Two-sample tests

Comparison of two populations

- **Paired**

  We can create pairs from the elements of the two populations.

- **Independent populations**

  We take samples from the two populations independently. No possibility to make pairs!

Independent samples

Two independent samples

two samples are taken independently, the number of observations may be different

- **Populations:**

  Y population
  X population
  } same variable, same unit of measure

- **samples:**

$$y_1, y_2, \ldots, y_{n_Y}$$
$$x_1, x_2, \ldots, x_{n_X}$$

- **Characteristics:**

Population	Population			Sample			
	Expected value	Prop	Variance	Size	Mean	Prop	Variance
Y	μ_Y	P_Y	σ_Y^2	n_Y	$\bar{y}$	p_Y	s_Y^2
X	μ_X	P_X	σ_X^2	n_X	$\bar{x}$	p_X	s_X^2

Comparison of expected values

- **Aim:** examination of the difference of two expected values

- **Hypotheses:**

$$H_0^T : \mu_Y - \mu_X = \delta_0$$

$$H_1 : \mu_Y - \mu_X \neq \delta_0$$
$$H_1 : \mu_Y - \mu_X > \delta_0$$
$$H_1 : \mu_Y - \mu_X < \delta_0$$

Comparison of expected values

1. Two-sample Z-test

Conditions:

- Normally distributed populations
- We know the sd of populations

Test:

$$z = \frac{(\bar{y} - \bar{x}) - \delta_0}{\sqrt{\frac{\sigma_Y^2}{n_Y} + \frac{\sigma_X^2}{n_x}}}$$

It follows standard normal distribution

Comparison of expected values

2. Two-sample t-test

Conditions:

- Normally distributed populations
- We estimate the sd from samples
- We assume that the standard deviations are the same for the 2 samples

Test:

$$t = \frac{(\bar{y} - \bar{x}) - \delta_0}{s_c \sqrt{\frac{1}{n_Y} + \frac{1}{n_X}}}$$

It follows Student-t distribution
Degress of freedom:

$$\nu = n_Y + n_X - 2$$

weight

$$s_c = \sqrt{\frac{(n_Y - 1)s_Y^2 + (n_X - 1)s_X^2}{n_Y + n_X - 2}} = \sqrt{\frac{\sum d_Y^2 + \sum d_X^2}{n_Y + n_X - 2}}$$

Comparison of expected values

3. Asymptotic Z-test

Conditions:

- **Samples are large enough**

Test:

$$z = \frac{(\bar{y} - \bar{x}) - \delta_0}{\sqrt{\frac{s_Y^2}{n_Y} + \frac{s_X^2}{n_X}}}$$

It follows standard normal distribution

BÉRTARIFA survey 2019

Populations: employees with an economics degree

Field	Sample		
	Size	Mean	Sd
Finance	203	906275	576393
HR	119	657765	466438

Asympltotic Z-test

In field of finance, the wages are higher on average compared to HR

alternative

$$H_0^T : \mu_{F} = \mu_{HR} \qquad H_0 : \mu_{F} \leq \mu_{HR} \qquad H_1^j : \mu_{F} > \mu_{HR}$$

$$c_a = z_{0,95} = 1,645$$

→ sample size > 100
asymtotic case

$$z = \frac{\bar{y}_{F} - \bar{y}_{HR}}{\sqrt{\frac{s^2_{F}}{n_{F}} + \frac{s^2_{HR}}{n_{HR}}}} = \frac{906275 - 657765}{\sqrt{\frac{576393^2}{203} + \frac{466438^2}{119}}} = \frac{248510}{58863} = 3,82$$

Decision: **we reject H0 and H0T, so the wages are higher in the finance sector**

Example

A team of health researchers wants to compare two different weight loss methods. The study involved 37 overweight people who were randomly assigned to two groups. Each procedure was used for one month. (The procedures are assumed to have a normal distribution of weight loss.)Can it be stated at the 5% significance level that, on average, more weight can be lost in one month with Procedure I than with Procedure II?

+- test

Proc	Sample		
	Size	Mean	Sd
I.	21	11,86	2,128
II.	16	9,88	1,928

Example

- **Normality is TRUE**
- **Pop. Standard deviations are unknown**

For small samples, this can only be handled if we know (or are willing to believe) that the unknown variances are equal.
Since in the present example we have no information whether the unknown variances are equal, we have to perform a variance equality test before the two-sample t-test!!!

Solution

Proc	Sample		
	Size	Mean	Sd
I.	21	11,86	2,128
II.	16	9,88	1,928

Hypotheses: $H_0^T: \mu_I = \mu_{II} \quad H_0: \mu_I \leq \mu_{II} \quad H_1: \mu_I > \mu_{II}$

Test:

$$s_c = \sqrt{\frac{20 \cdot 2{,}128^2 + 15 \cdot 1{,}928^2}{21+16-2}} = \sqrt{4{,}181} = 2{,}045$$

$$t = \frac{(\bar{y}-\bar{x})-\delta_0}{s_c\sqrt{\frac{1}{n_Y}+\frac{1}{n_X}}} = \frac{\bar{y}_I - \bar{y}_{II}}{s_c\sqrt{\frac{1}{n_I}+\frac{1}{n_{II}}}} = \frac{11{,}86-9{,}88}{2{,}045\sqrt{\frac{1}{21}+\frac{1}{16}}} = \frac{1{,}98}{0{,}679} = 2{,}92$$

Critical values $c_f = t_{0{,}95}(35) \approx t_{0{,}95}(40) = 1{,}68$

Decision: $c_f = 1{,}68 < 2{,}92$

We reject the null hypotheses, the first method is more effective!

Comparison of variances

Condition: normally distributed populations

Hypotheses: $H_0 : \sigma_Y = \sigma_X$ **vagy:** $H_0 : \sigma_Y^2 = \sigma_X^2$

$$H_1 : \sigma_Y \neq \sigma_X$$

$$H_1 : \sigma_Y > \sigma_X$$

$$H_1 : \sigma_Y < \sigma_X$$

Comparison of variances

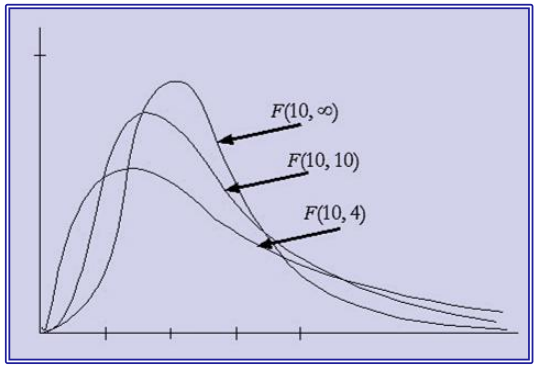
- **Test:** $$F = \frac{s_Y^2}{s_X^2}$$
- **If H_0 is true, it follows F distribution with df:**

  $\nu_1 = n_Y - 1$ és $\nu_2 = n_X - 1$
- **We decide which group will be in the denominator or in the numerator**

ν_1: numerator

ν_2: denominator

F-distribution



- Ratio of two independent χ^2-distributed random variables.
- The density function is not symmetrical.
- Connection between degrees of freedom and critical values

$$F_{1-p}(\nu_1, \nu_2) = \frac{1}{F_p(\nu_2, \nu_1)}$$

Comparison of proportions

$$H_0^T : P_Y - P_X = \varepsilon_0$$

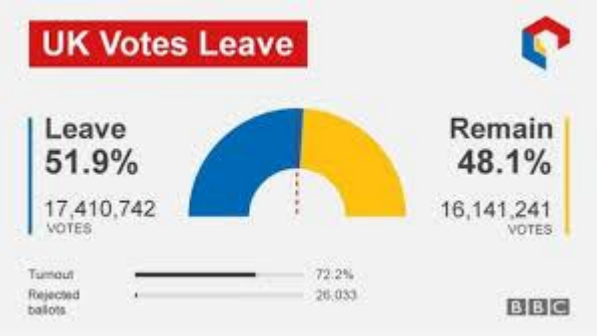
Condition:

If the sample is large enough, test function follows standard normal distribution

Test:

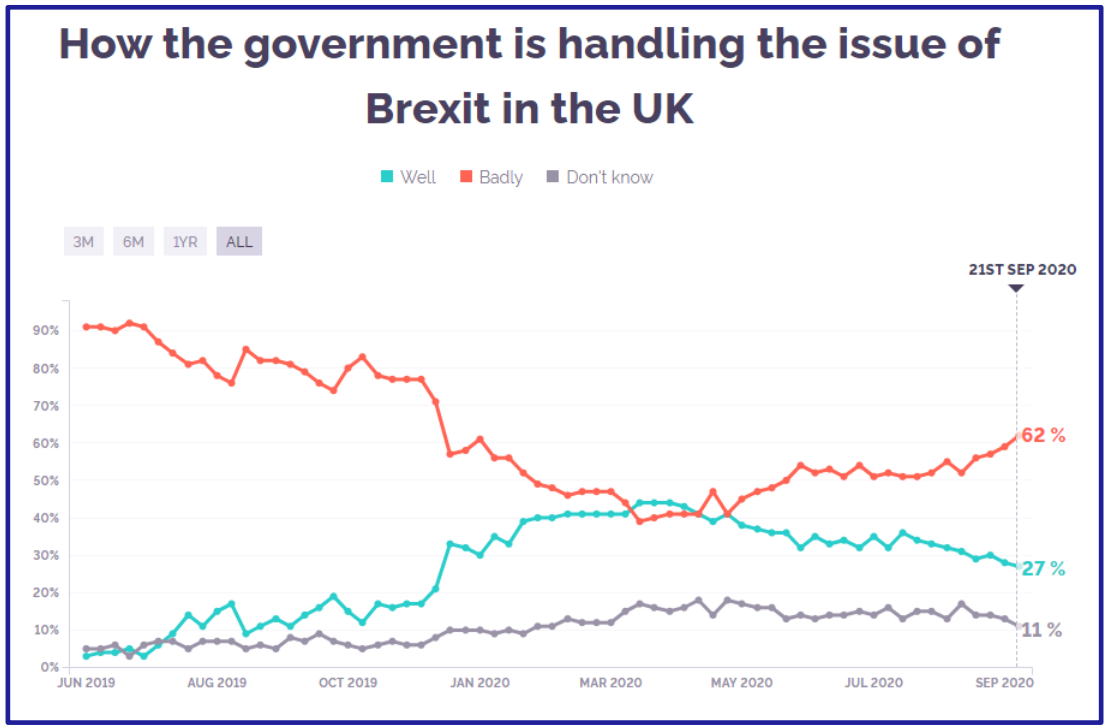
$$z_{\varepsilon_0} = \frac{(p_Y - p_X) - \varepsilon_0}{\sqrt{\frac{p_Y q_Y}{n_Y} + \frac{p_X q_X}{n_X}}}$$

Where: $q_Y = 1 - p_Y$ *és* $q_X = 1 - p_X$



2016. June 23.

BREXIT



Can it be stated at the 5% significance level that the proportion of people who think the government's handling of the UK's exit from the EU is rather bad increased by more than 20 percentage points between 23 March 2020 (1) and 21 September 2020 (2)?

$n_1 = 1633$ $p_1 = 0,39$ $n_2 = 1635$ $p_2 = 0,62$

Solution

Hypotheses:

$$H_0^T: \; P_2 - P_1 = 0,2$$

$$H_0: \; P_2 - P_1 \leq 0,2 \qquad H_1: \; P_2 - P_1 > 0,2$$

Test:

$$z_{\varepsilon_0} = \frac{(p_2 - p_1) - \varepsilon_0}{\sqrt{\frac{p_2 q_2}{n_2} + \frac{p_1 q_1}{n_1}}} = \frac{0,62 - 0,39 - 0,2}{\sqrt{\frac{0,62 \cdot 0,38}{1635} + \frac{0,39 \cdot 0,61}{1633}}} = \frac{0,03}{0,017} = 1,76$$

Critical value

$$c_f = z_{0,95} = 1,645$$

Decision:

$$1,76 > c_f = 1,645$$

***We reject the null hypotheses*. The proportion of people who think the government's handling of the UK's exit from the EU is rather bad increased by more than 20 percentage points between 23 March 2020 and 21 September 2020.**

Comparison of two proportions

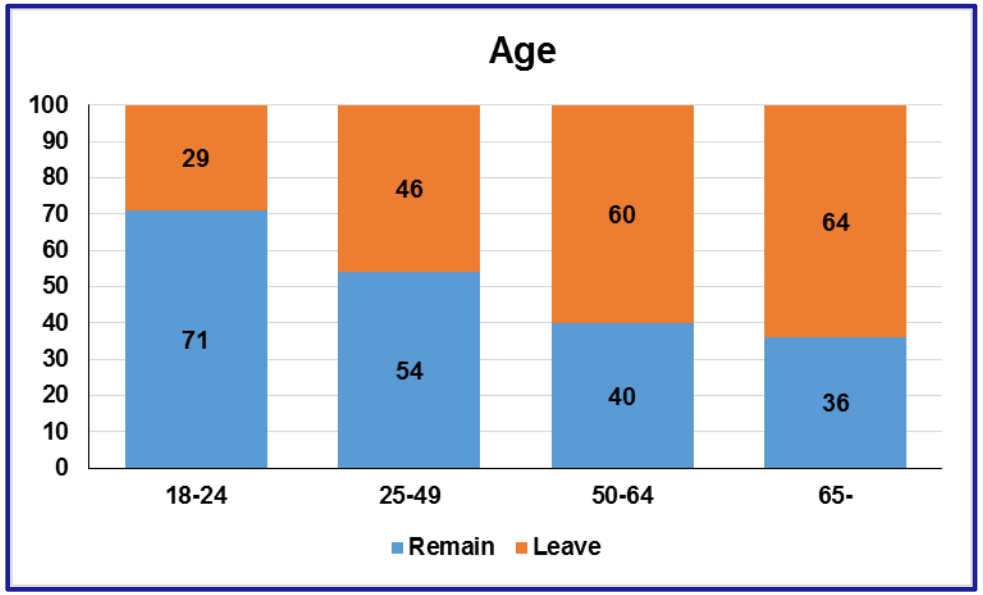
$$H_0^T : P_Y - P_X = \varepsilon_0 \qquad \varepsilon_0 = 0$$

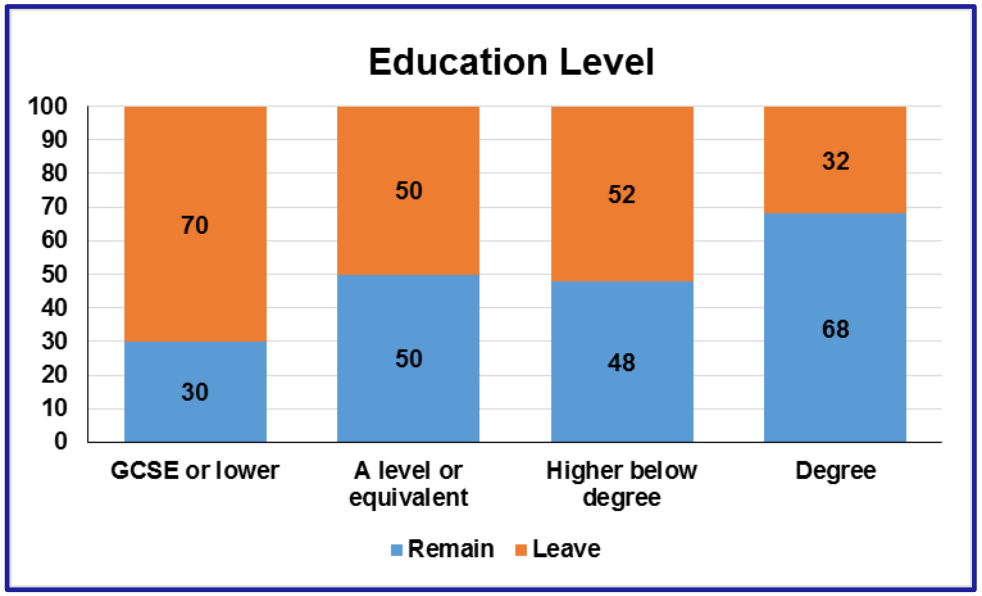
Test:

$$z_0 = \frac{p_Y - p_X}{\sqrt{\bar{p} \cdot \bar{q}\left(\frac{1}{n_Y} + \frac{1}{n_X}\right)}}$$

Where: $\bar{p} = \frac{k_Y + k_X}{n_Y + n_X} = \frac{n_Y p_Y + n_X p_X}{n_Y + n_X}$ $\qquad \bar{q} = 1 - \bar{p}$

EU referendum Vote





YouGov Survey Results Sample Size: 5122 UK Adults
Fieldwork: 23rd - 24th June 2016

EU referendum Vote

$n_M = 2477$ $p_M = 0,53$ $n_W = 2645$ $p_W = 0,51$

Can it be stated at the 5% significance level that men voted to leave in higher proportions than women?

YouGov Survey Results Sample Size: 5122 UK Adults
Fieldwork: 23rd - 24th June 2016

Solution

$$\bar{p} = \frac{\overbrace{2477 \cdot 0,53}^{1313} + \overbrace{2645 \cdot 0,51}^{1349}}{2477 + 2645} = \frac{2662}{5112} = 0,5197$$

Hypotheses:

$$H_0^T: \; P_M = P_W \qquad H_0: \; P_M \leq P_W \qquad H_1: \; P_M > P_W$$

Test

$$Z = \frac{p_M - p_W}{\sqrt{\bar{p} \cdot \bar{q}\left(\frac{1}{n_M} + \frac{1}{n_W}\right)}} = \frac{0,53 - 0,51}{\sqrt{0,5197 \cdot 0,4803\left(\frac{1}{2477} + \frac{1}{2645}\right)}} = \frac{0,02}{0,01397} = 1,43$$

Ciritcal values:

$$c_f = z_{0,95} = 1,645$$

Decision

$$1,43 < c_f = 1,645$$

We fail to reject the technical H0, there is no difference between the proportion of men and women