Hypothesis testing:
Errors, Power, Effect size, and Assumptions

Data Analysis for Psychology in R 1

Semester 2, Week 5

Dr Umberto Noè

Department of Psychology
The University of Edinburgh

Learning objectives

1. Understand what are Type I and Type II errors in hypothesis testing.
2. Recognise the significance level as measuring the tolerable chance of committing a Type I error.
3. Recognise the effect of sample size on power.
4. Be able to check the assumptions underlying the t-test for a population mean.

Part A

Errors and Power

Where we're going to

- Hypothesis testing lets us determine whether, for example, an observed difference between a sample mean and an hypothesised value is real or just due to random sampling variation.
- However, statistical significance sometimes may lead us to wrong conclusions!
- It is possible to make two kinds of wrong decisions:
    - rejecting a true null hypothesis
    - not rejecting a false null hypothesis
- This week we will discuss common pitfalls of hypothesis testing, as well as the factors that influence the probability of committing these errors.

Errors in hypothesis testing

- Whether your decision is either to (a) reject the null hypothesis or (b) not reject the null hypothesis, you might be making an error.
- The reasoning of hypothesis tests is often compared to that of a court trial. The possibilities in such a trial are given below.

- Possible outcomes of testing $H_0 : \text{Defendant} = \text{Not guilty}$:

		Defendant is actually	
		Not guilty	***Guilty***
Court decision	***Not guilty***	Correct	Error
	Guilty	Error	Correct

Errors in hypothesis testing

- In our system of justice, convicting an innocent person is considered worse than letting a guilty person go.

- Possible outcomes of testing $H_0 : \text{Defendant} = \text{Not guilty}$:

		Defendant is actually	
		Not guilty	***Guilty***
Court decision	***Not guilty***	Correct	Error
	Guilty	Worse Error	Correct

Errors in hypothesis testing

- Similarly, there are two types of errors in hypothesis testing.

Type I error

- You could convict an innocent
- i.e., you could reject "not guilty" when the person is truly "not guilty"
- i.e., you could reject the null hypothesis when it is true

Type II error

- You could fail to convict a guilty defendant
- i.e., you could fail to reject "not guilty" when the defendant is guilty
- i.e., you could fail to reject the null hypothesis when it is false

- Like convicting an innocent person, the error of rejecting a true null hypothesis is considered more serious, and so a null hypothesis isn't rejected unless the evidence against it is convincing beyond reasonable doubt.

Errors in hypothesis testing

- Whether your decision is either to reject the null hypothesis or to not reject the null hypothesis, you might be making an error.
  - You make a **Type I error** when you reject a true null hypothesis
  - You make a **Type II error** when you don't reject a false null hypothesis

- Possible outcomes of testing $H_0 : \mu = \mu_0$

		H_0 is actually	
		True	***False***
Your decision	***Do not reject H_0***	Correct	Type II error
	Reject H_0	Type I error	Correct

When is a Type II error worse?

- Sometimes, we are more worried about committing a Type II error than a Type I error.
  - This is application-specific, but it happens less often.
  - You should think about your particular study to see if this is the case.
- Example: developing a rapid test for diabetes:

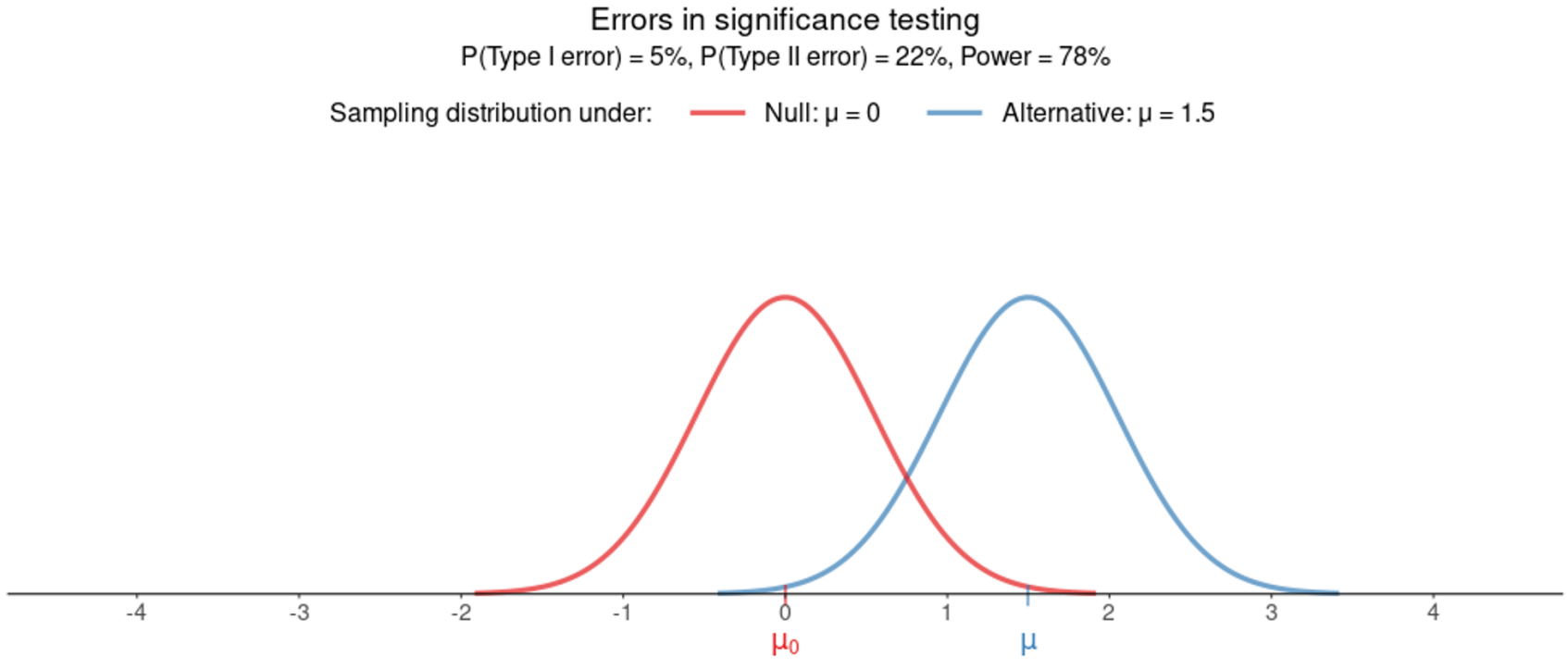
$$H_0 : \text{Person} = \text{Non-diabetic}$$

		Person is actually	
		Non-diabetic	***Diabetic***
Test result	***Non-diabetic***	Correct	Worse Error Type II error
	Diabetic	Error Type I error	Correct

When is a Type II error worse?

- Null hypothesis: Patient is not diabetic
- Alternative hypothesis: Patient is diabetic
- **Type I error** = **False Positive**:
  - a test that indicates a patient has diabetes when in reality they don't.
- **Type II error** = **False Negative**:
  - A test that indicates the patient does not have diabetes when in fact they do have it. That is, a test that fails to detect an actual diabetic.
  - More of a concern, as a person will miss out on treatment.

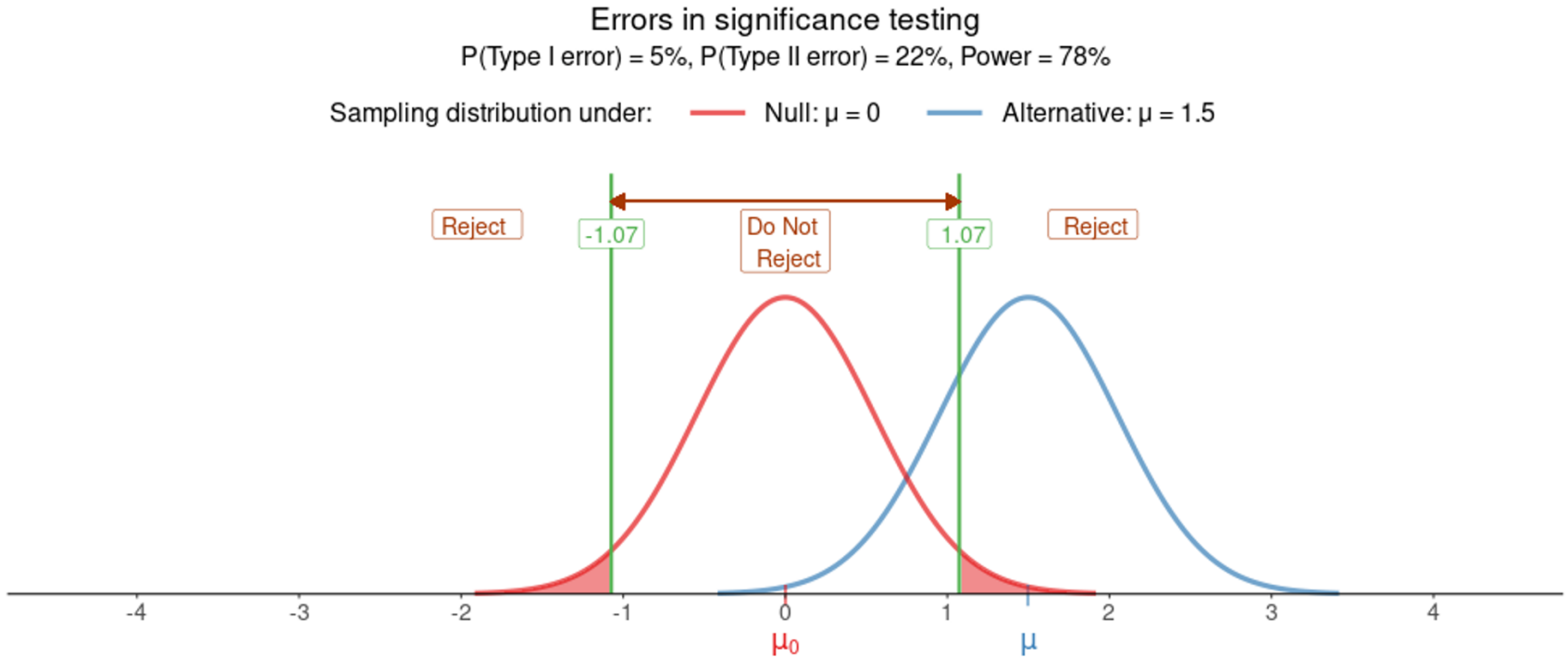
Visually



- Null hypothesis tests the claim that $\mu = 0$
- True value of $\mu = 1.5$

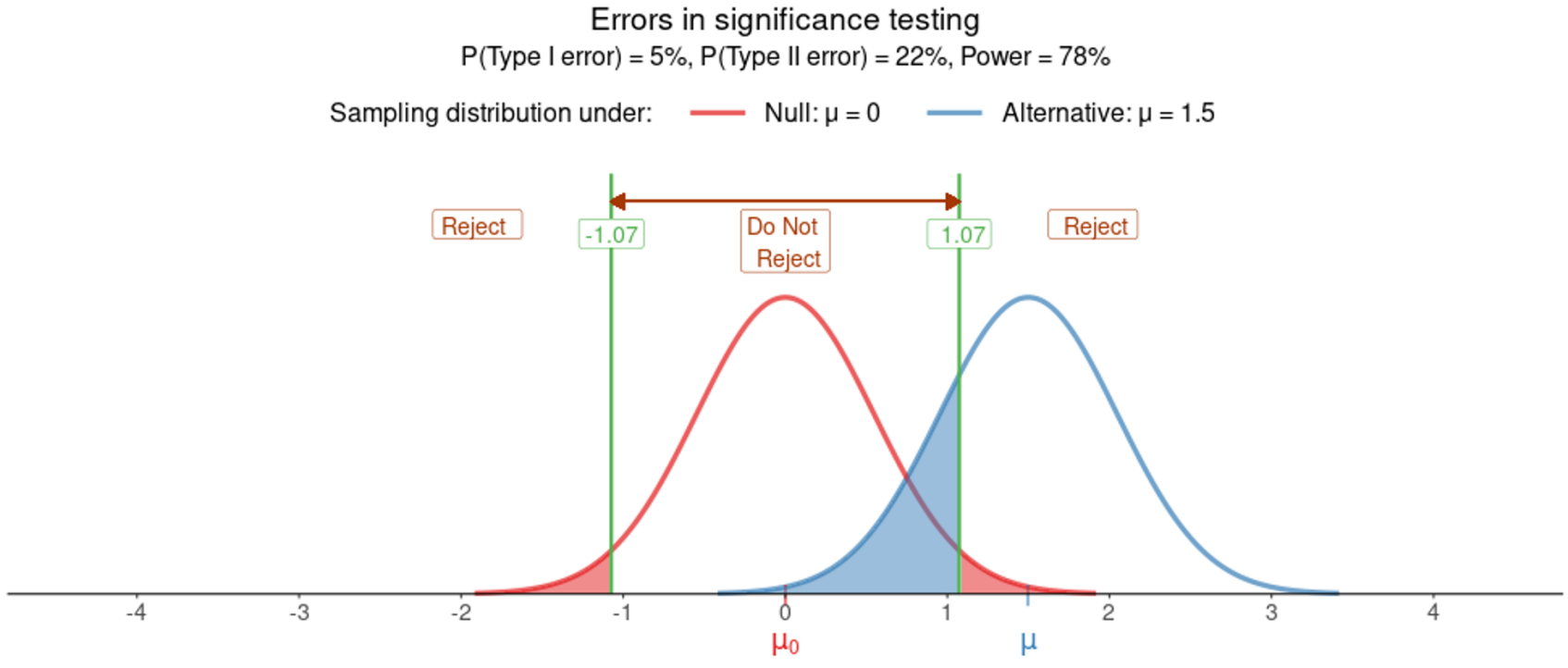
- $n = 30$
- $\sigma = 3$
- $\alpha = 0.05$

Visually



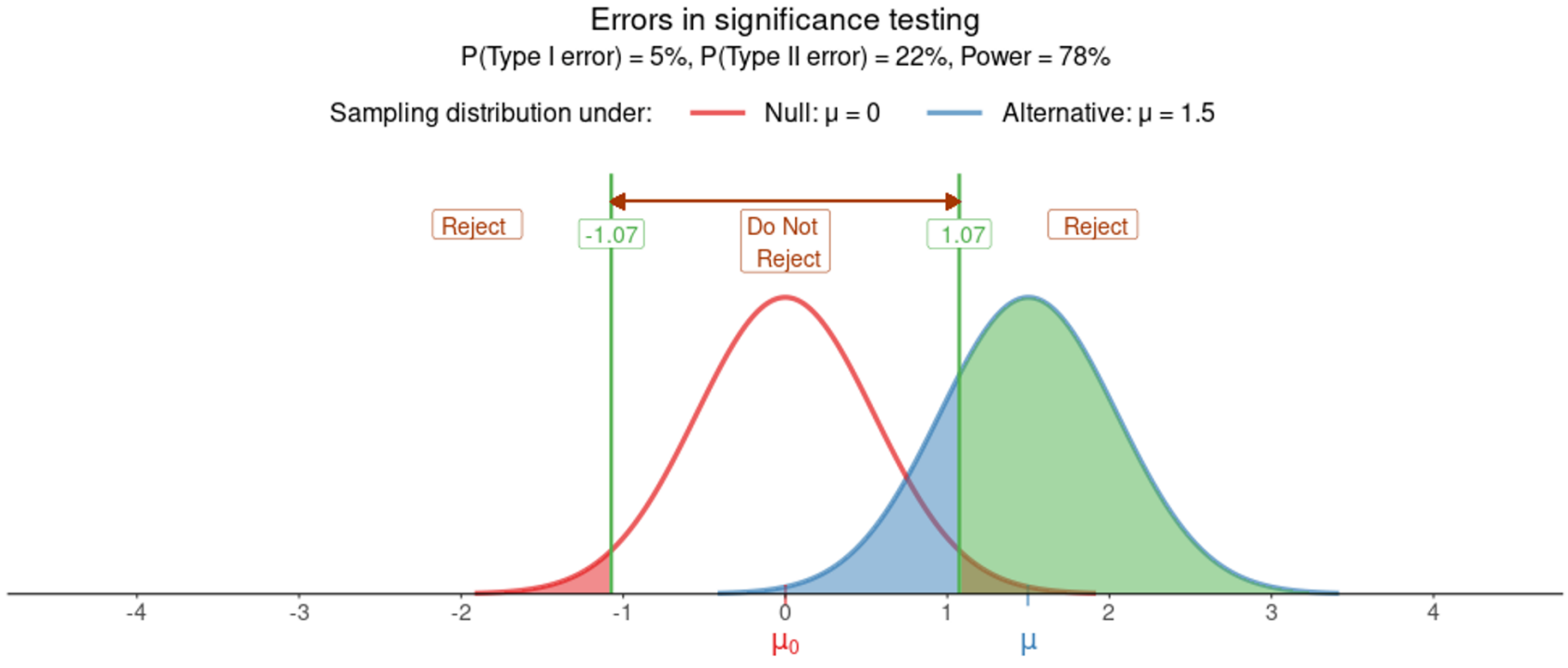
- In red:
  $\alpha = 0.05$ = P(Rejecting H_0 | H_0 is true) = P(Type I error)

Visually



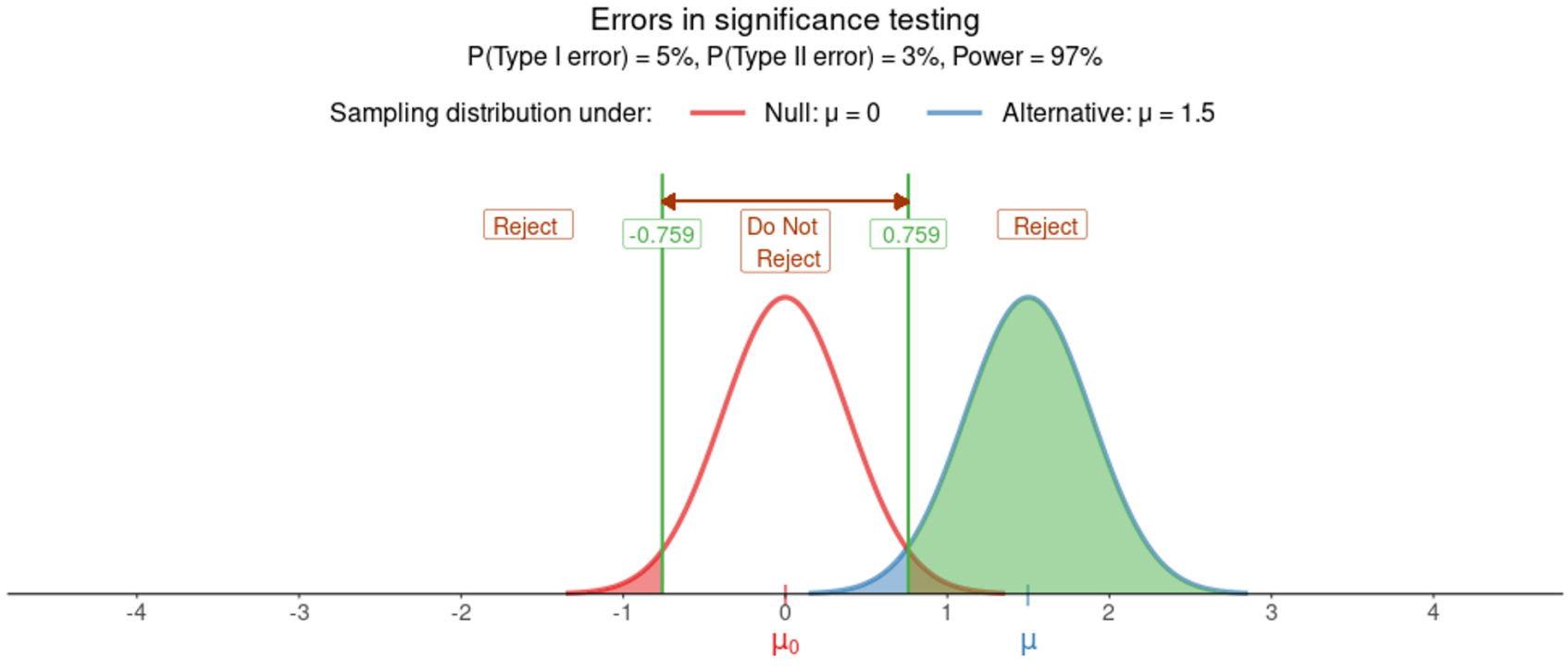
- In blue:
  β = P(Not rejecting H_0 | H_0 is false) = P(Type II error)

Visually



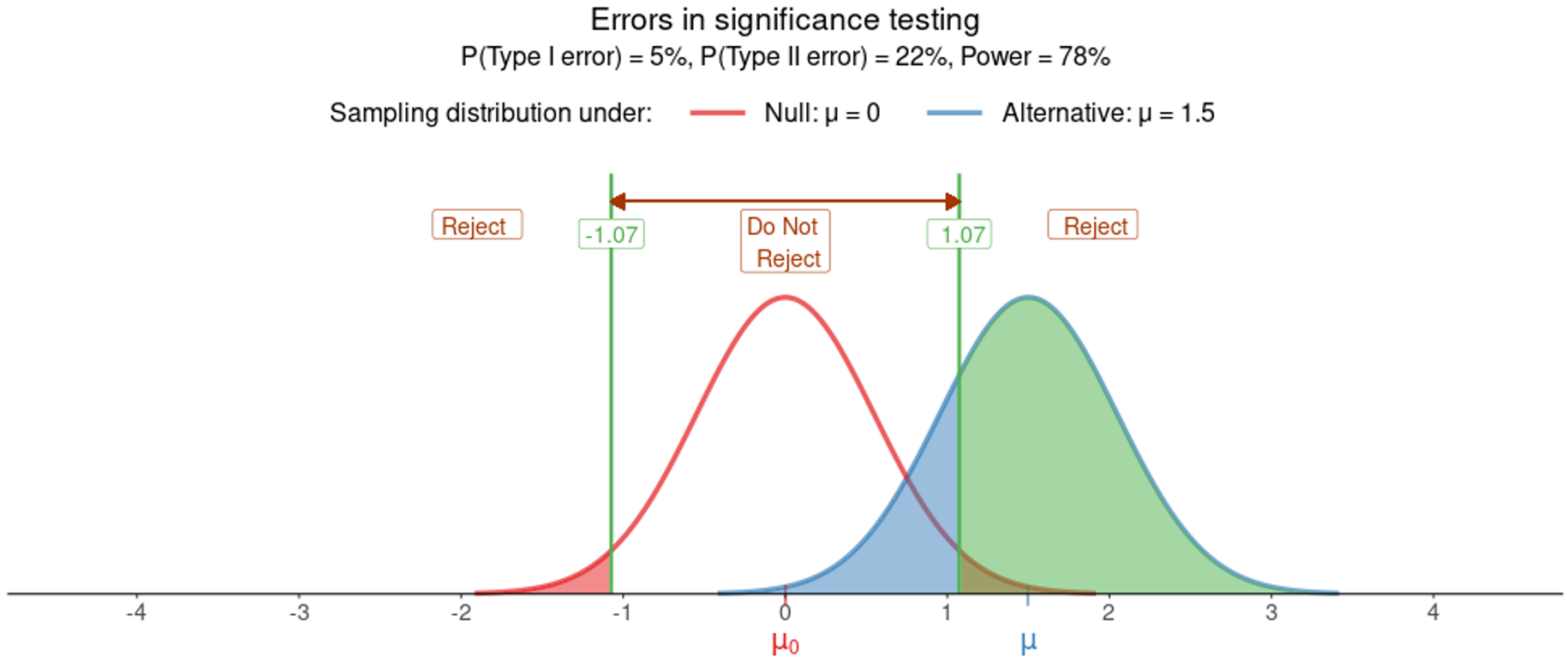
- In green:
  Power = $1 - \beta$ = P(Rejecting H_0 | H_0 is false)

Visually

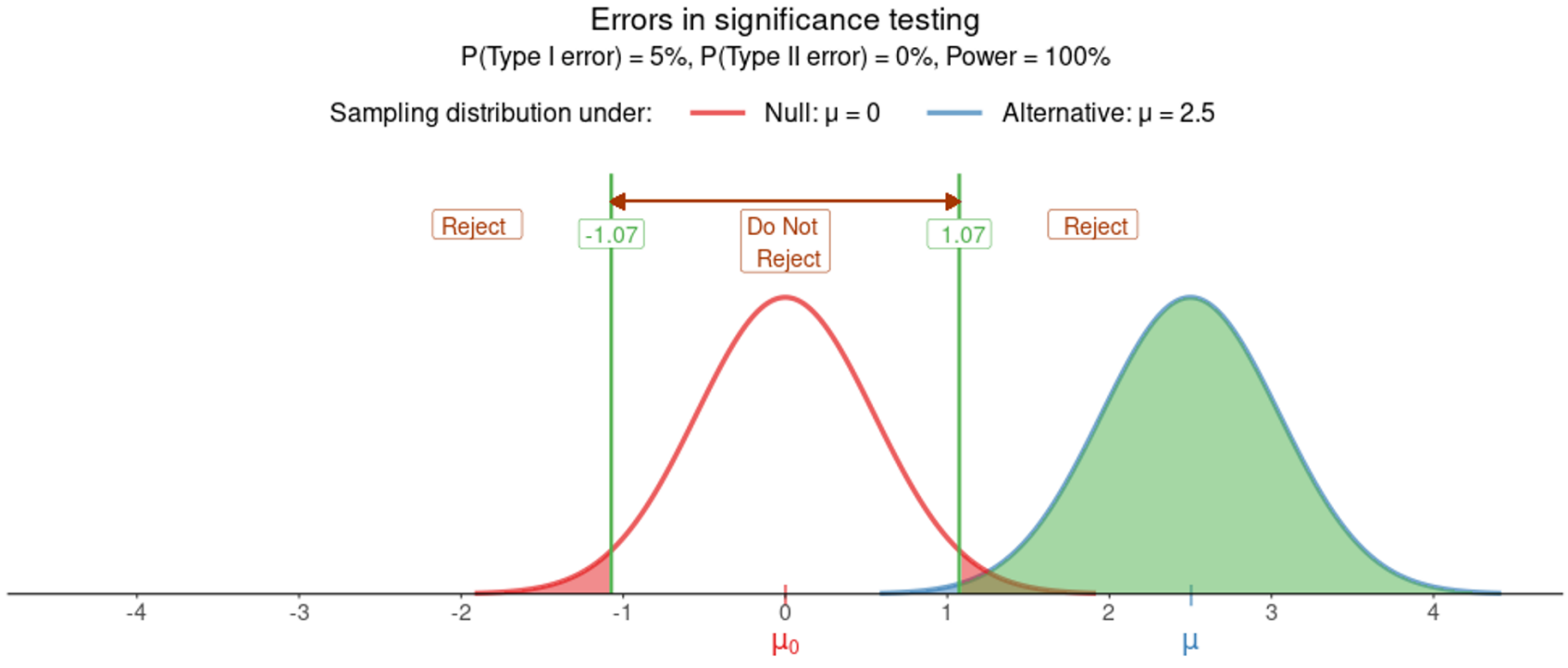


- Effect of increasing the sample size. Remember: the sampling distributions will get narrower because $SE = \frac{\sigma}{\sqrt{n}}$
- In the previous slide $n = 30$, in this slide $n = 60$

Visually

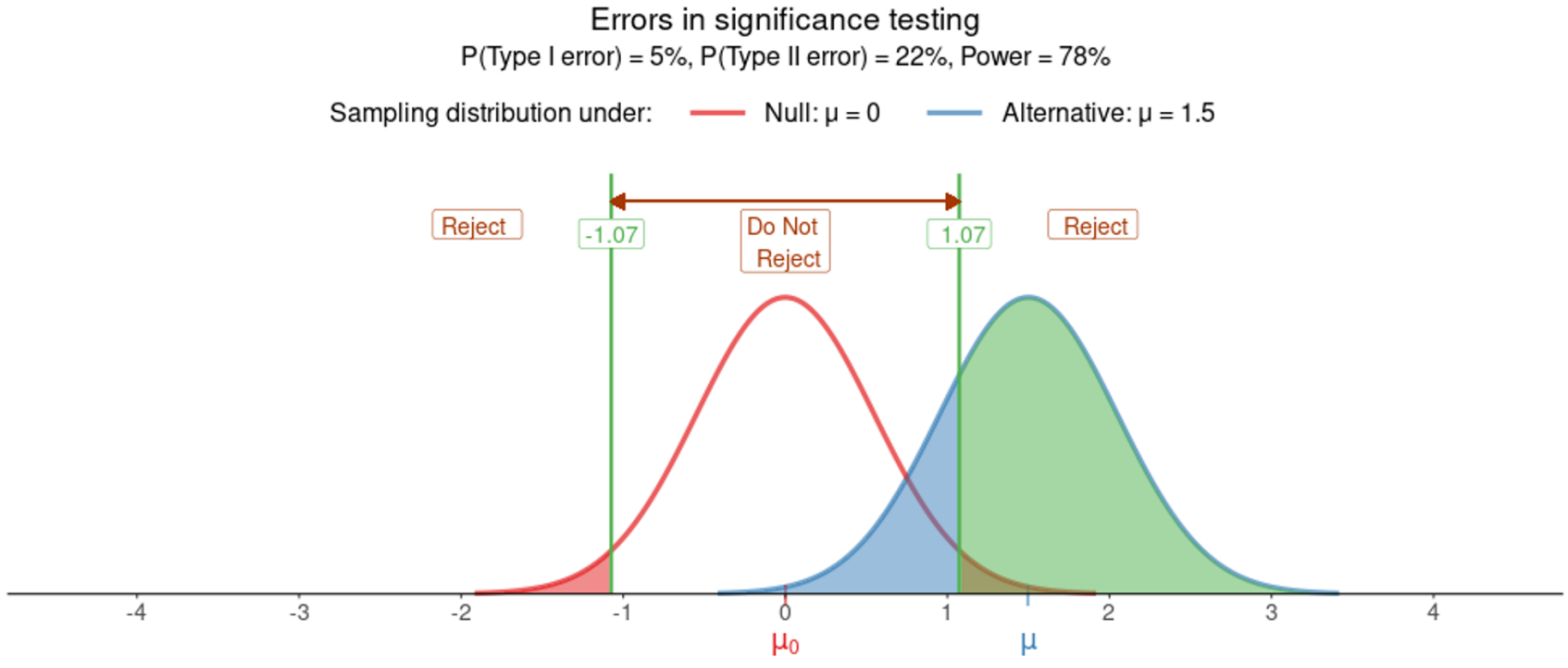

Errors in significance testing
P(Type I error) = 5%, P(Type II error) = 22%, Power = 78%
Sampling distribution under:
Null: μ = 0
Alternative: μ = 1.5
Reject
-1.07
Do Not Reject
1.07
Reject
-4
-3
-2
-1
0
1
2
3
4
μ0
μ

Visually

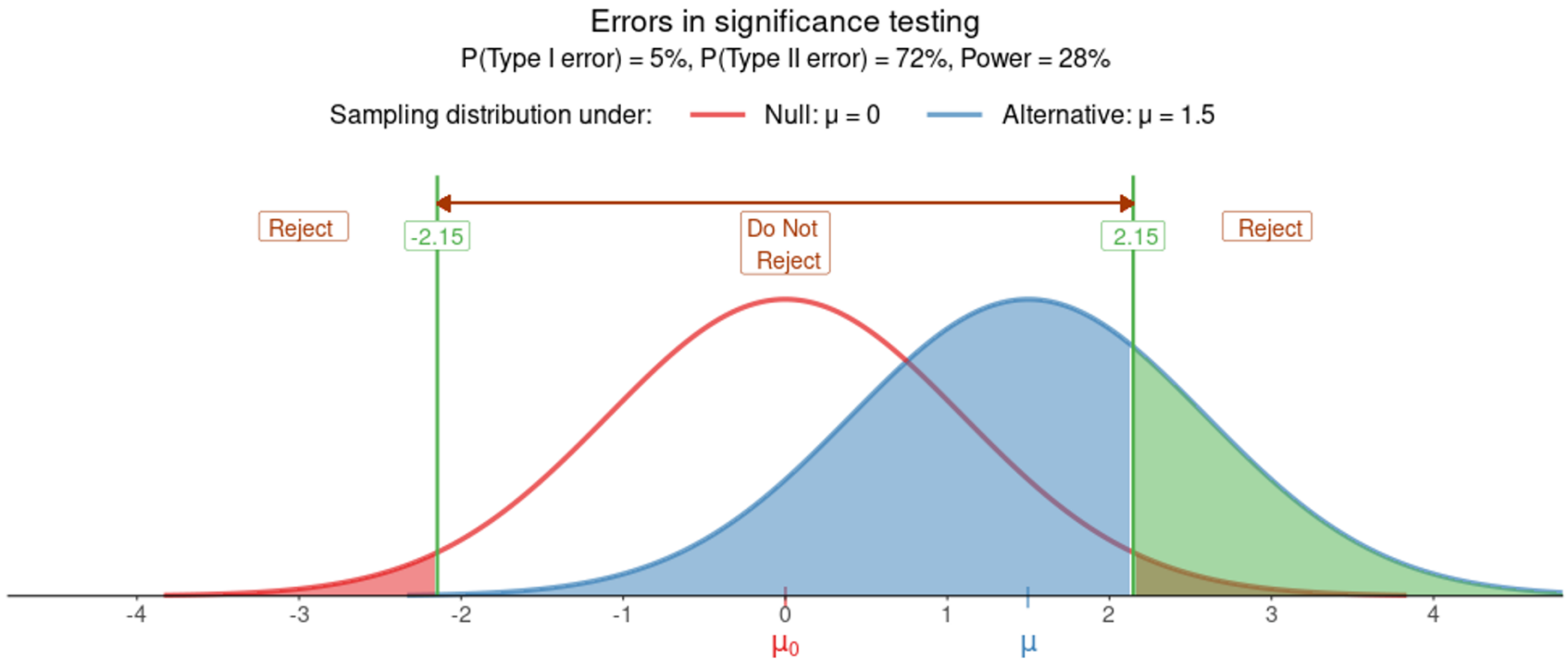


- Effect of increasing the distance between the alternative and the null.
- Power = $1 - \beta$ = P(Rejecting H_0 | H_0 if false)

Visually

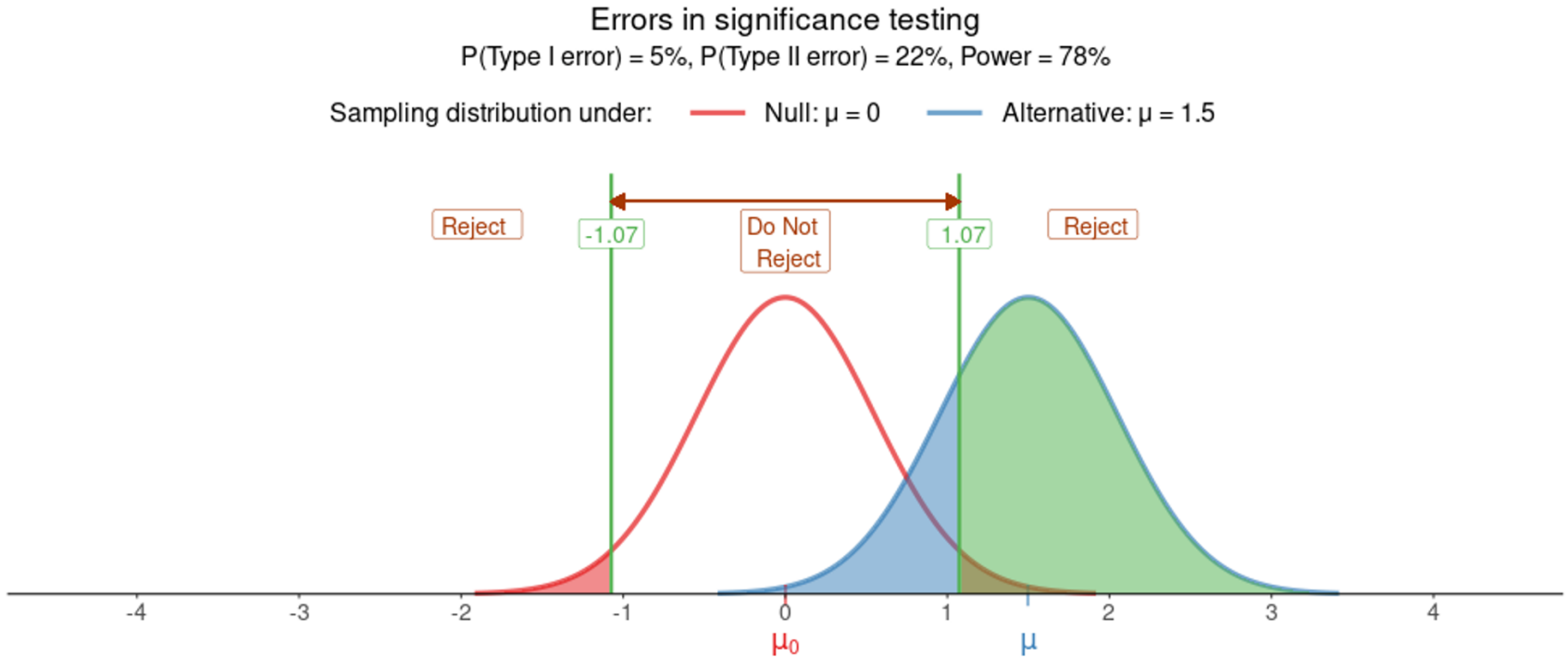

Errors in significance testing
P(Type I error) = 5%, P(Type II error) = 22%, Power = 78%
Sampling distribution under:
Null: μ = 0
Alternative: μ = 1.5
Reject
-1.07
Do Not Reject
1.07
Reject
-4
-3
-2
-1
0
1
2
3
4
μ0
μ

Visually

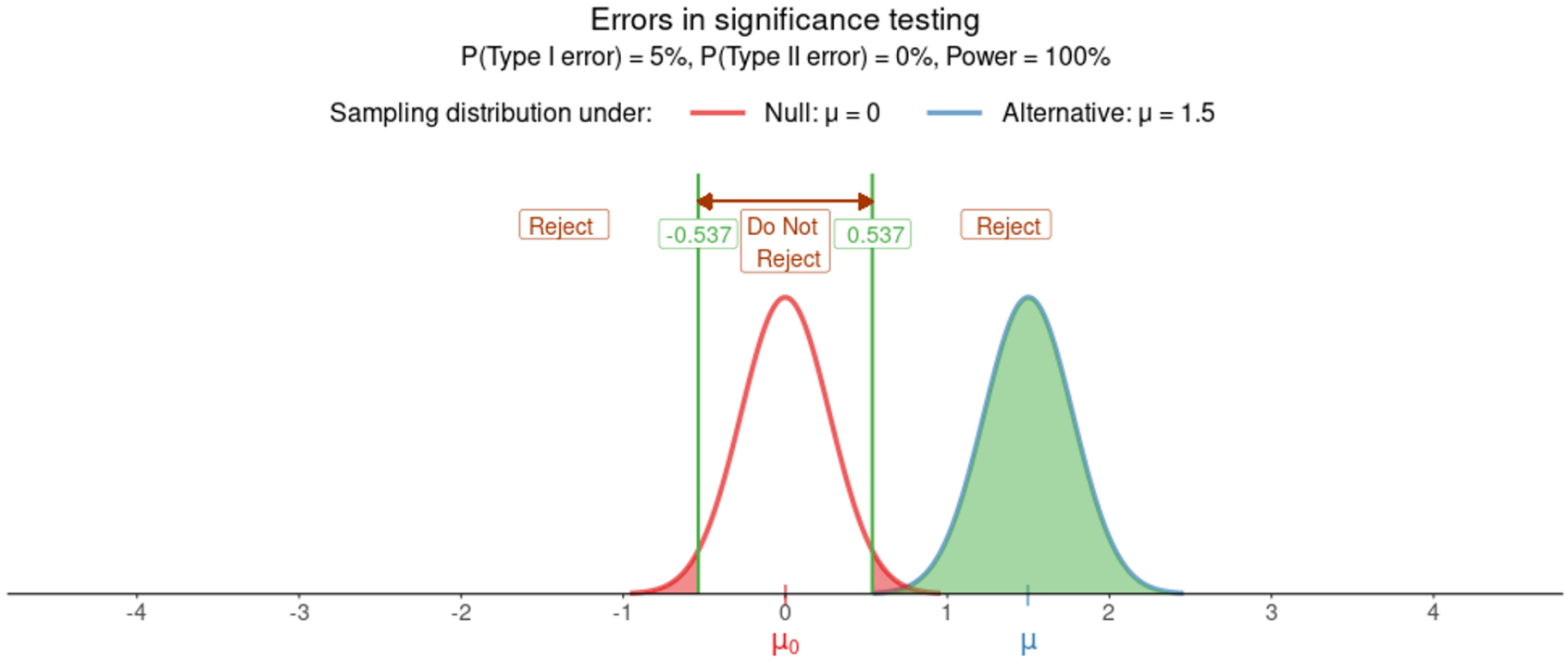


- Effect of increasing the population standard deviation, σ. Here it was doubled.
- Sampling distribution will have a larger spread because $SE = \frac{\sigma}{\sqrt{n}}$.

Visually

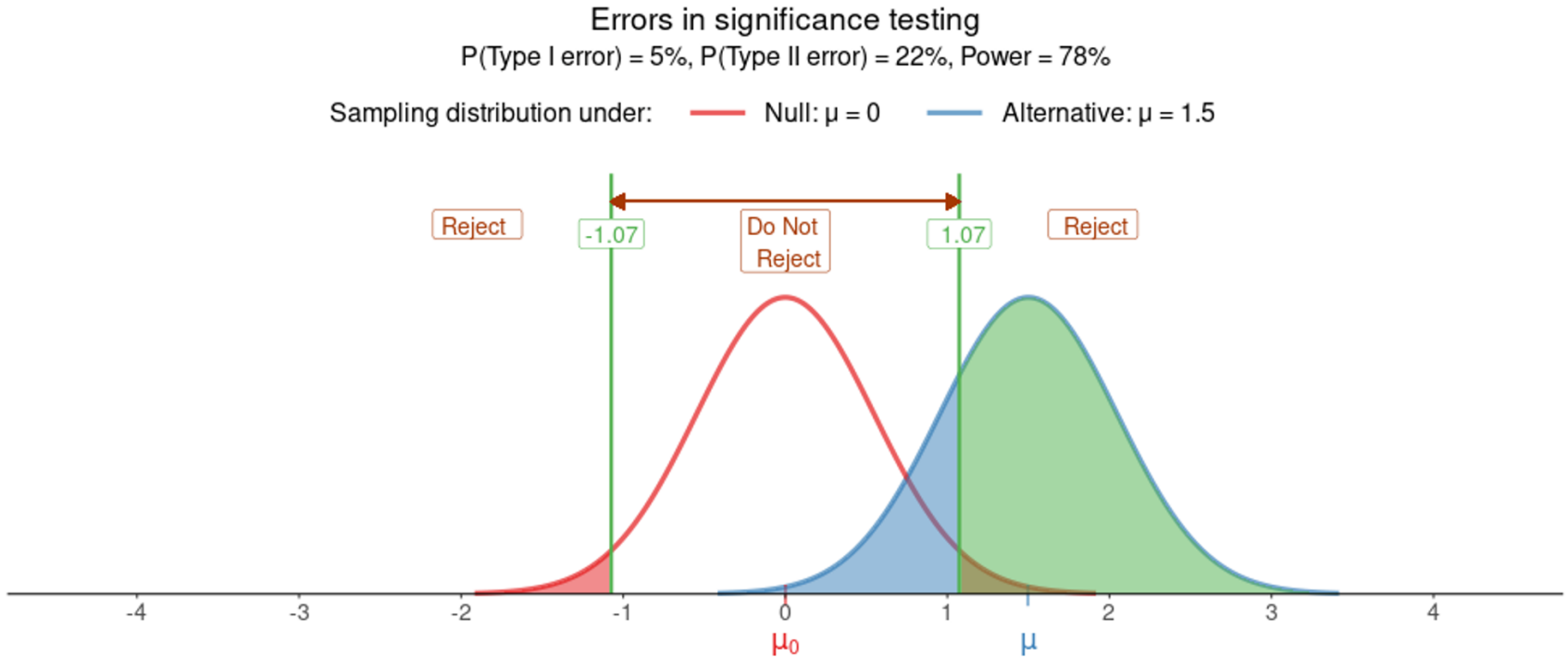

Errors in significance testing
P(Type I error) = 5%, P(Type II error) = 22%, Power = 78%
Sampling distribution under:
Null: μ = 0
Alternative: μ = 1.5
Reject
-1.07
Do Not Reject
1.07
Reject
-4
-3
-2
-1
0
1
2
3
4
μ0
μ

Visually

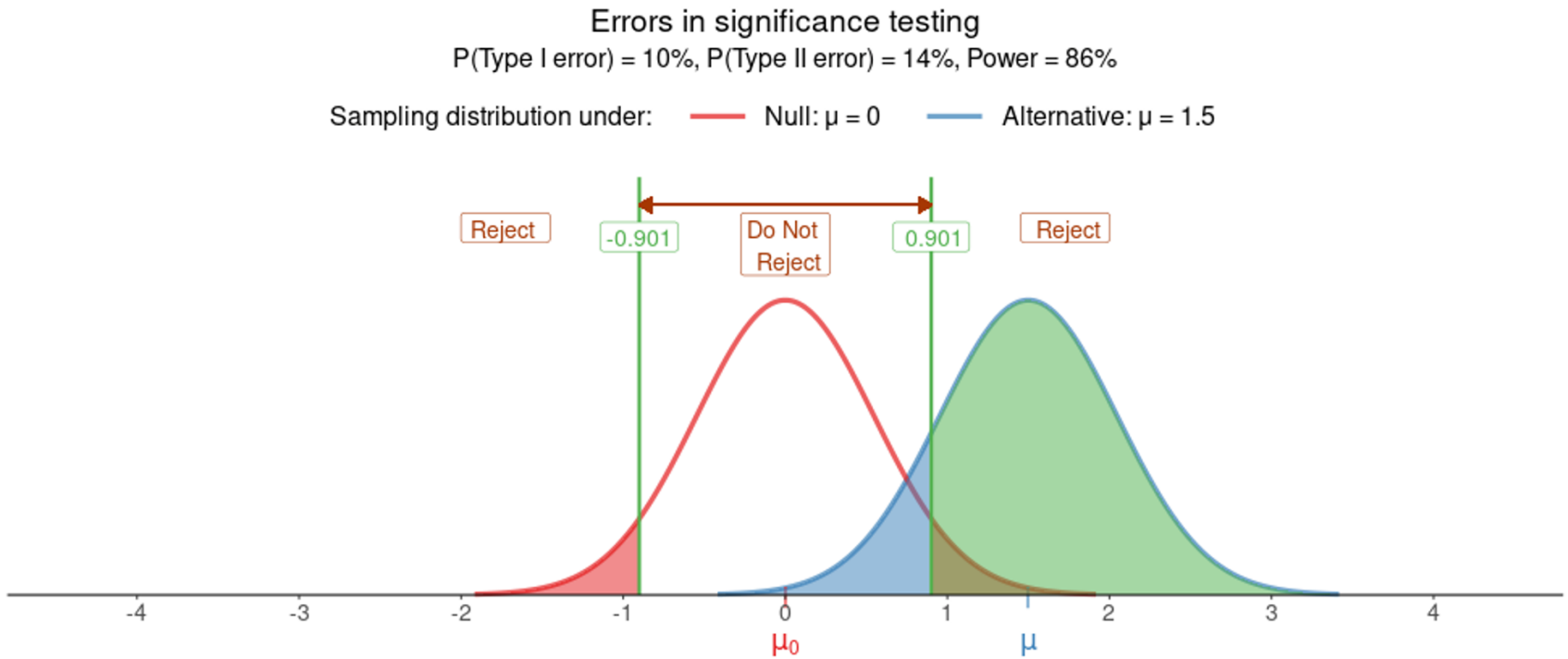


- Effect of decreasing the population standard deviation σ. Here it was halved.
- Sampling distribution will have a lower spread because $SE = \frac{\sigma}{\sqrt{n}}$.

Visually


Errors in significance testing
P(Type I error) = 5%, P(Type II error) = 22%, Power = 78%
Sampling distribution under:
Null: μ = 0
Alternative: μ = 1.5
Reject
-1.07
Do Not Reject
1.07
Reject
-4
-3
-2
-1
0
1
2
3
4
μ0
μ

Visually



- Effect of increasing α from 0.05 (previous slide) to 0.1 (this slide).

In symbols

Probability of Type I error

- The significance level α represents the tolerable probability of committing a Type I error

$$\alpha = P(\text{Reject } H_0 \mid H_0 \text{ is true}) = P(\text{Type I error})$$

- If you are worried about committing a Type I error, then your best strategy is to have a low significance level.

Probability of Type II error

- The probability of committing a Type II error is denoted

$$\beta = P(\text{Do not reject } H_0 \mid H_0 \text{ is false}) = P(\text{Type II error})$$

- If the null hypothesis is false, setting a low significance level increases the probability of making a Type II error.

In symbols

Power

The power of a test is the probability that the test correctly rejects a **false** null hypothesis.

$$\begin{aligned}\text{Power} &= P(\text{Reject } H_0 \mid H_0 \text{ is false}) \\ &= 1 - P(\text{Do not reject } H_0 \mid H_0 \text{ is false}) \\ &= 1 - \beta\end{aligned}$$

Note:

Recall, instead, that the probability of rejecting a **true** null hypothesis is the significance level:

$$\alpha = P(\text{Reject } H_0 \mid H_0 \text{ is true})$$

Be careful to not confuse the two!

Recap

		H_0 is actually	
		True	**False**
Your decision	**Don't reject H_0**	Correct Prob = $1 - \alpha$	Type II error Prob = β
	Reject H_0	Type I error Prob = α	Correct Prob = $1 - \beta$

Factors affecting power

- Power increases as the sample size increases, all else being held constant.

  This is because the distributions of the sample statistics become "narrower", and there will be less statistics on the left of the critical value.

- Power increases as the value of α increases, all else being held constant.

- Power increases when the true value of the parameter is farther from the hypothesised value in the null.

  - I.e., power increases as the effect size increases (more on this later).

- In practice you **cannot** change the distance of the true parameter value from the null, so you can increase power by either taking a larger sample size, or making α larger (the latter however is not good practice).

Significance level and errors

IDEALLY

While we wish to avoid both types of errors ...

IN REALITY

... in reality we have to accept some trade-off between them.

- If we make it very hard to reject H_0, we could reduce the chance of making a Type I error, but then we would make Type II errors more often.
- On the other hand, making it easier to reject H_0 would reduce the chance of making a Type II error, but increase the chance of making a Type I error and we would end up rejecting too many H_0's that were actually true.
- This balance is set by how easy or hard it is to reject H_0, which is exactly determined by the significance level!

Part B

Effect size

Effect size

- Effect size is related to the magnitude of the the difference between the true population mean μ and the hypothesised value μ_0
- We saw that a statistically significant result may not be important at all, i.e. may not have much real-world value.
    - Importance is related to the practical distance between the hypothesised value and the true population mean. I.e., it is related to the effect size.
    - In practice we don't know the true value of the population mean μ, so to calculate effect size we typically replace μ with its estimate $\bar{x}$.
    - A difference between $\bar{x}$ and μ_0 can be statistically significant and yet be too small in actual units to be of much importance
    - Remember in the body temperature example the sample mean was $\bar{x} = 36.81$ °C, which was found to be significantly different from the hypothesised value of 37 °C. However, the difference is tiny and not important in practice.

Formal effect size index: Cohen's D

- Cohen's D was introduced as a measure of "effect size", to report whether the result may be of real-world value or not.
- Consider a one-sample t-test for a population mean:

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

- For a test of **one population mean**, Cohen's D is defined as:

$$D = \frac{\bar{x} - \mu_0}{s}$$

that is, the difference between the sample and the hypothesised mean, measured in units of the **standard deviation**. (Careful: not the standard error!)

Effect size

- Below are some rough guidelines on how to interpret the size of the effect.
- These are not exact labels, but a loose guidance based on empirical research.

Verbal label	Magnitude of D in absolute value
Small (or weak)	≤ 0.20
Medium (or moderate)	≈ 0.50
Large (or strong)	≥ 0.80

- What about in between the categories?
    - Use "small to medium" or "medium to large" effect size.

Why not simply the difference?

- Why not just the difference $\bar{x} - \mu_0$?
- It depends on the units of measurement of the data.
- **Scenario 1**:

$$d = 5 - 3 = 2$$

- **Scenario 2**:

$$d = 500000 - 499998 = 2$$

- Clearly a difference of $d = 2$ in Scenario 1 is more substantial and of higher practical impact. However, a difference of $d = 2$ in Scenario 2 is less substantial and of pretty much no practical impact.
- Dividing the difference by the SD of the data (which is in the same unit of measurement as the data itself), gives you a measure that does not depend on the unit of measurement. Hence why you should use Cohen's D.

Part C

Example on power

Example on power

- Suppose the population standard deviation is $\sigma = 5$ and you will take a sample of size $n = 15$.
  - The $SE = \sigma/\sqrt{n} = 5/\sqrt{15} = 1.291$
- The true population mean is $\mu = 3$.
  - The sampling distribution of the mean will be $N(3, 1.291)$
- You want to perform a test to check whether the population mean is 0 or different from 0.

$$H_0 : \mu = 0 \quad \text{vs} \quad H_1 : \mu \neq 0$$

  - The value 0 corresponds to μ_0 in the null hypothesis.
  - The sampling distribution of the mean will be $N(0, 1.291)$
- We do not reject H_0 if the observed mean falls in the middle 95% of the $N(0, 1.291)$ distribution:

```
qnorm(c(0.025, 0.975), mean = 0, sd = 1.291)
```

```
## [1] -2.53  2.53
```

Example on power

- What's the power of the test?
- I.e. what's the probability of rejecting H_0 when H_0 is indeed false?
- If H_0 is false, then the sample means follow a $N(3, 1.291)$ distribution
- So it's the probability (in that distribution) to the right of 2.53

```
1 - pnorm(2.53, mean = 3, sd = 1.291)
```

```
## [1] 0.6421
```

```
pnorm(2.53, mean = 3, sd = 1.291, lower.tail = FALSE)
```

```
## [1] 0.6421
```

- Power = 0.64

Part D

The t-test assumptions

The t-test assumptions

- Check the technical conditions before reporting and interpreting any t-test results. If those are violated, the results may be incorrect.
- The results from a t-test for a population mean are valid when:
    1. The obtained sample data are a random sample from the population of interest
        - (This is called "independence" by some authors.)
    2. **Either** the population follows a normal distribution **or** the sample size is sufficiently large ($n \geq 30$ as a guideline)
        - (This is called "normality" by some authors, but the goal is normality of the sampling distribution of the mean)

Random sample

- Consider whether the sample was randomly selected from the population of interest before generalising the test conclusion to that population.
    - Each unit should have been sampled independently of the others.
    - The sample should be representative of the population to avoid sampling bias.

Normality

- Normality
- But of what?
- **The sampling distribution of the sample mean**! This is what we ultimately want to follow a normal distribution.
- All our formulas for confidence intervals and hypothesis testing started from the prerequisite that, when the population data are known, the sampling distribution of the mean is normal:

$$\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

- From this prerequisite, we derived a similar distribution to the standard normal distribution, called the **t-distribution**. We used the t-distribution when the population data are not known.

Key question

- When is the sample mean normally distributed?

The t-test assumptions

Key question

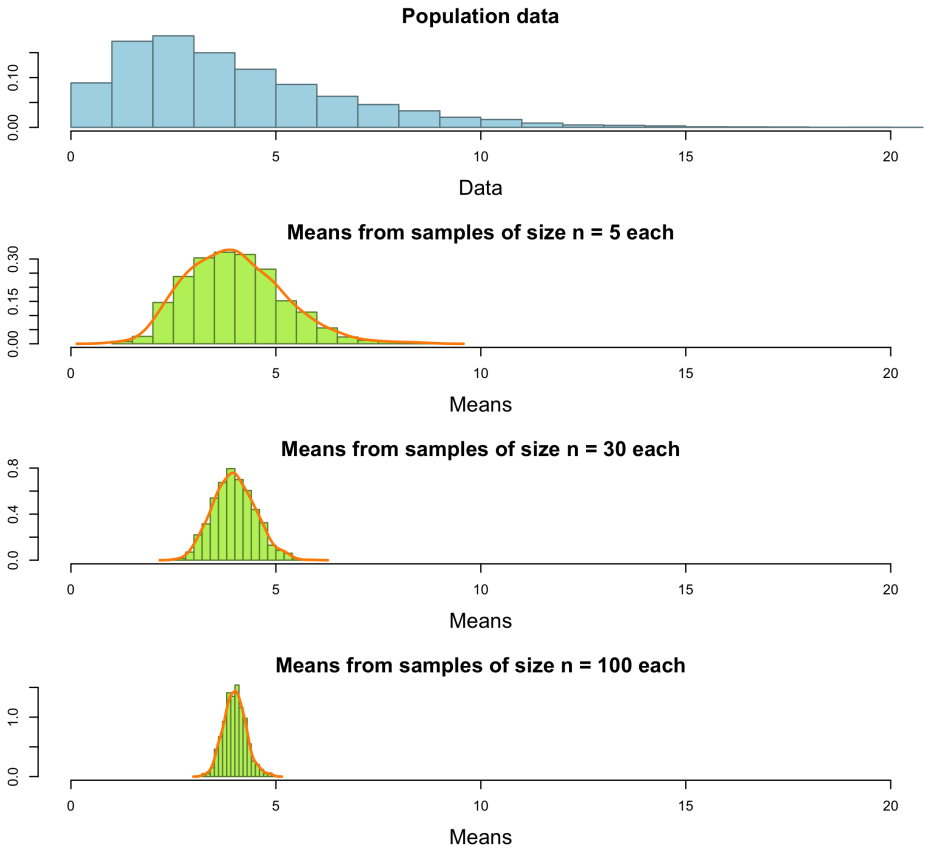
- When is the sample mean normally distributed?

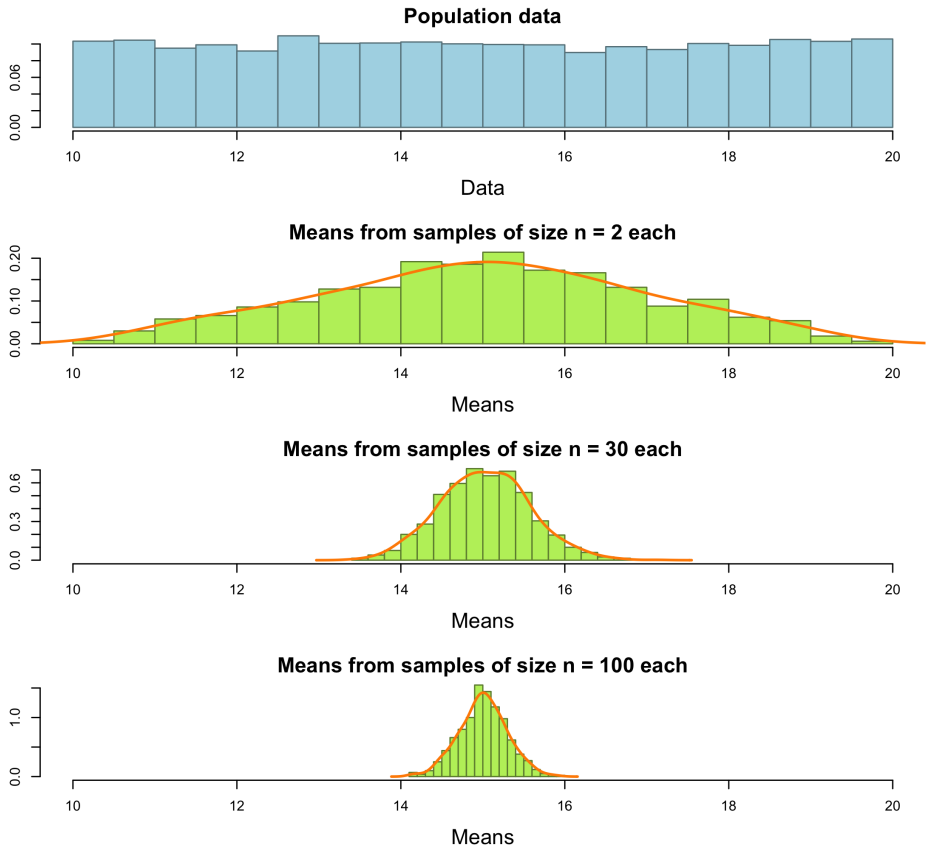
Answer

When EITHER one of these holds:

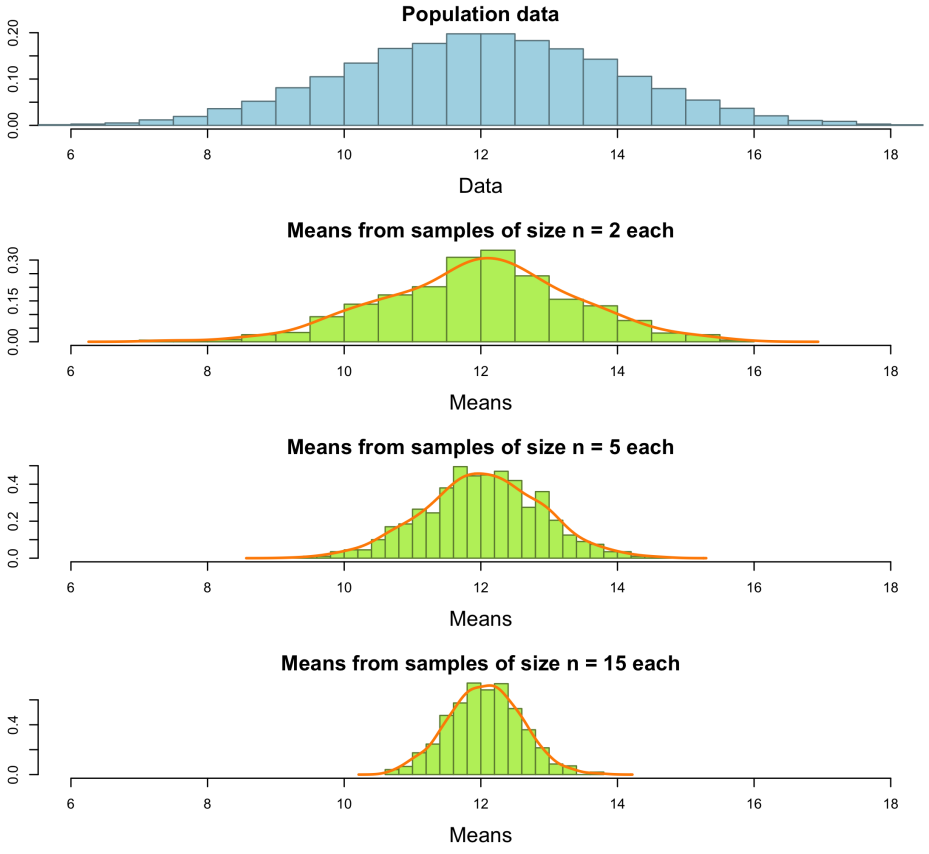
- The sample size is large enough ($n \geq 30$ as a guideline)
    - irrespectively of the distribution of the population data
- The population data follow a normal distribution
    - irrespectively of the sample size

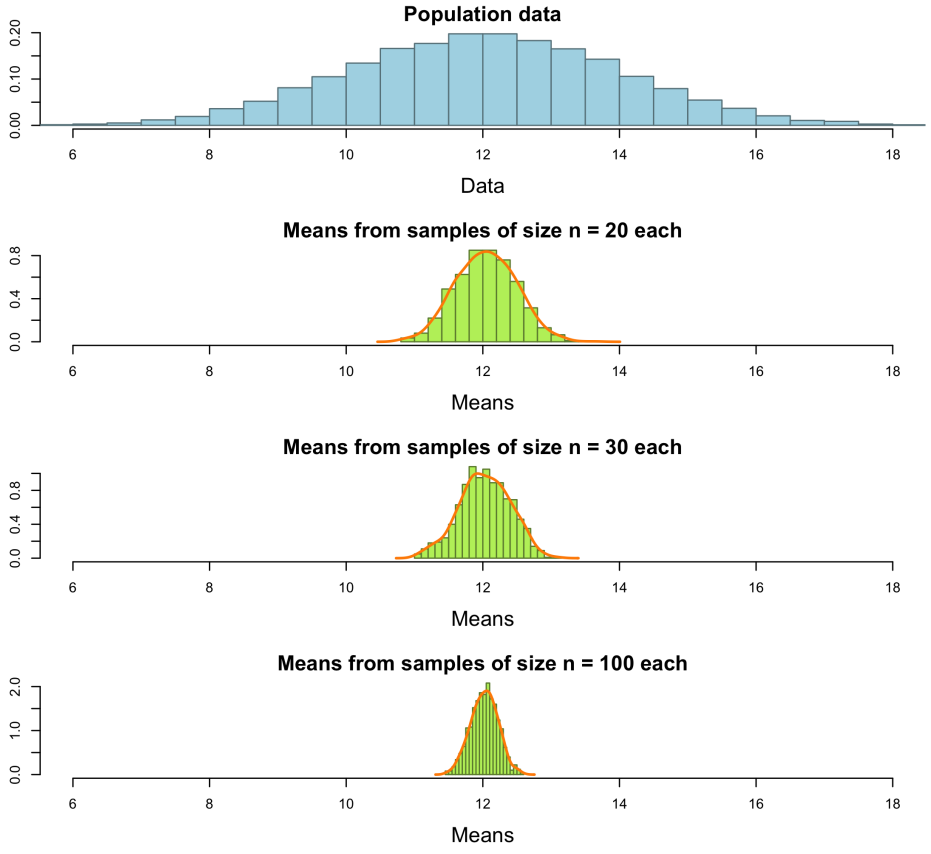
Large enough sample size





Population data normally distributed





Checking for normality of the population data

- Example data: a sample of 20 IQ scores:

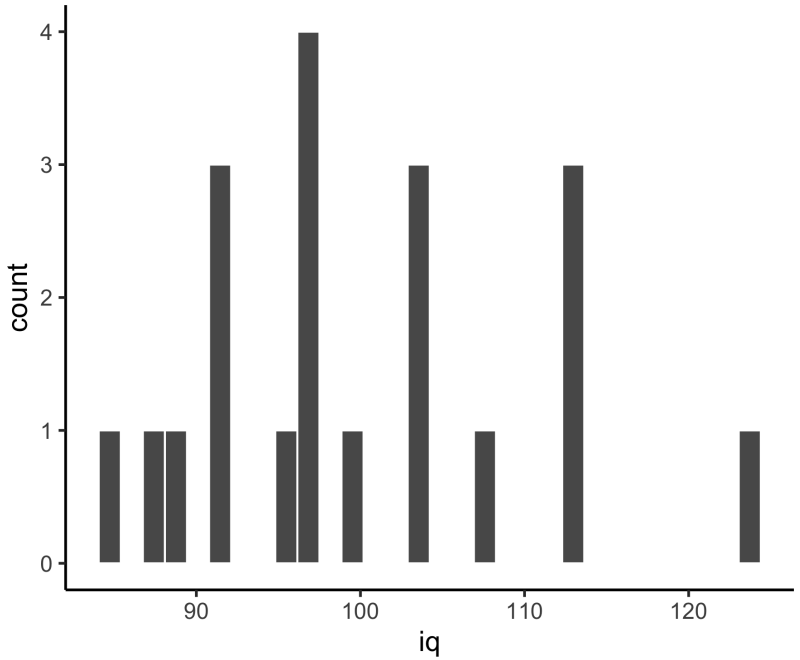
```
head(iq_sample)
```

```
## # A tibble: 6 × 1
##      iq
##   <dbl>
## 1   113
## 2    97
## 3   113
## 4   113
## 5   104
## 6    85
```

Checking for normality of the population data

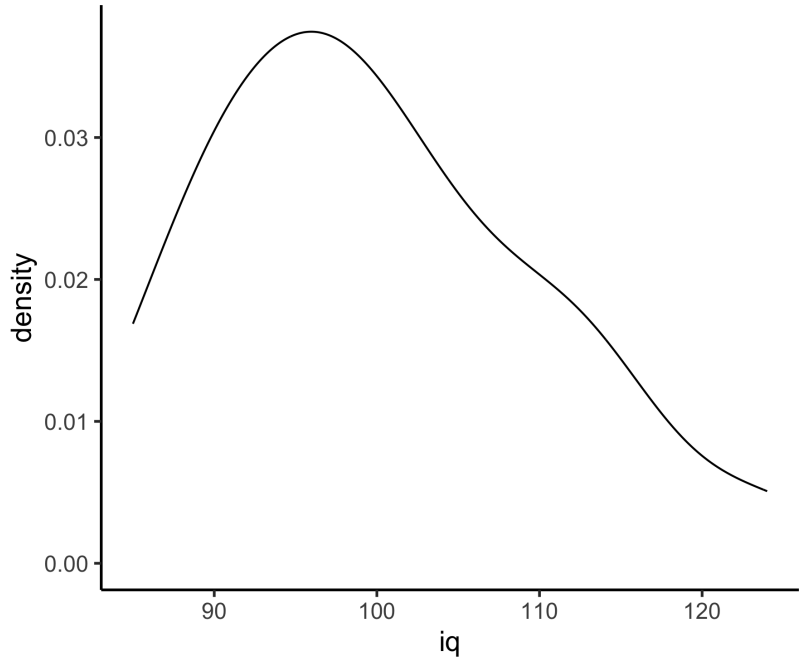
Histogram

```
ggplot(iq_sample, aes(x = iq)) +
    geom_histogram(color = 'white')
```



Density plot

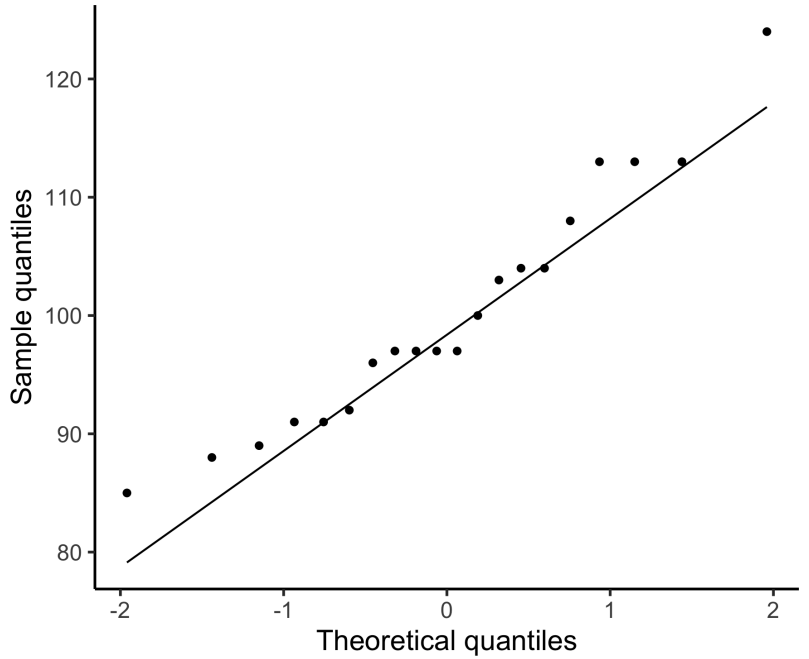
```
ggplot(iq_sample, aes(x = iq)) +
    geom_density()
```



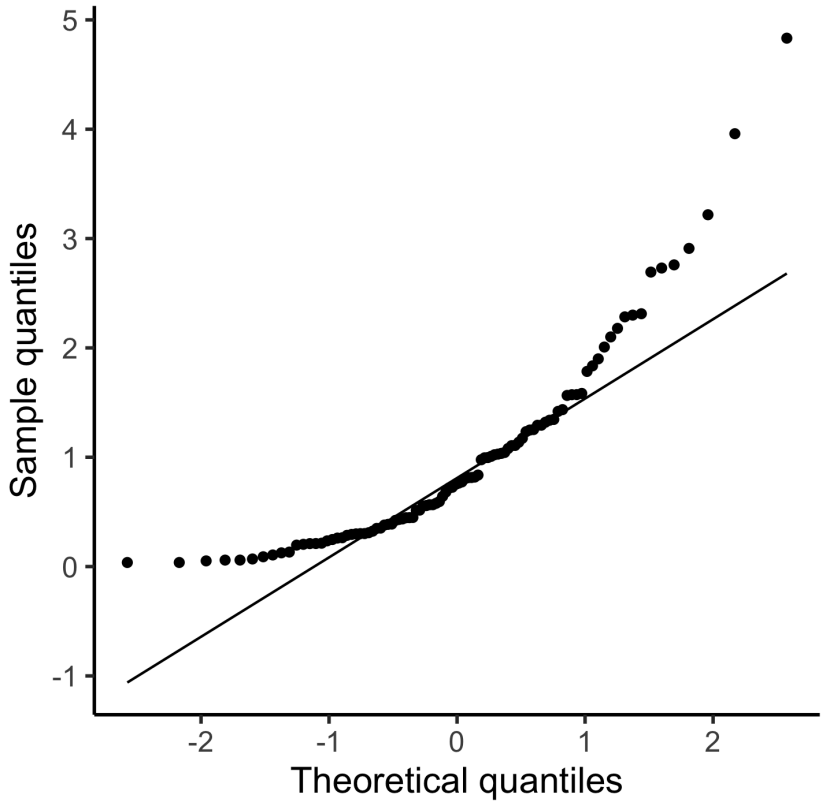
Checking for normality of the population data

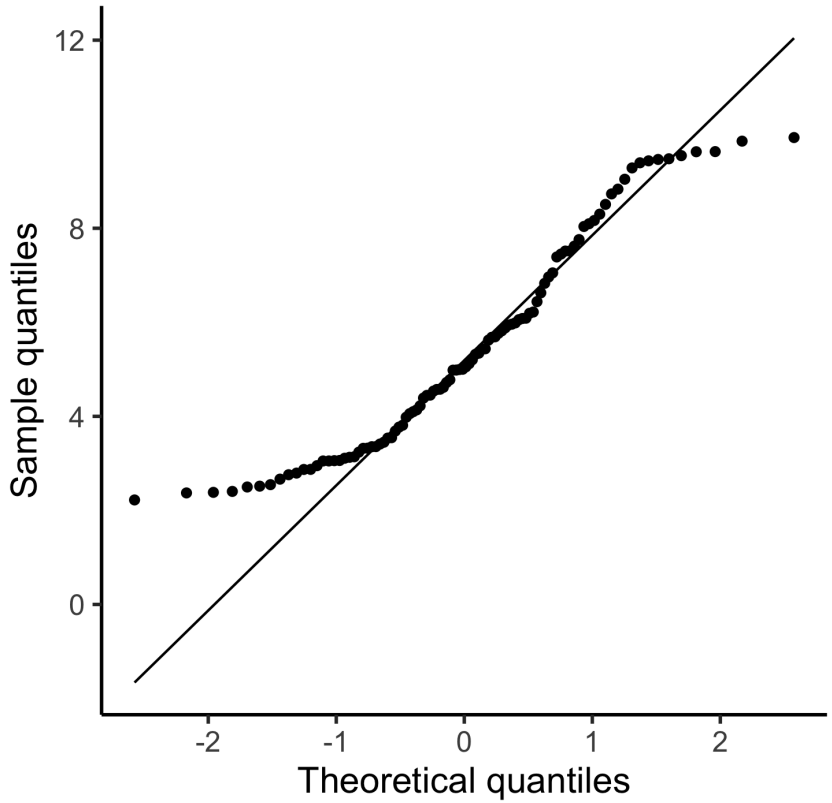
Quantile-Quantile plot (qq-plot): the points should roughly follow the line.

```
ggplot(iq_sample, aes(sample = iq)) +
    geom_qq() +
    geom_qq_line() +
    labs(x = "Theoretical quantiles",
         y = "Sample quantiles")
```



Problematic QQ-plots





Shapiro-Wilk normality test

Another hypothesis test, but this time:

- the null hypothesis states that the population data follow a normal distribution
- the alternative hypothesis states that the population data do not follow a normal distribution

```
shapiro.test(iq_sample$iq)
```

```
##
##     Shapiro-Wilk normality test
##
## data:  iq_sample$iq
## W = 0.95, p-value = 0.3
```

At the 5% significance level, we performed a Shapiro-Wilk test against the null hypothesis of normality of the population data: $W = 0.95, p = 0.3$. The sample data did not provide sufficient evidence to reject the null hypothesis of normality in the population.