The classification of outliers psychology essay

The concern over the outliers is one of the challenge existed for at least several hundred years. Outliers are the observations those are apart from the bulk of data. Edgeworth (1887) wrote that discordant observations those appeared differently from other observations with which they are combined. Almost every data set has the outliers in different percentages. Grubbs (1969) said that an outlier is one that appears to deviate significantly from other values of data.

Sometimes outliers may not be noticed but most of the times they can change the entire statistical data analysis. As Peter (1990) explored those observations which do not follow the pattern of the majority of the data are called outliers. At the earlier stage of the data analysis, summary statistics such as the sample mean and variance, outliers can cause totally different conclusion. For example a hypothesis may or may not be rejected due to outliers. In fitting regression line outliers can significantly change the slope. The detection of outliers before analyzing the data analysis is not done then it may lead to model misspecification, biased parameter estimation and incorrect results. It is therefore important to identify the outliers prior to proceed further for analysis and modeling.

An observation (or subset of observations) that appears to be inconsistent with the rest of data set is called an outlier (Barnet1995). The exact definition of an outlier depends on the assumption regarding the data structure and the methods which are applied to detect the outliers.

Outliers are observations that appear to be unusual with respect to the rest of the data.

Classification of Outliers

Outliers are classified into one of four classes. First, an outlier may arise from procedural error, such as a data entry error or a mistake in coding. These outliers should be identified in the data cleaning stage, but if overlooked, they should be eliminated or recorded as missing values. Second, an outlier is the observation that occurs as the result of an extraordinary event, which is an explanation for the uniqueness of the observation. In this case the researcher must decide whether the extraordinary event should be represented in the sample. If so, the outlier should be retained in the analysis; if not, it should be deleted. Third, outliers may represent extraordinary observations for which the researcher has no explanation. Although these are the outliers most likely to be omitted, they may be retained if the researcher feels they represent a valid segment of the population. Finally, outliers may be observations that fall within the ordinary range of values on each of the variables but are unique in their combination of values across the variables. In these situations, the researcher should be very careful in analyzing why these observations are outliers. Only when specific evidence is available that discounts an outlier as a valid member of the population should it is deleted.

Outliers may be “ real” or “ ericaceous”. “ Real” outliers are observations whose actual values are very different from those observed for rest of the data and violate plausible relationships among variables. “ Erroneous” outliers are observations those are distorted due to misreporting errors in the data-collection process.

Data set either come from homogeneous groups or from heterogeneous groups, have different characteristics regarding a specific variable, outliers occurred by incorrect measurements including data entry errors or by coming from a different population than the rest of the data. If the measurements in correct, it represent a rare event.

Outliers are often caused by human error, such as errors in data collection, recording, or entry. Data from an interview can be recorded incorrectly, upon data entry. Outliers may cause from intentional or motivated misreporting.

Many times the outliers come when participants purposefully report incorrect data to experimenters or surveyors. A participant may make a conscious effort to sabotage the research or may be acting from other motives. Depending on the details of the research, one of two things can happen: inflation of all estimates, or production of outliers. If all subjects respond the same way, the distribution will shift upward, not generally causing outliers. However, if only a small sub sample of the group responds this way to the experimenter, or if multiple researchers conduct interviews, then outliers can be created.

Another cause of outliers is sampling error. It is possible that a few members of a sample were inadvertently drawn from a different population than the rest of the sample.

Outliers can be caused from standardization failure like the weak research methodology, unusual phenomena; faulty equipment is another common cause of outliers. By these causes data can be legitimately discarded if the researchers are not interested in studying the particular phenomenon in question.

One type of data entry error is implausible or impossible values, for they make no sense when considering the expected range of the data. An out-of-range value is often easy to identify since it will most likely lie well outside the bulk of the data.

Another common cause for the occurrence of outliers is the rare event. Extreme observations that for some correct reason are just fine, but do not fit within the typical range of other data values

There are many possible sources of outliers. Firstly, purely deterministic reasons those include: reading or measurement error, recording error and execution error.

Secondly, some reasons are pointed out by Beckman and cook (1983) they arrange the reasons of outliers into three broad categories. These are global model weaknesses, local model weaknesses and natural variability.

When we replace the present model with a new are revised model for the entire sample. Measurement of response variables are in the wrong scale is called Global model weakness.

Local model weaknesses are applied only on the outlying observations and not to the model as a whole. And Natural variability is the variation over the population rather than any weakness of the model. These reasons are uncontrollable and reflect the properties of distribution of a correct basic model describing the generation of the data.

The outliers occurs due to entry error or a mistake in coding should be identified in the data cleaning stage, but if overlooked, they should be eliminated or recorded as missing values.

1. 3 Problematic effects of outliers

Outliers of either type may influence on the results of statistical analysis, so they should be identified by using some suitable and reliable detection methods prior to performing data analysis. When potential outlier(s) is encountered, the first suspicion may be that such observations resulted from a mistake or other extraneous effect, and should be discarded. However, if the outlier in " real" it may be contained some important information about the underlying population of real values. Non judicious removal of observation that appears to be outliers may results in underestimation of the uncertainty present in the data.

In the presence of outliers, any statistical test based on sample means and variances can be distorted. There will be Bias or Distortion of estimates and it will give wrong results. The inflated sum of squares makes it unlikely and will partition sources of variation in the data into meaningful components.

The decision point of a significance test, p-value, is also distorted. Statistical significance is changed due to presence of a few or even one unusual data value.

The strong building of the statistical methods is based on weak legs of assumptions. Incorrect assumptions about the distribution of the data can also lead to the presence of suspected outliers. If the data may have a different structure than the researcher originally assumed, and long or shortterm trends may affect the data in unanticipated ways. Depending upon the goal of the research, the extreme values may or may not represent an aspect of the inherent variability of the data.

Outliers can represent a nuisance, error, or legitimate data. They can also be inspiration for inquiry. Before discarding outliers, researchers need to consider whether those data contain valuable information that may not necessarily relate to the intended study, but has importance in a more global sense.

-

The considerable effects of outliers are bias or distortion of Estimates, inflated sum of square and ended analysis of the entire data set at faulty conclusions. The key features of descriptive data analysis like the mean, variance and regression coefficient are highly affected by outliers.

1. 4 Aspects of outlier

There are two considerable aspects. The first aspect explains that, outliers have a negative effect on data analysis. Outliers generally cause to increase error variance and reduce the power of statistical tests. Outliers violate the assumption of normality. Outliers can seriously influence estimates.

The second aspect of outliers in that they are correct, and they may be provides useful information about data set. It the outliers are most information points they should not be automatically discarded without justification. In this case the analyses perform the analysis both with and without these outliers, and examine their specific influence on the results. If this influence is minor, then it may not matter whether or not they are omitted. If their influence is substantial, then it is probably best to present the results of both analysis, and simply alert the researcher to the fact that these points may be questionable.

The data set may contain outliers and influential observation. It is thus important for the data analyst to be able to identify such observation; if the data set contains a single outlier or influential observation then identification of such an observation in relatively simple. On the other hand, if the data set contain more than one outlier or influential observations the identification of such observation becomes more difficult. This is due to the marking and swamping effects. Masking occurs when an outlying subset goes undected because of the presence of adjacent subset of outliers. Swamping occurs when " good" observations are incorrectly identified as outliers because of the presence of other outliers.

An outlier is the observation that occurs as the result of an extraordinary event. In this case the researcher must decide about that event. If it represents the sample then that outlier should be retained in the analysis. If that event should not represent the sample it should be deleted.

Some time outliers may represent extraordinary observations but the researcher can not explain it. These types of the outlier may be omitted but sometime the may be retained if the researcher feels that they represent a valid segment of the population.

Both the detection and the suitable treatment of outliers are therefore important. In the present scenario of modern sciences where the messy data sets are generated, potentially troublesome outlier detection method(s) should be researched and presented at one place The main feathers of such identify criteria is that imperative to correctly identify outliers amongst large masses of data, so that experts can be alerted to the possibility of trouble and investigate the matter in detail.

Outliers can provide useful information about the process. An outlier can be created by a shift in the location (mean) or in the scale (variability) of the process. Though an observation in a particular sample might be a candidate as an outlier, the process might be shifted.

Numbers of treatments are taken in order to deal with outlier(s) involved studies.

Accommodation of outliers uses techniques to mitigate their harmful effects. One of its strength is that accommodation of outliers does not need to precede identification. These techniques can be used with prior information that outlier exist.

One very effective way to work with data is to use nonparametric methods which are robust in the presence of outliers. Nonparametric statistical method fit into this type of analyses and should be more widely applied to continuous or interval data than their current use.

Often the observed data set do not follow the any of the specified distribution then it is better to transform the data by applying appropriate transformation(s) so that data set could follow the specific distribution.

Only as a last resort should outliers be deleted, and then only if they are found to be errors they can not be corrected or lie so far outside the range of the remainder of the data that they distort statistical inferences

Our goal in this thesis is firstly to collect the outliers detection methods in univariate and bivariate/ multivariate studies followed the Gaussian and Non-Gaussian distributions and secondly to modify them accordingly.

1. 5 Univariate Outliers

In unvariate data sets, the study of outlier(s) is relatively simple but demands careful attention. Outliers are those values located distant from the bulk of the data and can often be revealed from simple plot of the data, such as scatter plot, stem-and-leaf plot, QQ-plot, etc.

Sometimes univariate outliers are not easy to identify as would appear at first sight. Barnet and Lewis (1994) indicate that an outlying observation, or outlier, is one that appears differently and deviate markedly from other members of the sample, in which it occur. A common rule for outlier identification might be to calculate the sample mean and standard deviation, and classify all those points as outliers which are at 2 or 3 standard deviations away from the mean. It is an unfortunate reality that the presence of two or more outliers could leave some or most of the outliers invisible to this method. If there is one or more distant outlier and one or more not so distant outlier in the same direction, the more distant outlier(s) could significantly shift the mean in that direction, and also increase the standard deviation, to such an extent that the lesser outlier(s) falls less than 2 or 3 standard deviations from the sample mean, and goes undetected. This is called the masking effect, and results in this particular method and all related methods being unsuitable for use as outlier identification techniques. It is illustrated with an example, borrowed from Becker and Gather [1999].

Consider a data set of 20 observations taken from an N (0, 1) distribution: -2. 21, -1. 84, -0. 95, -0. 91, -0. 36, -0. 19, -0. 11, -0. 10, 0. 18, 0. 30, 0. 31, 0. 43, 0. 51, 0. 64, 0. 67, 0. 72, 1. 22, 1. 35, 8. 1, 17. 6, where the latter two observations were originally 0. 81 and 1. 76, but the decimal points were entered at the wrong place. It seems clear that these 2 observations should be labeled as outliers; let us apply the above method. The mean of this data set is 1. 27 while the standard deviation is 4. 35. Two standard deviations from the mean, towards the right, would be 9. 97, while three standard deviations would be 14. 32. Both criteria regard the point, 8. 1, as expected with reasonable probability and do not consider it an outlier. Additionally, the three standard deviation boundary for detecting outliers seems rather extreme for an N (0, 1) dataset, surely a point would not have to be as large as 14. 32 to be classified as an outlier. The masking effect occurs quite commonly in practice and we conclude that outlier methods based on classical statistics are unsuitable for general use, particularly in situations requiring non-visual techniques such as multivariate data. It is worth noting, however, that if instead of the sample mean and standard deviation, robust estimates of location and scale were used (such as the sample median, and median absolute deviation, MAD), both outliers would be detected without difficulty.

1. 6 Multivariate Outliers

Multivariate outliers are the challenges that do not occur with univariate data sets. For instance, visual methods simply do not work in case of multivariate case studies. Even plotting the data in bivariate form with a systematic rotation of coordinate pairs will not help. It is possible (and occurs frequently in practice) that points which are outliers in bivariate space, are not outliers in either of the two univariate subsets. Generalization to higher dimensions leads to the fact that a multivariate outlier does not have to be an outlier in any of its univariate or bivariate coordinates, at least not without some kind of transformation

A successful method of identifying outliers in all multivariate situations would be ideal, but is unrealistic. By successful, we mean both highly sensitive, the ability to detect genuine outliers, and highly specific, the ability to not mistake regular points for outliers.