SPSS - Etude de cas

Contexte

Contexte : Un entraîneur souhaite mieux comprendre les facultés de résistance d'athlètes reprenant les entraînements après une période de repos forcé suite à une blessure

Il voudrait savoir si la résistance des athlètes [mesurée par le temps de récupération après un marathon (temps nécessaire pour atteindre x pulsations par minute) et le fait que les athlètes ont ou non fait des arrêts durant la course (course avec arrêt ou sans arrêt)] peut être déterminée par la durée du repos, le sexe de l'athlète et les vitamines prises durant le mois de préparation.

Il pense que cette étude pourra l'aider à améliorer ses entraînements pour être plus adaptés aux athlètes.

Remarque : Il s'agit d'une étude fictive

Données

Données: Pour chaque athlète, nous avons les informations suivantes :

Colonne	Nom SPSS	Nom de variable	Label	Values
Col1	V1	Date	Date de la mesure	
Col2	V2	Identifiant	Identifiant de l'athlète	
Col3	V3	Sexe	Sexe de l'athlète	1=Homme 2=Femme
Col4	V4	Vitamine	Vitamine prise par l'athlète	1=Vitamine A 2=Vitamine B 3=Vitamine C
Col5	V5	Absence	Nombre de jours de repos	
Col6	V6	Recup1	Nombre de seconde pour récupérer après le marathon 1	
Col7	V7	Recup2	Nombre de seconde pour récupérer après le marathon 2	
Col8	V8	Recup3	Nombre de seconde pour récupérer après le marathon 3	
Col9	V9	Arret1	Marathon 1 réalisé avec ou sans arrêt	1=Sans arrêt 2=Avec arrêts
Col10	V10	Arret2	Marathon 2 réalisé avec ou sans arrêt	0=Sans arrêt 1=Avec arrêts
Col11	V10	Fausse_Date	Date inventée	

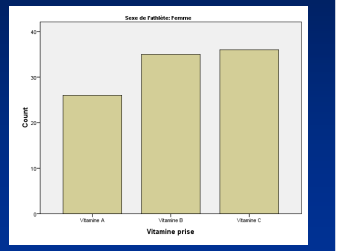
Analyse d'une variable qualitative

Visualisation graphique

- Ex : Visualiser la répartition des 3 types de vitamines chez les femmes
  → Pour sélectionner les femmes : *SPSS : Data → Select Cases → if ...*
- Diagramme en barres

SPSS : Graphs → Legacy Dialogs → Bar (Simple)

→ Une barre par catégorie
→ Fréquence ou pourcentage



- Diagramme de Pareto

SPSS : Analyze → Quality Control → Pareto Charts

→ Une barre par catégorie
→ Les barres sont ordonnées selon leur hauteur

Tableaux: Statistiques descriptives

- Ex: Résumer les proportions observées de la variable Arret1
  → Table de fréquence :
  SPSS : Analyze → Descriptive Statistics → Frequencies

Marathon 1 réalisé avec ou sans arrêt

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Sans arrêt	97	55,4	55,4	55,4
	Avec arrêts	78	44,6	44,6	100,0
	Total	175	100,0	100,0	

- Pour caractériser une variable ordinale
  → Utiliser la médiane ou le mode :
  SPSS : Analyze → Descriptive Statistics → Frequencies (Statistics)

Inférence : Test sur une proportion

Test binomial sur une proportion

- Ex: Tester si la proportion «avec arrêts» versus «sans arrêt» est la même

SPSS : Analyze → Non Parametric Tests → Binomial

Binomial Test

		Category	N	Observed Prop.	Test Prop.	Asymp. Sig. (2-tailed)
Marathon 1 réalisé avec ou sans arrêt	Group 1	Sans arrêt	97	,55	,50	,173[a]
	Group 2	Avec arrêts	78	,45		
	Total		175	1,00		

a. Based on Z Approximation.

→ H_0: proportions identiques ($\pi_A = \pi_B = 0.5$)
H_1: proportions différentes ($\pi A \neq \pi B \neq 0.5$)
P-valeur=0.173 => *P*-valeur>0.05 => On ne rejette pas H_0
=> On peut considérer que le nombre d'athlètes qui arrêtent au moins une fois durant le marathon est équivalent au nombre qui ne s'arrêtent pas

→ Ce test ne peut être appliqué que lorsque la variable d'intérêt ne peut prendre que 2 valeurs (ex:"avec" versus "sans")

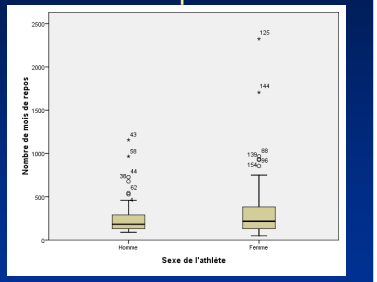
Analyse d'une variable quantitative en fonction d'au moins une variable qualitative

Visualisation graphique

- Ex : Visualiser la durée de repos (absence) en tenant compte du sexe

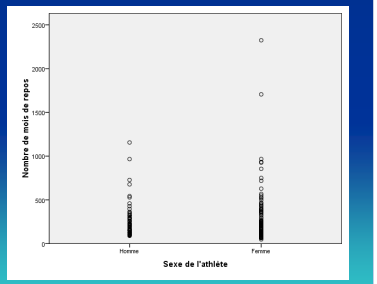
Boxplot

SPSS : Graphs → Chart Builder → Boxplot



Scatter/Dot

SPSS : Graphs → Chart Builder → Scatter/Dot



Tableaux: Statistiques descriptives

- Ex: Résumer la variable temps de repos (Absence) en fonction du sexe

SPSS : Analyze → Descriptive Statistics → Explore
ou Analyze → Reports → Case Summaries

Case Summaries

Nombre de mois de repos

Sexe ...	N	Mean	Median	Minimum	Maximum	Std. Deviation	Variance
Homme	78	245,09	182,50	89	1156	185,057	34246,161
Femme	97	311,21	217,00	48	2324	323,320	104535,895
Total	175	281,74	202,00	48	2324	271,875	73916,057

Inférence : Test sur 2 moyennes

Tests de comparaison de 2 moyennes indépendantes

- Ex: Tester si le log du temps de Recup1 diffère en moyenne selon le sexe
- Test-t pour 2 échantillons indépendants
  - → **Utilisation :** Lorsque les données à comparer sont indépendantes
  - → **Conditions :** Normalité des distributions, égalité des variances et indépendance des observations (transformations possibles pour la normalité)

  SPSS : Analyze → Compare Means → Independent-Samples T Test

- Tests non-paramétrique (normalité non respectée ou données ordinales)
  - → **Utilisation :** Quand les données ne se distribuent pas normalement dans au moins un des groupes ou qu'il s'agit de données ordinales
  - → **Tests :** Test de Mann-Whitney (≈Wilcoxon Rank-Sum), test de la médiane

  SPSS : Analyze → Nonparametric Tests → 2 Independent Samples (Mann-Whitney)
  SPSS : Analyze → Nonparametric Tests → k Independent Samples (Median)

Modélisation : Régression logistique

- Comment juger si le modèle est bon ?
  → En regardant les pseudo R^2

Mesures d'ajustement

Critère d'Akaike : AIC = -2 ln(L)+2*(nb de param)
Critère de Schwartz : SIC = -2 ln(L)+(nb de param)*ln(nb d'obs)
Rapport de vraisemblance: -2LL = -2*ln(max de vraisemblance)

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	227,409[a]	,050	,067

Mesures de la taille de l'effet : Pseudo R^2

Cox & Snell R^2 : Difficile à interpréter (max<1)
Nagelkerke R^2 : [0 ; 1] = Mesure de la force de l'association

Modélisation : Régression logistique

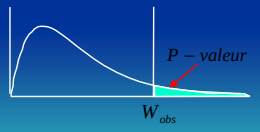
- Comment teste-t-on la significativité des paramètres ?
  → Test de Wald

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B) Lower	Upper
Step 1	LogAbsence	,703	,244	8,311	1	,004	2,019	1,252	3,256
	Constant	-4,178	1,330	9,871	1	,002	,015		

- But: tester H_0: $\beta_i = 0$ contre H_1: $\beta_i \neq 0$
- La statistique de Wald est définie par:
- On rejette H_0 si la p-valeur ($P(\chi^2_1 > W_{obs})$) est inférieure à un seuil fixé

$$W_{obs} = \frac{b_i^2}{s^2(b_i)} \sim \chi_1^2 \text{ sous } H_0$$



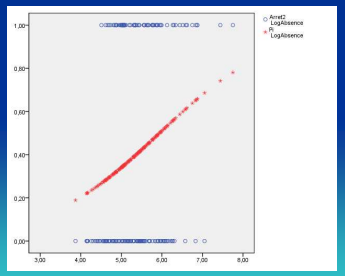
Modélisation : Régression logistique

- Comment rapporter le modèle estimé ?
  → Sous la forme d'une équation - catégorie de référence : Y=1 :

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B) Lower	Upper
Step 1	LogAbsence	,703	,244	8,311	1	,004	2,019	1,252	3,256
	Constant	-4,178	1,330	9,871	1	,002	,015		

$$\ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -4.178 + 0.703 \times LogAbsence \quad \rightarrow \quad \hat{\pi} = \frac{\exp(-4.178 + 0.703 \times LogAbsence)}{1 + \exp(-4.178 + 0.703 \times LogAbsence)}$$



Modélisation : Régression logistique

- Que représentent le « Odds » et le « Odds ratio » ?
  → Odds (ou cotes)
  - Ex: Le risque pour un athlète de s'arrêter durant le 2ème marathon (Arret2) en sachant qu'il a eu 18 mois de convalescence (Absence)

$$\frac{\pi}{1-\pi} = \frac{\text{Probabilité de s'arrêter au moins 1 fois sachant le LogAbsence}}{\text{Probabilité de ne pas s'arrêter sachant le LogAbsence}}$$

$$\hat{\pi} = \frac{\exp(-4.178 + 0.703 \times Log(18*30))}{1 + \exp(-4.178 + 0.703 \times Log(18*30))} = 0.561$$

→ Pour un athlète qui a eu une convalescence de 18 mois, la probabilité qu'il s'arrête au moins une fois durant le 2ème marathon est estimée à 56%

Modélisation : Régression logistique

- Que représentent le « Odds » et le « Odds ratio » ?

  → Odds Ratio (ou rapport de cotes)

  - Ex: Le risque relatif pour un athlète avec un temps de convalescence de X+1 de s'arrêter durant le 2ème marathon par rapport à un athlète avec un temps de convalescence de X (LogAbsence)

$$OR = \frac{\frac{\pi_1}{(1-\pi_1)}}{\frac{\pi_2}{(1-\pi_2)}} = \frac{\text{Probabilité de s'arrêter au moins 1 fois sachant le temps de convalescence} = X+1}{\text{Probabilité de ne pas s'arrêter sachant le temps de convalescence} = X+1} \Big/ \frac{\text{Probabilité de s'arrêter au moins 1 fois sachant le temps de convalescence} = X}{\text{Probabilité de ne pas s'arrêter sachant le temps de convalescence} = X}$$

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1	LogAbsence	,703	,244	8,311	1	,004	2,019	1,252	3,256
	Constant	-4,178	1,330	9,871	1	,002	,015		

$OR=\exp(\beta_1)$

Un athlète avec un temps de convalescence d'une unité en plus au niveau du LogAbsence a 2,019 fois plus de chance de s'arrêter au moins une fois durant le marathon 2

Modélisation : Régression logistique

- Comment se mettre dans de bonnes conditions pour obtenir un modèle qui soit valide ?

  → En repérant les outliers et points influents (via l'analyse des résidus, standardized residuals, leverage, Cook)

  → En incluant toutes les variables influentes dans le modèle et uniquement celles-là

  → En vérifiant que la relation entre VI et log odds de VD est linéaire

  → En vérifiant l'absence de multicolinéarité

  → En utilisant des échantillons de taille suffisante

  → En s'assurant que les conditions d'application des tests χ^2 sont respectées

  …

Modélisation : Régression logistique

- Comment teste-t-on la significativité des paramètres ?
  - → Test de Wald

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B) Lower	Upper
Step 1	Vitamine			2,442	2	,295			
	Vitamine(1)	,974	,629	2,395	1	,122	2,647	,772	9,084
	Vitamine(2)	,592	,474	1,562	1	,211	1,808	,714	4,579
	LogAbsence	1,179	,399	8,735	1	,003	3,250	1,487	7,100
	Constant	-7,261	2,423	8,979	1	,003	,001		

  - → Le temps de convalescence semble important pour prédire le fait qu'un athlète s'arrête ou non durant le 2ème marathon
  - → La vitamine prise par l'athlète ne semble pas avoir d'effet sur le fait que l'athlète s'arrête durant le marathon

Modélisation : Régression logistique

- Comment rapporter le modèle estimé ?
  - → Sous la forme d'une équation par niveau de la variable qualitative (Imaginons que la variable Vitamine soit gardée dans le modèle) :

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B) Lower	Upper
Step 1	Vitamine			2,442	2	,295			
	Vitamine(1)	,974	,629	2,395	1	,122	2,647	,772	9,084
	Vitamine(2)	,592	,474	1,562	1	,211	1,808	,714	4,579
	LogAbsence	1,179	,399	8,735	1	,003	3,250	1,487	7,100
	Constant	-7,261	2,423	8,979	1	,003	,001		

  - → La probabilité qu'un athlète s'arrête durant le marathon 2

$$\text{VitamineA} : \exp(-7.261 + 0.974 + 1.179 \times \textit{LogAbsence})$$
$$\text{VitamineB} : \exp(-7.261 + 0.592 + 1.179 \times \textit{LogAbsence})$$
$$\text{VitamineC} : \exp(-7.261 + 1.179 \times \textit{LogAbsence})$$

Modélisation : Régression logistique

- Comment se mettre dans de bonnes conditions pour obtenir un modèle qui soit valide ?

Mêmes conditions qu'exposé précédemment :

  - → En repérant les outliers et points influents (via l'analyse des résidus, standardized residuals, leverage, Cook)
  - → En incluant toutes les variables influentes dans le modèle et uniquement celles-là
  - → En vérifiant que la relation entre VI et log odds de VD est linéaire
  - → En vérifiant l'absence de multicolinéarité
  - → En utilisant des échantillons de taille suffisante
  - → En s'assurant que les conditions d'application des tests χ^2 sont respectées

  …