Malaysian Family Physician 2006; Volume 1, Number 2&3

Online version: http://www.ejournal.afpm.org.my/

Research Notes

HOW TO ANALYSE YOUR RESEARCH DATA? ILLUSTRATIONS WITH HANDS-ON EXERCISES USING SPSS.

Loh Keng Yin[1] *MMed (FamMed, UKM)*, Teng Cheong Lieng[1] *MMed (FamMed, UM) FRACGP*, Wong Kam Cheong[2] *MBBS, MSc.*
[1]International Medical University, Malaysia; [2]University of Queensland, Australia

Address for correspondence: Dr Loh Keng Yin, Senior Lecturer, International Medical University, Jalan Rasah, 70300 Seremban, Malaysia. Tel: 06-7677798, Fax: 06-7677709, Email: kengyin_loh@imu.edu.my

Loh KY, Teng CL, Wong KC. How to analyse your research data. Illustrations with hands-on exercises using spss. Malaysian Family Physician. 2006;1(2&3):77-81

INTRODUCTION

Statistical analysis for a quantitative study is often perceived to be the most difficult step by a novice researcher. On the other hand, some researchers tend to over-analyse their research data in search of the illusive "significant" p-value. Some of these problems and pitfalls can be reduced if the researchers give some thoughts to their research objectives.[1]

Another issue that trouble the novice is how much statistical knowledge one needs to have. There is no straight answer to this question; we feel that the information provided in this article is probably the bare minimum needed by most, if not all, researchers embarking on a research project. What about performing your own statistical analysis using statistical software? Although ability to handle statistical software is desirable, it is not mandatory as it is now possible to outsource to people who can do this properly. The researcher should, however, be able to tell the statistician what analysis is needed and to interpret statistical results. Take note that the statistician cannot undo the errors in the data (e.g. inadequate research design, inappropriate definition of research variables, inaccurate measurement during data collection, or data entry errors) – hence great care must be exercised during these earlier steps of research process.

There are several statistical packages available to assist you in data analysis. SPSS (Statistical Package for Social Sciences) software is applied in the following example. You can download a free trial version from www.spss.com (prior registration necessary)

DESCRIPTIVE STATISTICS

We shall start by this example: You have conducted a survey of 160 diabetic patients in your clinic. The mean HbA_{1c} of these patients was 8.9% (SD=2.2, range 5.2-15.7). The gender breakdown is males 44.4%, and females 55.6%. The ethnicity breakdown is Malays 28.1%, Chinese 41.3%, and Indians 30.6%. Other summarised data are given in Table 1. You may request the data file from the Editor at tengcl@gmail.com

The SPSS commands for obtaining the above statistics for 'HbA_{1c}' are as follows:
From the menus choose:
- Analyze
- Descriptive Statistics
- Descriptive
- Select the variable 'hba1c'
- Click 'OK'

The SPSS commands for obtaining the above statistics for 'gender' and 'race' are as follows:
From the menus choose:
- Analyze
- Descriptive Statistics
- Frequencies
- Select the variable 'sex' and 'race'
- Click 'OK'

Table 1: Mean HbA1c by gender and ethnicity

Characteristics	HbA1c, % (SD)
Gender	
Male (n=71)	8.7 (1.9)
Female (n=89)	8.9 (2.3)
Ethnic group	
Malays (n=45)	9.6 (2.5)
Chinese (n=66)	8.4 (2.0)
Indians (n=49)	8.7 (1.9)
All patients (n=160)	8.9 (2.2)

The SPSS commands for obtaining the statistics 'HbA_{1c}' breakdown by 'Gender' in Table 1 are as follows:
From the menus choose:
- Analyze
- Reports
- Case Summaries
- Select 'hba1c' into the 'Variable'
- Select 'sex' into the 'Grouping Variable'
- Click on the icon 'Statistics' and select 'Number of Cases', 'Mean', and 'Standard Deviation'
- Uncheck these two boxes 'display cases' and 'Limit cases to first 100'
- Click 'OK'

The SPSS commands for obtaining the statistics 'HbA_{1c}' breakdown by 'Race' in Table 1 are as follows:
From the menus choose:
- Analyze
- Reports
- Case Summaries
- Select 'hba1c' into the 'Variable'
- Select 'race' into the 'Grouping Variable'
- Click on the icon 'Statistics' and select 'Number of Cases', 'Mean', and 'Standard Deviation'
- Uncheck these two boxes 'display cases' and 'Limit cases to first 100'
- Click 'OK'

Type of data

The first thing to take note is the type of data (or variables) you have collected.

- *Categorical data.* There are basically two kinds of data in this groups:
  - Nominal data (named categories), e.g. gender (male/female), ethnicity (Malay, Chinese, Indian), outcome (dead/alive), etc. The nominal data are summarised by percentages.
  - Ordinal data (ordered categories), e.g. tumour staging (Stage 1, 2, 3, 4), disease severity (mild, moderate, severe), Likert scale (5-point scale, 1-5), etc. The ordinal data are summarised by median value.
- *Continuous data.* Continuous data is sometime referred to as interval data. These data take the form of a range of number, and may or may not have decimals, e.g. age, HbA1c, weight, height, haemoglobin level, etc. The continuous data are summarised by mean and standard deviation (SD).

Another way of looking at the data is defining the dependent and independent variables:

- *Dependent variable* is the variable of interest
- *Independent variable* is the grouping variable

Let us say, you want to find out if HbA1c differ by gender or ethnicity. Then HbA1c is the dependent variable, and gender and ethnicity are independent variables.

Summarised data

We have seen already earlier different type of data are summarised differently. This summary of data, plus a graphical display of the data (e.g. in graph and scatter plot) is a very useful way of having a sense of your data before you embark on formal statistical analysis (the so-called "eye-balling the data").

Hypothesis testing, sample and population

One of the reasons for conducting the above study is that you have observed that diabetic patients of certain ethnic group appeared to have poorer diabetic control. Rather than stating that "Malay diabetic patients have poor diabetic control", we should state that "in the population of all diabetic patients, there is no difference in glycaemic control by ethnicity" (this is the so-called Null Hypothesis). By drawing a representative sample of diabetic patients from the population, you then seek to disprove the Null Hypothesis. This process of drawing conclusion on a population from a sample is called inferential statistics.

INFERENTIAL STATISTICS: PARAMETRIC TESTS

t-test

If you want to find out if HbA1c differ by gender, a statistical output can appear as follow: means HbA1c for males and females are 8.7% (SD=1.9) and 8.9% (SD=2.3) respectively, t= - 0.711, df=158, p=0.478. As HbA1c is a continuous variable (and presumably normally distributed), we use t-test for *two groups comparison of means* (males vs females). Since the p value is more than 0.05 (the conventional cut-off for statistical significance), we can interpret the result as no statistical significant difference or "no real difference in HbA1c in male and female diabetic patients".

The SPSS commands for obtaining the above t-test statistics are as follows:
From the menus choose:
- Analyze
- Compare Means
- Independent-Samples T Test
- Select 'hba1c' into the 'Test Variable'
- Select 'sex' into the 'Grouping Variable'
- Click on the icon 'Define Groups'
- Type in '1' in 'Group 1' and '2' in 'Group 2'
- Click 'OK'

Note: Group 1 is male, group 2 is female.

ANOVA

For *three or more groups comparison of means* (Malays, Chinese and Indians), we use ANOVA (F test): means for HbA1c in Malays, Chinese and Indians are 9.6% (SD=2.5), 8.4% (SD=2.0) and 8.7% (SD=1.9) respectively, F=4.524, p=0.012. In this case, there is statistical significant difference (p is less than 0.05) in the HbA1c among these three ethnic groups.

The SPSS commands for obtaining the above F-test statistics are as follows:
From the menus choose:
- Analyze
- Compare Means
- One-Way ANOVA
- Select 'hba1c' into the 'Dependent List'
- Select 'race' into the 'Factor'
- Click 'OK'

Note: One-Way ANOVA is chosen because there is 'one factor' (i.e. ethnicity) in the example. Although there are three ethnic groups in the study, it is considered one factor because all these groups can be grouped into one factor i.e. 'ethnicity' for analysis.

Pearson's correlation

If you want to find out if there is a linear relationship between the HbA1c and fasting blood glucose (FBS) in these diabetic patients, then the appropriate statistical test is Pearson's correlation, its value is denoted by "r" (Figure 1):

- Correlation value (r) is between 0 (no relationship whatsoever) to 1 (perfect straight line relationship).
- Correlation value (r) can be positive or negative depending on the direction of the relationship (e.g. one variable increases while the other decreases will have a negative correlation).
- In our analysis of HbA1c and FBS, we obtained r=0.539, p<0.001. There is a moderate amount of linear correlation which is significant as the P value is smaller than 0.05

The SPSS commands for obtaining the above Pearson correlation statistics are as follows:
From the menus choose:

- Analyze
- Correlate
- Bivariate
- Select 'hba1c' and 'fbs' into the 'Variables'
- Click 'OK'

Figure 1. Scatterplots
Figure 1A. (A) HbA1c vs FBS,

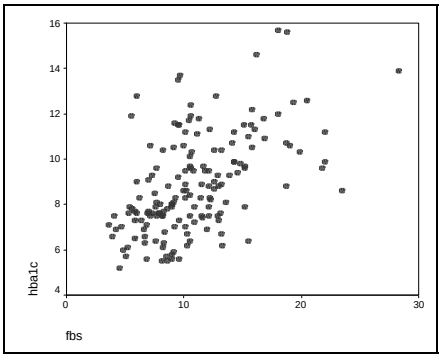
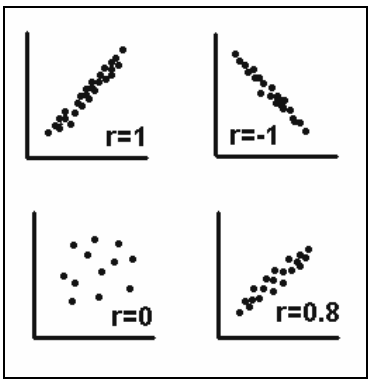


Figure 1B. Examples of scatterplots



Word of caution and further analysis

- Parametric tests should not be used if (1) the dependent variable does not follow a normal distribution, (2) the variance (square of SD) of dependent variables in the subgroups are very different from each other i.e. if the variances of HbA1c among Malay, Chinese, and Indian in the above example vary significantly, (3) the values of dependent variable in one subgroup is affecting the value of dependent variable in the other subgroup i.e. if the HbA1c of one ethnic group is affecting another group which seems to be unlikely in the above example. There are statistical methods for assessing if these assumptions for parametric test are violated. Parametric tests are considered robust test, i.e. minor deviation from the assumptions does not aversely affect the validity of the statistical test.
- In the case of the above ANOVA, if you want to tease out the difference further (say between Malays vs Chinese, or Chinese vs Indians), then you will need further statistical analysis, e.g. post-hoc multiple comparisons (Bonferroni method).
- Paired t-test should be used if you have two sets of data for the dependent variable that are linked , e.g. HbA1c for these 160 patients is measured twice (before and after diabetic education).
- Multiple regression is needed if you want to determine how one dependent variable (must be a continuous variable, e.g. HbA1c) can be predicted by a combination of several independent variables (all must be continuous variables, e.g. BMI, FBS, age, etc).

INFERENTIAL STATISTICS: NON-PARAMETRIC TESTS

Let us say, you have decided to regroup your HbA1c data into three categories: optimal (<7.0%), fair (7-9%), poor (>9%). The summarised data is shown in Table 2.

Table 2: Proportion of HbA1c categories by gender and ethnicity

Characteristics	HbA1c Optimal	Fair	Poor
Gender			
Male [n=71]	11 (15.5)	32 (45.1)	28 (39.4)
Female [n=89]	18 (20.2)	34 (38.2)	37 (41.6)
Ethnic group			
Malays [n=45]	6 (13.3)	17 (37.8)	22 (48.9)
Chinese [n=66]	14 (21.2)	28 (42.4)	24 (36.4)
Indians [n=49]	9 (18.4)	21 (42.9)	19 (38.8)
All patients	29 (18.1)	66 (41.3)	65 (40.6)

Note: The above results are percentages and the numbers in parenthesis are the standard deviations.

The SPSS commands for obtaining the above statistics are as follows:
From the menus choose:
- Analyze
- Descriptive Statistics
- Crosstabs
- Select 'sex' and 'race' into the 'Rows'
- Select 'hba1c category' into the 'Column'
- Click on the icon 'Cells', check the box 'Percentages Row', and click 'Continue'
- Click 'OK'

Chi-square test

Table 2 is in reality a cross-tabulation of gender (and ethnicity) with HbA1c categories. Chi-square test is used to compare frequencies (counts) in two or more groups.

- When you perform a chi-square test to look for association between gender and HbA1c categories, you will get this output: χ^2=0.984, df=2, p=0.611. Since p is >0.05, you may conclude that the HbA1c categories did not differ significantly among the two genders.
- When you perform a chi-square test to look for association between ethnicity and HbA1c categories, you will get this output: χ^2=2.196, df=4, p=0.700. Since p is >0.05, you may conclude that the HbA1c categories did not differ significant among the three ethnic groups.
- You may be wondering why there is no statistical significant difference in the chi-square test among the ethnic groups and HbA1c categories, but the ANOVA found statistical significant different in the mean HbA1c by ethnicity. By categorizing 160 HbA1c values into three groups only, you have actually lost quite a lot of useful information, thus producing a "non-significant" chi-square test.

The SPSS commands for obtaining the above Chi-square statistics are as follows:
From the menus choose:
- Analyze
- Descriptive Statistics
- Crosstabs
- Select 'sex' and 'race' into the 'Rows'
- Select 'hba1c category' into the 'Column'
- Click on the icon 'Statistics', check the box 'Chi Square', and click 'Continue'
- Click 'OK'

Other non-parametric tests

Non-parametric tests can be used in situations where the parametric tests are inappropriate, e.g. the dependent variable is not normally distributed (highly skewed data, ordinal data), sample size of study is small (<30), or when the assumptions of parametric tests may be violated (e.g. variances in subgroups highly unequal). When the independent variable has two groups, we use Mann-Whitney U test. When the independent variable has three or more groups, we use Kruskal-Wallis test. In both tests mentioned above, the "mean rank" of values in the dependent variables is compared instead of the arithmetic means (as in t-test or ANOVA).

Word of caution and further analysis

- Chi-square is a "big sample test", i.e. the sample size should be relatively large (to the extent that the "expected count" in each of the cells in the chi-square contingency table should be more than 5), otherwise this test becomes invalid. If your dependent or independent variables have more than two groups, you may attempt to combine some of the groups together to obtain a bigger sample size within a group. Nonetheless, if you have a study with small sample size, and both dependent and independent variables only have two groups, you may opt to perform Fisher Exact test.
- When the dependent variable is actually paired data, McNemar test is the correct test rather than Chi-square test. For example, when a group of patients with migraine are assessed twice (before and after intervention) for the presence of visual aura (present or absent).
- When the dependent variable has binary outcome (two possible responses, e.g. dead/alive), and you wish to determine the association with many other independent variables (may be continuous or categorical), logistic regression can be performed.

CONCLUSION

Choosing the correct statistical tests for your analysis depends on a good grasp of your research question (e.g. properly established research objectives), some understanding of the measurement you have made (is the variable continuous or categorical), the complexity of your analysis (one variable, 2 variables or multiple variables) and what the statistical test can or cannot do (its assumptions, its statistical output, etc). Table 3 is a summary of the commonly used statistical tests and their link to the characteristics of the variables.

REFERENCES

1. Khoo EM. Research questions and research objectives. *The Family Physician.* 2005;13(3):25-26 [PDF]

FURTHER READINGS

1. Dawson B and Trapp R G. Basic & clinical biostatistics. 4th edition. Mc Graw Hill, Boston, 2004.
2. Taylor G and Harris M. Medical statistics made easy. 1st edition. Taylor & Francis Group, UK, 2004.
3. Wong KC, Phua KL. Statistics Made Simple for Healthcare and Social Science Professionals and Students. University Putra Malaysia. 2006 (http://eprint.uq.edu.au/archive/00003913/).

Table 3: Choice of statistical tests and the characteristics of variables

Dependent variable	Independent variable	Parametric test	Non-parametric test
Categorical: 2 or more groups	Categorical: 2 or more groups	-	Chi-square Fisher Exact test*
Categorical: 2 group	Categorical or continuous	-	Logistic regression
Continuous: 1 group	Categorical: 2 groups	t-test	Mann-Whitney U test Wilcoxon rank sums test
Continuous: 1 group	Categorical: 3 or more groups	One-way ANOVA	Kruskal-Wallis test
Continuous: 1 group	Continuous: 1 group	Pearson's correlation Simple linear regression	Spearman's correlation
Continuous: 1 group	Continuous: many groups	Multiple regression	-
Analysis for paired data			
Categorical: 2 groups	Categorical: 2 groups	-	McNemar test
Continuous: 1 group	Categorical: 2 groups	Paired t-test	Mann-Whitney U test

*For 2x2 table only

Numerical illiteracy

"Some feel that the numerical illiteracy of doctors is being exploited by drug companies and by authors of NSF and NICE guidelines as well as by editors of mainstream journals…"
Peter Trewby, commenting on "Foundation of Evidence-based Medicine" (by Mios Jenicek). [PDF]
http://www.jrsm.org/cgi/reprint/96/10/515.pdf

Improving Medical Statistics and the Interpretation of Medical Studies
This website gives "examples of the misuse of statistics and inappropriate conclusions in the medical literature".

Rice Virtual Lab in Statistics
This website has an online statistics textbook, simulations and case studies.

Web Pages that Perform Statistical Calculations!
In this website, there are more than 600 links to online statistics books, tutorials, downloadable software, and related resources.

STATS - STeve's Attempt to Teach Statistics
Steve Simon's list of references and his simple way of explaining difficult statistical concepts.