 0.002818636
```

Thus, under the null hypothesis, this is a 0.28% chance of observing a sample mean at least as small as what we saw. This is a very tiny probability, and suggests that the null hypothesis may not be true and that there is evidence in favor of the alternative hypothesis.

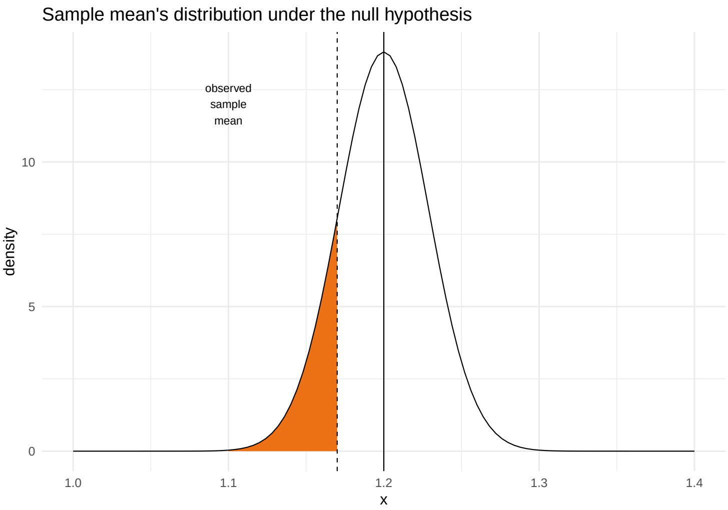
This probability is known as the p-value.

Example, extended.

Suppose instead that for this example, we chose a sample such that $\bar{x} = 1.17$.

- What is the null and alternative hypotheses?
  - H_0:
  - H_a:
- Let's look at the distribution under the null hypothesis and add a line at $\bar{x}$.

Example, extended.



Example, extended.

- Calculate the probability of observing an observation of 1.17 or lower.

```
pnorm(q = 1.17, mean = 1.2, sd = 0.0289)
```

```
## [1] 0.1496205
```

- Do you think you could observe this $\bar{x}$ if the null hypothesis were true?

So far

- We have considered a one-sided H_a.
- We computed the p-value using the Normal distribution of the sampling mean
- Next: Two-sided alternative hypotheses

Two sided testing

Hypothesis testing with a two-sided hypothesis

Suppose you are interested in the Aspirin content in a sample of pills. You have been told that the population of Aspirin tablets is Normally distributed and has a known standard deviation of 5 mg. Furthermore, you have a SRS of $n = 10$ pills. You need to be able to detect if there is evidence that the average Aspirin content in your sample is different from the population average and would be concerned if it appears to be higher or lower. Here:

$$H_0 : \mu = 325mg$$

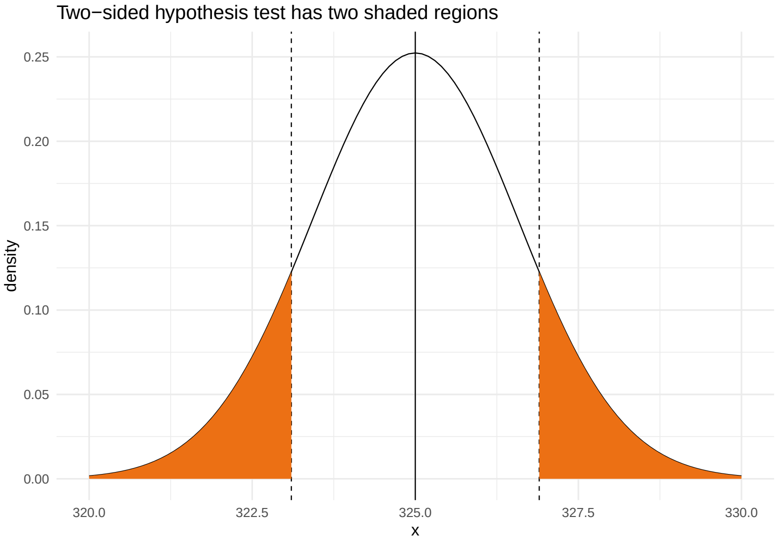
$$H_a : \mu \neq 325mg$$

This is a two-sided alternative hypothesis because we're interested in knowing if the sample mean appears to be either higher or lower.

Hypothesis testing with a two-sided hypothesis

1. Check the conditions.
2. Calculate the sampling distribution, assuming the null hypothesis is true:
   - $\mu = 325$
   - $sd(\bar{x}) = \sigma/\sqrt{n} = 5/\sqrt{10} = 1.581139$
3. Calculate $\bar{x}$ if provided data. Sketch the sampling distribution and add a vertical line at $\bar{x}$. Shade the region corresponding to H_a. If the hypothesis is two-sided, add another vertical line that is the same distance as $\bar{x}$ is from μ but on the other side of the distribution.

Two-sided alternative: add vertical lines at $\bar{x}$ and the equivalent distance from the null on the other side of the distribution



Calculation for the hypothesis testing with a two-sided hypothesis

```
pnorm(q = 326.9, mean = 325, sd = 1.581139, lower.tail = F)
```

```
## [1] 0.1147466
```

```
pnorm(q = 326.9, mean = 325, sd = 1.581139, lower.tail = F)*2
```

```
## [1] 0.2294932
```

- Why do we need to multiply the probability by 2?
- Why do we need to set `lower.tail=F`?

Interpret this probability. Does it provide evidence against the null hypothesis or, in the contrary, does this observation seem to follow under the null hypothesis?

Hypothesis testing using Z tests

Do you remember z-scores?

We briefly considered z-scores of the form:

$$z = \frac{x - \mu}{\sigma}$$

The z-score tells you how far x is from μ in terms of units of standard deviation.

Originally, we considered z-scores to examine how far a baby's birthweight deviated from the average birthweight we expected at a given gestational age.

You could make a statement like: This birthweight is 2.5 standard deviations below the mean, which is very small.

Do you remember z-scores?

We can generalize the z-score formula to look specifically at z-scores for $\bar{x}$, and use its μ and the standard deviation of the sampling distribution:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Recall that after standardization of a Normal variable, z~N(0,1)

Go back to the earlier example looking at phosphorus

$\bar{x} = 1.12$

$\mu = 1.2$

$\sigma/\sqrt{n} = 0.0289$ Thus:

$$z = \frac{1.12 - 1.2}{0.0289} = -2.768166$$

That is, z is 2.77 standard errors below the mean.

Go back to the earlier example looking at phosphorus

What is the probability of observing a z-score of -2.768166 (or lower) on the standard Normal distribution?

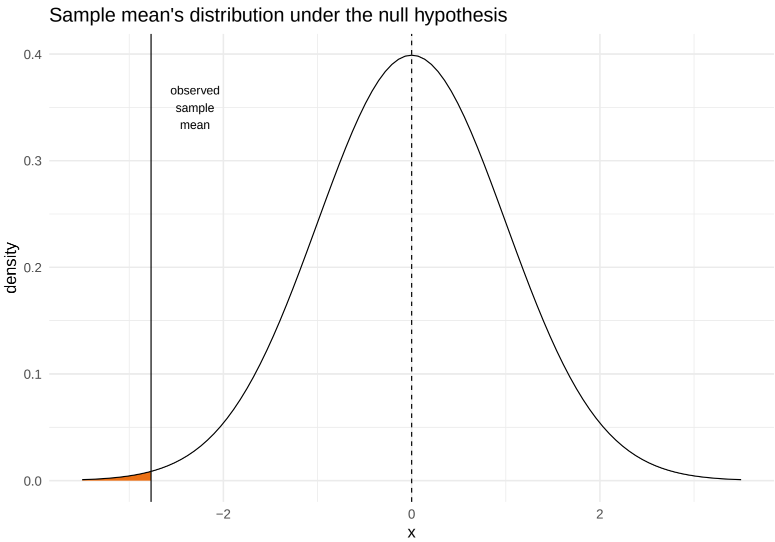
```
pnorm(q = -2.768166, mean = 0, sd = 1)
```

```
## [1] 0.002818637
```

The calculation is the same as before. We just standardized the units!

Go back to the earlier example looking at phosphorus

And we can see that the graph looks the same as well



Definitions

Test statistic

The Z value we calculated in our example is a kind of test statistic.

Test Statistic: Measures how far the data diverge from the null hypothesis H_0. Large values of the statistic show that the data are far from what we would expect if H_0 were true.

The Z is drawn from comparisons to a standard normal - we will learn about other tests statistics in part III

Formalizing what we meant by “test statistic”

The Z test for a population mean: Draw a SRS from a Normal(μ, σ) population, where μ is unknown, but σ is known. A test statistic and a p-value are obtained to test the null hypothesis that μ has a specified value:

H_0:

$$\mu = \mu_0$$

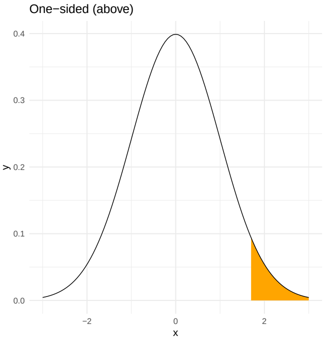
The one-sample z test statistic is:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Z test for a population mean (continued)

As the variable Z follows the standard Normal distribution (i.e., N(0,1)), the p-value for a test of H_0 against

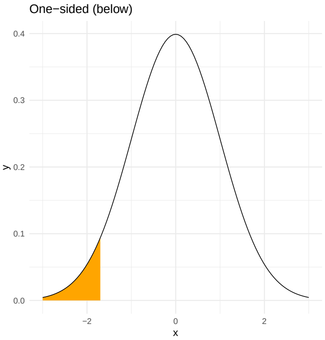
H_a: $\mu > \mu_0$ is $P(Z \geq z)$



Z test for a population mean (continued)

As the variable Z follows the standard Normal distribution (i.e., N(0,1)), the p-value for a test of H_0 against

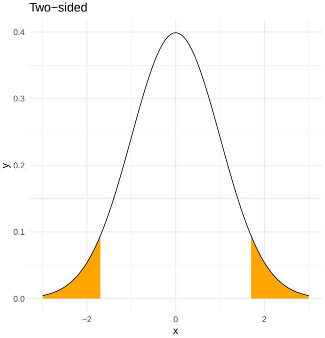
H_a: $\mu < \mu_0$ is $P(Z \leq z)$



Z test for a population mean (continued)

As the variable Z follows the standard Normal distribution (i.e., N(0,1)), the p-value for a test of H_0 against

H_a: $\mu \neq \mu_0$ is $2 \times P(Z \geq |z|)$



P-value

P-value: The probability, assuming that H_0 is true, that the test statistic would take a value at least as extreme (in the direction of H_a) as that actually observed. The smaller the p-value, the stronger the evidence against H_0 provided by the data.

Interpretation of the p-value

Remember!

We never conclude that the null hypothesis is true, only that there is "no evidence" against the null hypothesis.

We also never conclude that the alternative hypothesis is true, only that there is evidence to reject the null hypothesis.

Definition: Significance level and statistically significant

- The significance level, α, is an arbitrary threshold that can be used when reporting whether a p-value is "statistically significant".
- If the p-value is as small or smaller than α, people say that the data are statistically significant at level α.

Definition: Significance level and statistically significant

- Many researchers dislike the use of a significance level because it is a completely arbitrary cutpoint. If you are going to report p-values, it is much better to report the magnitude of the p-value than to only report whether the p-value is statistically significant or not.
- Both p-value $= 0.03$ and p-value $= 0.004$ are "significant at $\alpha = 0.05$", but the latter provides more evidence against the null hypothesis.
- Thus it is more informative to report the p-value you calculated then to only make a statement regarding statistical significance.

Reminder Statistical significance $\neq$ clinical significance

- If a finding is "statistically significant" this does not mean it is "clinically significant", or of a meaningful magnitude.
- For example, you might find that $\bar{x} = 1.19$ is statistically different from $\mu = 1.2$ (say $\alpha = 0.05$), but the estimated difference is only 1.2-1.19 $=$ 0.01. This might not be a meaningful difference.
- Whether the difference is of a meaningful magnitude in practice is not determined by the data, but based on judgement of decision-makers.
- What is a meaningful reduction in the number of drinks per day for an alcoholic?
- What is a meaningful reduction in depressive symptoms associated with a new very expensive treatment vs. a current cheaper treatment?

Relationship between confidence intervals and test statistics

- A two-sided test statistic at significance level α can be carried out from a confidence interval with confidence level $C = 1 - \alpha$.
- For example, if a 95% confidence interval does not contain the null value, then this implies that the p-value for the test that $H_0 : \bar{x} = \mu$ has a p-value $< \alpha = 0.05$ and is therefore statistically significant at the 5% level.

Example: Relationship between confidence intervals and test statistics

Recall earlier in the slides that we calculated the 95% CI for the mean height of girls, based on a sample of girls in a Midwestern school district. This CI was from 100.56 to 111.12.

Suppose you wanted to test whether the mean height was different form $H_0 : \mu = 113$ cm. This mean height is outside of the 95% CI, so we know that the p-value corresponding to the two-sided hypothesis test would be $< 5\%$, and we could conclude that it is statistically significant for $\alpha = 0.05$.

Relationship between confidence intervals and test statistic

CIs and p-values provide similar information, because you can deduce directly whether a test will be < 0.05 from a 95% CI.

However, if you only know a p-value you cannot derive the CI.

The CI is better because it puts a range around the magnitude of the value of the parameter.

Parting Humor