Université catholique de Louvain

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

LINGE1214

Exercices : Solutions, 2022/23

Auteurs :
Hortense DOMS
Charlotte JAMOTTON
Stéphane LHAUT
Christian HAFNER

Chapitre 1

Probabilités et résultats asymptotiques

1.1 À préparer avant la séance

1.1: On calcule, par symétrie de la loi normale,

$$\begin{aligned} P(-a \leq X \leq a) &= P(X \geq -a) - P(X \geq a) \\ &= 1 - P(X \geq a) - P(X \geq a) \\ &= 1 - 2P(X \geq a) \end{aligned}$$

Par conséquent, on cherche a tel que

$$\begin{aligned} & 1 - 2P(X \geq a) = 0.874 \\ \Leftrightarrow\ & P(X \geq a) = 0.063 \\ \Leftrightarrow\ & a = 1.53 \end{aligned}$$

En effet, si on note z_α le quantile supérieur d'ordre $\alpha \in (0,1)$, on cherche $a = z_{0.063}$. La fonction `qnorm(p=0.063, mean = 0, sd = 1, lower.tail = F)` en R ou la table statistique de la normale permettent de trouver $a = 1.53$.

1.2:

(a) Notons que $\overline{Y}_n \sim N(10, 4/n)$. Donc,

$$\begin{aligned} P(9.6 < \overline{Y}_n < 10.4) &= P\left(\frac{9.6-10}{\sqrt{4/n}} < \frac{\overline{Y}_n - 10}{\sqrt{4/n}} < \frac{10.4-10}{\sqrt{4/n}}\right) \\ &= P\left(-0.2\sqrt{n} < Z < 0.2\sqrt{n}\right) \text{ où } Z \sim \mathcal{N}(0,1). \end{aligned}$$

Pour $n = 20$ on trouve $P(-0.894 < Z < 0.894) = 0.627$ et pour $n = 100$, $P(-2 < Z < 2) = 0.954$.

(b) Soit $\epsilon > 0$. On calcule

$$\begin{aligned} P(10 - \epsilon < \overline{Y}_n < 10 + \epsilon) &= P\left(\frac{-\epsilon}{\sqrt{4/n}} < \frac{\overline{Y}_n - 10}{\sqrt{4/n}} < \frac{\epsilon}{\sqrt{4/n}}\right) \\ &= P\left(-\frac{\epsilon}{2}\sqrt{n} < Z < \frac{\epsilon}{2}\sqrt{n}\right), \end{aligned}$$

où $Z \sim \mathcal{N}(0,1)$. On déduit

$$\lim_{n\to+\infty} P(10-\epsilon < \overline{Y}_n < 10+\epsilon) = \lim_{n\to+\infty} P\left(-\frac{\epsilon}{2}\sqrt{n} < Z < \frac{\epsilon}{2}\sqrt{n}\right) = P(-\infty < Z < +\infty) = 1.$$

(c) On rappelle que $\overline{Y}_n \to_p 10$ si et seulement si pour tout $\epsilon > 0$, on a $\lim_{n\to\infty} P(|\overline{Y}_n - 10| \leq \epsilon) = 1$. Or, on remarque que, par définition de la valeur absolue,

$$P(10-\epsilon < \overline{Y}_n < 10+\epsilon) = P(-\epsilon < \overline{Y}_n - 10 < \epsilon) = P(|\overline{Y}_n - 10| < \epsilon).$$

Le point (b) montre donc la convergence en probabilité souhaitée.

(d) Les hypothèses de la loi des grands nombres sont les suivantes :

1. Les variables aléatoires $Y_1, Y_2, \ldots$ sont i.i.d.
2. L'espérance $\mathbb{E}[|Y_1|]$ est finie.

1.3: Notons T_i le temps d'arrêt observé au jour i. Par hypothèse, $\mathbb{E}[T_i] = 4$ et Var$[T_i] = 0.8^2$.

(a) Les hypothèses du théorème central limite sont les suivantes :

1. Les variables aléatoires $T_1, T_2, \ldots$ sont i.i.d.
2. La variance Var$[T_1]$ est finie.

La seconde hypothèse est directement supposée dans l'énoncé. La première hypothèse doit être faite explicitement car rien ne garantit à priori que les temps observés sont i.i.d.

(b) Si l'on suppose que le théorème central limite s'applique, alors

$$\sqrt{n}\frac{\overline{T}_n - 4}{0.8} \to_d Z \sim N(0,1),$$

et on peut calculer approximativement que

$$\begin{aligned} P\left(1 \leq \overline{T}_{30} \leq 5\right) &= P\left(\sqrt{n}\frac{1-4}{0.8} \leq \sqrt{n}\frac{\overline{T}_{30}-4}{0.8} \leq \sqrt{n}\frac{5-4}{0.8}\right) \\ &\approx P\left(\sqrt{n}\frac{1-4}{0.8} \leq Z \leq \sqrt{n}\frac{5-4}{0.8}\right) \approx 1. \end{aligned}$$

1.2 À préparer pendant et après la séance

1.2.1 Probabilités

1.4: La variable aléatoire X est binomiale avec paramètres $n = 45$ et $p = 1/3$. L'espérance est donc $\mathbb{E}(X) = np = 15$ et la variance Var$[X] = np(1-p) = 10$.

1.5: C'est une v.a. exponentielle avec paramètre $\lambda = 1$. La probabilité se calcule

$$P(X > 1) = \int_1^\infty e^{-x}dx = [-e^{-x}]_1^\infty = [0 + e^{-1}] = 1/e.$$

1.6: On identifie les paramètres $\mu = 2$ et $\sigma^2 = 9$.

1.7:

$$Cov(X, Y) = Cov(X, X^2) = \mathbb{E}[X^3] - \mathbb{E}(X)\mathbb{E}(X^2).$$

Le deuxième terme est zéro puisque $\mathbb{E}(X) = 0$. Le premier terme est également zéro puisque c'est le coefficient de skewness d'une loi symétrique, qui vaut zéro. Il est évident que, dans ce cas, les deux v.a. ne sont pas indépendantes malgré qu'elles aient une corrélation égale à zéro.

Code R pour l'illustration numérique :

```
set.seed(2022) #permet que les nombres aleatoires generes ulterieurement soient identiques a chaque execution du code
X = rnorm(100)
Y = X^2
cov(X,Y)
```

1.8: Oui, comme le vecteur aléatoire (X, Y) est Gaussien, une corrélation nulle implique l'indépendance entre X et Y. Pour le voir, il suffit de vérifier que la densité jointe du vecteur aléatoire se factorise en le produit des densités marginales de X et de Y quand $\rho = 0$. Cela ne rentre pas en contradiction avec l'exercice précédent car dans celui-ci $Y = X^2$ n'est pas une variable aléatoire normale, donc le vecteur (X, Y) n'est certainement pas Gaussien.

1.9:

```
set.seed(2022) #permet que les nombres aleatoires generes ulterieurement soient identiques a chaque execution du code
X = rnorm(10000)
Y = cos(X)
hist(Y)
```

1.2.2 Loi des grands nombres

1.10:

(a) Oui, car on peut calculer que $\mathbb{E}[X_1] = 1$ et les hypothèses de la loi des grands nombres sont vérifiées, donc la convergence en probabilité annoncée est vraie.

(b) Soit $\epsilon > 0$. Comme on calcule que $\text{Var}[\overline{X}_n] = 1/n$, l'inégalité de Chebyshev assure que

$$P(|\overline{X}_n - 1| \geq \epsilon) \leq \frac{1}{n\epsilon^2} \xrightarrow{n\to\infty} 0.$$

Cela montre la convergence en probabilité $\overline{X}_n \to_p 1$.

1.11: Non. Les probabilités des évènements individuels ne changent pas, peu importe la valeur prise par les couleurs obtenues précédemment. Asymptotiquement les fréquences de "rouge" et "noir" doivent s'équilibrer par la loi des grands nombres, mais en aucun cas les probabilités ne sont modifiées.

1.12:

(a) Pour un nombre n fixé de lancers, on pourrait avoir la situation, certes très peu probable, que $n = n_A$. Dans ce cas, $n_B = 0$ et on a bien $n_A - n_B > 0$.

(b) Quand $n \to \infty$, la loi des grands nombres assurent que les fréquences empiriques convergent vers la vraie probabilité. Dès lors,

$$\frac{n_A}{n} \to_p P(A) = \frac{1}{36} \qquad \frac{n_B}{n} \to_p P(B) = \frac{2}{36}.$$

Il s'ensuite que

$$\frac{n_A}{n} - \frac{n_B}{n} \to_p -\frac{1}{36}.$$

1.2.3 Théorème central limite

1.13:

(a) Notons X_i les variables aléatoires considérées. La distribution des X_i n'est pas connue mais les variables sont supposées i.i.d. et

$$\mathbb{E}[X_1] = 14 \qquad \mathsf{Var}[X_1] = 4.$$

Par conséquent, on calcule,

$$\mathbb{E}[\overline{X}_n] = 14 \qquad \mathsf{Var}[\overline{X}_n] = 4/n.$$

Le théorème central limite nous assure que si n est assez grand, alors on a approximativement $\overline{X}_n \sim_a N(14, 4/n)$. On calcule alors,

$$\begin{aligned} P(\overline{X}_{100} > 14.5) &= P\left(\frac{\overline{X}_{100} - 14}{\sqrt{4/100}} > \frac{14.5 - 14}{\sqrt{4/100}}\right) \\ &\approx P(Z > 2.5) \text{ où } Z \sim \mathcal{N}(0,1) \\ &= 0.0062. \end{aligned}$$

(b) Nous voulons trouver $U, L \in \mathbb{R}$ tels que

$$P(L < \overline{X}_{100} < U) = 0.95.$$

Comme nous savons que la loi de $\overline{X}_{100}$ devrait être proche d'une loi normale par le TCL, nous avons

$$0.95 = P(L < \overline{X}_{100} < U) \approx P\left(\frac{L - 14}{\sqrt{4/100}} < Z < \frac{U - 14}{\sqrt{4/100}}\right),$$

où $Z \sim \mathcal{N}(0,1)$. En cherchant dans la table statistique de la loi normale centrée réduite les quantiles [1] inférieurs d'ordre 0.025 et 0.975, on trouve que

$$\frac{L - 14}{\sqrt{4/100}} = -1.96 \text{ et } \frac{U - 14}{\sqrt{4/100}} = 1.96.$$

1. Ce choix est motivé par la notion d'intervalle de confiance qui sera abordée plus tard dans le cours. En effet, on souhaite que $P\left(\frac{L-14}{\sqrt{4/100}} > Z\right) = \frac{1-0.95}{2} = 0.025$ et que $P\left(Z > \frac{U-14}{\sqrt{4/100}}\right) = 0.025$.

Notons qu'en choisissant ces quantiles, nous avons bien

$$P\left(\frac{L-14}{\sqrt{4/100}} < Z < \frac{U-14}{\sqrt{4/100}}\right) = 0.95.$$

En résolvant les équations, on trouve $U = 14.392$ et $L = -13.608$.

1.14: Notons $X_1, \ldots, X_{100}$ les 100 tailles mesurées par l'anthropologue. Alors les X_i sont i.i.d. de distribution inconnue avec $\mathbb{E}[X_1] = \mu$ et Var$[X_1] = 2.5^2$. Nous cherchons

$$P(-0.5 < \overline{X}_{100} - \mu < 0.5)$$

où $\overline{X}_{100}$ satisfait

$$\mathbb{E}[\overline{X}_{100}] = \mu \text{ et Var}[\overline{X}_{100}] = \frac{2.5^2}{100}.$$

Par le théorème central limite, on sait que la distribution de $\overline{X}_{100}$ est proche de la loi normale, dès lors,

$$\begin{aligned} P(-0.5 < \overline{X}_{100} - \mu < 0.5) &\approx P\left(-\frac{0.5}{\sqrt{2.5^2/100}} < Z < \frac{0.5}{\sqrt{2.5^2/100}}\right) \text{ où } Z \sim \mathcal{N}(0,1) \\ &= P(-2 < Z < 2) \\ &\approx P(-1.96 < Z < 1.96) = 0.95. \end{aligned}$$

1.15: Notons Y_i la résistance en kg d'une chaise de son nouveau modèle prise au hasard. Par le théorème central limite, $\overline{Y}_n \sim_a N(\mu, \frac{\sigma^2}{n})$, avec $\sigma = 5$. Donc, asymptotiquement, on a aussi

$$Z = \frac{\overline{Y}_n - \mu}{5/\sqrt{n}} \sim_a N(0,1).$$

On peut chercher avec R ou avec les tables les quantiles de la normale centrée réduite pour trouver que

$$P(-1.96 \leq Z \leq 1.96) = 0.95.$$

Ceci implique que

$$P(|\overline{Y}_n - \mu| \leq 1.96\frac{5}{\sqrt{n}}) = 0.95.$$

Il est demandé qu'à cette probabilité de 95%, la marge d'erreur soit bornée par 1, c'est-à-dire

$$1.96\frac{5}{\sqrt{n}} \leq 1$$

ce qui donne le résultat $n \geq 97$.

1.16: Si les travailleurs de ce groupe ethnique ont un salaire similaire aux autres employés de l'entreprise et si leurs salaires sont notés $X_1, \ldots, X_{64}$, alors on devrait avoir

$$\mathbb{E}[X_1] = 7 \qquad \text{et Var}[X_1] = 0.5^2.$$

Par le théorème central limite on peut calculer, par la même méthode que dans les exercices précédents, que si cette hypothèse est vraie,

$$P(\overline{X}_{100} \leq 6.90) = 0.0548.$$

Cela veut donc dire que, sous l'hypothèse que les salaires proviennent de la même distribution, la probabilité d'observer un salaire moyen dans le groupe ethnique plus faible que celui observé est seulement d'approximativement 5%. Autrement dit, la valeur observée est peu probable sous l'hypothèse que les salaires sont similaires. Le calcul effectué ici correspond au calcul d'une p-valeur unilatérale, une notion qui sera abordée ultérieurement dans ce cours.

1.17:

(a) Non, les concentrations sont des nombres strictement positifs et ne suivent donc certainement pas une loi normale. Par contre, comme la variance est supposée finie, si l'on récolte un échantillon i.i.d. $C_1, \ldots, C_n$ de concentrations, alors $\overline{C}_n$ suit approximativement une distribution normale si n est assez grand, en vertû du théorème central limite.

(b) Un calcul tout à fait similaire aux exercices précédents permet de montrer, par le théorème central limite,

$$P(\overline{C}_{100} > 14) = 0.0132.$$

1.18:

(a) Utilisant l'indice, on sait que $\sum_{i=1}^{10} X_i \sim \textit{Gamma}(10, 1)$. Dès lors, la probabilité peut être calculée de manière exacte car

$$P(\overline{X}_{10} > 1.5) = P\left(\sum_{i=1}^{10} X_i > 15\right)$$

et cette dernière probabilité peut être calculée via la fonction `pgamma` de R. Cela donne `pgamma(15,10,1,lower.tail=FALSE)` $= 0.06985366$.

(b) Comme les $X_1, \ldots, X_n$ sont i.i.d. de variance finie $\text{Var}[X_1] = 1$, on peut appliquer le théorème central limite qui assure que

$$P(\overline{X}_{10} > 1.5) = P\left(\sqrt{10}\frac{\overline{X}_{10} - 1}{1} > \sqrt{10}\frac{1.5 - 1}{1}\right) \approx P(Z > 0.5\sqrt{10}),$$

où $Z \sim \mathcal{N}(0, 1)$. La fonction `pnorm` de `R` permet de calculer la dernière probabilité et donne 0.05692315. L'erreur absolue entre les probabilité calculées au (a) et au (b) est alors de 0.01293051, approximativement 1.3%.

(c) En suivant exactement le même raisonnement que dans le cas $n = 10$, on calcule que la probabilité exacte est 0.02187347 et l'approximation du théorème central limite donne 0.01267366. L'erreur absolue est maintenant inférieure à 1%. L'approximation est devenue meilleure, comme attendu en vue du caractère asymptotique du théorème central limite.

Chapitre 2

Estimation

2.1 À préparer avant la séance

2.1: Notons que

$$\mathbb{E}[\bar{X}] = \frac{1}{n}\sum_{i=}^{n} \mathbb{E}[X_i] = \mathbb{E}[X_1].$$

Il suffit de calculer l'espérance de X_1, en faisant la substitution $y = x - \theta$:

$$\begin{aligned}
\mathbb{E}[X_1] &= \int_{\theta}^{\infty} x e^{-(x-\theta)} dx \\
&= \int_{0}^{\infty} (y+\theta) e^{-y} dy \\
&= \underbrace{\int_{0}^{\infty} y e^{-y} dy}_{1} + \theta \underbrace{\int_{0}^{\infty} e^{-y} dy}_{1} \\
&= 1 + \theta,
\end{aligned}$$

ce qui démontre le résultat.

2.2: Nous obtenons pour la différence des moyennes :

$$\bar{Y}_A - \bar{Y}_B = 0.006666667 - 0.01 = -0.003333333$$

et pour la différence des variances :

$$S_A^2 - S_B^2 = 0.004522222 - 0.000566667 = 0.003955556$$

L'estimateur de la différence des moyennes est non biaisé (il suffit d'utiliser la linéarité de l'espérance). L'estimateur de la différence des variance est quant à lui biaisé et son biais est donné par

$$B(S_A^2 - S_B^2) = \mathbb{E}[S_A^2 - S_B^2] - (\sigma_A^2 - \sigma_B^2) = -\frac{1}{6}(\sigma_A^2 - \sigma_B^2).$$

Le signe de ce biais dépend de la différence des vraies variances, inconnues. Cette différence peut être positive ou négative. Nous ne pouvons donc pas en déduire le signe du biais.

Pour la préférence de l'investissement, notons que l'actif B possède un rendement moyen (estimé) supérieur à celui de l'actif A, et en même temps une variance plus faible. Si on est averse au risque (et c'est typiquement le cas), on préférerait ici un investissement dans l'actif B.

2.3: La différence des proportions est estimée par

$$\widehat{p}_H - \widehat{p}_F = \frac{157}{200} - \frac{172}{200} = -0.075.$$

La variance de cet estimateur est estimée par

$$\frac{\widehat{p}_H(1-\widehat{p}_H)}{200} + \frac{\widehat{p}_F(1-\widehat{p}_F)}{200} = 0.0008439 + 0.000602 = 0.0014459$$

2.2 À préparer pendant et après la séance

2.4: On obtient $\bar{Y} = 1.816$ et

$$S^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \bar{Y}^2 = 3.306 - 1.816^2 = 0.008224.$$

Cet estimateur est biaisé. Un estimateur non-biaisé serait :

$$(S')^2 = \frac{n}{n-1}S^2 = \frac{5}{4}0.008224 = 0.01028.$$

Si la taille est mesurée en centimètres, S^2 devient :

$$S^2 = (100)^2 0.008224 = 82.24$$

2.5: D'abord, le biais vaut

$$B(S^2) = -\frac{\sigma^2}{n} = -\frac{1}{20}$$

Ensuite, la variance selon l'approximation

$$\mathbb{V}\mathrm{ar}(S^2) \approx \frac{\sigma^4(K-1)}{n} = \frac{2}{20} = \frac{1}{10}$$

ce qui donne l'erreur moyenne quadratique

$$MSE(S^2) = B(S^2)^2 + \mathbb{V}\mathrm{ar}(S^2) = \frac{1}{400} + \frac{1}{10} = 0.1025$$

2.6: L'estimateur $\hat{p}$ est non-biaisé avec variance

$$\mathbb{V}\mathrm{ar}(\hat{p}) = \frac{p(1-p)}{n}$$

où n est le nombre total de personnes dans l'échantillon. L'autre estimateur proposé est

$$\hat{p}_h = \frac{f_h}{n_h}$$

où n_h est le nombre d'hommes, et f_h le nombre d'hommes qui fument. En supposant que la proportion d'hommes fumeurs dans la population est égale à celle des femmes fumeuses, c'est également un estimateur non-biaisé. Par contre, sa variance vaut

$$\mathbb{V}\mathrm{ar}(\hat{p}_h) = \frac{p(1-p)}{n_h} = 2\frac{p(1-p)}{n}$$

puisque $n_h = n/2$ par hypothèse. L'efficacité relative se calcule alors

$$\mathsf{eff}(\hat{p}_h, \hat{p}) = \frac{\mathbb{V}\mathrm{ar}(\hat{p})}{\mathbb{V}\mathrm{ar}(\hat{p}_h)} = 1/2.$$

L'estimateur $\hat{p}_h$ a une efficacité relative par rapport à $\hat{p}$ de 0.5, ou 50%, il est donc inefficace. Ceci s'explique par le fait qu'il n'utilise que la moitié de l'échantillon, les hommes, ce qui correspond à une perte d'information.

2.7:

(a) Rappelons que la MSE se décompose comme la somme du biais au carré et de la variance. Pour $\hat{\mu}_1 = \overline{T}_n$, on calcule

$$\mathbb{E}[\hat{\mu}_1] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} T_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[T_i] = \frac{n\mu}{n} = \mu,$$

où on a utilisé le caractère i.i.d. des T_i et, similairement,

$$\mathbb{V}\mathrm{ar}[\hat{\mu}_1] = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\mathrm{ar}[T_i] = \frac{\sigma^2}{n}.$$

On conclut que l'estimateur est non biaisé et que la MSE de $\hat{\mu}_1$ est donnée par sa variance, i.e.

$$\mathrm{MSE}[\hat{\mu}_1] = \mathbb{V}\mathrm{ar}[\hat{\mu}_1] = \frac{\sigma^2}{n}.$$

Pour $\hat{\mu}_2 = 10$, on calcule de manière immédiate (l'estimateur n'est pas stochastique)

$$\mathbb{E}[\hat{\mu}_2] = 10$$

et

$$\mathbb{V}\mathrm{ar}[\hat{\mu}_2] = 0.$$

Il s'ensuit que la MSE de $\hat{\mu}_2$ est donné par son biais au carré, i.e.

$$\mathrm{MSE}[\hat{\mu}_2] = (10-\mu)^2.$$

Ensuite, on montre que l'estimateur $\hat{\mu}_1$ est consistant pour μ : une condition suffisante pour avoir la convergence en probabilité $\hat{\mu}_1 \to_p \mu$ est que

$$\mathbb{E}[\hat{\mu}_1] \to \mu \quad \text{et} \quad \mathbb{V}\mathrm{ar}[\hat{\mu}_1] \to 0, \text{ quand } n \to \infty,$$

ou, de manière équivalente,

$$\mathrm{MSE}[\hat{\mu}_1] \to 0, \text{ quand } n \to \infty.$$

Au regard des calculs effectués ci-dessus, ces conditions sont satisfaites. Par contre, observons que l'estimateur $\hat{\mu}_2$ n'est pas consistant pour μ.

(b) Pour $\mu = \sigma = 10$, on calcule directement grâce au point (a) que

$$\mathrm{MSE}[\hat{\mu}_1] = 1 \quad \text{et} \quad \mathrm{MSE}[\hat{\mu}_2] = 0.$$

Dans ce cas, l'estimateur $\hat{\mu}_2$ sera donc préféré à l'estimateur $\hat{\mu}_1$ car sa MSE est plus petite. Pour $\mu = 9$ et $\sigma = 5$, on a

$$\mathrm{MSE}[\hat{\mu}_1] = \frac{1}{4} \quad \text{et} \quad \mathrm{MSE}[\hat{\mu}_2] = 1.$$

Dans ce cas, le critère de la MSE est donc favorable à l'estimateur $\hat{\mu}_1$.

(c) L'estimateur $\hat{\mu}_2 = 10$ ne dépend pas des données collectées. Il ne peut dès lors pas être un bon estimateur de μ car, même si l'on disposait d'une information de plus en plus grande sur μ (i.e. $n \to \infty$), sa MSE ne diminuerait pas, à l'inverse de la MSE de $\hat{\mu}_1$. En fait, les seuls cas où l'estimateur $\hat{\mu}_2$ sera plus performant que $\hat{\mu}_1$ sont les situations où μ est très proche de 10 (ce qui n'est pas connu d'avance). Cela peut se vérifier algébriquement de la manière suivante : les seules valeurs de $\mu \in \mathbb{R}$ telles que

$$\mathrm{MSE}[\hat{\mu}_2] < \mathrm{MSE}[\hat{\mu}_1]$$

sont les valeurs qui satisfont

$$(10 - \mu)^2 < \frac{\sigma^2}{n},$$

ou, de manière équivalente,

$$\mu \in \left[10 - \frac{\sigma}{\sqrt{n}}, 10 + \frac{\sigma}{\sqrt{n}}\right],$$

une région qui tend vers le singleton $\{\mu = 10\}$ quand $n \to \infty$.

2.8:

(a) Rappelons que la MSE satisfait la décomposition fondamentale suivante

$$\mathrm{MSE}[\hat{F}_n(2)] = B[\hat{F}_n(2)]^2 + \mathbb{V}\mathrm{ar}[\hat{F}_n(2)].$$

On calcule chaque terme indépendamment. Pour le biais, on a

$$\begin{aligned}
\mathbb{E}[\hat{F}_n(2)] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} 1\{X_i \leq 2\}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[1\{X_i \leq 2\}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} P(X_1 \leq 2) = F(2),
\end{aligned}$$

où la dernière ligne suit du caractère i.i.d. des données X_i et du fait que la variable aléatoire $1\{X_i \leq 2\}$ suit une loi de Bernouilli $Be(F(2))$. On déduit que $B[\hat{F}_n(2)]^2 = 0$.

Pour la variance, on a, de manière similaire,

$$\begin{aligned}\mathbb{V}\mathrm{ar}[\hat{F}_n(2)] &= \mathbb{V}\mathrm{ar}\left[\frac{1}{n}\sum_{i=1}^{n} 1\{X_i \leq 2\}\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}\mathrm{ar}\left[1\{X_i \leq 2\}\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n} F(2)(1-F(2)) = \frac{F(2)(1-F(2))}{n}.\end{aligned}$$

On déduit que

$$\mathrm{MSE}[\hat{F}_n(2)] = \frac{F(2)(1-F(2))}{n}.$$

(b) La consistance de l'estimateur $\hat{F}_n(2)$ suit de la condition suffisante $\mathrm{MSE}[\hat{F}_n(2)] \to 0$ quand $n \to \infty$.

(c) L'estimateur $\hat{F}_n(2)$ est de la forme $\overline{Y}_n$ pour les variables i.i.d. $Y_i := 1\{X_i \leq 2\} \sim Be(F(2))$. Il suit donc du TCL (et des calculs précédents de la moyenne et de la variance de l'estimateur) que

$$\sqrt{n}\big(\hat{F}_n(2) - F(2)\big) \to_d N(0, F(2)(1-F(2))).$$

De manière équivalente,

$$\hat{F}_n(2) \sim_a N\left(F(2), \frac{F(2)(1-F(2))}{n}\right).$$

2.9: (a) Nous avions déjà montré que $\mathbb{E}(\hat{\mu}_1) = \mu$, et donc $B(\hat{\mu}_1) = 0$. Ensuite,

$$\begin{aligned}\mathbb{E}[\hat{\mu}_2] &= \mathbb{E}\left[\frac{1}{2n}(Y_1 + Y_n) + \frac{n-1}{n(n-2)}\sum_{i=2}^{n-1} Y_i\right] \\ &= \frac{1}{2n}\underbrace{\mathbb{E}(Y_1 + Y_n)}_{2\mu} + \frac{n-1}{n(n-2)}\sum_{i=2}^{n-1}\underbrace{\mathbb{E}(Y_i)}_{\mu} \\ &= \mu\frac{1}{n} + \frac{n-1}{n(n-2)}(n-2)\mu = \mu\end{aligned}$$

et nous obtenons que cet estimateur est également sans biais : $B(\hat{\mu}_2) = 0$.

(b) Calculons les variances des estimateurs :

$$\begin{aligned}\mathbb{V}\mathrm{ar}[\hat{\mu}_1] &= \frac{5}{3} \\ \mathbb{V}\mathrm{ar}[\hat{\mu}_2] &= \frac{1}{4n^2}(\underbrace{\mathbb{V}\mathrm{ar}(Y_1)}_{5} + \underbrace{\mathbb{V}\mathrm{ar}(Y_n)}_{5}) + \left(\frac{n-1}{n(n-2)}\right)^2 \sum_{i=2}^{n-1}\underbrace{\mathbb{V}\mathrm{ar}(Y_i)}_{5} \\ &= 5\left\{\frac{1}{2n^2} + \left(\frac{n-1}{n(n-2)}\right)^2 (n-2)\right\} \\ &= 5\left(\frac{1}{2n^2} + \frac{(n-1)^2}{n^2(n-2)}\right)\end{aligned}$$

On obtient

$$\mathbb{V}\mathrm{ar}[\hat{\mu}_2] = \begin{cases} 5\left(\frac{1}{2\times 9} + \frac{2^2}{3^2}\right) = \frac{5}{2}, & n = 3 \\ 5\left(\frac{1}{2\times 16} + \frac{3^2}{4^2\times 2^2}\right) = \frac{55}{64}, & n = 4 \\ 5\left(\frac{1}{2\times 25} + \frac{4^2}{5^2\times 3^2}\right) = \frac{41}{90}, & n = 5 \end{cases}$$

et pour les efficacités relatifs :

$$eff(\hat{\mu}_2, \hat{\mu}_1) = \begin{cases} \frac{5/3}{5/2} = \frac{2}{3} < 1, & n = 3 \\ \frac{5/3}{55/64} = \frac{64}{33} > 1, & n = 4 \\ \frac{5/3}{41/90} = \frac{450}{123} > 1, & n = 5 \end{cases}$$

Nous observons que pour $n = 3$, l'estimateur $\hat{\mu}_1$ est plus efficace, tandis que pour $n \geq 4$ c'est $\hat{\mu}_2$ qui devient plus efficace. L'efficacité relative augmentera vers l'infinie lorsque $n \to \infty$.

(c) L'estimateur $\hat{\mu}_1$ n'est pas consistant. En effet, l'estimateur $\hat{\mu}_1$ n'est pas fonction de n, donc, pour n'importe quel $\epsilon > 0$, il n'y a aucune chance que la probabilité $P(|\hat{\mu}_1 - \mu| \geq \epsilon) \to 0$ quand $n \to \infty$. Dès lors, on a pas la convergence en probabilité $\hat{\mu}_1 \to_p \mu$ et l'estimateur n'est pas consistant.

L'estimateur $\hat{\mu}_2$ est consistant puisqu'il est non-biaisé et sa variance tend vers zéro lorsque $n \to \infty$.

2.10:

(a) On propose l'estimateur empirique

$$\hat{p}_N - \hat{p}_S = \frac{\sum_{i=1}^{n_N} N_i}{n_N} - \frac{\sum_{i=1}^{n_S} S_i}{n_S},$$

n_N (resp. n_S) correspond au nombre d'employés nicaraguayens (resp. suédois) et les N_i (resp. S_i) sont des variables aléatoires i.i.d. qui valent 1 si l'employé i parmi les employés nicaraguayens (resp. suédois) possède un balcon, 0 sinon.
La loi des grands nombres assure que

$$\hat{p}_N \to_p p_N, n_N \to \infty \text{ et } \hat{p}_S \to_p p_S, n_S \to \infty,$$

i.e. les estimateurs $\hat{p}_N$ et $\hat{p}_S$ sont consistants pour leur proportion respective. Il suit des propriétés P1 et P4 de la slide 2-23 que l'estimateur $\hat{p}_N - \hat{p}_S$ est consistant pour $p_N - p_S$.

(b) On vérifie facilement que l'estimateur $\hat{p}_N - \hat{p}_S$ est sans biais. En effet, notons que $N_i \sim Be(p_N)$ et $S_i \sim Be(p_S)$. Il s'ensuit que

$$\mathbb{E}[\hat{p}_N - \hat{p}_S] = \frac{1}{n_N}\sum_{i=1}^{n_N}\mathbb{E}[N_i] - \frac{1}{n_S}\sum_{i=1}^{n_S}\mathbb{E}[S_i] = p_N - p_S,$$

où on a utilisé le caractère identiquement distribué des N_i et S_i. Par conséquent,

$$\mathrm{MSE}[\hat{p}_N - \hat{p}_S] = \mathbb{V}\mathrm{ar}[\hat{p}_N - \hat{p}_S] = \mathbb{V}\mathrm{ar}[\hat{p}_N] + \mathbb{V}\mathrm{ar}[\hat{p}_S] = \frac{p_N(1-p_N)}{n_N} + \frac{p_S(1-p_S)}{n_S}.$$

(c) Si l'on suppose que $p_S = p_N$, il suit du théorème central limite que

$$\frac{\hat{p}_N - \hat{p}_S - (p_N - p_S)}{\sqrt{\frac{p_N(1-p_N)}{n_N} + \frac{p_S(1-p_S)}{n_S}}} = \frac{\hat{p}_N - \hat{p}_S}{\sqrt{\frac{p_N(1-p_N)}{n_N} + \frac{p_S(1-p_S)}{n_S}}} \to_d Z \sim N(0,1).$$

Si l'on souhaite pouvoir calculer la probabilité $P(\hat{p}_N - \hat{p}_S \leq -1/3)$ en utilisant cette approximation, il faut que le dénominateur de cette dernière quantité soit connu. Ce n'est pas le cas ici car p_N et p_S sont inconnus. En fait, cela n'est pas un problème car on peut les remplacer par leur estimateur respectif $\hat{p}_N$ et $\hat{p}_S$ sans compromettre la convergence en distribution vers $Z \sim N(0,1)$. Cela est une conséquence du théorème de Slutsky (slide 2-25). On a donc

$$\begin{aligned} P(\hat{p}_N - \hat{p}_S \leq -1/3) &= P\left(\frac{\hat{p}_N - \hat{p}_S)}{\sqrt{\frac{\hat{p}_N(1-\hat{p}_N)}{n_N} + \frac{\hat{p}_S(1-\hat{p}_S)}{n_S}}} \leq \frac{-1/3}{\sqrt{\frac{\hat{p}_N(1-\hat{p}_N)}{n_N} + \frac{\hat{p}_S(1-\hat{p}_S)}{n_S}}} \right) \\ &\approx \Phi\left(\frac{-1/3}{\sqrt{\frac{\hat{p}_N(1-\hat{p}_N)}{n_N} + \frac{\hat{p}_S(1-\hat{p}_S)}{n_S}}} \right) \\ &= \Phi(-3.57), \end{aligned}$$

où Φ désigne la fonction de répartition d'une loi normale centrée réduite. La fonction `pnorm` de `R` permet de calculer $\Phi(-3.57) \approx 0.0001$. Autrement dit, la probabilité que notre estimateur de la différence de proportion soit plus petit que la valeur qu'on a observée dans notre échantillon est très faible.
Cela veut dire que, sous l'hypothèse d'égalité des traitements (i.e. $p_S = p_N$), la vraisemblance de nos données est très petite. On peut donc penser que le patron de l'entreprise ne traite pas de manière égale l'ensemble de ses employés.

Chapitre 3

Intervalles de confiance

3.1 À préparer avant la séance

3.1: (a)

$$\bar{x}_n = \frac{0 \times 223 + 1 \times 135 + \cdots + 5 \times 14}{500} = 1.084$$
$$s_n^2 = \frac{1}{500-1}\left(223 \times (0 - 1.084)^2 + \cdots + 37 \times (3 - 1.084)^2\right) = 1.72$$
d'où $s_n = 1.31$

(b)

$$\begin{aligned} IC_{92\%}(\mu) &= \left[\bar{x}_n \pm z_{0.08/2} \times \frac{s_n}{\sqrt{n}}\right] \\ &= \left[1.084 \pm z_{0.04} \times \frac{1.31}{\sqrt{500}}\right] \\ &= \left[1.084 \pm 1.76 \times \frac{1.31}{\sqrt{500}}\right] \\ &= [0.98, 1.19] \end{aligned}$$

(c) On a $\hat{p} = \frac{37+22+14}{500} = \frac{73}{500} = 0.146$. D'où

$$\begin{aligned} IC_{97\%}(p) &= \left[\hat{p} \pm z_{0.03/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] \\ &= \left[0.146 \pm z_{0.015}\sqrt{\frac{0.146 \times (1-0.146)}{500}}\right] \\ &= [0.11, 0.18] \end{aligned}$$

(d)

$$2.17 \times \sqrt{\frac{0.146 \times (1-0.146)}{n}} = 0.01 \Rightarrow n = 5871.25 \approx 5872$$

3.2: Le $01/10$ on a $\widehat{p} = 75/130 = 0.58$. Le $15/10$ on a $\widehat{p} = 642/1056 = 0.61$. Ainsi, les intervalles de confiance de niveau 95% associés à ces estimations sont calculables via la formule

$$IC_{95\%}(p) = \left[\hat{p} \pm z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right].$$

Pour le $01/10$, on trouve

$$\begin{aligned} IC_{95\%}(p) &= \left[0.58 \pm 1.96\sqrt{\frac{0.58 \times (1-0.58)}{130}}\right] \\ &= [0.492, 0.662], \end{aligned}$$

et pour le $15/10$, on trouve

$$IC_{95\%}(p) = [0.58, 0.64].$$

Bien que les estimations ponctuelles des proportions d'électeurs en faveur de la proposition sont assez similaires aux deux dates, ce qui change est principalement l'incertitude associée à ces estimations dû à l'augmentation de la taille d'échantillon lors du deuxième sondage.

3.2 À préparer pendant et après la séance

3.3: On a $\hat{p} = \frac{42}{60} = 0.7$. D'où

$$\begin{aligned} IC_{99\%}(p) &= \left[\hat{p} \pm z_{0.005} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] \\ &= \left[0.7 \pm 2.576\sqrt{\frac{0.7 \times (1-0.7)}{60}}\right] \\ &= [0.55, 0.85]. \end{aligned}$$

3.4: L'estimateur ponctuelle de la moyenne est $\bar{Y} = 425$. Un intervalle de confiance au niveau $1 - \alpha = 95\%$ est donné par

$$[\bar{Y} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}] = [425 \pm 1.96\frac{17}{\sqrt{50}}] = [420.3; 429.7].$$

Notez qu'ici l'écart type de la distribution sous-jacente était supposé connu (17), sinon on aurait dû l'estimer.

La valeur supposée de 450g des pots de café n'est pas compris dans l'intervalle de confiance. Nous verrons au chapitre 5 que ceci est équivalent à un rejet de l'hypothèse que le vrai poids moyen des pots est égal à 450g.

3.5: On importe d'abord les données en suivant la procédure indiquée dans l'indice. Les données sont alors stockées dans l'objet R `data`.

(a) Pour rappel, la formule pour un intervalle de confiance d'une moyenne à niveau $1-\alpha$ avec écart-type inconnu est $IC_{1-\alpha}(\mu) = \left[\bar{X} - z_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right]$ où S est l'estimateur (biaisé) de l'écart-type. Dans notre cas, on a $n = 49$ et $\alpha = 0.02$, et on calcule
— $\bar{X}$ = `mean(data)` = 12.12 ;
— $z_{0.01}$ = `qnorm(0.01,lower.tail=FALSE)` = 2.33 ;
— S = `sqrt(48/49*var(data))` = 4.65.
On déduit que l'IC est donné par $[10.58, 13.67]$.
L'intervalle de confiance est basé sur le TCL et le théorème de Slutsky qui garantissent que

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \to_d N(0,1).$$

L'application de ces résultats suppose que :

(i) Les PER sont i.i.d. Cela n'est pas évident car cela suppose notamment que les PER des 49 actions ne dépendent aucunement les uns des autres. Cela est difficile à croire si, par exemple, certaines actions correspondent à des entreprises évoluant sur un même marché.

(ii) La variance des PER est finie. Cela peut ne pas être le cas si la queue de la distribution des PER s'avère être trop épaisse (trop de masse de probabilité aux extrémités), au quel cas nous ne pouvons plus nous baser sur le TCL pour construire l'IC.

(b) Pour rappel, la formule pour un intervalle de confiance d'une variance à niveau $1-\alpha$ avec kurtosis inconnu est donnée $IC_{1-\alpha}(\sigma^2) = \left[S^2 - z_{\frac{\alpha}{2}}\sqrt{\frac{S^4(\widehat{K}-1)}{n}}, S^2 + z_{\frac{\alpha}{2}}\sqrt{\frac{S^4(\widehat{K}-1)}{n}}\right]$ où S^2 est l'estimateur (biaisé) de la variance et $\widehat{K}$ est l'estimateur du kurtosis. Pour calculer ce dernier, il n'y a pas de fonction de base similaire à `var()` directement disponible en R. On utilise alors la fonction `kurtosis()` disponbible dans le package `moments` après installation (`install.packages("moments")`) et chargement de ce dernier (`library(moments)`). On calcule alors
— S^2 = `48/49*var(data)` = 21.64 ;
— $z_{0.025}$ = `qnorm(0.025,lower.tail=FALSE)` = 1.96 ;
— $\widehat{K}$ = `kurtosis(data)` = 11.41.
On déduit que l'IC à 95% est donné par $[2.09, 41.19]$. L'incertitude associée à l'estimation de la variance est plus grande et car la taille d'échantillon est relativement faible par rapport à la variabilité conséquente des données.
L'intervalle de confiance est basé sur le TCL et le théorème de Slutsky qui garantissent que

$$\frac{S^2 - \sigma^2}{\sqrt{S^4(\widehat{K}-1)/n}} \to_d N(0,1).$$

Comme pour le point précédent, l'application de ces résultats suppose le caractère i.i.d. des données. De plus, comme les données apparaissent au carré dans l'estimateur S^2, l'application du TCL nécessite également que $\mathbb{E}[X^4] < \infty$, une condition plus difficile à satisfaire que $\mathbb{E}[X^2] < \infty$.

3.6: (a) On estime la proportion en utilisant l'estimateur empirique (sur l'échantillon total) donné par

$$\hat{p} = \frac{27 + 35}{100} = 0.62.$$

Pour trouver l'intervalle à 95%, on utilise la formule du cours (slide 3-7) avec $\theta = p$ et $\sigma_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. On trouve

$$\text{IC}_{95\%}(p) = \left[\hat{p} \pm z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right].$$

En remplaçant $\hat{p} = 0.62$ et $n = 100$, on trouve

$$\text{IC}_{95\%}(p) = [0.52, 0.72].$$

La taille augmente si $\alpha = 1\%$ car $z_{\alpha/2}$ augmente (on passe de 1.96 à 2.576).

(b) La différence des proportions s'estime simplement en utilisant la différence des estimateurs empiriques au sein de chaque groupe, i.e., elle est estimée par

$$\hat{p}_H - \hat{p}_F = \frac{35}{50} - \frac{27}{50} = 16\%.$$

Ensuite, l'idée est exactement la même qu'au (a) : on applique la même formule avec $\theta = p_H - p_F$ et $\sigma_{\hat{p}_H - \hat{p}_F} = \sqrt{\frac{\hat{p}_H(1-\hat{p}_H)}{n_H} + \frac{\hat{p}_F(1-\hat{p}_F)}{n_F}}$. Cela donne l'intervalle :

$$\text{IC}_{95\%}(p_H - p_F) = \left[(\hat{p}_H - \hat{p}_F) \pm z_{0.025}\sqrt{\frac{\hat{p}_H(1-\hat{p}_H)}{n_H} + \frac{\hat{p}_F(1-\hat{p}_F)}{n_F}}\right].$$

En remplaçant les quantités par leur valeur observée, on trouve

$$\text{IC}_{95\%}(p_H - p_F) = [-0.03, 0.35].$$

Si l'on avait effectué l'enquête avec $n_H = n_F = 70$ et que les estimations pour les proportions restaient les mêmes, la taille de l'intervalle de confiance aurait diminuée car la variance empirique $\sigma_{\hat{p}_H - \hat{p}_F}$ aurait diminué.

3.7: (a) On applique simplement la formule du cours (slide 3-7) avec $\theta = \mu$ et $\sigma_{\hat{\mu}} = \frac{\hat{\sigma}}{\sqrt{n}}$ et $\alpha = 5\%$, i.e.

$$\text{IC}_{95\%}(\mu) = \left[\hat{\mu} \pm z_{0.025}\frac{\hat{\sigma}}{\sqrt{n}}\right].$$

En remplaçant les éléments par leur version calculée sur l'échantillon, on trouve

$$\text{IC}_{95\%}(\mu) = [1.43, 1.63].$$

L'intervalle contient la valeur 1.5 mm. Nous verrons au chapitre 5 que cela implique le non-rejet de l'hypothèse nulle dans un test bilatéral à 95% donné par

$$\begin{cases} H_0 & : \mu = 1.5 \\ H_1 & : \mu \neq 1.5. \end{cases}$$

(b) On sait du cours (slide 3-11) qu'un intervalle de confiance à 99% pour la variance est donné par

$$\mathsf{IC}_{99\%}(\sigma^2) = \left[\hat{\sigma}^2 \pm z_{0.005}\sqrt{\frac{\hat{\sigma}^4(\hat{K}-1)}{n}}\right].$$

En remplaçant les éléments par leur version calculée sur l'échantillon, on trouve

$$\mathsf{IC}_{99\%}(\sigma^2) = [0.22, 0.28].$$

3.8:

$$\begin{aligned} IC_{95\%}(\mu) &= \left[\hat{\mu} \pm z_{0.025} \times \frac{\hat{\sigma}}{\sqrt{n}}\right] \\ &= \left[1274 \pm 1.96 \times \frac{326}{\sqrt{100}}\right] \\ &= [1210.1, 1337.9] \end{aligned}$$

Chapitre 4

Propriétés des estimateurs et méthodes d'estimation

4.1 À préparer avant la séance

4.1: (a) Ici nous avons un seul paramètre à estimer donc il suffit de calculer le premier moment μ_1'. Ce dernier est égal à $E[X] = \frac{(\theta-1)+(\theta+1)}{2} = \theta$. D'autre part le premier moment de l'échantillon m_1' est égal à $\frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$. La méthode des moments donne donc $\theta = \bar{X}$ et donc l'estimateur $\hat{\theta} = \bar{X}$.

(b) Avec les données de l'énoncé, nous obtenons $\hat{\theta} = 12.492$.

4.2: (a) La vraisemblance est donnée par

$$\begin{aligned} L(p) = \prod_{i=1}^{n} f(y_i) &= \prod_{i=1}^{n} p(1-p)^{y_i-1} \\ &= p^n (1-p)^{\sum y_i - n} \end{aligned}$$

La log-vraisemblance est donnée par

$$\begin{aligned} \ell(p) = \ln L(p) &= \ln p^n + \ln\left((1-p)^{\sum y_i - n}\right) \\ &= n \ln p + (\sum y_i - n) \times \ln(1-p) \end{aligned}$$

La dérivée en p nous donne

$$\frac{d\ell(p)}{dp} = \frac{n}{p} - \frac{\sum y_i - n}{1-p}$$

et elle s'annule en $\hat{p} = \frac{n}{\sum y_i}$

(b) Avec ces données, on trouve $\hat{p} = 0.19$.

(c) Le paramètre p de la distribution géométrique de Y est la probabilité de succès d'un lancer individuel. Il est en fait connu et vaut $p = 1/6 = 1.666\ldots$ qui est relativement proche du MLE obtenu au point (a) calculé sur les 5 données du point (b).

4.2 À préparer pendant et après la séance

4.3: (a) Voir slides 4-9, 4-10 : $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

(b) Considérons la fonction bijective $g(x) = 1 - x$. La propriété d'invariance nous dit que $g(\hat{p}) = \frac{n-\sum_{i=1}^{n} Y_i}{n}$ est le maximum de vraisemblance de $g(p) = 1 - p$.

4.4: Introduisons la notation suivante :

$$\mathbf{1}\{x \geq \theta\} = \begin{cases} 1 & \text{si } x \geq \theta \\ 0 & \text{sinon,} \end{cases}$$

Cette notation permet d'exprimer la fonction de densité plus simplement : $f(x) = e^{-(x-\theta)}\mathbf{1}\{x \geq \theta\}$. On appelle $\mathbf{1}\{A\}$ la fonction indicatrice de l'évènement A.

En ayant supposé avoir observé $X_1, ..., X_n$, la vraisemblance de θ s'écrit :

$$L(\theta) = \prod_{i=1}^{n} f(X_i) = \prod_{i=1}^{n} e^{-(x_i-\theta)}\mathbf{1}\{x_i \geq \theta\} = e^{-\sum_{i=1}^{n} X_i + n\theta}\mathbf{1}\{\theta \leq \min(X_1, \ldots, X_n)\},$$

où la dernière égalité vient du fait que $L(\theta) = 0$ dès que $\theta > X_i$ pour un certain X_i, autrement dit dès que $\theta > \min(X_1, \ldots, X_n)$.

Cette fonction n'étant pas différentiable en θ (à cause de la présence de la fonction indicatrice qui n'est pas continue), la maximisation de L ne peut se faire par la procédure habituelle. Au lieu de cela, on observe simplement que L est croissante en θ sur la partie $\theta \leq \min(X_1, \ldots, X_n)$ et que $L \equiv 0$ sur la partie $\theta > \min(X_1, \ldots, X_n)$. On déduit que le point où L est maximale est $\widehat{\theta} = \min(X_1, \ldots, X_n)$.

4.5: (a) On calcule

$$\mathbb{E}[X_1] = \int_{\mathbb{R}} x f(x) dx = \theta \int_0^1 x^\theta dx = \frac{\theta}{\theta + 1}.$$

La méthode des moments suppose alors de résoudre l'équation

$$\frac{\hat{\theta}_n}{\hat{\theta}_n + 1} = \overline{X}_n,$$

pour $\hat{\theta}_n$. Cela donne

$$\hat{\theta}_n = \frac{\overline{X}_n}{1 - \overline{X}_n}.$$

(b) Il suit de la loi des grands nombres que

$$\overline{X}_n \to_p \frac{\theta}{\theta + 1}.$$

On déduit de la propriété (P4) avec la fonction continue $x \in (0, 1) \mapsto g(x) = \frac{x}{1-x}$ que

$$\hat{\theta}_n = g(\overline{X}_n) \to_p g\left(\frac{\theta}{\theta + 1}\right) = \ldots = \theta.$$

4.6: (a) On calcule

$$\begin{aligned}\mathbb{E}[X_1] &= \sum_{n=1}^{\infty} nP(X_1 = n) = \sum_{n=1}^{\infty} n\frac{\lambda^n}{n!}\exp(-\lambda) \\ &= \exp(-\lambda)\lambda \sum_{n=1}^{\infty} \frac{\lambda^{n-1}}{(n-1)!} \\ &= \exp(-\lambda)\lambda \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = \exp(-\lambda)\lambda\exp(\lambda) = \lambda,\end{aligned}$$

où, à la dernière ligne, nous avons utilisé la définition de la fonction exponentielle en termes de sa série de Taylor, i.e.,

$$\exp(x) = \sum_{k=0}^{\infty} \frac{x^k}{k!}.$$

(b) On déduit directement de la méthode des moments qu'un estimateur de λ est donné par

$$\hat{\lambda}_n = \overline{X}_n.$$

4.7: La vraisemblance jointe de l'échantillon est donnée par

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\theta}} \exp\left(-\frac{1}{2\theta}(x_i - \theta)^2\right) = (2\pi\theta)^{-n/2} \exp\left(-\frac{1}{2\theta}\sum_{i=1}^{n}(x_i - \theta)^2\right).$$

On déduit que la log-vraisemblance jointe de l'échantillon est donnée par

$$\ell(\theta) = \ln L(\theta) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln(\theta) - \frac{1}{2\theta}\sum_{i=1}^{n}(x_i - \theta)^2,$$

L'estimateur de maximum de vraisemblance, noté $\hat{\theta}$, satisfait l'équation $\frac{d\ell}{d\theta}(\hat{\theta}) = 0$. On calcule pour tout $\theta > 0$,

$$\frac{d\ell}{d\theta}(\theta) = -\frac{n}{2\theta} + \frac{1}{2\theta^2}\sum_{i=1}^{n}(x_i - \theta)^2 + \frac{1}{\theta}\sum_{i=1}^{n}(x_i - \theta).$$

Dès lors, demander que la dérivée s'annule revient à demander que

$$-\frac{n}{2} + \frac{1}{2\theta}\sum_{i=1}^{n}(x_i - \theta)^2 + \sum_{i=1}^{n} x_i - n\theta = 0,$$

ce qui peut se réécrire, après simplification,

$$\theta^2 + \theta - \frac{1}{n}\sum_{i=1}^{n} x_i^2 = 0.$$

C'est une équation du second ordre pour θ dont la solution négative doit être rejetée car $\theta = \mathbb{V}\mathrm{ar}[X] > 0$. On conclut que

$$\hat{\theta} = \frac{-1 + \sqrt{1 + \frac{4}{n}\sum_{i=1}^{n} x_i^2}}{2}.$$

4.8: La vraisemblance est donnée par

$$L(a) = \frac{1}{a^n}\big(\prod_{i=1}^{n} x_i\big)\exp\big(-\frac{\sum_{i=1}^{n} x_i^2}{2a}\big)$$

On en déduit que la log-vraisemblance est donnée par

$$l(a) = \sum_{i=1}^{n} \ln x_i - \frac{\sum_{i=1}^{n} x_i^2}{2a} - n\ln a$$

La dérivée en a nous donne

$$\frac{dl(a)}{da} = \frac{\sum_{i=1}^{n} x_i^2}{2a^2} - \frac{n}{a}$$

et elle s'annule en $\hat{a} = (2n)^{-1}\sum_{i=1}^{n} x_i^2$.
En utilisant les données du problème, on obtient $\hat{a} = 2.418$

4.9: La fonction de densité de X_i est

$$f(x_i;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\mathsf{exp}\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right].$$

La fonction de vraisemblance associée est définie par :

$$L(\mu,\sigma) = \sigma^{-n}(2\pi)^{-n/2}\mathsf{exp}\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right].$$

et donc la log-vraisemblance est :

$$\ell(\mu,\sigma) = \ln L(\mu,\sigma) = -n\ln\sigma - \frac{n}{2}\ln(2\pi) - \frac{\sum(x_i-\mu)^2}{2\sigma^2}$$

La dérivée en μ vaut

$$\frac{d\ell(\mu,\sigma)}{d\mu} = \frac{1}{\sigma^2}\sum(x_i-\mu)$$

et s'annule en $\hat{\mu} = n^{-1}\sum_{i=1}^{n} x_i$.
En se basant sur l'échantillon donné, $\hat{\mu} = 181.7$ cm.

Chapitre 5

Tests d'hypothèses

5.1 À préparer avant la séance

5.1: (a) L'erreur de type I consiste à rejeter l'hypothèse nulle H_0 quand celle-ci est vraie. Dans notre cas, cela reviendrait à conclure que le médicament est strictement moins efficace que les 80% d'efficacité prévus par l'entreprise, alors que ce dernier chiffre était en fait juste.
Le α est la probabilité d'une telle erreur. On calcule

$$\alpha = P(RH_0|H_0) = P(Y \leq 30|p = 0.8).$$

Notez qu'ici, Y compte simplement le nombre de patients endormis après la prise du médicaments parmi les 40 participants. Donc $Y \sim Bin(40, p)$. Dès lors, α est simplement la probabilité qu'une variable aléatoire de distribution binomiale $Bin(40, 0.8)$ soit plus petite ou égale à 30. On peut calculer cette probabilité directement (soit en utilisant une table de la loi binomiale, soit en utilisant la fonction `pbinom` de `R`). On trouve

$$\alpha \approx 0.268.$$

Le test considéré donne une probabilité de commettre une erreur de type I d'environ 27%. Une autre possibilité pour effectuer le calcul de α est d'utiliser le Théorème Centrale Limite pour $\hat{p} = Y/40$ que nous connaissons bien. C'est cette procédure qu'il faudra utiliser à l'examen car vous ne disposerez pas de table de la loi binomiale. Le calcul se fait alors de la manière suivante :

$$\begin{aligned}
\alpha &= P(RH_0|H_0) = P(Y \leq 30|p = 0.8) \\
&= P(\hat{p} \leq 30/40|p = 0.8) = P(\hat{p} \leq 0.75|p = 0.8) \\
&= P\left(\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{40}}} \leq \frac{0.75 - p}{\sqrt{\frac{p(1-p)}{40}}}\Big| p = 0.8\right) \\
&\approx P(Z \leq -0.79),
\end{aligned}$$

avec $Z \sim N(0, 1)$, et cette dernière probabilité se calcul directement en utilisant la fonction `pnorm` en `R` ou via les tables. On trouve $\alpha \approx 0.21$, l'erreur provenant de la nature

asymptotique du Théorème Centrale Limite.

Remarque supplémentaire à titre informatif : il existe une autre méthode équivalente pour calculer le α. Il s'agit de l'approximation de la loi binomiale $Bin(n, p)$ (qui correspond à la distribution de Y) par une loi normale donnée par $N(np, np(1-p))$.

(b) L'erreur de type II consiste à ne pas rejeter l'hypothèse nulle H_0 quand celle-ci est fausse. Dans notre cas, cela reviendrait à conclure que les données ne permettent pas de rejeter les propos de l'entreprise alors que la vraie efficacité du médicament était strictement inférieur aux 80% annoncés.
Le β est la probabilité d'une telle erreur. Notez qu'ici $H_a = \{p : p < 0.8\}$ n'est pas simplement un singleton comme H_0, i.e., plusieurs valeurs de p sont possibles sous H_a. Il faut donc spécifier sous quel $p \in H_a$ on souhaite calculer β. On prend $p = 0.6$. Il vient alors

$$\beta_{p=0.6} = P(\overline{R}H_0|p = 0.6) = 1 - P(RH_0|p = 0.6) = 1 - P(Y \leq 30|p = 0.6).$$

De nouveau, il y a 2 possibilités : soit on utilise le résultat exact qui se base sur le fait que $Y \sim \textit{Bin}(40, p)$, soit on utilise le Théorème Centrale Limite pour $\hat{p} = Y/40$. Les calculs sont similaires à ceux du (a) et ne sont pas refaits en détails ici. La deuxième procédure (asymptotique) donne $\beta_{p=0.6} \approx 0.026$.

5.2: (a) Le chercheur souhaite montrer que $p_1 > p_2$. Dès lors, le test qu'il va effectuer est le suivant :

$$H_0 : p_1 = p_2 \quad \text{contre} \quad H_a : p_1 > p_2.$$

De manière équivalente, on effectue le test

$$H_0 : p_1 - p_2 = 0 \quad \text{contre} \quad H_a : p_1 - p_2 > 0.$$

(b) La p-valeur (unilatérale), notée p, est donnée par

$$p = P(T \geq t|H_0),$$

où T est la statistique de test et t est sa valeur observée sur l'échantillon. Ici, comme on test une différence de proportions, T est donné par

$$T = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}},$$

qui satisfait

$$T \rightarrow_d \mathcal{N}(0, 1) \quad \text{sous } H_0.$$

On calcule $t = 1.5$. Dès lors, on déduit (comme les échantillons sont grands, on peut supposer que T est normal)

$$p = P(T \geq 1.5|H_0) = 0.067.$$

Pour $\alpha = 5\%$, on ne rejette donc pas H_0 car $p > \alpha$. Il n'y a pas suffisamment de preuves dans les données pour conclure à une différence significative entre les deux proportions.

5.2 À préparer pendant et après la séance

5.3: (a) Par les considérations expliquées à la question 5.1, cela revient à chercher c tel que

$$P(Y \leq c|p = 0.8) = 0.01.$$

Pour cela, comme à la question précédente, on utilise le Théorème Centrale Limite. La probabilité du membre de gauche peut se réécrire de la manière suivante :

$$\begin{aligned} P(Y \leq c|p = 0.8) &= P(\hat{p} \leq c/40|p = 0.8) \\ &= P\left(\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{40}}} \leq \frac{c/40 - p}{\sqrt{\frac{p(1-p)}{40}}}\Big|p = 0.8\right) \\ &\approx P\left(Z \leq \sqrt{40}\frac{c/40 - 0.8}{\sqrt{0.8 \times 0.2}}\right). \end{aligned}$$

Comme on veut que cette dernière probabilité soit égale à 0.01, il faut donc (par définition de la fonction quantile Φ^{-1})

$$\sqrt{40}\frac{c/40 - 0.8}{\sqrt{0.8 \times 0.2}} = \Phi^{-1}(0.01).$$

En isolant c, on trouve alors

$$c = 40 \times \left(0.0632 \times \Phi^{-1}(0.01) + 0.8\right) \approx 26.12.$$

Bien entendu, comme c représente un nombre de patients en pratique, on prendra $c = 26$ en pratique.

(b) Le raisonnement est identique à celui de la question (b) de l'exercice précédent. On a

$$\beta_{p=0.6} = 1 - P(RH_0|p = 0.6) = 1 - P(Y \leq 26|p = 0.6).$$

On calcule, sur base de l'approximation normale,

$$\beta_{p=0.6} \approx 0.26.$$

La probabilité de commettre une erreur de type II a augmenté par rapport à la question 5.1 car notre nouveau test admet une probabilité de commettre une erreur de type I plus faible.

5.4: (a) Les étudiants ayant, jusque là, suivi un enseignement similaire il n'y a priori aucune raison de croire que l'une ou l'autre moyenne soit supérieure à celle de l'autre groupe. Dès lors, le test que nous souhaitons effectuer est le suivant :

$$H_0 : \mu_1 = \mu_2 \quad \text{contre} \quad H_a : \mu_1 \neq \mu_2,$$

où μ_1 est la moyenne du groupe des étudiants qui vont suivre un enseignement traditionnel et μ_2 la moyenne du groupe des étudiants vont suivre un enseignement plus interactif. Notez que le test peut se réécrire de la manière suivante

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{contre} \quad H_a : \mu_1 - \mu_2 \neq 0,$$

qui est une forme plus pratique pour les calculs, comme nous le verrons dans la suite.

(b) Il s'agit d'une alternative bilatérale.

(c) La statistique de test que nous utiliserons tester l'égalité entre deux moyennes est donnée par

$$T = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}},$$

où les S_i^2 et les n_i désignent, respectivement, la variance estimée et la taille du groupe i pour $i = 1, 2$. Sous H_0, il suit du théorème centrale limite et du théorème de Slutsky (car l'estimateur S_i^2 est consistant pour la vraie variance σ_i^2) que

$$T \to_d \mathcal{N}(0, 1).$$

Comme on sait que la p-valeur permet d'effectuer la décision liée à notre test d'hypothèse, on va commencer par calculer celle-ci pour ensuite décider au rejet ou non de H_0 avec $\alpha = 1\%$. La p-valeur bilatérale est donnée par

$$p = P(|T| \geq |t| \,|H_0) = P(T \geq |t| \,|H_0) + P(T \leq -|t| \,|H_0),$$

où t est la valeur observée de T calculée sur les données. Ici, on calcule que $t = 1.66$. Pour calculer la p-valeur du test nous avons besoin de la distribution de T sous H_0. Comme nous avons un échantillon suffisamment grand et grâce aux arguments précédents, nous pouvons ici supposer que sous H_0, on a $T \sim \mathcal{N}(0, 1)$. Dès lors, par symétrie de la loi normale, nous avons

$$p = 2P(T \geq 1.66|H_0) = 0.097.$$

Comme $p > \alpha = 1\%$, on conclut au non-rejet de H_0, i.e., les données ne permettent pas de conclure qu'une différence significative existe entre les moyennes des deux groupes.

5.5: (a) Pour une différence de moyenne, un intervalle de confiance de niveau 95% est donné par (voir cours)

$$\text{IC}_{95\%}(\mu_1 - \mu_2) = \left[\hat{\mu}_1 - \hat{\mu}_2 \pm z_{0.975}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right],$$

où on a gardé les mêmes notations que dans la solution de l'exercice précédent. On calcule, sur base de notre échantillon,

$$\text{IC}_{95\%}(\mu_1 - \mu_2) = [-3.08, -0.92].$$

(b) L'intervalle calculé en (a) ne contient pas 0. Par le raisonnement expliqué à la slide 5-24 du cours, on conclut, pour un test de niveau $\alpha = 5\%$, avec les mêmes hypothèses que celles de l'exercice 5.3 (a), au rejet de H_0. Cette fois-ci, une différence significative existe entre les moyennes des deux groupes dans les résultats du post-test.

5.6: (a) Le test à effectuer est identique à celui de la question 5.3 (a), i.e.

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{contre} \quad H_a : \mu_1 - \mu_2 \neq 0,$$

où on note μ_i la moyenne respective du groupe des coureurs et des cyclistes. La statistique de ce test est, comme à la question 5.3 (c),

$$T = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}},$$

qui est asymptotiquement normale sous H_0, par le Thérèome Centrale Limite et le Théorème de Slutsky. Au niveau de confiance 95%, comme notre test est bilatéral, on compare notre statstique de test calculée sur notre échantillon, notée t, aux quantiles de la loi normale d'ordre 0.025 et 0.975. Ceux-ci sont données, respectivement, par -1.96 et 1.96 (par symétrie). On calcule également $t \approx 6.09$. Cette valeur est bien dans la région de rejet RR $= \{x : x < -1.96 \text{ ou } x > 1.96\}$ déterminée par nos deux quantiles. On rejette donc l'hypothèse nulle au profit de l'hypothèse alternative H_a pour un tel niveau de confiance.

(b) La p-valeur (bilatérale) du test est donnée par

$$p = P(|T| \geq |t| \, | H_0) = 2P(T \geq |t| \, | H_0),$$

où on a utilisé la symétrie de la loi normale pour obtenir la seconde égalité. On rappelle que $t \approx 6.09$. Dès lors, on calcule, soit en utilisant les tables de la loi normale, soit via `R`, que

$$p = 2 \times 5.65 \times 10^{-10} \approx 0.$$

Les données produisent une p-valeur quasiment nulle. Cela veut dire que dans notre test, on rejettera H_0 pour n'importe quel niveau raisonnable α (par exemple : $\alpha = 1\%, 5\%, 10\%$).

5.7: (a) La statistique à considérer pour ce test est

$$T = \sqrt{n} \frac{S^2 - 0.01}{0.01\sqrt{\hat{K} - 1}},$$

qui est asymptotiquement normale sous H_0 (voir slide 5-31). Comme le test est unilatéral à droite, on rejettera l'hypothèse nulle si t, la valeur observée de T, est dans la région de rejet RR $= \{x : x > 1.64\}$ déterminée par le quantile d'ordre 0.95 de la loi normale. On calcule $t = 5.96$. Clairement, t est donc bien dans la région de rejet et on conclut, au niveau de confiance 95%, que la variabilité de sa machine est supérieure aux 1% annoncés.

(b) La p-valeur unilatérale à droite est donnée par

$$p = P(T > t | H_0).$$

Comme $T \sim N(0, 1)$ sous H_0 et $t = 5.96$, on calcule

$$p \approx 1.26 \times 10^{-9},$$

la p-valeur est quasiment nulle. Les données sont très significatives.

Chapitre 6

Analyse des données catégorielles

6.1 À préparer avant la séance

6.1: On souhaite tester au seuil $\alpha = 5\%$:

$$\begin{cases} H_0 : & \text{Indépendance entre la tranche d'âge et la violence du programme regardé à la TV} \\ H_{\mathrm{a}} : & \text{Dépendance entre la tranche d'âge et la violence du programme regardé à la TV.} \end{cases}$$

La statistique de test observée est :

$$X^2_{\mathrm{obs}} = \frac{\left(8 - \frac{41\times 26}{81}\right)^2}{\frac{41\times 26}{81}} + \frac{\left(12 - \frac{27\times 41}{81}\right)^2}{\frac{27\times 41}{81}} + \cdots + \frac{\left(7 - \frac{40\times 28}{81}\right)^2}{\frac{40\times 28}{81}} = 11.169.$$

La région de rejet étant donnée par :

$$\mathrm{RR} = \{X^2 > \chi^2_{(2-1)\times(3-1);0.05}\} = \{X^2 > 5.99147\},$$

comme $X^2_{\mathrm{obs}} = 11.169 \in \mathrm{RR}$, on rejette H_0 au seuil $\alpha = 5\%$. On montre statistiquement, au niveau de confiance 95%, qu'il y a bien une dépendance entre la violence des émissions regardées et l'âge du spectateur.

6.2: On commence par estimer le paramètre λ qui apparaît dans la fonction de masse de probabilité d'une distribution de Poisson. On l'estime par maximum de vraisemblance

$$\hat{\lambda}^{MV}_n = \bar{Y}_n = \frac{57\times 0 + 22\times 1 + 7\times 2 + 2\times 3 + 0}{88} = 0.477.$$

Ensuite, on considère la statistique de test du chi-carré d'ajustement donnée par

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - np_i^{(0)})^2}{np_i^{(0)}},$$

avec $k = 3$ catégories où l'on a regroupé les 3 dernières catégories ensemble afin d'avoir un minimum suffisant d'observations par catégorie (en effet, on pourra vérifier que $np_i^{(0)} \geq 5$ pour

chaque $i = 1, 2, 3$ avec un tel regroupement). On a alors respectivement, $n_1 = 57$, $n_2 = 22$ et $n_3 = 9$ observations. On déduit aussi que $n = 88$. Ensuite, on calcule les probabilités $p_i^{(0)}$ sous H_0. La probabilité d'appartenir à la première catégorie sous H_0 est donnée par

$$p_1^{(0)} = P(Y = 0) = \exp(-\lambda) = \exp(-0.477) = 0.62.$$

De manière similaire, on calcule

$$p_2^{(0)} = P(Y = 1) = \lambda \exp(-\lambda) = 0.477 \times \exp(-0.477) = 0.296.$$

Pour les catégories 3, on a

$$p_3^{(0)} = P(Y \geq 2) = 1 - [P(Y = 0) + P(Y = 1)] = 1 - (p_1^{(0)} + p_2^{(0)}) = 0.084.$$

On peut alors calculer

$$X^2_{\text{obs}} = \frac{(57 - 88 \times 0.62)^2}{88 \times 0.62} + \frac{(22 - 88 \times 0.296)^2}{88 \times 0.296} + \frac{(9 - 88 \times 0.084)^2}{88 \times 0.084} = 1.09.$$

Asymptotiquement, on a $X^2 \overset{H_0}{\sim} \chi^2_{3-1-1} = \chi^2_1$. En effet, on a du estimer un paramètre (λ) pour le calcul des probabilités $p_i^{(0)}$. Il faut donc réduire le nombre de degrés de liberté $k - 1$ de un. On compare donc X^2_{obs} au quantile supérieur d'ordre 5% d'une distribution du chi-carré avec 1 degré1 de liberté, donné par 3.84146. Comme X^2_{obs} est inférieur à cette valeur, on ne rejette pas H_0. Pour un tel niveau de confiance, les données semblent bien coller au modèle Poissonnien.

6.2 À préparer après et pendant la séance

6.1: Posons X le type d'ECG du patient (négatif ou positif) et Y le type de complication (mortelle ou non mortelle). On souhaite tester, au seuil $\alpha = 5\%$,

$$\begin{cases} H_0 : & X \text{ est indépendant de } Y \\ H_a : & X \text{ n'est pas indépendant de } Y. \end{cases}$$

On effectue le test sur base de la statistique X^2 vue au cours pour un test d'indépendance. Ici, on calcule

$$X^2_{\text{obs}} = \frac{\left(166 - \frac{426 \times 167}{469}\right)^2}{\frac{426 \times 167}{469}} + \frac{\left(1 - \frac{43 \times 167}{469}\right)^2}{\frac{43 \times 167}{469}} + \frac{\left(260 - \frac{426 \times 302}{469}\right)^2}{\frac{426 \times 302}{469}} + \frac{\left(42 - \frac{302 \times 43}{469}\right)^2}{\frac{302 \times 43}{469}} = 22.8705$$

La région de rejet étant donnée par :

$$\text{RR} = \{X^2 > \chi^2_{(2-1)\times(2-1);0.05}\} = \{X^2 > 3.84\},$$

comme $X^2_{\text{obs}} = 22.8705 \in \text{RR}$, on rejette H_0 au seuil $\alpha = 5\%$. On conclut donc qu'il y a bien une dépendance au niveau des variables aléatoires sous-jacentes, car on rejette l'hypothèse nulle d'indépendance entre les deux variables. Notez qu'on a seulement établi une dépendance entre X et Y. Aucune relation de cause à effet n'a été suggérée.
Remarque : on peut montrer également que p-valeur = $P(X^2 > X^2_{obs}) = P(X^2 > 22.8705) < 0.05$. Ce qui conduit aussi au rejet de H_0 au niveau $\alpha = 5\%$. Les deux méthodes sont équivalentes.

6.2: Dans cet exercice, on souhaite faire un test d'ajustement pour tester l'hypothèse de normalité des vitesses rencontrées sur l'avenue, avec des paramètres de la loi normale données par $\mu = 70$ et $\sigma = 4$. Notez que ce problème est différent de celui rencontré dans un test d'ajustement classique pour lequel la loi sous H_0 est une loi discrète avec un nombre fini de valeurs possibles. Ici on considère une loi continue (la loi normale). Néanmoins, en considérant des intervalles de vitesses au lieu de considérer les valeurs individuelles, nous pouvons quand même appliquer les méthodes du chapitre 6. Cependant, il faut être prudent car ce découpage en intervalles peut influencer le résultat du test que l'on va effectuer.

La statistique que l'on considère est celle habituelle pour un test d'ajustement du chi-carré

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - np_i^{(0)})^2}{np_i^{(0)}}.$$

Ici, on a $k = 5$ catégories avec, respectivement, $n_1 = 12$, $n_2 = 14$, $n_3 = 78$, $n_4 = 40$ et $n_5 = 6$ observations (donc $n = 150$). Il reste à calculer les probabilités $p_i^{(0)}$ d'appartenir à la catégorie i sous H_0. On explique le raisonnement pour la première catégorie, les autres calculs sont identiques. Sous H_0, si la vitesse des automobilistes est noté Y, on a $Y \overset{H_0}{\sim} N(70, 16)$. Donc,

$$p_1^{(0)} = P(40 \leq Y \leq 55) = P\left(\frac{40-70}{4} \leq Z \leq \frac{55-70}{4}\right) = P(Z \geq -7.5) - P(Z \geq -3.75),$$

où $Z \sim N(0, 1)$. Les dernières probabilités sont calculable à partir de la commande `R` suivante :

`pnorm(-7.5,lower.tail=FALSE)-pnorm(-3.75,lower.tail=FALSE).`

On trouve $p_1^{(0)} = 8.84 \times 10^{-5}$. Pour les autres catégories de vitesses, on trouve les valeurs suivantes : $p_2^{(0)} = 0.11$, $p_3^{(0)} = 0.79$, $p_4^{(0)} = 0.11$ et $p_5^{(0)} = 8.84 \times 10^{-5}$. Nous sommes alors en mesure de calculer la statistique de test X^2

$$\begin{aligned} X^2_{\text{obs}} &= \frac{(12 - 150 \times 8.84 \times 10^{-5})^2}{150 \times 8.84 \times 10^{-5}} + \frac{(40 - 150 \times 0.11)^2}{150 \times 0.11} + \frac{(78 - 150 \times 0.79)^2}{150 \times 0.79} \\ &\quad + \frac{(40 - 150 \times 0.11)^2}{150 \times 0.11} + \frac{(6 - 150 \times 8.84 \times 10^{-5})^2}{150 \times 8.84 \times 10^{-5}} \\ &= 13586.86. \end{aligned}$$

Cette valeur est à comparer avec le quantile supérieur d'ordre $\alpha = 1\%$ d'une loi du chi-carré avec $k - 1 = 4$ degrés de liberté. Notez que ce nombre de degré de libertés aurait diminué de 2 si on avait estimé le paramètre $(\mu, \sigma) \in \mathbb{R}^2$ par maximum de vraisemblance sur base de nos données. On trouve

$$\chi^2_{4;0.01} = 13.2767.$$

Clairement, notre statistique de test observée est dans la région de rejet et on rejette H_0 au profit de l'hypothèse alternative. Les données ne sont pas normales.

6.3: Pour tester l'indépendance entre les deux variables aléatoires catégorielles considérées dans le problème, on utilise la statistique de test

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(n_{ij} - \frac{r_i \times c_j}{n}\right)^2}{\frac{r_i \times c_j}{n}},$$

où on a utilisé les notations du cours pour les quantités intervenant dans la table de contingence. Ici, on a $r = c = 3$, donc $X^2 \overset{H_0}{\sim} \chi^2_4$. On calcule

$$X^2_{\text{obs}} = \frac{\left(250 - \frac{500\times450}{1000}\right)^2}{\frac{500\times450}{1000}} + \ldots + \frac{\left(25 - \frac{200\times300}{1000}\right)^2}{\frac{200\times300}{1000}} = 219.3056.$$

Notez qu'il y a $r \times c = 3 \times 3 = 9$ termes dans la double somme. Cette valeur est à comparer avec le quantile supérieur d'ordre $\alpha = 5\%$ d'une distribution du chi-carré avec 4 degrés de libertés qui est donné par 9.48773. Clairement, notre statistique de test observée est dans la région de rejet et on rejette H_0 au profit de l'hypothèse alternative. Les données ne sont pas indépendantes.

6.4: (a) La statistique que l'on considère est celle habituelle pour un test d'ajustement du chi-carré

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - np_i^{(0)})^2}{np_i^{(0)}}.$$

Ici, on a $k = 5$ catégories de chaussures avec, respectivement, $n_1 = 22$, $n_2 = 21$, $n_3 = 29$, $n_4 = 17$ et $n_5 = 11$ étudiants (chaque fois, on somme le nombre de filles et de garçons car le genre ne joue pas de rôle dans la sous-question (a)). On déduit donc aussi que $n = 100$. De plus, les probabilités sous H_0 sont directement données dans l'énoncé, on ne doit pas les calculer à partir d'un modèle hypothétique comme dans l'exercice 6.3 ou l'exercice 6.6. On peut alors calculer

$$X^2_{\text{obs}} = \frac{(22 - 100 \times 0.2)^2}{100 \times 0.2} + \ldots + \frac{(11 - 100 \times 0.1)^2}{100 \times 0.1} = 0.833.$$

Cette valeur est à comparer avec le quantile supérieur d'ordre $\alpha = 5\%$ d'une distribution chi-carrée χ^2_4. Celui-ci est égal à 9.48773. La statistique de test observée n'est donc pas dans la région de rejet et nous ne pouvons pas rejeter H_0 pour un tel niveau de confiance.

(b) Maintenant, on considère bien les données séparément pour chacun des genres et on test l'indépendance entre les préférences de chaussure et le genre de l'étudiant, en utilisant la statistique de test habituel

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(n_{ij} - \frac{r_i \times c_j}{n}\right)^2}{\frac{r_i \times c_j}{n}},$$

qui suit asymptotiquement une loi du chi-carré avec $(r-1) \times (c-1) = 1 \times 4 = 4$ degrés de liberté. Notez qu'ici, comme à l'exercice 6.2, les sommes du nombre d'observations sur

chaque ligne et chaque colonne ne sont pas fournies et sont à calculer par vous-mêmes. On trouve

$$X^2_{\text{obs}} = \frac{\left(12 - \frac{50\times 22}{100}\right)^2}{\frac{50\times 22}{100}} + \ldots + \frac{\left(4 - \frac{50\times 11}{100}\right)^2}{\frac{50\times 11}{100}} = 2.82.$$

Cette quantité étant inférieure à $\chi^2_{4;0.05} = 9.48773$, on ne rejette pas H_0, l'hypothèse d'indépendance.

Chapitre 7

Modèles linéaires

7.1: (a) Ajuster le modèle aux données revient à estimer les coefficients β_0 et β_1 de la fonction linéaire qui modélise l'espérance conditionnelle $\mathbb{E}[Y|X = x] = \beta_0 + \beta_1 x$. Les formules provenant de la procédure des moindres carrés sont les suivantes :

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \text{et} \qquad \hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2},$$

où il est implicitement supposé que les sommes dans l'expression de $\hat{\beta}_1$ portent sur tout l'échantillon, i.e., les sommes vont de $i = 1$ à $i = n$. On commence par calculer $\hat{\beta}_1$ car l'expression de $\hat{\beta}_0$ dépend de l'estimateur de pente. Afin de pouvoir calculer l'estimateur de pente, nous réécrivons son expression dans une forme qui ne fait intervenir que des quantités données dans l'énoncé. Cela se fait comme suit :

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \\
&= \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x}) x_i} \\
&= \frac{\sum x_i y_i - n^{-1} \sum x_i \sum y_i}{\sum x_i^2 - n^{-1} (\sum x_i)^2}, \\
&= \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sum x_i^2 - n \bar{x}^2} \\
&= \frac{S_{xy}}{S_{xx}}
\end{aligned}$$

où le passage de la première ligne à la seconde ligne suit du fait que, dans le numérateur,

$$\sum (x_i - \bar{x}) \bar{y} = 0,$$

et, dans le dénominateur,

$$\sum (x_i - \bar{x}) \bar{x} = 0.$$

La preuve de ces deux égalités est laissée au lecteur et suit, dans chacun des cas, de la définition même de $\bar{x}$. Toutes les quantités intervenant dans l'expression obtenue pour $\hat{\beta}_1$

sont maintenant données dans l'énoncé et on trouve

$$\hat{\beta}_1 = \frac{841605.3 - 0.01 \times 2151.065 \times 39300.52}{46834.86 - 0.01 \times (2151.065)^2} = -6.69.$$

Comme on pouvait s'y attendre, on estime un effet négatif de la taille de la classe sur les résultats du test.
Pour $\hat{\beta}_0$, on a directement que

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 0.01 \times 39300.52 + 6.69 \times 0.01 \times 2151.065 = 536.91.$$

L'interprétation de ce coefficient n'est pas très pertinente ici car cela correspond au score attendu au test pour une classe de 0 étudiant.

(b) Si le score attendu est noté y, on a simplement

$$y = 536.91 - 6.69 \times 22 = 389.73.$$

(c) Si on note y_1 le score attendu pour le classe des 19 élèves et y_2 celui attendu pour la classe des 23 élèves, on a que la variation attendue du score est

$$y_2 - y_1 = (536.91 - 6.69 \times 23) - (536.91 - 6.69 \times 19) = 6.69 \times (19 - 23) = -26.76.$$

Comme la taille de la classe a un effet négatif sur la classe, on s'attend à une diminution des scores.

7.2: La sortie obtenue après exécution du code est présentée sur la figure 7.1.

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-23.079  -9.134  -1.573   7.811  32.326

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 536.9598    10.4057   51.60   <2e-16 ***
x            -6.6922     0.4808  -13.92   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.42 on 98 degrees of freedom
Multiple R-squared:  0.6641,    Adjusted R-squared:  0.6606
F-statistic: 193.7 on 1 and 98 DF,  p-value: < 2.2e-16
```

FIGURE 7.1 – Sortie de la fonction `summary` d'un modèle linéaire ajusté avec la fonction `lm`.

On observe, en rouge, les estimations ponctuelles des paramètres β_0 et β_1. Elles sont identiques à celles calculées « à la main » dans l'exercice précédent. Les valeurs obtenues sont relativement proches des vraies valeurs des paramètres du modèle sur base duquel on a généré les données.

Notons que la variance des estimateurs est une fonction croissante de σ^2 la variance (constante) du terme d'erreur, qui ici est relativement grande. Cela explique pourquoi les valeurs estimées des paramètres ne sont pas parfaites.
Beaucoup d'autres informations sont disponibles également :

- Dans la section « Residuals », on observe divers quantiles associés aux résidus $e_i = Y_i - \hat{Y}_i$.
- Dans la section « Coefficiens », on observe, dans l'ordre des colonnes : les estimations ponctuelles des estimateurs d'intercepte et de pente, leur écart-type estimé, la valeur du t-test (une version non-asymptotique du test normal vu au cours, quand on suppose la normalité de l'erreur ϵ) associé au test $H_0 : \beta_i = 0$ contre $H_a : \beta_i \neq 0$ et enfin, la p-valeur associée à ce test. Les étoiles apparaissant à la droite indiquent pour quels α la p-valeur est significative, selon le code expliqué juste en dessous.
- Dans le dernier paragraphe, en dessous, on retrouve différentes informations globales sur le modèle dont le R^2 vu au cours.

7.3: On souhaite effectuer le test suivant

$$H_0 : \beta_1 = 0 \qquad \text{contre} \qquad H_a : \beta_1 \neq 0.$$

Pour cela, on se base sur la statistique de test

$$T = \frac{\hat{\beta}_1 - \beta_1^{(0)}}{\hat{\sigma}\sqrt{c_{11}}} = \frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{S_{xx}}},$$

qui suit, asymptotiquement, une distribution normale centrée réduite sous H_0. On calcule

$$T_{\text{obs}} = \frac{0.155}{0.202} = 0.767.$$

Cette valeur est à comparer, en valeur absolue, au quantile supérieur $z_{0.025}$ de la distribution $N(0, 1)$. Celui-ci est donné par 1.96. Comme T_{obs} est en dessous de cette valeur, on ne peut pas rejeter H_0 pour un tel α. Le coefficient de pente n'est pas significativement différent de zéro pour un niveau de confiance 95%.

7.4: Rappelons que $e_i := Y_i - \hat{Y}_i$, où $\hat{Y}_i := \hat{\beta}_0 + \hat{\beta}_1 X_i$, avec

$$\hat{\beta}_0 := \bar{Y} - \hat{\beta}_1 \bar{X} \qquad \text{et} \qquad \hat{\beta}_1 := \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}.$$

Dès lors, on calcule

$$\begin{aligned}
\sum e_i^2 &= \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2 \\
&= \sum (Y_i - \bar{Y} + \hat{\beta}_1 \bar{X} - \hat{\beta}_1 X_i)^2 \\
&= \sum [(Y_i - \bar{Y}) - \hat{\beta}_1 (X_i - \bar{X})]^2 \\
&= \sum (Y_i - \bar{Y})^2 + \hat{\beta}_1^2 \sum (X_i - \bar{X})^2 - 2\hat{\beta}_1 \sum (X_i - \bar{X})(Y_i - \bar{Y}) \\
&= \sum (Y_i - \bar{Y})^2 + \hat{\beta}_1^2 \sum (X_i - \bar{X})^2 - 2\hat{\beta}_1^2 \sum (X_i - \bar{X})^2 \\
&= \sum (Y_i - \bar{Y})^2 - \hat{\beta}_1^2 \sum (X_i - \bar{X})^2,
\end{aligned}$$

comme annoncé.

7.5: (a) Pour l'ajustement du modèle, on se base sur les formules qui ont été démontrées au cours et dans l'exercice 7.1, à savoir

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$
$$\hat{\beta}_1 = \frac{\sum x_i y_i - n^{-1} \sum x_i \sum y_i}{\sum x_i^2 - n^{-1} (\sum x_i)^2}.$$

On trouve

$$\hat{\beta}_0 = 9 \qquad \text{et} \qquad \hat{\beta}_1 = -0.25.$$

(b) Comme conseillé dans l'indice, on se base sur la formule

$$R^2 = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}.$$

Notez que, en utilisant la formule de l'exercice précédent pour la sommes des résidus au carré, on peut réécrire le R^2 comme

$$R^2 = 1 - 1 + \hat{\beta}_1^2 \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \hat{\beta}_1^2 \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}.$$

On calcule

$$\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - n^{-1} (\sum_{i=1}^n X_i)^2 = 400$$
$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n Y_i^2 - n^{-1} (\sum_{i=1}^n Y_i)^2 = 150.$$

On déduit,

$$R^2 = 0.25^2 \times \frac{400}{150} = 1/6.$$

Le modèle explique environ 17% de la variabilité de la demande.

(c) Le test est le même que celui décrit dans l'exercice 7.3. La statistique utilisée pour ce test est

$$T = \frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{S_{xx}}} \sim_a N(0,1).$$

Rappelons que

$$\hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^n e_i^2} = \sqrt{\frac{1}{n} \left(\sum_{i=1}^n (Y_i - \bar{Y})^2 - \hat{\beta}_1^2 \sum_{i=1}^n (X_i - \bar{X})^2 \right)}$$
$$\sqrt{S_{xx}} = \sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}.$$

On trouve

$$\hat{\sigma} = 1.58 \qquad \text{et} \qquad \sqrt{S_{xx}} = 20.$$

Donc,

$$T_{\text{obs}} = \frac{-0.25}{1.58/20} = -3.16.$$

On compare cette valeur, en valeur absolue, au quantile supérieure $z_{0.025} = 1.96$ et on observe que $|T_{\text{obs}}| > 1.96$ donc notre décision est le rejet de H_0. Le coefficient β_1 est significativement différent de zéro, au niveau de confiance 95%.

(d) Nous rappelons du cours (slide 7-21) qu'un intervalle de confiance de niveau $1 - \alpha$ pour $\mathbb{E}[Y|X = x^*]$ est donné par

$$\left[\hat{\beta}_0 + \hat{\beta}_1 x^* \pm z_{\alpha/2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}\right].$$

Dans notre cas, cela donne

$$\left[9 - 0.25 \times 5 \pm 1.96 \times 1.58 \times \sqrt{\frac{1}{50} + \frac{(5-4)^2}{400}}\right] = [7.285, 8.215].$$

L'intervalle est relativement petit car la valeur $x^* = 5$ se trouve assez proche de la moyenne $\bar{x}$ et, au vue de la formule, l'intervalle de confiance est plus petit quand on se trouve proche de la moyenne des x_i.