TP 11 : Intervalles de confiance et tests statistiques

Exemple introductif :
Soit la série statistique S=[3, 6, 3, 2, 3, 7, 3, 4, 6, 4, 4, 5, 7, 3, 5, 3, 8, 4, 2, 5, 3, 5,4, 3, 4, 4, 5, 1, 6, 7].
Cette série statistique représente la mesure d'un même caractère X chez 30 individus.
Le but est de savoir si l'hypothèse " la moyenne du caractère étudié est $\mu = 4.5$" est raisonnable.

Formalisme mathématique :
On introduit $X_1, ..., X_{30}$ 30 variables aléatoires indépendantes suivant la même loi que X (= loi du caractère étudié). On pourra noter μ et σ^2 la moyenne et la variance théoriques inconnues de la loi de X.

La moyenne empirique est la variable aléatoire $M = \frac{1}{30}\sum_{k=1}^{30} X_k$, et l'écart-type corrigé empirique est la variable aléatoire $S' = \sqrt{\frac{1}{29}\sum_{k=1}^{30}(X_k - M)^2}$.

La série statistique est constituée des 30 observations $x_1, ..., x_{30}$ des variables aléatoires X_k correspondantes.

1. Rappeler l'espérance et la variance théoriques de la moyenne empirique M en fonction de μ et σ^2.
2. Calculer (à l'aide de python) la moyenne empirique observée de l'échantillon (=série statistique) $\overline{m} = \frac{1}{30}\sum_{k=1}^{30} x_k$, l'écart-type corrigé observé de l'échantillon $s' = \sqrt{\frac{1}{29}\sum_{k=1}^{30}(x_k - \overline{m})^2}$, ainsi que l'écart-type théorique de la moyenne de l'échantillon $\frac{s'}{\sqrt{30}}$ (cf 1.)
3. A l'aide du théorème central limite (3e forme, rappelé ci-dessous) déterminer l'intervalle de confiance à 95% sur μ. Autrement dit, trouver un intervalle du type $[\overline{m} - a, \overline{m} + a]$ tel que $P(\mu \in [\overline{m} - a, \overline{m} + a]) \geq 0.95$.
   Rappels :
   (a) Théorème central limite 3e forme : la loi de $\dfrac{M - \mu}{\frac{S'}{\sqrt{30}}}$ est proche de la loi $\mathcal{N}(0, 1)$.
   (b) si $Z \hookrightarrow \mathcal{N}(0, 1)$, $P(-1.96 < Z < 1.96) = 0.95$.
4. Que pensez-vous de l'hypothèse "$\mu = 4.5$" : est-elle plausible ?
   Vous venez de faire ainsi votre premier "test statistique", au risque $\alpha = 0.05$: attention, le risque mesure la probabilité de dire "non" au test, alors que la réponse théorique est "oui" et non l'inverse.
   Il faut prendre conscience que de répondre "oui" au test ne signifie pas que $\mu = 4.5$; cela signifie seulement que l'hypothèse $\mu = 4.5$ est plausible donc envisageable.
5. Refaire le test avec l'échantillon suivant et comparer :
   [4, 8, 4, 4, 4, 2, 1, 2, 3, 2, 2, 4, 6, 4, 2, 4, 1, 3, 1, 3, 3, 1, 4, 7, 4, 5, 4, 1, 3, 3].
   Pour information, les deux séries statistiques ont été obtenues en simulant 30 fois une loi de Poisson de paramètre 4 (donc de moyenne théorique 4).

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 TP, nous alllons étudier différents tests statistiques.

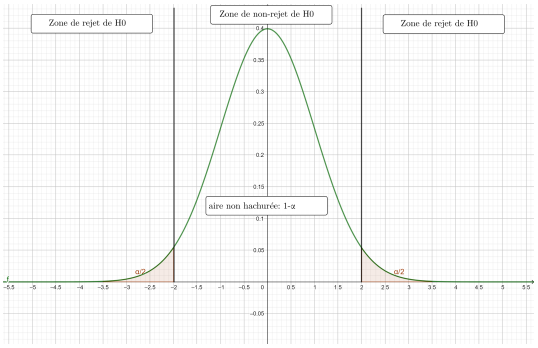
Test de conformité à la moyenne dans le cas où l'effectif de la population est grand (≥ 30).

C'est le test qui correspond à l'exemple introductif.

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 (d'où la 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 $P(\left(\mu \in \left]M_n - u\frac{\sigma}{\sqrt{n}}, M_n + u\frac{\sigma}{\sqrt{n}}\right[\right) \simeq 1 - \alpha)$.
  L'intervalle $\left]M_n - u\frac{\sigma}{\sqrt{n}}, M_n + u\frac{\sigma}{\sqrt{n}}\right[$ est appelé intervalle de confiance pour μ au risque α.
- 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 $S_n'^2 = \dfrac{1}{n-1}\sum_{k=1}^{n}(X_k - M_n)^2$. 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.

Exercice 1:

1. Ecrire une fonction `Estimateurs` prenant en entrée une liste L des valeurs observées de l'échantillon, et donnant en sortie la moyenne empirique observée m et la variance empirique corrigée observée s^2.
2. Ecrire une fonction `Test_Moy` prenant en entrée une liste L des valeurs observées de l'échantillon de grande taille et une valeur théorique m_0, puis donnant en sortie un booléen : *False* si l'hypothèse est rejetée à la suite de l'application d'un test de conformité de la moyenne au risque $\alpha = 5\%$ et *True* sinon.
3. Cas où X suit une loi géométrique de paramètre $p \in]0,1[$:
   (a) Ecrire une fonction `Genere_Geo` prenant en entrée un entier N et un réel p de]0,1[, et donnant en sortie une liste L de N valeurs aléatoires de X.
   (b) Dans le cas où $p = 0.1$, simuler différents échantillons de loi géométrique avec $N = 100$ et effectuer pour chacun d'eux un test de conformité de la moyenne d'hypothèse $H_0 : \mu = \frac{1}{0.1}$. L'accepte-t-on souvent ? Refaire ces tests, lorsque les échantillons sont de taille $N = 30$. Constater.
   (c) *Bonus :* faire afficher lors des tests précédents l'intervalle de non-rejet, pour réaliser qu'il augmente lorsque N diminue, et pour voir où M_n^* se situe dans cet intervalle.
   (d) Effectuer alors des tests de conformité de la moyenne d'hypothèse $H_0 : \mu = m_0$ avec $m_0 \neq \frac{1}{0.1}$ à choisir. On pourra prendre $N = 30$ et commencer par choisir une valeur m_0 proche de celle de $1/p$ avant d'en choisir une un peu moins proche ...

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 la loi normale de paramètres μ, σ^2 devient :
norm.pdf(x,μ,σ), *norm.cdf(x,μ,σ)*, *norm.ppf(x,μ,σ)*, *norm.rvs(μ,σ)*.
Rappel : pour simuler une loi normale, on peut aussi utiliser la syntaxe `rd.gauss` de la bibliothèque `random`.

Exercice 2:

Tester les commandes ci-dessus. Retrouver en particulier le quantile $u = 1.96$ pour obtenir le risque $\alpha = 5\%$. Que devient u lorsque $\alpha = 1\%$, $\alpha = 10\%$?

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, $\dfrac{M_n - \mu}{\frac{\sigma}{n}} \hookrightarrow \mathcal{N}(0,1)$.
  On est ainsi ramené au cas précédent.

  lorsque la variance σ^2 est inconnue alors $\dfrac{M_n - \mu}{\frac{S'_n}{n}} \hookrightarrow \mathcal{T}(n-1)$, 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 $\mu_0 \notin]m_n - u\frac{s'_n}{\sqrt{n}}, m_n + u\frac{s'_n}{\sqrt{n}}[$ avec u tel que $P(-u < T < u) = 1-\alpha$ 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 vaille 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 à d degrés de liberté : importation du module *from scipy.stats import t*

t.pdf(x,d) *t.cdf(x,d)* *t.ppf(x,d)* *t.rvs(d)*

Exercice 3:

1. Déterminer la valeur de u telle que $P(-u < T < u) = 0.95$ où $T \hookrightarrow \mathcal{T}(10)$.
2. Même question lorsque $T \hookrightarrow \mathcal{T}(20)$, puis $T \hookrightarrow \mathcal{T}(30)$.
   Que pouvez-vous en déduire sur la longueur des intervalles de non-rejet lorsque le nombre de degrés de liberté augmente ? Etait-ce prévisible ?
3. Ecrire une fonction `Graphe_Student` prenant en entrée un réel d et affichant le graphe d'une densité d'une loi de Student à d degrés de liberté.

Exercice 4:
Une biologiste étudie un type d'algue qui attaque les plantes marines. La toxine contenue dans cette algue est obtenue sous forme d'une solution organique. Elle mesure la quantité de toxine par gramme de solution, et obtient les mesures suivantes, exprimées en milligrammes :

1 1.2 0.8 0.6 1.1 1.2 0.9 1.5 0.9 1 1.2

On suppose que la quantité de toxine par gramme de solution suit une loi gaussienne, de variance inconnue. L'hypothèse que la biologiste souhaite tester est : H_0 "$m = 1$".

1. Déterminer, à l'aide des données, l'intervalle de non-rejet pour le test de conformité de la moyenne de Student au risque $\alpha = 5\%$. Conclure quant à l'hypothèse testée par la biologiste.
2. Dans cette question, on suppose que la variance est connue et égale à 0.05.
   Reprendre la question précédente, en adaptant le test. Comparer les deux intervalles.

Parmi les tests "classiques", on peut encore citer un test de conformité à une distribution discrète donnée, un test de comparaison de 2 moyennes dans le cas gaussien (cf document TIPE), un test de conformité de la variance (sous H_0 "$\sigma = \sigma_0$", on a le résultat mathématique $\frac{nS_n^2}{\sigma_0} \hookrightarrow \chi^2(n-1)$), et un test d'indépendance.