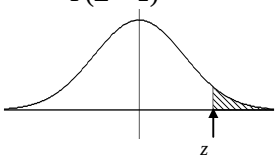
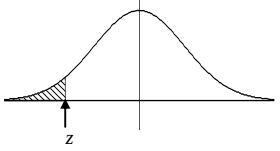
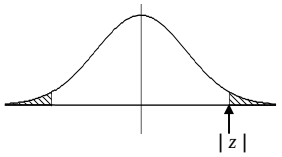
HYPOTHESIS TESTING

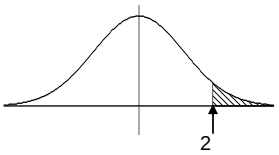
H_0	Test Statistic	H_1	p-value
$\mu = \mu_0$	Large sample size $n \geq 30$ $z = \dfrac{\overline{x} - \mu_0}{\sigma / \sqrt{n}}$	$\mu > \mu_0$	$P(Z > z)$ 
		$\mu < \mu_0$	$P(Z < z)$ 
		$\mu \neq \mu_0$	$P(\lvert Z \rvert > \lvert z \rvert)$ 
	Small sample size $n < 30$ $t = \dfrac{\overline{x} - \mu_0}{\hat{\sigma} / \sqrt{n}}$	$\mu > \mu_0$	$P(T > t)$ with df $= n - 1$
		$\mu < \mu_0$	$P(T < t)$
		$\mu \neq \mu_0$	$P(\lvert T \rvert > \lvert t \rvert)$
$p = p_0$	Large sample size $z = \dfrac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}}$	$p > p_0$	$P(Z > z)$
		$p < p_0$	$P(Z < z)$
		$p \neq p_0$	$P(\lvert Z \rvert > \lvert z \rvert)$
$\mu_1 = \mu_2$	Large sample size $n \geq 30$ $z = \dfrac{\overline{x}_1 - \overline{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$	$\mu_1 > \mu_2$	$P(Z > z)$
		$\mu_1 < \mu_2$	$P(Z < z)$
		$\mu_1 \neq \mu_2$	$P(\lvert Z \rvert > \lvert z \rvert)$
	Small sample size $n < 30$ $t = \dfrac{\overline{x}_1 - \overline{x}_2}{\hat{\sigma}_p \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ where pooled variance $\hat{\sigma}_p^2 = \dfrac{(n_1 - 1)\hat{\sigma}_1^2 + (n_2 - 1)\hat{\sigma}_2^2}{n_1 + n_2 - 2}$	$\mu_1 > \mu_2$	$P(T > t)$ with df $= n_1 + n_2 - 2$
		$\mu_1 < \mu_2$	$P(T < t)$
		$\mu_1 \neq \mu_2$	$P(\lvert T \rvert > \lvert t \rvert)$
	Matched pairs $t = \dfrac{\overline{d}}{\hat{\sigma}_d / \sqrt{n}}$	$\mu_1 > \mu_2$	$P(T > t)$ with df $= n - 1$
		$\mu_1 < \mu_2$	$P(T < t)$
		$\mu_1 \neq \mu_2$	$P(\lvert T \rvert > \lvert t \rvert)$
$p_1 = p_2$	Large sample size $z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\,\hat{q}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$ where $\hat{p} = \dfrac{x_1 + x_2}{n_1 + n_2}$	$p_1 > p_2$	$P(Z > z)$
		$p_1 < p_2$	$P(Z < z)$
		$p_1 \neq p_2$	$P(\lvert Z \rvert > \lvert z \rvert)$

Large-Sample Test about a Population Mean (sample size ≥ 30)

We claim that the mean chin-up scores of children is more than 2. A random sample of children produced the following data:

Sample size	$n = 48$
Sample mean	$\overline{x} = 2.5$
Population variance	$\sigma^2 = 3$

Do the data support our claim? Use a significance level of $\alpha = 0.05$.



Test statistic $z = \dfrac{\overline{x} - 2}{\sigma / \sqrt{n}} = \dfrac{2.5 - 2}{\sqrt{3} / \sqrt{48}} = 2$

p-value = $P(Z > 2) = 0.0228$

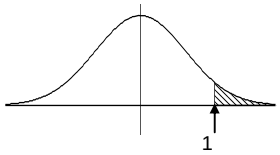
Since p-value = 0.0228 < 0.05, we have **sufficient evidence** at 5% level of significance to conclude that the mean chin-up score of children is **significantly more than 2**.

Small-Sample Test about a Population Mean (sample size < 30)

Same scenario but with a small sample:

Chin-up scores	2, 2, 5
Sample size	$n = 3$
Sample mean	$\overline{x} = \dfrac{2 + 2 + 5}{3} = 3$
Estimate of Population variance	$\hat{\sigma}^2 = \dfrac{2^2 + 2^2 + 5^2 - 3(3)^2}{3 - 1} = 3$

Degree of freedom $\nu = n - 1 = 3 - 1 = 2$



Test statistic $t = \dfrac{\overline{x} - 2}{\hat{\sigma} / \sqrt{n}} = \dfrac{3 - 2}{\sqrt{3} / \sqrt{3}} = 1$

p-value = $P(T(2) > 1) = 0.211$

Since p-value = 0.211 > 0.05, we have **insufficient evidence** at 5% level of significance to conclude that the mean chin-up score of children is **significantly more than 2**.

Large-Sample Test about a Population Proportion

We claim that more than half of all children can pass a certain fitness test. A random sample of children produced the following data:

Sample size	$n = 36$
No. of children who passed	$x = 21$

Do the data support our claim? Use a significance level of $\alpha = 0.05$.

Test statistic $z = \dfrac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}} = \dfrac{\frac{7}{12} - \frac{1}{2}}{\sqrt{\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) / 36}} = 1$

p-value = $P(Z > 1) = 0.159$

Since p-value = 0.159 > 0.05, we have **insufficient evidence** at 5% level of significance to conclude that the proportion of children who passed is **significantly more than half.**

Note: To ensure that the sample size is large enough, we need to check that the interval

$\hat{p} \pm 2\sqrt{\hat{p}\hat{q}/n} = \dfrac{7}{12} \pm 2\sqrt{\left(\dfrac{7}{12}\right)\left(\dfrac{5}{12}\right) / 36} = [0.419, 0.748]$ does not contain 0 or 1.

Large-Sample Test of Difference between 2 Population Means (sample size ≥ 30)

We claim that boys can do more chin-ups than girls. 2 random samples of children produced the following data:

	Boys	Girls
Sample size	$n_1 = 54$	$n_2 = 54$
Sample mean	$\overline{x}_1 = 4$	$\overline{x}_2 = 3$
Population variance	$\sigma_1^2 = 3$	$\sigma_2^2 = 3$

Do the data support our claim? Use a significance level of $\alpha = 0.05$.

Test statistic $z = \dfrac{\overline{x}_1 - \overline{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} = \dfrac{4 - 3}{\sqrt{\dfrac{3}{54} + \dfrac{3}{54}}} = 3$

p-value = $P(Z > 3) = 0.00135$

Since p-value = 0.00135 < 0.05, we have **sufficient evidence** at 5% level of significance to conclude that the mean chin-up score of boys is **significantly higher** than that of girls.

Small-Sample Test of Difference between 2 Population Means (sample size < 30)

Same scenario but with small samples:

	Boys	Girls
Scores	2, 5, 5	2, 2, 5
Sample size	$n_1 = 3$	$n_2 = 3$
Sample mean	$\overline{x}_1 = \dfrac{2 + 5 + 5}{3} = 4$	$\overline{x}_2 = \dfrac{2 + 2 + 5}{3} = 3$
Estimate of Population variance	$\hat{\sigma}_1^2 = \dfrac{2^2 + 5^2 + 5^2 - 3(4)^2}{3 - 1} = 3$	$\hat{\sigma}_2^2 = \dfrac{2^2 + 2^2 + 5^2 - 3(3)^2}{3 - 1} = 3$

Degree of freedom $\nu = 3 + 3 - 2 = 4$

Pooled variance $\hat{\sigma}_p^2 = \dfrac{(n_1 - 1)\hat{\sigma}_1^2 + (n_2 - 1)\hat{\sigma}_2^2}{n_1 + n_2 - 2} = \dfrac{2\,(3) + 2\,(3)}{4} = 3$

Test statistic $t = \dfrac{\overline{x}_1 - \overline{x}_2}{\sqrt{\hat{\sigma}_p^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} = \dfrac{4 - 3}{\sqrt{3\left(\dfrac{1}{3} + \dfrac{1}{3}\right)}} = 0.707$

p-value = $P(T(4) > 0.707) = 0.259$

Since p-value = 0.259 > 0.05, we have **insufficient evidence** at 5% level of significance to conclude that the mean chin-up score of boys is **significantly higher** than that of girls.

Test of Difference between 2 Population Means using Matched Pairs

We claim that children can do more chin-ups after taking up a special diet. A random sample of children produced the following data:

	No. of chin-ups		Difference
	Before	**After**	***d***
Ann	2	4	2
Bob	4	6	2
Clare	5	4	–1

Do the data support our claim? Use a significance level of $\alpha = 0.05$.

Sample mean $\overline{d} = \frac{2 + 2 - 1}{3} = 1$

Estimate of population variance $\hat{\sigma}_d^2 = \frac{2^2 + 2^2 + 1^2 - 3(1)^2}{3 - 1} = 3$

Degree of freedom $\nu = 3 - 1 = 2$

Test statistic $t = \frac{\overline{d}}{\hat{\sigma}_d / \sqrt{n}} = \frac{1}{\sqrt{3} / \sqrt{3}} = 1$

p-value = $P(T(2) > 1) = 0.211$

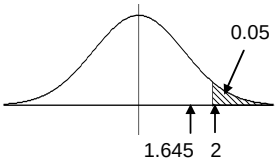
Since p-value = 0.211 > 0.05, we have **insufficient evidence** at 5% level of significance to conclude that the mean score is **significantly higher** after taking up the special diet.

Large-Sample Test of Difference between 2 Proportions

We claim that boys do better than girls in a fitness test. 2 random samples of children produced the following data:

	Boys	**Girls**
Sample size	$n_1 = 32$	$n_2 = 32$
No. of children who passed	$x_1 = 20$	$x_2 = 12$

Do the data support our claim? Use a significance level of $\alpha = 0.05$.



$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{20 + 12}{32 + 32} = \frac{32}{64} = \frac{1}{2}$

Test statistic $z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\,\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{\frac{5}{8} - \frac{3}{8}}{\sqrt{\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{32} + \frac{1}{32}\right)}} = 2$

p-value = $P(Z > 2) = 0.0228$

Since p-value = 0.0228 < 0.05, we have **sufficient evidence** at 5% level of significance to conclude that the proportion of boys who passed is **significantly higher** than the corresponding proportion for girls.

Note: To ensure that sample sizes are large enough, we need to check that the intervals

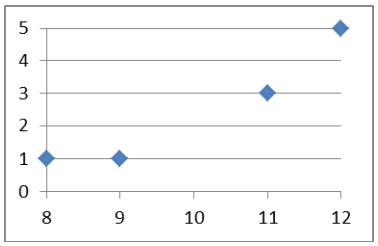
$\hat{p}_1 \pm 2\sqrt{\hat{p}_1\hat{q}_1 / n} = \frac{5}{8} \pm 2\sqrt{\left(\frac{5}{8}\right)\left(\frac{3}{8}\right) / 32} = [0.454, 0.796]$ and

$\hat{p}_2 \pm 2\sqrt{\hat{p}_2\hat{q}_2 / n} = \frac{3}{8} \pm 2\sqrt{\left(\frac{3}{8}\right)\left(\frac{5}{8}\right) / 32} = [0.204, 0.546]$ do not contain 0 or 1.

Correlation Analysis

The ages and chin-up scores of 4 children are as follows:

	Age x	Chin-up score y	xy	x^2	y^2
Ann	8	1	8	64	1
Bob	9	1	9	81	1
Clare	11	3	33	121	9
Dave	12	5	60	144	25
Sum	**40**	**10**	**110**	**410**	**36**



Are x and y linearly correlated? Use a significance level of $\alpha = 0.05$.

$$SS_{xy} = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n} = 110 - \frac{(40)(10)}{4} = 110 - 100 = 10$$

$$SS_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 410 - \frac{40^2}{4} = 410 - 400 = 10$$

$$SS_{yy} = \Sigma y^2 - \frac{(\Sigma y)^2}{n} = 36 - \frac{10^2}{4} = 36 - 25 = 11$$

$$\text{Correlation coefficient } r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} = \frac{10}{\sqrt{(10)(11)}} = 0.953$$

Since 0.953 > the critical value $r_{0.025,4} = 0.95$, we have **sufficient evidence** at 5% level of significance to conclude that x and y are **linearly correlated**.

Linear Regression

The primary goal of regression analysis is to predict the unknown value of one variable using the known value of another by means of a regression equation. Let $y_1, y_2, ..., y_n$ be the observed values corresponding to the values $x_1, x_2, ..., x_n$. The least square line $y = a + bx$ is the line that minimises the sum of the squares of the prediction errors ie. $S = \sum_{i=1}^{n} (y_i - a - bx_i)^2$.

$$\frac{\partial S}{\partial a} = -2\Sigma(y_i - a - bx_i) = -2(\Sigma y_i - na - b\Sigma x_i) = 0$$

$$a = \frac{\Sigma y_i}{n} - b\frac{\Sigma x_i}{n} = \bar{y} - b\bar{x}$$

$$\frac{\partial S}{\partial b} = -2\Sigma x_i(y_i - a - bx_i) = 0$$

$$\Sigma x_i y_i - a\Sigma x_i - b\Sigma x_i^2 = 0$$

$$\Sigma x_i y_i - (\bar{y} - b\bar{x})n\bar{x} - b\Sigma x_i^2 = 0$$

$$b(\Sigma x_i^2 - n\bar{x}^2) = \Sigma x_i y_i - n\bar{x}\bar{y}$$

$$b = \frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{\Sigma x_i^2 - n\bar{x}^2} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{SS_{xy}}{SS_x}$$

Example:

Same scenario. Find the estimated regression line. Use the estimated regression line to predict the average chin-up score of 10-year old children.

$$\bar{x} = \frac{40}{4} = 10, \quad \bar{y} = \frac{10}{4} = 2.5$$

$$b = \frac{110 - 4(10)(2.5)}{410 - 4(10)^2} = \frac{110 - 100}{410 - 400} = \frac{10}{10} = 1$$

$$a = 2.5 - 1(10) = -7.5$$

Therefore the estimated regression line is $y = -7.5 + x$.

When $x = 10$, the predicted value of y is 2.5.