Introduction

· The goal is to compare if two groups are significantly different.

· There are two options regarding "Two Sample Hypothesis Testing":

+ Matched pairs: the two samples are 'dependent',

+ Independent samples: → equal variances → unequal variances

Matched Pairs

· Data in the two samples is dependent. Some examples: → Same number of observations in both samples

+ Clinical trials: same participants, before / after. ⓐ

+ Measurements using various instruments (for example, car speed measured using radar and speed-meter). ⓑ

+ Test (blind) of various products by some customers. ⓒ

ⓐ

	Blood pressure	
	Before	After
Juan	18.4	12.2
Ramon	16.5	11.5

ⓑ

	Instrument	
	Radar	Speed-meter
Car A	120,5	119,4
Car B	115.2	112.7

ⓒ

	Coca Cola	
	Zero	Classic
Tester 1	7	9
Tester 2	4	5

· Is equivalent to make a 1 sample hypothesis test:

+ Two-sided 'equality' test.

+ Testing against constant mean 0, taking difference of samples:

$$H_0 : \mu_{diff} = 0 \; ; \; H_1 : \mu_{diff} \neq 0$$

Independent Samples

· Data in the two samples is independent. Some examples: → No need to have same number of elements.

+ Clinical trials with two different medicine and with different people. ⓐ

+ Measurement of speed of different cars in different roads. ⓑ

+ A/B testing for websites, users are shown a version of the website and click / no-click behavior is measured. ⓒ

ⓐ Drug A: [11.1, 12.2, 13.3, 18.7, etc.]

Drug B: [12.2, 13.5, 19.4, etc.]

Blood pressure

ⓑ Road A: [100.5, 90.2, 85.7, ..., etc.]

Road B: [70.4, 110.5, 95.7, ..., etc.]

km/h

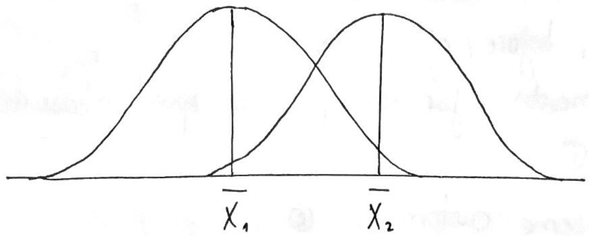
ⓒ Version A: [1, 0, 0, 1, 0, 0, 1, 0, 0, 1]

Version B: [0, 1, 0, 0, 1, 0, 0, 1, 0] Suscribed

- In this case: $H_0: \mu_1 = \mu_2$; $H_1: \mu_1 \neq \mu_2$
- There are two variants of the test:
  a) Assuming 2 samples have equal variances, s can be computed pooling all samples:

$$S_{pooled} = \sqrt{\frac{(n_1 - 1) \cdot S_1^2 + (n_2 - 1) \cdot S_2^2}{n_1 + n_2 - 2}} \qquad t = \frac{\bar{X}_1 - \bar{X}_2}{S_{pooled} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

  b) Assuming 2 samples have unequal variances:
  This version is called Welch's test: $t = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$
- Equivalent to something similar to:



Chi - squared contingency test

- Test if 2 categorical variables are independent (or not).
- Operates over a contingency matrix.
- Widely used as feature selection algorithm (ML algorithms). "H_0: independent / H_1: related"

$X_i^2 \uparrow\uparrow$, $p \downarrow\downarrow$, reject H_0, variables are related

Full example (Pokemon), finish in lab, this is just an introduction:

①

Legendary \ Water Type	True	False	
True	4	61	65
False	108	627	735
	112	688	800

First step, compute contingency matrix, observed values

$P_{water} \cap P_{legend} = P_{water} \cdot P_{legend}$

② Second step, compute 'expected values':

$P_{water} \cap P_{legent} \cdot Total$ $\frac{65 \cdot 112}{800} = 9,1$	$\frac{65 \cdot 688}{800} = 55,9$
$\frac{112 \cdot 735}{800} = 102,9$	$\frac{688 \cdot 735}{800} = 632,1$

③ Compute Observed - Expected

$-5.1 = 4 - 9.1$	$5.1 = 61 - 55.9$
$5.1 = 108 - 102.9$	$-5.1 = 627 - 632.1$

④ Compute $(\text{Observed} - \text{Expected})^2$

$(-5.1)^2 = 26.01$	$5.1^2 = 26.01$
$5.1^2 = 26.01$	$(-5.1)^2 = 26.01$

⑤ $\frac{(O-E)^2}{E}$

$\frac{26.01}{9.1} = 2.858$	$\frac{26.01}{55.9} = 0.465$
$\frac{26.01}{102.9} = 0.252$	$\frac{26.01}{632.1} = 0.041$

⑥ Compute $X_i^2 = \sum_{\text{cells}} \frac{(O-E)^2}{E} = 2.858 + 0.465 + 0.252 + 0.041 = 3.61$

this is the test statistic

degrees of freedom = $(\text{rows} - 1) \cdot (\text{colums} - 1) = 1$

$p > 0.05$ → Can not reject null hypothesis. ≡

Look distribution tables...