P-values and Power

Summary of concepts so far

Null hypothesis: status quo, "nothing unusual is going on"

Test statistic: summarizes the evidence in the data; extreme values indicate evidence against the null; needs to be compared to a reference (null) distribution that describes its behavior under the null

p-value probability of observing a value of the test statistics that is at least as extreme as the observed value

We **reject the null hypothesis** when the p-value is small, typically smaller than 5%

Learning about the p-value

In the example for testing the null hypothesis of independence in a contingency table, suppose `marital_status` and `sex` really were independent and our marginal frequencies were good estimates.

Let's simulate some tables under independence and calculate p-values for those.

Aside: How would we do this (more in lab next Tuesday)?

Getting the reference (null) distribution is key for assessing the test statistic - Aside

This is where some thinking is required. Recall the example of the contingency table that shows Marital Status and Sex for 1000 subjects. Here is the population version:

	Div/Sep	Married	Never Married	Widowed	
F	π_{11}	π_{12}	π_{13}	π_{14}	q_1
M	π_{21}	π_{22}	π_{23}	π_{24}	q_2
	p_1	p_2	p_3	p_4	1

To get the reference distribution of the test statistic, we need to simulate 1000 data from that table **under the null hypothesis**, then evaluate the test statistic on it.

How can one simulate this? First, the basics in R:

Sampling with replacement from a number of objects

Suppose I want to choose one of 3 (say) categories, where the 3 categories have probabilities 0.1, 0.3, 0.6:

```
p <- c(0.1, 0.3, 0.6)
sample(1:3,1,replace=T,p)
```

```
[1] 3
```

If I want to have a sample of size n, then I could use

```
n <- 1000
p <- c(0.1, 0.3, 0.6)
sample(1:3,n,replace=T,p)
```

So the output would be a vector of 1000 numbers, each of which is 1, 2 or 3, and each number is simulated with the given probability.

Sampling with replacement from a number of objects

Alternatively, I can simulate from the *multinomial distribution*:

```
n <- 1000
p <- c(0.1, 0.3, 0.6)
rmultinom(1,n,p)
```

```
     [,1]
[1,]   94
[2,]  314
[3,]  592
```

The output is a vector of length 3, corresponding to the number of times 1 has come up, number of times 2 has come up... So these three numbers sum to $n = 1000$.

This output in terms of counts is what we need for our contingency table.

Sampling a contingency table assuming indepedence

	Div/Sep	Married	Never Married	Widowed	
F	π_{11}	π_{12}	π_{13}	π_{14}	q_1
M	π_{21}	π_{22}	π_{23}	π_{24}	q_2
	p_1	p_2	p_3	p_4	1

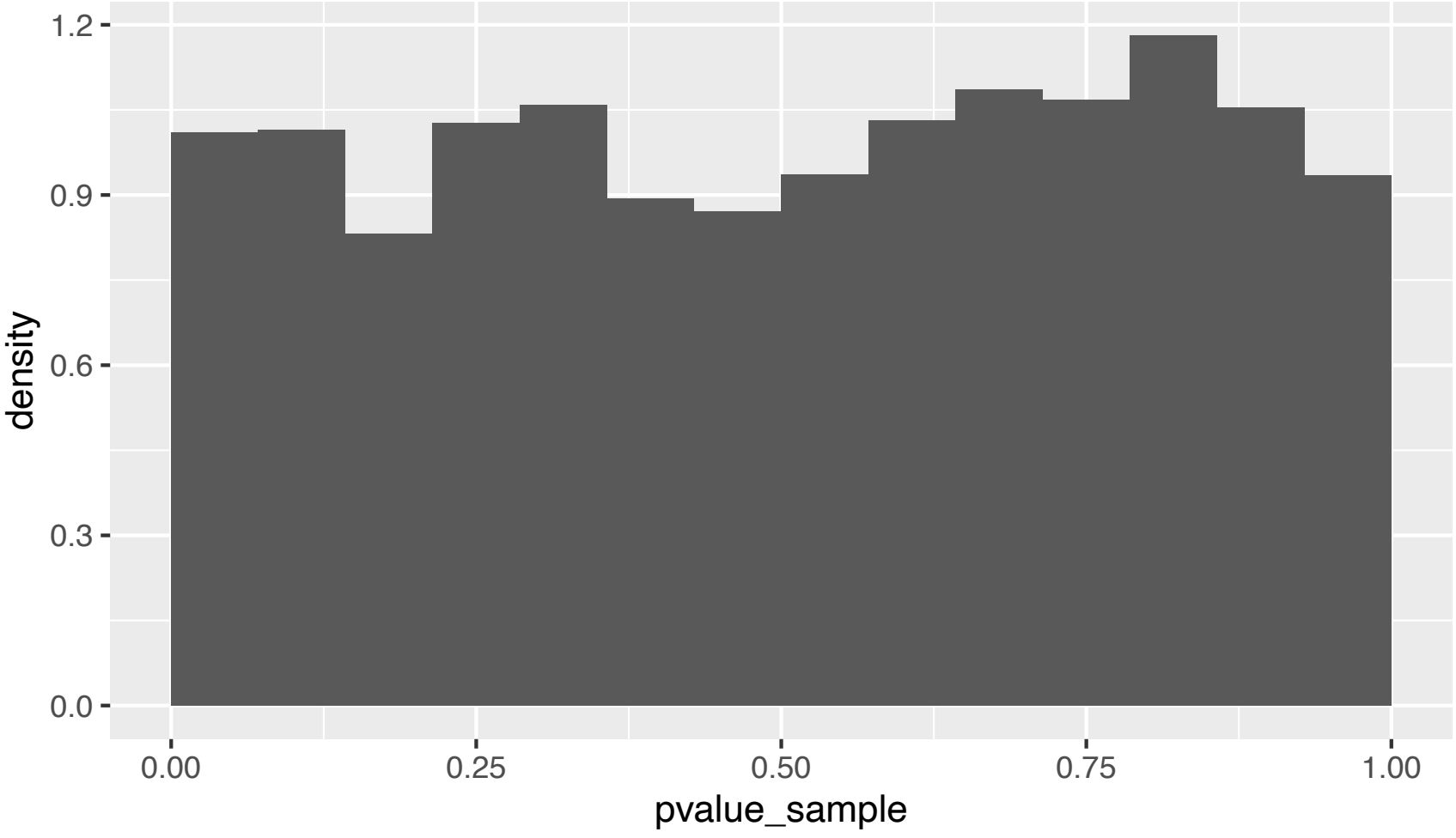
We have 8 categories, which we can identify with 1:8. (R lets you do that more elegantly, see the lab.). We have to sample these 8 categories using the 8 probabilities π_{jk}.

But we don't know the π_{jk}. Here is where the thinking is required:

- We could use the plug-in principle and estimate the π_{jk} from the data. But this does not guarantee that the resulting $\hat{\pi}_{jk}$ will obey the null hypothesis of independence.
- To enforce independence, we require $\pi_{jk} = p_j q_k$. So the idea is to estimate the marginal probabilities p_j and q_k from the data, then obtain the desired $\hat{\pi}_{jk}$ via $\hat{\pi}_{jk} = \hat{p}_j \hat{q}_k$, then use these $\hat{\pi}_{jk}$ to simulate the entries of the table.

Learning about the p-value

Let's calculate the p-value for all the generated tables under independence:



What is a *p-value*?

- The histogram shows that the p-values are roughly evenly distributed between 0 and 1. This is called the **uniform distribution** on [0, 1].
- This is true in general: Under the null hypothesis, p-values follow a uniform distribution on $[0, 1]$.
- This is an example of the general result referred to as the probability integral transform (you don't need to know this, just putting it here in case of interest).

What is a *p-value*?

- A random variable X has a **uniform distribution** on the interval $[0, 1]$ if $P[X \leq x] = x$.
- Specifically, this means that there is a 5% chance that the p-value is less than 5%, a 10% chance that it is less than 10%, etc.

What can we do with a *p-value*?

Suppose we decide that we reject the null hypothesis of independence if the *p-value* less than 0.05.

What do we know?

- Because of the uniform distribution, 5% of the time, even the null hypothesis is true, we reject it. For example, for the null hypothesis that martial status and gender are independent, 5% of the time we will say that there is strong evidence that they are not independent, even if the null hypothesis of independence is true.
- Thresholding the p-value at 5% controls the **type I** (rejecting the null when the null is true) **error rate** at 5%.
- In other words, suppose we test **many** hypotheses by thresholding the p-value at 5%, and in **every case**, the null was actually true, then in about 5 percent of those cases, we will **incorrectly reject the null**.

Recap of p-values and hypothesis testing

So far, we have discussed doing inference in the following way:

1. State a null hypothesis H_0.
2. Define a test statistic and compute it on the data.
3. Quantify evidence about the hypothesis by comparing to a reference distribution (the distribution of the test statistic given that the null hypothesis is true). We often do this by computing a p-value.

- We talked about **Type I errors**: rejecting the null hypothesis when the null hypothesis is true, and said that if we reject the null when the p-value $p < \alpha$, then we will have a type I error rate of α, which we call **level** of the test.
- But what happens when the null **isn't true**?

Recap of p-values and hypothesis testing

So far, we have discussed doing inference in the following way:

1. State a null hypothesis H_0.
2. Define a test statistic and compute it on the data.
3. Quantify evidence about the hypothesis by comparing to a reference distribution (the distribution of the test statistic given that the null hypothesis is true). We often do this by computing a p-value.

- We talked about **Type I errors**: rejecting the null hypothesis when the null hypothesis is true, and said that if we reject the null when the p-value $p < \alpha$, then we will have a type I error rate of α, which we call **level** of the test.
- But what happens when the null **isn't true**?

The alternative hypothesis H_1 (or H_a)

- Often, this is just the complement (or negation) of the null (or two-sided test). For example, when we were testing for independence of X and Y, the alternative was just that X and Y were **dependent**. In the regression setting, we were just testing $\beta_1 = 0$ vs. $\beta_1 \neq 0$.
- Sometimes the alternative is more specific. For example in a one-sided test, we might want to test $\mu = 0$ vs. $\mu > 0$.

The alternative hypothesis H_1 (or H_a)

- We don't want to reject the null when the null is true. However, we would like to reject the null when the alternative is true as often as possible.
- Recall that failing to reject the null when the alternative is true is called a **Type II error**: false negative.
- The **power** of a test is the probability of rejecting the null hypothesis when it is false

$$\text{Power} = 1 - \mathbb{P}(\text{ type II error})$$

Putting all the terms together

We need to decide between H_0 and H_1 and there are two ways in which we can go wrong

	H_0 is true	H_1 is true
Reject H_0	Type I error	-
Accept H_0	-	Type II error

- A statistical test is a rule depending on the data that makes us decide for H_0 or H_1
- The probability of type I error is called **level** of the test
- 1−P(type II error) is called **power** of the test

Example

Suppose x is normally distribution with mean μ and variance 1,

$$x \sim \mathcal{N}(\mu, 1)$$

We will visualize the power of the one-sided test

$$H_0 : \mu = 0 \qquad \text{vs} \qquad H_1 : \mu > 0$$

at fixed level $\alpha = 0.05$ for different values of μ.

Let's construct a test

We have already seem a couple of test statistics, for example the z-statistic or the chi-square test (in contingency table setting).

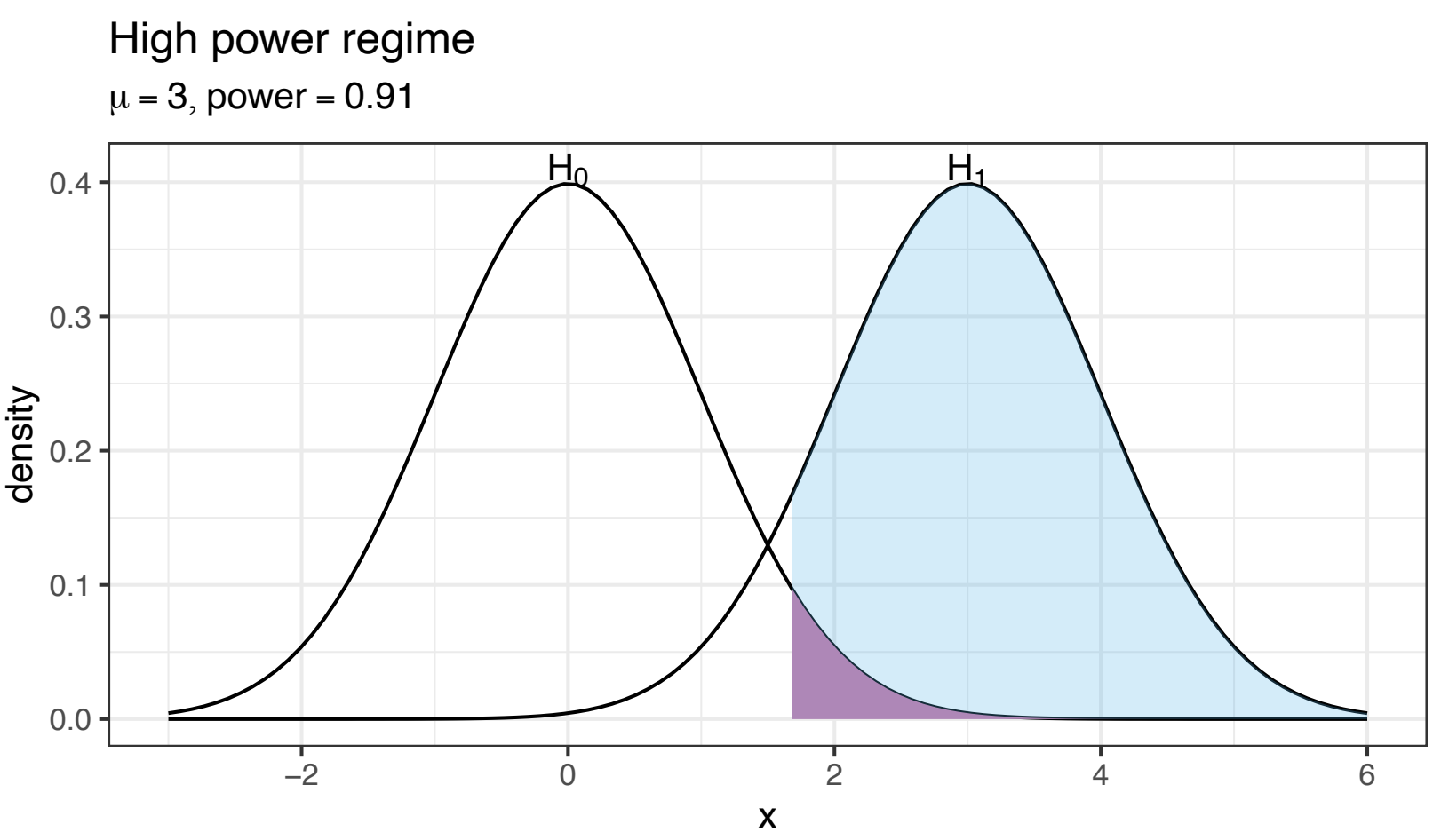
In our setting: For simplicity, let's use X itself as a test statistic. We have a variance of 1, so we have a "simple" problem.

Which values of X "contradict' ' the null hypothesis?

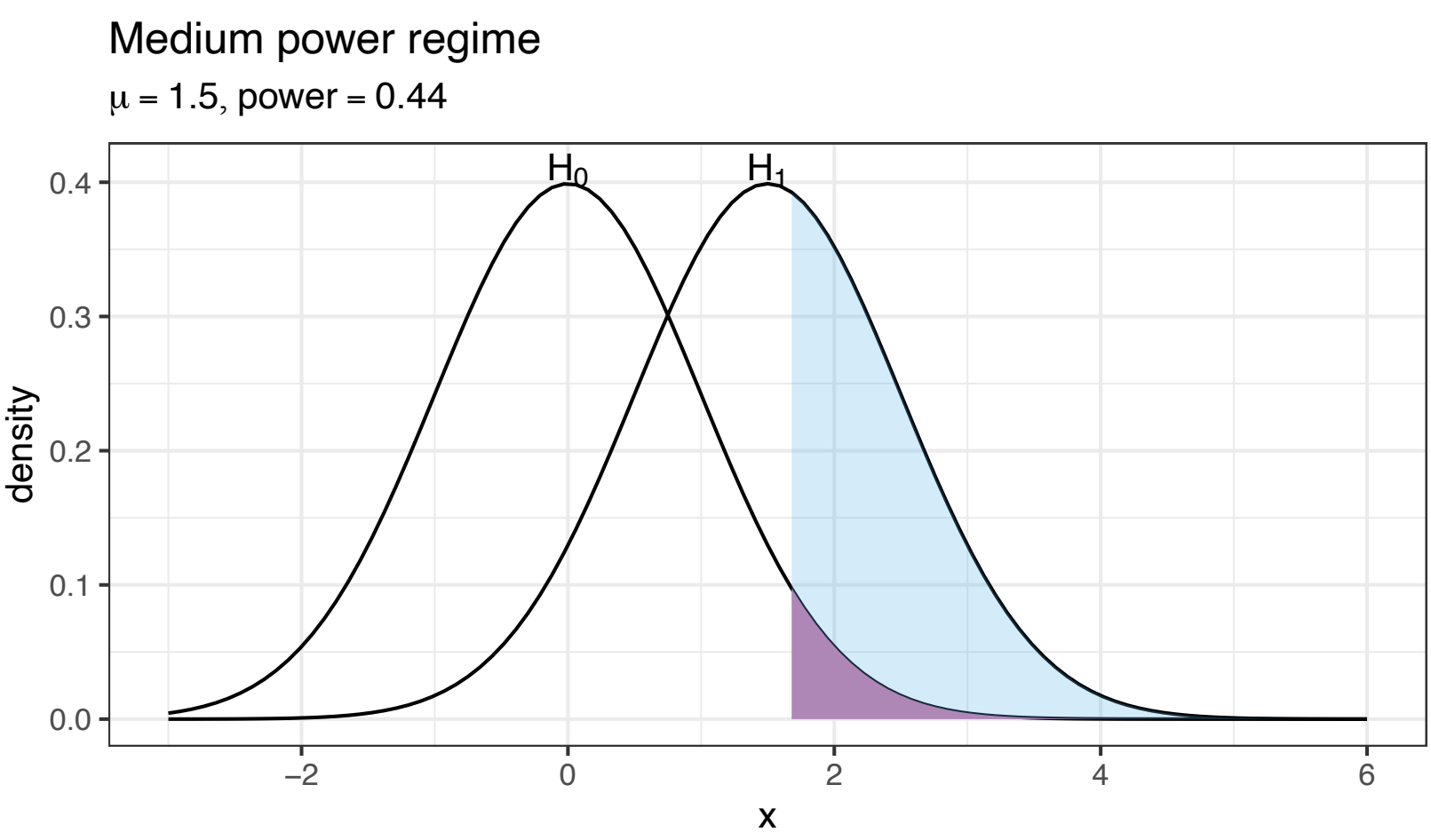
- If we observe a large X, then it seems that the null becomes "unlikely". The larger X is, the more unlikely the null seems.
- $\mathbb{P}(X \geq x_{\text{obs}} | \mu = 0)$ is the p-value (i.e. the probability of observing a value at least as large as the observed one, under the null hypothesis (i.e. $\mu = 0$)).

Let's construct a test

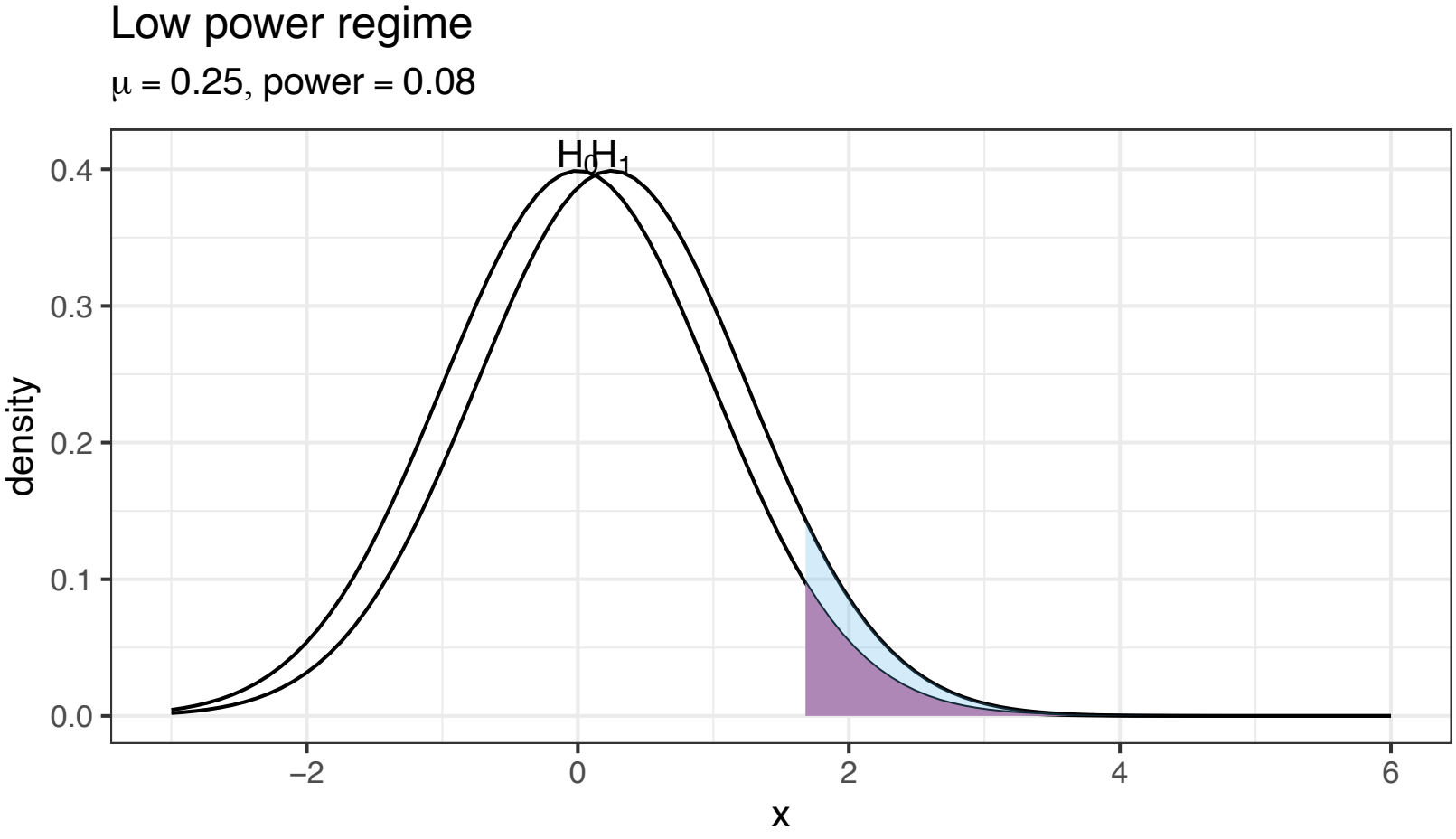
- Decision rule: reject H_0 when p-value is smaller than 0.05.
- Equivalent decision rule: reject H_0 when $x_{\text{obs}} > 1.65$.
- Where does 1.65 come from? This is the value of a Normal distribution with mean 0 and variance 1 (i.e. what we have under the null) such that the probability of observing a value at least as large as 1.65 is 5%.



Power is 0.91 and $\mathbb{P}(\text{type II error}) = \mathbb{P}(\text{false negative}) = 0.09$



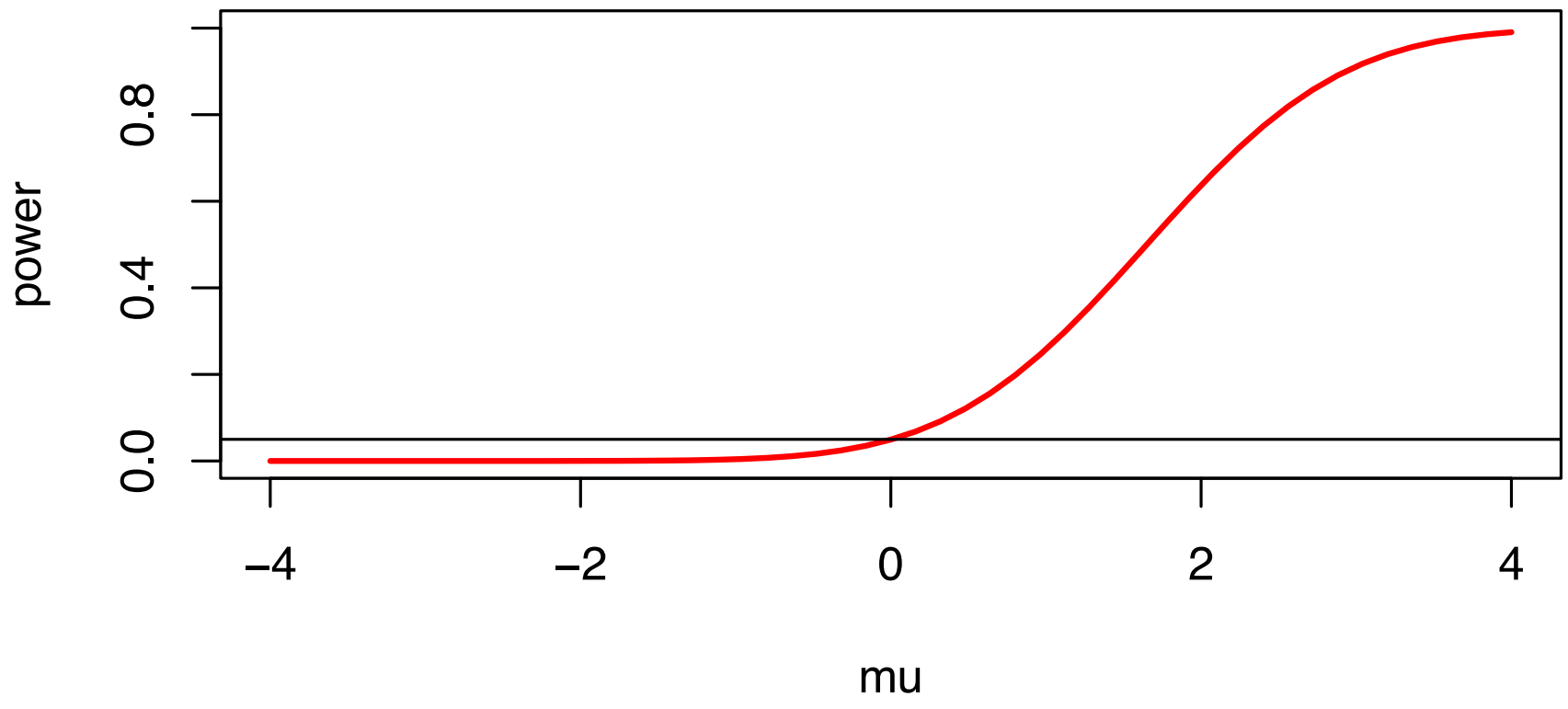
Power is 0.44 and $\mathbb{P}(\text{type II error}) = \mathbb{P}(\text{false negative}) = 0.56$



Power is 0.08 and $\mathbb{P}(\text{type 2 error}) = \mathbb{P}(\text{false negative}) = 0.92$

Power for one sided test of level 0.05

```
mu = seq(-4, 4, length=51)
power = 1 - pnorm(1.65, mu, 1)
plot(mu, power, type='l', lwd=2, col='red', ylim=c(0,1))
abline(h=0.05)
```

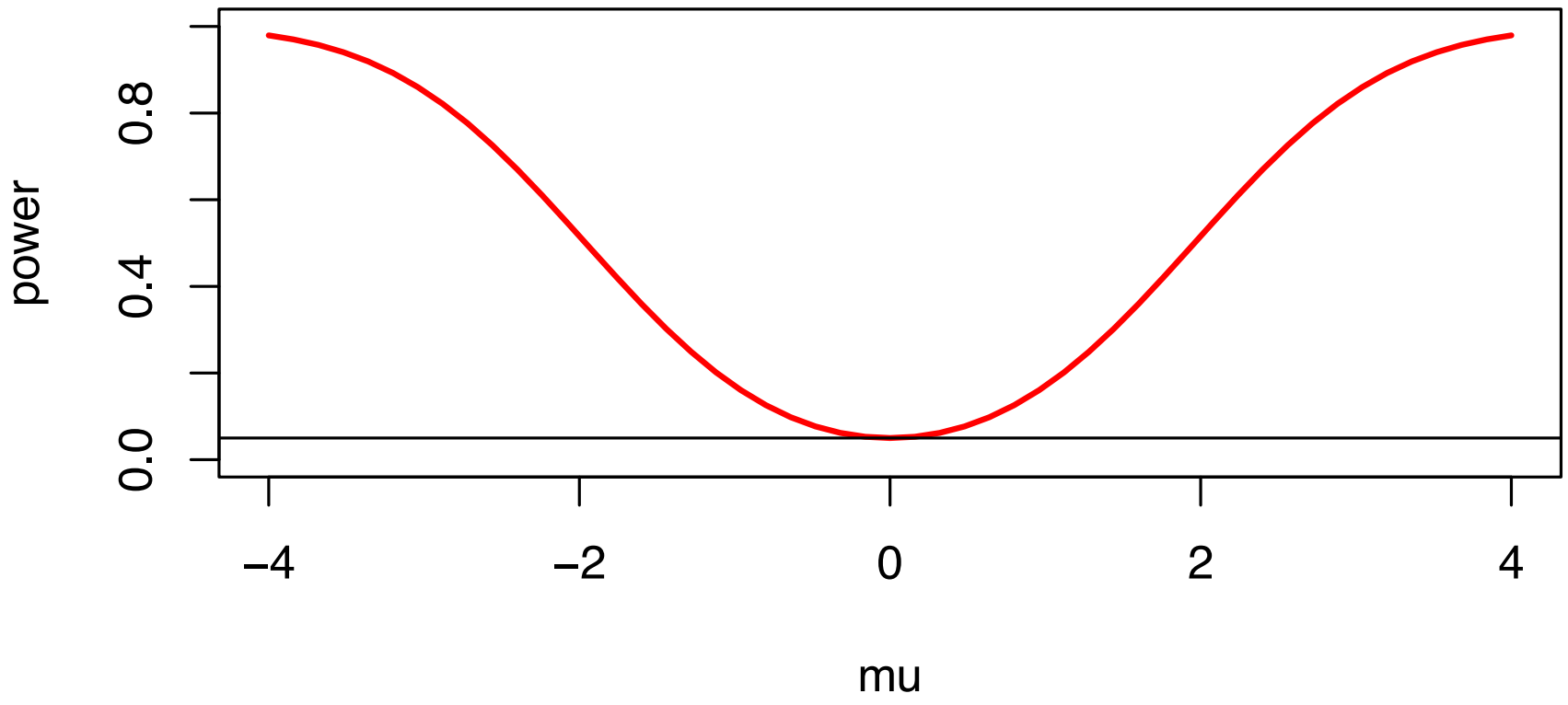


Recap: Two-sided tests

- When testing a null hypothesis about a single parameter like the mean μ, e.g. $H_0 : \mu = 0$, it is usually appropriate to define the alternative hypothesis H_1 to be *two-sided*.
- In other words, a two-sided H_1 is defined as $\mu > 0$ OR $\mu < 0$.
- For example, when testing the effectiveness of a new drug B vs an existing drug A, μ could be defined as the response rate for drug B minus the response rate for drug A.
- Then $\mu > 0$ means drug B is better than drug A, and $\mu < 0$ means drug A is better than drug B. Before we run an experiment, either outcome is a possibility.

Power for two sided test of level 0.05

```
power = pnorm(-1.96, mu, 1) + 1 - pnorm(1.96, mu, 1)
plot(mu, power, type='l', lwd=2, col='red', ylim=c(0,1))
abline(h=0.05)
```



Notes

Our observations make sense:

- We have trouble distinguishing the null from the alternative when they correspond to very similar values of μ.
- When the corresponding values are very different, then it is easier to distinguish the null and the alternative, because the data we see tend to look very different from what we would expect under the null hypothesis.