Correction TD 3 : Tests de χ^2

Correction exercice 1.

On veut tester l'hypothèse "le dé est équilibré", c'est à dire, en notant $p_i = \mathbb{P}(X = i)$ pour $i = 1, \cdots, 6$,

$$H_0 : p = \left(\frac{1}{6}, \cdots, \frac{1}{6}\right)$$
$$H_1 : p \neq \left(\frac{1}{6}, \cdots, \frac{1}{6}\right)$$

Sous l'hypothèse H_0, on a

$$S := \sum_{k=1}^{6} \frac{\Big(O_i - E_i\Big)^2}{E_i} \sim \chi^2(5),$$

où $E_i = p_i N_{tot} = 100$.
Donc, d'après la table de quantiles, on a, sous H_0,

$$\mathbb{P}(S > 11.07) = 0.05.$$

L'expérience donne

$$s = \frac{(50-100)^2}{100} + \cdots + \frac{(149-100)^2}{100} \approx 91.02.$$

Or, $91.02 > 11.07$, on rejette donc le test, et on peut affirmer de façon statistiquement significative, au niveau de 5%, que le dé n'est pas équilibré.

Correction exercice 2.

Cet exercice est très similaire au précédent :
On veut tester l'hypothèse "les chiffres sont équiprobables", c'est à dire, en notant $p_i = \mathbb{P}(X = i)$ pour $i = 1, \cdots, 10$,

$$H_0 : p = \left(\frac{1}{10}, \cdots, \frac{1}{10}\right)$$
$$H_1 : p \neq \left(\frac{1}{10}, \cdots, \frac{1}{10}\right)$$

Sous l'hypothèse H_0, on a

$$S := \sum_{k=1}^{10} \frac{\Big(O_i - E_i\Big)^2}{E_i} \sim \chi^2(9),$$

où $E_i = p_i N_{tot} = 100$.
Donc, d'après la table de quantiles, on a, sous H_0,

$$\mathbb{P}(S > 16.91) = 0.05.$$

L'expérience donne

$$s = \frac{(120-100)^2}{100} + \cdots + \frac{(77-100)^2}{100} \approx 17.38.$$

Or, $17.38 > 11.07$, on rejette donc le test, et on peut affirmer de façon statistiquement significative, au niveau de 5%, que le générateur ne génère pas des nombres uniformément.

Correction exercice 3.

Cet exercice est très similaire au précédent :
On veut tester l'hypothèse, en notant $p_i = \mathbb{P}(X = i)$ pour $i = 1, \cdots, 5$,

$$H_0 : p = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right)$$
$$H_1 : p \neq \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right)$$

Sous l'hypothèse H_0, on a

$$S := \sum_{k=1}^{5} \frac{\Big(O_i - E_i\Big)^2}{E_i} \sim \chi^2(4),$$

où E_i sont donnés dans le tableau suivant :

i	1	2	3	4	5
O_i	111	113	118	81	77
E_i	125	125	125	62.5	62.5

Donc, d'après la table de quantiles, on a, sous H_0,

$$\mathbb{P}(S > 9.49) = 0.05.$$

L'expérience donne

$$s = \frac{(111 - 125)^2}{125} + \cdots + \frac{(77 - 62.5)^2}{62.5} \approx 11.95.$$

Or, $11.95 > 9.49$, on rejette donc le test, et on peut affirmer de façon statistiquement significative, au niveau de 5%, que $p \neq \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right)$.

Correction exercice 4.

Cet exercice est très similaire au précédent :
On veut tester l'hypothèse $X \sim \mathcal{B}in\Big(4, \frac{1}{3}\Big)$, c'est à dire, en notant $p_i = \mathbb{P}(X = i)$ pour $i = 0, \cdots, 4$,

$$H_0 : p = \Big(0.198, 0.395, 0.296, 0.099, 0.012\Big)$$
$$H_1 : p \neq \Big(0.198, 0.395, 0.296, 0.099, 0.012\Big)$$

Sous l'hypothèse H_0, si l'approximation χ^2 est raisonnable, on a

$$S := \sum_{k=0}^{4} \frac{\Big(O_i - E_i\Big)^2}{E_i} \sim \chi^2(4),$$

où E_i sont donnés dans le tableau suivant :

i	0	1	2	3	4
O_i	67	122	94	28	13
E_i	64	128	96	32	4

En réalité, garder une "case" avec aussi peu d'observation rend l'approximation χ^2 peu raisonnable, il est donc plus sage de regroupper les deux dernières catégories :

$$H_0 : p = \Big(0.198, 0.395, 0.296, 0.099 + 0.012\Big)$$
$$H_1 : p \neq \Big(0.198, 0.395, 0.296, 0.099 + 0.012\Big)$$

Sous l'hypothèse H_0, si l'approximation χ^2 est raisonnable, on a

$$S := \sum_{k=0}^{3} \frac{\Big(O_i - E_i\Big)^2}{E_i} \sim \chi^2(3),$$

où E_i sont donnés dans le tableau suivant :

i	0	1	2	$3-4$
O_i	67	122	94	$28+13$
E_i	64	128	96	$32+4$

Donc, d'après la table de quantiles, on a, sous H_0,

$$\mathbb{P}(S > 7.81) = 0.05.$$

L'expérience donne

$$s = \frac{(67-64)^2}{64} + \cdots + \frac{(41-36)^2}{36} \approx 11.95.$$

Or, $1.16 < 7.81$, on ne peut donc pas rejeter le test, et on ne peut pas exclure, au niveau de 5%, que $X \sim \mathcal{B}in\Big(4, \frac{1}{3}\Big)$.

Correction exercice 5.

On procède comme à l'exercice précédent, avec la nuance qu'il faut ici estimer tout d'abord le paramètre de la loi de Poisson (et on enlèvera un degré de liberté).

Que ce soit par estimation des moments, ou par maximum de vraisemblance, on a vu que le paramètre d'une Poisson peut s'estimer par $\hat{\Lambda} = \bar{X}$.

L'estimation ponctuelle donne

$$\hat{\lambda} = \frac{18+18+\cdots+1}{52} \approx 1.19$$

On calcule d'abord les effectifs attendus, pour $\mathcal{P}(\hat{\lambda})$:

i	0	1	2	3	4	≥ 5
$\mathbb{P}(X=i)$	0.304	0.362	0.216	0.086	0.026	0.007
E_i	15.8	18.8	11.2	4.5	1.4	0.3

Pour que l'approximation χ^2 soit valide, il faut au minimum regroupper les 3 dernières catégories (on applique la règle empirique $E_i > 5$), ce qui donne :

i	0	1	2	≥ 3
$\mathbb{P}(X=i)$	0.304	0.362	0.216	$0.086+0.026+0.007$
E_i	15.8	18.8	11.2	$4.5+1.4+0.4$

Autrement dit, en prenant en compte les effectifs observés

i	0	1	2	≥ 3
O_i	18	18	8	8
$\mathbb{P}(X=i)$	0.304	0.362	0.216	0.118
E_i	15.8	18.8	11.2	6.2

On peut alors effectuer le test statistique : On pose

H_0 : X suit une loi de Poisson

H_1 : X ne suit pas une loi de Poisson

Sous l'hypothèse H_0, si l'approximation χ^2 est raisonnable, on a (en prenant en compte l'estimation d'un paramètre)

$$S := \sum_{k=0}^{3} \frac{\Big(O_i - E_i\Big)^2}{E_i} \sim \chi^2(4-1-1) = \chi^2(2),$$

Donc, d'après la table de quantiles, on a, sous H_0,

$$\mathbb{P}(S > 5.99) = 0.05.$$

L'expérience donne

$$s = \frac{(18-15.8)^2}{15.8} + \cdots + \frac{(8-6.2)^2}{6.2} \approx 1.8.$$

Or, $1.8 < 5.99$, on ne peut donc pas rejeter le test, et on ne peut pas exclure, au niveau de 5%, que X suive une loi de Poisson.

Correction exercice 6.

Pour tester une variable continue, on doit diviser l'intervalle en classes (potentiellement après avoir estimé les paramètres de la loi). Evidemment, dans cette méthode, il y aura de l'arbitraire, en choisissant le nombre de classes. Pour que l'approximation χ^2 soit raisonnable, il faut (selon les auteurs, c'est arbitraire aussi) que chaque classe (ou 80% des classes selon les auteurs) aient un effectif attendu d'au moins 5 (10 selon les auteurs). Ici, il y a donc plusieurs façons de faire, et je vais choisir 3 classes (compromis entre 2 classes, qui ne teste pas beaucoup la forme de la loi, et 4 classes, qui serait dans la limite de l'approximation χ^2.)

On peut donc commencer par estimer le paramètre. Pour la loi exponentielle, on a vu que l'estimation par moments ou par maximum de vraisemblance donnent toutes deux $\hat{\Lambda} = \frac{1}{\bar{X}}$. Pour les observations données, l'estimation ponctuelle donne

$$\hat{\lambda} = \frac{1}{\bar{x}} = 0.0045.$$

Pour diviser en trois parties, le plus simple, c'est de prendre 3 parties équiprobable pour la loi $\mathcal{E}(\hat{\lambda})$, ce qui donne :

$$\begin{aligned}
\mathbb{P}(\mathcal{E}(\hat{\lambda}) > q_1) = \frac{2}{3}) &\Leftrightarrow e^{-\hat{\lambda} q_1} = \frac{2}{3} \\
&\Leftrightarrow q_1 = -\frac{\ln\left(\frac{2}{3}\right)}{\hat{\lambda}} \\
\mathbb{P}(\mathcal{E}(\hat{\lambda}) > q_2) = \frac{1}{3}) &\Leftrightarrow e^{-\hat{\lambda} q_2} = \frac{1}{3} \\
&\Leftrightarrow q_2 = -\frac{\ln\left(\frac{1}{3}\right)}{\hat{\lambda}}
\end{aligned}$$

L'application numérique donne

$q_1 \approx 90$

$q_2 \approx 244$

On a plus qu'à remplir le tableau des effectifs observés et théoriques :

	$X \leq 90$	$90 \leq X \leq 244$	$244 \leq X$
O_i	9	6	5
p_i	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
E_i	6.67	6.67	6.67

On peut alors effectuer le test statistique : On pose

H_0 : X suit une loi exponentielle

H_1 : X ne suit pas une loi exponentielle

Sous l'hypothèse H_0, si l'approximation χ^2 est raisonnable, on a (en prenant en compte l'estimation d'un paramètre)

$$S := \sum_{k=1}^{3} \frac{\Big(O_i - E_i\Big)^2}{E_i} \sim \chi^2(3-1-1) = \chi^2(1),$$

Donc, d'après la table de quantiles, on a, sous H_0,

$$\mathbb{P}(S > 3.84) = 0.05.$$

L'expérience donne

$$s = \frac{(9-6.67)^2}{6.67} + \cdots + \frac{(5-6.67)^2}{6.67} \approx 1.3.$$

Or, $1.3 < 3.84$, on ne peut donc pas rejeter le test, et on ne peut pas exclure, au niveau de 5%, que X suive une loi exponentielle.

Pour refaire le test avec un niveau de 0.1, il suffit de recalculer le quantile : Donc, d'après la table de quantiles, on a, sous H_0,

$$\mathbb{P}(S > 2.7) = 0.1,$$

et ici, cela ne change pas la conclusion.

Remarque : Pour certaines valeurs de S, changer le niveau change la conclusion (en particulier, augmenter le niveau du test, autorise a rejeter plus facilement à tord, et donc à conclure plus facilement H_1 à tord. C'est pour cela, que l'on peut préférer donner la p-valeur $\mathbb{P}(S > s)$, et conclure plus prudemment (c'est indicatif) :

- $p \leq 10^{-6}$: Ordre de grandeur exigé en physique pour conclure
- $p \leq 10^{-3}$: Très forte présemption en faveur de l'hypothèse nulle
- $p \leq 10^{-2}$: Forte présemption en faveur de l'hypothèse nulle
- $p \leq 10^{-1}$: Faible présemption en faveur de l'hypothèse nulle

Correction exercice 7.

On raisonne comme à l'exercice 5 : que ce soit par estimation des moments, ou par maximum de vraisemblance, on a vu que le paramètre d'une Poisson peut s'estimer par $\hat{\Lambda} = \bar{X}$.

L'estimation ponctuelle donne

$$\hat{\lambda} = \frac{50*0 + 74*1 + \cdots + 1*5}{200} \approx 2.29$$

On calcule d'abord les effectifs attendus, pour $\mathcal{P}(\hat{\lambda})$:

i	0	1	2	3	4	5	> 5
$\mathbb{P}(X=i)$	0.101	0.232	0.266	0.203	0.116	0.053	0.029
E_i	20.3	46.4	53.1	40.5	23.2	10.6	5.9

Pour que l'approximation χ^2 soit valide, si on applique la règle empirique $E_i > 5$, on n'a pas besoin de regroupper ici, ce qui donne :

i	0	1	2	3	4	5	> 5
$\mathbb{P}(X=i)$	0.101	0.232	0.266	0.203	0.116	0.053	0.029
E_i	20.3	46.4	53.1	40.5	23.2	10.6	5.9
O_i	50	74	50	21	4	1	0

On peut alors effectuer le test statistique : On pose

H_0 : X suit une loi de Poisson
H_1 : X ne suit pas une loi de Poisson

Sous l'hypothèse H_0, si l'approximation χ^2 est raisonnable, on a (en prenant en compte l'estimation d'un paramètre)

$$S := \sum_{k=0}^{6} \frac{\Big(O_i - E_i\Big)^2}{E_i} \sim \chi^2(7-1-1) = \chi^2(5),$$

Donc, d'après la table de quantiles, on a, sous H_0,

$$\mathbb{P}(S > 11.07) = 0.05.$$

L'expérience donne

$$s \approx 100.24.$$

Or, $100.24 > 11.07$, on peut donc rejeter le test, et on peut exclure, au niveau de 5%, que X suive une loi de Poisson.

Correction exercice 8.

Dans cet exercice, on va formuler un test d'adéquation, en notant X le groupe sanguin observé dans notre échantillon, $p = \Big(\mathbb{P}(X = i)\Big)_{i=1,\cdots,4}$, et $p_0 = \big(0.47, 0.43, 0.07, 0.02\big)$ la répartition dans la population totale, on peut formuler le test

$H_0 : p = p_0$
$H_1 : p \neq p_0$

Remarque : En réalité, cette formulation est assez-proche de la formulation d'un test d'indépendance entre la variable "maladie", et la variable "groupe sanguin", que l'on formulerait si l'on avait accès à un échantillon de groupes sanguins de patient·e·s non malades.
On peut alors remplir le tableau d'effectifs :

i	0	A	B	AB
p_0	0.47	0.43	0.07	0.03
E_i	94	86	14	4
O_i	104	76	18	2

Pour que l'approximation soit raisonnable, on peut appliquer la règle empirique $E_i \geq 5$ et regroupper des catégories (par exemple les deux dernières, on obtient alors) :

i	0	A	B/AB
p_0	0.47	0.43	0.1
E_i	94	86	18
O_i	104	76	20

Sous l'hypothèse H_0, si l'approximation χ^2 est raisonnable, on a

$$S := \sum_{k=1}^{3} \frac{\Big(O_i - E_i\Big)^2}{E_i} \sim \chi^2(3-1) = \chi^2(2),$$

Donc, d'après la table de quantiles, on a, sous H_0,

$$\mathbb{P}(S > 5.99) = 0.05.$$

L'expérience donne

$$s \approx 2.22.$$

Or, $2.22 \leq 5.99$, on ne peut donc pas rejeter le test, ni exclure, au niveau de 5%, que la fréquence de la maladie soit indépendante du groupe sanguin.

Correction exercice 9.

On note A la variable "âge" (jeune ou âgé), et R la variable "résultats" (bons ou pas), on veut tester l'indépendance entre les variables :

H_0 : A est indépendante de R
H_1 : A n'est indépendante de R

On peut écrire la table de contingence des effectifs observés O_{ij} :

$R\backslash A$	jeune	âgé	Total
Bons	40	50	90
pas bons	30	50	80
Total	70	100	170

On peut aussi calculer la table d'effectifs attendus associés E_{ij} :

$R\backslash A$	jeune	âgé	Total
Bons	37.1	52.9	90
pas bons	32.9	47.1	80
Total	70	100	170

Sous l'hypothèse H_0, on a

$$S := \sum_{i,j=1}^{2} \frac{\Big(O_{ij} - E_{ij}\Big)^2}{E_{ij}} \sim \chi^2\Big((2-1)(2-1)\Big) = \chi^2(1),$$

Donc, d'après la table de quantiles, on a, sous H_0,

$$\mathbb{P}(S > 2.7) = 0.1.$$

L'expérience donne

$$s = \frac{(40-37.1)^2}{37.1} + \cdots + \frac{(50-47.1)^2}{47.1} \approx 0.843.$$

Or, $0.843 \leq 2.7$, on ne peut donc pas rejeter le test, ni exclure, au niveau de 5%, que les variables soient indépendantes.

Correction exercice 10.

On note S la variable "Sexe", et C la variable "cause d'abscence" (maladie ou autre), on veut tester l'indépendance entre les variables :

H_0 : S est indépendante de C
H_1 : S n'est indépendante de C

On peut écrire la table de contingence des effectifs observés O_{ij} :

$S\backslash C$	Maladie	Autre	Total
Homme	1800	1700	3500
Femme	1200	600	1800
Total	3000	2300	5300

On peut aussi calculer la table d'effectifs attendus associés E_{ij} :

$S\backslash C$	Maladie	Autre	Total
Homme	1981	1519	3500
Femme	1019	781	1800
Total	3000	2300	5300

Sous l'hypothèse H_0, on a

$$S := \sum_{i,j=1}^{2} \frac{\Big(O_{ij} - E_{ij}\Big)^2}{E_{ij}} \sim \chi^2\Big((2-1)(2-1)\Big) = \chi^2(1),$$

Donc, d'après la table de quantiles, on a, sous H_0,

$$\mathbb{P}(S > 3.84) = 0.05.$$

L'expérience donne

$$s = \approx 112.3.$$

Or, $112.3 > 3.84$, on peut donc rejeter le test, et conclure de façon statistiquement significative, au niveau de 5%, que les variables ne sont indépendantes, donc qu'il y a un lien entre le sexe et les causes d'absences.

Remarque : Dans la définition formelle d'un test, le "niveau" α du test correspond à la probabilité de rejeter sour l'hypothèse H_0. Dans certains ouvrage francophones, vous pourrez lire "niveau de confiance" pour $1 - \alpha$. Ceci explique la confusion qu'on peut voir entre le sujet de TD et la correction.

Correction exercice 11.

On procède comme à l'exercice précédent :

On note N la variable "Niveau hierarchique", et S la variable "Origine sociale", et on veut tester l'indépendance entre les variables :

H_0 : N est indépendante de S

H_1 : N n'est indépendante de S

On peut écrire la table de contingence des effectifs observés O_{ij} :

$N\backslash S$	Agricole	Cadres	Ouvriers/Employés	autres	Total
Ouvriers/Employés	11	12	145	52	220
Chefs d'équipe	8	6	71	23	108
Cadres	1	27	14	30	72
Total	20	45	230	105	400

On peut aussi calculer la table d'effectifs attendus associés E_{ij} :

$N\backslash S$	Agricole	Cadres	Ouvriers/Employés	autres	Total
Ouvriers/Employés	11	25	126	58	220
Chefs d'équipe	5	12	62	28	108
Cadres	4	8	41	19	72
Total	20	45	230	105	400

On pourraît directement calculer la statistique de test sur ces données, mais il semble que l'approximation χ^2 risque d'être mauvaise vu le nombre de cases avec de petits effectifs. On se propose donc de regroupper (au choix, par exemple les deux dernières lignes). On obtient alors comme effectif observés:

$N\backslash S$	Agricole	Cadres	Ouvriers/Employés	autres	Total
Ouvriers/Employés	11	12	145	52	220
Chefs d'équipe et Cadres	9	33	85	53	180
Total	20	45	230	105	400

Et pour les effectifs attendus

$N\backslash S$	Agricole	Cadres	Ouvriers/Employés	autres	Total
Ouvriers/Employés	11	25	126	58	220
Chefs d'équipe et Cadres	9	20	104	47	180
Total	20	45	230	105	400

Sous l'hypothèse H_0, on a

$$S := \sum_{i=1}^{2}\sum_{j=1}^{4} \frac{\Big(O_{ij} - E_{ij}\Big)^2}{E_{ij}} \sim \chi^2\Big((2-1)(4-1)\Big) = \chi^2(3),$$

Donc, d'après la table de quantiles, on a, sous H_0,

$$\mathbb{P}(S > 9.83) = 0.02.$$

L'expérience donne

$$s \approx 21.88.$$

Or, $21.88 > 9.83$, on peut donc rejeter le test, et conclure de façon statistiquement significative, au niveau de 2%, que les variables ne sont indépendantes, donc qu'il y a un lien entre le niveau hierarchique et les causes d'absences.

Correction exercice 12.

On procède comme à l'exercice précédent :
On note T la variable "Traitement", et F la variable "Fructification", et on veut tester l'indépendance entre les variables :

$H_0 : T$ est indépendante de F
$H_1 : T$ n'est indépendante de F

On peut écrire la table de contingence des effectifs observés O_{ij} :

$T\backslash F$	Pas de fruits	Au moins un fruit	Total
A	203	156	359
B	266	113	379
C	258	128	386
D	196	185	381
Total	932	582	

On peut aussi calculer la table d'effectifs attendus associés E_{ij} :

$T\backslash F$	Pas de fruits	Au moins un fruit	Total
A	220	139	359
B	232	147	379
C	237	149	386
D	234	147	381
Total	932	582	

Sous l'hypothèse H_0, on a

$$S := \sum_{i=1}^{2} \sum_{j=1}^{4} \frac{\Big(O_{ij} - E_{ij}\Big)^2}{E_{ij}} \sim \chi^2\Big((2-1)(4-1)\Big) = \chi^2(3),$$

Donc, d'après la table de quantiles, on a, sous H_0,

$$\mathbb{P}(S > 7.81) = 0.05.$$

L'expérience donne

$$s \approx 36.63.$$

Or, $36.63 > 7.81$, on peut donc rejeter le test, et conclure de façon statistiquement significative, au niveau de 5%, que les variables ne sont indépendantes, donc qu'il y a un lien entre le traitement et la fructification.