جامعة نيويورك أبوظبي

NYU | ABU DHABI

PSYCH-UH 1004Q: Statistics for Psychology

Class 8: The logic of null hypothesis testing

Prof. Jon Sprouse
Psychology

The basic logic of Null Hypothesis Testing

(Fisher's approach)

Defining the null hypothesis

Null Hypothesis Testing (NHT) begins by defining two hypotheses: the null hypothesis and the alternative hypothesis (or experimental hypothesis):

H_0: The **null hypothesis**. This states that there is no effect in your data (e.g., no difference between conditions in an experiment).

H_A or **H_1** The **alternative hypothesis** or **experimental hypothesis**. This is a hypothesis that states that there is an effect in your data (e.g., a difference between conditions). This is sometimes also called. Your book calls it H_A, but you will also see it called H_1.

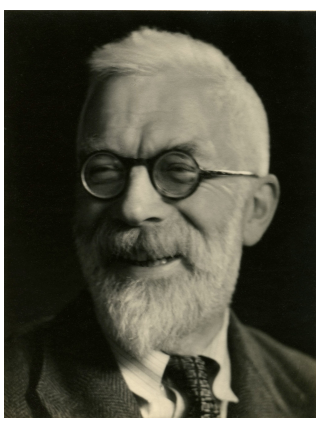
So let's start with the obvious - as the name suggests, NHT focuses on the null hypothesis. In fact, the alternative hypothesis (or experimental hypothesis) doesn't really factor into the mathematical steps of null hypothesis testing.

This is counterintuitive. The alternative hypothesis is the one you care about. It is the one that is scientifically interesting. Your gut is going to make you want to learn things about the alternative hypothesis. That is natural. That means you are a scientist. But logic is not always intuitive.

It is all about falsification

NHT uses the logic of **falsification**. In NHT, you seek to **reject** the null hypothesis.

H_0: The **null hypothesis**. This states that there is no effect in your data (e.g., no difference between conditions in an experiment).



Ronald A. Fisher (1890-1962) was a British geneticist who developed the first coherent framework of NHT.

"Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis." - Fisher (1966)

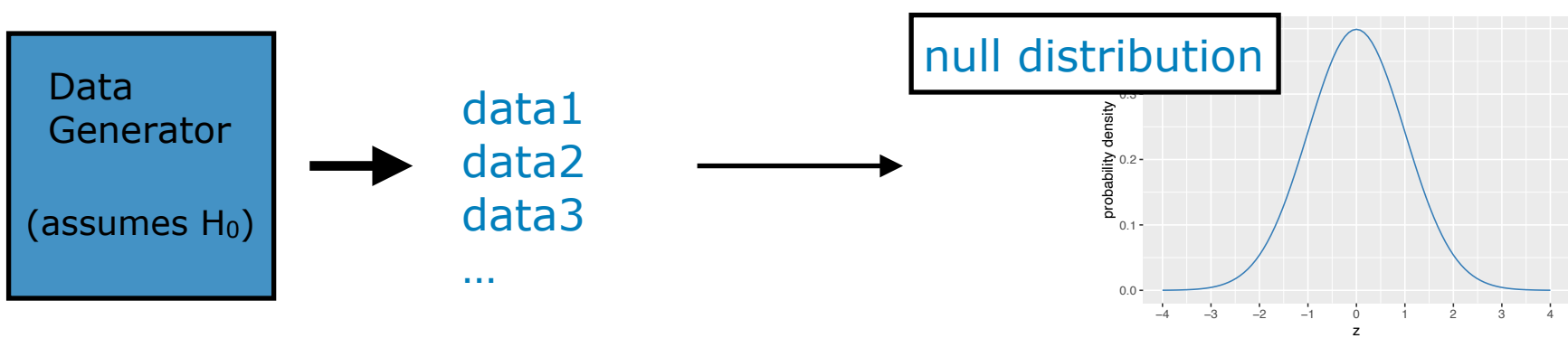
To repeat - the goal of NHT is to reject the null hypothesis. NHT makes no statements about the alternative hypothesis (either disproving or proving). You should sear this into your brain.

The mathematical part of NHT

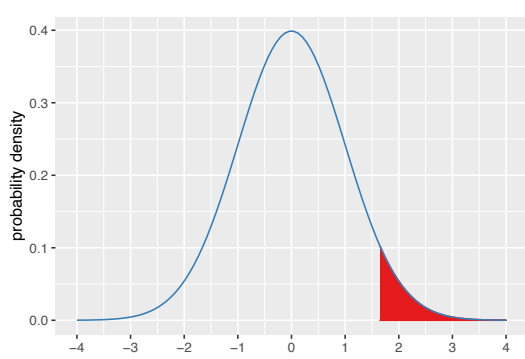
The mathematical part of NHT has three steps:

1. Run an experiment to collect the observed data. Calculate a statistic from it, like the mean or a z-score (or others we will learn later like *t* or *F*).

2. Assume that the null hypothesis is true, and generate **all possible data sets** that could arise (using the same sample size as your experiment). We summarize it as a distribution called the null distribution.



3. Look up the probability of your observed data or data more extreme in the null distribution. This is a conditional probability.

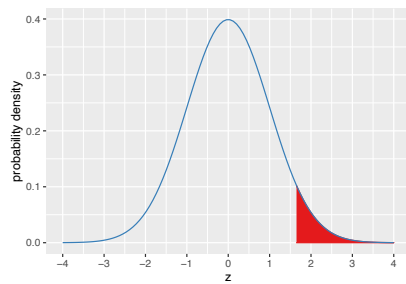


$$\mathbf{P(data \mid H_0)} = \frac{\text{observed data}}{\text{generated data}}$$

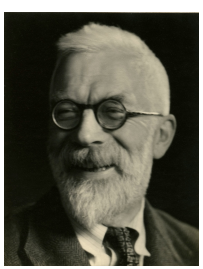
The logical part of NHT

The **mathematical** part of NHT yields a conditional probability - the probability of obtaining the observed data or data more extreme under the assumption that the null hypothesis is true. We call this a ***p*-value**.

$$p\text{-value} = P(\text{data} \mid H_0) = \frac{\text{observed data}}{\text{generated data}}$$



The **logical** part of NHT interprets the *p*-value.



Interpreting *p*-values is actually a fairly philosophical act. We will start with Fisher's philosophy, because he started NHT. His interpretation can be captured in a statement called **Fisher's disjunction** (a disjunction is a statement with "or" in it):

If the *p*-value is sufficiently low, then you can conclude either:

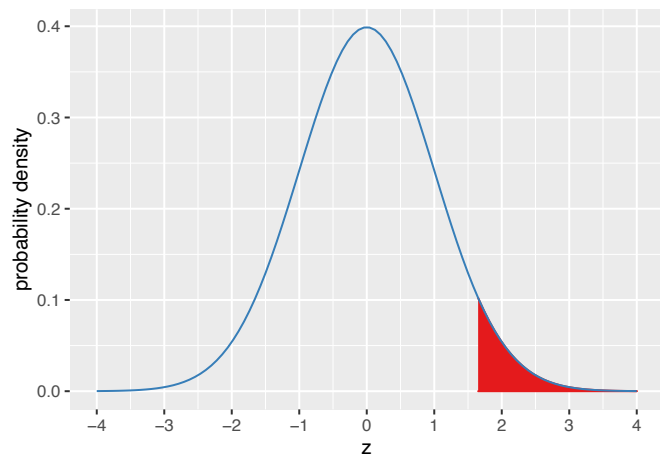
(i) the null hypothesis is incorrect, or
(ii) a rare event occurred.

Really thinking about it

If the *p*-value is sufficiently low, then you can conclude either:

(i) the null hypothesis is incorrect, or
(ii) a rare event occurred.

Option 1 says the null hypothesis is incorrect. The idea is that if our observed data has a low probability, then maybe the hypothesis we used to generate the probability distribution is incorrect. We don't expect low probability things to happen very often. So, when we see one, we should question whether we actually understand the universe correctly.



Option 2 says that the null hypothesis is correct, and we just happened to observe a rare event. Rare events do occur. They just occur rarely. We just got lucky, and happened to see one.

Notice that these two are opposites - one says H_0 is incorrect, the other says H_0 is correct. Fisher's disjunction says that we can't know which one is true. We only know that one of the two is true.

Rejecting H_0

If p(data|H_0), called the p-value, is sufficiently low, then you can conclude either: (i) the null hypothesis is incorrect, or (ii) a rare event occurred.

The obvious question is how low does a *p*-value need to be in order to reject the null hypothesis?

For Fisher, *p*-values are directly interpretable as the strength of evidence against the null hypothesis. This is because they are frequentist probabilities. So they have a very natural interpretation - they tell us how frequently we would expect our data or data more extreme if the null hypothesis were true.

Fisher wanted each scientist to decide when they would reject the null hypothesis. So, for Fisher, there is no correct answer to this. He did make the suggestion that a *p*-value less than .05 would probably be a good cutoff. And the field of psychology has generally followed that suggestion.

But, crucially, for Fisher, *p*-values are a continuous measure of evidence against the null hypothesis. So there is not much difference between p=.06 and p=.05. There is no hard or fast cutoff that you can use.

Does rejecting H_0 mean that we proved H_A /H_1?

If p(data|H_0), called the p-value, is sufficiently low, then you can conclude either: (i) the null hypothesis is incorrect, or (ii) a rare event occurred.

It is tempting to say that we proved the alternative hypothesis (aka the experimental hypothesis). But science is about precision. So we need to be precise.

NHT doesn't work with H_A /H_1. We never mathematically formulate an H_A /H_1. And none of the steps in the mathematical or logical portions of NHT make reference to H_A /H_1. Therefore we cannot draw any direct conclusions about it. Our conclusion is just that we reject H_0.

But all is not lost. The logical implication of rejecting the null hypothesis is that some H_A /H_1 is likely correct. In other words, we have ruled out one hypothesis (the null hypothesis), so now we know that it will be more productive to explore alternative hypotheses.

But it is important to remember that we did not test H_A /H_1 directly.

… or a rare event occurred.

If p(data|H_0), called the p-value, is sufficiently low, then you can conclude either: (i) the null hypothesis is incorrect, or (ii) a rare event occurred.

Don't forget the other half of the disjunction.

Rare events do happen. They just happen rarely. For example, if the *p*-value of the observed data is .05, that tells us that data equal to or more extreme than the observed data will occur in around 1 out of 20 experiments over the long run.

There is no way to know if the null hypothesis should be rejected or if a rare event occurred. All we can do is interpret the *p*-value based on our best scientific judgment. This is just part of a broader issue in science - we can never know the truth. We can only use our best judgment based the evidence.

Failing to reject H_0

If p(data|H_0), called the p-value, is sufficiently low, then you can conclude either: (i) the null hypothesis is incorrect, or (ii) a rare event occurred.

Though it is not stated directly in the disjunction, we have to consider what happens if the *p*-value is not low. Let's say it is p=.36. What can we conclude?

We conclude that we have failed to reject the null hypothesis.

This probably sounds like a really clumsy phrase. And it is. But it is also precise. And science is about precision.

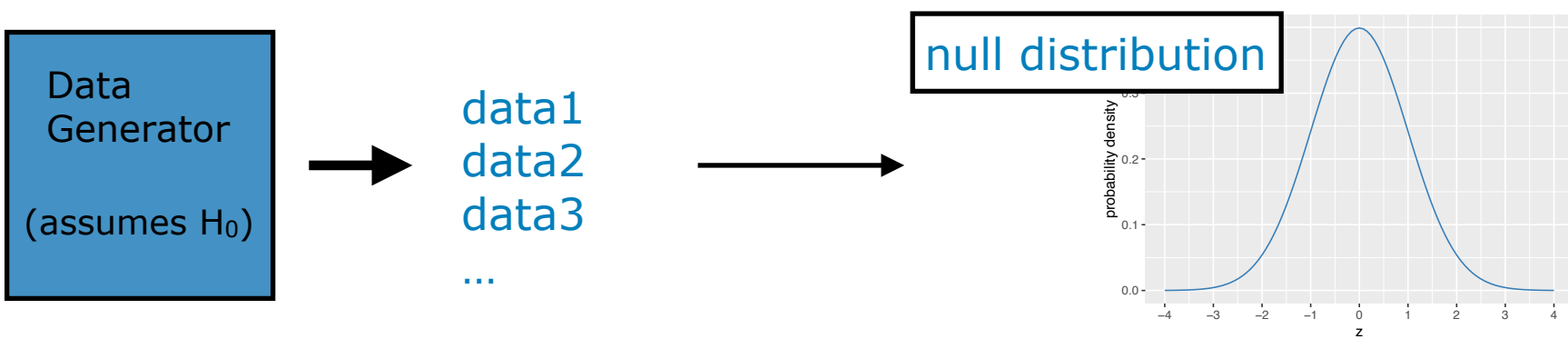
You may be tempted to say that this proves the null hypothesis. But **we cannot prove the null hypothesis in NHT because we assume it is true** in our mathematical steps (the generation of the null distribution). You can't prove something that is already assumed to be true. (And, also, NHT only endorses falsification!)

A brief note on ways to generate data under the null hypothesis

The mathematical part of NHT requires generating the null distribution

When you think of statistics, you probably think of math. The reason is that the mathematical step of NHT is a little complicated. Here it is again:

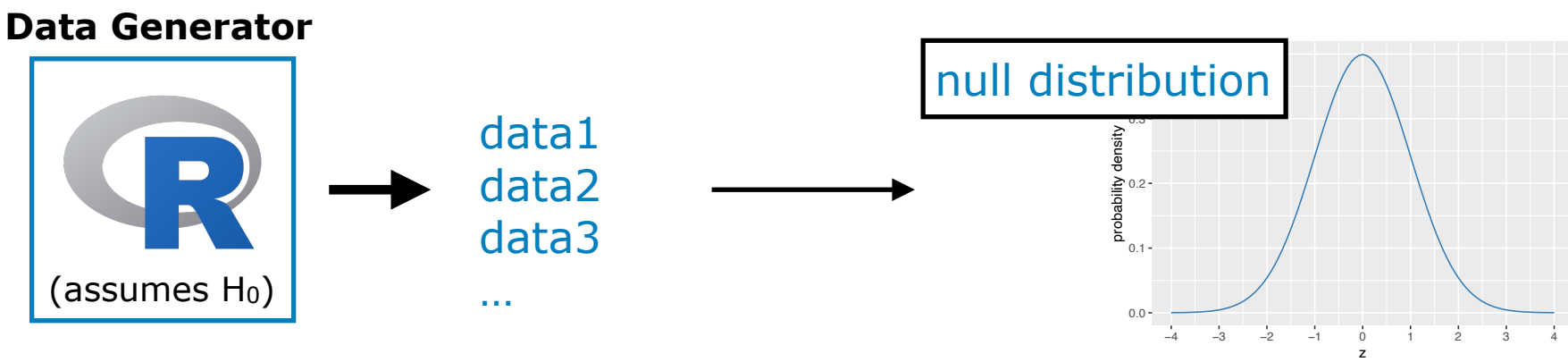
Assume that the null hypothesis is true, and generate **all possible data sets** that could arise (using the same sample size as your experiment). We summarize it as a distribution called the null distribution.



The question is: How can we generate the hypothetical data under the null hypothesis? Remember, these aren't real experiments. We aren't actually running them. We couldn't do that physically. We just need to know what would happen, hypothetically, if the null hypothesis were true and we ran all possible experiments. It turns out that there are two main methods of doing this - the intuitive way (simulating the data) and the less intuitive way (analytic methods using calculus).

The intuitive way - simulating the data

Living now in the 21st century, the obvious way to generate hypothetical data is to **simulate it**. We've already been doing that from time to time in this course. It means using a computer language, like R, to generate the data.



We don't focus on simulation methods in this course. But if you'd like to learn more about simulation methods, the words you want to search for are: **randomization methods** (sometimes called permutation methods) and **bootstrap methods**. Those are the two major approaches to simulating data within Null Hypothesis Testing.

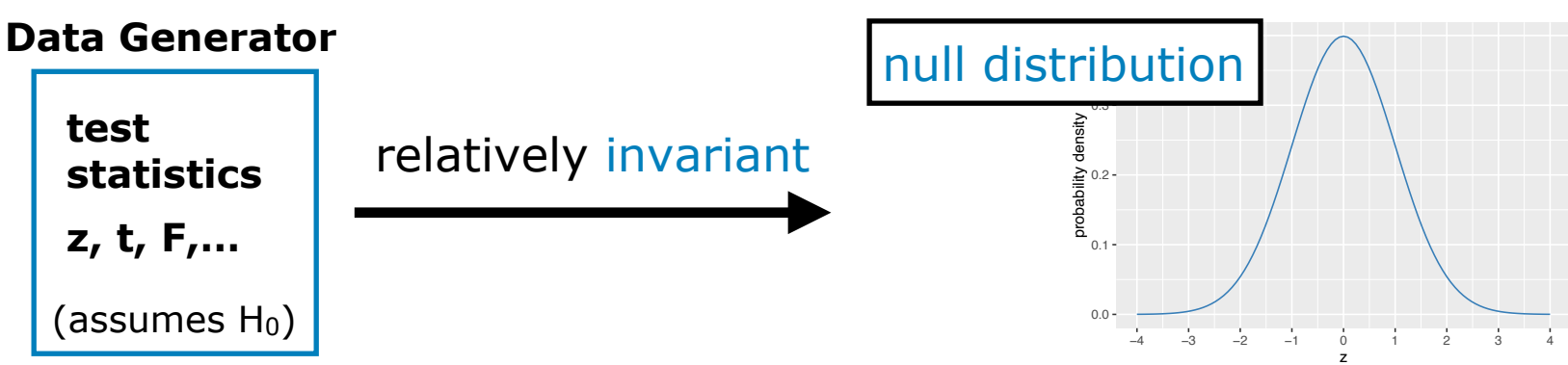
The less intuitive way - analytic methods

Null Hypothesis Testing was developed primarily in the first half of the 20th century (with Fisher doing the bulk of the work in 1920s and 1930s).

Because that was before computers, simulation methods simply were not feasible for anything beyond a very, very small experiment. Otherwise scientists would spend all of their time calculating fake data for every experiment that they ran. So the founders of statistics had to be clever.

Here is what they did: They discovered special statistics (single numbers that describe a sample), called **test statistics**, that had relatively invariant distributions when the null hypothesis is true. What "invariant" means is that the distributions of these statistics do not vary based on the scales of the measurements - height, stress, SAT scores, temperature of stars, etc.

This means that we can calculate the distributions once, print them in books, and use them over and over. It is a really clever solution!



Why do we still learn analytic methods?

Now that we do have access to computers, and can simulate data, why do we still learn analytic methods? And why do we learn them first?

Part of this is practical - so that you can connect with the literature that came before now. There is over 100 years of science that uses analytic methods. If we didn't maintain analytic methods, we'd lose access to that knowledge!

But the deeper reason is conceptual. Analytic methods are not cheats or shortcuts. They reveal something deep about data sets and the properties that data sets have under the null hypothesis.

In an intro course like this, we may not get to see that depth yet. This course builds the foundation. Just like any skill, or any academic discipline, we need to build the foundation before we can really see the full complexity. (For 80s nerds - we are waxing the cars and painting the fence before we can learn karate.)

But I want you to know that there is another layer to this that you can explore once you are finished with this course. I will try to point to it a bit when I can, and maybe at the end of the course we will have time to discuss the directions you can go next with statistics.

How to **learn** a new statistical test

(This is pretty much what we will do over and over again as we explore NHT together)

How to learn a new statistical test

What is the **scientific question** it attempts to answer? This tells us which kinds of questions we can apply it to.

What is the **mathematical question** it asks? This tells us the information it is going to give us to answer the scientific question.

What is the **null distribution** of the test statistic? These are well-known. We just need to know which one we are working with.

How is the test statistic **calculated** for our observed data? This is the calculation that we will need to perform. This is the math bit of statistics.

How do we **compare** the test statistic to the null distribution? There are really two answers - look it up in a book, or use a built-in R function. (I suppose you could simulate it yourself, but at that point, you might as well use simulation methods directly.)

Null Hypothesis Testing with the **z-test**

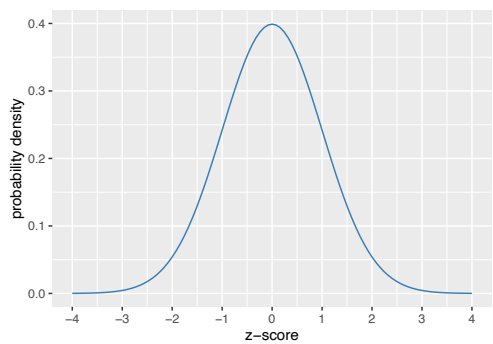
(We've already seen this, we just need to make it explicit!)

The **scientific** question

The first thing to learn about any null hypothesis test is what kind of **scientific question** it can answer.

The z-test asks: Does our sample come from a known population, or does it come from a different population?

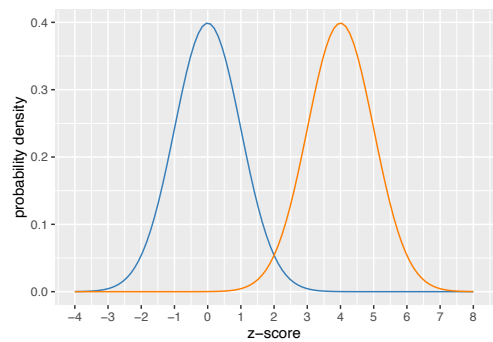
Comes from known population



[our sample]

-or-

Comes from another population



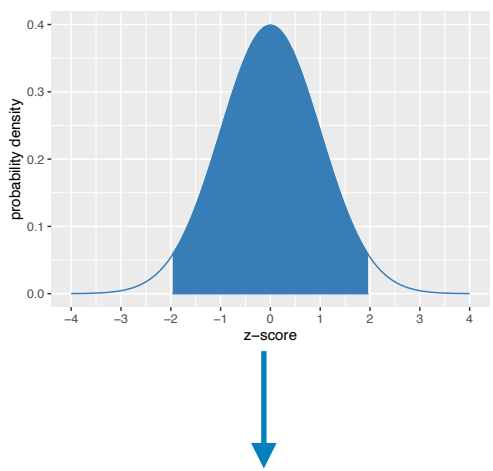
[our sample]

Notice that only certain scientific questions will fit: This has to be a question about one sample, and it has to be about something where we know the parameters of the population. (These are rare, but it is the simplest test, so we start with it!)

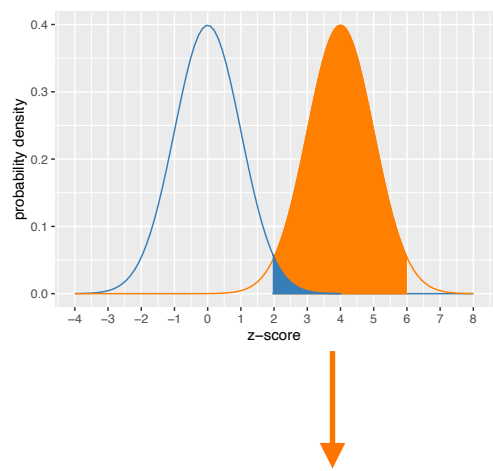
The **mathematical** question

The second thing to learn about any null hypothesis test is what kind of **mathematical question** it answers. It will always be a *p*-value in NHT. But it is important to notice exactly what the *p*-value is quantifying.

The z-test asks: What is the probability of obtaining our sample or one more extreme if it came from the known population. (This is a *p*-value.)



If our sample comes from the known population, the probability should be high.



If our sample comes from the other population, the probability that it is part of the known population should be low — notice the overlap is in the extreme tail of the known population.

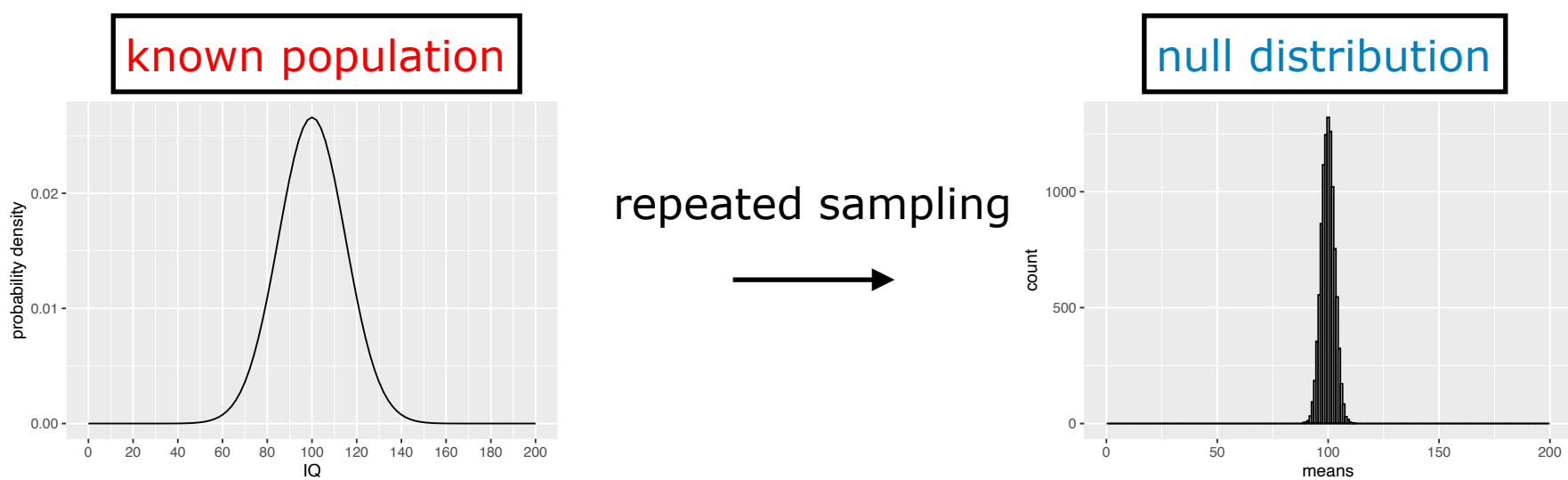
The **null distribution** for the statistic

The third thing to learn about any null hypothesis test is what the null distribution is for the test statistics (z, t, F, etc). Here we are talking about z as a test statistic. So we need to figure out what the distribution of z is under the null hypothesis.

Our null hypothesis is that the sample that we've observed comes from a known population. We can use the sample mean as our statistic, and convert it into a sample z-score.

So the null distribution would be all possible sample means from that known population. We've seen that before — it is called the **sampling distribution of the mean**.

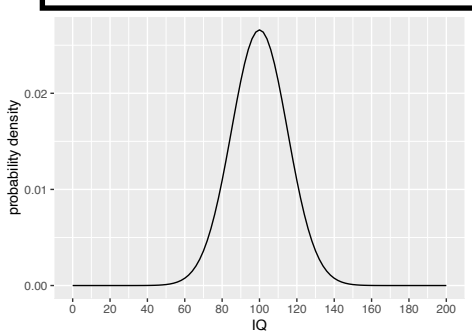
So, our null distribution is the sampling distribution of the mean.



The **null distribution** for the statistic

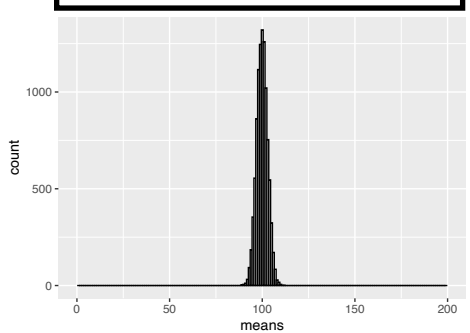
But we can go a step further. We know that the sampling distribution of the mean is a normal distribution in many cases (specifically, when the population is normal, or when then sample size is >30). So our null distribution is a normal distribution.

known population



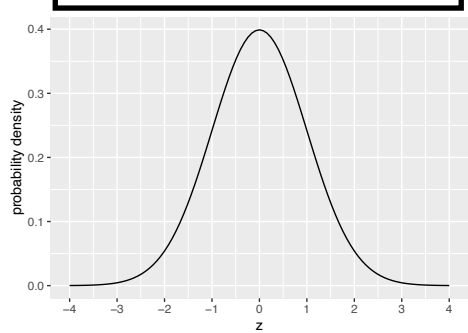
repeated sampling

null distribution



z-transform

null distribution



This means that we can convert our null distribution to the standard normal distribution. So, when we use z as a test statistic, our null distribution is the standard normal distribution (so we can use table A1 or pnorm for p-values!). In other words, the distribution of the z-statistic is a standard normal distribution.

The **calculation** of the (observed) test statistic

The fourth piece of knowledge you need is how to calculate the test statistic from your sample(s). Luckily, we already know this one. We just didn't talk about it this way. We are going to calculate the z-score for a sample mean.

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \qquad \text{... and remember:} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

The example used in the book is SAT scores, which have a known population mean of 500 and standard deviation of 100 (this is specified by the company that creates the SAT).

In the example, someone collects a sample of 25 scores that have a mean of 530. The question is whether this sample likely comes from the known population of SAT scores, or whether the sample comes from a different population (like the population of people really good at math). Plugging the numbers in, we get:

$$\sigma_{\bar{x}} = \frac{100}{\sqrt{25}} = 20 \qquad z = \frac{530 - 500}{20} = 1.5$$

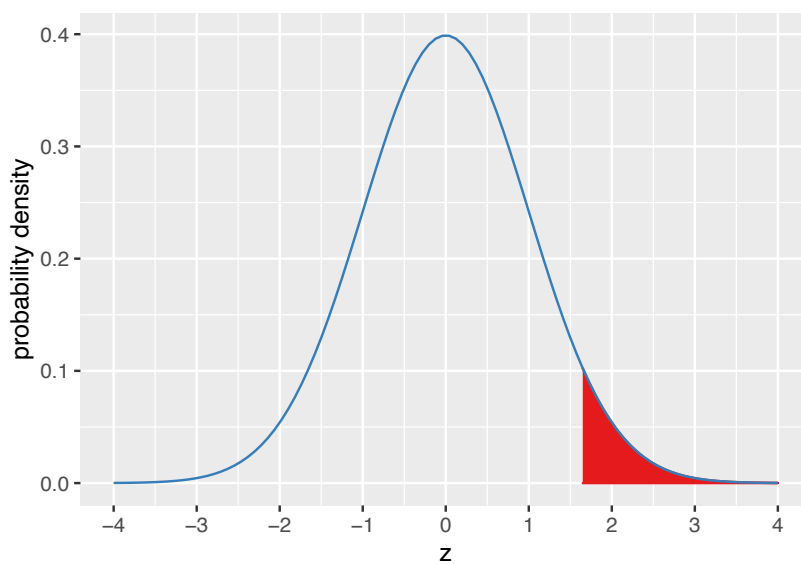
Compare the observed test statistic to its null distribution

The last piece of information we need is how to compare the observed statistic to its null distribution.

Option 1: The tables in the appendix of our book. There are tables for each of the different test statistics that are covered in the book.

Option 2: Built-in functions in R. R has functions for each of the test statistics as well.

Whichever option you use, the procedure is roughly the same. You use the test statistic, in this case the z-score, to find the probability of obtaining that score or one more extreme (operationalized as the area under the curve of the distribution)!



And if you look up 1.5, the answer is p=.067

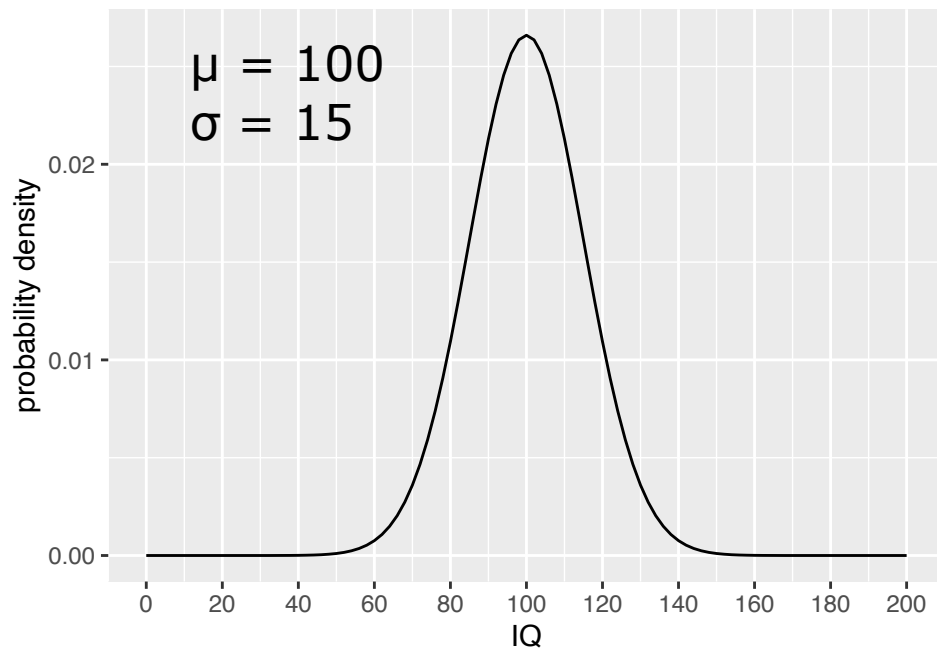
Putting it all together with another example

Another standardized measure: “IQ”

IQ stands for “intelligence quotient”. It is typically measured through a battery of tests. There are a lot of controversies surrounding IQ measures, so I don’t want us to dive too deeply into them. But they are a good example for z-tests because they are a standardized measure (similar to SAT scores) — the population mean and standard deviation are known:

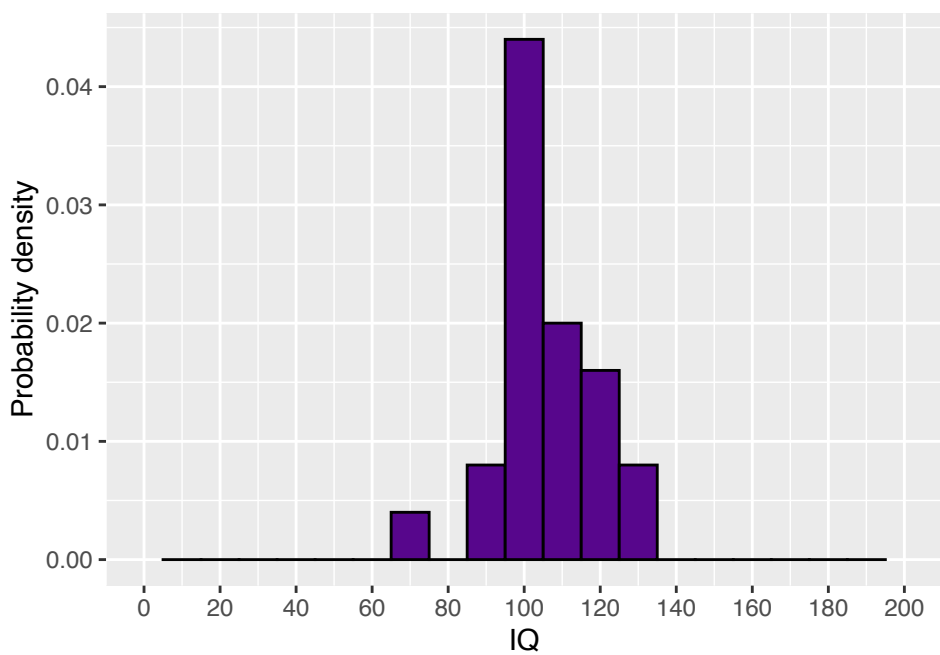
Given that we know the population mean and the population standard deviation, we can use z-tests to evaluate any experiments that we run on IQ.

Population mean and sd:



The effect of practice of IQ tests

A researcher has a hypothesis that practicing IQ tests will cause people to perform better on the test. The researcher recruits a sample of 25 participants. They bring each participant into the lab on 4 consecutive days. For the first three days, the participants take an IQ test as practice (a different test each day). On the fourth day, they take the critical IQ test. The researcher records the 25 values.



scores:

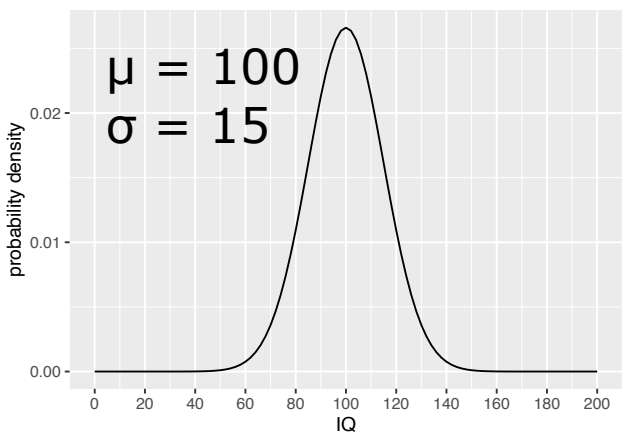
72 93 93 96 98
99 100 101 101 102
103 103 104 105
107 109 110 113
115 118 119 122
125 126 127

$\bar{x} = 106$

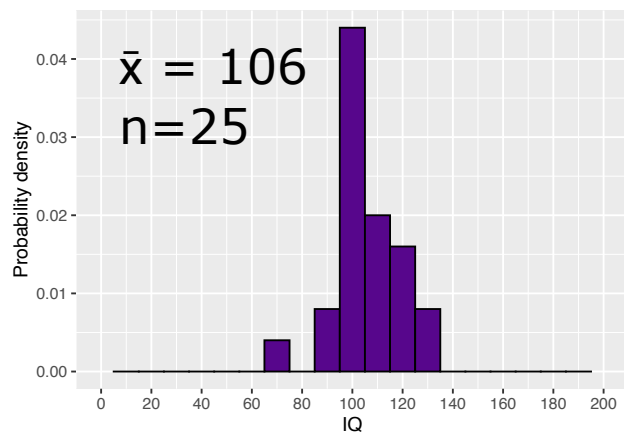
What are we asking?

Let's look at what we've done so far. We have a known population of IQ scores with a mean of 100 and a standard deviation of 15. We have selected a sample of 25 scores, and "treated" them with our condition — they practiced taking IQ tests.

Population without treatment



Sample with treatment

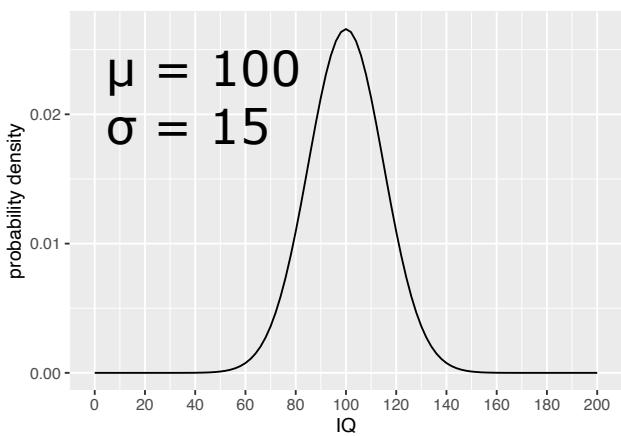


What we want to know is if it likely that our sample came from that population (= the treatment had no effect, so they still come from the general population), or if it is unlikely that our sample came from that population (= the treatment had an effect, such that we should now think of these people as coming from a different population - people who are better at IQ tests).

For NHT, we generate all possible experiments with sample size 25 under the null hypothesis

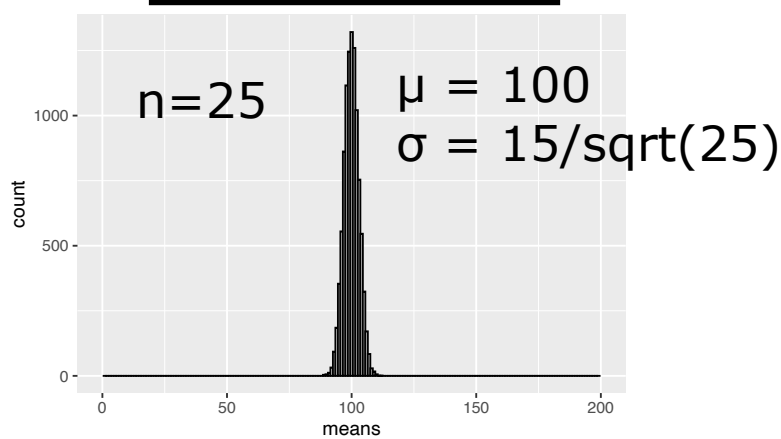
We know two ways to do this. The first is to simulate it. We can simply draw samples of size 25 from this population over and over. This is our null distribution:

Population without treatment



sample repeatedly

null distribution



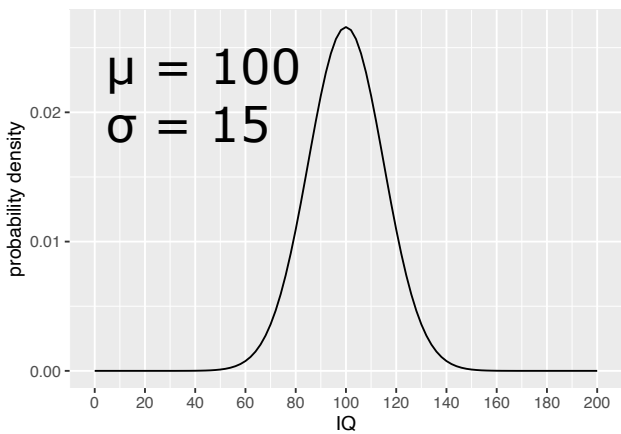
Note that our null distribution is just the sampling distribution of the mean!

We could, if we wanted, stop there and simply look up our sample mean (106) in the null distribution that we just calculated. Modern computers would allow us to do that. We can use pnorm() with a mean of 100 and a standard deviation of 3, and get a p-value for our mean of 106.

For NHT, we generate all possible experiments with sample size 25 under the null hypothesis

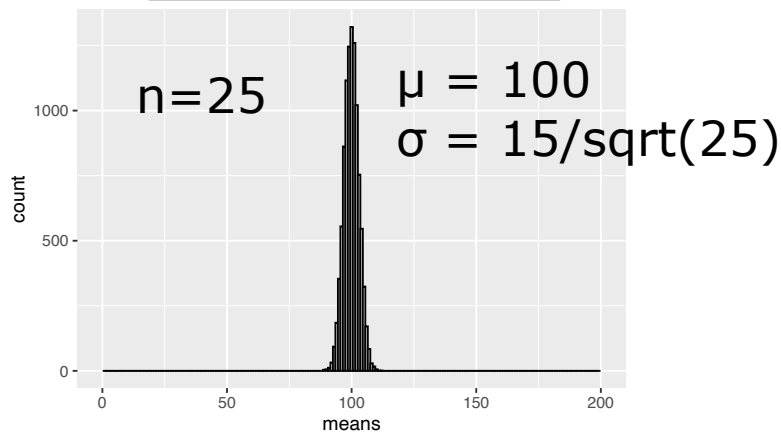
We know two ways to do this. The first is to simulate it. We can simply draw samples of size 25 from this population over and over. This is our null distribution:

Population without treatment



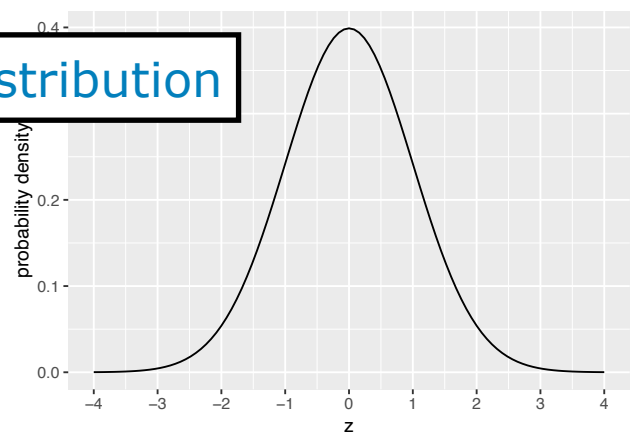
sample repeatedly

null distribution



But we also know an analytic shortcut. The null distribution is just the sampling distribution of the mean, and the sampling distribution of the mean is normal, so we can use the standard normal distribution instead.

null distribution



The analytic shortcut we can take is to calculate the z for our sample

This is what our z-test does. It is an analytic shortcut that shows us the z-score of our sample relative to our known population without having to simulate anything:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \qquad \text{… and remember:} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

So let's plug in our numbers:

$$z = \frac{106 - 100}{3} \qquad \text{… and remember:} \quad \sigma_{\bar{x}} = \frac{15}{\sqrt{25}}$$

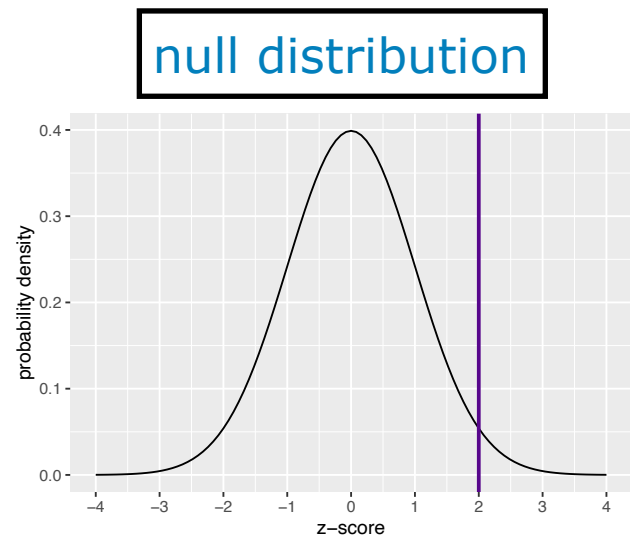
So the z for our sample is 2!

And then compare it to the known null distribution (in this case, the z distribution)

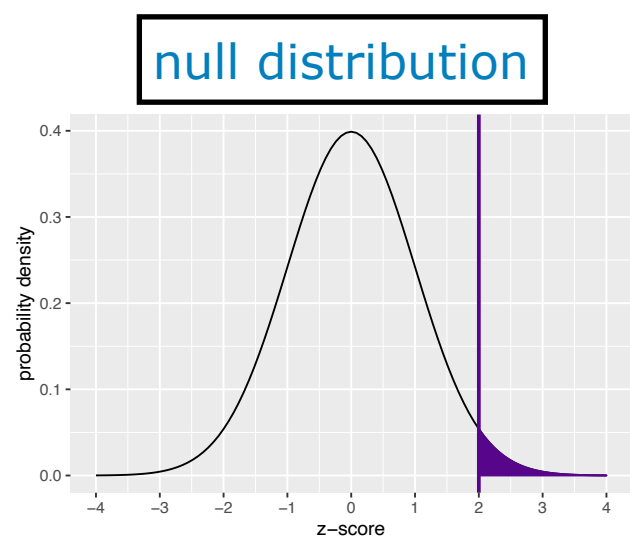
To complete the z-test, we just need to find our sample's z in the null distribution:

$$z = \frac{106 - 100}{3}$$ … and remember: $$\sigma_{\bar{x}} = \frac{15}{\sqrt{25}}$$

$z = 2$



And then we ask the NHT question: What is the likelihood of observing our value (z=2) or one more extreme, under the null hypothesis (in the null distribution)?



And the answer (using either approach) is p=.023

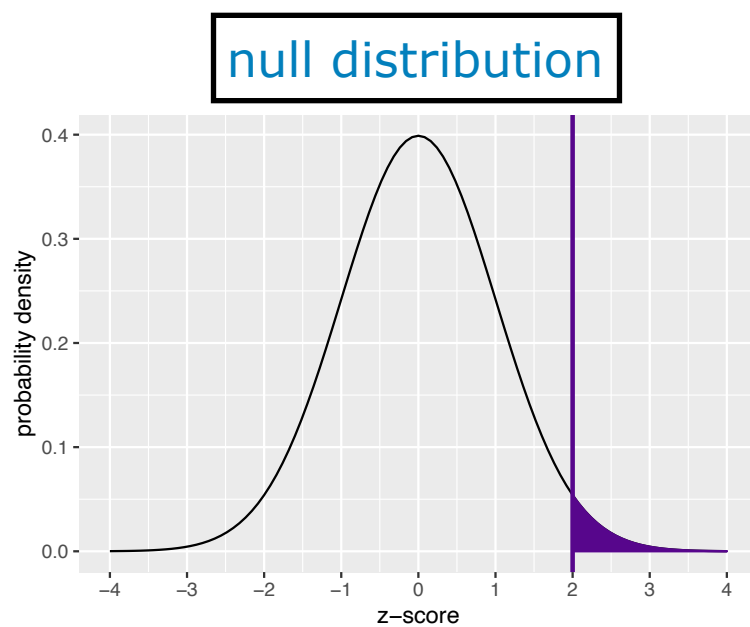
Then we can apply logic

So far, we have only learned Fisher's logic. Next time, we will learn Neyman-Pearson's logic. But for today, let's apply Fisher's disjunction to our results:

If p(data|H_0), called the *p*-value, is sufficiently low, then you can conclude either: (i) the null hypothesis is incorrect, or (ii) a rare event occurred.

For our question about the effect of practice on IQ scores, we found that our *p*-value is .023.

That means either the null hypothesis is false (practice DOES impact IQ scores), or something very rare happened in our experiment (we just happened to sample people who are unnaturally good at IQ tests — from the tail of the null distribution).



The z-test in practice!

If we saw this on homework or an exam

IQ stands for "intelligence quotient". It is typically measured through a battery of tests. There are a lot of controversies surrounding IQ measures, so I don't want us to dive too deeply into them. But they are a good example for z-tests because they are a standardized measure (similar to SAT scores) — the population mean and standard deviation are known:

scores:

72 93 93 96 98
99 100 101 101 102
103 103 104 105
107 109 110 113
115 118 119 122
125 126 127

$\bar{x} = 106$

We'd first remember our formula:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \quad \text{… and remember:} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Then plug in our numbers:

$$z = \frac{106 - 100}{3} \quad \text{… and remember:} \quad \sigma_{\bar{x}} = \frac{15}{\sqrt{25}}$$

If we saw this on homework or an exam

IQ stands for "intelligence quotient". It is typically measured through a battery of tests. There are a lot of controversies surrounding IQ measures, so I don't want us to dive too deeply into them. But they are a good example for z-tests because they are a standardized measure (similar to SAT scores) — the population mean and standard deviation are known:

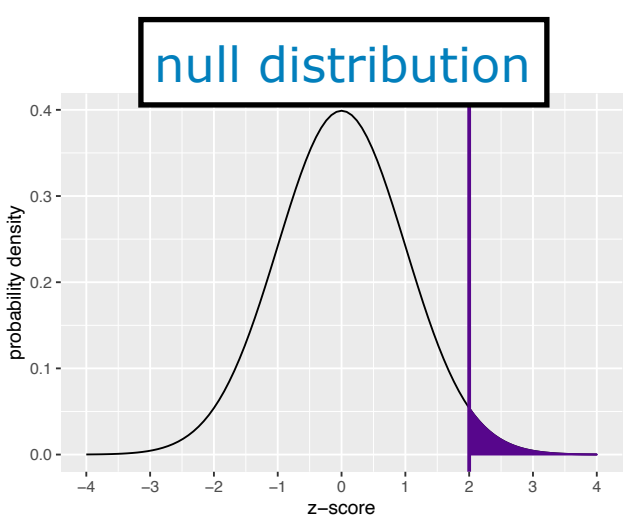
scores:

72 93 93 96 98
99 100 101 101 102
103 103 104 105
107 109 110 113
115 118 119 122
125 126 127

$\bar{x} = 106$

Then we'd look up our z in Table A1 in the book, or use pnorm() in R:

$z = 2$ $p = .023$



1.98	.4761	.0239
1.99	.4767	.0233
2.00	.4772	.0228
2.01	.4778	.0222

```
> pnorm(2)
[1] 0.9772499
> 1-pnorm(2)
[1] 0.02275013
>
```