Business Statistics 41000

Hypothesis Testing

Mladen Kolar

Homeless guys don't wear nice shoes

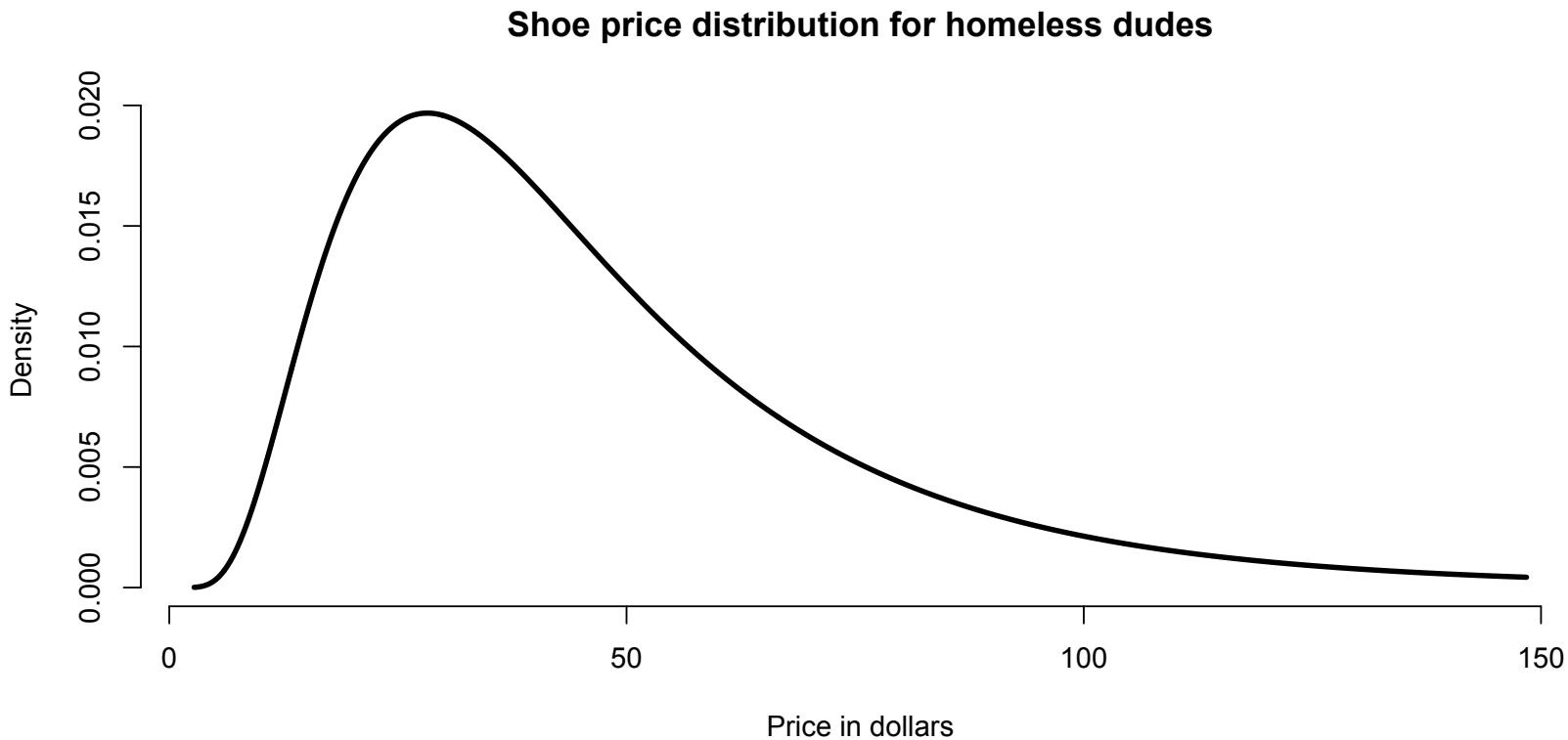
The guy asking for your change outside Alinea is wearing a posh pair of loafers. Would you be willing to conclude that he's not actually homeless on the basis of this evidence?

To make things numerical, assume we recognize the shoes and know for a fact they cost \$285, or 5.65 log dollars.

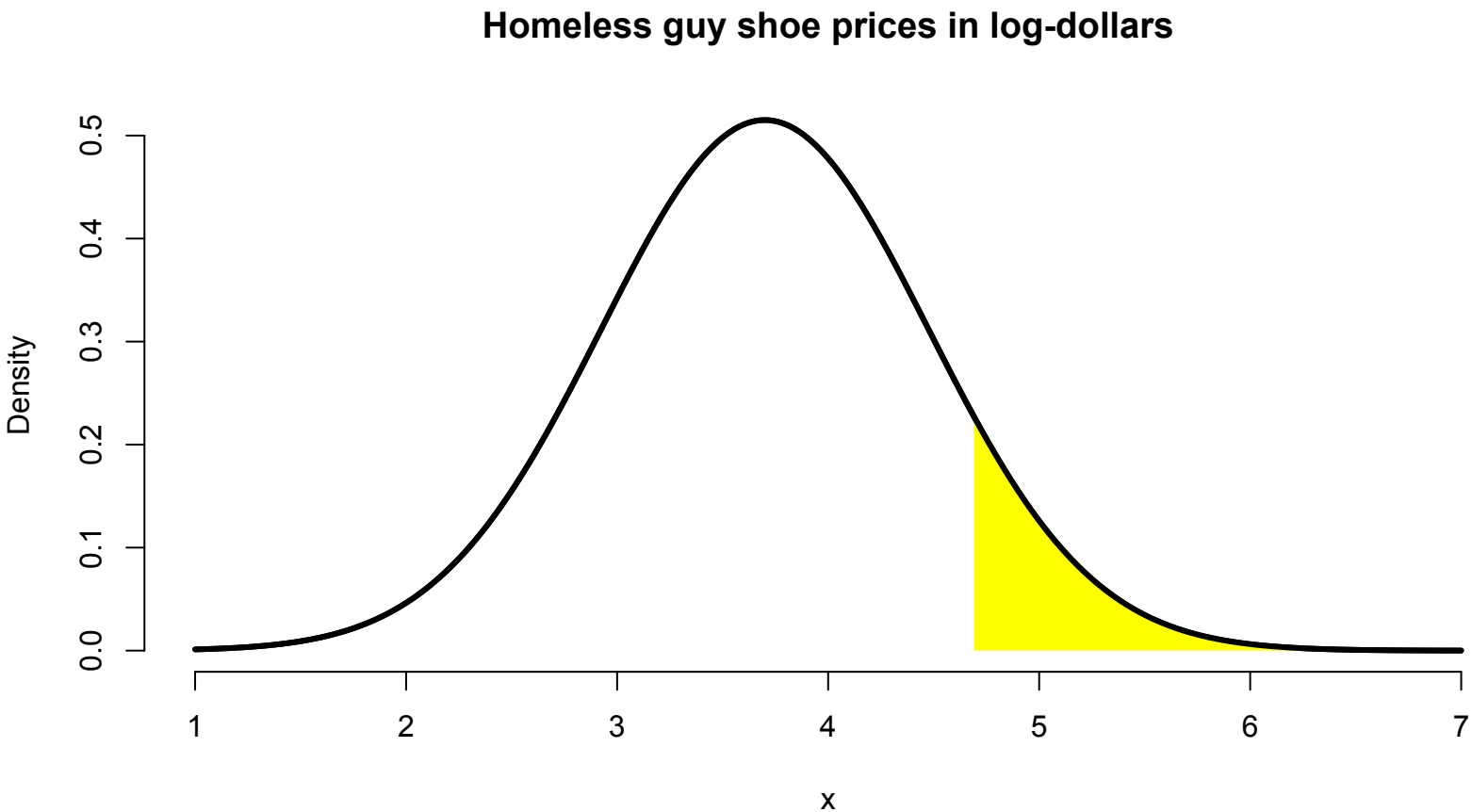
Also assume that the distribution of log-prices of homeless guys' shoes is described by $X \sim N(3.7, 0.6^2)$. Then we find $P(X > 4.69) = 0.05$ (using `qnorm(0.05,3.7,0.6,lower.tail = FALSE)`).

So, if we call out all supposed homeless guys with shoes worth more than $\exp(4.69) = \$108$ we'll only do so incorrectly 5% of the time.

Homeless guys don't wear nice shoes



Homeless guys don't wear nice shoes



Homeless guys don't wear nice shoes

To turn this problem into a **hypothesis testing problem**, we must phrase the question in terms of probability distributions and their parameters.

Assume the data we observe - the shoe price - was a draw from a normal probability distribution with an unknown mean, μ, and a known variance (for now), $\sigma^2 = 0.6^2$.

If the guy were homeless, then $\mu = 3.7$. So we want to test the hypothesis

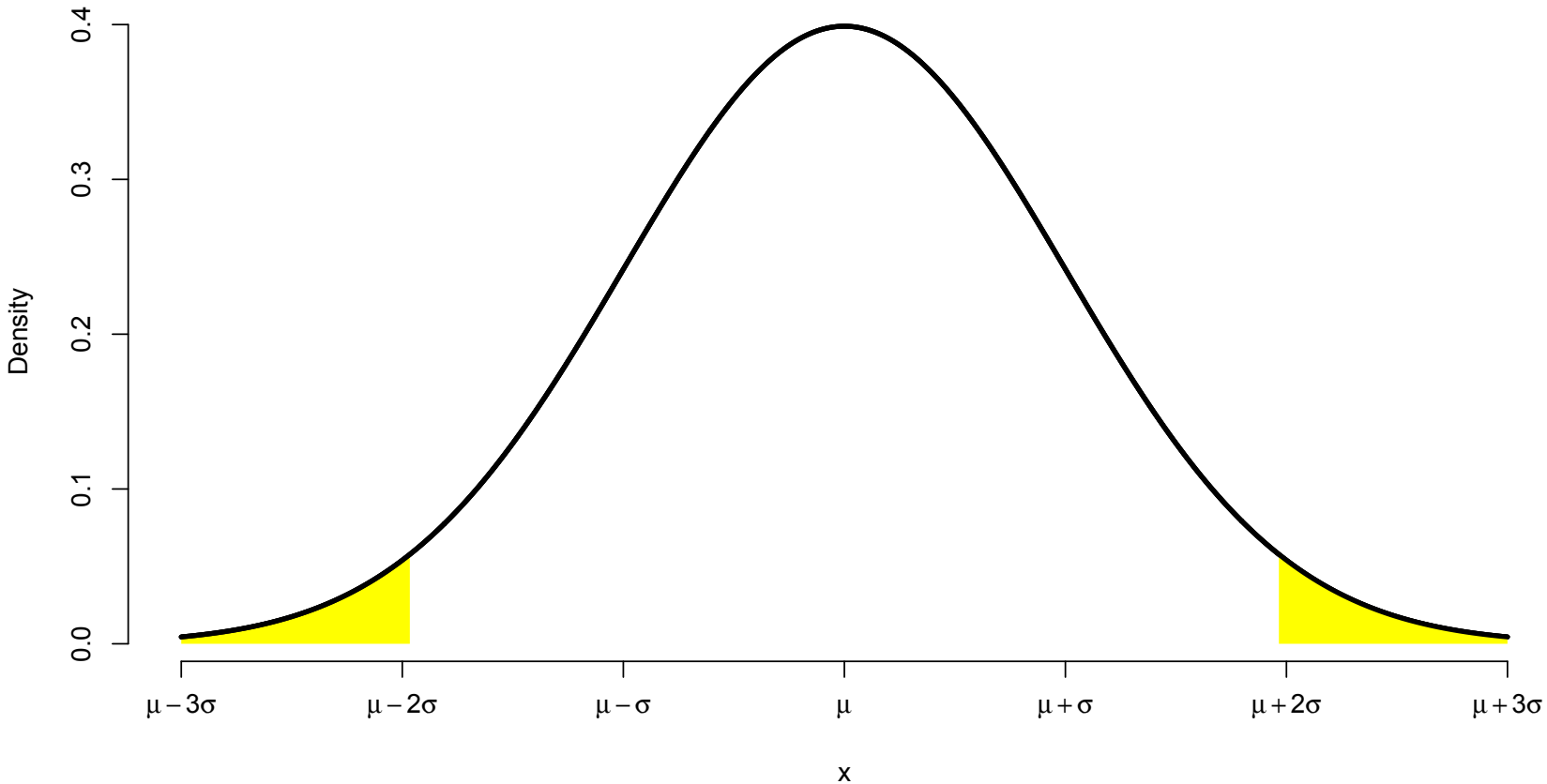
$$H_0 : \mu = \mu_0$$

where $\mu_0 = 3.7$.

Logic of hypothesis tests

Consider a normal random variable

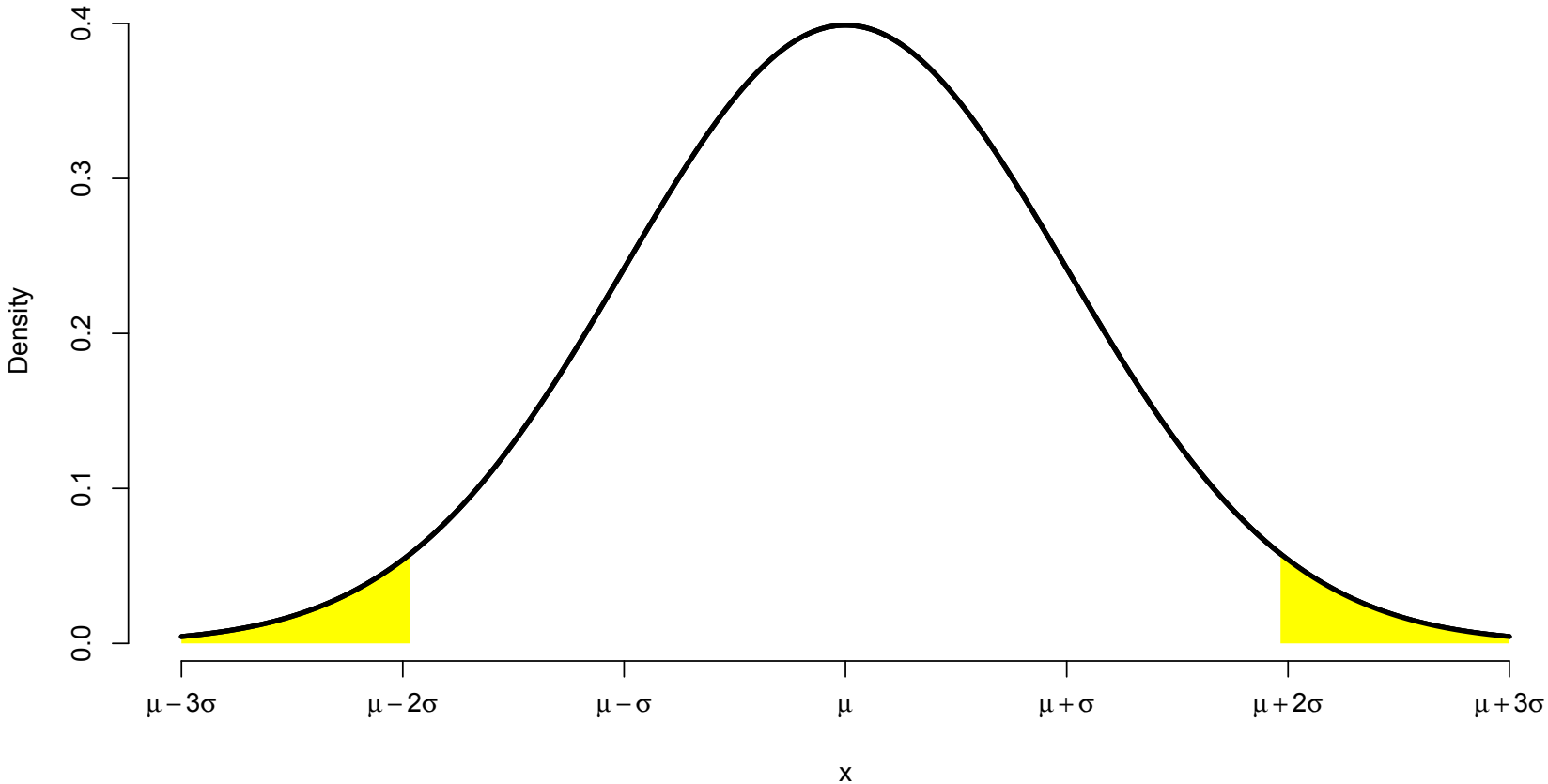
$$X \sim N(\mu, \sigma^2).$$



Logic of hypothesis tests

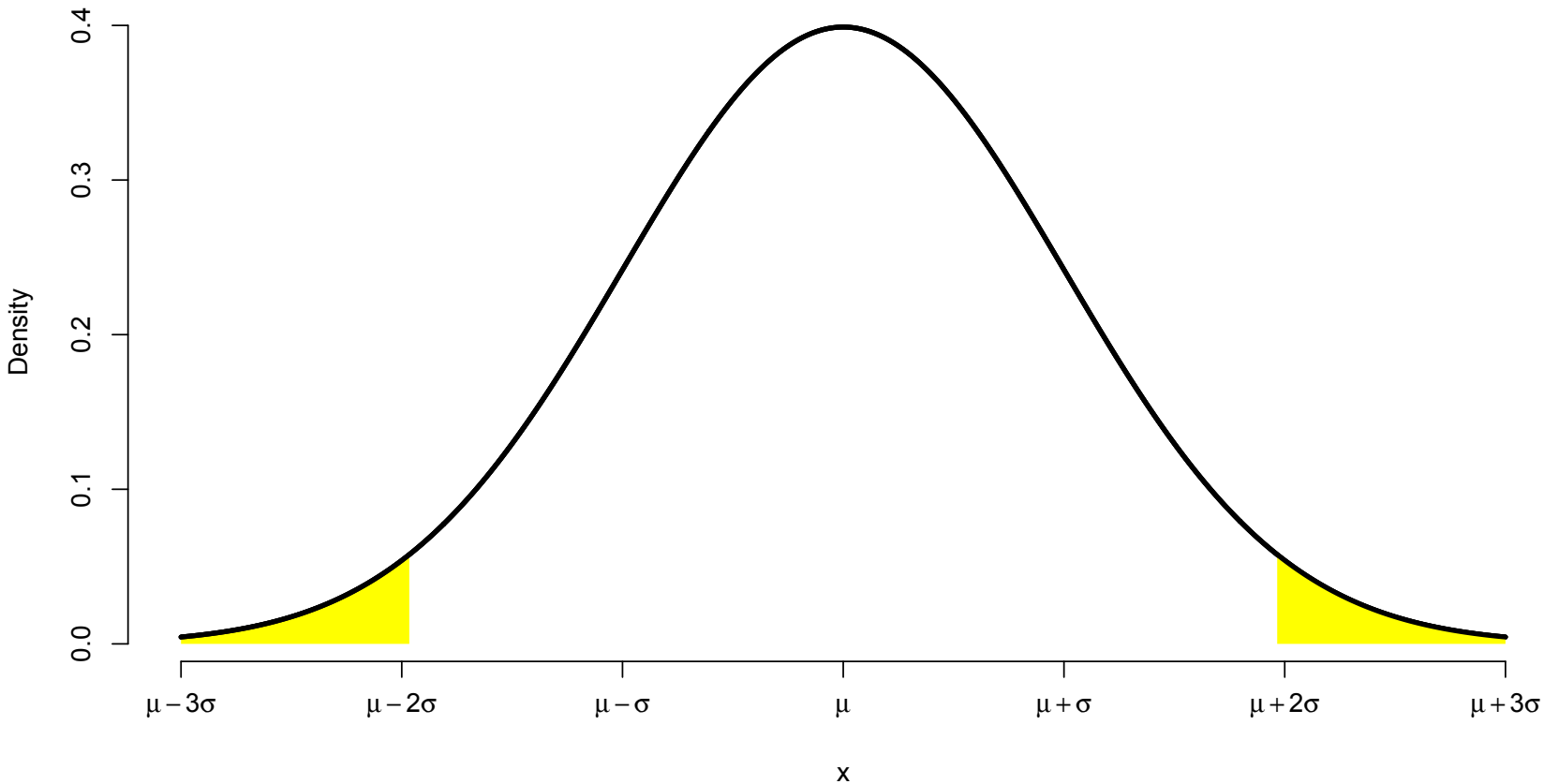
Imagine observing a single draw of this random variable, call it x.

Assume the variance σ^2 is known, but the mean parameter μ is not.



Logic of hypothesis tests

Intuitively, this single observed value x tells us something about the unknown parameters: more often than not, the observed value will tend to be *close* to the parameter value.



Then again, sometimes it will not be. But it will only *rarely* be too far off.

Logic of hypothesis tests

Assume we have a guess in mind for our true parameter value. We denote this guess by μ_0, pronounced "mew-naught".

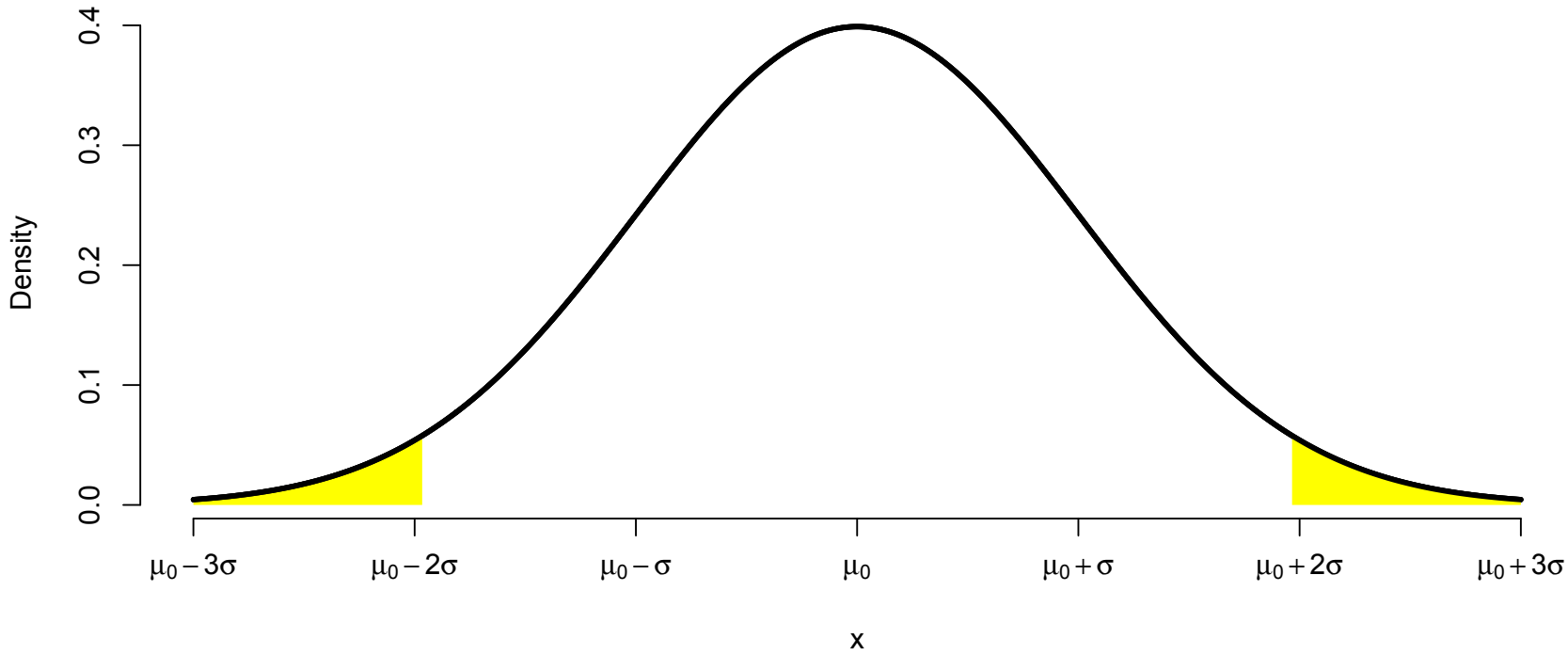
We refer to this as the "null hypothesis", which we write:

$$H_0 : \mu = \mu_0.$$

The symbol μ is the "true value" and μ_0 is the "hypothesized value".

Logic of hypothesis tests

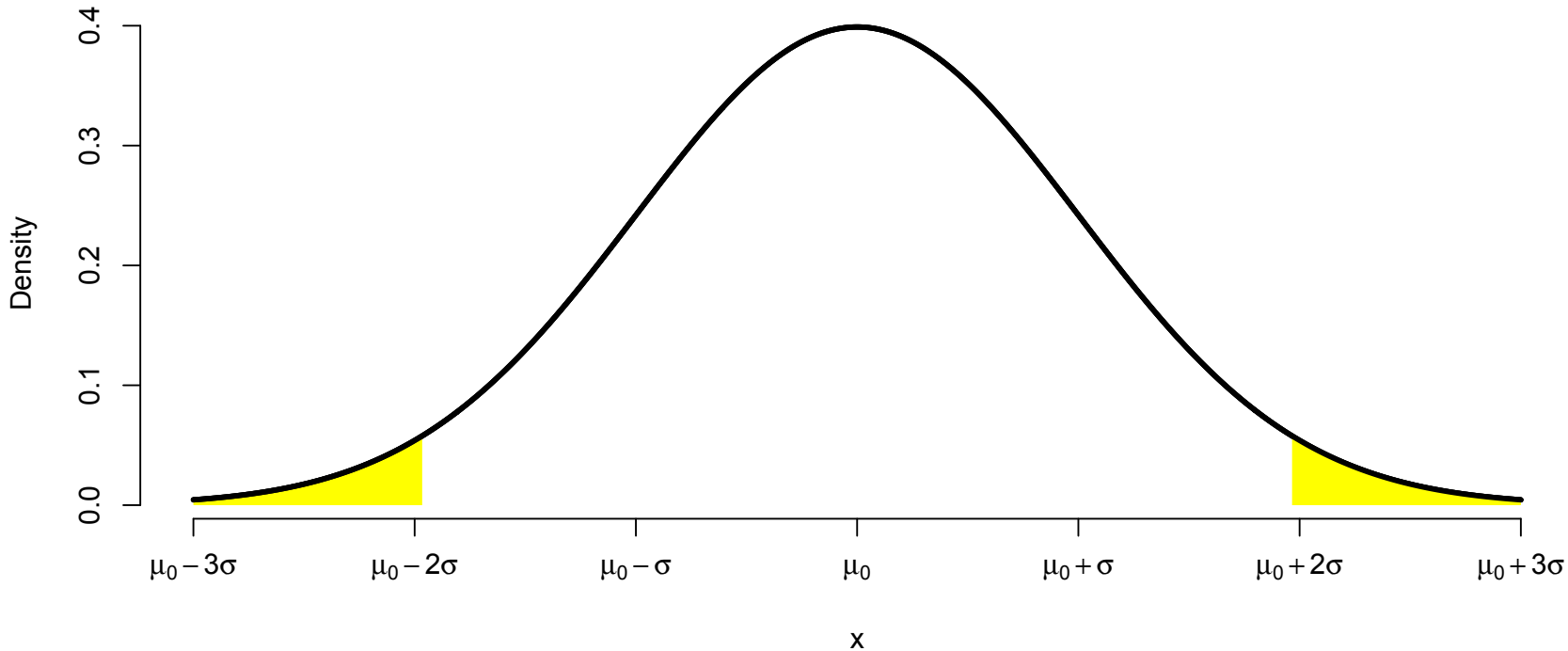
Hypothesis testing asks the following question: if the true value were μ_0, is my data in an unlikely region?



If we consider it too unlikely, we decide not to believe our hypothesis and we "reject the null hypothesis".

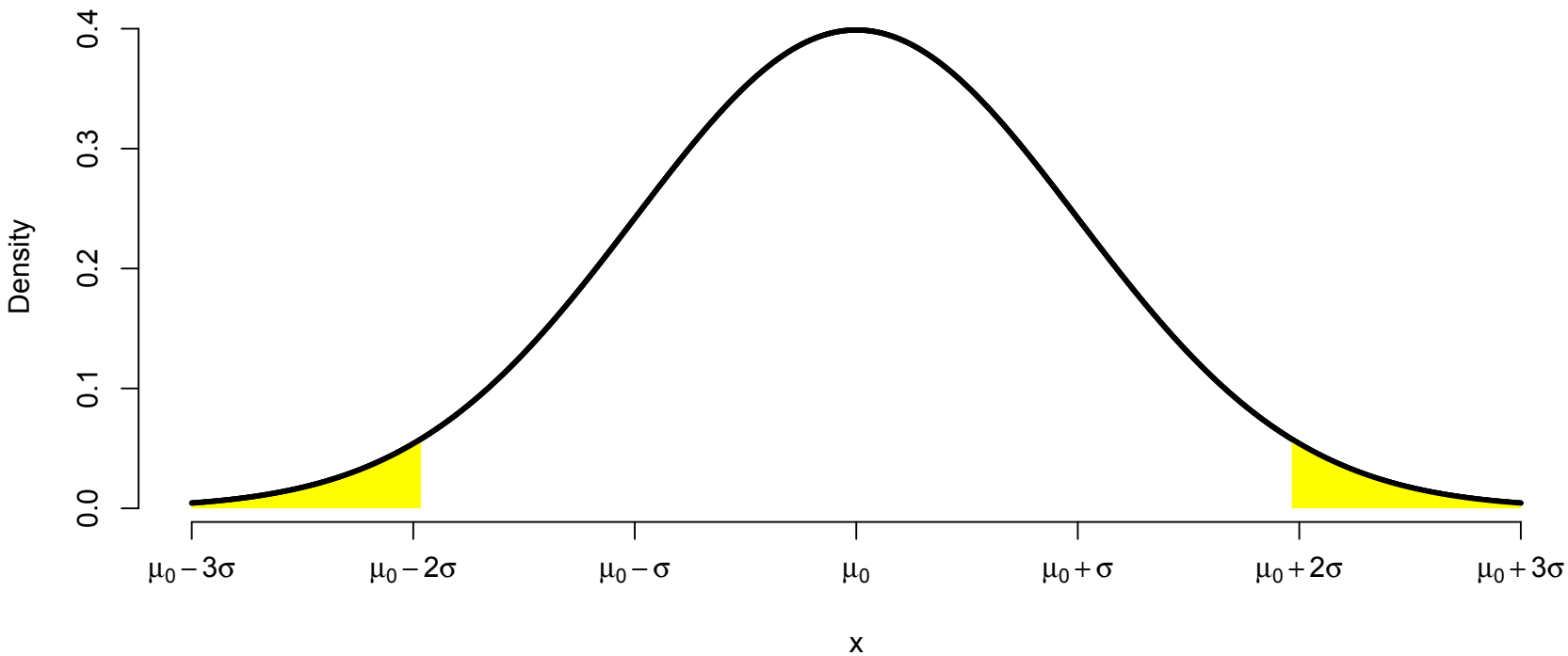
Logic of hypothesis tests

On the other hand, if the data falls in a likely region, we decide our hypothesis was plausible and we "fail to reject" the null hypothesis.



Level of the test

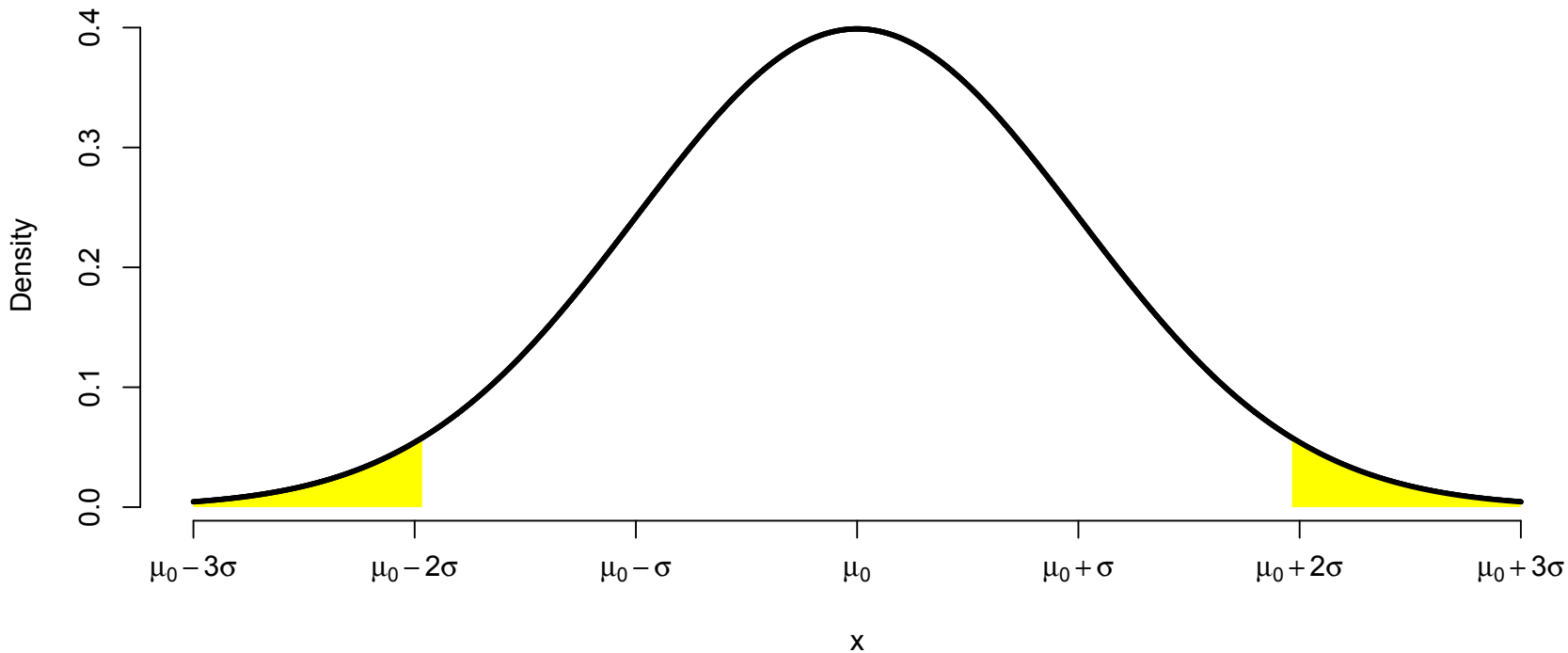
Where do we put the rejection region? In general, it depends on the problem (more on that in a minute).



But one thing is always true: the probability of the rejection region (the area under the curve) dictates how often we will **falsely reject** the null hypothesis. This is called the **level** of the test.

Level of the test

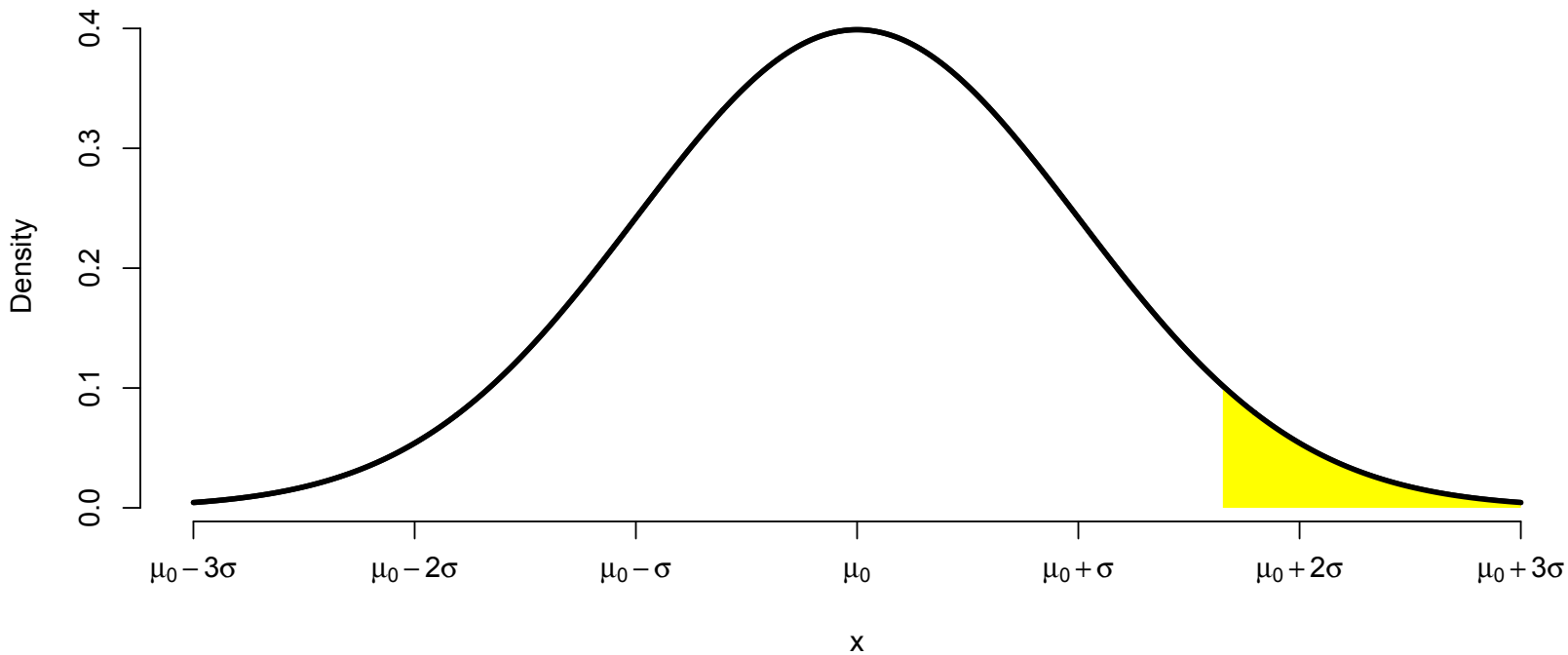
Because when the null hypothesis is true, we still end up in unusual areas sometimes. How often this happens is exactly the level of the test.



Where to put the rejection region

One way to think about rejection regions is in terms of alternative hypotheses, such as

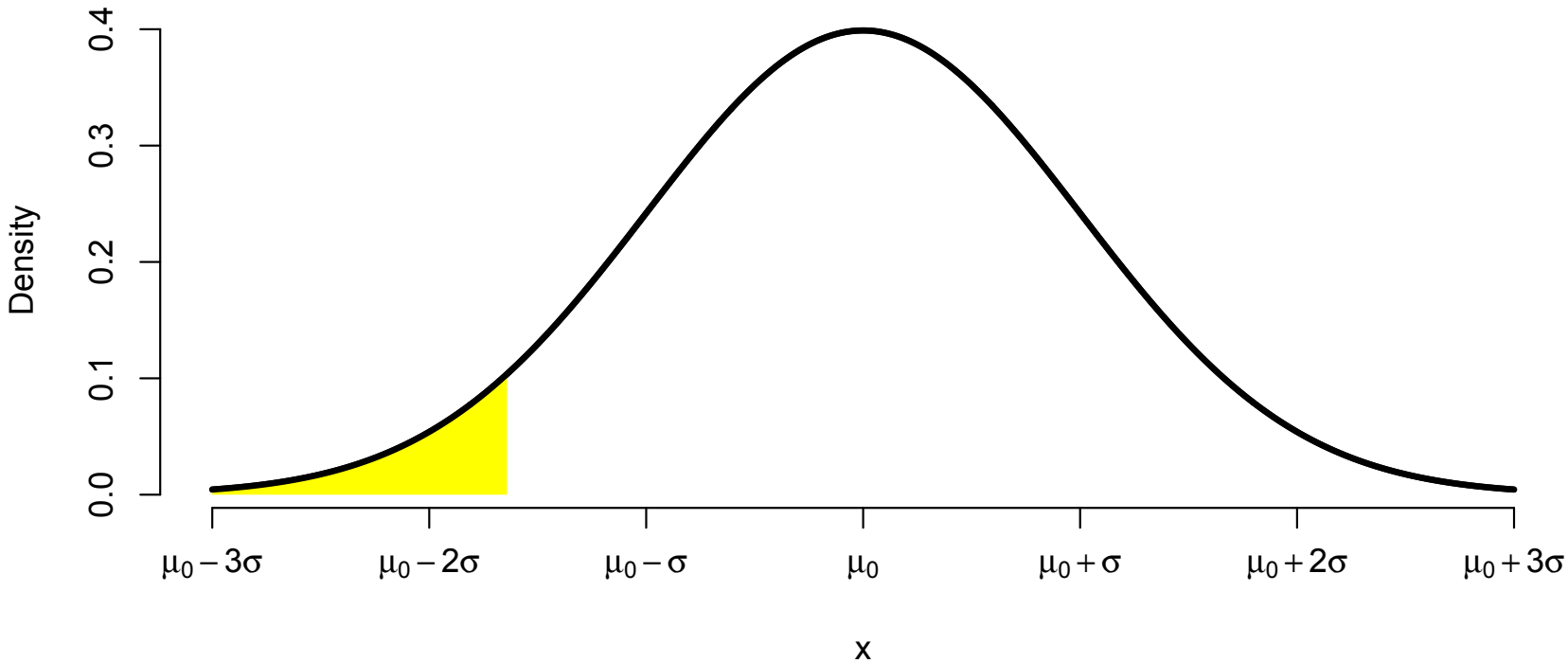
$$H_A : \mu > \mu_0.$$



I prefer to think of it the other way around: where we place our rejection region dictates what the alternative hypothesis is, because it determines what counts as *unusual*.

Where to put the rejection region

For $H_A : \mu < \mu_0$ the rejection region is on the other side.



In all of the pictures so far, the level of the test has been $\alpha = 0.05$. There is nothing special about that number.

More than one observation

To apply this logic to more than one data point, we simply collapse our data into a single number, or statistic, and figure out the **sampling distribution** of this statistic. Then we proceed as before.

In this lecture we will use sample means as our **test statistic**.

Recall from lecture 6 that if we have n i.i.d. samples from $N(\mu, \sigma^2)$ then $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is distributed as

$$\bar{x} \sim N(\mu, \sigma^2/n)$$

Did something change?

You have implemented a new incentive policy with your sales force and you want to measure if the new policy is translating to increased sales.

Previously sales hover around \$50,000 a week, with a standard deviation of \$6,000. The first five weeks have produced the following sales figures (in thousands of dollars) of

$$[61, 52, 48, 43, 65].$$

Do you reject the null hypothesis that nothing has changed?

Did something change?

Our test statistic is $\bar{x}$. Under the null hypothesis

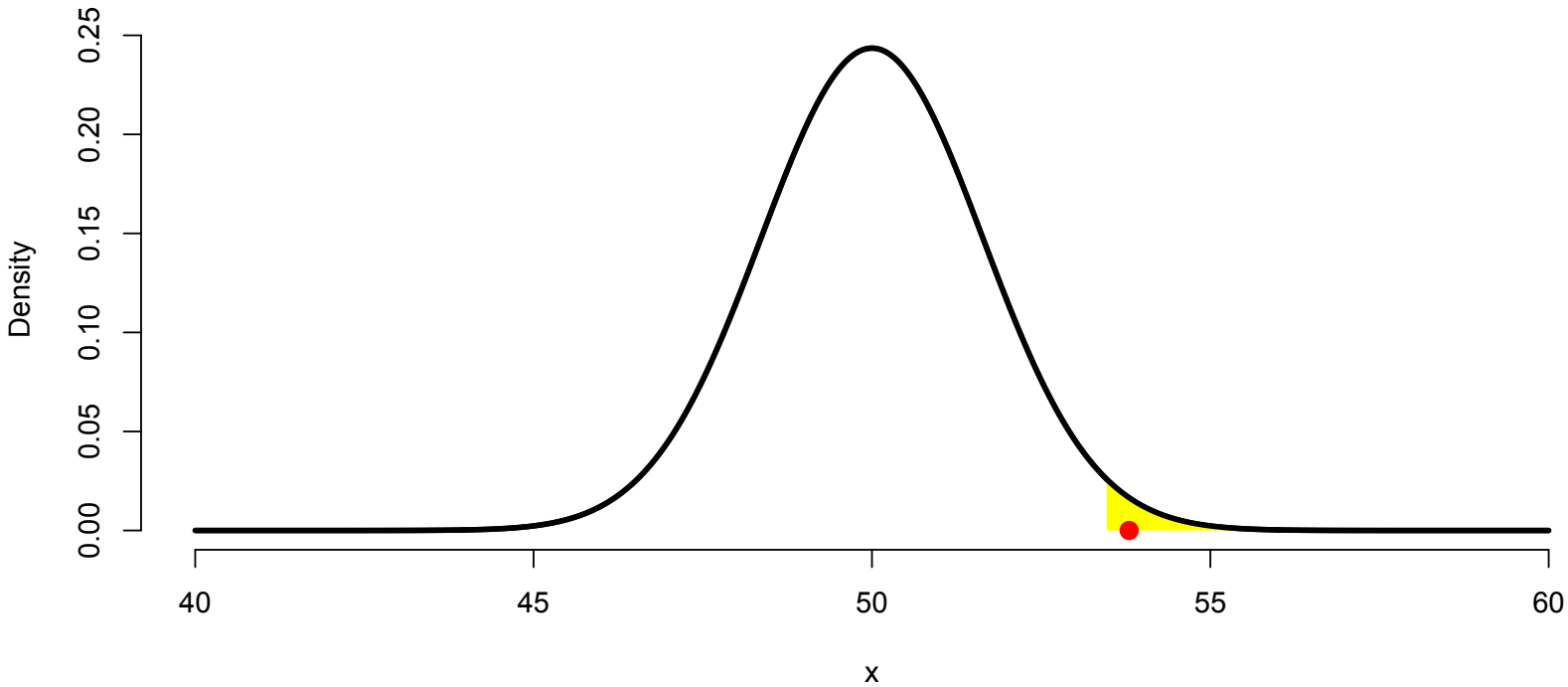
$$\bar{x} \sim N(50, 6^2/5).$$

We observe $\bar{x} = 53.8$

We want our "unusual" region to be unusually high sales. At a level of 10% the rejection region starts at 53.44, so we reject.

At a level of 5%, the rejection region starts at 54.4, so we fail to reject.

Did something change?



The empirical or sample mean falls in the 10% rejection region (but not the 5% rejection region).

p-values

The largest level at which we would reject our observed value is called the p-value of the data.

In other words, the p-value is the probability of seeing data as, or more, extreme than the data actually observed.

So the p-value will change depending on the shape of the rejection region.

So a p-value larger than the level of a test, implies that you fail to reject.
A p-value smaller than the level of a test implies you reject.

Application to a proportion

We ask $n = 50$ cola drinkers if they prefer Coke to Pepsi; 28 say they do. Can we reject the null hypothesis that the two brands have evenly split the local market?

We can approach this problem using a normal approximation to the binomial. Under the null distribution, the proportion of Coke drinkers has an approximate

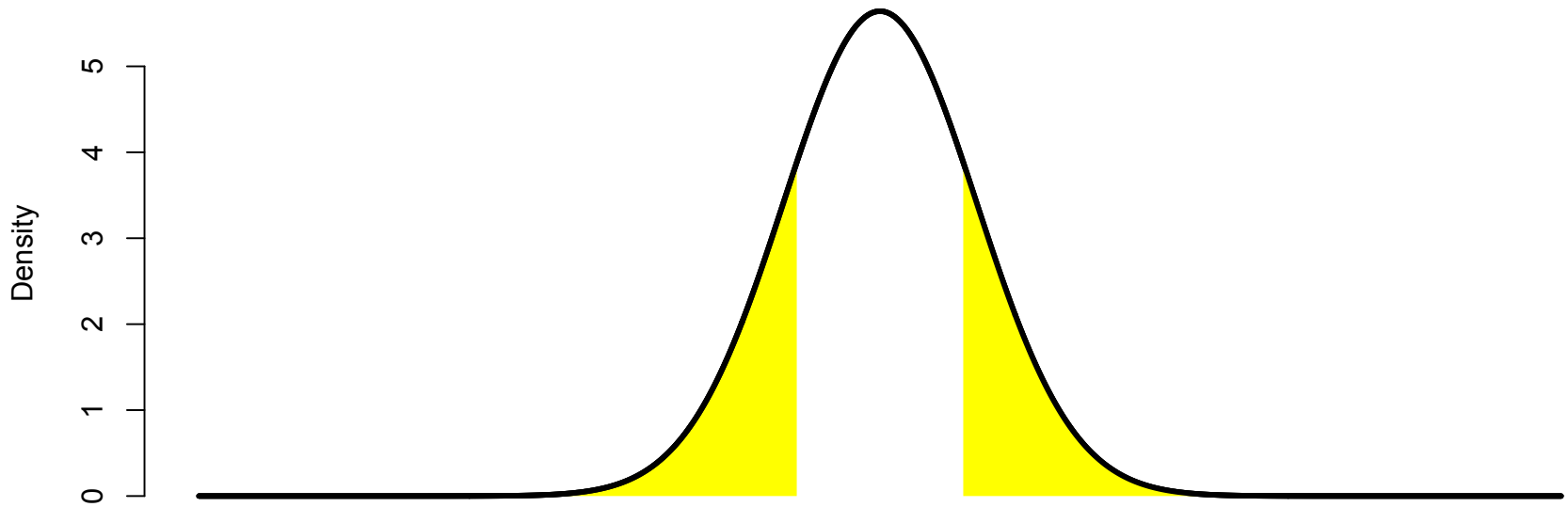
$$N(0.5, 0.5^2/50)$$

distribution.

We observe $\hat{p} = 28/50 = 0.56$. The p-value is the area under the curve less than 0.44 and greater than 0.56.

Coke vs Pepsi

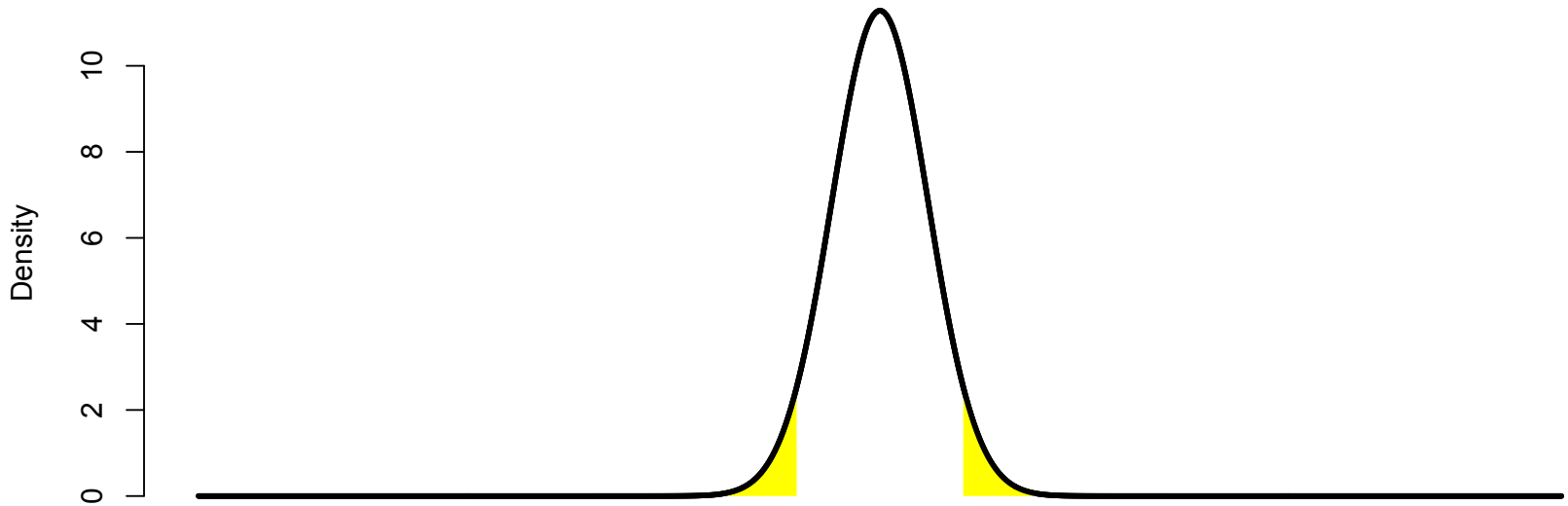
The p-value is an uncompelling 40%.



What would happen if we had the same observed proportion of Coke drinkers, but a sample size of 200?

Coke vs Pepsi

At $n = 200$ we have reduced our standard deviation by a factor of 2.



Our p-value drops to 0.09.

Variance unknown

So far we have been considering normal hypothesis tests when the variance σ^2 is known. Very often it is unknown.

But if we have a sample of reasonable size (say, more than 30), then we can use a plug-in estimate without much inaccuracy.

That is, we use the empirical standard deviation (the sample standard deviation) as if it were our known standard deviation: we treat s^2 as if it were σ^2.

Are mountain people taller?

It is claimed that individuals from the Eastern mountains are much taller, on average, than city dwellers who are known to have an average height of 67 inches.

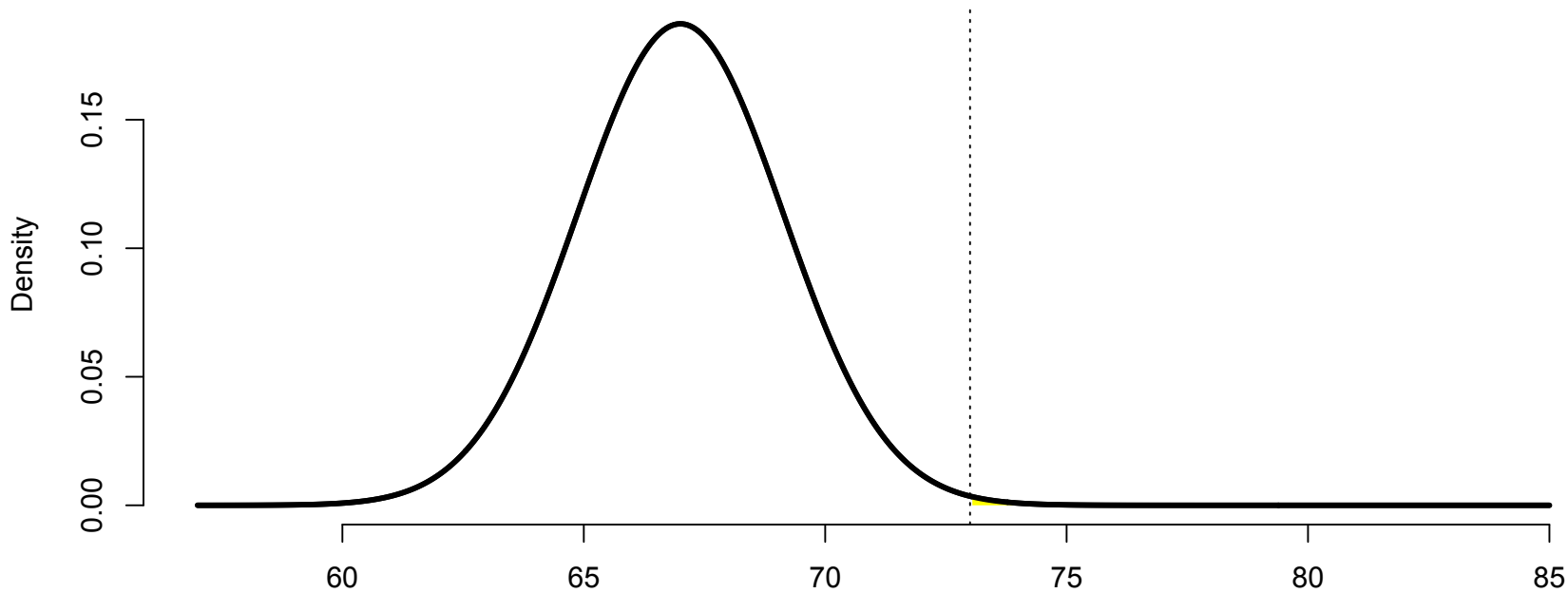
Of course, some mountain people are hobbits, so obviously there is a lot of variability.

Based on a sample of 35 mountain people we measured, we find $\bar{x} = 73$ and $s = 12.6$.

Can we reject the null hypothesis that there is no difference in height?

Are mountain people taller?

We assume that our test statistic is distributed $\bar{x} \sim N(67, 12.6^2/35)$ under the null distribution.



Our p-value is 0.00242.

z-scores

In normal hypothesis tests where the rejection region is in the tail, we're essentially measure the distance of our observed measurement from the mean under the null distribution. "How far is too far" is determined by the level of our test and by the standard deviation under the null.

To get a sense of how far into the tail an observation is, we can standardize our observation.

If $X \sim N(\mu, \sigma^2)$, then $\frac{X-\mu}{\sigma} \sim N(0, 1)$. Applying this idea to a normal test statistic tells us how many standard deviations away from the mean our observed value is.

In this last example we would get $z = \frac{\bar{x}-67}{12.6/\sqrt{35}} = 2.82$.

Difference of two means

A common use of hypothesis testing is to compare the means between two groups based on observed data from each group.

For example, we may want to compare a drug to a placebo pill in terms of how much it reduces a patient's weight. In this case we have

$$X_i \sim N(\mu_x, \sigma_x^2)$$

and

$$Y_j \sim N(\mu_y, \sigma_y^2)$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, m$.

Our test statistic in this case will be $\bar{x} - \bar{y}$, the difference in the observed sample means.

Better than a placebo?

Our test is

$$\begin{aligned} H_0 &: \mu_x - \mu_y = 0, \\ HA &: \mu_x > \mu_y, \end{aligned}$$

which defines a rejection region in the right tail.

The test statistic has null distribution of

$$\bar{x} - \bar{y} = \bar{D} \sim N(0, \sigma_x^2/n + \sigma_y^2/m)$$

which we approximate as

$$N(0, s_x^2/n + s_y^2/m).$$

Better than a placebo?

We observe that 34 patients receiving treatment have a mean reduction in weight of 5 pounds with standard deviation of 4 pounds. The 60 patients in the placebo group show a mean reduction in weight of 3 pounds with a standard deviation of 6 pounds.

Can we reject the null hypothesis at the 5% level?

In this case

$$z = \frac{(5-3)-0}{\sqrt{4^2/34 + 6^2/60}} = 1.9333$$

so we reject at the 5% level because $P(Z > 1.933) < 5\%$.

If this were a 5% two-sided test, would we reject?

Difference in proportions

Suppose we try to address the Coke/Pepsi local market share with a different kind of survey in which we conduct two separate polls and ask each person either "Do you regularly drink Coke?" or "Do you regularly drink Pepsi?"

With this set up we want to know if $p_x = p_y$.

Suppose we ask 40 people the Coke question and 53 people the Pepsi question. In this case the observed difference in proportions has approximate distribution

$$\hat{p}_x - \hat{p}_y \sim N\left(0, \sqrt{\frac{p_x(1-p_x)}{40} + \frac{p_y(1-p_y)}{53}}\right)$$

Difference in proportions

In practice we have to use

$$N\left(0, \sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{40} + \frac{\hat{p}_y(1-\hat{p}_y)}{53}}\right)$$

If 30/40 people say that they regularly drink Coke and 30 out of 53 people say they regularly drink Pepsi, do we reject the null hypothesis at the 10% level?

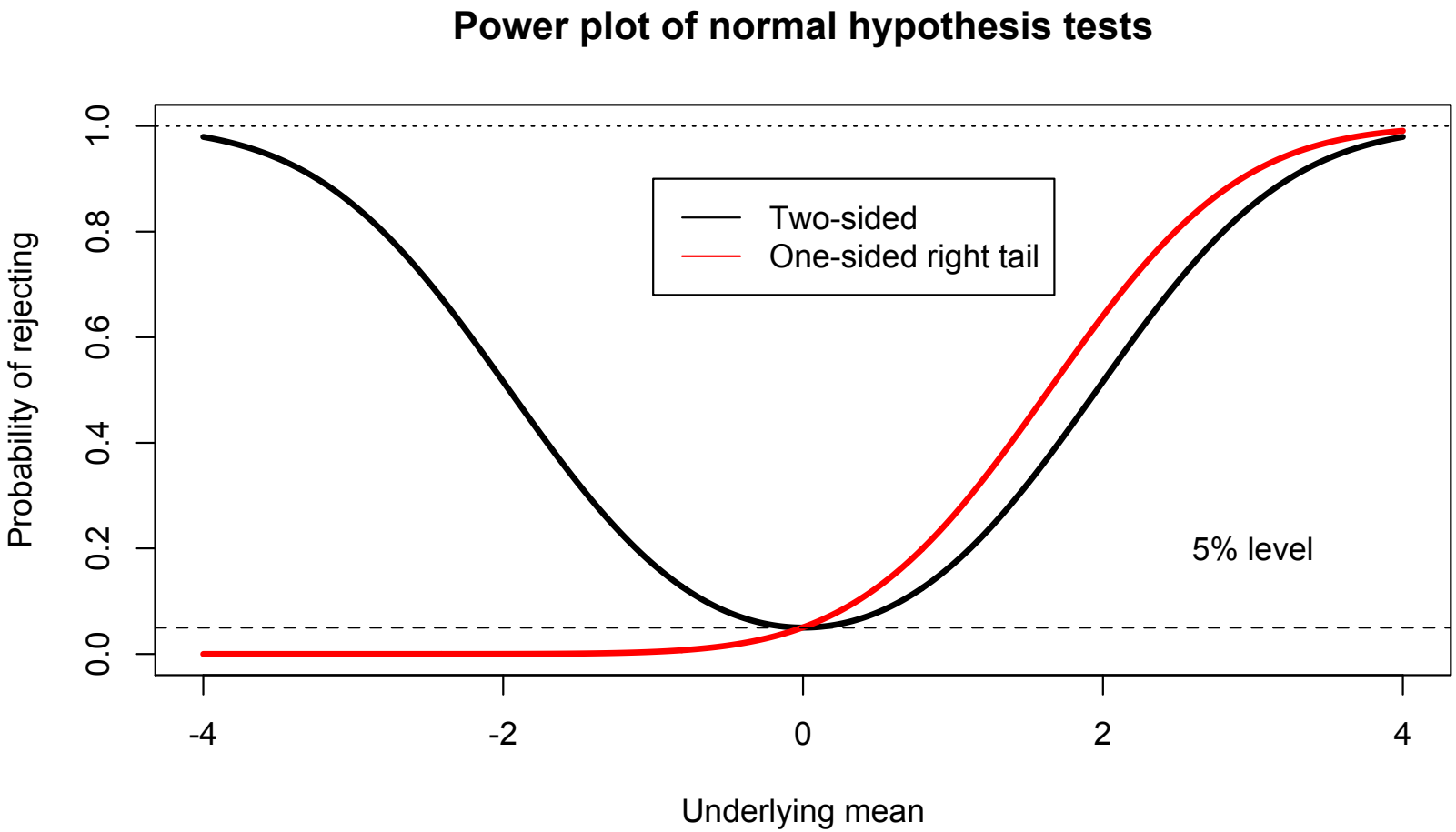
We find

$$\sqrt{\frac{0.75(1-0.75}{40} + \frac{0.566(1-0.566)}{53}} = 0.09655$$

so $z = \frac{(0.75-0.566)-0}{0.09655} = 1.905$

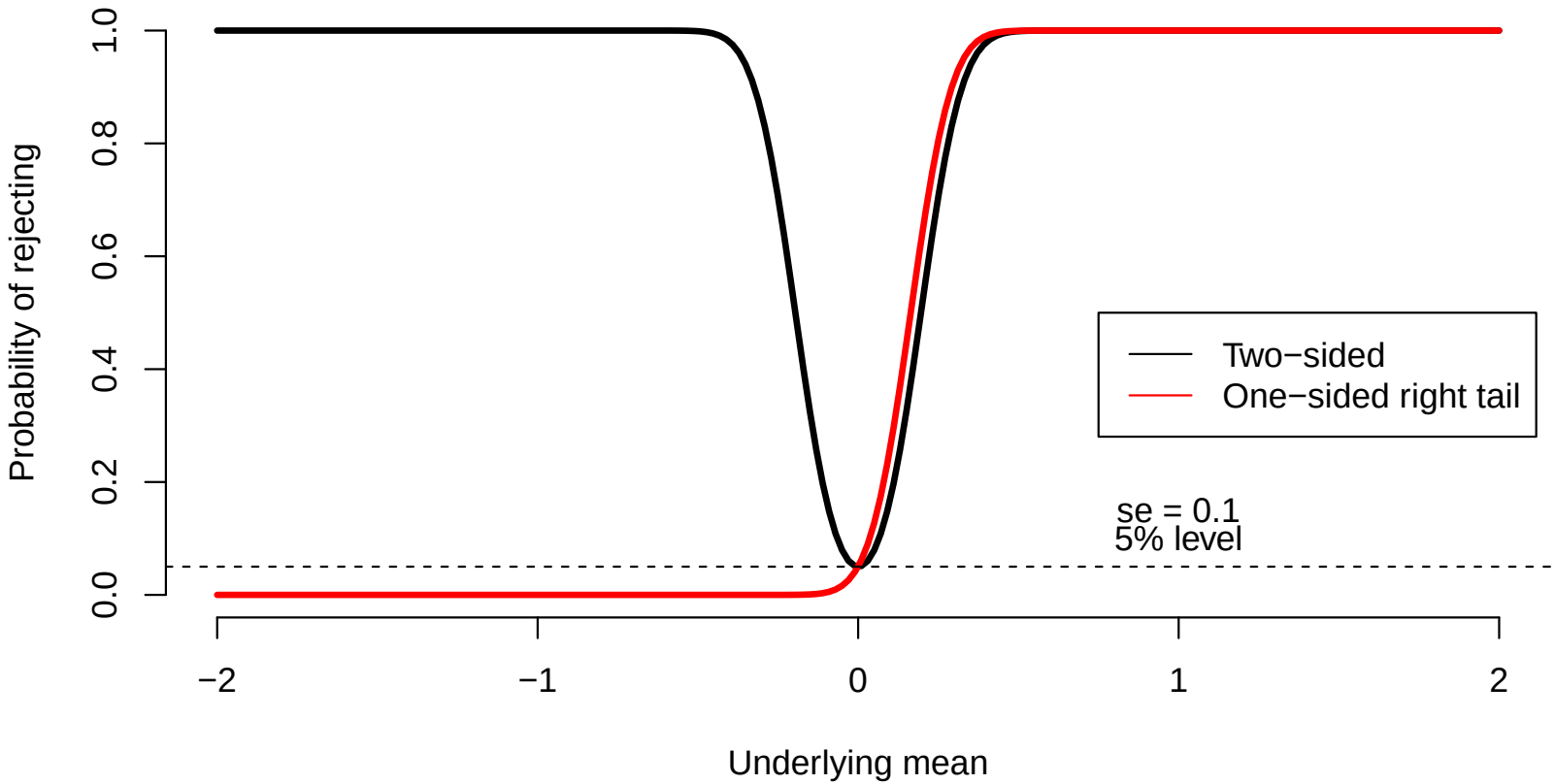
Do we reject at the 5% level?

Power



The probability of rejecting the null hypothesis is called the **power** of the test. It will depend on the actual underlying value. The **level** of a test is precisely the power of the test when the null hypothesis is true.

Power



The power function gets more "pointed" around the null hypothesis value as the sample size gets larger (which makes the standard error smaller).