

CME Article

Biostatistics 307.

Conjoint analysis and canonical correlation

Y H Chan



Conjoint Analysis is used to understand how a person's preference for a product/service is being governed by the attributes of that product/service. A **Utility score** is derived from the responses of each individual which measures the unique subjective judgment of preference (this includes both the tangible and intangible features of that product/service). **Part-worth utility scores** are used to determine the importance of each attribute for that product/service.

For example, assuming we are interested to develop XYZ Health Service in which three attributes are considered to be important; namely: hospital (A or B), cost (\$X or \$Y) and doctor (non-consultant or consultant). For three attributes with two levels each, we have $(2 \times 2 \times 2) =$ eight combinations (Table I).

Table I. Eight combinations of XYZ Health Services.

Combination	Hospital	Cost (\$)	Doctor
1	A	X	Non-consultant
2	A	X	Consultant
3	A	Y	Non-consultant
4	A	Y	Consultant
5	B	X	Non-consultant
6	B	X	Consultant
7	B	Y	Non-consultant
8	B	Y	Consultant

Participants are then asked to rank-score, 1 to 8 (8 being the best ranked) or to rate on a preference scale of 1 to 10 (10 being the most preferred) for each combination. Let us use the former rank-score as an illustration. Table II shows the rank-score for each combination of two respondents.

Table II. Rank-score of two respondents.

Com- bination	Hospital	Cost (\$)	Doctor	Subject 1	Subject 2
1	A	X	Consultant	8	8
2	A	X	Non-consultant	7	7
3	A	Y	Consultant	4	6
4	A	Y	Non-consultant	3	5
5	B	X	Consultant	6	2
6	B	X	Non-consultant	5	4
7	B	Y	Consultant	2	1
8	B	Y	Non-consultant	1	3

How does Conjoint analysis "weigh" the preference of each individual? With rank-scores from 1 to 8, the mean rank-score will be $4.5 = (1 + 2 + \dots + 8)/8$. Table III shows the average ranks and deviations for Subject 1.

Table III. Average ranks and deviations for Subject 1.

Attribute level	Ranks across combinations	Average rank of level	Deviations from mean rank-score
Hospital			
A	8, 7, 4, 3	5.5	1.0
B	6, 5, 2, 1	3.5	-1.0
Cost			
\$X	8, 7, 6, 5	6.5	2.0
\$Y	4, 3, 2, 1	2.5	-2.0
Doctor			
Consultant	8, 6, 4, 2	5.0	0.5
Non-consultant	7, 5, 3, 1	4.0	-0.5

The part-worth Utility score for each attribute is then calculated as follows (Table IV shows resultant values):

Step 1. Square the Deviations and sum across all levels (=10.5).

Step 2. Calculate the Standardising Value = total number of levels (= 2+2+2)/Sum of Squared Deviations = $6/10.5 = 0.571$.

Yong Loo Lin School
of Medicine
National University
of Singapore
Block MD11
Clinical Research
Centre #02-02
10 Medical Drive
Singapore 117597

Y H Chan, PhD
Head
Biostatistics Unit

Correspondence to:
Dr Y H Chan
Tel: (65) 6874 3698
Fax: (65) 6778 5743
Email: medcyh@
nus.edu.sg

- Step 3. Calculate the Standardised Deviation = Standardising Value X Squared Deviations from mean rank-score.
- Step 4. Calculate the part-worth Utility = Square root of the absolute value of Standardised Deviation.
- Step 5. Calculate Range of Utility = Maximum – Minimum of the Utility for each Attribute level. Sum all Ranges of Utility (=5.29).
- Step 6. Calculate the Attribute Importance = Range of Utility of the Attribute/Sum of all Ranges of Utility.

The Utility score of Subject 1 for this XYZ Health Service is 5.29. This Subject values Cost as the most influencing attribute for his using this service, followed by type of hospital but it does not matter to him whether he is being seen by a consultant or not. Within each attribute comparison, he prefers Hospital A, at \$X and (if possible) seen by a consultant. Likewise, we can determine the scores of Subject 2 (Table V).

For Subject 2, the most influencing attribute is Hospital (preferable Hospital A), followed by Cost \$X and a non-consultant. His overall utility for this service is 4.83 which is lower than that of Subject 1 (in comparison).

We can aggregate the results of both Subjects for each attribute’s contribution to the XYZ Health Service (Table VI). Subjects using this Service are equally (about similar) influenced by Hospital and Cost. One extension of conjoint analysis is to group subjects with similar part-worth utilities or importance values, and use logistic regression⁽¹⁾ or discriminant analysis⁽²⁾ for further investigations.

Table VI. Aggregated results of Subjects 1 & 2.

Attribute	Mean	SD	Min	Max
Hospital	47.62	26.93	28.57	66.67
Cost	41.07	22.72	25.00	57.14
Doctor	11.30	4.2	8.33	14.28

Table IV. Calculating the utility score for each attribute of Subject 1.

Attribute level	Deviations from mean rank-score	Step 1 Squared deviations	Step 3 Standardised deviation	Step 4 Part-worth utility	Step 5 Range of utility	Step 6 Attribute importance
Hospital					1.512	28.6%
A	1.0	1.0	0.571	0.756		
B	-1.0	1.0	-0.571	-0.756		
Cost					3.022	57.1%
\$X	2.0	4.0	2.284	1.511		
\$Y	-2.0	4.0	-2.284	-1.511		
Doctor					0.756	14.3%
Consultant	0.5	0.25	0.143	0.378		
Non-consultant	-0.5	0.25	-0.143	-0.378		
Sum = 10.5					Sum = 5.29	

Table V. Utility score of Subject 2.

Attribute level	Deviations from mean rank-score	Step 1 Squared deviations	Step 3 Standardised deviation	Step 4 Part-worth utility	Step 5 Range of utility	Step 6 Attribute importance
Hospital					3.224	66.7%
A	2.0	4.0	2.60	1.612		
B	-2.0	4.0	-2.60	-1.612		
Cost					1.208	25.0%
\$X	0.75	0.5625	0.365	0.604		
\$Y	-0.75	0.5625	-0.365	-0.604		
Doctor					0.400	8.3%
Consultant	-0.25	0.0625	-0.04	-0.20		
Non-consultant	0.25	0.0625	0.04	0.20		
Sum = 9.25					Sum = 4.83	

NB: Standardising value = 6/9.25 = 0.649

We can perform the above analysis (with more subjects, definitely) using SPSS categories. Alternatively, the PROC TRANSREG in SAS could be used. The SAS codes for the above example are given in Appendix I (The author does not have the SPSS categories licence to generate the above results).

The above design, Table I, is balanced where each level of each attribute (= 4), each combination between two levels of two attributes (= 2) and each 3-way combinations (= 1) appear equal number of times. For example A, B, \$X, \$Y, consultant and non-consultant each appears four times in the table and each 2-way (A X \$X, A X \$Y, etc) combination appears two times. This is termed as a 100% efficient and uncorrelated design.

What happens if we have four attributes A-D with A having three levels and B to D having two levels each. In total, we have $3 \times 2 \times 2 \times 2 = 54$ combinations – definitely not a feasible study! In this case, what is the smallest subset of the 54 combinations to be generated to have a 100% efficient and uncorrelated design? This subset must be divisible by four (= 2X2, combination of any 2 level attributes) and six (= 2X3, combination of a 2-level and the 3-level attributes) and the smallest is 12 (the least common multiple [LCM]). The SAS codes (Appendix 2), will help to generate a balanced uncorrelated design.

The above conjoint analysis (Traditional method) makes the assumption that the total Utility score is simply the addition of each part-worth utility scores. A choice-based conjoint analysis needs to be performed when one believes that there are interaction effects in which the total Utility score is derived. For example, a respondent usually assigns a higher score for a consultant. However, a bad experience with Hospital B consultants make him assign higher scores for non-consultants in this case. Here, we have a hospital-doctor interaction.

We have merely discussed the main use of conjoint analysis, readers are encouraged to get further details on this growing technique of analysis from any standard statistical text. A biostatistician would be of assistance to do the full analysis for the checking of the conjoint model and assessing the overall fit.

CANONICAL CORRELATION

The bivariate correlation between two quantitative variables and the relationship of a quantitative dependent with many independents were discussed in Biostatistics 104⁽³⁾ and 201⁽⁴⁾, respectively. Consider the situation where the routine laboratory blood test consisting of the renal panel (urea, creatinine,

sodium, potassium, glucose) and liver panel (total bilirubin, alkaline phosphatase (ALP), aspartate transaminase (AST), alkaline transaminase (ALT), lactate dehydrogenase (LDH)) of 400 subjects were collected. Table VII shows the bivariate correlation between the two panels.

Table VII. Bivariate correlations between renal and liver panel.

	ALT	AST	Total bilirubin	ALP	LDH
Urea	-0.0392	0.1306	0.0305	0.1510	0.1153
Sodium	-0.0315	-0.2277	0.0990	-0.0551	-0.0812
Potassium	-0.1227	0.3499	0.0104	-0.0676	0.6018
Creatinine	-0.0115	0.1039	-0.0590	0.2379	0.0660
Glucose	0.0134	0.0224	-0.0831	0.1660	0.0426

We can extend the analysis to a “multivariate” nature by performing a Canonical Correlation where one is interested to determine the optimised linear relationship between two sets of variables. We can perform the analysis with SPSS categories CANCELL procedure (sorry – no licence). The SAS codes are:

```
proc cancell all;
var urea sodium potassium creatinine glucose;
with alt ast bilirubin_total alp ldh;
run;
```

Table VIII is the summarised results from the large amount of output listings in a canonical correlation analysis.

For this example, five (given by the smaller number of variables in the two sets) Canonical Correlations will be produced displayed in the order of intercorrelational importance.

Each canonical correlation measures the bivariate correlation between two Canonical Variates (linear composites of the two sets of variables) whose **Cross-loadings** (other loadings are also given in the output listings - Raw weights, Standardised or Canonical weights and Canonical loadings, leave them alone!). Each cross-loading gives the correlation of each variable to the opposite canonical variate. For Canonical Variate 1, Potassium has the highest effect (0.6333) on the renal panel as a whole whereas LDH (0.6008) and AST (0.3367) contribute the most weightage to the liver panel. The rest of the loadings for the canonical variates 2 to 4 are not worth mentioning. The **Redundancy Index** gives the amount of variance explained by one canonical variate to the other. For example, the renal panel explains

Table VIII. Canonical correlations between renal and liver panels.

Canonical correlation	0.6409 (p<0.001)	0.3322 (p<0.001)	0.3002 (p<0.001)	0.0817 (p=0.5877)	0.0220 (p=0.6617)
Renal panel canonical variate					
	1	2	3	4	5
Urea	0.0976	0.1911	0.0423	0.0637	-0.0029
Sodium	0.0100	-0.2303	0.1676	0.0249	-0.0075
Potassium	0.6333	-0.0057	-0.0394	-0.0054	0.0009
Creatinine	0.0504	0.2411	0.1383	0.0172	-0.0101
Glucose	0.0549	0.1045	0.1543	-0.0040	0.0175
Redundancy index	0.0833	0.0317	0.0149	0.0010	0.0001
Liver panel canonical variate					
	1	2	3	4	5
ALT	-0.1268	0.0191	0.0025	-0.0368	0.0192
AST	0.3367	0.2073	-0.1451	-0.0247	0.0022
Bilirubin	-0.0092	0.0209	-0.1788	0.0591	0.0076
ALP	-0.0406	0.2219	0.1800	0.0069	0.0094
LDH	0.6008	0.0720	-0.0658	-0.0104	0.0022
Redundancy index	0.0984	0.0196	0.0180	0.0011	0.0001

Appendix 1. SAS codes for conjoint analysis.

```

* Creating the value labels for each Attribute level;
proc format;
value hosp
  1 = 'Hospital A'
  2 = 'Hospital B';
value cost
  1 = 'Cost $X'
  2 = 'Cost $Y';
value doctor
  1 = 'Consultant'
  2 = 'Non-Consultant';
run;

* To generate Table I;

data service;
format hosp hosp.cost cost.doctor doctor.;
input hosp cost doctor subj1 subj2;
datalines;
1 1 1 8 8
1 1 2 7 7
1 2 1 4 6
1 2 2 3 5
2 1 1 6 2
2 1 2 5 4
2 2 1 2 1
2 2 2 1 3
;
proc print;
run;

* Performing the Conjoint Analysis - giving Tables IV & V;

proc transreg utilities outtest = allfile separators='';
model identity(Subj1-Subj2) = class (hosp cost doctor/ zero = sum);
run;

data all;
set allfile (keep=_depvar_ Importance Label);
if n(Importance);
label = substr(label, 1, index(label, ' '));
run;

* Aggregated results - giving Table VI;

proc transpose out = allfile1 (drop=_);
by _depvar_;
id label;
run;

proc means;
run;

```

only 8.33% of the liver panel, whereas the liver panel explains only 9.84% of the renal panel.

In conclusion, for the canonical correlation analysis to be of value, firstly the Redundancy Index should be at least in the 70% mark, followed by a significant canonical correlation value of 0.6 and above.

For our next article, we shall discuss Structural Equation Modeling in Biostatistics 308.

REFERENCES

1. Chan YH. Biostatistics 202. Logistic regression analysis. Singapore Med J 2004; 45:149-53.
2. Chan YH. Biostatistics 303. Discriminant analysis. Singapore Med J 2005; 46:54-61.
3. Chan YH. Biostatistics 104. Correlational analysis. Singapore Med J 2003; 44:614-9.
4. Chan YH. Biostatistics 201. Linear regression analysis. Singapore Med J 2004; 45:55-61.

Appendix 2. SAS codes for generation of a balanced uncorrelated design.

```

%mktx (2 2 2 3, n = 12)
%mkltab (vars=A B C D, out = sasuser.design)
%mkteval;

proc print data=sasuser.design;
run;

```

SINGAPORE MEDICAL COUNCIL CATEGORY 3B CME PROGRAMME

Multiple Choice Questions (Code SMJ 200510A)

Questions 1-3: Consider the response of a 3rd subject for our Conjoint analysis example given in the following table:

Combination	Hospital	Cost (\$)	Doctor	Subject 1
1	A	X	Consultant	1
2	A	X	Non-consultant	2
3	A	Y	Consultant	3
4	A	Y	Non-consultant	4
5	B	X	Consultant	7
6	B	X	Non-consultant	5
7	B	Y	Consultant	8
8	B	Y	Non-consultant	6

- | | True | False |
|---|--------------------------|--------------------------|
| Question 1. Are the following statements true or false? | | |
| (a) The Sum Squared deviations = 10.5 | <input type="checkbox"/> | <input type="checkbox"/> |
| (b) The Sum Squared deviations = 9.25 | <input type="checkbox"/> | <input type="checkbox"/> |
| (c) Sum of all Ranges of Utility = 4.83 | <input type="checkbox"/> | <input type="checkbox"/> |
| (d) Sum of all Ranges of Utility = 4.83 | <input type="checkbox"/> | <input type="checkbox"/> |
| Question 2. Are the following statements true or false? | | |
| (a) Most important attribute is Hospital. | <input type="checkbox"/> | <input type="checkbox"/> |
| (b) Most important attribute is Cost. | <input type="checkbox"/> | <input type="checkbox"/> |
| (c) Most important attribute is Doctor. | <input type="checkbox"/> | <input type="checkbox"/> |
| (d) Least important attribute is Hospital. | <input type="checkbox"/> | <input type="checkbox"/> |
| Question 3. Are the following statements true or false? | | |
| (a) Least important attribute is Cost. | <input type="checkbox"/> | <input type="checkbox"/> |
| (b) Least important attribute is Doctor. | <input type="checkbox"/> | <input type="checkbox"/> |
| (c) Preferred Hospital B over A. | <input type="checkbox"/> | <input type="checkbox"/> |
| (d) Preferred Cost X over Y. | <input type="checkbox"/> | <input type="checkbox"/> |
| Question 4. Given that we have four attributes A-D with A & B having three levels and C & D having two levels each, what is the smallest subset required for a 100% efficient and uncorrelated design? | | |
| (a) 18. | <input type="checkbox"/> | <input type="checkbox"/> |
| (b) 24. | <input type="checkbox"/> | <input type="checkbox"/> |
| (c) 30. | <input type="checkbox"/> | <input type="checkbox"/> |
| (d) 36. | <input type="checkbox"/> | <input type="checkbox"/> |
| Question 5. The following indicates that a canonical correlation has interpretive value: | | |
| (a) Redundancy index of at least 0.7. | <input type="checkbox"/> | <input type="checkbox"/> |
| (b) Canonical correlation of at least 0.6. | <input type="checkbox"/> | <input type="checkbox"/> |
| (c) P-value for Canonical correlation is significant. | <input type="checkbox"/> | <input type="checkbox"/> |
| (d) None of the above. | <input type="checkbox"/> | <input type="checkbox"/> |

Doctor's particulars:

Name in full: _____

MCR number: _____ Specialty: _____

Email address: _____

Submission instructions:

A. Using this answer form

- Photocopy this answer form.
- Indicate your responses by marking the "True" or "False" box
- Fill in your professional particulars.
- Post the answer form to the SMJ at 2 College Road, Singapore 169850.

B. Electronic submission

- Log on at the SMJ website: URL <<http://www.sma.org.sg/cme/smj>> and select the appropriate set of questions.
- Select your answers and provide your name, email address and MCR number. Click on "Submit answers" to submit.

Deadline for submission: (October 2005 SMJ 3B CME programme): 12 noon, 25 November 2005

Results:

- Answers will be published in the SMJ December 2005 issue.
- The MCR numbers of successful candidates will be posted online at <<http://www.sma.org.sg/cme/smj>> by 20 December 2005.
- All online submissions will receive an automatic email acknowledgment.
- Passing mark is 60%. No mark will be deducted for incorrect answers.
- The SMJ editorial office will submit the list of successful candidates to the Singapore Medical Council.