International Journal of Research in Economics and Social Sciences (IJRESS)
Available online at: http://euroasiapub.org
Vol. 7 Issue 2, February- 2017, pp. 111~118
ISSN(o): 2249-7382 | Impact Factor: 6.939 | Thomson Reuters Researcher ID: L-5236-2015

Exploring Data in Behavioral Science

Eshrat Ara
Research Scholar
Department of Psychology, University of Kashmir

Abstract

Data Analysis is the heart of the research process. Often the joy of data analysis prompts the researchers to rush for data analysis, fitting the statistical models to the data. This approach often leads to the process of GIGO (garbage in, garbage out) and the product comes out to be false hits and misses. Before fitting the statistical models to the data, the data ought to be fit for that statistical model, which needs extra efforts on part of the researcher to analyze the data for the preliminary analyses to understand the trends in the data. The present article describes some of those preliminary analyses procedures, so as to use the statistical tool in a fair way and to come with accurate results, avoiding misinterpretation and misuse of the data in the research.

Keywords: *Research Process, Preliminary Data Analysis, Testing Assumptions*

The research process begins with an observation of any phenomenon that we want to understand. From our observation we generate theories from which we can make predictions (hypotheses) that we can test. To test our predictions, we need data. First we collect some relevant data and then analyses those data. The analysis of the data may support our theory or give us cause to modify the theory. As such, the processes of data collection and analysis and generating theories are intrinsically linked: theories lead to data collection/analysis and data collection/analysis informs theories. When numbers are involved the research involves quantitative methods, but we can also generate and test theories by analyzing language (such as conversations, magazine articles, media broadcasts and so on). This involves qualitative methods.

So the final stage of the research process is to analyse the data we have collected. When the data are quantitative this involves both looking at the data graphically to see what the general trends in the data are, and also fitting statistical models to the data. Due to the joys of data analysis, researchers often rush their analysis, but rushing the analysis leads to the messy and incoherent outcomes. Before fitting any statistical model to the data the researchers should go for preliminary analyses to see the trends in the data. Some of the important preliminary analyses are explored here:

Data Screening

Data screening analyses examines the data for errors in data entry. Improbable values (falling outside the actual range of the measure or scale) can be detected and corrected. Frequencies, histograms, and minimum and maximum scores are used to check for errors in data entry. Means and standard deviations are checked for any improbable values. Values that fell outside the range of possible values can be compared to the original measures in order to see if the anomalous values are the result of data entry error or incomplete measures. If values fell outside the range of possible values due to data entry error, these can be corrected.

Missing Data Analysis

As researchers we strive to collect complete sets of data, but it is often the case that we have missing data. However, just because we have missed out on some data for a participant doesn't mean that we have to ignore the data we do have. We can *exclude cases listwise*; if participants have missing values for any variable, then they can be excluded from the whole analysis (but can lead to huge loss of data, sometimes). Another option is to *exclude cases on a pair-wise basis*, which means that if participants have score missing for a particular variable, then their data are excluded only from calculations involving the variable for which they have no score. However, if we do this many of our variables may not make sense, and we can end up with absurdities (can prove a bad option). Another possibility is to replace the missing score with the average score and include in the analysis. The problem with this choice is that it is likely to suppress the true value of the standard deviation (more importantly the SE). If the sample is large and the number of missing values is small then this is not a serious consideration. However, if there are many missing values, this choice is potentially dangerous because smaller standard errors are more likely to lead to significant results that are a product of the data replacement rather than a genuine effect. Then we can go for major operation, i.e., the final option is to use the Complex Missing Value Analysis Procedures.

Testing Assumptions

We have two types of statistical tests namely parametric and non-parametric tests. Parametric tests are based on the normal distribution. A parametric test is one that requires data from one of the large catalogue of distributions that statisticians have described and for data to be parametric certain assumptions must be true. If we use a parametric test when our data are not parametric then the results are likely to be inaccurate. Therefore, it is very important that we check the assumptions before deciding which statistical test is appropriate. Most parametric tests based on the normal distribution, have four basic assumptions that must be met for the test to be accurate.

Normally Distributed Data: This is a tricky and misunderstood assumption because it means different things in different contexts. In short, the rationale behind hypothesis testing relies on having something that is normally distributed (in some cases it's the sampling distribution, in others the errors in the model) and so if this assumption is not met then the logic behind hypothesis testing is flawed.

Homogeneity of Variance: This assumption means that the variances should be the same throughout the data. In designs in which we test several groups of participants this assumption means that each of these samples comes from populations with the same variance. In correlational designs, this assumption means that the variance of one variable should be stable at all levels of the other variable.

Interval Data: Data should be measured at least at the interval level.

Independence: This assumption, like that of normality, is different depending on the test we are using. In some cases it means that data from different participants are independent, which means that the behaviour of one participant does not influence the behaviour of another. In some cases, however, this assumption also relates to the errors in the model being uncorrelated (e.g., regression).

The Assumption of Normality

As defined earlier, this assumption is not easy to understand – it does not mean that our data are normally distributed. In many statistical tests (e.g. the t-test) we assume that the sampling distribution is normally distributed. This is a problem because we don't have access to this distribution – we can't simply look at its shape and see whether it is normally distributed. However, we know from the *central limit theorem* that if the sample data are approximately normal then the sampling distribution will be also. Therefore, people tend to look at their sample data to see if they are normally distributed. If so, then they assume that the sampling distribution (which is what actually matters) is also. We also know from the central limit theorem that in big samples the sampling distribution tends to be normal anyway – regardless of the shape of the data we actually collected (and remember that the sampling distribution will tend to be normal regardless of the population distribution in samples of 30 or more). As our sample gets bigger then, we can be more confident that the sampling distribution is normally distributed. The assumption of normality is also important in research using regression (or general linear models). General linear models, assume that errors in the model (basically, the deviations) are normally distributed. In both cases it might be useful to test for normality. We can look for normality visually, look at values that quantify aspects of a distribution (i.e. skew and kurtosis) and compare the distribution we have to a normal distribution to see if it is different (i.e., testing the significance).

Testing Normality

From a statistical perspective, the violation of normality makes the hypothesis tests more conservative (i.e., reduces the Type I error rate) and hence reduces the power of the statistical test. Assessment of the assumption of normality is typically identified using graphs or statistics. Skewness and kurtosis values, which measure symmetry and peakedness respectively, should be zero if the data is normally distributed. If the sample is large (100 or more) it is recommended that the skewness and kurtosis values be inspected to see how far they deviate from zero as opposed to calculating their significance. Graphs that tell us about the distribution of our data are histograms, P-P plots, box plots, density plots, stem & leaf diagrams, etc.

Frequency distributions are a useful way to look at the shape of a distribution (Fig. 1, 3, & 5). There is another useful graph that we can inspect to see if a distribution is normal called a P–P plot (probability–probability plot). This graph plots the cumulative probability of a variable

against the cumulative probability of a particular distribution (in this case a normal distribution) (Fig. 2, 4, & 6).

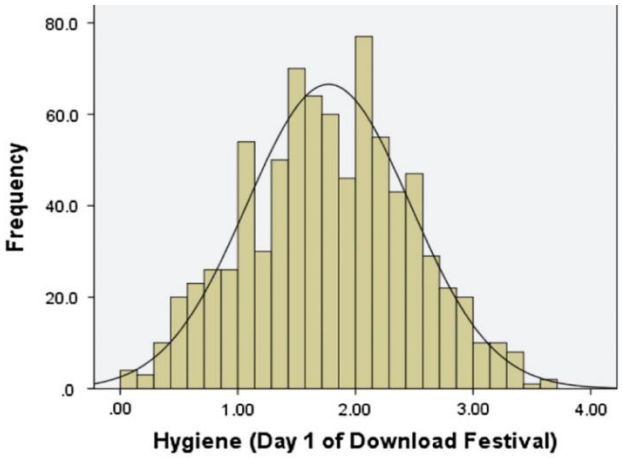

Figure 1

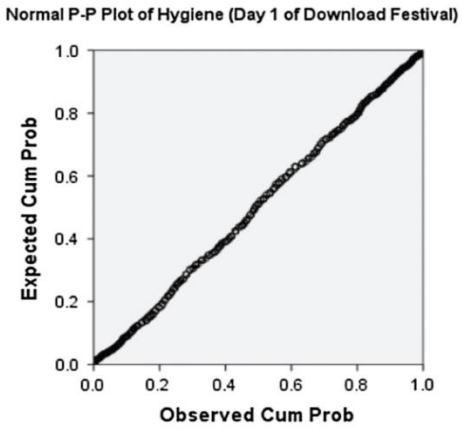

Figure 2

Figure 1 & 2: Graphs showing Normal Distribution

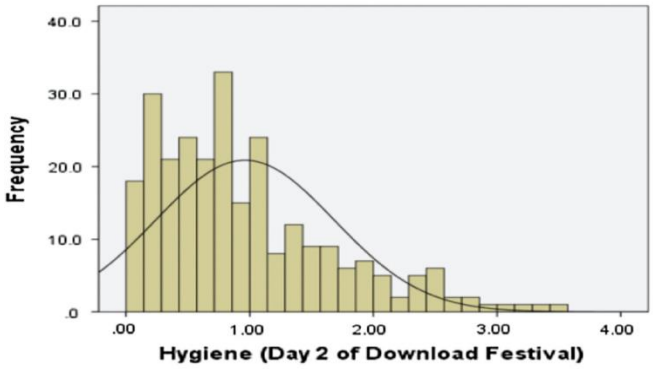

Figure 3

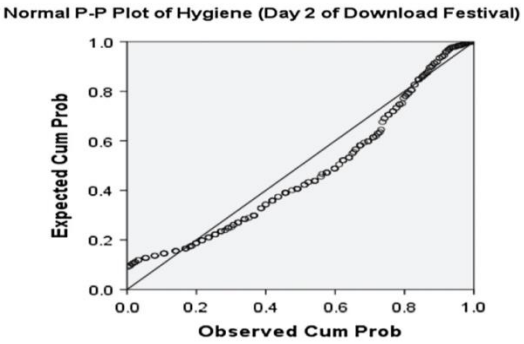

Figure 4

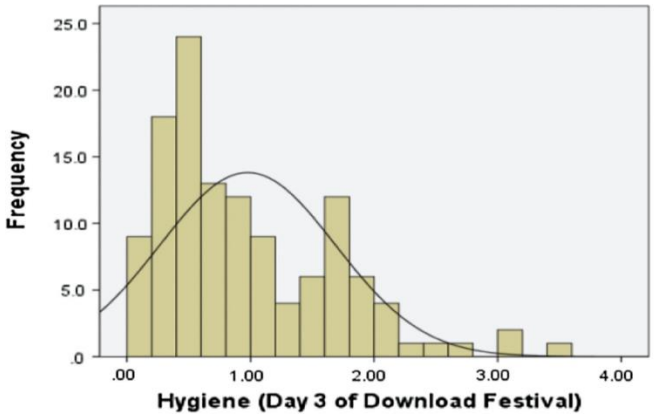

Figure5

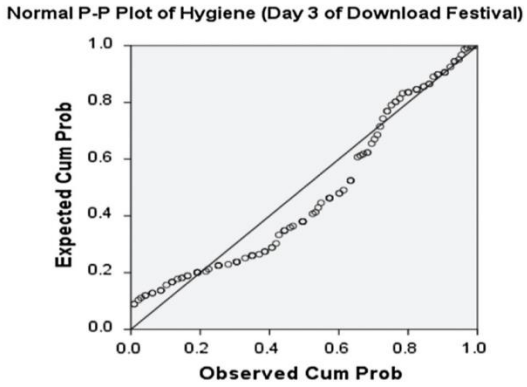

Figure 6

Figure 3, 4, 5, & 6: Graphs showing Skewed Distribution

Boxplots or box–whisker diagrams are really useful ways to display our data. At the centre of the plot is the median, which is surrounded by a box the top and bottom of which are the limits within which the middle 50% of observations fall (the inter-quartile range). Sticking out of the top and bottom of the box are two whiskers which extend to the most and least extreme scores respectively. Box plots are best tools for detecting the outliers in a dataset (see Fig. 7 & 8).

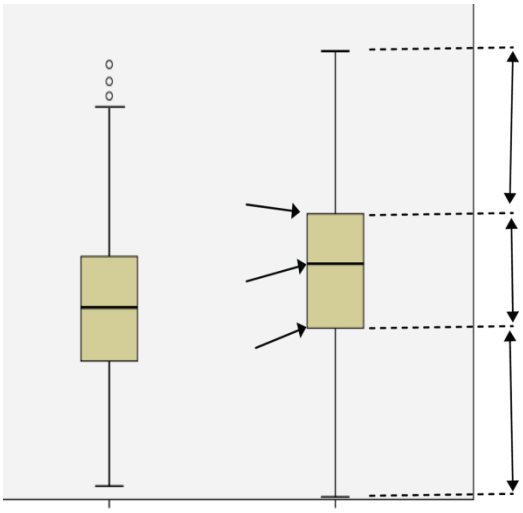

Figure 7

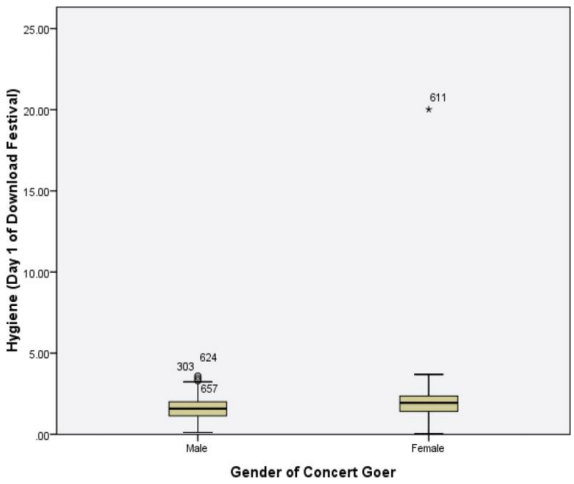

Figure 8

Exploring data through Descriptive Statistics

The graphs/charts provide a simple way to plot the frequency distribution of scores as a bar chart, a pie chart or a histogram. It is all very well to look at histograms, but they are subjective and open to abuse. Therefore, having inspected the distributions visually, we can move on to look at ways to quantify the shape of the distributions and to look for outliers. The statistics allows us several ways in which a distribution of scores can be described, such as measures of central tendency (*mean, mode, median*), measures of variability (*range, standard deviation, variance, quartile splits*), measures of shape (*kurtosis and skewness*). To describe the characteristics of the data we should select the mean, mode, median, standard deviation, variance and range. To check that a distribution of scores is normal, we need to look at the values of kurtosis and skewness.

Table 1: Showing an Example

Statistics

		Hygiene (Day 1 of Download Festival)	Hygiene (Day 2 of Download Festival)	Hygiene (Day 3 of Download Festival)
N	Valid	810	264	123
	Missing	0	546	687
Mean		1.7711	.9609	.9765
Std. Error of Mean		.02437	.04436	.06404
Median		1.7900	.7900	.7600
Mode		2.00	.23	.44[a]
Std. Deviation		.69354	.72078	.71028
Variance		.481	.520	.504
Skewness		-.004	1.095	1.033
Std. Error of Skewness		.086	.150	.218
Kurtosis		-.410	.822	.732
Std. Error of Kurtosis		.172	.299	.433
Range		3.67	3.44	3.39
Minimum		.02	.00	.02
Maximum		3.69	3.44	3.41
Percentiles	25	1.3050	.4100	.4400
	50	1.7900	.7900	.7600
	75	2.2300	1.3500	1.5500

a. Multiple modes exist. The smallest value is shown

Testing Assumption of Normality through statistics

Positive values of skewness indicate too many low scores in the distribution, whereas negative values indicate a build-up of high scores. Positive values of kurtosis indicate a pointy and heavy-tailed distribution, whereas negative values indicate a flat and light tailed distribution. The further the value is from zero, the more likely it is that the data are not normally distributed. We can convert these scores to z-scores by dividing by their standard error. If the resulting score is greater than 1.96 then it is significant ($p < .05$). Significance tests of skew and kurtosis should not be used in large samples (because they are likely to be significant even when skew and kurtosis are not too different from normal).

Significance Tests for Normality

The Kolmogorov–Smirnov (K–S) test can be used to see if a distribution of scores significantly differs from a normal distribution. If the K–S test is significant ($p < .05$) then the scores are significantly different from a normal distribution. Otherwise, scores are approximately normally distributed. The Shapiro–Wilk test does much the same thing, but it has more power to detect differences from normality (so, we might find this test is significant when the K–S test is not). In large samples these tests can be significant even when the scores are only slightly different from a normal distribution. Therefore, they should always be interpreted in conjunction with histograms, P–P or Q–Q plots, and the values of skewness and kurtosis.

Table 2: Showing an Example:

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Percentage on SPSS exam	.102	100	.012	.961	100	.005
Numeracy	.153	100	.000	.924	100	.000

a. Lilliefors Significance Correction

Converting Raw Scores into Standardized Scores

The data can also be converted into z-scores. Standardized scores exceeding 3.29, p = .001, may be potential outliers. These data points can be investigated further by examining the 5% trimmed mean. This value is generated when the top and bottom 5% of cases are removed. The original mean can then be compared to the trimmed mean. Accordingly decision is taken for their retention or removal. Remember outlier cases can't be removed, without strong reasons to believe that they are not indicative of the intended population.

Homogeneity of Variance

This assumption means that as we go through levels of one variable, the variance of the other should not change. If we have collected groups of data then this means that the variance of our outcome variable or variables should be the same in each of these groups. If we have collected continuous data (such as in correlational designs), this assumption means that the variance of one variable should be stable at all levels of the other variable.

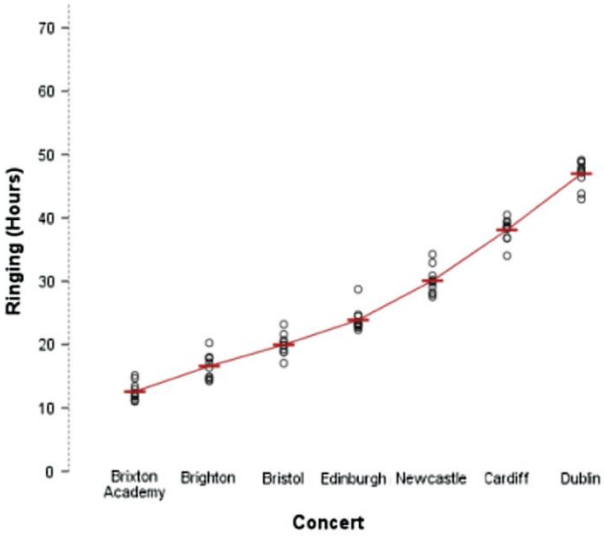

Figure 9

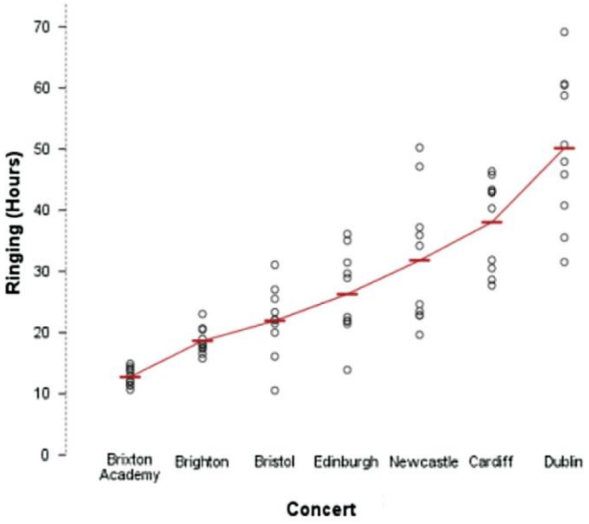

Figure 10

Figure 9 & 10: Graphs Illustrating Data with Homogeneous (left) & Heterogeneous (right) Variances

Testing Homogeneity of Variance

In short, Homogeneity of variance is the assumption that the spread of scores is roughly equal in different groups of cases, or more generally that the spread of scores is roughly equal at different points on the predictor variable. When comparing groups, this assumption can be tested with Levene's test and the Variance Ratio (Hartley's FMax). If Levene's test is significant (Sig. is less than .05) then the variances are significantly different in different groups. Otherwise, homogeneity of variance can be assumed. In large samples Levene's test can be significant even when group variances are not very different. Therefore, it should be interpreted in conjunction with the Variance Ratio. The Variance Ratio is the largest group variance divided by the smallest. This value needs to be smaller than the Hartley's critical values.

Table 3: Showing an Example:

Test of Homogeneity of Variance

		Levene Statistic	df1	df2	Sig.
Percentage on SPSS exam	Based on Mean	2.584	1	98	.111
	Based on Median	2.089	1	98	.152
	Based on Median and with adjusted df	2.089	1	94.024	.152
	Based on trimmed mean	2.523	1	98	.115
Numeracy	Based on Mean	7.368	1	98	.008
	Based on Median	5.366	1	98	.023
	Based on Median and with adjusted df	5.366	1	83.920	.023
	Based on trimmed mean	6.766	1	98	.011

Summary

A researcher needs to understand the basic principles of data analysis. Without a solid understanding, there is a risk of misinterpretation and misuse of data. If a researcher wants to interpret his/her research findings in a meaningful and accurate manner, then he/she must analyze critically the data that he/she has collected in his/her research. The preliminary analysis is mostly neglected in the data analyses by the researchers, which is very important before fitting the statistical models to the data. The researchers should analyze the data for assumptions before rushing for main analysis. The statistics is a play and can lead to accurate results if played fairly, otherwise can lead to false hits and misses.

Bibliography

- *Discovering Statistics Using SPSS* by Andy Field (2009), New Delhi, MR: Sage Publications India Pvt. Ltd.;
- *Introduction to Statistics in Psychology* by Dennis Howitt and Duncan Cramer (2011), Pearson Education Limited; *Tests Measurements and Research Methods in Behavioural Sciences* (3rd ed.) by A. K. Singh;
- *Research Methods in Psychology* (2003, 6th ed.) by John J Shaughnessy, Eugene B. Zechmeister, & Jeanne S. Zechmeister, McGraw-Hill;
- *Statistical Reasoning In Psychology and Education* (4th ed.) by Bruce M. King & Edward W. Minium, Wiley International.