T-test and ANOVA

1. Student's t-test

The Student t-test is a hypothesis testing tool by comparing means. A *one sample t-test* tests the hypothesis that the mean of the variable is the same as the theoretical expected value of the variable.

A *two sample t-test* tests the hypothesis that the mean of the variable is the same in two different samples.

A one sample t-test allows us to test whether a sample mean (of a normally distributed interval variable) significantly differs from a hypothesised value about the population.

Below are the assumptions of a one sample t-test

(1) Your dependent variable should be scale in nature.
(2) Your observed data should be unrelated (i.e. independent of each other).
(3) There should be no significant outliers.
(4) The dependent variable should be approximately normally distributed.

Definition 1.1. The following formula is used to calculate the *t-statistic* for the one sample t-test

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}},$$

where $\bar{x}$ is the sample mean, μ is the population mean, s is the standard deviation of the sample and n is the number of individuals in the sample. The degrees of freedom is equal to $n - 1$.

Example 1.2. Say we know that the hypothesised population mean height is 68 inches. Then we take a sample of 25 people at random and find that mean height in this sample is 70 inches with a standard deviation of 4 inches. Based on our observations can we conclude that the value of the hypothesised population mean is correct? First we need to formulate our null and alternative hypothesis.

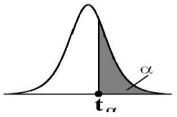
- Our null hypothesis, H_0, is that the population mean is 68 inches.
- The alternative hypothesis, H_1, is the population mean is greater than 68 inches.

Solution: $\bar{x} = 70$, $\mu = 68$, $\sigma = 4$, $n = 25$ and $df = 24$.

$$t = \frac{70 - 68}{4/\sqrt{25}} = \frac{2}{4/5} = 2.5.$$

Now we need to use the tables and degrees of freedom to see if we reject or accept H_0. We will choose a 0.05 level of significance.

Percentage Points of the t Distribution; $t_{\nu,\alpha}$
$P(T > t_{\nu,\alpha}) = \alpha$



	α													
ν	**0.40**	**0.30**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.02**	**0.015**	**0.01**	**0.0075**	**0.005**	**0.0025**	**0.0005**
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706	15.895	21.205	31.821	42.434	63.657	127.322	636.590
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303	4.849	5.643	6.965	8.073	9.925	14.089	31.598
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182	3.482	3.896	4.541	5.047	5.841	7.453	12.924
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776	2.999	3.298	3.747	4.088	4.604	5.598	8.610
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571	2.757	3.003	3.365	3.634	4.032	4.773	6.869
6	0.265	0.553	0.906	1.134	1.440	1.943	2.447	2.612	2.829	3.143	3.372	3.707	4.317	5.959
7	0.263	0.549	0.896	1.119	1.415	1.895	2.365	2.517	2.715	2.998	3.203	3.499	4.029	5.408
8	0.262	0.546	0.889	1.108	1.397	1.860	2.306	2.449	2.634	2.896	3.085	3.355	3.833	5.041
9	0.261	0.543	0.883	1.100	1.383	1.833	2.262	2.398	2.574	2.821	2.998	3.250	3.690	4.781
10	0.260	0.542	0.879	1.093	1.372	1.812	2.228	2.359	2.527	2.764	2.932	3.169	3.581	4.587
11	0.260	0.540	0.876	1.088	1.363	1.796	2.201	2.328	2.491	2.718	2.879	3.106	3.497	4.437
12	0.259	0.539	0.873	1.083	1.356	1.782	2.179	2.303	2.461	2.681	2.836	3.055	3.428	4.318
13	0.259	0.538	0.870	1.079	1.350	1.771	2.160	2.282	2.436	2.650	2.801	3.012	3.372	4.221
14	0.258	0.537	0.868	1.076	1.345	1.761	2.145	2.264	2.415	2.624	2.771	2.977	3.326	4.140
15	0.258	0.536	0.866	1.074	1.341	1.753	2.131	2.249	2.397	2.602	2.746	2.947	3.286	4.073
16	0.258	0.535	0.865	1.071	1.337	1.746	2.120	2.235	2.382	2.583	2.724	2.921	3.252	4.015
17	0.257	0.534	0.863	1.069	1.333	1.740	2.110	2.224	2.368	2.567	2.706	2.898	3.222	3.965
18	0.257	0.534	0.862	1.067	1.330	1.734	2.101	2.214	2.356	2.552	2.689	2.878	3.197	3.922
19	0.257	0.533	0.861	1.066	1.328	1.729	2.093	2.205	2.346	2.539	2.674	2.861	3.174	3.883
20	0.257	0.533	0.860	1.064	1.325	1.725	2.086	2.197	2.336	2.528	2.661	2.845	3.153	3.850
21	0.257	0.532	0.859	1.063	1.323	1.721	2.080	2.189	2.328	2.518	2.649	2.831	3.135	3.819
22	0.256	0.532	0.858	1.061	1.321	1.717	2.074	2.183	2.320	2.508	2.639	2.819	3.119	3.792
23	0.256	0.532	0.858	1.060	1.319	1.714	2.069	2.177	2.313	2.500	2.629	2.807	3.104	3.768
24	0.256	0.531	0.857	1.059	1.318	1.711	2.064	2.172	2.307	2.492	2.620	2.797	3.091	3.745
25	0.256	0.531	0.856	1.058	1.316	1.708	2.060	2.167	2.301	2.485	2.612	2.787	3.078	3.725
26	0.256	0.531	0.856	1.058	1.315	1.706	2.056	2.162	2.296	2.479	2.605	2.779	3.067	3.707
27	0.256	0.531	0.855	1.057	1.314	1.703	2.052	2.158	2.291	2.473	2.598	2.771	3.057	3.690
28	0.256	0.530	0.855	1.056	1.313	1.701	2.048	2.154	2.286	2.467	2.592	2.763	3.047	3.674
29	0.256	0.530	0.854	1.055	1.311	1.699	2.045	2.150	2.282	2.462	2.586	2.756	3.038	3.659
30	0.256	0.530	0.854	1.055	1.310	1.697	2.042	2.147	2.278	2.457	2.581	2.750	3.030	3.646
40	0.255	0.529	0.851	1.050	1.303	1.684	2.021	2.123	2.250	2.423	2.542	2.704	2.971	3.551
60	0.254	0.527	0.848	1.045	1.296	1.671	2.000	2.099	2.223	2.390	2.504	2.660	2.915	3.460
120	0.254	0.526	0.845	1.041	1.289	1.658	1.980	2.076	2.196	2.358	2.468	2.617	2.860	3.373
∞	0.253	0.524	0.842	1.036	1.282	1.645	1.960	2.054	2.170	2.326	2.432	2.576	2.807	3.291

From the t-table on above we should reject the null hypothesis if $t > 1.711$. Since $t = 2.5 > 1.711$, we can reject the null hypothesis and conclude that the population mean must be greater than 68 inches.

Now that we know how to do it by hand we'll go through the procedure in Minitab. Here are the steps to test the assumptions in of the one sample t-test in Minitab;

- We'll assume we meet the first two assumptions.
- To detect outliers we make a boxplot for the height.
  (1) Minitab will use $*$ to highlight any outliers.
  (2) In Minitab Graph $\rightarrow$ Boxplot $\rightarrow$ One $Y \rightarrow$ Simple.
  (3) Enter ``$\mathrm{Height}$" into the ``Graph Variable" box.
  (4) Then click ``OK".
- Now that we've seen we don't have any outliers we need to check that our dependent variable (Height) is approximately normally distributed.
- To check if our data is approximately normally distributed we need to look at the histograms of the height.
  (1) Go to Graph $\rightarrow$ Histogram $\rightarrow$ With Fit.
  (2) Enter ``$\mathrm{Height}$" into the ``Graph Variable" box.
  (3) Then Click ``OK".

Now that we've checked the assumptions we can carry out the t-test with the following steps.

- Go to **Stat** $\rightarrow$ **Basic Statistics** $\rightarrow$ **1-Sample t...**.
- Put the "Height" variable in the text box below "one or more sample, each in a column".
- Click on "Perform hypothesis test".
- Beside "Hypothesized mean" enter 68.
- Click "OK".

We see that we get the same results

2. Two sample t-test

The following formula is used to calculate the **t-statistic** for the two sample t-test, when both samples have size n

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\left(s_1^2 + s_2^2\right)/n}},$$

where $\overline{x}_1$ is the mean of the first sample, $\overline{x}_2$ is the mean of the second sample, s_1 is the standard deviation of the first sample sample and s_2 is the standard deviation of the second sample sample. The degrees of freedom is equal to $2n - 2$.

Example 2.1. Suppose that we measured the biomass (milligrams) produced by bacterium ``A" and bacterium ``B" in shake flasks containing glucose as substrate. We had 4 replicate flasks of each bacterium. The measurements for mean and standard deviation for bacteria A and B were $(512.25, 8.808)$ and $(501.75, 8.884)$ respectively. Can we say there is a significant variation between the amount of biomass produced by each bacteria?

Solution: Clearly $\overline{x}_1 = 512.25$, $S_1 = 8.808$, $\overline{x}_2 = 501.75$, $S_2 = 8.884$ and $n = 4$. Then

$$t = \frac{512.25 - 501.75}{\sqrt{\frac{8.808^2 + 8.884^2}{4}}} = 1.678623717.$$

Here's the steps to perform a two sample t-test in Minitab.

- We'll assume we meet the first three assumptions (mainly because if we've broken those there's nothing we can do about it at the analysis stage!).
- To detect outliers we make a boxplot for the growth in each group.
  (1) Minitab will use $*$ to highlight any outliers.
  (2) In Minitab Graph $\rightarrow$ Boxplot $\rightarrow$ One Y $\rightarrow$ With Groups.
  (3) Enter ``Growth" into the ``Graph Variable" box and ``Bacteria" into "Categorical variables".
  (4) Then click ``OK".
- Now that we've seen we don't have any outliers we need to check that our dependent variable (Growth) is approximately normally distributed.
  (1) To check if our data is approximately normally distributed we need to look at the histograms of the growth in each group.
  (2) Go to Graph $\rightarrow$ Histogram $\rightarrow$ With Fit and Groups.
  (3) Enter ``Growth" into the ``Graph Variable" box and ``Bacteria" into "Categorical variables".
  (4) Then Click OK.

- The last assumption we need to check is homogeneity of variance.
  (1) This means that we don't want there to be a significant difference in the variance of our two samples.
  (2) We use Levene's test to check this assumption and we want a result greater than 0.05.
  (3) Go to Stat $\rightarrow$ Two Sample $\rightarrow$ Variances.
  (4) Put ``$\mathrm{Growth}$" into the ``$\mathrm{Samples}$" box and ``$\mathrm{Bacteria}$" into the "Samples IDs" box.
  (5) Click "OK".

3. One-way ANOVA

ANOVA provides a statistical test of whether or not the means of several groups are equal. In other words an ANOVA test is a more general form of a t-test.

A one-way analysis of variance (ANOVA) is used when you have a categorical independent variable (with two or more categories) and a normally distributed interval dependent variable and you wish to test for differences in the means of the dependent variable broken down by the levels of the independent variable.

The assumptions of a one-way ANOVA is as follows:

(1) Your dependent variable should be scale in nature.
(2) Your independent variable should be nominal in nature and have two or more categories.
(3) There should be independence of observations.
(4) There should be no significant outliers.
(5) The dependent variable should be approximately normally distributed.
(6) There should be homogeneity of variance.

You can test these assumptions the same way we did for the two sample t-test.

Example 3.1. Suppose that we measured the biomass (milligrams) produced by bacterium ``$\mathrm{A}$", bacterium ``B" and bacterium ``C" in shake flasks containing glucose as substrate. We had 4 replicate flasks of each bacterium. The measurements for A, B and C were $\{500, 512, 520, 517\}$, $\{490, 510, 500, 507\}$ and $\{475, 450, 490, 495\}$, respectively. Can we say there is a significant variation between the amount of biomass produced by each bacteria?

To solve this problem using a t-test we would need to run the test three times!

To perform a one-way ANOVA follow these steps;

(1) Compute the **sum of squares within**, SS_w.
(2) Compute the **sum of squares between**, SS_b.
(3) Use the previous two steps to get the total sum of squares, SS_t.
(4) Calculate the **degrees of freedom** within and between, df_w, df_b.
(5) Divide SS_w by df_w to get the mean square within, MS_w.
(6) Divide SS_b by df_b to get the mean square between, MS_b.
(7) Divide MS_w by MS_b to get the F-ratio.

To compute the sum of squares within for each group we subtract the mean for that group from each value in the group, square the result then add all the values up. We know from the previous example that $\overline{x}_A = 512.25$, $\overline{x}_B = 501.75$ and

$$\overline{x}_C = \frac{475 + 450 + 490 + 495}{4} = 477.5$$

So now we can calculate the sum of squares within for each group.

$$SS_w(A) = (500 - 512.25)^2 + (512 - 512.25)^2 + (520 - 512.25)^2 + (517 - 512.25)^2$$
$$SS_w(A) = (-12.25)^2 + (-0.25)^2 + (7.75)^2 + (4.75)^2 = 232.75$$

$$SS_w(B) = (490 - 501.75)^2 + (510 - 501.75)^2 + (500 - 501.75)^2 + (507 - 501.75)^2$$
$$SS_w(B) = (-11.75)^2 + (8.25)^2 + (-1.75)^2 + (5.25)^2 = 236.75$$

$$SS_w(C) = (475 - 477.5)^2 + (450 - 477.5)^2 + (490 - 477.5)^2 + (495 - 477.5)^2$$
$$SS_w(C) = (-2.5)^2 + (-27.5)^2 + (12.5)^2 + (17.5)^2 = 1225$$

$$SS_w = 232.75 + 236.75 + 1225 = 1694.5$$

To calculate the sum of squares between first we have to calculate to overall mean. Then we subtract the overall mean from the mean of each group, square it then multiply by the number of values in the group. Finally we add these all up.

$$\overline{x} = \frac{500 + 512 + 520 + 517 + 490 + 510 + 500 + 507 + 475 + 450 + 490 + 495}{12} = 497.17$$

$$SS_b(A) = 4 \times (512.25 - 497.17)^2 = 909.63$$

$$SS_b(B) = 4 \times (501.75 - 497.17)^2 = 83.91$$

$$SS_b(C) = 4 \times (477.5 - 497.17)^2 = 1547.64$$

$$SS_b = 909.63 + 83.91 + 1547.64 = 2541.18$$

$$SS_t = 2541.18 + 1694.5 = 4235.68$$

- Getting the degrees of freedom is much easier than calculating the sum of squares.
- The degrees of freedom between is one less than the number of groups, so in this example $df_b = 3 - 1 = 2$.
- The degrees of freedom within is the total number of values we have minus the number of groups $df_w = 12 - 3 = 9$
- The total degrees of freedom is $df_t = df_w + df_b = 9 + 2 = 11$.
- The mean square between is equal to $MS_b = \frac{SS_b}{df_b} = \frac{2541.18}{2} = 1270.59$.
- The mean square between is within to $MS_w = \frac{SS_w}{df_w} = \frac{1694.5}{9} = 188.28$.
- The F-ratio is

$$F = \frac{MS_b}{MS_w} = \frac{1270.59}{188.28} = 6.75.$$

F - Distribution ($\alpha = 0.05$ in the Right Tail)

df_2 \ df_1	Numerator Degrees of Freedom								
Denominator Degrees of Freedom	1	2	3	4	5	6	7	8	9
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385
3	10.128	9.5521	9.2766	9.1172	9.0135	8.9406	8.8867	8.8452	8.8123
4	7.7086	9.9443	6.5914	6.3882	6.2561	6.1631	6.0942	6.0410	6.9988
5	6.6079	5.7861	5.4095	5.1922	5.0503	4.9503	4.8759	4.8183	4.7725
6	5.9874	5.1433	4.7571	4.5337	4.3874	4.2839	4.2067	4.1468	4.0990
7	5.5914	4.7374	4.3468	4.1203	3.9715	3.8660	3.7870	3.7257	3.6767
8	5.3177	4.4590	4.0662	3.8379	3.6875	3.5806	3.5005	3.4381	3.3881
9	5.1174	4.2565	3.8625	3.6331	3.4817	3.3738	3.2927	3.2296	3.1789
10	4.9646	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355	3.0717	3.0204
11	4.8443	3.9823	3.5874	3.3567	3.2039	3.0946	3.0123	2.9480	2.8962
12	4.7472	3.8853	3.4903	3.2592	3.1059	2.9961	2.9134	2.8486	2.7964
13	4.6672	3.8056	3.4105	3.1791	3.0254	2.9153	2.8321	2.7669	2.7144
14	4.6001	3.7389	3.3439	3.1122	2.9582	2.8477	2.7642	2.6987	2.6458
15	4.5431	3.6823	3.2874	3.0556	2.9013	2.7905	2.7066	2.6408	2.5876
16	4.4940	3.6337	3.2389	3.0069	2.8524	2.7413	2.6572	2.5911	2.5377
17	4.4513	3.5915	3.1968	2.9647	2.8100	2.6987	2.6143	2.5480	2.4943
18	4.4139	3.5546	3.1599	2.9277	2.7729	2.6613	2.5767	2.5102	2.4563
19	4.3807	3.5219	3.1274	2.8951	2.7401	2.6283	2.5435	2.4768	2.4227
20	4.3512	3.4928	3.0984	2.8661	2.7109	2.5990	2.5140	2.4471	2.3928
21	4.3248	3.4668	3.0725	2.8401	2.6848	2.5727	2.4876	2.4205	2.3660
22	4.3009	3.4434	3.0491	2.8167	2.6613	2.5491	2.4638	2.3965	2.3419
23	4.2793	3.4221	3.0280	2.7955	2.6400	2.5277	2.4422	2.3748	2.3201
24	4.2597	3.4028	3.0088	2.7763	2.6207	2.5082	2.4226	2.3551	2.3002
25	4.2417	3.3852	2.9912	2.7587	2.6030	2.4904	2.4047	2.3371	2.2821
26	4.2252	3.3690	2.9752	2.7426	2.5868	2.4741	2.3883	2.3205	2.2655
27	4.2100	3.3541	2.9604	2.7278	2.5719	2.4591	2.3732	2.3053	2.2501
28	4.1960	3.3404	2.9467	2.7141	2.5581	2.4453	2.3593	2.2913	2.2360
29	4.1830	3.3277	2.9340	2.7014	2.5454	2.4324	2.3463	2.2783	2.2229
30	4.1709	3.3158	2.9223	2.6896	2.5336	2.4205	2.3343	2.2662	2.2107
40	4.0847	3.2317	2.8387	2.6060	2.4495	2.3359	2.2490	2.1802	2.1240
60	4.0012	3.1504	2.7581	2.5252	2.3683	2.2541	2.1665	2.0970	2.0401
120	3.9201	3.0718	2.6802	2.4472	2.2899	2.1750	2.0868	2.0164	1.9588
∞	3.8415	2.9957	2.6049	2.3719	2.2141	2.0986	2.0096	1.9384	1.8799

Since we had $\mathrm{df_b} = 2$, we go to column two and since $\mathrm{df}_w = 9$ we go to row nine. This a critical F-value of 4.2565. The F-ratio we had was $6.75 > 4.2565$ so we can reject the null hypothesis.

	SS	df	MS	F
Between	2541.18	2	1270.59	6.75
Within	1694.5	9	188.28	
Total	4235.68	11		

To perform a one way ANOVA in Minitab, follow these steps.

(1) Click on the **Stat** tab.
(2) Go to **ANOVA**, then click on **One-Way**
(3) In the text box beside **Response**, enter the Variable of interest.
(4) In the text box beside **Factor**, enter the grouping variable.
(5) Click ok.

The Analysis of Variance table below shows the results of our analysis

Source	DF	Adj SS	Adj MS	F-value	P-value
Bacteria	2	2541.17	1270.58	6.75	0.0162
Error	9	1694.50	188.28		
Total	11	4235.67			

However if we checked the assumptions (as you should) in Minitab you'll see that this example breaks the assumption that the dependant variable is approximately normally distributed. Luckily there is another test we can perform in this case, the Kruskal-Wallis test. Below are the assumptions of the Kruskal-Wallis test.

(1) The dependent variable should be either ordinal or scale in nature.
(2) The dependent variable should be nominal with two or more categories.
(3) There should be independence of observations.

To perform a Kruskal-Wallis test follow these steps:

(1) $Stat \rightarrow Nonparametrics \rightarrow Kruskal - Wallis.$
(2) $Growth$ is the response variable and $Bacteria$ is the factor variable.

Bacteria	N	Median	Ave Rank	Z
A	4	514.5	9.9	2.29
B	4	503.5	6.8	0.17
C	4	482.5	2.9	-2.46
Overall	12		6.5	

H = 7.57	DF = 2	P=0.023
H = 7.62	DF=2	P=0.022

$*$NOTE$*$ One or more small samples.

Minitab is telling us that our sample is still too small, in practice you should gather more data!

4. Two-way ANOVA

A two way ANOVA is used to determine if there is any interaction between two independent variables on a single dependent variable. Below are the assumptions of a two-way ANOVA:

(1) Your dependent variable should be scale in nature.
(2) Your two independent variables should be nominal in nature and have two or more categories.
(3) There should be independence of observations.
(4) There should be no significant outliers.
(5) The dependent variable should be approximately normally distributed for each combination of the groups of independent variables.
(6) There should be homogeneity of variance for each combination of the groups of independent variables.

Example 4.1. We are interested in the effect the type of bacteria and the solution used have on the amount of biomass produced. Here again we have bacteria "A", "B" and "C" and the solutions are "X", "Y" and "Z". Below are the measurements for each bacteria along with the solution used in each case.

$$A = \{(505, Y), (512, Y), (520, Y), (517, Y), (520, Y), (523, Y), (520, Y), (517, Y), (523, Z)\}$$
$$B = \{(495, X), (505, X), (500, X), (505, X), (495, X), (495, X), (505, Z), (500, X), (500, X)\}$$
$$C = \{(480, Z), (480, Z), (485, X), (485, Z), (475, Z), (490, Z), (480, Z), (480, X), (470, Y)\}$$

To perform this test in Minitab follow these steps:

- Go to **Stat**, then **ANOVA**, **General Linear Model** and finally **Fit General Linear Model**.
- Put the dependent variable in responses.
- Enter the two independent variables in Factors table.
- Then click on **Model**.
- Click on your independent variables under the **Factors and covariates** box while holding shift. Then click add.
- Click OK, then click OK again in the General Linear model box.

This gives us the table:

Source	DF	Adj SS	Adj MS	F-value	P-value
Bacteria	2	4069.11	2034.55	83.63	$<$0.0001
Solution	2	151.42	75.71	3.11	0.0637
Error	23	559.52	24.33		
Lack-of-Fit	2	43.30	21.65	0.88	0.4292
Pure Error	21	516.22	24.58		
Total	27	6836.11			

From the p-values we can see that the type of bacteria has a significant effect on the biomass produced ($p < 0.001$) but the type of solution doesn't ($p = 0.0637$).