ENSAE-ENSAI Formation continue [Cepe]

Certificat de
chargé d'études statistiques

Sessions de révision – Exercices pratiques

Martin Chevalier (INSEE) – *Version corrigée*

La plupart des exercices pratiques proposés dans le cadre des sessions de révision des 2, 3 et 4 mai s'appuie 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 ainsi l'avantage de comporter une très large variété de variables qualitatives et quantitatives. Elle se prête ainsi à tous les outils et méthodes au programme des sessions de révision des 2, 3 et 4 mai.

Les fichiers de l'enquête PISA 2012 sont librement téléchargeables sur le site de l'OCDE[1]*. Le fichier « élèves » réduit*[2] *ainsi que le code ayant servis à la production des sorties statistiques utilisées dans les exercices pratiques sont fournis 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.

Session 1 : Statistique descriptive

Question 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). L'enquête permet d'obtenir les deux tris à plat suivants :

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

ST01Q01	Frequency	Percent	Cumulative Frequency	Cumulative Percent
8	3441.067	1.67	3441.067	1.67
9	59428.85	28.79	62869.91	30.45
10	135518.9	65.64	198388.8	96.10
11	7890.986	3.82	206279.8	99.92
12	165.3195	0.08	206445.1	100.00

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

b. Interprétez les résultats du tri à plat de gauche. Comment recoderiez-vous cette variable pour en mener l'étude ?

c. Comparez les résultats des deux tris à plat. À quoi les différences observées sont-elles dues à votre avis ? Quel tri à plat privilégieriez-vous et pourquoi ?

Correction

a. Il s'agit de 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. La principale différence entre les deux tableaux tient à la valeur des effectifs (colonne **Frequency**) : la somme des effectifs vaut 1 368 dans le tableau de gauche contre 206 445,1 dans le tableau de droite. Cette différence provient très certainement de l'utilisation des poids de sondage de l'enquête pour produire le tableau de droite :

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

En termes de pourcentages, les écarts sont néanmoins assez faibles : de fait les poids de sondage de l'enquête sont peu dispersés et conduisent donc à des écarts faibles par rapport aux estimations non-pondérées. On a tendance à privilégier néanmoins les estimations pondérées, dans la mesure où elles seules permettent de généraliser les résultats calculés à partir de l'échantillon à l'ensemble de la population d'étude.

Question 1.2 La variable PV1MATH correspond au score synthétique de l'élève aux évaluations de mathématique. Le tableau suivant en synthétise la distribution (non-pondérée) :

Analysis Variable : PV1MATH					
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 ces résultats ?

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 498,6. On remarque que les valeurs maximales et minimales sont du même ordre de grandeur que la moyenne et que la variance est particulièrement faible. 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 :

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

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

$$CV(\text{PV1MATH}) = \frac{s_{\text{PV1MATH}}}{\bar{\text{PV1MATH}}} = \frac{96,09}{498,18} = 19,29\ \%$$

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

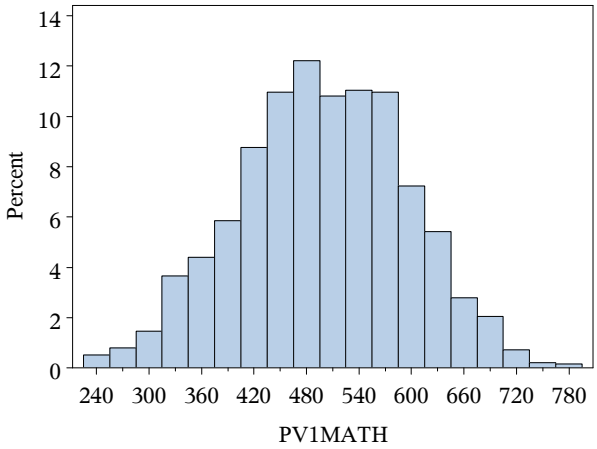
a. Quel est le nom de ces deux graphiques ? En quoi leur analyse confirme-t-elle les résultats de la question précédente quant à l'influence des valeurs extrêmes ?

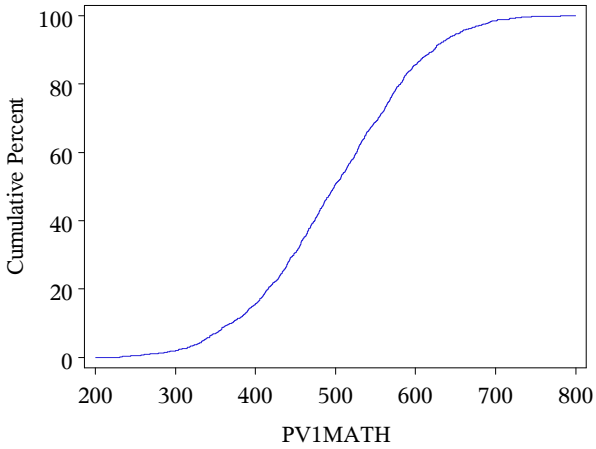
b. Utilisez ces graphiques pour déterminer (approximativement) la valeur des quartiles de la variable PV1MATH ainsi que celle des premier et neuvième déciles.

c. Quelle procédure auriez-vous pu utiliser pour obtenir directement les quantiles de la variable PV1MATH [3] ?

Correction

3. C'est en fait cette même procédure qui a été utilisée pour produire ces graphiques.





a. Ces deux graphiques sont l'histogramme (à gauche) et la fonction de répartition (à droite) de `PV1MATH`. L'histogramme confirme la très grande régularité de la distribution de `PV1MATH` (qui semble même gaussienne, *cf.* **Question 3.1**), 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` : 400 (D1), 450 (Q1), 550 (Q3), 620 (D9).

c. Il s'agit de la **`PROC UNIVARIATE`** :

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

Question 1.4 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	retard 0	retard 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

Remarque : La modalité `"1"` de la variable `ST04Q01` correspond aux femmes et la modalité `"2"` aux hommes, à rebours des conventions françaises. Un label a été appliqué pour plus de clarté.

a. Quelles statistiques déjà présentes dans la question 1 retrouvez-vous dans ce tableau ? Comment les désigne-t-on dans le contexte du tri croisé ?

b. Interprétez un pourcentage de cellule, un pourcentage en ligne et un pourcentage en colonne.

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

Correction

a. Les effectifs et pourcentages en pied de colonne sont calculés à partir de ceux du tableau de gauche de la **Question 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 % hommes et femmes confondus. 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 commenter 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.* **Question 2.4**).

Question 1.5 Les variables `PV1READ` et `PV1SCIE` correspondent respectivement aux scores synthétiques de l'élève aux évaluations de compréhension de l'écrit et de sciences. 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	1.00000	0.86722 <.0001	0.89773 <.0001
PV1READ	0.86722 <.0001	1.00000	0.88809 <.0001
PV1SCIE	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. Des indicateurs analogues existent-ils ? Quel est leur intérêt et comment les calculer ?

Correction

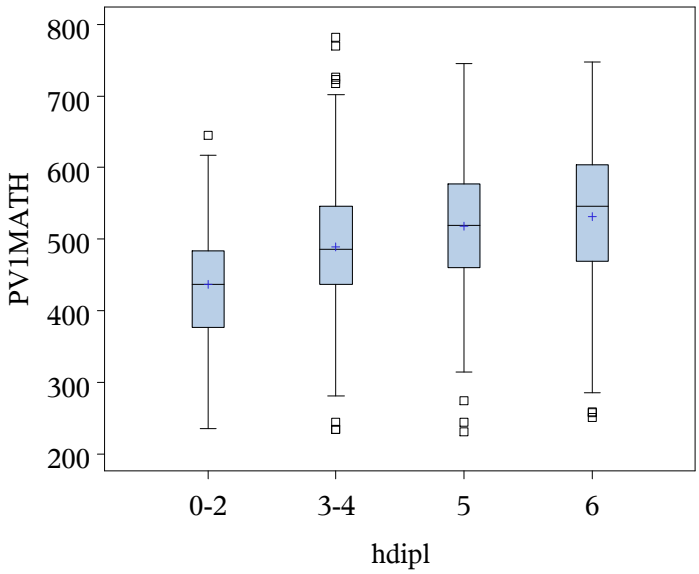
a. Il s'agit de la **PROC CORR** :

```
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.* **Question 2.5**), 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 : coefficient de corrélation des rangs de Spearman et τ de Kendall (*cf.* support de révision). Dans les deux cas, l'objectif est d'obtenir une mesure d'association qui soit moins sensible aux valeurs extrêmes que le coefficient de corrélation de Pearson.

Question 1.6 On dispose par ailleurs du plus haut niveau d'étude des parents que l'on regroupe en quatre modalités [4] : 0-2 Aucun, primaire ou collège ; 3-4 Fin d'études secondaires ou post-secondaire non-supérieur ; 5 Premier cycle d'études supérieures ; 6 Seconde cycle 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. Quel nom porte ce type de graphique et avec quelle procédure le construire ?

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 ?

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

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;
```

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 (quo correspond au graphique présenté), les valeurs extrêmes au-delà des moustaches sont singularisées par un point, une croix 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.* session 3).

Session 2 : Statistique inférentielle

Question 2.1 À partir des informations suivantes, construisez l'intervalle de confiance à 95 % de la moyenne des scores de mathématiques, compréhension de l'écrit et sciences :

Variable	N	Mean	Std Dev
PV1MATH	1368	498.1789377	96.0865993
PV1READ	1368	509.9585724	108.6969627
PV1SCIE	1368	503.4514466	98.0487839

Quelle procédure permet d'obtenir directement ces intervalles de confiance ?

Correction En appliquant le théorème central limite, on a pu montrer que l'intervalle de confiance à 95 % de la moyenne d'une variable Y est :

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

On applique donc cette formule aux variables `PV1MATH`, `PV1READ` et `PV1SCIE` :

- $IC_{95\ \%}(\text{PV1}\bar{\text{M}}\text{ATH}) = 498,18 \pm 1,96 \times \dfrac{96,09}{\sqrt{1368}} = [493,09; 503,27]$
- $IC_{95\ \%}(\text{PV1}\bar{\text{R}}\text{EAD}) = 509,96 \pm 1,96 \times \dfrac{108,70}{\sqrt{1368}} = [504,20; 515,72]$
- $IC_{95\ \%}(\text{PV1}\bar{\text{S}}\text{CIE}) = 503,45 \pm 1,96 \times \dfrac{98,05}{\sqrt{1368}} = [498,25; 508,65]$

Il est possible de mener automatiquement ces calculs avec SAS à l'aide d'une **`PROC TTEST`** :

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

Question 2.2 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 région critique. Quelle est la valeur critique ?

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 ?

Correction

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

b. Le test est unilatéral : la région critique au niveau 95 % est de la forme $W_{5\ \%} =]q; +\infty[$ où q est un quantile de la loi suivie par la statistique de test sous H_0. D'après les tables en annexe, $q = q^{\chi^2_8}_{95\ \%} = 15,51$ donc $W_{5\ \%} =]15,51; +\infty[$.

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

d. Au seuil de 1 %, la région critique devient : $W_{1\ \%} =]q^{\chi^2_8}_{99\ \%}; +\infty[=]20,09; +\infty[$. Ainsi $Z \notin W_{1\ \%}$ donc on ne peut pas rejeter H_0 au seuil de 1 %. Le choix d'un seuil plus prudent pour l'erreur de première espèce modifie notre conclusion.

Question 2.3 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 région critique correspondante. Est-il possible de l'écrire 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. Le test est bilatéral : la région critique au niveau 95 % est de la forme $W_{5\ \%} =]-\infty; q_1[\cup]q_2; +\infty[$ où q_1 et q_2 sont des quantiles de la loi suivie par la statistique de test sous H_0. Ici, $q_1 = q^{\mathcal{T}_4}_{2,5\ \%} = -2,78$ et $q_1 = q^{\mathcal{T}_4}_{97,5\ \%} = 2,78$ donc $W_{5\ \%} =]-\infty; -2,78[\cup]2,78; +\infty[$.
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 rejeter H_0 au seuil de 1 %.

Question 2.4 On cherche à tester l'indépendance des variables croisées dans la **Question 1.4** (sexe et retard scolaire au moment de l'enquête). Le tableau de résultat est le suivant :

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 sa 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 la **Question 1.4**. 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 valeur critique 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 % ? En aurait-il été de même si cette p-valeur avait valu 0,023 ?

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.

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 la **Question 1.4** 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 valeur 2,71. $D^2 > 2,71$ 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 la **Question 1.4** 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.

Question 2.5 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 qui accompagne la **Question 1.5**.

a. Rappelez comment est posé le test d'indépendance de deux variables quantitatives ainsi que le comportement de sa 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. Par ailleurs, interprétez 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

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

Tous les éléments sont présents dans la sortie présentée dans la **Question 1.5** pour calculer cette statistique de test, aussi :

$$t = 0,86722 \times \sqrt{\frac{1368-2}{1-0,86722^2}} = 64,37$$

Ce test est un test bilatéral : on cherche donc la valeur du quantile à 99,5 % d'une loi de Student à $1368 - 2 = 1366$ 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).

Session 3 : Analyse de variance

Question 3.1 Hypothèses de l'ANOVA Dans cette question, on souhaite tester sur les données de l'enquête PISA 2012 les hypothèses de normalité et d'homogénéité. Les résultats des tests de Shapiro-Wilk et de Bartlett menés sur l'ANOVA du score synthétique en mathématiques (PV1MATH) selon le sexe (ST04Q01) sont les suivants :

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.99814	**Pr < W**	0.1322
Kolmogorov-Smirnov	**D**	0.020463	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.074578	**Pr > W-Sq**	0.2464
Anderson-Darling	**A-Sq**	0.474044	**Pr > A-Sq**	0.2453

Bartlett's Test for Homogeneity of PV1MATH Variance			
Source	**DF**	**Chi-Square**	**Pr > ChiSq**
ST04Q01	1	4.6070	0.0318

a. Rappelez l'hypothèse nulle et l'hypothèse alternative de ces deux tests.

b. Commentez la p-valeur du test de Shapiro-Wilk : que concluez-vous ? Êtes-vous surpris du résultat (en repensant aux questions de la session 1) ?

c. Commentez la statistique de test et la p-valeur du test de Bartlett. Que concluez-vous ? Quelle est la conséquence de ce résultat sur l'analyse de la variance ?

Correction

a. Le test de Shapiro-Wilk teste la normalité d'une distribution :

H_0 : {La distribution est normale} contre H_1 : {La distribution n'est pas normale}

Le test de Bartlett teste l'égalité des variances au sein de différents sous-groupes :

$$H_0 : \sigma_1^2 = \ldots = \sigma_K^2 = \sigma^2 = \text{constante} \quad \text{contre} \quad H_1 : \exists k \in \{1, \ldots, K\} \quad \sigma_k^2 \neq \sigma^2$$

b. La p-valeur du test de Shapiro-Wilk est 0,1322. Elle est supérieure à 0,10 aussi on ne peut pas rejeter au seuil de 10 % l'hypothèse H_0 de normalité de la distribution de la variable PV1MATH. La seule lecture de l'histogramme de cette variable, présenté à la **Question 1.5**, suffit à souligner sa proximité avec une distribution gaussienne.

c. La p-valeur du test de Bartlett est 0,0318 : on peut rejeter l'hypothèse nulle d'égalité des variances au seuil de 5 % mais pas au seuil de 1 %. Ce résultat doit nous amener à privilégier l'hypothèse de variances inégales, tout en examinant par sécurité si l'hypothèse de variances égales conduit à des résultats très différents.

Question 3.2 ANOVA selon un facteur dichotomique : TTEST 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

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

b. Interprétez le test d'égalité des variances. Comparez avec les résultats de la **Question 3.1**.

c. Interprétez le test d'égalité des moyennes sous l'hypothèse d'inégalité des variances en utilisant la statistique de test. 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'égalité des variances. Que retenez-vous d'un point de vue qualitatif ?

<u>Correction</u>

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

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

b. La p-valeur du test d'égalité des variances est exactement la même que celle du test de Bartlett à la **Question 3.1** (le test mené est rigoureusement équivalent). On est donc amené à privilégier l'hypothèse de variances inégales, tout en examinant si l'hypothèse de variances égales change ou pas les résultats.

c. Sous l'hypothèse de variances inégales, la statistique de test vaut $t = -1,56$. On sait que sous H_0 elle suit une loi de Student à $n - 2 = 1368 - 2 = 1366$ degrés de liberté. La région critique à 5 % est donc

$$W_{5\ \%} =]-\infty; q^{\mathcal{T}^{1366}}_{2,5\ \%}[\cup]q^{\mathcal{T}^{1366}}_{97,5\ \%}; +\infty[=]-\infty; -1,96[\cup]1,96; +\infty[$$

Donc $t \notin W_{5\ \%}$: on ne peut donc pas rejeter H_0 au seuil de 5 %. Ceci est (par définition) cohérent avec la p-valeur du test, qui est de $0,1194 > 0,05$. De fait, on ne peut pas même rejeter H_0 au seuil de 10 %. En pratique on conclut que le lien entre sexe et score en mathématiques ne peut pas être considéré comme statistiquement significatif.

d. La p-valeur du test sous l'hypothèse d'égalité des variances est de 0,1181 : il est donc également impossible dans ce cas de rejeter H_0 au seuil 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).

Question 3.3 ANOVA selon un facteur polytomique On cherche à analyser la variance du score synthétique en mathématiques (`PV1MATH`) selon le plus haut diplôme des parents (`hdipl`). Les résultats sont les suivants :

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1060169.54	353389.85	43.56	<.0001
Error	1289	10457452.19	8112.84		
Corrected Total	1292	11517621.72			

R-Square	Coeff Var	Root MSE	PV1MATH Mean
0.092048	17.89535	90.07131	503.3223

Source	DF	Anova SS	Mean Square	F Value	Pr > F
hdipl	3	1060169.538	353389.846	43.56	<.0001

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

b. Rappelez la formule de la statistique de test utilisée dans le test de l'ANOVA. Repérez-la dans les sorties puis utilisez les autres résultats pour la recalculer.

c. On souhaite mener le test au seuil de 5 %. Quels sont les degrés de liberté de la loi de Fisher que suit la statistique de test sous H_0 ? Quelle est la valeur critique du test ? Que concluez-vous ? Vérifiez que votre conclusion est bien cohérente avec l'interprétation de la p-valeur.

Correction

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

```
PROC ANOVA DATA = d.pisa12;
    CLASS hdipl;
    MODEL PV1MATH = hdipl;
RUN;
```

b. La statistique de test de l'ANOVA est la suivante :

$$F = \frac{SCE/(K-1)}{SCR/(n-K)}$$

où K est le nombre de modalités de la variable explicative. Cette statistique vaut dans l'exemple 43,56 et figure dans le premier et le troisième tableau de la sortie.

Les éléments des colonnes **Sum of Squares** et **DF** du premier tableau permettent de la recalculer :

$$\text{SCE} = 1060169,54 \quad K-1=3 \quad \text{SCR} = 10457452,19 \quad n-K=1289$$

Le numérateur et le dénominateur de la statistique de test figurent également directement dans la colonne **Mean Square**.

c. Sous H_0 la statistique de test F suit une loi de Fisher à $K-1=3$ et $n-K=1289$ degrés de liberté. Le test est unilatéral, aussi sa valeur critique à 5 % est le quantile à 95 % d'une $F_{3,1289}$, c'est-à-dire d'après les tables en annexe environ 2,61. F > 2,61 donc on peut très largement rejeter l'hypothèse nulle d'égalité des variances au seuil de 5 %. On peut directement arriver à la même conclusion en interprétant la p-valeur, qui est bien inférieure à 0,05.

Remarque : Si on procède à l'analyse de la variance du score synthétique en mathématiques selon le sexe en utilisant la même procédure que dans la **Question 3.3**, on obtient les résultats suivants (partiels) :

Source	DF	Anova SS	Mean Square	F Value	Pr > F
ST04Q01	1	22553.10330	22553.10330	2.45	0.1181

Comparez la p-valeur avec celles obtenues à la **Question 3.2**. Que remarquez-vous ? Comparez également les statistiques de test : remarquez-vous une relation entre t et F ?

Correction

- La p-valeur est exactement la même que celle obtenue avec la **PROC TTEST** sous l'hypothèse d'égalité des variances. Cela illustre bien le fait que l'analyse de variance dans le cas général (que mène la **PROC ANOVA**) est équivalente à la **PROC TTEST** sous l'hypothèse d'égalité des variances.
- Plus encore, on peut remarquer que la statistique de test F de la **PROC ANOVA** (2,45) est (aux arrondis près) le carré de la statistique de test t de la **PROC TTEST** (-1,56). Cette relation entre les statistiques de test explique la relation entre les p-valeurs, dans la mesure où on sait que si une variable aléatoire suit une loi de Student à p degrés de liberté, alors son carré suit une loi de Fisher à 1 et p degrés de liberté.

Session 4 : Régression linéaire

L'objectif des questions de cette session est d'identifier certains déterminants des résultats au test standardisé de mathématiques (variable `PV1MATH`). Les variables explicatives sont intégrées une à une dans le modèle, dans l'ordre :
- le nombre d'heures de travail personnel consacré aux mathématiques chaque semaine : variable `mhours` ;
- le sexe : variable `ST04Q01` dichotomisée avec les variables `femme` et `homme` ;
- le plus haut niveau d'étude atteint par les parents : variable `hdipl` dichotomisée avec les variables `hdipl02`, `hdipl34`, `hdipl5` et `hdipl6` (*cf.* **Question 1.6** pour la signification des modalités de cette variable).

Question 4.1 Régression linéaire simple On estime tout d'abord le modèle :

$$\text{PV1MATH} = \beta_0 + \beta_1 \times \text{mhours} + u$$

dont les résultats sont les suivants :

Number of Observations Read	1368
Number of Observations Used	724
Number of Observations with Missing Values	644

Root MSE	89.90534	**R-Square**	0.0226
Dependent Mean	511.60180	**Adj R-Sq**	0.0212
Coeff Var	17.57330		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	476.46714	9.23439	51.60	<.0001
mhours	1	10.29129	2.52157	4.08	<.0001

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

b. Que remarquez-vous quant au nombre d'observations effectivement utilisées dans le modèle ? Quelle est selon vous l'origine de ce phénomène ?

c. Quelle est la valeur du R^2 du modèle ? Est-ce une valeur faible ou une valeur élevée ?

d. Interprétez la valeur de $\hat{\beta}_1$. Sachant que la variance de `mhours` vaut 1,74 et la covariance de `PV1MATH` et `mhours` vaut 18,10, recalculez-la manuellement.

e. Rappelez l'hypothèse nulle et l'hypothèse alternative du test de significativité de la valeur du coefficient β_1. Quelle est sa statistique de test et quelle loi suit-elle sous l'hypothèse nulle ?

f. Menez le test de significativité de β_1 au seuil de 5 %. Que concluez-vous ? Vérifiez que cette conclusion est cohérente avec l'interprétation de la p-valeur du test.

Correction

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

```
PROC REG DATA = d.pisa12;
    MODEL PV1MATH = mhours;
RUN; QUIT;
```

b. Le premier tableau renseigne sur le nombre d'observations effectivement utilisées dans le modèle, ici 724 sur 1368. C'est une diminution assez forte du nombre d'observations disponibles. Les 644 observations restantes ont été écartées du modèle en raison de valeurs manquantes à la variable `mhours` : il s'agit très certainement de cas de non-réponse.

c. Le R^2 du modèle est 0,0226 : le modèle ne rend compte que de 2,26 % de la variance du score synthétique en mathématiques, ce qui est très faible.

d. $\hat{\beta}_1 = 10,29$: en moyenne, à une heure de travail personnel en mathématique supplémentaire par semaine est associé un score au test standardisé de mathématiques de 10,29 points supérieurs. Dans le cas de la régression linéaire simple, ce coefficient est facilement obtenu par :

$$\hat{\beta}_1 = \frac{Cov(\text{PV1MATH}, \text{mhours})}{V(\text{mhours})} = \frac{18,10}{1,74} \approx 10,29 \text{ (aux arrondis près)}$$

e. Le test de significativité de β_1 est posé de la façon suivante :

$$H_0 : \beta_1 = 0 \quad \text{contre} \quad H_1 : \beta_1 \neq 0$$

On sait que sous H_0 :

$$t_{\beta_1} = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \hookrightarrow \mathcal{T}_{n-(K+1)}$$

où $se(\hat{\beta}_1)$ est l'erreur standard de $\hat{\beta}_1$.

f. Le test est un test bilatéral : on rejette donc H_0 au seuil de 5 % dès lors que $|t| > q_{97,5}^{\mathcal{T}_{n-(K+1)}}$. $n-(K+1) = 724-(1+1) = 722$ et $q_{97,5}^{\mathcal{T}_{722}} = 1,96$. Or on lit dans le tableau que $t_{\beta_1} = 4,08$ donc $|t_{\beta_1}| > 1,96$: on peut rejeter H_0 au seuil de 5 %. En pratique, on conclut que $\hat{\beta}_1$ est significatif au seuil de 5 % et donc que la relation entre travail personnel et score au test standardisé en mathématiques est statistiquement significative. Ceci est cohérent avec le fait que la p-valeur du test est inférieure à 0,0001 donc *a fortiori* inférieure à 0,05.

Question 4.2 On intègre la variable de sexe au modèle :

$$\text{PV1MATH} = \beta_0 + \beta_1 \times \text{mhours} + \beta_2 \times \text{femme} + u$$

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	142478	71239	8.81	0.0002
Error	721	5828065	8083.30726		
Corrected Total	723	5970543			

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	479.81434	9.84016	48.76	<.0001
mhours	1	10.32120	2.52181	4.09	<.0001
femme	1	-6.58919	6.69062	-0.98	0.3250

a. Rappelez l'hypothèse nulle et l'hypothèse alternative du test de significativité globale des coefficients d'une régression. Quelle est sa statistique de test et quelle loi suit-elle sous l'hypothèse nulle ?

b. Repérez la valeur de cette statistique de test et menez le test à un niveau de confiance de 95 %. Que concluez-vous ? Vérifiez que cette conclusion est bien cohérente avec l'interprétation de la p-valeur du test.

c. Pourquoi n'a-t-on pas intégré directement la variable de sexe (`ST04Q01`) dans le modèle ? Pourquoi la variable indicatrice `homme` n'apparaît-elle pas ?

d. Interprétez la valeur de $\hat{\beta}_2$. Ce coefficient est-il significativement différent de 0 aux seuils de 10 %, 5 %, 1 % ?

Correction

a. Le test de significativité globale des coefficients de cette régression à deux variables explicatives est posé de la façon suivante :

$$H_0 : \beta_1 = 0 \text{ et } \beta_2 = 0 \quad \text{contre} \quad \beta_1 \neq 0 \text{ ou } \beta_2 \neq 0$$

Sous H_0, on sait que :

$$F = \frac{SCE/K}{SCR/(n-(K+1)} \hookrightarrow F_{K,n-(K+1)}$$

avec K le nombre de variables explicatives, ici $K = 2$.

b. La statistique de test figure dans le tableau **Analysis of Variance** dans la colonne **F Value** soit 8,81. Le test est unilatéral aussi la valeur critique est le quantile à 95 % d'un loi de Fisher à 2 et $n-(K+1) = 724-(2+1) = 721$ degrés de liberté soit entre 3,00 et 3,09. F est donc supérieure à la valeur critique du test au seuil de 5 % : on peut rejeter H_0 au seuil de 5 %. En pratique, on conclut que le modèle proposé est globalement explicatif.

c. La variable `ST04Q01` est une variable qualitative à deux modalités (`1` et `2`) : l'intégrer telle quelle dans la régression n'a pas de sens, il convient de la dichotomiser. Intégrer les deux variables indicatrices correspondantes (`homme` et `femme`) aurait pour conséquence d'induire une situation de colinéarité parfaite avec la constante du modèle : par définition, la somme de `homme` et `femme` vaut 1 pour tous les individus. Une des deux variables indicatrices (`homme`) est donc omise du modèle : c'est la modalité de référence.

d. $\hat{\beta}_2 = -6,59$: à temps de travail personnel en mathématiques égal, le fait d'être une femme (plutôt qu'un homme) est associé en moyenne à un score synthétique en mathématiques plus faible de 6,59 points. La p-valeur du test de significativité est néanmoins largement supérieure à 0,10 : on ne peut donc pas rejeter l'hypothèse de nullité de ce coefficient. En pratique on conclut que ce coefficient n'est pas significatif.

Question 4.3 On intègre au modèle la variable de plus haut niveau d'étude atteint par les parents par le biais de ses indicatrices :

$$\text{PV1MATH} = \beta_0 + \beta_1 \times \text{mhours} + \beta_2 \times \text{femme} + \beta_3 \times \text{hdipl02} + \beta_4 \times \text{hdipl5} + \beta_5 \times \text{hdipl6} + u$$

Root MSE	85.35541	**R-Square**	0.1239
Dependent Mean	511.60180	**Adj R-Sq**	0.1178
Coeff Var	16.68395		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	463.89103	10.08136	46.01	<.0001
mhours	1	9.38701	2.39872	3.91	<.0001
femme	1	-5.39314	6.37038	-0.85	0.3975
hdipl02	1	-45.01388	11.93042	-3.77	0.0002
hdipl5	1	25.66931	8.56069	3.00	0.0028
hdipl6	1	50.44090	7.67687	6.57	<.0001

a. Comparez le R^2 obtenu dans ce dernier modèle à celui du modèle linéaire simple de la **Question 4.1**. Comment expliquez-vous cette évolution ?

b. Pourquoi ne pas avoir intégré au modèle la variable indicatrice `hdipl34` ? Ce choix de modalité de référence vous paraît-il judicieux ?

c. Interprétez la valeur de $\hat{\beta}_3$, $\hat{\beta}_4$ et $\hat{\beta}_5$. Ces coefficients peuvent-ils être considérés comme statistiquement significatifs aux seuils statistiques usuels ?

Correction

a. Le R^2 de cette régression est 0,1239 : le modèle rend compte de 12,39 % de la variance du score synthétique en mathématiques. C'est beaucoup mieux que dans le premier modèle (des variables pertinentes ont été ajoutées à la modélisation), mais cela reste relativement peu satisfaisant (même dans une perspective purement explicative on peut souhaiter obtenir un R^2 supérieur à 30 %).

b. La variable `hdipl` est une variable qualitative : elle est donc intégrée dans le modèle par le biais des variables indicatrices qui lui correspondent. Néanmoins, si les quatre variables étaient intégrées conjointement, on serait dans une situation de colinéarité parfaite avec la constante du modèle. Une des variables indicatrices est donc omise du modèle : c'est la modalité de référence.

   Ici la modalité de référence correspond aux diplômes équivalents au baccalauréat. C'est une modalité relativement centrale qui permet à ce titre de distinguer les écarts avec des niveaux de diplôme inférieurs et supérieurs. Ce choix de modalité de référence est à cet égard judicieux.

c. L'interprétation des coefficients relatifs aux différentes modalités d'une variable qualitative est à effectuer par rapport à la modalité de référence :
   - $\hat{\beta}_3 = -45,01$: à temps de travail personnel en mathématiques et sexe égaux, le fait pour un individu d'avoir des parents dont le plus haut diplôme est inférieur au baccalauréat est associé en moyenne à un score synthétique en mathématiques plus faible de 45,01 points par rapport à un individu dont les parents ont comme plus haut diplôme un diplôme exactement équivalent au baccalauréat.
   - $\hat{\beta}_4 = 25,67$: à temps de travail personnel en mathématiques et sexe égaux, le fait pour un individu d'avoir des parents dont le plus haut diplôme est équivalent à BAC + 2 est associé en moyenne à un score synthétique en mathématiques plus élevé de 25,67 points par rapport à un individu dont les parents ont comme plus haut diplôme un diplôme exactement équivalent au baccalauréat.
   - $\hat{\beta}_5 = 50,44$: à temps de travail personnel en mathématiques et sexe égaux, le fait pour un individu d'avoir des parents dont le plus haut diplôme est supérieur à BAC + 2 est associé en moyenne à un score synthétique en mathématiques plus élevé de 50,44 points par rapport à un individu dont les parents ont comme plus haut diplôme un diplôme exactement équivalent au baccalauréat.

Session 5 : Régression logistique

Dans cette session, on examine certaines variables susceptibles d'influencer le retard scolaire : sexe, diplôme des parents, conditions de vie. Les conditions de vie sont abordées à travers les variables `chambre`, `bureau`, `ordi` et `manuel` qui indiquent respectivement si la personne interrogée dispose d'une chambre individuelle, d'un bureau, d'un ordinateur et de manuels scolaires. On estime ainsi le modèle :

$$\begin{aligned} \text{retard} &= \beta_0 + \beta_1 \times \text{femme} + \beta_2 \times \text{hdipl02} + \beta_3 \times \text{hdipl5} + \beta_4 \times \text{hdipl6} \\ &\quad + \beta_5 \times \text{chambre} + \beta_6 \times \text{bureau} + \beta_7 \times \text{ordi} + \beta_8 \times \text{manuel} + u \end{aligned}$$

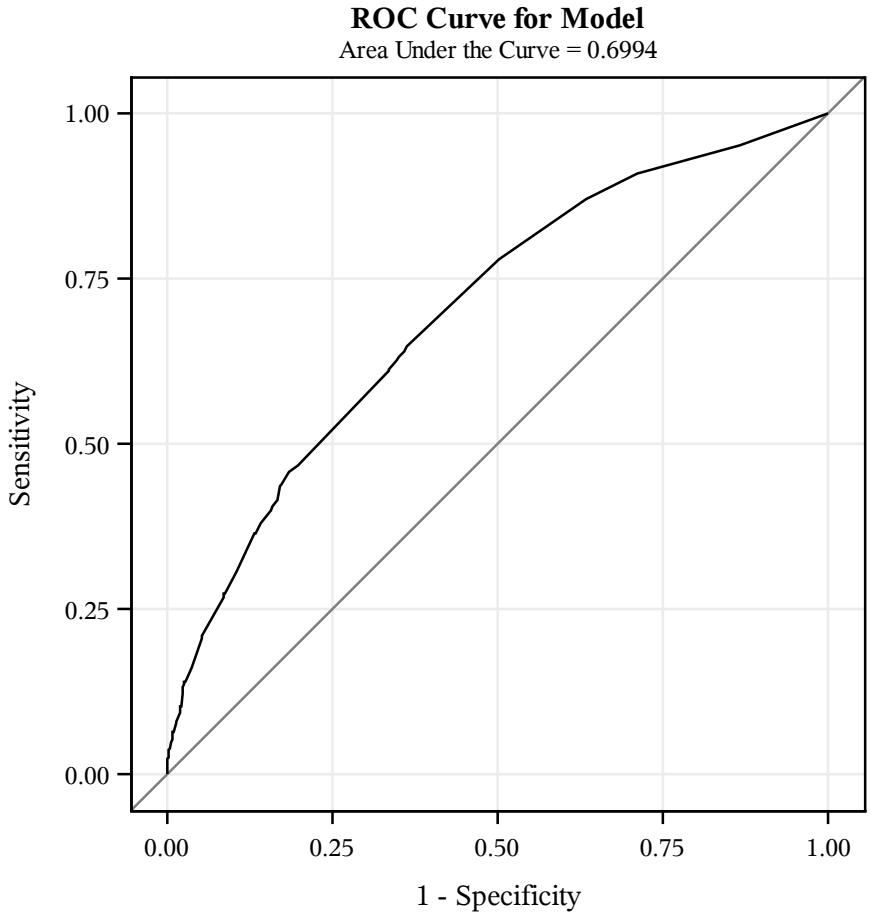
Question 5.1 Indicateurs de qualité du modèle Les indicateurs de qualité du modèle sont les suivants :

Number of Observations Read	1368
Number of Observations Used	1329

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1601.784	1477.287
SC	1606.977	1524.017
-2 Log L	1599.784	1459.287

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	140.4972	8	<.0001
Score	144.1774	8	<.0001
Wald	121.5095	8	<.0001

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	66.3	Somers' D	0.399
Percent Discordant	26.4	Gamma	0.430
Percent Tied	7.3	Tau-a	0.164
Pairs	363440	c	0.699



a. Comparez la valeur des différents indicateurs construits à partir de la log-vraisemblance. Êtes-vous en mesure de recalculez l'AIC et le SC à partir de -2 Log L ?

b. Identifiez la statistique du test de significativité globale par le ratio de vraisemblance. Êtes-vous en mesure de la recalculez à partir des autres informations de la sortie ? Interprétez ce test.

c. Quel est le pourcentage de concordance ? Vous paraît-il élevé ? Interprétez un des points de la courbe ROC. Que pensez-vous de son allure générale ?

Correction

a. Les trois indicateurs construits à partir de la log-vraisemblance ont des valeurs proches : 1459,287 pour -2 Log L, 1477,287 pour l'AIC et 1524,017 pour le SC. Pour chacun de ces indicateurs, plus la valeur est petite mieux c'est (quand on compare avec un autre modèle). Il est possible de recalculez l'AIC et le SC à partir des -2 Log L :

$$\text{AIC} = \text{-2 Log L} + 2(K+1) = 1459,287 + 2(8+1) = 1477,287$$

et

$$\text{SC} = \text{-2 Log L} + ln(n)(K+1) = 1459,287 + ln(1329)(8+1) = 1524,017$$

b. La statistique du test de significativité globale par le ratio de vraisemblance est 140,4972. On peut la recalculez en exploitant les deux colonnes du tableau **Model**

Fit Statistics :

$$LR = -2ln\left(\frac{L^0}{L_n}\right) = -2\ell^0 - (-2\ell_n) = 1599,784 - 1459,287 = 140,497$$

où ℓ^0 est la log-vraisemblance du modèle ne contenant que la constante (que l'on lit dans la colonne **Intercept Only**). La p-valeur du test est inférieure à 1 % donc on peut rejeter l'hypothèse de nullité globale de tous les coefficients au seuil de 1 % : le modèle est globalement explicatif.

c. Le pourcentage de concordance est 66,3, ce qui est relativement élevé. On interprète le point d'abscisse 0,25 et d'ordonnée approximative 0,52. Avec un taux de 75 % de négatifs classés négatifs (spécificité) le modèle parvient à classer en positif 52 % des individus effectivement positifs (sensitivité). De manière générale, le pouvoir prédictif du modèle est assez peu satisfaisant. L'aire sous la courbe est assez faible, de l'ordre de 0,6994.

Question 5.2 Interprétation des coefficients Ce modèle conduit à l'estimation des coefficients suivants :

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	2.8617	0.6259	20.9012	<.0001
femme	1	-0.3969	0.1305	9.2557	0.0023
hdipl02	1	0.8532	0.2099	16.5179	<.0001
hdipl5	1	-0.6175	0.1774	12.1178	0.0005
hdipl6	1	-0.5306	0.1606	10.9106	0.0010
chambre	1	-0.9312	0.1908	23.8313	<.0001
bureau	1	-0.5728	0.4541	1.5911	0.2072
ordi	1	-1.3739	0.4042	11.5540	0.0007
manuel	1	-0.7896	0.1777	19.7335	<.0001

a. Rappelez l'hypothèse nulle et l'hypothèse alternative du test de significativité des coefficients (par exemple pour β_1 associé à la variable `femme`). Quelle est la statistique de test et quelle loi suit-elle sous H_0 ? Menez le test au seuil de 5 %. Interprétez également la p-valeur.

b. Quelles sont les variables significatives aux seuils statistiques usuels ?

Correction

a. Le test de significativité de β_1 s'écrit :

$$H_0 : \beta_1 = 0 \quad \text{contre} \quad H_1 : \beta_1 \neq 0$$

Sous H_0, on sait que

$$z_{\beta_1} = \left(\frac{\hat{\beta}_1}{ase(\hat{\beta}_1)} \right)^2 \xrightarrow[n \to +\infty]{\mathcal{L}} \chi^2_1$$

où $ase(\hat{\beta}_1)$ est l'erreur standard asymptotique de β_1 (colonne **Standard Error** des sorties). Dans le cas de β_1, on a donc :

$$z_{\beta_1} = \left(\frac{-0,3969}{0,1305} \right)^2 = 9,25$$

Cette valeur est directement lisible dans la colonne **Wald Chi-Square**.

D'après la table du χ^2 en annexe, le quantile à 95 % d'une loi du χ^2 à 1 degré de liberté est 3,84. Ainsi la statistique de test a une valeur supérieure à la valeur critique à 95 % du test : on peut donc rejeter l'hypothèse de nullité de β_1 au seuil de 5 %. Concrètement, cela revient à affirmer que $\hat{\beta}_1$ est significativement différent de 0 au seuil de 5 %.

Par ailleurs, la p-valeur associée au coefficient β_1 vaut 0,0023. 0,0023 < 0,05 donc on retrouve bien que l'on peut rejeter l'hypothèse nulle au seuil de 5 %. Comme 0,0023 < 0,01, on peut également rejeter l'hypothèse nulle au seuil plus exigeant de 1 %.

b. En interprétant directement les p-valeurs, on constae que toutes les variables du modèle si ce n'est `bureau` sont significatives au seuil de 1 % : la p-valeur correspondant à leur test de significativité est systématiquement inférieure à 0,01.

Question 5.3 Interprétation des *odds-ratio* Le modèle produit enfin le tableau d'*odds-ratio* suivant :

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
femme	0.672	0.521	0.868
hdipl02	2.347	1.555	3.542
hdipl5	0.539	0.381	0.764
hdipl6	0.588	0.429	0.806
chambre	0.394	0.271	0.573
bureau	0.564	0.232	1.373
ordi	0.253	0.115	0.559
manuel	0.454	0.320	0.643

a. Interprétez la valeur de l'*odds-ratio* associé à la variable `hdipl02` (référez-vous à la question 1.6 pour connaître la signification des modalités de la variable `hdipl`). Pouvez-vous déterminer si l'association entre faible diplôme des parents et retard scolaire est statistiquement significative à partir de ce tableau ?

b. Interprétez la valeur de l'*odds-ratio* associé au fait que l'individu dispose d'un bureau. L'association avec le retard scolaire est-elle significative ?

Correction

a. L'*odds-ratio* associé à la variable `hdipl02` vaut 2,347. La modalité `0-2` de la variable `hdipl` correspond aux individus dont les parents ont un niveau d'étude inférieur au bac. La modalité de référence de la variable `hdipl` dans le modèle est la modalité `3-4` qui correspond aux individus dont les parents ont un niveau d'étude équivalent au bac.

On interprète donc cet *odds-ratio* de la façon suivante : toutes les autres variables du modèle contrôlées par ailleurs, les individus dont les parents ont un niveau d'étude inférieur au bac ont 2,347 fois plus de chances d'être en retard scolaire que les individus dont les parents ont un niveau d'étude équivalent au bac.

Ce tableau ne comporte ni de p-valeur ni de statistique de test. En revanche y figure l'intervalle de confiance de l'*odds-ratio* : si 1 appartient à son intervalle de confiance à 95 %, alors on ne peut pas considérer qu'un *odds-ratio* est significatif au seuil de 5 %.

Dans le cas de `hdipl02`, l'intervalle de confiance à 95 % est [1,555 ;3,542] et ne contient donc pas 1 : comme on l'a déjà vu précédemment avec le test (qui est tout à fait équivalent), le coefficient et donc l'*odds-ratio* associés à `hdipl02` sont significatifs au seuil de 5 %.

b. L'*odds-ratio* associé à la variable `bureau` a la valeur 0,564. On pourrait l'interpréter de la façon suivante : les individus qui ont un bureau ont 0,564 fois plus de chances d'être en retard scolaire que ceux qui n'en ont pas. Cependant, afin de se ramener à un rapport des cotes supérieur à 1, on inverse la proposition : les individus qui ont un bureau ont $1/0,564 = 1,773$ fois *moins* de chances d'être en retard scolaire que ceux qui n'en ont pas.

Par ailleurs $1 \in [0,232; 1,373]$ donc on ne peut cependant pas affirmer que cet *odds-ratio* est significativement différent de 1 au seuil de 5 %.

Session 6 : Compléments

Question 6.1 Introduire une variable et son carré À partir des données de l'EEC 2012T4 (données du support de révision), on estime le modèle :

$$\text{salaire} = \beta_0 + \beta_1 \times \text{age} + \beta_2 \times \text{age}^2 + \beta_3 \times \text{femme} + u$$

qui conduit à l'estimation (les erreurs-standards sont indiquées entre parenthèses) :

$$\text{salaire} = \underset{(422)}{-1473} + \underset{(21)}{160} \times \text{age} - \underset{(0,26)}{1,71} \times \text{age}^2 - \underset{(75)}{423} \times \text{femme} + u$$

Le modèle est estimé sur 647 observations et son R^2 est 0,1409.

a. Quel est le salaire prédit par le modèle pour une femme de 30 ans ? un homme de 50 ans ? Pouvez-vous déterminer l'âge pour lequel le salaire est maximal d'après le modèle ?

b. Quelle est la différence moyenne de salaire associée à une année supplémentaire à 30 ans ? à 50 ans ?

c. Posez le test de significativité jointe des coefficients β_1 et β_2. Quelle est la statistique traditionnellement associée à ce type de test et quelle loi suit-elle sous l'hypothèse nulle ?

d. Sachant que le R^2 du modèle contraint vaut 0,0334, quelle est la valeur de la statistique de test ? Que concluez-vous ?

e. Comment indiquer à SAS de mener directement ce test ?

<u>**Correction**</u>

a. Pour déterminer le salaire prédit par le modèle pour un âge et un sexe donnés, il suffit de remplacer les variables par leur valeur :
   - femme de 30 ans : $-1473 + 160 \times 30 - 1,71 \times 30^2 - 423 \times 1 = 1365$
   - homme de 50 ans : $-1473 + 160 \times 50 - 1,71 \times 50^2 - 423 \times 0 = 2252$

   La valeur maximale du salaire pour le modèle est atteinte pour l'âge qui conduit au maximum de la fonction

   $$\text{age} \mapsto -1,71 \times \text{age}^2 + 160 \times \text{age}$$

   c'est-à-dire $160/(2 \times 1,71) = 47$ ans.

b. La différence de salaire associée à une année supplémentaire correspond à :

   $$\frac{\delta \text{salaire}}{\delta \text{age}} = \beta_1 + 2\beta_2 \text{age}$$

   - à 30 ans : une année supplémentaire est associée à un salaire mensuel supérieur en moyenne de $160 - 2 \times 1,71 \times 30 = 57$ € ;
   - à 50 ans : une année supplémentaire est associée à un salaire mensuel supérieur en moyenne de $160 - 2 \times 1,71 \times 50 = -11$ € (soit une diminution moyenne de 11 €).

c. On pose le test de sigificativité jointe de β_1 et β_2 :

$$H_0 : \beta_1 = 0 \text{ et } \beta_2 = 0 \quad \text{contre} \quad \beta_1 \neq 0 \text{ ou } \beta_2 \neq 0$$

Sous H_0, on sait que la statistique

$$F = \frac{(R^2_{nc} - R^2_c)/q}{(1 - R^2_{nc})/(n - (K+1))} \hookrightarrow F_{q,n-(K+1)}$$

où R^2_c est le R^2 du modèle contraint (celui qui est estimé sous H_0), R^2_{nc} le R^2 du modèle non-contraint (le modèle d'origine), q le nombre de contraintes, n le nombre d'observations et K le nombre total de variables du modèle non-contraint.

Dans le cas présent, le modèle contraint est le modèle estimé sous l'hypothèse que $\beta_1 = \beta_2 = 0$, c'est-à-dire :

$$\text{salaire} = \beta_0 + \beta_3 \times \text{femme} + u$$

d. Le test comporte 2 contraintes ($q = 2$), le modèle non-contraint comprend 3 variables ($K = 3$) et les estimations portent sur 647 observations ($n = 647$). La statistique de test est donc calculée de la façon suivante :

$$F = \frac{(0,1409 - 0,0334)/2}{(1 - 0,1409)/(647 - (3+1))} = 40,23$$

Le test étant unilatéral, sa valeur critique à 95 % est la valeur du quantile à 95 % d'une loi de Fisher à $q = 2$ et $n - (K+1) = 643$ degrés de liberté soit $q^{F_{2,643}}_{0,95} \approx 4,70$ d'après la table en annexe.

La valeur de la statistique de test (40,23) excède largement la valeur critique du test à 95 % : on peut donc srejeter l'hypothèse nulle au seuil de 5 %. Autrement dit on peut affirmer que la variable d'âge est statistiquement liée à la variable de salaire avec un risque d'erreur inférieur à 5 %.

e. Pour mener ce test directement dans SAS, il suffit d'utiliser l'instruction `TEST` de la **`PROC REG`** :

```
PROC REG DATA = ee12t4;
    MODEL salred = sexe2 age age2;
    TEST age = age2 = 0;
RUN;
```

Test 1 Results for Dependent Variable SALRED				
Source	**DF**	**Mean Square**	**F Value**	**Pr > F**
Numerator	2	36357827	40.25	<.0001
Denominator	643	903274		

L'écart de la statistique de test calculée « à la main » par rapport à la valeur calculée par SAS est dû aux arrondis dans la valeur des R^2.

Question 6.2 Introduire la variable explicative ou une variable expliquée en logarithme On estime plusieurs spécifications de modèles de régression linéaire simple faisant intervenir la variable expliquée salaire et la variable explicative age. Dans chaque cas, interprétez le coefficient de age.

a. $\text{salaire} = 984 + 19 \times \text{age} + u$

b. $ln(\text{salaire}) = 6,77 + 0,01 \times \text{age} + u$

c. $ln(\text{salaire}) = 5,05 + 0,61 \times ln(\text{age}) + u$

Correction

a. La variable expliquée et la variable explicative sont en niveau (pas de logarithme) et $\hat{\beta}_1 = 19$ donc on interprète de la façon suivante : en moyenne, être âgé d'un an de plus est associé à un salaire mensuel supérieur de l'ordre de 19 euros.

b. La variable expliquée est en logartihme et la variable explicative en niveau et $\hat{\beta}_1 = 0,01$ donc on interprète de la façon suivante : en moyenne, être âgé d'un an de plus est associé à un salaire mensuel supérieur de l'ordre de 1 %.

c. La variable expliquée et la variable explicative sont en logartihme donc on interprète de la façon suivante : en moyenne, avoir un âge de 1 % supérieur est associé à un salaire mensuel supérieur de l'ordre de 0,61 %.

   Remarque : Très étrange dans le cas de l'âge, cette formulation fait davantage sens avec des grandeurs économiques : élasticité de la consommation au revenu, etc.

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.

Table 4 : Quantiles de la loi de Fisher

Le quantile de niveau γ d'une variable aléatoire X suivant une loi de Fisher à q et p degrés de liberté noté $q_\gamma^{F_{q,p}}$ est défini par :

$$F_X(q_\gamma^{F_{q,p}}) = \mathbb{P}(X \leq q_\gamma^{F_{q,p}}) = \gamma$$

où F_X est la fonction de répartition de X.

On présente les quantiles à 0,95 et 0,99 d'une loi de Fisher pour les degrés de liberté q et p usuels.

Quantiles de niveau $\gamma = 0,95$ d'une loi de Fisher $F_{q,p}$ à q et p degrés de liberté.

p \ q	1	2	3	4	5	6	7	8	9	10
1	161	199	216	225	230	234	237	239	241	242
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,03
100	3,94	3,09	2,70	2,46	2,31	2,19	2,10	2,03	1,97	1,93
1000	3,85	3,00	2,61	2,38	2,22	2,11	2,02	1,95	1,89	1,84

Lecture : 95% des valeurs d'une variable aléatoire suivant une loi de Fisher $F_{1,100}$ inférieures à 3,94.

Quantiles de niveau $\gamma = 0,99$ d'une loi de Fisher $F_{q,p}$ à q et p degrés de liberté.

p \ q	1	2	3	4	5	6	7	8	9	10
1	4052	4999	5403	5625	5764	5859	5928	5981	6022	6056
2	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39	99,40
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80
50	7,17	5,06	4,20	3,72	3,41	3,19	3,02	2,89	2,78	2,70
100	6,90	4,82	3,98	3,51	3,21	2,99	2,82	2,69	2,59	2,50
1000	6,66	4,63	3,80	3,34	3,04	2,82	2,66	2,53	2,43	2,34

Lecture : 99% des valeurs d'une variable aléatoire suivant une loi de Fisher $F_{1,100}$ sont inférieures à 6,90.