ENSAE-ENSAI Formation continue [Cepe]

Certificat de
chargé d'études statistiques

Révision du niveau 1 – Exercices d'application

Martin Chevalier – *Version corrigée*

Ces exercices d'application reposent sur des exploitations de l'enquête PISA (Program for International Student Assessment) 2012. Réalisée tous les trois ans par l'OCDE dans une soixantaine de pays, cette enquête vise à mesurer les acquis des élèves de 15 ans.

En plus des scores aux tests standardisés de mathématiques, compréhension de l'écrit et sciences, cette enquête comporte de très nombreuses informations sur l'origine sociale des élèves, leurs conditions d'enseignement ainsi que leur rapport aux enseignants et à l'école.

Du point de vue de la formation, cette enquête présente l'avantage de comporter une très large variété de variables qualitatives et quantitatives. Elle se prête à tous les outils et méthodes au programme de la session de révision des 28 et 29 novembre 2016.

Les fichiers de l'enquête PISA 2012 sont librement téléchargeables sur le site de l'OCDE[1]*. Le fichier « élèves » réduit*[2] *ayant servi à la production des sorties statistiques utilisées dans ces exercices pratiques est fourni aux stagiaires.*

1. `http://www.oecd.org/pisa/pisaproducts/pisa2012database-downloadabledata.htm`
2. L'échantillon est restreint aux données collectées en France, rééchantillonnées à hauteur de 30 % pour limiter la taille du fichier final.

Analyse univariée : statistiques et représentations

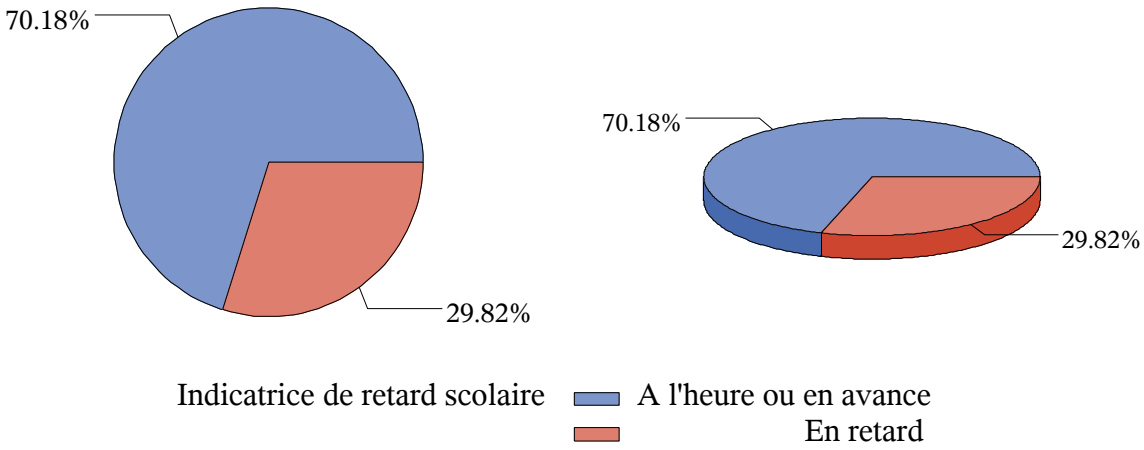
Exercice 1.1 La variable `ST01Q01` correspond à la classe dans laquelle se trouve l'élève au moment de l'enquête (la 10ème classe correspond à la seconde en France).

Classe de l'élève au moment de l'enquête				
ST01Q01	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
8	22	1.61	22	1.61
9	386	28.22	408	29.82
10	910	66.52	1318	96.35
11	49	3.58	1367	99.93
12	1	0.07	1368	100.00

a. Comment désigne-t-on les tableaux du type de celui ci-dessus et comment a-t-il été produit ?

b. Interprétez les résultats. Proposez un recodage pour étudier le retard scolaire.

c. Comment désigne-t-on les représentations graphiques ci-dessous ? Laquelle vous semble la plus appropriée ? Quelle procédure utiliseriez-vous pour les construire ?



Correction

a. Il s'agit d'un tri à plat produit par la **PROC FREQ** :

```
PROC FREQ DATA = d.pisa12;
    TABLES ST01Q01;
RUN;
```

b. Plus des deux tiers des élèves de 15 ans interrogés (66,5 %) sont « à l'heure » scolairement, c'est-à-dire en classe de seconde. 3,6 % sont en avance, 29,8 % sont en retard. Étant donnée la distribution de cette variable, on est amené à la recoder en deux modalités : en retard (`8` ou `9`) ou pas (`10`, `11`, `12`).

c. On a dans les deux cas affaire à un diagramme circulaire que l'on produit à l'aide de la **PROC GCHART** de SAS. Le graphique de droite présente une vue en perspective : à ce titre, les surfaces sont déformées. On privilégie donc le graphique de gauche.

Code SAS utilisé pour produire l'exercice

```
PROC FREQ DATA = d.pisa12;
    TABLES ST01Q01;
RUN;
PROC FORMAT;
    VALUE retard
        1 = "En retard"
        0 = "A l'heure ou en avance"
    ;
RUN;
PROC GCHART DATA = d.pisa12;
    PIE retard / PERCENT = ARROW LEGEND DISCRETE;
    PIE3D retard / PERCENT = ARROW LEGEND DISCRETE;
    FORMAT retard retard.;
RUN;
```

Exercice 1.2 La variable PV1MATH correspond au score synthétique aux évaluations de mathématiques. Le tableau suivant en synthétise la distribution :

Analysis Variable : PV1MATH Score aux évaluations de mathématiques					
N	Mean	Median	Variance	Minimum	Maximum
1368	498.1789377	498.6836000	9232.63	230.8070000	781.4379000

a. Quelle procédure a été utilisée pour produire ce tableau ?

b. Interprétez la valeur de la moyenne. En analysant les autres statistiques présentées, que pensez-vous de l'influence des valeurs extrêmes ?

c. Calculez l'écart-type et le coefficient de variation de la variable PV1MATH.

Correction

a. Il s'agit de la **PROC MEANS** :

```
PROC MEANS DATA = d.pisa12 N MEAN MEDIAN VAR MIN MAX;
    VAR PV1MATH;
RUN;
```

b. Le score moyen en mathématiques des élèves de l'échantillon est d'environ 498,2. On remarque que les valeurs maximales et minimales sont du même ordre de grandeur que la moyenne. On est ainsi amené à penser que les valeurs extrêmes sont assez peu susceptibles d'affecter la moyenne.

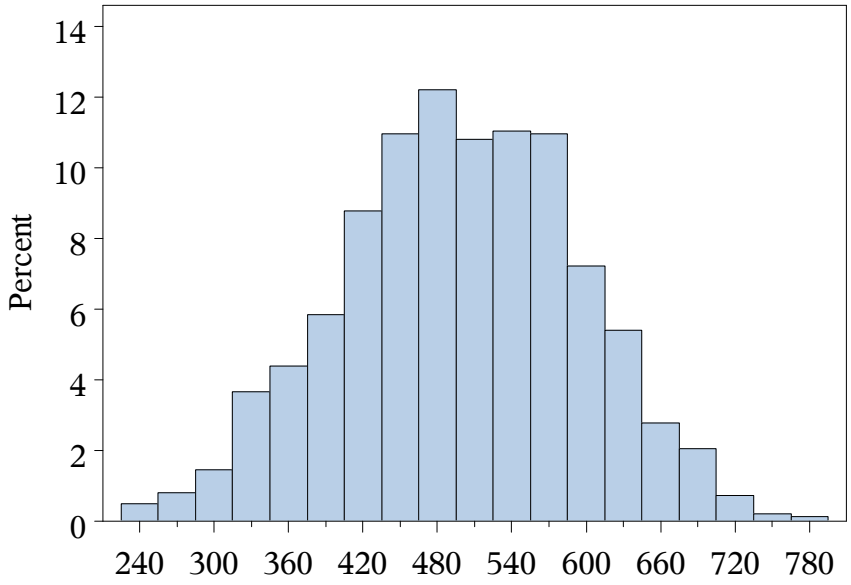
c. L'écart-type est directement obtenu à partir de la variance :

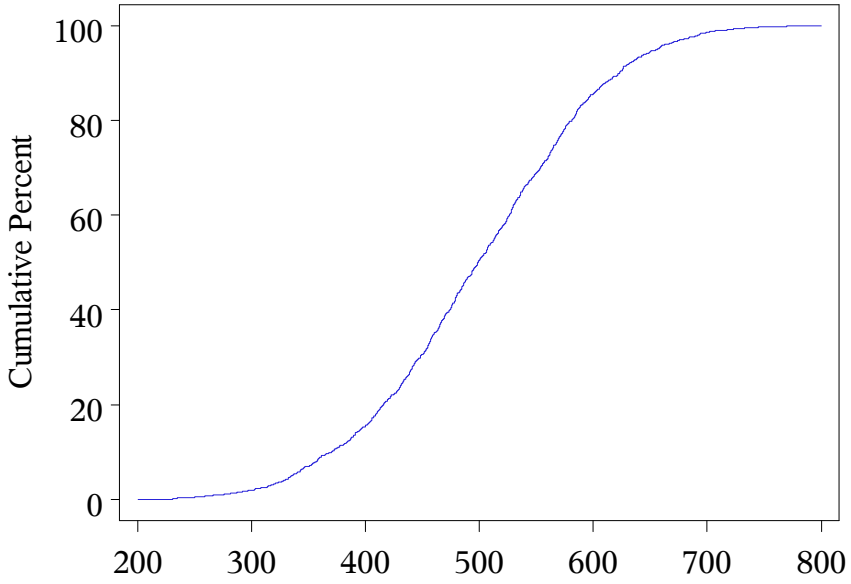
$$\hat{\sigma}_{\text{PV1MATH}} = \sqrt{V(\text{PV1MATH})} = \sqrt{9232,63} = 96,09$$

Le coefficient de variation est obtenu à partir de l'écart-type et de la moyenne :

$$CV(\text{PV1MATH}) = \frac{\hat{\sigma}_{\text{PV1MATH}}}{\overline{\text{PV1MATH}}} = \frac{96,09}{498,18} = 19,29\ \%$$

Exercice 1.3 Les deux graphiques suivants représentent la distribution de la variable `PV1MATH`.





a. Comment désigne-t-on ces deux types de graphiques ? En quoi leur analyse confirme-t-elle les résultats de l'exercice précédent quant à l'influence des valeurs extrêmes ?

b. Utilisez ces graphiques pour déterminer (approximativement) la valeur des quartiles de la variable PV1MATH ainsi que celle des premier et neuvième déciles.

c. Quelle procédure auriez-vous pu utiliser pour obtenir directement les quantiles de la variable PV1MATH ?

Correction

a. Ces deux graphiques sont l'histogramme et la fonction de répartition de PV1MATH. L'histogramme indique la très grande régularité de la distribution de PV1MATH, ce qui confirme la faible influence des valeurs extrêmes sur la moyenne.

b. Il est possible de lire directement sur la fonction de répartition la valeur (approximative) des quantiles de PV1MATH : 370 (D1), 450 (Q1), 550 (Q3), 620 (D9).

c. La **PROC MEANS** et la **PROC UNIVARIATE** peuvent être utilisées pour calculer les quantiles d'une distribution :

```
PROC MEANS DATA = d.pisa12 N MEAN P10 Q1 MEDIAN Q3 P90;
    VAR PV1MATH;
RUN;

PROC UNIVARIATE DATA = d.pisa12;
    VAR PV1MATH;
RUN;
```

Code SAS utilisé pour produire l'exercice

```
PROC UNIVARIATE DATA = d.pisa12 NOPRINT;
    VAR PV1MATH;
    HISTOGRAM;
    CDFPLOT;
RUN;
```

Analyse univariée : inférence

Exercice 2.1 En annexe figurent les tables des quantiles de la loi normale centrée réduite, de la loi du χ^2 et de la loi de Student.

a. Retrouvez la valeur du quantile à 97,5 % d'une loi normale centrée réduite. Comparez-la à la valeur du quantile à 2,5 %. Comment expliquez-vous leur relation ?

b. Retrouvez la valeur du quantile à 95 % d'une loi du χ^2 à 5 degrés de liberté. Comparez-la à la valeur du quantile à 5 %. Expliquez la différence avec la loi normale.

c. Retrouvez la valeur du quantile à 99 % d'une loi de Student à 10 degrés de liberté. Comparez-la aux valeurs des quantiles à 99 % respectivement d'une loi de Student à 100 degrés de libertés et d'une loi normale centrée réduite. Que traduit ce résultat ?

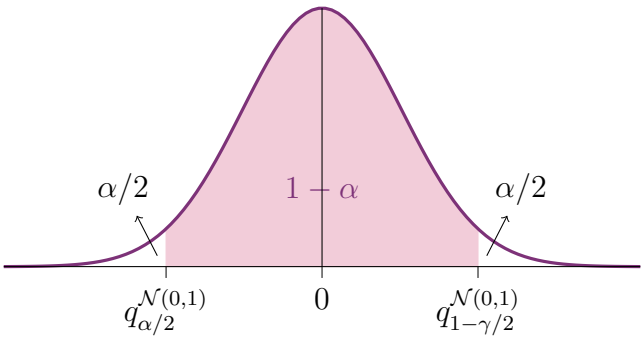
d. Êtes-vous en mesure de déterminer à partir de ces tables la valeur du quantile à 97,5 % d'une loi normale d'espérance 4 et d'écart-type 1 ? D'espérance 0 et d'écart-type 4 ?

Correction

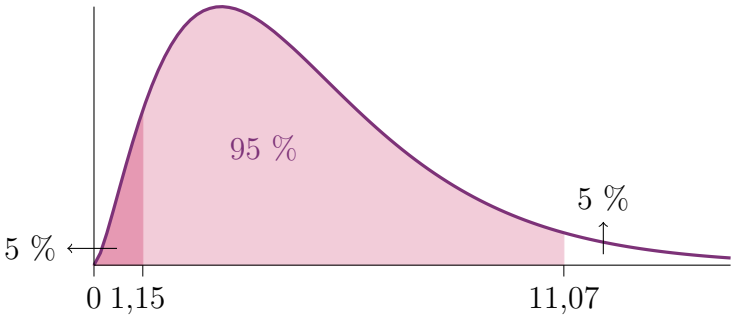
a. On peut lire dans la table 1 que le quantile à 97,5 % d'une loi normale centrée réduite est 1,96. Le quantile à 2,5 % d'une loi normale centrée réduite est quant à lui -1,96. De façon générale

$$\forall \gamma \in [0;1] \qquad q_{\gamma}^{\mathcal{N}(\mu,\sigma^2)} = -q_{1-\gamma}^{\mathcal{N}(\mu,\sigma^2)}$$

Ceci est dû à la parité de la densité de la loi normale :



b. On peut lire dans la table 2 que le quantile à 95 % d'une loi du χ^2 à 5 degrés de liberté est 11,07. Le quantile à 2,5 % d'une loi du χ^2 à 5 degrés de liberté est quant à lui 1,15. Contrairement à la loi normale, la densité d'une loi du χ^2 n'est pas paire : les quantiles n'ont aucune raison d'entretenir les mêmes relations que pour la loi normale ou la loi de Student.



c. On peut lire dans la table 3 que le quantile à 99 % d'une loi de Student à 10 degrés de liberté est 2,76. Les quantiles de même niveau d'une loi de Student à 100 degrés de liberté et d'une loi normale centrée réduite sont respectivement 2,36 et 2,33. Ces valeurs illustrent le rapprochement entre la loi de Student et la loi normale centrée réduite quand le nombre de degrés de liberté tend vers l'infini.

d. Les quantiles de toutes les lois normales peuvent être déterminés à partir de ceux de la loi normale centrée réduite. En effet, si

$$Z \hookrightarrow \mathcal{N}(\mu, \sigma^2) \quad \text{alors} \quad \frac{Z-\mu}{\sigma} \hookrightarrow \mathcal{N}(0,1)$$

Par définition de $q_\gamma^{\mathcal{N}(0,1)}$

$$\mathbb{P}\left(\frac{Z-\mu}{\sigma} \leq q_\gamma^{\mathcal{N}(0,1)}\right) = \gamma$$

donc

$$\mathbb{P}\left(Z \leq \mu + \sigma \times q_\gamma^{\mathcal{N}(0,1)}\right) = \gamma$$

Autrement dit $q_\gamma^{\mathcal{N}(\mu,\sigma^2)} = \mu + \sigma \times q_\gamma^{\mathcal{N}(0,1)}$. On détermine ainsi :

— le quantile à 97,5 d'une loi normale de moyenne 4 et d'écart-type 1 vaut $4 + 1 \times 1,96 = 5,96$;

— le quantile à 97,5 d'une loi normale de moyenne 0 et d'écart-type 4 vaut $0 + 4 \times 1,96 = 7,84$.

Exercice 2.2 En plus du score synthétique aux évaluations de mathématiques `PV1MATH`, l'enquête comporte les scores synthétiques aux évaluations de compréhension de l'écrit et de sciences, respectivement `PV1READ` et `PV1SCIE`.

a. Utilisez les informations de l'exercice 1.2 pour calculer l'erreur standard et l'intervalle de confiance à 95 % de la moyenne du score en mathématiques.

b. Quelles options de la **`PROC MEANS`** ont été utilisées pour obtenir le tableau ci-dessous ?

Variable	Label	N	Mean	Std Dev	Std Error	Lower 95% CL for Mean	Upper 95% CL for Mean
PV1READ	Score aux évaluations de compréhension de l'écrit	1368	509.9585724	108.6969627	2.9388293	504.1934684	515.7236765
PV1SCIE	Score aux évaluations de sciences	1368	503.4514466	98.0487839	2.6509355	498.2511041	508.6517892

Sachant que par construction la moyenne des scores est 500 au niveau international, que pouvez-vous conclure quant au niveau des élèves scolarisés en France en mathématiques, compréhension de l'écrit et sciences ?

c. Calculez l'intervalle de confiance à 99 % de la moyenne du score en compréhension de l'écrit. Cela conduit-il à nuancer la conclusion de la question précédente ?

Correction

a. L'application du théorème central limite conduit à définir l'intervalle de confiance à 95 % de la moyenne de la variable X comme

$$IC_{95\ \%}(\bar{X}) = \left[\bar{X} - 1,96 \times \frac{\hat{\sigma}_X}{\sqrt{n}}; \bar{X} + 1,96 \times \frac{\hat{\sigma}_X}{\sqrt{n}}\right]$$

Application numérique : d'après le tableau de l'exercice 1.2, $\bar{PV1MATH} = 498,2$, $\hat{\sigma}_{PV1MATH} = \sqrt{9232,6} = 96,1$ et $n = 1\ 368$.

Ainsi l'erreur standard de la moyenne de `PV1MATH` vaut $se_{PV1MATH} = \dfrac{96,1}{\sqrt{1\ 368}} = 2,6$ et son intervalle de confiance à 95 % est :

$$IC_{95\ \%}(\bar{PV1MATH}) = [498,2 - 1,96 \times 2,6; 498,2 + 1,96 \times 2,6] = [493,1; 503,3]$$

b. Il faut utiliser les options `STDERR`, `CLM` et `ALPHA = 0.05` pour obtenir ce tableau. Pour ce qui concerne l'interprétation :

— L'intervalle de confiance à 95 % de la moyenne du score en mathématiques a été calculé à la question précédente et vaut [493,1 ;503,3]. Autrement dit, même si le score moyen en mathématiques calculé sur l'échantillon vaut 498,7 et est donc inférieur au niveau de référence de 500, on ne peut pas raisonnablement exclure que dans l'ensemble de la population il soit en fait supérieur à 500. En termes qualitatifs, on retient que le score moyen en mathématiques des élèves scolarisés en France n'est **pas significativement différent du score de référence au niveau international**.

— L'intervalle de confiance à 95 % de la moyenne du score en compréhension de l'écrit est calculé par la **`PROC MEANS`** et vaut [504,2 ;515,7]. La valeur de référence 500 n'appartient donc pas à l'intervalle de confiance à 95 % de la moyenne : on peut donc raisonnablement estimer que la moyenne du score en compréhension de l'écrit est bien supérieure à 500 dans l'ensemble de la population (on a seulement 5 % de chances de se tromper en l'affirmant). En termes qualitatifs, on retient que le score moyen en compréhension de l'écrit des élèves scolarisés en France est **significativement supérieur au score de référence au niveau international**.

— L'intervalle de confiance à 95 % de la moyenne du score en sciences est calculé par la **`PROC MEANS`** et vaut [498,3 ;508,7]. Autrement dit, même si le score moyen en sciences calculé sur l'échantillon vaut 503,5 et est donc supérieur au niveau de référence de 500, on ne peut pas raisonnablement exclure que dans l'ensemble de la population il soit en fait inférieur à 500. En termes qualitatifs, on retient que le score moyen en sciences des élèves scolarisés en France n'est **pas significativement différent du score de référence au niveau international**.

c. Pour calculer l'intervalle de confiance à 99 % de la moyenne du score en compréhension de l'écrit, on peut repartir de l'erreur standard. En effet, on sait que

$$IC_{99\ \%}(\bar{X}) = \left[\bar{X} - q^{\mathcal{T}_{n-1}}_{99,5\ \%} \times se_X ; \bar{X} + q^{\mathcal{T}_{n-1}}_{99,5\ \%} \times se_X\right]$$

avec ici $se_X = se_{PV1READ} = 2,9$ (colonne **Std. Error** de la sortie SAS). Par ailleurs on lit en annexe que le quantile à 99,5 % d'une loi de Student à $1\ 368 - 1 = 1\ 367$ degrés de liberté est 2,58. Ainsi :

$$IC_{99\ \%}(\bar{PV1READ}) = [510,0 - 2,58 \times 2,9 ; 510,0 + 2,58 \times 2,9] = [502,5 ; 517,5]$$

L'intervalle de confiance à 99 % de la moyenne du score en compréhension de l'écrit est donc [502,5 ;517,5]. La valeur de référence 500 n'appartient donc pas à l'intervalle de confiance à 95 % de la moyenne : on peut donc **très** raisonnablement estimer que la moyenne du score en compréhension de l'écrit est bien supérieure à 500 dans l'ensemble de la population (on a **seulement 1 % de chances de se tromper** en l'affirmant). La conclusion qualitative de la question précédente n'est pas remise en question : quel que soit le seuil statistique choisi, on peut bien estimer que le score moyen des élèves scolarisés en France en compréhension de l'écrit est **significativement supérieur à la valeur de référence au niveau international** .

Code SAS utilisé pour produire l'exercice

```
PROC MEANS DATA = d.pisa12 N MEAN STD STDERR CLM ALPHA = 0.05;
    VAR PV1READ PV1SCIE;
RUN;
```

Analyse bivariée : statistiques et représentations

Exercice 3.1 La variable `ST01Q01` est recodée en deux modalités dans la variable `retard` qui vaut 1 si l'élève est « en retard » (s'il est dans une classe inférieure à la seconde au moment de l'enquête) et 0 sinon. Cette variable est croisée avec la variable `ST04Q01`, qui code le sexe des élèves.

Frequency Expected Cell Chi-Square Percent Row Pct Col Pct	Table of ST04Q01 by retard		
ST04Q01(Sexe (1 = Femme, 2 = Homme))	retard(Indicatrice de retard scolaire)		
	0	1	Total
Femme	528 501.05 1.4493 38.60 73.95 55.00	186 212.95 3.41 13.60 26.05 45.59	714 52.19
Homme	432 458.95 1.5822 31.58 66.06 45.00	222 195.05 3.7229 16.23 33.94 54.41	654 47.81
Total	960 70.18	408 29.82	1368 100.00

a. Quelles options de la **PROC FREQ** ont été utilisées pour obtenir ce tableau ? Quelles statistiques de l'exercice 1.1 y retrouvez-vous ? Comment les désigne-t-on dans le contexte du tri croisé ?

b. Interprétez le pourcentage de cellule, le pourcentage en ligne et le pourcentage en colonne relatifs de la case « Homme × `retard = 1` ».

c. Utilisez l'ensemble des informations du tableau pour identifier et justifier des sur- ou sous-représentations manifestes.

Correction

a. Les options de la **PROC FREQ** utilisées pour produire ce tableau sont `EXPECTED` et `CELLCHI2`. On retrouve ici en pied de colonne les effectifs et pourcentages des données qui figurent dans le tableau de l'exercice 1.1. Dans le contexte du tri croisé, on les qualifie d'effectifs et de pourcentages marginaux.

b. On interprète les informations de la case « Homme × `retard = 1` » :
   - 16,23 % des individus de l'échantillon sont des hommes en retard scolaire ;
   - 33,94 % des hommes de l'échantillon sont en retard scolaire ;
   - 54,41 % des individus en retard scolaire de l'échantillon sont des hommes.

c. 33,94 % des hommes de l'échantillon sont en retard scolaire, contre 29,82 % des individus de l'ensemble de l'échantillon. 222 individus sont des hommes en retard scolaires alors que, sous l'hypothèse d'indépendance entre les deux variables, il devraient

être 195. La contribution à la statistique du χ^2 de la case « Homme × `retard = 1` » est la plus importante du tableau (3,7229). Tous ces éléments conduisent à conclure à une surreprésentation sensible des individus en retard scolaire parmi les hommes. Son caractère statistiquement significatif peut être examiné à l'aide d'un test du χ^2 (*cf.* exercice 4.5).

Code SAS utilisé pour produire l'exercice

```
PROC FORMAT;
    VALUE ST04Q01_
        1 = "Femme"
        2 = "Homme"
    ;
RUN;
PROC FREQ DATA = d.pisa12;
    TABLES ST04Q01*retard / EXPECTED CELLCHI2;
    FORMAT ST04Q01 ST04Q01_.;
RUN;
```

Exercice 3.2 Le tableau suivant représente le coefficient de corrélation de Pearson calculés entre les trois scores pris deux-à-deux :

Pearson Correlation Coefficients, N = 1368 Prob > |r| under H0: Rho=0			
	PV1MATH	**PV1READ**	**PV1SCIE**
PV1MATH Score aux évaluations de mathématiques	1.00000	0.86722 <.0001	0.89773 <.0001
PV1READ Score aux évaluations de compréhension de l'écrit	0.86722 <.0001	1.00000	0.88809 <.0001
PV1SCIE Score aux évaluations de sciences	0.89773 <.0001	0.88809 <.0001	1.00000

a. Quelle procédure a été utilisée pour produire ces résultats ?

b. Interprétez la valeur des indicateurs. Que peut-on conclure ?

c. Quel indicateur alternatif aurait-on pu utiliser ? Quel est son intérêt et comment le calculer ?

Correction

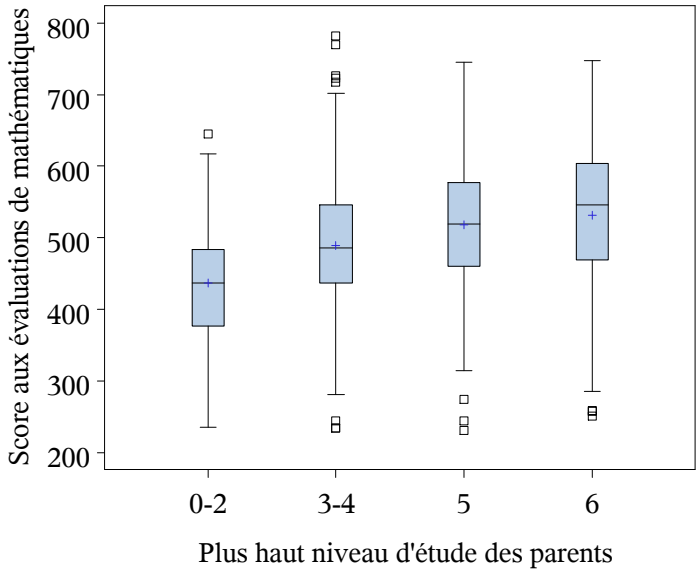
a. Il s'agit de la **PROC CORR** :

```
PROC CORR DATA = d.pisa12;
    VAR PV1MATH PV1READ PV1SCIE;
RUN;
```

b. Les trois coefficients de corrélation linéaire de Pearson présentés dans le tableau sont très élevés : score en mathématiques et en compréhension de l'écrit sont corrélés à hauteur de 0,86722, scores en mathématiques et en sciences à hauteur de 0,89773 et scores en compréhension de l'écrit et en sciences à hauteur de 0,88809. Sans même avoir rigoureusement testé la significativité statistique de ces coefficients (*cf.* exercice 4.6), ces valeurs particulièrement élevées permettent de conclure à une association forte entre les trois scores mesurés par l'enquête PISA 2012.

c. D'autres mesures d'association entre variables quantitatives existent, en particulier le coefficient de corrélation des rangs de Spearman. Par rapport au coefficient de corrélation de Pearson, il est moins sensible aux valeurs extrêmes. On le calcule en utilisant l'option `SPEARMAN` de la **`PROC CORR`**.

Exercice 3.3 On dispose également du plus haut niveau d'étude des parents que l'on regroupe en quatre modalités[3] : 0-2 Aucun, primaire ou collège ; 3-4 Fin d'études secondaires ou post-secondaire non-supérieur ; 5 Premier cycle d'études supérieures ; 6 Second cycle universitaire et au-delà. Le graphique suivant représente la relation entre score en mathématiques (variable `PV1MATH`) et plus haut diplôme des parents regroupé (variable `hdipl`).



a. Comment désigne-t-on ce type de graphique et avec quelle procédure le construiriez-vous ?

b. Explicitez la signification des éléments constituant une boîte et interprétez le graphique.

c. Quelle mesure d'association correspond à ce type de représentation ? Sa valeur est 0,092048 : qu'en pensez-vous ?

Correction

a. Ce graphique représente des « boîtes à moustaches » ou « boîtes de Tukey ». Il est possible de les produire avec une **`PROC BOXPLOT`** précédée d'un tri selon la variable qualitative (en abscisse) :

```
PROC SORT DATA = d.pisa12;
    BY hdipl;
RUN;
PROC BOXPLOT DATA = d.pisa12;
    PLOT PV1MATH * hdipl / BOXSTYLE = SCHEMATIC;
RUN;
```

3. La nomenclature originale est ISCED 1997 (`http://www.uis.unesco.org/Library/Documents/isced97-fr.pdf`).

b. Les traits horizontaux de la boîte de Tukey correspondent aux trois quartiles de la distribution de la variable quantitative pour une modalité donnée de la variable qualitative. La croix à l'intérieur de la boîte correspond à la moyenne de la variable. Les moustaches peuvent soit aller jusqu'aux valeurs extrêmes, soit avoir une longueur maximale de $1,5 \times (Q3 - Q1)$. Dans le second cas (qui correspond au graphique présenté), les valeurs extrêmes au-delà des moustaches sont singularisées par des points (ici des carrés) et parfois par leur identifiant.

c. La mesure d'association correspondant au croisement d'une variable quantitative avec une variable qualitative est le rapport de corrélation. Une valeur de 0,092048 est relativement faible mais peut parfois suffire à conclure que les deux variables sont significativement liées. C'est tout l'objet de l'analyse de la variance que de mettre en œuvre des tests d'association entre variables dans ce type de configuration (*cf.* module 3, 30 janvier-1er février).

Code SAS utilisé pour produire l'exercice

```
PROC SORT DATA = d.pisa12;
    BY hdipl;
RUN;
PROC BOXPLOT DATA = d.pisa12;
    PLOT PV1MATH*hdipl / BOXSTYLE = SCHEMATIC;
RUN;
PROC ANOVA DATA = d.pisa12;
    CLASS hdipl;
    MODEL PV1MATH = hdipl;
RUN;
```

Analyse bivariée : inférence

Exercice 4.1 On cherche à tester l'hypothèse

$$H_0 : \beta = c \quad \text{contre} \quad H_1 : \beta \neq c$$

On sait que sous H_0, une statistique Z suit une loi de Student à 4 degrés de liberté.

a. Ce test est-il un test bilatéral ou unilatéral ?

b. On souhaite mener ce test au niveau de confiance de 95 %. Utilisez les quantiles de la loi de Student en annexe pour déterminer la zone de rejet correspondante. Comment l'écririez-vous sous la forme $[q; +\infty[$?

c. Le calcul de Z donne la valeur 7,35. Que concluez-vous ?

d. Votre conclusion serait-elle modifiée si le test était mené à 99 % ?

Correction

a. L'hypothèse alternative est une réunion d'intervalles symétrique par rapport à la valeur ponctuelle testée par l'hypothèse nulle : le test est un test bilatéral.

b. i. Le test est bilatéral : la zone de rejet au niveau 95 % est de la forme $W =]-\infty; q_1[\cup]q_2; +\infty[$ où q_1 et q_2 sont des quantiles d'une loi de Student à 4 degrés de liberté.

ii. Pour obtenir un test au niveau de confiance de 5 %, il faut prendre $q_1 = q^{\mathcal{T}_4}_{2,5\ \%} = -2,78$ et $q_2 = q^{\mathcal{T}_4}_{97,5\ \%} = 2,78$ donc $W =]-\infty; -2,78[\cup]2,78; +\infty[$.

iii. La loi de Student étant symétrique, on a toujours la relation : $q_{1-\alpha/2} = -q_{\alpha/2}$. De ce fait, il est équivalent de vérifier $Z \in]-\infty; q_{\alpha/2}[\cup]q_{1-\alpha/2}; +\infty[$ et de vérifier $|Z| > q_{1-\alpha/2}$. Dans le cas présent, cela revient à vérifier si $|Z| > 2,78$.

c. Z = 7,35 donc $|Z| > 2,78$: on peut rejeter H_0 au seuil de 5 %.

d. Au seuil de 1 %, on cherche à vérifier si $|Z| > q^{\mathcal{T}_4}_{99,5\ \%} \Leftrightarrow |Z| > 4,60$. C'est bien le cas donc on peut également rejeter H_0 au seuil de 1 %.

Exercice 4.2 On cherche à tester l'hypothèse

$$H_0 : \beta = c \quad \text{contre} \quad H_1 : \beta > c$$

On sait que sous H_0, une statistique Z suit une loi du χ^2 à 8 degrés de liberté.

a. Ce test est-il un test bilatéral ou unilatéral ?

b. On souhaite mener ce test au niveau de confiance de 95 %. Utilisez les quantiles de la loi du χ^2 en annexe pour déterminer la zone de rejet.

c. Le calcul de Z donne la valeur 17,35. Que concluez-vous ?

d. Afin d'être plus prudent, on préfère en fait ne tolérer un risque de première espèce que de 1 % : cela modifie-t-il votre conclusion ? Qu'en déduisez-vous quand à la p-valeur associée à ce test ?

Correction

a. L'hypothèse alternative est un intervalle de la forme $]c;+\infty[$: le test est un test unilatéral.

b. i. Le test est unilatéral : la zone de rejet au niveau 95 % est de la forme $W =]q;+\infty[$ où q est un quantile d'une loi du χ^2 à 8 degrés de liberté.

ii. Pour obtenir un test au niveau de confiance de 95 %, il faut prendre $q = q_{95\ \%}^{\chi_8^2} = 15,51$ donc $W =]15,51;+\infty[$.

c. Z = 17,35 : $Z \in W$ donc on peut rejeter H_0 au seuil de 5 %.

d. Au seuil de 1 %, la zone de rejet devient : $W_{1\ \%} =]q_{99\ \%}^{\chi_8^2};+\infty[=]20,09;+\infty[$. Ainsi $Z \notin W_{1\ \%}$ donc on ne peut pas rejeter H_0 au seuil de 1 %. On peut donc rejeter l'hypothèse H_0 au seuil de 5 % mais pas au seuil de 1 % : la p-valeur est comprise entre 0,01 et 0,05 (elle vaut en fait 0,027).

Exercice 4.3 Dans la continuité de l'exercice 2.2, on cherche à tester si le score moyen des élèves scolarisés en France en mathématiques, compréhension de l'écrit et sciences est significativement différent de la valeur de référence 500 pour les seuils statistiques classiques (10 %, 5 % et 1 %).

a. Posez le test bilatéral d'égalité de la moyenne du score de mathématiques à 500. Quelle est la statistique de test correspondante et quelle loi suit-elle sous l'hypothèse d'égalité H_0 ?

b. En utilisant les informations de l'exercice 1.2, menez ce test au seuil de 10 %. Que concluez-vous ?

c. Réinterprétez les résultats et conclusions de l'exercice 2.2 dans le cadre de la théorie des tests statistiques.

d. Quelle procédure utiliser pour obtenir les tableaux suivants ? Interprétez la p-valeur du test : en quoi ce résultat complète-t-il ceux obtenus aux exercices précédents ?

Variable: PV1READ (Score aux évaluations de compréhension de l'écrit)

N	Mean	Std Dev	Std Err	Minimum	Maximum
1368	510.0	108.7	2.9388	69.5906	803.9

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
510.0	504.2	515.7	108.7	104.8	112.9

DF	t Value	Pr > |t|
1367	3.39	0.0007

Correction

a. On pose le test :

$$H_0 : PV1\bar{M}ATH = 500 \quad \text{contre} \quad H_1 : PV1\bar{M}ATH \neq 500$$

Sous l'hypothèse d'égalité H_0 on sait que la statistique

$$t_{PV1MATH} = \frac{PV1\bar{M}ATH - 500}{\hat{\sigma}_{PV1MATH}/\sqrt{n}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{T}_{n-1}$$

b. Application numérique : d'après le tableau de l'exercice 1.2, $PV1\bar{M}ATH = 498,2$, $\hat{\sigma}_{PV1MATH} = \sqrt{9232,6} = 96,1$ et $n = 1\ 368$. Ainsi

$$t_{PV1MATH} = \frac{498,2 - 500}{96,1/\sqrt{1\ 368}} = -0,7$$

Par ailleurs ce test est bilatéral : pour être mené au seuil de 10 %, il faut définir sa zone de rejet comme

$$W =]-\infty; q_{0,05}^{\mathcal{T}_{1\ 367}}[\cup]q_{0,95}^{\mathcal{T}_{1\ 367}}; +\infty[$$

c'est-à-dire

$$W =]-\infty; -1,64[\cup]1,64; +\infty[$$

Comme $t_{PV1MATH} = -0,7 \notin W$, on ne peut pas rejeter l'hypothèse nulle d'égalité au seuil de 10 % (ni à *fortiori* aux seuils plus prudents de 5 % et 1 %).

c. Construire un intervalle de confiance à un niveau donné et tester l'égalité à une valeur théorique sont deux opérations rigoureusement équivalentes. On peut donc réinterpréter les résultats de l'exercice 2.2 dans le cadre de la théorie des tests statistiques :

— d'une part, 500 appartient à l'intervalle de confiance à 95 % de la moyenne du score en sciences : on ne peut donc pas rejeter au seuil de 5 % l'hypothèse d'égalité de la moyenne du score en sciences à la valeur de référence de 500 ;

— d'autre part, 500 n'appartient pas à l'intervalle de confiance à 99 % de la moyenne du score en compréhension de l'écrit : on peut donc rejeter au seuil de 1 % l'hypothèse d'égalité de la moyenne du score en compréhension de l'écrit à la valeur de référence de 500.

d. C'est la **PROC TTEST** qu'il faut utiliser pour obtenir ces tableaux, avec l'option H0 :

```
PROC TTEST DATA = d.pisa12 H0 = 500;
    VAR PV1READ;
RUN;
```

La p-valeur du test est 0,0007 : la probabilité de se tromper en affirmant que le score moyen des élèves scolarisés en France en compréhension de l'écrit est significativement différent de la valeur de référence 500 est inférieure à 0,1 %. On peut donc bien rejeter l'hypothèse nulle d'égalité aux seuils de 5 % et 1 % comme déterminé par le calcul des intervalles de confiance, mais aussi au seuil encore plus restrictif de 0,1 %.

Exercice 4.4 On cherche à tester l'égalité des moyennes du score synthétique en mathématiques (PV1MATH) selon le sexe (ST04Q01). Les résultats sont les suivants :

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	1366	-1.56	0.1181
Satterthwaite	Unequal	1327.9	-1.56	0.1194

Equality of Variances				
Method	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Folded F	653	713	1.18	0.0318

Remarque : Le test d'égalité des moyennes à mettre en œuvre varie selon que la variance de la variable d'intérêt au sein des deux groupes testés est égale ou non. De ce fait, l'interprétation de ce type de test est à effectuer en deux temps : (1) test de l'hypothèse d'égalité des variances ; (2) test de l'hypothèse d'égalité des moyennes avec le test correspondant (variances égales ou inégales).

a. Quelle procédure a été utilisée pour produire ces résultats ?
b. Peut-on rejeter l'hypothèse d'égalité des variances au seuil de 5 % ? de 1 % ?
c. Interprétez le test d'égalité des moyennes sous l'hypothèse d'égalité des variances en utilisant la statistique de test et la table de quantile appropriée. Vérifiez que votre conclusion est bien cohérente avec l'interprétation de la p-valeur.
d. Comparez avec le test d'égalité des moyennes sous l'hypothèse d'inégalité des variances. Que retenez-vous d'un point de vue qualitatif ?

Correction

a. Il s'agit de la **PROC TTEST** :

```
PROC TTEST DATA = d.pisa12;
    CLASS ST04Q01;
    VAR PV1MATH;
RUN;
```

b. La p-valeur du test d'égalité des variances est de 0,0318. On peut donc rejeter l'hypothèse nulle d'égalité des variances au seuil de 5 % (0,0318 $<$ 0,050) mais pas au seuil de 1 % (0,0318 $>$ 0.010).

c. Sous l'hypothèse d'égalité des variances et sous l'hypothèse nulle d'égalité des moyennes, la statistique de test

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\hat{\sigma}(\frac{1}{n1} + \frac{1}{n_2})} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{T}_{n_1+n_2-2}$$

avec $\hat{\sigma} = \dfrac{(n_1-1)\hat{\sigma}_1^2 + (n_2-1)\hat{\sigma}_2^2}{n_1+n_2-2}$ .La correspondance des degrés de liberté confirme qu'il s'agit de la statistique calculée à la première ligne (**Pooled**) dont la valeur est -1,56.
On mène ce test bilatéral au seuil de 5 % : la zone de rejet est donc

$$W =]-\infty; q_{0,025}^{\mathcal{T}_{1\ 366}}[\cup]q_{0,975}^{\mathcal{T}_{1\ 366}}; +\infty[$$

c'est-à-dire

$$W =]-\infty; -1,96[\cup]1,96; +\infty[$$

.

Comme $t = -1,56 \notin W$, on ne peut pas rejeter l'hypothèse nulle d'égalité au seuil de 5 %. Ce résultat est bien cohérent avec la p-valeur du test, qui vaut 0,1181 et est donc supérieure à 0,05.

d. La p-valeur du test sous l'hypothèse de variances inégales est de 0,1194 : il est donc également impossible dans ce cas de rejeter H_0 au seuil de 5 % (et même de 10 %). Qualitativement, le test conduit donc au même résultat, que l'hypothèse d'égalité des variances soit vérifiée ou pas (c'est souvent le cas en pratique).

Exercice 4.5 On cherche à tester l'indépendance des variables croisées dans l'exercice 3.1 (sexe et retard scolaire au moment de l'enquête). Le tableau de résultat est le suivant :

Statistics for Table of ST04Q01 by retard

Statistic	DF	Value	Prob
Chi-Square	1	10.1644	0.0014
Likelihood Ratio Chi-Square	1	10.1642	0.0014
Continuity Adj. Chi-Square	1	9.7907	0.0018
Mantel-Haenszel Chi-Square	1	10.1570	0.0014
Phi Coefficient		0.0862	
Contingency Coefficient		0.0859	
Cramer's V		0.0862	

a. Rappelez comment est posé le test d'indépendance de deux variables qualitatives ainsi que le comportement de la statistique de test D^2 sous l'hypothèse nulle.

b. Quelle est la valeur de cette statistique ? Vérifiez qu'il est bien possible de la retrouver à partir du seul tableau de l'exercice 3.1. Quelle option utiliser pour faire apparaître le tableau de résultat ci-dessus ?

c. En utilisant les quantiles de la loi du χ^2 en annexe, déterminez la zone de rejet du test au seuil de 10 %. Que concluez-vous ?

d. Interprétez la p-valeur du test : est-il possible de rejeter l'hypothèse nulle à un seuil plus prudent que 10 % ?

Correction

a. Le test d'indépendance de deux variables qualitatives X et Y est posé de la façon suivante :

H_0 : {X et Y sont indépendantes } contre H_1 : {X et Y ne sont pas indépendantes}

Ce faisant, le risque de première espèce correspond au fait d'affirmer à tort que X et Y ne sont pas indépendantes, ce qui correspond à une approche « prudente ». Sous H_0, la statistique de test D^2 suit une loi du χ^2 à $(P-1)(Q-1)$ degrés de liberté, où P et Q sont le nombre de modalités des variables X et Y respectivement.

b. La lecture de la sortie permet d'identifier que la valeur de cette statistique est 10,1644. On peut la retrouver à partir du tableau de l'exercice 3.1 en sommant les contributions à la statistique du χ^2 de toutes les cellules (ligne **Cell Chi-Square**). Pour faire apparaître ce tableau, il suffit d'utiliser l'option `CHISQ` :

```
PROC FREQ DATA = d.pisa12;
    TABLES ST04Q01*retard / CHISQ;
    FORMAT ST04Q01 ST04Q01_.;
RUN;
```

c. Le test est unilatéral : on cherche donc le quantile à 90 % d'une loi du χ^2 à $(2-1)\times(2-1) = 1$ degré de liberté. La lecture des tables en annexe conduit à la zone de rejet $W =]2,71 : +\infty[$. $D^2 \in W$ donc on peut rejeter H_0 au seuil de 10 %. Pour une erreur de première espèce de 10 % (ce qui est assez élevé), on peut estimer que les variables de sexe et de retard scolaire sont statistiquement liées. Les surreprésentations identifiées à l'exercice 3.1 sont donc statistiquement significatives au seuil de 10 %.

d. La p-valeur du test est de 0,0014. Elle est donc inférieure à 0,05 et à 0,01 : on peut donc en réalité rejeter H_0 aux seuils de 5 % et même de 1 %. L'association entre sexe et retard scolaire peut donc être jugée très significative.

Code SAS utilisé pour produire l'exercice

```
PROC FREQ DATA = d.pisa12;
    TABLES ST04Q01*retard / CHISQ;
    FORMAT ST04Q01 ST04Q01_.;
RUN;
```

Exercice 4.6 On cherche à tester la significativité de la corrélation entre les scores en mathématiques et en compréhension de l'écrit. L'ensemble des informations nécessaires figurent dans le tableau de l'exercice 3.2.

a. Rappelez comment est posé le test d'indépendance de deux variables quantitatives ainsi que le comportement de la statistique de test t sous l'hypothèse nulle.

b. Calculez la statistique de test et, en utilisant les quantiles de la loi de Student en annexe, menez le test correspondant au seuil de 1 %.

c. Interprétez par ailleurs la p-valeur et concluez.

Correction

a. Le test d'indépendance de deux variables quantitatives X et Y est posé de la façon suivante :

$$H_0 : r_{X,Y} = 0 \quad \text{contre} \quad H_1 : r_{X,Y} \neq 0$$

Ce faisant, le risque de première espèce correspond au fait d'affirmer à tort que la corrélation entre X et Y est différente de 0, ce qui correspond à une approche « prudente ». Sous H_0, la statistique de test t suit une loi de Student à $n-2$ degrés de liberté, où n est le nombre d'observations intervenant dans le calcul de la corrélation.

b. On sait que sous l'hypothèse nulle,

$$t = r_{X,Y} \times \sqrt{\frac{n-2}{1-r_{X,Y}^2}} \hookrightarrow \mathcal{T}_{n-2}$$

Tous les éléments nécessaires figurent dans la sortie présentée dans l'exercice 3.2 pour calculer cette statistique de test, aussi :

$$t = 0,86722 \times \sqrt{\frac{1\,368-2}{1-0,86722^2}} = 64,37$$

Ce test est un test bilatéral : on cherche donc la valeur du quantile à 99,5 % d'une loi de Student à $1\,368 - 2 = 1\,366$ degrés de liberté. La lecture des tables conduit à la valeur 2,58. 64,37 > 2,58 donc on rejette très largement H_0 au seuil de 1 %. Comme le laissait anticiper sa valeur très élevée, la corrélation entre score au test de mathématiques et score au test de compréhension de l'écrit est très significative.

c. L'interprétation de la p-valeur conduit (par définition) au même résultat : celle-ci est inférieure à 0,0001 et donc *a fortiori* à 0,01 ; on peut donc bien rejeter l'hypothèse H_0 au seuil de 1 % (et en fait également à des seuils plus prudents).

Annexe : Tables statistiques usuelles

Table 1 : Quantiles de la loi normale centrée réduite

Le quantile de niveau γ d'une variable aléatoire X suivant une loi normale centrée réduite noté $q_\gamma^{\mathcal{N}(0,1)}$ est défini par :

$$\Phi(q_\gamma^{\mathcal{N}(0,1)}) = \mathbb{P}(X \leq q_\gamma^{\mathcal{N}(0,1)}) = \gamma$$

où Φ est la fonction de répartition de la loi normale centrée réduite.

γ	**0,005**	**0,01**	**0,025**	**0,05**	**0,1**	**0,9**	**0,95**	**0,975**	**0,99**	**0,995**
$q_\gamma^{\mathcal{N}(0,1)}$	-2,58	-2,33	-1,96	-1,64	-1,28	1,28	1,64	1,96	2,33	2,58

Lecture : 97,5% des valeurs d'une variable aléatoire suivant une loi normale centrée réduite sont inférieures à 1,96.

Table 2 : Quantiles de la loi du χ^2

Le quantile de niveau γ d'une variable aléatoire X suivant une loi du χ^2 à p degrés de liberté noté $q_\gamma^{\chi^2_p}$ est défini par :

$$F_X(q_\gamma^{\chi^2_p}) = \mathbb{P}(X \leq q_\gamma^{\chi^2_p}) = \gamma$$

où F_X est la fonction de répartition de X.

p \ γ	**0,005**	**0,01**	**0,025**	**0,05**	**0,1**	**0,9**	**0,95**	**0,975**	**0,99**	**0,995**
1	0,00	0,00	0,00	0,00	0,02	2,71	3,84	5,02	6,63	7,88
2	0,01	0,02	0,05	0,10	0,21	4,61	5,99	7,38	9,21	10,60
3	0,07	0,11	0,22	0,35	0,58	6,25	7,81	9,35	11,34	12,84
4	0,21	0,30	0,48	0,71	1,06	7,78	9,49	11,14	13,28	14,86
5	0,41	0,55	0,83	1,15	1,61	9,24	11,07	12,83	15,09	16,75
6	0,68	0,87	1,24	1,64	2,20	10,64	12,59	14,45	16,81	18,55
7	0,99	1,24	1,69	2,17	2,83	12,02	14,07	16,01	18,48	20,28
8	1,34	1,65	2,18	2,73	3,49	13,36	15,51	17,53	20,09	21,95
9	1,73	2,09	2,70	3,33	4,17	14,68	16,92	19,02	21,67	23,59
10	2,16	2,56	3,25	3,94	4,87	15,99	18,31	20,48	23,21	25,19
20	7,43	8,26	9,59	10,85	12,44	28,41	31,41	34,17	37,57	40,00
30	13,79	14,95	16,79	18,49	20,60	40,26	43,77	46,98	50,89	53,67
40	20,71	22,16	24,43	26,51	29,05	51,81	55,76	59,34	63,69	66,77
50	27,99	29,71	32,36	34,76	37,69	63,17	67,50	71,42	76,15	79,49
100	67	70	74	78	82	118	124	130	136	140
1000	889	899	914	928	943	1058	1075	1090	1107	1119

Lecture : 95% des valeurs d'une variable aléatoire suivant une loi du χ^2 à 1 degré de liberté sont inférieures à 3,84.

Table 3 : Quantiles de la loi de Student

Le quantile de niveau γ d'une variable aléatoire X suivant une loi de Student à p degrés de liberté noté $F_X(q_\gamma^{\mathcal{T}_p}) = q_\gamma^{\mathcal{T}_p}$ est défini par :

$$F(q_\gamma^{\mathcal{T}_p}) = \mathbb{P}(X \leq q_\gamma^{\mathcal{T}_p}) = \gamma$$

où F_X est la fonction de répartition de X.

p \ γ	**0,005**	**0,01**	**0,025**	**0,05**	**0,1**	**0,9**	**0,95**	**0,975**	**0,99**	**0,995**
1	-63,66	-31,82	-12,71	-6,31	-3,08	3,08	6,31	12,71	31,82	63,66
2	-9,92	-6,96	-4,30	-2,92	-1,89	1,89	2,92	4,30	6,96	9,92
3	-5,84	-4,54	-3,18	-2,35	-1,64	1,64	2,35	3,18	4,54	5,84
4	-4,60	-3,75	-2,78	-2,13	-1,53	1,53	2,13	2,78	3,75	4,60
5	-4,03	-3,36	-2,57	-2,02	-1,48	1,48	2,02	2,57	3,36	4,03
6	-3,71	-3,14	-2,45	-1,94	-1,44	1,44	1,94	2,45	3,14	3,71
7	-3,50	-3,00	-2,36	-1,89	-1,41	1,41	1,89	2,36	3,00	3,50
8	-3,36	-2,90	-2,31	-1,86	-1,40	1,40	1,86	2,31	2,90	3,36
9	-3,25	-2,82	-2,26	-1,83	-1,38	1,38	1,83	2,26	2,82	3,25
10	-3,17	-2,76	-2,23	-1,81	-1,37	1,37	1,81	2,23	2,76	3,17
20	-2,85	-2,53	-2,09	-1,72	-1,33	1,33	1,72	2,09	2,53	2,85
30	-2,75	-2,46	-2,04	-1,70	-1,31	1,31	1,70	2,04	2,46	2,75
40	-2,70	-2,42	-2,02	-1,68	-1,30	1,30	1,68	2,02	2,42	2,70
50	-2,68	-2,40	-2,01	-1,68	-1,30	1,30	1,68	2,01	2,40	2,68
100	-2,63	-2,36	-1,98	-1,66	-1,29	1,29	1,66	1,98	2,36	2,63
1000	-2,58	-2,33	-1,96	-1,65	-1,28	1,28	1,65	1,96	2,33	2,58
$+\infty$	-2,58	-2,33	-1,96	-1,64	-1,28	1,28	1,64	1,96	2,33	2,58

Lecture : 97,5% des valeurs d'une variable aléatoire suivant une loi de Student à 1 degré de liberté sont inférieures à 12,71. On remarque que quand p tend vers $+\infty$, les quantiles de la loi de Student sont ceux de la loi normale centrée réduite.