Everything You Need To Know about Hypothesis Testing — Part II

In this second part, I will be focusing on the **Statistical Tests**. After we look at the distribution of data and perhaps conducting some descriptive statistics to find the mean, median, or mode, it is time to make inferences about the data. In statistics, we categorize "Hypothesis Testing" under Inferential Statistics. **Statistical Tests** allow us to make inferences because they can show whether an observed pattern is due to intervention or by chance.

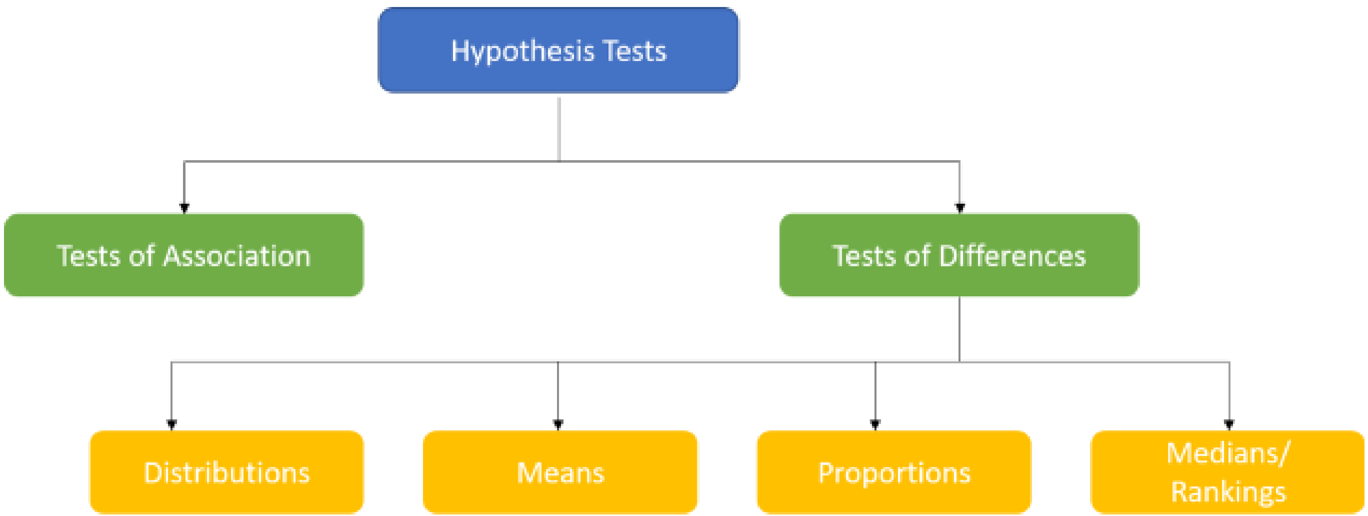
In general, we identify the most important features in the modeling process by observing their ***P-values,*** unless some critical business scenarios refrain us from considering other variables. If we analyze the output of any popular modeling techniques, be it Linear or Logistic regression or XGBoost or Random Forests, we check for ***P-Values*** and whether it is ***<0.05 (at a 95% confidence level)*** or not for the significance in the model. How are we getting these ***P-values*****?** ***What are we testing in the underlined modeling process? And What statistical tests performed?***

These questions will help you to understand what hypotheses we are testing on the predictor variables, which eventually help you in making business decisions. However, sometimes, analytics professionals get confused with ***Which Statistical Test should use in a given scenario? The*** choice of which statistical test to use depends on the research design, the distribution of the data, and the type of variable(s). But, by keeping a bit close attention to the problem at hand, it's easy to decide on the type of statistical test we should use to draw an inference.

Before we learn about all of these, let's try to understand some nuances about the Statistical Tests.

A Statistical Test intends to determine whether there is enough evidence to "reject" a null hypothesis or not. Not rejecting may be a good result if we want to continue to act as if we "believe" the null hypothesis is correct. Or it may be a disappointing result, possibly indicating we may not yet have enough data to "prove" something by rejecting the null hypothesis.

When we are performing a Hypothesis Testing, we must question ourselves on what are we testing? Are we testing an association between the two groups or a difference between them? Do we have one group of items to test or two groups? All these questions matter deciding upon choosing the right **Statistical Test**. In general, we can broadly classify the Hypothesis Tests in two ways.



Classification of Hypothesis Tests

Let's look at an example of testing an association between the two groups:

In my previous post, we discussed whether an Internet shopping service had been introduced or not. For those of you who didn't get a chance to read my last post, here is the example again:

Ex: A major department store is considering the introduction of an Internet Shopping service. The new service will introduce if more than 40 percent of the Internet users shop via the Internet. In this case, we performed hypothesis testing using ***Z-test***, and the Null and Alternate hypotheses are;

$$H_0: \pi \leq 0.40$$

$$H_1: \pi > 0.40$$

Now suppose that we are interested in determining whether **Internet usage is related to gender or not.**

Let's assume that those reporting 5 hours or less usage classified as light users, and the remaining are heavy users. The data are shown in the below cross-tabulation.

Cross-Tabulation: *It is a method to analyze the relationship between multiple variables quantitatively. Sometimes we call this as "Contingency Table" or "Cross tabs." Cross-Tabulation is usually performed on* ***Categorical data — data that can be divided into mutually exclusive groups.***

Let's consider the below Cross-Tabulation with Light and Heavy Internet users.

Gender and Internet Usage

	Gender		
Internet Usage	**Male**	**Female**	**Row Total**
Light	5	10	15
Heavy	10	5	15
Column Totals	15	15	

The above cross-tabulation includes a cell for every combination of the categories of the two variables. 10 respondents were females who reported light Internet usage. The marginal totals in this table indicate that of the 30 respondents with valid responses on both the variables, 15 reported light usage, and 15 were heavy users. In terms of gender, 15 respondents were females, and 15 were males.

To test the statistical significance of the observed association in a cross-tabulation, we use the ***Chi-Square Statistic (χ^2)***. It assists us in determining whether a systematic association exists between the two variables. In this case, we are trying to test the association between Internet usage (Light/Heavy) Vs. Gender.

The null hypothesis Ho is that there is no association between the variables.

How the chi-square test is computed: In general, the test is conducted by computing the cell frequencies that would be expected if no association were present between the variables, given the existing row and column totals. These expected cell frequencies, denoted f_e are then compared to the actual observed frequencies, f_o found in the cross-tabulation to calculate chi-square statistic. The greater the discrepancies between the expected and actual frequencies, the larger the value of the statistic. Assume that a cross-tabulation has ***r*** rows and ***c*** columns and a random sample of ***n*** observations. Then the expected frequency can be calculated as:

$f_e = (n_r\, n_c)/n$, where,

n_r = total number in the row,

n_c = total number in the column,

n = total sample size

For the Internet usage data, the expected frequencies for the cells, going from left to right and from top to bottom, are:

(15 X 15) / 30 = 7.50

(15 X 15) / 30 = 7.50

(15 X 15) / 30 = 7.50

(15 X 15) / 30 = 7.50

The value of χ^2 is calculated as follows:

$\chi^2 = \sum (f_o - f_e)^2 / f_e$

$\chi^2 = (5–7.5)^2/7.5 + (10–7.5)^2/7.5 + (10–7.5)^2/7.5 + (5–7.5)^2/7.5$

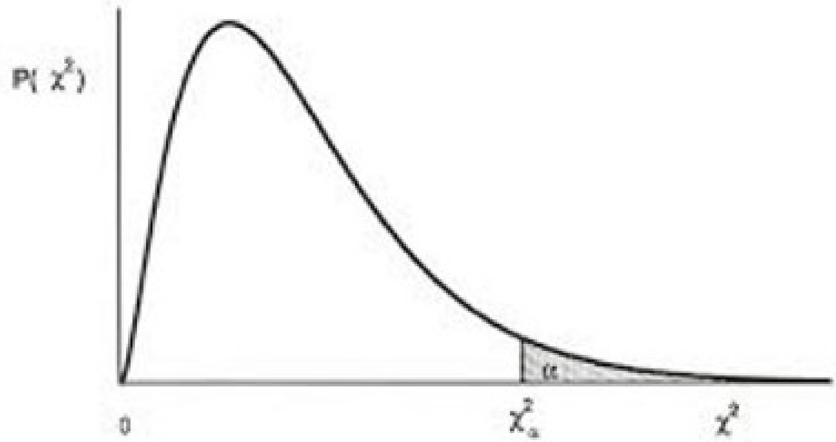
$\chi^2 = 0.833 + 0.833 + 0.833 + 0.833$

$\chi^2 = 3.333$

To determine whether a systematic association exists or not, the probability of obtaining a value of chi-square as large as or larger than the one calculated from the cross-tabulation is estimated (in other words, calculating p-value). An important characteristic of the chi-square statistic is the number of ***degrees of freedom (df)*** associated with it.

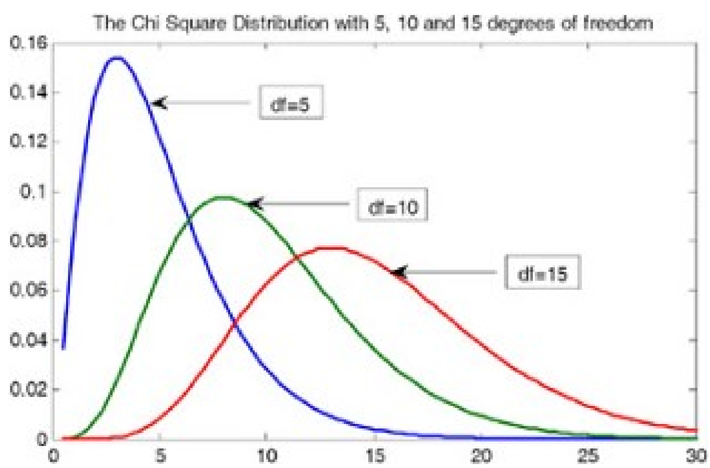
> ***Degrees of Freedom:*** *It refers to the number of values involved in the calculations that have the* **freedom** *to vary. In other words, the* **degrees of freedom,** *in general, can be defined as the total number of observations minus the numbers of independent constraints imposed on the observations.*

In the case of a chi-square statistic associated with a cross-tabulation, the number of degrees of freedom is equal to the product of number of rows *(r)* less one and the number of columns © less one. That is, df = *(r-1) * (c-1)*. The null hypothesis (Ho) of no association between the two variables will be rejected only when the calculated value of the test statistic is greater than the critical value of the chi-square distribution with the appropriate degrees of freedom, as shown in the below image.



chi-square distribution with the critical region

> *The chi-square distribution is a skewed distribution whose shape depends solely on the number of degrees of freedom. As the number of degrees of freedom increases, the chi-square distribution becomes more symmetrical.*



In this case, the degree of freedom is 1 (i.e., *(2–1) * (2–1)*). Using chi-square tables, we get the critical value as 3.84 at a 5% level of significance (You can see from below chi-square table)

Chi-square Distribution Table

d.f.	.995	.99	.975	.95	.9	.1	.05	.025	.01
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09
6	0.68	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81

7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81
19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57
22	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89
32	15.13	16.36	18.29	20.07	22.27	42.58	46.19	49.48	53.49
34	16.50	17.79	19.81	21.66	23.95	44.90	48.60	51.97	56.06
38	19.29	20.69	22.88	24.88	27.34	49.51	53.38	56.90	61.16
42	22.14	23.65	26.00	28.14	30.77	54.09	58.12	61.78	66.21
46	25.04	26.66	29.16	31.44	34.22	58.64	62.83	66.62	71.20
50	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15
55	31.73	33.57	36.40	38.96	42.06	68.80	73.31	77.38	82.29
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38
65	39.38	41.44	44.60	47.45	50.88	79.97	84.82	89.18	94.42
70	43.28	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.43
75	47.21	49.48	52.94	56.05	59.79	91.06	96.22	100.84	106.39
80	51.17	53.54	57.15	60.39	64.28	96.58	101.88	106.63	112.33
85	55.17	57.63	61.39	64.75	68.78	102.08	107.52	112.39	118.24
90	59.20	61.75	65.65	69.13	73.29	107.57	113.15	118.14	124.12
95	63.25	65.90	69.92	73.52	77.82	113.04	118.75	123.86	129.97
100	67.33	70.06	74.22	77.93	82.36	118.50	124.34	129.56	135.81

Chi-Square table at different α values

We got chi-square calculated value as 3.33. Because this is less than the critical value of 3.84, the null hypothesis of no association cannot be rejected, indicating that the association is not statistically significant at the 5% level. Note that this lack of significance is mainly due to the small sample size (which is 30 in this case). If instead, the sample size were 300, and each entry of the cross-tabulation were multiplied by 10, it can be seen that the value of the chi-square statistic would be multiplied by 10 and would be 33.33, which is significant at the 5% level.

> *The chi-square statistic should be estimated only on counts of data. When the data are in percentage form, they should first be converted to absolute counts or numbers. In addition, an underlying assumption of the chi-square test is that the observations are drawn*

> *independently. As a general rule, chi-square analysis should not be conducted when the expected or theoritical frequencies in any of the cells is less than five.*

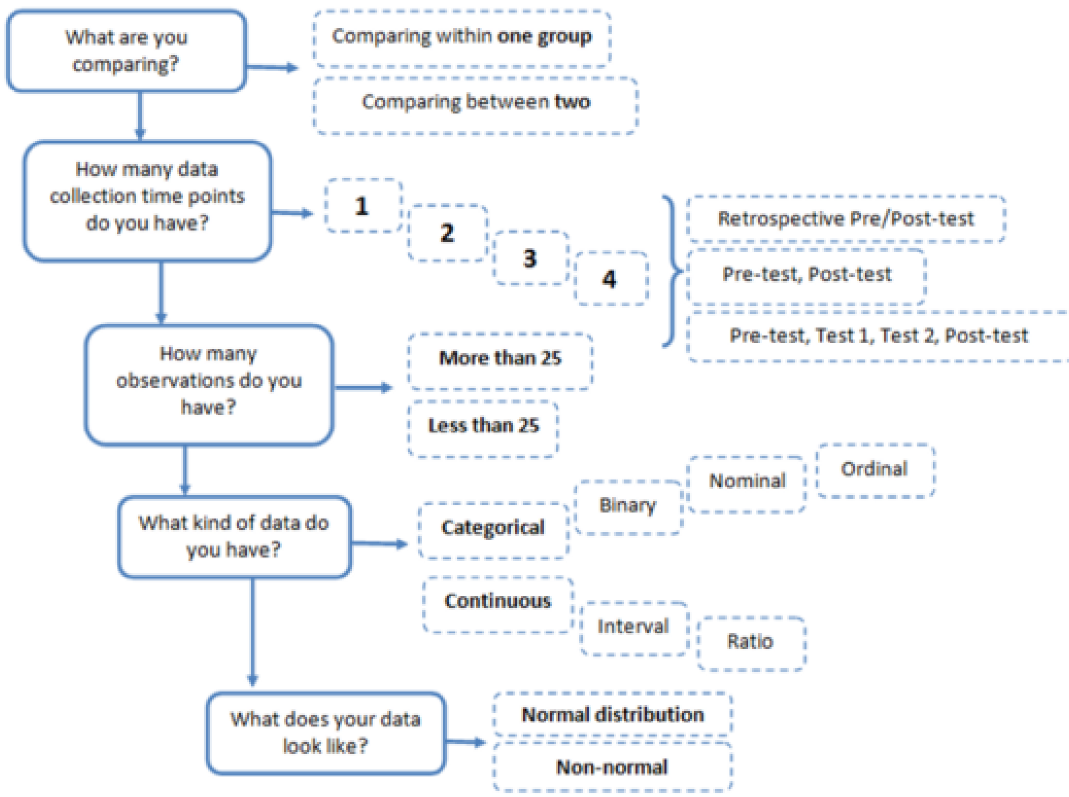
Like the chi-square test, other statistical tests can be used depending upon the problem at hand. Let's look at what are other statistical tests available and in what scenario we can use them.

Type of Test	Use
Correctional: these tests look for an association between variables	
Pearson Correlation	Tests for the strength of the association between two continuous variables
Spearman Correlation	Tests for the strength of association between two ordinal variables (does not rely on the assumption of normally distributed data)
Chi-Square	Tests for strength of the association between two categorical variables
Comparison of Means: these tests look for the difference between the means of variables	
Paired T-test	Tests for the difference between two variables from the same population (e.g., a pre- and posttest score)
Independent T-test	Tests for the difference between the same variable from different populations (e.g., comparing boys to girls)
ANOVA	Tests for the difference between group means after any other variance in the outcome variable is accounted for (e.g., controlling for gender, income or age)

Type of Test	Use
Regression: these tests assess if change in one variable predicts change in other variable	
Simple Regression	Tests how change in the predictor variable predicts the level of change in the outcome variable
Multiple Regression	Tests how change in the combination of two or more predictor variables predict the level of change in the outcome variable
Non-Parametric: these tests are used when the data does not meet the assumptions required for parametric tests	

Wilcoxon Rank-Sum Test	Tests for the difference between two independent variables; takes into account magnitude and direction of difference
Wilcoxon Sign-Rank Test	Tests for difference between two related variables; takes into account magnitude and direction of difference
Sign Test	Tests if two related variables are different; ignores the magnitude of change – only takes into account direction

The chart below provides a summary of the questions that need to be answered before the right test can be chosen.



Few examples of how to select the right statistical test given the study design:

Study Design	Groups	Time Points	Data Distribution	Variable Type	Test
Pretest/Posttest: Studying an eight-week tutoring component at an after-school program. Assessing student satisfaction of 40 participants. Comparing the same students' satisfaction using a pretest before tutoring and posttest after the tutoring component ends.	One	Two	Normal	**Ordinal:** 1=Very satisfied, 2=Satisfied, 3=Not at all Satisfied	**Paired T-Test**
Same as above.	One	Two	Non-Normal	Same as above.	**Wilcoxon Sign-Rank Test**
Pretest, Posttest, and Control Group: Comparing the satisfaction of two groups of students in different after-school programs.	Two	Two	Non-Normal	Same as above.	**Wilcoxon Rank-Sum Test**

Each group has 25 participants. Comparing the satisfaction scores using a pretest before the intervention and a posttest after the intervention.					
Pretest, Posttest: Assessing weight loss after a nutrition intervention among the one group of 50 students who receive the intervention. Would like to determine if there is a relationship between participation in the intervention and weight loss. Weight is measured before and after the intervention.	One	Two	Normal	**Continuous (ratio):** Weight in Pounds	**Paired T-Test**