UNIVERSITÀ DEGLI STUDI DI PADOVA

FACOLTÀ DI SCIENZE STATISTICHE

CORSO DI LAUREA
IN STATISTICA E GESTIONE DELLE IMPRESE

RELAZIONE FINALE

FULL PROFILE AND CHOICE-BASED CONJOINT ANALYSIS: METODOLOGIE E SOFTWARE

RELATORE: CH.MO PROF. FORTUNATO PESARIN
CORRELATORE: CH.MO PROF. LUIGI SALMASO

LAUREANDO: PAOLO BROGGIAN

ANNO ACCADEMICO 2003-04

Ai miei genitori

Indice

Prefazione

In questa tesi si tratta la *Conjoint Analysis* classica con approccio "Full Profile" e la *Choice-Based Conjoint* con sviluppo sia classico che bayesiano. La relazione è organizzata nel seguente modo:

- nel primo capitolo si espongono le metodologie classiche della *Conjoint Analysis* "Full Profile" e un approfondimento sui piani fattoriali 2^k;
- nel secondo capitolo si introducono le finalità di uno studio di *Conjoint Analysis* e si applicano le metodologie precedentemente esposte ad un caso reale. I dati saranno elaborati col software SPSS modulo "Conjoint Analysis";
- nel terzo capitolo si espongono le metodologie della *Choice-Based Conjoint* prima con approccio classico e successivamente con approccio bayesiano;
- nell'ultimo capitolo si espone l'operatività e le funzioni del "Sawtooth Software" programma concepito per analisi di *Choice-Based Conjoint*. Si spiegheranno le procedure di importazione e analisi con obbiettivi pratici senza analisi qualitative dei dati.

Cos'è la *Conjoint Analysis*

L'utilizzo del marketing quale strumento strategico ed operativo per competere nei settori economici, caratterizzati da una forte concorrenza, rappresenta ormai uno statement consolidato nella cultura imprenditoriale e manageriale.
La conoscenza dei mercati, in termini di competitività, di canali distributivi e di clienti finali, rappresenta un vantaggio competitivo che consente alle aziende di eccellere e di differenziarsi. Per questo è di fondamentale interesse, nella complessità propria che regola il mercato, conoscere i meccanismi che determinano le scelte del consumatore, ossia quell'insieme di azioni che precedono, accompagnano e seguono le decisioni di acquisto. Una risposta nasce dall'integrazione della *conjoint analysis*.
L'analisi della misurazione congiunta (*conjoint analysis*) è una tecnica statistica di analisi multivariata, a carattere "decomposizionale", di nuovissima concezione e sviluppo che prende in considerazione le preferenze del consumatore nella scelta di beni e servizi. I fondamenti concettuali della *Conjoint analysis*(da ora in poi CA) risiedono:

- nella teoria del comportamento del consumatore proposta da Kevin Lancaster, secondo il quale l'utilità d'uso di un bene deriva dalle singole caratteristiche che lo compongono;

- nei modelli di preferenza multiattributivi (*MultiAttribute Utility Theory*) di Fishbein-Rosemberg, basati sull'approccio compositivo (l'utilità totale di un prodotto discende dalle utilità dei singoli attributi che lo compongono).

Ipotesi di base è che ogni bene o servizio può essere descritto per mezzo di un insieme di caratteristiche che lo definiscono, da noi denominate attributi, che a loro volta si esplicano in sottocategorie, chiamate livelli. Il consumatore, quindi, associa ad ogni profilo ,cioè ad ogni combinazione di attributi e rispettivi livelli, un proprio gradimento che per assunto non è altro che l'utilità del prodotto o del servizio. L'utilità complessiva è data dalla combinazione delle varie utilità dei livelli tramite una certa regola di cui in seguito daremo risposta. Tramite la CA è possibile valutare:

- il grado di utilità corrispondente ad ogni livello o modalità di ciascuna caratteristica (*parth-worth*);
- l'importanza che ogni individuo attribuisce a ciacuna caratteristica di un prodotto o servizio.

Quindi l'obbiettivo del seguente lavoro è, avvalendoci dell'ottica decompositiva della CA, di individuare il processo di formazione delle preferenze ovvero ciò che spinge un consumatore ad acquistare un prodotto o a fruire di un servizio.

Capitolo 1

I passi della *Conjoint Analysis*

Il procedimento dalla CA prevede una serie di passi successivi strettamente legati l'uno all'altro che portano alla formazione della scala di gradimento dei vari profili.
Queste sono, schematizzando, le fasi:

- individuazione degli attributi del bene/servizio e dei relativi livelli;
- definizione di un piano degli esperimenti e dei relativi profili da sottoporre a giudizio diretto dei consumatori;
- selezione di un campione casuale di consumatori ai quali chiedere valutazioni di preferenza su ciascun profilo;
- scelta di un modello di utilità;
- stima dei parametri associati a ciascuna modalità degli attributi del prodotto (funzioni di "utilità parziale" degli attributi);
- stima dell'importanza relativa di ciascun attributo o fattore;
- valutazione dell'utilità totale corrispondente a profili non compresi nella rilevazione.

Qui di seguito si presentano le più importanti fasi di una *conjoint analysis*.

1.1 Selezione degli attributi e dei relativi livelli

Individuato il campo di indagine su cui effettuare l'analisi delle preferenze del consumatore, la prima fase della CA è neccessariamente rappresentato dalla selezione degli attributi e dei relativi livelli che caratterizano il bene o il servizio considerato.
Questo primo passo riveste un'importanza cruciale nella determinazione del grado di correttezza e di significatività dell'intera CA. Il procedimento si articola in due fasi:

- la decisione sulla numerosità di attributi e livelli;
- scelta qualitativa degli stessi.

Non esiste una regola precisa che possa aiutare in questo e tutto è affidato alla capacità e all'esperienza dell'analista.

1.2 Definizione di un piano degli esperimenti

Nella CA il piano delle combinazioni (stimoli sperimentali) viene predisposto sulla base della teoria della *programmazione degli esperimenti.*
La scelta della numerosità dei fattori discende dall'ipotesi formulata sulla forma della funzione di risposta considerata, che può essere di tipo *additivo* cioè ad effetti principali o *misto* cioè ad effetti principali e ad effetti di interazione. Nel modello additivo a "coefficienti separati" (*part-worths*) in corrispondenza di ogni modalità di un fattore viene stimato un coefficiente di "utilità parziale"; nel modello si stima, in aggiunta, un coefficiente per ciascuna combinazione ("iterazione") di modalità.
Quindi dati P fattori, a $m_1, m_2, \ldots, m_p$ modalità (qualitative o quantitative) il numero di possibili *combinazioni* delle stesse è dato dal prodotto cartesiano:

$$\prod_{p=1}^{P} m_p$$

Il numero di *coefficienti* di utilità da stimare nel modello ad effetti principali risulta, invece, pari a :

$$\sum_{p=1}^{P} m_{p-1} = \sum_{p=1}^{P} m_p - P$$

Così per esempio per 5 fattori a due livelli, cioè p=1,2,....,P=5, con m_p=2 si hanno 2^5=32 possibili combinazioni di livelli ed un numero di coefficienti di utilità (incognite) pari a $\sum_{p=1}^{5} m_p$-P=10-5=5. Ovvero necessitiamo almeno 5 combinazioni sperimentali per giungere a tante equazioni quante sono le incognite.

E' vero che un piano *fattoriale completo* (che consente di stimare gli effetti "principali" e di "interazione" delle varie modalità degli attributi del prodotto) permette la massima informazione tuttavia, allo stesso modo, ciò comporta un carico alquanto gravoso per il valutatore appena il numero dei fattori e/o delle corrispondenti modalità superano le poche unità. Per questo è opportuno considerare un sottoinsieme di combinazioni (piano frazionato), da scegliere con criterio stocastico, così pure da consentire l'estensione dei parametri stimati anche ai profili di prodotto non sottoposti a valutazione. I piani frazionati, però, non producono la stima dei parametri degli effetti di *interazione* dei fattori, che risultano "confusi" con alcuni effetti principali. Tra i piani frazionati, il piano *ortogonale*[1] consente di stimare, con il minor numero possibile di combinazioni (tramite un modello additivo), le utilità parziali dei livelli dei vari attributi, cioè gli effetti *principali*. Le combinazioni del piano frazionato si scelgono tra i punti "candidati" [2], cioè all'interno dei punti del disegno fattoriale completo o nell'ambito di un suo sottoinsieme (per esempio escludendo le combinazioni di prodotto non realizzabili, sul piano produttivo o quelle prive di interesse). Nel secondo caso non è possibile pervenire ad un piano ortogonale. Esistono vari algoritmi di generazione di disegni ortogonali; tali algoritmi sono in grado di selezionare un insieme di punti che ottimizzano un dato criterio di efficienza. [3]
Con la CA si individuano:

[1] Trattasi di un piano i cui fattori sono tra loro incorrelati.

[2] W.F.Kuhfeld, R.D.Tobias,M.Garratt(1994)

[3] Procedura "Orthoplan" di SPSS

- gli attributi più importanti del prodotto;
- il miglior *concept* di prodotto; ovvero la combinazione di modalità più "attraente" dei suoi attributi.

1.2.1 Un particolare disegno fattoriale: il piano 2^k

Il disegno fattoriale 2^k è il piano fattoriale di k fattori a 2 livelli ciascuno. Il presente piano è particolarmente utilizzato quando l'oggetto di studio presenta un'elevata numerosità di fattori. Con ciò, quindi, si minimizzano i livelli dei vari fattori rendendo di più facile lettura il problema. Si assume per altro che:

- i fattori sono fissi;
- il disegno è completamente randomizzato;
- le normali assunzioni di normalità sono soddisfatte.

Come vedremo il disegno 2^k trova una facile generalizzazione dai piani 2^2 e 2^3. Iniziamo, quindi, a presentare un piano 2^2. Denominati A e B i fattori con i relativi livelli + e - , n=il numero delle repliche e N=2^kn la numerosità totale si ha:

	A	B	AB
(1)	-	-	+
a	+	-	-
b	-	+	-
ab	+	+	+

l'effetto di A è dato da:

$$A = \overline{y}_{A^+} - \overline{y}_{A^-}$$

$$A = \frac{a + ab}{2n} - \frac{(1) + b}{2n}$$

$$A = \frac{1}{2n}(a + ab - (1) - b)$$

parallelamente l'effetto di B è:

$$B = \overline{y}_{B^+} - \overline{y}_{B^-}$$

$$B = \frac{ab + b}{2n} - \frac{a + (1)}{2n}$$

$$B = \frac{1}{2n}(ab + b - (1) - a)$$

infine l'effetto dell'interazione è:

$$AB = \frac{ab + (1)}{2n} - \frac{a + b}{2n}$$

$$AB = \frac{1}{2n}(ab + (1) - a - b)$$

in ogni piano 2^k è importante esaminare la direzione e l'intensità dell'effetto del fattore per determinare se lo stesso risulta importante o meno. L'analisi della

varianza generalmente ci fornisce la risposta.
Per calcolare la somma dei quadrati del singolo fattore ci si avvale della scomposizione della varianza. Infatti si prova che

$$S_A^2 = bn \sum_{i=1}^{a} (\overline{y}_{i..} - \overline{y}_{...})^2$$

da cui si ricava, essendo b il numero dei livelli del fattore B,

$$2n\Big[(\underbrace{\overline{y}_{A-} - \overline{y}_{...}}_{\frac{\overline{y}_{A+}+\overline{y}_{A-}}{2}})^2 + (\underbrace{\overline{y}_{A+} - \overline{y}_{...}}_{\frac{\overline{y}_{A+}+\overline{y}_{A-}}{2}})^2\Big]$$

$$2n\Big[(\frac{\overline{y}_{A+} - \overline{y}_{A-}}{2})^2 + (\frac{\overline{y}_{A+} - \overline{y}_{A-}}{2})^2\Big]$$

$$n(\overline{y}_{A+} - \overline{y}_{A-})^2$$

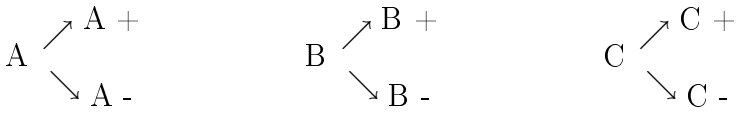
$$S_A^2 = n(\texttt{ effetto di A })^2$$

Ciò è vero anche per il fattore B e per l'iterazione dei due fattori, ovvero:

$$S_B^2 = n(\texttt{ effetto di B })^2$$

$$S_{AB}^2 = n(\texttt{ effetto di AB })^2$$

Ora si dimostra per un disegno 2^3:



	A	B	C	AB	AC	BC	ABC
(1)	-	-	-	+	+	+	-
a	+	-	-	-	-	+	+
b	-	+	-	-	+	-	+
ab	+	+	-	+	-	-	-
c	-	-	+	+	-	-	+
ac	+	-	+	-	+	-	-
bc	-	+	+	-	-	+	-
abc	+	+	+	+	+	+	+

l'effetto dei singoli fattori e delle interazioni risultano:

$$A = \frac{1}{4n}(a + ab + ac + abc - (1) - b - c - bc)$$

$$B = \frac{1}{4n}(b + ab + bc + abc - (1) - a - c - ac)$$

$$C = \frac{1}{4n}(c + ac + bc + abc - (1) - b - a - ab)$$

$$AB = \frac{1}{4n}((1) + ab + c + abc - a - b - ac - bc)$$

$$AC = \frac{1}{4n}((1) + b + ac + abc - a - ab - c - bc)$$

$$BC = \frac{1}{4n}((1) + a + bc + abc - b - ab - c - ac)$$

$$ABC = \frac{1}{4n}(a + b + b + abc - (1) - bc - ab - ac)$$

dall'equazione di scomposizione della varianza in un disegno a 3 fattori si ricava che la somma dei quadrati di A è:

$$A = bcn \sum_{i=1}^{a} (\overline{y}_{i..} - \overline{y}_{...})^2$$

quindi si avrà:

$$S_A^2 = 2n(\ \texttt{effetto di A}\ )^2$$
$$S_B^2 = 2n(\ \texttt{effetto di B}\ )^2$$
$$S_C^2 = 2n(\ \texttt{effetto di C}\ )^2$$
$$S_{AB}^2 = 2n(\ \texttt{effetto di AB}\ )^2$$
$$S_{AC}^2 = 2n(\ \texttt{effetto di AC}\ )^2$$
$$S_{BC}^2 = 2n(\ \texttt{effetto di BC}\ )^2$$
$$S_{ABC}^2 = 2n(\ \texttt{effetto di ABC}\ )^2$$

esposto il caso 2^2 e 2^3 si deduce che dato un piano 2^k dove k è il numero dei fattori:

$$effetto_i = \frac{1}{2^{k-1}n}(Contrasto_i) \qquad \forall\ i = A, B, ..., K \tag{1.1}$$

$$S_i^2 = (effetto_i)^2 n 2^{k-2} \qquad \forall\ i = A, B, ..., K \tag{1.2}$$

la 1.1 e la 1.2 valgono anche per le interazioni di qualsiasi grado, inoltre si definisce:

$$Contrasto_{AB...K} = (a \pm 1)(b \pm 1)...(k \pm 1)$$

considerando un disegno 2^3 il contrasto di AB è:

$$Contrasto_{AB} = (a-1)(b-1)(c+1) = abc + ab + c + (1) - ac - bc - a - b$$

1.2.2 I piani frazionati 2^{k-p}

Si è visto nel precedente paragrafo la grande utilità del piano 2^k, come già detto infatti il disegno permette di studiare l'effetto di tutte le combinazioni dei 2 distinti livelli di ciascuno dei k fattori presi in considerazione. Per effettuare ciò, sono necessarie minimo 2^k osservazioni, si può quindi notare che il numero delle osservazioni cresce rapidamente con il numero dei fattori sotto studio. Tuttavia è bene notare che lo sperimentatore potrebbe non essere interessato a stimare tutte le possibili interazioni tra i k fattori, ma solamente gli effetti principali e le interazioni di ordine basso; inoltre il costo ed i tempi necessari per eseguire un piano completo potrebbe risultare eccessivo. In questo contesto quindi nasce l'esigenza di introdurre i piani fattoriali frazionati 2^{k-p}, in cui solo una frazione del disegno completo viene considerata.
Per stimare gli effetti principali e gli effetti delle interazioni di ordine basso può essere sufficiente ricorrere ad un frazionamento del piano che utilizza 2^m trattamenti con m<k. Se p è il numero di frazionamenti del piano si individua un piano fattoriale frazionato 2^{k-p} con k fattori principali e 2^{k-p} osservazioni. Si definisce che un piano fattoriale S^{n-k}, con n fattori su S livelli e con S^{n-k} osservazioni, è univocamente determinato da k indipendenti parole detinitorie. Una parola è costituita da un insieme di lettere o numeri che denominano i fattori. Una parola è costituita da un insieme di lettere o numeri che denominano i fattori. I fattori vengono indicati con 1,2,3, ...,n o con A,B,...Il numero di lettere (o numeri) che costituisce una parola viene denominato lunghezza della parola e il gruppo formato dalla moltiplicazione delle parole è il gruppo dei contrasti definitori o relazione definitrice. Il vettore W=(A_i,...,A_n) viene denominato schema di lunghezza della parola, dove A_i indica il numero di parole di lunghezza i nel gruppo di contrasti definitori.

Esempio

Si consideri un esperimento a 3 fattori A,B,C ciascuno con 2 livelli; si hanno a disposizione 4 osservazioni invece di 8, si può allora effettuare un frazionamento

del piano completo 2^3. Si supponga di selezionare i trattamenti a,b,c,abc : il disegno 2^{3-1} è costituito dai soli trattamenti che hanno segno + in corrispondenza della colonna ABC (ABC è detto generatore della frazione, indicando con I la colonna identità con tutti +, si ottiene I=ABC detta relazione definente).

	I	A	B	C	AB	AC	BC	ABC
a	+	+	-	-	-	-	+	+
b	+	-	+	-	-	+	-	+
c	+	-	-	+	+	-	-	+
abc	+	+	+	+	+	+	+	+

	I	A	B	C	AB	AC	BC	ABC
ab	+	+	+	-	+	-	-	-
ac	+	+	-	+	-	+	-	-
bc	+	-	+	+	-	-	+	-
(1)	+	-	-	-	+	+	+	-

Le combinazioni lineari delle osservazioni, usate per stimare gli effetti principali A,B,C e le interazioni BC,AC,AB sono:

$$L_A = \frac{1}{2}(a - b - c + abc)$$

$$L_B = \frac{1}{2}(-a + b - c + abc)$$

$$L_C = \frac{1}{2}(-a - b + c + abc)$$

$$L_BC = \frac{1}{2}(a - b - c + abc)$$

$$L_AC = \frac{1}{2}(-a + b - c + abc)$$

$$L_AB = \frac{1}{2}(-a - b + c + abc)$$

Si nota che $L_A = L_{BC}$,$L_B = L_{AC}$,$L_C = L_{AB}$ in quanto il fattore A è alias dell'interazione BC, B è alias di AC, C è alias di AB. Due fattori o interazioni si dicono ALIAS (o confusi) se le colonne della matrice del disegno corrispondenti a tali fattori sono uguali.
Risulterà impossibile distinguere tra A e BC, tra B e AC e tra C e AB in quanto, per esempio, L_A andrà a stimare in modo confuso la somma degli effetti A+BC:

$$L_A \Rightarrow A + BC$$
$$L_B \Rightarrow B + AC$$
$$L_C \Rightarrow C + AB$$

L'effetto di un qualunque fattore e/o interazione F_i si dice stimabile in modo confuso se la colonna corrispondente a F_i è uguale alla colonna corrispondente F_j con $i \neq j$.
In generale un disegno fattoriale 2^m permette la stima di $2^m - 1$ effetti di cui m sono le stime attribuibili agli effetti dei fattori principali e $2^m - m - 1$ sono le stime attribuibili agli effetti delle interazioni.
Se nell'esempio si suppone che le interazioni a due fattori siano trascurabili, allora il piano 2^{3-1} ha prodotto le stime dei 3 effetti principali A,B,C. Se invece riteniamo che esse siano rilevanti, protremo andare a stimarle utilizzando una sucessione di piani fattoriali frazionati del piano completo, effettueremo delle nuove stime che poi andremo a combinare linearmente con quelle precendenti. Dunque, combinando una successione di due piani fattoriali frazionati possiamo isolare sia gli effetti principali sia le interazioni a due fattori. Riguardo la stima degli effetti e l'analisi della varianza si procede in modo del tutto analogo ai piani 2^k.
In un piano fattoriale frazionato 2^{k-p} ci sono 2^p modi di attribuire ai generatori i segni + e - ed esistono parimenti 2^p differenti frazioni del piano. Dato un esperimento a cui vogliamo applicare un piano fattoriale frazionato 2^{k-p} dobbiamo scegliere tra le varie frazioni quella che ci individua un "buon" piano,

ossia un piano che permette di individuare senza alcuna ambiguità un numero abbastanza elevato di effetti principali e di interazioni.
Quando però si hanno a disposizione un numero basso di osservazioni può accadere che si sia costretti a frazionare notevolmente il piano, a tal punto da ottenere effetti principali confusi con interazioni di secondo ordine. Quindi un piano risulterà tanto migliore quanti più effetti non sono confusi tra loro.

1.3 Scelta di un modello di utilità

E' intuitivo pensare che quanto un specifico profilo di un prodotto/servizio incontrerà il gradimento di un consumatore tanto più il suo uso fornirà utilità. La preferenza a sua volta può essere interpretata come funzione delle modalità o dei livelli delle caratteristiche rilevanti del prodotto/servizio. Per giungere a ciò è neccessario decidere la forma del "modello di composizione" (funzione di utilità individuale) che interpreta la formazione delle preferenze di un consumatore.
Varie alternative sono disponibili però, sicuramente, il modello più utilizzato è quello additivo, nel quale le utilità parziali dei singoli livelli di ogni attributo vengono sommate per ottenere l'utilità complessiva di un profilo.
In letteratura sono presenti i seguenti schemi principali:

a) il **modello vettore**;

b) il **modello punto ideale**;

c) il **modello part-worth**.

1) Indichiamo con x_{jk} il livello d'intensità (modalità) che il carattere quantitativo k-simo, k=1,2,...,K, presenta nella combinazione j-esima, j=1,2,...,J, e con y_j la preferenza/utilità assegnata a detto stimolo j, il modello vettoriale è così espresso:

$$y_j = \sum_{k=1}^{K} w_k x_{jk}$$

dove il termine w_k indica il peso d'importanza assegnato all'attributo k. I pesi w_k risultano normalmente diversi tra i cosumatori e manifestano la struttura delle loro preferenze[4].

2) Nel modello del punto ideale si ipotizza per ciascun valutatore esista un profilo "ideale" del prodotto, corrispondente al livello I_k con(k=1,2,...,K)

[4]Il modello è di tipo compositivo(Fishbein, 1967).

di ciascun attributo k.

Con il presente modello si vuole individuare, per il profilo analizzato, una misura di utilità crescente al decrescere della sua distanza dal profilo ideale. Tale misura è la distanza euclidea, cioè:

$$d_j^2 = \sum_{k=1}^{K} w_k (x_{jk} - I_k)^2$$

L'utilità y_j è correlata negativamente alla distanza d_j^2.

3) Il modello *part-worth*(a coefficienti separati) formula la preferenza/utilità y_j, per lo stimolo j-esimo, tramite una funzione discreta f(.), definita sulla combinazione di livelli degli attributi, qualitativi o quantitativi, secondo la seguente espressione:

$$y_j = \sum_{k=1}^{K} f_k(x_{jk})$$

Il presente modello risulta più flessibile date le diverse forme della funzione di preferenza/utilità, in relazione a diverse specificazioni di f(.).

1.4 Stima dei valori di utilità parziale

Esistono due modelli di *Conjoint analysis*, la CA metrica e la non metrica. Qui si tratterà solamente il primo tipo. Nella CA metrica la variabile risposta Y, espressione dei giudizi di preferenza di ciascun intervistato, viene utilizzata direttamente per la stima dei parametri tramite la regressione lineare multipla. La stima può essere calcolata tramite la funzione di utilità parziale (non lineare). Con riferimento ad un generico consumatore i e codificate le variabili esplicative qualitative (fattori) nel seguente modo:

$$d_{mkp} = \begin{cases} 1 & \text{se il profilo } m \text{ presenta l'attributo } k \text{ con livello } p \\ 0 & \text{in caso contrario} \end{cases}$$

dove:

m è la generica combinazione m=1,2,...,M

k il generico livello del fattore k, p=1,2,...,P_k
la funzione di utilità parziale (non lineare) del fattore k per il profilo m è la seguente:

$$u_{ik} = \sum_{p=1}^{P_k} w_{ikp} d_{mkp}$$

dove:

u_{ik} indica l'utilità che il fattore k procura al rispondente i-esimo;

w_{ikp} è il coefficiente di regressione che esprime l'importanza attribuita dall'i-esimo rispondente al fattore k, considerato al livello p. Si noti che, poiché $d_{mkp} = 1$ solo per un livello del fattore k, u_{ik} corrisponde all'utilità del livello medesimo per il fattore k, con riferimento al profilo m.

Con riguardo a tutti i K fattori, la funzione di utilità *totale* (R_{im}) dell'i-esimo rispondente, per il profilo m del prodotto si esprime secondo il seguente modello additivo stocastico:

$$R_{im} = \sum_{k=1}^{K} \sum_{p=1}^{P_k} w_{ikp} d_{mkp} + \varepsilon_{im}$$

dove:
w_{ikp} (part worths) è il generico parametro incognito;
ε_{im} è il termine di errore relativo al rispondente i sul profilo m.

1.4.1 OLS: Stima dei minimi quadrati ordinari

Il metodo più largamente utilizzato per la stima delle utilità parziali è sicuramente la regressione ai minimi quadrati.
Le tre principali assunzioni della CA metrica sono le seguenti:

- l'utilità totale di un prodotto j per un consumatore è funzione lineare della valutazione di tale prodotto;
- i giudizi dati da ciascun individuo sono misurati su scala ad intervallo;
- le valutazioni dei rispondenti indicano con quale probabilità i corrispondenti prodotti veranno acquistati.

Com è intuitivo pensare stiamo trattando un modello lineare multivariato matricialmente così definito:

$$y = X\beta + \varepsilon$$

dove:

- y è il vettore colonna, di dimensioni M×1, dei giudizi di valutazione (y_{im}) espressi dall'i-esimo rispondente (i=1,2,...,M) del prodotto;
- X indica la matrice del piano sperimentale, di dimensione M×$(\sum_{k=1}^{K} P_k - K + 1)$, delle variabili indicatrici d_{mkp} delle categorie degli attributi (contenente per riga i vettori delle modalità *dummy* delle diverse combinazioni sperimentali), cui è stato aggiunto il termine di intercetta;
- β è il vettore colonna, di dimensione $\sum_{k=1}^{K}(P_k - K + 1)$×1, dei coefficienti incogniti di utilità parziale;
- ε è il vettore colonna, di dimensione M×1, dei residui del modello per il rispondente i-esimo.

I parametri β vengono stimati sotto la condizioni che risulti minima la somma dei quadrati degli scarti tra i punteggi di valutazione osservati ed i punteggi calcolati (stimati); cioè formalmente:

$$\|y - X\beta\|^2$$

La stima di beta quindi è:

$$\widehat{\beta} = (X'X)^{-1}X'y \tag{1.3}$$

questa è l'unica soluzione del sistema lineare, inoltre tramite la (1.3) viene definito il vettore dei residui:

$$\widehat{\varepsilon} = y - X\widehat{\beta}$$

che costituisce una stima degli errori non osservabili ε.

1.5 Importanza relativa dei fattori

E' consuetudine valutare, anche se in modo semplicistico, l'importanza relativa dei fattori allo scopo di rendere comparabili le utilità parziali e ricavare valori di importanza compresi fra 0 e 1. Questi sono i passi per calcolare questi indici:

- si determina il campo di variazione cioè la differenza fra l'utilità parziale più elevata e l'utilità più bassa delle sue modalità;
- si sommano i campi di variazione di tutti i fattori;
- si calcola, per ciascun fattore, il rapporto tra il campo di variazione e la somma dei campi di variazione.

Formalmente:

$$I_{rFi} = \frac{(\text{utilità parziale più grande di} F_i - \text{utilità parziale più piccola di} F_i)}{\sum_{i=1}^{I}(\text{utilità parziale più grande di} F_i - \text{utilità parziale più piccola di} F_i)}$$

1.6 Stima di profili non compresi nella rilevazione

Combinando tutti i possibili livelli di k fattori, secondo un disegno fattoriale completo, si potrebbe ottenere un numero di profili di prodotto sproporzionatamente elevato, ciò misurato in termini di costo e di carico lavoro da sottoporre ai valutatori. E' consuetudine, per ovviare al problema, utilizzare un piano fattoriale frazionato, infatti, punto di forza della CA è la capacità di ottenere stime *ex-post* delle valutazioni "virtuale" delle combinazioni non sottoposte a giudizio. In pratica con la presente tecnica è possibile ottenere tutte le valutazioni di tutti i possibili profili di prodotto. La numerosità dei profili non sottoposti a valutazione è:

$$\sum_{k=1}^{K} P_k - M$$

la stima è ottenuta tramite il presente modello:

$$\widehat{y} = T\beta$$

dove:

$\widehat{y}$ è il vettore colonna, di dimensioni $\prod_{k=1}^{K} \times 1$, dei valori delle valutazioni stimate;

T è una matrice, di dimensioni $\prod_{k=1}^{K} P_k \times (\sum_{k=1}^{P_k} P_k - K + 1)$, omologa della matrice X, ma contenente le variabili indicatrici di tutti i possibili profili del prodotto;

β è il vettore colonna, di dimensioni $\sum_{k=1}^{P_k}(P_k - K + 1) \times 1$, dei coefficienti incogniti di utilità parziale per l'i-esimo valutatore.

1.7 I limiti della CA

Non bisogna dimenticare che, come tutte le metodologie statistiche, la CA può presentare limitazioni dovute soprattutto ad una poco attenta applicazione della tecnica. Queste sono le valutazioni da non sottovalutare:

- alcuni prodotti/servizi, soprattutto quelli ad alto contenuto d'immagine, non sono valutati analiticamente dai consumatori nelle loro caratteristiche rilevanti, pertanto il modello della CA può rappresentare solo con grande approssimazione il processo decisionale di acquisto;

- è importante ricordare che i risultati ottenuti sono fortemente condizionati dal ricorso ad alcune ipotesi volutamente semplificatrici, in particolare quella di uno schema additivo che collega la preferenza/utilità totale di ogni alternativa da valutare alle specifiche modalità o livelli degli attributi che la caratterizzano e quella di assenza di interazioni di qualunque ordine tra tali modalità o livelli;

- occorre prestare molta attenzione nell'estendere le conclusioni anche a modalità o a livelli dei caratteri che non entrano esplicitamente nel piano della rilevazione;
- non è corretto concludere che un basso valore di importanza relativa per un attributo riflette una sua scarsa rilevanza per i consumatori;
- può risultare azzardato prevedere le reazioni dei consumatori in base ai risultati dell'analisi svolta se non sono stati inseriti nel disegno sperimentale attributi chiave del prodotto.

Capitolo 2

Applicazioni della *CA*

In questo secondo capitolo si presenta, dopo una breve digressione sulle utilità della *Conjoint analysis*, lo studio di un caso reale elaborando i dati secondo i metodi fin qui illustrati.
Obbiettivo di qualunque azienda è di adeguare la propria offerta in funzione delle esigenze reali del mercato in cui intende operare. E' importante, quindi, poter valutare i bisogni e le caratteristiche dei probabili acquirenti. La segmentazione è quel processo mediante il quale l'azienda individua gruppi omogenei e distinti di consumatori, al fine di adeguare tanto i prodotti, quanto le strategie di marketing, alle differenze individuabili entro l'insieme delle esigenze manifestate dai consumatori e/o utilizzatori.

2.1 La *CA* nella segmentazione flessibile

La segmentazione del mercato, da un punto di vista operativo, passa attraverso l'attuazione di alcune fasi così riassumibili:

- definizione del problema e selezione della procedura di segmentazione;
- messa a punto del programma dell'indagine sul campo ai fini di raccogliere le informazioni necessarie alla realizzazione delle operazioni di segmentazione;

- elaborazione, interpretazione e impiego dei risultati.

Esistono vari modelli di segmentazione che vengono così classificati:

- a priori;
- a posteriori;
- flessibile.

In qualsiasi modello di segmentazione è importante individuare le **basi**, ovvero i caratteri rispetto ai quali viene eseguita la segmentazione, e i **descrittori**, ossia le variabili che servono per interpretare i profili dei segmenti. La scelta delle variabili che rientrano nelle due citate categorie non è un'operazione semplice.
Nei modelli di **segmentazione a priori** si procede alla suddivisione della popolazione obbiettivo a seconda delle modalità presentate da una o più basi, specificate appunto a priori. Un esempio è la classica suddivisione geografica della popolazione, quando si assume come base la ripartizione territoriale di residenza. I descrittori dei profili dei segmenti si individuano abitualmente attraverso tecniche statistiche di segmentazione binaria o multipla, quali l'**Automatic Interaction Detection** (AID) e il **CHi-squared Automatic Interaction Detection** (CHAID).
I modelli di **segmentazione a posteriori** si basano sull'applicazione di algoritmi di raggruppamento (*clustering*) differenziandosi dai precedenti solo in relazione al modo in cui viene selezionata la base di segmentazione. Il termine a posteriori evidenzia il fatto che i segmenti sono determinati attraverso la classificazione delle unità statistiche a seguito dei risultati di una **Cluster Analysis**, cioè a partire dal grado di dissomiglianza rispetto ad un insieme prescelto di variabili. Queste esprimono in generale i comportamenti dei consumatori, i loro bisogni, le loro attitudini, lo stile di vita, o altre caratteristiche di tipo psicologico o ancora i benefici attesi dall'uso di determinati prodotti. In questo caso non c'è una scelta a priori, e non sono prefissati nè il numero nè le tipologie dei gruppi da formare.
Con i modelli di **segmentazione flessibile** è possibile individuare un numero

anche elevato di segmenti alternativi, ciascuno costituito dall'insieme dei consumatori con profili di risposta simili in termini di preferenza per prodotti o marche già esistenti, oppure potenzialmente introducibili sul mercato. La segmentazione flessibile si fonda sull'integrazione dei risultati di uno studio che si avvale dell'applicazione della **conjoint analysis** e di una simulazione sul comportamento di scelta dei consumatori. Quindi dopo essersi soffermati sulle radici applicative della CA si affronterà nel proseguo lo studio di un caso reale.

2.2 Il problema reale

Lo studio condotto in quest'estelaborato prende libera ispirazione da un caso reale che fu sottoposto da un cliente ad un'impresa di marketing operativo. L'azienda è un piccolo studio con sede a Barcellona, città dove mi trovavo in qualità di studente "Erasmus", nella quale svolgevo la mansione di stagista.
Il caso fu oggetto di studio nel mese di marzo del 2003. Il cliente è proprietario di un negozio di capi d'abbigliamento di ultima moda sito all'angolo di via "Provença" con viale "Rocafort" area centro-nord di Barcellona. Si vuole fare uno studio di segmentazione sulle preferenze della clientela in termini di orari e giorni di apertura dell'esercizio commerciale. L'analisi volutamente si focalizza su giorni infrasettimanali eliminando la domenica e il lunedì considerati giorni di chiusura. In aggiunta si elimina la giornata di sabato, giorno che produrebbe un alto indice di preferenziabilità distogliendo i valutatori dal vero problema. Nella fattispecie operando in questo modo si otterrà una simulazione dell'affluenza dei potenziali clienti durante la settimana.

2.3 Selezione degli attributi

Il caso presenta due variabili di segmentazione o **basi** e tre variabili oggetto di studio o **descrittori**.
Le basi sono:

- **l'età** a due livelli: minori di 40 anni e maggiori di 40 anni;

- **il sesso** con uomini e donne.

Da ciò si descriveranno quattro segmenti d'individui che in un primo momento si analizzerà separatamente. Invece i descrittori o gli attributi del problema sono i seguenti:

- **il giorno** con le seguenti opzioni: martedì, mercoledì, giovedì e venerdì;

- **l'ora** con le seguenti fasce orarie: 10-12, 12-14, 16-18 e 18-20;

- **la direzione**, è stata introdotta questa variabile per intuire quale delle due entrate (disposte su due vie diverse) fosse preferita dalla clientela, essa presenta i seguenti livelli: Rocafort e Provença.

2.4 Definizione del piano sperimentale

Tutte le possibili combinazioni delle 10 modalità dei 3 attributi ammontano a $4 \times 4 \times 2 = 32$. Essendo P=3 e $\sum_{p=1}^{P} m_k - P = 10 - 3 = 7$, necessitiamo almeno di 7 combinazioni di livelli dei fattori per stimare 7 coefficienti del modello. Si sono scelte, attraverso la procedura "Ortoplan" di SPSS, 16 combinazioni che rappresentano un piano frazionato di tipo ortogonale. Il disegno contiene il minor numero possibile di combinazioni che permettono di stimare in modo indipendente le utilità parziali. Il piano si presenta in questo modo:

giorno	ora	direzione
venerdì	12-14	rocafort
mercoledì	12-14	provença
venerdì	18-20	rocafort
martedì	16-18	provença
giovedì	16-18	rocafort
venerdì	10-12	provença
martedì	10-12	provença
mercoledì	10-12	rocafort
venerdì	16-18	provença
giovedì	10-12	rocafort
mercoledì	18-20	provença
giovedì	18-20	provença
martedì	12-14	rocafort
mercoledì	16-18	rocafort
martedì	18-20	rocafort
giovedì	12-14	provença

2.5 I Dati e la codifica del modello

Non potendo adeguare i dati in mio possesso si è decisi di simularli. Si sono simulati un totale di 40 valutazioni dei 16 profili su una scala 0-10 (0=disinteresse completo per il profilo proposto; 10=interesse massimo), espresso con due valori decimali. I 40 punteggi sono divisi in quattro segmenti a seconda dell'età e del sesso preso in considerazione. Si è simulato da una distribuzione uniforme le medie dei vari profili per poi simulare i veri valori da distribuzioni normali con le citate medie e varianze dipendenti dal segmento preso in considerazione. In pratica all'interno dei segmenti ci sarà un'omogeneità della varianza che differenzierà i segmenti a seconda del loro valore. In appendice presento i dati relativi ai segmenti dei rispondenti. Le funzioni di utilità parziali associate alle modalità dei tre attributi sono state stimate con la tecnica della regressione lineare multipla su variabili *dummy* tramite il pacchetto statistico SPSS. La

codifica degli attributi comporta, quindi, che ogni modalità di ciascuna caratteristica sia espressa sotto forma di una variabile di tipo *dummy*, come spiegato nel primo capitolo; è possibile tuttavia omettere dalla specificazione della relazione lineare una modalità per ogni attributo, selezionabile arbitrariamente. Pertanto il modello di regressione da utilizzare in questo caso è, ad esempio:

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + b_6X_6 + b_7X_7$$

dove:

Y indica la valutazione dell'intervistato su ciascun stimolo;

X_1 è la variabile di tipo *dummy* generata dalla modalità martedì dell'attributo giorno;

X_2 è la variabile *dummy* generata dalla modalità mercoledì dell'attributo giorno;

X_3 è la variabile *dummy* generata dalla modalità giovedì dell'attributo giorno;

X_4 è la variabile *dummy* generata dalla modalità 10-12 dell'attributo ora;

X_5 è la variabile *dummy* generata dalla modalità 12-14 dell'attributo ora;

X_6 è la variabile dummy generata dalla modalità 16-18 dell'attributo ora;

X_7 è la variabile *dummy* generata dalla modalità provença dell'attributo direzione.

Da come si evince sono state omesse le modalità venerdì per quanto riguarda il giorno, 18-20 per l'ora e rocafort per direzione. Poichè ogni stimolo è dato da una specifica combinazione delle 12 modalità di questi 3 caratteri, può essere riespresso, dopo un'opportuna ricodifica, nei termini delle sette variabili *dummy* appena indicate. Le codifiche assunte da tali variabili, relativamente ad un profilo di prodotto/servizio, sono le seguenti: 1 se nello stimolo si presenta proprio la modalità alla quale si riferisce la variabile di tipo *dummy*; 0 se nello stimolo è presente non la modalità alla quale si riferisce la variabile di tipo

dummy , ma quella che corrisponde ad un'altra *dummy* inserita nel modello; -1 se nello stimolo è presente la modalità che è stata omessa in sede di specificazione del modello e alla quale non corrisponde quindi alcuna *dummy*. Con riferimento al carattere giorno, per esempio, si avrà allora $X_1 = 1$, $X_2 = 0$ e $X_3 = 0$ se nello stimolo appare la modalità martedì; $X_1 = 0$, $X_2 = 1$ e $X_3 = 0$ se nello stimolo appare la modalità mercoledì; $X_1 = 0$, $X_2 = 0$ e $X_3 = 1$ se nello stimolo appare la modalità giovedì e $X_1 = -1$, $X_2 = -1$ e $X_3 = -1$ se nello stimolo appare la modalità venerdì. In maniera analogo si procede per tutte le altre variabili ricavando la matrice X del nostro modello:

STIMOLI	**VARIABILI DI TIPO DUMMY**						
	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1	-1	-1	-1	0	1	0	-1
2	0	1	0	0	1	0	1
3	-1	-1	-1	-1	-1	-1	-1
4	1	0	0	0	0	1	1
5	0	0	1	0	0	1	-1
6	-1	-1	-1	1	0	0	1
7	1	0	0	1	0	0	1
8	0	1	0	1	0	0	-1
9	-1	-1	-1	0	0	1	1
10	0	0	1	1	0	0	-1
11	0	1	0	-1	-1	-1	1
12	0	0	1	-1	-1	-1	1
13	1	0	0	0	1	0	-1
14	0	1	0	0	0	1	-1
15	1	0	0	-1	-1	-1	-1
16	0	0	1	0	1	0	1

ottenuto ciò si è proceduto ad una prima analisi disgiunta dei segmenti.

2.6 Analisi disgiunta dei segmenti

Iniziamo con un'analisi disgiunta dei dati. Come detto in precedenza abbiamo individuato due classi d'età, maggiori o minori di anni 40, che combinate con le due classi del sesso formano le quattro classi descritte precedentemente. La sintassi per un'analisi CA in SPSS è molto semplice e queste sono le linee di comando:

```
CONJOINT PLAN="C:\percorso file"
/DATA="C:\percorso file" /SCORE=SCORE1 TO SCORE16 /SUBJECT=ID
/FACTORS=giorno ora dir
/PLOT=ALL
/PRINT=ALL
```

SPSS in questo modo fornira un'analisi CA con stime sia individuali che complessive delle utilità parziali dai dati dei rispondenti, inoltre si otterranno i valori d'importanza dei fattori e dei grafici riassuntivi. Si dovrà specificare sia il file contenente il piano fattoriale costruito in SPSS sia il file contenente i dati dei rispondenti. Infine è possibile specificare la relazione che lega uno specifico fattore con i dati. La scelta cadrà fra queste quattro alternative:

- **Discreto** il fattore è una variabile categoriale e non si fanno assunzioni circa la relazione fra attributo e puntuazione. Se non si specifica il tipo di relazione il programma assume di default un modello discreto;
- **Lineare** il modello lineare indica una relazione lineare tra il fattore e i valori, si specifica, inoltre, MORE a indicare che valori alti del fattore suppongono preferenze alte o LESS viceversa;
- **Ideal** modello che indica una relazione quadratica tra il fattore e i dati al decrescere della preferenza;
- **Antiideal** in questo caso si ha una relazione quadratica tra fattore e i dati all'incremento della preferenza.

Si presentano le utilità parziali stimate per le modalità dei tre attributi su un rispondente del segmento uomini minori di anni 40:

Importanza	Attributi e modalità	Coefficienti di utilità parziale
	Giorno	
44.89	martedì	-0.2831
	mercoledì	1.5669
	giovedì	-0.1756
	venerdì	-1.1081
	Ora	
19.59	10-12	0.4069
	12-14	-0.7606
	16-18	0.0669
	18-20	0.2869
	Direzione	
35.51	provença	1.0581
	rocafort	-1.0581
	costante	5.7631
	R^2	0.780

le stime fornite dal programma non sono altro che i risultati ottenuti con il metodo dei minimi quadrati ordinari, così come spiegato nel primo capitolo. La codifica applicata alle variabili *dummy* è stata disegnata in modo da far sì che la somma delle utilità parziali riferite a tutte le modalità di uno stesso attributo sia pari a zero. Pertanto l'utilità parziale relativa alla modalità di ogni caratteristica, non esplicitamente inclusa nel modello di regressione sotto forma di variabile *dummy*, viene calcolata come complemento a zero della somma dei parametri stimati in corrispondenza delle modalità incluse. Il modello sopra

stimato presenta un R^2 molto buono, a significare che i valori stimati non si discostano molto dai valori osservati. Il rispondente presenta il più elevato punteggio di utilità nel giorno di mercoledì con fascia oraria dalle 10 alle 12 e direzione provença. Confrontando i valori d'importanza relativa dei fattori si rileva la bassa importanza dell'attributo "ora"; ciò fa intuire che il rispondente sia più influenzato nella scelta dall'attributo "giorno" e "direzione". Risultano più interessanti le stime medie del segmento che si presantano di seguito:

Importanza	Attributi e modalità	Coefficienti di utilità parziale
	Giorno	
42.72	martedì	-0.1246
	mercoledì	-0.4023
	giovedì	-0.4611
	venerdì	0.9879
	Ora	
39.91	10-12	-0.6748
	12-14	0.2892
	16-18	-0.0168
	18-20	0.4024
	Direzione	
17.38	provença	-0.2187
	rocafort	0.2187
	costante	5.0143
	R^2	0.823

l'R^2 di questo modello è molto buono. Le importanze relative dei fattori "ora" e "giorno" risultano pressoché simili, mentre l'importanza del fattore "direzione" è

basso; si intuisce che il presente attributo influisce poco nella scelta dei rispondenti. Le combinazioni dei livelli delle modalità che massimizzano l'utilità totale è:

Livelli dei fattori	**Coefficiente di utilità**
Giorno apertura: venerdì	0.9879
Ora apertura: 18-20	0.4024
Direzione: Provença	0.2187
Costante	5.0143
Punteggio globale teorico:	6.5101

nel caso dei maschi minori di anni 40 il punteggio teorico è 6.5101. E' comunque inportante sottolineare che poco avrebbe influito sul punteggio teorico lo scambio, per esempio nell'attributo ora, del livello 12-14; lo *score* totale, infatti, sarebbe sceso di soli 0.1132 in confronto, per esempio, dello scambio con il livello 10-12 che è il meno preferito.
Il programma SPSS inoltre ci offre la possibilità di effettuare un controllo di validità esterna dei risultati ottenuti attraverso la scelta di un **campione di conferma** (holdout sample). Ad ogni consumatore intervistato si può chiedere infatti di fornire giudizi di preferenza per qualche altra alternativa di prodotto/servizio selezionata tra quelle non comprese all'interno del disegno fattoriale frazionato (in genere non più di tre o quattro). I punteggi così raccolti non contribuiscono a determinare l'insieme dei dati su cui avviene la stima dei parametri del modello, e dunque non entrano in gioco nel calcolo delle utilità parziali. Proprio a partire dai valori stimati delle utilità *part-worth* si può ricostruire algebricamente il giudizio di preferenza che spetterebbe, se il modello applicato fosse valido, alle alternative che fanno parte del campione di conferma. Il confronto tra punteggi rilevati sul campo e punteggi previsti sulla base dei parametri stimati per tali alternative mette a disposizione una misura sul grado di coerenza del modello per ogni intervistato.
Per l'analisi dei dati il programma offre validi strumenti grafici che si presentano per l'analisi del segmento fin qui analizzato:

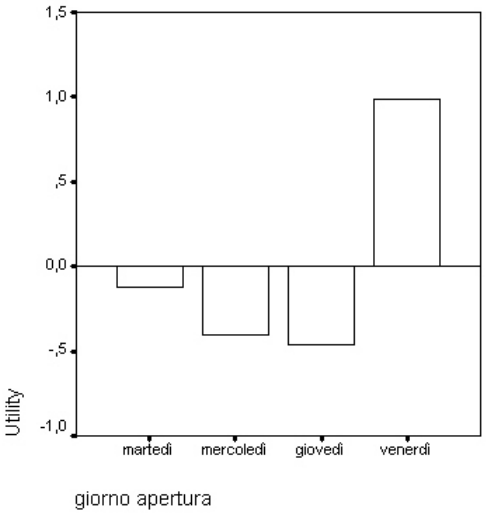


Figura 2.1: Utilità "giorno di apertura"

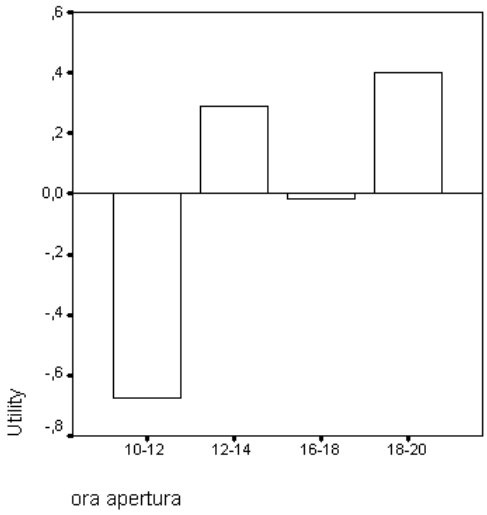


Figura 2.2: Utilità "ora di apertura"

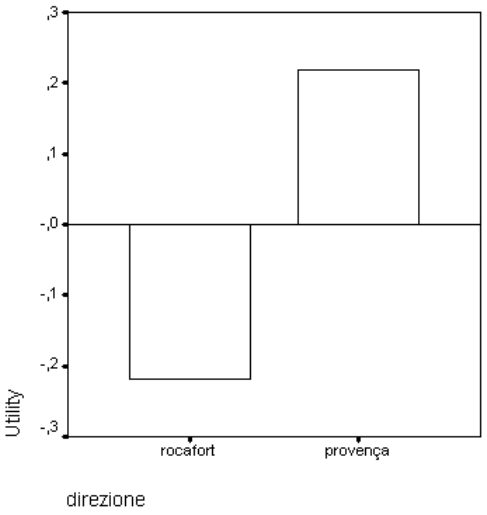


Figura 2.3: Utilità "direzione"

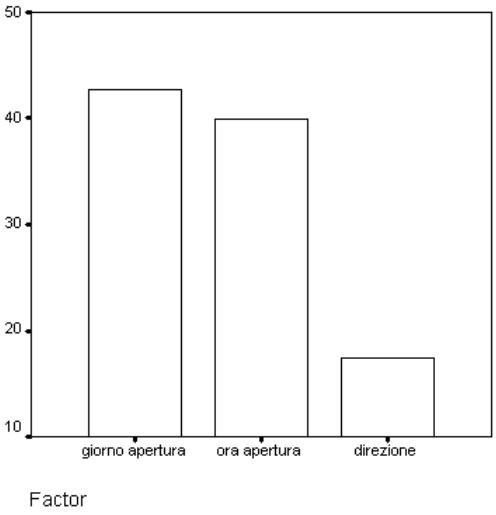


Figura 2.4: Importanza dei fattori

i primi tre grafici raffigurano le utilità medie di ognuno dei livelli di ciascun attributo. L'ultimo grafico rappresenta le importanze relative dei fattori. Passiamo alle utilità parziali dei vari profili:

Modalità degli attributi	utilità uomini<40	utilità uomini>40	utilità donne<40	utilità donne>40
Giorno	(42.72)	(58.15)	(52.44)	(56.36)
martedì	-0.1246	-0.3574	-0.126	-0.4040
mercoledì	-0.4023	-1.2052	-1.2568	-1.0670
giovedì	-0.4611	-0.1707	-0.3711	-0.9575
venerdì	0.9879	1.7333	1.6404	1.5685
Ora	(39.91)	(31.17)	(32.31)	(32.08)
10-12	-0.6748	-0.4997	-0.2631	-0.3338
12-14	0.2892	0.6306	0.5029	0.3840
16-18	-0.0168	-0.5544	-0.6843	-0.6760
18-20	0.4024	0.4236	0.4444	0.6258
Direzione	(17.38)	(10.68)	(15.25)	(11.56)
provença	-0.2187	-0.1463	-0.0081	-0.1601
rocafort	0.2187	0.1463	0.0081	0.1601
costante	5.0143	4.8662	4.7593	5.0745
R^2	0.823	0.983	0.871	0.949

in questo schema risultano chiare le differenze ma soprattutto le similitudini fra i quattro segmenti presi in analisi. Iniziamo la lettura notando che l'importanza relativa del fattore "direzione" presenta un valore basso in confronto al valore dei rimanenti due, a spiegarci che l'influenza del presente attributo, nella scelta di entrare o no del nostro negozio, è bassa o quasi nulla. Rimangono cruciali nella scelta del consumatore i due attributi "giorno di apertura" e "ora di apertura", essi sono quelli che maggiormente influiscono nella scelta del rispondente. Nel notare i valori dei quattro segmenti non vediamo grandi scostamenti se non leggere flessioni dei valori soprattutto sulle fasce orarie. I modelli stimati risul-

tano buoni grazie ad un valore degli R^2 superiori dell'80%, ma consideriamo i livelli degli attributi che massimizzano le utilità dei quattro segmenti:

Fattori	Livelli dei fattori e coefficienti di Utilità			
	uomini<40	uomini>40	donne<40	donne>40
Giorno apertura	venerdì	venerdì	venerdì	venerdì
	0.9879	1.7333	1.6404	1.5685
Ora apertura	18-20	12-14	12-14	18-20
	0.40240	0.63060	0.50290	0.62580
Direzione	provença	provença	provença	provença
	0.2187	0.1463	0.0081	0.1601
Costante	5.0143	4.8662	4.7593	5.0745
Punteggio globale	6.6233	7.3764	6.9107	7.4289

si evince che in tutti e quattro i segmenti viene prediletto il venerdì, esso è il livello che maggiormente influisce nel punteggio totale, grazie pure all'elevata importanza relativa dell'attributo che rappresenta. La "direzione" prediletta è "Provença" ma il suo apporta risulta quasi impercettibile. La differenziazione dei segmenti è data in modo pressoché impercettibile dall'attributo "ora". Di fatto esso varia tra i segmenti senza però influire pesantemente sull'utilità teorica finale. Si passa ad una successiva analisi congiunta ipotizzando un andamento delle utilità in linea con i dati presentati in questa prima analisi.

2.7 Analisi congiunta dei segmenti

Se nel paragrafo precedente si analizzavano i dati a seconda della loro classificazione per età e sesso ora, con l'analisi congiunta, si raggrupperanno tutte le osservazioni. In questo modo potremo notare le preferenze globali di tutti i rispondenti senza distinzione di sesso o età. L'analisi aggregata dei dati, effettuata con la stessa procedura in SPSS, trova le stime, nel nostro caso, dei 40 soggetti intervistati e delle utilità medie globali. Non trovando grandi differenze tra i segmenti nel precedente paragrafo è facile prevedere che il modello produca stime che non si discostano in modo significativo dalle precedenti. Si presentano qui di seguito le utilità medie:

Importanza	Attributi e modalità	Coefficienti di utilità parziale
	Giorno	
52.42	martedì	-0.2246
	mercoledì	-0.9828
	giovedì	-0.2751
	venerdì	1.4825
	Ora	
33.87	10-12	-0.4428
	12-14	0.4517
	16-18	-0.4829
	18-20	0.4740
	Direzione	
13.71	provença	-0.1333
	rocafort	0.1333
	costante	4.9286
	R^2	0.949

lo scenario non appare, in effetti, molto diverso da come si era previsto. Il giorno preferito era e rimane il "venerdì" così come la direzione "provença". L'attributo "ora", che in precedenza mi permetteva la differenziazione dei livelli, trova la massima utilità nel livello "18-20" seppur con differenza quasi impercettibile con l'utilità del livello "12-14". Il modello appare molto buono con un R^2 quasi al 95%. L'attributo più significativo resta il giorno mentre la direzione risulta quasi ininfluente nella valutazione totale, infatti le utilità parziali risultano prossime allo zero. Lo scenario che massimizza l'utilità totale è il seguente:

Livelli dei fattori	**Coefficiente di utilità**
Giorno apertura: venerdì	1.4825
Ora apertura: 18-20	0.4740
Direzione: Provença	0.1333
Costante	4.9286
Punteggio globale teorico:	7.0184

il punteggio teorico globale è 7.0184. Da rilevare che poco avrebbe influito sul punteggio finale la scelta del livello "12-14".

2.8 Conclusioni finali

Per concludere il capitolo si presentano qui di seguito alcune osservazioni rilevanti sull'analisi esposta. In primo luogo è da rilevare l'atipicità della CA qui svolta, di fatto, l'oggetto di studio non si presenta sotto forma di bene o servizio atti al soddisfacimento diretto di un bisogno del cliente. Il presente lavoro appare finalizzato ad un uso direttivo e di ottimizzazione di risorse umane. L'azienda, preso atto dei dati fin qui analizzati, potrà operare strategie in campo manageriale quale l'alloccazione efficiente di personale nel negozio o politiche flessibili nell'orario di apertura del locale stesso. E' utile notare che l'elaborato si presta come punto base per una stima potenziale delle vendite divise per giorno

e fasce orarie. Per esempio, nel nostro caso, è facile prevedere più vendite di venerdì rispetto al martedì. L'analisi svolta presenta, quindi, molteplici utilità che spaziano da una semplice domanda che l'azienda potrebbe porsi ("quando mi conviene aprire il negozio e invece quando mi è utile chiuderlo?") fino ad arrivare al desiderio di efficienza che si aspetta un cliente quando entra in un esercizio commerciale. Se il modello pretende (seppur nei limiti della CA presentati nel capitolo primo) di stimare affluenze di clienti fittizie nel citato negozio è pur vero che lo stesso modello non prevede stagionalità o ciclicità dei dati. Nel periodo, per esempio, dei saldi si dovrà tenere in considerazione un afflusso mediamente superiore rispetto ad un periodo estivo. Come ogni modello statistico dovremo quindi interpretare con un occhio critico le realtà che lo stesso intende prevedere.

Capitolo 3

La Choice-based Conjoint (CBC)

Nei primi due capitoli si è parlato dello sviluppo classico della *conjoint analysis*, ovvero con raccolta dei dati full-profile. Nella CA i rispondenti sono chiamati a valutare una serie di profili possibili di un certo prodotto/servizio oggetto di studio. La *Choice-based Conjoint* (d'ora in poi CBC) rientra in uno degli ultimi sviluppi della CA, in essa i rispondenti operano una vera e propria scelta fra diversi prodotti. Quindi se in una CA tradizionale i dati rappresentano punteggi (solitamente rappresentati su scala ad intervalli), nella CBC i dati sono delle vere e proprie scelte fra profili di prodotto diversi.
Recentemente questa metodologia ha attirato molto gli interessi del marketing per le seguenti ragioni:

- grande capacità di simulare il processo di scelta che accompagna il compratore, infatti la scelta di un prodotto fra un gruppo di diverse alternative è una semplice azione che frequentemente tutti noi facciamo al momento dell'acquisto;

- la CBC include fra le possibili opzioni una non-scelta lasciando, quindi, al rispondente la possibilità di non effettuare una scelta forzata fra i prodotti presentati. Il contributo offerto da tale opzione farà decrescere le utilità attese finali;

- le metodologie di CA più tradizionali operano l'assunzione semplificatrice

che solo gli effetti principali influiscano sull'utilità del prodotto, tuttavia siccome la CBC analizza i dati ad un livello "aggregato" e non individualmente per ogni rispondente, è possibile quantificare, oltre agli effetti principali, pure le interazioni;

- l'analisi dei dati risulta più semplice rispetto ad altri modelli di CA, infatti, effettuando semplici statistiche di base si possono già fare assunzioni sulle preferenze dei rispondente.

Tuttavia la CBC non è esente da limiti poichè le scelte dei rispondenti, così ottenute, risultano inefficienti per dedurre le preferenze. Ogni concetto di prodotto è descritto da tutti gli attributi considerati nello studio, e ogni set di scelte contiene diversi concetti. Il rispondente prima di dare una risposta per ogni set è venuto, quindi, a conoscenza di una quantità di informazioni superiore a quella disponibile nella realtà. Per questa ragione gli studi di CBC non sono tradizionalmente usati per stimare i valori di utilità parziale per ogni rispondente che, invece, generalmente vengono stimate con la CA tradizionale. Nella CBC i dati vengono raggruppati per specifici segmenti di mercato e i valori di utilità, prodotti per ogni gruppo, rappresentano mediamente le scelte fatte dagli individui che lo compongono e , come accade in altre metodologie di CA, gli stessi valori sono usati per simulare e predire le reazioni dei rispondenti ai profili di prodotto che non sono potuti apparire nel piano sperimentale. Nelle analisi CBC si assume che i rispondenti di un certo segmento rispondano in modo omogeneo, ciò può risultare a volte poco appropriato o non desiderabile. Sviluppi innovativi di CBC hanno riconosciuto eventuali eterogeneità di risposta fra individui di uguali segmenti, inoltre sono stati introdotti nuovi modelli di stima che riescono a produrci valori di utilità individuali.
La CBC non è appropiata per studi con un elevato numero di attributi. Ogni profilo presenta vari livelli interamente descritti dagli attributi dei prodotti, l'eccesso di informazione causa un processo di confusione nei rispondenti e, quindi, ad un sovraccarico di lavoro che può pregiudicare l'analisi finale; è bene limitare gli attributi ad una numerosità non superiore a sei.

3.1 Raccolta dei dati

La CBC si avvale di una raccolta dati molto differente rispetto alla CA più tradizionale, ciò implica pure che l'analisi degli stessi è diversa. Il questionario di una analisi CBC solitamente viene somministrato tramite PC, ma ciò non esclude la possibilità di uno sviluppo dello stesso per via cartacea. Ora presento un possibile scenario che può presentarsi ad un intervistato in un questionario:

Quale dei seguenti scenari di prodotto preferisce?

marca A prezzo 1 garanzia 1 anno	marca A prezzo 2 garanzia 3 anni	marca C prezzo 1 garanzia 3 anni	nessuno dei tre

nel nostro caso il candidato dovrà barrare la casella che ritiene preferibile fra le quattro possibili scelte.
I primi passi della CBC sono simili a quelli della CA infatti si dovrà, in un primo momento, selezionare gli attributi con i relativi livelli, successivamente si stabilirà un piano sperimentale e per finire si redigerà il questionario per poi somministrarlo ad un campione della popolazione obbiettivo. La stesura del questionario sarà importante, infatti se in una CA tradizionale lo stesso veniva redatto in relazione al piano fattoriale prestabilito, in una analisi CBC dovremo sì avvalerci del piano sperimentale ma dovremo, allo stesso tempo, aver molta cura nella scelta del numero di opzioni da sottoporre ad ogni passo del questionario. Riferendoci all'esempio precedente abbiamo optato che il rispondente possa scegliere fra tre possibili profili di prodotto più una "non"-scelta. I due aspetti che dovremo tener conto nella redazione del questionario sono:

- numero di profili di prodotto che si presenta al rispondente ad ogni "step";
- numero totale di "step" da presentare al rispondente.

Regole precise che possano aiutarci non ci sono. Di fatto, solo il proseguo dell'analisi può esserci d'aiuto a tracciare delle linee guida per comprendere le

conseguenze causate dalla scelte precedentemente fatte.
Da un punto di vista statistico la presentazione dei dati in questo modo non rappresenta un metodo efficace per tracciare le preferenze dei consumatori. Infatti, il consumatore valuta una molteplicità di profili di prodotto diversi, ma l'informazione che otterremo sarà "solo" sulla scelta da lui effettuata. In questo modo non sapremo quanto grande sarà questa sua scelta nei confronti dei profili di prodotto non presenti in quel prefissato item del questionario. L'informazione sarà maggiore quanto maggiore è il numero di possibili scelte che si presentano al rispondente. Questo purtroppo si scontra con gli interessi di un questionario breve e di facile lettura per i rispondenti. In generale si è stabilito che il numero di profili di prodotto oggetto di scelta non dovrebbe essere più di cinque, ma non meno di tre.
Stabilita la numerosità delle scelte poste ad ogni "step" del questionario ora è importante individuare il numero totale di item del questionario, ovvero quanti passi di diversi profili di prodotto si dovrà somministrare al rispondente. Anche in questo caso la risposta non è univoca e solitamente ci si avvale delle risposte date nella CA tradizionale. In questo punto si dovrà guardare il piano sperimentale scelto e quindi la numerosità degli attributi e dei relativi livelli del prodotto/servizio oggetto di studio. In generale è bene non sottoporre più di venti cicli di scelta ma non meno di dodici.
E' importante inoltre notare la possibilità di implementare come possibile scelta una "non"-scelta, ossia una risposta a cui il candidato può optare se non trova di suo gradimento i prolili di prodotto presentati. Grazie alla presente scelta si riesce a riflettere molto bene il mondo reale, infatti il rispondente non è chiamato a rispondere se i prodotti non lo soddisfano. In letteratura viene chiamata *alternativa costante* in quanto, al momento dell'analisi dei dati, questa viene presentata come una scelta a se stante e quindi costante.

3.2 Analisi dei dati per "*Counting Choices*"

Finita la raccolta dei dati si passa ad una prima analisi degli stessi. La *Counting Choices* è probabilmente la metodologia più semplice e intuitiva ed infatti si avvale di statistiche base descrittive. Essa calcola la proporzione di ciascun livello, basandosi sulle frequenze di scelta dei profili di prodotto che lo contengono, diviso le volte che lo stesso livello è incluso nei profili somministrati ai rispondenti. In pratica tramite questa procedura si ottiene la proporzione di scelta di ciascun attributo ma anche le proporzioni di scelta congiunte di due o più attributi. A livello statistico ci vengono fornite una serie di tabelle di contingenza dove gli attributi del prodotto sono le variabili rappresentate con le relative proporzioni di scelta e l'indice χ^2 che ci permette di stabilire l'influenza dello specifico attributo o dell'interazione nel modello. Per esempio, ipotizzato un prodotto univocamente descritto da due attributi cioè il prezzo e la marca, all'analista si potrà presentare:

Marca	%	**Prezzo**	%
A	0.387	\$ 4.00	0.132
B	0.207	\$ 3.50	0.175
C	0.173	\$ 3.00	0.254
D	0.165	\$ 2.50	0.372

χ^2(Marca)= 148,02 df=3 p<.01 χ^2(Prezzo)=151.32 df=3 p<.01

nel seguente caso si leggera che il 38.7% dei rispondenti ha scelto la marca A e solo il 16.5% la marca D si dirà inoltre che il prezzo \$2.50 è scelto dal 37.2% dei rispondenti contro il 13.2% del prezzo \$4.00. La statistica χ^2 ci permette di dire che entrambi gli attributi influiscono significativamente nella scelta del prodotto, ad un livello del 99%. Se fissiamo le frequenze di un nostro attributo $f_1, \ldots, f_d$, in questo caso $d = 4$, realizzazioni di $Mn_d(n, \pi)$, (ovvero realizzazioni di variabili casuali multinomiali con n osservazioni e d livelli) il nostro obbiettivo è di saggiare la seguente ipotesi nulla:

$$H_0 : \pi_1 = \pi_1^0, \ldots, \pi_{d-1}^0 = \pi_{d-1}^0, \tag{3.1}$$

dove $\pi_1^0, \ldots, \pi_{d-1}^0$ sono probabilità positive assegnate, con $\sum_{i=1}^{d-1} \pi_i^0 < 1$, contro l'ipotesi alternativa che almento una delle $d-1$ uguaglianze non sia soddisfatta. La statistica **log-rapporto di verosimiglianza** ci permette di testare l'ipotesi sopra citata, essa si presenta, in generale, così:

$$W(\pi^0) = 2\{l(\hat{\pi}) - l(\pi^0)\},$$

essa è la log-verosimiglianza calcolata per $\pi = \pi^0$ moltiplicata per -2. Chiaramente, valori grandi di $W(\pi^0)$ sono critici per l'ipotesi nulla. Sotto condizioni di regolarità la distribuzione nulla asintotica di $W(\pi^0)$ è chi-quadrato con $d-1$ gradi di libertà, dove $d-1$ è il numero di componenti scalari meno uno di π. Il test di verosimiglianza nel nostro caso risulta:

$$W(\pi^0) = 2\sum_{i=1}^{d} f_i log\frac{\hat{\pi}_i}{\pi_i^0}.$$

In pratica il test dice che le varie marche non sono ugualmente scelte. Ora si presentano i valori delle interazione dei due attributi:

	$4.00	**$3.50**	**$3.00**	**$2.50**
Marca A	0.262	0.320	0.398	0.370
Marca B	0.083	0.146	0.254	0.347
Marca C	0.104	0.100	0.163	0.221
Marca D	0.078	0.129	0.206	0.149

χ^2=14.99 df=9 non significativo.

Nel presente caso si leggerà che il 37% dei rispondenti hanno scelto la combinazione di prodotto di marca A ad un prezzo di $2.50 e invece solo il 7.8% dei rispondenti hanno scelto il prodotto di marca D ad un prezzo di $4.00. La statistica χ^2 nel presente caso ci saggerà l'ipotesi nulla di indipendenza stocastica fra le due variabili (nel nostro caso attributi del prodotto). Quindi l'ipotesi nulla è data da:

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j},\ i = 1, \ldots, r,\ j = 1, \ldots, c. \tag{3.2}$$

L'ipotesi alternativa è che almeno una delle uguaglianze nella (3.2) non sia soddisfatta. Sotto l'ipotesi nulla (3.2) tutte le r distribuzioni condizionate di riga π_{ij}/π_{i+} sono tra loro identiche e uguali alla distribuzione marginale $\pi_{+j}, j = 1, \ldots, c$. Così pure, sono tra loro identiche e uguali alla distribuzione marginale π_{i+} le c distribuzioni condizionate di colonna $\pi_{ij}\pi{+}j, i = 1, \ldots, r$. La (3.2) è un caso particolare della (3.1) dove le nostre probabilità sono espresse come funzioni regolari di un parametro θ con dimensione p inferiore a $d - 1$. Propiamente $\theta = (\pi_{1+}, \pi_{2+}, \ldots, \pi_{r-1+}, \pi_{+1}, \pi_{+2}, \ldots, \pi_{+c-1})$ con dimensione $r+c-2$. Le stime di massima verosimiglianza di π_{i+} e di π_{+j} sono f_{i+}/n e f_{i+}/n, rispettivamente. La statistica χ^2 assume quindi la forma

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(f_{ij} - f_{i+}f_{+j}/n)^2}{f_{i+}f_{+j}/n}$$

e la distribuzione asintotica nulla è $\chi^2_{(r-1)(c-1)}$. Il numero di gradi di libertà si ottiene ricordando che in questo caso il numero di celle della multinomiale è $d = rc - 1$, mentre la dimensione di θ è $p = r + c - 2$. Nel caso presentato il valore della statistica non si presenta significativo e quindi l'interazione dei due attributi non influisce nella scelta del prodotto grazie alla saggiata conformità all'indipendenza stocastica delle due variabili.
Come vediamo da questo semplice esempio, l'analisi dei dati attraverso la *Counting Choices* risulta molto intuitiva e facile da calcolare ma non risponde agli obbiettivi che ci siamo posti. Gli effetti degli attributi non sono calcolabili e quindi è impossibile, per esempio, prevedere il consenso attribuibile ad un specifico prodotto non presente nel piano sperimentale.

3.3 Analisi dei dati "*Logit*"

In questa metodologia, di più ampia complessità rispetto al precedente metodo di analisi, si suppone che ogni alternativa, i, in un set di numerosità prefissata, I, abbia un certo livello di preferibilità funzione lineare delle utilità parziali β dei livelli presenti in ciascun attributo k. Posta x_{ik} la matrice di variabili dummy che ci identificano i livelli in una determinata alternativa, la probabilità

che un individuo scelga l'opzione i dato un certo set è:

$$\hat{\pi}_i = \frac{e^{\sum_{k=1}^{k=K} \beta_k x_{ik}}}{\sum_{i=1}^{i=I} e^{\sum_{k=1}^{k=K} \beta_k x_{ik}}}, \tag{3.3}$$

ovvero non è altro che l'esponente dell'utilità propria diviso le somme delle utilità di tutte le alternative presenti nel set. Esprimiamo la (3.3) con l'uso del logaritmo:

$$\log\Big[\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\Big] = \sum_{k=1}^{K} \beta_k x_{ik}.$$

Ci troviamo, quindi, di fronte ad un modello lineare generalizzato (mlg). Partiamo da un risultato del modello lineare:

$$\mathrm{E}(Y) = \mu = \sum_{k=1}^{K} \beta_k x_k,$$

dove Y è la variabile dipendente. Ora definiamo η, parametro lineare prodotto da $x_1, x_2, \ldots, x_k$, come:

$$\eta = \sum_{k=1}^{K} \beta_k x_k.$$

In un modello lineare si avrà:

$$\eta = \mu$$

mentre in un mlg η è una funzione di μ, ovvero $\eta = g(\mu)$. In un modello logistico binomiale si ha:

$$\eta = \log[\mu/(1-\mu)]. \tag{3.4}$$

nel modello multinomiale logit si ha:

$$\eta_j = \log(\mu_j/(1-\mu_j)), \tag{3.5}$$

dove j rappresenta la j-esima variabile categoriale. Come vediamo la (3.5) non è altro che una generalizzazione della (3.4). Deduciamo che per la stima delle utilità ci serviamo di un modello logistico multinomiale. Ora il nostro obbiettivo è di spiegarne il motivo. Per iniziare, quindi, vediamo come appaiono i dati.

Ipotizzando che un prodotto sia univocamente descritto da due attributi chiamati X, Y con generiche modalità x_i, y_j si avrà la seguente tabella di frequenza (pur detta tabella di *contingenza*):

X Y	y_1	$\cdots$	y_j	$\cdots$	y_c	totale
x_1	f_{11}	$\cdots$	f_{1j}	$\cdots$	f_{1c}	f_{1+}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_i	f_{i1}	$\cdots$	f_{ij}	$\cdots$	f_{ic}	f_{i+}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_r	f_{r1}	$\cdots$	f_{rj}	$\cdots$	f_{rc}	f_{r+}
totale	f_{+1}	$\cdots$	f_{+j}	$\cdots$	f_{+c}	n

dove r è il numero di righe, c il numero di colonne, e f_{ij}, $i = 1, \ldots, r, j = 1, \ldots, c$, è la generica frequenza congiunta. Inoltre, $f_{i+} = \sum_{j=1}^{c} f_{ij}, i = 1, \ldots, r, f_{+j} = \sum_{i=1}^{r} f_{ij}, j = 1, \ldots, c$, sono le generiche frequenze marginali. Infine, $n = \sum_{i=1}^{r}\sum_{j=1}^{c} f_{ij} = \sum_{i=1}^{r} f_{i+} = \sum_{j=1}^{c} f_{+j}$. Fissato il nostro n e posto $d = rc$ la distribuzione del fenomeno segue un modello multinomiale quale:

$$p(f_{11}, \ldots, f_{rc}; \pi) = \frac{n!}{f_{11}! \cdots f_{rc}!} \pi_{11}^{f_{11}} \cdots \pi_{rc}^{f_{rc}},$$

con $0 < \pi_{ij} < 1$ per $i = 1, \ldots, r, j = 1, \ldots, c$ e $\sum_{i=1}^{r}\sum_{j=1}^{c} \pi_{ij} = 1$. Infine con $\pi_{i+} = \sum_{j=1}^{c} \pi_{ij}$ e $\pi_{+j} = \sum_{i=1}^{r} \pi_{ij}$ si indicano le probabilità marginali di riga e di colonna. La distribuzione di Y=$(Y_1, \ldots, Y_d)$ si dice multinomiale con indice n ($n = 1, 2, \ldots$) e vettore di probabilità $\pi = (\pi_1, \ldots, \pi_d)$ ($0 < \pi_i < 1$ per $i = 1, \ldots, d$ con $\sum_{i=1}^{d} \pi_i = 1$), e si scrive sinteticamente $Y \sim Mn_d(n, \pi)$, se Y è discreta con supporto:

$$S_Y = \{y = (y_1, \ldots, y_d) \in \mathbb{N}^d : \sum_{i=1}^{d} y_i = n\}$$

e funzione di probabilità, per $y \in S_Y$,

$$p_Y(y) = \frac{n!}{y_1! \cdots y_d!} \pi_1^{y_1} \cdots \pi_d^{y_d}, \tag{3.6}$$

essa rappresenta una generalizzazione di una v.c. binomiale a cui si riconduce per $d = 2$. Per l'analisi statistica interessa un modello che mostri come varia la probabilità π che y si manifesti al variare di una certa variabile esplicativa x. Un modello di regressione lineare posto direttamente sulle probabilità,

$$\pi_i = \theta_1 + \theta_2 x_i \qquad i = 1, \ldots, d$$

è in generale inadeguato, perchè conduce all'assurdo di poter indicare per certi eventi probabilità negative, e per altri probabilità maggiori di 1. Un modello di regressione lineare posto sui logaritmi delle quote di scelta, come ad esempio

$$\eta_i = \beta_1 + \beta_2 x_i \qquad i = 1, \ldots, d$$

dove $\eta_i = \log \frac{\pi_i}{1-\pi_i}$ è libero dall'obiezione appena sollevata, perchè il logaritmo della quota può assumere valori in tutto $\mathbb{R}$ e lo spazio parametrico per $\beta = (\beta_1, \beta_2)$ può essere assunto semplicemente $\mathbb{R}^2$. Abbiamo spiegato il motivo di un modello logistico applicato ai nostri dati supposti come eventi casuali generati da una distribuzione multinomiale.

3.3.1 Stima degli effetti dei livelli degli attributi

La stima degli effetti dei livelli degli attributi avviene tramite il calcolo dei $\hat{\beta}$ dell'equazione (3.3). Come è intuibile il compito non risulta facile. Infatti le equazioni di verosimiglianza della (3.3) non si prestano ad una soluzione esplicita e bisogna ricorrere a metodi di calcolo numerico. La funzione di log-verosimiglianza dedotta dalla (3.6) è:

$$l(\pi_1, \ldots, \pi_{d-1}; f_1, \ldots, f_d) = \sum_{i=1}^{d-1} f_i \log \pi_i + f_d \log(1 - \pi_1 - \ldots - \pi_{d-1}). \quad (3.7)$$

Dalla (3.7) si deduce che lo stimatore di massima verosimiglianza per π_i è:

$$\hat{\pi_i} = \frac{f_i}{n} \qquad i = 1, \ldots, d-1,$$

quindi, la funzione di massima verosimiglianza del modello risulta:

$$l_M = \sum_i (y_i \log \pi_i) + c$$

$$l_M = \sum_i y_i \log \mathbb{E}\{Y_i|n\} + c$$

$$l_M = \sum_i (y_i \log \mu_i - y_i) + c \tag{3.8}$$

tenuto presente che $\mathbb{E}\{Y_i|n\} = n\pi_i \propto \mu_i$. Posto quindi che $\mu_i = e^{\sum_i \beta_i x_i} = e_i^{\eta}$ la stima dei nostri β avverrà tramite un procedimento iterativo dove il vettore di partenza non è altro che il vettore nullo. Di passo in passo i β veranno incrementati di un vettore gradiente finchè i valori faranno convergere la (3.8) o, espresso in altri termini, il risultato del sucessivo passo porterà una variazione della verosimiglianza ininfluente.

3.4 Hierarchical Bayes Analysis

Presento in questo paragrafo l'ultimo modello di analisi CBC trattata in quest'elaborato. La Hierarchical Bayes (d'ora in poi HB) si avvale del modello gerarchico o a stadi di natura Bayesiana.
La principale differenza fra impostazione classica e bayesiana risiede, in sostanza, nell'uso della regola di Bayes e nelle conseguenze che ne derivano; nell'impostazione classica si ha un rifiuto del teorema di Bayes poichè l'interpretazione che si assegna della probabilità ha un significato puramente oggettivo e quindi legato alla nozione di frequenza. Si conclude basando il ragionamento induttivo riferendosi solo alla verosimiglianza e/o alle propietà dello spazio campionario. L'inferenza bayesiana invece si basa sul legame tra verosimiglianza e distribuzione finale che, insieme alle probabilità "iniziali", è il fulcro del teorema di Bayes.

3.4.1 Il modello statistico bayesiano

Un modello statistico probabilistico per un esperimento $e \in \xi$ è una terna $(\mathcal{Z}, A_{\mathcal{Z}}, \mathcal{P})$, dove $\mathcal{Z}$ è lo spazio campionario, $A_{\mathcal{Z}}$ è un'opportuna σ-algebra, di sottoinsiemi di $\mathcal{Z}$ che rende misurabile lo spazio campionario e $\mathcal{P}$ una famiglia

di misure di probabilità sullo stesso spazio misurabile. I dati z, che si ricavano da e, una volta che l'esperimento è stato condotto, costituiscono la premessa per inferire su $\mathcal{P}$. Le misure di probabilità $P \in \mathcal{P}$ possono essere indicizzate da un parametro θ, $\theta \in \Omega$:

$$\mathcal{P} = \{P_\theta^z \ : \ \theta \in \Omega\},$$

cioè ad ogni $\theta \in \Omega$ è associata una misura di probabilità P_θ, che assegna la probabilità ai membri di $A_{\mathcal{Z}}$, dove Ω è lo spazio dei parametri; θ può essere un numero reale, una n.upla ordinata di numeri reali, una funzione di ripartizione nel caso non parametrico, od altro.

E' ovvio che se non si avessero incertezze sui possibili valori di Ω, vale a dire Ω contiene un solo elemento non si avrebbe alcun problema statistico, poichè saremmo in grado di calcolare la probabilità di un qualsiasi evento connesso con l'esperimento. Supposto che le misure di probabilità siano dominate da una misura μ σ-finita, esse potranno essere descritte mediante la funzione di densità $p(z|\theta)$ rispetto a questa misura.

Di conseguenza se A è un qualunque evento della σ-algebra $A_{\mathcal{Z}}$ avremo

$$P(A|\theta) = \int_A p(z|\theta) d\mu(x) \qquad \forall \ A \in A_{\mathcal{Z}}, \forall \ \theta \in \Omega.$$

Si è, quindi, definito il modello statistico-probabilistico parametrico.

Nel modello bayesiano il riferimento iniziale cadrà su una variabile aleatoria (Θ, Z) con realizzazioni $(\theta, z) \in \Omega \times \mathcal{Z}$. Si assegna una legge di probabilità a *priori* su (Ω, A_Ω), che determina una sola legge di probabilità congiunta Ψ su $(\Omega \times \mathcal{Z}, A_\Omega \times A_{\mathcal{Z}})$, dove A_Ω è una opportuna σ-algebra di sottoinsiemi di Ω. Il modello bayesiano si presenterà in questo modo:

$$(\Omega \times \mathcal{Z}, A_\Omega \times A_{\mathcal{Z}}, \{\Psi(\theta, z)|(\theta, z) \in \Omega \times Z\}), \Omega \times \mathcal{Z} \subseteq \mathcal{R}^k \times \mathcal{R}^n (k, n \geq 1).$$

Le componenti del modello bayesiano sono:

e=$(\mathcal{Z}, A_{\mathcal{Z}}, P_\theta, \theta \in \Omega)$, dove $\{P_\theta, \theta \in \Omega\}$ è da intendersi come insieme delle leggi di probabilità condizionata di $Z/\Theta = \theta$.

(Ω, A_Ω, π), spazio di misura per Θ, con π legge di probabilità a *priori*.

Assunto che le leggi di probabilità componenti si esprimano attraverso densità

si definisce applicando il teorema di Bayes la legge di probabilità condizionata di $\Theta/Z = z$ o legge di probabilità *finale*:

$$\pi(\theta|z) = \frac{\pi(\theta)p(z/\theta)}{\int_\Omega \pi(\theta)p(z/\theta)d\theta}.$$

La legge di probabilità marginale (non condizionata) di $\mathcal{Z}$, nota come legge predittiva iniziale, risulta:

$$m(z) = \int_\Omega \Psi(\theta, z)d\theta = \int_\Omega \pi(\theta)p(z/\theta)d\theta.$$

La legge di probabilità marginale (non condizionata) di Θ, nota come legge a *priori*, è data da:

$$\pi(\theta) = \int_\mathcal{Z} \Psi(\theta, z)dz = \int_\mathcal{Z} \pi(\theta)p(z/\theta)dz$$

$$\pi(\theta/z) = \frac{1}{m(z)}\pi(\theta)p(z/\theta)$$

Posto c=$\frac{1}{m(z)}$

$$\pi(\theta/z) = c\pi(\theta)p(z/\theta)$$

ma essendo realizzato z nella fase post-sperimentale si avrà:

$$\pi(\theta/z) = c\pi(\theta)l(\theta)$$

$$\propto \text{(densità iniziale)x(verosimiglianza).}$$

La procedura bayesiana ha carattere dinamico poichè un nuovo esperimento trasforma la precedente legge finale in legge iniziale creando un aggiornamento informativo.

3.4.2 Il Modello Gerarchico

Il modello "gerarchico" rappresenta un particolare modello statistico bayesiano, $(f(x|\theta), \pi(\theta))$, dove la distribuzione a priori $\pi(\theta)$ è decomposta in distribuzioni condizionate ossia:

$$\pi_1(\theta|\theta_1), \pi_2(\theta_1|\theta_2), \ldots, \pi_n(\theta_{n-1}|\theta_n)$$

e $\pi(\theta)$ è ottenuta come marginale:

$$\pi(\theta) = \int_{\Theta_1 \times \ldots \times \Theta_n} \pi_1(\theta|\theta_1)\pi_2(\theta_1|\theta_2)\ldots\pi_{n+1}(\theta_n)d\theta_1\ldots d\theta_{n+1},$$

dove θ_i sono detti iperparametri dei livelli i con $1 \leq i \leq n$. Nel problema sotto studio fissando l'attenzione sull'i-esimo individuo la modellazione verrà effettuata in due livelli:

- ad un primo livello i nostri dati, come già spiegato nel precedente paragrafo, sono eventi generati da una distribuzione multinomiale logistica con funzione di probabilità data da:

$$p_k = \frac{\exp(\beta_i x_k)}{\sum_j \exp(\beta_i x_j)}$$

  dove:

  p_k è la probabilità che un singolo individuo scelga un particolare concetto di prodotto tra un set di risposte;

  x_j è la matrice delle variabili che descrivono i livelli della j-esima alternativa in quel set di risposte;

  β_i è il vettore delle utilità parziali dell'i-esimo individuo.

- ad un secondo livello i parametri β, che rappresentano le utilità parziali di un singolo individuo sono considerate un vettore di variabili aleatorie che si assume distribuito come una normale ed è rappresentato:

$$\beta_i \sim \text{Normal}(\alpha, \text{D})$$

  dove:

  α è il vettore delle medie della distribuzione delle utilità parziali di ciascun individuo

  D è la matrice di varianze e covarianze delle utilità parziali fra gli individui.

Il problema, quindi, si focalizzerà sulla stima dei parametri che governano le distribuzioni sopra citate.

3.4.3 Metodo iterativo di stima dei parametri

La stima dei parametri, quindi, avviene tramite un processo iterativo. Il presente metodo ci permette, per quanto possibile, la convergenza verso i veri parametri. Prima di tutto si identificano i valori iniziali:

- le stime iniziali di β sono le stime **OLS**, dove la variabile dipendente è la variabile dicotomica 1 e 0;
- le stime di α sono in media i β iniziali;
- la matrice D è la stima di varianze e covarianze dei β iniziali.

Premesso ciò il processo iterativo si sintetizza in questo modo:

- usando le presenti stime di β e D si genera una nuova stima di α. Si assume che α sia distribuita normalmente con media uguale alle medie di β e matrice di covarianze uguale a D diviso il numero dei rispondenti. La nuova stima di α viene calcolata tramite la citata distribuzione;
- usando la nuova stima di β e α, si calcola una nuova stima di D come sarà esposto nel paragrafo successivo;
- utilizzando le stime di α e D si generano nuove stime di β, la procedura si avvale del "Metropolis-Hastings Algorithm" (approfondimenti nel proseguo). Successivamente si incrementano i β finchè il modello non risulti sempre più adeguato ai dati. Quando ciò accade il processo iterativo converge.

Ad ogni passo si ristima un set di parametri (α,D o β) così permettendo il calcolo delle stime degli altri due set. La presente tecnica è chiamata 'Gibbs sampling', e converge alla distribuzione esatta dei tre set di parametri.

3.4.4 Stima degli alpha e di D

Il vettore degli alpha non è altro che un vettore di valori casuali distribuiti con media pari alla media dei correnti beta e con matrice di covarianza $\frac{1}{n}D$.

Il presente vettore, quindi, può essere calcolato grazie alla seguente procedura. Prendiamo α il vettore delle medie della distribuzione con D la sua matrice di covarianza. D può essere espresso come prodotto TT' dove T è una matrice triangolare inferiore che dicesi decomposizione di Cholesky della matrice D. Ora consideriamo il vettore u normalmente e indipendentemente distribuito con media zero e varianza unitaria, definiamo $v = Tu$. Per un'alta numerosità di n, $\frac{1}{n}\sum_n uu'$ tende all'identità, invece $\frac{1}{n}\sum_n vv'$ tenderà a D nel seguente modo:

$$\frac{1}{n}\sum_n vv' = \frac{1}{n}\sum_n Tuu'T' = T(\frac{1}{n}\sum_n uu')T' \Rightarrow TT' = D$$

dove per $\Rightarrow$ si intende la convergenza in media.
Quindi, per disegnare un vettore da una distribuzione normale multivariata con media α e matrice di covarianza D, dovremo calcolarci la decomposizione di Cholesky di D prendere T e moltiplicarlo per il vettore u. Il vettore finale $\alpha + Tu$ sarà distribuito con media α e matrice di covarianza D.
Ora stimiamo la matrice D definendo p il numero di parametri stimati per ciascun individuo n, si definisce inoltre $N = n + p$. La nostra prima stima di D è la matrice identità I di ordine p. Calcoliamo la matrice H come combinazione delle informazioni a priori delle correnti stime di α e β_i cioè:

$$\mathrm{H} = \mathrm{p}\ I + \sum_n (\alpha - \beta_i)(\alpha - \beta_i)'$$

ora calcoliamoci H^{-1} e la sua decomposizione di Cholesky:

$$H^{-1} = TT$$

successivamente si generano N vettori di valori casuali indipendenti distribuiti normalmente con media zero e variaza unitaria, u_i, moltiplichiamo ciascuno con T, e poi sommiamoli in questo modo:

$$S = \sum_n ((Tu_i)(Tu_i)'$$

per finire la stima di D non sarà altro che l'inversa della matrice S.

3.4.5 The Metropolis Hastings Algorithm

Con la presente procedura iterativa ci calcoleremo il set dei parametri β per ogni rispondente. Definiamo β_0 le stime iniziali dell'utilità parziali di un individuo. Generiamo un valore prova per le nuove stime che indicheremo β_n e testiamo se esse rappresentano un miglioramento. Se ciò avviene accetteremo le nuove stime, in caso contrario, invece, le rifiuteremo dipendentemente dal peggioramento rilevato. Prendendo β_n troviamo un vettore casuale d di "differenze" da una distribuzione normale con media zero e matrice di covarianza pari a D, definiamo $\beta_n = \beta_0 + d$. Fatto ciò calcoliamo le probabilità dei dati riferendoci alle utilità parziali di β_0 e β_n usando la formula del modello logistico multinomiale sopra presentata. Le presenti probabilità non sono altro che il prodotto di tutte le probabilità delle scelte che ogni individuo ha effettuato calcolate con il solito modello multinomiale logit. Chiameremo i presenti valori p_0 e p_n. Successivamente si calcola la densità relativa della distribuzione dei beta corrispondenti a β_0 e β_n, grazie alle correnti stime dei parametri α e D. Chiamiamo d_o e d_n i presenti valori. La densità relativa si calcola con la seguente formula:

$$\exp[-\frac{1}{2}(\beta - \alpha)' D^{-1}(\beta - \alpha)].$$

Infine calcoliamo il presente rapporto:

$$r = \frac{p_n d_n}{p_0 d_0}$$

le probabilità p_n e p_0 sono le probabilità date le stime dei parametri β_n e β_0. Le densità d_n e d_0 sono proporzionali alle probabilità calcolate con i valori di β_n e β_0 e rappresentano le distribuzioni a priori delle utilità parziali. Quindi r rappresenta il rapporto delle probabilità a posteriori di queste due stime date gli alpha e la matrice D stimate correntemente. In pratica r è un indice di bontà dei nuovi β. Se r è più grande o uguale all'unità, β_n ha una probabilità a posteriori più grande o uguale a quella di β_0 e quindi accetteremo β_n come nuove stime dei beta per quell'individuo. Se r è minore di uno β_n ha probabilità a posteriori inferiore a β_0. In questo caso useremo un processo casuale per decidere se accettare i nuovi beta o mantenere i nostri vecchi beta al massimo per un'altra

iterazione. La probabilità con cui accetteremo i nuovi beta è pari a r. Come vediamo sia p_0 che p_n entrano nel processo di calcolo e di fatto se uno dei due presenta un'altà densità, con le correnti stime di α e D, quest'ultimo avrà un vantaggio di scelta. Se le densità a priori non fossero considerate i β sarebbero scelti solamente massimizzando la verosimiglianza come al paragrafo (3.3.1) e ci ricondurremo a riprodurci le stime secondo il modello logistico multinomiale.

Capitolo 4

Utilizzo del Sawtooth Software

4.1 Introduzione

Questo capitolo spiega passo per passo l'utilizzo di CBC versione 2 dalla creazione di un nuovo studio, con l'inserimento della lista di attributi e livelli, alla creazione del questionario fino all'analisi dei risultati. L'unico scopo del presente esempio applicativo è imparare l'utilizzo del software senza la necessità di prestare particolare attenzione alla validità dello studio nel prevedere il comportamento dei consumatori. L'esempio riportato è inerente al settore dei televisori. Dopo aver identificato il problema e il campione dei rispondenti, uno dei primi steps della conjoint analysis è definire gli attributi e i livelli di ogni attributo:

Marca	**Dimensione dello schermo**
JVC	25"
SONY	26"
RCA	27"

Qualità del suono	**Prezzo**
Mono sound	Euro 300
Stereo sound	Euro 350
Sorround sound	Euro 400

Ciascun profilo di prodotto sarà descritto usando un unico livello per ogni attributo. Dopo aver definito attributi e livelli, si compila il resto del questionario e si pianifica il design. Si può decidere di condurre l'intervista a computer tramite un floppy disk consegnato agli intervistati oppure in modalità Paper&Pencil realizzando delle cards contenenti i profili di prodotto.

4.2 Intervista assistita tramite PC

Con il presente questionario di conjoint analysis, si vuole misurare la preferenza degli intervistati per differenti livelli di attributo e l'impatto dei livelli stessi sulla scelta del prodotto televisore. Per raggiungere tale scopo si è deciso di somministrare 15 "randomized tasks". In aggiunta ai "random choice tasks", si può decidere di imporre anche dei "fixed holdout tasks" per le seguenti ragioni:

- le aziende che decidono di condurre delle ricerche di mercato basate sulla conjoint analysis, spesso, non hanno esperienza della metodologia e sembrano avere delle perplessità sull'affidabilità e sull'accuratezza del market simulator. Dimostrando che il market simulator può predire in modo accurato le risposte date agli "holdout task", si fornisce una prova della validità del metodo;

- avere alcuni "holdout tasks" permette di confrontare l'accuratezza predittiva di diversi modelli di simulazione.

I "fixed task" non sono costruiti in modo casuale ma presentano livelli che devono essere specificati dall'analista. Nell' esempio si è deciso di sottoporre ai rispondenti un'unica domanda di controllo. Nella parte iniziale del questionario, inoltre, si è introdotto una schermata introduttiva in cui si ringrazia l'intervistato per aver dato la sua disponibilità e si è deciso di somministrare alcune domande di carattere demografico(generalità). Definiti i punti chiave citati precedentemente, si procede con la creazione dello studio congiunto con il supporto del software.

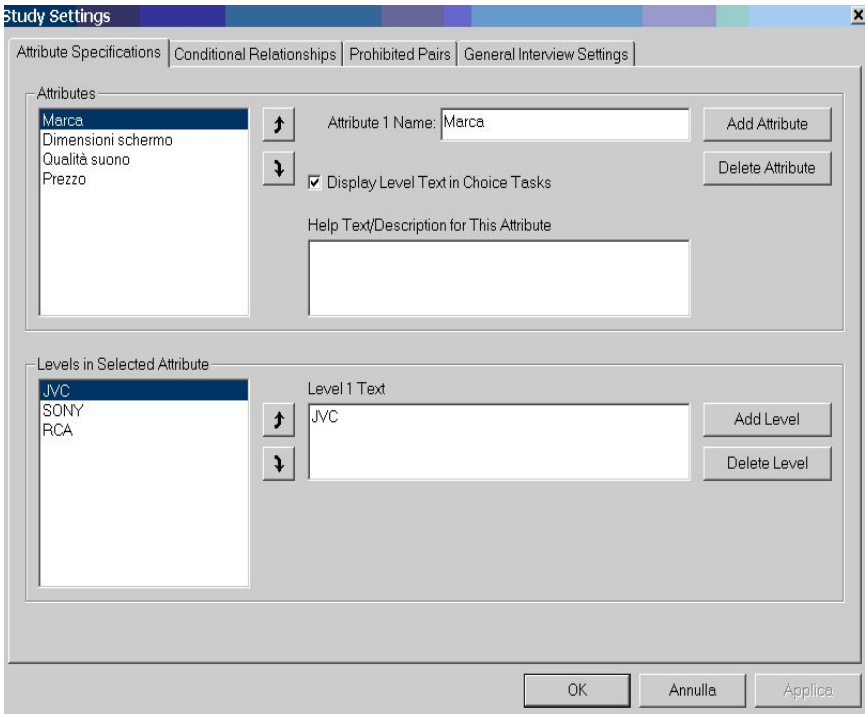
CREARE UN NUOVO STUDIO

File | New

Appare la finestra di dialogo *"New Study"* che permette di salvare lo studio congiunto che ci accingiamo a condurre. Salviamo quindi il file **study0.smt** nella cartella **Tutorial** cliccando **Save**

INSERIRE LA LISTA DEGLI ATTRIBUTI E DEI LIVELLI

Cliccare **A** o scegliere **Compose|Attributes**.
Si apre la finestra di dialogo **Study Parameters** dalla quale selezioniamo inizialmente **Attribute Specifications**.



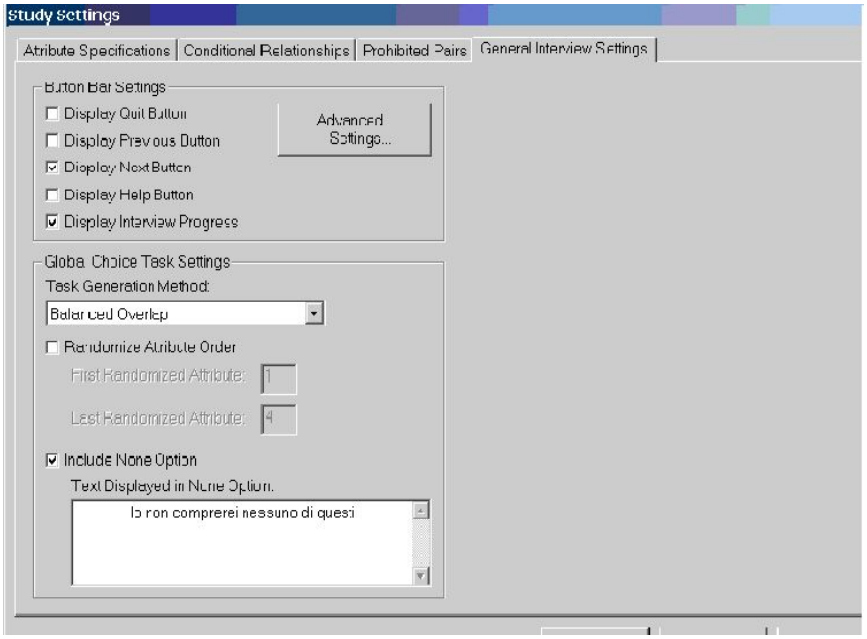
Per aggiungere il primo attributo (Marca), si clicca il **Add Attribute** e si digita il nome dell'attributo nell'apposito campo. Successivamente si clicca **Add Level** per inserire tutti i livelli dell'attributo specificato. Quando si è

pronti per aggiungere l'attributo successivo (Dimensione Schermo), si ripete la procedura sopra descritta fino al completo inserimento dei dati.

PARAMETRI DI STUDIO ADDIZIONALI

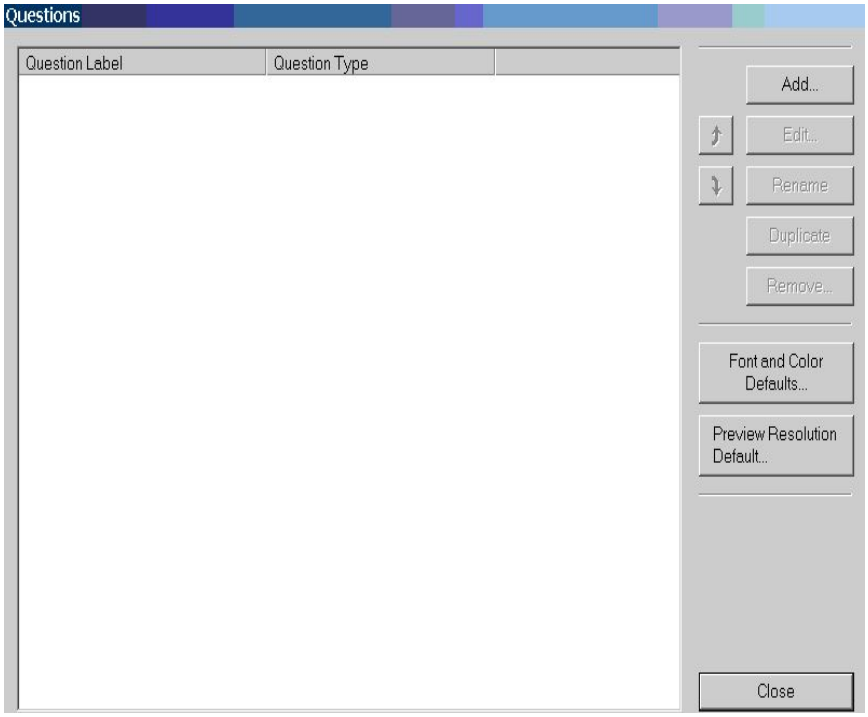
Dopo aver specificato la lista di attributi e livelli, si possono specificare altri parametri che governano il questionario. In questo esempio applicativo, né conditional relationship né level prohibitions sono richiesti.

Le *conditional relationships* possono essere utili se il prezzo di prodotti dipende dai livelli di un altro attributo, come ad esempio la marca. Così facendo, si evitano coppie proibite di certi livelli di prezzo e marca che rendono il design meno efficiente. Le *prohibitions*, invece, sono utilizzate per specificare che un livello di un certo prodotto non venga combinato con un livello di un altro attributo. Selezionare **General Interview Setting** permette di scegliere il tipo di design. Per default, è fissato il *Complete Enumeration* ma poiché nell'esempio si desidera misurare con precisione le interazioni two-way fra attributi, si è deciso di scegliere il metodo *Balanced Overlap*. Infine si può specificare l'opzione "None" specificando il testo che apparirà agli occhi dei rispondenti al momento dell'intervista:



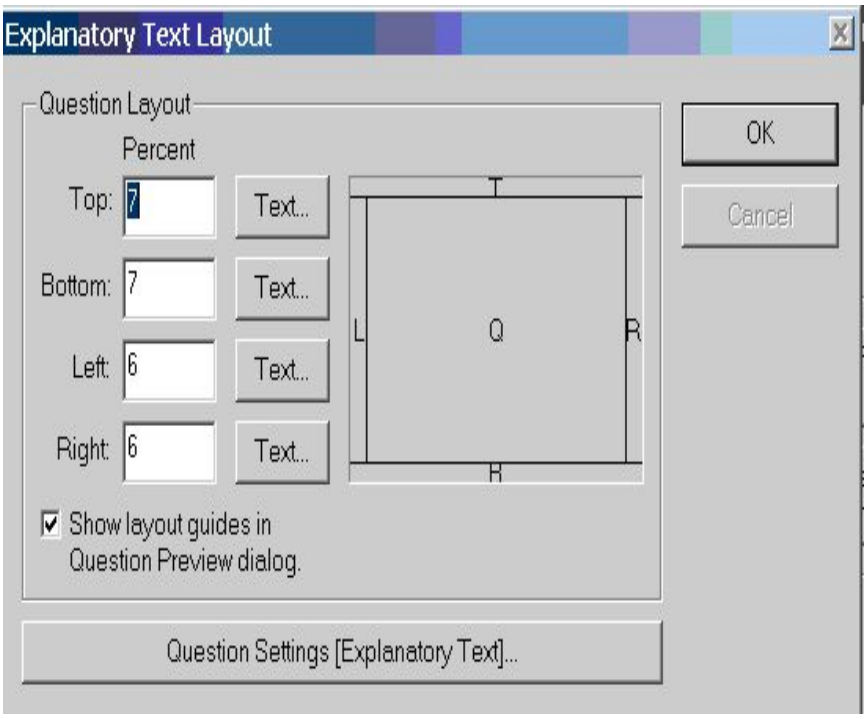
Cliccare **Ok** per chiudere la finestra e procedere.

SPECIFICARE EXPLANATORY TEXT E SELECT QUESTIONS
Cliccare **Q** o selezionare **Compose** | **Questions**. Si apre la finestra **Questions** inizialmente vuota:



EXPLANATORY TEXT

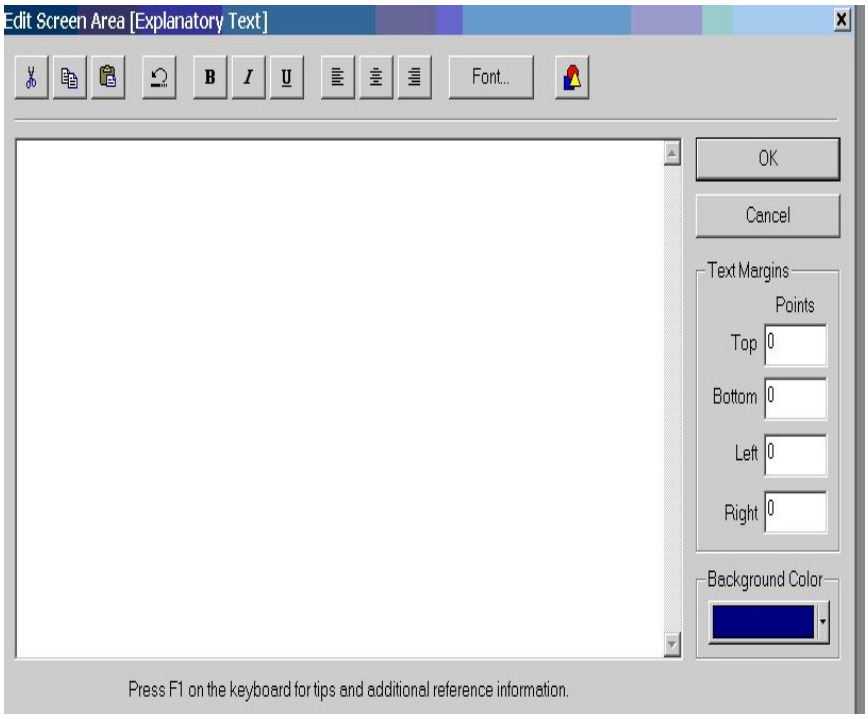
Si clicchi **Add** e appare una ulteriore finestra. Dovendo realizzare una schermata introduttiva all'inizio del questionario, si specifica **Explanatory text** e si sceglie l'etichetta **Introduzione**. Una *Explanatory Text Question* è una domanda che non presuppone alcuna risposta da parte dell'intervistato. Cliccare **Finish** per far comparire *Explanatory Text Layout*:



Tutte le domande in CBC sono divise in 5 aree:

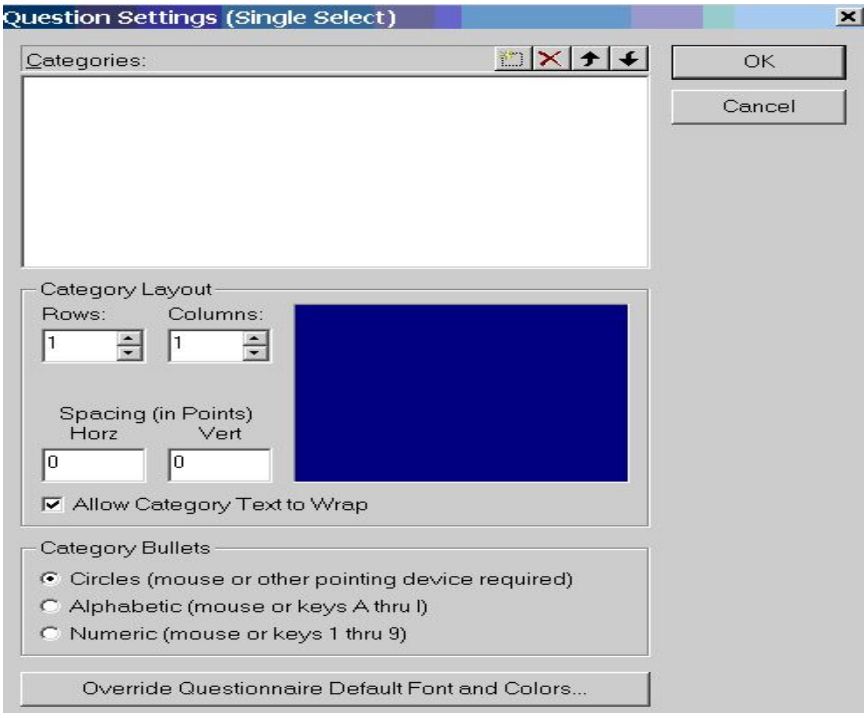
text area (top)		
Text area (sinistra)	Question area	Text area (destra)
text area (bottom)		

L'analista può specificare le dimensioni di ogni area, le dimensioni dei caratteri presenti sullo schermo, il colore dello sfondo, i margini di ogni area etc. Per la schermata introduttiva, ha senso considerare solo l'area centrale (Question area). Cliccare **Question Settings (Explanatory Text)** per far comparire la finestra **Edit Screen Area** la presenta schermata verrà visualizzata :



all'interno dello spazio bianco si digita il testo introduttivo specificando la dimensione dei caratteri (font), i colori e lo stile:
Grazie per aver deciso di partecipare al nostro questionario! Inizialmente dovrai rispondere a domande di carattere personale. Il questionario è inerente al prodotto TELEVISIONE . Ti verranno mostrate alcune alternative di televisori e ti sarà richiesto di scegliere il profilo che saresti disponibile a comprare. Sarebbe importante che tu rispondessi cercando di immedesimarti nella parte di chi vuole realmente acquistare un televisore. Puoi scegliere l'opzione "None" qualora non acquisteresti nessuno dei televisori che ti proponiamo.

SINGLE SELECT
Per realizzare le domande di carattere demografico/anagrafico si clicca **Add** della finestra **Questions** e si sceglie *Single Select*. Cliccando, poi, **Finish** si apre la finestra **Question Settings (Single select)**:



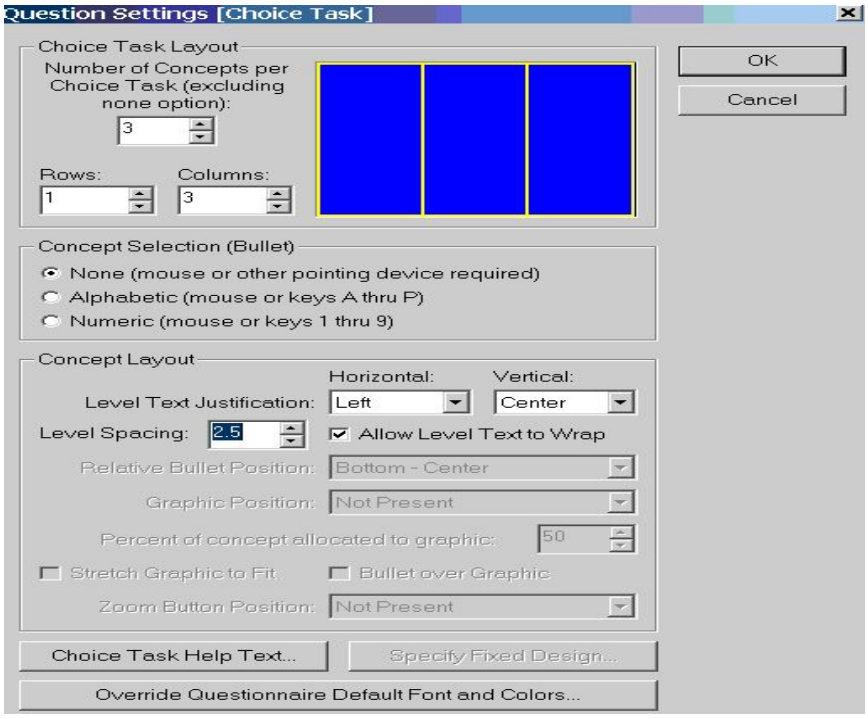
la finestra è divisa in 4 parti:

- *Categories* cliccare l'icona "New(Insert)" e digitare l'etichetta della prima categoria (meno di 30 anni). Per inserire le categorie rimanenti, premere invio e l'icona "New(Insert)";
- *Category Layout* si può specificare il numero di righe e colonne;
- *Category bullets* si è scelto l'opzione "cerchi";
- *Override questionnaire default* serve per cambiare la dimensione dei caratteri e i colori.

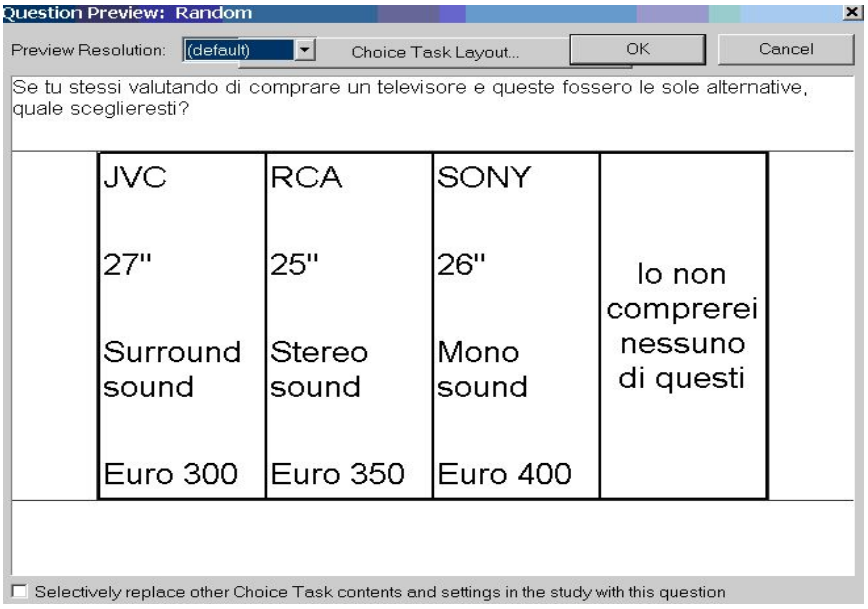
Specificati tutti i parametri desiderati, cliccare **Ok**.

SPECIFICARE I RANDOM CHOICE TASKS

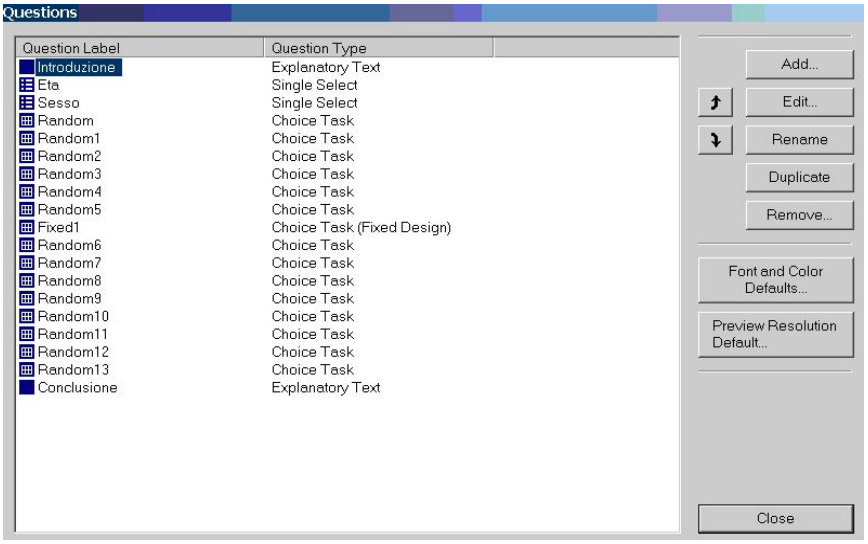
Cliccare il bottone **Add** della finestra **Questions**, selezionare *Choice task*, specificare **Random** come etichette per il primo task. Cliccare **Finish**. Compare la finestra **Choice task Layout** in cui si clicca **Question Setting (Choice**



Task) . La finestra **Question Settings (Choice Task)** permette di specificare il numero di concetti di prodotto per task (esclusa l'opzione None). Per l'esempio applicativo, si è deciso di sottoporre 3 concetti per profilo. *Concept Layout*: si è cambiato il *Level Text Justification Horizontal* da Center a Left in modo che i livelli di prodotto siano allineati a sinistra. *Level Spacing*: per avere una separazione ottimale fra i liveli di attributo entro un singolo concetto ho impostato 2,5. Cliccare **Ok** per chiudere la finestra aperta. Dalla finestra **Choice Task layout** si clicca **ok** per far comparire **Question Preview** e facendo doppio-click si inserisce il seguente testo: *Se tu stessi valutando di acquistare un televisore e queste fossero le sole alternative, quale sceglieresti? Fai la tua scelta cliccando col mouse sull'alternativa preferita.*



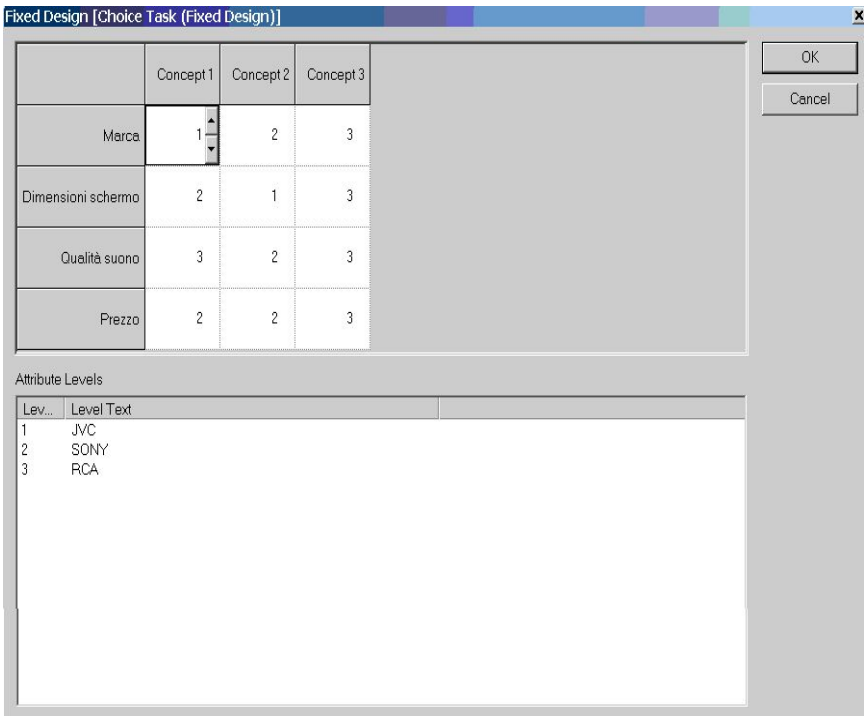
Per generare i rimanenti random tasks alla finestra **Questions** (bottone **Q**) evidenziare il choice task e cliccare **Duplicate**.



SPECIFICARE I FIXED TASKS

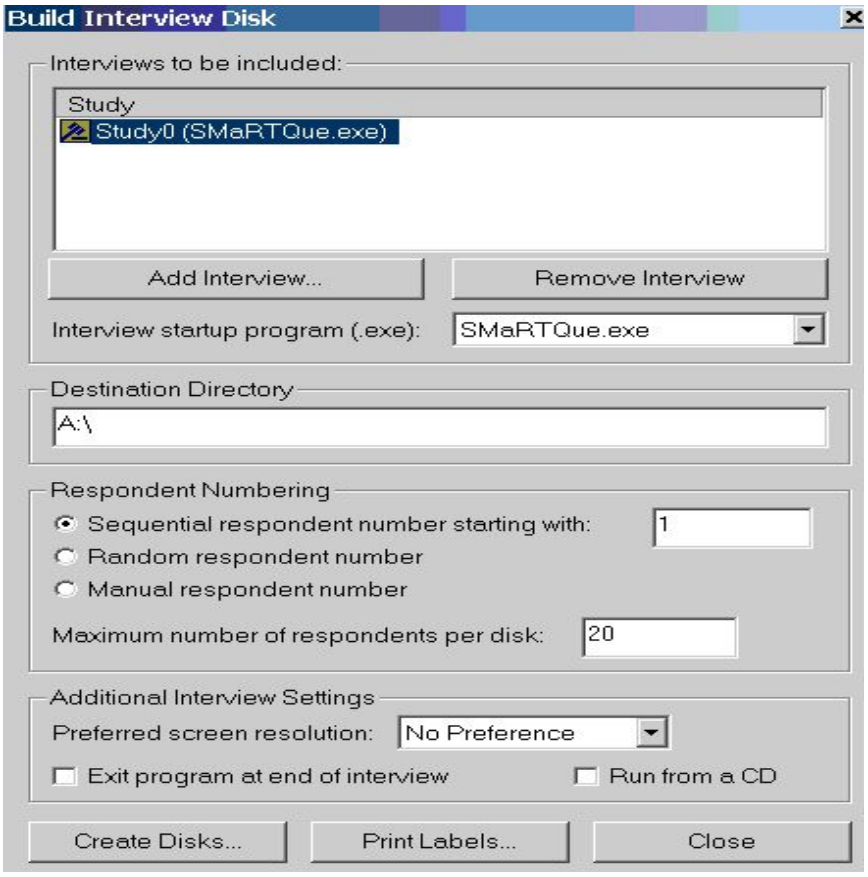
Ora si può creare uno o più tasks caratterizzati da concetti i cui livelli di ogni attributo sono specificati dall'analista. Per l'esempio proposto, si è deciso di introdurre un unico "Fixed task" dopo il sesto "Random Task" con lo scopo di controllare l'abilità del market simulator a predirre la risposta data nel Fixed holdout task. Ecco i passi da seguire:

- alla finestra Questions, evidenziare il sesto "Random Task";
- cliccare il bottone Add, scegliere *Choice Task (fixed design)* con etichetta *Fixed*;
- cliccare Questions settings *Choice Task (Fixed design)* per specificare i 3 concetti e la spaziatura fra livelli pari a 2,5;
- cliccare Specify fixed Design in modo da aprire la seguente finestra in cui specificare i livelli degli attributi di ogni concetto:



RUNNING/FIELDING IL QUESTIONARIO

Cliccare **R**, o **Compose** | **Run Questionnaire**, per visualizzare il questionario. Dopo aver salvato i dati, se sono state specificate delle *prohibitions* o se si desidera condurre un questionario **paper&pencil** è consigliabile testare l'efficienza del design scegliendo **T** (per maggiori dettagli vedere "QUESTIONARIO PAPER AND PENCIL"). Se si è soddisfatti del questionario creato, si procede col realizzare i floppy disk contenenti i questionari scegliendo **Field** | **Make Field Disk**

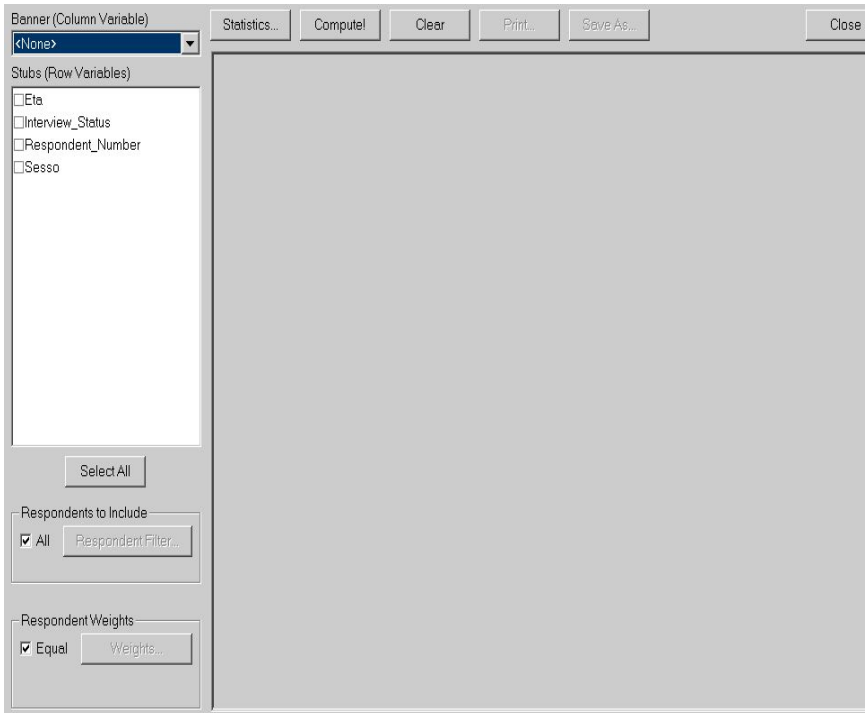


Quando compare la finestra **Build Interview Disk** si può specificare il numero di rispondenti per dischetto; in questo esempio si è mantenuto il valore di default (20). Successivamente si clicca **Create Disks** per realizzare due dischetti contenenti in totale 40 questionari che sono stati somministrati ai

rispondenti. Per la catture di tutti i dati provenienti dai *Field Disks* , si segue il seguente percorso: **Field** | **Accumulate Field Data**.

4.3 Ausilio di tabelle per un'analisi preliminare

Con **Analysis** | **Tables**

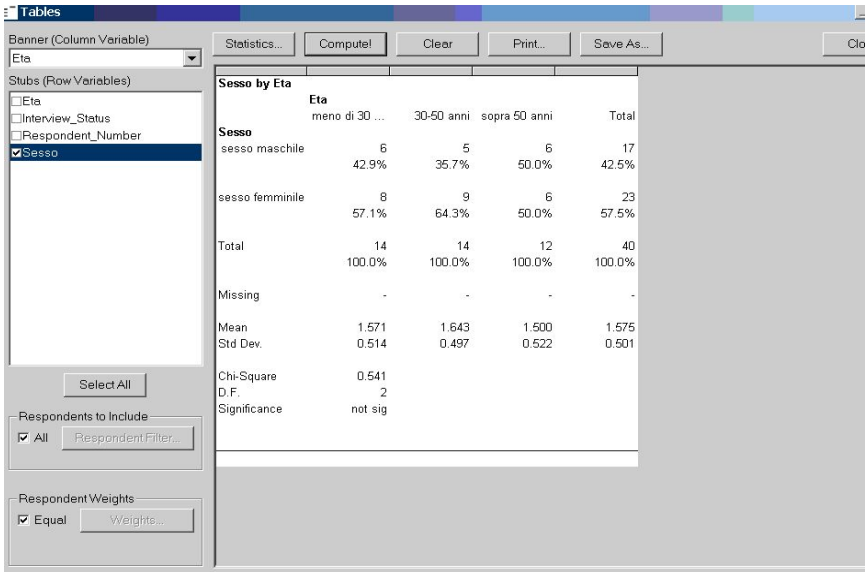


- si può esaminare la frequenza delle risposte per ogni categoria (relativa a età o sesso nel nostro esempio) che compare in un determinato task. Dopo aver selezionato le variabili che interessano, si clicca **Compute!** per ottenere i risultati dell'analisi;

```
Tables Report
Copyright 1999-2003 Sawtooth Software
Total
età
meno di 30 anni 14
```

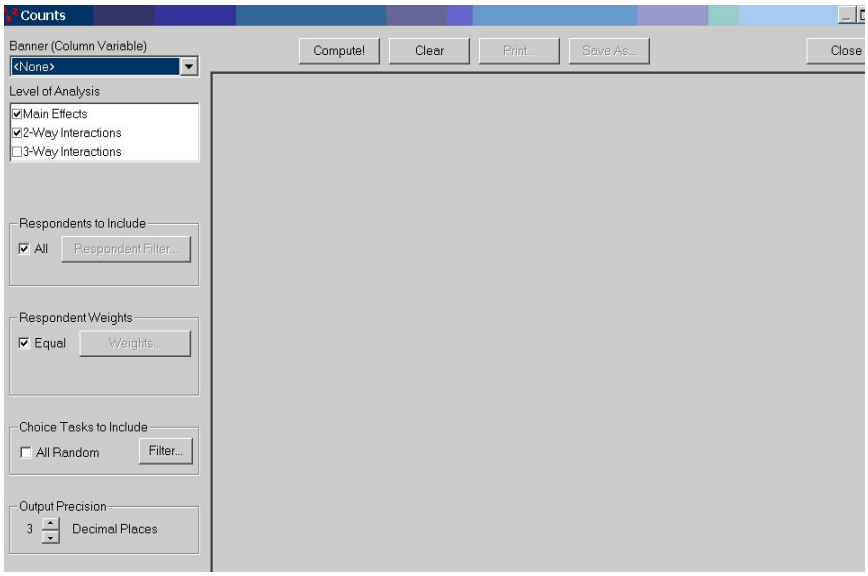
```
35,00%

30-50 anni 15
37,50%
sopra i 50 anni 11
27,50%
Total 40
100,00%
Missing -

Mean 1,93
Std Dev. 0,80
Total
Sesso
sesso maschile 17
42,50%
sesso femminile 23
57,50%
Total 40
100,00%
Missing -
Mean 1,58
Std Dev. 0,50
```

- si può anche esaminare l'interazione fra due variabili oggetto di studio delle domande Single Select (età e sesso) selezionando **Statistics** e,dopo aver specificato le opzioni desiderate e la variabili di riga (sesso) e di colonna (età), **Compute**!.



COUNTING ANALYSIS

Analysis | Counts



L'analisi riporta la percentuale delle volte in cui concetti con un livello di attributo sono stati scelti diviso il numero di volte in cui concetti con quel determinato livello appaiono (probabilità di scelta). I counts appartengono all'intervallo 0-1. Se, ad esempio, il count di un livello è 0,31 significa che i rispondenti hanno scelto quel livello il 31% delle volte. Per default, "Counts

Program" analizza tutti i "main-effects", gli effetti delle interazioni fra due attributi senza includere nell'analisi i "fixed tasks". Infatti il programma assume un "random design" dove ogni livello di attributo appare un ugual numero di volte combinato con ogni livello degli altri attributi. Cliccando *Compute*! , compaiono i risultati:

```
CBC System
Analyze by Counting Choices
Copyright 1993-2003 Sawtooth Software
Choice Tasks Included:   All Random

Marca
Total
Total Respondents 40
JVC 0,28
SONY 0,25
RCA 0,30
Within Att. Chi-Square 0,91
D.F. 2
Significance not sig

.
.
.
.

Dimensione dello schermo
Total
Total Respondents 40
25''0,33
26''0,28
27''0,23
```

Within Att. Chi-Square 3,99
D.F. 2
Significance not sig

.
.
.
.

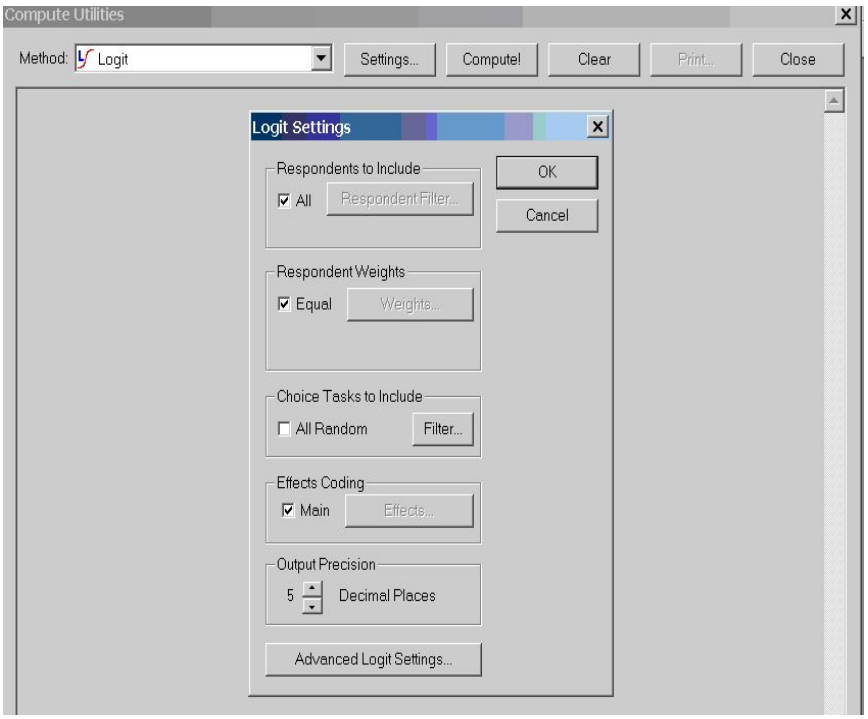
Qualità del suono x Prezzo
Total
Total Respondents 40
Mono sound Euro 300 0,52
Mono sound Euro 350 0,16
Mono sound Euro 400 0,20
Stereo sound Euro 300 0,45
Stereo sound Euro 350 0,13
Stereo sound Euro 400 0,14
Sorround sound Euro 300 0,52
Sorround sound Euro 350 0,21
Sorround sound Euro 400 0,16
Interaction Chi-Square 1,12
D.F. 4
Significance not sig
None
Total
Total Respondents 40
None chosen: 0,17

ANALIZZARE I DATI USANDO IL METODO "LOGIT"

Analysis | Compute Utilities

Si apre una finestra in cui cliccando **Settings** si possono specificare alcuni parametri dell'analisi:

- i rispondenti da inserire;
- i pesi da applicare alle diverse categorie di rispondenti (uomo/donna . . .);
- i task da includere;
- gli effetti da considerare (solo effetti principali oppure anche effetti di interazione).



Per far avviare l'analisi cliccare **Compute**! :

```
CBC System  Multinomial Logit Estimation
Copyright 1993-2000 Sawtooth Software
Name/Description: Logit Run
11:30:54AM Saturday, January 31, 2004
Main Effects
Tasks Included: All Random
    Total number of choices in each response category:
             1     66  27.50%
             2     74  30.83%
             3     59  24.58%
          NONE     41  17.08%
    Files built for 40 respondents.
    There are data for 240 choice tasks.
    Iter    1  Chi Square =      72.52902  rlh =       0.29078
    Iter    2  Chi Square =      75.39773  rlh =       0.29252
    Iter    3  Chi Square =      75.39872  rlh =       0.29252
    Iter    4  Chi Square =      75.39872  rlh =       0.29252
    Converged.
    Log-likelihood for this model =    -295.01128
    Log-likelihood for null model =    -332.71065
                                      ------------
Difference =      37.69936    Chi Square =      75.39872

       Effect         Std Err         t Ratio  Attribute Level
 1   -0.02139        0.10922       -0.19581    1 1 JVC
 2   -0.05690        0.11119       -0.51172    1 2 SONY
 3    0.07828        0.10700        0.73162    1 3 RCA

 4    0.20950        0.10560        1.98395    2 1 25''
 5    0.04028        0.10897        0.36961    2 2 26''
 6   -0.24977        0.11492       -2.17339    2 3 27''
```

```
 7    0.08520        0.10833        0.78653    3 1 Mono sound
 8   -0.15596        0.11242       -1.38732    3 2 Stereo sound
 9    0.07076        0.10699        0.66137    3 3 Sorround sound

10    0.75465        0.09904        7.61988    4 1 Euro 300
11  -0.38717        0.12312       -3.14469    4 2 Euro 350
12  -0.36748        0.12281       -2.99228    4 3 Euro 400

13        -0.31385        0.17655       -1.77771    NONE
Time for computation = 0 seconds.
```

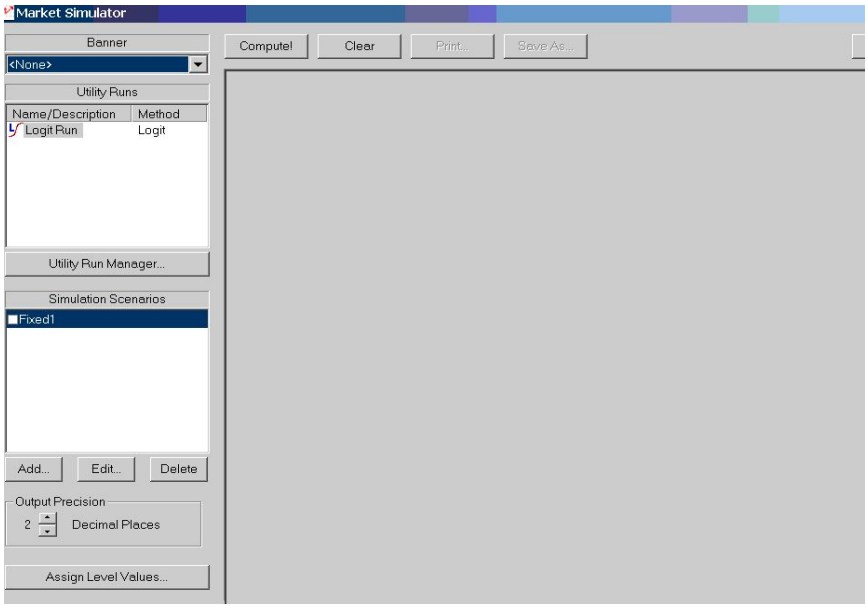
La colonna *Effect* contiene le utilità per ogni livello di ciascun attributo. Più grande è l'utilità, maggiore è la preferenza accordata al livello, inoltre, come sappiamo, la somma delle singole utilità entro un attributo è 1. La colonna *Std Err* mostra le deviazioni standard per ogni effetto. La colonna *t-ratio* ci calcola la statistica di nullità del parametro. Per finire il programma ci calcola la statistica log rapporto di verosimiglianza per testare se il modello stimato risulta significativamente diverso dal modello con tutti i parametri nulli.

4.4 Market Simulator

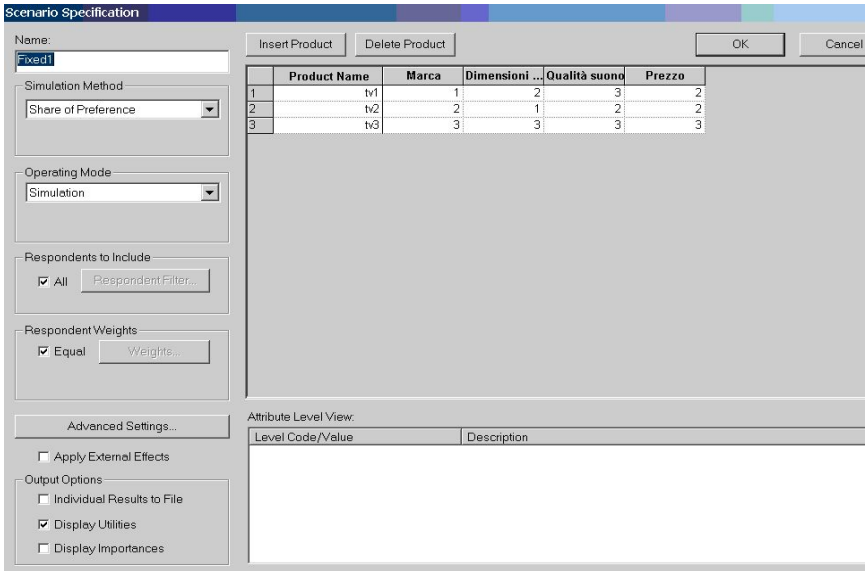
La presente funzione permette di simulare le utilità di possibili prodotti non presenti nel questionario.

Analysis | Market Simulator
Nel campo *Utilities Runs* si trovano i risultati della *Logit analysis* salvati precedentemente. Si clicca **Add** per creare scenari di simulazione.

Per controllare l'accuratezza del market simulator a prevedere le scelte dei consumatori, si considera il "fixed task". Nel *Name Field* si digita *Fixed 1* e poi si specificano i 3 prodotti inclusi nella domanda e il metodo di simulazione. Per questo particolare esempio si è scelto lo "Share of Preference" che lavora bene



soprattutto quando i prodotti specificati presentano in minima parte *Overlap* dei livelli. Si poteva utilizzare anche il "Randomized First Choice" che è preferibile nel caso di prodotti che condividono livelli. Per stimare lo share of preference dell'opzione NONE si seleziona Advanced Setting.



I risultati della simulazione si ottengono cliccando **Compute**! :

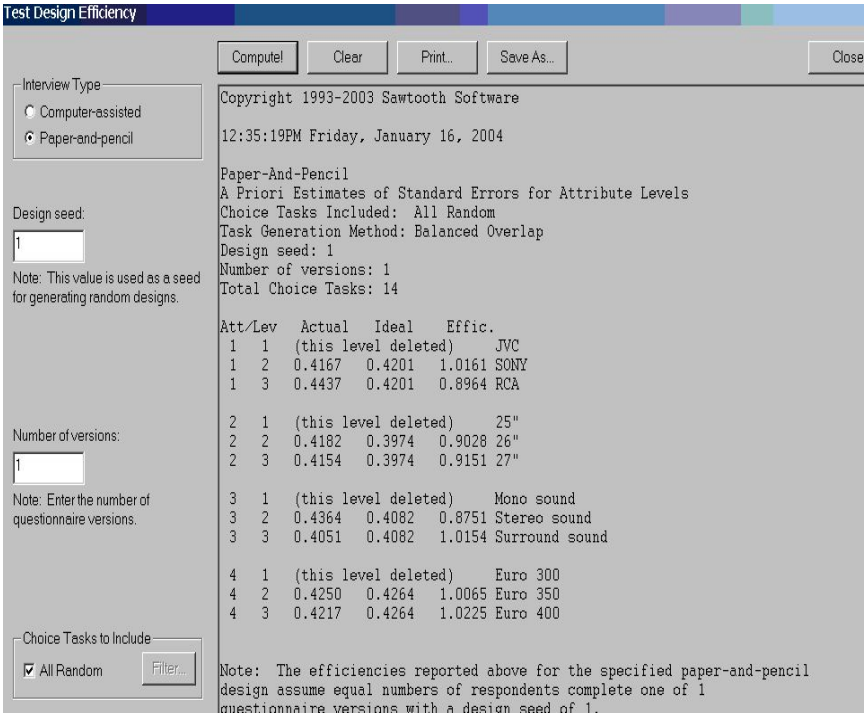
```
Sawtooth Software SMRT Market Simulator
Copyright 1999-2003
Scenario: Fixed1
Utility Run: Logit Run

Average Utility Values
Rescaling Method: Zero-Centered Diffs
Total
JVC -4,33
SONY -11,51
RCA 15,83

25''42,38
26''8,15
27''-50,52

Mono sound 17,23
Stereo sound -31,55
Sorround sound 14,31

Euro 300 152,65
Euro 350 -78,32
Euro 400 -74,34

None -63,49

Product Simulation Settings
Simulation Mode: Simulation
Simulation Method: Share of Preference
None Weight:
```

```
Exponent: 1,00
Product Specifications
    Marca  Schermo    Suono     Prezzo
TV1     1    2          3            2
TV2     2    1          2            2
TV3     3    3          3            3
Product Shares of Preference
TV1 36,31
TV2 33,08
TV3 30,61
```

4.5 Raccolta dati tramite Questionario Cartaceo

Il processo per la creazione del questionario su carta è simile a quello del questionario su computer per ciò che concerne l'inserimento degli attributi, dei livelli e la composizione dei "choice-tasks". Nonostante non si impieghi un tipico randomized design con il questionario su carta, si può tuttavia usare i random choice tasks generati da un "computer-based CBC study". Invece di dare a ogni rispondente un'unica versione del questionario, si possono creare poche versioni e assegnarne una a ciascun rispondente in modo casuale. Si raccomanda generalmente di includere un numero sufficiente di versioni del questionario in modo che il numero di "tasks" moltiplicato per il numero di versione del test sia maggiore o uguale a 80. Il seguente esempio applicativo rappresenta un semplice esercizio di utilizzo del software che ha permesso di conoscerne l'operatività. Dopo aver creato i "choice tasks" in modo del tutto simile al caso di una intervista interativa a computer, è consigliabile nel caso di questionario "paper-based" (in cui il numero di versioni è piccolo) testare l'efficienza del design usando **Compose** | **Test Design**. Dopo aver indicato il design seed e il numero di versioni differenti, CBC automaticamente testa il design e mostra i

risultati: Descriviamo ora in dettaglio l'output del test di efficienza del design.

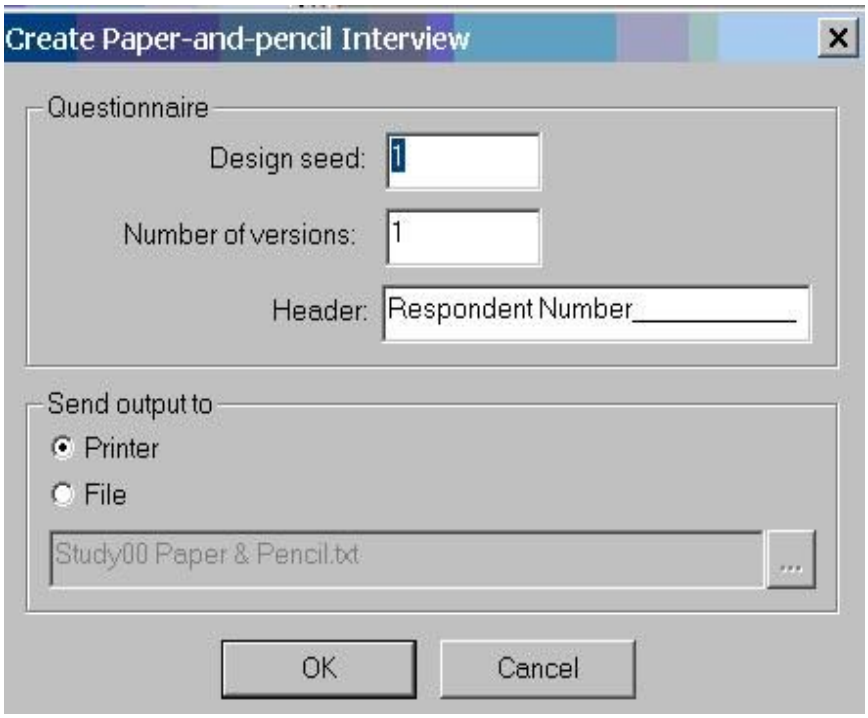


Per le stime, è necessario omettere un livello da ogni attributo e in particolare il primo livello è automaticamente cancellato dalla analisi. La prima colonna etichettata **Actual** fornisce gli standard errors per il data file analizzato. La colonna etichettata **Ideal** dà le stime del valore che dovrebbero assumere gli standard errors se il design fosse completamente ortogonale e avesse lo stesso numero di osservazioni. La colonna **Effic** fornisce l'efficienza relativa del design paragonato all'ipotetico piano ortogonale (è la radice quadrata del rapporto). Il valore di efficienza non è accettabile se si avvicina troppo a zero. E' sempre opportuno testare il design se si verifica una delle seguenti condizioni:

- alcune *prohibitions* sono incluse;
- sample size (rispondenti x tasks) è piccolo;
- il numero di versioni è piccolo.

STAMPARE IL QUESTIONARIO

Per stampare il questionario, si sceglie **Field | Create Paper -and-Pencil Interview**. Si apre la seguente finestra: Si specificano i seguenti campi:



- *Design seed*: si sceglie tipicamente "1"
- *Number of version*: per ottenere un piano che misura in modo efficiente i "main-effects" e le interazioni di primo ordine nel caso vengano somministrati 20 "choice tasks" per intervistato, si suggerisce di avere 4 versioni del questionario. Poichè l'esempio applicativo qui presentato ha solo valenza operativa, ci si limita ad utilizzare una unica versione del questionario che si somministrerà a 10 intervistati in modo da diminuire il tempo di raccolta dati.
- *Header*: si specifica il testo che si vuole far apparire sulla prima pagina di ogni versione del questionario.

Seleziono "**Printer**" per stampare il questionario.

REGISTRAZIONE DELLE RISPOSTE DEGLI INTERVISTATI

I dati dei rispondenti vanno registrati dentro un file di testo delimitato da spazi, virgole o tabulazioni. Per l'esempio proposto, si è usato il programma *Blocco Note*:

```
01 1 1 2 2 2 1 4 3 2 3 4 3 2 3 2 2 1 1
02 1 3 1 2 4 1 4 2 1 2 4 1 1 2 2 2 1 1
03 1 3 1 2 3 2 3 2 3 3 4 3 4 2 2 4 1 3
04 1 2 1 2 3 2 1 4 3 3 3 1 2 3 4 1 1 1
05 1 2 1 2 3 2 4 2 3 2 1 2 2 3 3 1 1 1
06 1 2 1 2 3 1 4 2 3 2 1 3 2 3 2 1 1 1
07 1 1 2 1 4 3 1 2 3 1 4 1 1 3 2 2 1 1
08 1 1 2 1 4 2 3 1 2 4 2 1 3 3 2 1 1 3
09 1 1 2 4 1 3 3 2 2 1 2 4 3 3 4 1 1 1
10 1 2 2 2 3 3 2 4 4 3 2 2 1 1 2 2 1 3
```

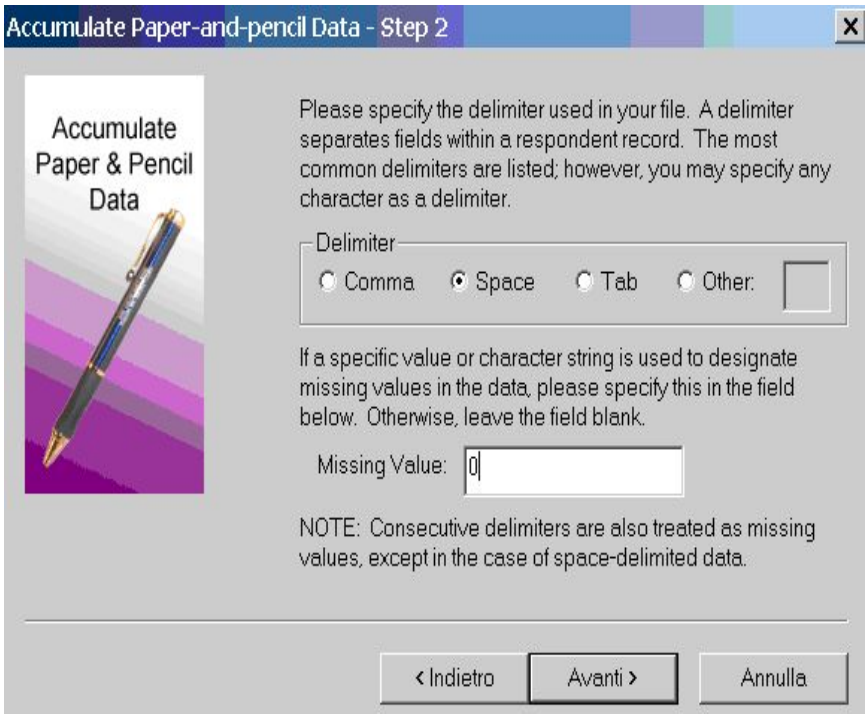
Il file di testo deve presentare solo valori numerici e le risposte di ogni intervistato occupano un'unica riga. Il numero del rispondente e il numero della versione sono i primi due campi, le risposte date alle *select* e *choice questions* sono inserite nello stesso ordine in cui i **tasks** appaiono nel questionario e codificate come "1" (se è scelto il primo concetto di prodotto/ prima categoria), "2" (se è scelto il secondo prodotto/ seconda categoria), etc. Se alcuni rispondenti saltano delle domande, si può utilizzare un valore come "0" per rappresentare la risposta mancante. L'esempio prevede 2 *single select questions* (domande di segmentazione) e 15 *choice task*.

CATTURA DEI DATI

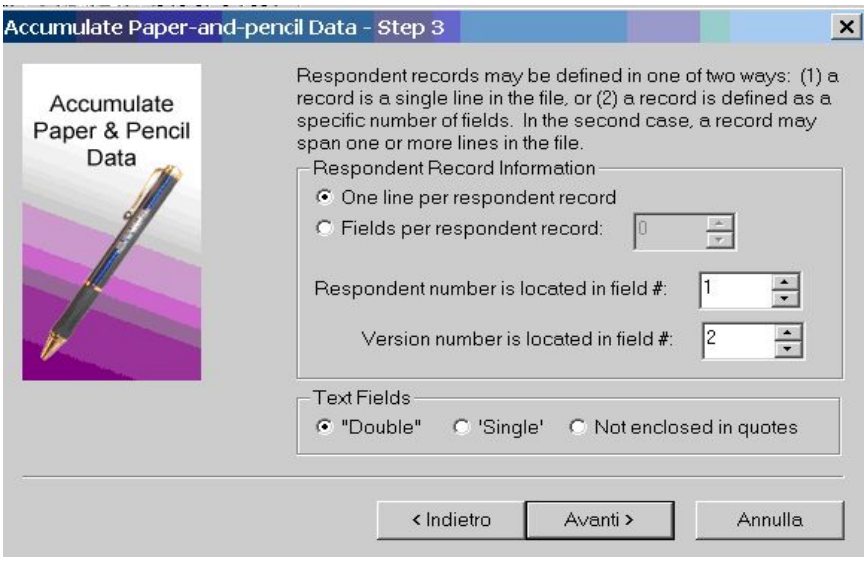
Selezionare **Field** e **Accumulate Paper & Pencil Data**

Step 1 : si ricerca il file di testo contenente le risposte dei rispondenti;

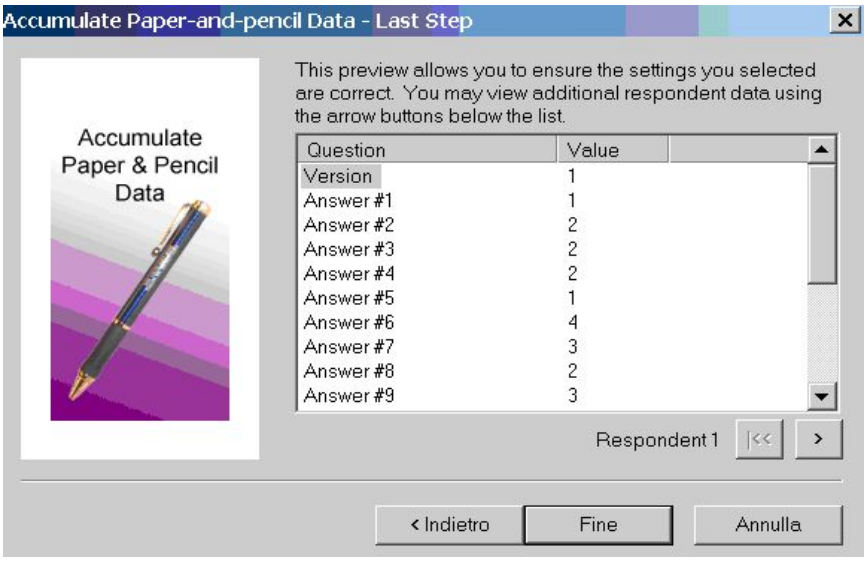
Step 2 : si specifica il delimitatore dei dati e il valore attribuito ai valori mancanti;



Step 3 : si specifica il numero di riga per ogni osservazione, la posizione del numero del rispondente e della versione del questionario nella riga di ogni osservazione; *Text Fields*: se il file delle risposte include variabili stringa, si può specificare se queste sono racchiuse fra doppi virgolette, singole o nessun tipo di virgolette;



Step 4 : la finestra presenta i valori dei campi del primo record per permettere all'analista di verificare se ha inserito i dati nel modo corretto. Per visualizzare i record relativi ai successivi intervistati, si clicca ">" .



Cliccare ora **Fine** per "fondere" il file delle risposte con il disegno . L'analisi dei dati del market simulator avviene nella stessa procedura del questionario assistito da personal computer, quindi si rimanda alla visione del paragrafo precedente.

4.6 Hierarchical Bayes Analysis

CBC/HB System è un software per stimare le "part-worths" tramite la choice-based conjoint analysis.
Il rispondente deve scegliere l'alternativa preferita in ogni "choice set" caratterizzato da diverse alternative di prodotto. Un'opzione avanzata include la possibilità di stimare i termini di interazione. CBC/HB usa dati che possono essere automaticamente esportati da CBC System di Sawtooth SW o dati in codice ASCII. I primi metodi per analizzare i dati di scelta combinavano i dati di tutti gli individui. Nonostante molti ricercatori da sempre affermano che le analisi aggregate nascondono importanti aspetti, solo recentemente sono disponibili metodi per analizzare i "choice" data a livello individuale. La tecnica Hierarchical Bayes è stata descritta in modo favorevole da molti articoli di *journals*. Allenby e Ginger (1995) , Lenk , DeSarbo, Green e Young (1996) hanno pubblicato articoli che trattano la stima di part worths individuali usando modelli HB. Questo approccio sembra essere estremamente promettente in quanto si ottiene una stima ragionevole delle utilità individuali a partire da pochi dati ricavati da ogni rispondente. Tuttavia, l'approccio è molto intenso dal punto di vista computazionale. Poiché negli ultimi anni i computers sono diventati più veloci, la Sawtooth Software è stata in grado di fornire un HB Software capace di trattare problemi in un tempo ragionevole (anche meno di 1 ora). Riassumendo:

- HB ha il vantaggio di fornire stime individuali delle utilità a partire da questionari in cui sono poche le scelte fatte da ciascun rispondente;
- HB ha lo svantaggio di essere pesante dal punto di vista dei calcoli in quanto richiede molte migliaia di interazioni.

Ovviamente il tempo richiesto per i calcoli diminuirà con il continuo aumento della velocità degli elaboratori. Un altro problema è la mancanza di software facili da usare. Sawtooth Software ha alleviato questo problema realizzando CBC/HB che può essere utilizzato in combinazione con CBC System in modo che le utilità siano stimate col minimo sforzo. CBC/HB System può essere usato

per analizzare i dati provenienti dal programma software CBC o da altre fonti. In questo caso potrà manipolare data sets che sono più grandi rispetto ai limiti imposti dai questionari CBC. Il programma presenta le seguenti limitazioni:

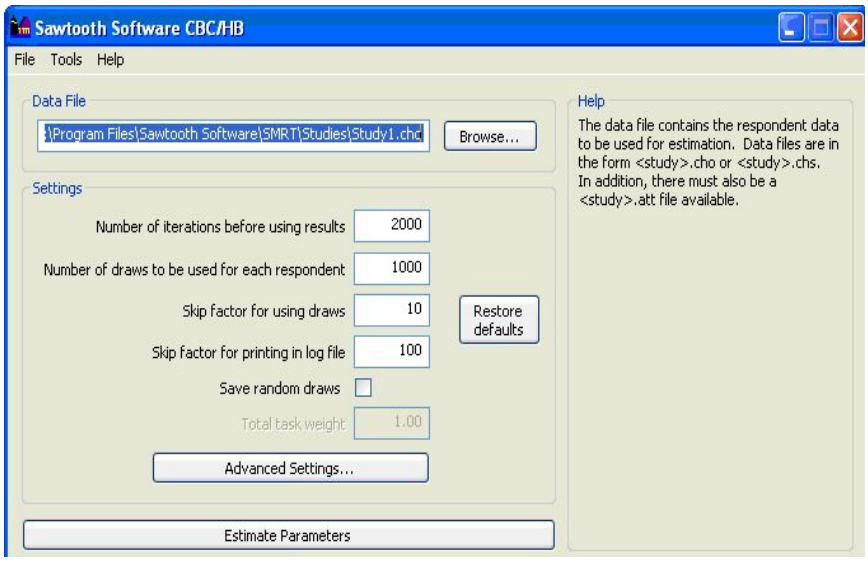
- il numero massimo di parametri che possono essere stimati per un individuo è 1000;
- il numero massimo di alternative in ogni "choice task" è 1000;
- il numero massimo di attributi e livelli è 1000;
- il numero massimo di "tasks" è 1000.

Utilizzare i valori delle part-worths individuali fornisce un enorme valore in termini di segmentazione, targeting e costruzione di simulazioni "what-if". Quindi CBC/HB combina in sé la validità di un "choice-based task" con la flessibilità di un'analisi a livello individuale che i ricercatori di marketing utilizzano frequentemente nella conjoint tradizionale. Valendosi del CBC System, si è realizzato uno studio congiunto (Study1) che si trova nella cartella Studies all'interno del programma. Il questionario è caratterizzato da 5 choice tasks con 2 concetti di prodotto ciascuno. Ora per l'analisi dei dati si utilitzza il modello "Hierarchical Bayes".
La schermata iniziale del programma CBC/HB rappresenta il menù principale.

SELEZIONARE IL FILE DI DATI
Per usare CBC/HB System, servono due file di dati che sono automaticamente prodotti da CBC System:

- il file studyname.att contenente le etichette dei livelli di attributo;
- il file studyname.cho contenente informazioni sul design o sui dati di scelta.

Per ritrovare il file .CHO è sufficiente ritornare allo studio creato con CBC e esportare il file delle risposte.



SETTING PARAMETERS

Il prossimo step è definire i valori dei parametri che governano la stima. I numeri mostrati in ogni campo sono valori di default che possono essere cambiati se si desidera:

- *numero di interazioni* prima che i risultati siano utilizzati: è il numero di iterazioni prima che la convergenza sia assunta: il valore di default è 2000, nonostante alcuni data sets richiedono meno iterazioni e altri molti di più. Una strategia è accettare il valore di default e monitorare il progresso della computazione in modo da fermarla se si raggiunge la convergenza;

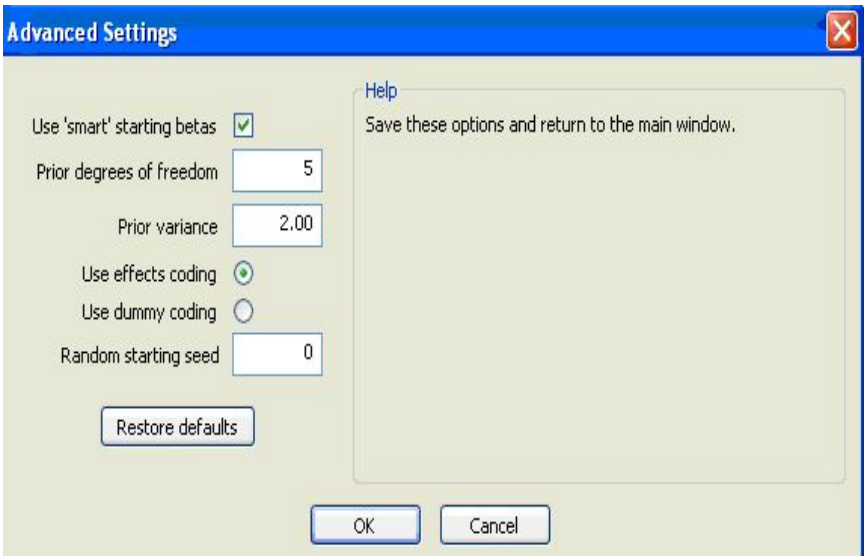
- *numero di draws* da usare per ogni rispondente: è il numero di iterazioni per il quale i risultati saranno disponibili per l'analisi;

- *skip factor for using draws*: indica con quale frequenza si prendono le iterazioni che servono per la stima (una iterazione ogni dieci);

- *skip factor for printing in log file*: controlla l'ammontare di dettagli che sono salvati nel file .LOG che riporta la storia di ogni iterazione.

ADVANCED SETTINGS

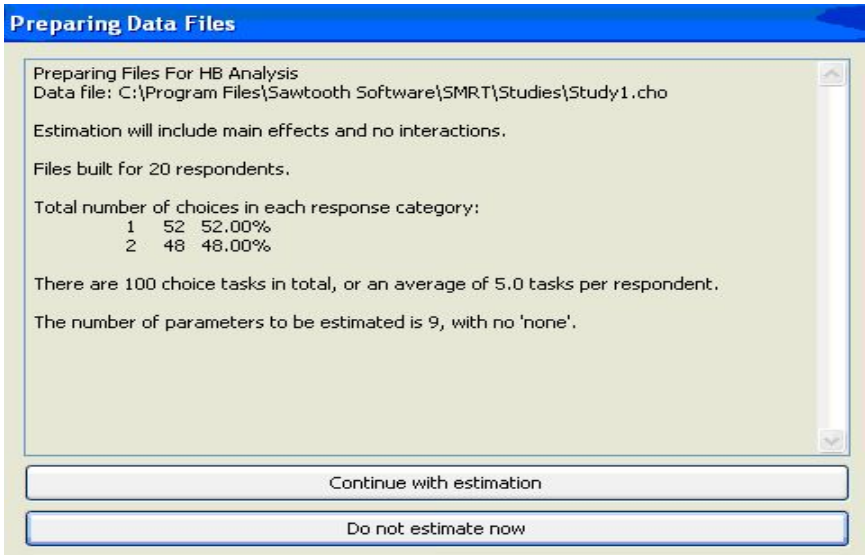
La maggior parte degli utilizzatori non cambiano i valori di default di questa finestra. Gli "advanced settings" possono fornire maggiore flessibilità e maggior controllo.



Use "smar" starting betas: le stime di partenza di beta e alpha definite dai realizzatori del software portano a una convergenza veloce. Purtroppo per data sets estremi , le stime di partenza possono ostacolare piuttosto che aiutare la convergenza. In questi casi l'utilizzatore del software può decidere di impostare a zero le stime di partenza di tutti i beta e alpha.
Prior degrees of freedom: sono i gradi di libertà addizionali per la matrice di covarianza a priori e possono assumere valori da 2 a 100000.
Prior variance: il valore di default è 2 per la varianza a priori di ogni parametro, ma l'utilizzatore può modificare questo valore (0-100). Aumentare la varianza a priori tende a dare maggior peso alla stima dei dati di ogni individuo e minor enfasi alla informazioni aggregate.
Using effects/dummy coding: con effects coding, l'ultimo livello di ogni attributo è omesso per evitare dipendenza lineare ed è stimato come la somma degli altri livelli entro l'attributo con segno negativo. Con dummy coding l'ultimo livello è omesso e assume un valore pari a 0 rispetto al quale sono stimati gli altri livelli entro l'attributo.

Effects/dummy coding producono gli stessi risultati nel caso di stime OLS o Logit, ma le utilità stimate usando effects/coding sono generalmente più facili da interpretare soprattutto se sono incluse le iterazioni. Per l'analisi HB i risultati possono dipendere dalla codifica.
Random Starting Seed: il valore "0" indica l'uso di un "random seed" basato sul computer clock.

PREPARARE I DATI
Una volta definiti tutti i parametri, si clicca **Estimate Parameters** e compare la seguente finestra:



Essa indica che solo i "main effects" sono inclusi. Se si vuole includere alcune interazioni o escludere qualche attributo, è necessario creare un file *study1.eff* con un Text Editor. Il file .EFF contiene informazioni sugli attributi da includere nell'analisi, sulle interazioni da considerare, sulla codifica delle variabili (part-worths o lineari). Se non viene realizzato il file .EFF tutti gli attributi sono inclusi, nessuna interazione è stimata e le variabili sono codificate come part-worths. Nella prima riga del file, si fa una lista degli effetti principali che si vogliono includere. Introdurre il segno negativo significa voler trattare le variabili come lineari. Le successive righe del file sono utilizzate per specificare

le two-way interactions. Ogni riga deve presentare i numeri corrispondenti agli attributi di cui si vuole conoscere gli effetti di interazione.
Esempio: il seguente file studyname.eff indica che:

- si vuole misurare gli effetti principali degli attributi 1,2,3;
- l'attributo 3 è codificato come lineare;
- il parametro "None" è incluso (presenza dello "0");
- si vuole stimare l'interazione per gli attributi 1 e 3.

1 2 -3 0
1 3

La finestra **Preparing Data Files** mostra inoltre:

- il numero di rispondenti;
- il numero di volte in cui i rispondenti hanno selezionato le alternative 1 e 2 dei "choice tasks";
- il numero di "task" per ogni questionario;
- il numero di parametri da stimare (dati P fattori con $m_1, m_2, m_3, \ldots, m_p$ livelli, il numero di parametri da stimare è dato :

$$\sum_{p=1}^{P}(m_p - p).$$

Se si è soddisfatti del modo in cui i dati sono stati preparati, si clicca **Continue with estimation** per iniziare le iterazioni HB.

SPECIFICARE COSTRAINTS (VINCOLI)
E' necessario preparare un file *studyname.con* con il proprio Text editor. Ogni riga del file deve contenere 3 valori:

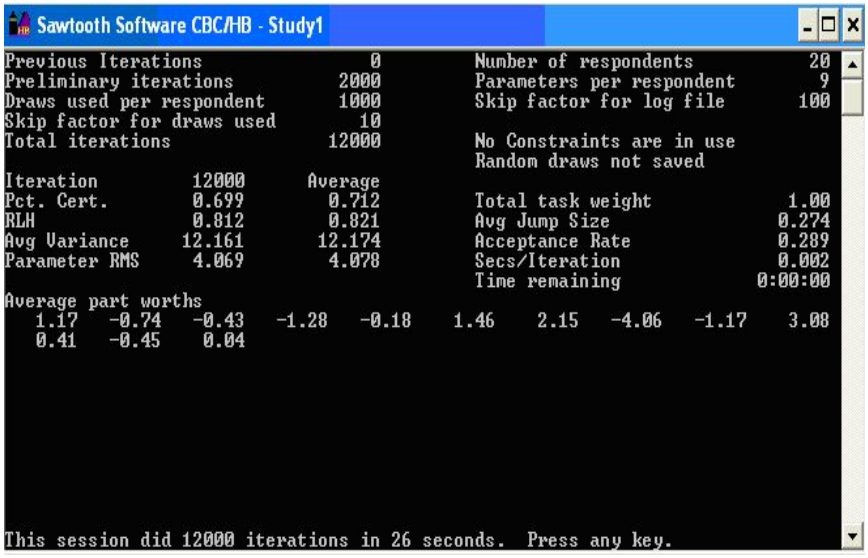
- il numero di attributo;

- il livello che dovrebbe ricevere l'utilità più alta;
- il livello che dovrebbe avere utilità più piccola rispetto il precedente.

Questo tipo di vincoli è particolarmente usato nel caso dell'attributo prezzo. *Esempio*: l'attributo 3 con 3 livelli è vincolato ad avere la part-worth del primo livello più grande di quella del secondo livello e quest'ultima più grande della part-worth del terzo livello.

3 1 2
3 2 3

MONITORARE LA COMPUTAZIONE



Descriviamo l'informazione che si ricava dalla schermata del monitoraggio. L'informazione in alto descrive i parametri definiti precedentemente : 2000 iterazioni iniziali, seguite da 10000 iterazioni, 1 iterazione su 10 è usata (skip) per il calcolo delle stime. Inoltre è indicato anche il numero dei rispondenti e dei parametri da stimare per ogni rispondente. Sulla sinistra dello schermo, sono riportate su due colonne alcuni indicatori statistici della bontà della stima :

- la prima colonna contiene il valore dell'iterazione precedente;

- la seconda colonna contiene il valore medio di tutte le iterazioni della sessione corrente.

La *"percent certainty"* e RLH derivano dalle probabilità dei dati. Si calcola la probabilità di scelta di ogni rispondente applicando un modello logit multinomiale che usa le stime correnti delle part-worths di ogni rispondente. La verosimiglianza è il prodotto di queste probabilità su tutti i rispondenti. Essendo il valore estremamente piccolo , si considera la "log verosimiglianza". La *percent certainty* indica quanto migliore è la soluzione corrente rispetto la 'soluzione perfetta'. La *percent certainty* varia fra 0 e 1: se è vicina a 1 la stima è buona. RLH è l'abbreviativo di "root likelihood" e rappresenta una misura della bontà della stima. RLH è la radice n-esima della verosimiglianza , dove n è il numero totale delle scelte fatte da tutti i rispondenti in tutti i tasks. Se RLH è pari a 1, la stima è perfetta. Le due statistiche finali "*Avg Variance*" e "*Parameter RMS* sono anch'essi indicatori della bontà delle stime .
Avg Variance è la media delle stime correnti delle varianze delle part-worths fra i rispondenti.
Parameter RMS è la radice della media dei quadrati di tutte le stime delle part-worths.
Sulla destra dello schermo sono mostrati:

- Average jump size: rappresenta la media delle ampiezze degli incrementi effettuati nell'algoritmo iterativo;

- Average acceptance ration: medie degli incrementi della verosimiglianza entro il quale si accettano le nuove stime.

Nella parte bassa della schermata sono presentate le utilità medie dei livelli per tutti i rispondenti.

UTILIZZARE I RISULTATI

Alla fine delle iterazioni, sono disponibili alcuni file contenenti i risultati. Il file studyname.csv contiene le part-worths per ogni rispondente e può essere aperto

direttamente con il programma "EXCEL".
Study1.csv presenta un record per ogni rispondente contenente:

- il numero del rispondente;
- il valore di RLH;
- il valore "0" per la compatibilità con altri moduli;
- il numero totale delle stime dei parametri (13 livelli);
- "0" se non è inclusa l'opzione None, "-1" altrimenti;
- i valori delle part-worths per ogni livello.

QUANTO BUONI SONO I RISULTATI ?
Molti articoli hanno discusso l'utilizzo della Hierarchical Bayes (HB) per le stime delle "part worths" individuali.

- Allenby, Arora e Ginger (1959) hanno mostrato come HB può essere usato in modo vantaggioso per introdurre informazioni a priori sulla monotonia delle part worths.
- Allenby e Ginger (1995) hanno mostrato che HB può essere usato per stimare le utilità individuali a partire da pochi dati provenienti da ogni individuo.
- Lenk, DeSarbo, Green e Young (1996) hanno mostrato che HB può stimare efficacemente le part-worths individuali quando ogni rispondente fornisce un numero di risposte superiore al numero di parametri da stimare.

Questi risultati suggeriscono che HB potrà diventare il metodo preferito per la stima delle utilità individuali. Purtroppo, HB ha da sempre richiesto lunghi tempi di computazione e per il suddetto motivo agli inizi degli anni '90 alcuni dubitavano della possibilità di applicarlo a situazioni pratiche del mondo reale. L'esempio di Allenby e Ginter ha interessato 600 rispondenti con la stima di

solo 14 parametri. La maggior parte della applicazioni commerciali coinvolge un data set di più grandi dimensioni. Poiché i computer di oggi sono diventati molto più veloci, è possibile ottenere stime HB per un problema esteso in tempi ragionevoli. Basandosi su alcuni studi reali, le conclusioni che si ricavano sulla bontà dei risultati indicano che il modello predice in maniera eccellente gli "holdout concept".

Bibliografia

[1] De Luca, Amedeo (1990). *Metodi statistici per le ricerche di mercato*. UTET Libreria, Torino.

[2] Brasini, Sergio (2002). *Statistica aziendale e analisi di mercato*. Il mulino, Bologna.

[3] Brasini, Sergio (1999). *Marketing e pubblicità: metodi di analisi statistica*. Il mulino, Bologna.

[4] Gustafsson, Anders (2000). *Conjoint measurements: methods and applications*. Springer, Berlino.

[5] Montgomery, Douglas C. (1991). *Design and analysis of experiments*. Wiley, New York.

[6] Pace, Luigi e Salvan, Alessandra (2001). *Introduzione alla Statistica 2: Inferenza, verosimiglianza, modelli*. CEDAM, Padova.

[7] Azzalini, Adelchi (2001). *Inferenza statistica, una presentazione basata sul concetto di verosimiglianza*. Springer, Milano.

[8] Liao, Tim Futing (1994). *Interpreting probability models: Logit, Probit and other Generalized Linear Models*. Sage, Thousand Oaks.

[9] Sawtooth Software, Inc. (1999). *Choice-based Conjoint (CBC) Techinical Paper*. Sequim, Wa.

[10] Sawtooth Software, Inc. (2000). *CBC Hierarchical Bayes Analysis Techinical*. Sequim, Wa.

[11] Sawtooth Software, Inc. (2003).*HB-Regression for Hierarchical Bayes Analysis*. Sequim, Wa.

[12] Piccinato, Ludovico (1996). *Metodi per le decisioni statistiche*. Springer-Verlag Italia, Milano.

[13] SPSS, Inc. (1997). *SPSS Conjoint 8.0*. SPSS Press, Chicago.