DATA SCIENCE

Statistics for Data Science

Hypothesis Testing



UNC GREENSBORO

WHY TEST HYPOTHESES

We observe a difference. Is it meaningful?

Question: Does a new website design improve user engagement?

DESIGN A	DESIGN B
5.0 min	5.4 min

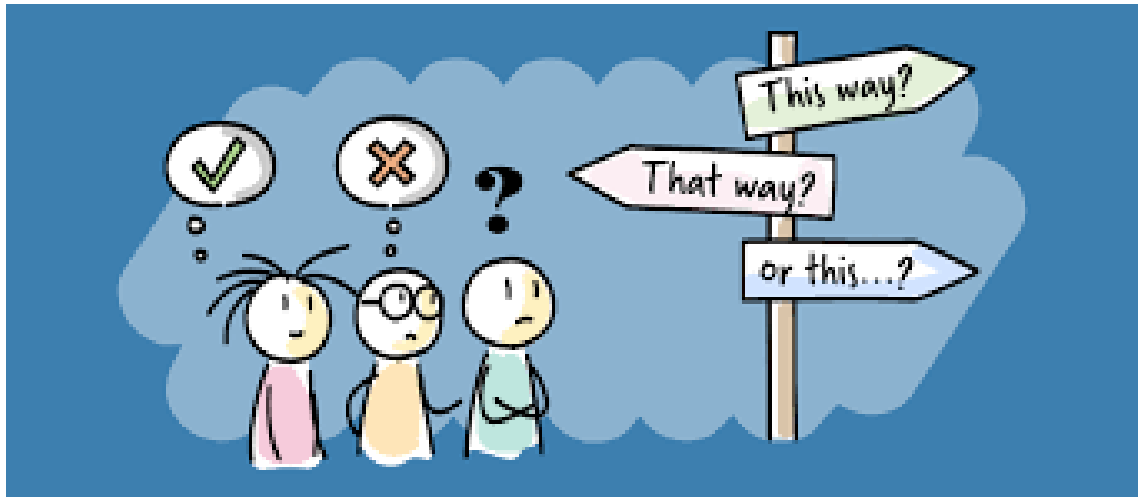
Observed difference: +0.4 min

Design B is better

- Is a 0.4 min difference large compared with the variation in scores?
- Did we collect enough observations?
- Could this difference have appeared by chance?

We don’t just describe data, we make decisions

- Is a new teaching method better?
- Do two groups behave differently?
- Are two variables related?



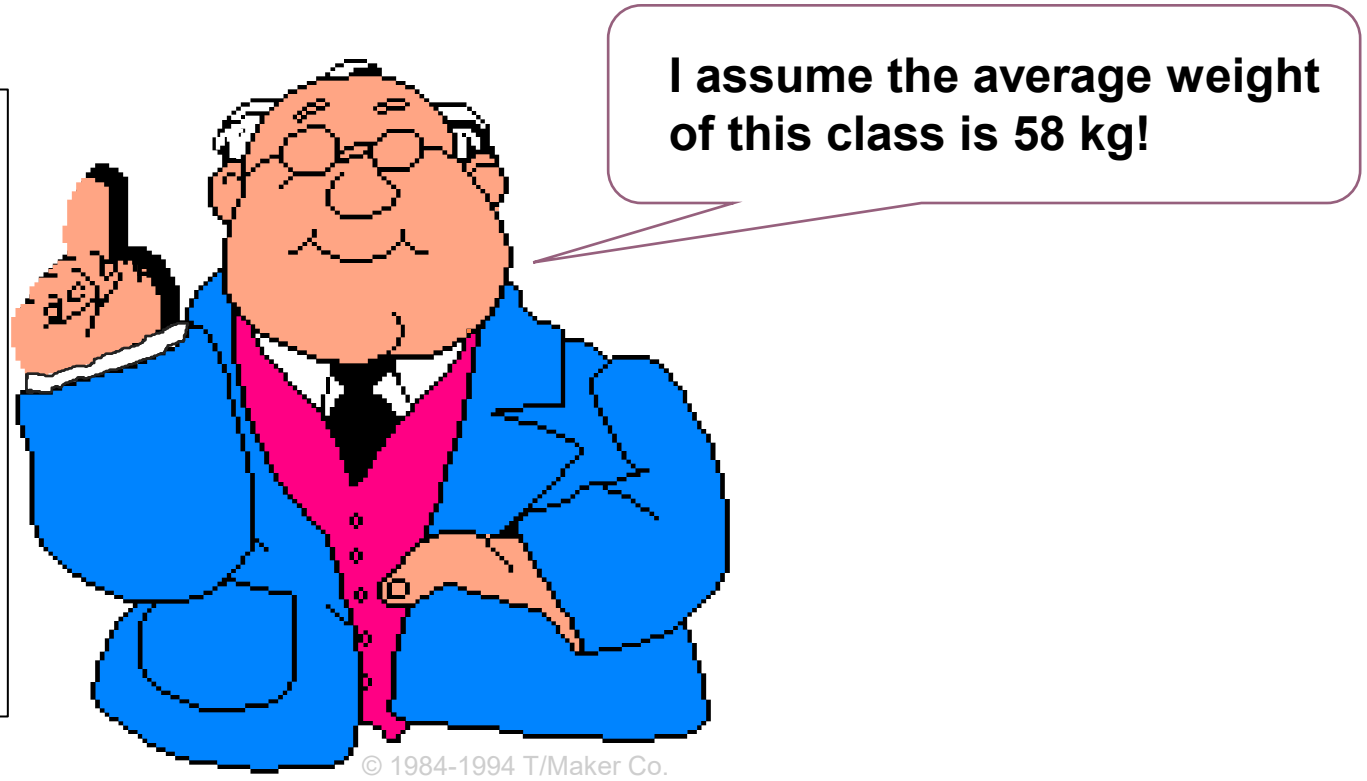
WHY TEST HYPOTHESES

Hypothesis testing evaluates evidence against a claim

A statistical method for asking whether sample data provide sufficient evidence against a claim about a population.

Hypothesis Testing Framework

- Making an assumption, called hypothesis, about a population parameter.
- Collecting sample data.
- Calculating a sample statistic.
- Using the sample statistic to evaluate the hypothesis.



We assess the strength of evidence. We rarely prove a claim with absolute certainty.

Two hypotheses describe competing population claims

NULL HYPOTHESIS H_0

- The assumption we wish/ want to test is called the **null hypothesis**. The symbol for null hypothesis is H_0.
- It usually assumes that there is no difference, effect, or association in the population.

- For example, in a clinical trial of a new drug, the null hypothesis might be that the new drug is no better, on average, than the current drug. We would write
  - H_0 : there is no difference between the two drugs on an average.



Begin with the assumption that the null hypothesis is TRUE.

(Similar to the notion of innocent until proven guilty)

NULL AND ALTERNATIVE

Two hypotheses describe competing population claims

ALTERNATIVE HYPOTHESIS H_1 OR H_A

- The alternative hypothesis is the opposite of the null hypothesis
- It usually assumes that there is a difference, effect, or association exists.

- For example, in a clinical trial of a new drug, the null hypothesis might be that the new drug is no better, on average, than the current drug. We would write
  - H_A: the new drug is better than the current drug, on average.

More examples

- A quality engineer would like to determine whether the production process he is charged of monitoring is still producing products whose mean response value is supposed to be μ_0 (process is in-control), or whether it is producing products whose mean response value is now different from the required value of μ_0 (process is out-of-control).

  - Statement 1 (Null): $\mu = \mu_0$ (process in-control)
  - Statement 2 (Alternative): $\mu \neq \mu_0$ (process out-of-control)

More examples

- An engineer would like to decide which of two computer chip manufacturers (say, Intel and Motorola) is more reliable in producing computer chips. If we denote by p_1 the proportion of defective chips for Intel, and p_2 the proportion of defective chips for Motorola, then the goal is to decide between the following competing statements:

  - Statement 1 (Null): $p_1 \leq p_2$ (Intel is more reliable)
  - Statement 2 (Alternative): $p_1 > p_2$ (Motorola is more reliable)

NULL AND ALTERNATIVE

A directional question gives a one-sided alternative

Research question: Does a new teaching method improve exam scores?

$H_0: \mu_{new} = \mu_{old}$

$H_1: \mu_{new} > \mu_{old}$

Why use “greater than” instead of “not equal to”?

NULL AND ALTERNATIVE

A test has two possible decisions

Reject H_0 in favor of H_A

The sample provides sufficient statistical evidence against H_0.

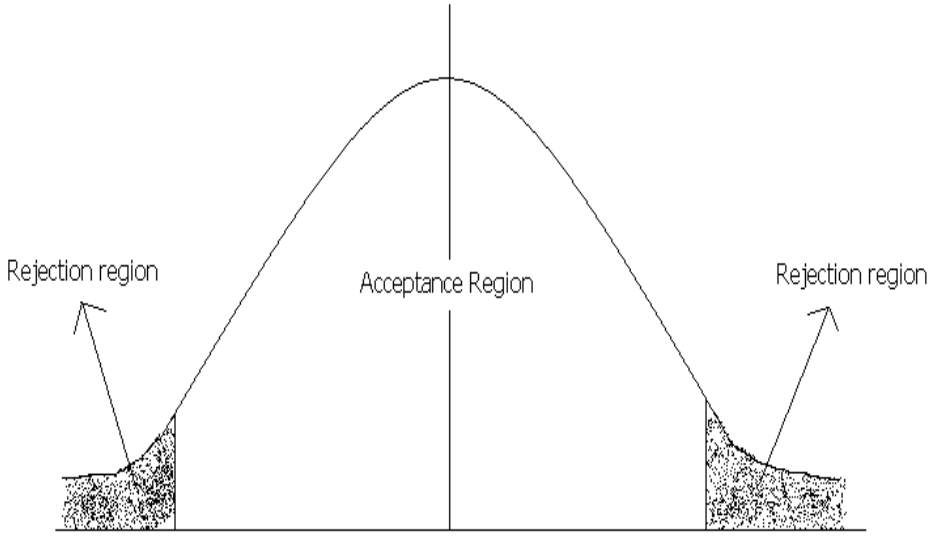
Fail to reject H_0

The sample does not provide sufficient evidence against H_0.

Failing to reject H_0 does not prove H_0 true.

The significance level sets a decision threshold

- Significance level: refers to the percentage of sample means that is outside certain prescribed limits. E.g. testing a hypothesis at 5% level of significance means
  - We reject the null hypothesis if it falls in the two regions of area 0.025.
  - Do not reject the null hypothesis if it falls within the region of area 0.95.



If our sample statistic(calculated value) fall in the non-shaded region (acceptance region), then it simply means that there is no evidence to reject the null hypothesis.

Choose Significance level before examining the results.

Type I error is a false alarm

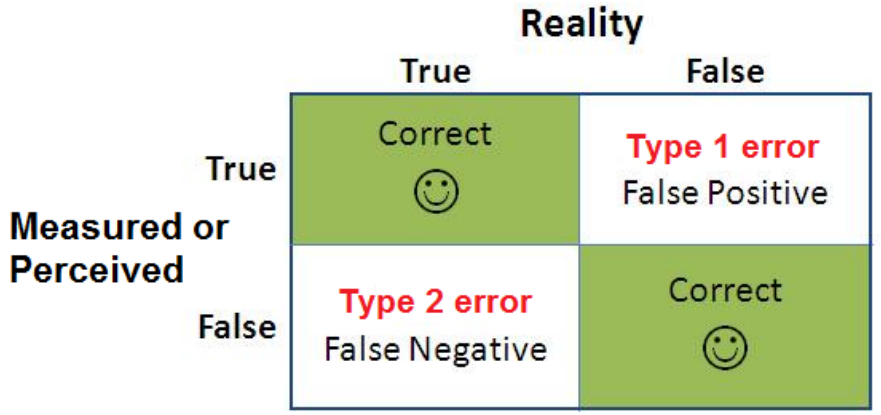
Definition

Type I error refers to the situation when we reject the null hypothesis when it is true (H_0 is wrongly rejected).

- For example:

  H_0: there is no difference between the two drugs on average.

  - Type I error will occur if we conclude that the two drugs produce different effects when actually there isn't a difference.
  - Prob(Type I error) = significance level = α



Type II error is a missed effect

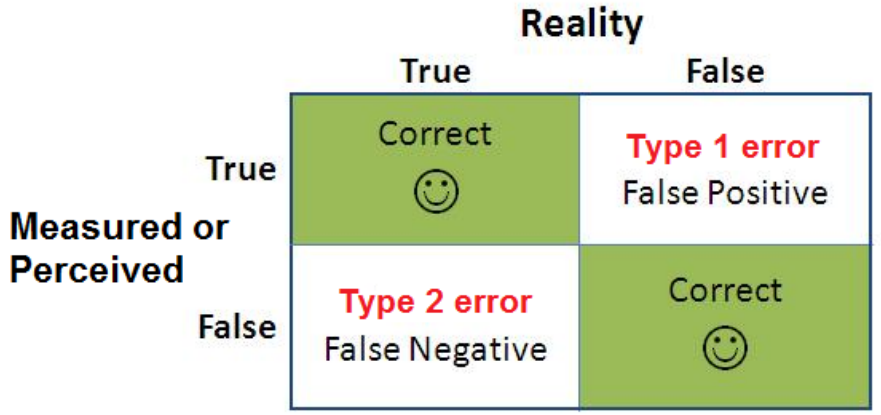
Definition

Type II error refers to the situation when we accept the null hypothesis when it is false.

- For example:

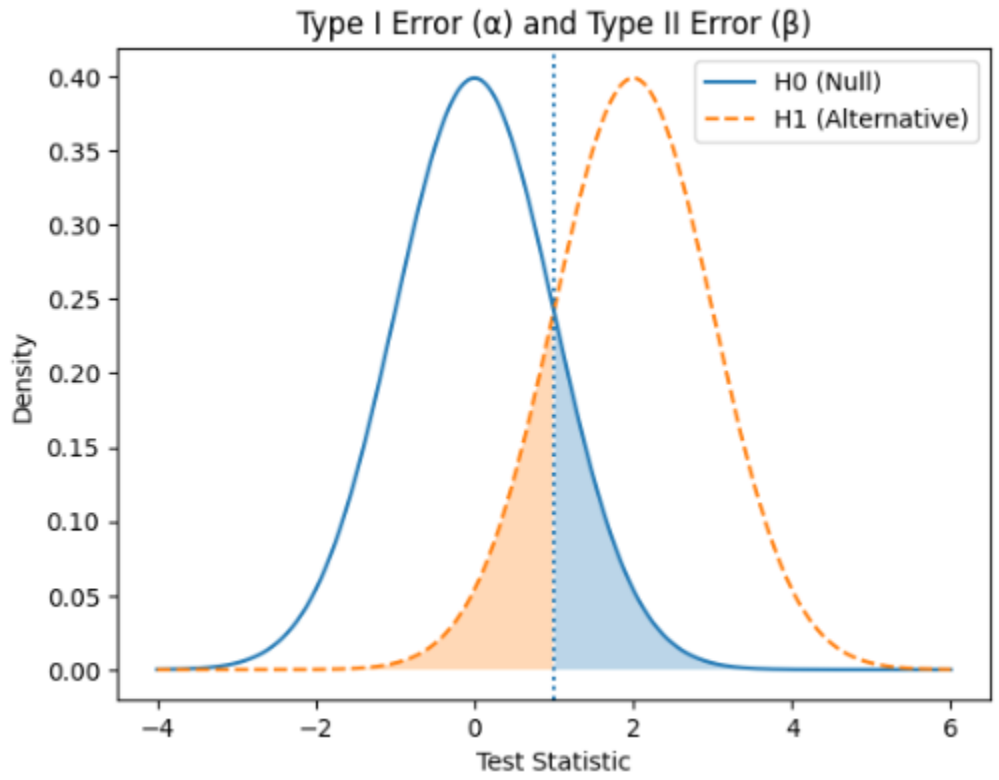
  H_0: there is no difference between the two drugs on average.

  - Type II error will occur if we conclude that the two drugs produce the same effect when actually there is a difference.
  - Prob(Type II error) = β



ERRORS AND SIGNIFICANCE

Type I and Type II Errors



Type I and Type II Errors

- Example

In court room, a defendant is considered not guilty as long as his guilt is not proven. The prosecutor tries to prove the guilt of the defendant. Only when there is enough charging evidence the defendant is condemned. In the start of the procedure, there are two hypotheses H_0: "the defendant is not guilty", and H_A : "the defendant is guilty". The first one is null hypothesis, and the second one is alternative hypothesis.

	Null Hypothesis (H_0) is true He is not guilty	**Alternative Hypothesis (H_1) is true He is guilty**
Accept Null Hypothesis	Right decision	Wrong decision Type II Error
Reject Null Hypothesis	Wrong decision Type I Error	Right decision

ONE-TAILED AND TWO-TAILED TESTS

The research question sets the direction of H_1

A company changes its recommendation algorithm.

Change?

H_1: $\mu_{new} \neq \mu_{old}$

Both directions matter.

Increase?

H_1: $\mu_{new} > \mu_{old}$

Only a higher value supports the claim.

Decrease?

H_1: $\mu_{new} < \mu_{old}$

Only a lower value supports the claim.

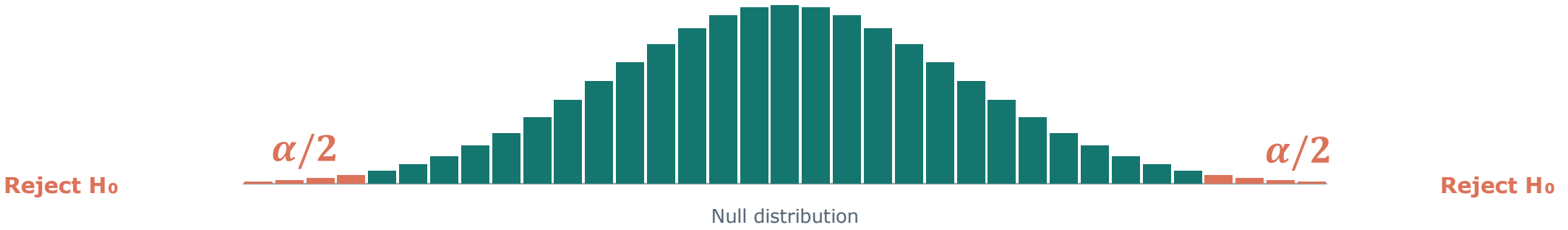
Write the alternative to match the question before analyzing results.

ONE-TAILED AND TWO-TAILED TESTS

A two-tailed test looks for a difference in either direction

H_0: $\mu_{new} = \mu_{old}$

H_1: $\mu_{new} \neq \mu_{old}$



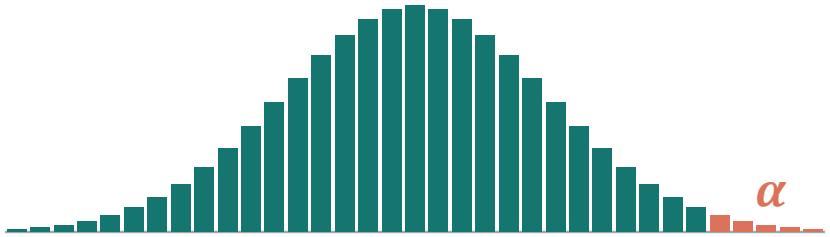
At α = 0.05, a symmetric two-tailed test allocates 0.025 to each tail.

ONE-TAILED AND TWO-TAILED TESTS

A one-tailed test places the rejection region in one tail

Right-tailed

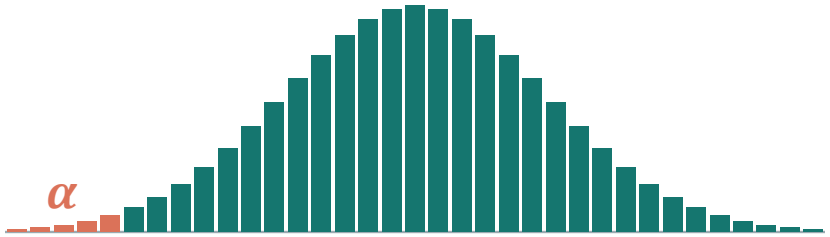
H_1: μnew > μold



Null distribution

Left-tailed

H_1: μnew < μold



Null distribution

At α = 0.05, the full 5% rejection probability is assigned to the specified tail.

Choose direction before looking at the sample result.

Examples

- The manufacturer of light bulbs wants to produce light bulbs with a mean life of 1000 hours. If the lifetime is shorter, he will lose customers to the competition and if it is longer then he will incur a high cost of production. He does not want to deviate significantly from 1000 hours in either direction. Thus, he selects the hypotheses as
    - H_0 : μ = 1000 hours and H_A : μ ≠ 1000 hours, and uses a two-tailed test.
- A wholesaler buys light bulbs from the manufacturer in large lots and decides not to accept a lot unless the mean life is at least 1000 hours.
    - H_0 : μ = 1000 hours and H_A : μ <1000 hours, and uses a lower-tailed test.

He rejects H_0 only if the mean life of sampled bulbs is significantly below 1000 hours. (he accepts H_A and rejects the lot)

TESTING PROCESS

Five steps make a test easier to follow

01 **State H_0 and H_1**

Define the population claims.

02 **Choose α**

Set the threshold in advance.

03 **Select a test**

Use outcome type, design, and assumptions.

04 **Calculate statistic and p**

Measure how incompatible the data are with H_0.

05 **Decide and interpret**

Compare p with α and answer in context.

CHOOSING A TEST

T-Tests

- T-Test: determine if there is a significant difference between the means of two groups and how they are related. There are :
  - One sample T- Test
    - Is the average delivery time different from the company's 30-minute target?
  - Independent T- Test
    - Do students taught with Method A and Method B have different average exam scores?
  - Paired T- Test
    - Does a training program improve students' scores?

One-sample	Independent	Paired
One group vs. a target	Two different groups	Two related measurements

T-TESTS

One-sample t-test

Research question: Does mean delivery time differ from 30 minutes?

Target μ_0	Sample mean $\bar{X}$	Sample SD s	Sample size n
30 min	32 min	5 min	25

H_0: $\mu = 30$

H_1: $\mu \neq 30$

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}} \quad \text{where} \quad s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

where

μ = Proposed constant for the population mean

$\bar{x}$ = Sample mean

n = Sample size (i.e., number of observations)

s = Sample standard deviation

$s_{\bar{x}}$ = Estimated standard error of the mean (s/sqrt(n))

One-sample t-test

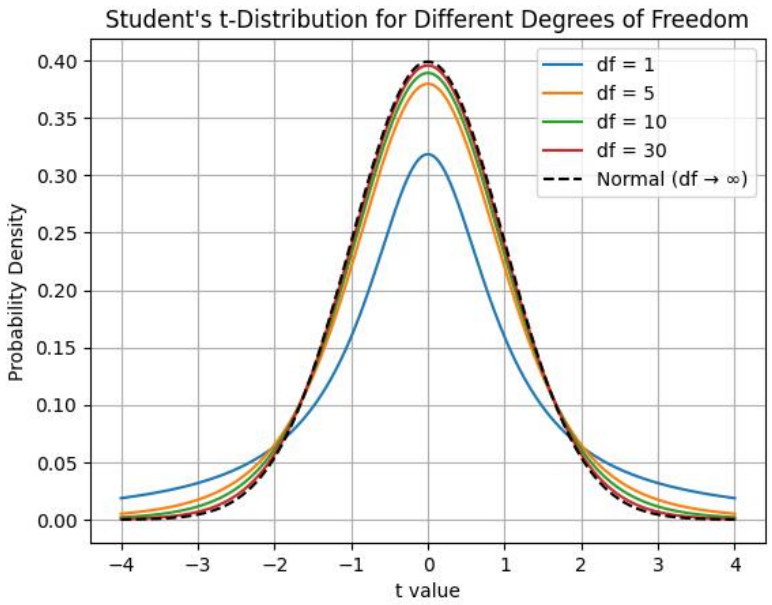
$$t = (32 - 30) / (5 / \sqrt{25}) = 2.0$$

df = n − 1 = 24

Two-tailed p ≈ 0.057

At α = 0.05, p > α. Fail to reject H_0.

The sample does not provide sufficient evidence that the population mean differs from 30 minutes.



T-TESTS

Independent t-test

Research question: Do students taught with Method A and Method B have different average exam scores?

Group	Student 1	Student 2	Student 3	Student 4	Student 5
Method A	78	82	75	85	80
Method B	84	86	79	89	82

Mean A = 80

Mean B = 84

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$s^2 = \frac{\sum_{i=1}^{n_1}(x_i - \bar{x}_1)^2 + \sum_{j=1}^{n_2}(x_j - \bar{x}_2)^2}{n_1 + n_2 - 2}$$

Two sample t-test

H_0: μA = μB

H_1: μA ≠ μB

Use an independent two-sample t-test.

Independent t-test

- T-Test: Independent Two-Sample t-test Example
  - Compare average exam scores of two groups:

    Group A: 78, 82, 85, 90, 88

    Group B: 72, 75, 78, 70, 74

$$H_0 : \mu_A = \mu_B$$
$$H_1 : \mu_A \neq \mu_B$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$s^2 = \frac{\sum_{i=1}^{n_1}(x_i - \bar{x}_1)^2 + \sum_{j=1}^{n_2}(x_j - \bar{x}_2)^2}{n_1 + n_2 - 2}$$

Two sample t-test

Independent t-test

- T-Test: Independent Two-Sample t-test Example
  - Compare average exam scores of two groups:

    Group A: 78, 82, 85, 90, 88

    Group B: 72, 75, 78, 70, 74

$$\bar{x}_A = \frac{78 + 82 + 85 + 90 + 88}{5} = 84.6$$

$$\bar{x}_B = \frac{72 + 75 + 78 + 70 + 74}{5} = 73.8$$

$$s_A^2 = \frac{\sum (x - \bar{x})^2}{n - 1} = 22.3$$

$$s_B^2 = 10.7$$

$$s_p^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2} = \frac{4(22.3) + 4(10.7)}{8} = 16.5$$

$$t = \frac{84.6 - 73.8}{\sqrt{16.5(1/5 + 1/5)}} = \frac{10.8}{\sqrt{6.6}} \approx 4.20$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$s^2 = \frac{\sum_{i=1}^{n_1}(x_i - \bar{x}_1)^2 + \sum_{j=1}^{n_2}(x_j - \bar{x}_2)^2}{n_1 + n_2 - 2}$$

Two sample t-test

T-TESTS

Paired t-test

Research question: Does a training program improve students' scores?

Student	Before	After	Difference
A	70	76	+6
B	80	83	+3
C	65	72	+7
D	75	78	+3
E	85	88	+3

d = After − Before

H_0: $\mu d = 0$

H_1: $\mu d \neq 0$

The same people provide both measurements.

Paired t-test

$$t = \frac{\overline{X}_D - \mu_0}{s_D/\sqrt{n}}.$$

ANOVA

ANOVA compares three or more group means

Suppose a school wants to compare three teaching methods.

Method	Average score
A	75
B	80
C	85

We could perform three pairwise t-tests:

- A vs. B
- A vs. C
- B vs. C

ANOVA tests whether there is evidence that the population means of three or more groups are not all equal.

ANOVA

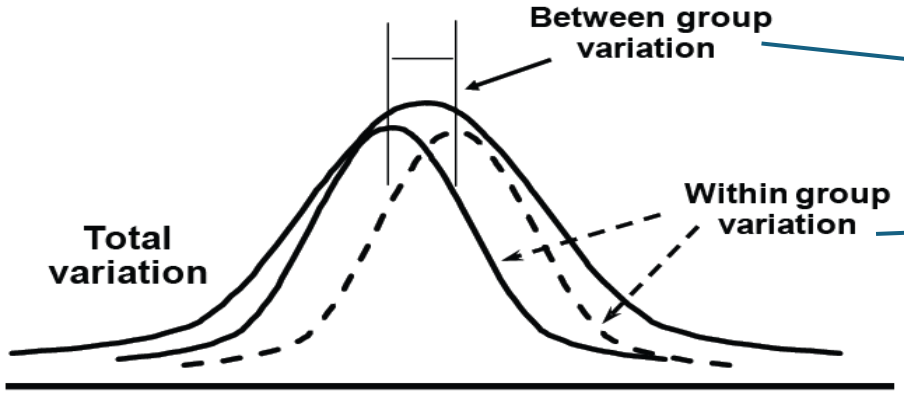
ANOVA compares between-group and within-group variation

Between-group variation

How far group means sit from the overall mean

Within-group variation

How much observations vary around their group mean



$$F = \frac{MSB}{MSW} = \frac{SSB/(k-1)}{SSW/(N-k)}$$

A relatively large F can indicate that group means differ more than expected from within-group variability.

The F-statistic is a ratio of two sources of variation.

ANOVA test: Example

Compare 3 teaching methods:

Group 1: 85, 88, 90

Group 2: 78, 82, 80

Group 3: 92, 94, 96

H_0: all means equal

H_1: at least one differs

$$\bar{x}_1 = 87.7$$
$$\bar{x}_2 = 80.0$$
$$\bar{x}_3 = 94.0$$

$$\bar{x} = 87.2$$

$$SSB = \sum n_i(\bar{x}_i - \bar{x})^2 = 294.6$$

$$SSW = SSW_1 + SSW_2 + SSW_3$$

$$= 12.67 + 8 + 8 = 28.67$$

$$F = \frac{\frac{294.6}{2}}{\frac{28.67}{6}} = 30.82$$

ANOVA

A significant ANOVA says at least one mean differs

H_0: $\mu_A = \mu_B = \mu_C$

H_1: Not all population means are equal

Example result

$p = 0.01$
$\alpha = 0.05$

$p < \alpha$, so reject H_0.

There is evidence that at least one teaching-method population mean differs.

ANOVA does not identify which pairs differ. Use a suitable follow-up comparison.

CHI-SQUARE TESTS

A chi-square test analyzes categorical counts

Device type and purchase outcome

Device	Purchased	Did not purchase	Total
Mobile	60	40	100
Desktop	40	60	100
Total	100	100	200

- Is purchase behavior associated with device type?
- Or could the observed difference in purchase rates be explained by random variation?

A chi-square test can be used to investigate this question using categorical count data.

chi-square test

H_0: Device type and purchase outcome are independent.

H_1: Device type and purchase outcome are associated.

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Where:

χ^2 = Chi Square obtained
$\sum$ = the sum of
O = observed score
E = expected score

Expected count

E = (row total × column total) / grand total

Mobile & purchased: E = 100 × 100 / 200 = 50

chi-square test

- Chi-Square Test: Example $\chi^2 = \sum \frac{(O-E)^2}{E}$ $E = \frac{(\text{row total})(\text{column total})}{N}$
- Test if gender and preference are independent.

	Like	Dislike	Total
Male	30	20	50
Female	20	30	50
Total	50	50	100

$$\chi^2 = \frac{(30-25)^2}{25} + \frac{(20-25)^2}{25} + \frac{(20-25)^2}{25} + \frac{(30-25)^2}{25}$$

$$= 1 + 1 + 1 + 1 = 4$$

CHOOSING A TEST

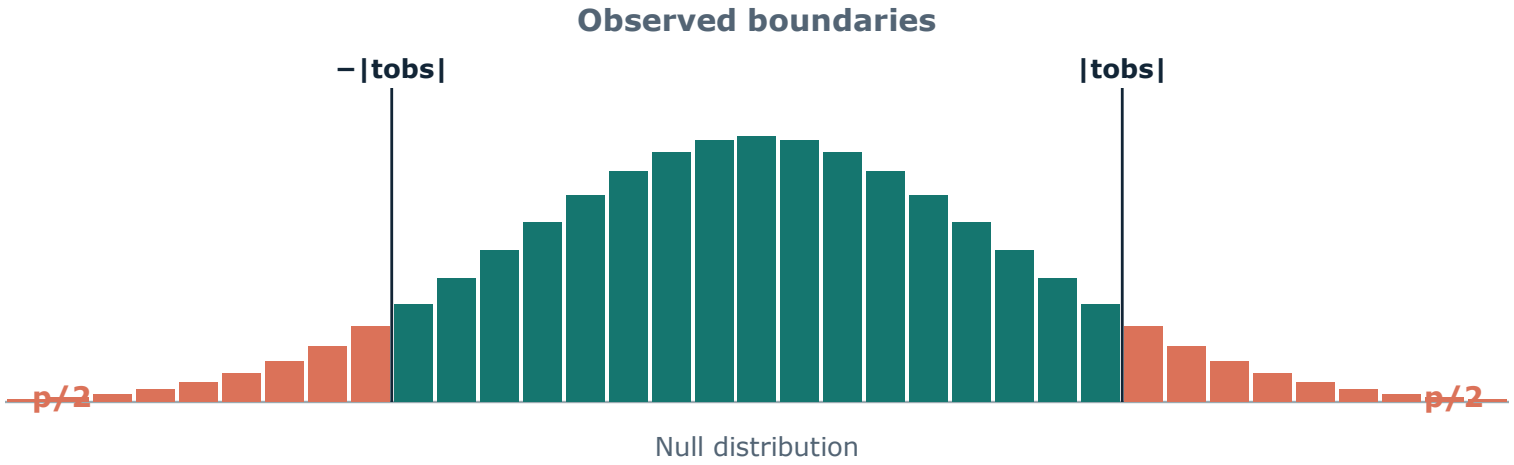
Different research questions call for different tests

Research question	Example	Common test
One sample mean vs. target	Delivery time vs. 30 min	One-sample t-test
Two independent means	Scores from two classes	Independent t-test
Two related measurements	Scores before and after	Paired t-test
Three or more means	Three teaching methods	ANOVA
Categorical association	Device type and purchase	Chi-square test

UNDERSTANDING THE P-VALUE

A p-value measures how extreme the data are under H_0

Assume H_0 and the test model are correct.



p-value: probability of a test statistic at least as extreme as observed under H_0.

Small p: relatively unusual under H_0. Large p: not especially unusual under H_0.

The p-value and α determine the test decision

- p-value: the probability of observing data at least as extreme as what we observed, assuming the null hypothesis is true.
  - If the p-value exceed 0.10, then H_0 is not rejected and we say that the result is not significant.
  - If the p-value is between 0.10 and 0.05, we usually say that the result is almost significant or tending towards significance.
  - If the p-value is between 0.05 and 0.01, we reject H_0 and conclude that the result is significant.
  - If the p-value is less than 0.01 then H_0 is rejected and conclude that the result is highly significant.
  - Or, we may compare the p-value with the level of significance: if it is smaller, reject H_0.

Example:

Suppose you test whether a new drug provides more pain relief than a placebo.

If the drug has no real effect, your test statistic will be close to what's expected under the null hypothesis, and the p-value will be high (close to 1).

If the drug truly works, your test statistic will differ more from the null expectation, and the p-value will drop.

The smaller the p-value, the less likely the results occurred by random chance, and the stronger the evidence that you should reject the null hypothesis.