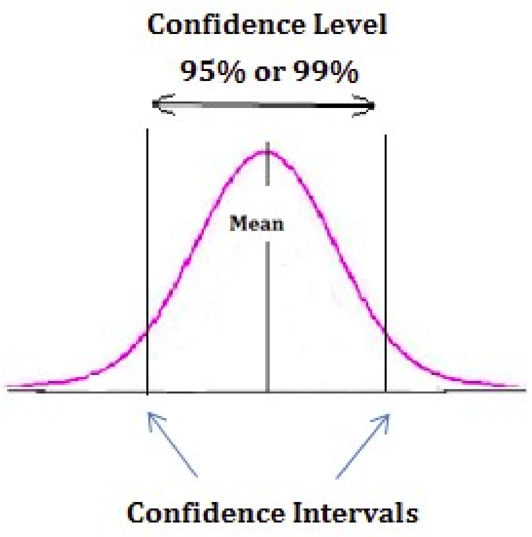
7. COMPARISON OF MEANS



7.1. *Methods for comparison of means*

- There are several *methods*:

  - comparison of a single observed *mean* with some *hypothesized* value [1];
  - comparison of two *means* arising from *paired* data [2];
  - comparison of two *means* from *unpaired* data [3].

- All can be made by using appropriate *confidence intervals* [4] and *t-tests*.

- The *single* mean and *paired* data cases are introduced first.

[1] another name "**one sample** t hypothesis test"

[2] e.g. when individual objects are measured **twice** (i.e. once for each type of measurement), or at two different times, etc.: e.g. to study the photosynthetic performance of ten plants in two environments in a greenhouse (***shady*** and ***sunny***), we could ***measure*** each individual plant ***twice***, once in the *shade* and once in the ***sun*** - the measures are ***paired*** by belonging to the ***same*** individual plant.

[3] another name "**two sample** t hypothesis test" and is probably the most **common**;

[4] a range of values we are fairly sure our true value lies in.

7.2. *Comparison of a single mean with a hypothesized value*

- Not very ***common*** in practice but it may be used (table) [1].

- E.g. these are the haemoglobin concentrations of 15 UK adult males admitted into an intensive care unit (ICU) [2].

8,1	10,1	12,3
9,7	11,7	11,3
11,9	9,3	13
10,5	8,3	8,8
9,4	6,4	5,4

- The population ***mean*** haemoglobin concentration in UK males is **15.0 g/dl** [3].

- Is there any ***evidence*** that critical illness is associated with an acute anemia?

[1] to compare a ***mean*** value from a ***sample*** with some ***hypothesized*** value (e.g. with some standards).

[2] Adapted from Whitley F, Ball J. Statistics review 5: Comparison of means. Crit. Care. 2002;6(5): 424–428.

[3] i.e. ***hypothesized*** value.

- The *mean* haemoglobin concentration of ICU men is **9.7 g/dl** [1].

- So, the question is:

- whether this difference is likely to be due to *random variation*?
- whether it is the result of some *systematic difference* between the men in the sample (ICU) and those in the general population?

- The best way to answer is:

- to calculate a *confidence interval* for the mean (1);
- to perform a *hypothesis test* (2).

[1] In practice any sample of 15 ICU men would be unlikely to have a mean haemoglobin of exactly 15.0 g/dl.

The **SD** of these data = **2.2 g/dl**.

A 95% *confidence interval* for the mean can be calculated using the **SE** = **SD**/√n [1]:

$$\mathbf{SE} = 2.2/\sqrt{15} = 0.56$$

The corresponding 95% *confidence interval* is as follows: [2]

$$9.7 \pm 2.14 \times 0.56 = 9.7 \pm 1.19 = (8.5,\ 10.9)$$ [3]

So, assuming that this sample is *representative*, it is likely that the true mean haemoglobin in the population of ICU adult male patients is between 8.5 and 10.9 g/dl.

[1] **SD** and **SE** both measure *variability*: high values of both indicate more dispersion, however, **SD** and **SE** are not the same:

SD quantifies the ***variability*** of data points around the *mean* in a given dataset (it tells us, on average, how far each data point is away from the mean).

SE quantifies the ***variability*** between samples drawn from the same population (indicates how different the sample mean is likely to be from the population mean).

[2] a ***confidence interval*** for μ (when the population σ is unknown) can be calculated as:

$$\bar{x} \pm t\frac{s}{\sqrt{n}}$$

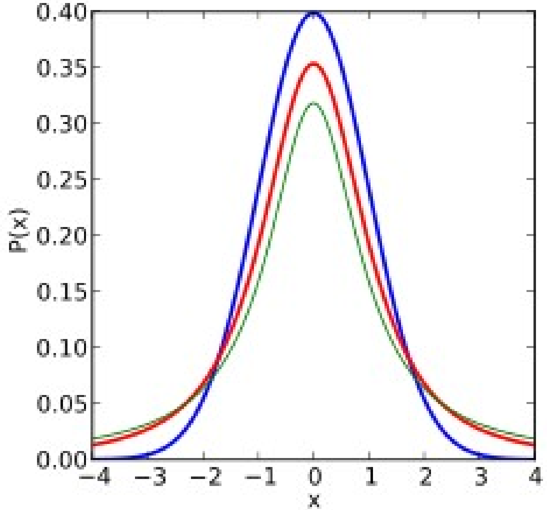
[3] the multiplier, in this case **2.14**, comes from the **t-distribution** because the sample size is ***small***.

$$t = \frac{sample\ mean - hypothesised\ mean}{SE\ of\ sample\ mean^{1}}$$

- The associated ***P value*** is obtained by *comparison* with the *t-distribution* [2] (regardless of sign) corresponding to smaller ***P values***.

- The *t-statistic* for the haemoglobin example is as follows:

$$t = \frac{9.7 - 15}{0.56} = \frac{-5.3}{0.56} = -9.54$$



Density of the t-distribution: 2 df – red; the standard normal distribution – blue; 1 df - green.

[1] it is the number of **SE**s that separate the sample ***mean*** from the hypothesized ***value***.

[2] the ***t-distribution*** (Student's t-distribution) is a ***probability*** distribution that is used to estimate population parameters when the sample size is ***small*** and/or when the population ***variance*** is unknown.

The shape of the t distribution is determined by the ***df***, which, in the case of the one sample t-test, is equal to the sample size minus 1, i.e. ***14*** (15-1).

Statistician William Sealy Gosset, known as "Student"

- The *observed* (t) mean haemoglobin concentration is **9.54**, which is *below* the *hypothesized* mean.

- *Tabulated* values indicate how likely this is to *occur*: for a sample size of **14** (15-1) *df* the *P* *value* is < than 0.0001 [1] (see next slide).

Conclusion:

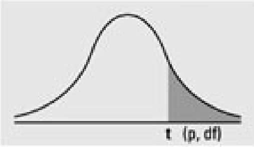
- it is *extremely* *unlikely* that the mean haemoglobin in this sample would differ from that in the general population to this extent by *chance* alone;

- this may *indicate* that there is a *truly* difference in haemoglobin concentrations in ICU men [2].

[1] the ***P value*** gives no indication of the ***size*** of any ***difference***; it merely indicates the probability that the difference arose by chance. In order to assess the ***magnitude*** of any difference, it is essential also to have the ***confidence*** ***interval*** calculated above.

[2] It is important to know how this sample of men was selected and whether they are representative of all UK men admitted to ICUs.

Numbers in each row of the table are values on a *t*-distribution with
(*df*) degrees of freedom for selected right-tail (greater-than) probabilities (*p*).



df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	43178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460
z	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905
CI	———	———	80%	90%	95%	98%	99%	99.9%

2.26

2.14

1.96

- *Conclusion*:

- the haemoglobin concentration in the *general* population of adult men in the UK is well *outside* this range;

- ICU men may *truly* have haemoglobin concentrations *lower* than the national average.

- Question: how likely it is that this *difference* is due to chance?

- Answer: we need a *hypothesis* test (one sample t-test) [1].

- H_0 is that *sample* mean (9.7 g/dl) is the *same* as *hypothesized* mean (15.0 g/dl).

- The *t statistic*, from which a ***P value*** is derived, is as follows.

[1] t-test formally examines how ***far*** the estimated mean haemoglobin of ICU men (9.7 g/dl), lies from the hypothesized value (15.0 g/dl).

7.3. *Comparison of two means arising from paired data*

- *Paired* data:
  - each data set has the *same number* of data points;
  - each *data* point in one data set is *related* to *one*, and only *one*, data point in the other data set [1].

E.g. table shows central venous O_2 saturation in 10 patients on admission and 6 hours after admission to an ICU [2].

Subject	On admission	6 h after admission	Difference (%)
1	39.7	52.9	13.2
2	59.1	56.7	-2,4
3	56.1	61.9	5.8
4	57.7	71.4	13.7
5	60.6	67.7	7.1
6	37.8	50	12.2
7	58.2	60.7	2.5
8	33.6	51.3	17.7
9	56	59.5	3.5
10	65.3	59.8	-5.5
Mean	**52.4**	**59.2**	**6.8**

[1] e.g. ***before***-***after*** drug test: you record the blood pressure of each subject in the study, ***before*** and ***after*** a drug is administered - each "***before***" measure is related only to the "***after***" measure from the same subject.

[2] Whitley F, Ball J. Statistics review 5: Comparison of means. Crit. Care. 2002;6(5): 424–428.

- The *means* were **52.4%** and **59.2%**, corresponding to an increase of **6.8%**.
- The question is:
  - whether this difference is likely to reflect a *truly effect* of admission and treatment;
  - or whether it is simply due to *chance*?
- **H_0 is that 52.4% = 59.2%.**

Subject	On admission	6 h after admission	Difference (%)
1	39.7	52.9	13.2
2	59.1	56.7	-2,4
3	56.1	61.9	5.8
4	57.7	71.4	13.7
5	60.6	67.7	7.1
6	37.8	50	12.2
7	58.2	60.7	2.5
8	33.6	51.3	17.7
9	56	59.5	3.5
10	65.3	59.8	-5.5
Mean	**52.4**	**59.2**	**6.8**

- The data are *paired* [1], and it is *important* to account for this pairing in the analysis.
- How to do this?
- To *concentrate* on the *differences between* the *pairs* of measurements rather than on the measurements themselves.
- H_0: the ***mean*** of the *differences* in central venous oxygen saturation = **0**.
- *T-test* therefore *compares* the observed *mean* of the differences with a hypothesized value of 0 [2].

Subject	On admission	6 h after admission	Difference (%)
1	39.7	52.9	13.2
2	59.1	56.7	-2,4
3	56.1	61.9	5.8
4	57.7	71.4	13.7
5	60.6	67.7	7.1
6	37.8	50	12.2
7	58.2	60.7	2.5
8	33.6	51.3	17.7
9	56	59.5	3.5
10	65.3	59.8	-5.5
Mean	**52.4**	**59.2**	**6.8**

[1] the two sets of observations are not independent of each other;

[2] i.e. the ***paired*** ***t-test*** is a special case of the single sample ***t-test*** described above.

- The ***t-statistic***:

$$t = \frac{sample\ mean\ of\ differences - 0}{SE\ of\ sample\ mean\ of\ differences}$$

$$t = \frac{sample\ mean\ of\ differences}{SE\ of\ sample\ mean\ of\ differences}$$

- The **SD** of the differences is **7.5,** and this corresponds to a **SE** of 7.5/√10 = **2.4.**

- The t-statistic: t = 6.8/2.4 = **2.87**, and this corresponds to a ***P value*** of 0.01 [1].

[1] based on a t-distribution with 10-1=9 df.

i. e. there is some evidence to suggest that admission to ICU and subsequent treatment may ***increase*** central venous oxygen saturation *beyond* the level expected by chance.

Subject	On admission	6 h after admission	Difference (%)
1	39.7	52.9	13.2
2	59.1	56.7	-2,4
3	56.1	61.9	5.8
4	57.7	71.4	13.7
5	60.6	67.7	7.1
6	37.8	50	12.2
7	58.2	60.7	2.5
8	33.6	51.3	17.7
9	56	59.5	3.5
10	65.3	59.8	-5.5
Mean	**52.4**	**59.2**	**6.8**

- Remember: ***P value*** gives no info. about the likely ***size*** of any ***effect***.

- This may be done by calculating a 95% ***confidence interval*** from the ***mean*** and **SE** of the differences:

$6.8 \pm 2.26 \times 2.4 = 6.8 \pm 5.34 = (1.4, 12.2)$ [1]

- So, the ***true increase*** in central venous O_2 saturation is probably ***between*** 1.4% and 12.2%.

- Although the ***increase*** may be small (1.4%), it is **unlikely** that the effect is to decrease saturation.

[1] the multiplier, in this case **2.26,** comes from the t-distribution because the sample size is ***small***. If n>30, z-tables are better.

Subject	On admission	6 h after admission	Difference (%)
1	39.7	52.9	13.2
2	59.1	56.7	-2,4
3	56.1	61.9	5.8
4	57.7	71.4	13.7
5	60.6	67.7	7.1
6	37.8	50	12.2
7	58.2	60.7	2.5
8	33.6	51.3	17.7
9	56	59.5	3.5
10	65.3	59.8	-5.5
Mean	**52.4**	**59.2**	**6.8**

7.4. *Comparison of two means arising from unpaired data*

- Comparison of two means arising from *unpaired data* [1] is most *common*.
- E.g. lets compare *early goal-directed therapy* (**EGDT**) with *standard therapy* (**ST**) in the treatment of septic shock [2].
- A total of 263 patients were randomized and 236 *completed* 6 hours of treatment.
- The *mean* arterial pressures after 6 hours of treatment in the **ST** and **EGDT** groups are shown in table [3].

	Mean arterial pressure (mmHg)	
	ST	**EGDT**
Number of patients	119	117
Mean	81	95
St. Dev.	18	19

[1] i.e. comparison of data from two ***independent*** groups;

[2] is a potentially fatal ***medical condition*** when organ injury or damage in response to infection leads to dangerously low blood pressure and abnormalities in cellular metabolism;

[3] Adapted from E. Whitley, and J. Ball. **Statistics review 5: Comparison of means.** Rivers E,et al. Early goal-directed therapy in the treatment of severe sepsis and septic shock. N Engl J Med 2001;345:1368–77.

- The *<u>mean</u>* arterial pressure was **14 mmHg** *<u>higher</u>* in the **EGDT**.

- The 95% *<u>confidence</u>* *<u>intervals</u>* for the mean arterial pressure in the two groups:

$$ST: 81 \pm 1.96 \times \frac{18}{\sqrt{119}} = 81 \pm 3.23 = [77.8; 84.2]$$

$$EGDT: 95 \pm 1.96 \times \frac{19}{\sqrt{117}} = 95 \pm 3.4 = [91.6; 98.4]$$

- There is no *<u>overlap</u>* between the two *<u>confidence</u>* *<u>intervals</u>* and there **may be** a *<u>difference</u>* between the two groups.

	Mean arterial pressure (mmHg)	
	ST	**EGDT**
Number of patients	119	117
Mean	81	95
St. Dev.	18	19

- Let us *estimate* the *size* of the difference.
- The only *difference* is in the *calculation* of the **SE**.
- In the *paired* case attention is focused on the *mean* of the *differences*.
- In the *unpaired* case interest is in the *difference* of the *means* [1].

	Mean arterial pressure (mmHg)	
	ST	**EGDT**
Number of patients	119	117
Mean	81	95
St. Dev.	18	19

[1] Because the ***sample sizes*** in the ***unpaired*** case may be (and indeed usually are) ***different***, the combined **SE** takes this into account and gives more weight to the ***larger sample size*** because this is likely to be more ***reliable***.

- The *pooled* **SD** for the *difference* in means:

$$SD_{difference} := \frac{\sqrt{(n_1 - 1) \times SD_1^2 + (n_2 - 1) \times SD_2^2}}{(n_1 + n_2 - 2)}$$

where SD_1 and SD_2 are the SDs in the two groups and n_1 and n_2 are the two sample sizes.

- The *pooled* **SE** for the *difference* in means:

$$SE_{difference} = SD_{difference} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- This **SE** can now be used to calculate a *confidence interval* for the difference in means and to perform an *unpaired* t-test, as above.

	Mean arterial pressure (mmHg)	
	ST	**EGDT**
Number of patients	119	117
Mean	81	95
St. Dev.	18	19

The pooled **SD** is:

$$SD_{difference} := \frac{\sqrt{(119-1)\times 18^2 + (117-1)\times 19^2}}{(119+117-2)}$$

$$SD_{difference} := \sqrt{\frac{38.232+41.876}{234}} = \sqrt{342.34} = 18.50$$

and the corresponding pooled **SE** is:

$$SD_{difference} = 18.50 \times \sqrt{\frac{1}{119}+\frac{1}{117}}$$

$$= 18.50 \times \sqrt{0.008+0.009}$$

$$= 18.50 \times 0.13 = 2.41$$

	Mean arterial pressure (mmHg)	
	ST	EGDT
Number of patients	119	117
Mean	81	95
St. Dev.	18	19

- The *difference* in *mean* arterial pressure between **EGDT** and **ST** groups is **14 mmHg**, with a corresponding 95% *confidence interval* of 14 ± 1.96 × 2.41= (9.3, 18.7) mmHg.

- The *confidence interval* suggests that the *true difference* is likely to be between 9.3 and 18.7 mmHg.

- To explore the likely role of chance in explaining this difference, an unpaired ***t-test*** can be performed.

	Mean arterial pressure (mmHg)	
	ST	**EGDT**
Number of patients	119	117
Mean	81	95
St. Dev.	18	19

- H_0 is that the difference in means = 0.
- As for the previous two cases, a *t-statistic* is calculated.

$$t = \frac{\textit{difference of sample means}}{\textit{SE of difference of sample means}}$$

- A *P value* may be obtained by comparison with the t-distribution on $n_1 + n_2 - 2$ df [1].
- t = 14/2.41 = **5.81**, with a corresponding ***P value*** < 0.0001 [2].
- *Conclusion*: there may be a *truly difference* between the two groups.

	Mean arterial pressure (mmHg)	
	ST	**EGDT**
Number of patients	119	117
Mean	81	95
St. Dev.	18	19

[1] the larger the t-statistic, the smaller the ***P value*** will be;

[2] **14 mmHg** is the ***difference*** in ***mean*** arterial pressure; it is extremely ***unlikely*** that a difference in mean arterial pressure of this magnitude would be observed just by chance.

7.5. *Comparison of two means: assumptions and limitations*

- The t-tests require certain *assumptions*.

- The *one sample* t-test: the data have an *approximately* Normal distribution.

- The *paired* t-test: the distribution of the **differences** are *approximately* Normal.

- The *unpaired* t-test: the data from the two samples are both *Normally* distributed [1].

- There are tests to *examine* whether a set of data are Normal or whether two **SD**s (or, two variances) are *equal* [2].

[1] and the **SDs** from the two samples are approximately ***equal***;

[2] the Kolmogorov-Smirnov test (K-S) and Shapiro-Wilk (S-W) test are designed to test normality by comparing your data to a normal distribution.

What to do if Normality is violated?

- The appropriate ***transformation*** [1] of the data may be used before performing any calculations ("Let the data decide").

- For "big" sample size (30, or even better 50), an impact to validity from non-normal data usually "small".

- If not possible, alternative tests can be used, e.g. ***nonparametric*** tests [2] .

- To explore differences in ***means*** across ***three*** or ***more*** groups, an analysis of ***variance*** (**ANOVA**) can be used.

[1] e.g. to apply a square root to each value or data may be log-transformed: your data may now be normal, but interpreting that data may be much more difficult;

[2] it is like a ***parallel*** universe to parametric tests, also called ***distribution-free*** tests: don't assume that your data follow a specific ***distribution***.

Can be performed well with ***skewed*** and ***nonnormal*** distributions. E.g. ***Mann-Whitney*** test instead of ***2-sample t test***; or ***Kruskal-Wallis, Mood's median test*** instead of ***One-Way ANOVA***, etc.