Hypothesis Testing: Inferential Statistics

It allows us to measure behavior in samples to learn more about populations. Because population are often too large and inaccessible. We use sample because sample is taken out from the population.

Example: The average height of the students in overall engineering colleges. Then we evaluate average on some sample expected to be the average of population.

Assume that probability of selecting any other sample is normally distributed.

The method in which we select sample to learn more characteristics about population is called hypothesis testing.

Hypothesis testing is systematic way to test claims about a group or population.

Hypothesis Testing is a method for testing a claim or hypothesis about a population using data measured in a sample.

Hypothesis Testing is the method of testing whether claims or hypothesis regarding the population are likely to be true.

Steps in Method of Hypothesis Testing:

1. Identify the Hypothesis, we feel to be tested
2. We select the criteria upon which we decide the hypothesis being tested is true
3. Select the random sample from the population and measure the sample mean
4. Compare what we observe in the sample to what we expect to observe   what we expect to observe is true.
5. If the discrepancy is small, we will likely decide the claim is true, if the discrepancy is large we will reject the claim.

Step 1: State the Hypothesis

Step 2: Set the criteria for a decision

Step 3: Compute the test statistic

Step 4: Make a decision

Step 1: State the Hypothesis: We begin by stating the value of the population mean in a null hypothesis, which presumes to be true. We assume that the null hypothesis is true. The basis of the decision is to determine whether this assumption is likely to be true.

Null Hypothesis denoted by H0 is a statement about a population parameter such as population mean which is assumed to be true.

Alternative Hypothesis: we state that assumption is wrong since Null hypothesis is assumed to be true It is our responsibility to prove that the Null Hypothesis is wrong.

Alternative Hypothesis statement contradicts the Null Hypothesis by stating the the actual value of population parameter is less than, greater than or not equal to the value stated in the null hypothesis

The Null Hypothesis is assumed to be true. A researcher conducts a study showing that this assumption is unlikely (rejects null hypothesis) or fails (retains null hypothesis) to do so.

Step 2:

Set a criterion for decision. To set a criteria for decision we state the level of significance for a test.

Level of significance is typically set at 5% in behavioural research studies. When the probability of obtaining a sample mean is less than the 5% if the null hypothesis were true, then we conclude that sample we selected is too unlikely and so we reject the null hypothesis.

Level of significance refers to criterion of judgement upon which decision is made regarding the value stated in a null hypothesis. The criterion is based on the probability of obtaining a statistic measured in a sample if the value stated in the null hypothesis were true.

The alternative hypothesis decides where to place the level of significance

The empirical rule tells us that atleast 95% of all sample means fall within about 2 standard deviation of the population mean.

The alternative hypothesis determines whether to place the level of significance in one or both tails of sampling distribution.

Step 3: Compute the test statistic

Test statistic is a mathematical formula that allow us to determine the likelihood of obtaining sample outcomes.

Step 4: Make a decision We use the value computed in the test statistic to make a decision about the null hypothesis. The decision is based on the probability of obtaining a sample mean given that the value stated in the null hypothesis is true. If the probability of obtaining a sample mean is less than 5% when the null hypothesis is true then the decision is to reject the null hypothesis. If the probability of obtaining a sample mean is greater than 5% when the null hypothesis is true then decision to retain null hypothesis.

Decision can be:

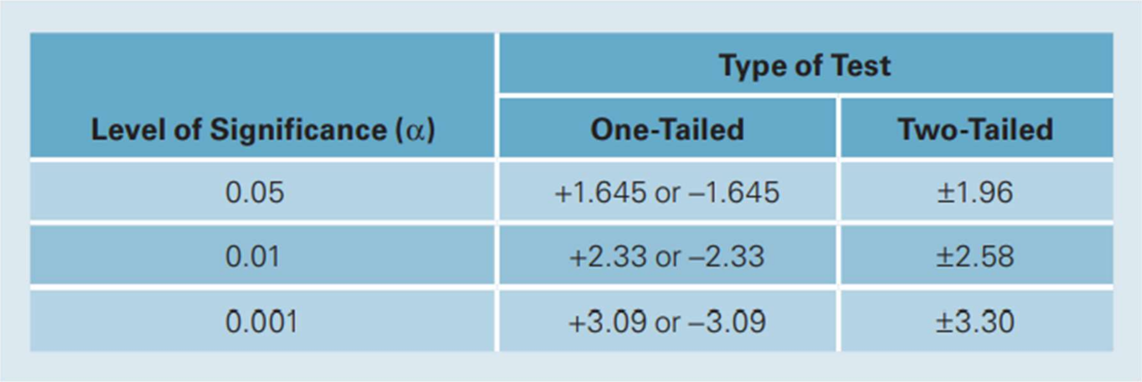
1. Reject the Null Hypothesis: The sample mean is associated with a low probability of occurrence when the null hypothesis is true
2. Retain the Null Hypothesis: The sample mean is associate with a high probability of occurrence when the null hypothesis is True.

The probability of obtaining a sample mean given that the value stored in the null hypothesis is true, is stated by the p Value. The p value is probability value IT varies from 0 to 1 can never be negative

A p value is the probability of obtaining a sample outcome given that the value stated in the null hypothesis is true. The p value for obtaining a sample outcome is compared to the level of significance

P value<= 5% we reject the null hypothesis

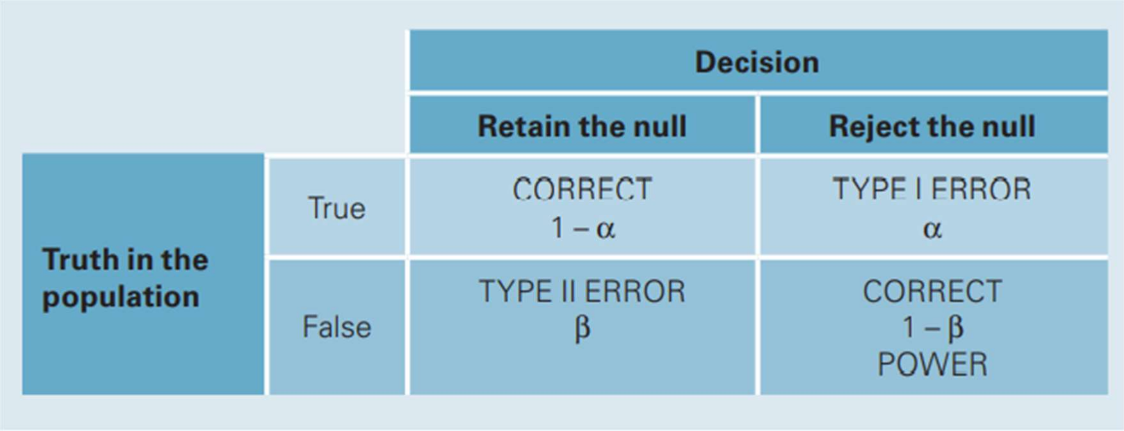
P value > 5% we retain the null hypothesis



Level of Significance (α)	Type of Test	
	One-Tailed	Two-Tailed
0.05	+1.645 or −1.645	±1.96
0.01	+2.33 or −2.33	±2.58
0.001	+3.09 or −3.09	±3.30

Type of Errors: When we decide to retain or reject Null hypothesis, we are observing sample not an entire population, it is possible that a conclusion may be wrong

1. The decision to retain the null hypothesis could be correct.
2. The decision to retain the null hypothesis could be incorrect.
3. The decision to reject the null hypothesis could be correct
4. The decision to reject the null hypothesis could be incorrect.



		Decision	
		Retain the null	Reject the null
Truth in the population	True	CORRECT $1 - \alpha$	TYPE I ERROR α
	False	TYPE II ERROR β	CORRECT $1 - \beta$ POWER

The incorrect decision is to retain a false null hypothesis. This decision is an example of a Type II error, or Beta error. the null hypothesis is true, we control for Type I error by stating a level of significance. The level we set, called the alpha level, is the largest probability of committing a Type I error that we will allow and still decide to reject the null hypothesis.

The usual line of reasoning is as follows:

1. There is an initial research hypothesis of which the truth is unknown.
2. The first step is to state the relevant null and alternative hypotheses. This is important, as mis-stating the hypotheses will muddy the rest of the process.

3. The second step is to consider the statistical assumptions being made about the sample in doing the test; for example, assumptions about the statistical independence or about the form of the distributions of the observations. This is equally important as invalid assumptions will mean that the results of the test are invalid.
4. Decide which test is appropriate, and state the relevant test statistic *T*.
5. Derive the distribution of the test statistic under the null hypothesis from the assumptions. In standard cases this will be a well-known result. For example, the test statistic might follow a Student's t distribution with known degrees of freedom, or a normal distribution with known mean and variance. If the distribution of the test statistic is completely fixed by the null hypothesis we call the hypothesis simple, otherwise it is called composite.
6. Select a significance level (α), a probability threshold below which the null hypothesis will be rejected. Common values are 5% and 1%.
7. The distribution of the test statistic under the null hypothesis partitions the possible values of *T* into those for which the null hypothesis is rejected—the so-called *critical region*—and those for which it is not. The probability of the critical region is α. In the case of a composite null hypothesis, the maximal probability of the critical region is α.
8. Compute from the observations the observed value t_{obs} of the test statistic *T*.
9. Decide to either reject the null hypothesis in favor of the alternative or not reject it. The decision rule is to reject the null hypothesis H_0 if the observed value t_{obs} is in the critical region, and not to reject the null hypothesis otherwise.

Chi-square test

Chi-square is used to test hypotheses about the distribution of observations in different categories. The null hypothesis (Ho) is that the observed frequencies are the same as the expected frequencies. If the observed and expected frequencies are the same, then $\chi^2 = 0$. If the frequencies you observe are different from expected frequencies, the value of χ^2 goes up.

Applications of Chi square:

- Chi-Square "Goodness of Fit" test: This is used when you have categorical data for one independent variable, and you want to see whether the distribution of your data is similar or different to that expected.

- Chi-Square Test of Association between two variables: This is appropriate to use when you have categorical data for two independent variables, and you want to see if there is an association between them.

Chi-Square "Goodness of Fit" test

This is used when you have **one independent variable**, and you want to compare an observed frequency-distribution to a theoretical expected frequency-distribution.

For the example described above, there is a single independent variable (in this example “**age group**”) with a number of different levels (17-20, 21-30, 31-40, 41-50, 51-60 and over 60). The statistical question is: do the frequencies you actually observe differ from the expected frequencies by more than chance alone?

In this case, we want to know whether or not our observed frequencies of traffic accidents occur equally frequently for the different ages groups (so that our theoretical frequency-distribution contains the same number of individuals in each of the age bands).

The way in which we would collate this data would be to use a *contingency table*, containing both the observed and expected frequency information.

	Age band						
	17-20	21-30	31-40	41-50	51-60	over 60	*Total:*
Observed frequency of accidents	25	15	5	5	5	5	*60*
Expected frequency of accidents	10	10	10	10	10	10	*60*

To work out whether these two distributions are significantly different from one another, we use the following Chi-square formula:

$$\chi^2 = \Sigma \frac{(O - E)^2}{E}$$

This translates into:

χ^2 = sum of (i.e., across categories) (observed frequency – expected frequency)^2
(divided by)
expected frequency

This may look complicated, but really it just means that you have to follow four simple steps, which are described on the next page.

Step One
Take each observed frequency and subtract from it its associated expected frequency (i.e., work out (O-E)):

25-10 = **15** 15-10 = **5** 5-10 = **-5** 5-10 = **-5** 5-10 = **-5** 5-10 = **-5**

Step Two
Square each value obtained in step 1 (i.e., work out $(O\text{-}E)^2$):
225 **25** **25** **25** **25** **25**

Step Three
Divide each of the values obtained in step 2, by its associated expected frequency (i.e., work out $\frac{(O\text{-}E)^2}{E}$):

$\frac{225}{10} = \mathbf{22.5}$ $\frac{25}{10} = \mathbf{2.5}$ $\frac{25}{10} = \mathbf{2.5}$ $\frac{25}{10} = \mathbf{2.5}$ $\frac{25}{10} = \mathbf{2.5}$ $\frac{25}{10} = \mathbf{2.5}$

Step Four
Add together all of the values obtained in step 3, to get your value of Chi-Square:

$\chi^2 = 22.5 + 2.5 + 2.5 + 2.5 + 2.5 + 2.5 = \mathbf{35}$

Assessing the size of our obtained Chi-Square value:

What you do, in a nutshell...

(a) Work out how many "degrees of freedom" (d.f.) you have.
(b) Decide on a probability level.
(c) Find a table of "critical Chi-Square values" (in most statistics textbooks).
(d) Establish the critical Chi-Square value for *this* particular test, and compare to your obtained value.

If your obtained Chi-Square value is **bigger** than the one in the table, then you conclude that **your obtained Chi-Square value is too large to have arisen by chance**; it is more likely to stem from the fact that there were real differences between the observed and expected frequencies. In other words, contrary to our null hypothesis, the categories did *not* occur with similar frequencies.

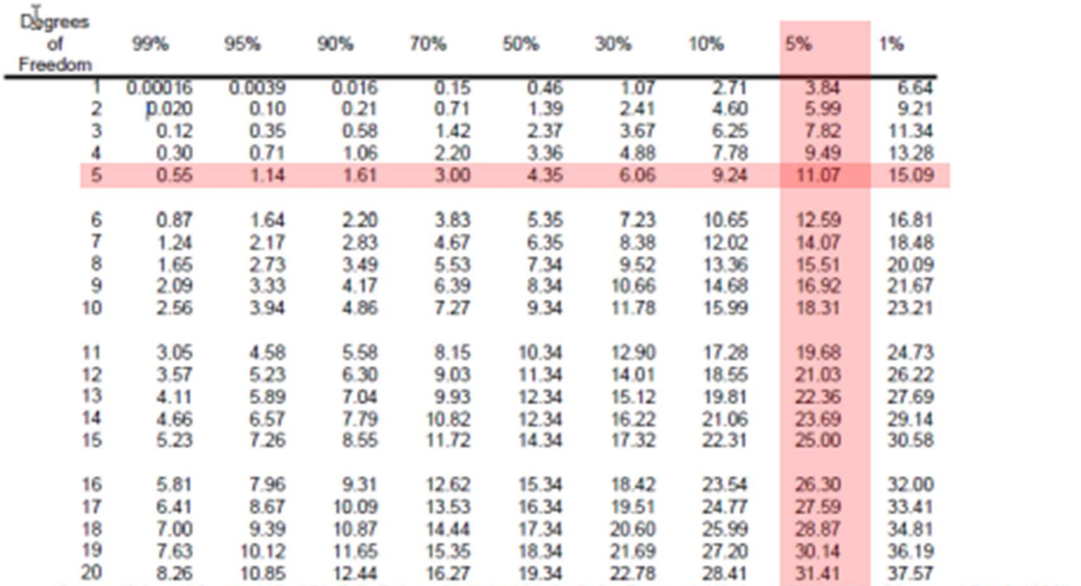
If, on the other hand, your obtained Chi-Square value is **smaller** than the one in the table, you conclude that there is no reason to think that the observed pattern of frequencies is not **due simply to chance** (i.e., we retain our initial assumption that the discrepancies between the observed and expected frequencies are due merely to random sampling variation, and hence we have no reason to believe that the categories did not occur with equal frequency).

For our worked example...

(a) First we work out our degrees of freedom. For the Goodness of Fit test, this is simply the number of categories minus one. As we have six categories, there are 6-1 = **5** degrees of freedom.

(b) Next we establish the probability level. In psychology, we use **p < 0.05** as standard – and this is represented by the **5% column**.

(c) We now need to consult a table of "critical values of Chi-Square". Here's an excerpt from a typical table:



Degrees of Freedom	99%	95%	90%	70%	50%	30%	10%	5%	1%
1	0.00016	0.0039	0.016	0.15	0.46	1.07	2.71	3.84	6.64
2	0.020	0.10	0.21	0.71	1.39	2.41	4.60	5.99	9.21
3	0.12	0.35	0.58	1.42	2.37	3.67	6.25	7.82	11.34
4	0.30	0.71	1.06	2.20	3.36	4.88	7.78	9.49	13.28
5	0.55	1.14	1.61	3.00	4.35	6.06	9.24	11.07	15.09
6	0.87	1.64	2.20	3.83	5.35	7.23	10.65	12.59	16.81
7	1.24	2.17	2.83	4.67	6.35	8.38	12.02	14.07	18.48
8	1.65	2.73	3.49	5.53	7.34	9.52	13.36	15.51	20.09
9	2.09	3.33	4.17	6.39	8.34	10.66	14.68	16.92	21.67
10	2.56	3.94	4.86	7.27	9.34	11.78	15.99	18.31	23.21
11	3.05	4.58	5.58	8.15	10.34	12.90	17.28	19.68	24.73
12	3.57	5.23	6.30	9.03	11.34	14.01	18.55	21.03	26.22
13	4.11	5.89	7.04	9.93	12.34	15.12	19.81	22.36	27.69
14	4.66	6.57	7.79	10.82	12.34	16.22	21.06	23.69	29.14
15	5.23	7.26	8.55	11.72	14.34	17.32	22.31	25.00	30.58
16	5.81	7.96	9.31	12.62	15.34	18.42	23.54	26.30	32.00
17	6.41	8.67	10.09	13.53	16.34	19.51	24.77	27.59	33.41
18	7.00	9.39	10.87	14.44	17.34	20.60	25.99	28.87	34.81
19	7.63	10.12	11.65	15.35	18.34	21.69	27.20	30.14	36.19
20	8.26	10.85	12.44	16.27	19.34	22.78	28.41	31.41	37.57

Source: Adapted from p.112 of Sir R.A. Fisher, *Statistical Methods fro Research Workers* (Edinburgh: Oliver and Boyd, 1958).

(d) The values in each column are "critical" values of Chi-Square. These values would be expected to occur by chance with the probability shown at the top of the column. The relevant value for *this* test is found at the intersection of the appropriate d.f. row and probability column. As our obtained Chi-Square has 5 d.f., we are interested in the values in the **5 d.f.** row. As the probability level is **p <.05**, we then need to look in the **5% column** (as .05 represents a chance level of 5 in 100... or 5%) to find the critical value for this statistical test. In this case, the critical value is **11.07**.

Finally, we need to compare our *obtained* Chi-Square to the critical value. If the obtained Chi-Square is larger than a value in the table, it implies that it is unlikely to have occurred by chance. Our obtained value of **35** is much larger than the critical value of 11.07. We can therefore be relatively confident in concluding that our *observed* frequencies are significantly different from the frequencies that we would *expect* to obtain if all categories were equally distributed. In other words, age *is* related to the amount of road traffic accidents that occur.

Chi-Square Test of Association between two variables

The second type of chi square test we will look at is the Pearson's chi-square test of association. You use this test when you have categorical data for **two** independent variables, and you want to see if there is an association between them.

For this example, let's stick with the theme of driving, but this time consider gender performance on driving tests. This time we have two categorical variables: Gender (two levels: Male vs Female) and Driving Test Outcome (two levels: Pass vs Fail).

In this case, the statistical question we want to know the answer to is whether driving test outcome is related to the gender of the person taking the test. Or in other words, we want to know if males show a different pattern of pass/fail rates than females.

To answer this question, we would start off by putting out data into a *contingency table*, this time containing only the *observed* frequency information. We can then use this table to calculate *expected* frequencies.

In this case, imagine the pattern of driving test outcomes looked like this:

	Male	**Female**
Pass	7	11
Fail	13	9

In this example, simply looking at the observed frequencies gives us an idea that the pattern of driving test outcomes may be different for the genders. It seems females have a more successful pass/fail rate than males. However, to test whether this observed difference is significant, we need to look at the outcome of a Chi-Square test. As with the one-variable Chi-Square test, our aim is to see if the pattern of observed frequencies is significantly different from the pattern of frequencies which we would expect to see by chance - i.e., what we would expect to obtain if there was no relationship between the two variables in question. With respect to the example above, "no relationship" would mean that the pattern of driving test performance for males was no different to that for females.

The Chi-Square formula is exactly the same as for the one-variable test described earlier; the only difference is in how you calculate the expected frequencies.

Step 1: Add numbers across columns and rows. Calculate total number in chart.

	Male	**Female**	
Pass	7	12	**= 19**
Fail	13	8	**= 21**
	= 20	**= 20**	**= 40**

Step 2: Calculate expected numbers for each individual cell (i.e. the frequencies we would expect to obtain if there were no association between the two variables). You do this by multiplying row sum by column sum and dividing by total number.

$$\textbf{Expected Frequency} = \frac{\textbf{Row Total x Column Total}}{\textbf{Grand Total}}$$

For example: using the first cell in table (Male/Pass);

$$\frac{19 \times 20}{40} = 9.5$$

...and the cell below (Male/Fail):

$$\frac{21 \times 20}{40} = 10.5$$

Do this for each cell in the table above.

Step 3: Now you should have an observed number and expected number for each cell. The observed number is the number already in 1st chart. The expected number is the number found in the last step (step 2). Redo the contingency table, this time adding in the expected frequencies in brackets below the obtained frequencies:

	Male	Female	
Pass	7 *(9.5)*	12 *(9.5)*	total=19
Fail	13 *(10.5)*	8 *(10.5)*	total=21
	total=20	total=20	grand total =40

Step 4: Now calculate Chi Square using the same formula as before:

$$\chi^2 = \Sigma \frac{(O-E)^2}{E}$$

...or...

$$\chi^2 = \textbf{Sum of } \frac{\textbf{(Observed - Expected)}^2}{\textbf{Expected}}$$

Calculate this formula for each cell, one at a time. For example, cell #1 (Male/Pass):

Observed number is: **7**
Expected number is: **9.5**

Plugging this into the formula, you have: $\mathbf{\frac{\underline{(7 - 9.5)^2}}{9.5} = 0.6579}$

Continue doing this for the rest of the cells.

Step 5: Add together all the final numbers for each cell, obtained in Step 4. **There are 4 total cells, so at the end you should be adding four numbers together for you final Chi Square number.**

In this case, you should have:

$$\mathbf{0.6579 + 0.6579 + 0.5952 + 0.5952 = 2.5062}$$

So, **0.095** is our obtained value of Chi-Square; it is a single-number summary of the discrepancy between our obtained frequencies, and the frequencies which we would expect if there was no association between our two variables. *The bigger this number, the greater the difference between the observed and expected frequencies.*

Step 6: Calculate degrees of freedom (*df*):

(Number of Rows – 1) x (Number of Columns – 1)
(2 – 1) x (2 – 1)
1 x 1
=1 *df* (degrees of freedom)

Assessing the size of our obtained Chi-Square value:

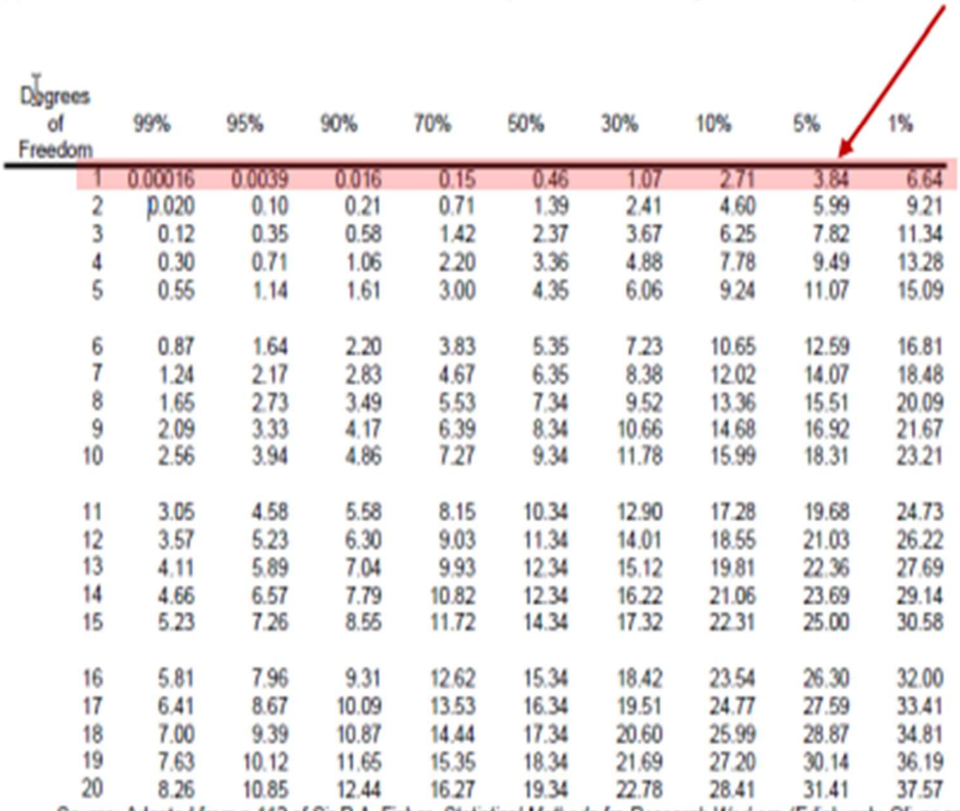
The procedure here is the same as for the Goodness of Fit test. We just need:

(a) Our "degrees of freedom" (d.f.) ✓
(b) A suitable probability level (p=0.05 in psychology) ✓
(c) A table of "critical Chi-Square values" ✓
(d) Establish the critical Chi-Square value for this test (at the intersection of the appropriate d.f. row and probability column), and compare to your obtained value.

As before, if the obtained Chi-Square value is **bigger** than the one in the table, then you conclude that **your obtained Chi-Square value is too large to have arisen by chance**. This would mean that the two variables are likely to be related in some way. **NB** - the Chi-Square test merely tells you that there is *some relationship* between the two variables in question: it does not tell you what that relationship is, and most importantly, it does not tell you anything about the causal relationship between the two variables.

If, on the other hand, your obtained Chi-Square value is **smaller** than the one in the table, you cannot reject the null hypothesis. In other words, you would conclude that your variables are unlikely to be associated.

For this example: Using the chart below, at the p = 0.05 significance level, with 1 *df*, the critical value can be established as **3.84**. Therefore, in order to reject the null hypothesis, the final answer to the Chi Square must be **greater or equal to 3.84**.



Degrees of Freedom	99%	95%	90%	70%	50%	30%	10%	5%	1%
1	0.00016	0.0039	0.016	0.15	0.46	1.07	2.71	3.84	6.64
2	0.020	0.10	0.21	0.71	1.39	2.41	4.60	5.99	9.21
3	0.12	0.35	0.58	1.42	2.37	3.67	6.25	7.82	11.34
4	0.30	0.71	1.06	2.20	3.36	4.88	7.78	9.49	13.28
5	0.55	1.14	1.61	3.00	4.35	6.06	9.24	11.07	15.09
6	0.87	1.64	2.20	3.83	5.35	7.23	10.65	12.59	16.81
7	1.24	2.17	2.83	4.67	6.35	8.38	12.02	14.07	18.48
8	1.65	2.73	3.49	5.53	7.34	9.52	13.36	15.51	20.09
9	2.09	3.33	4.17	6.39	8.34	10.66	14.68	16.92	21.67
10	2.56	3.94	4.86	7.27	9.34	11.78	15.99	18.31	23.21
11	3.05	4.58	5.58	8.15	10.34	12.90	17.28	19.68	24.73
12	3.57	5.23	6.30	9.03	11.34	14.01	18.55	21.03	26.22
13	4.11	5.89	7.04	9.93	12.34	15.12	19.81	22.36	27.69
14	4.66	6.57	7.79	10.82	12.34	16.22	21.06	23.69	29.14
15	5.23	7.26	8.55	11.72	14.34	17.32	22.31	25.00	30.58
16	5.81	7.96	9.31	12.62	15.34	18.42	23.54	26.30	32.00
17	6.41	8.67	10.09	13.53	16.34	19.51	24.77	27.59	33.41
18	7.00	9.39	10.87	14.44	17.34	20.60	25.99	28.87	34.81
19	7.63	10.12	11.65	15.35	18.34	21.69	27.20	30.14	36.19
20	8.26	10.85	12.44	16.27	19.34	22.78	28.41	31.41	37.57

Source: Adapted from p.112 of Sir R.A. Fisher, *Statistical Methods fro Research Workers* (Edinburgh: Oliver and Boyd, 1958).

The Chi Square calculation above was **2.5062.** This number is ***less*** than the critical value of **3.84**, so in this case the null hypothesis cannot be rejected. In other words, there does not appear to be a significant association between the two variables: males and females have a statistically similar pattern of pass/fail rates on their driving tests.

T Test:

A t-test is a type of inferential statistic used to determine if there is a significant difference between the means of two groups, which may be related in certain features. It is mostly used when the data sets, like the data set recorded as the outcome from flipping a coin 100 times, would follow a normal distribution and may have unknown variances.

A t-test is used as a hypothesis testing tool, which allows testing of an assumption applicable to a population. A t-test looks at the t-statistic, the t-distribution values, and the degrees of freedom to determine the statistical significance. To conduct a test with three or more means, one must use an analysis of variance.

Essentially, a t-test allows us to compare the average values of the two data sets and determine if they came from the same population.

T-Test Assumptions

1. The first assumption made regarding t-tests concerns the scale of measurement. The assumption for a t-test is that the scale of measurement applied to the data collected follows a continuous or ordinal scale, such as the scores for an IQ test.

2. The second assumption made is that of a simple random sample, that the data is collected from a representative, randomly selected portion of the total population.

3. The third assumption is the data, when plotted, results in a normal distribution, bell-shaped distribution curve.

4. The final assumption is the homogeneity of variance. Homogeneous, or equal, variance exists when the standard deviations of samples are approximately equal.

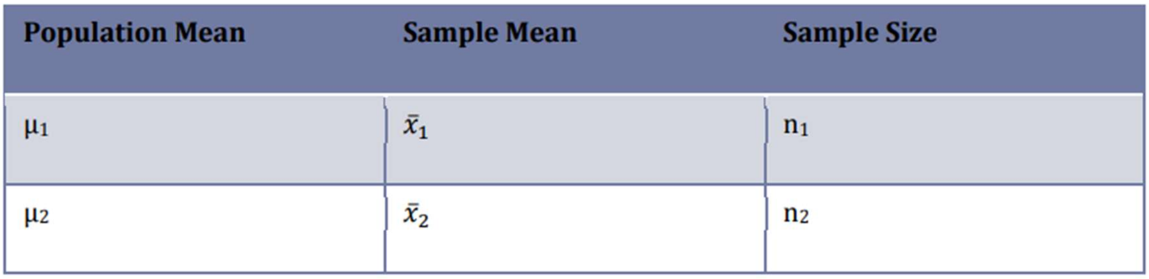
Two-Sample Problems

Researchers may want to compare two independent groups. With matched samples, the same individuals are tested twice, or pairs of individuals who are very similar in some respect are tested.

♣ Independent samples consist of two groups of individuals who are randomly selected from two different populations.

♣ The term “independent” is used because the individuals in one sample must be completely unrelated to the individuals in the other sample.

♣ Example: to find out if test scores are significantly different between males and females, researchers would need to randomly select a group of females and randomly select a group of males. There are two groups and they come from two separate populations (one population is males and the other separate population is females). Each of these populations has a mean for the variable.



Population Mean	Sample Mean	Sample Size
μ_1	$\bar{x}_1$	n_1
μ_2	$\bar{x}_2$	n_2

Conditions for Inference Comparing Two Means

Before conducting any statistical analyses, two assumptions must be met:

1) The two samples are random and they come from two distinct populations. The samples are independent. That is, one sample has no influence on the other. Matching violates independence, for example. Additionally, the same response variable must be measured for both samples.

2) Both populations are Normally distributed. The means and standard deviations of the populations are unknown. In practice, it is enough that the distributions have similar shapes and that the data have no strong outliers

The Two-Sample t Statistic

When data come from two random samples or two groups in a randomized experiment, the difference between the sample mean$(\bar{x}_1 - \bar{x}_2)$ is the best estimate of the difference between the populati$(\mu_1 - \mu_2)$.

♣ In other words, since the population means are unknown, the sample means must be used to make inferences.

♣ The inferences that are being made are based on the differences between the sample means. When the Independent condition is met, the standard deviation of the difference is:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Degrees of Freedom

♣ The shape of the t distribution is different for different sample sizes.

♣ Therefore, when making inferences about the difference between two population means, the size of the two samples must be taken into account.

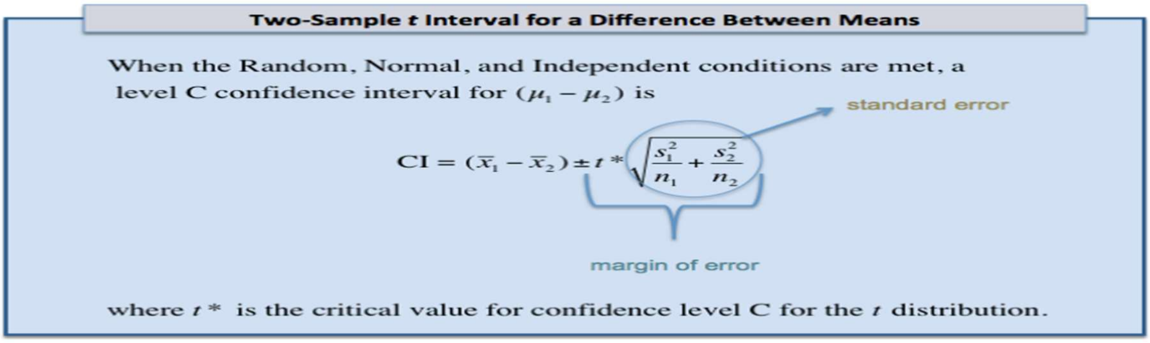
♣ This is because the t distribution is used to make these inferences.

degrees of freedom:

Choose the smaller of: df1=n1-1 and df2 =n2-1

Subtract 1 from each sample size.

Use the degrees of freedom that is the smallest (from the smaller sample size).



Two-Sample *t* Interval for a Difference Between Means

When the Random, Normal, and Independent conditions are met, a level C confidence interval for $(\mu_1 - \mu_2)$ is

$$\mathrm{CI} = (\bar{x}_1 - \bar{x}_2) \pm \underbrace{t^* \overbrace{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}^{\text{standard error}}}_{\text{margin of error}}$$

where t^* is the critical value for confidence level C for the t distribution.

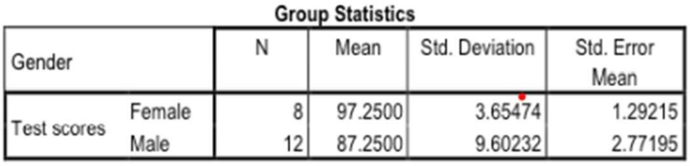
Question: *In the population, what is the difference between the females' average score and the males' average score on the test?*

Sample 1: Females:

n_1=8 x_1=97.25 s_1=3.65

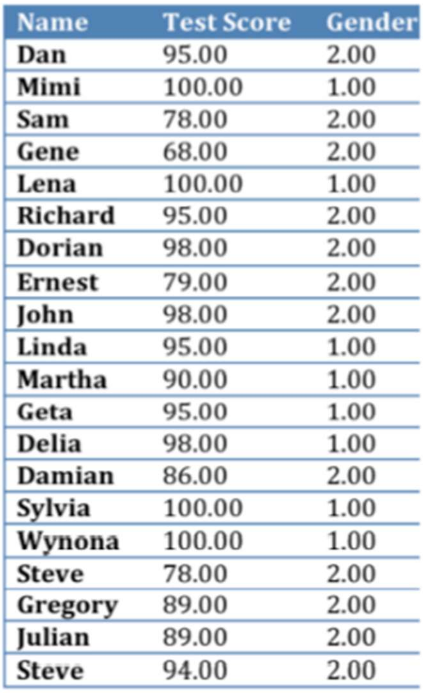
Sample 2: Males:

n_2=12 x_2=87.25 s_2=9.60

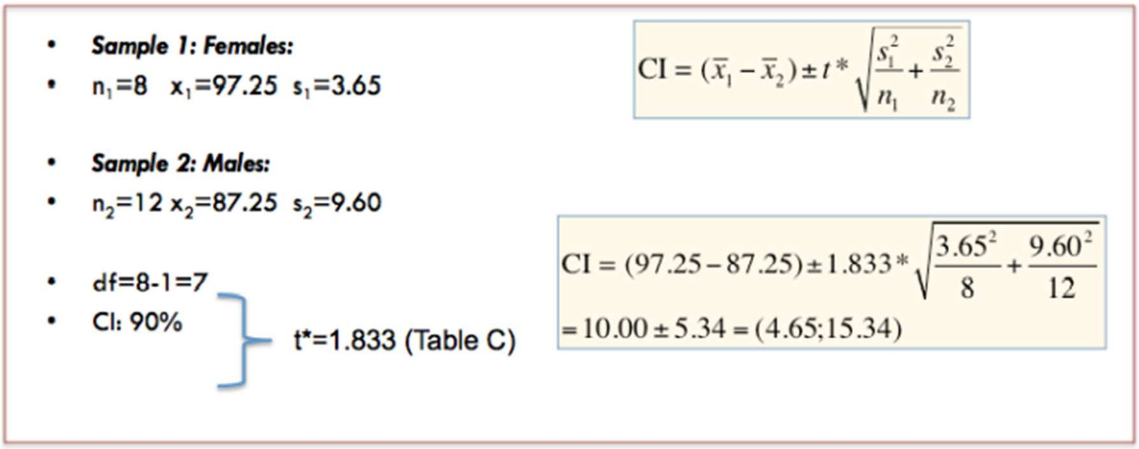


Group Statistics

Gender		N	Mean	Std. Deviation	Std. Error Mean
Test scores	Female	8	97.2500	3.65474	1.29215
	Male	12	87.2500	9.60232	2.77195



Name	Test Score	Gender
Dan	95.00	2.00
Mimi	100.00	1.00
Sam	78.00	2.00
Gene	68.00	2.00
Lena	100.00	1.00
Richard	95.00	2.00
Dorian	98.00	2.00
Ernest	79.00	2.00
John	98.00	2.00
Linda	95.00	1.00
Martha	90.00	1.00
Geta	95.00	1.00
Delia	98.00	1.00
Damian	86.00	2.00
Sylvia	100.00	1.00
Wynona	100.00	1.00
Steve	78.00	2.00
Gregory	89.00	2.00
Julian	89.00	2.00
Steve	94.00	2.00



- ***Sample 1: Females:***
- n_1=8 x_1=97.25 s_1=3.65
- ***Sample 2: Males:***
- n_2=12 x_2=87.25 s_2=9.60
- df=8-1=7
- CI: 90%

$$\left.\begin{matrix}\text{df}=8-1=7\\ \text{CI: }90\%\end{matrix}\right\} t^*=1.833\text{ (Table C)}$$

$$\text{CI} = (\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$\text{CI} = (97.25 - 87.25) \pm 1.833^* \sqrt{\frac{3.65^2}{8} + \frac{9.60^2}{12}}$$

$$= 10.00 \pm 5.34 = (4.65;15.34)$$

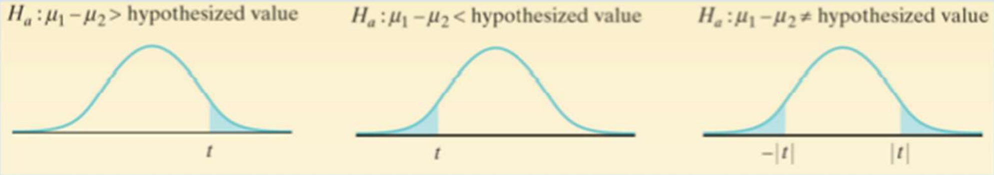
Interpretation: *It can be said with 90% confidence that difference between the females' average test score and the males' average test score ranges between 4.65 and 15.34.*

Two-Sample *t* Test for the Difference Between Two Means

Suppose the Random, Normal, and Independent conditions are met. To test the hypothesis $H_0 : \mu_1 - \mu_2 =$ hypothesized value, compute the t statistic

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

Find the P-value by calculating the probabilty of getting a t statistic this large or larger in the direction specified by the alternative hypothesis H_a. Use the t distribution with degrees of freedom approximated by technology or the smaller of $n_1 - 1$ and $n_2 - 1$.


$H_a : \mu_1 - \mu_2 >$ hypothesized value

$H_a : \mu_1 - \mu_2 <$ hypothesized value

$H_a : \mu_1 - \mu_2 \neq$ hypothesized value

- ***Sample 1: Females:***
- n_1=8 x_1=97.25 s_1=3.65

- ***Sample 2: Males:***
- n_2=12 x_2=87.25 s_2=9.60

- df=8-1=7
- α =0.05 (two-sided)

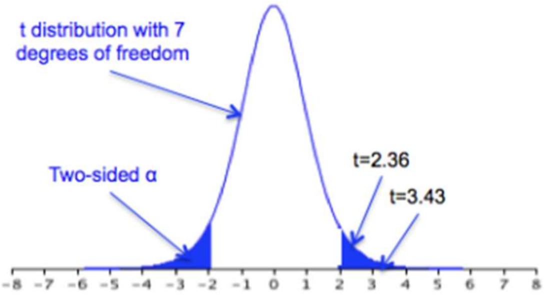
t_{cv}=2.365 (Table C)

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

$$t = \frac{(97.25 - 87.25)}{\sqrt{\frac{3.65^2}{8} + \frac{9.6^2}{12}}} = 3.43$$

$t > t_{cv}$ => reject Ho

Interpretation: *Data from the sample shows that, if the null hypothesis is true, the test statistic $t_{(7)}$=3.43 is greater than the critical value $t_{cv(7)}$=2.365, when the two-tailed α=0.05. Therefore, the null hypothesis was rejected.*



ANOVA:

ANOVA has some underlying assumptions which should be in place in order to make the results of calculations completely trustworthy.

They include:

(i) Subjects are chosen via a simple random sample.

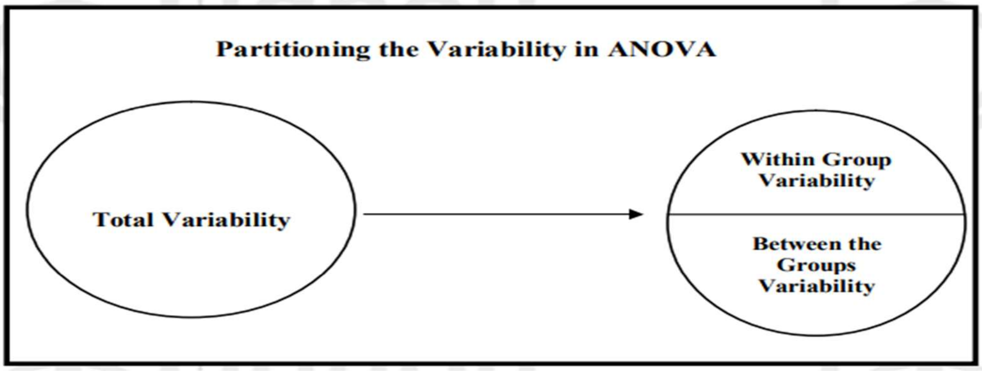
(ii) Within each group/population, the response variable is normally distributed.

(iii) While the population means may be different from one group to the next, the population standard deviation is the same for all groups.

Analysis of Variance (ANOVA) is a statistical method used to test differences between two or more means. It may seem odd that the technique is called “Analysis of Variance” rather than “Analysis of Means.”

An ANOVA conducted on a design in which there is only one factor is called a one-way ANOVA. If an experiment has two factors, then the ANOVA is called a two-way ANOVA. For example, suppose an experiment on the effects of age and gender on reading speed were conducted using three age groups (8 years, 10 years, and 12 years) and the two genders (male and female). The factors would be age and gender. Age would have three levels and gender would have two levels.

Analysis of Variance (ANOVA) is "Separation of variance ascribable to one group of causes from the variance ascribable to other group". So, by this technique, the total variation present in the data are divided into two components of variation one is due to assignable causes (between the groups variability) or other is variation due to chance causes (within group variability).


Partitioning the Variability in ANOVA
Total Variability
Within Group Variability
Between the Groups Variability

The analysis of variance technique solves the problems of estimating and testing to determine, whether to infer the existence of true difference among "treatment" means, among variety means and under certain conditions among other means with respect to the problem of estimation.

Analysis of variance technique can be classified as follows:

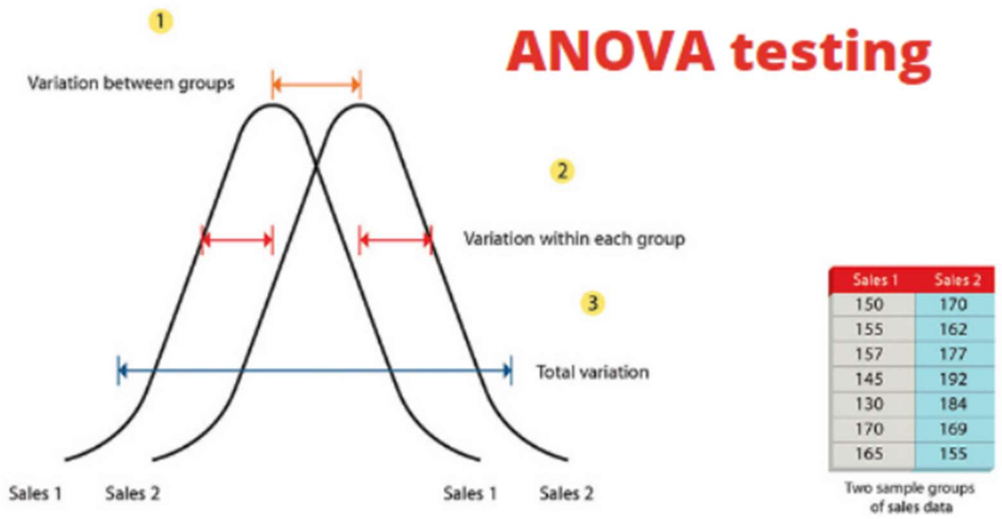
1. Parametric ANOVA.

2. Non Parametric ANOVA.

Parametric ANOVA can be classified as simply ANOVA if only one response variable is considered. If more than one response variables are under consideration than it is called multivariate analysis of variance (MANOVA).

If we consider, only one independent variable which affects the response / dependent variable then it is called One-way ANOVA. If the independent variables / explanatory variables are more than one i.e. n (say) then it is called n-way ANOVA. If n is equal to two than the ANOVA is called Two way classified ANOVA.

ANOVA is a type of hypothesis testing which is used to find out the experimental results by analyzing the variance of the different survey groups. It is usually used for deciding the result of the dataset.
Analysis of variance(ANOVA) is a statistical method to find out if the means of two or more groups are significantly different from each other. It checks the impact of one or more factors by comparing the means of different samples.
When we have two samples/groups we use a t-test to find out the mean between those samples but it is not that much reliable for more than two samples, therefore, we use ANOVA.



Sales 1	Sales 2
150	170
155	162
157	177
145	192
130	184
170	169
165	155

Two sample groups of sales data

Why do we use ANOVA testing?
In machine learning, the biggest problem is selecting the best features or attributes for training the model. We only require those features that are highly dependent on the response variable so that our model can able to predict the actual outcome after training the model. ANOVA is used to figure out the result when we have a continuous response variable and the target feature is categorical.

For example, we set up an experience of three groups of people, the very first group gets water drinks, second get some sugary juice and the third one like to take coffee or tea. Now, we need to test everyone's reaction time and want to know if there is any difference between the groups or not.

29	28	25
30	29	28
31	27	29
31	30	27
29	29	29

The null hypothesis tells that all the three groups have the same reaction time, we have three groups here to experiment and find out the result so we need to apply the ANOVA testing in case of two groups we could use the t-test when we experiment we would notice that the result won't be same.

The total variance of all these scores is made up of two parts:

1. The variance within the groups: As people have different reaction time in each group.
2. The variance between the groups: As the drinks are different which people prefer.

Example one:

10	11	12
12	14	13
18	19	17
24	23	25
36	38	37

As we can see here, there is a lot of variation in each sample/group, some of them are faster and some of them are slower but the groups are quite to one another, there is not much variation between the groups. So we can say that people are making a difference but not the type of drinks, in this case, we need to accept the null hypothesis we can’t reject that as the type of drink doesn’t put any effect on reaction time.

Example two:

29	17	10
29	18	11
30	19	12
31	19	13
31	20	13

Here we can see that there is not much difference within the groups but there is a lot of f=deifference between the groups. The people's reaction time doesn't make any effect on the groups, so here we will reject the null hypothesis.

In the example, we have seen a term hypothesis, what is the Hypothesis? ANOVA uses many terminologies with it.

Mean:

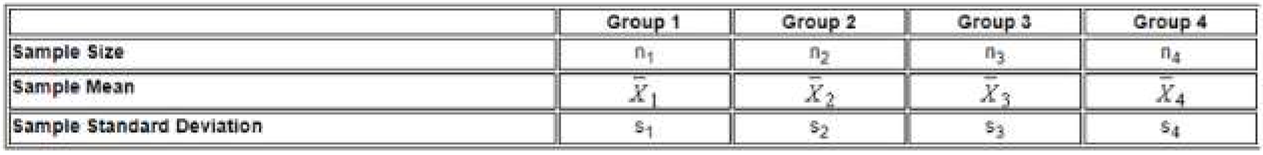
There are two types of mean that we used in ANOVA

1. Mean of each sample
2. Grand mean that is the mean of all the observation combined.

Hypothesis testing:

Hypothesis testing is statistical testing that is used to analyze the assumptions regarding the population parameters. There are two types of hypothesis in the hypothesis testing

1. Null hypothesis
2. Alternate hypothesis.



	Group 1	Group 2	Group 3	Group 4
Sample Size	n_1	n_2	n_3	n_4
Sample Mean	$\overline{X}_1$	$\overline{X}_2$	$\overline{X}_3$	$\overline{X}_4$
Sample Standard Deviation	s_1	s_2	s_3	s_4

Hypothesis in ANOVA is

- H0: μ1 = μ2 = μ3 …
- H1: Means are not all equal.

where k = the number of independent comparison groups.

Types of ANOVA

One way ANOVA:

The one-way ANOVA is used to find out the statistically significant difference between the mean of more than two independent groups.

More specifically it is used to test the null hypothesis.

In one-way ANOVA μ = group means and k is a number of groups, if one-way ANOVA returns the significant result, in this case, we accept the alternative hypothesis, this means that the mean of two groups is not equal.

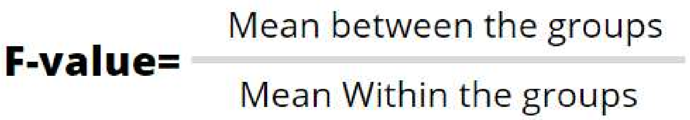
Two-way ANOVA:

A two-way is used to determine the effect of two nominal predictor features on a continuous outcome feature. It tests the effect of two independent variables on the expected outcome with the outcome itself.

F-value for ANOVA:

The F-value os ANOVA is a tool to help you to determine that, Is the variance between the means of two samples significantly different or not. The ratio of the between the groups and within the groups. It also helps us to find out the p-Value. The P-value is the probability of getting the result at least at the point where the null hypothesis should be true.

The formula for f-value:



$$\textbf{F-value=} \frac{\text{Mean between the groups}}{\text{Mean Within the groups}}$$

Regression:

Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable (s) (predictor). This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables. For example, relationship between rash driving and number of road accidents by a driver is best studied through regression.

There are multiple benefits of using regression analysis. They are as follows:

1. It indicates the significant relationships between dependent variable and independent variable.
2. It indicates the strength of impact of multiple independent variables on a dependent variable.

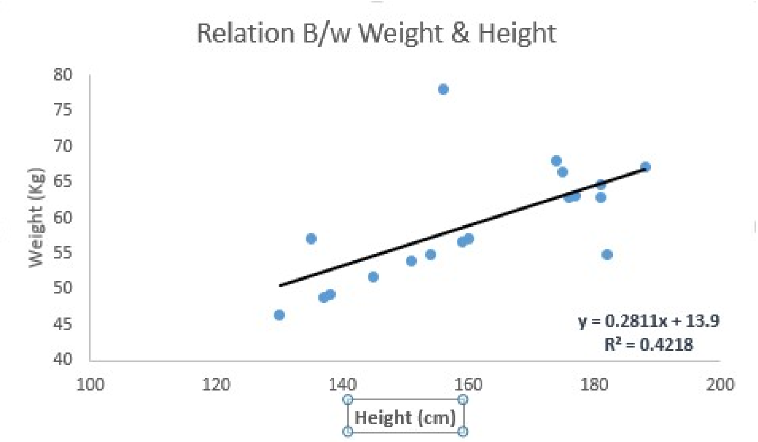
Regression analysis also allows us to compare the effects of variables measured on different scales, such as the effect of price changes and the number of promotional activities. These benefits help market researchers / data analysts / data scientists to eliminate and evaluate the best set of variables to be used for building predictive models.

Linear Regression

It is one of the most widely known modeling technique. Linear regression is usually among the first few topics which people pick while learning predictive modeling. In this technique, the dependent variable is continuous, independent variable(s) can be continuous or discrete, and nature of regression line is linear.

Linear Regression establishes a relationship between dependent variable (Y) and one or more independent variables (X) using a best fit straight line (also known as regression line).

It is represented by an equation Y=a+b*X + e, where a is intercept, b is slope of the line and e is error term. This equation can be used to predict the value of target variable based on given predictor variable(s).

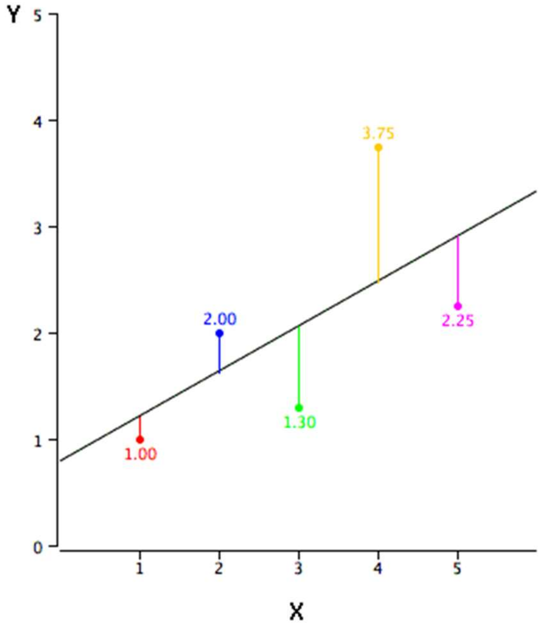


The difference between simple linear regression and multiple linear regression is that, multiple linear regression has (>1) independent variables, whereas simple linear regression has only 1 independent variable. Now, the question is "How do we obtain best fit line?".

How to obtain best fit line (Value of a and b)?

This task can be easily accomplished by Least Square Method. It is the most common method used for fitting a regression line. It calculates the best-fit line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line. Because the deviations are first squared, when added, there is no cancelling out between positive and negative values.

$$\underset{w}{min} \, ||Xw - y||_2{}^2$$



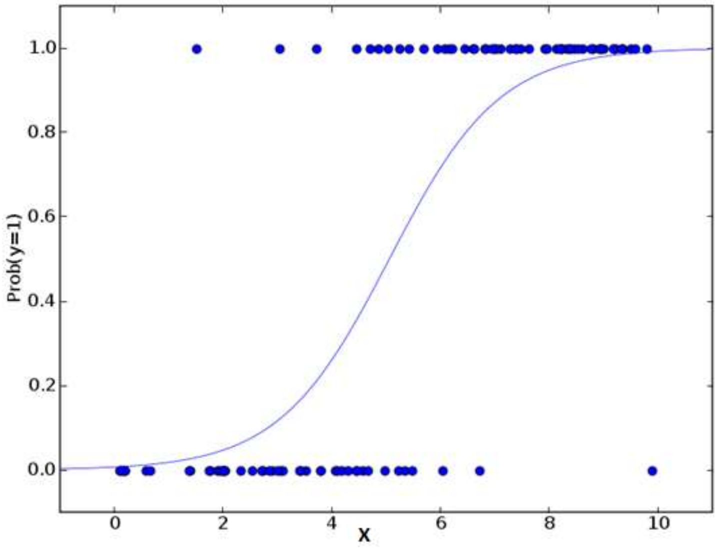
Logistic Regression

Logistic regression is used to find the probability of event=Success and event=Failure. We should use logistic regression when the dependent variable is binary (0/ 1, True/ False, Yes/ No) in nature. Here the value of Y ranges from 0 to 1 and it can represented by following equation.

odds= p/ (1-p) = probability of event occurrence / probability of not event occurrence
ln(odds) = ln(p/(1-p))
logit(p) = ln(p/(1-p)) = b0+b1X1+b2X2+b3X3....+bkXk

Above, p is the probability of presence of the characteristic of interest. A question that you should ask here is “why have we used log in the equation?”.

Since we are working here with a binomial distribution (dependent variable), we need to choose a link function which is best suited for this distribution. And, it is logit function. In the equation above, the parameters are chosen to maximize the likelihood of observing the sample values rather than minimizing the sum of squared errors (like in ordinary regression).



Important Points:

- Logistic regression is widely used for classification problems
- Logistic regression doesn’t require linear relationship between dependent and independent variables. It can handle various types of relationships because it applies a non-linear log transformation to the predicted odds ratio
- To avoid over fitting and under fitting, we should include all significant variables. A good approach to ensure this practice is to use a step wise method to estimate the logistic regression
- It requires large sample sizes because maximum likelihood estimates are less powerful at low sample sizes than ordinary least square
- The independent variables should not be correlated with each other i.e. no multi collinearity. However, we have the options to include interaction effects of categorical variables in the analysis and in the model.
- If the values of dependent variable is ordinal, then it is called as Ordinal logistic regression

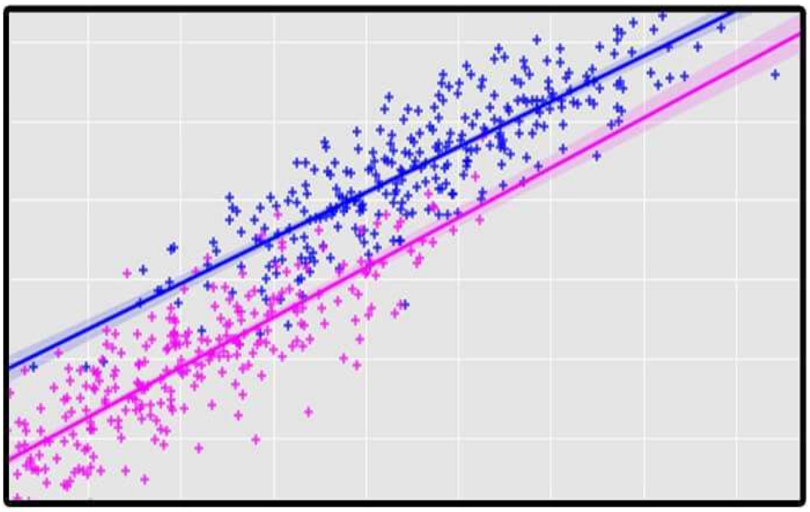
Multiple Linear Regression

Simple linear regression allows a data scientist or data analyst to make predictions about only one variable by training the model and predicting another variable. In a similar way, a multiple regression model extends to several more than one variable.

Simple linear regression uses the following linear function to predict the value of a target variable y, with independent variable x?.

$$y = b_0 + b_1x_1$$

Multiple Linear Regression



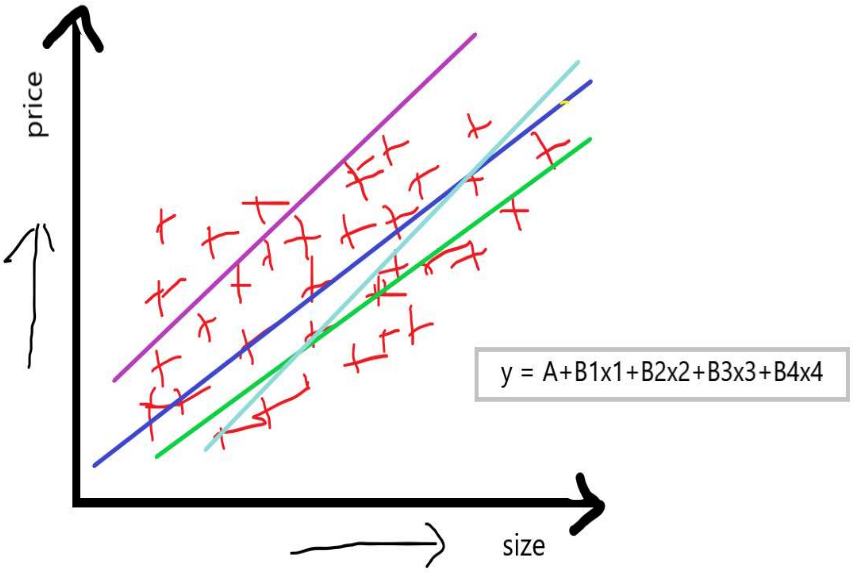
To minimize the square error we obtain the parameters b? and b? that best fits the data after fitting the linear equation to observed data.

Multiple linear regression (MLR/multiple regression) is a statistical technique. It can use several variables to predict the outcome of a different variable. The goal of multiple regression is to model the linear relationship between your independent variables and your dependent variable. It looks at how multiple independent variables are related to a dependent variable.

Multiple linear regression is what you can use when you have a bunch of different independent variables!

Multiple regression analysis has three main uses.

- You can look at the strength of the effect of the independent variables on the dependent variable.
- You can use it to ask how much the dependent variable will change if the independent variables are changed.
- You can also use it to predict trends and future values.
  - Multiple Linear Regression is basically indicating that we will be having many features Such as f1, f2, f3, f4, and our output feature f5. If we take the same example as above we discussed, suppose:
  - f1 is the size of the house.
  - f2 is bad rooms in the house.
  - f3 is the locality of the house.
  - f4 is the condition of the house and,
  - f5 is our output feature which is the price of the house.
  - Now, you can see that multiple independent features also make a huge impact on the price of the house, price can vary from feature to feature. When we are discussing multiple linear regression then the equation of simple linear regression y=A+Bx is converted to something like:
  - equation: $y = A+B_1x_1+B_2x_2+B_3x_3+B_4x_4$
  - “If we have one dependent feature and multiple independent features then basically call it a multiple linear regression.”



-
- Now, our aim to using the multiple linear regression is that we have to compute A which is an intercept, and B_1 B_2 B_3 B_4 which are the slops or coefficient concerning this independent feature, that basically indicates that if we increase the value of x_1 by 1 unit then B1 says that how much value it will affect int he price of the house, and this was similar concerning others B_2 B_3 B_4

Poisson Distribution:

Poisson Distribution is the discrete probability of count of events which occur randomly in a given interval of time. It is a limiting form of the binomial distribution in which n becomes very large and p is very very small (meaning the number of trials is very large while the probability of occurrences of outcome under observation is small.

Definition of Poisson Distribution

X =the number of events in a given interval

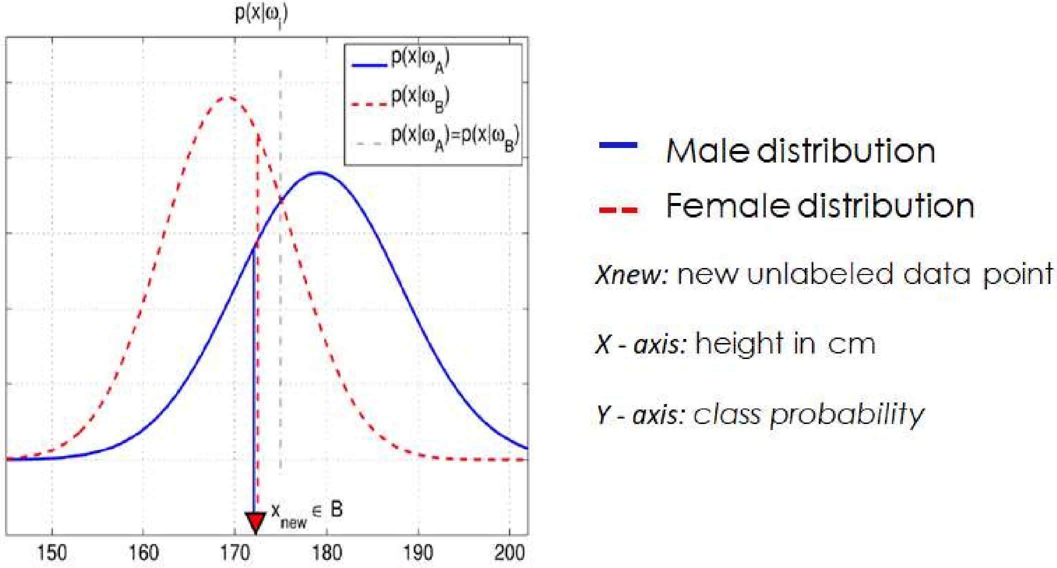
λ = mean number of events per interval

The probability of observing *x* events in a given interval is given by

$$\boxed{\mathbf{P(X = x)} = \mathbf{e}^{-\lambda}\frac{\lambda^{x}}{x!} \qquad x = 0, 1, 2, 3, 4, \ldots}$$

Maximum Likelihood:

The goal of maximum likelihood is to fit an optimal statistical distribution to some data. This makes the data easier to work with, makes it more general, allows us to see if new data follows the same distribution as the previous data, and lastly, it allows us to classify unlabelled data points.
imagine a binary classification problem between male and female individuals using height. Once we have calculated the probability distribution of men and woman heights, and we get a new data point (as height with no label), we can assign it to the most likely class, seeing which distribution reports the highest probability of the two.



Graphical representation of this binary classification problem

In the previous image this new data point (*xnew,* which corresponds to a height of 172 cm) is classified as female, as for that specific height value the female height distribution yields a higher probability than the male one.

how do we actually calculate these probability distributions?

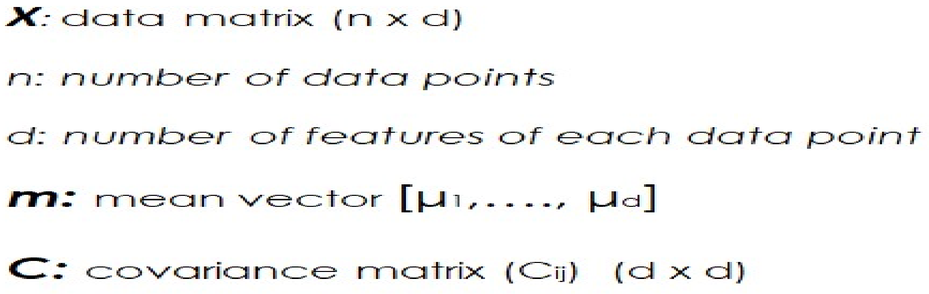
Calculating the distributions: estimating a parametric density function

As usual in Machine Learning, the first thing we need to start calculating a distribution is something to learn from: our precious data. We will denote our data vector of size n, as *X*. In this vector each of the rows is a data point with d features, therefore our data vector *X* is actually a vector of vectors: a matrix of size n x d; n data points with d features each.

Once we have collected the data which we want to calculate a distribution from, we need to start guessing.

there is a distribution which is most likely to fit best: Gaussian for features like temperature or height, exponential for features regarding time, like length of phone calls or the life of bacterial populations, or Poisson for features like the number of houses sold in a specific period of time.

Once this is done we calculate the specific parameters of the chosen distribution that best fit our data. For a normal distribution this would be the mean and the variance. As the gaussian or normal distribution is probably the easiest one to explain and understand, we will continue this post assuming we have chosen a gaussian density function to represent our data.



X: data matrix (n x d)

n: number of data points

d: number of features of each data point

m: mean vector $[\mu_1, \ldots, \mu_d]$

C: covariance matrix (C_{ij}) (d x d)

Data and parameters for our gaussian distribution

In this case, the number of parameters that we need to calculate is *d* means (one for each feature) and *d(d+1)/2* variances, as the Covariance matrix is a symmetrical *dxd* matrix.

$$d + \frac{d(d+1)}{2}$$

Total parameters we need to calculate for a normal distribution depending on the number of features

Lets call the overall set of parameters for the distribution θ. In our case this includes the mean and the variance for each feature. What we want to do now is obtain the parameter set θ that maximises the joint density function of the data vector; the so called *Likelihood function L(θ).* This likelihood function can also be expressed as *P(X|θ)*, which can be read as the conditional probability of X given the parameter set θ.

$$L(\theta) = p(X \mid \theta) = p(X(1), X(2), \ldots X(n) \mid \theta)$$

Likelihood function

In this notation X is the data matrix, and X(1) up to X(n) are each of the data points, and θ is the given parameter set for the distribution. Again, as the goal of Maximum Likelihood is to chose the parameter values so that the observed data is as likely as possible, we arrive at an optimisation problem dependent on θ.

To obtain this optimal parameter set, we take derivatives with respect to θ in the likelihood function and search for the maximum: this maximum represents the values of the parameters that make observing the available data as likely as possible.

$$\frac{\partial}{\partial \Theta} p(X | \Theta) = 0$$

Taking derivatives with respect to θ

Now, if the data points of X are independent of each other, the likelihood function can be expressed as the product of the individual probabilities of each data point given the parameter set:

$$L(\Theta) = p(X \mid \Theta) = \prod p(X(j) \mid \Theta)$$

Likelihood function if the data points are independent of each other

Taking the derivatives with respect to this equation for each parameter (mean, variance,etc…) keeping the others constant, gives us the relationship between the value of the data points, the number of data points, and each parameter.
Lets look at an example of how this is done using the normal distribution, and an easy male height dataset.

A deeper look into the maths of Maximum Likelihood using a normal distribution

Lets see an example of how to use Maximum Likelihood to fit a normal distribution to a set of data points with only one feature: height in centimetres. As we mentioned earlier, there are to parameters that we have to calculate: the mean and the variance.
For this, we have to know the density function for the normal distribution:

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Density function for the normal distribution.

Once we know this, we can calculate the likelihood function for each data point. For the first data point it would be:

$$p(X(1) \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}[x(1)-\mu]^2}$$

Likelihood equation for the first data point

For the whole data set, considering our data points as independent and we can therefore calculate the likelihood function as the product of the likelihoods of the individual points, it would be:

$$p(\boldsymbol{X} \mid \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}}\, e^{-\frac{1}{2\sigma^2}\Sigma[x(i)-\mu]^{2}}$$

Likelihood equation for the whole dataset

We can take this function and express it in a logarithmic way, which facilitates posterior calculations and yields exactly the same results.

$$\ln p(\boldsymbol{X} \mid \mu, \sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{n}[x(j)-\mu]^2$$

Same equation expressed in a logarithmic way

Finally, we set the derivative of the likelihood function with regards to the mean to zero, reaching an expression where we obtain the value of this first parameter:

$$\frac{\partial}{\partial\mu}\ln p(\boldsymbol{X} \mid \mu, \sigma^2) = \frac{1}{\sigma^2}\sum_{j=1}^{n}[x(j)-\mu] = 0$$

$$\mu = \hat{\mu}_{ML} = \frac{1}{n}\sum_{j=1}^{n}x(j)$$

Derivative of the likelihood function for the mean, and Maximum Likelihood value for this parameter

Surprise! The maximum likelihood estimate for the mean of the normal distribution is just what we would intuitively expect: the sum of the value of each data point divided by the number of data points.

Now that we have calculated the estimate for the mean, it is time to do the same for the other relevant parameter: the variance. For this, just like before, we take derivatives in the likelihood function with the goal of finding the value of the variance that maximises the likelihood of the observed data.

$$\frac{\partial}{\partial \sigma^2} \ln p(\boldsymbol{X} \mid \mu, \sigma^2) = 0$$

$$\hat{\sigma}^2_{ML} = \frac{1}{n} \sum_{j=1}^{n} [x(j) - \hat{\mu}_{ML}]^2$$

Maximum likelihood estimate for the variance

Maximum Likelihood estimate for male heights: a numeric example

Lets the very simple example we have mentioned earlier: we have a data set of male heights in a certain area, and we want to find an optimal distribution to it using Maximum Likelihood.

If we remember right, the first step for this (after collecting and understanding the data) is to choose the shape of the density function that we want to estimate. In our case, for height, we will use a Gaussian distribution, which we also saw in the general reasoning behind the maths of ML. Lets retake a look at the formula that defines such distribution:

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Density function for the normal distribution.

Also, lets recover the likelihood function for just one point of the data set.

$$p(X(1) \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}[x(1)-\mu]^2}$$

Likelihood equation for the first data point

Imagine our data vector X, in our case is the following:

[176,172,180,171,168,175,182,167,184,170]

Data vector of male heights

We have 10 data points (n = 10) and one feature for each data point (d=1). If in the formula shown above for each of the data points we substitute their actual values we get something like:

$$p(X(1) \mid \mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2\sigma^2}[176-\mu]^2}$$

$$p(X(2) \mid \mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2\sigma^2}[172-\mu]^2}$$

Likelihood of the first two data points

If in these formulas we choose a specific mean and variance value, we would obtain the likelihood of observing each of the height values (176 and 172 cm in our case) with those specific mean and variances. For example, if we pick a mean of 180 cm with a variance of 4 cm, we would get the following likelihoods for the two points shown above:

$$p(X(1) \mid 180,4) = \frac{1}{\sqrt{2\pi 4}}\, e^{-\frac{1}{8}[176-180]^2} = 0.03$$

$$p(X(2) \mid 180,4) = \frac{1}{\sqrt{2\pi 4}}\, e^{-\frac{1}{8}[172-180]^2} = 0.00007$$

Calculations of likelihood of observing points of 176 cm and 172 cm of height on a normal distribution with a mean of 180 cm and a variance of 4 cm

After this quick note, if we continue with the procedure to obtain the maximum likelihood estimate that best fits out data set, we would have to first calculate the mean. For our case it is very simple: we just sum up the values of the data points and divide this sum by the number of data points.

$$\mu = \hat{\mu}_{ML} = \frac{1}{n}\sum_{j=1}^{n} x(j) = 174{,}5 \text{ cm}$$

Maximum likelihood estimate for the mean of our height data set

If we do the same for the variance, calculating the squared sum of the value of each data point minus the mean and dividing it by the total number of points we get:

$$\hat{\sigma}^2_{ML} = \frac{1}{n}\sum_{j=1}^{n}[x(j) - \hat{\mu}_{ML}]^2 = 30{,}45$$

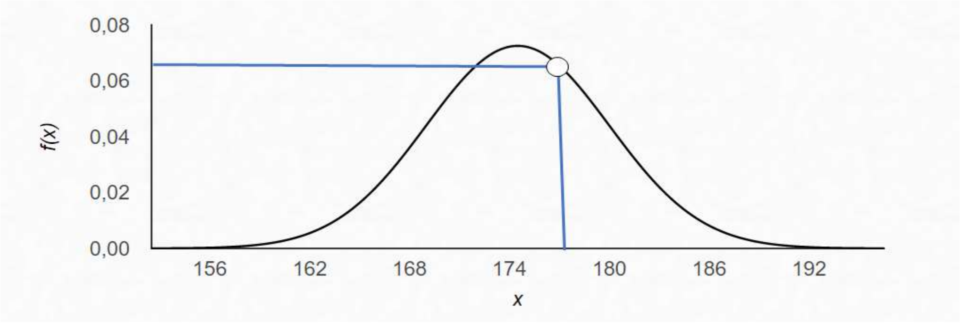
$$\hat{\sigma}_{ML} = 5{,}5 \text{ cm}$$

Variance and Standard deviation estimates for our height data set

That is it! Now we have calculated the mean and the variance, we have all the parameters we need to model our distribution. Now, when we get a new data point, for example, one with a height of 177 cm, we can see the likelihood of that point belonging to our data set:

$$p(177 \mid 174.5{,}30.45) = \frac{1}{\sqrt{2\pi 30.45}}\, e^{-\frac{1}{60.9}[176-180]^2} = 0.065$$

Likelihood of the new data point belonging to our data set



Representation of the obtained normal distribution and the likelihood of the new data point

Now, if we had another data set, with female heights for example, and we did the same procedure, we would have two height distributions: one for male and one for females.

With this, we could solve a binary classification problem of male and female heights using both distributions: when we get a new unlabelled height data point, we calculate the probability of that new data point belonging to both distributions, and assign it to the class (male or female) for which the distribution yields the highest probability.