Comparing Samples: the logic of hypothesis testing

Research Methods for Political Science

Thomas Chadefaux

Trinity College Dublin

This lecture

- More items on the logic of hypothesis tests
- t-tests

More on hypothesis testing

Another way to think about hypothesis tests

Think back about what we've learned:

$$z = \frac{\text{estimate} - \text{hypothesized value}}{\text{standard deviation of the estimate}}$$

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

Another way to think about hypothesis tests

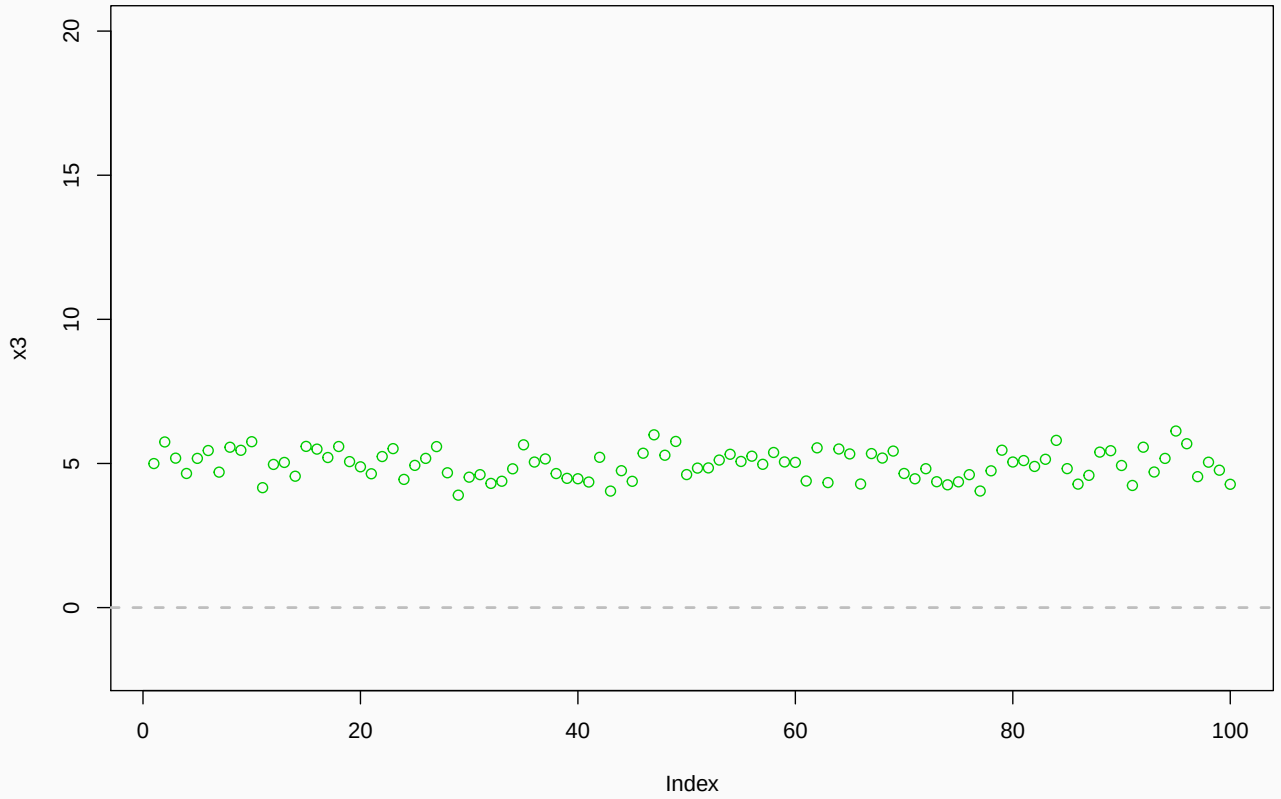
$$\text{test} = \frac{\text{Signal}}{\text{Noise}}$$

Sigmal = the magnitude of the estimate. E.g. difference in mean

Noise = the standard deviation of the estimate

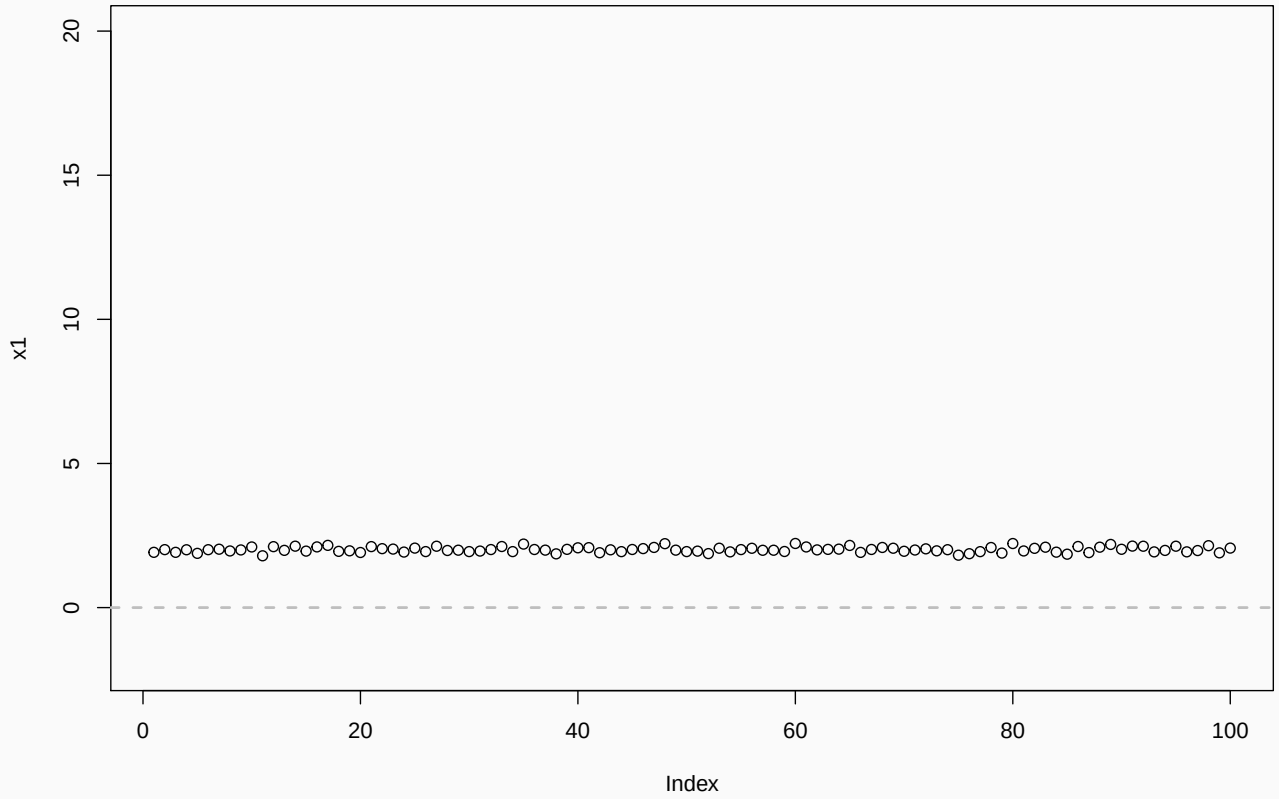
Another way to think about hypothesis tests

Ideal: high signal, low noise



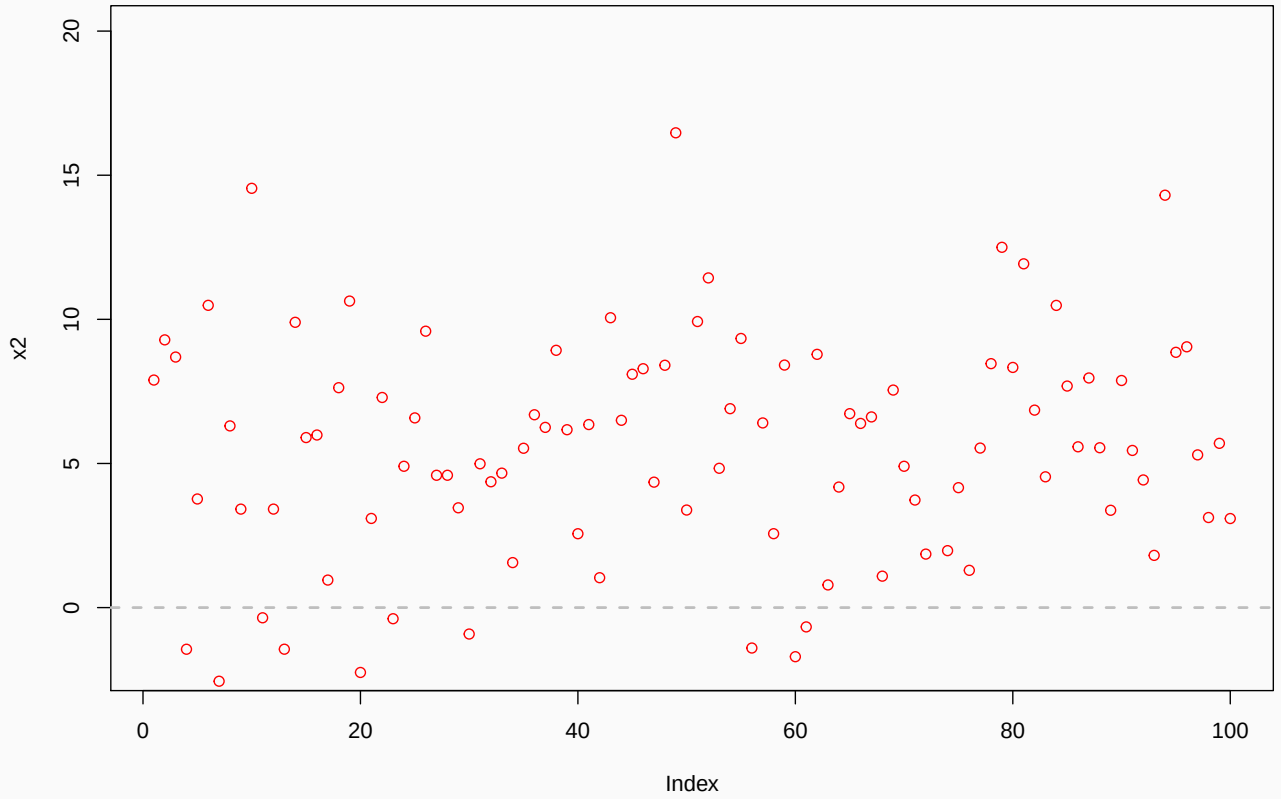
Another way to think about hypothesis tests

low signal, low noise



Another way to think about hypothesis tests

high signal, high noise



Accepting or rejecting the null hypothesis

Do you *accept* the null hypothesis? No

Why not? Because the test told you that your z is a *common* value given this null hypothesis (e.g., the null hypothesis that $\overline{IQ} = 100$). But it would also be a common value given a similar hypothesis (e.g., $\overline{IQ} = 100.5$).

So you do not *accept* the null hypothesis that $\overline{IQ} = 100$. You just **fail to reject it**.

One vs two-tailed tests

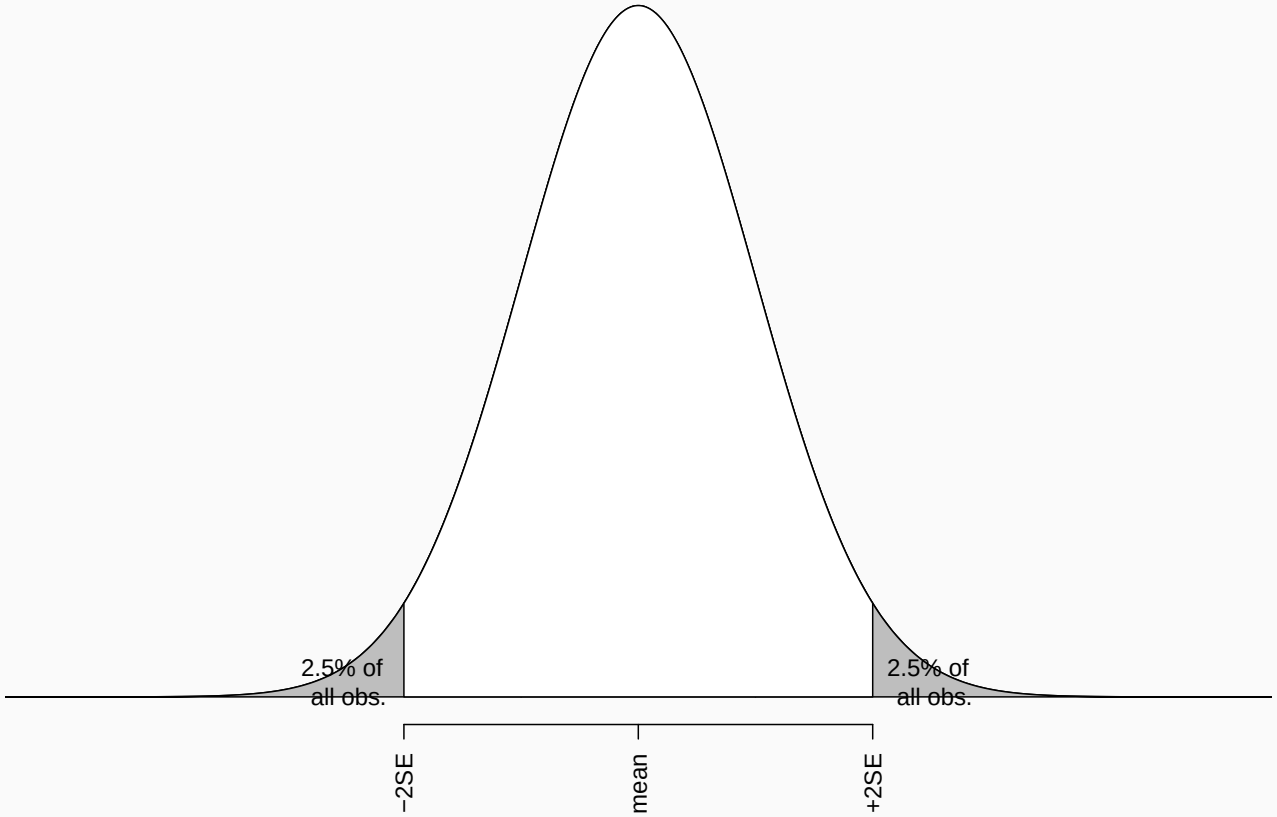
One-tailed vs two-tailed tests

Two-tailed tests: reject the hypothesis if z is particular large OR particularly low.

E.g.: H_0: the yearly income of TCD Graduates does not differ from the national average income of €30,000. Here we would reject the null hypothesis if TCD students earn more OR less than 30,000

Two-tailed tests

We reject the null hypothesis if z falls in the grey zone. Either the left or the right one.



One-tailed tests

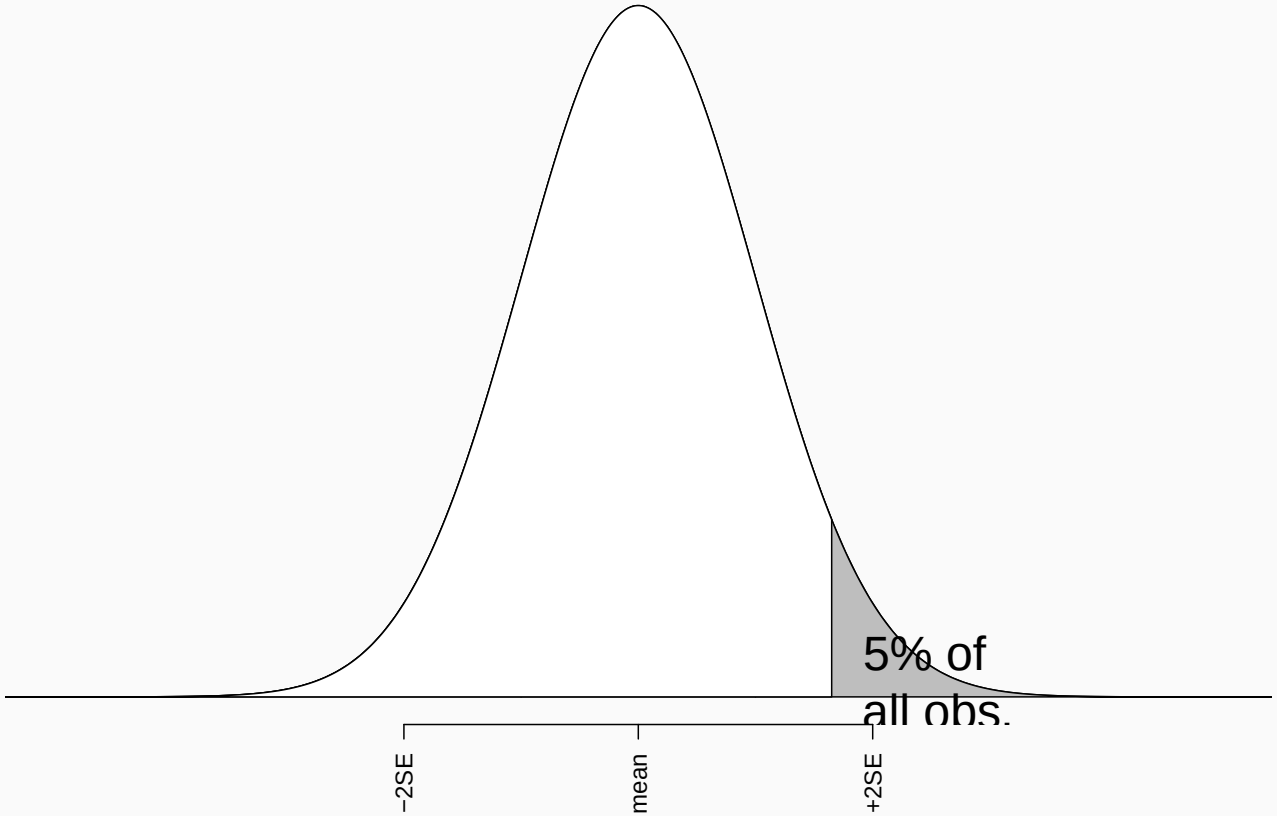
But perhaps this is not the hypothesis we want to test. Instead, we want to test whether TCD students earn MORE than the national average. So:

H_0 :TCD students earn the same amount as the national average

H_1 :TCD students earn more than the national average

One-tailed tests

We reject the null hypothesis if z falls in the grey zone. This time, the grey zone is only on the upper-tail.



One-tailed tests

Note that we reject the null hypothesis if what we observe is "rare" given the null hypothesis.

What we consider as rare, however, is now a particularly large income (no longer a particularly low one). So we reject only if we observe an outcome in the upper tail.

Since we still want a level of significance $\alpha = 0.05$, we assign all 5% to the upper-tail.

One-tailed tests

The threshold to rejecting the null hypothesis in a one-tailed test is therefore **lower**.

This could be a problem. I recommend only using two-tailed tests unless you have a very good reason not to.

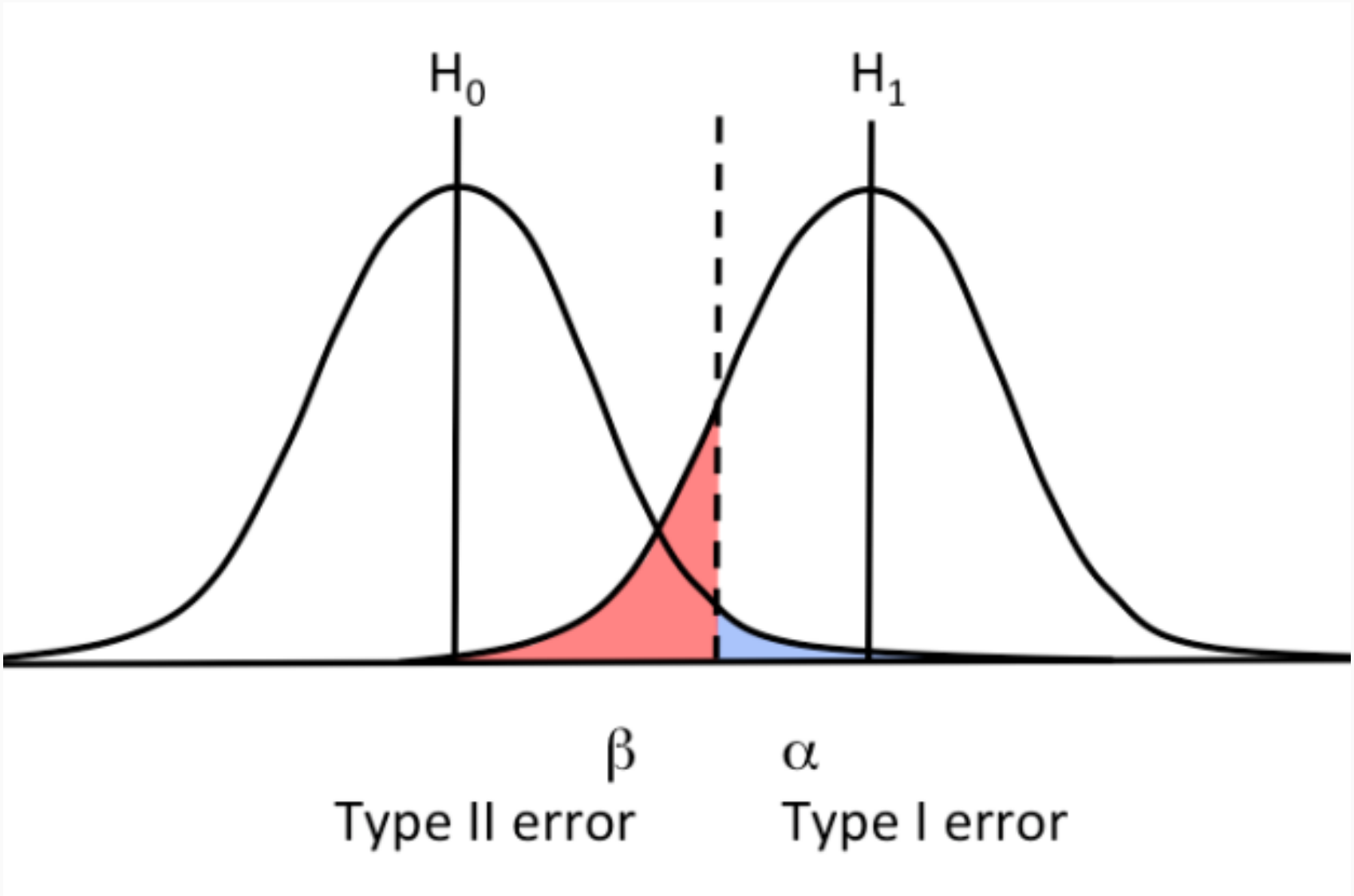
Types of errors

(From Witte)

Table 11.2
POSSIBLE OUTCOMES OF A HYPOTHESIS TEST

	STATUS OF H_0	
DECISION	**TRUE H_0**	**FALSE H_0**
Retain H_0	(1) Correct decision	(3) Type II error (miss)
Reject H_0	(2) Type I error (false alarm)	(4) Correct decision

Types of errors and power



Types of errors and power

(From Witte)

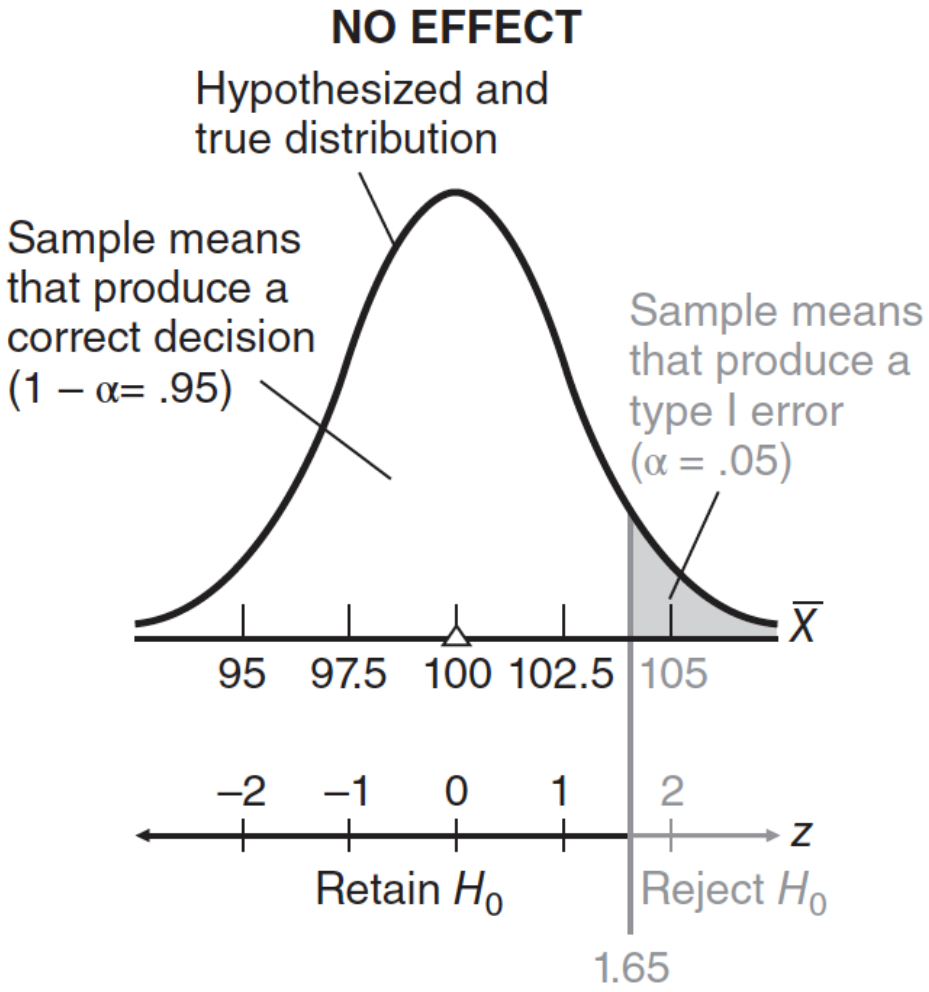
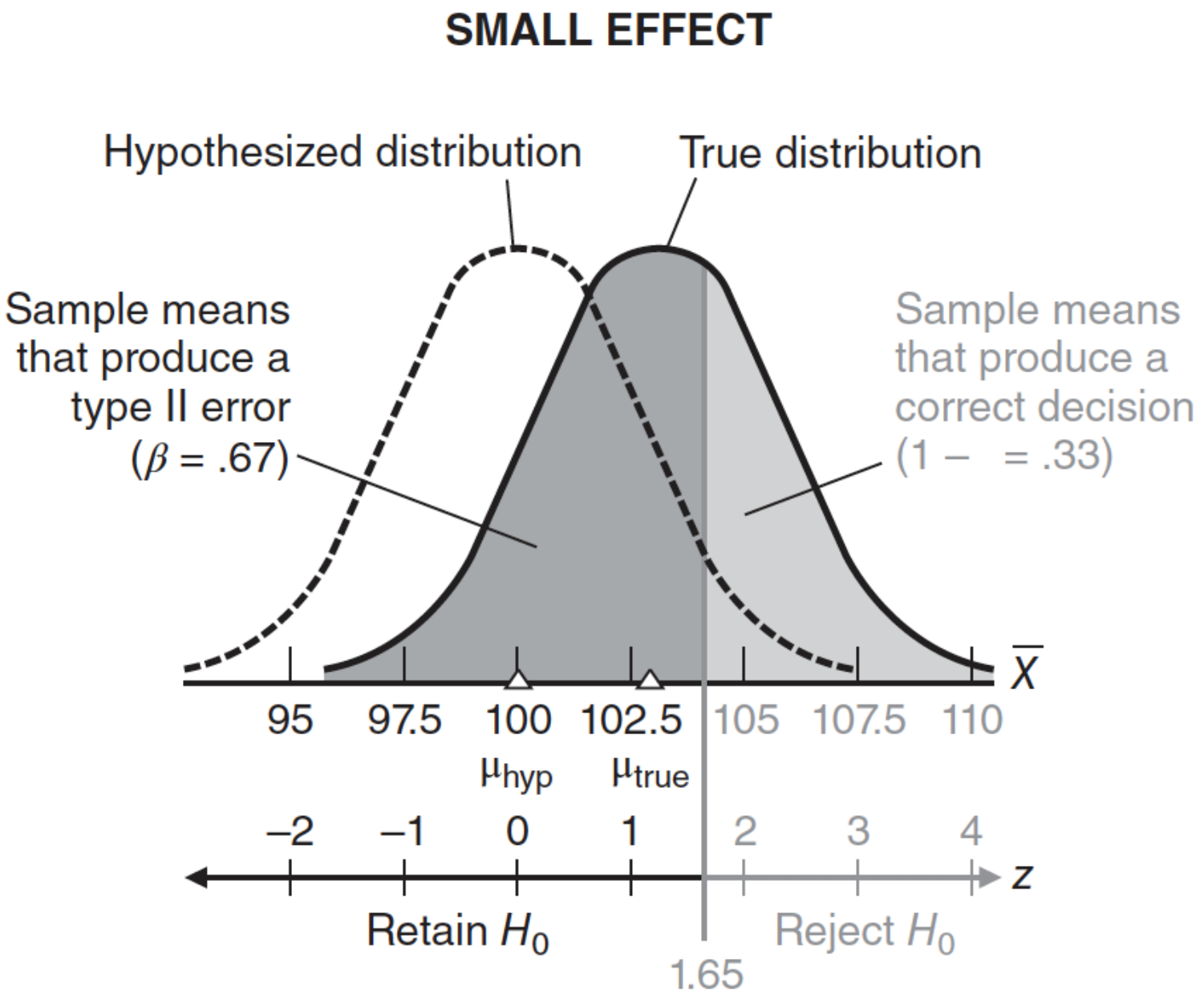


FIGURE 11.3

Hypothesized and true sampling distribution

Types of errors and power



Types of errors and power

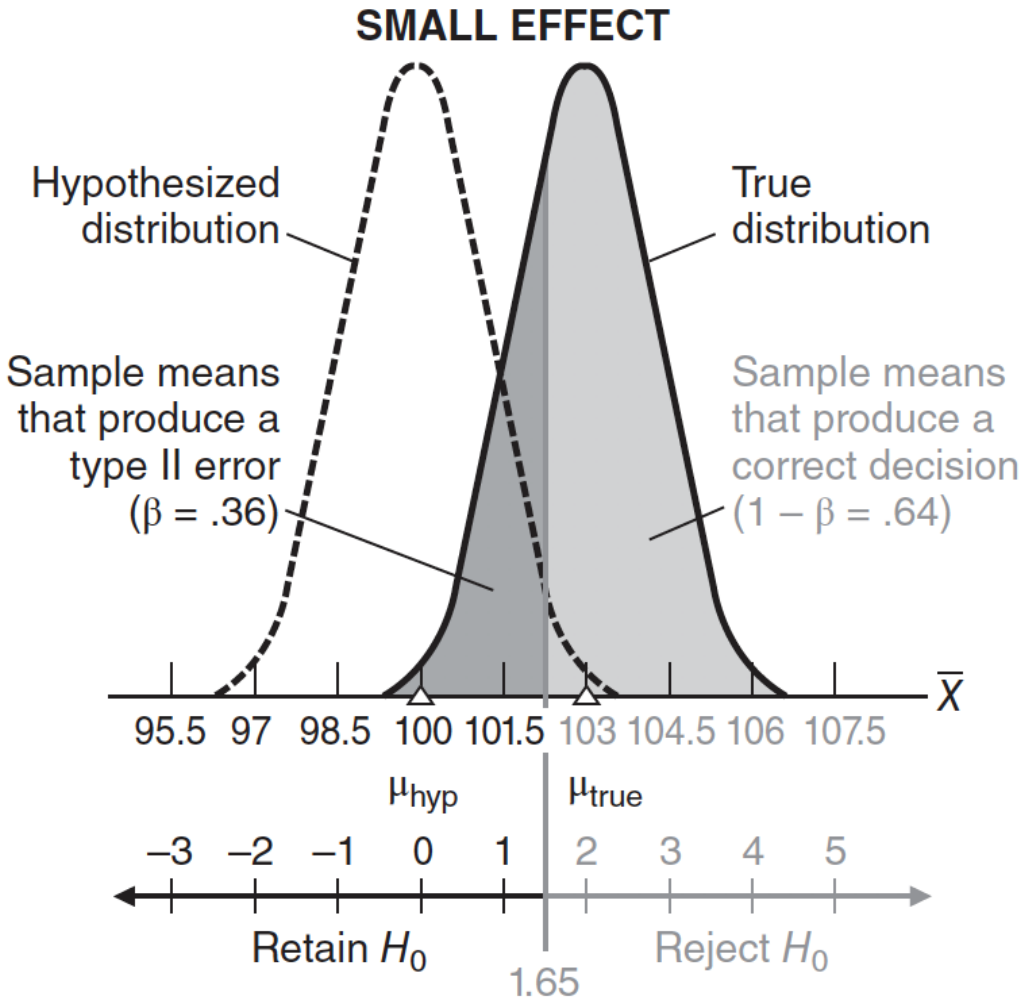


FIGURE 11.6

Hypothesized and true sampling distribution when H_0 is false because of a small effect but sample size is relatively large.

The p-value

The p-value

Suppose that we obtained a z-value of $z = 2.5$. We could just conclude that 2.5 is greater than the critical value of 1.96 associated with a significance level of 5% and hence reject the null hypothesis at the 5% significance level.

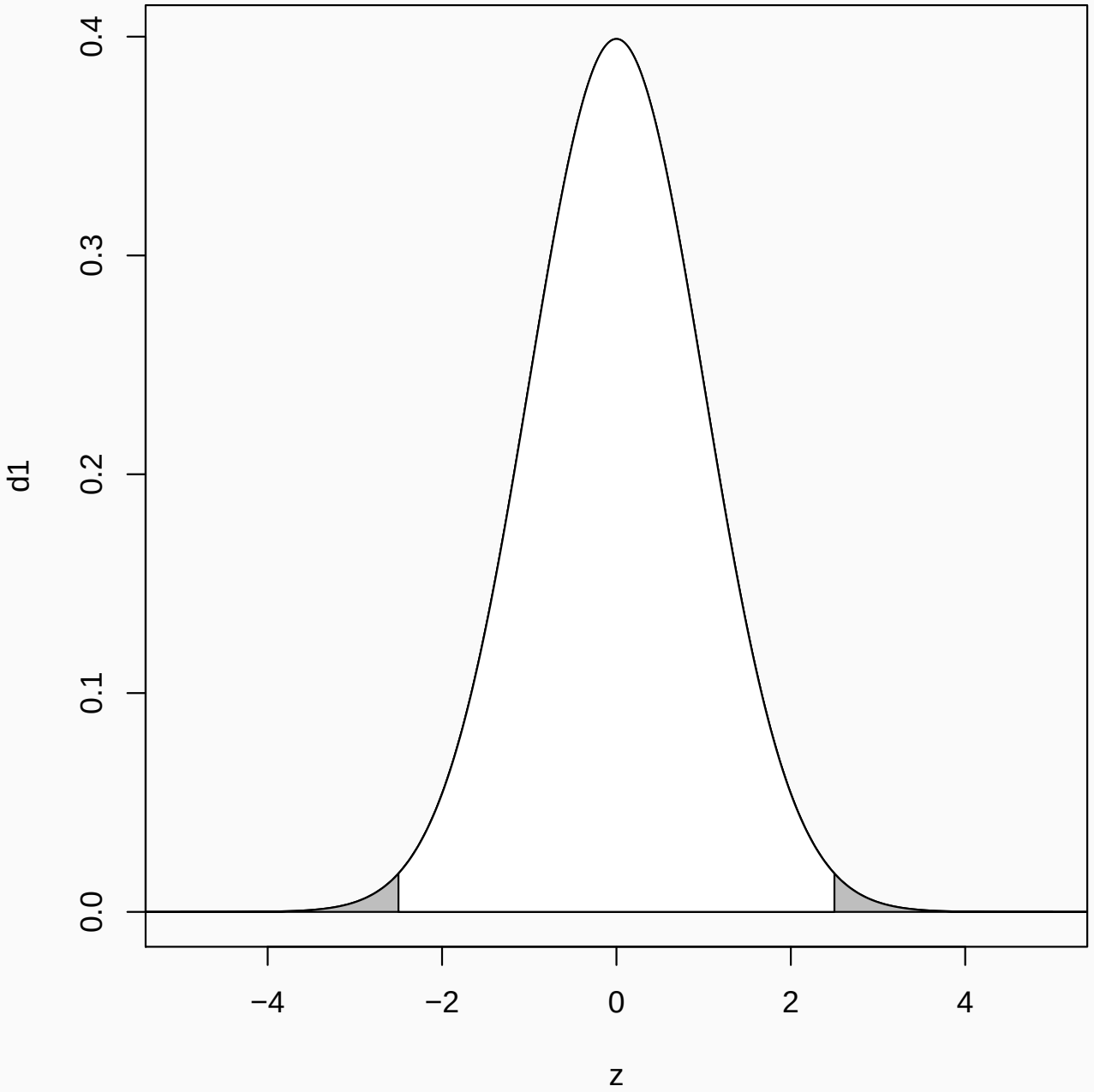
But often, rather than only making a decision (accept/reject) we'll want to assign a probability to observing a value this extreme if the null hypothesis were true

I.e., because we are using a two-sided alternative, we want to find:

$$P(Z \leq -2.5 \text{ OR } Z \geq 2.5),$$

where $z \sim N(0, 1)$

The p-value



The p-value

p-value
The p-value is the probability, assuming H_0 is true, that the test statistic would take a value as extreme as what is actually observed. The smaller the P-value, the stronger the evidence against H_0. (from Moore et al., p.413)

The key to calculating the P-value is the sampling distribution of the test statistic. In most cases we will only use the normal or t distribution (more on that later)

But how do I find $P(Z \leq 2.5 OR Z \geq 2.5$?

- Note that p is really the area under the sampling distribution curve (here the normal distribution) above 2.5 and below -2.5 (or whatever value you got for z).
- So let's ask R to calculate the area below -2.5:

```
## [1] 0.006209665
```

The p-value

```
## [1] 0.006209665
```

- This tells us that 0.6% of the area under the curve is below $z = -2.5$.
- Since the curve is symmetric, we also know that 0.6% is above $+2.5$. So in total, 1.2% of the data is within the grey area. In other words, if the null hypothesis were true, then an observation as extreme (or more) as ours (2.5) would occur in 1.2% of cases.

The p-value

The p-value is reported for a variety of tests. The z-test here, the t-test (more later), the χ^2-test (more later), the regression coefficients (more next term), etc.

It is an easy way to answer the question: if there was no difference (or not effect etc), how often would I get such an extreme outcome?

Remember: the lower the p-value, the more our outcome is "surprising" and hence the more confidence we have in the fact that there is a difference.

Hypothesis tests or confidence intervals?

Hypothesis tests or confidence intervals?

There is a very close relationship between the two. In fact, for the one-sample tests we have here, either one will give you the same answer.

If for example I find that the 95% confidence interval for my mean is $[1.3 - 4]$, then I know that 0 is not in the 95% CI, and therefore is not a reasonable estimate of the mean. Therefore, when I test my hypothesis that the mean $= 0$, I will find that I can reject the null hypothesis (i.e., all of these things will be true: $z > 1.96$, $p<0.05$)

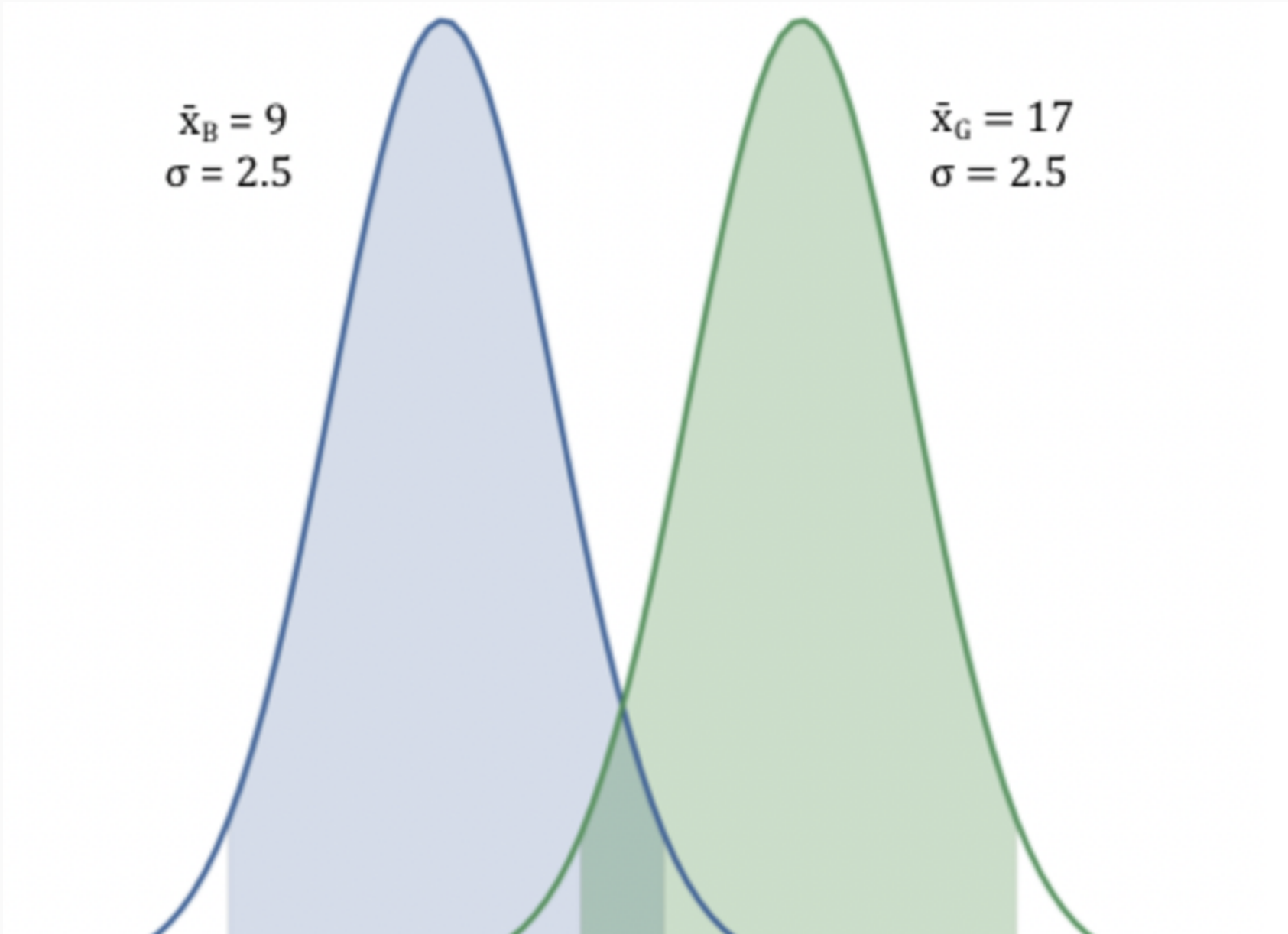
Hypothesis tests or confidence intervals?

So why bug you with hypothesis tests then?

Well, things get trickier with more complicated tests. When we compare two samples, for example, it will be harder to decide whether they are different or not by just looking at whether their CIs overlap or not.

(Hypothesis tests or confidence intervals? Jumping ahead)

E.g., the CIs overlap, yet their means are significantly different!



z-tests and t-tests

When is the z-test appropriate... and when it is not

The *z*-test is accurate if:

1. the population is normally distributed *or* the sample size is large enough to satisfy the requirements of the central limit theorem
2. the population standard deviation is known.

point # 2 above is problematic, because we usually do **not** know the **population** standard deviation. All we know is the **sample** standard deviation.

the t-test

Previously we used σ, i.e., the *population* standard deviation, to calculate the standard deviation of the sampling distribution, in order to obtain a z-score ($z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$).

Now we will just use s and get a statistic called t:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Why the t-test instead of the z-test? The z-test relies on the assumption that we *know* the population standard deviation.

But we usually do not.

We'll therefore use another test called the t-test

t-test

The good news is that the logic of the t-test is exactly the same as the one for the z-test. Nothing really new to learn.

t-test

Like the sampling distribution of z, the sampling distribution of t represents the distribution that would be obtained if a value of t were calculated for each sample mean for all possible random samples of a given size from some population

Why the t-test? Where does this come from?

The sample standard deviation is a biased estimator of the population standard deviation

In particular, the sample standard deviation is likely to be lower than the population standard deviation.

(Aside, can be skipped: Why is the sample SD smaller than the population SD?)

Too long to cover here, but see slides at the end if you are curious (not required).

Click *here*

Where the t-distribution comes from

The sample standard deviation is likely to be lower than the population standard deviation.

Consequence: Now imagine that the *population* standard deviation is 1.5, but the *sample* standard deviation is calculated to be 1.

Then a measurement of 1 greater than the mean is incorrectly estimated to be 1 standard deviations greater than the mean, rather than 2/3.

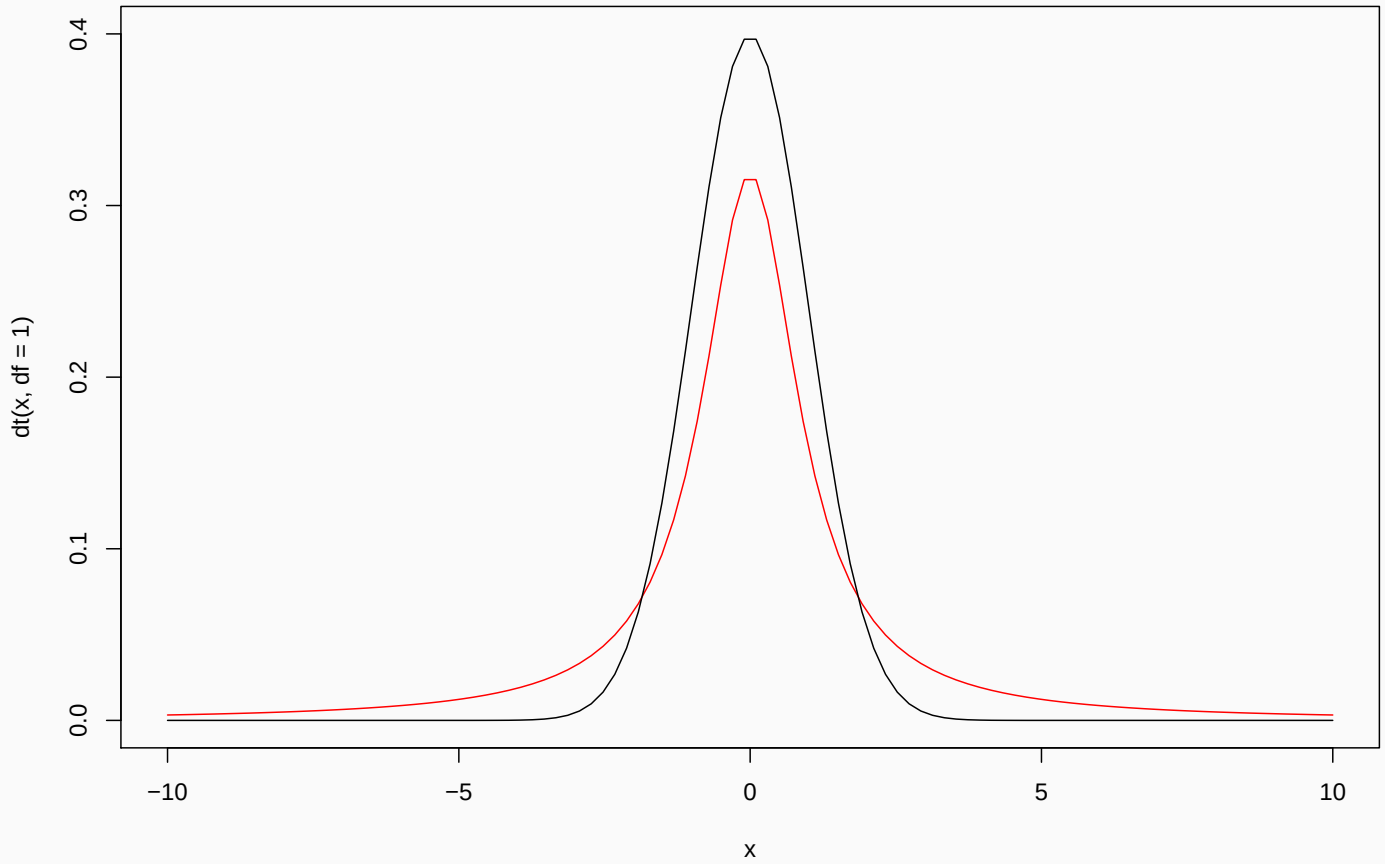
Where the t-distribution comes from

In other words: given that you underestimate the standard deviation, you will tend to observe more observations that are far away from the mean, and fewer observations that are closer to the mean.

If you don't account for this different distribution of sample means, you will tend to reject the null hypothesis too often, because you will see means that are "rare" more often.

Where the t-distribution comes from

In other words, you will observe a distribution that looks more like the red curve than the black one:



the t-distribution

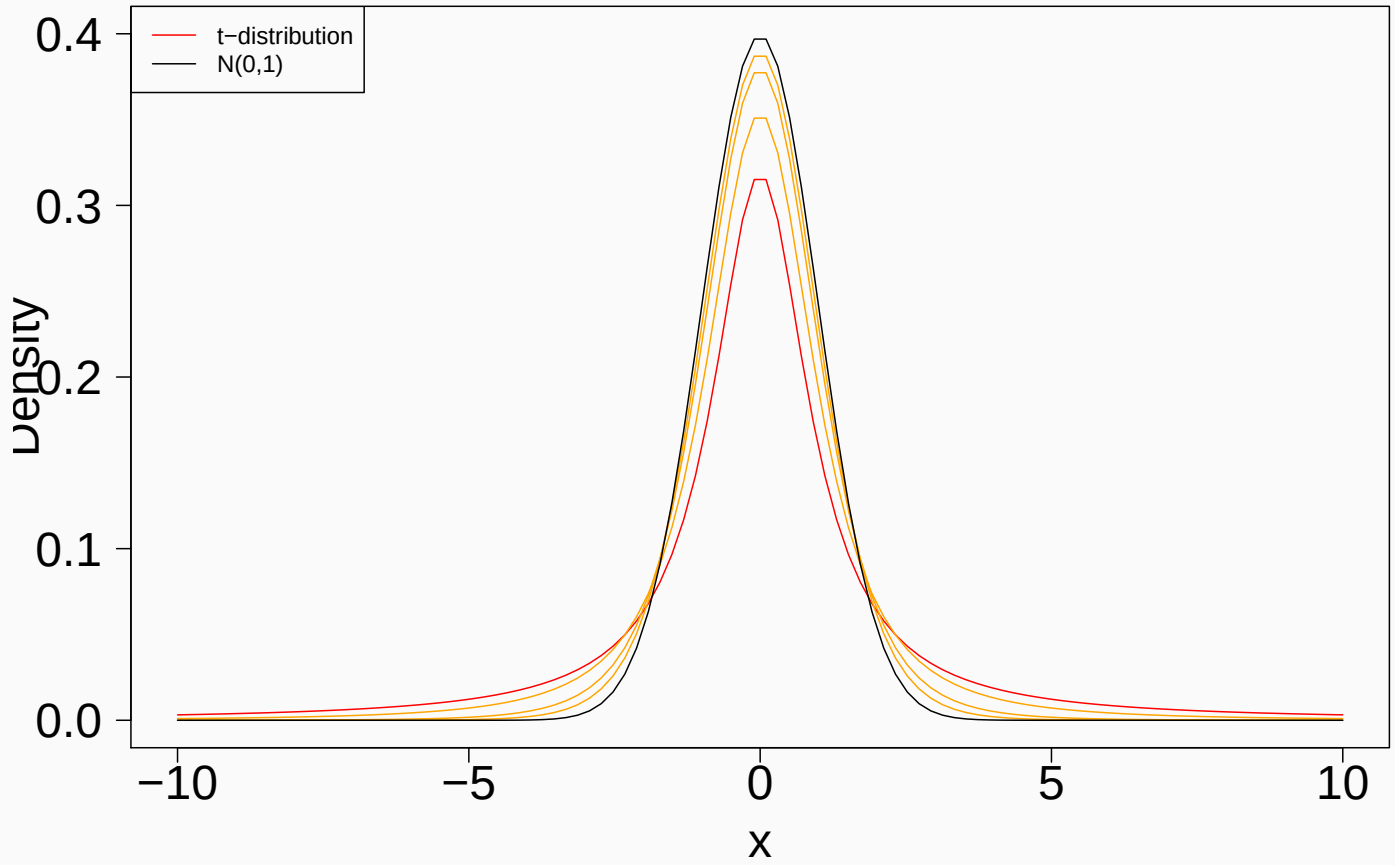
This new distribution is called the t-distribution.

We will use this distribution to test hypothesis, rather than the normal distribution (i.e., the distribution of z statistics)

The shape of that distribution depends on something called the number of degrees of freedom (more below). Intuitively, the more observations you have, the less biased (i.e., wrong) your estimate of the standard deviation will be.

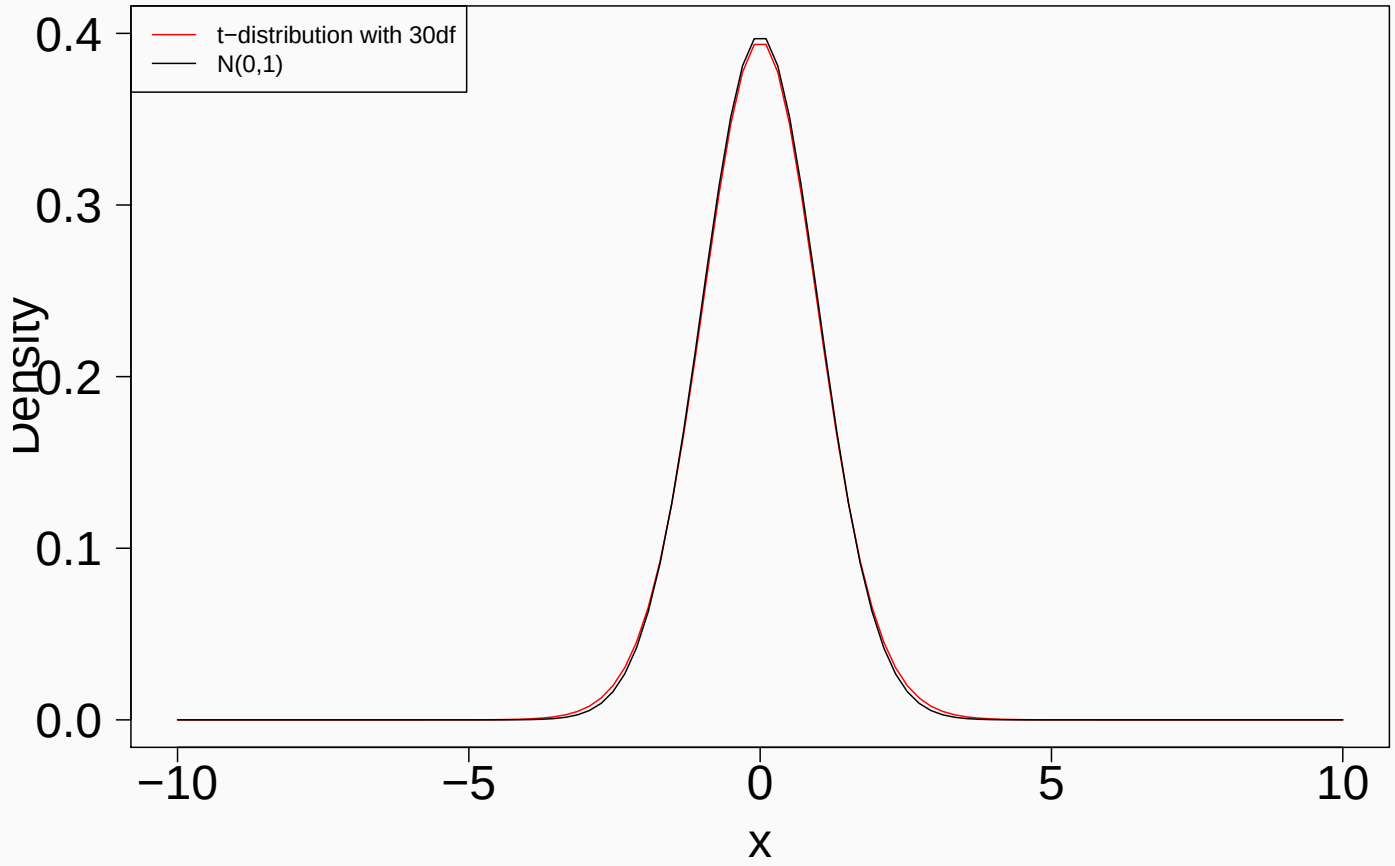
the t-distribution

t distribution with 1, 2, 5, 10 degrees of freedom:



the t-distribution

What matters is that as the number of observations (and hence degrees of freedom) increases, the t-distribution becomes increasingly normal. By the time you have about 30+ observations, the two are virtually indistinguishable.



A rabbit hole: degrees of freedom

Degrees of freedom

Degrees of freedom (df) indicate the number of values that are free to vary.

Hmm, that's not very helpful

Degrees of freedom: an example

Suppose I told you that I have a sample as follows:

$$1, 3, 5, x$$

now, I tell you that the mean of that sample is 3. Can you guess then what x is?

Degrees of freedom: an example

Yes, x MUST be 3. That's the only way the mean would be 3. In other words, x was not "free" to vary. It HAD to be 3.

Therefore in the sample above, given that the mean was 3, there were only 3 observations that were free to vary, NOT 4. I.e., there were n-1 degrees of freedom.

Degrees of freedom

The last value and the mean are entirely dependent on each other.

So, after estimating the mean, we have only $n - 1$ independent pieces of information, NOT n.

Out of the rabbit hole

Recap

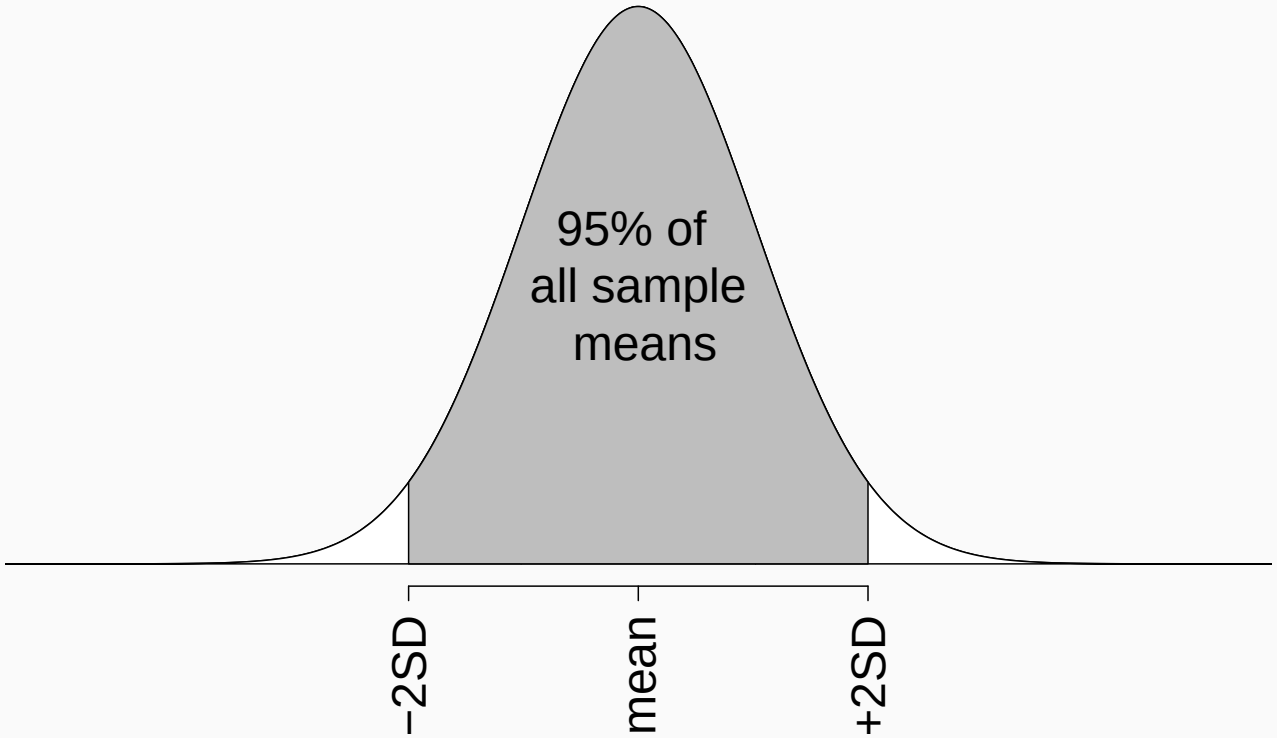
Recap

1. We observe a sample mean and want to know whether that sample mean is different from a hypothesized sample mean
2. We look at the distribution of sample means under the null hypothesis. How do we get that distribution? By knowing that:
   - the distribution will be t-distributed
   - the distribution will have standard deviation $\frac{s}{\sqrt{n}}$
3. We compare our sample mean to the sampling distribution and decide to reject H_0 or not, if t is greater than the critical value (more on how to find those below)
4. We can calculate the p-value

Finding critical values

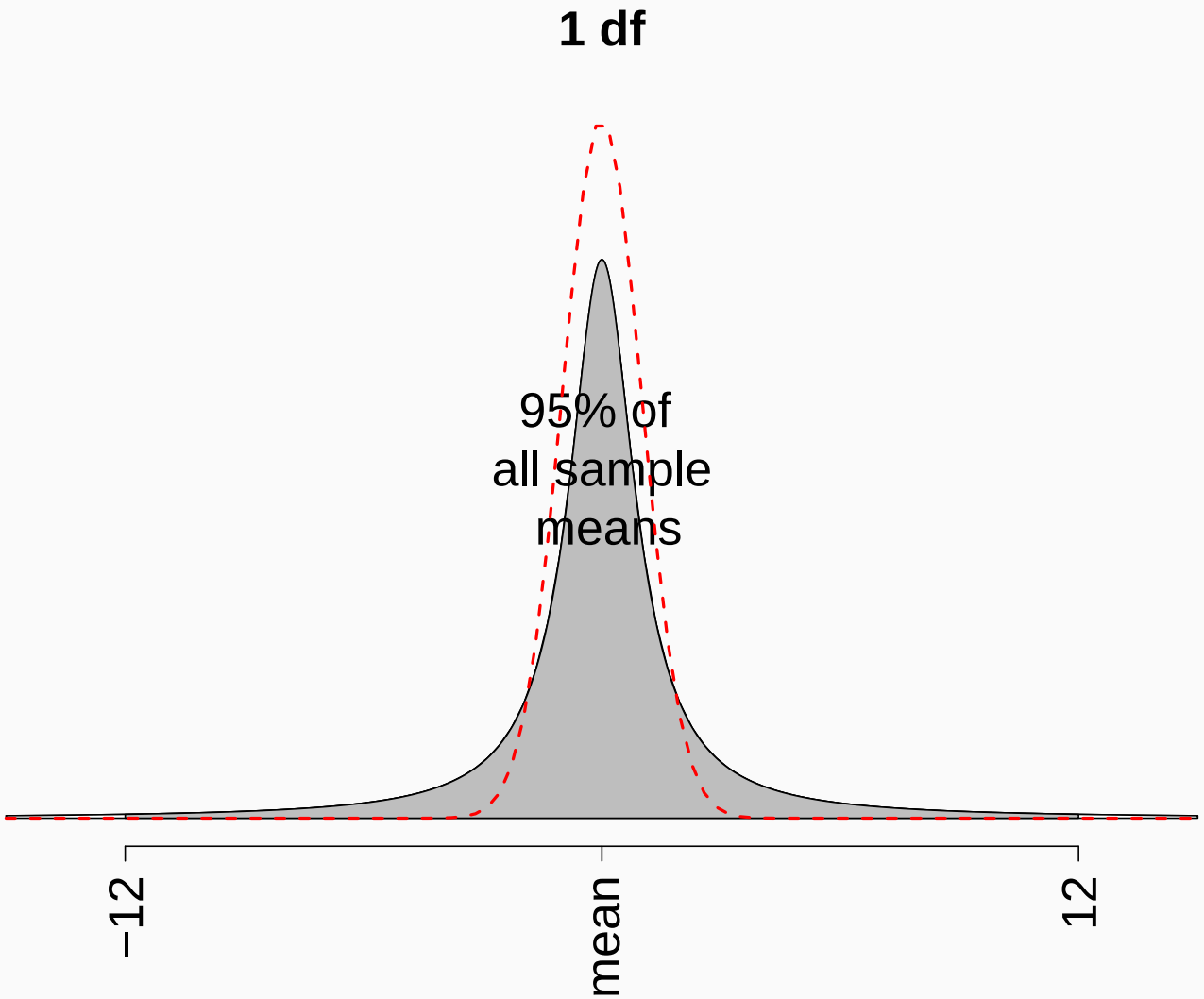
Finding critical values

So far I have given you the relevant critical values for the **z-test**. Mostly, we use 1.96 as a critical value for the two-tailed test at the 5% significance level.

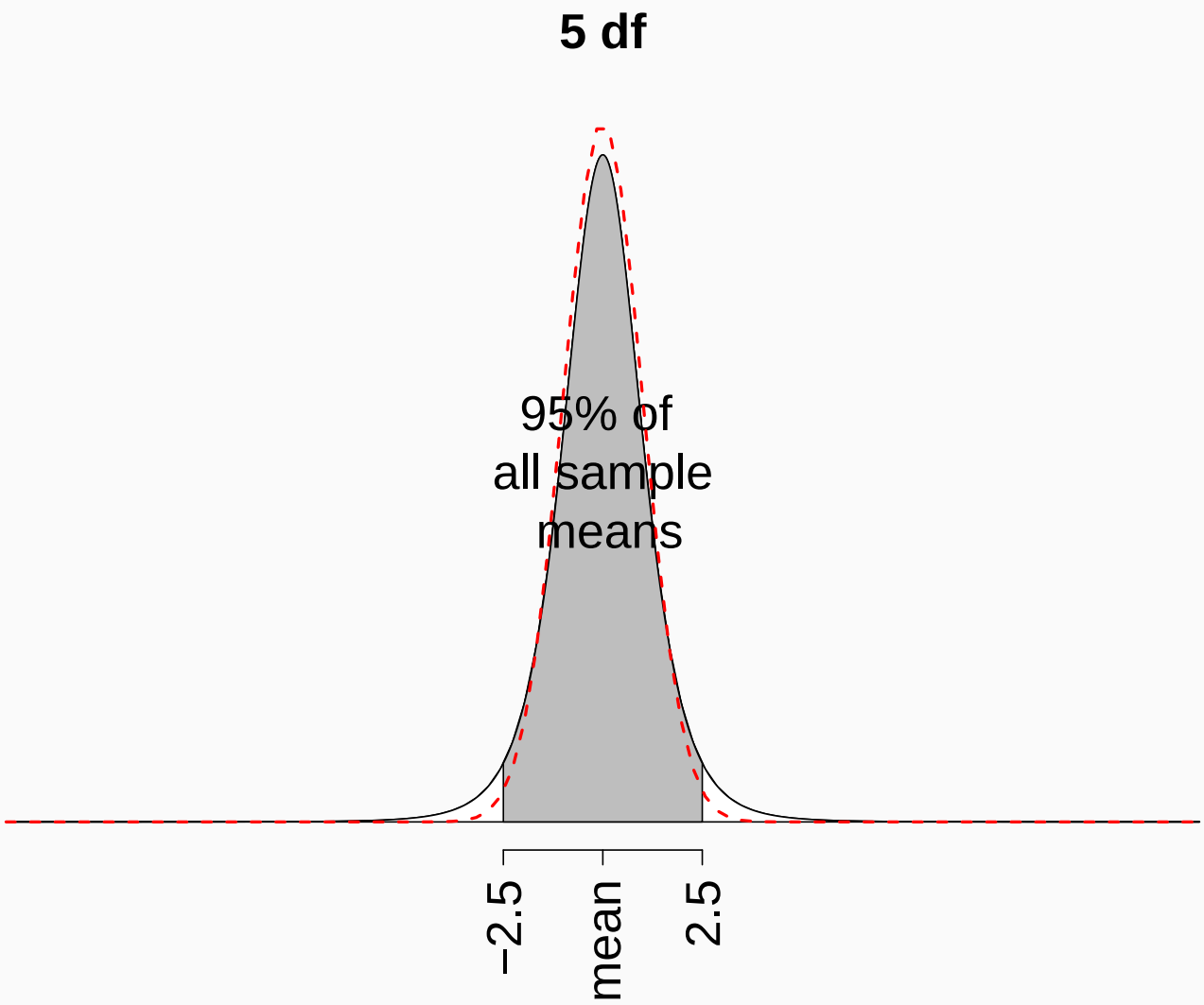


Finding critical values

For the t-test, things are bit trickier because the critical values depend on the degrees of freedom. To see why:



Finding critical values



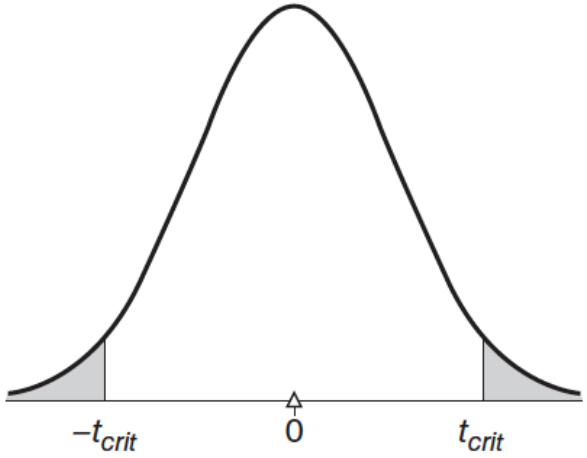
How to find critical values?

Two methods:

1. Statistical table
2. Ask R

How to find critical values? Method 1: Statistical tables

In the back of your book (or any stat book)



Two-tailed or Nondirectional Test
LEVEL OF SIGNIFICANCE

$p > .05$ $p < .05$ $p < .01$ $p < .001$

df	.05*	.01**	.001
1	12.706	63.657	636.62
2	4.303	9.925	31.598
3	3.182	5.841	12.924
4	2.776	4.604	8.610
5	2.571	4.032	6.869
6	2.447	3.707	5.959
7	2.365	3.499	5.408
8	2.306	3.355	5.041
9	2.262	3.250	4.781
10	2.228	3.169	4.587
11	2.201	3.106	4.437
12	2.179	3.055	4.318

How to find critical values? Use R

The function qt(x,...) gives you the t-value associated with a probability x in the lower tail. For example, we know that with many degrees of freedom the t distribution is the same as the normal distribution, so we should find the critical value for a significance level of 0.05 to be around -1.96

```
qt(p = 0.05/2, df = 1000)
```

```
## [1] -1.962339
```

Note: why did I divide 0.05 by 2? Because we want a two-tailed test

How to find critical values? Use R

Let's check that we get the same results as in the table:

```
qt(p = 0.05/2, df = 6)
```

```
## [1] -2.446912
```

So if we have 7 obs. (and hence 6 df), we'll know that the critical value of the t statistic is 2.44. I.e., if $|t| > 2.44$, we reject the null hypothesis.

Reminder: you can also find the p-value associated with that t:

```
pt(q = 2.44, df=6)
```

```
## [1] 0.9747642
```

t-test in R

t-test in R

very simple: t.test()

E.g. you observe a sample of 6 political candidates' yearly income (in thousands of euros):

you want to test the hypothesis that candidates' income significantly differs from the average income in Ireland, which is €47,000.

t-test in R

So:

$$H_0 : \text{income of candidates} = 47k$$

$$H_1 : \text{income of candidates} \neq 47k$$

Level of significance (α): 5% (equivalently, confidence level: 95%), and so critical value is:

```
## [1] -2.446912
```

t-test in R

Of course, R does it all for you:

```
##
##  One Sample t-test
##
## data:  incomes
## t = 1.5106, df = 5, p-value = 0.1913
## alternative hypothesis: true mean is not equal to 47
## 95 percent confidence interval:
##   29.34035 114.99299
## sample estimates:
## mean of x
##  72.16667
```

t-test in R

How to read this? First, R tells you

$$t = 1.5106$$

How did R get this number? First, calculate the standard error:

$$SE = \frac{SDincome}{\sqrt{n}} = 40.8/2.45 = 16.65$$

Now just apply the formula:

$$t = \frac{\text{meanIncome} - 47}{SE} = 25.17/16.65 = 1.51$$

t-test in R

Then R tells you: $df = 5$. The degrees of freedom are just the number of observations (here 6) minus the number of parameters to be estimated (here just one: the mean). So $6 - 1 = 5$df.

t-test in R

The most important info comes next: "p-value = 0.1913".

This is not smaller than 0.05, and therefore we fail to reject the null hypothesis (why?)

How did R get this p-value?

```
## [1] 0.1914287
```

(why multiply by 2? and why *minus* 1.51?)

t-test in R

Next, R kindly makes sure this is what you wanted: “alternative hypothesis: true mean is not equal to 47” (i.e., implicitly, R is telling you this is a two-sided test)

t-test in R

Finally, R gives you a “95 percent confidence interval”: 29.34035 114.99299

t-test in R

How did R calculate this? Remember that the confidence interval will be given by: $\overline{x} \pm t_{1-\alpha/2,df} \times SE$, where $t_{1-\alpha,df}$ is the critical value of the t-distribution with df degrees of freedom and at the $1 - \alpha/2$ confidence level.

t-test in R

We calculate $t_{1-\alpha/2,df}$ by using:

```
## [1] -2.570582
```

and so

$$CI = 72.16 \pm 2.57SE = 72.16 \pm 2.57 \times 16.65 = [29.37 - 114.95]$$

(note: v slight discrepancies due to rounding)

In practice

In practice (at least after this class), you will use the command t.test rather than calculate everything manually. But it is crucial that you understand what the output means and how R obtained it.

Appendix

Aside: Why is the sample SD smaller than the population SD?

Remember that the variance is estimated as:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

This is an unbiased estimator of the population variance σ^2.

However, note that the square root function is strictly concave such that (by Jensen's inequality if you care. Loosely, $x + y > \sqrt{(x^2 + y^2)}$):

$$E(\sqrt{s^2}) < \sqrt{E(s^2)}$$

For example, suppose we have a true population of: 1,2,3,4 with sd: we take a sample:

```
## [1] 0.7625066
```

```
## [1] 1.008172
```