Tests statistiques et TIPE

Principe d'un test statistique

Soit une population dont les individus possèdent un caractère X (mesurable ou dénombrable). Des paramètres de X ne sont pas connus : par exemple $E(X)$. Une hypothèse est formulée sur ce paramètre.
On souhaite porter un jugement sur cette hypothèse (est-elle raisonnable ?), en se basant sur des résultats d'un échantillon prélevé de cette population.

- **Un test d'hypothèse** est une démarche qui a pour but de fournir **une règle de décision** permettant, sur la base de résultats d'échantillon, de faire un choix entre deux hypothèses.
- **Hypothèse nulle, H_0 :** c'est l'hypothèse que l'on cherche à tester.
  Par exemple, on fait l'hypothèse que $E(X) = \mu_0$.
- **Seuil de signification d'un test d'hypothèse :** c'est le risque α consenti à l'avance de rejeter à tort l'hypothèse H_0, alors qu'elle est vraie.
  On utilisera en général $\alpha = 5\%$.
- **Statistique de test ou variable d'échantillonnage** : c'est une variable aléatoire T (en lien avec le problème initial) dont on connait la loi sous l'hypothèse H_0. Il existe alors deux domaines de $\mathbb{R}$, une **zone de rejet** notée R_{rejet} et une **zone de non-rejet** notée $R_{non-rejet}$ telles que $P(T \in R_{rejet}) = \alpha$ et $P(T \in R_{non-rejet}) = 1 - \alpha$.
- **Utilisation du test :** A partir d'un échantillon, on calcule une valeur observée de T, notée t_{obs}.
  Règle de décision : Si $t_{obs} \notin R_{non-rejet}$ c'est-à-dire $t_{obs} \in R_{rejet}$ alors l'hypothèse H_0 est rejetée.
  Sinon, l'hypothèse H_0 n'est pas rejetée.

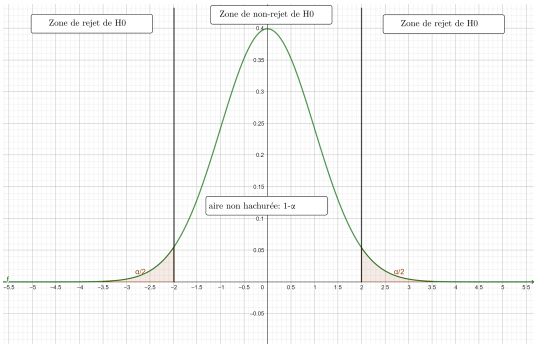
Dans ce document, nous alllons étudier différents tests statistiques. Pour information, ce document sera repris partiellement en TP Python.

Test de conformité à la moyenne dans le cas où l'effectif de la population est grand (≥ 30).

Nous observons un caractère quantitatif noté X d'une population, dite population-mère. X est une variable aléatoire, d'espérance $E(X) = \mu$ et de variance $V(X) = \sigma^2$.
Soit $n \in \mathbb{N}^*$, n grand. On prélève un échantillon de n individus de la population et on observe pour chacun d'eux la valeur du paramètre X : on a donc un $n-$uplet $(X_1, X_2, ..., X_n)$ de variables aléatoires, avec les $(X_k)_k$ toutes de même loi de probabilité égale à celle de X, et indépendantes.

- Prenons l'hypothèse nulle H_0 : $\mu = \mu_0$.
- La variable d'échantillonnage qui convient pour ce test est la moyenne empirique : $M_n = \dfrac{1}{n}\sum_{k=1}^{n} X_k$.
  On note $M_n^* = \dfrac{M_n - E(M_n)}{\sigma(M_n)} = \dfrac{M_n - \mu}{\frac{\sigma}{\sqrt{n}}}$ la moyenne empirique centrée-réduite.
- Résultat mathématique : d'après le théorème central limite, M_n^* suit une loi proche de la loi $\mathcal{N}(0,1)$.
  Donc pour tout $x \in \mathbb{R}_+^*$, $P(-x < M_n^* < x) \simeq P(-x < Z < x)$ avec $Z \hookrightarrow \mathcal{N}(0,1)$.

  Comment exploiter ce résultat ?
  On peut schématiser les régions de rejet et de non-rejet de l'hypothèse H_0 au risque α à l'aide du graphique d'une densité de la loi normale centrée réduite, loi proche de celle de M_n^*.
  Si la réalisation observée de M_n^* correspond à une abscisse de la zone de rejet, on rejette H_0, et sinon, on ne rejette pas H_0. Il faut le comprendre en : sous l'hypothèse H_0, c'est rare d'avoir une telle valeur observée (dans la zone de rejet). Soit on n'a pas eu de chance dans notre expérience (risque de probabilité α), soit c'est que l'hypothèse H_0 n'est pas la bonne (conclusion de notre test).



Plus précisément, notons u tel que $P(-u < Z < u) = 1-\alpha$. Alors $P(-u < M_n^* < u) \simeq 1-\alpha$.
Or $(-u < M_n^* < u) = (|M_n^*| < u) = \left(\left|\dfrac{M_n - \mu}{\frac{\sigma}{\sqrt{n}}}\right| < u\right) = \left(|M_n - \mu| < u\frac{\sigma}{\sqrt{n}}\right) = \left(\mu \in \left]M_n - u\frac{\sigma}{\sqrt{n}}, M_n + u\frac{\sigma}{\sqrt{n}}\right[\right)$,
Donc on a $P(\left(\mu \in \left]M_n - u\frac{\sigma}{\sqrt{n}}, M_n + u\frac{\sigma}{\sqrt{n}}\right[\right) \simeq 1-\alpha)$.

- Règle de décision et conclusion du test : on commence par calculer m_n, valeur observée de M_n.
  Puis, si $\mu_0 \notin \left]m_n - u\frac{\sigma}{\sqrt{n}}, m_n + u\frac{\sigma}{\sqrt{n}}\right[$, c'est-à-dire si l'écart $m_n - \mu_0$ n'appartient pas à $\left]-u\frac{\sigma}{\sqrt{n}}, u\frac{\sigma}{\sqrt{n}}\right[$ alors l'hypothèse H_0 est rejetée : la différence est anormalement élevée et les résultats de l'échantillon sont en contradiction avec H_0.
  Sinon, l'hypothèse H_0 n'est pas rejetée : la différence observée n'est pas significative et nous concluons qu'elle est imputable aux fluctuations de l'échantillonnage.

Exemple : Dans le cas où $\alpha = 5\%$, alors $u = 1.96$ et $P\left(\mu_0 \in \left]M_n - u\frac{\sigma}{\sqrt{n}}, M_n + u\frac{\sigma}{\sqrt{n}}\right[\right) \simeq 95\%$.
Ainsi la zone de rejet de H_0 est : $m_n - \mu_0 \notin \left]-1.96\frac{\sigma}{\sqrt{n}}, 1.96\frac{\sigma}{\sqrt{n}}\right[$
et la zone de non-rejet de H_0 est : $m_n - \mu_0 \in \left]-1.96\frac{\sigma}{\sqrt{n}}, 1.96\frac{\sigma}{\sqrt{n}}\right[$.

En pratique : σ est souvent inconnu.
Le résultat mathématique que l'on utilise est alors la version 3 du théorème central limite : dans M_n^*, on remplace σ par l'écart-type empirique corrigé S'_n défini par $S'^2_n = \dfrac{1}{n-1}\sum_{k=1}^{n}(X_k - M_n)^2$). Quand n est grand, la loi de cette variable M_n^* reste proche de la loi normale centrée réduite.
Donc les intervalles obtenus restent les mêmes à condition de remplacer σ par la valeur observée de S'_n.

Compléments : utilisation du module *scipy.stats*.
On peut alors utiliser les fonctions suivantes :
pdf : probability density function, correspond à une densité.
cdf : cumulative density function, correspond à la fonction de répartition.
ppf : percent point function, correspond à la réciproque de la fonction de répartition. (percentiles)
rvs : random variates, correspond à une valeur aléatoire (réalisation d'une variable aléatoire).

Pour travailler sur une loi normale, l'import est : *from scipy.stats import norm.*
La syntaxe pour utiliser ces fonctions associées à une loi normale de paramètres μ, σ^2 devient :
norm.pdf(x,loc=μ*,scale=*σ*)*, *norm.cdf(x,loc=*μ*,scale=*σ*)*
norm.ppf(x,loc=μ*,scale=*σ*)*, *norm.rvs(loc=*μ*,scale=*σ*)*.
Remarque : pour simuler une variable de loi normale, la syntaxe `rd.gauss` est plus accessible car elle nécessite seulement la bibliothèque `random`.

Test de conformité à la moyenne dans le cas gaussien

On garde les notations de la section précédente. On suppose dans cette section que X suit une loi normale.
On dispose d'une série statistique $x = (x_k)_{1\leq k\leq n}$ (=valeurs observées) que l'on suppose être une réalisation d'un n-échantillon $(X_1, ..., X_n)$ de X. Dans ce cadre d'étude, n peut être petit.
On note comme précédemment $m_n = \frac{1}{n}\sum_{k=1}^{n} x_k$ et $s'_n = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - m_n)^2}$.

- L'hypothèse H_0 est : $\mu = \mu_0$.
- Résultat mathématique utilisé :
  lorsque la variance σ^2 est connue, alors par stabilité de la loi normale, $\frac{M_n - \mu}{\frac{\sigma}{n}} \hookrightarrow \mathcal{N}(0,1)$.
  On est ainsi ramené au cas précédent.

  lorsque la variance σ^2 est inconnue alors $\frac{M_n - \mu}{\frac{S'_n}{n}} \hookrightarrow \mathcal{T}(n-1)$, où $\mathcal{T}(n-1)$ désigne la loi de Student à $n-1$ degrés de liberté.

- Règle de décision : en reprenant les calculs faits précédemment, on obtient :

  lorsque la variance σ^2 est connue : on rejette H_0 dès que $\mu_0 \notin]m_n - u\frac{\sigma}{\sqrt{n}}, m_n + u\frac{\sigma}{\sqrt{n}}[$ avec u tel que $P(-u < Z < u) = 1 - \alpha$ où $Z \hookrightarrow \mathcal{N}(0,1)$
  lorsque la variance σ^2 est inconnue : on rejette H_0 dès que $\mu_0 \notin]m_n - a\frac{s'_n}{\sqrt{n}}, m_n + a\frac{s'_n}{\sqrt{n}}[$ avec a tel que $P(-a < T < a) = 1 - \alpha$ où $T \hookrightarrow \mathcal{T}(n-1)$.

Compléments :
Une loi de Student à d degrés de liberté est une loi à densité, de densité $\forall t \in \mathbb{R}$, $f(t) = \dfrac{c_d}{\left(1 + \frac{t^2}{d}\right)^{\frac{d+1}{2}}}$ où c_d est une constante de normalisation (pour que l'intégrale vale 1).
Cette densité est une fonction paire, donc les zones de rejet et de non-rejet de l'hypothèse H_0 sont du même type que lorsqu'on se ramène à une loi normale.
Enfin, lorsque d est grand ($d \geq 30$), une loi de Student est très proche d'une loi normale centrée réduite.

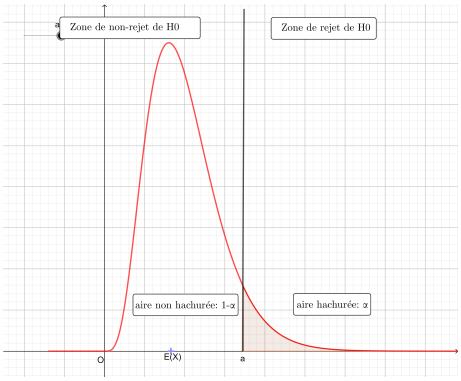
Pour une loi de Student de d degrés de liberté : Importation du module : *from scipy.stats import t.*
La syntaxe pour utiliser ces fonctions associées à une loi de Student de d degrés de liberté :
t.pdf(x,d), *t.cdf(x,d)* , *t.ppf(x,d)*, *t.rvs(d)*.

Test de conformité à une distribution discrète donnée.

Soit X une variable prenant un nombre fini m de valeurs notées $\{v_1, v_2, ..., v_m\}$.
On dispose d'une série statistique $x = (x_i)_{1\leq i\leq n}$ réelle que l'on suppose être une réalisation d'un n-échantillon $(X_1, ..., X_n)$ de X.
Pour tout k de 1 à m, on note f_k la fréquence d'apparition de la valeur v_k dans la série statistique $x = (x_i)_{1\leq i\leq n}$.

Soit $(p_1, p_2, .., p_m)$ une liste de probabilités telle que $\sum_{k=1}^{m} p_k = 1$

- L'hypothèse nulle H_0 à tester est : pour tout $k \in [\![1, m]\!]$, $P(X = v_k) = p_k$.
- Résultat mathématique admis : notons $F_k = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{(X_j = v_k)}$ la variable aléatoire égale à la fréquence d'apparition de la valeur v_k dans le n-échantillon $(X_1, ..., X_n)$.
  Alors sous l'hypothèse H_0, $\Delta_n = n\sum_{k=1}^{m} \dfrac{(F_k - p_k)^2}{p_k}$ suit une loi proche d'une loi du χ^2 à $m-1$ degrés de liberté.
- Règle de décision :
  On peut schématiser les régions de rejet et de non-rejet de l'hypothèse H_0 au risque α, en exploitant le graphe d'une densité d'une loi du χ^2 à $m-1$ degrés de liberté.
  Notons a tel que $P(Y < a) = 1 - \alpha$ avec $Y \hookrightarrow \chi^2(m-1)$



Alors on rejettera l'hypothèse H_0 si $s = n\sum_{k=1}^{m} \frac{(f_k - p_k)^2}{p_k}$ est strictement supérieur à a.

Compléments :
Une loi du χ^2 (prononcer "ki 2") à d degrés de liberté est une loi à densité, de densité du type : $\forall t \in \mathbb{R}$, $f(t) = \begin{cases} 0 & \text{si t<0} \\ C_d \times t^{\frac{d}{2}-1} e^{-\frac{1}{2}t} & \text{si } t \geq 0 \end{cases}$.
La loi du χ^2 est obtenue en sommant des variables aléatoires indépendantes normales centrées réduites au carré, autrement dit : si $(X_1, X_2, ..., X_d)$ sont d variables indépendantes etde même loi $\mathcal{N}(0,1)$, alors $K_d = \sum_{k=1}^{d} X_k^2$ suit la loi du χ^2 à d degrés de libertés.

Pour une loi du χ^2 à d degrés de liberté : Importation du module : *from scipy.stats import chi2.*
La syntaxe pour utiliser ces fonctions associées à une loi du χ^2 de d degrés de liberté :
chi2.pdf(x,d), *chi2.cdf(x,d)* , *chi2.ppf(x,d)*, *chi2.rvs(d)*.

Test de comparaison de 2 moyennes dans le cas gaussien

Nous observons un même caractère X quantitatif, suivant une loi normale, dans deux populations.
On prélève deux échantillons indépendants : un dans chacune des populations.
Modèle mathématique :
Population 1 :
$X_1 \hookrightarrow \mathcal{N}(\mu_1, \sigma_1)$, échantillon de taille n_1, de moyenne empirique $M_{n,1}$, et de variance empirique $S^2_{n,1}$
Population 2 :
$X_2 \hookrightarrow \mathcal{N}(\mu_2, \sigma_2)$, échantillon de taille n_2, de moyenne empirique $M_{n,2}$, et de variance empirique $S^2_{n,2}$.

- L'hypothèse nulle H_0 à tester est : "$\mu_1 = \mu_2$"

- Résultat mathématique admis :
  Lorsque les variances σ_1^2 et σ_2^2 sont connues, alors $\frac{M_{n,1} - M_{n,2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \hookrightarrow \mathcal{N}(0,1)$.
  Lorsque les variances σ_1^2 et σ_2^2 sont inconnues mais que les effectifs sont grands (n_1 et n_2 supérieurs à 30), alors

  $\frac{M_{n,1} - M_{n,2}}{\sqrt{\frac{S'^2_{n,1}}{n_1} + \frac{S'^2_{n,2}}{n_2}}} \hookrightarrow \mathcal{N}(0,1)$ où $S'^2_{n,1}$ et $S'^2_{n,2}$ sont les estimateurs corrigés de la variance.

  Lorsque les variances σ_1^2 et σ_2^2 sont inconnues mais proches (faire au jugé, ou faire un test de Fisher-Snedecor), et que n_1 et n_2 sont petits (inférieurs à 30), alors
  $\frac{M_{n,1} - M_{n,2}}{\sqrt{S^2(\frac{1}{n_1} + \frac{1}{n_2})}} \hookrightarrow \mathcal{T}(n_1 + n_2 - 1)$, loi de Student à $n_1 + n_2 - 1$ degrés de liberté.
  où S^2 est l'estimateur de la variance commune aux deux échantillons : $S^2 = \frac{n_1 S^2_{n,1} + n_2 S^2_{n,2}}{n_1 + n_2 - 2}$.

  Dans les autres cas, le test est plus compliqué, car le nombre de degrés de liberté de la loi de Student est

plus difficile à déterminer.

- Règle de décision :
  Lorsque les variances sont connues :
  on rejette H_0 dès que $\frac{m_{n,1}-m_{n,2}}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \notin [-u,u]$ avec u tel que $P(-u<Z<u)=1-\alpha$ où $Z \hookrightarrow \mathcal{N}(0,1)$.
  Lorsque les variances sont inconnues mais que les échantillons sont grands :
  on rejette H_0 dès que $\frac{m_{n,1}-m_{n,2}}{\sqrt{\frac{s'^2_{n,1}}{n_1}+\frac{s'^2_{n,1}}{n_2}}} \notin [-u,u]$ avec u tel que $P(-u<Z<u)=1-\alpha$ où $Z \hookrightarrow \mathcal{N}(0,1)$.
  Lorsque les variances sont inconnues mais proches, et que les échantillons sont petits :
  on rejette H_0 dès que $\frac{m_{n,1}-m_{n,2}}{\sqrt{s^2(\frac{1}{n_1}+\frac{1}{n_2})}} \notin [-u,u]$ avec s^2 valeur observée de S défini ci-dessus, et u tel que $P(-u<T<u)=1-\alpha$, où T suit la loi de Student à n_1+n_2-2 degrés de liberté.

Parmi les tests "classiques", on peut encore citer

— un test d'hypothèse sur une proportion (dans le cas où X suit une loi de Bernoulli)
— un test d'indépendance (test du χ^2 d'indépendance)
— un test pour tester la normalité (= le caractère gaussien) d'un échantillon : le test de Shapiro-Wilk pour des échantillons de taille inférieure à 50.