Confidence interval, sample-size formula and test statistic, concerning:

<u>Population mean :</u> (large sample, *n>30*):

$$\bar{x} \pm z_c \times \frac{s}{\sqrt{n}} \qquad n = \left(\frac{z_c \times s}{E} \right)^2 \qquad \frac{\bar{x} - \mu_0}{s} \times \sqrt{n}$$

Replace *s* by *F* if available (rare).

EXAMPLE: A random sample of 55 from a specific population yields the mean of 73.5 and standard deviation of 13.2 .
i) Construct 90% confidence interval for the population mean:

$73.5 \pm 1.645 \times 13.2 \div \sqrt{55} = 73.5 \pm 2.93$

Answer: (**70.57, 76.43**)

ii) How many more observations would we need to be within 1 unit of the correct answer (using the same level of confidence)?

$(1.645 \times 13.2 \div 1)^2 = 471.498$

Answer: 472 - 55 = 417

iii) Test H_0: : = 75 against H_1: : < 75 using " = 1%
Critical value: -2.326
Computed value: $(73.5-75) \times \sqrt{55} \div 13.2 = -0.843$
Not enough evidence to reject H_0.

Small sample: Need Normal population, use t_{n-1} instead of z.

EXAMPLE: Consider the following (random independent) sample: 72, 81, 69, 77, 73, from a specific Normal population.
i) Construct a 80% confidence interval for :.

Gx = 372, Gx^2 = 27764, $\bar{x}$ = 372 ÷ 5 =74.4

$$s = \sqrt{\frac{27764 - 372^2/5}{4}} = \sqrt{\frac{87.2}{4}} = 4.669$$

74.4 ± 1.533×4.669÷$\sqrt{5}$ = 74.4 ± 3.20 or (71.20, 77.60)

ii) Test H_0: : = 75 against H_1: : < 75 using " = 1%
Critical value: -3.747
Computed value: (74.4 -75)×$\sqrt{5}$÷4.669 = - **0.287**
Not enough evidence to reject H_0.

Population proportion *p* ($n\hat{p} > 5,\ n\hat{q} > 5$):

$$\hat{p} \pm z_c \times \sqrt{\frac{\hat{p} \times \hat{q}}{n}} \qquad \hat{p} \times \hat{q} \times \left(\frac{z_c}{E}\right)^2 \qquad \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 \times q_0}{n}}}$$

$$\frac{1}{4} \times \left(\frac{z_c}{E}\right)^2$$

EXAMPLE: In a sample of 72 Brock students, 37 use a bus to get to school.

i) The 95% CI for the proportion of all Brock students who use bus is: $\frac{37}{72} \pm 1.96 \times \sqrt{\frac{37 \times (72-37)}{72^3}}$ = 0.514 ± 0.115 or (39.9%, 62.9%)

ii) How many students (in total) do we need to sample to reduce the margin of error to ± 3%?

$\frac{37 \times (72-37)}{72^2} \times (1.96 \div 0.03)^2$ = 1066.3 ie. 1067

iii) Test H_0: p = ½ against H_1: p… ½ using " = 5%
Critical values: ± 1.96 Computed (observed) value of the test statistic: $\frac{37/72 - 1/2}{\sqrt{\frac{1}{4} \times \frac{1}{72}}} = 0.2357$ (cannot reject H_0).

<u>**Difference** in population means</u> (large samples)

$$\bar{x}_1 - \bar{x}_2 \pm z_c \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

EXAMPLE: A sample of 49 adult Canadian males had an average (sample mean) height of 179.2 cm, with the standard deviation of 7.4 cm. For a sample of 55 females, the average height was 167.8 cm, with the standard deviation of 9.4 cm.

i) Construct a 90% CI for the difference between the corresponding population means (male - female):

$$179.2 - 167.8 \pm 1.645 \times \sqrt{\frac{7.4^2}{49} + \frac{9.4^2}{55}}$$ = 11.4 ± 2.7 or

(8.7 cm,14.1 cm)

ii) Test H_0: : $_1$ = : $_2$ against H_1: : $_1$ > : $_2$ using '' = 1%
The critical value is **2.326**, the computed value of the test statistic:

$$\frac{179.2 - 167.8}{\sqrt{\frac{7.4^2}{49} + \frac{9.4^2}{55}}} = \frac{11.4}{1.6505} = 6.907$$

Conclusion: Yes, this represents a highly significant evidence that the female population mean is smaller than that of the male population.

<u>Small samples</u> (at least one # 30): Both populations must be *Normal* and have the *same* standard deviation F. Also, use *t* instead of *z*.

$$\bar{x}1 - \bar{x}2 \pm t_c \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \times \sqrt{\frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}}$$

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \times \sqrt{\frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}}}$$

and

EXAMPLE: Same as before, except now our male sample consists of only 172, 181, 177, 183, 169 cm, and the female sample is: 162, 158, 174, 171 cm.

i) Same as before.

$Gx_1 = 882$, $Gx_1^2 = 155724$, $Gx_2 = 665$, $Gx_2^2 = 110725$

This implies that $\bar{x}_1 - \bar{x}_2 = \frac{882}{5} - \frac{665}{4} = 10.15$ and

$$(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2 = \Sigma x_1{}^2 - \frac{(\Sigma x_1)^2}{n_1} + \Sigma x_2{}^2 - \frac{(\Sigma x_2)^2}{n_2}$$

$$= 155724 - \frac{882^2}{5} + 110725 - \frac{665^2}{4} = 307.95$$

So, we have:

$$10.15 \pm 1.895 \times \sqrt{0.2 + 0.25} \times \sqrt{307.95/7} = 10.15 \pm 4.45$$ or

(5.70, 14.60) cm

ii) Same as before.

Critical value of *t* is 2.998, computed value of the test statistic is:

$$\frac{10.15}{\sqrt{0.2+0.25} \times \sqrt{307.95/7}} = 2.281$$

Conclusion: Now, we cannot reject the null hypothesis !!

Difference in population *proportions* (both samples ‘large’ in the $n\hat{p}$ > 5, $n\hat{q}$ > 5 sense).

$$\hat{p}_1 - \hat{p}_2 \pm z_c \cdot \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \cdot \hat{p}_o \hat{q}_o}}$$

EXAMPLE: Out of a random sample of 51 Brock female students, 28 use bus to get to school, whereas for a sample of 42 male students the number is only 18.
i) Construct an 80% CI for the *populations'* difference:

$$\frac{28}{51}-\frac{18}{42}\pm 1.282\times\sqrt{\frac{28\times 23}{51^3}+\frac{18\times 24}{42^3}} = 0.1204 \pm 0.1325 \quad \text{or}$$

(-1.21%, 25.92%)

ii) Test, using 5% level of significance, whether the two population proportions are the same, against the $p_F > p_M$ alternate.

Critical value (z_c) is 1.645, the computed value of the test statistic is:
$$\frac{0.1204}{\sqrt{\left(\frac{1}{51}+\frac{1}{42}\right)\cdot\frac{46}{93}\cdot\frac{47}{93}}}=\frac{0.1204}{0.1042} = 1.156$$

Conclusion: We don't have sufficient evidence to prove that the proportion of female students who use bus is higher than that of their male counterparts (at 5% level of significance).

<u>Test for μ_d</u> (the population mean of paired differences). Test statistic:

$$\frac{\overline{d}}{s_d/\sqrt{n}}$$

EXAMPLE: A random sample of 7 cars was tested in terms of the distance travelled on one litre of regular, and one litre of high-octane gasoline. The results are given in the following table:

regular	7.31	11.23	8.60	9.42	10.08	10.70	8.18
hi-oct.	7.53	11.02	8.88	9.38	9.99	10.54	7.92

difference	-0.22	0.21	-0.28	0.04	0.09	0.16	0.26

We have to assume that the distribution of these differences is (at least approximately) Normal.

H_0: :$_d$ = 0 H_1: :$_d$ ... 0   " = 5%

Critical values: ±2.447 (Using *t* with 6 df.)

To compute the test statistic, we first need Gd = 0.26 and Gd^2 = 0.2738

Then:
$$\frac{0.26/7}{\sqrt{\dfrac{0.2738 - 0.26^2 / 7}{6 \times 7}}} = \frac{0.03714}{0.07930} = 0.468$$

Conclusion: No statistical evidence that either type of gasoline would be more economical than the other (at any practical level of significance).

<u>Regression</u>:

$$SS_x = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$$

$$SS_y = \Sigma y^2 - \frac{(\Sigma y)^2}{n}$$

$$SP_{xy} = \Sigma xy - \frac{(\Sigma x)\cdot(\Sigma y)}{n}$$

$b = SP_{xy} / SS_x$ and $a = \overline{y} - b \cdot \overline{x}$

Best (**least-squares**) straight (regression) **line**:

$$y = b \cdot x + a$$

Prediction interval for a new value of *y* taken at x_o:

Point estimate:

$$y_p \equiv a + b \cdot x_o$$

Residual standard error:

$$s_r \equiv \sqrt{\frac{SS_y - b \cdot SP_{xy}}{n-2}}$$

PI:

$$y_p \pm t_c \cdot s_r \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_o - \bar{x})^2}{SS_x}}$$

(Use *t* distribution with *n-2* df.)

To **test** $H_0\!: \ \$ = 0$ against ..... , use $\frac{b}{s_r} \times \sqrt{SS_x}$

(same distribution).

Correlation coefficient: $r \equiv \frac{SP_{xy}}{\sqrt{SS_x \cdot SS_y}}$

Coefficient of determination: r^2 (in %)