Analysing Questionnaires using Minitab

(for **SPSS** queries contact -)
Graham.Currell@uwe.ac.uk

Structure

As a starting point it is useful to consider a basic questionnaire as containing **three** main sections:

- Questions to identify **definite groups** that each subject falls into, e.g. male/female, age range, educational background, etc
- Questions to describe respondents **initial/background opinions, knowledge**, e.g.
- How often do you watch TV shows like CSI – less than once a month, once a month, once a week, more than once a week?
- Questions related to **specific issues to be assessed**, e.g. How strong do you think this evidence is? Are these two fingerprints a match?

As a first example, we have the following extract from a data set

	Grouping		Background			Responses				
Subject	**G1**	**G2**	**B1**	**B2**	**B3**	**R1**	**R2**	**R3**	**D1**	**D2**
1	M	2	2	1	1	3	2	4	1	0
2	F	3	1	1	3	1	3	3	0	1
↓										
99	M	2	2	3	1	3	2	7	1	0
100	F	1	1	3	3	1	1	8	0	0
Levels:	**M/F**	**1-3**	**1-2**	**1-3**	**1-3**	**1-4**	**1-4**	**1-10**	**0/1**	**0/1**

Data Set 1

In this data set there are 100 records corresponding to 100 respondents or *subjects*.

Factual data to **group the respondents**:
G1 is nominal data that could describe a binary group, e.g male/female
G2 is ordinal data that could describe a progressive group, e.g. age range
Background questions to describe initial opinions, knowledge:
B1 (1 or 2), *B2* (1 to 3), *B3* (1 to 3)
Response questions:
R1, R2 scale data on Likert scales 1 to 4
R3 scale data on Likert scale 1 to 10
D1, D2 logical/digital data
(*D1, D2* are actually derived from values of *R1, R2* greater than or less than 2.5)

A general, but very important point, is that **the number of conclusions that can be drawn and the power of any tests depend critically on the amount of data collected**. It is important to get as many responses as possible.

Designing a Likert scale response question

Many questions in a questionnaire invite the respondent to choose a response on a symmetrical Likert scale.

For example:

Do you agree with a particular a statement? Give your answer on the
scale between Disagree strongly is -3 to Agree strongly is +3: -3 -2 -1 0 +1 +2 +3
or between Disagree strongly is 1 to Agree strongly is 4: 1 2 3 4
(it does not make any difference to the analysis if you code answers 1 to 7 instead of -3 to +3)

One factor to consider is **whether you wish to have a neutral**, '0', answer (neither agree or disagree) or whether you force the respondent to choose '-' or '+'. This might depend on the question. Forcing '-' or '+' could give you a binary answer with the option of using simple, yes/no, proportions, but is it 'right' to deny the neutral option?

It is also important to consider how many levels you should include in your scale.
If you have **too few**, then you might find that almost all of the respondents give the same answer, e.g. on a scale of 1 to 4, everyone might reply '3', making it impossible to do detailed analysis.
If you have **too many**, then using the frequency of responses in each category can become too small for useful analysis, e.g. in using chi-squared.
Ideally a pilot study would reveal the ranges of answers that could be expected in a final questionnaire, allowing you to design the questions more sensitively, but this is not possible for your simple project.
An alternative option is to use a larger number of levels, e.g. 1 to 9 and anticipate combining levels, depending on the responses obtained, see the 69 responses below:

Original levels:	**-4**	**-3**	**-2**	**-1**	**0**	**+1**	**+2**	**+3**	**+4**
Counts:	1	2	1	4	9	22	19	8	3

New levels:	**-1**	**0**	**+1**	**+2**	**+3**
Counts:	8	9	22	19	11

Note that the above data reduction procedure should only be used for chi-squared frequency analysis.

Useful analytical techniques

At a basic level we could record the basic statistics of the data in each column, e.g.

The mean response to R3 was 4.13 with a standard deviation of 0.22
The proportion of '1' responses to D1 was 0.64
The proportion of '1' responses to D2 was 0.60

This might tell us the how subjects, in general, responded to each question, but it would not give any information about any relationships hidden within the data, e.g.

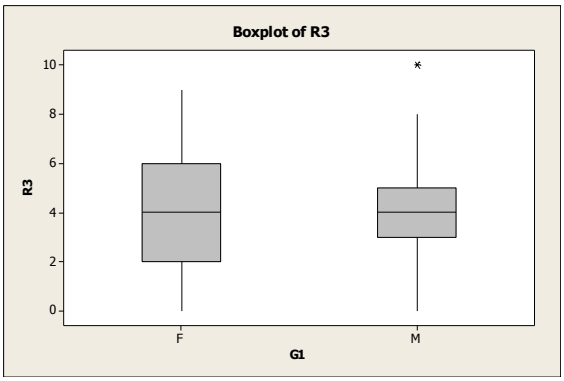
Does the mean response to R3 differ for different G2 groups?
Do the subjects who respond '1' for D1 also respond '1' for D2?

In the following sections we look at various techniques that you might find useful.

1. Using boxplots to explore raw data

Boxplots are an excellent way to explore your own data and to present raw data in your report.

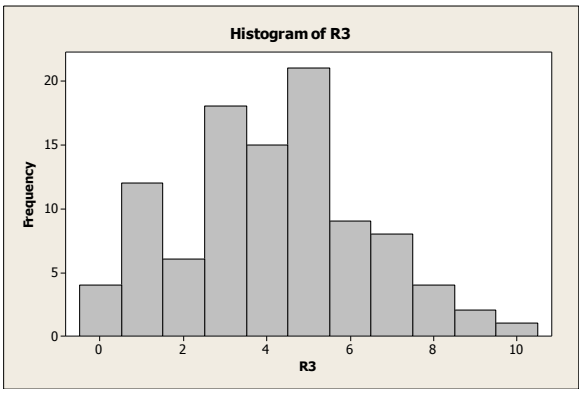
In the example, the boxplot of response R3, grouped separately for M and F responses to G1, suggests that there is no difference in the median value, but is there a difference in the spread or variance?
Is the outlier significant?



2. Using histograms to check data distributions

A histogram can be useful when you have several possible response levels to a question. It shows how the replies are distributed.
The example here shows a uni-modal response, with one main peak.

Sometimes you may see a bi-modal response with two different groups giving two different responses with two clear peaks.



3. Is there a difference in the **average responses**, R1, R2, R3 for the **two** values each of G1, B1?
Using **t-test** R1 and R3 shows significant differences in the mean responses for the two B1 groups with p = 0.009 and p = 0.004 respectively.
Mann-Whitney is preferable for data without normal distribution (Minitab needs data in two columns) giving p = 0.003 and p = 0.010 for the same tests as the t-tests above.

4. Is there a difference in the **average responses** R1, R2, R3 for the **three** values each of G2, B2, B3 etc ?
Using **One-Way ANOVA** to test whether the mean of R1 changes for different values of G2.

```
Source  DF      SS    MS     F      P
G2       2   10.80  5.40  4.47  0.014
Error   97  117.20  1.21
Total   99  128.00

                          Individual 95% CIs For Mean Based on
                          Pooled StDev
Level   N   Mean  StDev  -------+---------+---------+---------+--
1      40  2.500  1.240   (-------*--------)
2      40  3.200  0.966                    (--------*--------)
3      20  2.600  1.046  (-----------*-----------)
                          -------+---------+---------+---------+--
                               2.40      2.80      3.20      3.60
Pooled StDev = 1.099
```

Mean of R1 is different for levels 1 and 2 from G2.
Using **Kruskal-Wallis** also shows differences in R1 for different levels of G2, with p = 0.017

5. Are there any **interactions** between possible factors, i.e. does one factor have a different effect, depending on the value of another factor?
Using **GLM** to test whether R1 is affect by G2, B1 or an interaction between them

Analysis of Variance for R1, using Adjusted SS for Tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
G2	2	10.800	6.570	3.285	2.83	0.064
B1	1	6.883	8.091	8.091	6.97	0.010
G2*B1	2	1.212	1.212	0.606	0.52	0.595
Error	94	109.105	109.105	1.161		
Total	99	128.000				

These results give R1 dependent on B1 but not now significant for G2 (compare with 1-way ANOVA above) and shows no significance for any interaction between them.
Note: If you try to include too many factors or interactions, the available data is spread out and reduces the *power* of the *individual* analyses making a 'not significant' result more likely.
6. Is there a difference in the **variances of**, R1, R2, R3 for the **two** values each of G1, B1?
For example using the **2 variances test** for R3 for the two groups in G1

Method	DF1	DF2	Statistic	P-Value
F Test (normal)	54	44	1.46	0.194
Levene's Test (any continuous)	1	98	4.28	0.041

which is not significant for the F-test, but Levene's test suggests that there is a difference in the variance of F responses compared to those of M (see boxplot in 1. above)

7. Is there a difference in the **proportion** of '1' responses between D1 and D2?
Using the **2 proportions test**:

Fisher's exact test: P-Value = 0.662

There is no significant difference - see **9.** below for the measure of **agreement**.

8. Do the values of one answer **change in the same way** as those of another answer?

For example: You might have good reason, from **prior knowledge**, to believe that the values of R1 *change in a similar way* to the values of the respondents group B1.
A **test for correlation** between R1 and B1 produces *significant* correlation with p = 0.004
However be aware that correlation is **not** the same as 'cause and effect'.
It is possible that B1 does not directly influence R1 (or vice versa), but both might be influenced by a third factor (see next example below).

Alternatively, without any prior knowledge you might look **randomly** for any possible correlations between all of the answer.
A test for **bivariate correlations** between all of the answers gives:

Correlations: G2, B1, B2, B3, R1, R2, R3

	G2	B1	B2	B3	R1	R2
B1	-0.156					
	0.121					
B2	-0.024	-0.018				
	0.813	0.862				
B3	0.048	-0.416	-0.018			
	0.638	0.000	0.857			
R1	0.094	0.288	0.052	-0.306		
	0.350	0.004	0.607	0.002		
R2	0.084	0.126	-0.003	-0.239	0.799	
	0.407	0.211	0.977	0.017	0.000	
R3	-0.111	0.261	0.275	-0.095	0.030	-0.043
	0.272	0.009	0.006	0.348	0.765	0.668

Cell Contents: Pearson correlation
P-Value

In the above table the p-value is the lower value in the pair of values given for each combination of variables – check that the p-value for R1 and B1 is again given as 0.004.

Bonferroni correction:
You must be very careful with the above data because the chance of seeing a false significance (i.e. $p < 0.05$ *just by chance*) has increased considerably because you have calculated many ($n = 21$) p-values.
Where you are *randomly* looking for possible significant results amongst several p-values, the Bonferroni correction gives a new 95% confidence critical value:
Critical value = 0.05/n where n is the number of p-values calculated.

In this case the critical value would be 0.05/21 ~ 0.003
B1 and B3, R1 and B3, R1 and R2 show significant correlation, but not R1 and B1!

It is often possible that two variables might show correlation just because they are both correlated with the same third variable:
e.g. R1 is correlated with B3 and B3 is correlated with B1, so it is likely that R1 and B1 show some correlation (see previous example above).

SPSS can perform **partial** correlations where is possible to compensate for a third correlated variable when calculating the correlation between two variables.

9. How good is the **match, or agreement, between D1 and D2**

A test for correlation between D1 and D2 gives $p < 0.0005$, but it is sometimes important to know how strong the **agreement** is between D1 and D2. For example, D1 and D2 might be two assessments of the match of a fingerprint using different assessment protocols.

Using Attribute Agreement Analysis for D1, D2

```
Between Appraisers
Assessment Agreement
# Inspected  # Matched  Percent      95% CI
        100         86    86.00  (77.63, 92.13)

Cohen's Kappa Statistics
Response     Kappa   SE Kappa        Z  P(vs > 0)
0         0.703390  0.0996403  7.05929     0.0000
1         0.703390  0.0996403  7.05929     0.0000
```

K > 0.7 suggests a good match

There is a variety of other measures available to measure of the strength of association between questionnaire answers, which are grouped in the following table by the types of variables involved: nominal, ordinal and interval.

They are also divided into two types: symmetrical and directional. In a *directional* (or asymmetric) association we measure how a knowledge of one (independent) variable can be used to predict the variation of the other (dependent) variable. In a *symmetric* association, there is no sense of direction and we measure only the extent to which the two variables vary in similar ways.

For further information on these statistics, contact Graham.Currell@uwe.ac.uk

Variable pairs	Symmetric measures	Directional measures
Nominal / nominal	Phi, φ Cramer's V Kappa, κ	Lambda, λ
Ordinal / ordinal	Gamma, Γ Kendall's tau-b, τ Spearman's rho, ρ Coefficient of concordance	Somers' d
Interval / interval	Pearson's coefficient, r	Linear regression
Nominal / interval		Eta, η

Measures of association between questionnaire answers

10. Is the **distribution of answers** to one question related to (associated with) the way in which subjects answer another question?

For example: Is the choice that subjects make for D1 related to their answers to B3?

Using **Cross Tabulation** to count the numbers of respondents who fall into the 6 categories defined by 2 levels of D1 multiplied by 3 levels of B3, together with a **chi-squared** test for association:

Tabulated statistics: D1, B3

```
Rows: D1   Columns: B3
           1      2      3     All
0          6     12     18      36
       10.44  13.68  11.88   36.00
1         23     26     15      64
       18.56  24.32  21.12   64.00
All       29     38     33     100
       29.00  38.00  33.00  100.00
Cell Contents:      Count
                    Expected count
Pearson Chi-Square = 8.199, DF = 2, P-Value = 0.017
Likelihood Ratio Chi-Square = 8.242, DF = 2, P-Value = 0.016
```

This shows that subjects with B3=1 are more likely to choose D1=1 than those with B3=3.

11. Is an association between two answers **dependent on the level of a third** (or 4th) answer?

For example: Is an association between answers (as for D1 and B3 above) dependent on the group (e.g. G2) of the subject?

Using **cross tabulations** and **chi-squared**, ***layered*** by group G2:

Tabulated statistics: D1, B3, G2

Results for G2 = 1

```
Rows: D1   Columns: B3
           1       2       3     All
0          4       8       7      19
       4.275   8.550   6.175  19.000
1          5      10       6      21
       4.725   9.450   6.825  21.000
All        9      18      13      40
       9.000  18.000  13.000  40.000
Cell Contents:      Count
                    Expected count
Pearson Chi-Square = 0.311, DF = 2, P-Value = 0.856
Likelihood Ratio Chi-Square = 0.311, DF = 2, P-Value = 0.856
* NOTE * 2 cells with expected counts less than 5
```

Results for G2 = 2

```
Rows: D1   Columns: B3
           1      2      3    All
0          2      3      4      9
        3.60   3.15   2.25   9.00
1         14     11      6     31
       12.40  10.85   7.75  31.00
All       16     14     10     40
       16.00  14.00  10.00  40.00
Cell Contents:      Count
                    Expected count
Pearson Chi-Square = 2.683, DF = 2, P-Value = 0.261
Likelihood Ratio Chi-Square = 2.588, DF = 2, P-Value = 0.274
* NOTE * 3 cells with expected counts less than 5
```

Results for G2 = 3

```
Rows: D1   Columns: B3
           1      2       3     All
0          0      1       7       8
       1.600  2.400   4.000   8.000
1          4      5       3      12
       2.400  3.600   6.000  12.000
All        4      6      10      20
       4.000  6.000  10.000  20.000
Cell Contents:      Count
                    Expected count
Pearson Chi-Square = 7.778, DF = 2, P-Value = 0.020
Likelihood Ratio Chi-Square = 9.296, DF = 2, P-Value = 0.010
* NOTE * 5 cells with expected counts less than 5
```

It is only group G2 = 3 that shows a significant association (using 'likelihood' chi-squared) between D1 and B3 with the p-value less than 0.05/3 (using Bonferroni correction for 3 derived p-values – see Bonferroni correction below)
Note that as you try to get more information, the data is spread more thinly and a number of expected counts fall below 5.
You either need to restrict the categories that you create by the different levels in your answers, or you need to get more responses to the questionnaire.

12. What is the difference between finding an **association using chi-squared** or a **correlation** between different answers?

In most cases the choice of analysis is quite clear:

- Correlation is used for testing for a linear relationship between **two x-y variables**, and cannot be used with nominal data.
- Chi-squared is used for testing for differences in the spread of **numbers of data values** (frequencies) in different categories that can be described by nominal data.

However, it is also useful to illustrate the differences by considering a situation where it could be possible to use either correlation or chi-squared analysis
Two questions, Q1 and Q2, both have three possible answer levels, 1, 2, 3, and counting the results of 76 responses might give the two possible sets of results as Data A and Data B in the table below:
(For example the number 8 in the top middle of Data A shows that 8 respondents recorded '3' for Q2 and '2' for Q1)

3	2	8	10
2	8	20	8
1	10	8	2
Q2 / Q1	1	2	3

Data A

3	2	10	8
2	8	8	20
1	10	2	8
Q2 / Q1	1	2	3

Data B

Correlation:	$p < 0.0005$	$p = 0.132$
Chi-squared:	$p = 0.007$	$p = 0.007$

The only difference between data sets A and B is that the responses, '2' and '3' for Q1 have been reversed.

If we perform a correlation analysis for Data A, we get $p < 0.0005$, showing a high degree of correlation. We can see this in the data; a respondent who gives a high value for Q1 is also likely to give a high value for Q2 and vice versa. We can also imagine fitting a 'best fit' straight line going diagonally from bottom left to top right through the most data points.

However, in Data B, the cells with the most data points no longer follow a straight line; for example, a respondent who gives a '3' for Q1 is more likely to give a '2' than a '3' for Q2.
If we perform a correlation analysis for Data B, we get $p = 0.132$, which says that there is not enough evidence to claim that there is significant correlation. This agrees with our visual picture.

If we perform a chi-squared test we get the same significant differences in distribution with $p = 0.007$ for **both** data sets. The chi-squared test treats the answers as different nominal categories without any specific order, and consequently a change of order makes no difference to the result.

Chi-squared just tests for a **difference in distribution** of frequency values, whereas correlation looks specifically for a **continuing change** from one level to another.

13. Is it possible to **model relationships** between the different variables?

Use Stepwise regression to see if it is possible to predict values of R1 from the data in B1, B2, B3
Stepwise Regression: R1 versus G1, B1, B2, B3

```
Response is R1 on 3 predictors, with N = 100
Step            1      2
Constant    3.697  2.766
B3          -0.44  -0.32
T-Value     -3.18  -2.15
P-Value     0.002  0.034
B1                  0.44
T-Value             1.87
P-Value            0.065
```

Which gives an equation for R1: R1 = 2.77 - 0.32*B3 + 0.44*B1

R1 depends on both B3 and B1 in agreement with correlations.

Use Stepwise regression to see if it is possible to predict values of R3 as a function of other variables.

Stepwise Regression: R3 versus G1, B1, B2, B3, R1, R2

```
Response is R3 on 6 predictors, with N = 100
Step             1       2
Constant    2.5433  0.6638
B2            0.78    0.79
T-Value       2.83    2.98
P-Value      0.006   0.004
B1                    1.19
T-Value               2.83
P-Value              0.006
```

Which gives an equation for R3: R3 = 0.664 + 0.79*B2 + 1.19*B1

R3 depends on both B2 and B1 in agreement with correlations.

14 Is there any **clustering of answers or subjects hidden within our data**?

We can use another data set to illustrate this final set of analytical techniques:

Subject	G1	G2	B1	R1	R2	R3	R4
1	M	B	1	1	6	9	5
2	F	B	0	4	5	6	6
↓							
19	M	C	1	5	9	6	9
20	F	A	1	5	6	0	6
Levels:	**M/F**	**A/B/C**	**0/1**	**1 to 5**	**0 to 10**	**0 to 10**	**0 to 10**

Data Set 2

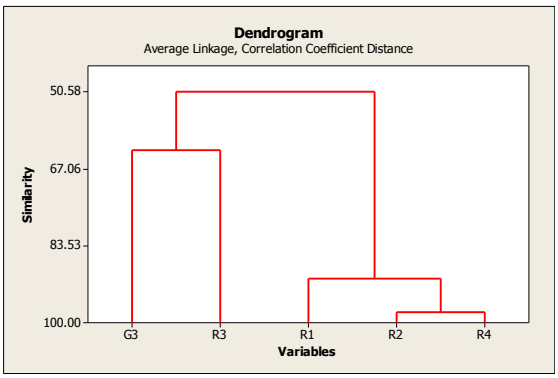
Data set 2 is a set of 20 responses, with two grouping questions, one background question and four response questions with the possible answer levels shown.
This data set has been created to illustrate the use of cluster and principal component analysis.

14a Is there are **similarities between the answers** to different questions?

The dendrogram in the diagram is based on correlation and shows the similarity between different variables (questionnaire answers).

R3 and R4 are very similar to each other and show some similarity to R1.
There is very little similarity with the other variables.

R3, R4 and R1 form a variables cluster.

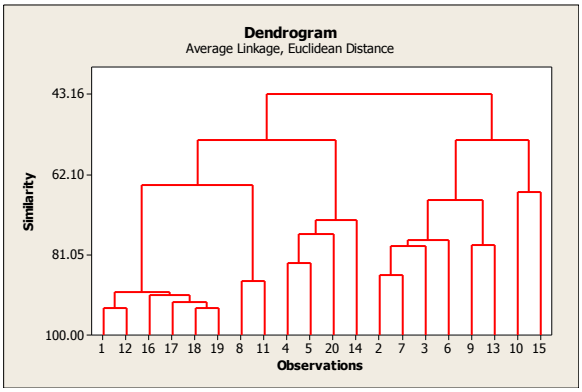


14b Is there any **clustering of subjects** showing similar sets of answers?

The dendrogram in the diagram is based on shows the similarity between different subjects responding to the questionnaire.

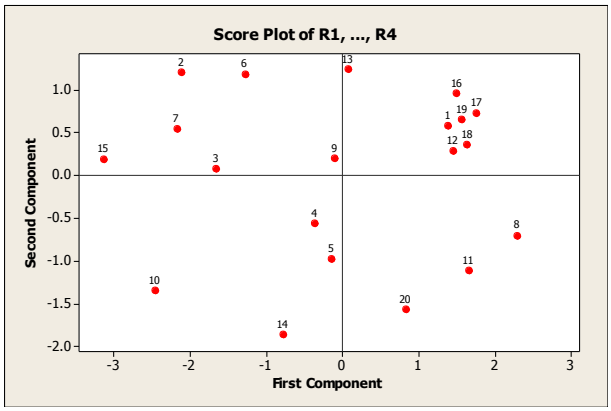
The clustering is not strong, but does show a weak cluster with subjects, 1, 12, 16, 17, 18 and 19.

We look for a long tail leading to several branches near the baseline.

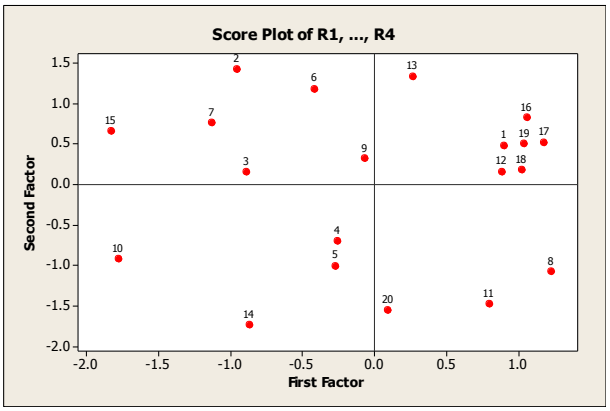


14c Can we **model the clustering of subjects**?

Using Principal Component Analysis and Factor Analysis it is possible to describe the multiple responses using just two main components or factors.
Every subject is then plotted on a two-dimensional plot using these two components or factors:



Principal Component Analysis



Factor Analysis with Quartimax rotation

Both diagrams show our cluster of subjects, 1, 12, 16, 17, 18 and 19, grouped together in the top right hand quadrants.

14d Using **Cluster analysis techniques**

It is important to realise that:

- The cluster analysis techniques described above are typically used to explore relationships, looking for hidden groupings within large amounts of data. This can then help the planning of future research.
- These techniques work best with large amounts of data. Note that in our example, there are responses over the full 1-5 and 0-10 ranges which gives the analysis enough information to be able to distinguish some groupings.

Although it is probable that cluster analysis will not be particularly appropriate for your project, it might be useful for you to demonstrate that you are aware of the technique even if you do not end up with startling discoveries!