Eco Internal Publication #ECO-10-430-011301

A New Statistical Test for Seasonality

by

Zurab Khviengia

Research and Development, Eco & Associates, Inc.

1855 W. Katella Ave, Suite 350, Orange, CA 92867

zkhviengia@ecoinc.info

Phone: 1(714)289-0995 Fax: 1(714)289-0965

Abstract

A new statistical test for seasonality is introduced for numeric and categorical variables. The concepts of seasonality and partial seasonality are discussed and their definitions offered. Examples demonstrating the application of this test are considered.

1 Introduction

A notion of seasonality plays an important role in various areas of research. Seasonal effects are present in a wide spectrum of environmental phenomena, influencing these phenomena and changing their characteristics. In most of the cases, statistical tests that can be applied to seasonal data serve the purpose of eliminating these effects rather than studying them directly. A typical example of these tests are seasonal trend tests that study trend and see seasonality as a hindrance that may hide or distort it [1]. There are only very few statistical tests that study seasonality and test for it directly, Hewitt's [2] and Edward's tests [3] being the two of the most widely used ones in this regard. In this paper, we offer a new test for seasonality and discusses its application. Prior to formulating the test itself, we shall discuss the notion of seasonality.

2 Seasonality

Seasonality as a variable does not have an exact definition. Many statistical quantities do, but the seasonality does not. In order to test for seasonality, we have to first define it. It is probably impossible to strictly define what seasonality is. In a most general sense, this is a dependency of one variable on another variable, categorical in nature, which we call 'season'. A dependent variable in this case is called to be seasonal. For example, the level of rainfall can be considered to be a seasonal variable. This general sense of seasonality is what is used in statistical analysis today. However, to test for seasonality, one has to have more rigorous and less qualitative definition of seasonality.

We define seasonality as a periodic, repetitive change of the same nature (increase or decrease) of a dependent on it variable in each corresponding season of each period. So if in each season, the dependent variable increases (or decreases) relative to the previous season, and maintains this pattern throughout all periods, then we call this variable seasonal, or say that seasonality is present in this variable. In many applications, a period is a year and seasons are either months or natural seasons. But in general, these can be any time periods. Weekly data or daily data can as easily be qualified to be seasonal as any other. There should simply be a readily identifiable period subdivided into an equal number of seasons. It is needless to say that there may be other definitions of seasonality, either more specific or more generic, which would lead to different statistical tests than the one we shall describe. For example, one may require not only that the change in a dependent variable be of the same type, but that the actual value in each season be at the same relative position with respect to the mean during the periods, either larger or smaller than that value. One may also restrict the definition even further by demanding the values of the variable to maintain their relative position across all periods. However, we shall adhere to our definition and use it to introduce a statistical test for seasonality.

3 Partial Seasonality

According to our definition, if two consecutive seasons show the same value, the variable cannot be considered seasonal. If there is no change, there is no seasonality, of course, assuming this absence of change persists over the periods. But all other seasons may show seasonality. To address this scenario, we introduce a notion of partial seasonality when not all seasons exhibit seasonal behavior of a function. Partial seasonality can be graded by the coefficient of partial seasonality which can be defined as the ratio of the number of seasons that do show a seasonal effect to the total number of seasons. Thus, if only one season out of four shows no seasonality, we say that the data are seasonal with ¾ partial seasonality coefficient. The notion of partial seasonality will cover the cases when data depend only on one or two seasons, like the sales of winter equipment during a year, for example. It would be unreasonable not to call these data seasonal.

4 Statistical Test

As in all statistical tests, we have to define a test statistic and the distribution that this statistic has for our notion of seasonality. We shall test a seasonal effect for each season separately. A test statistic for each of these tests will be binomially distributed. We shall assign the probability ½ to the change that we test for in the null hypothesis. We formulate the null and alternative hypotheses as:

Ho : Summer data do not show seasonal effect with '+' (that is, the probability of an increase p=1/2)

Ha: Summer data do show seasonal effect with '+' (p > 1/2)

Our binomial distribution will have a probability of ½ for 'increase' in summer data and ½ for all other occurrences (decrease or equality). For numerical data, the probability of equality is negligibly small and will be ignored in our analysis. Thus, if we analyze the data for ten years, eight years out of which show increase in summer, the P-value of binomial distribution will be

$$P\{y \geq 8 \mid p = \frac{1}{2}\} = \sum_{y=8}^{10} \binom{10}{y} \left(\frac{1}{2}\right)^{y} \left(\frac{1}{2}\right)^{10-y} = 0.01$$

That is at 1% of significance level, we can claim that our dataset shows seasonality in summer. Of course, we will arrive at the same conclusion if we choose a significance level to be greater than 0.01 as well. For each season the same test should be conducted. If we are able to reject the null hypothesis for each season, than we will be able to conclude that the dataset shows complete seasonality, but if only for some seasons, we shall claim that the dataset is partially seasonal with the corresponding coefficient of partial seasonality.

7 Seasonality Test for a Categorical variable

If the variable that is tested for seasonality is categorical, the probability of it remaining constant over the seasons can no longer be considered to be zero. It can be comparable to the probability of it changing over the consecutive seasons. Then we have to introduce the probability for no-change, as well as the probabilities for increase or decrease separately. We shall assume that the probability of equality is equal to 1/N, where N is the number of values the categorical variable can take. We assume that each value is equally probable. Then the corresponding probabilities for increase and decrease will be:

$$p_0 = \frac{1}{N}$$

$$p_+ = V^+/N$$

$$p_- = V^-/N$$

where V+ (V-) is the number of values greater (less) than the value in previous season.

In this case, we have to use the Poisson binomial distribution to test our null hypothesis [4]. First , one has to look at the dataset and see which of the outcomes have occurred most frequently: greater, less, or equal. That should form the basis for our null hypothesis. We take the most frequently occurring change and test it statistically. If no-change has occurred most frequently, then our probability is constant (1/N) and we can use Binomial distribution instead of the Poisson binomial distribution to perform the hypothesis testing. If the increase or decrease have occurred most frequently, we have to use the Poisson binomial distribution. The Poisson binomial distribution is a generalization of the binomial distribution to the case when the probability of each event is different. The Poisson binomial probabilities of success in k out of n events depend on the set of probability values for the tested event. For the basis of our calculations, we have used the recursive formula for the probability [4]. If we are not able to reject the null hypothesis and arrive at the conclusion that the values are increasing (or decreasing), we have to accept the null hypothesis and conclude that no seasonality is present in a given season. The rest of the procedure should be identical to the previously considered one. A notion of partial seasonality is as applicable to categorical variables as it was to numerical variables.

8 Examples

8.1 Example 1

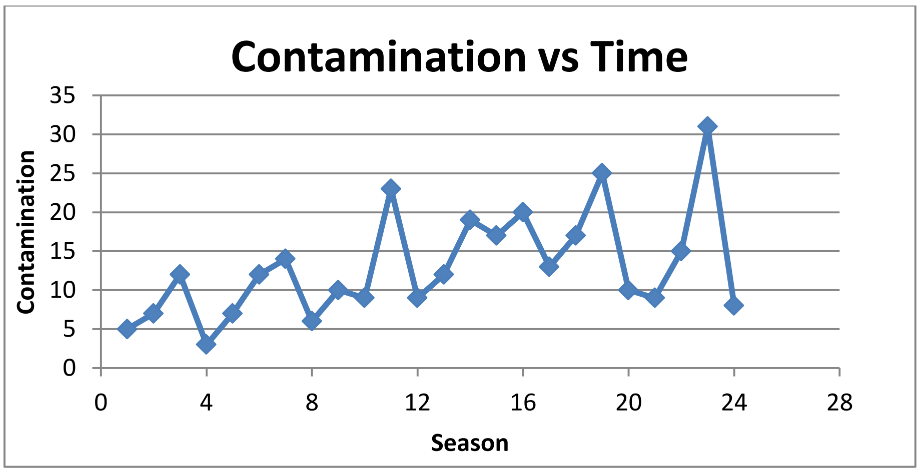
Below we shall consider an example to illustrate the usage of the seasonality test introduced above for numerical variable. Let us take a hypothetical dataset depicting the concentration of a chemical in each season for six consecutive years.

Table 1. Concentration Values

	Winter	Spring	Summer	Fall
2005	5	7	12	3
2006	7	12	14	6
2007	10	9	23	9
2008	12	19	17	20
2009	13	17	25	10
2010	9	15	31	8

These concentration values are plotted on the graph below. Even though the graph clearly indicates that some seasonality may be exhibited by the data, without quantitative analysis, it would be difficult to claim this with certainty.

Figure 1. Contamination vs Season



To apply binomial distribution, first the seasonal indicator table is built.

Table 2. Seasonal Indicators

	Winter	Spring	Summer	Fall
2005	0	+1	+1	-1
2006	+1	+1	+1	-1
2007	+1	-1	+1	-1
2008	+1	+1	-1	+1
2009	-1	+1	+1	-1
2010	-1	+1	+1	-1

Afterwards, for each season, a corresponding hypothesis is tested. Since winter, spring and summer indicate certain increase, for them the null hypothesis will be formulated for the increase. For fall, decrease in concentration will be tested. Corresponding P-values are given in Table 3 below where n indicates the total number of seasonal indicators for which the test is conducted out of the overall N=6 events.

Table 3. P-values for binomial distribution with N=6.

n (N=6)	α
3	0.34
4	0.11
5	0.02

One can conclude that at the significance lever of 5% (α=0.05), the spring, summer and fall data, for which n=5 and P = 0.02, show the seasonal increase during the given six years. One can state that the data are partially seasonal with the coefficient of partial seasonality being equal to 3/4. At the same time, at the significance level of 35% (α=0.35), the fall data, in addition to the data for the other three seasons, also exhibit seasonal increase in the concentration. At this level of significance, therefore, the entire dataset can be claimed to be seasonal.

8.2 Example 2

The test with categorical variables proceeds in a similar way to the one considered above. First, each season is tested for seasonality and then a conclusion is made whether the full data set is seasonal.

Let us consider the following hypothetical data set made of detected levels of contamination measured on the scale from 1 to 10 – 1 being the lowest and 10 being the highest possible concentration. We assume that the measurements are taken semiannually during the first and second halves of the year.

Table 4. Contamination Levels

	First Half	Second Half
2004	-	2
2005	1	8
2006	4	9
2007	3	3
2008	4	5
2009	4	10
2010	9	10

The table of seasonal indicators showing a relative change for each season compared to the previous one is given below.

Table 5. Seasonal Indicators

	First Half	Second Half
2005	-1	+1
2006	-1	+1
2007	-1	0
2008	+1	+1
2009	-1	+1
2010	-1	+1

To apply the Poisson binomial distribution, we need to calculate the probabilities of corresponding events. Assuming each outcome to be equally probable and assigning to each of them the probability of 1/10, we create a table of probabilities of the most frequently occurring indicator in the season for which the test is to be conducted – for the first half it will be a ‘decrease’ (-1), for the second half it will be an ‘increase’ (+1).

Table 6. Decrease and Increase Probabilities for Corresponding Seasons

	Decrease Probabilities (p^-)	Increase Probabilities (p^+)
2005	1/10	9/10
2006	7/10	6/10
2007	8/10	7/10
2008	2/10	6/10
2009	4/10	6/10
2010	9/10	1/10

Binomial probabilities for Poisson binomial distribution can be used to calculate corresponding P-values. For discrete distributions, P-values are the total probabilities of greater or equal than the observed number of successes, which are the analogues of the tail area of continuous distributions.

For the first half of the year, $P\{y \geq 5|\, p^-\} = P^-(5) + P^-(6)$. For the second half of the year, $P\{y \geq 5|\, p^+\} = P^+(5) + P^+(6)$. Using the corresponding values of the Poisson binomial distribution probabilities from Table 7 below, we find:

$$P\{y \geq 5|\, p^-\} = 0.062 + 0.004 \approx 0.07$$

$$P\{y \geq 5|\, p^+\} = 0.157 + 0.014 \approx 0.17$$

Table 7. Poisson Probabilities

k Successes	p^- Probabilities	Poisson Probabilities	p^+ Probabilities	Poisson Probabilities
0	-	0.003	-	0.002
1	1/10	0.042	9/10	0.028
2	7/10	0.213	6/10	0.139
3	8/10	0.405	7/10	0.316
4	2/10	0.271	6/10	0.345
5	4/10	0.062	6/10	0.157
6	9/10	0.004	1/10	0.014

As a result, we conclude that at the level of significance 0.2, our data are completely seasonal; whereas if hypothesis is tested at the level of significance 0.01, the data will show only partial seasonality with the coefficient of partial seasonality ½.

9 Conclusions

Introduced in this paper test for seasonality gives a simple way to measure this intuitive variable. Applicable to both numerical and categorical variables, this test provides an easy method to quantify and statistically analyze datasets for seasonality without linking it with any other test, be it trend analysis or other. The effect of the trend on the data can always be eliminated before this test is conducted. Seasonal trend analysis tests for a trend, but it does not indicate whether the seasonality is present or not. It accounts for it, but it does not give one a definite answer as of its existence. The above introduced statistical test for seasonality, on the other hand, can provide that answer for those who seek it.

References

[1] Hirsch, R.M. and Slack, J.R., 1984. A nonparametric trend test for seasonal data with serial dependence. *Water Resources Research*, *20*(6), pp.727-732.

[2] Hewitt, D., Milner, J., Csima, A. and Pakula, A., 1971. On Edwards' criterion of seasonality and a non-parametric alternative. *British journal of preventive & social medicine*, *25*(3), pp.174-176.

[3] Edwards, J.H., 1961. The recognition and estimation of cyclic trends. *Annals of human genetics*, *25*(1), pp.83-87.

[4] Chen, S.X. and Liu, J.S., 1997. Statistical applications of the Poisson-binomial and conditional Bernoulli distributions. *Statistica Sinica*, pp.875-892.