Chapitre ALEA.17.

Statistiques Descriptives & Inférentielles

Résumé & Plan

Les statistiques sont présentes dans beaucoup de domaines en Sciences, notamment dans l'exploitation de grosses quantités de données. Il existe plusieurs types de statistiques : nous en étudierons deux, les statistiques descriptives d'une part dont le but est d'étudier des séries de données, d'autre part les statistiques inférentielles dont l'objectif est de savoir si des données semblent provenir ou non de réalisations d'une certaine variable aléatoire.

Il existe trois types de mensonges : les mensonges simples, les sacrés mensonges et les statistiques.

— Mark Twain

Cadre

Dans tout ce chapitre, $(\Omega, \mathcal{T}, \mathbf{P})$ désignera un espace probabilisé sur lequel seront définis les objets aléatoires le cas échéant.

1. LA STATISTIQUE : POSITION DU PROBLÈME

1.1. Où apparaît l'aléatoire ?

L'aléatoire est présent dans toute expérience scientifique. Les deux grandes explications en sont :

- d'une part l'aléatoire «intrinsèque» lié à la complexité des individus et des phénomènes étudiés et au manque d'information dans le domaine.
- D'autre part, de l'aléatoire peut intervenir «expérimentalement», par mesures entachées d'erreur,[1] ou encore lorsque les moyens pour relever sont limités (*i.e.* par exemple un recensement dans une grande population).

La première source d'aléatoire est donc le manque d'informations, ou l'ignorance, la seconde est matérielle et peut donc être contrôlée et donnée elle-même lieu à une étude plus poussée. Il n'y a pas de raison de penser qu'une source d'aléatoire peut complètement être supprimée. En Biologie (ainsi qu'en Médecine), étant donné l'extrême complexité des systèmes étudiés, l'aléatoire est très présent.

Exemple 1 — On étudie l'influence d'un champignon sur une population d'hêtres dans deux parcelles de forêt. Plusieurs sources d'aléatoires vont être prise en compte.

- Plusieurs individus de la même espèce, de la même origine et du même âge ont des biomasses différentes : aléatoire «intrinsèque» à l'expérience);
- les parcelles ne sont pas identiques (aléatoire «extrasèque»);
- les échantillons de population ne possèdent pas le même nombre d'individus, les calculs vont donc être source d'aléatoire : aléatoire «expérimental», lié au protocole;
- la grandeur mesurée ne peut l'être qu'avec une précision finie.

Cette analogie aléatoire pousse à interpréter les mesures effectuées comme des réalisations d'une variables aléatoire X «cachée». Les Statistiques visent alors à étudier cet aléatoire, ou le prédire (Statistiques inférentielles) ou encore à le représenter (Statistiques descriptives)[2].

1.2. Statistiques descriptive

La statistique descriptive est la branche des statistiques qui regroupe les nombreuses techniques utilisées pour décrire un ensemble relativement important de données. Cette description peut être réalisée au moyen de calculs de grandeurs (moyenne, variance, écart-type, mode, *etc.*) ou au moyen de descriptions visuelles (graphiques par exemple). Nous nous poserons également une seconde question : on peut représenter deux séries de données sous forme d'un nuage de points, on peut alors se demander s'il est possible d'évaluer la corrélation entre les deux séries, et plus précisément la faculté de l'une à dépendre de l'autre.

[1] Rappelons que nous avons vu dans le Chapter ANA.8 une borne sur l'erreur commise dans une fonction de mesures chacune entachées d'erreurs

[2] Mais il existe d'autres types de statistiques !

1.3. Statistique inférentielle

Inférence statistique, ensemble des méthodes permettant de formuler en termes probabilistes un jugement sur une population à partir des résultats observés sur un échantillon extrait au hasard de cette population.

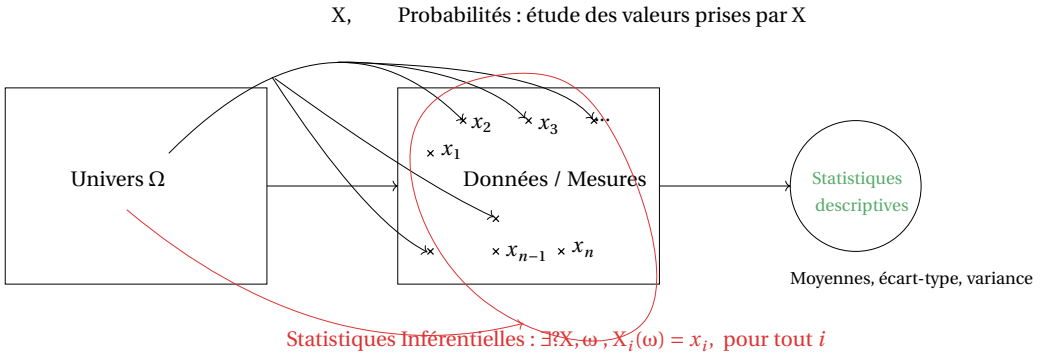
— Larousse

Commençons par expliquer ce que nous souhaitons faire en statistiques inférentielles.

1. Les probabilités ont pour objectif notamment l'étude de réalisations de ce que nous appelons *variables aléatoires*. Nous savons avec quelles probabilités ces variables prennent une certaine valeur (donnée de la loi) et nous étudions ensuite des paramètres : l'espérance, la variance, la covariance, intervalles de fluctuation *etc.*.
2. La *statistique inférentielle* souhaite effectuer la démarche inverse : *i.e.* savoir que si une série de données $x_1, \dots, x_n$ avec n un entier, peut être vue comme n réalisations d'une variable aléatoire suivant une certaine loi, autrement dit si le caractère mesuré semble suivre une certaine loi de probabilité. Bien évidemment, nous ne pourrons pas dire OUI ou NON avec certitude, nous pourrons juste dire OUI ou NON avec une bonne (si possible) probabilité, lorsque n est suffisamment grand.

L'objectif est donc de savoir s'il existe $(X_1, \dots, X_n)$ une famille de n variables aléatoires réelles **de même loi** telles que pour un certain $\omega \in \Omega$, $(x_1, \dots, x_n) = (X_1(\omega), \dots, X_n(\omega))$[3]. Nous les chercherons également indépendantes, ce qui semble être une hypothèse raisonnable (il n'y a pas de raisons que les relevés $x_1, \dots, x_n$ influent les uns sur les autres).

Pour des raisons techniques (application de la loi des grands nombres et du théorème central limite), nous supposerons que la loi commune des X_i possède une espérance μ et une variance σ^2.[4]



2. STATISTIQUES DESCRIPTIVES

Dans tout ce qui suit, nous verrons que toutes les notions définies ont une analogie dans le monde des probabilités (échantillon, espérance, variance, écart-type, covariance, *etc.*). Cette analogie sera explicitement précisée à chaque fois pour que les définitions deviennent naturelles. Pour plus de précisions, vous pourrez consulter la Remarque 2 ci-dessous.

Cadre

Dans cette section, les notations n, p, i, j désigneront des entiers même si cela n'est pas précisé.

[3] Ce point de vue est équivalent à l'existence de $\omega_1, \dots, \omega_n \in \Omega$ tels que $(x_1, \dots, x_n) = (X(\omega_1), \dots, X(\omega_n))$

[4] Ce qui, rappelons-le, n'est pas automatique en deuxième année pour les univers non finis

2.1. Univariées

2.1.1. Série statistique

Définition ALEA.17.1 | Population & Échantillon

Une *population* est un ensemble fini dont les éléments sont appelés des *individus*. Le nombre d'individus d'une population est appelé sa *taille*. Un sous-ensemble d'une population est appelé un *échantillon* de cette population.

Remarque 2.1 — Population ne signifie pas que l'on considère des personnes. Si vous réaliser plusieurs titrages pour votre T.I.P.E. on parlera alors de population de concentrations.

Très souvent, on ne mènera pas notre étude sur la population entière mais sur une sous-partie que l'on espère représentative.

Définition ALEA.17.2 | Caractère, Série statistique

1. Un *caractère* x de la population est une donnée *qualitative* ou *quantitative* attachée à chaque individu de la population. On notera x_i la valeur du caractère x pour un individu i.
2. Une *série statistique* de taille n, est une famille $(x_1, \dots, x_n)$ à n éléments.
3. La donnée des valeurs d'un caractère pour les individus d'un échantillon de taille n est
   - un caractère est dit *quantitatif* s'il prend des valeurs quantifiables, souvent des réels mais éventuellement des p-uplets ou des matrices.
   - Un caractère est dit *qualitatif* s'il correspond à une propriété qui ne se quantifie pas (c'est-à-dire couleur des cheveux, opinion politique, *etc.*)

Définition ALEA.17.3 | Modalités

On appelle *modalités* d'un caractère les valeurs possibles qu'il peut prendre.

Exemple 2 — *Série des poids* Le tableau ci-dessous regroupes des données de poids d'individus numérotés entre 1 et n.

Individu	1	2	…	n
Masse	62	80	…	74

Exemple 3 — *Diamètres de pièces* Le tableau ci-dessous regroupe les diamètres en cm de 48 pièces prélevées dans la production d'une machine.

19	26	23	20	22	24	20	24
22	20	21	19	21	22	19	20
21	21	22	21	23	22	21	24
25	23	22	19	20	26	24	25
23	26	25	25	21	22	25	24
23	22	24	24	25	23	25	22

Il s'agit donc d'un échantillon statistique de taille 48 dans la population des pièces fabriquées par la machine. Le caractère étudié est le diamètre de la pièce en centimètre.

Les modalités sont : 1.19, 1.20, 1.21, 1.22, 1.23, 1.24, 1.25 et 1.26.

2.1.2. Effectifs & Regroupements en classes

Une série statistique peut être donnée sous plusieurs formes.

- Soit on décide de fournir toutes les données de manière exhaustive,
- soit on décide de les regrouper en modalités et/ou classes, et on doit alors fournir

un tableau d'effectifs associés.

Données affichées de manière exhaustive. On peut simplement donner la valeur du caractère pour chaque individu. Voir les Exemples 2 et 3.

Données regroupées par modalité. Si la taille de l'échantillon est trop grande on préfèrera donner les nombres d'individus associés à chaque modalité. On appelle cela l'*effectif* associé à ladite modalité.

Exemple 4 — ***Âges d'enfants*** Dans cet exemple, une série d'âges sur des enfants entre 0 et 5 ans. On fournit alors les effectifs associés à chaque âge.

Nombre d'enfants	0	1	…	5
Effectif	20	31	…	3

Données regroupées en classes : lorsque le nombre de modalités est trop important. Parfois le nombre de modalités est trop grand, voire infini pour des modalités dites continues (c'est-à-dire à valeurs réelles). Il est alors pertinent de regrouper les modalités en classes disjointes, le plus souvent en des intervalles qui ne sont pas forcément de tailles égales. Inversement les modalités non regroupées en classe sont dites *ponctuelles*. Il peut également arriver que nos modalités comportent des classes et des modalités ponctuelles.

Exemple 5 — Par exemple, lorsque l'on étudie la démographie urbaine française on va regrouper les communes en classes selon leur nombre d'habitants :

- Hameaux de 1 à 99 habitants,
- Village de 100 à 1999 habitants,
- Ville de 200 à 99999 habitants,
- Agglomération à partir de 100000 habitants.

Effectifs & Effectifs cumulés croissants.

Définition ALEA.17.4

Soit une série statistique de taille n admettant un nombre fini de modalités ou à défaut de classes notées $a_1, \dots, a_p$.

1. Pour $j \in [\![1\,,\,p]\!]$, on définit l'*effectif* n_j associé à la valeur a_j comme étant le nombre d'individus i pour lesquels $x_i = a_j$.
2. Pour $j \in [\![1\,,\,p]\!]$, on définit la *fréquence de* f_j associée à la valeur a_j comme étant la proportion d'individus pour lesquels $x_i = a_j$, c'est-à-dire

$$f_j = \frac{n_j}{n}.$$

Proposition ALEA.17.1 | Formule des fréquences totales / effectifs totaux

Soit une série statistique de taille n admettant un nombre fini de modalités ou à défaut de classes notées $a_1, \dots, a_p$. Alors

$$\sum_{j=1}^{p} n_j = n \quad \text{et} \quad \sum_{j=1}^{p} f_j = 1.$$

Remarque 2.2 — Analogie avec les probabilités Reprenons les notations précédentes, *i.e.* une série statistique de taille n admettant un nombre fini de modalités ou à défaut de classes notées $a_1, \dots, a_p$. On note $n_1, \dots, n_p$ les effectifs correspondant. Alors si l'on considère une variable aléatoire réelle discrète X telle que[5] :

$$\mathrm{X}(\Omega) = \left\{a_1, \dots, a_p\right\}, \quad \mathbf{P}\left(\mathrm{X} = a_j\right) = \frac{n_j}{n}, \quad j \in [\![1\,,\,p]\!].$$

[5] on voit donc la série statistique comme la réalisation d'une variable réelle discrète, de loi la fréquence de ladite observation.

On vérifie sans peine que :

$$\sum_{j=1}^{p} \frac{n_j}{n} = 1, \quad \frac{n_j}{n} \geqslant 0$$

pour tout $j \in [\![1\,,\,p]\!]$, ce qui garantit l'existence de X. Alors la formule précédente est simplement dire que

$$\left\{X = a_j\right\}_{j\in[\![1\,,\,p]\!]}$$

est un système complet d'évènements. Toutes les notions que l'on va définir ci-après (espérance, variance, écart-type), correspondent en fait à :

$$\mathbf{E}(X)\,,\mathbf{Var}(X)\,,\sigma_X,\,\ldots$$

Dans le cas où les caractères étudiés sont des réels (ils peuvent donc être ordonnés), on va introduire les effectifs cumulés croissants et les fréquences cumulées croissantes.

Définition ALEA.17.5 | Effectifs cumulées croissants

On suppose ici les modalités $a_1, \ldots, a_p$ rangées dans l'ordre croissant ($a_1 < a_2 < \ldots < a_p$). Pour des intervalles $]a, b]$ et $]c, d]$ cela correspond à $a < b \leqslant c < d$. Soit $j \in [\![1\,,\,p]\!]$.

1. On définit *l'effectif cumulé croissant* associé à la modalité a_j comme le nombre d'observations $x_i \leqslant a_j$, c'est-à-dire

$$n_j^c = \#\left\{i \in [\![1\,,\,n]\!],\, x_i \leqslant a_j\right\}.$$

2. On définit la *fréquence cumulée (croissante)* associée à la modalité a_j comme la proportion d'observations $x_i \leqslant a_j$, c'est-à-dire

$$f_j^c = \frac{n_j^c}{n}.$$

Remarque 2.3 — Analogue probabiliste : la fonction de répartition.

Proposition ALEA.17.2 | Lien effectifs / effectifs cumulés

On suppose ici les modalités $a_1, \ldots, a_p$ rangées dans l'ordre croissant ($a_1 < a_2 < \ldots < a_p$). Pour des intervalles $]a, b]$ et $]c, d]$ cela correspond à $a < b \leqslant c < d$.

1. $\forall j \in [\![1\,,\,p]\!], \quad n_j^c = \sum_{k=1}^{j} n_k, \quad$ et $\quad f_j^c = \sum_{k=1}^{c} f_k,$
2. **(Les effectifs cumulés sont croissants)**

$$n_1^c \leqslant n_2^c \leqslant \ldots \leqslant n_p^c = n,$$

$$f_1^c \leqslant f_2^c \leqslant \ldots \leqslant f_p^c = 1.$$

Preuve

1. On écrit d'abord, avec $a_0 = -\infty$ pour convention, le découpage en tranches[a] suivant :

$$\left\{i \in [\![1\,,\,n]\!],\, x_i \leqslant a_j\right\} = \biguplus_{k=1}^{j} \left\{i \in [\![1\,,\,n]\!],\, a_{k-1} < x_i \leqslant a_k\right\}.$$

En passant au cardinal dans cette réunion disjointe, il vient : $n_j^c = \sum_{k=1}^{j} n_k$. Il suffit de diviser alors par n pour obtenir la version avec fréquences.

2. Il suffit de constater que pour $i, j \in [\![1\,,\,p]\!]$ tels que $1 \leqslant i < j \leqslant p$, nous avons l'inclusion $\left\{k \in [\![1\,,\,n]\!],\, x_k \leqslant a_i\right\} \subset \left\{k \in [\![1\,,\,n]\!],\, x_k \leqslant a_j\right\}$ puisque les classes sont ordonnées. Il suffit ensuite de passer au cardinal. On divise par n pour obtenir la version avec fréquences.

2.1.3. Représentation graphique

Les séries de données statistiques peuvent être représentés de plusieurs manières, en lieu et place de tableaux comme présentés *supra*.

Diagramme en bâtons & Histogrammes. Le diagramme en bâtons est adapté pour représenter des données ayant un nombre fini de modalités. Pour les données conti-

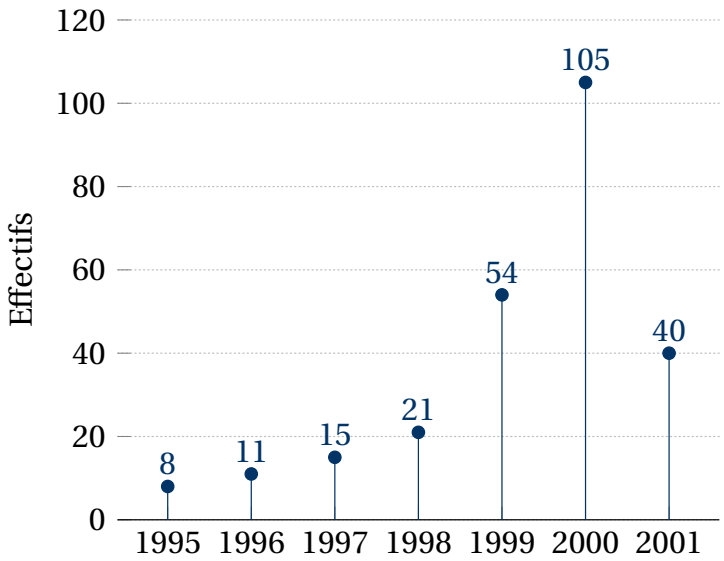
[a] Les tranches deviennent des égalités si aucun regroupement en classes n'a été fait.

nues regroupées en classes, on tracera plutôt ce que l'on appelle un *histogramme*.

Pour établir un diagramme en bâtons on va tracer pour chaque modalité un bâton (un rectangle long et fin) centré en a_j et de hauteur f_j ou n_j. Ce type de graphique est adapté aux données ponctuelles et aux données qualitatives.

Exemple 6 — *Diagramme en bâtons.* Par exemple, ici nous avons représenté le tableau d'effectifs d'âges suivant :

Âge	1994	1995	1996	1997	1998	1999	2000
Effectif	8	11	15	21	54	105	40



Lorsque le nombre de modalités est trop grand, et même infini, nous avons que l'on regroupait généralement les données en classes.

Pour chaque classe on va alors tracer un rectangle dont la largeur vaut l'amplitude de l'intervalle et dont *l'aire* est proportionnelle à la fréquence ou à l'effectif de la classe.

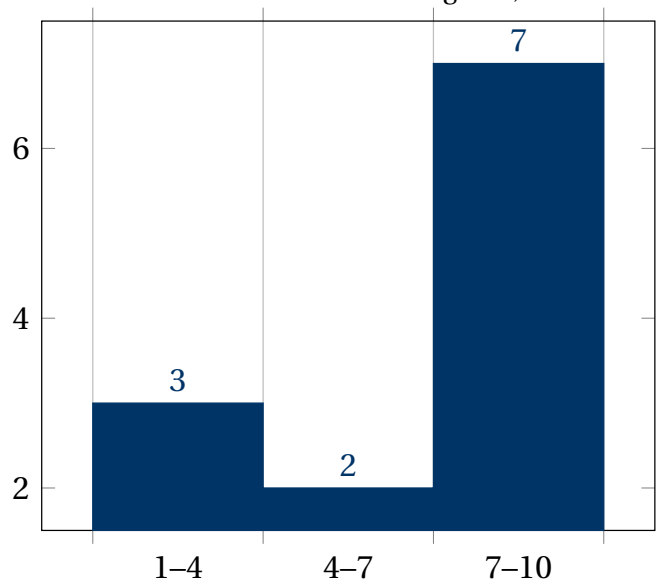
Attention

C'est l'aire du rectangle qui est importante, pas sa hauteur.

Exemple 7 — *Histogramme.* Par exemple, ici nous avons représenté le tableau d'effectifs d'âges suivant :

Données en classes	$[1,4[$	$[4,7[$	$[7,10[$
Effectifs	3	2	7

Sur cet exemple les classes ont toutes la même longueur, mais ceci n'est aucunement obligatoire.

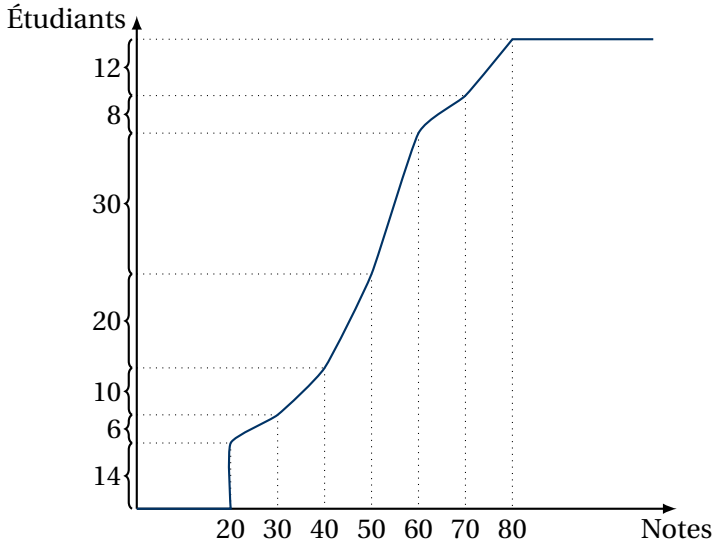


POLYGONE DES FRÉQUENCES CUMULÉES CROISSANTES. La procédure est différente pour des modalités ponctuelles ou regroupées par classes.

- Pour des modalités ponctuelles, on place les points $A_i(a_i, f_i^c)$. On va alors relier les points $A_i(a_i, f_i^c)$ à $B_i(a_{i+1}, f_i^c)$ puis B_i à $A_{i+1}(a_{i+1}, f_{i+1}^c)$. Schématiquement on trace un trait horizontal puis un trait vertical. *La courbe obtenue ressemble à celle d'une fonction de répartition de loi discrète.*
- Pour des modalités regroupées en classe $]\alpha_i, \alpha_{i+1}]$ (*resp.* $[\alpha_i, \alpha_{i+1}[$) on place les points A_i de coordonnées (α_{i+1}, f_i^c) (*resp.* (α_i, f_i^c)), à la droite (*resp.* gauche) de l'intervalle donc. On relie ensuite simplement les points A_i par une ligne brisée. *La courbe obtenue ressemble alors à celle d'une fonction de répartition de loi à densité.*

Remarque 2.4 — Cette représentation est pertinente si les individus sont répartis uniformément au sein de la classe. Si vous avez des raisons de penser que ce n'est pas le cas alors il faut séparer votre classe en plusieurs classes.

Exemple 8 — ***Notes de DS*** Par exemple, on peut essayer de représenter le polygone des fréquences cumulées.



Par exemple, ici nous avons représenté le tableau d'effectifs d'âges suivant :

Notes en classe	[30,40[	[40,50[	[50,60[	[60,70[	[70,80[	[80,90[	[90,100[
Effectifs	14	6	10	20	30	8	12

Lire graphiquement le nombre approximatif de notes inférieures à 30, et 56.

2.1.4. Caractéristiques de position

Les caractéristiques de positions ont leur analogue en probabilités : l'espérance et la médiane d'une variable aléatoire[6]. Ce sont des grandeurs dont la vocation est de mesurer la position des données qui constituent la série statistique. En revanche, le mode, défini ci-après, est très peu souvent défini en probabilités : mais ce serait alors, pour une variable aléatoire discrète X, un élément $x \in X(\Omega)$ qui maximise $\mathbf{P}(X = x)$.[7]

Mode & Classe modale.

Définition ALEA.17.6 | Mode & Classe modale

1. On appelle *mode* d'une série statistique x toute modalité de x dont l'effectif est maximal parmi les effectifs de toutes les modalités.
2. Lorsque les modes correspondent à des classes, on appelle alors *classe modale* la classe dont l'effectif est maximal.

Attention

Si vos classes sont de tailles différentes alors la classe modale n'est pas forcément la classe qui a le « plus haut » rectangle dans l'histogramme.

Remarque 2.5 — Il est possible qu'une série statistique admette plusieurs modes ou classes modales. D'un point de vue informatique, rechercher un mode revient à rechercher un maximum dans une liste.

Moyenne.

[6] Mais l'étude de l'existence d'une médiane, pour une variable aléatoire, n'est pas au programme.
[7] Rien ne garantit son existence en revanche.

Définition ALEA.17.7 | Moyenne pour des données ponctuelles

Soit $x = (x_1, \dots, x_n)$ une série statistique quantitative. La *moyenne* de la série, notée $\overline{x}$, est définie par

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Remarque 2.6 — Analogie L'espérance en probabilités.

Proposition ALEA.17.3

Soit $x = (x_1, \dots, x_n)$ une série statistique de modalités $(a_1, \dots, a_p)$ ponctuelles, alors :

$$\overline{x} = \frac{1}{n} \sum_{j=1}^{p} n_j a_j = \sum_{j=1}^{p} f_j a_j.$$

On dit alors que $\overline{x}$ est la *moyenne pondérée* des a_j par les fréquences f_j.

Preuve Constater simplement que dans $\sum_{i=1}^{n} x_i$, chaque x_i apparaît un nombre n_j de fois où n_j désigne l'effectif de x_i.

Quand on travaille avec des données regroupées par classes cette définition n'est pas utilisable. Dans cette situation on va alors considérer que les valeurs sont uniformément réparties dans les intervalles et prendre pour moyenne de la série la moyenne des milieux des intervalles pondérés par les effectifs.

Définition ALEA.17.8 | Moyenne pour des données regroupées en classes

Soit $x = (x_1, \dots, x_n)$ une série statistique quantitative. Supposons que les modalités $(a_j)_{j \in [\![1,p]\!]}$ de cette série correspondent à des intervalles $[b_j, c_j[$. On définit alors la *moyenne* par :

$$\overline{x} = \frac{1}{n} \sum_{j=1}^{p} n_j \left(\frac{b_j + c_j}{2} \right) = \sum_{j=1}^{p} f_j \left(\frac{b_j + c_j}{2} \right).$$

Proposition ALEA.17.4 | Propriétés de la moyenne

Soient x et y deux séries statistiques quantitatives dont les modalités sont à valeurs dans le même ensemble, et d'effectifs respectifs n et m.

1. **(Moyenne d'une série affine)** Soit $(a,b) \in \mathbf{R}^2$ et soit u la série statistique définie par

$$\forall i \in [\![1\,,\,n]\!], \qquad u_i = ax_i + b, \quad \text{alors :} \quad \overline{u} = a\overline{x} + b.$$

2. **(Moyenne d'un mélange)** Soit z la série statistique obtenue en «concaténant» les séries x et y, *i.e.*

$$z = (x_1, \dots, x_n, y_1, \dots, y_m), \quad \text{alors :} \quad \overline{z} = \frac{n\overline{x} + m\overline{y}}{n+m}.$$

Preuve

1. 

2. 

MÉDIANE. On suppose ici que nos modalités sont des réels. La moyenne est fortement influencée par les valeurs extrêmes, donc dans ce cas la donnée de la moyenne est assez peu instructive, et on privilégie une autre quantité appelée la médiane.

Remarque 2.7 — Cela signifie qu'il y a au moins autant d'éléments plus grands que d'éléments plus petits.

Définition ALEA.17.9 | Médiane pour des données ponctuelles

On appelle *médiane* d'une série statistique x de taille n tout réel m, «partageant la série d'observations en deux», *i.e.* tel que :

$$\#\left(\{i \in [\![1\,,\,n]\!],\, x_i \leqslant m\}\right) \geqslant \frac{n}{2} \quad \text{et} \quad \#\left(\{i \in [\![1\,,\,n]\!],\, x_i \geqslant m\}\right) \geqslant \frac{n}{2}.$$

En pratique on prend souvent comme médiane la valeur d'une modalité. Dans ce cas, un individu dont le caractère correspond à la médiane est dit être un *individu médian*.

Remarque 2.8 — Les symboles $\leqslant, \geqslant$ sont justifiés par le fait que $\frac{n}{2}$ n'est pas toujours un entier. Le signe d'égalité n'aurait dans ce cas aucun sens, puisqu'un cardinal est toujours entier.

Proposition ALEA.17.5 | Cas de modalités ordonnées

Soit x une série statistique de taille n dont les modalités sont données dans l'ordre croissant $a_1 < a_2 < \ldots < a_n$.

1. Si n est impair, alors : $a_{\frac{n+1}{2}}$ est une médiane.
2. Si n est pair, alors : tout nombre de l'intervalle $\left[a_{\frac{n}{2}}, a_{\frac{n}{2}+1}\right]$ est une médiane. En particulier,

$$\frac{a_{\frac{n}{2}} + a_{\frac{n}{2}+1}}{2} \quad \text{est } \underline{\textbf{une}} \text{ médiane.}[8]$$

Attention

Il y a en général plusieurs médianes.

[8] C'est souvent celle-ci qui, par abus de langage, nous appelons parfois *la* médiane.

Remarque 2.9 —

- La médiane a, par rapport à la moyenne, l'avantage d'être peu influencée par les valeurs extrêmes. Elle est alors plus représentative que la moyenne lorsque la série comporte des valeurs très grandes ou très petites.
  Par exemple, en France en 2014 le salaire moyen mensuel était de 1934 euros pour les femmes et 2389 euros pour les hommes tandis que le salaire médian mensuel était de 1619 euros pour les femmes et 1882 euros pour les hommes. La différence s'explique par le fait que les très hauts salaires, même s'ils sont peu nombreux, tirent la moyenne vers la haut.
- Sur le polygone des fréquences cumulées croissantes on lit une médiane assez simplement. Il suffit de chercher l'abscisse du point de la courbe d'ordonnée $\frac{1}{2}$.

Définition ALEA.17.10 | Médiane pour des données regroupées en classes

Soit x une série statistique dont les modalités sont regroupées en classes. On définit la *médiane* de x comme l'abscisse du point de la courbe des fréquences cumulées croissantes d'ordonnée $\frac{1}{2}$.

2.1.5. Caractéristiques de dispersion

L'idée des caractéristiques de dispersion est de donner une idée de la répartition de la série autour de sa moyenne ou de sa médiane. Les valeurs sont elles relativement proches de la moyenne ou existe-t-il des valeurs très grandes et très petites ? L'analogue probabilité est donc évident : l'espérance, la variance, l'écart-type ...

Valeurs extrêmes et étendue.

Définition ALEA.17.11

Soit x une série statistique de taille n à valeurs réelles.

1. Si les modalités sont en nombre fini, on appelle *valeurs extrêmes* de la série x les nombres :
$$\min_{i\in[\![1,n]\!]} x_i, \quad \text{et} \quad \max_{i\in[\![1,n]\!]} x_i.$$
Il s'agit donc des modalités maximales et minimales.
2. Si les modalités sont regroupées en classes, on appelle *valeurs extrêmes* de la série la borne supérieure de la classe maximale et la borne inférieure de la classe minimale.

On appelle *étendue* de la série statistique la différence entre la valeur maximale et la valeur minimale.

Remarque 2.10 — L'étendue est facile à déterminer mais ne délivre que très peu d'informations car elle est très fortement affectée par les valeurs extrêmes. Par exemple en France le revenu annuel se situe entre 0 euros et environ 7 millions, ce qui ne nous donne pas vraiment une idée de la répartition des salaires dans la population.

Quantiles. Plutôt que d'introduire des quantités qui découpent en deux une série statistique, on peut également partager en trois quatre *etc.*, de manière générale si on souhaite partager en parties de taille $n\cdot\alpha$ avec $\alpha\in[0,1]$, on parle alors de quantile d'ordre α.

Définition ALEA.17.12 | Quartile - $\alpha=\frac{1}{4}$

Soit x une série statistique de taille n à valeurs réelles. On appelle premier quartile de la série x tout réel Q_1 tel que
$$\#\left(\{i\in[\![1\,,\,n]\!],\, x_i\leqslant Q_1\}\right)\geqslant\frac{n}{4}\qquad\text{et}\qquad\#\left(\{i\in[\![1\,,\,n]\!],\, x_i\geqslant Q_1\}\right)\geqslant\frac{3n}{4}.$$
De-même on appelle troisième quartile de la série x tout réel Q_3 tel que
$$\#\left(\{i\in[\![1\,,\,n]\!],\, x_i\leqslant Q_3\}\right)\geqslant\frac{3n}{4}\qquad\text{et}\qquad\#\left(\{i\in[\![1\,,\,n]\!],\, x_i\geqslant Q_3\}\right)\geqslant\frac{n}{4}.$$

Remarque 2.11 — Il faut comprendre ces définitions de la même manière que pour la médiane : la première signifie qu'il y a au moins 1/4 des observations qui sont inférieures ou égales à Q_1, de-même pour les autres.

Remarque 2.12 — La médiane est donc aussi parfois appelée *deuxième quartile* ou *quantile d'ordre 2*.

On peut définir la notion de décile de manière similaire.

Définition ALEA.17.13 | Décile - $\alpha = \frac{1}{10}$

] Soit x une série statistique de taille n à valeurs réelles. Pour $j \in [\![1\,,9]\!]$, on appelle *j-ème décile* de la série x tout réel d_j tel que :

$$\#\left(\left\{i \in [\![1\,,n]\!],\, x_i \leqslant d_j\right\}\right) \geqslant \frac{jn}{10} \quad \text{et} \quad \#\left(\left\{i \in [\![1\,,n]\!],\, x_i \geqslant d_j\right\}\right) \geqslant \frac{(10-j)n}{10}.$$

Ces deux notions se généralisent avec la notion de quantiles.

Définition ALEA.17.14

Soit x une série statistique de taille n à valeurs réelles. Soit $t \in [0,1]$, on appelle *α-quantile* de la série x tout réel q_α tel que

$$\#\left(\left\{i \in [\![1\,,n]\!]\,,\, x_i \leqslant q_\alpha\right\}\right) \geqslant nt \quad \text{et} \quad \#\left(\left\{i \in [\![1\,,n]\!],\, x_i \geqslant q_\alpha\right\}\right) \geqslant (1-\alpha)n.$$

Comme pour les quartiles et les déciles il n'y pas unicité du α-quantile.

Remarque 2.13 —

- Le premier quartile est alors un $\frac{1}{4}$-quantile, *etc.*
- Si la série statistique est donnée via un regroupements en classes alors on déterminera un quantile d'ordre α en prenant l'abscisse d'un point du polygone des fréquences cumulées croissantes d'ordonnée α.

Définition ALEA.17.15 | Écarts

Soit x une série statistique de taille n à valeurs réelles. On appelle

1. *écart interquartile* la différence $Q_3 - Q_1$ noté parfois «IQR».
2. *intervalle interquartile* l'intervalle $[Q_1, Q_3]$.

De-même on appelle

1. *écart interdécile* la différence $d_9 - d_1$.
2. *intervalle interdécile* l'intervalle $[d_1, d_9]$

La moitié au moins de la population se trouve dans l'intervalle interquartile.

Visualisation graphique des quantiles : le diagramme de Tukey (ou « boîte à moustache »).

On peut représenter de manière graphique l'étendue, les quartiles et la médiane en dessinant un diagramme dit *diagramme de Tukey* conçu de la manière suivante :

- au centre une boite allant du premier au troisième quartile, séparée en deux par la médiane ;
- de chaque côté une moustache allant du minimum au premier quartile pour l'une, et du troisième quartile au maximum pour l'autre.

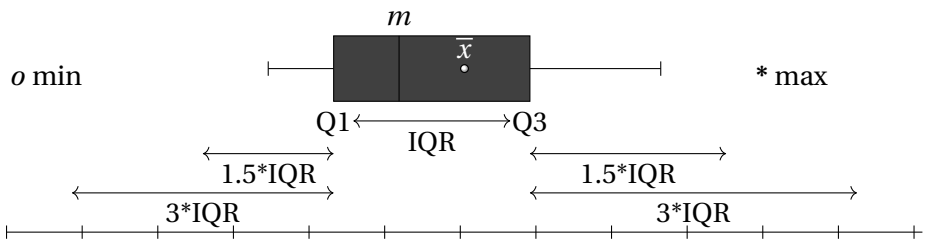


Fig. ALEA.17.1. : Représentation en boîte à moustache d'une série statistique

Variance et écart-type.

Définition ALEA.17.16 | Variance & Écart-Type

1. La *variance* d'une série statistique quantitative à valeurs réelles $x = (x_1, x_2, \dots, x_n)$ de nombre de modalités finies $a_1, \dots, a_p$, est le nombre $\mathbf{V}_x$ défini par :

$$\mathbf{V}_x = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{j=1}^{p} n_j(a_j - \overline{x})^2 = \sum_{j=1}^{p} f_j(a_j - \overline{x})^2.$$

2. L'*écart-type* d'une telle série, noté σ_x, est défini par :

$$\sigma_x = \sqrt{\mathbf{V}_x}.$$

Dans le cas d'une série regroupée en classes on prendra pour valeurs a_j les centres des classes.

$\mathbf{V}_x$ est la moyenne des carrés des écarts à la moyenne donc est toujours positive, d'où la bonne définition de l'écart-type.

Proposition ALEA.17.6 | Variance nulle

Soit une série statistique quantitative à valeurs réelles $x = (x_1, x_2, \dots, x_n)$ de nombre de modalités finies $a_1, \dots, a_p$, alors :

$$\mathbf{V}_x = 0 \iff \text{toutes les données de la série sont égales.}$$

Preuve On raisonne par exemple avec l'expression en fréquences de la variance.

$$\begin{aligned}
&\sum_{j=1}^{p} f_j(a_j - \overline{x})^2 = 0 \\
&\iff \forall i \in [\![1\,,\,p]\!],\quad f_j(a_j - \overline{x})^2 = 0, && \text{somme de termes positifs} \\
&\iff \forall i \in [\![1\,,\,p]\!],\quad (a_j - \overline{x})^2 = 0, && f_j \neq 0 \\
&\iff \forall i \in [\![1\,,\,p]\!],\quad a_j = \overline{x}.
\end{aligned}$$

C'est ce qu'on voulait.

Proposition ALEA.17.7 | König-Huygens

Soit x une série statistique quantitative à valeurs réelles. Alors :

$$\mathbf{V}_x = \overline{x^2} - \overline{x}^2.$$

Preuve C'est un calcul direct, comme pour les variables aléatoires.

$$\begin{aligned}
\mathbf{V}_x &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i^2 + \overline{x}^2 - 2x_i\overline{x}\right) = \frac{1}{n}\sum_{i=1}^{n}x_i^2 + \frac{1}{n}\sum_{i=1}^{n}\overline{x}^2 - 2\frac{\overline{x}}{n}\sum_{i=1}^{n}x_i \\
&= \overline{x^2} + \overline{x}^2 - 2\overline{x}^2 = \overline{x^2} - \overline{x}^2.
\end{aligned}$$

Remarque 2.14 — Interprétation

- Plus la variance est grande, plus la série s'éloigne de sa moyenne, et plus la série est donc «étalée». Inversement, plus la variance est proche de zéro et plus la série est concentrée autour de sa moyenne.
- La variance ne donne pas d'informations sur une éventuelle asymétrie de la série.

Remarque 2.15 — Homogénéité L'intérêt de l'écart-type par rapport à la variance est que l'écart-type s'exprime dans les mêmes unités que les modalités de la série. On pourra alors faire des calculs faisant intervenir modalités, moyenne et écart-type (par exemple dans des situations d'estimation de paramètres ou de test statistique d'hypothèses).

Proposition ALEA.17.8 | Propriétés de la variance

Soit x une série statistique quantitative réelle, $(a, b) \in \mathbf{R}^2$ et y la série statistique $y = ax + b$. On a alors

$$\mathbf{V}_y = a^2\mathbf{V}_x \qquad \text{et donc} \qquad \sigma_y = |a|\,\sigma_x.$$

Preuve On a

$$\begin{aligned}\mathbf{V}_y &= \frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(ax_i+b-(a\overline{x}+b)\right)^2\\ &= \frac{1}{n}\sum_{i=1}^{n}a^2(x_i-\overline{x})^2\\ &= a^2\mathbf{V}_x.\end{aligned}$$

Pour la deuxième, prendre simplement la racine dans l'expression.

2.1.6. En Python : calcul des grandeurs univariées

On suppose dans la suite que toutes les données d'une série à nombre de modalités fini sont contenues dans une liste L donnée en paramètre.

Espérance

```
def moyenne(L):
    """
    Renvoie l'espérance
    """
    S = 0
    for x in L:
        S += x
    return S/len(L)
```

Variance

```
from moyenne import *
def variance(L):
    """
    Renvoie la variance, version KH
    """
    esp = moyenne(L)
    V = 0
    for x in L:
        V += x**2
    return V/len(L) - esp**2
```

On peut également créer une fonction qui retourne une médiane, après avoir triée la liste. Celle-ci peut bien entendu être adaptées à tous les quantiles. On peut également, après recherche du minimum et du maximum, renvoyer l'étendue de la série.

Médiane

```
def mediane(L):
    """
    Cherche la médiane d'une liste, après tri rapide des
 ↪  observations
    """
    L = tri_rapide_rec(L)
    n = len(L)
    if n % 2 == 1:
        # Nombre impair d'observations, on prend le milieu
        return L[n//2]
    else:
        # Nombre pair d'observations, on prend la moyenne des deux
         ↪  termes du milieu
        return (L[n//2-1] + L[n//2])/2
```

On peut ensuite tester si la quantité retournée est bien une médiane, en contrôlant la définition.

```
def mediane_verif(L, m):
    """
    Renvoie True si m est bien une médiane de L
    """
    nb_inf = 0
```

```
    nb_sup = 0
    for x in L:
        if x >= m:
            nb_sup += 1
        if x <= m:
            nb_inf += 1
    return nb_sup >= len(L)/2 and nb_inf >= len(L)/2
```

Étendue d'une série

```
def etendue(L):
    """
    Cherche l'étendue de L, sans la trier
    """
    min = L[0]
    max = L[0]
    for i in range(1, len(L)):
        if L[i] < min:
            min = L[i]
        elif L[i] > max:
            max = L[i]
    return max - min
```

Test sur une série de données

Par exemple, si l'on considère des notes de DS.

```
>>> Notes_DS = [5.2, 7.4, 10.1, 12.8, 7.4, 10.0, 14.5, 6.5, 4.4,
 ↪  3.3, 11.8, 11.6, 8.0, 8.5, 10.2, 11.9, 8.3, 9.6, 10.0, 9.6,
 ↪  9.4, 10.2, 10.3, 15.1, 15.6, 13.2, 9.9, 9.6, 5.9, 14.5, 6.4,
 ↪  10.5, 6.7, 6.3, 10.9, 10.1, 7.5]
>>> moy = moyenne(Notes_DS)
>>> moy
9.545945945945943
>>> var = variance(Notes_DS)
>>> var
8.503564645726854
>>> med = mediane(Notes_DS)
>>> med
9.9
>>> mediane_verif(Notes_DS, med)
True
```

2.2. Bivariées

On a vu dans les sous-sections précédentes diverses manières d'extraire de l'information d'un échantillon statistique. Lorsque l'on ne dispose plus d'un seul mais de plusieurs échantillons statistiques, on peut, au delà de la simple étude des échantillons, étudier les éventuels liens entre eux, c'est l'objet de cette dernière sous-section. Pour des raisons de simplicité on se limitera à deux échantillons. Cette sous-section est essentiellement une reformulation dans le cadre déterministe des notions aléatoires déjà vues dans le Chapter ALEA.15 (covariance, coefficient de corrélation, *etc.*).

2.2.1. Série statistique

On va s'intéresser ici à deux caractères quantitatifs d'une même population. On notera n la taille de l'échantillon étudié et (x, y) les deux caractères étudiés. L'observation des valeurs des caractères se traduit par un échantillon d'éléments de $\mathbf{R}^2$, que l'on note :

$$\big((x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\big).$$

Définition ALEA.17.17 | Série statistique bivariée

1. Un *caractère* (x,y) de la population est une donnée *qualitative* ou *quantitative* attachée à chaque individu de la population. On notera (x_i,y_i) la valeur du caractère (x,y) pour un individu i.
2. Soient x et y deux séries statistiques de taille n. Alors on appelle *série statistique bivariée* une famille de $\mathbf{R}^2$ du type
$$\big((x_1,y_1),(x_2,y_2),\dots,(x_n,y_n)\big).$$
3. On appelle nuage point associé à l'échantillon (x,y) le tracé de tous les points de coordonnées $\mathrm{M}(x_k,y_k)$ pour $k\in[\![1\,,\,n]\!]$.

Définition ALEA.17.18 | Modalités

On appelle *modalités* d'un caractère bivarié les valeurs possibles qu'il peut prendre.

2.2.2. Effectifs

Notons $(a_1,\dots,a_p)$ les modalités du caractère x (éventuellement des classes) et $(b_1,\dots,b_q)$ les modalités du caractères y. Le plus souvent on regroupe les individus par modalités, on obtient alors le tableau d'effectifs suivant :

$x\backslash y$	b_1	…	b_j	…	b_q	**Totaux**
a_1	$n_{1,1}$	…	$n_{1,j}$	…	$n_{1,q}$	$n_{1,\bullet}$
⋮	⋮		⋮		⋮	⋮
a_i	$n_{i,1}$	…	$n_{i,j}$	…	$n_{i,q}$	$n_{i,\bullet}$
⋮	⋮		⋮		⋮	⋮
a_p	$n_{p,1}$	…	$n_{p,j}$	…	$n_{p,q}$	$n_{p,\bullet}$
Totaux	$n_{\bullet,1}$	…	$n_{\bullet,j}$	…	$n_{\bullet,q}$	n

$n_{i,j}$ est le cardinal de l'ensemble des individus présentant à la fois les modalités a_i et b_j. Pour $(i,j)\in[\![1\,,\,p]\!]\times[\![1\,,\,q]\!]$, on pose
$$n_{i,\bullet}=\sum_{j=1}^{q}n_{i,j}\qquad\text{et}\qquad n_{\bullet,j}=\sum_{i=1}^{p}n_{i,j}.$$

On a alors
$$n=\sum_{i=1}^{p}\sum_{j=1}^{q}n_{i,j}=\sum_{i=1}^{p}n_{i,\bullet}=\sum_{j=1}^{q}n_{\bullet,j}.$$

Définition ALEA.17.19 | Fréquence marginale & Effectif marginale

Pour $(i,j)\in[\![1\,,\,p]\!]\times[\![1\,,\,q]\!]$, on définit

- l'*effectif conjoint* $\big(resp.\ fréquence\ conjointe\big)$ des modalités (a_i,b_j) la suite double : $n_{i,j}=n_{i,j}$ $\left(resp.\ f_{i,j}=\frac{n_{i,j}}{n}\right)$.
- L'*effectif marginal de la modalité* a_i $\big(resp.\ fréquence\ marginale\ de\ la\ modalité\ a_i\big)$ par : $n_{i,\bullet}$ $\left(resp.\ f_{i,\bullet}=\frac{n_{i,\bullet}}{n}\right)$.
- L'*effectif marginal de la modalité* a_j $\big(resp.\ fréquence\ marginale\ de\ la\ modalité\ a_j\big)$ par : $n_{\bullet,j}$ $\left(resp.\ f_{\bullet,j}=\frac{n_{\bullet,j}}{n}\right)$.

Proposition ALEA.17.9 | Propriétés des fréquences/effectifs marginaux

Soit une série (x,y) statistique de taille n admettant un nombre fini de modalités ou à défaut de classes.

1. $\sum_{i=1}^{p}\sum_{j=1}^{q} f_{i,j} = \sum_{i=1}^{p} f_{i,\bullet} = \sum_{j=1}^{q} f_{\bullet,j} = 1$,
2. $f_{i,\bullet} = \sum_{j=1}^{p} f_{i,j}$ et $f_{\bullet,j} = \sum_{i=1}^{p} f_{i,j}$.

Preuve Diviser par n les égalités précédemment établies sur les effectifs.

Remarque 2.16 — Analogie avec les probabilités Reprenons les notations précédentes, *i.e.* une série statistique bivariée de taille $p \times q$ admettant un nombre fini de modalités ou à défaut de classes notées $(a_i, b_j)_{(i,j)\in[\![1\,,p]\!]\times[\![1\,,q]\!]}$. On note $(n_{i,j})_{(i,j)\in[\![1\,,p]\!]\times[\![1\,,q]\!]}$ les effectifs correspondant. Alors si l'on considère un vecteur aléatoire discret (X, Y) tel que[9] :

$$(\mathrm{X},\mathrm{Y})(\Omega) = \Big\{(a_i,b_j),\ (i,j)\in[\![1\,,p]\!]\times[\![1\,,q]\!]\Big\},$$

$$\mathbf{P}\Big(\mathrm{X}=a_i,\mathrm{Y}=b_j\Big) = \frac{n_{i,j}}{n},\quad (i,j)\in[\![1\,,p]\!]\times[\![1\,,q]\!].$$

On vérifie sans peine que :

$$\sum_{i=1}^{p}\sum_{j=1}^{q}\frac{n_{i,j}}{n} = 1,\quad \frac{n_{i,j}}{n}\geqslant 0$$

pour tout $(i,j)\in[\![1\,,p]\!]\times[\![1\,,q]\!]$, ce qui garantit l'existence de (X, Y). Alors les formules précédentes traduisent simplement que

$$\Big\{\mathrm{X}=a_i,\mathrm{Y}=b_j\Big\}_{[\![1\,,p]\!]\times[\![1\,,q]\!]}$$

est un système complet d'évènements, et d'autre part le lien entre loi marginale et loi conjointe. La covariance d'une série statistique que nous allons définir ci-après sera simplement **Cov** (X, Y).

[9] on voit donc la série statistique bivariée comme la réalisation d'un couple aléatoire discret, de loi la fréquence conjointe de ladite modalité.

2.2.3. Caractéristiques de position & dispersion

On définit les moyennes et variances de manière similaire au cas univarié, pour chacune des séries x et y. On peut les exprimer en fonction des notations propres aux séries bivariées.

Proposition ALEA.17.10 | Expression de la moyenne

1. **(Nombre fini de modalités)** Soit une série (x,y) statistique de taille n admettant un nombre fini de modalités ou à défaut de classes.
$$\overline{x} = \frac{1}{n}\sum_{i=1}^{p} n_{i,\bullet}a_i = \sum_{i=1}^{p} f_{i,\bullet}a_i,\qquad \overline{y} = \frac{1}{n}\sum_{j=1}^{q} n_{\bullet,j}b_j = \sum_{j=1}^{q} f_{\bullet,j}b_j.$$
2. **(Regroupement en classes)** On remplace les a_i dans les définitions des moyennes par les milieux des classes, **mais les formules précédentes ne sont plus vraies.**

Preuve Montrons **1** pour x. Les modalités de la série statistique univariée x sont les a_i, $i\in[\![1\,,p]\!]$. Chaque modalité apparaît avec une fréquence $\sum_{j=1}^{q} f_{i,j}$, qui est la proportion des modalités $(a_i,y_1),\dots,(a_i,y_q)$ dans la série statistique bivariée de départ. Ainsi,

$$\overline{x} = \sum_{i=1}^{n}\left(\sum_{j=1}^{q} f_{i,j}\right)a_i,$$

on reconnaît alors les fréquences partielles $\overline{x} = \sum_{i=1}^{n} f_{i,\bullet}a_i$. La formule est démontrée.

Définition ALEA.17.20 | Point moyen du nuage

Le point de coordonnées $(\overline{x},\overline{y})$ est appelé *point moyen du nuage* associé à la série bivariée (x,y).

Proposition ALEA.17.11 | Expression de la variance

1. **(Nombre fini de modalités)** Soit une série (x,y) statistique de taille n admettant un nombre fini de modalités ou à défaut de classes.

$$\mathbf{V}_x = \frac{1}{n}\sum_{k=1}^{n}(x_k-\overline{x})^2 = \frac{1}{n}\sum_{i=1}^{p} n_{i,\bullet}(a_i-\overline{x})^2 = \sum_{i=1}^{p} f_{i,\bullet}(a_i-\overline{x})^2,$$
$$\mathbf{V}_y = \frac{1}{n}\sum_{k=1}^{n}(y_k-\overline{y})^2 = \frac{1}{n}\sum_{j=1}^{q} n_{\bullet,j}(b_j-\overline{y})^2 = \sum_{j=1}^{q} f_{\bullet,j}(b_j-\overline{y})^2.$$

2. **(Regroupement en classes)** On remplace les a_i dans les définitions des moyennes par les milieux des classes, **mais les formules précédentes ne sont plus vraies.**

Remarque 2.17 — Là aussi, si les données sont regroupées en classes on prend pour a_i et b_j les centres des classes

Preuve Même preuve que pour la moyenne.

Définition ALEA.17.21 | Covariance

Soit une série (x,y) statistique de taille n admettant un nombre fini de modalités ou à défaut de classes. On appelle *covariance de x et de y*, notée $\mathbf{C}_{x,y}$, la quantité

$$\mathbf{C}_{x,y} = \frac{1}{n}\sum_{k=1}^{n}(x_k-\overline{x})(y_k-\overline{y}) \qquad \text{ou de manière équivalente}$$
$$\mathbf{C}_{x,y} = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{i,j}(a_i-\overline{x})(b_j-\overline{y}).$$

Remarque 2.18 — Interprétation

- Si $\mathbf{C}_{x,y} < 0$ alors x et y ont tendance à varier dans des sens opposés (quand l'un augmente l'autre diminue), si $\mathbf{C}_{x,y} > 0$ alors ils ont tendance à varier dans le même sens.
- On a $\mathbf{Cov}(x,x) = \mathbf{V}_x$.

Proposition ALEA.17.12 | König-Huygens

Soit une série (x,y) statistique de taille n admettant un nombre fini de modalités ou à défaut de classes. Alors :

$$\mathbf{C}_{x,y} = \overline{xy} - \overline{x}\cdot\overline{y}.$$

Remarque 2.19 — C'est la moyenne des produits moins le produits des moyennes

Preuve

$$\begin{aligned}
\boldsymbol{C}_{x,y} &= \frac{1}{n}\sum_{k=1}^{n}(x_k-\overline{x})(y_k-\overline{y}) \\
&= \frac{1}{n}\sum_{k=1}^{n}\left(x_k y_k - \overline{x}y_k - \overline{y}x_k + \overline{x}\cdot\overline{y}\right) \qquad \textit{développement du produit} \\
&= \frac{1}{n}\sum_{k=1}^{n} x_k y_k - \frac{1}{n}\sum_{k=1}^{n}\overline{x}y_k - \frac{1}{n}\sum_{k=1}^{n}\overline{y}x_k + \frac{1}{n}\sum_{k=1}^{n}\overline{x}\cdot\overline{y} \qquad \textit{linéarité de la somme} \\
&= \overline{xy} - \overline{x}\cdot\overline{y} - \overline{x}\cdot\overline{y} + \overline{x}\cdot\overline{y} \\
&= \overline{xy} - \overline{x}\cdot\overline{y}
\end{aligned}$$

Proposition ALEA.17.13 | Variance d'une somme

Soit une série (x,y) statistique de taille n admettant un nombre fini de modalités ou à défaut de classes. Alors :

$$\mathbf{V}_{x+y} = \mathbf{V}_x + 2\mathbf{C}_{x,y} + \mathbf{V}_y$$

Preuve

$$\begin{aligned}\mathbf{V}_{x+y} &= \overline{(x+y)^2} - \overline{x+y}^2 \\ &= \overline{x^2+2xy+y^2} - (\overline{x}+\overline{y})^2 \\ &= \overline{x^2} + 2\overline{xy} + \overline{y^2} - \overline{x}^2 - \overline{y}^2 - 2\overline{x}\cdot\overline{y} \qquad \text{linéarité de l'espérance} \\ &= \overline{x^2} - \overline{x}^2 + \overline{y^2} - \overline{y}^2 + 2\left(\overline{xy} - \overline{x}\,\overline{y}\right) \qquad \text{formule de König-Huygens} \\ &= \mathbf{V}_x + 2\mathbf{C}_{x,y} + \mathbf{V}_y.\end{aligned}$$

On a également une «inégalité de Cauchy-Schwarz» pour la covariance.

Définition/Proposition ALEA.17.1 | Inégalité de Cauchy-Schwarz/ Coefficient de corrélation.

Soit une série (x,y) statistique de taille n admettant un nombre fini de modalités ou à défaut de classes.

1. **(Inégalité de Cauchy-Schwarz)**

$$|\mathbf{C}_{x,y}| \leqslant \sqrt{\mathbf{V}_x}\sqrt{\mathbf{V}_y} \qquad \text{ou encore} \qquad |\mathbf{C}_{x,y}| \leqslant \sigma_x\sigma_y.$$

Si x,y sont d'écart-type non nul. On appelle *coefficient de corrélation entre x et y* la quantité $\rho(X,Y) = \frac{\mathbf{Cov}(X,Y)}{\sigma_X.\sigma_Y} \in [-1,1]$.

2. **(Cas d'égalité)** $\rho(x,y) = \pm 1 \iff \exists a,b \in \mathbf{R}$, $y = ax+b$, *i.e.* la série y dépend de x et de manière affine.

Preuve *On définit l'application*

$$\mathrm{P} \left| \begin{array}{ccc} \boldsymbol{R} & \longrightarrow & \boldsymbol{R}, \\ t & \longmapsto & \boldsymbol{V}_{x+ty}. \end{array} \right.$$

En développant la variance, d'après une proposition précédente, il vient :

$$\forall t \in \boldsymbol{R}, \qquad \mathrm{P}(t) = \sigma_x^2 + 2t\boldsymbol{C}_{x,y} + t^2\sigma_y^2.$$

L'application P *est donc une application polynomiale de degré* 2*. De plus, d'après les propriétés de la variance on a*

$$\forall t \in \boldsymbol{R}, \qquad \mathrm{P}(t) \geqslant 0.$$

Comme P *est de signe constant il ne peut pas admettre deux racines réelles distinctes, son discriminant est négatif ou nul, c'est-à-dire*

$$\Delta = 4\boldsymbol{C}_{x,y}^2 - 4\sigma_x^2\sigma_y^2 \leqslant 0.$$

D'où

$$\boldsymbol{C}_{x,y}^2 \leqslant \sigma_x^2\sigma_y^2.$$

Par croissance de la fonction racine carrée on a alors

$$|\boldsymbol{C}_{x,y}| \leqslant \sigma_x\sigma_y.$$

Pour le cas d'égalité, on constate qu'il est obtenu lorsque $\Delta = 0 \iff$ P *possède une racine réelle (double a fortiori), i.e. si et seulement si il existe* $t \in \boldsymbol{R}$, $\mathrm{P}(t) = \boldsymbol{V}_{x+ty} = 0$ *donc si et seulement si la série* $x + ty$ *est constante. C'est ce qu'il fallait montrer.*

2.2.4. Ajustement affine

Il est courant, en physique-chimie, en sciences industrielles, ou plus généralement dans toute discipline expérimentale (biologie, chimie, économie, ...), d'avoir à comparer des données expérimentales et de conjecturer une éventuelle dépendance linéaire entre deux paramètres donnés. Vous pourriez avoir ce besoin lors de vos TIPE. Notez qu'il est aussi possible d'étudier les dépendances polynomiales entre deux paramètres pour un degré quelconque, nous n'aborderons pas ce point ici.

L'idée de l'ajustement affine est la suivante : on dispose de deux séries statistiques (souvent expérimentales) x et y et on soupçonne qu'il existe une relation les liant de la forme $y = ax+b$: par exemple après avoir trouvé un coefficient de corrélation proche de un. On veut alors chercher la droite d'équation $y = ax+b$ qui passe «le mieux» par notre nuage de points. Parfois on sait que la relation existe et on veut déterminer a et b.

Plus précisément, soit $(x_i,y_i)_{1\leqslant i\leqslant n}$ avec $n \geqslant 1$ est un nuage de n points provenant de séries statistiques x,y. En regardant un dessin, nous voyons que si l'on approche le

nuage par la droite $y = ax + b$ avec $(a,b) \in \mathbf{R}^2$, alors l'écart entre cette droite et le nuage au point x_i, $i \in [\![1\,,\,n]\!]$, est donné par : $\quad y_i - ax_i - b$.

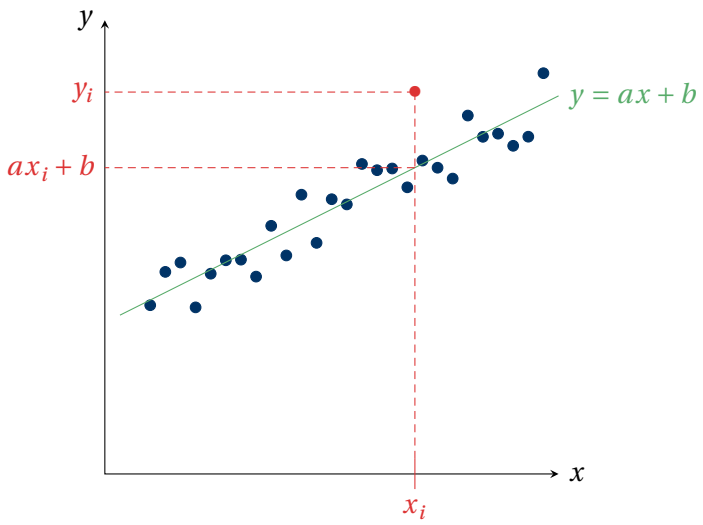


FIG. ALEA.17.2. : Problème de régression linéaire

Ainsi, on pourrait se poser par exemple la question de la minimisation en a, b des quantités suivantes :

$$\max_{1\leqslant i\leqslant n} \left|y_i - ax_i - b\right|, \quad \sum_{i=1}^{n} \left|y_i - ax_i - b\right|, \quad \mathrm{F}(a,b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2,$$

de sorte que quand l'une ou l'autre de ces trois quantités est petite, tous les écarts à la droite sont également petits. Dans le dernier cas, on parle de *minimisation au sens des moindres carrés* et c'est cette minimisation que nous allons essayer de réaliser car c'est pour celle-ci que les calculs sont les plus simples. Nous pouvons résoudre ce problème de deux manières. Il s'agira donc de minimiser la fonction de deux variables

$$\mathrm{F} \left| \begin{array}{ccc} \mathbf{R}^2 & \longrightarrow & \mathbf{R} \\ (a,b) \in \mathbf{R}^2 & \longmapsto & \displaystyle\sum_{i=1}^{n} (y_i - ax_i - b)^2, \end{array} \right.$$

i.e. de déterminer

$$\inf_{(a,b)\in\mathbf{R}^2} \sum_{i=1}^{n} (y_i - ax_i - b)^2 = \inf_{(a,b)\in\mathbf{R}^2} \mathrm{F}(a,b).$$

HEURISTIQUE. Faisons une heuristique. Si y est réellement affine en x, *i.e.* de la forme $y = ax + b$ avec $a, b \in \mathbf{R}$, alors d'après les propriétés de la covariance et de l'espérance déjà établie, nous avons :

$$\overline{y} = \overline{ax+b} = a\overline{x} + b \quad \Longrightarrow \quad \text{la droite passe par le point moyen, } b = \overline{y} - a\overline{x}$$

$$\mathbf{Cov}(y,x) = a\mathbf{Cov}(x,x) + 0 \quad \Longrightarrow \quad a = \frac{\mathbf{Cov}(y,x)}{\sigma_x^2}.$$

Il s'avère que le couple (a,b) obtenu précédemment, noté $(a^\star, b^\star)$ dans la suite, sera également la solution obtenue au sens des moindres carrés. C'est ce que nous montrons dès à présent.

Théorème ALEA.17.1 | Existence de la droite des moindres carrés

Soit (x,y) une série statistique double constituée d'une suite de couples $((x_k,y_k))_{1\leqslant k\leqslant n}$. Alors $(a^\star, b^\star)$ défini ci-dessous est l'unique minimum global de F :

$$a^\star = \frac{\mathbf{C}_{x,y}}{\sigma_x^2}, \qquad b^\star = \overline{y} - a^\star \overline{x}.$$

La droite de régression par la méthode des moindres carrés de y en x a donc pour équation :

$$y = \frac{\mathbf{C}_{x,y}}{\sigma_x^2}(x - \overline{x}) + \overline{y}.$$

Première démonstration : en optimisant des trinômes. Soit $(a,b)\in\mathbf{R}^2$. Comme

$$\frac{\partial \mathrm{F}}{\partial b}(a,b)=-2\sum_{k=1}^{n}(ax_k+b-y_k)=-2\left[a\sum_{k=1}^{n}x_k+b\underbrace{\sum_{k=1}^{n}1}_{n}-\sum_{k=1}^{n}y_k\right]=0$$

$$\iff \quad b=\frac{\sum_{k=1}^{n}y_k}{n}-a\frac{\sum_{k=1}^{n}x_k}{n}=\overline{y}-a\overline{x},$$

et que pour tout $a\in\mathbf{R}$, le graphe de $b\longmapsto \mathrm{F}(a,b)$ est une parabole orientée vers le haut, nous avons :

$$\forall a,b\in\mathbf{R},\quad \mathrm{F}(a,b)\geqslant \mathrm{F}(a,\overline{y}-a\overline{x}).$$

On considère ensuite :

$$f:a\in\mathbf{R}\longmapsto \mathrm{F}\left(a,\overline{y}-a\overline{x}\right).$$

Puisque

$$\begin{aligned}f(a)&=\sum_{k=1}^{n}\left[a(x_k-\overline{x})-(y_k-\overline{y})\right]^2\\&=a^2\sum_{k=1}^{n}(x_k-\overline{x})^2-2a\sum_{k=1}^{n}(x_k-\overline{x})(y_k-\overline{y})\\&\quad+\sum_{k=1}^{n}(y_k-\overline{y})^2\\&=a^2n\sigma_x^2-a\left(2n\mathbf{C}_{x,y}\right)+n\sigma_y^2.\end{aligned}$$

identité remarquable

polynôme de degré 2 en a

$$f'(a)=2an\sigma_x^2-2n\mathbf{C}_{x,y}.$$

Comme f est encore un trinôme de graphe une parabole orientée vers le haut, elle est minimale là où f' s'annule, *i.e.* en $a=\frac{\mathbf{C}_{x,y}}{\sigma_x^2}$. En résumé, nous avons montré :

$$\forall a,b\in\mathbf{R},\quad \mathrm{F}(a,b)\geqslant \mathrm{F}(a,\overline{y}-a\overline{x})\geqslant \mathrm{F}\left(\frac{\mathbf{C}_{x,y}}{\sigma_x^2},\overline{y}-\frac{\mathbf{C}_{x,y}}{\sigma_x^2}\overline{x}\right).$$

Cette inégalité prouve que $\left(\frac{\mathbf{C}_{x,y}}{\sigma_x^2},\overline{y}-\frac{\mathbf{C}_{x,y}}{\sigma_x^2}\overline{x}\right)$ est un minimum global de F. Un calcul de points critiques montre ensuite qu'il s'agit du seul minimum possible :

En conclusion : $\left(\frac{C_{x,y}}{\sigma_x^2}, \overline{y} - \frac{C_{x,y}}{\sigma_x^2}\overline{x}\right)$ est l'unique minimum (global) de F sur $\mathbf{R}^2$.

Seconde démonstration : en utilisant une projection orthogonale. Nous notons $E = \mathbf{R}^2$. Interprétons autrement la quantité

$$\inf_{(a,b)\in\mathbf{R}^2} \sum_{i=1}^{n} (y_i - ax_i - b)^2 = \inf_{(a,b)\in\mathbf{R}^2} F(a,b).$$

On constate que :

$$\begin{aligned}\inf_{(a,b)\in\mathbf{R}^2} \sum_{i=1}^{n} (y_i - ax_i - b)^2 &= \inf_{(a,b)\in\mathbf{R}^2} \left\|(y_1,\dots,y_n) - a(x_1,\dots,x_n) - b(1,...,1)\right\|^2 \\ &= \inf_{(a,b)\in\mathbf{R}^2} \|Y - aX - b\mathbb{1}\|^2 = \inf_{Z\in F} \|Y - Z\|^2 = d(Y,F)^2\end{aligned}$$

où $X = (x_1,\dots,x_n)$, $Y = (y_1,\dots,y_n)$, $\mathbb{1} = (1,\dots,1)$ et $F = \left\{aX + b, (a,b) \in \mathbf{R}^2\right\} = \text{Vect}(X, \mathbb{1})$.

Nous avons interprété le problème initial comme une distance minimale à une partie, qui elle-même s'exprime en fonction de la projection orthogonale de Y sur F :

$$\inf_{(a,b)\in\mathbf{R}^2} \sum_{i=1}^{n} (y_i - ax_i - b)^2 = \left\|Y - p_F(Y)\right\|^2.$$

Pour terminer, il reste à trouver $Y' = p_F(Y)$. En utilisant la définition d'une projection, on résout alors :

$$\left\{\begin{array}{rcl} Y' & \in & F \\ Y' - Z & \in & F^\perp \end{array}\right. \iff \left\{\begin{array}{rcl} Y' & = & aX + b, \quad a,b \in \mathbf{R} \\ Y - Y' & \perp & (1,\dots,1) \\ Y - Y' & \perp & X \end{array}\right.$$

$$\iff \left\{\begin{array}{rcl} Y' & = & aX + b, \quad a,b \in \mathbf{R} \\ Y - aX - (1,\dots,1) & \perp & (1,\dots,1) \\ Y - aX - (1,\dots,1) & \perp & X. \end{array}\right.$$

On aboutit alors au système sur $(a,b) \in \mathbf{R}^2$:

$$\left\{\begin{array}{rl} \sum_{i=1}^n (y_i - ax_i - b) & = 0 \quad \left(\rightleftarrows \frac{\partial F(a,b)}{\partial b} = 0\right) \\ \sum_{i=1}^n (y_i - ax_i - b)x_i & = 0 \quad \left(\rightleftarrows \frac{\partial F(a,b)}{\partial a} = 0\right), \end{array}\right.$$

qui a pour unique solution le couple $(a^\star, b^\star)$ trouvée précédemment (en cherchant les points critiques). Seulement, la théorie sur la projection orthogonale déroulée plus haut garantit que la solution obtenue cette fois-ci est un minimum : inutile donc de chercher à l'établir. On récupère de suite l'unicité par unicité de la projection orthogonale sur un sous-espace vectoriel de dimension finie.

Remarque 2.20 — À propos du vocabulaire Pourquoi parle-t-on de «régression linéaire»? La réponse est une erreur de traduction. Le mathématicien anglais Sir Galton étudiait les tailles des fils (y_j) en fonction de la taille de leur père (x_j) ; et a constaté un «retour à la moyenne». En effet, les grands individus ont en moyenne des enfants plus petits qu'eux et les petits individus ont des enfants plus grand qu'eux. En Anglais le terme «retour à la moyenne» est «régression to the mean», ce terme a ensuite été mal transposé au Français.

Qualité d'une régression linéaire. Comment évaluer la «justesse» d'un ajustement? Pour y répondre on définit un nouvel indicateur statistique : le coefficient de détermination, plutôt que le seul coefficient de corrélation.

Définition/Proposition ALEA.17.2 | Coefficient de détermination d'une régression

Soit (x, y) une série statistique double constituée d'une suite de couples $((x_k, y_k))_{1\leqslant k\leqslant n}$. On appelle *coefficient de détermination de x et y*, noté $r^2(x, y)$, la quantité définie par :

$$r^2(x,y) = \rho(x,y)^2 = \frac{\mathbf{C}_{x,y}^2}{\mathbf{V}_x \mathbf{V}_y} \in [0,1].$$

Attention

Ce n'est donc pas le coefficient de corrélation, mais son carré.

Preuve Puisque $\rho(x, y) \in [-1, 1]$, son carré est bien dans $[0, 1]$.

Remarque 2.21 — Interprétation Ainsi $r^2(x, y) = 1$ correspond à une adéquation parfaite tandis que $r^2(x, y)$ proche de 0, équivalent à $\rho(x, y)$ proche de 0, indique une faible liaison linéaire ce qui peut signifier qu'il n'y a pas de lien entre x et y ou bien que x et y sont liés par une relation non-affine. En général, on considère une régression linéaire comme «satisfaisante» lorsque

$$\boxed{r^2(x,y) \geqslant 0.9}.$$

Autres ajustements se ramenant à une régression linéaire. On peut penser à beaucoup d'ajustements. Par exemple :

1. si l'on souhaite tester la relation $y = \lambda e^{\alpha x}$, avec $(\alpha, \lambda) \in \mathbf{R} \times \mathbf{R}^{+\star}$ avec y série statistique strictement positive, on peut constater qu'elle est équivalente à $(\ln y = \ln\lambda + \alpha x$: on fait alors la régression sur $(x, \ln y)$. Comment retrouver α, λ à partir de $a^\star, b^\star$? 🖌
2. si l'on souhaite tester la relation $y = a\ln x + b$, avec $(a, b) \in \mathbf{R}$, on peut faire alors la régression sur $(\ln x, y)$.

2.2.5. En Python : calcul des grandeurs bivariées

On suppose dans la suite que toutes les données d'une série à nombre de modalités fini sont contenues dans une liste L donnée en paramètre.

Covariance

```
def covariance(L, M):
    """
    Renvoie la covariance des deux séries
    """
    Prod = [L[i]*M[i] for i in range(len(M))]
    return esperance(Prod) - esperance(L)*esperance(M)
```

Coefficient de corrélation

```
import math as ma
def coeff_cor(L, M):
    """
    Renvoie le coefficient de corrélation des deux séries
    """
    return covariance(L,
     ↪ M)/(ma.sqrt(variance(L))*ma.sqrt(variance(M)))
```

3. STATISTIQUES INFÉRENTIELLES

Rappelons que l'inférence statistique (*cf.* introduction) consiste à savoir si une série statistique x peut être vue comme plusieurs réalisations d'une même variable aléatoire, et estimer les valeurs des paramètres de cette loi (espérance, variance, *etc.*). Consulter l'introduction pour plus de détails.

3.1. Estimation ponctuelle

Dans cette première sous-section, l'inférence va se situer dans l'aspect suivant : on essaie d'estimer les paramètres d'une loi à l'aide de réalisations $X_1, \dots, X_n$ (des variables aléatoires formant ce qu'on appellera un *échantillon* dans la suite) de cette loi.

Définition ALEA.17.22 | n-échantillon et estimateur

1. Un *n-échantillon* est un n-uplet $(X_1, \dots, X_n)$ de variables aléatoires indépendantes et de même loi (i.i.d.). Une *observation (ou réalisation) de* $(X_1, \dots, X_n)$ est $(X_1(\omega), \dots, X_n(\omega))$ pour un certain $\omega \in \Omega$. La loi commune s'appelle la *loi parente* ou la *loi mère*. On parle d'*échantillon gaußien* lorsque tous les X_i, $i \in [\![1, n]\!]$ suivent une loi normale.
2. Un *estimateur d'un paramètre* θ inconnu est une suite de variables aléatoires $\left(\widehat{\theta_n}\right)$, où $\widehat{\theta_n}$ est une fonction de $(X_1, \dots, X_n)$, *i.e.* $\widehat{\theta_n} = \varphi_n(X_1, \dots, X_n)$ avec $\varphi_n : \mathbf{R}^n \longrightarrow \mathbf{R}$.

Le plus souvent le paramètre θ sera l'espérance ou la variance de la loi de X, une médiane, ou toute autre valeur caractéristique d'une loi de probabilité.

Attention à la terminologie « estimateur »

On s'attend à ce qu'un estimateur « estime » θ au sens propre du terme *i.e.* qu'il en soit très proche. La définition précédente ne garantit pas cela : les propriétés vraiment intéressantes d'un estimateur seront énoncées *infra*.

Cadre

Dans toute la suite de cette section, on se fixe un n-échantillon $(X_1, \dots, X_n)$ de même loi qu'une variable aléatoire réelle X. On suppose de plus que cette loi commune dépend d'un paramètre $\theta \in \mathbf{R}$.

Exemple 9 — *Uniforme en zéro* Par exemple, $(X_1, \dots, X_n)$ et X où $X_1, \dots, X_n$ sont i.i.d. de loi $\mathcal{U}[0, \theta]$, et X de même loi indépendante des X_i. C'est un n-échantillon, et $\widehat{\theta_n} = X_1 + \dots + X_n$ est un estimateur de θ. Calculer $\mathbf{E}\left(\widehat{\theta_n}\right)$.

Définition ALEA.17.23 | Qualité d'un estimateur

Soit $\left(\widehat{\theta_n}\right)$ un estimateur de θ.

1. On appelle *erreur d'estimation* la **variable aléatoire** $\widehat{\theta_n} - \theta$.
2. Supposons que $\widehat{\theta_n}$ admet un moment d'ordre un. On appelle *biais*, l'espérance de l'erreur d'estimation :

$$b(\theta_n) = \mathbf{E}\left(\widehat{\theta_n} - \theta\right) = \mathbf{E}\left(\widehat{\theta_n}\right) - \theta.^{11}$$

On dit qu'un estimateur est :
- *sans biais* (ou qu'*il estime* θ *de manière non biaisée*) si $b\left(\theta_n\right) = 0$ pour tout $n \in \mathbf{N}$, il est dit *biaisé* dans le cas contraire.
- Il est dit *asymptotiquement sans biais* si $b\left(\theta_n\right) \xrightarrow{n\to\infty} 0$.

3. Supposons que $\widehat{\theta_n}$ admet un moment d'ordre deux. On appelle *risque quadratique* la quantité :

$$r_n(\theta) = \mathbf{E}\left((\widehat{\theta_n} - \theta)^2\right).$$

On dit qu'un estimateur est *plus efficace qu'un autre* si son risque quadratique est moins élevé.

4. Un estimateur $\left(\widehat{\theta_n}\right)$ est dit *convergeant vers* θ si :

$$\forall \varepsilon > 0, \quad \mathbf{P}\left(\left|\widehat{\theta_n} - \theta\right| \geqslant \varepsilon\right)^{12} \xrightarrow{n\to\infty} 0,$$

i.e. s'il converge en probabilité vers le paramètre θ.

Exemple 10 — ***Élection présidentielle*** À l'approche du second tour d'une élection présidentielle, on interroge une personne au hasard et on note $X = 1$ si elle se prononce pour le candidat A et $X = 0$ si c'est pour le candidat B. Alors X suit une loi de Bernoulli de paramètre $\theta \in [0,1]$ inconnu[10] qui correspond à la proportion de français qui votent pour A.

On questionne cinq individus sur leurs intentions de vote et on obtient les résultats suivants (en notant 1 ou 0 selon que le choix se porte sur le candidat A ou B) : $1, 0, 0, 1, 0$. Ces résultats observés correspondent, pour tout $\theta \in [0,1]$, à la réalisation d'un 5 - échantillon $(X_1, \dots, X_5)$ de loi mère $\mathscr{B}(\theta)$.

Proposition ALEA.17.14 | Décomposition biais/variance

Soit $\left(\widehat{\theta_n}\right)$ un estimateur du paramètre θ admettant un moment d'ordre deux. Alors pour tout entier $n \in \mathbf{N}$:

$$r_n(\theta) = \mathbf{Var}\left(\widehat{\theta_n}\right) + \left(b\left(\theta_n\right)\right)^2.$$

[10] Sauf bien entendu si on se dit capable de poser la question à l'ensemble de la population, ce qui n'est pas réalisable pour un institut de sondage

[11] par linéarité de l'espérance

[12] Donc plus on ajoute d'observation, plus la probabilité de s'écarter de θ est faible

Au même titre que la formule de König-Huygens s'utilise le plus souvent pour calculer une variance lorsque l'on a déjà calculé une espérance, la décomposition biais/variance s'utilise pour calculer un risque lorsque l'on a déjà calculé le biais.

Preuve *(Point clef — **Développer le carré**)*

Proposition ALEA.17.15 | Condition suffisante de convergence

Soit $\left(\widehat{\theta_n}\right)$ un estimateur de θ admettant un moment d'ordre deux, et tel que : $\lim\limits_{n\longrightarrow\infty} r_n(\theta) = 0$. Alors :

$$\left(\widehat{\theta_n}\right)_{n\in\mathbf{N}} \text{ est convergeant vers } \theta \in \mathbf{R}.$$

Preuve *(Point clef — **On applique l'inégalité de Markov pour les moments d'ordre deux**)*

Étude des estimateurs de la moyenne/variance/écart-type empiriques. On va essentiellement s'intéresser dans la suite aux estimateurs de l'espérance et de la variance, appelés moyenne et variance empirique. Rappelons l'expression de ces estimateurs vue dans le **??**.

Définition ALEA.17.24 | Moyenne/Variance empirique

Soient $X_1, \dots, X_n$ une famille de $n \in \mathbf{N}^\star$ variables aléatoires réelles. On appelle *moyenne empirique des* X_i *(resp. variance empirique)* les variables aléatoires réelles

$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \left(\textit{resp.} \qquad \sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X_n})^2\right).$$

On appelle *écart-type empirique* la variable aléatoire

$$\sigma_n \underset{\textbf{(défi.)}}{=} \sqrt{\sigma_n^2}.$$

Nous avions également établi une version König-Huygens en développant le carré *via* une identité remarquable.

Proposition ALEA.17.16 | Version König-Huygens

Soient $X_1, \dots, X_n$ une famille de $n \in \mathbf{N}^\star$ variables aléatoires réelles. Alors :

$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X_n}^2.$$

Corollaire ALEA.17.1 | Lien entre moyenne/variance empirique pour des Bernoulli

Soient $X_1, \dots, X_n$ une famille de $n \in \mathbf{N}^\star$ variables aléatoires réelles de loi $\mathscr{B}(p)$ avec $p \in [0,1]$ et $n \in \mathbf{N}^\star$. Alors

$$\sigma_n^2 = \overline{X_n} - \overline{X_n}^2.$$

Il n'y a en règle général pas de lien entre espérance et variance empirique (ce qui semble logique). C'est ici un cas très particulier qui provient du fait suivant : la somme des carrés des X_i est égale à la somme, puisqu'une Bernoulli est à valeurs dans $\{0,1\}$.

Preuve

Passons maintenant aux qualités de ces estimateurs.

Proposition ALEA.17.17 | Qualité de la moyenne/variance empirique

On suppose que $X_1, \dots, X_n$ admettent une espérance μ et une variance $\sigma^2 > 0$, $n \in \mathbf{N}^\star$.

1. **(Biais de la moyenne empirique)**

$$\mathbf{E}\left(\overline{X_n}\right) = \mu, \quad \text{et} \quad \mathbf{Var}\left(\overline{X_n}\right) = \frac{\sigma^2}{n}.$$

En particulier, la moyenne empirique $\left(\overline{X_n}\right)$ est un estimateur sans biais de l'espérance.

2. **(Biais de la variance empirique)**

$$\mathbf{E}\left(\sigma_n^2\right) = \frac{n-1}{n}\sigma^2, \quad \text{et donc :} \quad \mathbf{E}\left(\frac{n}{n-1}\sigma_n^2\right) = \sigma^2.$$

En particulier, la variance empirique $\left(\sigma_n^2\right)$ est un estimateur biaisé de la variance.

Preuve Il s'agit ici de calculer les biais de $\overline{X_n}$ et σ_n^2.

1. *Soit $n \in \mathbf{N}^\star$. Nous avons d'une part par linéarité de l'espérance :*

$$\boldsymbol{E}\left(\overline{X_n}\right) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{E}(X_i) = \frac{1}{n}(n\,\mu) = \mu,$$

et d'autre part par argument d'indépendance des X_i :

$$\boldsymbol{Var}\left(\overline{X_n}\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

2. Nous allons utiliser la version König-Huygens de l'estimateur σ_n^2, *i.e.*

$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X_n}^2.$$

$$\begin{aligned}
\boldsymbol{E}\left(\sigma_n^2\right) &= \boldsymbol{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X_n}^2\right), \\
&= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{E}\left(X_i^2\right) - \boldsymbol{E}\left(\overline{X_n}^2\right), && \text{linéarité de l'espérance} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\sigma^2+\mu^2\right) - \left(\boldsymbol{Var}\left(\overline{X_n}\right) + \boldsymbol{E}\left(\overline{X_n}\right)^2\right), && \text{lien moment d'ordre deux et variance} \\
&= \sigma^2 + \mu^2 - \left(\frac{\sigma^2}{n} + \mu^2\right) && \text{d'après } \boldsymbol{1} \\
&= \frac{n-1}{n}\sigma^2.
\end{aligned}$$

Nous déduisons, par linéarité de l'espérance une version dite « corrigée » de l'estimateur de la variance, *i.e.* un estimateur sans biais appelé *variance corrigée*.

Définition ALEA.17.25 | Variance empirique corrigée

On appelle *variance empirique corrigée des* X_i *(resp. écart-type corrigé)* les variables aléatoires définies pour tout $n \geqslant 2$ par :

$$\sigma_n^{2,\text{cor}} = \frac{n}{n-1}\sigma_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X_n})^2 \quad \left(\textit{resp.} \qquad \sigma_n^{\text{cor}} = \sqrt{\sigma_n^{2,\text{cor}}}\right).$$

Proposition ALEA.17.18 | Qualité de la variance corrigée

On suppose que $X_1, \dots, X_n$ admettent une espérance μ et une variance $\sigma^2 > 0$. Alors pour tout $n \geqslant 2$:

$$\mathbf{E}\left(\sigma_n^{2,\text{cor}}\right) = \sigma^2.$$

Ainsi, la variance empirique corrigée $\sigma_n^{2,\text{cor}}$ est un estimateur sans biais de la variance.

Preuve

Proposition ALEA.17.19 | Convergence

On suppose que $X_1, \dots, X_n$ admettent une espérance μ et une variance $\sigma^2 > 0$, $n \in \mathbf{N}^\star$. Alors :

L'estimateur $\left(\overline{X_n}\right)_{n\in\mathbf{N}^\star}$ $\left(\textit{resp.}\ (\sigma_n^2)_{n\geqslant 2}\right)$ converge vers μ (*resp.* σ^2).

Preuve *(Point clef — **Les risques quadratiques convergent vers zéro**)*

Nous admettons la convergence de σ_n^2.

Exemple 11 — Reprenons l'Exemple 10, on note $(X_1, \dots, X_n)$ un $n \in \mathbf{N}^\star$-échantillon associé. Un estimateur naturel pour θ est $\overline{X_n}$. On peut envisager d'autres estimateurs, par exemple

$$A_n = \frac{2}{n(n+1)}\sum_{k=1}^{n} kX_k, \quad \text{ou encore} \quad B_n = p.$$

Lequel des trois est le meilleur du point de vue du risque quadratique ? On commencera par calculer leur biais, puis leur risque quadratique. Commentez les résultats obtenus.

1. Qu'estime de manière non biaisée $2\overline{X_n}$?

2. Montrer que $\widehat{\theta_n}$ est une variable aléatoire à densité pour tout $n \geqslant 1$.

3. L'estimateur $\left(\widehat{\theta_n}\right)$ est-il biaisé ? Donner alors un estimateur non biaisé de θ.

Notez que des observations de ces trois estimateurs ont été donnés dans l'énoncé :

- $\overline{X_5} = \frac{1}{5}(1+0+0+1+0) = \frac{2}{5}$,
- $A_5 = \frac{1}{3}$,
- $B_5 = 0$.

Exemple 12 — ***Uniforme en zéro*** On prolonge l'Exemple 9. Si $(X_1, \dots, X_n)$ est un $n \in \mathbf{N}^\star$-échantillon de loi mère $\mathscr{U}[0,\theta]$ avec $\theta \in \mathbf{R}$. On note $\widehat{\theta_n} = \max(X_1, \dots, X_n)$.

4. Entre l'estimateur précédent et $2\overline{X_n}$, lequel choisir?

3.2. Estimation par Intervalle de confiance

Une fois le calcul d'un estimateur effectué, on ne peut pas se contenter d'une valeur estimée : il faut mesurer l'erreur commise entre la valeur inconnue et l'estimation. En effet, même avec un risque quadratique faible, on n'est jamais à l'abri de tomber sur un « mauvais » échantillon qui nous donnerait une mauvaise estimation du paramètre.

Cette estimation d'erreur est précisément la vocation de l'estimation par intervalle confiance qui est plus précise que la seule donnée d'un estimateur : nous allons donner des intervalles, contenant le paramètre à estimer, avec très forte probabilité. [13]

Définition ALEA.17.26 | Intervalle de confiance

Soit $\alpha \in]0,1[$ et $(X_1,\dots,X_n)$ un échantillon de variables aléatoires indépendantes et de même loi que X, loi dépendant d'un paramètre θ.

1. On appelle *intervalle de confiance de niveau* $1-\alpha$ *(on dit aussi de probabilité de confiance* $1-\alpha$*) pour le paramètre* θ, tout intervalle aléatoire noté $I_{X_1,\dots,X_n}$ dépendant des $X_1,\dots,X_n$ tel que :
$$\mathbf{P}\left(\theta \in I_{X_1,\dots,X_n}\right) \geqslant 1-\alpha.$$
2. On appelle *intervalle de confiance asymptotique de niveau* $1-\alpha$ *pour le paramètre* θ (on dit aussi de *probabilité de confiance* $1-\alpha$) toute suite $(I_{X_1,\dots,X_n})_n$ d'intervalles aléatoires telle que :
$$\lim_{n\longrightarrow\infty} \mathbf{P}\left(\theta \in I_{X_1,\dots,X_n}\right) \geqslant 1-\alpha.$$

Remarque 3.1 —

- Très souvent, on recherche un intervalle de confiance de θ sous la forme d'un intervalle centré en une estimation ponctuelle de θ. Par exemple, pour l'intervalle de confiance de la moyenne μ d'un échantillon, il sera centré en $\overline{X_n}$ le plus souvent.
- La plupart du temps, c'est ce niveau de risque de 0.05 qui est utilisé, et qui est communément accepté par exemple en sciences humaines. Mais dans des domaines plus sensibles où l'on n'a pas vraiment de droit à l'erreur (aérospatiale, physique nucléaire, etc), on travaille avec des niveaux de risque de 0.01, voir moins.

[13] En passant au complémentaire, la probabilité que le paramètre soit en dehors de cet intervalle sera donc très petite.

Proposition ALEA.17.20 | Stabilité par élargissement

Soit $\alpha \in]0,1[$ et $(X_1,\dots,X_n)$ un échantillon de variables aléatoires indépendantes et de même loi que X, loi dépendant d'un paramètre θ. Alors :
si $I_{X_1,\dots,X_n}$ est un intervalle de confiance de niveau $1-\alpha$ pour θ, alors pour tout intervalle $J_{X_1,\dots,X_n} \supset I_{X_1,\dots,X_n}$, $J_{X_1,\dots,X_n}$ est encore un intervalle de confiance de niveau $1-\alpha$ pour θ.

Preuve

On commence par deux exemples de recherche d'un intervalle de confiance non-asymptotique.

Exemple 13 — ***Pesons-nous avec Cruella. Intervalle de confiance non asymptotique obtenu par propriétés des lois normales.*** Pamela est un mannequin célèbre dont le poids est strictement surveillé par Cruella. Cette charmante dame a investi un jour dans l'achat d'une balance Harmonia afin de connaître précisément le poids de sa protégée. Horreur : elle a constaté sur l'emballage de la balance que les fabricants (d'honnêtes artisans suisses) admettaient que leur outil de mesure (nul n'est parfait) pouvait commettre des erreurs de mesure dont l'écart-type valait 0,1 kg, néanmoins l'étiquette précise que les mesures $(X_1,\dots,X_n)$ avec $n \in \mathbf{N}^\star$ sont gaußiennes. En effet les pièces détachées ne sont pas toutes exactement identiques, leur montage n'est jamais parfait et le transport à travers les Alpes endommage parfois les balances. Ne faisant ni une ni deux, Cruella, a, dès le lendemain, dévalisé le magasin en investissant dans l'achat de 99 nouvelles balances Harmonia et a forcé Pamela à sauter sur les 100 balances pendant que Cruella relevait scrupuleusement les 100 mesures. Résultat moyen des pesées : 55,4 kg. Donner à Cruella un intervalle de confiance pour le poids moyen de Pamela sur ce type de balance, de probabilité de confiance 0,95.

Remarque 3.2 — Nous n'avons pas eu besoin d'utiliser le théorème central limite ici, car l'échantillon de départ était déjà gaußien. Sinon, de manière générale, on utilisera la Proposition ALEA.17.21 ci-après.

Exemple 14 — *Intervalle de confiance non asymptotique obtenu via la loi faible des grands nombres* On considère une pièce dont on souhaite savoir si elle est truquée ou non. Pour cela, on peut la lancer autant de fois que l'on veut. Mathématiquement, étant donné $n \in \mathbf{N}^*$, on observe la réalisation d'un échantillon $(X_1, \dots, X_n)$ de variables aléatoires indépendantes et de même loi $\mathscr{B}(p)$, où $p \in]0;1[$ est **connu**.

On cherche alors un intervalle $[a,b]$, dont les bornes dépendent des observations, mais pas de p, et tel que la probabilité que le paramètre inconnu p appartienne à cet intervalle soit égale à 0,95. On note $\widehat{p_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

1. À l'aide de l'inégalité de Bienaymé-Tchebychev, justifier que pour tout $\varepsilon > 0$:
$$\mathbf{P}\left(|\hat{p}_n - p| > \varepsilon\right) \leq \frac{1}{4n\varepsilon^2}.$$
2. Déterminer une valeur de ε telle que $\mathbf{P}\left(|\hat{p}_n - p| > \varepsilon\right) \leq \alpha$.
3. En déduire que $\mathbf{P}\left(|\hat{p}_n - p| \leq \frac{1}{2\sqrt{\alpha}\sqrt{n}}\right) \geq 1-\alpha$ et donner un intervalle de confiance de niveau $\alpha = 0,05$.
4. On effectue 2000 lancers de la pièce et on trouve $\hat{p}_n = 0,57$. Que peut-on conclure ?

Intervalle de confiance asymptotique obtenu par le théorème central limite On cherche à présent un intervalle de confiance asymptotique pour μ à l'aide du théorème central limite. Puisque (X_n) est une suite de variables aléatoires i.i.d. possédant une variance, on peut lui appliquer le théorème central limite : pour tout $(a,b) \in \mathbf{R}^2$,

$$\lim_{n\to+\infty} \mathbf{P}\left(a \leq \sqrt{n}\frac{\overline{X_n}-\mu}{\sigma_n} \leq b\right) = \Phi(b)-\Phi(a).$$

On souhaite obtenir un intervalle de confiance asymptotique au niveau de confiance $1-\alpha$ avec $\alpha \in]0,1[$. Pour cela, on doit donc choisir a, b de sorte que

$$\Phi(b)-\Phi(a) \geq 1-\alpha.$$

Il y a une infinité de façon de choisir a et b. On choisit couramment $a=-b$ (intervalle symétrique), et donc

$$\Phi(b)-\Phi(a) = \Phi(b)-\Phi(-b) = \Phi(b)-(1-\Phi(b)) = 2\Phi(b)-1.$$

On cherche donc b de sorte que :

$$2\Phi(b)-1 = 1-\alpha \iff \Phi(b) = 1-\frac{\alpha}{2} \iff b = \Phi^{-1}\left(1-\frac{\alpha}{2}\right),$$

car, rappelons-le, nous avons montré que $\Phi : \mathbf{R} \longrightarrow [0,1]$ est bijective. Avec ce choix, nous donc montré que :

$$\lim_{n\to+\infty} \mathbf{P}\left(-\Phi^{-1}\left(1-\frac{\alpha}{2}\right) \leq \sqrt{n}\frac{\overline{X_n}-\mu}{\sigma_n} \leq \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right) = 1-\alpha,$$

$$\lim_{n\to+\infty} \mathbf{P}\left(-\frac{\sigma_n}{\sqrt{n}}\Phi^{-1}\left(1-\frac{\alpha}{2}\right) \leq \mu-\overline{X_n} \leq \frac{\sigma_n}{\sqrt{n}}\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right) = 1-\alpha.$$

manipulation d'encadrement

Il reste ensuite à ajouter $\overline{X_n}$ de chaque côté de l'encadrement, on obtient alors la proposition suivante.

Proposition ALEA.17.21 | IC asymptotique pour la moyenne donné par le théorème central limite

Soit (X_n) une suite de variables aléatoires indépendantes et de même loi, on note μ leur espérance commune, et admettant une variance et soient $\alpha \in]0,1[$, $n \in \mathbf{N}^\star$. Alors :

$$\mathbf{P}\left(\overline{X_n} - \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\frac{\sigma_n}{\sqrt{n}} \leq \mu \leq \overline{X_n} + \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\frac{\sigma_n}{\sqrt{n}}\right) \xrightarrow{n\to\infty} 1-\alpha.$$

Ainsi,

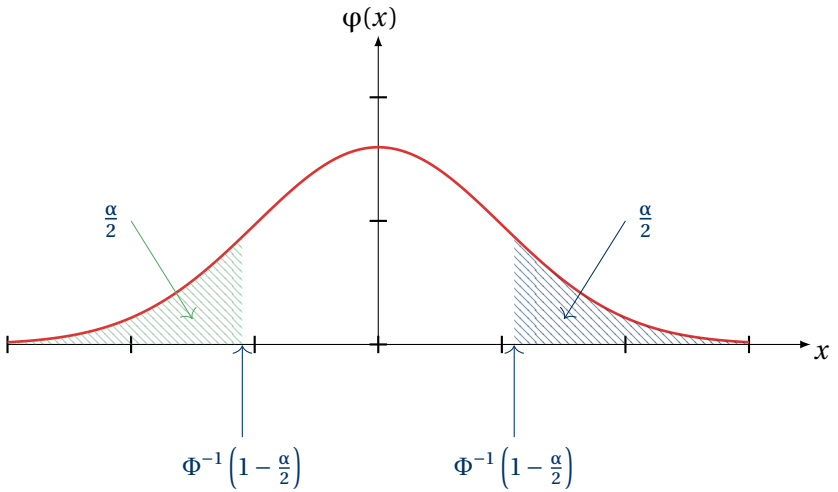
$$\left[\overline{X_n} - \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\frac{\sigma_n}{\sqrt{n}} \leq \mu \leq \overline{X_n} + \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\frac{\sigma_n}{\sqrt{n}}\right]$$

est un intervalle de confiance asymptotique de niveau α pour l'espérance μ.

Attention

Il faut savoir refaire la démarche qui précède l'énoncé de cette proposition.

Remarque 3.3 — Visualisation de $\Phi^{-1}\left(1-\frac{\alpha}{2}\right)$ Représentons sur la densité Gaußienne la quantité $\Phi^{-1}\left(1-\frac{\alpha}{2}\right)$.



Pour connaître les valeurs de $\Phi^{-1}\left(1-\frac{\alpha}{2}\right)$ on utilisera la table de la loi normale (voir le tableau en fin de ce chapitre). On peut cependant garder à l'esprit les valeurs remarquables suivantes que l'on utilisera la plupart du temps :

- $\Phi^{-1}\left(1-\frac{0.05}{2}\right) = 1.96$ donc pour un risque $\alpha = 0.05$,
- $\Phi^{-1}\left(1-\frac{0.01}{2}\right) = 2.57$ donc pour un risque $\alpha = 0.01$.

Remarque 3.4 — Commentaires

1. Comment varie la taille de l'intervalle de confiance en fonction de n ?

2. Comment simplifier cet intervalle de confiance lorsque l'écart-type σ est supposé connu ?

3. Pourquoi utiliser plutôt la seconde forme du théorème central limite ?

Résumons les techniques pour obtenir un intervalle de confiance.

Méthode Résultats probabilistes pour établir un intervalle de confiance : le théorème central limite et l'inégalité de Bienaymé-Tchebychev

1. **(Si le « type de loi » (bernoulli, gaußienne *etc.*) du n-échantillon est connu)**
   On arrive parfois à calculer les probabilités $\mathbf{P}\left(\theta \in I_{X_1,\dots,X_n}\right)$ explicitement pour n'importe quel intervalle $I_{X_1,\dots,X_n}$, les intervalles de confiance obtenus ne sont alors **pas asymptotiques**.
   Pour des échantillons gaußiens, on a deux cas de figure :
   - si σ est connue, on centre/réduit la moyenne empirique et on utilise la propriété de stabilité de la loi normale (*cf.* exemple de Cruella).
   - [H.P] Si σ est inconnue, on peut avoir recours à la loi de Student[14] : voir Remarque 6 ci-après pour une définition.
2. **(Si la loi de départ n'est pas connue)** [15] On utilise soit :
   - le théorème central limite en centrant réduisant la moyenne empirique (en approchant la variance par la version empirique si elle n'est pas connue), cela nous donne un intervalle de confiance seulement asymptotique,
   - soit l'inégalité de Bienaymé-Tchebychev.

Exemple 15 — Afin d'étudier la proportion p d'élèves satisfaits par le nouveau carambar bi-goût, on en interroge 100, et 56 d'entre eux déclarent être satisfaits par ce nouveau modèle. Donner un intervalle de confiance à 95 % pour p.

Exemple 16 — Sur 250 ampoules, on observe une durée de vie moyenne de 600 heures et un écart-type de 50 heures. Donner un intervalle de confiance à 99 % de l'espérance de vie d'une ampoule.

Remarque 3.5 — Différence avec les intervalles de « fluctuation ». On se donne par exemple une variable aléatoire $X \hookrightarrow \mathcal{N}(m, \sigma^2)$. Alors en utilisant la propriété de stabilité de la loi normale , on a :

$$\mathbf{P}(m - 1,96\sigma \leq X \leq m + 1,96\sigma) = 0,95.$$

L'intervalle $[m-1,96\sigma, m+1,96\sigma]$ est appelé « intervalle de *fluctuation* » pour X, car X

[14] C'est la loi obtenue en remplaçant σ inconnue par σ_n^{cor} dans la centrée/réduite de la moyenne empirique d'un échantillon gaußien

[15] On sera dans ce contexte l'immense majorité du temps

prend 95 % de ses valeurs dans cet intervalle, dont les bornes sont des nombres réels (dépendant des paramètres de la loi de X). Au contraire, un intervalle de *confiance* contient une valeur réelle mais inconnue, et ce sont ses bornes qui sont des variables aléatoires. Nous n'estimons pas la même chose dans les deux cas, mais on passe de l'un à l'autre par de simples manipulations sur les encadrements.

3.3. Test de conformité à la moyenne

Comme nous venons de le voir, l'une des fonctions des statistiques est de proposer, à partir d'observations d'un phénomène aléatoire (ou modélisé comme tel) une estimation de la loi de ce phénomène (ou plus précisément des paramètres associés). C'est ce que nous avons fait en construisant des intervalles de confiance. Les statistiques servent aussi à prendre des décisions. Peut-on considérer qu'un médicament est plus efficace qu'un placebo ? Le nombre de consultations de Google par seconde suit-il une loi de Poisson ? Les gènes pilotant la couleur des yeux et celle des cheveux sont-ils sur les mêmes chromosomes ? Il y a deux points communs (au moins) à toutes ces questions : leurs réponses sont des oui-non et le phénomène sous-jacent est aléatoire. Les tests statistiques vont permettre d'apporter une réponse à des questions manichéennes en contrôlant l'aléa inhérent à la situation.
En statistique les deux éventualités sont appelées des hypothèses et sont notées $\mathscr{H}_0$ (hypothèse nulle) et $\mathscr{H}_1$ (hypothèse alternative : « $\mu \neq \mu_0$ » dans l'exemple qui suit).

Commençons directement par le test de conformité à la moyenne avant de présenter le vocabulaire général des tests statistiques.

Principe du test d'adéquation à la moyenne. Le seul test qui est au programme de BCPST est le *test d'adéquation à la moyenne*. On considère un n-échantillon $(X_1, \dots, X_n)$ possédant une variance $\sigma^2 > 0$, et une espérance μ. Notons $\mathscr{H}_0$ « $\mu = \mu_0$ » : c'est l'hypothèse que la moyenne commune des X_i est $\mu = \mu_0$ avec $\mu_0 \in \mathbf{R}$. Si $\mathscr{H}_0$ est vraie, alors nous avons vu le résultat suivant.

Proposition ALEA.17.22 | Application du théorème central limite pour trouver une zone de rejet

Si $\mathscr{H}_0$ est vraie, alors pour tout $\alpha \in \mathbf{R}$, $\mathbf{P}\left(\left|\frac{\overline{X_n}-\mu_0}{\frac{\sigma_n}{\sqrt{n}}}\right| > \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right) \xrightarrow{n\to\infty} \alpha$.

Preuve

Ainsi, en se fondant sur ce résultat probabiliste, une stratégie de décision pour accepter ou rejeter l'hypothèse $\mathscr{H}_0$ serait la suivante.

Définition/Proposition ALEA.17.3 | Test d'adéquation à la moyenne

1. si pour $n \geqslant 30$,[16] $\mu_0 \in \left[\overline{X_n} - \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\frac{\sigma_n}{\sqrt{n}}, \overline{X_n} + \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\frac{\sigma_n}{\sqrt{n}}\right]$, alors on ne rejète pas l'hypothèse $\mathscr{H}_0$ avec risque d'erreur[17] α,
2. si pour $n \geqslant 30$, $\mu_0 \notin \left[\overline{X_n} - \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\frac{\sigma_n}{\sqrt{n}}, \overline{X_n} + \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\frac{\sigma_n}{\sqrt{n}}\right]$, alors on rejette l'hypothèse.

On appelle ce test le *test d'adéquation à la moyenne*.

[16] Un consensus pour que la convergence dans le théorème central limite soit suffisamment précise

Fixons un peu de vocabulaire :

- La variable aléatoire $\widehat{\theta_n} = \frac{\overline{X_n}-\mu_0}{\frac{\sigma_n}{\sqrt{n}}}$ est appelée *statistique de test.*
- L'intervalle
$$\left]-\infty, -\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right[\bigcup \left]\Phi^{-1}\left(1-\frac{\alpha}{2}\right), +\infty\right[$$
est appelé *la zone de rejet* ; car c'est lorsque la statistique de test est dans cette zone de rejet que l'on rejette $\mathscr{H}_0$.
- Le risque d'erreur (de première espèce) est la probabilité de rejeter $\mathscr{H}_0$ alors qu'elle est vraie, c'est ce risque que l'on souhaite le plus petit possible. Dans notre test d'adéquation, il s'agit d'α qui est petit.
- L'hypothèse $\mathscr{H}_1$ qui est ici $\mu \neq \mu_0$ est appelée *hypothèse alternative.* Dans nos exemples, l'hypothèse alternative $\mathscr{H}_1$ sera toujours ${}^{\complement}\mathscr{H}_0$ — l'hypothèse contraire de $\mathscr{H}_0$.

Résumé du test d'adéquation

1. **(But)** Tester une valeur possible de moyenne $\mathscr{H}_0$ « $\mu = \mu_0$ ».
2. **(Résultat probabiliste qui fonde le test)** le théorème central limite.
3. **(Décision)** On rejette $\mathscr{H}_0$ si $\mu_0 \notin \left[\overline{X_n} - \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\frac{\sigma_n}{\sqrt{n}}, \overline{X_n} + \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\frac{\sigma_n}{\sqrt{n}}\right]$. Sinon on ne rejète pas $\mathscr{H}_0$.

Remarque 3.6 — Que faire quand $n < 30$? Une solution : la loi de Student lorsque l'échantillon est gaußien Dans le cas où $n < 30$[18] **et** que l'échantillon est gaußien, on peut utiliser une version du test précédent faisant appel à la *loi de* Student.

Plus précisément, on appelle loi de Student à $k \in \mathbf{N}^\star$ degrés de liberté la loi d'une variable aléatoire $\mathscr{T}_k$ définie par :
$$\mathscr{T}_k = \frac{N}{\sqrt{\left(N_1^2 + \ldots + N_k^2\right)/k}},$$
où $(N_1, \ldots, N_k)$ est un k-échantillon de loi $\mathscr{N}(0,1)$, et N également de loi $\mathscr{N}(0,1)$ indépendante des N_i, $i \in [\![1\,,\,k]\!]$.

Soit donc $(X_1, \ldots, X_n)$ notre échantillon de moyenne commune μ. Alors la centrée-réduite de $\overline{X_n}$ où l'écart-type σ est remplacé par sa version empirique corrigée est :
$$\sqrt{n}\frac{\overline{X_n} - \mu}{\sigma_n^{\text{cor}}}.$$

Alors, un théorème non trivial[19] permet de montrer que cette variable aléatoire suit une loi $\mathscr{T}_{n-1}$. Cette loi étant tabulée, on peut alors en déduire facilement des intervalles de confiance comme nous l'avions fait avec le théorème central limite.

Généralités sur les tests statistiques.

Définition ALEA.17.27 | Test statistique, hypothèse nulle, zone de rejet

1. Une *hypothèse statistique* est un énoncé concernant un n-échantillon (valeur d'un paramètre, nature de la distribution, *etc.*).
2. Un *test statistique* est une démarche ayant pour but de fournir une règle de décision permettant, en se fondant sur l'observation d'un échantillon, de faire un choix entre deux hypothèses statistiques. L'*hypothèse nulle* est l'hypothèse fixant *a priori* une condition sur le paramètre, on la note $\mathscr{H}_0$. Toute autre hypothèse est appelée hypothèse alternative, on la note $\mathscr{H}_1$.
3. On appelle *région de rejet* (*resp. statistique de test*) une partie I de $\mathbf{R}$ (*resp.* un estimateur $\left(\widehat{\theta_n}\right)$ associé au n-échantillon considéré) qui conduit à rejeter $\mathscr{H}_0$ pour $\mathscr{H}_1$ si $\widehat{\theta_n} \in I$. Dans le cas contraire, on dit que l'on *ne rejète pas* $\mathscr{H}_0$.
4. On appelle *risque de première espèce* (ou *niveau du test*) la probabilité de rejeter $\mathscr{H}_0$ à tort (alors qu'elle est vraie).

[17] *i.e.* rejeter à tort l'hypothèse $\mathscr{H}_0$ alors qu'elle était vraie
[18] et même pour tout n dans ce cas
[19] théorème de Cochran

Définition ALEA.17.28 | Test statistique et niveau

Un test statistique est un algorithme qui conduit à accepter $\mathcal{H}_0$ ou à rejeter $\mathcal{H}_0$ à partir d'observations d'un phénomène aléatoire. On appelle *niveau du test* la probabilité de rejeter $\mathcal{H}_0$ alors qu'elle est vraie.

C'est le niveau du test que l'on souhaite le plus faible possible. En revanche, on se préoccupe en général peu de la probabilité d'accepter $\mathcal{H}_0$ alors que $\mathcal{H}_1$ est vraie. L'objectif d'un test est avant tout de valider l'hypothèse $\mathcal{H}_0$. Ne pas rejeter $\mathcal{H}_0$ veut simplement dire que les observations ne sont pas incompatibles avec cette hypothèse.

Remarque 3.7 — Dissymétrie des hypothèses Retenez l'analogie avec la justice qui pose comme principe la présomption d'innocence. On souhaite contrôler en priorité la probabilité d'envoyer un innocent en prison en négligeant pour l'instant celle de relâcher un coupable. Dans cet exemple $\mathcal{H}_0$ est «la personne est innocente» et $\mathcal{H}_1$ est «la personne est couple». Le risque de première espèce correspond donc au rejet de $\mathcal{H}_0$ (personne envoyée en prison) alors qu'elle est vraie (personne innocente).

Plutôt que de dire «on ne rejète pas l'hypothèse», on devrait dire «on ne rejète pas l'hypothèse avec un risque α de se tromper» (*i.e.* d'accepter l'hypothèse alors qu'elle est fausse). De-même, plutôt que de dire «on rejette l'hypothèse», on devrait dire «on rejette l'hypothèse avec un risque $1-\alpha$ de se tromper» (*i.e.* de rejeter l'hypothèse alors qu'elle est vraie).

Méthode Démarche générale d'un test statistique

1. Poser l'hypothèse (nulle) $\mathcal{H}_0$ que l'on souhaite tester.
2. Trouver un résultat de probabilité qui donne deux résultats différents selon que $\mathcal{H}_0$ est vraie ou non (dans le test d'adéquation *supra*, c'était le théorème central limite).
3. Donner la stratégie de décision, en fonction du résultat énoncé.

Remarque 3.8 — Il est illusoire, à cause de l'aléatoire sous-jacent au n-échantillon, de vouloir prendre à coup sûr la bonne décision. C'est pourquoi on se laisse une marge d'erreur. En général, on choisit $\alpha = 0,05$ ou $\alpha = 0,01$.

Exemple 17 — Chez le petit lapin, la durée moyenne de gestation est de 30 jours. On étudie un échantillon de 66 familles de gros lapins, pour lesquelles on observe une durée moyenne de gestation de 30,83 jours avec un écart-type de 4,07 jours. Peut-on conclure que la durée de gestation est significativement différente chez les petits et les gros lapins?

★★★ **Fin du chapitre** ★★★

Exemple 18 — Une étude commerciale, réalisée sur 100 personnes, montre que 49 % des internautes ont moins de 39 ans, alors que d'après le recensement cette tranche d'âge représente 41 % de la population française. Peut-on conclure que la population des internautes est plus jeune que la population française? La différence observée est-elle révélatrice d'un phénomène ou provient-elle des fluctuations d'échantillonnage? On pourra considérer le cas d'un niveau de confiance de 90 % puis de 95 %.

Soit $(X_1,\dots,X_{100})$ *un* 100*-échantillon, de loi* $\mathscr{B}(p)$ *avec* $p\in[0,1]$*, c'est pour un individu donné la probabilité qu'il ait moins de 39 ans. On sait par hypothèse que* $\overline{X_{100}}=\frac{49}{100}$*, puis par formule de* König-Huygens *que* $\sigma_{100}^2=\frac{1}{100}\sum_{i=1}^{100}X_i^2-\overline{X_{100}}^2=\overline{X_{100}}-\overline{X_{100}}^2=0.2499$.
Ainsi :
— un intervalle de confiance de niveau 95 *% pour* p *est :*

$$\left[0,49-1,96\times\frac{\sqrt{0.2499}}{\sqrt{100}};0,49+1,96\times\frac{\sqrt{0.2499}}{\sqrt{100}}\right]\approx[0.392,0.588]\,,$$

— un intervalle de confiance de niveau 90 *% pour* p *est :*

$$\left[0,49-1.64\times\frac{\sqrt{0.2499}}{\sqrt{100}};0,49+1.64\times\frac{\sqrt{0.2499}}{\sqrt{100}}\right]\approx[0.408,0.572]\,.$$

On constate que : $0,41$ *est dans les deux intervalles de confiance. On accepte donc l'hypothèse* $\mathscr{H}_0$ *«* $p=\frac{41}{100}$ *», i.e. que parmi les internautes et l'ensemble de la population la fréquence de personnes de moins de 39 ans est sensiblement la même. La petite différence observée provient de fluctuations d'échantillonnage, avec une étude commerciale menée sur plus de personnes nous aurions peut-être eu un résultat différent.*

Annexe : tables de valeurs pour la fonction de répartition de la loi normale centrée réduite $\mathcal{N}(0,1)$.

Méthode Obtenir $\Phi(x)$ pour un certain $x \in \mathbf{R}$ à l'aide d'une table

Si l'on souhaite avoir, par exemple, $\Phi(0,96)$, on :

1. se place sur la ligne «0.9»,
2. se place ensuite sur la colonne «0.06».
3. On obtient alors la valeur désirée. Dans cet exemple, $\Phi(0,96) = 0,8315$.

Méthode Chercher $x \in \mathbf{R}$ tel que $\Phi(x) = \alpha$ à l'aide d'une table, $\alpha \in [0,1]$

Si l'on souhaite avoir, par exemple, $x \in \mathbf{R}$ tel que $\Phi(x) = 0.975$.

1. On cherche dans la grille l'endroit où se trouve une valeur suffisamment proche de $\alpha = 0.975$.
2. Dans cet exemple, on constate que $\Phi(1.96) = 0.975$.

Si l'on souhaite avoir, par exemple, $x \in \mathbf{R}$ tel que $\Phi(x) = 0.160$.

1. En parcourant la table, on constate que 0.160 n'y apparaît pas.
2. On reformule alors la condition en passant au complémentaire :

$$1 - \Phi(x) = 0,84 = \Phi(-x).$$

3. On cherche donc dans la table 0.84, on trouve alors

$$0.84 = \Phi(1.00) \quad \text{donc} \quad -x = 1.00, \quad x = -1.00.$$

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.090
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224

0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

$$\Phi(x) = \mathbf{P}(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \quad \text{et} \quad \Phi(-x) = 1 - \Phi(x).$$

On retiendra en particulier la valeur typique $\Phi(1.96) = 0.975$, de sorte que :

$$\mathbf{P}(X \in [-1.96; 1.96]) = 2\Phi(1.96) - 1 = 0.95$$

★★★ **Fin du chapitre** ★★★

4. EXERCICES

4.1. Descriptives

Exercice ALEA.17.1 | Un médecin effectue des recherches sur l'efficacité d'un nouveau béta-bloquant. Cette famille de médicaments est destinée à diminuer le rythme cardiaque des malades atteints de tachycardie (pouls supérieur à 100 battements par minute au repos). Il a donc séparé les malades en 2 groupes : le groupe A reçoit le traitement d'un nouveau médicament, le groupe B reçoit un placébo. Voici les résultats.

- A : 74 - 91 - 91 - 84 - 95 - 93 - 95 - 102 - 81 - 116 - 88 - 95,
- B : 94 - 95 - 113 - 95 - 104 - 113 - 94 - 144 - 105 - 153.

1. Calculer l'étendue et la médiane pour chacune de ces deux séries.
2. Construire le diagramme de Tuckey de ces deux séries.
3. L'effet du médicament semble-t-il satisfaisant ?

Exercice ALEA.17.2 | L'indice moyen d'un salaire a évolué de la façon suivante :

Année	1	2	3	4	5	6	7
Indice	165	176	193	202	222	245	253

1. Représenter cette série statistique par un nuage de points.
2. Déterminer la droite de régression linéaire de l'indice en fonction de l'année.
3. Prévoir l'indice à l'année 9.

Exercice ALEA.17.3 | **Ajustement d'ordre deux – Extrait Agro—Véto 2019** Dans le cas de la régression linéaire on cherche à approcher un nuage de points à l'aide d'une droite, donc un polynôme de degré 1. Dans cet exercice, on essaie de généraliser à un polynôme de degré 2.

Soient $n \in \mathbf{N}^\star$, et $x = (x_1, \dots, x_n)$, $y = (y_1, \dots, y_n)$ deux éléments de $\mathbf{R}^n$, on suppose que les x_i sont deux à deux distincts. Soit de plus

$$\mathrm{F} \left| \begin{array}{ccc} \mathbf{R}^3 & \longrightarrow & \mathbf{R}, \\ (a,b,c) & \longmapsto & \sum_{i=1}^n \left(y_i - a - bx_i - cx_i^2\right)^2. \end{array} \right.$$

1. Expliquer ce que représente géométriquement $\mathrm{F}(a,b,c)$ pour tout $(a,b,c) \in \mathbf{R}^3$. On pourra placer sur le graphique un nuage de points $(x_i, y_i)_{1 \leqslant i \leqslant n}$ pour n petit.
2. Justifier que F admet des dérivées partielles dans toutes les directions, et calculer $\operatorname{grad} \mathrm{F}(a,b,c)$ pour tout $(a,b,c) \in \mathbf{R}^3$.
3. On note ici $\mathrm{Y} = {}^{\mathrm{T}} y$ et $\beta = \begin{pmatrix} a \\ b \\ c \end{pmatrix}$ avec $(a,b,c) \in \mathbf{R}^3$.

   3.1) Chercher une matrice $\mathrm{T} \in \mathfrak{M}_{n,3}(\mathbf{R})$ telle que :

$$\begin{cases} \frac{\partial \mathrm{F}}{\partial a}(a,b,c) &= -2 \left\langle \mathrm{Y} - \mathrm{T}\beta \middle| \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \right\rangle, \\ \frac{\partial \mathrm{F}}{\partial b}(a,b,c) &= -2 \left\langle \mathrm{Y} - \mathrm{T}\beta \middle| \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \right\rangle, \\ \frac{\partial \mathrm{F}}{\partial c}(a,b,c) &= -2 \left\langle \mathrm{Y} - \mathrm{T}\beta \middle| \begin{pmatrix} x_1^2 \\ \vdots \\ x_n^2 \end{pmatrix} \right\rangle, \end{cases}$$

   où $\langle . | . \rangle$ désigne le produit scalaire euclidien sur $\mathfrak{M}_{n,1}(\mathbf{R})$.

   3.2) En écrivant matriciellement les produits scalaires précédents, déduire que tout point critique $\left(\widehat{a}, \widehat{b}, \widehat{c}\right) \in \mathbf{R}^3$ de F vérifie :

$$\left({}^{\mathrm{T}}\mathrm{TT}\right)\widehat{\beta} = {}^{\mathrm{T}}\mathrm{TY}, \quad \text{avec :} \quad \widehat{\beta} = \begin{pmatrix} \widehat{a} \\ \widehat{b} \\ \widehat{c} \end{pmatrix}.$$

4. On souhaite ensuite trouver une condition pour que la matrice ${}^{\mathrm{T}}\mathrm{TT}$ soit inversible.

4.1) Préciser le format de ${}^{\mathrm{T}}\mathrm{TT}$.
4.2) Montrer que $\mathrm{Ker}(\mathrm{T}) = \mathrm{Ker}\left({}^{\mathrm{T}}\mathrm{TT}\right)$.
4.3) Que vaut $\dim \mathrm{Ker}\left({}^{\mathrm{T}}\mathrm{T.T}\right) + \dim \mathrm{Rg}\left({}^{\mathrm{T}}\mathrm{T.T}\right)$? Justifier.
4.4) Que vaut $\dim \mathrm{Ker}\,(\mathrm{T}) + \dim \mathrm{Rg}\,(\mathrm{T})$? Justifier.
4.5) En déduire que ${}^{\mathrm{T}}\mathrm{T.T}$ est inversible si et seulement si T est de rang 3.

5. Dans ce cas-là, donner une expression de $\widehat{\beta}$.

4.2. Estimateurs

Exercice ALEA.17.4 | Barycentre de deux estimateurs

1. Dans une population de porcs, on veut estimer le gain moyen quotidien (GMQ) noté μ, on suppose que les gains sont d'écart-type $\sigma \in \mathbf{R}$. À cet effet, on choisit deux échantillons indépendants dans cette population. On observe deux échantillons : l'un $(X_1, \dots, X_{10})$ de 10 individus, et $(Y_1, \dots Y_{30})$ de 30 individus. On propose deux estimateurs de μ :

$$T_1 = \frac{\overline{X_{10}} + \overline{Y_{30}}}{2} \text{ et } T_2 = \frac{10\overline{X_{10}} + 30\overline{Y_{30}}}{40}.$$

On cherche à déterminer le meilleur des deux estimateurs.

1.1) Calculer le biais de chaque estimateur pour le paramètre μ. Cela permet-il de les départager ?
1.2) Calculer les variances de T_1 et T_2 en fonction de la variance σ^2 du gain quotidien de la population. Conclure.

2. (Généralisation) Soient T_1 et T_2 deux estimateurs de $\mu \in \mathbf{R}$, sans biais et indépendants. Pour tout $a \in \mathbf{R}$, on pose $\Theta_a = aT_1 + (1-a)T_2$.
2.1) Soit $a \in \mathbf{R}$. Calculer le biais de Θ_a pour le paramètre μ.
2.2) Parmi tous les Θ_a, $a \in \mathbf{R}$, lequel a le plus petit risque quadratique ? Est-ce cohérent avec la première question ?

Exercice ALEA.17.5 | Soit $(X_1, \dots, X_n)$ un n-échantillon de loi de Bernoulli de paramètre $p \in]0,1[$, $n \geqslant 1$. On pose $S_n = \sum_{k=1}^{n} X_k$ ainsi que

$$\overline{X_n} = \frac{S_n}{n}, \quad T_n = \frac{S_n + 1}{n+2}.$$

Comparer les biais, ainsi que les risques quadratiques de $\overline{X_n}$ et T_n en tant qu'estimateurs de p. Peut-on privilégier l'un de ces estimateurs par rapport à l'autre ?

Exercice ALEA.17.6 | Soit $(X_1, \dots, X_n)$ un n-échantillon, $n \geqslant 2$, d'une loi $\mathscr{P}(\lambda)$ de paramètre $\lambda > 0$ inconnu. On souhaite estimer le paramètre $\theta = e^{-\lambda}$. Pour $k \in [\![1\,,\,k]\!]$, on pose $Y_k = \mathbb{1}_{\{X_k = 0\}}$, et on introduit $\overline{Y_n} = \frac{1}{n}\sum_{k=1}^{n} Y_k$, $S_n = \sum_{k=1}^{n} X_k$.

1. 1.1) Montrer que $\overline{Y_n}$ est un estimateur sans biais de θ.
1.2) Calculer $\mathbf{Var}\left(\overline{Y_n}\right)$. Que dire de $\mathbf{P}\left(\left|\overline{Y_n} - \theta\right| \geqslant \varepsilon\right)$ pour tout $\varepsilon > 0$? *On dit que* $\overline{Y_n}$ *est un estimateur convergeant vers* θ.
2. Pour $j \in \mathbf{N}$, calculer $\varphi(j) = \mathbf{P}(X_1 = 0 \,|\, S_n = j)$.
3. 3.1) Montrer que $T_n = \varphi(S_n)$ est un estimateur sans biais de θ.
3.2) Calculer $\mathbf{Var}\,(T_n)$ et en déduire que T_n converge vers θ.
4. Lequel des deux vous semble plus efficace ?

Exercice ALEA.17.7 | Estimer l'amplitude d'une uniforme Soit $(X_1, \dots, X_n)$ un $n \in \mathbf{N}^\star$-échantillon de loi $\mathscr{U}[a,b]$ avec $(a,b) \in \mathbf{R}^2$. On souhaite estimer $b-a$.

1. Qu'estime de manière non biaisée $2\overline{X_n}$?
2. Déterminer une densité de $\max(X_1, \dots, X_n)$ et $\min(X_1, \dots, X_n)$, puis calculer leur espérance.
3. Déterminer quel paramètre estime $\widehat{\theta_n} = \max(X_1, \dots, X_n) - \min(X_1, \dots, X_n)$ de manière non biaisée. Commentez.

Exercice ALEA.17.8 | Loi exponentielle généralisée Soient $a \in \mathbf{R}$, $b \in]0, +\infty[$, et

$$f \left| \begin{array}{ccc} \mathbf{R} & \longrightarrow & \mathbf{R} \\ x & \longmapsto & \left\{ \begin{array}{ll} 0 & \text{si } x \leqslant a, \\ \frac{1}{b} e^{-\frac{x-a}{b}} & \text{sinon.} \end{array} \right. \end{array} \right.$$

1. Vérifier que f est une densité de probabilité.
2. Soit X une variable aléatoire de densité f, on note $X \hookrightarrow \mathscr{E}(a,b)$.
   - **2.1)** Déterminer la loi de $\frac{X-a}{b}$.
   - **2.2)** En déduire l'espérance et la variance de X.
3. Soit $(X_1,\dots,X_n)$ un n-échantillon de X. On pose $Y_n = \min(X_1,\dots,X_n)$.
   - **3.1)** Montrer que $Y_n \hookrightarrow \mathscr{E}\left(a,\frac{b}{n}\right)$.
   - **3.2)** Montrer que Y_n est un estimateur convergeant vers a.
4. On pose $Z_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - Y_n)$, $\quad U_n = \sum_{i=1}^{n} X_i$.
   - **4.1)** Calculer $\mathbf{E}(Z_n)$, $\mathbf{Var}(Z_n)$, en fonction de $\mathbf{Cov}(U_n,Y_n)$, b et $n \in \mathbf{N}$ pour tout entier n.
   - **4.2)** En utilisant l'inégalité $\left|\rho(U_n,Y_n)\right| \leqslant 1$, montrer que : $\lim\limits_{n\longrightarrow\infty} \mathbf{Var}(Z_n) = 0$.
   - **4.3)** En déduire que Z_n est un estimateur convergeant de b.

Exercice ALEA.17.9 | **Améliorons les résultats du baccalauréat** Lors de l'examen national du baccalauréat, $N \in \mathbf{N}^\star$ candidats obtiennent des moyennes générales entre 0 et 20. Des tests statistiques d'adéquation ont montré que ces moyennes se répartissent selon une $\mathcal{N}(12,2^2)$. On considère donc dans la suite $X_1,\dots,X_N$ une suite de variables aléatoires réelles indépendantes de même loi $\mathcal{N}(12,2^2)$. Le candidat $i \in [\![1,N]\!]$ est déclaré admis (on note A_i l'évènement associé) si :

1. il a obtenu une moyenne supérieure ou égale à 10,
2. il a obtenu une moyenne entre 8 et 10, et les oraux de rattrapage lui ont permis d'atteindre 10. On suppose que ceci se produit avec une probabilité $\alpha \in]0,1[$.

On note dans tout l'exercice Φ la fonction de répartition de la loi $\mathcal{N}(0,1)$, $Y_N = \sum_{i=1}^{N} \mathbb{1}_{A_i}$ et P_N la variable aléatoire donnant la proportion de candidats admis. L'objectif du ministre de l'Éducation Nationale est d'obtenir un taux de réussite de 90 % — légèrement supérieur à l'année précédente — et souhaiterait ainsi connaitre le paramètre α lui permettant d'arriver à ses fins.

1. Soit $i \in [\![1,N]\!]$. Montrer que :

$$\mathbf{P}(A_i) = 1 - \Phi(-1) + \alpha(\Phi(-1) - \Phi(-2)).$$

On notera $\beta_\alpha = 1 - \Phi(-1) + \alpha(\Phi(-1) - \Phi(-2))$ dans la suite.

2. Donner le lien entre P_N et Y_N. Que représente Y_N et quelle est sa loi ?
3. On suppose N très grand. Par quelle loi normale peut-on approcher la loi de Y_N ? Justifier.
4. Montrer alors que : $\quad \mathbf{P}(P_N \geqslant 0.9) \approx 1 - \Phi\left(\frac{\sqrt{N}(0,9-\beta_\alpha)}{\sqrt{\beta_\alpha(1-\beta_\alpha)}}\right)$.
5. Tracer la fonction $\alpha \in [0,001,0,999] \longmapsto 1 - \Phi\left(\frac{\sqrt{N}(0,9-\beta_\alpha)}{\sqrt{\beta_\alpha(1-\beta_\alpha)}}\right)$ à l'aide de Python par exemple pour $N = 1000$. *Indication : On pourra faire appel à la commande* `norm.cdf` *pour obtenir* Φ, *après avoir réalisé l'import* `from scipy.stats import norm`. Commenter.
6. Quel algorithme vous permettrait d'obtenir une valeur approchée de α, à 10^{-3} près, qui permettrait d'atteindre une probabilité de 95 % pour l'évènement $\{P_N \geqslant 0.9\}$? Le mettre en place, et donner une valeur approchée de α pour $N = 1000$ candidats.
7. À la session qui suit, le nouveau ministre de l'Éducation Nationale Jean-Michel B. (qui souhaite garder son anonymat) décide de faire comme son prédécesseur et d'obtenir un taux de réussite supérieur à 90%. Il a à sa disposition un N-échantillon $(X_1,\dots,X_N)$ des notes obtenues à la session précédente (où le taux de réussite était supérieur à 90 %) mais ne connaît pas *a priori* la valeur de α ayant permis une telle prouesse. Proposer une démarche statistique, à l'aide d'un estimateur classique, permettant d'obtenir une estimation de α. On supposera à nouveau N grand.

Solution (exercice ALEA.17.9)

1. Le fonctionnement de l'examen invite ici à conditionner. Nous avons, pour $i \in [\![1\,,\,N]\!]$,

$$\begin{aligned}\mathbf{P}(A_i) &= \mathbf{P}(A_i|X_i > 10)\,\mathbf{P}(X_i > 10) + \mathbf{P}(A_i|8 \leqslant X_i \leqslant 10)\,\mathbf{P}(8 \leqslant X_i \leqslant 10),\\ &= \mathbf{P}(A_i|X_i > 10)\,\mathbf{P}(X_i > 10) + \mathbf{P}(A_i|8 \leqslant X_i \leqslant 10)\,\mathbf{P}(8 \leqslant X_i \leqslant 10),\\ &= 1.\mathbf{P}(X_i > 10) + \alpha\mathbf{P}(8 \leqslant X_i \leqslant 10).\end{aligned}$$

Or, $\quad X_i^\star \underset{\textbf{(défi.)}}{=} \frac{X_i-12}{2} \hookrightarrow \mathcal{N}(0,1)$ par propriété de stabilité de la loi normale, donc

$$\begin{aligned}&= \mathbf{P}(X_i^\star > -1) + \alpha\mathbf{P}(-2 \leqslant X_i^\star \leqslant -1)\\ &= \boxed{1 - \Phi(-1) + \alpha(\Phi(-1) - \Phi(-2)).}\end{aligned}$$

2. L'interprétation de Y_N est la suivante : cette variable aléatoire correspond au nombre d'admis au baccalauréat sur les N candidat(e)s. Or, P_N est par définition la proportion, donc $\boxed{P_N = \frac{Y_N}{N}}$. De l'interprétation qui précède, nous avons également qu'Y_N est le nombre de succès dans une succession d'expériences de Bernoulli (l'admission ou non d'un candidat) indépendantes (on suppose que la réussite d'un candidat n'influe par sur les autres). Donc : $\boxed{Y_N \hookrightarrow \mathscr{B}(N, 1-\Phi(-1)+\alpha(\Phi(-1)-\Phi(-2)))}$.

3. Puisque N est assez grand, le théorème de Moivre-Laplace d'approximation de la loi normale nous livre l'approximation suivante : la loi de $\boxed{Y_N \text{ est proche d'un } \mathscr{N}(N\beta_\alpha, N\beta_\alpha(1-\beta_\alpha))}$. Rappelons que ce théorème est bien applicable dans ce contexte puisque les variables aléatoires $\mathbb{1}_{A_i}$ sont indépendantes de même loi et possèdent une variance.

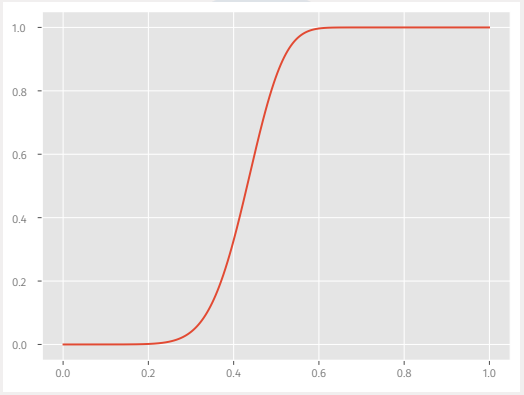
4.

$$\begin{aligned}\mathbf{P}(P_N \geqslant 0.9) &= \mathbf{P}(Y_N \geqslant N0.9)\\ &= \mathbf{P}\left(\frac{Y_N - N\beta_\alpha}{\sqrt{N\beta_\alpha(1-\beta_\alpha)}} \geqslant \frac{0.9N - N\beta_\alpha}{\sqrt{N\beta_\alpha(1-\beta_\alpha)}}\right)\\ &\approx 1-\Phi\left(\frac{0.9N - N\beta_\alpha}{\sqrt{N\beta_\alpha(1-\beta_\alpha)}}\right)\\ &= \boxed{1-\Phi\left(\frac{\sqrt{N}(0.9-\beta_\alpha)}{\sqrt{\beta(1-\beta_\alpha)}}\right)}.\end{aligned}$$

5. Passons au tracé de la fonction. On commence par définir en Python la fonction $\alpha \longmapsto \beta_\alpha$.

```
from scipy.stats import norm
import numpy as np
import matplotlib.pyplot as plt
def beta(alpha):
    return 1 + alpha*(norm.cdf(-1) - norm.cdf(-2)) -
    ↪ norm.cdf(-1)
N = 1000
X = np.linspace(0.001, 0.999, 10**3)
```

```
Y = 1 - norm.cdf(np.sqrt(N)*(0.9-beta(X))/np.sqrt(beta(X)*(1
↪ -beta(X))))
plt.plot(X, Y)
```



Nous constatons que la fonction est croissante. Ceci s'explique aisément : plus on laisse passer de candidats après l'oral de rattrapage, plus la proportion d'admis sera grande ! À l'aide de ce graphique nous pourrions trouver une valeur approchée de β_α permettant de réaliser la condition $\mathbf{P}(P_N \geqslant 0.9) \geqslant 0,95$. On en déduit alors une valeur approchée de α. Nous allons faire autrement dans la question suivante.

6. Il s'agit de résoudre l'équation suivante en β_α :

$$1-\Phi\left(\frac{\sqrt{N}(0.9-\beta_\alpha)}{\sqrt{\beta(1-\beta_\alpha)}}\right) = 0.95,$$

ou de manière équivalente

$$g(\alpha) = 0.05 - \Phi\left(\frac{\sqrt{N}(0.9-\beta_\alpha)}{\sqrt{\beta(1-\beta_\alpha)}}\right) = 0$$

Nous allons donc :

- créer dans Python la fonction g, apparaissant dans le membre de gauche,
- puis mettre en place un algorithme de Dichotomie pour obtenir la valeur désirée de β_α.
- Enfin on en déduira facilement une valeur approchée pour α.

```python
from scipy.stats import norm
import numpy as np
N = 1000
def beta(alpha):
    return 1 + alpha*(norm.cdf(-1) - norm.cdf(-2)) -
    ↪ norm.cdf(-1)
def g(alpha):
    return 0.05 - norm.cdf(np.sqrt(N)*(0.9
    ↪ -beta(alpha))/np.sqrt(beta(alpha)*(1-beta(alpha))))
def dicho(a, b, f, prec):
    """
    Retourne une valeur approchée d'un zéro de f entre a et b
 ↪  avec précision prec
    """
    while b-a > prec:
        c = (a+b)/2
        if f(a)*f(c) <= 0:
            b = c
        else:
            a = c
        return (a+b)/2
alpha_suffisant = dicho(0.1, 0.9, g, 10**(-3))
```

Voici ce que l'exécution nous donne : 0.7.

7. Il suffit de regarder les notes entre 8 et 10 des candidats, puis de calculer la fréquence notée f dans la suite de ceux ayant été admis. En remarque (non demandée), donnons une explication Mathématique. Constatons que :

- $\sum_{j=1}^{N} \mathbb{1}_{\{8\leqslant X_i\leqslant 10\}}$ est le nombre de candidats au rattrapage,
- $\sum_{i=1}^{N} \mathbb{1}_{A_i\cap\{8\leqslant X_i\leqslant 10\}}$ est le nombre de candidats admis au rattrapage.

La fréquence mentionnée précédemment est donc

$$f = \frac{1}{\sum_{j=1}^{N} \mathbb{1}_{\{8\leqslant X_i\leqslant 10\}}} \sum_{i=1}^{N} \mathbb{1}_{A_i}\mathbb{1}_{\{8\leqslant X_i\leqslant 10\}}$$
$$= \frac{N}{\sum_{j=1}^{N} \mathbb{1}_{\{8\leqslant X_i\leqslant 10\}}} \times \frac{\sum_{i=1}^{N} \mathbb{1}_{A_i\cap\{8\leqslant X_i\leqslant 10\}}}{N}.$$

Rappelons que l'espérance d'une indicatrice est égale à la probabilité de l'évènement. Ainsi, d'après la loi faible des grands nombres, le premier terme est «proche de» $\frac{1}{\mathbf{P}(8\leqslant X_i\leqslant 10)}$. De-même, le second est «proche de» $\mathbf{P}(A_i\cap 8\leqslant X_i\leqslant 10) = \mathbf{P}(A_i|8\leqslant X_i\leqslant 10)\,\mathbf{P}(8\leqslant X_i\leqslant 10) = \alpha\mathbf{P}(8\leqslant X_i\leqslant 10)$. Donc, *in fine*,

$$f \approx \frac{\alpha}{\mathbf{P}(8\leqslant X_i\leqslant 10)}\mathbf{P}(8\leqslant X_i\leqslant 10) = \boxed{\alpha}.$$

Seconde remarque : les manipulations du «proche de» nécessitent des connaissances sur la convergence en probabilités, qui dépassent très largement le cadre du programme.

4.3. Intervalles de confiance

Exercice ALEA.17.10 | Contrairement à ce que pense Popeye, l'épinard n'est pas l'aliment le plus riche en fer. La lentille, par exemple, en apporte davantage. On a procédé à des analyses sur 10 échantillons de lentilles (de loi parente X) et d'épinards (de loi parente Y), et on a relevé la teneur en fer (en mg pour 100 g de produit frais).

Echantillon	1	2	3	4	5	6	7	8	9	10
Epinard (~ Y)	2.64	2.75	2.82	2.72	2.56	2.59	2.83	2.70	2.67	2.62

Lentille (~ X)	9.02	9.08	8.82	8.94	8.95	9.11	9.14	9.02	9.04	8.85

1. Calculer la teneur moyenne en fer, l'écart-type associé, la médiane et les quartiles pour les épinards et les lentilles, sur les relevés statistiques donnés dans le tableau.
2. Déterminer un intervalle de confiance à 95 % pour la teneur moyenne en fer des épinards et des lentilles, *i.e.* pour l'espérance de X et Y.
3. Proposer une représentation graphique illustrant le propos initial.

Exercice ALEA.17.11 | Une usine fabrique des câbles. On suppose que la charge maximale supportée par un câble, exprimée en tonnes, est une variable aléatoire qui suit une loi normale $\mathcal{N}(\mu, 0,5^2)$. Une étude portant sur 50 câbles a donné une moyenne des charges maximales supportées égales à 12,2 tonnes.

1. Déterminer l'intervalle de confiance à 99% de la charge maximale moyenne de tous les câbles fabriqués par l'usine.
2. Déterminer la taille minimale de l'échantillon étudié pour que la longueur de l'intervalle de confiance à 99% soit inférieure ou égale à $0,2$?

Solution (exercice ALEA.17.11)

1. Ici l'échantillon est déjà gaussien, donc il n'y a pas besoin d'appliquer le théorème central limite. Notons $(X_i)_i$ une suite i.i.d. de même loi $\mathcal{N}(\mu, 0,5^2)$. Alors, en notant $\overline{X_n} = \frac{X_1+\cdots+X_n}{n}$ pour tout $n \geqslant 1$, on sait par indépendance que $\overline{X_n}$ suit aussi une loi normale. De plus, par linéarité $\mathbf{E}\left(\overline{X_n}\right) = \mu$ et par indépendance, $\mathbf{Var}\left(\overline{X_n}\right) = \frac{0,5^2}{n}$. Donc par stabilité de la loi normale, pour tout $a \in \mathbf{R}$,

$$\mathbf{P}\left(-a \leqslant \frac{\overline{X_n}-\mu}{\sqrt{\frac{0,5^2}{n}}} \leqslant a\right) = 2\Phi(a) - 1.$$

De manière équivalente, on a :

$$\mathbf{P}\left(-a\frac{0.5}{\sqrt{n}} \leqslant \mu - \overline{X_n} \leqslant a\frac{0.5}{\sqrt{n}}\right) = \mathbf{P}\left(-a\frac{0.5}{\sqrt{n}} + \overline{X_n} \leqslant \mu \leqslant a\frac{0.5}{\sqrt{n}} + \overline{X_n}\right) = 2\Phi(a) - 1.$$

Il reste à choisir a de sorte que $2\Phi(a) - 1 = 0.99$. D'après le cours, on a $a = 2.57$. On obtient l'intervalle de confiance de niveau 99% :

$$\boxed{[12.02, 12.38]}.$$

2. De manière générale, pour n observations, on a l'intervalle

$$\left[-2.57\frac{0.5}{\sqrt{n}} + 12.2, 2.57\frac{0.5}{\sqrt{n}} + 12.2\right].$$

On résout donc :

$$2 \times 2.57\frac{0.5}{\sqrt{n}} \leqslant 0.2 \iff \frac{2.57}{\sqrt{n}} \leqslant 0.2 \iff \sqrt{n} \geqslant 12.85.$$

On trouve après calcul un nombre d'observations $\boxed{n = 166}$.

Exercice ALEA.17.12 | **Principe CMR : capture/marquage/recapture** Dans cet exercice, il va être question d'évaluer, de diverses manières, le nombre L (inconnu) de souris dans la cantine. Pour cela, on a prélevé dans la cantine 200 souris que l'on a marquées avant de les relâcher. Il y a donc maintenant dans la cantine L souris dont 200 sont marquées. En capturant de nouveau un certain nombre (on prendra 100) de souris on observe $\overline{X_{100}} = 0,6$, on va devoir estimer au mieux le nombre L.

1. On capture une souris, quelle est la probabilité p qu'elle soit marquée ?
2. Quel estimateur sans biais et convergeant de p connaissez-vous ?
3. En utilisant l'inégalité de Bienaymé-Tchebychev, proposer un intervalle de confiance non-asymptotique dans lequel le nombre L a une probabilité supérieure à 0.95 % de se trouver. On utilisera la majoration classique $p(1-p) \leq \frac{1}{4}$.
4. En utilisant le théorème central limite, trouver un intervalle de confiance asymptotique dans lequel le nombre L a une probabilité supérieure à 0.95 % de se trouver.
5. Comparer les deux intervalles de confiance.

Exercice ALEA.17.13 | **Agro—Véto, Sujet 3, 2018**

On pourra utiliser pour les programmes Python *la fonction* `linalg.matrix_rank()` *du module* numpy, *qui permet de déterminer le rang d'une matrice, comme le montre l'exemple suivant :*

```
>>> import numpy as np
>>> A = np.array([[1 ,2 ,1] , [2 ,3 ,2], [3 ,5 ,3]])
>>> np.linalg.matrix_rank(A)
2
```

La dernière ligne affiche le rang de la matrice $\begin{pmatrix}1&2&1\\2&3&2\\3&5&3\end{pmatrix}$, *c'est-à-dire* 2. *On pourra aussi utiliser la fonction* `randint()` *du module* `random`. *Pour a et b deux entiers* `randint(a,b)` *retourne un entier équiprobablement entre a et b (a et b étant inclus).*
On considère la matrice :

$$A=\begin{pmatrix}3&1&1\\1&0&1\\-3&0&-1\end{pmatrix}.$$

1. **1.1)** Écrire une fonction Python prenant en arguments deux vecteurs de taille 3 et renvoyant un booléen (`True` ou `False`) indiquant s'ils sont colinéaires. *(On pourra représenter les vecteurs par des listes).*
   1.2) Écrire une fonction Python `vecteurs_propres(u)` prenant en argument un vecteur de taille 3 et renvoyant un booléen (`True` ou `False`) indiquant s'il est un vecteur propre de A.
2. **2.1)** Vérifier que -1, 1, 2 sont valeurs propres de A et préciser pour chacune un vecteur propre associé.
   2.2) La matrice A est-elle diagonalisable ?
3. Soient $X_1,\dots,X_n$, n variables aléatoires indépendantes suivant la loi de Bernoulli de paramètre $p\in]0;1[$. On note :

$$M_n=\frac{1}{n}\sum_{k=1}^{n}X_k,\quad M_n^\star=\frac{M_n-p}{\sqrt{\frac{p(1-p)}{n}}}.$$

   3.1) Donner, pour $\alpha\in\mathbf{R}_+^\star$, l'approximation de la probabilité $P([-\alpha<M_n^\star<\alpha])$ donnée par le théorème central limite.
   3.2) En déduire que $\left[M_n-\frac{1}{\sqrt{n}},M_n+\frac{1}{\sqrt{n}}\right]$ est un intervalle de confiance de p au seuil de 95%. *On pourra admettre que,* $\forall x\in[0;1]$, $x(1-x)\leq\frac{1}{4}$ *et si* Φ *désigne la fonction de répartition d'une variable suivant une loi normale centrée réduite, alors* $\Phi(1;96)\approx 0,975$.
4. On note N_V le nombre de vecteurs propres de A dont les coefficients sont des entiers de $[\![-5\,,5]\!]$.
   4.1) Expliquer comment le programme suivant permet d'estimer la valeur de N_V :

```
def simul ():
    u = [ randint ( -5 ,5) for k in range (3) ]
    return vecteurs_propres(u)
n = 10000 #Valeur de n a definir.
nb = 0
for k in range (n):
    if simul ():
        nb += 1
print(round (nb/n *11**3)) # round (x) = l'entier le plus
 ↪ proche de x.
```

   4.2) Comment choisir n pour que l'on soit sûr à 95% de la valeur affichée ?
   4.3) Commenter le résultat obtenu.

Solution (exercice ALEA.17.13)

On considère la matrice :

$$A=\begin{pmatrix}3&1&1\\1&0&1\\-3&0&-1\end{pmatrix}.$$

1. 1.1) On rappelle que deux vecteur u et v sont colinéaires si et seulement si $\text{Rg}(u,v) < 2$.

```
def colineaires(u, v):
    a = np.array([u, v])
    return np.linalg.matrix_rank(a) < 2 #retourne un
    ↪ booléen
```

Nous pouvons alors tester si $[1,1],[2,2]$ sont colinéaires : True.

1.2) Le vecteur u est vecteur propre de A si et seulement si u est non nul et Au est colinéaire à u. Il suffit alors de tester la condition $u \neq 0$ et la colinéarité entre le produit Au et u :

```
def vecteurs_propres(u):
    return u != [0,0,0] and colineaires(np.dot(A,u), u)
```

2. 2.1) On pose $\text{X} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$; $\text{AX} = -\text{X} \iff (\text{A}+\text{I})\text{X} = 0 \iff \begin{cases} 4x+y+z &= 0 \\ x+y+z &= 0 \\ -3x &= 0 \end{cases} \iff$

$\begin{cases} x &= 0 \\ z &= -y \end{cases}$ Le système admet d'autres solutions que $(0,0,0)$ donc -1 est valeur propre de A et

$$\boxed{\text{E}_{-1}(\text{A}) = \text{Vect}\left(\begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}\right).}$$

$$\text{AX} = \text{X} \iff \begin{cases} 2x+y+z &= 0 \\ x-y+z &= 0 \\ -3x-2z &= 0 \end{cases} \iff \begin{cases} y+2x+z &= 0 \\ 3x+2z &= 0 \\ -3x-2z &= 0 \end{cases} \iff$$

$\begin{cases} y &= -1/2x \\ z &= -3/2x \end{cases}$ De même, on en déduit que 1 est valeur propre de A et

$$\boxed{\text{E}_1(\text{A}) = \text{Vect}\left(\begin{pmatrix} -2 \\ 1 \\ 3 \end{pmatrix}\right).}$$

$$\text{AX} = 2\text{X} \iff \begin{cases} x+y+z &= 0 \\ x-2y+z &= 0 \\ -3x-3z &= 0 \end{cases} \iff \begin{cases} y+x+z &= 0 \\ 3x+3z &= 0 \\ -3x-3z &= 0 \end{cases} \iff \begin{cases} y &= 0 \\ z &= -x \end{cases}$$

Ainsi 2 est bien valeur propre de A et

$$\boxed{\text{E}_2(\text{A}) = \text{Vect}\left(\begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}\right).}$$

Comme A ne peut avoir plus de trois valeurs propres, on en déduit que $\boxed{\text{Spec}(\text{A}) = \{-1,1,2\}}$.

2.2) A est carrée d'ordre 3 et possède 3 valeurs propres distinctes donc $\boxed{\text{A est diagonalisable}}$.

3. Soient $\text{X}_1,\dots,\text{X}_n$, n variables aléatoires indépendantes suivant la loi de Bernoulli de paramètre $p \in]0;1[$. On note : $\text{M}_n = \frac{1}{n}\sum_{k=1}^{n}\text{X}_k$ et $\text{M}_n^\star = \frac{\text{M}_n - p}{\sqrt{\frac{p(1-p)}{n}}}$.

3.1) X_k suit la loi de Bernoulli $\mathbf{B}(p)$ donc $\mathbf{E}(\text{X}_k) = p$ et $\mathbf{Var}(\text{X}_k) = p(1-p)$ pour tout $k \in [\![1\,,\,n]\!]$. On en déduit, par linéarité de l'espérance que $\mathbf{E}(\text{M}_n) = \frac{1}{n}\sum_{k=1}^{n}\mathbf{E}(\text{X}_k) = \frac{1}{n}\sum_{k=1}^{n} p = p$. De plus, comme les variables aléatoires $\text{X}_1,\dots,\text{X}_n$ sont indépendantes, on a :
$\mathbf{Var}(\text{M}_n) = \frac{1}{n^2}\sum_{k=1}^{n}\mathbf{Var}(\text{X}_k) = \frac{p(1-p)}{n}$ et par conséquent $\sigma_{\text{M}_n} = \sqrt{\frac{p(1-p)}{n}}$. Ainsi $\text{M}_n^\star = \frac{\text{M}_n - \mathbf{E}(\text{M}_n)}{\sigma(\text{M}_n)}$: $\text{M}_n^\star$ correspond à la variable centrée réduite associée à M_n. Comme les les variables aléatoires $\text{X}_1,\dots,\text{X}_n$ sont indépendantes et de même loi, on peut appliquer le théorème central limite : on en déduit que $\boxed{\text{M}_n^\star \text{ suit approximativement la loi normale } \mathcal{N}(0,1)}$. Ainsi, pour $\alpha \in \mathbf{R}_+^\star$, $\mathbf{P}([-\alpha < \text{M}_n^\star < \alpha]) \approx \Phi(\alpha) - \Phi(-\alpha)$ où Φ désigne la fonction de répartition de la loi normale centrée réduite. Sachant qu'on a $\Phi(-\alpha) = 1 - \Phi(\alpha)$ et $\text{P}(-\alpha \leq \text{X} \leq \alpha) = \text{P}(-\alpha < \text{X} < \alpha)$ pour toute variable X à densité, on peut aussi écrire que

$$\boxed{\mathbf{P}([-\alpha \leq \text{M}_n^\star \leq \alpha]) \approx 2\Phi(\alpha) - 1}.$$

3.2)

$$\mathbf{P}\left(p \in \left[\mathrm{M}_n - \frac{1}{\sqrt{n}}; \mathrm{M}_n + \frac{1}{\sqrt{n}}\right]\right) = \mathbf{P}\left(-\frac{1}{\sqrt{n}} \leq \mathrm{M}_n - p \leq \frac{1}{\sqrt{n}}\right)$$
$$= \mathbf{P}\left(-\frac{1}{\sqrt{p(1-p)}} \leq \mathrm{M}_n^\star \leq \frac{1}{\sqrt{p(1-p)}}\right)$$

Comme $\sqrt{p(1-p)} \leq \sqrt{1/4} = 1/2$ alors $-\frac{1}{\sqrt{p(1-p)}} \leq -2$ et $2 \leq \frac{1}{\sqrt{p(1-p)}}$.[20] Donc : $\mathbf{P}\left(-\frac{1}{\sqrt{p(1-p)}} \leq \mathrm{M}_n^\star \leq \frac{1}{\sqrt{p(1-p)}}\right) \geq \mathrm{P}\left(-2 \leq \mathrm{M}_n^\star \leq 2\right)$. Or $\mathbf{P}\left(-2 \leq \mathrm{M}_n^\star \leq 2\right) = 2\Phi(2) - 1 \geq 2\Phi(1,96) - 1 = 0,95$ car la fonction Φ est croissante sur $\mathbf{R}$. Dès lors, on en déduit que $\mathbf{P}\left(-\frac{1}{\sqrt{p(1-p)}} \leq \mathrm{M}_n^\star \leq \frac{1}{\sqrt{p(1-p)}}\right) \geq 0,95$. $\mathbf{P}\left(p \in \left[\mathrm{M}_n - \frac{1}{\sqrt{n}}; \mathrm{M}_n + \frac{1}{\sqrt{n}}\right]\right) \geq 95\%$ ce qui signifie que

$\boxed{\left[\mathrm{M}_n - \frac{1}{\sqrt{n}}, \mathrm{M}_n + \frac{1}{\sqrt{n}}\right] \text{ est un intervalle de confiance de } p \text{ au seuil de } 95\%}$.

4. 4.1) On note $\mathrm{N_V}$ (respectivement p) le nombre (respectivement la proportion) de vecteurs propres de A qui appartiennent à $[\![-5,5]\!]^3$. Comme $\#\,[\![-5,5]\!]^3 = 11^3$, alors $p = \frac{\mathrm{N_V}}{11^3}$ soit $\mathrm{N_V} = p \times 11^3$. On considère l'épreuve de Bernoulli qui consiste à choisir au hasard un vecteur de $[\![-5\,,\,5]\!]^3$, puis à renvoyer 1 si le vecteur en question est un vecteur propre de la matrice A (la probabilité de succès est notre paramètre de Bernoulli). On réalise $n = 10000$ fois dans des conditions indépendantes cette expérience (ce qui est réalisé dans la boucle `for`). On note $\mathrm{X}_k = 1$ si le k-ième vecteur tiré est vecteur propre de A, $\mathrm{X}_k = 0$ sinon. Les variables aléatoires X_k sont indépendantes et de même loi $\mathscr{B}(p)$. On sait alors que la variable aléatoire $\mathrm{M}_n = \frac{1}{n}\sum_{k=1}^{n} \mathrm{X}_k$ est un estimateur sans biais de p. Le nombre `nb/n` en sortie de boucle correspond à une réalisation de la variable M_n et donne une estimation de p. En multipliant par 11^3 et en arrondissant à l'entier le plus proche (car $\mathrm{N_V} \in \mathbf{N}$)[21], on obtient donc une estimation de $\mathrm{N_V}$.

4.2) On a vu précédemment que $\mathbf{P}\left(\mathrm{M}_n - \frac{1}{\sqrt{n}} \leq p \leq \mathrm{M}_n + \frac{1}{\sqrt{n}}\right) \geq 0,95$. Cela équivaut à : $\mathbf{P}\left(11^3\mathrm{M}_n - \frac{11^3}{\sqrt{n}} \leq \mathrm{N_V} \leq 11^3\mathrm{M}_n + \frac{11^3}{\sqrt{n}}\right) \geq 0,95$. Si on choisit n tel que $\frac{11^3}{\sqrt{n}} \leq 0,5$, on aura donc

$$\mathbf{P}\left(11^3\mathrm{M}_n - 0,5 \leq \mathrm{N_V} \leq 11^3\mathrm{M}_n + 0,5\right) \geq 0,95.$$

Par ailleurs, l'entier N_n le plus proche de $11^3\mathrm{M}_n$ vérifie

$$11^3\mathrm{M}_n - 0,5 \leq \mathrm{N}_n \leq 11^3\mathrm{M}_n + 0,5.$$

On en déduit que l'écart entre N_n et $\mathrm{N_V}$ est inférieur ou égal à 1 (avec une probabilité d'au moins 95%). Or $\frac{11^3}{\sqrt{n}} \leq 0,5 \iff \sqrt{n} \geq 2 * 11^3 \iff n \geq 4 \times 11^6$. Donc en choisissant $n \geq 4 \times 11^6$ (soit $n \geq 7086244$), la valeur affichée `round(nb/n*11**3)` donne une estimation de $\mathrm{N_V}$ à 95%.

4.3) On reprend le programme du début de la question, en remplaçant n par 7086244. On obtient $\boxed{22}$ après exécution. Calculons la valeur exacte de $\mathrm{N_V}$ afin de la comparer à 22 : D'après l'étude réalisée en seconde question, les vecteurs propres de A à coefficients entiers sont de la forme $(0,k,-k)$ ou $(-2k,k,3k)$ ou $(k,0,-k)$ avec $k \in \mathbb{Z}^\star$. Comme il y a 10 entiers non nuls compris entre -5 et 5, on dénombre :

- 10 vecteurs propres $(0,k,-k)$ éléments de $[\![-5\,,\,5]\!]^3$
- 10 vecteurs propres $(k,0,-k)$ éléments de $[\![-5\,,\,5]\!]^3$

Reste à dénombrer ceux qui sont de la forme $(-2k,k,3k)$ avec $k \neq 0$. Il faut que l'on ait : $\begin{cases} 0 < |2k| \leq 5 \\ 0 < |k| \leq 5 \\ 0 < |3k| \leq 5 \end{cases}$

Les seuls entiers k qui conviennent sont -1 et 1.

Il y a donc 2 vecteurs propres de la forme $(-2k,k,3k)$ qui appartiennent à $[\![-5\,,\,5]\!]^3$.

On en déduit que $\boxed{\mathrm{N_V} = 10 + 10 + 2 = 22}$ ce qui correspond à la valeur obtenue par estimation dans la question précédente.

[20] On « remontre » ici, dans ce cas particulier, que l'on peut épaissir tout intervalle de confiance de seuil $1-\alpha$, la version épaissie reste un intervalle de confiance de seuil $1-\alpha$.

[21] Attention, cette fonction n'est pas la partie entière, qui est `int()` dans Python

4.4. Tests

Exercice ALEA.17.14 | On considère un test de dépistage d'une maladie qui a donné les résultats suivants lors d'une phase de test sur 1600 individus sains et 1600 individus malades :

	Malades	Sains
Tests positifs	1522	92
Tests négatifs	78	1508

On considère p_1 la probabilité d'avoir un test positif pour une personne saine et p_2 la probabilité d'avoir un test négatif pour une personne malade.

1. Déterminer un intervalle de confiance pour p_1 au risque 5 %. On arrondira au millième.
2. Combien doit on avoir au plus de personnes malades testées négatives pour valider l'hypothèse $p_2 = 0.001$? On arrondira les calculs à 10^{-4}.

Exercice ALEA.17.15 |

1. Afin d'étudier le pourcentage p de consommateurs satisfait par le produit A, on interroge 100 consommateurs et 56 déclarent être satisfaits. Est-ce suffisant pour continuer l'exploitation du produit A ? *Indication : on cherchera un intervalle de confiance à* 95%.
2. En supposant qu'on garde la même moyenne empirique de $0,56$, et le même risque $\alpha = 0,05$, combien de personnes doit-on interroger pour prendre une décision ?