


How to compare apples and oranges? : Part I

 clevertap.com/blog/how-to-compare-apples-and-oranges-part-i

How often have you come across the idiom “Comparing apples and oranges”. It is a great analogy to articulate that two things can’t be compared due to the fundamental difference between them. As an analyst, you deal with such difference and make sense of it on a daily basis.

Let’s take an example and understand some ways to compare apples and oranges.

We will attempt to understand ways to compare apples and oranges by transforming the data and its key metrics. Please read this [article](#), if you need to compare relationship of different types of variables visually.

We will start with numerical variables. Consider you have the below dataset and you need to compare the variables:

First 10 Rows			Data Summary		
Salary (US\$)	Experience (years)	Age (years)	Salary (US\$)	Experience (years)	Age (years)
1940	3	24	Min : 191	Min : 0.000	Min : 23.00
2270	5	32	1st Qu : 1874	1st Qu : 2.750	1st Qu : 26.00
4059	6	38	Median : 2442	Median : 5.000	Median : 31.00
2571	8	33	Mean : 2492	Mean : 5.085	Mean : 31.29
2630	7	39	3rd Qu : 3068	3rd Qu : 7.000	3rd Qu : 36.00
4216	9	40	Max : 5742	Max : 15.000	Max : 45.00
1235	0	27			
1814	2	25			
2055	4	28			
3725	5	35			

The first thing to hit you is all the above 3 variables in the dataset are different. But, the job of an analyst is to bring order in chaos by making sense of data. Some of the questions that may come to your mind:

- Which variables vary the most?
- How related are the variables?
- Can I use these variables in my predictive model directly?

Let’s take the above questions one by one and attempt to answer them.

a) Which variables vary the most ?

We need to compare the variation between variables to answer the above question. But, first we need to know how the variables vary or the measure of their dispersion. Variance is a popular measure of dispersion or variation.

$$\sigma^2 = \frac{\sum(x - \mu)^2}{N}$$

where $\sigma^2 = \text{Variance}$; $\mu = \text{Mean of } x_i$; $n = \text{Number of observations}$

Steps involved in calculating the Variance:

Step 1) Calculate the mean (average) of the variable,

Step 2) Subtract mean from each of the observation and square it,

Step 3) Sum up the values obtained from Step 2,

Step 4) Divide the value obtained in Step 3 by the number of observations.

The above steps attempt to capture the average deviation of each observation of the variable from the mean. As a result of step 2 and step 3, a positive or a negative deviation from the mean only increases the variance. The variables in the example have their own unit measurements; Salary in dollars, Experience in years and Age in years. The unit of measurement for Variance is square of its respective unit (see Step 2 above). Standard Deviation is obtained by taking the square root of the Variance. This results in the measurement unit of Standard Deviation to be same as the original unit of measurement of the variable.

	Salary	Experience	Age
Variance	889525 US\$	10.70 years	34.67 years
Standard Deviation	943.15 US\$	3.27 years	5.88 years

The above table gives us the variance and the standard deviation for the 3 variables. Since we have the measure of variation for all 3 variables, can we compare the variation among them? The answer is an emphatic “NO” since the variables are not measured on the same scale or unit.

Can we make some modification to the standard deviation to make it comparable? What if we make standard deviation unitless? This is where mean comes to our rescue. Mean is measured in the same unit as that of the Standard Deviation. Dividing standard deviation by mean achieves our objective and the result is known as Coefficient of variation.

$$\text{Coefficient of variation} = \frac{\sigma}{\mu} * 100$$

where $\sigma = \text{Standard Deviation}$; $\mu = \text{Mean}$

	Salary	Experience	Age
Coefficient of Variation	37.88	64.33	18.82

From the above table, it seems that Experience shows more variation than the other variables. Coefficient of variation has enabled us to compare the degree of variation in the variables even though they have drastically different means and scale.

ii) How related are the variables ?

Analyzing relationship between variables requires key metrics as well as visualization either to support the metrics or understand the variables or its relationship better. One

metric, which is the most common and popular to understand relationship between numerical variables, is Correlation. To understand correlation, let's look at covariance first:

$$Cov(x, y) = \frac{\sum (x - \bar{x})(y - \bar{y})}{N}$$

*where $Cov(x, y)$ = Covariance of x and y ;
 \bar{x} = Mean of x ; \bar{y} = Mean of y ; N = Number of observations*

The formula is essentially a variation of the variance formula with the first 2 steps of calculating variance replaced with the following 2 steps:

Step 1: Calculate the mean of the 2 variables

Step 2: Subtract respective means from the observations of the respective variables and multiply the results obtained

Rest of the steps remains the same as that of Variance. In this case, the steps attempt to capture the average deviation of 2 variables from their respective means simultaneously. Covariance can either be positive, if the variables move together or negative, if they move in the opposite direction (see step 2).

	Salary,Experience	Salary,Age	Age,Experience
Covariance	2653 US\$. Years	4190 US\$. Years	16 Years . Years

Covariance suffers from the same problem that we faced with Variance due to the units attached to it. So the unit for covariance between Salary and Age will be dollars x years. We need to get rid of the units and standardize covariance to enable comparison. If you have noticed, the unit of the covariance is a multiplication of the units of the 2 variables. What if we divide the covariance by the respective standard deviations of the 2 variables? Since standard deviation is measured in the unit of the variable, multiplying the standard deviation of the 2 variables will yield the same measurement unit as that of covariance, thereby resulting in unitless measure, when we divide both.

$$\rho_{x,y} = \frac{Cov(x, y)}{\sigma_x \cdot \sigma_y}$$

*where $\rho_{x,y}$ = Correlation Coefficient of x and y ;
 $Cov(x, y)$ = Covariance of x and y ; σ_x = Standard Deviation of x ;
 σ_y = Standard Deviation of y*

Correlation coefficient is bounded with lower bound of -1 and upper bound of +1. The closer the metric is to both the bounds, higher is the movement of the variables together or against each other. If the correlation coefficient is closer to -1 (negatively correlated), the variables move against each i.e. if one moves up the other falls. Likewise, if the correlation coefficient is closer to 1 (positively correlated), the variables move together i.e. if one moves up, the other follow.

	Salary,Experience	Salary,Age	Age,Experience
Correlation Coefficient	0.86	0.75	0.83

As per the table, there seems to be a strong positive relationship among all the variables with the highest between Salary and Experience, since the correlation coefficient is positive for all and close to +1.

iii) Can I use these variables in my predictive model directly ?

When you compare variables, you look at relative metrics rather than absolute metrics. You won't be looking at coefficient of variation of Salary in isolation but compare it with the other 2 variables. The first 2 questions attempted to compare the individual metric. But, what if the goal was to compare the observations of the entire dataset rather than an individual metric of the variables. Building predictive models requires the entire dataset as the input.

Can you use the variables in the dataset as the input directly, especially if you use machine learning algorithms? Do you want your algorithms to give importance to variables just because their value is relatively high than other variables? In our example, Salary is on the highest scale. So, if we give all the variables in our dataset as an input to K-means, a popular machine learning algorithm, the algorithm will tend to give more importance to 'Salary' and the resultant clusters will be formed, probably segmenting just Salary and not the other variables in conjunction. This is because the algorithm just sees the values of the variables. So if K-means clusters the observations based on the numeric distance between observations, it is logical that Salary gets a higher weight in determining how the observations get clustered since its value is much higher than the other 2 variables.

What's the way out? We have looked at standardizing individual metrics till now and not the entire data. Let's look at a method to standardize the entire variable data using z-score.

$$z = \frac{x - \mu}{\sigma}$$

where z = Standard score; μ = Mean of x ; σ = Standard Deviation of x

In the above formula, we are subtracting each observation from its mean and dividing the result by the standard deviation. If you look at the formula closely, the resultant z-score is unitless as the units get canceled out due to the division.

First 10 Rows			Data Summary		
Salary	Experience	Age	Salary	Experience	Age
-0.585	-0.637	-1.238	Min : -2.4396	Min : -1.5544	Min : -1.4079
-0.235	-0.026	0.121	1st Qu : -0.6546	1st Qu : -0.7138	1st Qu : -0.8984
1.662	0.280	1.140	Median : -0.0530	Median : -0.0260	Median : -0.0493
0.084	0.891	0.290	Mean : 0.0000	Mean : 0.0000	Mean : 0.00003
0.146	0.585	1.309	3rd Qu : 0.6114	3rd Qu : 0.5854	3rd Qu : 0.8000
1.828	1.197	1.479	Max : 3.4460	Max : 3.0309	Max : 2.3284
0.497	1.197	-0.389			
-1.333	-1.554	-0.729			
-0.719	-0.943	-1.068			
-0.463	-0.332	-0.559			

The table on the left shows the z-scores for the first 10 rows of the variables and the table to the right shows the data summary of the variables, post transformation.

	Salary	Experience	Age
Mean	0	0	0
Standard Deviation	1	1	1

As seen from the above table, z-scores for the variables have transformed their mean to zero and standard deviation to 1.

There are situations where you require standardization, especially in machine learning techniques like PCA, K-means, etc but, at the same time, you might be required to transform the data into a particular range like in image processing, where pixel intensities have to be normalized to fit within a certain range (i.e., 0 to 255 for the RGB color range). Also, neural network algorithms may use data that are on a 0-1 scale in a way to avoid bias. This bias may arise due to the observations that are at the extreme end of the range or are outliers. To avoid such issues, a transformation technique, which bounds the data within a range, is required. This can be achieved with Normalization.

$$normal\ score_x = \frac{x - x_{min}}{x_{max} - x_{min}}$$

where $normal\ score_x$ = normalized score of x ;
 x_{max} = Maximum of x ; x_{min} = Minimum of x

Just like standardization, normalized data too is unitless. Let's see what does our original dataset look like after normalization.

First 10 Rows

Salary	Experience	Age
0.315	0.200	0.045
0.375	0.333	0.409
0.697	0.400	0.682
0.429	0.533	0.455
0.439	0.467	0.727
0.725	0.600	0.773
0.499	0.600	0.273
0.188	0.000	0.182
0.292	0.133	0.091
0.336	0.267	0.227

Data Summary

Salary	Experience	Age
Min : 0.0000	Min : 0.0000	Min : 0.0000
1st Qu : 0.3033	1st Qu : 0.1833	1st Qu : 0.1364
Median : 0.4055	Median : 0.3333	Median : 0.3636
Mean : 0.4145	Mean : 0.3390	Mean : 0.3768
3rd Qu : 0.5184	3rd Qu : 0.4667	3rd Qu : 0.5909
Max : 1.0000	Max : 1.0000	Max : 1.0000

The table on the left shows the normalized scores for the first 10 rows of the variables and the table to the right shows the data summary of the variables, post transformation. All the variables are bounded between 0 and 1.

So how would you select between Standardization and Normalization? Depending on the objective of the technique, the method has to be selected. For techniques like PCA or K-means, you would like to retain the unbounded nature and the variation in the data while at the same time make the data unitless and relatively on same scale (range of transformed variable values). In such a case, Standardization is your best bet. There are times when you don't want your data to be unbounded like in the case of Standardization so that your technique does not give a bias to observations, which are towards the higher/lower side of the range (potential outliers). Normalization reduces the impact of outliers as the range of the data is strictly between 0 and 1.

Closing Thoughts

Comparing different variables should first involve identification of the purpose of such comparison depending on which the appropriate technique to transform the variables or the key metric should be selected. With the help of an example, we looked at coefficient of variation, correlation, standardization and normalization as some of the ways to compare and use different numerical variables for analysis and build predictive models on.

In the ensuing part, we will discuss how to compare categorical variables.

How to compare apples and oranges ? : Part II

 clevertap.com/blog/how-to-compare-apples-and-oranges-categorical-variables-part-ii

In the [previous article](#), we looked at some of the ways to compare different numerical variables. In this article, we shall look at techniques to compare categorical variables with the help of an example.

Assume you have been given a dataset totaling 10,000 rows containing user information on Operating System, Gender and whether the user has transacted over a particular period.

First 10 Rows			Data Summary		
Gender	OS	Transact	Gender	OS	Transact
M	Android	no	F : 3456	Android : 5653	no : 7940
F	iOS	no	M : 6544	Windows : 957	yes : 2060
M	iOS	no		iOS : 3390	
F	Android	yes			
F	iOS	no			
M	Android	no			
M	iOS	yes			
F	Android	yes			
M	Windows	no			
M	Android	yes			

All the variables mentioned above are categorical variables. It seems 35% of the users are Female and 65% Male. Female Android users constitute 25% and Male Android users constitute 75% of the Android users. If there is no association between Gender and OS, you will expect that the percentage composition of Female Android and Male Android users (25% & 75%) will be similar to that of the percentage composition between Female and Male users (35% and 65%). The same holds true for Windows and iOS users. But, is there a way to conclude if the observed difference is big enough to concur that the percentage composition indeed is not similar? In short, we are trying to ask the question '**Is there an association between the categorical variables – Gender and OS?**'.

Since we are dealing with categorical variables, we cannot use the techniques like coefficient of variation or correlation coefficient used to analyze numerical variables.

In order to compare categorical variables, we have to work with frequency of levels/attributes of such variables. From the above table, we know that the frequency of 'Android' in OS is 5653 users of which Male users

are 1385 and Female users are 4268 as can be seen in the first row of the table. We need to use this frequency to compare the categorical variables. The above table is a Contingency table where we are analyzing 2 categorical variables. A contingency table is essentially a display format used to analyze and record the relationship between two or more categorical variables. Let's further analyze the contingency table:

Contingency Table			
	M	F	Row Total
Android	1385	4268	5653
Windows	330	627	957
iOS	1741	1649	3390
Column Total	3456	6544	10000

From the above table, it seems that the break-up of Gender is different across Operating System. For example:

- Android users constitute 56.53% of the total users. But, if we segregate the users based on Gender, we get different percentages for Males and Females on Android.
- Male Android users constitute 40.08% of the Male users whereas Female Android users constitute 65.22% of the Female users.

		M	F	Row Total
Android	r	1385	4268	5653
	c	24.50%	75.50%	100.00%
	c	40.08%	65.22%	56.53%
Windows	r	330	627	957
	c	34.48%	65.52%	100.00%
	c	9.55%	9.58%	9.57%
iOS	r	1741	1649	3390
	c	51.36%	48.64%	100.00%
	c	50.38%	25.20%	33.90%
Column Total	r	3456	6544	10000
	c	34.56%	65.44%	100.00%
	c	100.00%	100.00%	100.00%

* r = Row percentage c = Column percentage

The question that may arise is why there is a difference in the frequency percentages when we look at levels in a single category compared to the combination of levels of more than 1 categorical variable. Is there an association between the Gender and OS resulting in the difference? Is this deviation in percentages statistically significant to conclude the presence of some association?

Statistical significance

We often come across the term 'statistical significance' or 'random chance'. But what does it mean intuitively?

Imagine you are tossing 2 coins, A and B 10 times. Coin 'A' landed heads 3 times whereas Coin 'B' landed heads 5 times. Does it mean that Coin 'A' is an unfair coin where chances of landing tails are more than heads? You know intuitively that the difference could have occurred simply due to luck or by chance. But, what if you have tossed the coins 1000 times and Coin 'A' landed heads 100 times whereas Coin 'B' landed heads 550 times? Would you still attribute this difference to chance or some other underlying factors such as the shape of the coins? We can answer this difference with the help of statistical tests.

Coming back to our discussion on our example of User data, we will attempt to answer the difference seen in the contingency table with the help of Hypothesis testing.

Hypothesis testing

Claim 1: Gender is independent of Operating System (No Association)

Claim 2: Gender is not independent of Operating System (Association)

The above 2 claims/statements are essentially what we test in hypothesis testing. We deal with hypothesis on a daily basis. We might have hypothesis on political issues, social issues, financial issues, etc. For example, we might have a hypothesis on whether it will rain today?

In any hypothesis, you will have a default or null hypothesis referred to as H_0 (Claim 1), which is your default belief and an alternate hypothesis referred to as H_1 (Claim 2), which is against your default belief. The null hypothesis is the statement being tested. Usually the null hypothesis is a statement of “no effect” or “no difference”.

So, in our example, we would expect the percentage composition of Gender to be the same for Android, Windows and iOS users (Null Hypothesis). The Alternate Hypothesis is that we don't expect it to be the same. Here, the word 'same' does not imply that the percentage composition has to be exactly equal but it means that there is no statistical difference. We run some appropriate statistical tests to determine it i.e. whether to accept or reject the null hypothesis.

But, prior to that, we need to understand 3 statistical concepts, (i) p-value (ii) chi-square statistic (iii) degrees of freedom

i) What is p-value ?

Assuming you have a hypothesis (in the above case, Gender and OS are independent of each other), the p-value helps you evaluate if the null hypothesis is true. Statistical test use p-value to determine whether to accept or reject the null hypothesis. It measures how compatible your data is with your null hypothesis or the chance that you are willing to take in being wrong. For example, a p value of 0.05 and 0.1 means you are willing to let 5% and 10% of your predictions be wrong respectively.

In other words, p-value is the probability of observing the effect by chance in your data, assuming the null hypothesis is true. So, lower the p-value, lower the probability of observing the effect by chance or at random, and higher the probability of rejecting your default or null hypothesis. In practice, depending on your area of study, generally you have cut-off levels like 1% and 5% for p-values below which you could conclude that the effect is not random or by chance and the null hypothesis could be rejected.

ii) What is Chi-square statistic ?

Before understanding chi-square statistic, let's understand the concept of observed and expected frequencies.

Observed			
	M	F	Row Total
Android	1385	4268	5653
Windows	330	627	957
iOS	1741	1649	3390
Column Total	3456	6544	10000

Expected			
	M	F	Row Total
Android	1954	3699	5653
Windows	331	626	957
iOS	1172	2218	3390
Column Total	3456	6544	10000

Observed frequencies are the actual frequencies as seen in the data and shown in the contingency table above as 'Observed'. This is the same contingency table, we had introduced earlier. Expected frequencies are frequencies we could expect if there was absolutely no association and shown in the contingency table above as 'Expected'.

How did we calculate the expected frequency? Expected frequency is calculated from the observed or actual frequencies.

$$\text{Expected Cell Frequency} = \frac{\text{row total containing the cell} * \text{column total containing the cell}}{\text{total number of observations}}$$

$$\text{Expected cell frequency for Male Android Users} = \frac{5653 * 3456}{10000} = 1954$$

Let's analyze the expected frequencies further:

In the above table, the row and column percentages are quite similar and there doesn't seem to be a difference in percentages due to the influence of Gender on OS or vice-versa.

Chi-square test is a statistical test commonly used to compare observed data with the data we would expect to obtain according to a specific hypothesis. In our example, we would

have expected 1954 of 5653 Android users to be Male but actual or observed were 1385. So is this deviation of 569 users statistically significant? Were the deviations (differences between observed and expected) the result of chance, or were they due to other factors? The chi-square test helps us answers this by calculating the chi-square statistic.

		Expected		
		M	F	Row Total
Android	r	1954	3699	5653
	c	34.56%	65.44%	100.00%
	c	56.53%	56.53%	56.53%
Windows	r	331	626	957
	c	34.56%	65.44%	100.00%
	c	9.57%	9.57%	9.57%
iOS	r	1172	2218	3390
	c	34.56%	65.44%	100.00%
	c	33.90%	33.90%	33.90%
Column Total	r	3456	6544	10000
	c	34.56%	65.44%	100.00%
	c	100.00%	100.00%	100.00%

* r = Row percentage c = Column percentage

$$\text{Chi - square statistic } (\chi^2) = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

That is, chi-square statistic is the sum of the squared difference between observed and the expected data, divided by the expected data in all possible categories.

iii) What is Degrees of Freedom ?

The degrees of freedom is the number of values in a calculation that we can vary. Let's understand degrees of freedom with the help of an example.

Example 1:

Suppose you know that the mean for a data with 10 observations is 25 and that variable has many such sets of 10 observations. So, for a new set of 10 observations, we have the freedom to set the value of 9 observations i.e. you can have the freedom to select any 9 values. But, you won't have the freedom to set the value for the 10th observation. This is because the mean of the data has to be equal to 25. So the value of the 10th observation has to be equal to (25 * 10 – sum of the values of 9 observations). Hence, the degrees of freedom in this case is 9.

Example 2:

Contingency table

In order to run a chi-square test on the contingency table, the row total and the column total is like the mean and the other cells in the contingency table are like the observations in the Example 1. In the above contingency table, we can only freely select 2 values so that the row and column totals are not changed. Hence degrees of freedom is 2. The formula to calculate it for a contingency table with 2 categorical variables is $(r - 1) * (c - 1)$, which for our case is $(3 - 1) * (2 - 1) = 2$

Steps to calculate p-value

In order to accept or reject the Null Hypothesis, we need to calculate the p-value. p-value is calculated in the following 3 steps:

Step 1) Calculate Chi-square statistic

Step 2) Calculate the degrees of freedom

Step 3) Find the p-value corresponding to chi-square statistic with corresponding degrees of freedom in the chi-square distribution table.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188

The above table is an excerpt of a chi-square distribution table. The first column contains degrees of freedom. The cells of each row give the critical value of chi-square for a given p-value (column heading) and a given number of degrees of freedom. For a given degrees of freedom, higher the chi-square statistic (cell value), lower the p-value.

OS & Gender

In our example, the Chi-square statistic (χ^2) for OS and Gender using the chi-square statistic formula = 675.86.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188

In the chi-square distribution table, $\chi^2_{0.005}$ statistic is 10.597 at 2 degrees of freedom. Hence, the p-value has to be less than 0.005. This can be easily solved using a computer rather than manually.

p-value or $P(\chi^2 > 675.86)$ at 2 degrees of freedom $< 2.2e-16$ or almost zero.

We have to compare this p-value with an assumed cut-off level of 5% or 1% known as alpha or significance level. The assumed alpha value helps to conclude if the statistic is observed by chance or by any other factor. The p-value calculated is less than the assumed alpha. Hence, we can say that based on the evidence, we fail to accept or reject the Null Hypothesis and conclude that Gender and OS are not independent.

Gender & Transact

$$\chi^2 = 0.11647$$

Contingency Table of Gender and Transact
--

$P(\chi^2 > 675.86)$ at 1 degree of freedom = 0.7329

Since p-value is greater than the alpha value of 0.05, we fail to reject the Null Hypothesis and conclude that Gender and Transact are independent.

OS & Transact

$$\chi^2 = 24.581$$

Contingency Table of OS and Transact

$P(\chi^2 > 24.581)$ at 2 degrees of freedom = $4.595e-06$ or almost zero.

Since p-value is less than the alpha value of 0.05, we reject the Null Hypothesis and conclude that OS and Transact are not independent.

Closing Thoughts

To sum up, we have been able to compare 2 categorical variables with the help of contingency table and chi-square test. The same concept can be extended to compare more than 2 categorical variables together. The [next article](#) will deal with ways to compare mixed type of variables i.e. when we have to deal with numerical and categorical together.

How to compare apples and oranges ? : Part III

 clevertap.com/blog/how-to-compare-apples-and-oranges-part-iii

In the [part 1](#) and [part 2](#) of the series, we looked at ways to compare numerical variables and categorical variables. Let's now look at techniques to compare mixed type of variables i.e. numerical and categorical variables together. Please read this [article](#) to visually analyze the relationship between mixed type of variables.

We will work with the same dataset that we worked on in [part 2](#) with the addition of a numerical variable, Revenue.

First 10 Rows

Gender	OS	Transact	Revenue
M	Android	no	0
F	iOS	no	0
M	iOS	no	0
F	Android	yes	1954
F	iOS	no	0
M	Android	no	0
M	iOS	yes	3014
F	Android	yes	2014
M	Windows	no	0
M	Android	yes	2343

Data Summary

Gender	OS	Transact	Revenue
F : 3456	Android : 5653	no : 7940	Min. : 0.0
M : 6544	Windows : 957	yes : 2060	1st Qu. : 0.0
	iOS : 3390		Median : 0.0
			Mean : 479.8
			3rd Qu. : 0.0
			Max. : 3530.7

Let us now compare the numerical variable, Revenue with the categorical variables. But, again the techniques learnt in [part 1](#) and [part 2](#) wouldn't help in this case due to the difference in types of variables involved. If you look at the levels of the categorical variables closely, Gender and Transact has 2 levels whereas OS has 3 levels.

Some of the questions that may arise are?

- Can we get some insights for Revenue for each of the levels of the categorical variable?
- Is there a significant difference in Revenue between levels of each categorical variable or are they the same?

Let's address the above questions in detail:

- Can we get some insights for Revenue for each of the levels of the categorical variable?

AVERAGE REVENUE

Gender	OS	Transact
Female : 521	Android : 437	no : 0
Male : 458	Windows : 298	yes : 2329
	iOS : 603	

STANDARD DEVIATION OF REVENUE

Gender	OS	Transact
Female : 1046	Android : 832	no : 0
Male : 927	Windows : 714	yes : 509
	iOS : 1205	

The above tables show the mean and the standard deviation of Revenue for each of the levels in the categorical variables. It is amply clear that the above revenue statistics for users across different levels in the categorical variables are different.

Few Insights from Average Revenue

- iOS users are highly valuable compared to Windows users since the average spend of iOS users is more than double of Windows users and 38% higher than Android users.
- Female users on an average have spent 13.7% more than Male users.

Similarly, we can draw out insights based on standard deviation and extend the analysis further based on other summary statistics like median, quantiles, etc.

b) Is there a difference in Revenue between levels of each categorical variable or are they same?

It is amply clear from points discussed in (a) that there is a difference. But, the important question is there a statistical difference, based on which one could take appropriate actions?

Revenue & Gender

Let's begin with comparing Revenue and Gender, which has 2 levels. We will use the mean/average Revenue of Male and Female users to check whether the Revenue is impacted by Gender of the person i.e. is the higher average spend of Female users over Male users statistically significant? We will use Hypothesis Testing discussed in [part 2](#) to answer the same. Please refer [part 2](#) to understand some of the concepts discussed hereinafter.

H0: There is no difference in the means (Mean Revenue of Male Users = Mean Revenue of Female Users)

H1: There is a difference in means (Mean Revenue of Male Users \neq Mean Revenue of Female Users)

We will use a statistical test known as **t-test** to test our hypothesis. Similar to chi-square test, we will calculate the t-statistic, degrees of freedom and get the p-value to compare it against the assumed alpha to conclude whether we can accept or reject the null hypothesis.

t-statistic = 4.0026

p-value for the t-statistic at 1 degree of freedom = 0.1559

Since p-value is greater than the assumed alpha of 0.05, we fail to reject the null hypothesis and conclude that there is no difference in Mean Revenue of Male and Female Users i.e. **Gender of the user does not impact the Revenue of the app**.

Revenue & OS

Can we use t-test to compare the mean Revenue across levels of OS? OS has 3 levels. In t-tests, you compare 2 means at a time. So, you will have to compare 3 combinations of means for OS since it has 3 levels. For those 3 combinations, you will need to test 3 hypotheses as enlisted below:

i) H_0 : Mean Revenue of Android users = Mean Revenue of iOS users

ii) H_0 : Mean Revenue of Android users = Mean Revenue of Windows users

iii) H_0 : Mean Revenue of iOS users = Mean Revenue of Windows users

Each of the tests has an assumed alpha of say 5%. So essentially, the 5% chance of a wrong prediction for each test gets accumulated resulting in approximately 15% chance of error (14.3%* to be precise). The error rate compounds for multiple t-tests.

***Combined alpha** = $1 - (0.95 * 0.95 * 0.95) = 14.3\%$

But, we didn't want the alpha to be greater than 5%. Hence, multiple t-tests won't work in this case. In cases where we have to compare more than 2 means, ANOVA is a better option.

Analysis of Variance (ANOVA)

In order to understand if the levels in the categorical variables affect Revenue, you need to test the following hypothesis:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \dots = \mu_n$$

H1 : Atleast one of the means (μ) is different

In simple terms, before identifying the categorical level for which the mean Revenue is different, you ideally would want to first know if there exists a difference in at least one of the means of the categorical levels.

$$ANOVA = \frac{\text{Variance Between the means}}{\text{Variance Within the means}}$$

where,

Variance Between the means = sum of variances between the means of the levels and the overall mean;

Variance Within the means = sum of the variances of the observations for each level

Let's understand the above formula with help of an example:

Suppose you need to analyze the marks of students in 5 divisions of a class. More specifically, you need to know whether the average marks of at least one division significantly differs from the average marks of the entire class. ANOVA tries to answer it by breaking down the source of variation.

Total Variation = *Variance Between the means* + *Variance Within the means*

The numerator, in the ANOVA formula, calculates the variation in average marks of each division from the average marks of the entire class. The denominator calculates the variation of the marks of each student in each division from the average marks for that division. The ratio of the numerator to denominator helps us to conclude if the average marks of at least one of the divisions is significantly different from the average marks of the entire class. The ratio, which is further and further away from 1, implies that such difference could be statistically significant.

We used the chi-square statistic in [part 2](#) to ascertain the p-value. Here, we will use the F statistic to ascertain p-value for the hypothesis test concerning OS and Revenue.

F-statistic = 50.22

We need to check the F-statistic against the F distribution table or use statistical software to arrive at the p value.

The p-value for the ANOVA between OS and Revenue yielded a value $< 2e-16$ or almost zero. Hence, we can reject the null hypothesis and conclude that the **type of OS has an effect on Revenue from the user**. We can dig deeper and use statistical tests like Tukey's HSD to understand which particular level(s) of the categorical variable has statistically different mean.

What we have looked above is a one-way ANOVA where we analyzed the relationship between a numerical and categorical variable. We can extend this to analyze a numerical variable and more than 1 categorical variable at the same time.

Closing Thoughts

In this series, we have introduced the below techniques to compare variables and make better sense of data.

Numerical Variables	Categorical Variables	Mixed Type
Coefficient of Variation	Chi-square test	t-test
Correlation Coefficient		Anova
Standardization		
Normalization		

Comparing variables, which differ in scale, type, etc. is not an option but a need that could bring out patterns and valuable insights in data. Based on the type of variables, the context and the objective of such comparison, it is prudent to judiciously use the right technique to compare them, else it may lead to poor results.