Statistiques et Séries Chronologiques,
Probabilités et processus aléatoires
Université de Lorraine

Enoncés des exercices de TP et TD

Clément Dell'Aiera

Table des matières

Première partie

Statistiques et séries chronologiques

Chapitre 1

Régression linéaire

La régression linéaire sous R

Ce numéro rappelle les notions nécessaires à l'interprétation d'une sortie R de la fonction *lm*. Dans toute la suite, *regression* désigne un objet de type *lm* que l'on a appellé grâce à $lm(Y \sim X_1 + ... + X_n, data = ...)$. Si l'on note X_{-i}, le moins signifie que l'on calcule la quantité X sans tenir compte de l'observation i.

Dans le modèle de régression

$$Y = X\beta + \epsilon,$$

la commande *summary(regression)*, où *regression* est un objet de la classe *lm*, renvoie plusieurs tableaux.
On note $\hat{y}_j = \sum h_{ij} x_j$, i.e.

$$h_{ij} = \frac{1}{n} + \sum \frac{(x_i - \hat{x})(x_j - \hat{x})}{\sum (x_j - \hat{x})^2}.$$

Le premier tableau,*Residuals*, est destiné à donner une idée de la répartition des résidus en affichant les quantiles. Je vous recommande d'afficher tout de même les résidus, et d'observer leur distribution. Mieux, vous pouvez utiliser les résidus studentisés. A priori, bien que tous soient centrés, les résidus n'ont pas même variance (même sous hypothèse d'homoscédasticité !) : $Var[\epsilon_j] = \sigma^2(1 - h_{jj})$. Pour les rendre comparables, on pourrait les réduire, mais si l'on remplace la variance par la variance estimée, celle-ci dépend de l'information contenue dans x_j, ce qui empêche une quantification de l'effet que x_j a sur les coeffcients de la régression. Pour palier à ce problème, on introduit

$$\hat{\sigma}^2_{-j} = \frac{1}{n-3}[(n-2)\hat{\sigma}^2 - \frac{\epsilon_j}{1 - h_{jj}}]$$

qui n'est rien d'autre que la variance estimé sur le modèle où l'on a supprimé l'observation x_j. Les résidus studentisés sont définis par $T_j = \frac{\epsilon_j}{\sigma_{-j}(1-h_{jj})} \sim T(n-3)$ et suivent une loi de Student à $n-3$ degré de liberté sous des hypothèses raisonnables. Pour détecter une anomalie dans les données, on peut vérifier que les résidus studentisés se répartissent de manière uniforme sur l'intervalle $[-2; 2]$

(sous hypothèse d'homoscédasticité). Repérer des formes suspectes est un moyen facile pour repérer les valeurs aberrantes. On peut par exemple taper :
qqnorm(studres(regression)) ; qqline(studres(regression)).
Une autre méthode pour détecter les valeurs aberrantes : utiliser la distance de Cook. Elle est définie par

$$D_j = \frac{\sum_j (\hat{y}_{-i,j} - \hat{y}_j)^2}{2\hat{\sigma}^2},$$

et mesure l'influence d'une observation sur l'ensemble des prévisions (qu'on veut petite !). Encore une règle du pouce : si $D_j > 1$, on enlève l'observation j. Vous pouvez le faire automatiquement en tapant *plot(regression,which = 4)*.
Le deuxième est nommé *Coefficients* :

	Estimate	Std. Error	t-value	$Pr(>\|t\|)$
β_j	$\hat{\beta}_j$	$\hat{\sigma}_j$	$\hat{t}_j = \frac{\hat{\beta}_j}{\hat{\sigma}_j}$	p-value

Les trois premières colonnes s'expliquent elles-mêmes, mais à quoi servent les 2 dernières ? A effectuer un test de significativité. Plus précisément, on sait que $\hat{t}_j$ suit une loi de Student à $N-k$ degrés de liberté sous l'hypothèse $H_0|\beta_j = 0$ contre $H_1|\beta_j \neq 0$. La quatrième colonne donne donc la valeur de cette statistique, et la dernière sa p-value, définie comme la valeur seuil de confiance α qui fait basculer le test. (Rappelez vous que, mécaniquement, si α diminue assez, on finit par accepter H_0.) Une règle appliquée par les statisticiens est la suivante :

$p < 0.01$	suspicion très forte contre H_0
$0.01 - 0.05$	suspicion forte contre H_0
$0.05 - 0.1$	suspicion faible contre H_0
> 0.1	peu ou pas de suspicion contre H_0

Donc, si $Pr(>|t|)$ est petit, on rejette l'hypothèse $\beta_j = 0$, ce qui signifie que le coefficient est significatif. R ajoute même des petites étoiles à côté des coefficients les plus significatifs.

Reste encore à observer plusieurs indicateurs. L'erreur *Residual standard error* est calculée comme un estimateur de σ sous l'hypothèse de matrice variance-covariance égale à $\sigma^2 I_n$. Il nous reste encore le R^2 et le R^2 ajusté. Rappelons que le coefficient R^2 peut s'interpréter comme le cosinus de l'angle entre le vecteur des observations Y_j et son projeté pour la norme $\mathcal{L}^2$ sur l'espace linéaire engendré par les observations X_j, soit

$$R^2 = \frac{\sum(\hat{y}_j - \overline{y})^2}{\sum(y_j - \overline{y})^2}.$$

Cet indicateur à des valeurs comprises entre 0 et 1, la proximité avec 1 indiquant une bonne adéquation du modèle aux données. Toutefois, son interprétation est sujette à caution : sa valeur augmente mécaniquement avec l'ajout de variables explicatives. En particulier, pour comparer la qualité de deux modèles au nombre de variables explicatives distinct, on lui préférera le R^2 ajusté, qui prend en compte ce nombre noté k :

$$RR^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}.$$

1.1 Exercice 1

1. Que font les commandes *pnorm*, *qnorm* , *rnorm*, *dnorm* ? Utilisez l'aide de *R*.
2. Tracer la fonction de densité de la loi normale. Faites varier les paramètres et afficher les différentes courbes sur le même graphique.
3. Simuler 2 vecteurs X et Y contenant chacun $N = 100$ variables indépendantes identiquement distribuées suivant une loi normale $\mathcal{N}(0, 1)$.
4. Afficher les points de coordonnées $(X[j], Y[j])$ dans le plan, pour j allant de 1 à 100.
5. Tracer la fonction de répartition empirique des $X[j]$.
6. Soit $\mathcal{E}$ une v.a. de loi exponentielle de paramètre 1, et U une v.a. suivant une loi uniforme sur $[0, 2\pi]$. On pose

$$(X, Y) = (\sqrt{\mathcal{E}} \cos(U), \sqrt{\mathcal{E}} \sin(U)).$$

   Quelle est la loi du couple (X, Y) ? (Vous pouvez le prouvez, ou observer grâce à *R* ce qu'il se passe en simulant ces variables et en les traçant.)

1.2 Exercice 2

1. Une table est déjà en mémoire dans *R* : la table *stackloss*. Analyser la rapidement.
2. Tracer *stack.loss* en fonction de *Air.Flow*. Qu'en pensez-vous ?
3. Effectuer la régression linéaire de *stack.loss* en fonction des autres variables. Quelles sont celles qui sont significatives ?

1.3 Exercice 3

Le fichier *ozone.dta* contient les variables suivantes, pour une série de journées (qui sont ici nos individus) :

— l'identifiant de la journée,
— le maximum d'ozone (variable maxO3)
— l'heure à laquelle le maximum d'ozone a été obtenu (heure),
— les températures à 6h, 9h, 12h, 15h, 18h (resp. T6 à T18)
— la nébulosité à 6h, 9h, 12h, 15h, 18h (resp. Ne6 à Ne18)
— la projection du vent sur l'axe est-ouest à 12h (Vx),
— le maximum d'ozone de la veille (maxO3v).

Le but est de modéliser la valeur des pics d'ozone en fonction de grandeurs physiques facilement mesurables (température, heure, nébulosité, vent) afin d'avoir des approximations de la qualité de l'air faciles et rapides à obtenir.

1. Importer la table, et afficher un résumé de ce qu'elle contient.
2. Tracer *maxO3* en fonction de *T12*, puis effectuer un régression linéaire. Ajouter la droite de régression sur le graphique. Soignez la présentation.
3. Afficher les résultats de la régression.

4. Extraire les résidus et tracer leur densité estimée.
5. Effectuer la régression de *maxO3* sur toutes les variables, et supprimer récursivement celles qui ne sont pas significatives, jusqu'à ce qu'elles le soient toutes.

1.4 Exercice 4

1. Sur le site *data.gouv.fr*, vous pourrez trouver des tables de données publiques en libre accès. Choisissez un thème qui vous intéresse, puis une table en conséquence. Télécharger là.
2. Les tables sont souvent au format *.xls* : vous aurez besoin d'installer un package pour pouvoir les lire. La commande pour ce faire est *install.packages("nom du package")*. Installer le package *gdata*.
3. Analyser votre table.

1.5 Exercice 5

1. Télécharger et installer le package *ISwR* (Introductory Statistics with R).
2. Utiliser la commande *summary* pour analyser rapidement la table *bp.obese*. L'échantillon provient d'un échantillon de population mexicaine en Californie, et la table décrit 3 variables : le sexe (femme = 1, homme =0), le ratio d'obésité (*obese*) et la pression sanguine systolique en *mm* de mercure (*bp*).
3. Représenter les données dans un graphe, en utilisant des symboles différents pour les hommes et les femmes.
4. Expliquer la pression sanguine en fonction du ration d'obésité, puis du ratio d'obésité et du sexe.
5. Tracer sur un même graphe les courbes correspondant aux régressions dans les 2 modèles. Soignez la présentation (couleurs différentes, légende,...)

1.6 Exercice 6

1. Télécharger la librairie *MASS*.
2. Analyser rapidement la table *cats* et afficher les variables les unes en fonctions des autres par paires.
3. Effectuer une régression linéaire selon le modèle $Hwt \sim Bwt * Sex$. Cela apporte-t-il quelque-chose par rapport au modèle $Hwt \sim Bwt + Sex$?
4. Visualiser les composantes de votre régression.
5. En extraire les prédiction, les coefficients, les résidus, lés résidus studentisés, et la formule du modèle.
6. Tracer le *qqplot* des résidus studentisés ainsi que la première bissectrice.
7. Tracer le graphe des résidus contre les prédictions.
8. Tracer le graphe des distance de Cook.
9. Observer les attributs que vous donne *summary*

10. Afficher le R^2 ajusté de la régression, le nombre de degrés de liberté résiduels, la matrice de variance-covariance des paramètres estimés.

Chapitre 2

Tests statistiques

2.1 Théorie des tests

2.1.1 Exemples

1. Une entrerise vend des biens dont elle assure que la durée de vie dépasse 10000 heures. Vous êtes engagés pour vérifier la qualité d'iceux. Vous disposez d'un échantillon de 30 de ces biens. La durée de vie moyenne calculée sur l'échantillon vaut 9900 heures, on suppose que l'écart-type est connu et vaut 120 heures. Pouvez-vous rejeter leur assertion à à un niveau de confiance de 0.05%.
   Répondez à la même question si l'écart-type n'est plus connu, et que l'écart-type obtenu sur l'échantillon vaut 125 heures.
   Calculez la puissance du test, c'est-à-dire la probabilité de l'erreur de seconde espèce.
2. Une firme agroalimentaire assure qu'un cookie qu'elle produit ne contient pas plus de 2 grammes d'un certain composé (graisse, colorant,...). Vous achetez un paquet, contentant 35 cookiees, et mesurez une teneur moyenne de 2.1 grammes. En supposant que l'écart-type de l'échantillon est de 0.25, pouvez-vous incriminez la firme à un niveau de confiance de 0.05%. Même question si l'on ne connaît que l'écart-type empirique de 0.3. Calculez la puissance du test.
3. Lors des dernières élections, les médias affirment qu'au moins 60% des citoyens ont voté. Vous interrogez 148 citoyens de façon à obtenir un échantillon représentatif de la population (vous êtes statisticien après tout). Vous obtenez que 85 des personnes interrogées ont voté. A 0.05%, votre test concorde-t-il avec l'affirmation des médias ?

2.1.2 Analyse of Variance ou procédure ANOVA

1. Une institution de santé publique veut comparer l'effet de trois traitement contre la grippe. Pour cela, 18 hopitaux sont choisis de façon aléatoire, répartis par groupes de 6, chacun appliquant un et un seul des trois traitement. Voici le nombre de personnes guéries au bout d'une semaine de traitement :

Trait. 1	Trait. 2	Trait. 3
22	52	16
42	33	24
44	8	19
52	47	18
45	43	34
37	32	39

(a) Une méthode pour rentrer les données dans R : les taper dans un fichier *.txt* puis utiliser *read.table* pour créer un objet de la classe *data frame.* (Faîtes-le)

(b) Utiliser un test ANOVA pour répondre à la problématique de l'institution.

(c) Un pays frontalier, lui aussi touché par l'épidémie, décide de répliquer l'expérience avec le même nombre d'hôpitaux et les mêmes traitements. Chaque hôpital est selectionné de façon aléatoire, et dois appliquer les trois traitements pendant trois semaines, chaque traitement pendant une semaine, l'ordre des traitements étant lui aussi aléatoire. Voici le résultat :

Trait. 1	Trait. 2	Trait. 3
22	52	16
42	33	24
44	8	19
52	47	18
45	43	34
37	32	39

2.2 Régression sur variables qualitatives

Le but de cet exercice est d'expliquer la concentration en ozone *O3* en fonction de la température *T12* et de la direction du vent *vent* dans la table *ozone.txt.*

1. Télécharger la table, et effectuer des régressions selon les différents modèles.
2. Tester l'égalité des pentes.
3. Tester l'égalité des ordonnées à l'origine.
4. Analyser les résidus.

2.2.1 ANOVA à 1 facteur

Nous souhaitons modéliser la concentration en ozone en fonction de la direction du vent.

1. Tracer une boîte à moustaches de la variable $O3$ par rapport aux quatres modalités de la variable *vent.* Le vent semble-t-il avoir une influence sur la concentration en ozone ?
2. On se place dans un modèle d'analyse de la variance à un facteur

$$y_{ij} = \mu + \alpha_j + \epsilon_{ij}$$

(a) Effectuer la regression linéaire de *O3* sur *vent* sous la contrainte $\mu = 0$.
(b) Effectuer la regression linéaire de *O3* sur *vent* sous la contrainte $\alpha_1 = 0$.
(c) Effectuer la regression linéaire de *O3* sur *vent* sous la contrainte $\sum n_i \alpha_i = 0$.
(d) Effectuer la regression linéaire de *O3* sur *vent* sous la contrainte $\sum n_i \alpha_i = 0$.

3. Analyser les résidus afin de constater que l'hypothèse d'homoscédasticité est vérifiée. Pour cela, tracer un boxplot des résidus en fonction de *vent*, les résidus en fonction de $\hat{O}3$, leurs quantiles théoriques ainsi que la distribution des résidus par modalité de *vent*.

2.2.2 ANOVA à 2 facteurs

Nous voulons maintenant modéliser la concentration en ozone par le vent et la nébulosité, variable à 2 modalités : SOLEIL et NUAGEUX.

1. Procéder à un examen graphique qui puisse déterminer si l'interaction des facteurs influe sur la variable à expliquer. (voir ce qu'est un *profil*)
2. On suppose la gaussianité des résidus.
   (a) Tester le modèle avec interaction : **mod1**.
   (b) Tester le modèle sans interaction : **mod2**.
   (c) Tester le modèle sans effet du facteur *nebulosité* : **mod3**.
3. Grâce à la commande ANOVA de R, effectuer des analyses de la variance entre les modèles **mod1**, **mod2** et **mod3**.
4. Répondez à la problématique.

Chapitre 3

Partiel

3.1 Follicule de rose

Cet exercice est fortement inspiré d'un exercice du livre *Initiation à la statistique avec R* de F. Bertrand et M. Maumy-Bertrand aux édition Dunod. Vous renderez une copie sur laquelle chaque question sera détaillée, et vous enverrez votre fichier *.R* à mon adresse mail ainsi que tous vos graphiques. Votre fichier *.R* devra être nommé *NomPrénom.R*. Lorsque l'on demande de donner des conclusions, cela signifie que l'on s'attend à une **phrase rédigée** sur la copie, ainsi que du code **commenté** dans le fichier *.R*. De plus, les graphiques et autres tableaux ne peuvent qu'être bénéfiques à la compréhension, et donc à la notation, de votre travail.

Cet énoncé s'intéresse à quantifier l'effet de la taille des follicules de roses sur leurs masse, ainsi que l'effet possible de l'espèce. Dans la première partie, on ne s'intéresse qu'à la relation entre taille et masse, la seconde essaie d'incorporer l'espèce (qui est une variable qualitative).

1. Télécharger le package *BioStatR* et importer la librairie ainsi que la table *Mesures*. Décrire rapidement la table *Mesures*. (Nombre de variables, principales caractéristiques des variables,etc.)
2. Tracer le nuage de points de la taille en fonction de la masse des follicules de roses.
3. Pouvez-vous soupçonner une relation linéaire entre ces deux variables ?
4. Effectuer la régression linéaire de la taille sur la masse. Tracer la droite de régression sur le graphique. Soignez la présentation (titre, couleurs,etc).
5. Donner un intervalle de confiance pour les coefficients du modèle que vous obtenez. Comment avez-vous trouvé ces informations ? En bonus, vous pouvez tracer la zone de confiance pour la droite de régression sur votre graphique.
6. Donnez une prédiction plausible de la taille d'un follicule si sa masse est de 42 grammes.
7. Comment appellez-vous l'écart entre la valeur observée d'une observation et celle prédit par le modèle ?

8. La droite des moindres carrés passe-t-elle par le point moyen $(\overline{x}_n, \overline{y}_n)$? Si oui, démontrer-le.
9. Donner le coefficient de détermination R^2. A quoi sert ce coefficient ? Quelle est la différence entre le R^2 et le R^2 ajusté ?
10. Donner la part de variance expliquée et totale.
11. Donnez la valeur de l'estimation de σ^2.
12. Effectuer les deux tests suivants au risque 0.05% :
    — $H_0 \mid \beta_0 = 0$ contre $H_1 \mid \beta_0 \neq 0$.
    — $H_0 \mid \beta_1 = 0$ contre $H_1 \mid \beta_1 \neq 0$.

Cette partie vise à quantifier l'effet de l'espèce sur la taille du follicule.

1. Effectuer une régression croisée qui tient compte de la variable *espece*. Donner une synthèse rapide du résultat.
2. Comment pourriez-vous tester si l'espèce à un effet sur la taille ?
3. Proposer une méthode, le mettre en oeuvre et donnez vos conclusions.

3.2 Problème du voyageur de commerce et algorithme du recuit simulé

D'après le livre de Michel Benaïm et Nicole El Karoui, *Promenade aléatoire, Chaîne de Markov et simulations, martingales et stratégies*, exemple 3.1.8.

La méthode du recuit simulé est un algorithme d'optimisation proche de celui de Metropolis, et consiste à se promener aléatoirement sur l'espace (fini) des états d'un système selon une loi construite de façon à ce que la promenade converge vers un état qui minimise une certaine fonctionnelle.

On se donne une fonction $h :]0; \infty[\to]0; 1]$ telle que

$$h(x) = xh(\frac{1}{x}),$$

par exemple $min(1, x)$ ou bien $\frac{x}{1+x}$. La fonction $V : E \to \mathbb{R}_+$ est la fonction "coût" à minimiser.

La terminologie "recuit simulé" vient d'une technique métallurgique consistant à faire fondre de façon répétée le métal puis à le faire lentement refroidir pour en améliorer les propriétés. En effet, on va se donner un schéma de décroissance d'une quantité analogue à la température, que nous noterons T_n. Ce schéma est crucial pour que l'algorithme converge vers un minimum global de la fonction V et ne reste pas piégé dans un de ses minima locaux. Vous pourrez choisir l'un des schémas suivants :

— décroissance logarithmique :

$$T_n = \frac{C}{\log(n)}$$

— recuit par palier :

$$T_n = \frac{1}{k} \text{ pour } e^{(k-1)C} \leq n < e^{kC}$$

Voici l'algorithme :

Initialiser X_0. Choisir le nombre de pas de la marche aléatoire m. Pour n allant de 1 à $m-1$, répéter :

1. Choisir un voisin y de X_n aléatoirement.
2. Tirer $U \sim \mathcal{U}_{[0,1]}$.
3. Si $U < h(\exp(\frac{1}{T_n})(V(X_n) - V(y))\frac{N(y)}{N(X_n)})$, accepter $X_{n+1} = y$, sinon refuser i.e. $X_{n+1} = X_n$.

Le problème auquel nous allons appliquer cet algorithme est celui d'un commerçant devant visiter un ensemble fini $E = X_1, ..., X_N$ de villes, une et une seule fois. Pour minimiser son temps et son argent, il souhait trouver le chemin l qui minimise la distance, soit, avec nos notations :

$$V(l) = \sum_{j=1}^{N-1} d(X_{l(j)}, X_{l(j+1)}),$$

où l'on voit un chemin comme une permutation de l'ensemble E. Nous travaillerons avec des villes disposées aléatoirement dans le carré $[0;1] \times [0;1]$.

1. Créer une fonction qui calcule le coût d'un chemin donné l.
2. Créer une fonction qui affiche un chemin donné l.
3. Choisir une loi de transition sur les chemins, et l'implémenter.
4. Implémenter l'algorithme du recuit simulé sur ce problème, à l'aide d'une fonction si possible. Afficher le chemin obtenu, ainsi que l'évolution de la longueur du chemin en fonction du nombre d'itérations . Qu'en pensez vous ? De combien d'itérations avez-vous besoin pour obtenir un chemin plausible ?

Deuxième partie

Probabilités et processus aléatoires

Chapitre 4

Estimation

4.1 Généralités sur l'estimateur du maximum de vraisemblance

1. Rappeler les propriétés de l'EMV.
2. Soient X_j des variables exponentielles indépendantes de paramètre $\theta > 0$, non-observées, et T un instant de censure. Soit $\mathcal{E}^n$ l'expérience engendrée par l'observation du n-échantillon $X_j^* = \min\{T, X_j\}$. Donner une mesure qui domine le modèle et calculer sa vraisemblance.
3. Montrer que l'estimateur du maximum de vraisemblance ne dépend pas du choix de la mesure dominante.

4.2 Exemples de calculs de maximum de vraisemblance

Pour chaque loi, on considère un n-échantillon tiré de façon i.i.d selon cette loi. Proposer un espace des paramètres donnant un modèle identifiable. Donner une mesure dominante si possible. Calculer la vraisemblance du modèle, ainsi que la log-vraisemblance, donner les équations de vraisemblance et déterminer , s'il existe, un estimateur du maximum de vraisemblance.

1. Modèle gaussien standard, de densité par rapport à la mesure de Lebesgue
$$f_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{1}{2\sigma^2}(x-\mu)^2 \quad , \theta = (\mu, \sigma).$$
2. Modèle de Bernoulli
$$\mathbb{P}_\theta(X = 1) = 1 - \mathbb{P}(X = 0) = \theta.$$
3. Modèle de Laplace, où $\sigma > 0$ est connu, de densité par rapport à la mesure de Lebesgue
$$f_\theta(x) = \frac{1}{2\sigma} \exp\left(-\frac{|x-\theta|}{\sigma}\right).$$

4. Modèle uniforme, de densité par rapport à la mesure de Lebesgue
$$f_\theta(x) = \frac{1}{\theta} 1_{[0,\theta]}(x).$$
5. Modèle de Cauchy, de densité par rapport à la mesure de Lebesgue
$$f_\theta(x) = \frac{1}{\pi(1-(x-\theta)^2)}.$$
6. Modèle de translation. On considère la densité
$$h(x) = \frac{e^{-\frac{|x|}{2}}}{2\sqrt{2\pi|x|}}.$$
Le modèle de translation par rapport à la densité h est le modèle dominé par la mesure de Lebesgue sur $\mathbb{R}$ de densités
$$f_\theta(x) = h(x-\theta) \quad , x \in R, \theta \in \mathbb{R}.$$

4.3 Méthode des moments

1. Calculer des estimateurs des moments d'ordre 1 et 2 pour l'expérience engendrée par l'observation d'un n-échantillon de variables exponentielles de paramètre $\theta > 0$. Donner l'asymptotique des ces deux estimateurs.
2. On considère le modèle de translation associé à la famille des lois de Cauchy :
$$f_\theta(x) = \frac{1}{\pi(1+(x-\theta)^2)} \quad , x \in \mathbb{R}.$$
On note g la fonction signe, qui vaut 1 si $x > 0$, -1. Trouver un estimateur pour $\theta \mapsto \mathbb{E}[g(X_1)]$ et donner ses propriétés.

4.4 Estimation de la fonction de répartition

On se donne un n-échantillon X_1,..., X_n i.i.d suivant une loi donnée par la même fonction de répartition (f.d.r) F sur $\mathbb{R}$. $\mathcal{F}$ dénote l'ensemble des fonctions de répartition sur $\mathbb{R}$.

1. Décrire l'expérience statistique.
2. Le modèle est-il dominé ?
3. On veut estimer $F(x) = \mathbb{P}(X \leq x)$.
   (a) On pose $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n 1_{X_i \leq x}$. Calculer $\mathbb{E}[\hat{F}_n(x)]$ et $V[\hat{F}_n(x)]$.
   (b) Montrer que $\hat{F}_n(x)$ converge presque-sûrement vers $F(x)$.
   (c) Montrer que, si $l(x,y) = (x-y)^2$ est la perte quadratique,
$$\sup_{F \in \mathcal{F}} \mathbb{E}[l(\hat{F}_n(x), F(x))] = \frac{1}{4n}.$$
   (d) En déduire que $\hat{F}_n(x)$ converge uniformément en norme $\mathcal{L}^2$ vers $F(x)$, et donc en probabilité.

4. (a) Montrer que

$$\mathbb{P}(|\hat{F}_n(x) - F(x)| > t) \leq \frac{1}{t^2} Var[\hat{F}_n(x)] \leq \frac{1}{4nt^2}$$

   (b) Soit $\alpha \in]0;1[$. Déterminer

$$t_{\alpha,n} = \inf\{t > 0 \ : \ \frac{1}{4nt^2} \leq \alpha\}$$

   et en déduire un intervalle de confiance pour $F(x)$ de niveau $1-\alpha$.

   (c) Comment interpréter $I_{n,\alpha}$? Quelle est sa précision ?

5. On pose $\xi_n = \sqrt{n}\frac{\hat{F}_n(x) - F(x)}{\sqrt{\hat{F}_n(x)(1-\hat{F}_n(x))}}$.
   (a) Déterminer la limite en loi de ξ_n.
   (b) On note $J_{n,\alpha}$ l'intervalle $[-\phi^{-1}(1-\frac{\alpha}{2}); \phi^{-1}(1-\frac{\alpha}{2})]$.Calculer la limite de $\mathbb{P}(\xi_n \in J_{n,\alpha})$ lorque n tend vers ∞.
   (c) Donner un intervalle de confiance asymptotique pour $J_{n,\alpha}$, ainsi que sa précision asymptotique.

6. Soient Y_j des variables aléatoires réelles indépendantes centrées : $\mathbb{E}Y_j = 0$ et bornées : $a_j \leq Y_j \leq b_j$. On veut démontrer ce que l'on appelle l'*inégalité de Hoeffding* : pour tout $t > 0$,

$$\mathbb{P}(\sum Y_j \geq t) \leq e^{-st} \prod e^{\frac{s^2(b_j-a_j)^2}{8}} \quad \forall s > 0.$$

   On pose $\Phi_Y(s) = \log \mathbb{E}[e^{s(Y-\mathbb{E}Y)}]$.
   (a) Montrer que

$$\Phi_Y''(s) = e^{-\Phi_Y(s)}\mathbb{E}[Y^2 e^{sY}] - e^{-2\Phi_Y(s)}(\mathbb{E}[Ye^{sY}])^2.$$

   (b) On définit une nouvelle mesure de probabilité par $\mathbb{Q}(A) = e^{-\Phi_Y(s)}\mathbb{E}[e^{sY}1_A]$ pour tout borélien A. Comment interpréter $\Phi_Y''(s)$ dans ce cadre ?
   (c) Montrer alors que $\Phi_Y(s) \leq s^2\frac{(b-a)^2}{8}$.
   (d) En déduire l'inégalité de Hoeffding.

7. (a) Soient X_j des v.a. de Bernoulli de paramètre p et $\overline{X}_n = \frac{1}{n}\sum_{j=1}^n X_j$, montrer que

$$\mathbb{P}(|\overline{X}_n - p| \geq t) \leq 2e^{-2nt^2}.$$

   (b) En déduire un intervalle de confiance de niveau $1-\alpha$ pour $F(x)$.

8. Comparer les différents intervalles de confiance que vous avez obtenu.

4.5 Intervalles de confiance

1. Soient $\mu \in \mathbb{R}$, $\sigma > 0$ et $\alpha \in (0,1)$. On observe un n-échantillon $\underline{x} = (X_1, ..., X_n)$ de variables iid de loi normale $\mathcal{N}(\mu, \sigma^2)$.
   (a) Donner un intervalle de confiance de niveau α pour μ, si σ^2 est connu, puis si σ^2 est inconnu.
   (b) Donner un intervalle de confiance de niveau α pour σ^2, si μ est connu, puis si μ est inconnu.
2. On observe un n-échantillon $\underline{Y} = (Y_1, ..., Y_n)$ de variables iid suivant une loi de Bernoulli de paramètre $0 < p < 1$ inconnu. Donner un intervalle de confiance pour p au niveau α.

4.6 Estimateur du maximum de vraisemblance

On observe un n-échantillon $\underline{x} = (X_1, ..., X_n)$ de variables iid suivant une loi $\mathbb{P}_\theta$ de densité :

$$f_\theta(x) = \frac{2}{\sqrt{\pi}\theta^{3/2}} x^2 e^{-\frac{x^2}{\theta}}, \text{ pour } \theta > 0.$$

1. Décrire le modèle statistique engendré par $\underline{x}$.
2. Calculer un estimateur du maximum de vraisemblance, que l'on notera $\hat{\theta}$.
3. Examiner les qualités suivantes de $\hat{\theta}$: efficacité, biais et convergence.

On observe un n-échantillon $\underline{x} = (X_1, ..., X_n)$ de variables iid suivant une loi exponentielle de paramètre $\lambda > 0$, de densité

$$g_\lambda(x) = \lambda e^{-\lambda x} 1_{x \geq 0}.$$

1. Décrire le modèle statistique engendré par $\underline{x}$.
2. Calculer la vraismeblance du modèle.
3. Donner un estimateur du maximum de vraisemblance pour λ.

4.7 Estimation de la fonction de répartition

On se donne un n-échantillon X_1,..., X_n i.i.d suivant une loi donnée par la même fonction de répartition (f.d.r) F sur $\mathbb{R}$. $\mathcal{F}$ dénote l'ensemble des fonctions de répartition sur $\mathbb{R}$.

1. Décrire l'expérience statistique.
2. Le modèle est-il dominé ?
3. On veut estimer $F(x) = \mathbb{P}(X \leq x)$.
   (a) On pose $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n 1_{X_i \leq x}$. Calculer $\mathbb{E}[\hat{F}_n(x)]$ et $V[\hat{F}_n(x)]$.
   (b) Montrer que $\hat{F}_n(x)$ converge presque-sûrement vers $F(x)$.
   (c) Montrer que, si $l(x, y) = (x - y)^2$ est la perte quadratique,
   $$\sup_{F \in \mathcal{F}} \mathbb{E}[l(\hat{F}_n(x), F(x))] = \frac{1}{4n}.$$
   (d) En déduire que $\hat{F}_n(x)$ converge uniformément en norme $\mathcal{L}^2$ vers $F(x)$, et donc en probabilité.
4. (a) Montrer que
   $$\mathbb{P}(|\hat{F}_n(x) - F(x)| > t) \leq \frac{1}{t^2} Var[\hat{F}_n(x)] \leq \frac{1}{4nt^2}$$
   (b) Soit $\alpha \in]0; 1[$. Déterminer
   $$t_{\alpha,n} = \inf\{t > 0 \ : \ \frac{1}{4nt^2} \leq \alpha\}$$
   et en déduire un intervalle de confiance pour $F(x)$ de niveau $1 - \alpha$.

   (c) Comment interpréter $I_{n,\alpha}$? Quelle est sa précision ?

5. On pose $\xi_n = \sqrt{n}\frac{\hat{F}_n(x)-F(x)}{\sqrt{\hat{F}_n(x)(1-\hat{F}_n(x))}}$.
   (a) Déterminer la limite en loi de ξ_n.
   (b) On note $J_{n,\alpha}$ l'intervalle $[-\phi^{-1}(1-\frac{\alpha}{2});\phi^{-1}(1-\frac{\alpha}{2})]$.Calculer la limite de $\mathbb{P}(\xi_n \in J_{n,\alpha})$ lorque n tend vers ∞.
   (c) Donner un intervalle de confiance asymptotique pour $J_{n,\alpha}$, ainsi que sa précision asymptotique.
6. Soient Y_j des variables aléatoires réelles indépendantes centrées : $\mathbb{E}Y_j = 0$ et bornées : $a_j \leq Y_j \leq b_j$. On veut démontrer ce que l'on appelle l'*inégalité de Hoeffding* : pour tout $t > 0$,
$$\mathbb{P}(\sum Y_j \geq t) \leq e^{-st} \prod e^{\frac{s^2(b_j-a_j)^2}{8}} \quad \forall s > 0.$$
   On pose $\Phi_Y(s) = \log \mathbb{E}[e^{s(Y-\mathbb{E}Y)}]$.
   (a) Montrer que
$$\Phi_Y''(s) = e^{-\Phi_Y(s)}\mathbb{E}[Y^2 e^{sY}] - e^{-2\Phi_Y(s)}(\mathbb{E}[Ye^{sY}])^2.$$
   (b) On définit une nouvelle mesure de probabilité par $\mathbb{Q}(A) = e^{-\Phi_Y(s)}\mathbb{E}[e^{sY}1_A]$ pour tout borélien A. Comment interpréter $\Phi_Y''(s)$ dans ce cadre ?
   (c) Montrer alors que $\Phi_Y(s) \leq s^2\frac{(b-a)^2}{8}$.
   (d) En déduire l'inégalité de Hoeffding.
7. (a) Soient X_j des v.a. de Bernoulli de paramètre p et $\overline{X}_n = \frac{1}{n}\sum_{j=1}^n X_j$, montrer que
$$\mathbb{P}(|\overline{X}_n - p| \geq t) \leq 2e^{-2nt^2}.$$
   (b) En déduire un intervalle de confiance de niveau $1-\alpha$ pour $F(x)$.
8. Comparer les différents intervalles de confiance que vous avez obtenu.

4.8 Théorème de Cochran et applications

Voici un énoncé simplifié du théorème de Cochran.

Théorème 1 (Cochran)**.** Soit $E = (\mathbb{R}^n, \langle\ ,\ \rangle)$ l'espace euclidien usuel, et F un sous-espace vectoriel de E de dimension $p \leq n$. Notons P la projection orthogonale sur le sous-espace F. Soit $\underline{x}$ un vecteur gaussien de E centré réduit. Les vecteurs $P\underline{x}$ et $P^\perp\underline{x}$ sont indépendants, gaussiens, centrés et de matrice de variance-covariance respectives P et $P^\perp$.
Les variables aléatoires $||P\underline{x}||^2$ et $||P^\perp\underline{x}||^2$ sont indépendantes et suivent une loi du χ^2 à p et $n-p$ degrés de liberté, respectivement.

1. Démontrer le théorème.
2. Soient X_j un n-échantillon gaussien i.i.d d'espérance μ et de variance σ^2. On note
$$\overline{X}_n = \frac{1}{n}\sum_{j=1}^n X_j \text{ et } s_n^2 = \frac{1}{n-1}\sum_{j=1}^n (X_j - \overline{X}_n)^2.$$
   Démontrer que $\overline{X}_n$ suit un loi normale $\mathcal{N}(\mu, \frac{\sigma^2}{n})$ et que $(n-1)\frac{s_n^2}{\sigma^2}$ suit une loi du χ^2 à $n-1$ degrés de liberté. En déduire la loi de $\sqrt{n}\frac{\overline{X}_n-\mu}{s_n^2}$.

4.9 Un modèle non-linéaire

Soit $f : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^n$ une fonction de classe $\mathcal{C}^2$ que l'on suppose connue. Soit le modèle

$$y = f(X, \alpha) + \epsilon \text{ et } \epsilon \sim \mathcal{N}(0, \sigma^2 I_n).$$

On cherche à estimer

$$\theta = (\alpha, \sigma^2) \in \Theta \subset \mathbb{R}^{k+1}.$$

On note $L(\alpha) = ||y - f(X, \alpha)||^2$.

1. Définir le modèle, et calculer la vraisemblance.
2. Montrer que maximiser la vraisemblance est équivalent à minimiser $L(\alpha)$.
3. Calculer l'information de Fisher du modèle.

4.10 Maximum de vraisemblance et séries temporelles

Soient $\lambda \in \mathbb{R}$ tel que $|\lambda| < 1$, $c \in \mathbb{R}$ et $\sigma^2 > 0$. On observe un échantillon $\{Y_t\}_{t \leq T}$ que l'on pense suivre le modèle $AR(1)$

$$Y_t = c + \lambda Y_{t-1} + \epsilon_t \text{ où les } \epsilon_t \sim \mathcal{N}(0, \sigma^2)$$

sont des variables i.i.d.
On cherche à estimer

$$\theta = (c, \lambda, \sigma^2)^T \in \Theta \subset \mathbb{R}^3.$$

1. Calculer $\mathcal{L}(Y_1; \theta)$, $\mathcal{L}(Y_t | Y_{t-1}; \theta)$, et en déduire $\mathcal{L}(Y_2, Y_1; \theta)$.
2. Calculer la vraisemblance du modèle $\mathcal{L}(Y_1, ..., Y_n | \theta)$.
3. Calculer la matrice de variance-covariance du processus $AR(1)$ gaussien. On la note Ω.
4. Réécrire la log-vraisemblance du modèle en utilisant Ω. Quel est le lien avec la question 2 ?
5. Déterminer un estimateur du maximum de vraisemblance.
6. Refaire l'exercice pour le modèle $MA(1)$ gaussien

$$Y_t = c + \epsilon_t - \theta \epsilon_{t-1} \text{ où } \epsilon_t \sim \mathcal{N}(0, \sigma^2) \text{ i.i.d.}$$

7. En cas de forme de motivation extrême, le faire pour le modèle $ARMA(p, q)$ gaussien

$$Y_t = c + \sum_{j=1}^{p} \lambda_j Y_{t-j} + \epsilon_t + \sum_{j=1}^{q} \theta_j \epsilon_{t-j}$$

avec $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ i.i.d.

4.11 Test du χ^2

1. Soit $(X_k, Y_k)_{k=1,...,n}$ un n-échantillon d'une loi $Q = (q_{ij})_{(i,j)\in\{1,...,I\}^2}$ sur $\{1, ..., I\}^2$, dont les marginales sont égales. Soit, pour tout $k = 1, ..., n$, le vecteur aléatoire

$$Z_k = (1_{X_k=i} - 1_{Y_k=i})_{1\leq i\leq I}.$$

   (a) Quelle est la matrice de covariance Γ de Z_k ?

   (b) On suppose Γ inversible et on note son inverse

$$\Gamma^{-1} = (\Gamma^{ij})_{1\leq i,j\leq I-1}.$$

   Soient, pour tout $i, j = 1, ..., I$,

$$\begin{aligned} N_{ij} &= \textstyle\sum_{k=1}^n 1_{X_k=i,Y_k=j} \\ N_{i.} &= \textstyle\sum_{k=1}^n 1_{X_k=i} \\ N_{.j} &= \textstyle\sum_{k=1}^n 1_{Y_k=j} \end{aligned}$$

   Quelle est la loi asymptotique de

$$\frac{1}{n}\sum_{1\leq i,j\leq I}(N_{i.} - N_{.i})(N_{j.} - N_{.j})\Gamma^{ij} \text{ ?}$$

2. On ne suppose plus a priori que les marginales soient égales. On observe (X_k, Y_k) décrit comme ci-dessus. Soit V la matrice $(V_{ij})_{1\leq i,j\leq I-1}$ définie par

$$nV_{ii} = N_{i.} + N_{.i} - 2N_{ii},$$

   et pour tout $i \neq j$,

$$nV_{ij} = -(N_{ij} + N_{ji}).$$

   (a) Montrer que, sous l'hypothèse d'égalité des marginales, V converge vers Γ.

   (b) Soit $(V^{ij})_{1\leq i,j\leq I-1}$ l'inverse de V . Montrer que

$$\Delta = \frac{1}{n}\sum_{i,j}(N_{i.} - N_{.i})(N_{j.} - N_{.j})V_{ij}$$

   converge vers une loi $\chi^2(I - 1)$.

3. Quel test peut-on construire ?
4. Appliquer ce test aux données suivantes. On évalue le degré de vision des deux yeux de 7477 femmes agées de 30 à 40 ans en le classifiant en 4 groupes (1 à 4, du meilleur au pire). On obtient

oeildroit — oeilgauche	1	2	3	4
1	1520	266	124	66
2	234	1512	432	78
3	117	362	1772	205
4	36	82	179	492

Chapitre 5

Tests

5.1 Principe de Neyman : décision à 2 points

1. Soit f la densité d'une loi de probabilité sur $\mathbb{R}$, et $\mathcal{E}$ l'expérience statistique engendré par un n-échantillon de loi $p_\theta(x) = f(x-\theta)$. On suppose que $\Theta = \{0, \theta_0\}$ avec $\theta_0 \neq 0$. On veut tester $H_0|\theta = 0$ contre $H_1|\theta = \theta_0$.
   (a) Décrire l'expérience statistique et donner la vraisemblance du modèle.
   (b) Donner la zone de rejet du test de Neyman-Pearson de niveau α associé à H_0 et H_1.
2. L'expérimentateur observe une seule réalisation d'une v.a. X de loi de Poisson de paramètre $\theta > 0$. On veut tester $H_0|\theta = \theta_0$ contre $H_1|\theta = \theta_1$, où $\theta_0 \neq \theta_1$.
   (a) Donner la zone de rejet du test de Neyman-Pearson de niveau α associé.
   (b) Sachant que $\mathbb{P}_{\theta_0}(X > 9) = 0.032$ et $\mathbb{P}_{\theta_1}(X > 8) = 0.068$, donner une zone de rejet explicite pour $\alpha = 0.05 = 5\%$. Le test est-il optimal ?

5.2 Neyman-Pearson : familles à rapport de vraisemblance monotone

1. Soit $\mathcal{E}$ l'expérience statistique engendrée par un n-échantillon de loi normale $\mathcal{N}(\theta, \sigma^2)$, où σ^2 est connu, et $\theta \in \Theta = \mathbb{R}$. On souhaite tester $H_0|\theta = \theta_0$ contre $H_1|\theta = \theta_1$, où $\theta_0 < \theta_1$.
   (a) Décrire le modèle ainsi que la vraisemblance. On choisira la mesure de Lebesgue comme mesure dominante.
   (b) Calculer le rapport de vraisemblance
   $$\frac{f(\theta_1, Z)}{f(\theta_0, Z)}.$$
   (c) Donner la zone de rejet pour le test de Neyman-Pearson associé.

2. Pour la même expérience statistique, on a un test optimal (uniformément plus puissant) de H_0 contre H_1 donné par la région de rejet

$$\mathcal{R} = \{\overline{X}_n > c\}$$

où c est solution de $\mathbb{P}_{\theta_0}(\overline{X}_n > c) = \alpha$.

   (a) Calculer explicitement la valeur de la constante $c = c(\theta_0, \alpha)$.

   (b) Calculer la puissance de ce test.

5.3 Exercice

L'expérimentateur observe 2 échantillons indépendants $X_1, ..., X_n$ et $Y_1, ..., Y_m$ de tailles distinctes $n \neq m$, de lois respectives $\mathcal{N}(\mu_1, \sigma_1^2)$ et $\mathcal{N}(\mu_2, \sigma_2^2)$. Il souhaite tester

$$H_0 \ : \ \mu_1 = \mu_2 \quad \text{contre} \quad H_1 : \ \mu_1 \neq \mu_2.$$

Si $s_{n,1}^2 = \frac{1}{n}\sum_{j=1}^n (X_j - \overline{X}_n)^2$ et $s_{m,2}^2 = \frac{1}{m}\sum_{j=1}^m (Y_j - \overline{Y}_m)^2$, construire un test basé sur la statistique

$$T_{n,m} = \frac{\overline{X}_n - \overline{Y}_m}{\sqrt{s_{n,1}^2 + s_{m,2}^2}}$$

et étudier sa consistance.

5.4 Neyman-Pearson : loi exponentielle

On observe un n-échantillon $\underline{x} = (X_1, ..., X_n)$ de variables iid de loi exponentielle de paramètre $\lambda > 0$, de densité

$$x \mapsto \lambda \exp(-\lambda x) 1_{x \leq 0}.$$

1. Rappeler l'espérance et la variance (les calculer si besoin) d'une loi exponentielle de paramètre λ. On rappelle que $2\lambda \sum_{j=1}^n X_j$ suit alors une loi du χ^2 à $2n$ degrés de liberté.
2. Ecrire le modèle statistique engendré par l'observation $\underline{x}$.
3. Calculer l'estimateur du maximum de vraisemblance $\hat{\lambda}^{MV}$ de λ.
4. Montrer que $\hat{\lambda}^{MV}$ est asymptotiquement normal, et calculer sa variance limite.
5. Soient $0 < \lambda_0 < \lambda_1$. Construire un test d'hypothèse de

$$H_0 : \lambda = \lambda_0 \text{ contre } H_1 : \lambda = \lambda_1$$

   de niveau α et uniformément plus puissant. Expliciter le choix du seuil définissant la région critique. Montrer que le test est consistant, i.e. que l'erreur de seconde espèce du test tend vers 0 lorsque $n \to \infty$.

5.5 Tests du χ^2

On considère une variable qualitative X, à valeur dans un ensemble fini $E = \{1, .., d\}$. Les lois de probabilité de telles v.a. sont entièrement décrites par le vecteur de probabilité $(p_1, ..., p_d)^T$, où $p_j = \mathbb{P}(X = j)$. On confondera donc les lois de probabilités de E avec

$$\mathfrak{M}_d = \{p = (p_1, ..., p_d)^T : 0 \geq p_j \geq 1 \text{ et } \sum p_j = 1\}.$$

1. **Test d'adéquation du χ^2.** On observe un n-échantillon de loi p et l'on souhaite tester $p = q$ contre $p \neq q$, où $q \in \mathfrak{M}_d$ est une loi fixée.
   (a) Décrire le modèle statistique.
   (b) On définit les fréquences empiriques
   $$\hat{p}_{n,l} = \frac{1}{n} \sum_{j=1}^{n} 1_{X_j = l} \quad \text{pour } l = 1, ..., d.$$
   Donner la limite du vecteur $\hat{p}_n = (\hat{p}_{n,l})_{l=1,d}^T$ pour la topologie de la convergence en probabilité sous $\mathbb{P}_p$.
   (c) On définit
   $$U_n(p) = \sqrt{n}(\frac{\hat{p}_{n,l} - p_l}{\sqrt{p_l}})_{l=1,d}^T.$$
   Donner la limite en loi de chaque composante de $U_n(p)$ sous $\mathbb{P}_p$. Que peut-on dire a priori de la limite en loi de $U_n(p)$? Pourquoi ?
   (d) On définit
   $$Y_l^j = \frac{1}{\sqrt{p_l}}(1_{X_j = l} - p_l).$$
   Si Y_j désigne le vecteur $(Y_1^j, ..., Y_d^j)$, montrer que $\frac{1}{\sqrt{n}} \sum Y_j = U_n(p)$.
   (e) Calculer $E[Y_l^j]$, et $E[Y_l^j Y_{l'}^j]$. Que valent les composantes de la matrice $V(p) = I_d - \sqrt{p}\sqrt{p}^T$, où $\sqrt{p} = (\sqrt{p_1}, ..., \sqrt{p_d})^T$?
   (f) En déduire la limite en loi sous $\mathbb{P}_p$ de $U_n(p)$ et de $||U_n(p)||^2$, le carré de sa norme euclidienne.
   (g) Soient $p, q \in \mathfrak{M}_d$ tels que les coefficients de q soient tous non nuls. On définit :
   $$\chi^2(p, q) = \sum_{l=1}^{d} \frac{(p_l - q_l)^2}{q_l}.$$
   Cette quantité est appelée "distance du χ^2" bien que ce ne soit pas une distance ! Toutefois, $\chi^2(p, q) = 0$ ssi $p = q$.
   Montrer que $n\chi^2(\hat{p}_n, p) = ||U_n(p)||^2$.
   (h) On définit, pour $\alpha \in (0, 1)$, la zone de rejet
   $$\mathcal{R}_{n,\alpha} = \{n\chi^2(\hat{p}_n, p) \geq q_{1-\alpha,d-1}^{\chi^2}\},$$
   où $q_{1-\alpha,d-1}^{\chi^2}$ est le quantile d'ordre $1 - \alpha$ de la loi du χ^2 à $d - 1$ degrés de liberté.
   Montrer que le test associé est asymptotiquement de niveau α et est asymptotiquement consistant.

(i) Application numérique. On décrit ici l'expérience de Mendel. Le croisement des pois fait apparaître 4 phénotypes, distibués selon une loi multinomiale de paramètre

$$q = (\frac{9}{16}, \frac{3}{16}), \frac{3}{16}, \frac{1}{16}).$$

Pour $n = 556$ observations, Mendel rapporte les observations suivantes : les phénotypes se répartissent selon $(315, 101, 108, 32)$. Sachant que le quantile d'ordre 0.95 de la loi du χ^2 à 3 degrés de liberté vaut 0.7815, accepter vous le test $p = q$ contre $p \neq q$.