Entreprise Guide ou comment rendre presque séduisantes les procédures statistiques de SAS

Présentation au
Club des Utilisateurs SAS de Québec
30 octobre 2006

Jean Hardy
Services Conseils Hardy

LES SERVICES CONSEILS HARDY CONSULTING SERVICES

Plan

- Tirer un échantillon d'une table existante
  - Échantillon simple
  - Échantillon stratifié
- Diviser une table en *k* sous-groupes
- Générer aléatoirement des données
- Régression linéaire – un exemple de base
- Conclusion

LES SERVICES CONSEILS HARDY CONSULTING SERVICES

Tirer un échantillon d'une table existante (1)

- Utilité:
  - Tester une requête ou un processus en cours de développement, avec une partie de la table initiale
  - Réaliser des études impraticables avec toutes les données (en assurance par exemple)
  - Pour fins d'analyse plus sophistiquée (techniques de ré-échantillonnage diverses par exemple)
- La syntaxe de **SURVEYSELECT** demeure déroutante, car utilisée rarement

LES SERVICES CONSEILS HARDY CONSULTING SERVICES

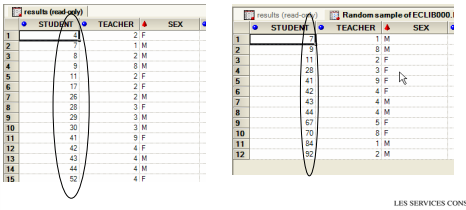
Tirer un échantillon d'une table existante (2)

- Plusieurs types d'échantillons:
  - échantillon simple, où *n* observations ou encore *m* % des observations sont extraites;
  - échantillon stratifié (1 ou *k* var. catégorielles);
  - plusieurs réplications si nécessaire (pour ré-échantillonnage ou "bootstrap" par exemple).

LES SERVICES CONSEILS HARDY CONSULTING SERVICES

Échantillon simple (1)

- Exemple – tirer au hasard 40% des rangées d'une table SAS nommée *RESULTS* (*n*=28)



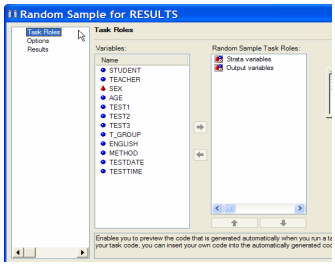
LES SERVICES CONSEILS HARDY CONSULTING SERVICES

Échantillon simple (2)

- Sélectionner **Data > Random Sample**
- Au besoin, dans **Task Roles**, restreindre les variables copiées à l'aide de la zone **Output Variables**
- Dans **Options** (via panneau de navigation), choisir *Percent of rows* plutôt que *Rows* dans la zone **Sample size**, puis fournir le % à tirer
- S'assurer que la zone **Sample method** contient *Single random sample without…*

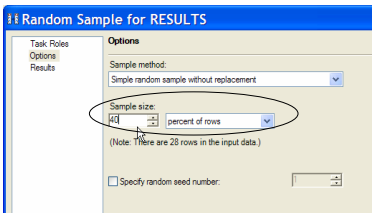
LES SERVICES CONSEILS HARDY CONSULTING SERVICES

Échantillon simple (3)



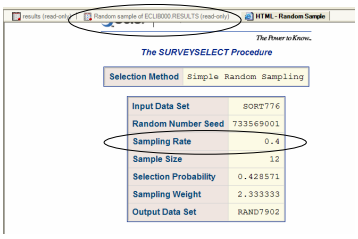
LES SERVICES CONSEILS HARDY CONSULTING SERVICES

Échantillon simple (4)



LES SERVICES CONSEILS HARDY CONSULTING SERVICES

Échantillon simple (5)



LES SERVICES CONSEILS HARDY CONSULTING SERVICES

Échantillon simple (6)

- Code SAS généré:

```
PROC SURVEYSELECT DATA=WORK.SORT776
    OUT=SASUSER.RAND7902(LABEL="Random sample of ECLIB000.RESULTS")
    METHOD=SRS
    RATE=0.4
    ;
RUN;
```

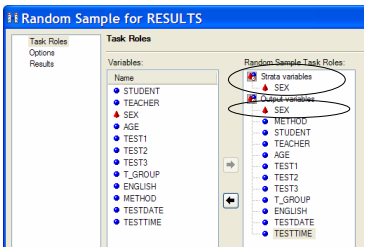
LES SERVICES CONSEILS HARDY CONSULTING SERVICES

Échantillon stratifié (1)

- Le nombre ou la proportion d'observations est tirée au sein de chaque niveau d'une ou plusieurs variables catégorielles (les *strates*)
- Procéder comme auparavant, en plaçant dans la zone *Strata* la ou les variable(s) de stratification
- Il faut ici copier toutes les variables requises dans la table en sortie, y compris la variable de stratification

LES SERVICES CONSEILS HARDY CONSULTING SERVICES

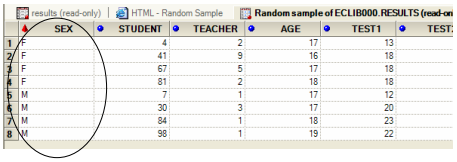
Échantillon stratifié (2)



LES SERVICES CONSEILS HARDY CONSULTING SERVICES

Échantillon stratifié (3)

- La proportion de *25* % représente ici 4 observations de chaque sexe (*F*: 4/15, *M*: 4/13)



results (read-only) | HTML - Random Sample | Random sample of ECLIB000.RESULTS (read-on

	SEX	STUDENT	TEACHER	AGE	TEST1	TEST
1	F	4	2	17	13	
2	F	41	9	16	18	
3	F	67	5	17	18	
4	F	81	2	18	18	
5	M	7	1	17	12	
6	M	30	3	17	20	
7	M	84	1	18	23	
8	M	98	1	19	22	

LES SERVICES CONSEILS HARDY CONSULTING SERVICES

Échantillon stratifié (4)

- Code SAS généré:

```
PROC SORT
    DATA=ECLIB000.RESULTS
    OUT=WORK.SORT1147
    ;
    BY SEX;
RUN;

PROC SURVEYSELECT DATA=WORK.SORT1147
    OUT=SASUSER.RAND816(LABEL='Random sample of ECLIB000.RESULTS')
    METHOD=SRS
    RATE=0.25
    ;
    STRATA SEX;
    ID SEX METHOD STUDENT TEACHER AGE TEST1 TEST2 TEST3 T_GROUP ENGLISH TESTDATE T
RUN;
```

LES SERVICES CONSEILS HARDY CONSULTING SERVICES

Échantillon stratifié (5)

- L'énoncé **ID** est facultatif, en son absence toutes les variables sont copiées dans la nouvelle table.
- Pas d'énoncé **VAR**!
- Pour tirer un nombre fixe d'observations plutôt qu'une proportion, remplacer **RATE=** par **SAMPSIZE=**.
- À défaut d'employer souvent la procédure, Enterprise Guide constitue une bonne alternative.

LES SERVICES CONSEILS HARDY CONSULTING SERVICES

Diviser une table en *k* sous-groupes (1)

- Diviser un échantillon de taille *n* en *k* sous-groupes suivant la valeur d'une variable continue
- Sous-groupes de taille égale ou presque
- Les *k*/*n* valeurs les plus petites dans le premier sous-groupe, les *k*/*n* valeurs les plus grandes dans le dernier sous-groupe, etc.
- Au besoin, procéder ainsi au sein de chaque niveau d'une variable catégorielle
- La table SAS initiale n'a pas à être triée

LES SERVICES CONSEILS HARDY CONSULTING SERVICES

Diviser une table en *k* sous-groupes (2)

- Exemple – un fichier contient 225 observations que l'on veut diviser en 4 sous-groupes suivant la valeur de la variable *TOT_SCORE*:

	SEX	AGE	TOT_SCORE
1	F	64	74.8
2	F	22	75.9
3	F	29	55.9
4	F	54	57.3
5	F	26	53.9
6	F	46	74.8
7	M	65	86
8	M	41	61.8
9	F	34	61.1
10	M	56	59.2
11	M	35	57.1
12	F	40	68.4

La variable peut contenir plusieurs valeurs identiques

LES SERVICES CONSEILS HARDY CONSULTING SERVICES

Diviser une table en *k* sous-groupes (3)

- Groupes numérottés de *0* à *k-1*:

	SEX	AGE	TOT_SCORE	Rank for Variable TOT_SCORE
1	F	64	74.8	3
2	F	22	75.9	3
3	F	29	55.9	0
4	F	54	57.3	0
5	F	26	53.9	0
6	F	46	74.8	3
7	M	65	86	3
8	M	41	61.8	1
9	F	34	61.1	1
10	M	56	59.2	1
11	M	35	57.1	0
12	F	40	68.4	2

LES SERVICES CONSEILS HARDY CONSULTING SERVICES

Diviser une table en *k* sous-groupes (4)

- Sous-groupes constitués:
  - d'un nombre presqu'identique d'observations
  - de plages de valeurs évidemment non-superposées

The MEANS Procedure

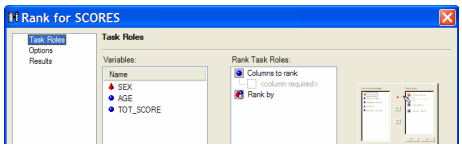
Analysis Variable : TOT_SCORE

Rank for Variable TOT_SCORE	N Obs	Mean	Std Dev	Minimum	Maximum
0	56	51.8	5.0	40.7	57.5
1	56	61.2	1.9	58.1	64.6
2	57	68.8	2.2	64.7	72.4
3	56	78.3	4.9	72.5	93.5

LES SERVICES CONSEILS HARDY CONSULTING SERVICES

Diviser une table en *k* sous-groupes (5)

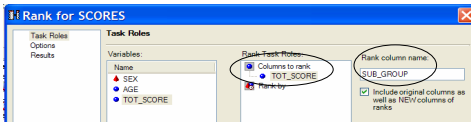
- Sélection **Data > Rank**



LES SERVICES CONSEILS HARDY CONSULTING SERVICES

Diviser une table en *k* sous-groupes (6)

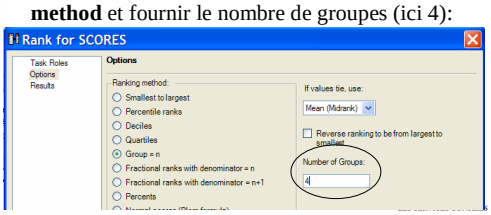
- Amener la variable *TOT_SCORE* dans le champ **Columns to rank** et fournir un nom à celle qui contiendra les rangs *(SUB_GROUP* ici):



LES SERVICES CONSEILS HARDY CONSULTING SERVICES

Diviser une table en *k* sous-groupes (7)

- Cliquer sur **Options** (panneau de navigation), sélectionner **Groups=n** dans la zone **Ranking method** et fournir le nombre de groupes (ici 4):



Diviser une table en *k* sous-groupes (8)

- Code SAS généré:

```
PROC RANK DATA = WORK.SORT285
    GROUPS=4
    TIES=MEAN
    OUT=SASUSER.RANK8446(LABEL="Rank Analysis for ECLIB000.SCORES");
    VAR TOT_SCORE;
RANKS SUB_GROUP ;
/*
   End of task code.
   ---------------------------------------------------------------- */
RUN; QUIT;
```

Générer aléatoirement des données (1)

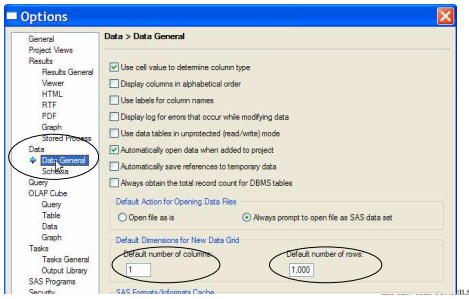
- Utilité:
  - Fabriquer des données de test
  - Obtenir des données répondant à certains critères (ex. un % fixe d'observations comportant des erreurs)
- Exemple: une table contenant *1000* observations avec une variable numérique faite de *0* et de *1* où 70% des observations contiennent la valeur *1*

Générer aléatoirement des données (2)

- Générer d'abord une table SAS contenant le nombre d'observations désirées
  - Ne pas utiliser directement la séquence **File > New > Data** car cela crée une nouvelle table ayant par défaut 10 rangées et 8 colonnes
  - Pour fixer le nombre de rangées voulues, sélectionner **Tools > Options > Data > Data General**

LES SERVICES CONSEILS HARDY CONSULTING SERVICES

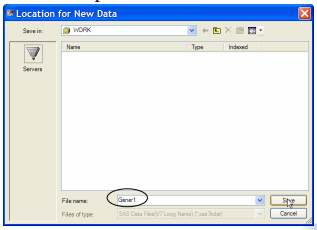
Générer aléatoirement des données (3)



LES SERVICES CONSEILS HARDY CONSULTING SERVICES

Générer aléatoirement des données (4)

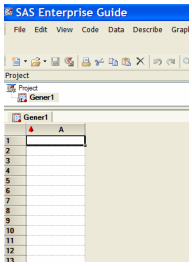
- Générer ensuite une table vide à l'aide de **File > New > Data**, puis la nommer



LES SERVICES CONSEILS HARDY CONSULTING SERVICES

Générer aléatoirement des données (5)

- La nouvelle table contient une variable (caractère, nommée *A* par défaut) et 1000 observations, tel que demandé



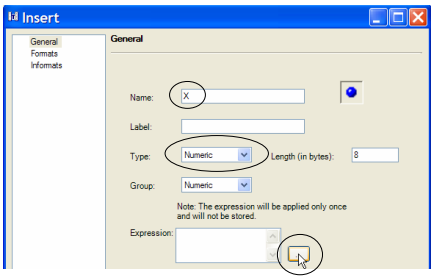
LES SERVICES CONSEILS HARDY CONSULTING SERVICES

Générer aléatoirement des données (6)

- Insérer ensuite une nouvelle colonne à l'aide de **Data > Columns > Insert**
- Fournir pour la nouvelle variable:
  - son nom
  - son type (ici numérique)
  - l'expression servant à la définir
  - cliquer ici sur le bouton situé à droite de la zone **Expression** pour choisir la fonction à utiliser

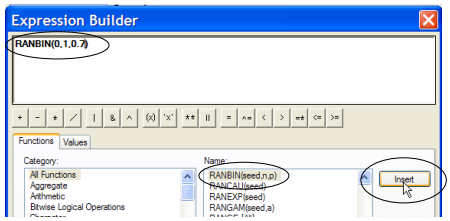
LES SERVICES CONSEILS HARDY CONSULTING SERVICES

Générer aléatoirement des données (7)



LES SERVICES CONSEILS HARDY CONSULTING SERVICES

Générer aléatoirement des données (8)



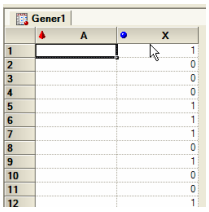
LES SERVICES CONSEILS HARDY CONSULTING SERVICES

Générer aléatoirement des données (9)

- L'expression `RANBIN(0,1,0,7)` signifie:
  - *0* : utiliser l'horloge interne de l'ordinateur pour initialiser la série de nombres
  - *1* : génère des entiers entre *0* et *1* seulement (*3* comme 2ième paramètre générera des entiers compris entre *0* et *3* par exemple)
  - *0,7* : 70 % des valeurs générés seront des *1* et 30% seront des *0* (en raison de la distribution dite *binomiale* utilisée)

LES SERVICES CONSEILS HARDY CONSULTING SERVICES

Générer aléatoirement des données (10)



The FREQ Procedure

X	Frequency	Percent
0	308	30.80
1	692	69.20

LES SERVICES CONSEILS HARDY CONSULTING SERVICES

Générer aléatoirement des données (11)

- Une méthode du même genre permettra de générer:
  - des âges, par exemple des entiers compris entre *18* et *85* (68 valeurs possibles, chacune aussi probable que les autres pour simplifier);
  - le premier caractère d'un code postal situé au Québec ou au Nouveau-Brunswick (lettre *E*, *G*, *H* ou *J*)

LES SERVICES CONSEILS HARDY CONSULTING SERVICES

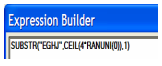
Générer aléatoirement des données (12)

Age
37
72
70
80
20
73
30
58
36
57
67
64
41
54

Age	Frequency	Percent
18	13	1.30
19	19	1.90
20	19	1.90
21	16	1.60
22	17	1.70
23	13	1.30
82	12	1.20
83	10	1.00
84	6	0.60
85	13	1.30

LES SERVICES CONSEILS HARDY CONSULTING SERVICES

Générer aléatoirement des données (13)



PC_1
J
J
G
E
G
E
J
J
J
E
G
E
H
H

PC_1	Frequency	Percent
E	258	25.80
G	245	24.50
H	228	22.80
J	269	26.90

LES SERVICES CONSEILS HARDY CONSULTING SERVICES

Générer aléatoirement des données (14)

- Ces expressions sont complexes, mais qui a dit qu'il faut les connaître par coeur?
- Un document Word, bien organisé, peut servir de "procédurier" et les expressions en sont copiées au besoin.

LES SERVICES CONSEILS HARDY CONSULTING SERVICES

Régression multiple (1)

- Évaluer le degré de relation entre une variable dépendante continue et une série de variables prédictrices
- Exemple: prédire la consommation de protéines per capita de 51 pays à partir d'indicateurs socio-économiques et agricoles
- Pays: membres du G20 + autres pays parmi les plus développés en Europe et sur autres continents (Chili, Thaïlande, Malaysie, Qatar, Koweït, Iran, Israël, EAU, etc.)

LES SERVICES CONSEILS HARDY CONSULTING SERVICES

Régression multiple (2)

COUNTRY	G20_MEMB	GDP_PCAP	ENERG_PRO	ENERG_CON	RD_GDP	RD_P_CAP	POOR_SHA	RNG_TEMP	PCT_URB	POP_DENS	BI
Argentina	1	10,513	3.1	145	0.39	44	9.1	17	90.1	14	
Australia	1	26,659	16.8	210	1.55	405	5.9	17	92.2	3	
Austria	0	28,093	1.2	134	2.21	645	8.1	17	65.8	97	
Bahamas	0	14,650	0.0	250	0.50	150	5.5	8	89.5	22	
Bahrain	0	13,674	19.9	596	0.19	32	5.0	17	90.3	932	
Belarus	0	8,745	0.4	468	0.64	35	8.4	28	69.8	50	
Belgium-Luxe	0	29,861	1.7	185	2.23	615	8.0	17	96.9	325	
Brazil	1	7,710	0.8	146	1.04	77	2.4	8	83.0	21	
Bulgaria	0	6,517	1.8	350	0.49	35	6.7	28	67.5	68	
Canada	1	29,641	16.1	278	2.00	588	7.0	28	80.5	3	
Chile	0	9,870	0.4	168	0.54	52	3.3	8	87.0	21	
China	1	4,566	0.8	219	1.23	56	4.7	17	38.9	137	
Croatia	0	9,729	1.1	188	1.14	116	8.3	17	58.1	78	
Cyprus	0	11,721	0.0	182	0.27	49	6.1	17	69.1	86	
Czech Republi	0	15,338	3.9	272	1.30	205	10.0	17	74.1	130	
Denmark	0	28,896	6.8	124	2.51	778	8.3	17	85.4	125	
Finland	0	25,735	2.3	268	3.46	905	9.6	28	60.9	15	
France	1	25,904	2.9	171	2.27	611	7.2	17	76.2	111	
Germany	1	26,189	2.2	161	2.64	686	8.5	17	88.0	231	
Greece	0	18,495	1.2	146	0.65	116	7.1	17	61.1	83	
Hungary	0	13,567	1.7	187	1.01	135	9.5	28	64.9	106	

LES SERVICES CONSEILS HARDY CONSULTING SERVICES

Régression multiple (3)

- Prédicteurs économiques:
  - Produit intérieur brut per capita, R-D per capita
  - Tonnes d'énergie produite et consommée per capita
- Prédicteurs sociaux:
  - Espérance de vie
  - % de richesse du pays détenue par 20% + pauvres
  - Densité de population, % de la pop. Urbanisée
  - Appartenance aux principales religions (qq pays ont plus d'une appartenance, tel Canada, Israël, Singapour, Maylaisie)

LES SERVICES CONSEILS HARDY CONSULTING SERVICES

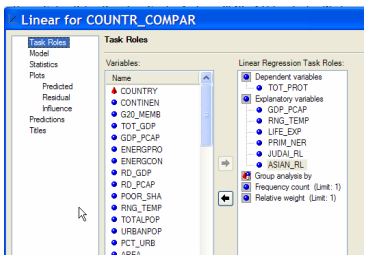
Régression multiple (4)

- Prédicteurs agricoles:
  - Quantité de viande bovine produite per capita
  - Captures de poisson per capita
  - Superficies en culture per capita
  - Écart entre les températures moyennes des saisons chaudes et froides

LES SERVICES CONSEILS HARDY CONSULTING SERVICES

Régression multiple (5)

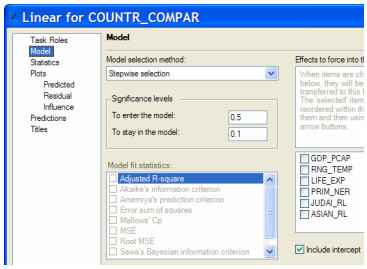
Rôle des variables



LES SERVICES CONSEILS HARDY CONSULTING SERVICES

Régression multiple (6)

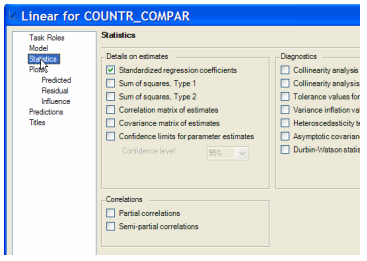
Construc. du modèle



LES SERVICES CONSEILS HARDY CONSULTING SERVICES

Régression multiple (7)

Statis-tiques calculées



LES SERVICES CONSEILS HARDY CONSULTING SERVICES

Régression multiple (8)

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-10.05809	26.80296	12.70745	0.14	0.7092
GDP_PCAP	0.00057597	0.00021270	661.71546	7.33	0.0095
LIFE_EXP	0.59575	0.32969	294.64671	3.27	0.0775
PRIM_NER	0.56274	0.20717	665.80476	7.38	0.0093
JUDAI_RL	24.78189	9.78360	578.98153	6.42	0.0149
ASIAN_RL	-16.65154	4.26581	1374.98797	15.24	0.0003

LES SERVICES CONSEILS HARDY CONSULTING SERVICES

Conclusion

- EG aide à exploiter, pour la gestion de données, les capacités de procédures peu usuelles (RANK, SURVEYSELECT, STANDARD, TRANSPOSE), en remplacement d'étapes DATA
- S'ajoute donc à ses capacités reconnues d'interface facilitant l'exploitation des procédures contenues dans les modules SAS/GRAPH et SAS/STAT

LES SERVICES CONSEILS HARDY CONSULTING SERVICES

Contacts

Jean Hardy
Services Conseils Hardy Inc.
4715 des Replats, suite 260
Québec G2J 1B8
(418) 626-1666
jhardy@schardy.qc.ca

LES SERVICES CONSEILS HARDY CONSULTING SERVICES