# The Significance Level as a Long-Run Error Rate

NHH
TECH3

**Significance level**

Reject $H_0$ if $P - value \leq \alpha$

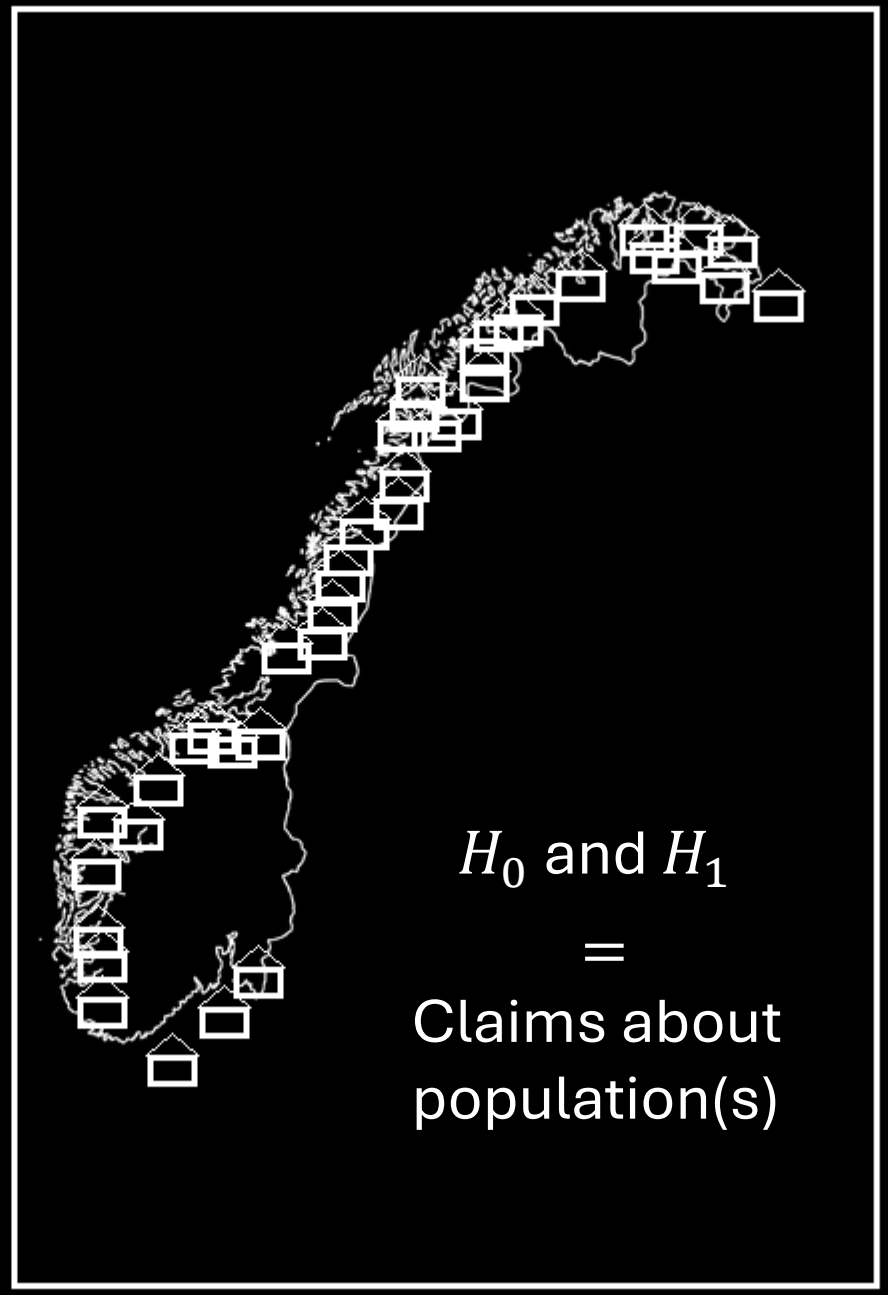
NHH
TECH3

When doing hypothesis testing..

# MISTAKES WILL SOMETIMES BE MADE

NHH
TECH3

Population

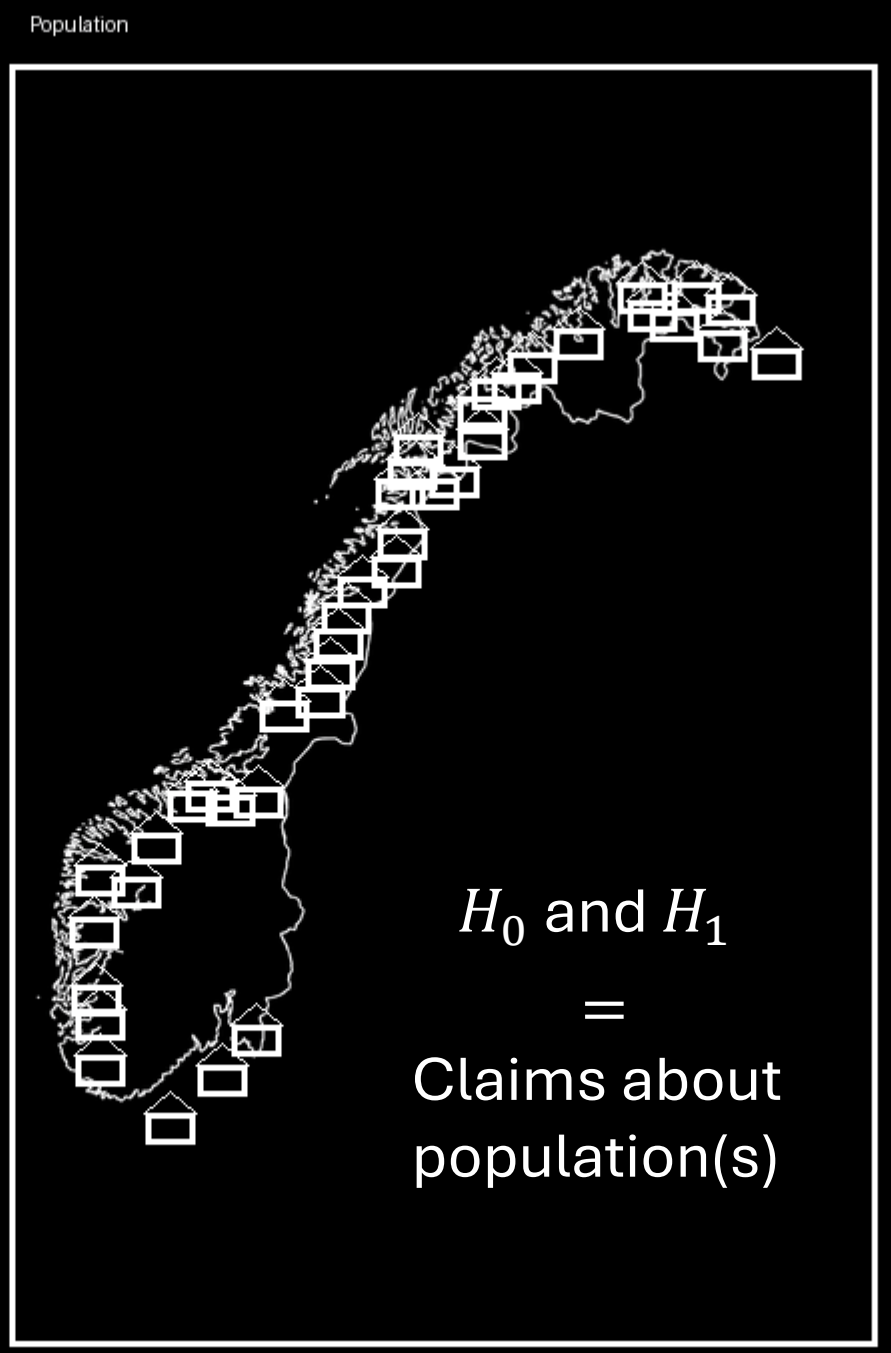


$H_0$ and $H_1$

=

Claims about population(s)

Our conclusions about $H_0$ and $H_1$ are based on:



J. Neyman and Pearson 1933:

“…Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not often be wrong”

NHH
TECH3

States of reality:

Possible decisions

| | $H_0$ is true | $H_0$ is false |
|---|---|---|
| Reject $H_0$ | Type I Error | Correct decision |
| Retain $H_0$ | Correct decision | Type II Error |

NHH
TECH3

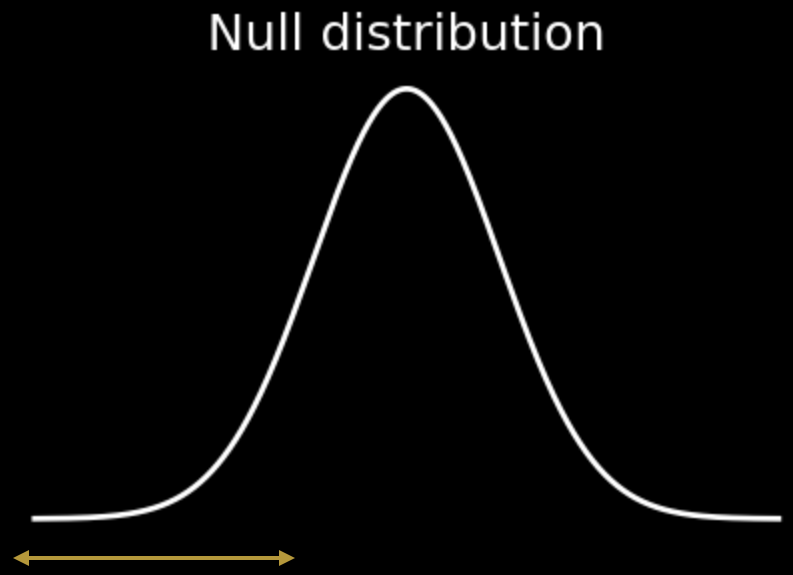
Decision rule:

Reject $H_0$ if $P - value \leq \alpha$

NHH
TECH3

Decision rule:

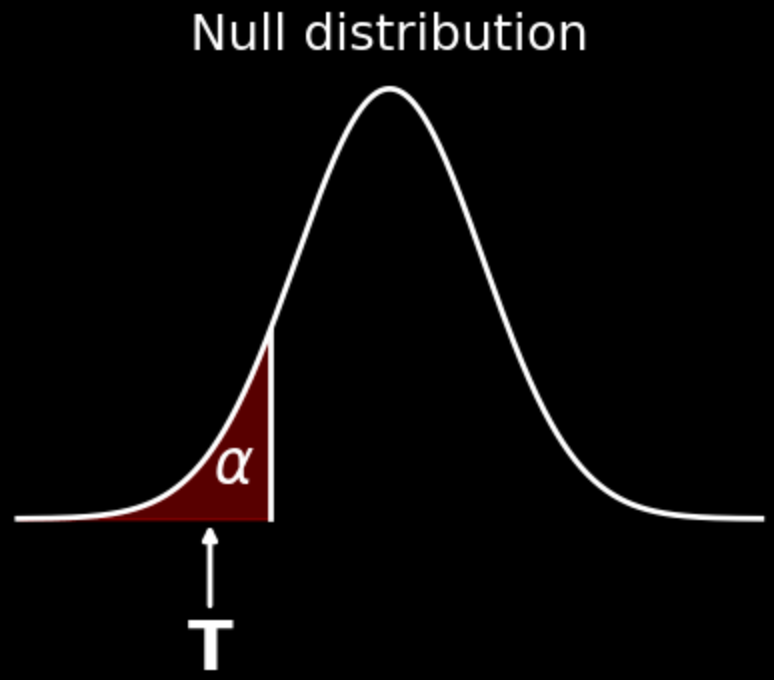
Reject $H_0$ if $P - value \leq \alpha$




NHH
TECH3

Decision rule:

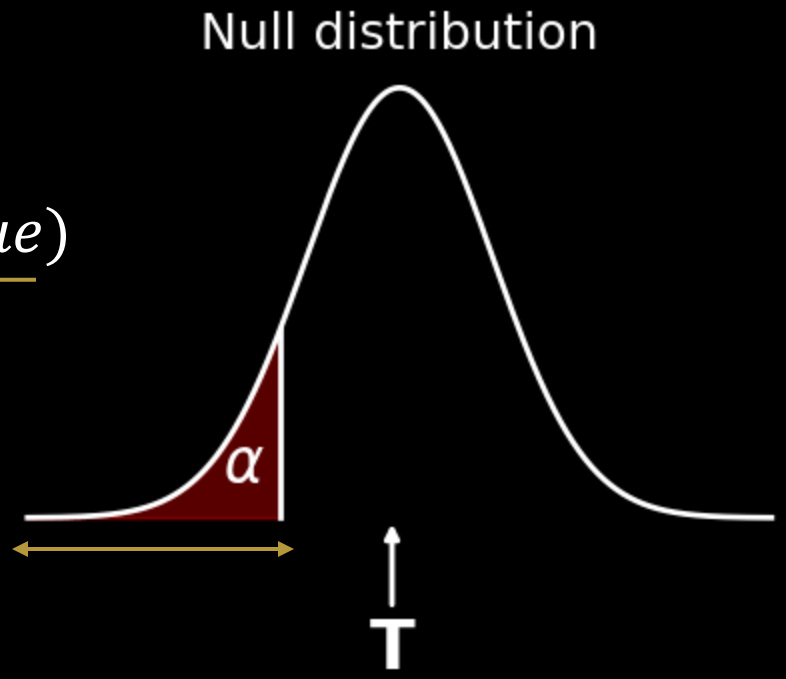
Reject $H_0$ if $P - value \leq \alpha$



$P(Type\ I\ Error) = P(Reject\ H_0 | H_0\ true)$

Decision rule:

Reject $H_0$ if $P - value \leq \alpha$

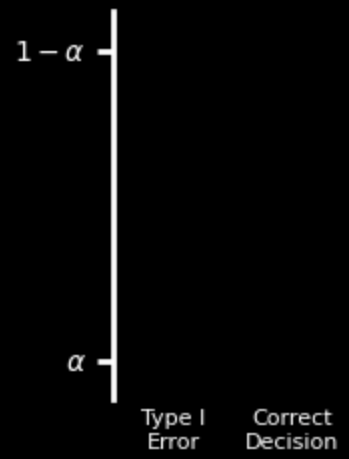


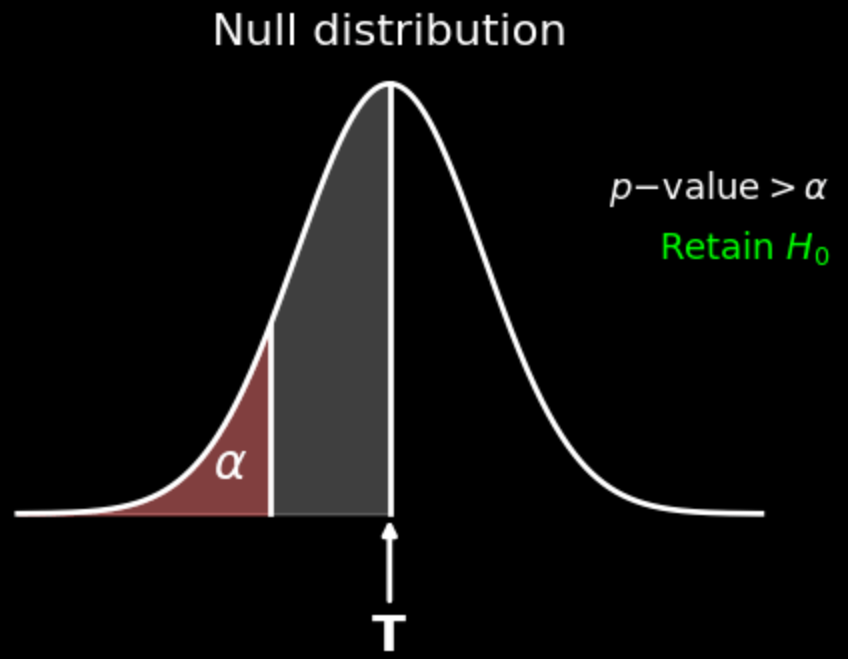

If $H_0$ is true:

1. Reject $H_0$ = Type I Error
2. T follows the Null distribution

$\Rightarrow P(Type\ I\ Error) = \alpha$

NHH TECH3




$$\Rightarrow P(Type\ I\ Error) = \alpha$$

NHH
TECH3

**Hypotheses:**

$H_0$ : The batch of airbags is **defective**.

$H_A$ : The batch of airbags is **safe.**

**Action:** If we reject $H_0$ the batch is approved for installation.

Type I error: Approving a batch of defective airbags → **Risk of injury or death**.

Very low $\alpha$ → Ensures defective airbags rarely pass quality checks.

NHH TECH3

**Hypotheses:**

$H_0$ : The batch of airbags is **defective**.

$H_A$ : The batch of airbags is **safe.**

**Common misconception:**

• "A proportion α of all approved batches are defective."

**The significance level is a conditional error rate:**

• α is the proportion of *defective batches that we mistakenly approve* in the long run.

NHH TECH3

**1000 A/B tests with hypotheses:**

$H_0 : p_{new} = p_{current}$

$H_A : \ p_{new} > p_{current}$

$$1000 \times 0.05 = 50 \text{ Significant results}$$

«Multiple testing problem»

NHH TECH3

Why not set $\alpha = 0,$ so that $P(Type\ I\ Error) = 0$?

| | $H_0$ is true | $H_0$ is false |
|---|---|---|
| Reject $H_0$ | Type I Error | Correct decision |
| Retain $H_0$ | Correct decision | Type II Error |

NHH
TECH3

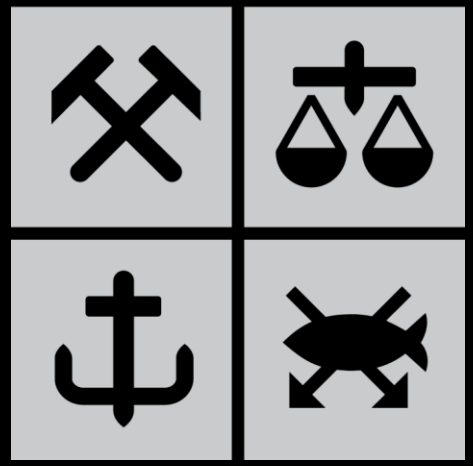
# Neyman–Pearson framework:

- *We cannot avoid mistakes.*
- *We can aim to control how often we make them in the long run.*

NHH
TECH3

NHH

TECH3



Sondre Hølleland
Geir Drage Berentsen