VALIDATION DES RESULTATS

DU LOGICIEL DE STATISTIQUES STATEL

Le NIST (National Institute of Standards and Technology) et son laboratoire ITL (Information Technology Labolatory), situés aux USA, ont élaboré des jeux de données statistiques, accompagnés de leurs résultats certifiés, dans le but d'aider les utilisateurs de logiciels statistiques à évaluer la qualité de leur outil.

Afin d'assurer les utilisateurs de StatEL de la véracité des calculs effectués, nous avons entrepris de confronter les résultats présentés par notre logiciel aux résultats certifiés du laboratoire ITL, mis à disposition de la communauté du Net.

Vous pouvez visiter le site "Statistical Reference Datasets" du NIST à la page suivante : http://www.itl.nist.gov/div898/strd/

Nous n'avons utilisé que les jeux de données applicables aux tests disponibles dans le logiciel StatEL, à savoir :

- En statistiques univariées :
  - **PiDigits** (Niveau de difficulté facile)
  - **Lottery** (Niveau de difficulté facile)
  - **NumAcc2** (Niveau de difficulté moyen)
  - **NumAcc3** (Niveau de difficulté moyen)
  - **NumAcc4** (Niveau de difficulté élevé)

- En Anova à 1 facteur :
  - **SiRstv** (Niveau de difficulté facile)
  - **SmLs01** (Niveau de difficulté facile)
  - **AtmWtAg** (Niveau de difficulté moyen)
  - **SmLs06** (Niveau de difficulté moyen)
  - **SmLs07** (Niveau de difficulté élevé)
  - **SmLs08** (Niveau de difficulté élevé)
  - **SmLs09** (Niveau de difficulté élevé)

- En régression linéaire simple et multiple
  - **Norris**
  - **Longley**

Les résultats (ci-après) sont présentés sous la forme de captures d'écran des résultats disponibles sur le site du NIST et ceux affichés par StatEL sur le tableur Excel.

STATISTIQUES UNIVARIEES

- Jeu **PiDigits** (Niveau de difficulté facile)

Résultats certifiés :

Certified Values

Dataset Name: PiDigits

Procedure: Univariate Summary Statistics
Certification Method & Definitions

Data: 1 Response Variable (y)
5000 Observations
Lower Level of Difficulty
Observed Data

Model: 3 Parameters (μ, σ, ρ_1)

$$y_i = \mu + \epsilon_i$$

Parameter	Certified Estimate
μ	4.53480000000000
σ	2.86733906028871
ρ_1	-0.00355099287237972

Résultats StatEL :

Résultats	
Nb sujets	5000
Minimum	0
Maximum	9
Etendue	9
Moyenne	4,5348
CV (Coefficie	0,632
SEM (Erreur	0,0406
IC 95% (Inter	4,455 < m < 4,614
Ecart-Type	2,86733906
Variance	8,222
Médiane	5
Quartile 25%	2
Quartile 75%	7
Inter Quartile	5
Asymétrie (C	0,122
Aplatisseme	-1,453

- Jeu **Lottery** (Niveau de difficulté facile)

Résultats certifiés :

Certified Values

Dataset Name: Lottery

Procedure: Univariate Summary Statistics
Certification Method & Definitions

Data: 1 Response Variable (y)
218 Observations
Lower Level of Difficulty
Observed Data

Model: 3 Parameters (μ, σ, ρ_1)

$$y_i = \mu + \epsilon_i$$

Parameter	Certified Estimate
μ	518.958715596330
σ	291.699727470969
ρ_1	-0.120948622967393

Résultats StatEL :

Résultats	
Nb sujets	218
Minimum	4
Maximum	999
Etendue	995
Moyenne	518,958716
CV (Coefficie	0,562
SEM (Erreur	19,76
IC 95% (Inter	480,2 < m < 557,7
Ecart-Type	291,699727
Variance	85089
Médiane	522,5
Quartile 25%	272,8
Quartile 75%	779,3
Inter Quartile	506,5
Asymétrie (C	-0,0891
Aplatisseme	-1,257

- Jeu **NumAcc2** (Niveau de difficulté moyen)

Résultats certifiés :

Certified Values

Dataset Name: NumAcc2

Procedure: Univariate Summary Statistics
Certification Method & Definitions

Data: 1 Response Variable (y)
1001 Observations
Average Level of Difficulty
Generated Data

Model: 3 Parameters (μ, σ, ρ_1)

$$y_i = \mu + \epsilon_i$$

Parameter	Certified Estimate
μ	1.2 (exact)
σ	0.1 (exact)
ρ_1	-0.999 (exact)

Résultats StatEL :

Résultats	
Nb sujets	1001
Minimum	1,1
Maximum	1,3
Etendue	0,2
Moyenne	1,200
CV (Coefficie	0,0833
SEM (Erreur	3,16E-03
IC 95% (Inter	1,194 < m < 1,206
Ecart-Type	0,1
Variance	0,0100
Médiane	1,2
Quartile 25%	1,1
Quartile 75%	1,3
Inter Quartile	0,2
Asymétrie (C	2,96E-13
Aplatisseme	-2,242

- Jeu **NumAcc3** (Niveau de difficulté moyen)

Résultats certifiés :

Certified Values

Dataset Name: NumAcc3

Procedure: Univariate Summary Statistics
Certification Method & Definitions

Data: 1 Response Variable (y)
1001 Observations
Average Level of Difficulty
Generated Data

Model: 3 Parameters (μ, σ, ρ_1)

$$y_i = \mu + \epsilon_i$$

Parameter	Certified Estimate	
μ	1000000.2	(exact)
σ	0.1	(exact)
ρ_1	-0.999	(exact)

Résultats StatEL :

Résultats		
Nb sujets	1001	
Minimum	1000000,1	
Maximum	1000000,3	
Etendue	0,200	
Moyenne	1000000,2	
CV (Coefficie	1,00E-07	
SEM (Erreur	3,16E-03	
IC 95% (Inter	1000000,19 < m < 1000000,21	
Ecart-Type	0,100	
Variance	0,0100	
Médiane	1000000,2	
Quartile 25%	1000000,1	
Quartile 75%	1000000,3	
Inter Quartile	0,2	
Asymétrie (C	1,67E-08	
Aplatisseme	-2,242	

- Jeu **NumAcc4** (Niveau de difficulté élevé)

Résultats certifiés :

Certified Values

Dataset Name: NumAcc4

Procedure: Univariate Summary Statistics
Certification Method & Definitions

Data: 1 Response Variable (y)
1001 Observations
Higher Level of Difficulty
Generated Data

Model: 3 Parameters (μ, σ, ρ_1)

$$y_i = \mu + \epsilon_i$$

Parameter	Certified Estimate	
μ	10000000.2	(exact)
σ	0.1	(exact)
ρ_1	-0.999	(exact)

Résultats StatEL :

Résultats			
Nb sujets	1001		
Minimum	10000000,1		
Maximum	10000000,3		
Etendue	0,200		
Moyenne	10000000,2		
CV (Coefficie	1,00E-08		
SEM (Erreur	3,16E-03		
IC 95% (Inter	10000000,19 < m < 10000000,21		
Ecart-Type	0,100		
Variance	0,0100		
Médiane	10000000,2		
Quartile 25%	10000000,1		
Quartile 75%	10000000,3		
Inter Quartile	0,2		
Asymétrie (C	-2,55E-06		
Aplatisseme	-2,242		

ANOVA 1 FACTEUR

- Jeu **SiRstv** (Niveau de difficulté facile)

Résultats certifiés :

Certified Values

Dataset Name: SiRstv

Procedure: Analysis of Variance
Certification Method & Definitions

Data: 1 Factor
5 Treatments
5 Replicates/Cell
25 Observations
3 Constant Leading Digits
Lower Level of Difficulty
Observed Data

Model: 6 Parameters ($\mu, \tau_1, \ldots, \tau_5$)
$y_{ij} = \mu + \tau_i + \epsilon_{ij}$

Source of Variation	Certified Degrees of Freedom	Certified Sums of Squares	Certified Mean Squares	Certified F Statistic
Between Instrument	4	5.11462616000000E-02	1.27865654000000E-02	1.18046237440255E+00
Within Instrument	20	2.16636560000000E-01	1.08318280000000E-02	
Certified R-Squared			1.90999039051129E-01	
Certified Residual Standard Deviation			1.04076068334656E-01	

Résultats StatEL :

Résultats de l'ANOVA	
Var. Factorielle :	0,0127866
Var. Résiduelle :	0,0108318
F :	1,18
F lim :	2,866
p <	0,35

- Jeu **SmLs01** (Niveau de difficulté facile)

Résultats certifiés :

Certified Values

Dataset Name: SmLs01

Procedure: Analysis of Variance
Certification Method & Definitions

Data: 1 Factor
9 Treatments
21 Replicates/Cell
189 Observations
1 Constant Leading Digit
Lower Level of Difficulty
Generated Data

Model: 10 Parameters $(\mu, \tau_1, \cdots, \tau_9)$
$y_{ij} = \mu + \tau_i + \epsilon_{ij}$

Source of Variation	Certified Degrees of Freedom	Certified Sums of Squares	Certified Mean Squares	Certified F Statistic
Between Treatment	8	1.68000000000000E+00	2.10000000000000E-01	2.10000000000000E+01
Within Treatment	180	1.80000000000000E+00	1.00000000000000E-02	
	Certified R-Squared		4.82758620689655E-01	
	Certified Residual Standard Deviation		1.00000000000000E-01	

Résultats StatEL :

Résultats de l'ANOVA	
Var. Factorielle	0,21
Var. Résiduelle	0,01
F :	21
F lim :	1,99
p <	0,00001

- **AtmWtAg** (Niveau de difficulté moyen)

Résultats certifiés :

Certified Values

Dataset Name: AtmWtAg

Procedure: Analysis of Variance
Certification Method & Definitions

Data: 1 Factor
2 Treatments
24 Replicates/Cell
48 Observations
7 Constant Leading Digits
Average Level of Difficulty
Observed Data

Model: 3 Parameters (μ, τ_1, τ_2)
$y_{ij} = \mu + \tau_i + \epsilon_{ij}$

Source of Variation	Certified Degrees of Freedom	Certified Sums of Squares	Certified Mean Squares	Certified F Statistic
Between Instrument	1	3.63834187500000E-09	3.63834187500000E-09	1.59467335677930E+01
Within Instrument	46	1.04951729166667E-08	2.28155932971014E-10	
Certified R-Squared			2.57426544538321E-01	
Certified Residual Standard Deviation			1.51048314446410E-05	

Résultats StatEL :

Résultats de l'ANOVA	
Var. Factorielle	3,63834189E-09
Var. Résiduelle	2,28155933E-10
F :	15,95
F lim :	4,052
p <	0,00023

- Jeu **SmLs06** (Niveau de difficulté moyen)

Résultats certifiés :

Certified Values

Dataset Name: SmLs06

Procedure: Analysis of Variance
Certification Method & Definitions

Data: 1 Factor
9 Treatments
2001 Replicates/Cell
18009 Observations
7 Constant Leading Digits
Average Level of Difficulty
Generated Data

Model: 10 Parameters ($\mu, \tau_1, \ldots, \tau_9$)
$y_{ij} = \mu + \tau_i + \epsilon_{ij}$

Source of Variation	Certified Degrees of Freedom	Certified Sums of Squares	Certified Mean Squares	Certified F Statistic
Between Treatment	8	1.60080000000000E+02	2.00100000000000E+01	2.00100000000000E+03
Within Treatment	18000	1.80000000000000E+02	1.00000000000000E-02	
	Certified R-Squared		4.70712773465067E-01	
	Certified Residual Standard Deviation		1.00000000000000E-01	

Résultats StatEL :

Résultats de l'ANOVA	
Var. Factorielle :	20,01
Var. Résiduelle :	0,01
F :	2001
F lim :	1,939
p <	0,00001

- Jeu **SmLs07** (Niveau de difficulté élevé)

Résultats certifiés :

Certified Values

Dataset Name: SmLs07

Procedure: Analysis of Variance
Certification Method & Definitions

Data: 1 Factor
9 Treatments
21 Replicates/Cell
189 Observations
13 Constant Leading Digits
Higher Level of Difficulty
Generated Data

Model: 10 Parameters $(\mu, \tau_1, \cdots, \tau_9)$
$y_{ij} = \mu + \tau_i + \epsilon_{ij}$

Source of Variation	Certified Degrees of Freedom	Certified Sums of Squares	Certified Mean Squares	Certified F Statistic
Between Treatment	8	1.68000000000000E+00	2.10000000000000E-01	2.10000000000000E+01
Within Treatment	180	1.80000000000000E+00	1.00000000000000E-02	
	Certified R-Squared		4.82758620689655E-01	
	Certified Residual Standard Deviation		1.00000000000000E-01	

Résultats StatEL :

Résultats de l'ANOVA	
Var. Factorielle :	0,2098975126
Var. Résiduelle :	0,0100005469
F :	20,99
F lim :	1,99
p <	0,00001

Remarque :
Il existe une légère différence entre les 2 résultats sur ce jeu de données dont les valeurs sont des nombres à 13 chiffres.

- Jeu **SmLs08** (Niveau de difficulté élevé)

Résultats certifiés :

Certified Values

Dataset Name: SmLs08

Procedure: Analysis of Variance
Certification Method & Definitions

Data: 1 Factor
9 Treatments
201 Replicates/Cell
1809 Observations
13 Constant Leading Digits
Higher Level of Difficulty
Generated Data

Model: 10 Parameters $(\mu, \tau_1, \cdots, \tau_9)$
$y_{ij} = \mu + \tau_i + \epsilon_{ij}$

Source of Variation	Certified Degrees of Freedom	Certified Sums of Squares	Certified Mean Squares	Certified F Statistic
Between Treatment	8	1.60800000000000E+01	2.01000000000000E+00	2.01000000000000E+02
Within Treatment	1800	1.80000000000000E+01	1.00000000000000E-02	
		Certified R-Squared	4.71830985915493E-01	
		Certified Residual Standard Deviation	1.00000000000000E-01	

Résultats StatEL :

Résultats de l'ANOVA	
Var. Factorielle :	2,009018674
Var. Résiduelle :	0,010000545
F :	200,9
F lim :	1,944
p <	0,00001

Remarque :
Il existe une légère différence entre les 2 résultats sur ce jeu de données dont les valeurs sont des nombres à 13 chiffres.

- Jeu **SmLs09** (Niveau de difficulté élevé)

Résultats certifiés :

Certified Values

Dataset Name: SmLs09

Procedure: Analysis of Variance
Certification Method & Definitions

Data: 1 Factor
9 Treatments
2001 Replicates/Cell
18009 Observations
13 Constant Leading Digits
Higher Level of Difficulty
Generated Data

Model: 10 Parameters $(\mu, \tau_1, \cdots, \tau_9)$
$y_{ij} = \mu + \tau_i + \epsilon_{ij}$

Source of Variation	Certified Degrees of Freedom	Certified Sums of Squares	Certified Mean Squares	Certified F Statistic
Between Treatment	8	1.60080000000000E+02	2.00100000000000E+01	2.00100000000000E+03
Within Treatment	18000	1.80000000000000E+02	1.00000000000000E-02	
	Certified R-Squared		4.70712773465067E-01	
	Certified Residual Standard Deviation		1.00000000000000E-01	

Résultats StatEL :

Résultats de l'ANOVA	
Var. Factorielle :	20,02466590
Var. Résiduelle :	0,01000055
F :	2002
F lim :	1,939
p <	0,00001

Remarque :
Il existe une légère différence entre les 2 résultats sur ce jeu de données dont les valeurs sont des nombres à 13 chiffres.

REGRESSION LINEAIRE SIMPLE

- Jeu **Norris**

Résultats certifiés :

Certified Values

Dataset
Name: Norris

Procedure: Linear Least Squares Regression
Certification Method & Definitions

Data: 1 Response Variable (*y*)
1 Predictor Variable (*x*)
36 Observations
Lower Level of Difficulty
Observed Data

Model: $y = \beta_0 + \beta_1 x + \epsilon$

Certified Regression Statistics

Parameter	Estimate	Standard Deviation of Estimate
β_0	-0.262323073774029	0.232818234301152
β_1	1.00211681802045	0.429796848199937E-03

Residual Standard Deviation	0.884796396144373
R-Squared	0.999993745883712

Résultats StatEL :

Modèle de régression liéaire :	
y = 1,00211681802045*x - 0,262323073773882	
R² (Coefficient de Détermination) :	**99,999375%**
Ce coefficient exprime le pourcentage de variatio	
R²aj (Coef. de Détermination ajusté) :	**99,999356%**
Ce coefficient permet de comparer différents mod	

REGRESSION LINEAIRE MULTIPLE

- Jeu **Longley**

Résultats certifiés :

Certified Values

Dataset Name: Longley

Procedure: Linear Least Squares Regression
Certification Method & Definitions

Data: 1 Response Variable (y)
6 Predictor Variable ($x_1, \ldots, x_6$)
16 Observations
Higher Level of Difficulty
Observed Data

Model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \epsilon$

Certified Regression Statistics

Parameter	Estimate	Standard Deviation of Estimate
β_0	-3482258.63459582	890420.383607373
β_1	15.0618722713733	84.9149257747669
β_2	-0.358191792925910E-01	0.334910077722432E-01
β_3	-2.02022980381683	0.488399681651699
β_4	-1.03322686717359	0.214274163161675
β_5	-0.511041056535807E-01	0.226073200069370
β_6	1829.15146461355	455.478499142212

Residual Standard Deviation	304.854073561965
R-Squared	0.995479004577296

Résultats StatEL :

Modèle de régression linéaire multiple :

y = 15,06 * x1 - 0,03582 * x2 - 2,02 * x3 - 1,033 * x4 - 0,0511 * x5 + 1829 * x6 - 3482259

Coefficient de corrélation multiple (exprime l'intensité de la liaison linéaire entre la variable

r = 0,99774746

Significativité de r par rapport à 0 : p < 0,00001

Coefficient de détermination multiple (exprime le pourcentage de variation expliquée par le

R^2 = 99,55%

Coefficient de détermination ajusté (permet de comparer différents modèles ne disposant pas

R^2aj = 99,25%

Significativité des estimateurs du modèle :

Hypothèses de travail :

H0 = l'estimateur étudié n'est pas significativement différent de 0.

H1 = l'estimateur étudié est significativement différent de 0, au risque d'erreur < p.

Variable	*Estimateur*	*Ecart-type*	*t*	*p*	*Intervalle de Confiance (95%)*
Constante	-3482259	890420	-3,911	0,0036	[-5496529 ; -1467988]
x1	15,06	84,91	0,1774	0,86	[-177 ; 207,2]
x2	-0,03582	0,03349	-1,07	0,31	[-0,1116 ; 0,03994]
x3	-2,02	0,4884	-4,136	0,0025	[-3,125 ; -0,9154]
x4	-1,033	0,2143	-4,822	0,00094	[-1,518 ; -0,5485]
x5	-0,0511	0,2261	-0,2261	0,83	[-0,5625 ; 0,4603]
x6	1829	455,5	4,016	0,003	[798,8 ; 2860]

CONCLUSIONS

Hormis dans le cas particulier des tests relatifs aux données à 13 chiffres où les calculs d'Anova divergent très légèrement, les résultats présentés par le logiciel StatEL sont en tous points identiques aux résultats certifiés par le NIST, sur les jeux de données utilisés.

En tout état de cause, si vous êtes attachés à une précision très importante des résultats lors d'études avec des données dont les valeurs dépassent la dizaine de chiffres, il est déconseillé d'utiliser le logiciel StatEL, sauf si vous êtes en mesure d'apporter un traitement particulier à vos données (cf. dernier paragraphe : Solution).

Les écarts constatés sont dus aux erreurs dites «cancellation error», «truncation error» et/ou «accumulation error». Elles sont liées à la façon dont les variables sont codées et aux procédures de calcul qui laissent, au fur et à mesure des itérations, de nombreuses décimales de côté.

Ces erreurs sont sans effet sur les résultats avec des données « classiques ». Elles sont en revanche un peu plus notables avec des nombres de grande valeur (tels que les nombres à 13 chiffres utilisés dans les exemples précédents), et s'accroissent avec la taille de l'échantillon. Ils apparaissent également lorsque les données analysées n'ont qu'une faible variation relative.

Solution :
Une façon de résoudre manuellement le problème, et donc d'utiliser StatEL même avec ce type de données, consiste à soustraire de chaque valeur la partie constante sur toutes les données. Ainsi, dans nos exemples, il suffirait de soustraire 1000000000000 de 1000000000000,1 afin de ne travailler qu'avec 0,1 et de faire de même pour toutes les mesures étudiées.