Notions de statistiques pour les TIPE

BCPST Spé 2 Lycée Champollion Grenoble

Table des matières

I Vocabulaire

Variable ou charactère Une **variable** est une **caractéristique mesurable** qui peut prendre différentes valeurs. Taille, concentration d'un élément chimique,allèle, temps, réussite à un test. On distingue les variables en fonction du type de donnée.

Variable qualitative/catégorielle/nominale Ce sont des variables qui ne sont pas quantifiables : allèle, couleur des yeux, rang dans un classement. On ne peut pas calculer de moyenne sur ce type de variable.

Variable numérique discrète Ce sont des variables qui sont quantifiables et dont les valeurs ne peuvent prendre qu'un nombre fini (ou en théorique dénombrable) de valeurs : nombre d'individu dans une population, nombre de cellule ...

Variable numérique continue C'est une variable quantifiable qui peut prendre toutes les valeurs réelles dans un intervalle donné : taille, masse, concentration, ...

Données statistiques brutes/ échantillon/ Série Ensemble des données sur lesquelles on effectue le travail statistique.

- Si on choisit une partie de la population étudiée et que les mesures ne sont faites que sur ces individus, les données forment un **échantillon** que l'on espère représentatif.
- Si les données représentent l'évolution d'un phénomène au cours du temps, c'est alors une **série temporelle**.
- Si les données sont comptabilisées par classes et fréquences on a alors une **série statistique**. Si ce travail de classement n'a pas été effectué on dit que la **série est brute**.

Statistique descriptive/représentation graphique On cherche à donner une description des données pour cela on calcule des descripteurs : moyenne, écart-type ...

Attention : Les calculs sont effectués sur l'ensemble des données.

On peut aussi représenter les données sous forme de tableau ou de graphes (en barre , histogramme, évolution des données en fonction du temps)

Statistique inférentielle/ estimation On suppose que le phénomène que l'on étudie suit une loi de probabilité classique, et on cherche à l'aide des mesures à estimer les paramètres de cette loi.
Exemple :

- On dispose d'une pièce et on cherche à estimer le paramètre p probabilité d'obtenir un pile.
- On estime que la taille d'un individu suit une loi normale dont on cherche à estimer l'espérance et la variance en mesurant un échantillon de la population.

Souvent on ne fait pas d'hypothèse sur la loi suivie par le phénomène mais on cherche à estimer les caractéristiques comme l'espérance et la variance
Exemple :

- On veut connaître la densité de myosotis dans les Alpes en fonction de l'altitude. On ne peut pas recenser l'ensemble de ces plantes sur l'ensemble des Alpes, on choisit donc quelques parcelles où l'on effectue des comptages, pour en déduire des densités moyennes.
- On veut connaître le taux de prévalence d'une mutation dans une population. On ne peut pas tester tous les individus, on choisit donc un échantillon, on compte le nombre d'individus portant la mutation et on calcule la fréquence des individus portant cette mutation, ce qui nous donne une estimation de la fréquence de la mutation dans la population globale.

Attention : Dans ce cas là contrairement aux statistiques descriptives les calculs ne sont pas effectués sur l'ensemble de la population (ce qui serait impossible) mais sur un échantillon. L'une des questions qui se pose est de connaître la qualité de notre estimation en fonction de la taille de notre échantillon.

Tests statistiques Lorsque l'on effectue des expériences on cherche à valider ou invalider une hypothèse. Un **test statistique** permet de savoir si les données mesurées permettent ou non de valider l'hypothèse faite.

Ma pièce est elle truquée?

Tel allèle influence t'elle la survie des individus.

II Rappels de statistiques descriptives pour une variable qualitative

II.1 Descripteur de position la moyenne

Soit $(x_i)_{i\in[\![1,n]\!]}$ une série brute et $((y_i,n_i))_{i\in[\![1,p]\!]}$ la série classée associée où y_i est une modalité et n_i un effectif.

Définition 1 (Moyenne)**.**

$$\overline{x} = \frac{\sum\limits_{i\in[\![1,n]\!]} x_i}{n} = \frac{\sum\limits_{i\in[\![1,p]\!]} n_i y_i}{\sum\limits_{i\in[\![1,p]\!]} n_i}$$

Outils pour un calcul de moyenne

Python/numpy méthode `mean()` ou fonction `numpy.mean`.
Python/pandas méthode `mean()`
Excel `MOYENNE()`

II.2 Descripteurs de dispersion

II.2.a Variance et écart-type

Définition 2 (Variance et écart-type)**.**
*La **variance** de la série est*

$$\sigma_x^2 = \frac{\sum\limits_{i\in[\![1,n]\!]} (x_i-\overline{x})^2}{n} = \frac{\sum\limits_{i\in[\![1,p]\!]} n_i(y_i-\overline{x})^2}{\sum\limits_{i\in[\![1,p]\!]} n_i} = \overline{x^2}-\overline{x}^2$$

L'écart-type est la racine de la variance, c'est la notion la plus utilisée en statistique, on peut remarquer que l'écart-type à la même unité que les données de la série.
L'écart-type est un descripteur de dispersion relatif, il mesure la dispersion autour de la moyenne, plus l'écart-type est grand plus les valeurs s'éloigne de la moyenne en positif et négatif. Une série dont l'acart-type est nul est constante.

Proposition 1 (König-Huygens)**.**

$$\sigma_x^2 = \overline{x^2}-\overline{x}^2$$

Définition 3 (Écart-type corrigé)**.**

$$\sigma'^2_x = \frac{\sum\limits_{i\in[\![1,n]\!]} (x_i-\overline{x})^2}{n-1}$$

et

$$\sigma'_x = \sqrt{\frac{n}{n-1}}\sigma_x$$

Attention : L'écart-type corrigé n'est pas au programme de première année, on ne peut pas utiliser la formule de König-Huygens mais cet écart-type corrigé est utile dans certaines formules, et c'est un estimateur non baisée de l'écart-type d'un échantillon. (*cf* annexe)

Comment choisir quel écart-type utiliser

- Si le nombre de données est grand cela n'a pas d'importance.
- Si vous voulez calculer un écart-type sur la population totale de votre étude utilisez l'écart-type non corrigé.
- Si vous étudiez un échantillon pour estimer l'écart-type de la population globale utilisez l'écart-type corrigé.

Outils pour un calcul d'écart-type

Python/numpy fonction `numpy.std()` ou la méthode `.std()` par défaut calcule l'écart-type non corrigé. Pour obtenir l'écart-type corrigé utilisez l'option[a] `ddof=1`. `np.std(tableau,ddof=1)` ou `tableau.std(ddof=1)`

Python/pandas Méthode `.std()` PAr défaut calcule l'écart-type corrigé. Pour calculer l'écart-type corrigé utilisez l'option `ddof=0` par exemple `tables.std(ddof=0)`

Excel `ECARTYPE.STANDARD` pour l'écart type corrigé et `ECARTYPE.PEARSON` pour l'écart-type non corrigé.

a. ddof : degres of freedom

III Incertitude et barre d'erreur

III.1 But

On fait plusieurs mesures et on calcule la moyenne des résultats. On voudrait savoir si on peut avoir confiance au résultat obtenu, on ne s'intéresse pas aux incertitudes de mesure, mais aux incertitude dues à l'échantillonnage.

Outils pour afficher une barre d'erreur

Il est conseillé, quel que soit le logiciel utilisé, de calculer soit même la barre d'erreur, et de mettre dans la légende de la figure le type de barre d'erreur utilisé.

III.2 Écart-type

On calcule la moyenne $\overline{x}$ et l'écart-type et on donner comme incertitude l'intervalle

$$[\overline{x}-\sigma_x;\overline{x}+\sigma_x]$$

On peut montrer que pour une variable aléatoire Z suivant une loi normale d'espérance m et d'écart-type σ

$$P(m-\sigma\leqslant Z\leqslant m+\sigma)=2\Phi(1)-1\simeq 0.68 \qquad \text{et} \qquad P(m-2\sigma\leqslant Z\leqslant m+2\sigma)=2\Phi(2)-1\simeq 0,95$$

On peut voir cet intervalle comme une description de la dispersion d'une population.

III.3 Erreur standard à la moyenne

$$\left[\overline{x}-\frac{\sigma_x}{\sqrt{n}};\overline{x}+\frac{\sigma_x}{\sqrt{n}}\right] \text{ ou } \left[\overline{x}-\frac{\sigma'_x}{\sqrt{n}};\overline{x}+\frac{\sigma'_x}{\sqrt{n}}\right]$$

avec l'écart-type ou l'écart-type corrigé.

III.4 Grande séries (plus de trente données)

On veut obtenir un information du type « Si je considère que la valeur que je cherche à mesurer est dans l'intervalle $[\overline{x}-a;\overline{x}+a]$ quelle est la probabilité d'avoir raison/de me tromper ». Si on note m la moyenne théorique

$$P\left(m\in\left[\overline{x}-t\frac{\sigma_x}{\sqrt{n}};\overline{x}+t\frac{\sigma_x}{\sqrt{n}}\right]\right)\geqslant 1-\alpha$$

La probabilité que la valeur recherchée soit dans cet intervalle doit être plus grand que le niveau de confiance $1-\alpha$.

cas $t=1$ L'écart standard à la moyenne correspond à un intervalle de confiance au niveau de confiance 0.68.

cas $t=1.96$ **ou** $t=2$ le niveau de confiance est 95%

Cas général pour avoir un niveau de confiance de $1-\alpha$ avec $\alpha\in]0;1[$ proche de 0 il faut choisir t tel que $\Phi(t)=\dfrac{1+\alpha}{2}$

Exemple 1 (Mercure)**.**
Dans [BY] les auteurs ont mesuré le taux de mercure dans les tissus musculeux de truites d'un lac d'Alaska (en ng par g de muscle) [3720,2290,3200,4000, 2810 3370, 2960, 3210, 1630, 2010, 4600, 1980, 2090, 2570, 13400, 2440, 2660, 1680, 2790, 3310, 632, 1670, 1930, 1500, 2320, 1360, 1280, 910, 1180, 1670, 1180, 1170, 2950, 3630, 1890, 739, 2120, 1210, 1600, 2580]

- Il y a 40 mesures
- La moyenne est de 2506.025
- l'écart-type est de 1976.82
- l'écart à la moyenne est de $\dfrac{1976.82}{\sqrt{40}}\simeq 312$
- L'intervalle de confiance à 95% est $[2506.025-1.96\times 312;2506.025-1.96\times 312]=[2194;2818]$

III.5 Petites séries (loi de Student)

L'intervalle de confiance est encore de la forme

$$\left[\overline{x}-t\frac{\sigma'_x}{\sqrt{n}};\overline{x}+t\frac{\sigma'_x}{\sqrt{n}}\right]$$

Remarque : Comme le nombre de données est faible on utilise l'écart-type corrigé.

La probabilité que la valeur recherchée soit dans cet intervalle doit être plus grand que le niveau de confiance $1-\alpha$. Cette fois on ne peut plus utiliser une loi normale, on fait le raisonnement précédent mais en considérant que $\overline{X_n}^*$ suit une loi de Student à $n-1$ degrés de libertés où n est la taille de la série.

Exemple 2 (Poissons)**.**
Dans [McD] on trouve un extrait d'une étude où les auteurs ont mesurés le nombre de Naseux noir (*Rhinichthys atratulus*) sur des sections de 75 mètres le long de cours d'eau du Maryland. [76,102,12,39,55,93,98,53,102].

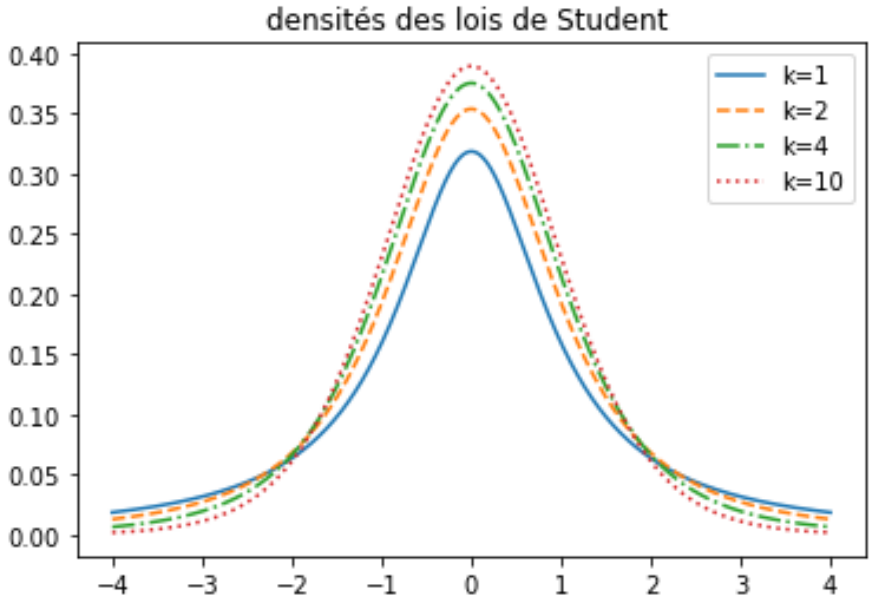
- La taille de la série est 9, on va utiliser une loi de Student à 8 degrés de libertés.
- La moyenne est $m=70.0$ et l'écart-type corrigé est $\sigma'\simeq 32$.
- L'erreur standard est $\dfrac{\sigma'}{\sqrt{9}}\simeq 10.7$.

- Comme précédemment

$$P(m \in \left[\overline{x} - t\frac{\sigma'}{\sqrt{n}}; \overline{x} + t\frac{\sigma'}{\sqrt{n}}\right]) = P(-t \leqslant T \leqslant t)$$

  mais cette fois ci T suit une loi de Student
- Par symétrie $P(-t \leqslant T \leqslant t) = 0.95$ si et seulement si $2P(T > t) = 1 - 0.95$
- On cherche sur les tables que, l'on trouve [Lar], et la valeur approximative de t qui convient est 2.3
- L'intervalle de confiance au niveau de confiance 95% est

$$[45.39; 94.61]$$



On remarque que si n devient grand les densités des lois de Student ressemble fortement à des Gaussiennes

III.6 Incertitude sur une fréquence

Si on cherche à déterminer une fréquence on utilise l'intervalle de confiance à 95%, f désignant la fréquence calculée sur l'échantillon et n sa taille

$$\left[f - 1.96\sqrt{\frac{f(1-f)}{n}}; f + 1.96\sqrt{\frac{f(1-f)}{n}}\right]$$

Il faut vérifier

- La taille de l'échantillon n est plus grande que 30
- **ou bien** La taille de l'échantillon n et la fréquence de l'échantillon vérifient $nf \geqslant 5$ et $n(1-f) \geqslant 5$

Exemple 3 (Gaucher droitier)**.**

IV Tests statistiques

IV.1 Un exemple de test exact

Exemple 4 (Lancer d'une pièce)**.**
On lance une pièce inconnue 100 fois on obtient 98 piles on décide donc que cette pièce est truquée, mais il est possible même avec une pièce non truquée d'obtenir un tel résultat ou un résultat encore plus déséquilibré. On cherche à éliminer l'hypothèse

$$\mathscr{H}_0 \text{ « La pièce est non truquée la probabilité d'obtenir pile est } 1/2 \text{ »}$$

Si on note X la variable aléatoire qui donne le nombre de pile obtenu lors de 100 lancers identiques et indépendants avec une pièce non truquée, $X \hookrightarrow \mathscr{B}(100, 0.5)$

La probabilité d'obtenir 98 piles ou plus est

$$\begin{aligned} p &= P(X=100) + P(X=99) + P(X=98) \\ &= \binom{100}{100}\left(\frac{1}{2}\right)^{100} + \binom{100}{99}\left(\frac{1}{2}\right)^{100} + \binom{100}{98}\left(\frac{1}{2}\right)^{100} \\ &= \left(1 + 100 + \frac{100\times 99}{2}\right)\left(\frac{1}{2}\right)^{100} \\ &\simeq 3.9 \times 10^{-27} \end{aligned}$$

Exercice 1 (Un dé)**.**
On lance 600 fois un dé et on obtient 50 fois la face 1. Quelle est la probabilité de se tromper si on considère que le dé est truqué ? Pour calculer la valeur numérique vous pouvez utilisez python. (On trouve $2.04935265772009 \times 10^{-09}$)

IV.2 Hypothèses, valeur p et type de Tests

Définition 4 (Hypothèse nulle $\mathscr{H}_0$)**.**
*C'est l'hypothèse que l'effet observé ne soit qu'au hasard et qu'il n'y a pas de différence significative dans les données. Sa négation est l'**hypothèse alternative**.*

Dans l'exemple 4 l'hypothèse nulle est « la pièce est non truquée », dans l'exercice 1 « la pièce est non truquée »

Définition 5 (p-value/valeur p)**.**
C'est la probabilité[1] *en supposant l'hypothèse nulle vérifiée d'obtenir un résultat au moins aussi extrême que celui observé lors de la mesure des données. On veut souvent que cette valeur soit inférieure à 0.05 voir 0.01. C'est donc la probabilité de se tromper en rejetant l'hypothèse nulle.*

On peut vouloir tester trois choses

- Peut on considérer que notre série de valeurs est significativement différente d'une valeur théorique
- Peut on considérer deux séries de données comme différentes ?
- Deux séries de données sont elles indépendantes ?

1. sous une modélisation statistique donnée

Dans l'exemple 4 nous avons vérifié que notre série de lancers avait une moyenne ou fréquence de pile significativement différente de celle supposé pour une pièce non-truquée. Il rentre dans la première catégorie

On se trouve vite limité, même avec un ordinateur, pour faire des tests exacts comme celui présenté précédemment.

1. Calculer une certaine valeur construite à partir donnés c'est **la statistique de test**.
2. **Si l'hypothèse nulle est vraie**, cette statistique doit suivre une loi donnée (normale, student , χ^2). Il faut souvent faire des approximations (liées aux théorèmes limites vus en fin d'année) et d'autres hypothèses, par exemple des séries suivent une loi normale, les variances sont identiques,.... Les démonstrations de ces résultats sont hors de notre portée cette année.
3. Utiliser les tables des fonctions de répartition pour calculer la valeur p.

Bien sur toutes ces étapes peuvent être réalisées automatiquement avec les bons outils, le but ici est de comprendre le fonctionnement de ces tests.

IV.3 Test de Student

Le but de ce test est de distinguer si deux séries ont des moyennes significativement différentes. On fait l'hypothèse que les deux variables à distinguer suivent des lois normales[2] identiques, ou du moins des lois proches.
Attention : *Pour une première approche ; on peut se contenter de calculer la moyenne et une incertitude comme dans la partie précédente, si les intervalles de confiance sont disjoints, on peut considérer que les moyennes sont disjointes.*

Exemple 5 (Populations de dahuts)**.**
Après de longs mois dans le massif de Belledonne un biologiste a mesuré la masse des Dahuts [Bou24].

Dextrogyre	Lévrogyre
43.6	58.8
56.3	60.8
57.1	61.
63.3	61.7
63.8	60.8
56.7	62.2
64.3	59.4
68.8	62.8
82.4	60.
54.3	62.2
70.3	62.1
50.	59.9
60.1	60.7
63.8	61.8
44.3	62.2
54.4	

Mesure de masse (kg) sur les dahuts lévogyres et dextrogyres

On voudrait savoir si les sous-espèces ont des masses significativement différentes

$\mathscr{H}_0$ « les deux sous espèces ont en moyenne la même masse »

2. On peut tester cette hypothèse avec un test de Henry

```
from scipy.stats import f_oneway
X1=np.array([43.6, 56.3, 57.1, 63.3, 63.8, 56.7, 64.3,
68.8, 82.4, 54.3, 70.3,
50. , 60.1, 63.8, 44.3, 54.4])
X2=np.array([58.8, 60.8, 61. , 61.7, 60.8, 62.2, 59.4,
 62.8, 60. , 62.2, 62.1,
59.9, 60.7, 61.8, 62.2])

_,pvalue=f_oneway(X1,X2)
print(pvalue)
```

On obtient un résultat de 0.56273904197224, on ne peut pas rejeter l'hypothèse nulle.

Outils pour un test Student

Python Disponible dans `scipy.stats` sous le nom `f_oneway`. renvoie deux valeurs : la valeur de la statistique et la valeur p.

Excel `TEST.STUDENT`

IV.4 ANOVA

Une généralisation du test précédent, à utiliser dans le cas où les données ont trois ou plus catégories, est le test ANOVA, qui permet de savoir si l'une des moyennes est significativement différentes des autres. Par contre ce type de test ne permet pas de savoir quelle série a une moyenne significativement des autres.

Exemple 6 (Dahuts dans différents massifs autour de Grenoble)**.**
D'après les auteurs de [Bou24]

Belledonne	Vanoise	Oisan	Chartreuse
53.3	56.5	59.1	53.3
60.6	61.9	62.	60.6
52.9	65.3	55.5	52.9
60.8	61.5	55.5	60.8
57.8	60.2	53.3	57.8
54.	56.1	58.6	54.
55.	59.3	61.4	55.
57.7	53.7	65.4	57.7
58.1	62.2	59.8	58.1
51.8	57.2	64.7	51.8
45.2	57.9	61.7	45.2
57.	70.8	58.7	57.
58.7	53.8		
61.1	62.9		
56.5			
69.2			

masse (kg) des dahuts dans différents massifs des Alpes du Nord

$\mathscr{H}_0$ la moyenne des masses des populations est égale dans tous les massifs.

```
B=np.array([53.3, 60.6, 52.9, 60.8, 57.8, 54. , 55. ,
 57.7, 58.1, 51.8, 45.2,57. ])
V=np.array([56.5, 61.9, 65.3, 61.5, 60.2, 56.1, 59.3,
 53.7, 62.2, 57.2, 57.9,70.8, 58.7, 61.1])
O=np.array([59.1, 62. , 55.5, 55.5, 53.3, 58.6, 61.4,
 65.4, 59.8, 64.7, 61.7,58.7, 53.8, 62.9, 56.5, 69.2])
C=np.array([53.3, 60.6, 52.9, 60.8, 57.8, 54. , 55. ,
 57.7, 58.1, 51.8, 45.2,57. ])

from scipy.stats import f_oneway
_,pvalue=f_oneway(B,V,O,C)
print(pvalue)
>>>0.003583032559099313
```

On peut rejeter l'hypothèse nulle. Mais il faut faire des tests plus poussés pour savoir quelle population est différentes des autres. En pratique vous pouvez vous contenter alors de choisir celle qui à la moyenne significativement différente des autres.

Outils pour un test ANOVA

Python/scipy Disponible dans `scipy.stats` sous le nom `f_oneway`. renvoie deux valeurs : la valeur de la statistique et la valeurs p.

Excel ANOVA ONEWAY, il semblerait qu'il faille rajouter un module https://support.microsoft.com/fr-fr/office/utiliser-l-utilitaire-d-analyse-pour-effectuer-une-analyse-de-donn%C3%A9es-complexe-6c67ccf0-f4a9-487c-8dec-bdb5a2cefab6

IV.5 Test d'indépendance : test exact de Fischer

Les calculs utilisent la loi hypergéométrique vu en DL.

Exemple 7 (Survenu d'un accident vasculaire)**.**
On trouve dans [Ber+06] les données suivantes pour la survenue d'un AVC ischémique chez les femmes prenant préventivement de l'aspirine ou non

	Aspirine	Placebo
AVC ischémique	176	230
Pas d'AVC	21035	21018

L'hypothèse nulle est

$\mathscr{H}_0$ « La prise d'aspirine n'a pas influence sur la survenue d'un AVC ischémique. »

```
from scipy.stats import fisher_exact
Donnees=[[176, 230], [21035, 21018]]
res=fisher_exact(Donnees)
print(res.pvalue)
0.008149655416292926
```

Outils Python

python/scipy `scipy.stats.fisher_exact` renvoie une estimation de la proportion et la p-value

Excel Il n'existe pas de fonction à utiliser directement, mais on peut utiliser `HYPGEOM.DIST` voir : https://www.statology.org/fishers-exact-test-excel/.

IV.6 Test d'indépendance : Test du $\chi 2$

Exemple 8 (Oiseaux)**.**
Dans [LAT+12] les auteurs ont recensé les populations d'oiseaux sur des rives de rivières de Californie. Ces oiseaux sont séparés selon leur espèces et le le type de végétation de la zone riparienne : dégradée ou restaurée.

	Dégradée	Restaurée
Roitelet à couronne rubis	677	198
Bruant à couronne blanche	408	260
Bruant de Lincoln	270	187
Bruant à couronne dorée	300	89
Orite	198	91
Bruant chanteur	150	50
Tohi tacheté	137	32
Troglodyte de Bewick	106	48
Grive solitaire	119	24
Junco ardoisé	34	39
Chardonneret mineur	57	15
Autre	457	125

L'hypothèse nulle est

$\mathscr{H}_0$ le type de végétation n'a pas d'influence sur la composition de la population d'oiseaux.

```
from scipy.stats import chi2_contingency
Oiseaux=[[677,198],[408,260],[270,187],[300,89],[198,91],[150,50],[137,32],[106,48],[119,24],[34,39],[57    ,15],[457,125]]
res=chi2_contingency(Oiseaux)
print(res.pvalue)
```

On obtient une p-value de $1.6046724030592576 \times 10^{-26}$, on peut rejeter l'hypothèse nulle, mais on n'a pas déterminé quelle est l'influence positive ou négative de la végétation sur chacune des espèces.

Outils pour un test du χ^2

Python/scipy.stats Disponible dans `scipy` sous le nom `chi2_contingency`

Excel `TEST.KHIDEUX`

A Ressources

- `www.biostathandbook.com` [McD] En anglais, de nombreux tests décrits avec leur limites et des applications tirées de la biologie, pas de théorie, les applications sont en excel , R et SAS qui sont des langages dédiés aux statistiques. Certains exemples et exercices de ce polycopié en sont tirés.
- Sur les différents type d'incertitudes `https://cahier-de-prepa.fr/bcpst2-lakanal/download?id=1024`. On y trouve des conseils pour tracer les barres d'erreur.

B Démonstrations

Proposition 2 (L'écart-type empirique corrigé)**.**
Soit $(X_1, X_2, \dots, X_n)$ *n variable aléatoire identiquement distribuées et indépendante telles que*

$$\forall i \in [\![1, n]\!] \qquad E(X_i) = m \qquad V(X_i) = \sigma^2$$

on pose

$$\overline{X_n} = \frac{1}{n}\sum_{k=1}^{n} X_k$$
$$S_n = \frac{1}{n}\sum_{k=1}^{n} (X_k - \overline{X}_n)^2$$
$$S'_n = \frac{n}{n-1} S_n = \frac{1}{n-1}\sum_{k=1}^{n} (X_k - \overline{X}_n)^2$$

alors

$$E(S_n) = \frac{n-1}{n}\sigma^2 \qquad E(S'_n) = \sigma^2$$

Démonstration :

1. On sait que (classique) $S_n = \frac{1}{n}\sum_{k=1}^{n} X_k^2 - \overline{X}_n^2$
2. Montrons que $E(S_n) = \frac{n-1}{n}\sigma^2$. En notant μ l'espérance des X_i et σ leur écart type On a déjà calculé

   $$E(\overline{X_n}) = \mu \qquad V(\overline{X_n}) = \frac{\sigma^2}{n}$$

   en utilisant la formule de Koening-Huyggens sur $\overline{X_n}$

   $$E(\overline{X_n}^2) = V(\overline{X_n}) + E(\overline{X_n})^2 = \frac{\sigma^2}{n} + \mu^2$$

   On a de même

   $$E(X_k^2) = V(X_k) + E(X_k)^2 = \sigma^2 + \mu^2$$

   $$\begin{aligned}
   E(S_n) &= E\left(\frac{1}{n}\sum_{k=1}^{n} X_k^2 - \overline{X}_n^2\right) && \text{question précédente}\\
   &= \frac{1}{n}\sum_{k=1}^{n} E\left(X_k^2\right) - E\left((\overline{X}_n^2\right) && \text{linéarité}\\
   &= \frac{1}{n}\sum_{k=1}^{n} \left(\sigma^2 + \mu^2\right) - \left(\frac{\sigma^2}{n} + \mu^2\right)\\
   &= \left(\sigma^2 + \mu^2\right) - \left(\frac{\sigma^2}{n} + \mu^2\right)\\
   &= \frac{n-1}{n}\sigma^2
   \end{aligned}$$

3. $$E(S'_n) = \frac{n}{n-1} E(S_n) = \sigma^2\sigma$$

Proposition 3 (Intervalle de confiance)**.**
Soit $(X_1, X_2, \ldots, X_n)$ *suivant toutes une loi identique , de variance* σ^2 *et d'espérance* m*. on note*

$$\overline{X_n} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

alors pour tout $t \in \mathbb{R}$

$$P\left(m \in \left[\overline{X_n} - t\frac{\sigma}{\sqrt{n}}; \overline{X_n} + t\frac{\sigma}{\sqrt{n}}\right]\right) \simeq 2\Phi(t) - 1$$

Démonstration :

1. On montre que
$$E(\overline{X_n}) = m \qquad V(\overline{X_n}) = \frac{\sigma^2}{n}$$
2. On admet (théorème central limite) que l'on peut assimiler $\overline{X_n}$ à une variable aléatoire d'espérance m et d'écartype $\frac{\sigma}{\sqrt{n}}$

$$\begin{aligned}
P\left(m \in \left[\overline{X_n} - \frac{\sigma}{\sqrt{n}}; \overline{X_n} + t\frac{\sigma}{\sqrt{n}}\right]\right) &= P\left(\overline{X_n} - t\frac{\sigma}{\sqrt{n}} \leqslant m \leqslant \overline{X_n} + t\frac{\sigma}{\sqrt{n}}\right) \\
&= P\left(m - t\frac{\sigma}{\sqrt{n}} \leqslant \overline{X_n} \leqslant m + t\frac{\sigma}{\sqrt{n}}\right) \\
&= P\left(-t \leqslant \frac{\overline{X_n} - m}{\frac{\sigma}{\sqrt{n}}} \leqslant t\right) \\
&= P\left(-t \leqslant \overline{X_n}^* \leqslant t\right) && \text{où } \overline{X_n}^* \text{ suit la loi normale centrée réduite} \\
&= \Phi(t) - \Phi(-t) && \text{fonction de répartition d'une variable aléatoire suivant une loi normale centrée réduite} \\
&= 2\Phi(t) - 1
\end{aligned}$$

∎

C Bibliographie

Références

[Ber+06] Jeffrey S. BERGER et al. « Aspirin for the Primary Prevention of Cardiovascular Events in Women and MenA Sex-Specific Meta-analysis of Randomized Controlled Trials ». In : *JAMA* 295.3 (jan. 2006), p. 306-313. ISSN : 0098-7484. DOI : 10.1001/jama.295.3.306. eprint : https://jamanetwork.com/journals/jama/articlepdf/202217/jrv50028.pdf. URL : https://doi.org/10.1001/jama.295.3.306.

[LAT+12] STEVEN C. LATTA et al. « Use of Data on Avian Demographics and Site Persistence during Overwintering to Assess Quality of Restored Riparian Habitat ». In : *Conservation Biology* 26.3 (2012), p. 482-492. DOI : https://doi.org/10.1111/j.1523-1739.2012.01828.x. eprint : https://conbio.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1523-1739.2012.01828.x. URL : https://conbio.onlinelibrary.wiley.com/doi/abs/10.1111/j.1523-1739.2012.01828.x.

[Bou24] J.L. Auchan de BOUFFON. « Le Dahut : recherche à hue et à dia sur un mythe alpin ». In : *Annales de Cryptozoologie appliquée* 42.? (Jan. 2024), p. 3, 12.

[BY] Gabriel BARTZ et D. B. YOUNG. *Mercury Concentrations in Resident Lake Fish Sampled from Katmai National Park and Preserve in 2021*. URL : https://irma.nps.gov/DataStore/Reference/Profile/2300703.

[Lar] Fabrice LARRIBE. *Tables statistiques*. URL : http://fabricelarribe.uqam.ca/Enseignement/#s3/.

[McD] John H. MCDONALD. *Handbook of Biological Statistics*. URL : http://www.biostathandbook.com/.