CY-Tech 2020 - 2021	Statistiques **TD 5 (Corrigé)**	Mathématiques ING2-GI

Exercice 1 -

Une chaîne d'agences immobilières cherche à vérifier que le nombre de biens vendus par agent par mois suit une loi de Poisson de paramètre $\lambda = 1,5$.
On observe 52 agents pendant un mois et on obtient la répartition suivante :

Nbre ventes par mois	0	1	2	3	4	5
Nbre d'agents	18	18	8	5	2	1

1. Etablir le test permettant de répondre à la question.
2. Même question mais sans connaitre le paramètre λ.

Correction -

1) Soit X le nombre de biens vendus par mois. On a un échantillon de taille $n = 52$ agents. La question est de savoir si l'échantillon suit une loi $\mathcal{P}(\lambda)$ avec $\lambda = 1.5$, donc test d'adéquation :

$\begin{cases} (H_0) & \text{l'échantillon suit } \mathcal{P}(\lambda) \\ (H_1) & \text{l'échantillon ne suit pas } \mathcal{P}(\lambda) \end{cases}$

Si (H_0) est vraie, l'effectif théorique de la valeur x_i est donné par :

$$n_{i,th} = n \times P(X = x_i) = n \times e^{-\lambda} \times \frac{\lambda^{x_i}}{(x_i)!} = 52 \times e^{-1,5} \times \frac{(1,5)^{x_i}}{(x_i)!}$$

On peut donc compléter le tableau précédent avec les effectifs théoriques :

Nbre ventes par mois	0	1	2	3	4	5	> 5
Nbre d'agents $(n_{i,obs})$	18	18	8	5	2	1	0
$P(X = x_i)$	0.22	0.33	0.25	0.13	0.05	0.01	0.004
Effectif théorique $(n_{i,th})$	11.60	17.40	13.05	6.53	2.45	0.73	0.23

Etant donné que certains effectifs sont < 5, nous devons regrouper certaines modalités afin de pouvoir effectuer un test du khi-deux.

Nbre ventes par mois	0	1	2	> 2
Nbre d'agents $(n_{i,obs})$	18	18	8	8
$P(X = x_i)$	0.22	0.33	0.25	0.19
Effectif théorique $(n_{i,th})$	11.60	17.40	13.05	9.94
Distance du χ^2	3.53	0.02	1.96	0.38

La distance du χ^2 entre les deux distributions est :

$$d_{\chi^2} = \sum_{i=1}^{4} \frac{(n_{i,th} - n_{i,obs})^2}{n_{i,th}} = \frac{(18 - 11.6)^2}{11.6} + \frac{(18 - 17.4)^2}{17.4} + \frac{(8 - 13.05)^2}{13.05} + \frac{(8 - 9.94)^2}{9.94}$$

$$d_{\chi^2} = 5.88$$

Nous nous retrouvons donc avec 4 modalités. Sous (H_0), la distance du khi-deux suit justement une loi du khi-deux à $4 - 1 = 3$ degrés de liberté : $D_{\chi^2} \sim \chi^2_3$.
La région critique a la forme : $W = \{D_{\chi^2} > C\}$.
Avec un risque de $\alpha = 5\%$, on trouve un seuil de $C = 7.82$.
Nous avons $d_{\chi^2} < C$, et donc devons valider (H_0). L'échantillon suit bien la loi de Poisson indiquée.
2) Lorsque le paramètre λ de la loi de Poisson n'est pas connu, sa valeur est estimée à l'aide de l'échantillon lui même.

Cela a pour effet d'enlever un degré de liberté à la loi du khi-deux suivie par la distance sous l'hypothèse (H_0). Nous aurons donc : $D_{\chi^2} \sim \chi^2_2$
Le seuil à 5% est alors de : $C = 5.99$
La valeur de λ qui sera retenue est celle donnée par la moyenne empirique $\overline{X}$.
Ici $\overline{x} = \frac{1}{52}(0 \times 18 + 1 \times 18 + 2 \times 8 + 3 \times 5 + 4 \times 2 + 50 \times 1) = 1.19$.
On refait donc les calculs avec $\lambda = 1.19$

Nbre ventes par mois	0	1	2	> 2
Nbre d'agents $(n_{i,obs})$	18	18	8	8
$P(X = x_i)$	0.30	0.36	0.22	0.12
Effectif théorique $(n_{i,th})$	15.82	18.83	11.20	6.15
Distance du χ^2	0.30	0.04	0.91	0.55

Ce qui donne $d_{\chi^2} = 1.81$.
Nous avons toujours $d_{\chi^2} < C$, et donc devons valider (H_0). L'échantillon suit bien la loi de Poisson indiquée. La conclusion est plus affirmative car la distance est nettement plus petite que le seuil.
-

Exercice 2 -

Nous disposons de $n = 10$ valeurs prises par une variable X dont nous voudrions nous assurer qu'elle est gaussienne. Nous utiliserons pour cela le **test de Kolmogorov-Smirnov**.
Voici les données ordonnées par ordre croissant :

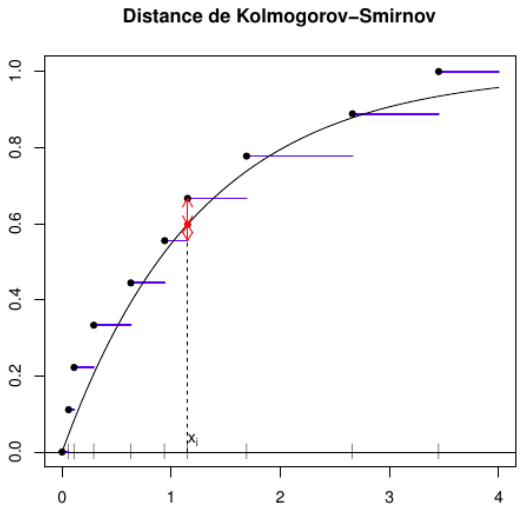
x_i	10.8	10.9	11.9	13.5	15.9	16.6	17.4	17.9	18.7	23

La moyenne empirique est de $\overline{x} = 15.66$ et l'écart-type empirique corrigé est de $s^* = 3.90$

1. A quelle loi normale précise va-t-on vérifier l'adéquation de notre échantillon ?
2. La fonction de répartition empirique est donnée par $F_n(x) = \dfrac{Card\{i,\ x_i \leqslant x\}}{n}$

   Lorsque les x_i sont ordonnés, $F_n(x_i) = \dfrac{i}{n}$.

   Cette fonction empirique sera comparée à la fonction de répartition théorique F de la loi normale citée dans la question précédente. Celle-ci est obtenue, après centrage et réduction, à partir de la table de la loi normale standard $(F(x_i) = F_Z(\frac{x_i - 15.66}{3.9}))$.



   Complétez le tableau suivant :

i	1	2	3	4	5	6	7	8	9	10
x_i	10.8	10.9	11.9	13.5	15.9	16.6	17.4	17.9	18.7	23
$F(x_i)$										
$F_n(x_i) = \frac{i}{n}$										
$\left\|F(x_i) - \frac{i}{n}\right\|$										
$\left\|F(x_i) - \frac{i-1}{n}\right\|$										

3. La variable de décision est de ce test d'adéquation est
   $D_{KS} = \underset{i=1,..,n}{Max} \left\{ \left|F(x_i) - \frac{i}{n}\right| , \left|F(x_i) - \frac{i-1}{n}\right| \right\}$
   Calculer d_{KS} la valeur prise dans notre échantillon.
4. La région critique a la forme : $W = \{D_{KS} > C\}$, et le seuil C est déterminé grâce à la table ci-jointe.

   Quelle conclusion obtient-on si on se fixe un risque $\alpha = 5\%$?

Correction : -

1. Loi normale attendue : $\mathcal{N}(15.66; (3.9)^2)$.
2. $F_n(x_i) = \frac{i}{n}$.
   $F(x_i) = F_Z(\frac{x_i - 15.66}{3.9}))$.

i	1	2	3	4	5	6	7	8	9	10
x_i	10.8	10.9	11.9	13.5	15.9	16.6	17.4	17.9	18.7	23
$F(x_i)$	0.106	0.111	0.167	0.290	0.525	0.595	0.672	0.717	0.782	0.970
$F_n(x_i) = \frac{i}{n}$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
$\left\|F(x_i) - \frac{i}{n}\right\|$	0.007	0.089	**0.132**	0.110	0.025	0.005	0.028	0.083	0.118	0.030
$\left\|F(x_i) - \frac{i-1}{n}\right\|$	0.107	0.011	0.032	0.010	0.125	0.095	0.072	0.017	0.018	0.070

3. La variable de décision est de ce test d'adéquation est
   $D_{KS} = \underset{i=1,..,n}{Max} \left\{ \left|F(x_i) - \frac{i}{n}\right| , \left|F(x_i) - \frac{i-1}{n}\right| \right\}$.
   Ici, nous obtenons : $d_{KS} = 0.132$
4. Avec un risque $\alpha = 5\%$, la table nous donne un seuil de $C = 0.409$.
   $d_{KS} = 0.132 < 0.409 = C$, donc l'échantillon suit bien la loi normale annoncée.

Exercice 3 -

On souhaite valider un programme simulant une variable aléatoire de loi normale $\mathcal{N}\left(109, (0.5)^2\right)$.
Pour cela on regroupe les données en classe et on applique un test du khi-deux comme s'il s'agissait d'une variable discrète.
On commence par générer $n = 500$ valeurs qui se répartissent de la manière suivante :

1. Peut-on considérer que le programme est correct avec un risque $\alpha = 5\%$, $\alpha = 1\%$? (on pourra s'aider d'un tableur pour calculer les effectifs théoriques).
2. Supposons maintenant que μ et σ^2 soient inconnus mais estimés sur l'échantillon. Peut-on considérer que le programme est correct ?

Correction -

Classe	Effectif	Classe	Effectif
]107.8 , 108]	5	]109 , 109.2]	75
]108 , 108.2]	5	]109.2 , 109.4]	85
]108.2 , 108.4]	30	]109.4 , 109.6]	55
]108.4 , 108.6]	40	]109.6 , 109.8]	30
]108.6 , 108.8]	70	]109.8 , 110]	10
]108.8 , 109]	85	]110 , 110.2]	10

1) Nous allons tester $\begin{cases}(H_0) & \text{l'échantillon suit } \mathcal{N}\left(109,(0.5)^2\right)\\ (H_1) & \text{l'échantillon ne suit pas } \mathcal{N}\left(109,(0.5)^2\right)\end{cases}$

Pour déterminer l'effectif théorique de la classe $]a\ ,b]$ sous l'hypothèse (H_0) : $X \sim \mathcal{N}\left(109,(0.5)^2\right)$, on calcule la probabilité $P(a < X \leqslant b)$ grâce à la table de la loi normale, puis on multiplie par $n = 500$.
Pour tenir compte du fait que la loi normale prend toutes les valeurs réelles possibles, la première classe $]a, b]$ doit être remplacée par $]-\infty, b]$. De même la dernière classe devrait être remplacée par $[a, +\infty[$.
La variable de décision est D_{χ^2} la distance du khi-deux entre la distribution observée et la distribution théorique de la loi normale $\mathcal{N}\left(109,(0.5)^2\right)$.
La région critique est de la forme, $W = \{D_{\chi^2} > C\}$ et sous l'hypothèse (H_0) , D_{χ^2} suit une loi du chi-deux à $(12-1) = 11$ d.d.l. car il y a 12 classes (modalités).
Si on prend un rique $\alpha = 5\%$. On lit dans la table χ^2_{11} , $C = 19.68$.
Pour calculer la valeur de la statistique du test sur l'échantillon,voici le tableau
des effectifs théoriques (obtenu grâce à un tableur).

Classe	Borne inf. a	Borne sup. b	Eff.	$P(a \leqslant X \leqslant b)$	Eff. théo	dist. χ^2
]107.8 , 108]	$-\infty$	108	5	0.023	11.38	3.573
]108 , 108.2]	108	108.2	5	0.032	16.02	7.585
]108.2 , 108.4]	108.2	108.4	30	0.060	30.14	0.001
]108.4 , 108.6]	108.4	108.6	40	0.097	48.39	1.456
]108.6 , 108.8]	108.6	108.8	70	0.133	66.36	0.200
]108.8 , 109]	108.8	109	85	0.155	77.71	0.684
]109 , 109.2]	109	109.2	75	0.155	77.71	0.095
]109.2 , 109.4]	109.2	109.4	85	0.133	66.36	5.235
]109.4 , 109.6]	109.4	109.6	55	0.097	48.39	0.902
]109.6 , 109.8]	109.6	109.8	30	0.060	30.14	0.001
]109.8 , 110]	109.8	110	10	0.032	16.02	2.265
]110 , 110.2]	110	$+\infty$	10	0.023	11.38	0.166

On note que tous les effectifs théoriques sont supérieurs à 5, il n'y a donc pas besoin de regrouper de classes.
La distance du khi-deux obtenue est :
$d_{\chi^2} = \sum_{i=1}^{12} \frac{(n_{i,th}-n_{i,obs})^2}{n_{i,th}} = \frac{(11.38-5)^2}{11.38} + ... + \frac{(11.38-10)^2}{11.38}$
$d_{\chi^2} = 3.573 + 7.585 + ... + 0.166 = 22.16$
Nous avons $d_{\chi^2} = 22.16 > C$, et donc devons valider (H_1). L'échantillon ne suit pas la loi de normale indiquée.
Cela ne veut pas dire qu'il n'est pas gaussien. Il est possible que ce soit les valeurs des paramètres qui ne conviennet pas.
2) Si maintenant on suppose que les paramètres $\mu = 109$ et $\sigma^2 = (0.5)^2$ sont des valeurs estimées sur l'échantillon alors il faut enlever un degré de liberté par paramètre estimé.
La statistique du test suit donc une loi du chi-deux à $(12-1-2) = 9$ d.d.l.
Le seuil est alors égal à $C = 16,92$.
On a toujours $d_{\chi^2} = 22.16 > C$, donc la décision ne change pas mais on peut dire maintenant que l'échantillon n'est pas gaussien.
-

Exercice 4 -

Pour comparer l'efficacité de deux médicaments semblables mais de prix très différents (un médicament

d'origine et un générique), la sécurité sociale a effectué une enquête sur les guérisons obtenues avec les deux traitements.
Les résultats sont présentés dans le tableau suivant :

	Médicament d'origine	Générique
Guérison	156	44
Non guérison	44	6

Peut-on considérer que l'efficacité du médicament est indépendante de son coût ?

-

Correction -

Distribution observée :

	Médicament d'origine	Générique	Total
Guérison	156	44	200
Non guérison	44	6	50
Total	200	50	250

Distribution théorique (en cas d'indépendance) :

	Médicament d'origine	Générique	Total
Guérison	160	40	200
Non guérison	40	10	50
Total	200	50	250

La statistique du test est D_{χ^2} la distance du chi-deux entre la distribution observée et la distribution théorique dans le cas de l'indépendance des variables.
La région critique est de la forme, $W = \{D_{\chi^2} > C\}$ et sous l'hypothèse (H_0) , D_{χ^2} suit une loi du khi-deux à $(2-1)(2-1) = 1$ d.d.l.
Si on prend un risque $\alpha = 5\%$. On lit dans la table, $C = 3.84$.

$$d_{\chi^2} = \sum_{i=1}^{4} \frac{(n_{i,th} - n_{i,obs})^2}{n_{i,th}} = \frac{(160-156)^2}{160} + \frac{(40-44)^2}{40} + \frac{(40-44)^2}{40} + \frac{(10-6)^2}{10}$$

On obtient : $d_{\chi^2} = 2.5$
$d_{\chi^2} < 3,84$ donc on garde (H_0). On peut donc dire que le type de médicament n'a pas d'impact sur la guérison.

-

Exercice 5 -

Pendant 200 durées d'une minute, on a noté le nombre de voitures arrivant au poste de péage sur l'autoroute dans le tableau ci-dessus.

Nbre voitures par mn	0	1	2	3	4	5	6	7	8	9	10	11	Total
Effectif observé	1	15	30	46	38	30	16	13	5	3	2	1	200

N.B. *1 seule durée d'une minute a vu passe 0 voiture. 38 durées d'une minute ont vu passer exactement 4 voitures.*
On vous demande de déterminer la loi de probabilité de cet échantillon.

-

Correction -

Soit X le nombre de voitures arrivant au poste de péage par minute. On a un échantillon de taille $n = 200$.
On suppose que $X \sim \mathcal{P}(\lambda)$ suit une loi de Poisson et on estime son paramètre par la moyenne empirique des observations, $\overline{x} = \dfrac{800}{200} = 4$.
La question est donc de savoir si l'échantillon suit une loi $\mathcal{P}(4)$.
Si cela était le cas alors l'effectif théorique de la valeur k est donné par :

$n_{théo} = n \times P(X = k) = ne^{-\lambda}\frac{\lambda^k}{k!} \ = \ 200 \times e^{-4}\frac{4^k}{k!}$
On peut donc compléter le tableau avec ces effectifs théoriques et ensuite calculer la distance (du khi-deux) entre les 2 distributions, l'observée et la théorique.

Nbre voitures par mn (k)	0	1	2	3	4	5	6	7	8	9	10	11	>11
Effectif observé	1	15	30	46	38	30	16	13	5	3	2	1	0
$P(X = k)$	0.02	0.07	0.15	0.2	0.2	0.16	0.1	0.06	0.03	0.01	0.01	0	0
Effectif théorique	3.66	14.7	29.3	39.1	39.1	31.3	20.8	11.9	5.95	2.65	1.06	0.38	0.18

Les effectifs des valeurs extrêmes étant petits (< 5), on procède à un regroupement.
On regroupe les modalités 0 et 1 d'un côté, et toutes les modalités $\geqslant 8$ de l'autre.
On obtient alors :

Nbre voitures par mn (k)	0-1	2	3	4	5	6	7	$\geqslant$8
Effectif observé	16	30	46	38	30	16	13	11
$P(X = k)$	0.09	0.15	0.2	0.2	0.16	0.1	0.06	0.05
Effectif théorique	18.3	29.3	39.1	39.1	31.3	20.8	11.9	10.2
Distance χ^2	0.29	0.02	1.23	0.03	0.05	1.12	0.1	0.06

La région critique est de la forme, $W = \{D_{\chi^2} > C\}$ et sous l'hypothèse (H_0) , D_{χ^2} suit une loi du khi-deux à $(8-1-1) = 6$ d.d.l. En effet, $\lambda = 4$ ayant été obtenu par estimation à partir de l'échantillon, on doit enlever un degré de liberté supplémentaire.
Si on prend un risque $\alpha = 5\%$. On lit dans la table, $C = 12.6$.
$d_{\chi^2} = \sum_{i=1}^{4} \frac{(n_{i,th} - n_{i,obs})^2}{n_{i,th}} = \frac{(16-18.3)^2}{18.3} + ... + \frac{(11-10.2)^2}{10.2}$
On obtient : $d_{\chi^2} = 3.196$
$d_{\chi^2} < C$ donc on garde (H_0). On peut donc dire que X suit bien la loi de Poisson de paramètre $\lambda = 4$.
-

Exercice 6 -

Le tableau ci-dessous donne la répartition par taille (cm) de $n = 2700$ salariés masculins par catégorie socio-professionnelle (CSP) : :

Taille - CSP	Ouvrieres	Employes	Cadres	Total
Moins de 165 cm	325	66	22	413
165 - 170 cm	488	110	51	649
170 - 175 cm	636	158	123	917
175 cm et plus	451	146	124	721
Total	1900	480	320	2700

Peut-on considérer que la taille soit indépendante de la CSP ?
-

Correction -

Distribution théorique (en cas d'indépendance) :

Taille - CSP	Ouvrieres	Employes	Cadres	Total
Moins de 165 cm	290.63	73.42	48.95	413
165 - 170 cm	456.70	115.38	76.92	649
170 - 175 cm	645.30	163.02	108.68	917
175 cm et plus	507.37	128.18	85.45	721
Total	1900	480	320	2700

La statistique du test est D_{χ^2} la distance du chi-deux entre la distribution observée et la distribution théorique dans le cas de l'indépendance des variables.

La région critique est de la forme, $W = \{D_{\chi^2} > C\}$ et sous l'hypothèse (H_0) , D_{χ^2} suit une loi du khi-deux à $(4-1)(3-1) = 6$ d.d.l.
Si on prend un risque $\alpha = 5\%$. On lit dans la table, $C = 12.59$.

$$d_{\chi^2} = \sum_{i=1}^{4} \frac{(n_{i,th} - n_{i,obs})^2}{n_{i,th}} = \frac{(325 - 290.63)^2}{290.63} + ... + \frac{(124 - 85.45)^2}{85.45}$$

On obtient : $d_{\chi^2} = 59.09$
$d_{\chi^2} > 3,84 \implies$on doit donc rejeter (H_0) et valider (H_1). Les deux variables ne sont pas indépendantes. Il y a un lien entre catégorie socio-professionnelle et taille.

Annexe : Table de Kolmogorov-Smirnov -

$n \backslash^{\alpha}$	0.001	0.01	0.02	0.05	0.1	0.15	0.2
1		0.99500	0.99000	0.97500	0.95000	0.92500	0.90000
2	0.97764	0.92930	0.90000	0.84189	0.77639	0.72614	0.68377
3	0.92063	0.82900	0.78456	0.70760	0.63604	0.59582	0.56481
4	0.85046	0.73421	0.68887	0.62394	0.56522	0.52476	0.49265
5	0.78137	0.66855	0.62718	0.56327	0.50945	0.47439	0.44697
6	0.72479	0.61660	0.57741	0.51926	0.46799	0.43526	0.41035
7	0.67930	0.57580	0.53844	0.48343	0.43607	0.40497	0.38145
8	0.64098	0.54180	0.50654	0.45427	0.40962	0.38062	0.35828
9	0.60846	0.51330	0.47960	0.43001	0.38746	0.36006	0.33907
10	0.58042	0.48895	0.45662	0.40925	0.36866	0.34250	0.32257
11	0.55588	0.46770	0.43670	0.39122	0.35242	0.32734	0.30826
12	0.53422	0.44905	0.41918	0.37543	0.33815	0.31408	0.29573
13	0.51490	0.43246	0.40362	0.36143	0.32548	0.30233	0.28466
14	0.49753	0.41760	0.38970	0.34890	0.31417	0.29181	0.27477
15	0.48182	0.40420	0.37713	0.33760	0.30397	0.28233	0.26585
16	0.46750	0.39200	0.36571	0.32733	0.29471	0.27372	0.25774
17	0.45440	0.38085	0.35528	0.31796	0.28627	0.26587	0.25035
18	0.44234	0.37063	0.34569	0.30936	0.27851	0.25867	0.24356
19	0.43119	0.36116	0.33685	0.30142	0.27135	0.25202	0.23731
20	0.42085	0.35240	0.32866	0.29407	0.26473	0.24587	0.23152
25	0.37843	0.31656	0.30349	0.26404	0.23767	0.22074	0.20786
30	0.34672	0.28988	0.27704	0.24170	0.21756	0.20207	0.19029
35	0.32187	0.26898	0.25649	0.22424	0.20184	0.18748	0.17655
40	0.30169	0.25188	0.23993	0.21017	0.18939	0.17610	0.16601
45	0.28482	0.23780	0.22621	0.19842	0.17881	0.16626	0.15673
50	0.27051	0.22585	0.21460	0.18845	0.16982	0.15790	0.14886
OVER 50	$\frac{1.94947}{\sqrt{n}}$	$\frac{1.62762}{\sqrt{n}}$	$\frac{1.51743}{\sqrt{n}}$	$\frac{1.35810}{\sqrt{n}}$	$\frac{1.22385}{\sqrt{n}}$	$\frac{1.13795}{\sqrt{n}}$	$\frac{1.07275}{\sqrt{n}}$