greatlearning

Hypothesis Testing

1. One of the very common use of hypothesis in machine learning exercises will be to assess whether an observation, for e.g. An apparent link between sale of beer and sale of baby diapers, is a statistical chance or is it real

2. Null Hypothesis says “No relationship” while alternate hypothesis “There is a relationship”

3. The data that shows the apparent relationship, would have certain characteristics such as central values, spread, shape of the curve, direction of the spread etc.

4. Statistical techniques are employed that assess the probability of collecting such data (with the observed characteristics) from real world data if there was no such relation ship in the real world

5. The probability is expressed in terms of P value. Usually a P of less than .05 (5%) is considered as evidence of relationship

6. P value is linked directly to the confidence level which by default is 95% i.e. the confidence in rejecting null hypothesis when it should be

7. There are various ways of conducting the hypothesis testing to get the P value. One of the ways (used for categorical outputs) is Chi-square test

Power of Statistical Test

Hypothesis Testing

Type 1 error: Reject H0 when it is true

1. you think the new procedure introduced as part of process improvement has increased productivity but in long run the productivity remains same as it was before the procedure was introduced
2. While preparing to model loan status, you find strong link between income and loan status in the sample data and believe it to be true in population while it is not so
3. Significance level (α) or Type 1 error rate: is the probability of making this type of error I (**P value**)
4. This **P value** is usually set to 0.05 as a standard. This translates to 5% chance of incorrectly rejecting H0

Type 2 error: Failing to reject H0 when it is false

1. You think the new procedure introduced as part of process improvement has not had any effect but if you had used it long enough you would have noticed the productivity actually increased compare to before the procedure was introduced
2. While preparing to model loan status, you do not find strong link between income and loan status in the sample data and believe there is no relation between them in population while a larger sample would have shown the relation
3. The value β is the probability of a type 2 error or type 2 error rate

Power: 1−β: probability of correctly rejecting H0 when it is fails

Hypothesis Testing

1. In data science modeling, when a decision about a query point (test record) has to be taken using the model, it is similar to statistical testing

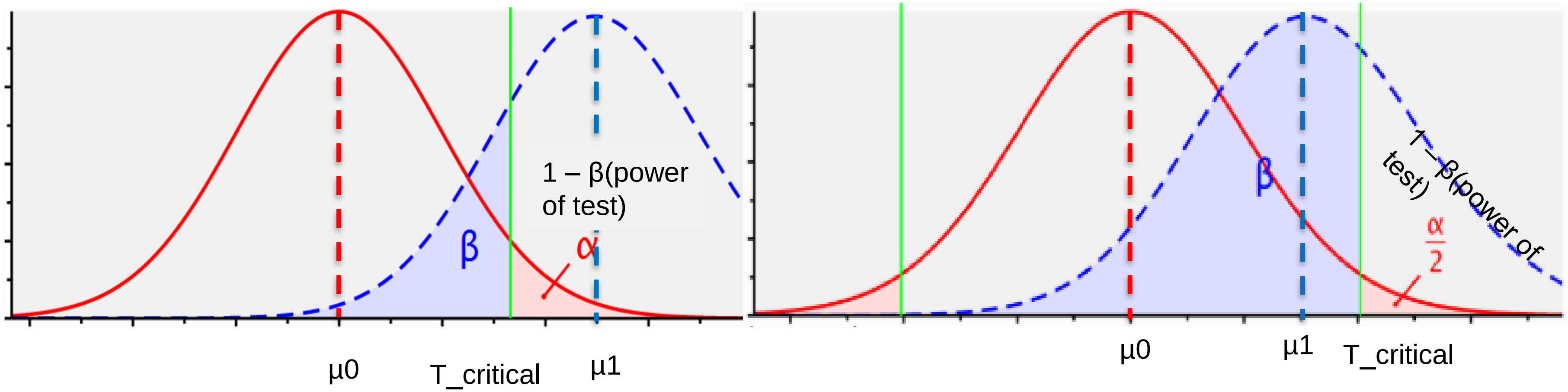
2. The choice is, either it belongs to the normal class or not. Claiming that it likely belongs to the normal class is like accepting H0 while claiming otherwise is rejecting H0

3. Thus we can represent the entire discussion about hypothesis and errors in a confusion matrix

Decision about H0

	Confusion Matrix	Predicted Positive	Predicted Negative
State of H0	Actual Positive H0 is false	True Positive (Correctly rejected H0) Prob = 1- β	False Negative (failed to reject H0) β error / Type II error
	Actual Negative H0 is True	False Positive (incorrectly rejected H0) α error / Type I error	True Negative (correctly accepted H0) Prob = 1 - α

Power of a hypothesis test



Power: The probability a test will reject H0 when it is supposed to. The estimated probability (power of test) is a function of sample size, variability, level of significance, and the difference between the null and alternative hypotheses.

Power of a hypothesis test

- The power of a statistical test is function of –
  a. Goes up as $|\mu_0 - \mu_1|$ **a.k.a. effect** (magnitude of the differences in means) increases
  b. Goes up as sample size *n* goes up
  c. Goes up as standard deviation of the sample goes down
  d. Goes up as α value increases (probability of committing type 1 error increases)

- To keep the chances of making a correct decision high –
  a. the probability of a Type I error (α, the level of significance of a hypothesis test) is kept low, usually 0.05 or less,
  b. the power of the test (1-β, the probability of rejecting H0 when H1 is true) is kept high, usually 0.8 or more.
  c. In order to achieve a desirable power for a fixed level of significance, the sample size will generally need to increase
  d. Increase in sample size will increasingly reflect the characteristics of the population and thus help in addressing both Type I and Type II error simultaneously

Power Analysis

Power analysis is the process of estimating one of the four parameters that the power of a test is dependent on, given values for three other parameters

It helps in design of experiment (sample size, power of test likely) and in analysis of the predictions of a model using hypothesis testing

Effect size, a parameter we need to know for power analysis

1. Measure of the minimum magnitude of the signal (difference between an observation and the Null distribution) that can be detected by the test. For e.g. difference between the sample mean and population mean
2. Effect size is calculated using a specific statistical measure, such as Pearson’s correlation coefficient for the relationship between variables or Cohen’s d for the difference between groups. It describes the difference in means in terms of the number of standard deviations

Power Analysis(Contd…)

Power analysis is the process of estimating one of the four parameters that the power of a test is dependent on, given values for three other parameters

It helps in design of experiment (sample size, power of test likely) and in analysis of the predictions of a model using hypothesis testing

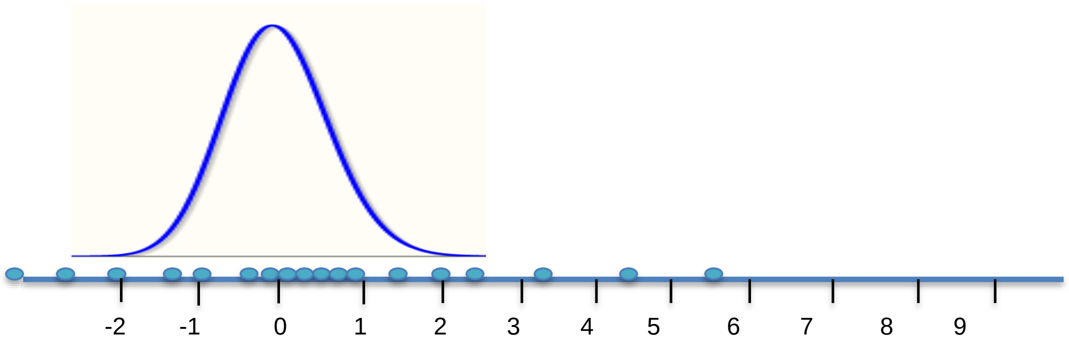
Effect size, a parameter we need to know for power analysis

1. Measure of the minimum magnitude of the signal (difference between an observation and the Null distribution) that can be detected by the test. For e.g. difference between the sample mean and population mean
2. Effect size is calculated using a specific statistical measure, such as Pearson’s correlation coefficient for the relationship between variables or Cohen’s d for the difference between groups. It describes the difference in means in terms of the number of standard deviations

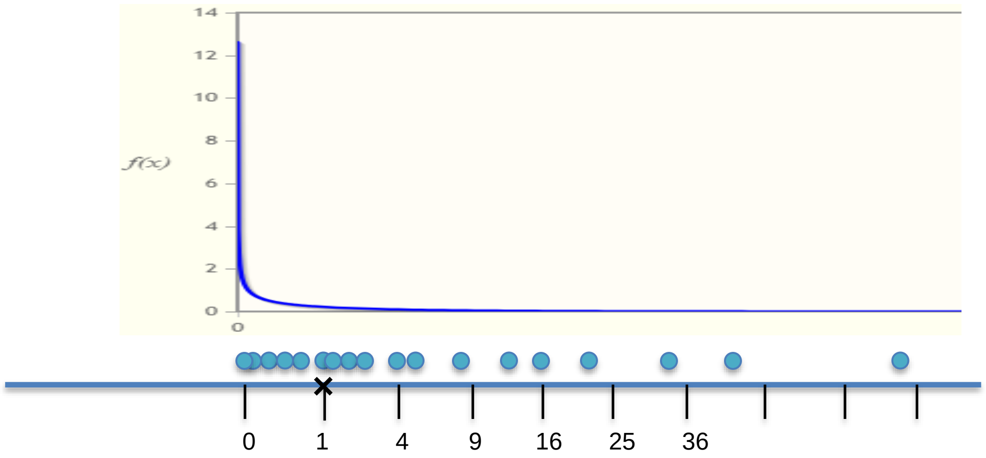
Chi-Square For Hypothesis Testing

1. The Chi-square distribution is the distribution of the sum of squared standard normal deviates
2. The degrees of freedom of the distribution is equal to the number of standard normal deviates being summed
3. Therefore, Chi-square with one degree of freedom, written as χ2(1), is simply the distribution of a single normal deviate squared
4. The mean of a Chi-square distribution is its degrees of freedom
5. Chi-square distributions are positively skewed, with the degree of skew decreasing with increasing degrees of freedom
6. As the degrees of freedom increases, the Chi-square distribution approaches a normal distribution
7. The Chi-square distribution is very important because many test statistics are approximately distributed as Chi-square.

Chi-Square For Hypothesis Testing



Chi-Square Distribution

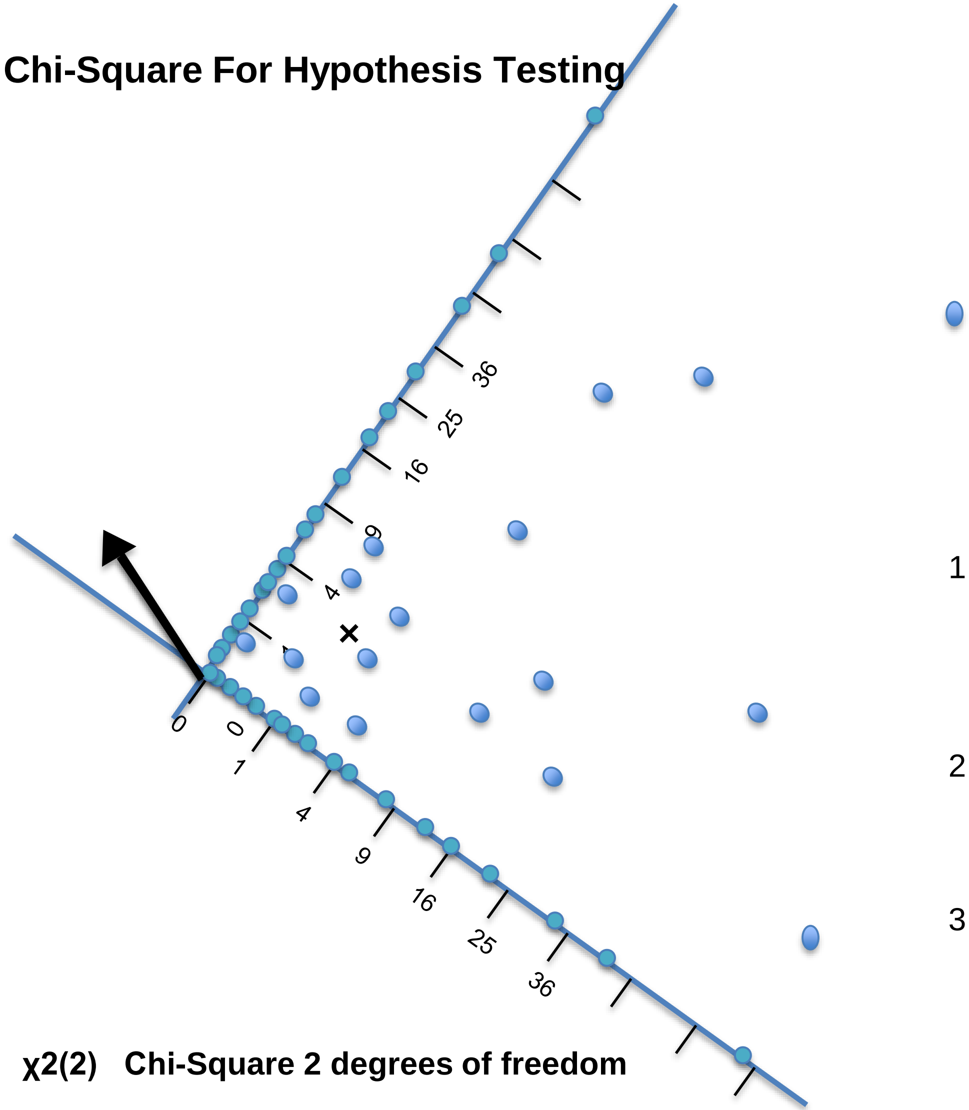


χ2(1) Chi-Square 1 degree of freedom

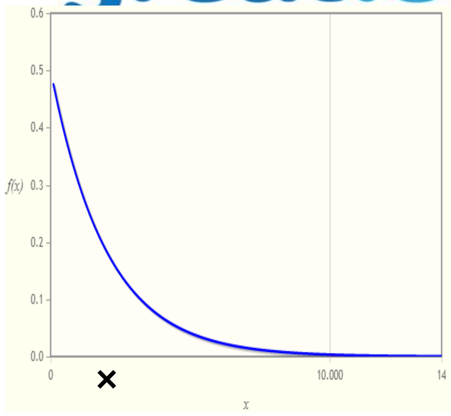
Ref: https://www.di-mgt.com.au/chisquare-calculator.html

1. Standard Normal Distribution –
   a. Centered on 0
   b. Standard Dev = Variance = 1
   c. The distribution is symmetric
2. When we study the distribution pattern of variance of the values we square the difference of value from mean which is 0
3. So, square the number X and plot the density function for X^2
4. When we do that, it is like shrinking the scale. i.e. pushing all data points towards 0
5. We also have the –ive points from SND become positive
6. Result of 3 and 5, data points get cluttered close to zero
7. Hence density plot becomes steep as we approach 0
8. This is chi-square density distribution in 1 dimension
9. The mean of this distribution is 1 shown by black cross

Chi-Square For Hypothesis Testing



χ2(2) Chi-Square 2 degrees of freedom



X^2 + Y^2

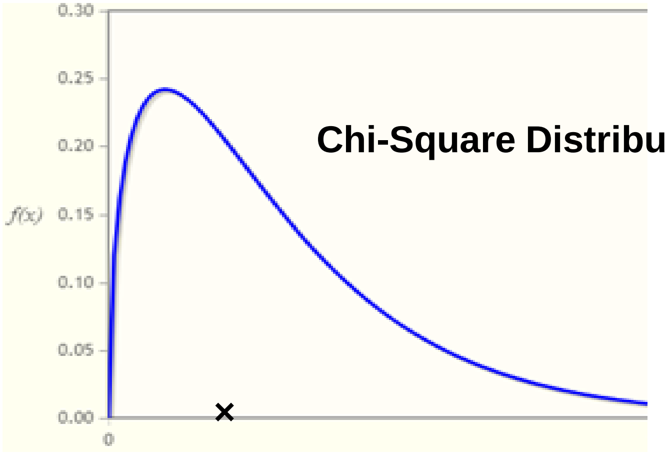
1. Chi-Square in two dimensions
   a. Centered on 0
   b. Standard Dev = Variance = 1
   c. The distribution is symmetric

2. Note carefully, the density plot is not in terms of individual variables but sum of square of the variables

3. It is like creating a new dimension from the two given dimensions

Ref: https://www.di-mgt.com.au/chisquare-calculator.html

Chi-Square For Hypothesis Testing

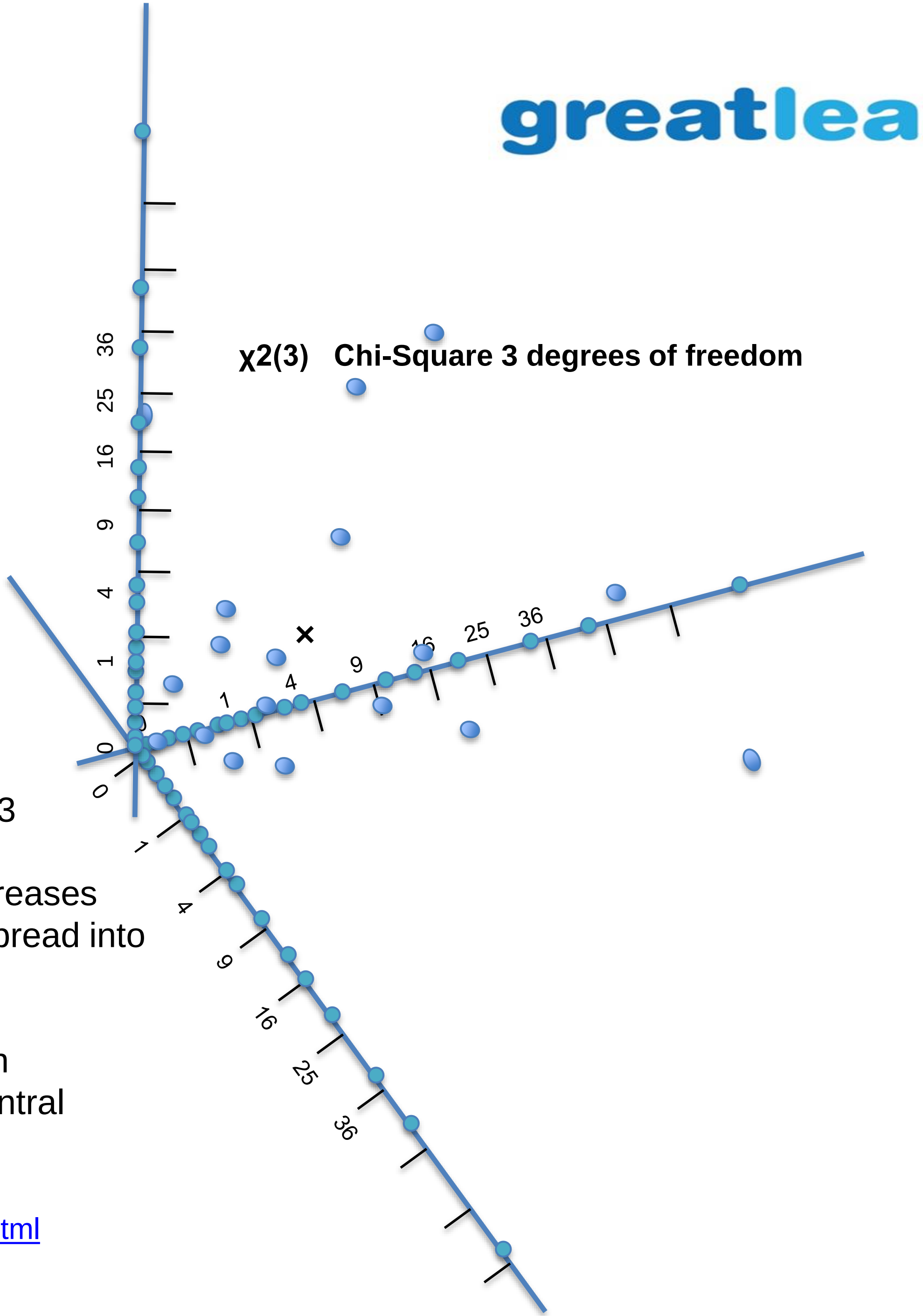


Chi-Square Distribution

X^2 + Y^2 + Z^2

1. Chi-Square in three dimensions has mean of 3
2. The density of distribution near the origin decreases with increase in the dimensions as the data spread into the space
3. For higher dimensions, the density distribution approaches normal distribution around the central value (the dark cross)

Ref: https://www.di-mgt.com.au/chisquare-calculator.html



χ2(3) Chi-Square 3 degrees of freedom

Chi-Square For Hypothesis Testing

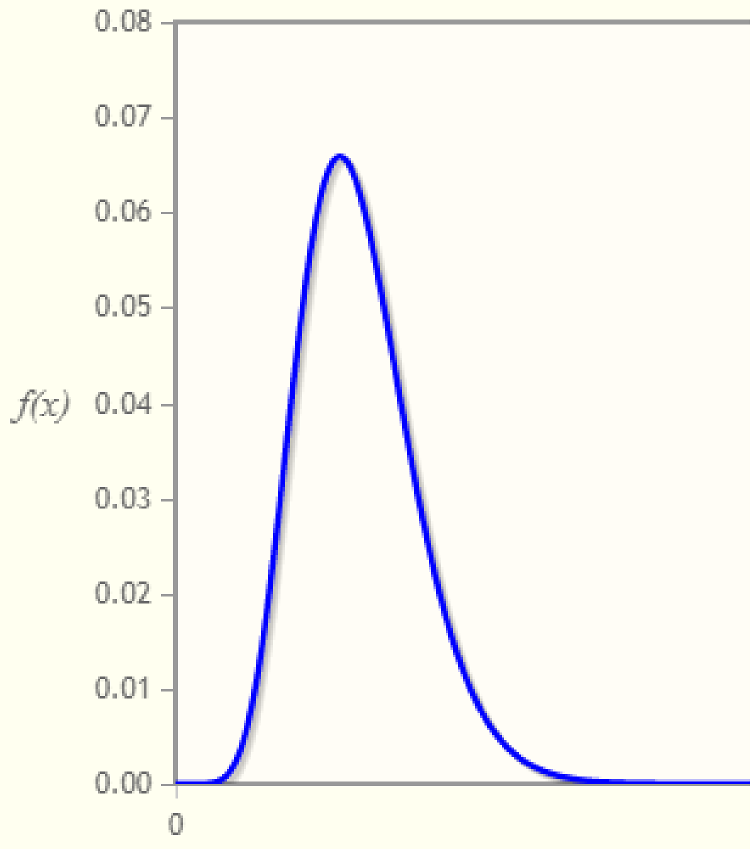
1. As the number of dimensions increase, the degree of freedom for data points to scatter into increases

2. As a result the density of distribution changes near the origins

3. The density peak shifts away and the distribution overall approaches a normal distribution

4. The mean of the distribution continues to be number of degrees of freedom

5. The variance is 2*DOF



Chi-Square Distribution

Ref: https://www.di-mgt.com.au/chisquare-calculator.html

Chi-Square For Hypothesis Testing

1. Chi-square test is one of the most common ways to examine relationships between two or more categorical variables
2. It involves calculating a metric called the Chi-square statistic which follows the Chi-square distribution
3. There are several types of Chi-square tests, we will cover the most common which is Pearson's Chi-square test
4. Further, there are three ways of using the Chi-square test, which are sometimes treated identically
   a. Chi-square test for independence – to test the hypothesis that two variables are independent of each other
   b. Chi-square test for equality of proportions – to test the hypothesis that the distribution of some variable is same in all populations. For e.g. proportion of defaulters: non-defaulters is same in both the genders
   c. Chi-square test for goodness of fit – to test hypothesis that the distribution of categorical variable within a population follows a specific pattern of proportions
5. We will focus on Chi-square test for independence. The other two tests are similar but the way null hypothesis is formed is different

Chi-Square For Hypothesis Testing

Example 1 –

We have data on smoking status and diagnosis with lung cancer from a random sample of adults. Both the variables i.e. smoking status and diagnosis are dichotomous (binary i.e. have values "yes" or "no")

	Lung Cancer - Yes	Lung Cancer - No	Total
Smoker - Yes	60	300	360
Smoker - No	10	390	400
Total	70	690	760 (G total)

From data it seems the two variables have a connection. 20% of the smokers (60 / 360) and only 2.5% of non-smokers (10/400) have been diagnosed with lung cancer. But looks can be misleading…

The Chi-squared test relies on the differences between the observed and expected frequencies in each category.

Let Eij, represent expected values of the two variables are independent of one another. Eij = ith (row total X jth column total) / grand total

Chi-Square For Hypothesis Testing

Example 1 (Contd…)–

	Lung Cancer - Yes	Lung Cancer - No	Total
Smoker - Yes	60	300	360
Smoker - No	10	390	400
Total	70	690	760 (G total)

1. E1,1 = expected value of (smoker = ‘yes’ and lung cancer = ‘yes’) = 360 * (70 / 760) = 33.16
2. E1,2 = expected value of (smoker = ‘yes’ and lung cancer = ‘no’) = 360 * (690/760) = 326.84
3. E2,1 = expected value of (smoker = ‘no’ and lung cancer = ‘yes’) = 400 * (70 / 760) = 36.84
4. E2,2 = expected value of (smoker = ‘no’ and lung cancer = ‘no’) = 400* (690 / 760)= 363.16

	Lung Cancer - Yes	Lung Cancer - No	Total
Smoker - Yes	60 (33.16)	300 (326.84)	360
Smoker - No	10 (36.84)	390 (363.16)	400
Total	70	690	760 (G total)

Expected values in brackets

Chi-Square For Hypothesis Testing

Example 1 (Contd…)–

	Lung Cancer - Yes	Lung Cancer - No	Total
Smoker - Yes	60 (33.16)	300 (326.84)	360
Smoker - No	10 (36.84)	390 (363.16)	400
Total	70	690	760 (G total)

Expected values in brackets

1. Hypothesis –
   1. Null (H0) – smoking and lung cancer diagnosis are independent
   2. Alternate (H1) – smoking and lung cancer diagnosis are not independent (Note: we are not stating H1 as “smoking leads to cancer”…. Such causations, unless established should not be stated in hypothesis definitions)
2. Chi-square test –

$$X^2 = \sum \frac{(O-E)^2}{E} = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

$$\text{df} = (r-1)(c-1)$$

	Lung Cancer - Yes	Lung Cancer - No
Smoker - Yes	(60 - 33.16)^2 / 33.16 = **21.7**	(300 - 326.84)^2 / 326.84 = 2.204 = **2.2 rounded**
Smoker - No	(10 - 36.84)^2 / 36.84 = 19.5544 = **19.6 approx**	(390 -363.16)^2 / 363.16 = 1.9836 = **2.0 approx**

= 45

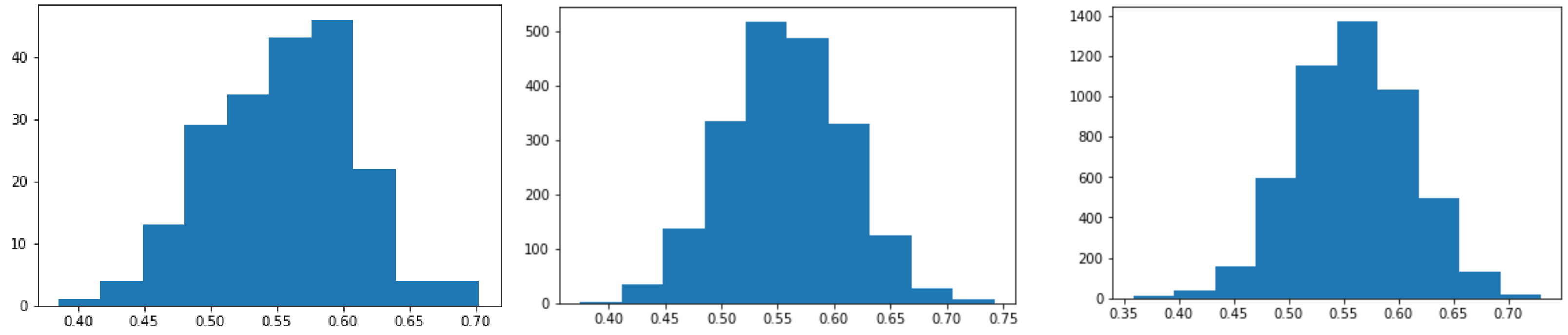
3. Degree of freedom is number of categories in a variable -1 for all the variables, often shown as number of rows – 1, number of cols -1 = (2-1) * (2 -) = 1
4. Having calculated the chi-square value and degrees of freedom, we consult a chi-square table to check whether the chi-square statistic of 45 exceeds the critical value for the Chi-square distribution. The critical value for alpha of .05 (95% confidence) is 3.84
5. Since the statistic is much larger than 3.84, we have sufficient evidence to reject the H0

Inferential Statistics and Modelling

Central Limit Theorem (Sampling Distribution of Means)

1. Concept – If we draw infinite samples of size n from a distribution and plot the means of those samples, we get sampling distribution of the mean. This is a theoretical concept as drawing infinite samples is not practical.

2. The sampling distribution of the means of those samples will become approximately normally distributed with mean µ and standard deviation σ/√ N as the sample size (N) and the number of samples taken become larger, irrespective of the shape of the population distribution



3. The mean value of the distribution will be a close estimate of the population mean µ

Normal Deviate Z Test

1. **Objective** – Often we are required to compare two distribution means of data. For e.g. the given sample data in a data science project and the production data. That the sample means is different from population mean is not the point, the question is, is the difference too small to assign it to statistical fluctuations in sampling or is it too large to be ignored i.e. due to something more than statistical fluctuation

2. **Approach** - Z norm dev = (M – µ) / std_error. M is sample mean, µ is population mean and std_error is standard deviation of the distribution of sample means. Look up the probability in a normal curve table

   a. It considers the sample mean as one instance of the sample means in the sample distribution of means (CLT) with population mean being µ. We have population mean and sample mean. What we need is the standard error of means

   b. Std_error is estimated as - standard_dev of population / sqrt(sample size). Standard_error is the standard deviation of the sample distribution of means (Central Limit Distribution)

   c. The formula looks very similar to Z score calculation because both Z score calculation and Z Norm_dev because both are examples of test of statistical significance

3. **Pre-requisites –** We need to know the population mean (µ) and population standard deviation (std_dev)

Ref: Statistical_Tests.ipynb

Two Sample T-Test

1. **Objective** – Determine whether the null hypothesis, that “the true mean difference between the paired samples is zero” is likely to be true or false. In other words, we can formulate our hypothesis as –

$H0$: $\mu d = 0$
$H1$: $\mu d \neq 0$ (two-tailed)
$H1$: $\mu d > 0$ (upper-tailed)
$H1$: $\mu d < 0$ (lower-tailed)

We will be doing two tailed test

2. μd refers to the differences in the mean of the paired samples, it’s value in population is unknown

3. **Assumptions** –
   a. The dependent variable must be continuous (interval/ratio).
   b. The observations are independent of one another.
   c. The dependent variable should be approximately normally distributed.
   d. The dependent variable should not contain any outliers.

4. **Approach** -
   a. Assuming the two samples have come from two different populations, we create a sampling distribution of the differences between the means. This is only an imaginary step
   b. Under Null hypothesis (H0) this distribution will have mean of 0 because if H0 is true, there is no difference in the population of the two given samples
   c. Calculate the difference in mean of the given samples. This is one instance of the differences from the imaginary distribution
   d. Calculate standard error of differences between the means
   e. Estimate the T-statistic and generate the P values

Ref: Statistical_Tests.ipynb