Notions de statistiques pour les TIPE

BCPST Spé

Lycée Champollion Grenoble

Table des matières

Vocabulaire **2**

I Décrire les données **3**
I.1 Indicateur de position : la moyenne . . . 3
I.2 Indicateurs de dispersion : Variance et écart-type . . . 3

II Manipuler et représenter les données **4**
II.1 Manipuler les données d'une feuille de calcul excel avec python . . . 4
II.1.a Pandas . . . 4
II.1.b Opencsv . . . 5
II.2 Représenter les données. . . . 5
II.3 Exemples de représentations. . . . 6
II.4 Histogramme : Visualiser la variabilité d'une population . . . 6
II.4.a Boites à moustaches : Visualiser la variabilité des populations. . . . 6
II.4.b Nuage de points : chercher une corrélation dans une série double. . . . 7
II.4.c Diagramme en barre : représenter des valeurs moyennes . . . 8

III Choisir un type d'incertitude ou de barres d'erreur **8**
III.1 Écart-type . . . 8
III.2 Erreur type (ou erreur standard) de la moyenne . . . 8
III.3 Intervalle de confiance . . . 9
III.3.a Grandes séries (plus de trente données) . . . 9
III.3.b Petites séries (moins de trente mesures) : loi de Student . . . 10
III.3.c Incertitude sur une fréquence ou une proportion . . . 10

IV Tester une hypothèse **10**
IV.1 Vocabulaire : Hypothèses, valeur p . . . 11
IV.2 Test de Student . . . 11
IV.3 Test d'indépendance : Test du $\chi 2$ . . . 12

Ressources **13**

Références **13**

Vocabulaire

Variable ou caractère Une **variable** est une **caractéristique mesurable** qui peut prendre différentes valeurs. Taille, concentration d'un élément chimique, allèle, temps, réussite à un test. Il faut distinguer 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. Il n'est pas possible 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'individus dans une population, nombre de cellules ...

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 le travail statistique est effectué.

- Si nous choisissons une partie de la population étudiée et que les mesures ne sont faites que sur ces individus, les données forment un **échantillon** que nous espérons 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 c'est alors une **série statistique**. Si ce travail de classement n'a pas été effectué, la série est dite **brute**.

Statistique descriptive/représentation graphique Pour donner une description des données nous calculons des indicateurs : moyenne, écart-type ... Il est aussi possible de représenter les données en histogramme, nuage de points ...

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

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

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

- Nous disposons d'une pièce et nous cherchons à estimer le paramètre p probabilité d'obtenir un pile.
- Nous supposons que la taille d'un individu suit une loi normale dont nous cherchons à estimer l'espérance et la variance en mesurant un échantillon de la population.

Souvent nous ne ferons pas d'hypothèse sur la loi suivie par le phénomène mais nous voulons à estimer les caractéristiques comme l'espérance et la variance.
Exemple :

- Nous cherchons à connaître la densité de myosotis nain (*Eritrichium nanum*) dans les Alpes en fonction de l'altitude. Nous ne pouvons pas recenser l'ensemble de ces plantes sur l'ensemble des Alpes, nous choisissons donc quelques parcelles où l'on effectue des comptages, pour en déduire des densités moyennes.
- Nous voulons connaître le taux de prévalence d'une mutation dans une population. Il n'est pas réaliste de tester tous les individus, nous choisissons un échantillon, nous comptons le nombre d'individus portant la mutation et nous calculons 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. Se pose alors la question de quantifier la qualité de notre estimation.

Tests statistiques Lorsque l'on effectue des expériences pour chercher à valider ou invalider une hypothèse, un **test statistique** permet de savoir si les données mesurées permettent ou non de valider l'hypothèse faite.
Exemples d'hypothèses que l'on peut chercher à tester :

Ma pièce est elle truquée ?

Tel allèle a-t-il une influence sur la survie des individus ?

I Décrire les données

I.1 Indicateur 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)**.**

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

Outils pour un calcul de moyenne

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

Attention : L'utilisation une **fonction** a pour syntaxe `fonction(L)`, une **méthode** `L.methode()`

I.2 Indicateurs de dispersion : 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 statistiques, nous pouvons remarquer que l'écart-type a la même unité que les données de la série.
L'écart-type est un indicateur de dispersion relatif, il mesure la dispersion autour de la moyenne, plus l'écart-type est grand plus les valeurs s'éloignent de la moyenne en positif et négatif. Une série dont l'écart-type est nul est constante.

Proposition 1 (König-Huygens)**.**

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

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

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

Attention : L'écart-type corrigé n'est pas au programme de première année, la formule de König-Huygens n'est pas valide mais cet écart-type corrigé est utile dans certaines formules, et c'est un estimateur non baisée de l'écart-type d'un échantillon.

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 non corrigé il faut préciser 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 : delta degres of freedom

II Manipuler et représenter les données

II.1 Manipuler les données d'une feuille de calcul excel avec python

Vous allez sûrement sauvegarder vos données dans une feuille de calcul excel ou équivalent [1], ou au format csv [2] qui est un format très simple pour représenter les tableaux de données et qui peut être manipulé par un tableur comme excel.

Vous pouvez traiter les données dans excel, mais nous vous conseillons d'utiliser python pour le traitement statistiques.

Avantages Python

- Vos données sont séparées de leur traitement. Si vous gérez bien votre programme, une augmentation du nombre de résultats ne nécessite qu'une nouvelle exécution pour être traitée.
- Les documentations sont plus précises, les formules utilisées par les fonctions sont souvent disponibles dans l'aide.
- Vous pouvez programmer d'autres traitements que ceux statistiques : obtenir une accélération avec une liste de positions, appliquer des formules à un ensemble de données, de façon plus lisible et donc plus facile à corriger.
- Vous montrez que vous savez mettre en pratique les cours d'informatique, en utilisant un langage régulièrement utilisé par les chercheurs.

II.1.a Pandas

Ouvrir un fichier csv ou excel

```
import pandas as pd
Table=pd.read_excel('Mesdonnes.xlsx')
Table=pd.read_csv('Mesdonnes.csv')
```

1. N'oubliez pas de sauvegarder vos données régulièrement, sur un cloud ET sur un support physique
2. *comma separated values*

Manipulation des tables

Les données se manipulent simplement, pour accéder à une colonne la syntaxe est : `Table['Nom de la colonne']`

Avantages Permet de manipuler les tables plus facilement, avec un nom explicite pour les colonnes, il existe plein d'outils d'extraction et de traitement de données.

Inconvénient Il faut apprendre des nouvelles syntaxes.

II.1.b Opencsv

Un outil plus basique pour manipuler des fichiers de données est `opencsv`

Ouvrir un fichier csv ou excel

```
import csv
Table=csv.reader(open('Mesdonnes.csv'),delimiter=',',quotechar='"')
Table=csv.reader(open('Mesdonnes.xlsx'))
```

Il faut parfois adapter les options, pour connaître celles à utiliser faites attention lors de la sauvegarde du fichier.

II.2 Représenter les données.

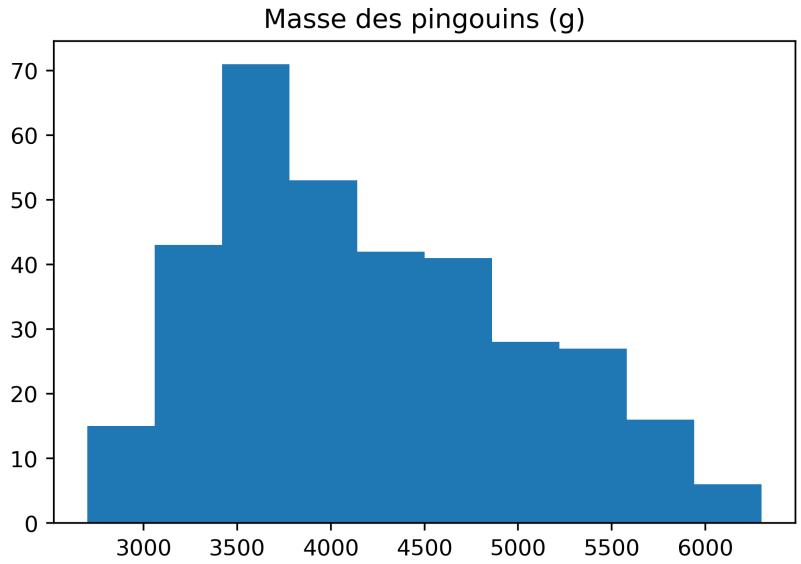
Pour représenter vos données si vous décidez d'utiliser python, vous pouvez utiliser la bibliothèques `matplotlib.pyplot`, voici une liste non exhaustive des commandes qui peuvent vous servir.

- `matplotlib.pyplot.scatter` Nuage de points la commande `numpy.polyfit` permet de calculer les coefficients de la droite de régression linéaire.
- pylmatplotlib.pyplot.boxplot Permet de représenter des données sous forme de "boites à moustaches", permet de visualiser la variabilité d'un échantillon ou d'une population.
- `matplotlib.pyplot.bar` Permet de tracer les diagrammes en barres. Il est possible de tracer des diagrammes avec plusieurs barres, des barres empilées.
- `matplotlib.pyplot.errorbar` Permet de choisir et d'afficher les barres d'erreurs dans les diagrammes précédents.
- `matplotlib.pyplot.hist` Histogramme de répartition, il est possible de représenter des effectifs ou des pourcentages.
- `matplotlib.pyplot.pie` Diagramme en camembert ou tarte.
- `matplotlib.pyplot.violinplot` Diagramme en "violon" pour représenter la distribution d'une population.

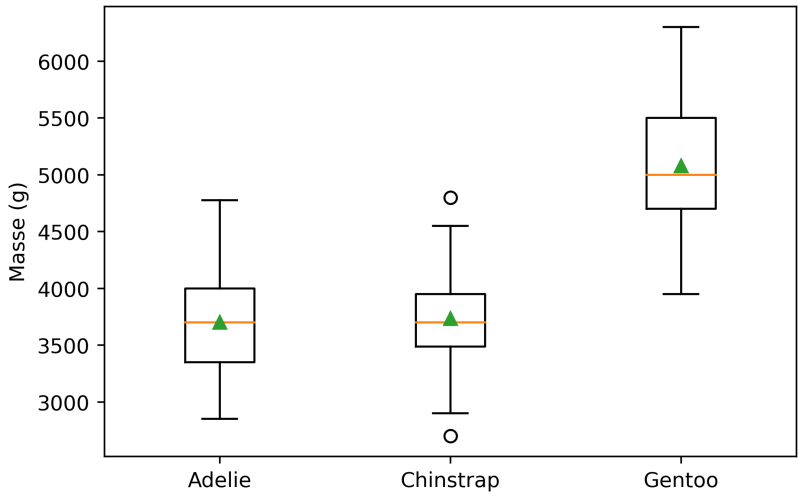
Des exemple d'utilisation de ces commandes sont disponibles dans le fichier `.py`. Les graphiques obtenus peuvent être personnalisés, grand choix de couleurs, de forme des points, légendes.... Il ne faut pas hésiter **à augmenter la qualité des images** obtenues lors de la préparation du rapport :
commande `plt.rcParams['figure.dpi'] = 400`.

II.3 Exemples de représentations.

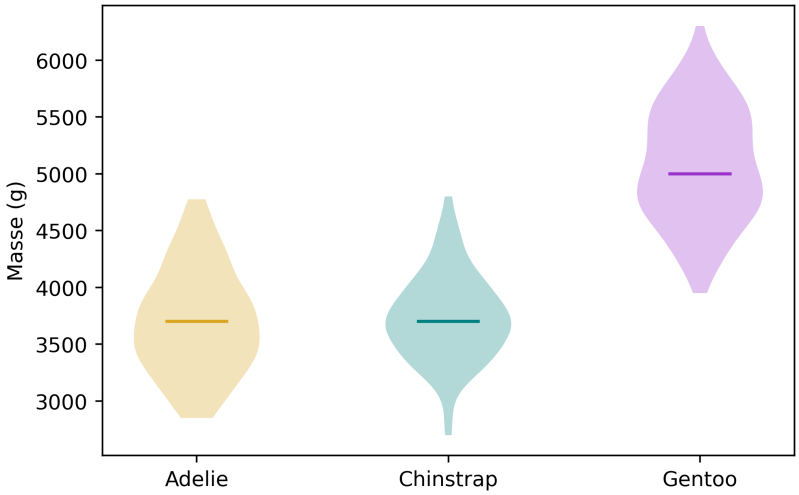
II.4 Histogramme : Visualiser la variabilité d'une population



II.4.a Boites à moustaches : Visualiser la variabilité des populations.

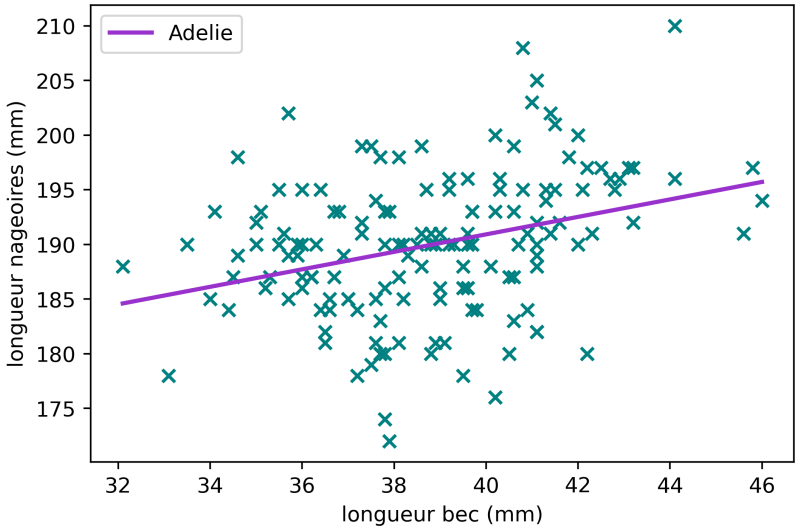


La boite centrale est délimitée par le premier quartile et le troisième quartile. La barre centrale représente la médiane, le triangle la moyenne. Les "moustaches" indiquent l'étendue de la série de données. Les cercles (population de Chinstrap) représentent les données qui sont trop éloignées des autres et en dehors de l'étendue (en anglais *outliers*, en français *données aberrantes*).
Attention : Il existe plusieurs options pour définir l'étendue et donc les *outliers*, vous devez indiquer sur vos documents quelle définition vous avez choisie.
On peut aussi utiliser un diagramme en "violon" pour représenter la distribution d'une population.



La barre centrale représente la médiane.

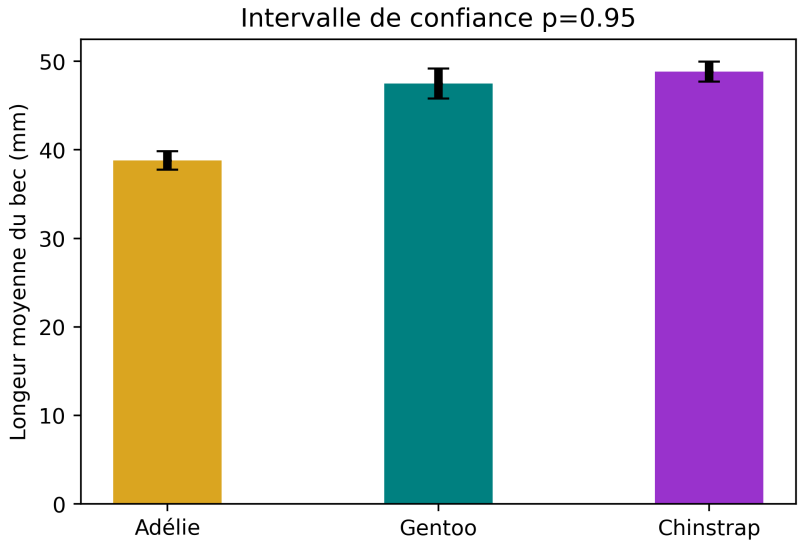
II.4.b Nuage de points : chercher une corrélation dans une série double.



On a fait apparaître les couples de mesures longueur du bec / longueur des nageoires mesurées chez une population de manchots Adélie.

Attention : En plus de la droite de corrélation linéaire, existe d'autres façons de rechercher une corrélation.

II.4.c Diagramme en barre : représenter des valeurs moyennes



On a représenté sur un même diagramme les longueurs moyennes des becs de trois populations de manchots de trois espèces différentes. On a fait aussi apparaître une mesure de l'incertitude.**Attention : Vos documents doivent toujours indiquer quelle mesure de l'incertitude vous vous faîtes figurer sur le graphique.**

III Choisir un type d'incertitude ou de barres d'erreur

Il est exigé que vous représentiez des barres d'erreurs sur vos diagrammes, mais il vous faudra faire un choix.

Outils pour afficher une barre d'erreur

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

III.1 Écart-type

Vous pouvez calculer la moyenne $\overline{x}$ et l'écart-type et donner comme incertitude l'intervalle

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

Vous démontrerez au second semestre que pour une variable aléatoire Z suivant une loi normale d'espérance m et d'écart-type σ

$$P(m-\sigma\leqslant Z\leqslant m+\sigma)\simeq 0,68 \qquad \text{et} \qquad P(m-2\sigma\leqslant Z\leqslant m+2\sigma)\simeq 0,95.$$

Cet intervalle donne une indication sur la dispersion d'une population.

III.2 Erreur type (ou erreur standard) de la moyenne

On définit l'erreur type par :

$$\frac{\sigma_x}{\sqrt{n}} \text{ ou } \frac{\sigma'_x}{\sqrt{n}}$$

avec l'écart-type ou l'écart-type corrigé.
L'intervalle d'incertitude s'écrit

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

On utilise cet intervalle lorsque l'on effectue plusieurs fois la même expérience de physique ou de chimie et que l'on veut donner une mesure de l'incertitude.

Outils pour calculer l'erreur type

Python/scipy.stats La fonction et la méthode `sem` renvoient l'erreur type, en utilisant par défaut l'écart-type corrigé (`ddof=1`). Pour utiliser l'écart-type non corrigé, il faut inclure l'option `ddof=0`.

Excel Aucune commande spécifique ne semble être disponible, vous pouvez utiliser `STDEV (plage d'échantillonnage) / SQRT (COUNT (plage d'échantillonnage))` en adaptant la fonction pour l'écart-type.

III.3 Intervalle de confiance

Si vous avez calculé la moyenne d'un caractère d'une population sur un échantillon, et obtenu une valeur $\overline{x}$, cette moyenne empirique n'est qu'une estimation de la vraie moyenne m.

Le but est d'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 ? ».

III.3.a Grandes séries (plus de trente données)

Les résultats de la fin d'année nous permettrons de montrer que pour $n \geqslant 30$, il existe t_α tel que

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

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

cas $t=1$ On retrouve l'erreur type de la moyenne qui donne intervalle de confiance au niveau de confiance 0,68.

cas $t=1.96$ **ou** $t\simeq 2$ Le niveau de confiance est de 95%.

Exemple 1 (Mercure)**.**
Dans [BY] les auteurs ont mesuré le taux de mercure dans les tissus musculaires de truites prélevées dans un lac d'Alaska (mesures en ng de mercure 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.
- L'écart-type est de 1976.
- L'écart à la moyenne est de $\dfrac{1976.82}{\sqrt{40}} \simeq 312$.
- Un intervalle de confiance à 95% est $[2506.025-1.96\times 312;\ 2506.025-1.96\times 312] \simeq [1894;3117]$.

Voici comment utiliser les commandes Python (ici pour un niveau de confiance de 0,68) :

```
import scipy.stats as stats
import pandas as pd
Mercure=pd.read_csv('poissons.csv')
X=list(Mercure['Total_Hg(nano g/g)'])
I=stats.norm.interval(confidence=0.68,loc=np.mean(X),scale=stats.sem(X))
>>>(1885.6069240325446, 3126.4430759674556)
```

Pour calculer une intervalle de confiance

- Vous pouvez procéder comme dans l'exemple et utiliser les fonctions permettant de calculer la moyenne, l'écart-type et appliquer la formule.
- `stats.norm.interval(confidence=0.95,loc=np.mean(x),scale=stats.sem(x))` renvoie les bornes de l'intervalle de confiance. Vous pouvez choisir le niveau de confiance que l'on veut, mais en principe on utilise 0.95 ou plus rarement 0.99.

Attention : Les paramètres de cette fonction ont changé avec les dernières versions du module `scipy.stats`, il ne faut pas hésiter à consulter l'aide.

III.3.b Petites séries (moins de trente mesures) : 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 il faut utiliser l'écart-type corrigé.

Cette fois la statistique ne suit plus une loi normale, une loi de Student à $n-1$ degrés de libertés où n est la taille de la série. Le paramètre t n'est donc pas le même que pour les grandes séries.

Exemple 2 (Poissons)**.**
Dans [McD] propose un extrait d'une étude où les auteurs ont mesuré le nombre de Naseux noirs (*Rhinichthys atratulus*) sur des sections de 75 mètres le long de cours d'eau du Maryland.

Ils ont obtenus [76,102,12,39,55,93,98,53,102].

```
import scipy.stats as stats
D=[76,102,12,39,55,93,98,53,102]
stats.t.interval(confidence=0.95, df= len(D)-1, loc=np.mean(D), scale=stats.sem(D))
>>>(45.33665361224985, 94.66334638775015)
```

Pour calculer une intervalle de confiance (Student)

- Vous pouvez utiliser les fonctions permettant de calculer la moyenne, l'écart-type et appliquer la formule en utilisant une table, consulter une explication détaillée dans [Lak] et [Lar].
- `stats.t.interval(confidence=0.95,df= len(x)-1,loc=np.mean(x),scale=stats.sem(x))` renvoie les bornes de l'intervalle de confiance. Il faut adapter le niveau de confiance que l'on veut, mais en principe on utilise 0.95 ou plus rarement 0.99.

III.3.c Incertitude sur une fréquence ou une proportion

Si vous cherchez à déterminer une fréquence vous pouvez utiliser 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 que

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

IV Tester une hypothèse

Le but de cette partie est de présenter deux types de tests.

Exemple 3 (Pièce truquée)**.**
Nous cherchons à savoir si une pièce donnée est équilibrée ou non. Nous effectuons 10 lancers et nous obtenons 8 piles et 2 faces.

L'hypothèse nulle est "la pièce est non truquée" l'hypothèse alternative "la pièce est truquée". Un calcul simple montre que **si nous lançons une pièce non truquée** la probabilité lors de 10 tirages d'obtenir 8 piles **ou plus** est d'environ 0.054. Il y a environ 6% de chance qu'une telle situation arrive avec une pièce non truquée. Nous interprétons ce résultat en disant que la probabilité de faire une erreur en rejetant l'hypothèse nulle est 6%.

IV.1 Vocabulaire : Hypothèses, valeur p

Définition 4 (Hypothèse nulle $\mathscr{H}_0$)**.**
*C'est l'hypothèse "l'effet observé n'est dû qu'au hasard". Sa négation est l'**hypothèse alternative** qui est souvent celle que l'on "aimerait" voir vérifiée.*

Définition 5 (p-value/valeur p)**.**
C'est la probabilité[3] *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**.*

Nous pouvons tester trois types de choses :

- La moyenne d'une série de valeurs est-elle significativement différente d'une valeur théorique ?
- Les moyennes de deux séries de données peuvent-elles être considérées différentes ?
- Deux séries de données sont-elles indépendantes ?

Les étapes lorsque l'on fait les calculs sont :

1. Calculer une certaine valeur construite à partir des données : 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.

Mais pour chacun de ces tests on peut trouver des fonctions python qui font l'ensemble des calculs.

IV.2 Test de Student

Nous voulons tester si deux séries ont des moyennes **significativement différentes**. Nous faisons l'hypothèse que les deux variables à distinguer suivent des lois normales[4] identiques, ou du moins des lois proches.

Exemple 4 (Populations de dahuts)**.**
Après de longs mois passés dans le massif de Belledonne un biologiste a mesuré la masse d'un échantillon de dahuts (*Dahus lateralis var belladonnanis*)[Bou24].

Dextrogyre	Lévrogyre	Dextrogyre	Lévrogyre
43.6	58.8	82.4	60.
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	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	

TABLE 1 – Mesure de masse (kg) sur les dahuts lévrogyres et dextrogyres

Nous voulons savoir si les sous-espèces ont des masses moyennes significativement différentes.

3. sous une modélisation statistique donnée
4. Il est possible tester cette hypothèse avec un test de Henry [Lak]

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

```
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)
```

Le résultat obtenu est environ 0.56, il ne faut 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.3 Test d'indépendance : Test du $\chi 2$

Attention : Si vous avez un échantillon de petite taille, il vaut mieux utiliser un test de Fisher exact avec la commande `scipy.stats.fisher_exact`.

Exemple 5 (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

TABLE 2 – Population d'oiseaux

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)
```

La p-value calculée est environ 1.6×10^{-26}, il est possible de rejeter l'hypothèse nulle, mais nous n'avons 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`

Ressources

- www.biostathandbook.com [McD] En anglais, de nombreux tests décrits avec leurs limites et des applications tirées de la biologie. Peu de résultats théoriques sont exposés, 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.Vous y trouverez des conseils pour tracer les barres d'erreur.

Références

[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.

[Lak] Lycée Lakanal. *Traitement statistique des données pour le TIPE*. URL : https://cahier-de-prepa.fr/bcpst2-lakanal/download?id=1024.

[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/.