Université catholique de Louvain

Faculté des sciences économiques, sociales, politiques et de communication (ESPO)

LINGE1214

Exercices, 2022/23

Auteurs :
Hortense DOMS
Charlotte JAMOTTON
Stéphane LHAUT
Christian HAFNER

Chapitre 1

Probabilités et résultats asymptotiques

Les exercices de ce chapitre sont essentiellement consacrés à quelques rappels de probabilité, à la loi des grands nombres et au théorème central limite. Les concepts seront brièvement rappelés lors de la séance d'exercices ainsi que les fonctions R suivantes, utiles à la résolution des exercices : `pnorm()` (calcul d'une probabilité), `qnorm()` (calcul d'un quantile) et `rnorm()` (génération de nombres aléatoires). Certains des exercices ci-après proviennent du livre « Mathematical Statistics with Applications, 7^{th} Edition » de Wackerly, Mendenhall et Scheaffer.

1.1 À préparer avant la séance

1.1: Soit $X \sim N(0,1)$. Déterminez $a > 0$ tel que

$$P(-a \leq X \leq a) = 0.874.$$

Expliquez comment solutionner le problème avec une table statistique et avec R.

1.2: Supposons que $Y_1, ..., Y_n$ soient des variables aléatoires i.i.d., chacune distribuée selon $\mathcal{N}(10, 4)$.

(a) En utilisant une table statistique appropriée, calculez $P(9,6 < \overline{Y}_n < 10,4)$ lorsque $n = 20$ et $n = 100$. Utilisez ensuite RStudio au lieu de la table statistique et comparez vos réponses.

(b) Soit $\epsilon > 0$. Montrez que $\lim_{n\to\infty} P(10 - \epsilon < \overline{Y}_n < 10 + \epsilon) = 1$.

(c) Utilisez votre réponse obtenue en b) pour argumenter que $\overline{Y}_n \to_p 10$, c'est-à-dire que la moyenne arithmétique converge en probabilité vers 10.

(d) Le résultat du point (c) est connu sous le nom de loi des grands nombres. Quelles hypothèses sous-jacentes permettent d'utiliser un tel résultat ?

1.3: Le temps d'arrêt par jour pour une installation informatique a une moyenne de 4 heures et un écart-type de 0.8 heures. Supposons que l'on souhaite calculer des probabilités concernant le temps d'arrêt moyen sur une période de 30 jours.

(a) Quelles hypothèses doivent être faites si l'on souhaite utiliser le théorème central limite ?

(b) Sous les hypothèses effectuées en (a), calculez la probabilité que le temps d'arrêt moyen sur une période de 30 jours soit compris dans l'intervalle $[1, 5]$.

1.2 À préparer pendant et après la séance

1.2.1 Probabilités

1.4: Un examen QCM consiste de 45 questions avec 3 réponses possibles pour chaque question, dont une seule d'entre elles est correcte. Soit X le nombre de réponses correctes d'une personne ayant répondu au hasard à chacune des questions. Calculez $\mathbb{E}[X]$ et Var$[X]$.

1.5: Soit X une variable aléatoire avec densité de probabilité donnée par $f(x) = \exp(-x)1_{[0,\infty)}(x)$. Calculez la probabilité $P(X > 1)$.

1.6: Considérons la densité Gaussienne suivante :

$$f(x) = \frac{1}{\sqrt{18\pi}} \exp\left(-\frac{x^2 - 4x + 4}{18}\right).$$

Déterminez les paramètres μ et σ^2 de cette Gaussienne.

1.7: Soit $X \sim N(0,1)$ et $Y = X^2$. Montrez que $\mathbb{C}ov(X,Y) = 0$. Cela implique-t-il que X et Y^2 sont indépendants ? Générez $n = 100$ valeurs de X en R (via la fonction `rnorm()`), générez les valeurs de Y associées, et ensuite calculez la covariance empirique entre les deux. Est-elle proche de 0 ?

1.8: Soit (X,Y) un vecteur Gaussien de paramètre de corrélation ρ. Si $\rho = 0$, cela implique-t-il que X et Y sont indépendants ? Cela rentre-t-il en contradiction avec votre réponse à l'exercice précédent ?

1.9: Supposons que $X \sim N(0,1)$. Quel est la densité de probabilité de $Y = \cos(X)$? Répondez à la question numériquement en simulant 10 000 fois une distribution normale standard et réalisez un histogramme des valeurs de Y associées.
[Indice : pensez aux fonctions `rnorm()` et `hist()` en R.]

1.2.2 Loi des grands nombres

1.10: Supposons que $X_1, \ldots, X_n$ soit un échantillon i.i.d. provenant d'une distribution exponentielle de moyenne 1.

(a) Est-il vrai que $\overline{X}_n \to_p 1$? Pourquoi ?

(b) Démontrez la convergence en probabilité mentionnée au point (a) en partant directement de la définition.
[Indice : l'inégalité de Chebyshev assure que pour une variable aléatoire X de variance finie, on a, pour tout $\epsilon > 0$, que $P(|X - \mathbb{E}[X]| \geq \epsilon) \leq \frac{\text{Var}[X]}{\epsilon^2}$.]

1.11: À la roulette, admettons qu'il est tombé dix fois de suite "rouge". Si c'est un jeu de roulette correcte, ça voudrait dire que la probabilité d'obtenir "noir" doit désormais être plus grande que celle d'obtenir "rouge" afin de garantir que les fréquences relatives s'approchent de 1/2 selon la loi des grands nombres ? Pourquoi ?

1.12: Vous lancez deux dés n fois, $n \geq 1$, chaque fois notant la somme des deux dés. L'événement A soit défini par la somme des deux dés égale à 12, et l'événement B par la somme égale à 11. Vous comptez le nombre de fois que A et B se produisent, n_A et n_B, et calculez les fréquences relatives associés, n_A/n et n_B/n.

(a) Pour un petit n (par exemple $n = 10$), est-il possible que $n_A - n_B > 0$?

(b) Lorsque vous augmentez n, vers quelle valeur converge $n_A/n - n_B/n$?

1.2.3 Théorème central limite

1.13: La résistance à la rupture du verre trempé est en moyenne de 14 (mesurée en milliers de livres par pouce carré) et son écart type est de 2.

(a) Quelle est la probabilité que la résistance moyenne à la rupture de 100 morceaux de ce verre, choisis aléatoirement, dépasse $14,5$? Donnez deux réponses : la première à l'aide des tables statistiques et la seconde à l'aide de RStudio.

(b) Trouvez un intervalle qui inclut, avec une probabilité de $0,95$, la résistance moyenne à la rupture de 100 morceaux de ce verre choisis aléatoirement. Donnez deux réponses : la première à l'aide des tables statistiques et la seconde à l'aide de Rstudio.

1.14: Un anthropologue souhaite estimer la taille moyenne des hommes pour une certaine race de personnes. Supposons que l'écart-type de la population est de $2,5$ pouces. Si l'anthropologue échantillonne aléatoirement 100 hommes, trouvez la probabilité que la différence entre la moyenne de l'échantillon et la vraie moyenne de la population ne dépasse pas $0,5$ pouce.

1.15: Un fabricant de chaises veut déterminer la résistance moyenne (en kg) d'un nouveau modèle qu'il produit. Il sait que la résistance d'une chaise est une variable aléatoire dont l'écart-type est égal à 5 kg. Combien de chaises doit-il tester s'il veut que la différence entre la résistance moyenne de l'échantillon et la moyenne de la population soit inférieure à 1 avec une probabilité d'au moins $0,95$? Utilisez R pour effectuer vos calculs.

1.16: Les travailleurs employés dans une grande entreprise de services ont un salaire moyen de $7,00$€ de l'heure avec un écart type de $0,50$€. L'industrie compte 64 travailleurs appartenant à un certain groupe ethnique. Ces travailleurs ont un salaire moyen de $6,90$€ de l'heure. Est-il raisonnable de supposer que le taux de salaire du groupe ethnique est équivalent à celui d'un échantillon aléatoire de travailleurs parmi ceux employés dans l'industrie des services ?

[Indice : calculez la probabilité d'obtenir une moyenne d'échantillon inférieure ou égale à $6,90$€ par heure].

1.17: Les concentrations de monoxyde de carbone sur une heure dans les échantillons d'air d'une grande ville sont en moyenne de 12 ppm (parties par million), avec un écart-type de 9 ppm.

(a) Pensez-vous que les concentrations en monoxyde de carbone dans les échantillons d'air de cette ville soient normalement distribuées ? Pourquoi ?

(b) Calculez la probabilité que la concentration moyenne de 100 échantillons aléatoires dépasse 14 ppm.

1.18: Supposons que $X_1, \ldots, X_n$ soit un échantillon i.i.d. provenant d'une distribution exponentielle de moyenne 1.

(a) Calculez la probabilité suivante via R, avec $n = 10$:

$$P(\overline{X}_n > 1.5).$$

[Indice : une somme de n variables aléatoires i.i.d. provenant d'une distribution exponentielle d'intensité λ suit une distribution Gamma de paramètres (n, λ). Tout comme pour une Gaussienne, on peut calculer les probabilités associées à une distribution Gamma via la fonction `pgamma()` en R.]

(b) Calculez une approximation de la probabilité précédente sur base du théorème central limite et comparez le résultat obtenu avec la réponse obtenue en (a).

(c) Ré-effectuez vos calculs des points (a) et (b) avec $n = 20$. L'approximation devient-elle meilleure ?

Chapitre 2

Estimation

2.1 À préparer avant la séance

2.1: Soit $X_1, ..., X_n$ un échantillon aléatoire d'une population de densité

$$f(x) = \begin{cases} e^{-(x-\theta)} & \text{si } x > \theta \\ 0 & \text{sinon,} \end{cases}$$

pour un paramètre $\theta \in \mathbb{R}$.
Montrez que $\overline{X}_n$ est un estimateur non biaisé de $1 + \theta$.

2.2: Vous avez deux investissements financiers, A et B. Lors des six derniers mois, les rendements de l'actif A étaient les valeurs suivantes :

$$0.05, 0.07, -0.12, -0.04, 0.06, 0.02$$

tandis que ceux de l'actif B étaient :

$$0.02, 0.01, -0.03, 0.04, -0.01, 0.03$$

Estimez la différence de moyennes ainsi que la différence de variances entre les deux actifs. Les deux estimateurs sont-ils biaisés ? Si oui, peut-on dire si le biais est positif ou négatif ? Auriez-vous une préférence pour votre investissement ?

2.3: Pour estimer la différence entre les proportions d'hommes et de femmes vacciné.e.s contre un virus, on prélève deux échantillons indépendantes des $n_1 = 200$ hommes et $n_2 = 200$ femmes. Parmi ces personnes, 157 hommes et 172 femmes ont été vacciné.e.s. Estimez la différence entre les proportions, et estimez la variance de cet estimateur.

2.2 À préparer pendant et après la séance

2.4: On souhaite estimer la moyenne et la variance de la taille des hommes en Belgique avec un échantillon de cinq hommes, tirés au hasard. Les valeurs obtenues sont :

$$1.75, 1.82, 1.69, 1.95, 1.87$$

Utilisez S^2 pour l'estimation de la variance. L'estimateur est-il biaisé ? Quel serait un estimateur non-biaisé ? Si les valeurs n'avaient pas été exprimées en mètres mais en centimètres, quelle aurait été l'estimateur de la variance ?

2.5: Supposez $X \sim N(\mu, \sigma^2)$, c'est-à-dire, la variance de X est σ^2, et la kurtosis $K = 3$. Avec $n = 20$ réalisations de X, calculez l'erreur quadratique moyenne (MSE) de l'estimateur de la variance S^2 pour le cas où $\sigma^2 = 1$, en utilisant l'approximation du théorème 2.3 du syllabus.

2.6: On s'intéresse à la proportion p de fumeurs dans la population. Avec un échantillon de taille n, on obtient un estimateur ponctuel $\widehat{p}$. Quelqu'un prétend qu'un meilleur estimateur serait de prendre la proportion de fumeurs uniquement des hommes dans l'échantillon. On supposant que la proportion de femmes et d'hommes qui fument est égale dans la population, calculez l'efficacité relative de cet estimateur par rapport à $\widehat{p}$, si le nombre de femmes et d'hommes dans l'échantillon est le même. Lequel est préférable ?

2.7: Supposons que le temps moyen de défaillance technique d'un ordinateur soit donné par T années. Notons μ sa moyenne et $\sigma > 0$ son écart-type. On souhaite estimer μ. Pour cela, basé sur un échantillon i.i.d. $T_1, \ldots, T_n$, on considère deux estimateurs pour cette quantité :

$$\hat{\mu}_1 = \overline{T}_n \quad \text{et} \quad \hat{\mu}_2 = 10.$$

(a) Calculer l'erreur quadratique moyenne (MSE) pour chacun des estimateurs. L'estimateur $\hat{\mu}_1$ est-il consistant pour μ ?

(b) Supposons que nous disposions de $n = 100$ données. Sur base des MSE calculées au point (a), quel estimateur préférez-vous si on sait que $\mu = \sigma = 10$? Si $\mu = 9$ et $\sigma = 5$?

(c) De manière générale, pensez-vous que l'estimateur $\hat{\mu}_2$ est un estimateur intéressant ? Pourquoi ?

2.8: Supposons que nous disposions de données i.i.d. $X_1, \ldots, X_n$ de loi inconnue. Notons F la fonction de répartition associée. On souhaite estimer la probabilité $F(2) = P(X_1 \leq 2)$. Sans hypothèse sur F, la manière la plus directe de le faire est certainement de compter le nombre de données plus petites ou égales à 2 et de diviser ce nombre par n. Cela correspond à prendre l'estimateur

$$\hat{F}_n(2) = \frac{1}{n} \sum_{i=1}^{n} 1\{X_i \leq 2\}.$$

(a) Calculez l'erreur quadratique moyenne (MSE) de cet estimateur.

(b) Est-ce un estimateur consistant pour la probabilité d'intérêt ? Pourquoi ?

(c) (Pour aller plus loin) Le théorème central limite permet-il d'obtenir une distribution asymptotique pour l'estimateur $\hat{F}_n(2)$? Si oui, quelle est-elle ? [Indice : appliquer le TCL sur les variables aléatoires $Y_i := 1\{X_i \leq 2\}$.]

2.9: Supposons que $Y \sim N(\mu, 5)$, avec μ inconnue. Basé sur un échantillon i.i.d. $Y_1, \ldots, Y_n$, $n \geq 3$, on considère les deux estimateurs

$$\hat{\mu}_1 = \frac{1}{3}(Y_1 + Y_2 + Y_3)$$

et

$$\hat{\mu}_2 = \frac{1}{2n}(Y_1 + Y_n) + \frac{n-1}{n(n-2)} \sum_{i=2}^{n-1} Y_i$$

(a) Calculez le biais des deux estimateurs.

(b) Calculez l'efficacité relative de $\hat{\mu}_2$ par rapport à $\hat{\mu}_1$ pour $n = 3$, $n = 4$, et $n = 5$. Vers quelle valeur tend cette efficacité relative lorsque n tend vers infinie ?

(c) Les estimateurs proposés sont-ils consistants ? Pourquoi ?

2.10: Une entreprise suédoise emploie 100 travailleurs. Parmi ces employés, 40 sont d'origine nicaraguayenne, et les autres sont suédois. Les nicaraguayens se plaignent fréquemment auprès de leur patron que leurs conditions de travail sont moins bonnes que celles de leurs homologues suédois. Ils se basent sur le fait que sur les 45 employés de l'entreprise qui possèdent un bureau avec balcon, seuls 10 sont nicaraguayens. On notera, respectivement, p_S et p_N la proportion (inconnue) de suédois et de nicaraguayens ayant un bureau avec balcon.

(a) Proposez un estimateur consistant pour la différence des proportions $p_N - p_S$. Pourquoi votre estimateur est-il consistant ?

(b) Calculer l'erreur quadratique moyenne (MSE) de votre estimateur.

(c) Supposons que $p_S = p_N$. Sur base d'une approximation asymptotique, calculez la probabilité $P(\hat{p}_N - \hat{p}_S \leq -1/3)$. Votre résultat semble-t-il indiquer que les nicaraguayens ont de moins bonnes conditions de travail ?

Chapitre 3

Intervalles de confiance

3.1 À préparer avant la séance

3.1: Lors d'une enquête, on a demandé à un échantillon aléatoire de 500 belges le nombre de fois qu'ils s'étaient rendu au cinéma au cours du dernier mois écoulé. On a obtenu les résultats décrits par le tableau suivant :

x_j	0	1	2	3	4	5
n_j	223	135	69	37	22	14

où n_j désigne le nombre d'individus ayant déclaré s'être rendus x_j fois au cinéma au cours du dernier mois écoulé pour $j \in \{0, \ldots, 5\}$.

(a) Donnez une estimation de la moyenne et de l'écart-type du nombre de fois par mois que les belges se rendent au cinéma.

(b) Donnez un intervalle de confiance à 92% pour la moyenne.

(c) Donnez une estimation de la proportion des belges qui se rendent plus de deux fois par mois au cinéma et construisez un intervalle de confiance de niveau 0,97 pour cette proportion.

(d) On souhaiterait obtenir avec le même niveau de confiance $(= 97\%)$ une estimation de cette proportion à $\pm 0,01$ près ? Quelle taille d'échantillon approximative devrait-on considérer ?

3.2: On souhaite évaluer la proportion d'électeurs (fréquence en population) qui approuveront une certaine proposition lors d'un prochain référendum. Deux sondages aléatoires ont été effectués et ils ont donné les résultats suivants :

- Le 1er octobre, 75 des 130 personnes interrogées approuvaient la proposition.
- Le 15 octobre, 642 des 1056 personnes interrogées approuvaient la proposition.

Pour chaque date :

(a) Donnez une estimation de la proportion des électeurs qui sont en faveur de la proposition.

(b) Donnez un intervalle de confiance à 95% pour cette proportion.

3.2 À préparer pendant et après la séance

3.3: Le gestionnaire d'un self-service souhaite tester la satisfaction des usagers de ce restaurant. Il sélectionne un échantillon aléatoire de 60 usagers. Parmi eux, 42 se déclarent globalement satisfaits. Calculez un intervalle de confiance à 99% pour la proportion de satisfaits sur l'ensemble des usagers.

3.4: Une marque de café vend des pots de café indiquant $450g$. Sur un échantillon indépendant de 50 pots, on obtient une moyenne de $425g$. En supposant que le poids d'un pot suit une distribution normale d'écart-type $\sigma = 17g$, donnez un intervalle de confiance bilatéral à 95% pour la vraie moyenne μ.

3.5: De nombreux fonds d'investissement utilisent comme critère d'investissement dans une action la comparaison entre son *price earning ratio* (PER) et le PER du S&P500 (il s'agit d'un indicateur pour voir si une action est surévaluée ou non).

Les 49 données suivantes correspondent aux PER's d'actions sélectionnées par un certain fond d'investissement durant une année :

$$
\begin{aligned}
&6.8, 5.6, 8.5, 8.5, 8.4, 7.5, 9.3, 9.4, 7.8, 7.1,\\
&9.9, 9.6, 9.0, 9.4, 13.7, 16.6, 9.1, 10.1, 10.6, 11.1,\\
&8.9, 11.7, 12.8, 11.5, 12.0, 10.6, 11.1, 6.4, 12.3, 12.3,\\
&11.4, 9.9, 14.3, 11.5, 11.8, 13.3, 12.8, 13.7, 13.9, 12.9,\\
&14.2, 14.0, 15.5, 16.9, 18.0, 17.9, 21.8, 18.4, 34.3
\end{aligned}
$$

(a) Trouvez un intervalle de confiance à 98% pour le PER moyen. Interpréter le résultat obtenu et préciser les hypothèses supposées.

(b) Construisez un intervalle de confiance approximative de 95% pour la variance du PER, en estimant la kurtosis K par la kurtosis empirique des données. Interpréter le résultat obtenu et préciser les hypothèses supposées.

[Indice : Utiliser R pour faciliter les calculs. Importer le fichier "data.txt" en R : File > Import Dataset > From Text (base)... > Select file "data.txt" to import > Open. Transformer la base de données en vecteur dans un R Script : data <- as.vector(unlist(data))]

3.6: Une enquête vise à déterminer la proportion d'étudiants consommant de l'alcool au moins une fois par semaine tout en effectuant une comparaison entre les garçons et les filles. Cent étudiants ont été interrogés au total, cinquante garçons et cinquante filles. Les résultats de l'enquête sont présentés dans le tableau suivant.

	Filles	Garçons
Alcool au moins 1x/semaine	27	35

(a) Estimez la proportion d'étudiants consommant de l'alcool au moins une fois par semaine et donner un intervalle à 95% pour cette proportion. La taille de votre intervalle augmente-t-elle si l'on demande un niveau de confiance de 99% ? Pourquoi ?

(b) Estimez la différence de proportion entre les hommes et les femmes et donner un intervalle à 95% pour cette différence. La taille de votre intervalle augmente-t-elle si l'on avait effectué l'étude avec 70 garçons et 70 filles et que l'estimation des proportions restait identique ? Pourquoi ?

3.7: Une machine produit de petites visses utilisées pour la fixation de certains composants informatiques. Les visses doivent être environ de longueur 1.5 mm. Sur base d'un échantillon aléatoire de taille 100, on estime (par les estimateurs classiques de moyenne et de variance) que la moyenne de longueur des visses produites est de 1.53 mm et que l'écart-type est de 0.5 mm.

(a) Donnez un intervalle de confiance à 95% pour la moyenne des longueurs des visses produites. Cet intervalle contient-il la valeur 1.5 mm ?

(b) On nous rapporte que $\hat{K} = 1.2$, où $\hat{K}$ est un estimateur consistant pour le kurtosis des données. Sur base du TCL, proposez un intervalle de confiance à 99% pour la variance.

3.8: Une entreprise souhaite estimer le solde créditeur moyen de ses clients à partir d'un échantillon aléatoire de 100 comptes clients. A partir des relevées sur cet échantillon, on obtient une estimation ponctuelle de la moyenne de 1274 euro, ainsi qu'une estimation ponctuelle de l'écart-type de 326 euro. Construisez un intervalle de confiance de 95% pour la moyenne du solde créditeur.

Chapitre 4

Méthodes d'estimation

4.1 À préparer avant la séance

4.1: Soit $X_1, ..., X_n$ un échantillon i.i.d. suivant une loi uniforme sur l'intervalle $[\theta - 1, \theta + 1]$ pour un certain paramètre θ.

(a) En utilisant la méthode des moments, trouver un estimateur pour θ.

(b) Utiliser cet estimateur sur l'échantillon suivant :

$$11.72, 12.81, 12.09, 13.47, 12.37$$

4.2: Un dé est lancé successivement jusqu'à ce que l'on obtienne un 6. Le nombre de lancers, Y, suit une distribution géométrique avec fonction de probabilité

$$f(y) = P(Y = y) = p(1 - p)^{y-1}, \qquad y = 1, 2, 3, \ldots$$

où p est un paramètre, $0 < p < 1$.

(a) Avec un échantillon $Y_1, \ldots, Y_n$, dériver le MLE de p.

(b) On répète l'expérience cinq fois et on obtient des valeurs pour Y de 3, 7,5,8, et 4. Estimer p avec la méthode de maximum vraisemblance.

(c) Quelle valeur attendez-vous pour la valeur théorique de p dans cet exemple ?

4.2 À préparer pendant et après la séance

4.3: Soit $X_1, ..., X_n$ un échantillon aléatoire indépendant obtenu à partir de n lancers d'une pièce biaisée dont la probabilité de tomber sur face est p, avec $X_i = 1$ si le $i-$ème lancer tombe sur face et $X_i = 0$ sinon.

(a) Trouvez le MLE $\hat{p}$ de p.

(b) En utilisant la propriété d'invariance, déduire de (a) un MLE pour $q = 1 - p$.

4.4: Soit $X_1, ..., X_n$ un échantillon aléatoire indépendant d'une population de densité

$$f(x) = \begin{cases} e^{-(x-\theta)} & \text{si } x > \theta \\ 0 & \text{sinon,} \end{cases}$$

pour un certain paramètre θ.

Montrer que le MLE de θ est égal à $\min(X_1, \ldots, X_n)$.

4.5: Soit $X_1, ..., X_n$ un échantillon i.i.d. d'une population de densité

$$f(x) = \begin{cases} \theta x^{\theta-1} & \text{si } 0 < x < 1, \theta > 0 \\ 0 & \text{sinon,} \end{cases}$$

(a) Proposez un estimateur pour θ par la méthode des moments.

(b) Montrez que cet estimateur est consistant pour θ. [Indice : pensez à la loi des grands nombres et aux propriétés des estimateurs consistants.]

4.6: Soit $X_1, ..., X_n$ un échantillon i.i.d. d'une loi de Poisson de paramètre λ, i.e.,

$$f(n) = P(X_1 = n) = \frac{\lambda^n}{n!} \exp(-\lambda), \qquad n \in \mathbb{N}.$$

(a) Calculez $\mathbb{E}[X_1]$.

(b) Déduisez un estimateur de λ par la méthode des moments.

4.7: Soit $X_1, ..., X_n$ un échantillon i.i.d. d'une loi normale $\mathcal{N}(\theta, \theta)$ pour quelque $\theta > 0$. Calculez le MLE de θ.

4.8: La hauteur maximale H de la crue annuelle d'un fleuve est observée. La densité de probabilité de la variable aléatoire H est représentée par :

$$f(x) = \begin{cases} \frac{x}{a} \exp(-\frac{x^2}{2a}) & \text{si } x > 0 \\ 0 & \text{sinon,} \end{cases}$$

où a est un paramètre inconnu. Durant une période de 8 ans, on a observé les hauteurs (supposées être indépendantes) de crues suivantes en mètres : 2.1, 2.8, 1.7, 0.9, 1.8, 2.5, 2.2, 2,9.

(a) Trouver le MLE $\hat{a}$ du paramètre a.

(b) En utilisant les données du problème, que vaut votre estimateur $\hat{a}$?

4.9: Supposons que la taille d'hommes belges sélectionnés au hasard soit normalement distribuée avec une moyenne μ et un écart type σ inconnus. Un échantillon aléatoire de 10 hommes belges correspond aux tailles suivantes (en cm) : 177, 189, 175, 181, 189, 175, 178, 182, 192, 179.

(a) Par la méthode du maximum de vraisemblance, proposez un estimateur $\hat{\mu}$ pour le paramètre μ, la taille moyenne des hommes belges.

(b) Que vaut la valeur de $\hat{\mu}$ pour l'échantillon donné ?

Chapitre 5

Tests d'hypothèse

5.1 À préparer avant la séance

5.1: Une entreprise prépare un médicament contre l'insomnie. Celle-ci prétend que son médicament est efficace à 80%. Cependant, après examination du produit, les autorités suspectent que ce taux d'efficacité a été gonflé. Afin de vérifier les informations communiquées par l'entreprise, on souhaite effectuer un test :

$$H_0 : p = 0.8 \quad \text{contre} \quad H_a : p < 0.8.$$

Pour cela, on se base sur un échantillon de 40 personnes (on considère cet échantillon suffisamment grand pour effectuer une approximation normale). On administre le médicament à toute la cohorte et on observe le nombre de patients, noté Y, qui se sont endormis suite à la prise du produit.

En supposant qu'on utilise une région de rejet donnée par $\{Y \leq 30\}$, répondez aux questions suivantes :

(a) En termes du problème considéré, qu'est-ce qu'une erreur de type I ? Calculez α.

(b) En termes du problème considéré, qu'est-ce qu'une erreur de type II ? Calculez β quand $p = 0.6$.

5.2: Un chercheur Américain en sciences politiques pense que la fraction p_1 de Républicains en faveur de la peine de mort est plus grande que la fraction p_2 de Démocrates en faveur de cette sentence. Pour vérifier ses pensées, il interroge 200 membres de chacun des partis. Il observe que 46 Républicains et 34 Démocrates sont en faveur de la peine de mort.

(a) Formalisez l'hypothèse du chercheur en un test statistique.

(b) Calculez la p-valeur du test formalisé en (a). Rejette-t-on H_0 si $\alpha = 5\%$?

5.2 À préparer pendant et après la séance

5.3: On considère à nouveau la situation décrite dans l'exercice 5.1.

(a) Trouvez la région de rejet de la forme $\{Y \leq c\}$ telle que $\alpha = 0.01$.

(b) Pour la région de rejet calculée en (a), trouvez β quand $p = 0.6$.

5.4: On examine les résultats d'une étude de 2001 de Leonard, Speziale et Pernick comparant 2 méthodes pour enseigner la biologie : une méthode traditionnelle et une méthode plus interactive. Des pré-tests ont été donnés aux étudiants qui suivront une des deux méthodes pour leur enseignement. Les statistiques descriptives des résultats sont disponibles pour 368 étudiants qui suivront un enseignement traditionnel et pour 372 étudiants qui suivront un enseignement plus interactif.

(a) Sans regarder les données, s'attend-on à observer une différence entre les moyennes des résultats du pré-test de ceux qui suivront un enseignement traditionnel et de ceux qui suivront un enseignement plus interactif ? Sur base de votre réponse à cette question, quelle hypothèse alternative choisiriez-vous pour un test où l'on souhaite tester l'hypothèse nulle d'égalité entre les moyennes des deux groupes ?

(b) Votre hypothèse alternative est-elle unilatérale ou bilatérale ?

(c) On lit que la moyenne des pré-tests de ceux qui suivront un enseignement traditionnel est 14.06 tandis que pour ceux suivront un enseignement plus interactif, on observe une moyenne de 13.38. Au niveau des écarts-types, on observe respectivement 5.45 et 5.59. Les données appuient-elles l'hypothèse d'égalité des moyennes au sein des deux groupes ? Testez formellement cette hypothèse pour $\alpha = 1\%$ et calculez également la p-valeur de votre test.

5.5: On considère à nouveau la situation décrite dans l'exercice précédent.
Un post-test est effectué après les deux enseignements respectifs. La moyenne du post-test chez les étudiants qui ont suivi l'enseignement traditionnel est de 16.50 et l'écart-type est donné par 6.96. Chez les étudiants qui ont suivi un enseignement plus interactif, on observe respectivement les valeurs 18.50 et 8.03 pour ces mêmes statistiques.

(a) Donnez un intervalle de confiance à 95% pour la différence de moyennes entre les deux groupes.

(b) L'intervalle de confiance calculé en (a) fournit-il une preuve que les résultats moyens aux post-tests sont différents entre les deux groupes d'étudiants ? Expliquez votre raisonnement.

5.6: Dans le tableau ci-dessous, on présente quelques données provenant d'une étude de 1993 réalisée par Susan Beckham et ses collègues. Dans cette étude, on mesure la pression dans la loge intérieure (partie de la jambe) chez 100 coureurs et 100 cyclistes au repos.

	Coureurs		Cyclistes	
	Moyenne	Écart-type	Moyenne	Écart-type
Repos	14.5	3.92	11.1	3.98

(a) Y-a-t-il suffisamment de preuves dans les données pour supporter, avec un degré de confiance de 95%, que la pression moyenne mesurée chez les coureurs et les cyclistes au repos diffère ?

(b) Calculez la p-valeur de votre test.

5.7: Un entrepreneur affirme que la machine de son entreprise qui permet d'emballer du savon en poudre dans des cartons de 5kg effectue cette tâche avec une variance maximale de 1%.

On estime la variance ainsi que le kurtosis sur base d'un échantillon de 100 cartons. On trouve $S^2 = 0.018$ et $\hat{K} = 2.8$.

(a) Effectuez le test suivant avec $\alpha = 5\%$:

$$H_0 : \sigma^2 = 0.01 \quad \text{contre} \quad H_a : \sigma^2 > 0.01.$$

Quelle conclusion obtenez-vous ?

(b) Calculez la p-valeur de votre test.

Chapitre 6

Analyse de données catégorielles

6.1 À préparer avant la séance

6.1: Une étude vise à comprendre la relation entre le degré de violence des émissions regardées à la télévision et l'âge du spectateur. Les résultats sont présentés dans le tableau ci-dessous. Ils concernent 81 personnes.

	Âge		
Violence des émissions	16-34	35-54	55 et plus
Faible	8	12	21
Élevée	18	15	7

Les données indiquent-elle une dépendance entre l'âge et le degré de violence ? Effectuez un test avec $\alpha = 5\%$.

6.2: Une compagnie d'assurance propose des contrats d'assurance automobile. Chaque année, elle recense le nombre de sinistres déclarés en fonction de leur gravité. Elle obtient les résultats suivants pour les sinistres ayant causés des décès humains :

Nombre de sinistres mortels	0	1	2	3	4 et plus
Nombre d'années	57	22	7	2	0

Testez, au niveau de confiance 95%, l'hypothèse que les données proviennent d'une distribution de Poisson, caractérisée par les probabilités suivantes : pour $Y \sim Po(\lambda)$, on a

$$P(Y = n) = \frac{\lambda^n}{n!} \exp(-\lambda), \text{ pour tout } n \in \mathbb{N}.$$

Indice : commencer par estimer λ par maximum de vraisemblance (cela diminuera de un le nombre de degrés de liberté de la distribution asymptotique de la statistique de test sous H_0). Ensuite, comme on a trop peu d'observations dans les catégories "Nombre de sinistres = 3" et "Nombre de sinistres = 4 et plus", groupez ces catégories pour le calcul de la statistique de test et utilisez donc $k = 3$, au lieu de 5. Faites attention à bien adapter le nombre de degrés de liberté.

6.2 À préparer pendant et après la séance

6.3: Les résultats d'une étude suggèrent que l'électrocardiogramme (ECG) d'une victime potentielle d'attaque cardiaque peut être utilisé pour prédire les complications hospitalières futures. L'étude inclut 469 patients à risque d'infarctus du myocarde (attaque cardiaque). Chaque patient est classé selon la nature positive ou négative de son ECG et selon que la personne ait subi des complications hospitalières dangereuses pour sa vie ou non. Les résultats sont décrits dans le tableau suivant.

	Type de complications		
Type d'ECG	Mortelles	Non mortelles	Total
Négatif	166	1	167
Positif	260	42	302
Total	426	43	469

Y a-t-il suffisamment de preuves dans les données pour indiquer si la nature des complications dépend de l'état initial de l'ECG ? Effectuez un test avec $\alpha = 5\%$.

6.4: Les vitesses de 150 automobilistes roulant sur une avenue d'une grande ville ont été mesurées en km/h. On obtient les données suivantes :

Vitesse	$40-55$	$55-65$	$65-75$	$75-85$	>85
Occurence	12	14	78	40	6

Testez à un seuil $\alpha = 1\%$ si les vitesses sont distribuées normalement avec une moyenne de 70 km/h et un écart-type de 4.
Indice : calculez pour chaque intervalle la probabilité $p_i^{(0)}$ d'obtenir une valeur dans cet intervalle sous l'hypothèse de normalité. Utilisez ces probabilités pour effectuer un test d'ajustement du χ^2.

6.5: Une enquête menée sur un échantillon de 1000 Belges répartis selon leur région s'intéresse à l'opinion à propos d'une mesure envisagée par le gouvernement. Les résultats obtenus sont les suivants :

	Pour	Contre	Sans opinion	Total
Flamands	250	50	200	500
Wallons	150	75	75	300
Bruxellois	50	125	25	200
Total	450	250	300	1000

Testez à un seuil $\alpha = 5\%$ si l'opinion d'un citoyen est indépendante de la région où il vit.

6.6: Une enquête menée auprès de 50 filles et 50 garçons étudiant dans une certaine université à propos des préférences en matière de chaussures a donné les résultats suivants :

	Bottes	Chaussures en cuir	Sneakers	Sandales	Autres
Filles	12	9	12	10	7
Garçons	10	12	17	7	4

(a) Soient $p_i, i = 1, \ldots, 5$ les vraies proportions d'étudiants préférant le type i de chaussures parmi ceux listés ci-dessus (on suppose qu'on lit le tableau de gauche à droite). Effectuez le test d'hypothèses suivant en utilisant un seuil $\alpha = 5\%$:

$$\begin{cases} H_0 : p_1 = 0.20,\ p_2 = 0.20,\ p_3 = 0.30,\ p_4 = 0.20,\ p_5 = 0.10. \\ H_1 : \text{au moins une des probabilités est différente de sa valeur hypothétique.} \end{cases}$$

(b) Testez à un seuil $\alpha = 5\%$ si la préférence de chaussures est indépendante du genre.

Chapitre 7

Modèles linéaires

7.1: Supposons qu'un chercheur se propose d'estimer la régression du score de tests (Y) sur l'effectif des classes (X). Pour un échantillon de $n = 100$ classes, les données sont résumées par

$$\sum_{i=1}^{n} x_i y_i = 841605.3, \quad \sum_{i=1}^{n} x_i^2 = 46834.86$$
$$\sum_{i=1}^{n} x_i = 2151.065, \quad \sum_{i=1}^{n} y_i = 39300.52$$

(a) Ajustez le modèle suivant aux données :

$$Y = \beta_0 + \beta_1 x + \epsilon.$$

(b) Déterminez la valeur attendue du score des tests pour une classe de 22 élèves.

(c) Supposons que la classe comportait 19 élèves l'année dernière et 23 élèves cette année. Quelle est la variation attendue du score des tests induite par cette variation de l'effectif de la classe ?

7.2: Maintenant nous "jouons Dieu" et créons les données nous-mêmes en R par simulation. Entrez les commandes suivantes (sans les commentaires après chaque commande) :

```
set.seed(1234)    # initialiser le générateur aléatoire
ep = rnorm(100)*11.5    # générer 100 v.a. N(0, sqrt(11.5))
x = runif(100,17,26)    # générer 100 v.a. uniforme ∈ [17, 26]
y = 520 - 5.82*x + ep    # générer les v.a. de la réponse Y.
```

Ici nous connaissons donc les vrais paramètres de la régression de Y sur X, puisque nous les avons fixés à $\beta_0 = 520$ et $\beta_1 = -5.82$. Mais si une autre personne, ne connaissant pas le vrai lien entre X et Y, disposait du jeu de données $\{(X_i, Y_i)\}_{i=1,\ldots,100}$, il pourrait estimer ce lien avec les formules comme dans l'exercice 7.1, ou en utilisant la commande `lm()` en R comme suit :

```
m = lm(y~x)    # estimer la régression de y sur x
summary(m)    # sortir un résumé de la régression estimée
```

Interprétez la sortie de cette estimation, et en particulier, comparez le résultat avec les vrais paramètres, et avec vos estimateurs de la question 7.1. Que constatez-vous ?

7.3: Le nombre d'octane Y d'un pétrole raffiné est une variable alétatoire. On cherche à savoir si Y dépend de la température x du processus de raffinage. Une expérience sur un échantillon de taille 31 a obtenu le modèle linéaire suivant :

$$\hat{y} = 9.36 + 0.155x.$$

De plus on sait que $(0.202)^2 = \frac{\hat{\sigma}^2}{S_{xx}}$. Tester à un seuil $\alpha = 5\%$ l'hypothèse que la pente est significativement différente de 0.

7.4: Démontrez que

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 - \widehat{\beta}_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2$$

(Indice : partez de la définition des résidus, $e_i := Y_i - \widehat{Y}_i$. Ensuite, remplacez les valeurs ajustées $\widehat{Y}_i$ par $\widehat{\beta}_0 + \widehat{\beta}_1 X_i$, et puis $\widehat{\beta}_0$ par l'expression obtenue pour l'estimateur moindres carrés, et ensuite simplifiez.)

7.5: Une entreprise considère une fonction de demande linéaire, $\mathbb{E}[Y|X] = \beta_0 + \beta_1 X$, où Y est la demande à un prix X. Par une analyse du marché on obtient $n = 50$ observations (X_i, Y_i), résumées par :

$$\sum_{i=1}^{n} x_i y_i = 1500, \quad \sum_{i=1}^{n} x_i^2 = 1200$$
$$\sum_{i=1}^{n} x_i = 200, \quad \sum_{i=1}^{n} y_i = 400, \quad \sum_{i=1}^{n} y_i^2 = 3350$$

(a) Calculez les estimateurs moindres carrés du modèle.

(b) Calculez le coefficient de détermination, R^2.
(Indice : obtenez la somme des résidus au carré en utilisant l'équation de l'exercice précédent, et notez que $R^2 = 1 - \sum_{i=1}^{n} e_i^2 / \sum_{i=1}^{n} (Y_i - \bar{Y})^2$.)

(c) Testez l'hypothèse $H_0 : \beta_1 = 0$ contre $H_a : \beta_1 \neq 0$ à un niveau de test $\alpha = 5\%$.

(d) Construisez un intervalle de confiance pour $\mathbb{E}[Y|X]$ à l'endroit $X = 5$ pour le même α qu'à la sous-question précédente.