cont... Inferences: One and two means

Comparing two means from two independent populations

We wish to compare the means μ_1 and μ_2, from two separate populations. The characteristics of each population are as follows:

Population	Mean	Standard Deviation
1	μ_1	σ_1
2	μ_2	σ_2

In order to do this, we take an SRS from each population and measure the same response variable for each sample. The characteristics of each sample are as follows:

Population	Sample Size	Sample Mean	Sample Standard Deviation
1	n_1	$\bar{y}_1$	s_1
2	n_2	$\bar{y}_2$	s_2

To compare the two means, we wish to analyse the difference between μ_1 and μ_2, i.e. $\mu_1 - \mu_2$ .To do this, we use the difference between the corresponding sample means, $\bar{y}_1 - \bar{y}_2$, as an estimate.

1. Because $\mu_{\bar{y}_1} = \mu_1$ and $\mu_{\bar{y}_2} = \mu_2$, then the mean of all the differences of $(\bar{y}_1 - \bar{y}_2)$ is $(\mu_1 - \mu_2)$.
2. Every time we select a sample from each population we won't always get the same difference in sample means, the standard deviation of the difference in sample means is: $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.
3. The distribution of $(\bar{y}_1 - \bar{y}_2)$ is normal if σ_1 and σ_2 are known. (*z* procedures can then be used)

If σ_1 and σ_2 are unknown (usually the case), then we estimate these parameters by s_1 and s_2 and use t-distributions with

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{SE(\bar{y}_1 - \bar{y}_2)} \text{ where } SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

Two Sample Hypothesis Test for means (σ_1 and σ_2 unknown)

Follow the usual procedure:

Formulate the ***null hypothesis H_o*** and the ***alternative hypothesis H_a.*** Usually, they are of the form:

Null Hypothesis $\quad H_o: \mu_1 - \mu_2 = 0$

versus

Alternative Hypothesis $\quad H_a: \mu_1 - \mu_2 > 0$

or $\quad \mu_1 - \mu_2 < 0$

or $\quad \mu_1 - \mu_2 \neq 0$

Calculate the value of the **test statistic** with sample data using:

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Find the ***P*-value** based on the test statistic. Use degrees of freedom: $\boldsymbol{df = n_1 + n_2 - 2}$.
(Note that this technically isn't the correct rule for the *df*. See the footnote on page 480 for the correct one. SPSS uses the correct formula. When solving these problems by hand, just use $\boldsymbol{df = n_1 + n_2 - 2}$.)

Compare the P-value with the **significance level** α, make a decision and write a conclusion.

Assumptions and Conditions

For using *t* procedures (confidence intervals and hypothesis tests) with two independent means, the **assumptions** for each sample are that the **data values are independent** from each other, and that the **sample** has been taken from a population that is **normally distributed**.

The main **conditions** to check are: the data are from a random sample or randomised experiment, each sample is smaller than 10% of the population and that graphical displays of each data set appear normal or at least not strongly skewed with outliers, or each sample size is 'large'.

In addition, the two populations/groups must be independent of each other.

Example 1

Two banks wish to compare the amount that their credit card customers charge to their cards each year. A sample is taken of customers from each bank with the following results:

Bank	Sample Size, n	Sample Mean, $\bar{y}$	Sample Standard Deviation, s
1	15	\$1987	\$392
2	15	\$2056	\$413

Do the data show a significant difference between the mean amounts charged by customers from each bank?

Let μ_1 = average credit charge/customer from Bank A.
Let μ_2 = average credit charge/customer from Bank B.

$$H_o\colon \mu_1 = \mu_2 \qquad \textit{or} \qquad \mu_1 - \mu_2 = 0$$

$$H_a\colon \mu_1 \neq \mu_2 \qquad\qquad \mu_1 - \mu_2 \neq 0$$

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$= \frac{(1987 - 2056) - 0}{\sqrt{392^2/15 + 413^2/15}}$$

$$= \frac{-69}{147.02}$$

$$t = -0.47$$

Degrees of freedom = $n_1 + n_2 - 2 = 28$.

Using Table T, we find:

P-value > 0.20

Fail to reject H_0. We have no evidence to conclude that the mean amounts charged by customers at each bank are significantly different.

Two Sample Confidence Intervals(σ_1 and σ_2 unknown)

To estimate the difference in parameters $\mu_1 - \mu_2$ for a confidence level C, we calculate:-

$$(\bar{y}_1 - \bar{y}_2) \pm t^*_{df} \times SE(\bar{y}_1 - \bar{y}_2) = (\bar{y}_1 - \bar{y}_2) \pm t^*_{df} \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

We use Table T to find t^*_{df} .

Example 2

A university is investigating the level of income that students earn over summer vacation to fund their education. A sample study produced the following:

Find a 90% confidence interval for $\mu_m - \mu_f$

Sex	n	$\bar{y}$	s
Male	675	\$1884.52	\$1368.37
Female	621	\$1360.39	\$1037.46

μ_{m} = average summer earnings for male uni students

μ_f = average summer earnings for female uni students

90% Confidence Interval for $\mu_m - \mu_f$:

$$df = n_m + n_f - 2 = 1294 \quad t^*_{df} = 1.645$$

$$(\bar{y}_m - \bar{y}_f) \pm t^*_{df} \times \sqrt{\frac{s_m^2}{n_m} + \frac{s_f^2}{n_f}}$$

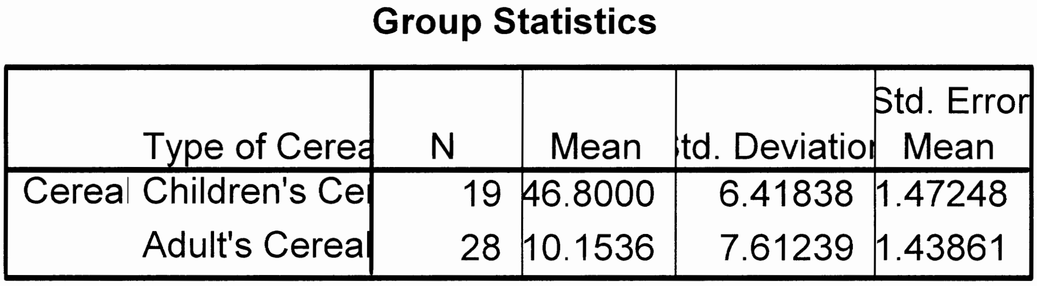
$$= (1884.52 - 1360.39) \pm 1.645\sqrt{\frac{1368.37^2}{675} + \frac{1037.46^2}{621}}$$

$$= (414, 635)$$

We are 90% confident that the difference between average summer earnings of male and female students is between \$414 and \$635. (Males earn between \$414 and \$635 more than females.)

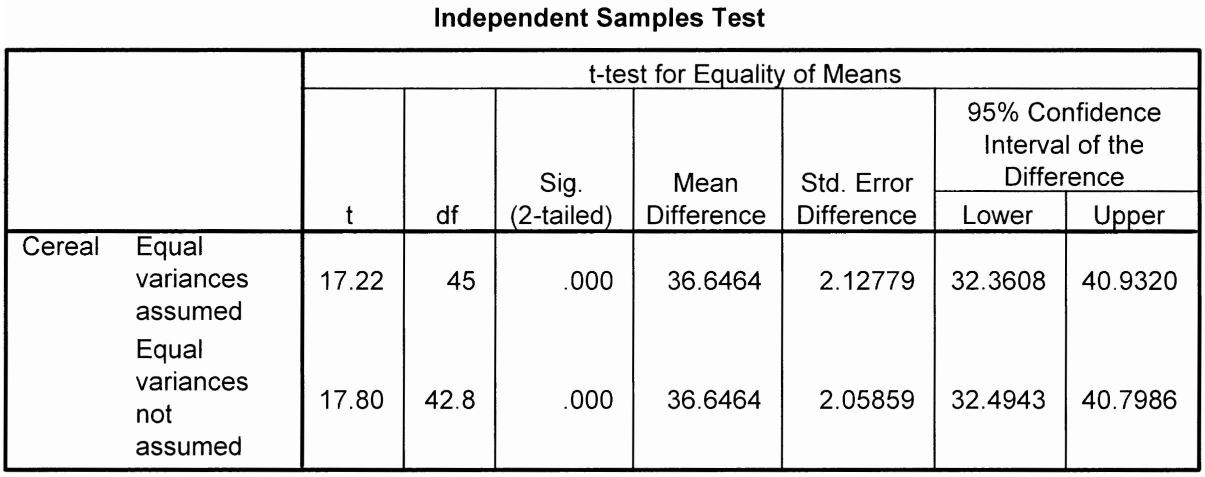
SPSS can be used to calculate confidence intervals, test statistic values and P-values

E.g.: SPSS output for Q7 from Chapter 24



Group Statistics

	Type of Cerea	N	Mean	td. Deviatio	Std. Error Mean
Cereal	Children's Ce	19	46.8000	6.41838	1.47248
	Adult's Cereal	28	10.1536	7.61239	1.43861



Independent Samples Test

		t-test for Equality of Means						
							95% Confidence Interval of the Difference	
		t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Cereal	Equal variances assumed	17.22	45	.000	36.6464	2.12779	32.3608	40.9320
	Equal variances not assumed	17.80	42.8	.000	36.6464	2.05859	32.4943	40.7986

t Procedures - Paired Samples (not independent)

This procedure is useful when we are analysing samples where each subject has been observed (measured) twice. For instance, each subject may be given two different treatments and we observe the difference in the response to the treatments for each subject.

More commonly, matched pair designs relate to before and after tests where we observe a subject's response variable both before and after a treatment is imposed. We then analyse the difference in the two values for each subject.

Hence, we are analyzing a single sample of data (a sample of differences within pairs of observations). Inference testing involves calculating the sample mean of the differences, and making inferences regarding the population mean difference.

Example 3

A medical researcher wishes to determine if the contraceptive pill has the undesirable side effect of reducing the blood pressure of the user.

The study involves recording the initial blood pressures of 15 university aged women. After they use the pill regularly for six months, their blood pressures are again recorded. The researcher wishes to draw inferences about the effect of the pill on blood pressure.

For each subject, we observe a pair of measurements: one before using the pill and the other after using the pill. The paired differences ***d*** **= before - after** are computed.

Table 1: Blood Pressure

Subject	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Before (1)	70	78	72	76	76	76	72	78	82	64	74	92	74	68	84
After (2)	68	70	62	70	66	58	68	52	64	72	74	60	74	72	74
Diff (*d*)	**2**	**8**	**10**	**6**	**10**	**18**	**4**	**26**	**18**	**-8**	**0**	**32**	**0**	**-4**	**10**

The observed mean of the differences is:

$$\bar{d} = \frac{\Sigma d}{n} = 8.8$$

The observed standard deviation is:

$$s_d = \sqrt{\frac{\Sigma(d-\bar{d})^2}{n-1}} = 10.98$$

Do the data substantiate the claim that the use of the pill reduces blood pressure?

Null Hypothesis $H_o : \mu_d = 0 = \Delta_0$

Alternative Hypothesis $H_a : \mu_d > 0$

Set the level of significance $\alpha = 0.05$

$df = n - 1 = 14$

Test statistic

$$t = \frac{\bar{d} - \Delta_0}{\frac{s_d}{\sqrt{n}}}$$

$$= \frac{8.8}{\frac{10.98}{\sqrt{15}}} = \frac{8.8}{2.84} = 3.10$$

$$P\text{-value} = P(\bar{d} > 8.8)$$
$$= P(t > 3.10) \qquad \text{with } df = 14$$
$$P\text{-value} < 0.005$$

Decision/Conclusion: Since the ***P*-value** $< \alpha$, we have evidence to reject $H_{o,}$ in favour of H_a. The pill appears to have a significant effect in reducing blood pressure.

Assuming the paired differences constitute a random sample from a normal population a **95% Confidence Interval** for the mean difference μ_d is given by:

point estimate ± margin of error

$$\bar{d} \pm t^*_{n-1} \cdot \frac{s_d}{\sqrt{\boldsymbol{n}}}$$

$$df = n - 1 = 14 \qquad t^*_{n-1} = 2.145$$

$$\bar{d} \pm t^*_{n-1} \times \frac{s_d}{\sqrt{\boldsymbol{n}}}$$
$$= 8.8 \pm 2.145 \cdot \frac{10.98}{\sqrt{15}}$$
$$= 8.8 \pm 6.08$$
$$= (2.72, 14.88)$$

Therefore, we are 95% confident that the mean decrease in blood pressure is between 2.72 and 14.88 units.