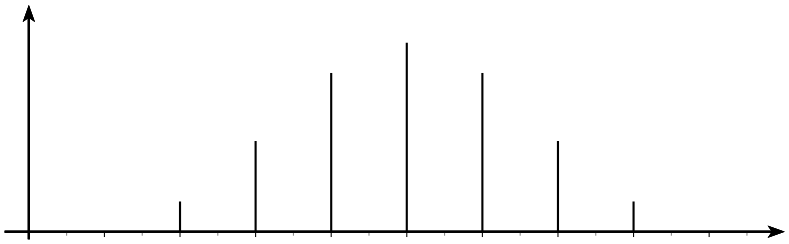
Statistiques

On veut étudier un ensemble appelé **population** dont les éléments sont appelés **individus**.
On s'intéresse à certaines propriétés de ces individus, les **caractères**, appelés aussi **variables statistiques**.
Ces caractères peuvent être quantitatifs ou qualitatifs.
Si c'est impossible d'étudier la population dans sa totalité (pour des questions de coût ou de logistique), on prélève une n-liste d'individus, appelée **échantillon** de taille n .

1 Statistique univariée

L'observation d'un caractère x sur un échantillon de taille n se traduit par une n-liste $(x_1,...x_n)$ appelée **série statistique**.
Pour une meilleure lisibilité des données, on les regroupe par valeur (modalité) : $x'_1, x'_2, \dots, x'_p$.
Pour tout $i \in [\![1, p]\!]$, on appelle **effectif** associé à la valeur x'_i le nombre n_i d'individus de l'échantillon pour lesquels le caractère observé prend la valeur x'_i.
L'entier $n = \sum_{i=1}^{p} n_i$ est appelé effectif total de la série statistique.
Le rapport $f_i = \dfrac{n_i}{n}$ est appelé **fréquence** d'observation de x'_i.
Les sommes $\sum_{j=1}^{i} f_i$ pour $1 \leqslant i \leqslant p$ sont appelées **fréquences cumulées**.
Pour représenter une série statistique on reporte en abscisse les valeurs x'_i équidistantes les unes des autres.
Pour obtenir un **diagramme en bâtons** on reporte en ordonnée une longueur proportionnelle à l'effectif n_i.



Pour obtenir un **histogramme** on trace des rectangles dont la hauteur est proportionnelle à l'effectif n_i.

1.1 Caractéristiques de position

On appelle **moyenne** (empirique) d'une série statistique quantitative le réel : $\bar{x} = \dfrac{1}{n}\sum_{k=1}^{n} x_k$.
On appelle **médiane** d'une série statistique quantitative toute valeur qui partage la série en deux parties de même effectif : il y a autant de valeurs inférieures que de valeurs supérieures à une médiane.
Si $x_1 \leqslant \cdots \leqslant x_n$, alors :
Si n est impair, $n = 2q+1$, il y a une seule médiane : x_{q+1}.
Si n est pair, $n = 2q$, tout réel de l'intervalle $[x_q, x_{q+1}]$ est une médiane. On peut choisir $\frac{1}{2}(x_q + x_{q+1})$.
On appelle **mode** toute valeur x_i telle que la fréquence f_i correspondante soit maximale.

1.2 Caractéristiques de dispersion

On appelle **variance** (empirique) d'une série statistique quantitative le réel : $s_x^2 = \dfrac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2 = \dfrac{1}{n}\sum_{k=1}^{n} x_k^2 - \bar{x}^2$.
On appelle **écart type** de la série le réel $s_x = \sqrt{s_x^2}$.
Les **quartiles** permettent de séparer une série statistique en quatre groupes de même effectif (à une unité près).
Un quart des valeurs sont inférieures au premier quartile Q_1.
Un quart des valeurs sont supérieures au troisième quartile Q_3.
Les **déciles** permettent de séparer une série statistique en dix groupes de même effectif (à une unité près).

Un dixième des valeurs sont inférieures au premier décile D_1.
Un dixième des valeurs sont supérieures au neuvième décile D_9.
On arrondit les nombres $n/4, 3n/4, n/10, 9n/10$ aux entiers supérieurs.

2 Statistique bivariée

L'observation de deux caractères quantitatifs x et y sur un échantillon de taille n se traduit par un n-uplet d'éléments de $\mathbb{R}^2$: $\big((x_1,y_1),\dots,(x_n,y_n)\big)$ appelé série statistique double.

2.1 Corrélation

On appelle **covariance** de x et y le réel : $s_{x,y} = \dfrac{1}{n}\sum_{k=1}^{n}(x_k-\bar{x})(y_k-\bar{y}) = \dfrac{1}{n}\sum_{k=1}^{n}x_k y_k - \bar{x}\bar{y}$.

On appelle **coefficient de corrélation linéaire** de x et y le réel : $r_{x,y} = \dfrac{s_{x,y}}{s_x s_y}$. Il vérifie $|r_{x,y}| \leqslant 1$.

2.2 Nuage de points

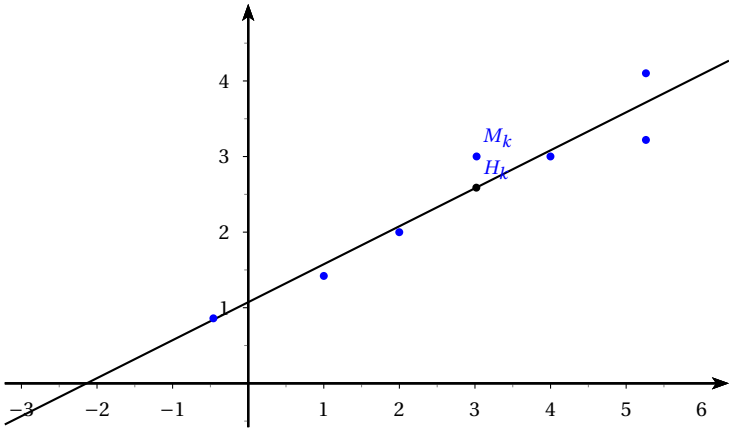
Dans le plan rapporté à un repère $(O,\vec{\imath},\vec{\jmath})$ portons les points $M_1(x_1,y_1),\dots,M_n(x_n,y_n)$.
L'ensemble des n points s'appelle **nuage de points**.
Le point de coordonnées $(\overline{x},\overline{y})$ est le **point moyen** du nuage.

2.3 Droite de régression

On se place dans $\mathbb{R}^n$ muni du produit scalaire usuel.
À chaque point $M_k(x_k,y_k)$ du nuage de points on fait correspondre le point H_k de la droite d'équation $y=ax+b$ ayant même abscisse que M_k.
Soient les vecteurs de $\mathbb{R}^n$: $\vec{x}=(x_1,\dots,x_n)$ et $\vec{y}=(y_1,\dots,y_n)$. Posons $\vec{w}=(1,\dots,1)$ et $F=\text{vect}(\vec{x},\vec{w})$.
On cherche $(a,b)\in\mathbb{R}^2$ qui minimise $f(a,b)=\sum_{k=1}^{n}(M_kH_k)^2=\sum_{k=1}^{n}(y_k-(ax_k+b))^2=\|\vec{y}-(a\vec{x}+b\vec{w})\|^2$.
Ce problème a une solution unique : le couple (a_0,b_0) tel que $a_0\vec{x}+b_0\vec{w}$ soit la projection orthogonale de $\vec{y}$ sur F.

$(\widehat{a},\widehat{b})$ est déterminé par : $\begin{cases}(a_0\vec{x}+b_0\vec{w}).\vec{x}=\vec{y}.\vec{x}\\(a_0\vec{x}+b_0\vec{w}).\vec{w}=\vec{y}.\vec{w}\end{cases}$ On trouve $\begin{cases}a_0=\dfrac{s_{x,y}}{s_x^2}\\ b_0=\overline{y}-a_0\overline{x}\end{cases}$.

La droite d'équation $y=a_0x+b_0$ est appelée droite de régression de y par rapport à x (ou droite des moindres carrés).
Elle passe par le point moyen de coordonnées $(\bar{x},\bar{y})$.
$f(a_0,b_0)=(\vec{y}-(a_0\vec{x}+b_0\vec{w})).(\vec{y}-(a_0\vec{x}+b_0\vec{w}))=(\vec{y}-(a_0\vec{x}+b_0\vec{w})).\vec{y}=n(s_y^2-a_0s_{x,y})=ns_y^2(1-r_{x,y}^2)$.
Plus $r_{x,y}$ est proche de 1, meilleur est l'ajustement.



Si $r_{x,y}>0$ alors $s_{x,y}>0$ et $a_0>0$. Le nuage est ascendant : plus la valeur de x est grande, plus celle de y est grande en moyenne.
Si $r_{x,y}<0$, alors $a_0<0$ et le nuage est descendant.

3 Statistique inférentielle

3.1 Estimation ponctuelle

Définition 1.
On appelle **n- échantillon** de la variable aléatoire X un n-uplet $(X_1,\dots,X_n)$ de variables aléatoires indépendantes et de même loi que X.
On appelle **réalisation** de l'échantillon $(X_1,\dots,X_n)$ la valeur observée de l'échantillon.
C'est un n-uplet $(x_1,\dots,x_n)=(X_1(\omega),\dots,X_n(\omega))$ de $\mathbb{R}^n$.

Définition 2.
Soit θ un paramètre de la loi de X. (En général θ sera l'espérance ou la variance de X).
On appelle **estimateur** noté T_n de l'échantillon $(X_1,\dots,X_n)$ de X toute fonction de $(X_1,\dots,X_n)$ donnant des informations sur le paramètre θ.
La valeur t_n de T_n obtenue à partir d'un échantillon observé est appelée **estimation** du paramètre.
On appelle **erreur d'estimation** la différence $T_n-\theta$ entre l'estimateur et la valeur du paramètre.
On appelle **biais** l'espérance de l'erreur d'estimation $E(T_n)-\theta$.
Un estimateur est sans biais si $E(T_n)=\theta$.

Définition 3. Estimateurs de la moyenne et de la variance
X est una variable aléatoire d'espérance μ et de variance σ^2. Soit $(X_1,\dots,X_n)$ un n-échantillon de X.
On appelle **moyenne empirique** de l'échantillon l'estimateur :

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

On appelle **variance empirique** de l'échantillon l'estimateur :

$$S_n^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-M_n)^2=\frac{1}{n}\sum_{i=1}^{n}X_i^2-M_n^2.$$

Théorème 1.
$E(M_n)=\mu$: *M_n est un estimateur sans biais de μ.*
$V(M_n)=\dfrac{\sigma^2}{n}$: *La dispersion autour de μ diminue quand n augmente.*
$E(S_n^2)=\dfrac{n-1}{n}\sigma^2$: *S_n^2 est un estimateur biaisé de la variance de X.*
Un meilleur estimateur (sans biais) est $\hat{S}_n^2=\dfrac{n}{n-1}S_n^2$.
$\hat{S}_n$ est appelé écart type corrigé.

Preuve
$E(M_n)=\dfrac{1}{n}\sum_{k=1}^{n}E(X_k)=\mu$
$V(M_n)=\dfrac{1}{n^2}\sum_{k=1}^{n}V(X_k)=\dfrac{\sigma^2}{n}$ (var indépendantes)
$E(S_n^2)=\dfrac{1}{n}\sum_{i=1}^{n}E(X_i^2)-E(M_n^2)=\dfrac{1}{n}\sum_{i=1}^{n}(V(X_i)+E(X_i)^2)-\left(V(M_n)+E(M_n)^2\right)=\dfrac{1}{n}\sum_{i=1}^{n}(\sigma^2+\mu^2)-\left(\dfrac{\sigma^2}{n}+\mu^2\right)=\dfrac{n-1}{n}\sigma^2$

3.2 Estimation par intervalle de confiance

Théorème 2. ***Théorème central limite : 2e forme***
Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires définies sur un même espace de probabilité $(\Omega,\mathscr{A},P)$ indépendantes, de même loi, admettant une espérance μ et une variance σ^2. Alors quels que soient les réels a et b tels que $a<b$,

$$\lim_{n\to+\infty}P\left(a<\frac{M_n-\mu}{\frac{S_n}{\sqrt{n}}}\leqslant b\right)=\int_a^b\frac{\mathrm{e}^{-\frac{x^2}{2}}}{\sqrt{2\pi}}dx=\Phi(b)-\Phi(a).$$

Conséquence

$\forall u>0,\ \lim\limits_{n\to\infty} P\left(M_n - u\dfrac{S_n}{\sqrt{n}} < \mu \leqslant M_n + u\dfrac{S_n}{\sqrt{n}}\right) = P\left(-u < \dfrac{M_n-\mu}{\frac{S_n}{\sqrt{n}}} \leqslant u\right) = \Phi(u) - \Phi(-u) = 2\Phi(u) - 1$

On fixe le risque α ($\alpha \in]0,1[$). Il existe un unique $u_{1-\frac{\alpha}{2}} > 0$ tel que $2\Phi(u_{1-\frac{\alpha}{2}}) - 1 = 1-\alpha$:

Le nombre $u_{1-\frac{\alpha}{2}} = \Phi^{-1}\left(1-\dfrac{\alpha}{2}\right)$ est appelé quantile d'ordre $1-\dfrac{\alpha}{2}$ de la loi normale centrée réduite.

$\lim\limits_{n\to\infty} P\left(\mu \in \left[M_n - u_{1-\frac{\alpha}{2}}\dfrac{S_n}{\sqrt{n}}, M_n + u_{1-\frac{\alpha}{2}}\dfrac{S_n}{\sqrt{n}}\right]\right) = 1-\alpha$

Définition 4.

On dit que l'intervalle aléatoire $\left[M_n - u_{1-\frac{\alpha}{2}}\dfrac{S_n}{\sqrt{n}}, M_n + u_{1-\frac{\alpha}{2}}\dfrac{S_n}{\sqrt{n}}\right]$ est un intervalle de confiance (asymptotique) de μ au niveau de confiance $1-\alpha$ (ou de risque α).

En particulier, si $X_1, \ldots, X_n$ suivent une loi de Bernoulli de paramètre p, alors l'intervalle

$\left[M_n - u_{1-\frac{\alpha}{2}}\sqrt{\dfrac{M_n(1-M_n)}{n}}, M_n + u_{1-\frac{\alpha}{2}}\sqrt{\dfrac{M_n(1-M_n)}{n}}\right]$ est un intervalle de confiance de p au niveau de confiance $1-\alpha$.

Par abus de langage, on appelle aussi intervalle de confiance toute réalisation de cet intervalle.

Si n est "grand", une proportion de $1-\alpha$ d'intervalles de confiance ainsi construits contiennent μ.

En python

Pour accéder concrètement au nombre u, on pourra faire appel à la bibliothèque `scipy.stats` qui fournit les fonctions suivantes :

`norm.cdf()` qui donne la fonction de répartition d'une loi normale centrée réduite,

`norm.ppf()` qui donne la fonction réciproque de la précédente (également nommée fonction des quantiles).

4 Test statistique

4.1 Notion de test statistique

Construire un test de l'hypothèse nulle H_0 contre l'hypothèse alternative H_1, c'est établir un critère de décision permettant de choisir entre l'hypothèse H_0 et H_1.

H_0 est l'hypothèse privilégiée : c'est celle que l'on garde si le résultat de l'expérience n'est pas clair. Il y a analogie entre un test d'hypothèse et un procès : tout suspect est présumé innocent et l'accusation doit apporter la preuve de sa culpabilité. Quand on accepte H_0 on ne prouve pas qu'elle est vraie, on accepte de conserver H_0 parce qu'on n'a pas pu accumuler suffisamment de preuves contre elle.

4.2 Test de conformité sur la moyenne

On veut tester l'**Hypothèse nulle** H_0 **:** $\mu = \mu_0$ contre l'**Hypothèse alternative** H_1 **:** $\mu \neq \mu_0$

On fixe un risque α.

Si H_0 est vérifiée, alors $\lim\limits_{n\to\infty} P\left(\left|\dfrac{M_n-\mu_0}{\frac{S_n}{\sqrt{n}}}\right| > u_{1-\frac{\alpha}{2}}\right) = \alpha$

On rejette l'hypothèse H_0 si la valeur observée de $\dfrac{M_n-\mu_0}{\frac{S_n}{\sqrt{n}}}$ est en dehors de l'intervalle $\left[-u_{1-\frac{\alpha}{2}}, u_{1-\frac{\alpha}{2}}\right]$.

4.3 Comparaison de deux moyennes (Pour les TIPE)

On veut tester l'**Hypothèse nulle** H_0 **:** $\mu_1 = \mu_2$ contre l'**Hypothèse alternative** H_1 **:** $\mu_1 \neq \mu_2$
On fixe un risque α.

Si H_0 est vérifiée, alors $\lim\limits_{n\to+\infty} P\left(\dfrac{\overline{X}-\overline{Y}}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \leqslant u_{1-\frac{\alpha}{2}}\right) = 1-\alpha$.

On rejette l'hypothèse H_0 si la valeur observée de $\dfrac{\overline{X}-\overline{Y}}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} > u_{1-\frac{\alpha}{2}}$ est en dehors de l'intervalle $\left[-u_{1-\frac{\alpha}{2}}, u_{1-\frac{\alpha}{2}}\right]$.