Comparaison d'un échantillon de mesures à une distribution théorique

Objectifs

Les objectifs de ce TP numérique sont :

i) de tester la méthode quantile-quantile sur divers échantillons ;

ii) d'évaluer l'adéquation d'un échantillon empirique à une loi de probabilité par un test d'hypothèse (test de Kolmogorov-Smirnov)

iii) de tester plusieurs échantillons.

La programmation sera réalisée en `python 3`. Il est demandé d'écrire un script faisant appel à des fonctions élémentaires (fonctions graphiques ou de calcul). Dans votre répertoire de travail, créer les répertoires `proba`, `plots` et `data` (s'ils n'existent pas). Les scripts python et les modules de fonctions seront écrits dans des fichiers séparés, placés dans le répertoire `proba`. Les scripts importeront les modules de fonctions (directive `import`). Les données sont dans le répertoire `data`, les fichiers graphiques seront placés dans le répertoire `plots`.

Outils python

`python` dispose de nombreux outils statistiques notamment dans les modules `numpy` (alias `np`) et `scipy.stats` (alias `scs`).
Dans ce TP nous aurons besoin des fonctions suivantes de la bibliothèque `scipy.stats` :
– `x=scs.<loi>.rvs(...size=1000)` pour générer 1000 tirages d'une v.a. suivant la loi `<loi>`.
– `p=scs.<loi>.cdf(x)` renvoie la probabilité p qu'une v.a. suivant la distribution «`loi`» soit inférieure à x (i.e. fonction de répartition)
– `x=scs.<loi>.ppf(p)` renvoie l'inverse de la fonction de répartition, i.e. la valeur seuil x associée à une probabilité p qu'une v.a. suivant la distribution «`loi`» soit inférieure à x.
On consultera avec avantage la documentation de ces fonctions avant usage `help(fonction)` ou `?fonction`.

Travail à faire

a) Lecture/importation de données

1. Créer le script `stat_tp.py`. Importer les modules numpy (alias `np`), scipy.stats (alias `scs`) et matplotlib.pyplot (alias `mpl`). Charger éventuellement les bibliothèques de fonctions que vous avez créées précédemment (vous aurez besoin de la fonction «`fonction_repartition`»).

2. Charger le fichier «`tpparis.dat`» dans le répertoire `data`. Examiner le fichier `tpparis.dat`. Que contient-il ?
Stocker les données dans le tableau `data` : `data = np.loadtxt("../data/tpparis.dat")`
Ranger la première colonne dans le tableau `an`, les 12 colonnes suivantes dans le tableau 2D `tp`.
Tracer sur une même figure les températures des mois de janvier et juillet. Sauvegarder les figures en format `pdf` et imprimer les.

b) Estimations ponctuelles et intervalles de confiance

1. Estimer les températures moyennes des mois de janvier et juillet.
2. Calculer les intervalles de confiances à 95% de ces deux températures.

3. Calcul des anomalies de température. Création du vecteur `atp` contenant les anomalies de températures. Les anomalies sont définies comme les fluctuations de température relativement aux moyennes mensuelles (données centrées).

1. Soustraire de chaque colonne la moyenne de la colonne.
2. Créer le vecteur `atp` contenant les anomalies (fonction `reshape`).
3. Tracer l'histogramme des anomalies de température.

c) Comparaison d'un échantillon à une loi donnée : méthode graphique : QQ plot

La fonction `qq_plot` (tracé Quantiles-Quantiles)
– calcule la fonction de répartition empirique d'un vecteur X passé en argument (données `atp` ou autre) (utilisez la fonction `fonction_repartition` que vous avez créée lors d'un précédent TP) ;
– calcule les quantiles de la loi normale centrée-réduite associés aux valeurs de la fonction de répartition empirique (méthode `ppf`) ;
– trace les quantiles du vecteur X en fonction des quantiles de la loi normale, i.e. $F_{\text{exp}}^{-1}(p)$ en fonction de $F_{\text{th}}^{-1}(p)$.

Coder la fonction `qq_plot` dans une fonction.

4. Définir une séquence X de n nombres (pseudo) aléatoires normalement distribués de moyenne 5 et d'écart type 2, n étant la taille du vecteur `atp`.
`scs.norm.rvs(size=len(atp), loc=..., scale=...)`

5. Afficher le QQ-plot pour le vecteur X. Interpréter ce graphique : déduire visuellement la moyenne et l'écart type de la séquence X.

6. Afficher le QQ-plot pour le vecteur `atp`. Conclusions ?

7. Comparer le vecteur `atp` à une loi exponentielle de paramètre $\alpha = 1$ (i.e. remplacer les quantiles de la loi normale centrée-réduite par les quantiles de la loi exponentielle de paramètre $\lambda = 1$).

d) Test d'ajustement : comparaison d'un échantillon à une loi donnée

Test de Kolmogorov-Smirnov

Le test de **Kolmogorov-Smirnov** est un test d'hypothèse utilisé pour déterminer si un échantillon est distribué suivant une loi de probabilité donnée, ou bien si deux échantillons suivent la même loi. On teste l'hypothèse H_0={l'échantillon suit la loi} contre l'hypothèse H_1={l'échantillon ne suit pas la loi}. Le principe consiste à comparer la fonction de répartition empirique $F_n(x_i)$ à la fonction de répartition théorique (de la loi supposée) $F(x_i)$. Kolmogorov a montré qu'en cas d'adéquation, le sup de la différence entre les deux fonctions de répartitions suit une loi assymptotique ne dépendant pas de la loi supposée :

$$\Pr\left[\sqrt{n}D_n > c\right] \xrightarrow[n\to\infty]{} \alpha(c) = 2\sum_{r=1}^{+\infty}(-1)^{r-1}\exp(-2c^2r^2)$$

où $D_n = \sup_x |F_n(x) - F(x)|$. Autrement dit, la probabilité α que la statistique $\sqrt{n}D_n$ soit supérieure au seuil c ne dépend que de c. Le principe du test consiste à choisir un seuil c suffisament grand, tel que la probabilité α de dépasser ce seuil sous l'hypothèse H_0 soit petite. La probabilité $\alpha(c)$, choisie à

priori (5% par exemple), est associée de façon univoque au seuil c. Elle définit le risque de se tromper, i.e. de rejeter H_0, alors que H_0 est vraie.
Pratiquement, on calcule la probabilité associée au seuil $\sqrt{n}D_n$. Si cette probabilité est inférieure au risque α choisi à priori, la différence $\sqrt{n}D_n$ est supérieure au seuil critique $c(\alpha)$. On ne peut retenir H_0 dans ce cas. Dans le cas contraire, on ne peut rejeter H_0.

8. Écrire une fonction `test_KS` permettant de :
- centrer et réduire les données d'un vecteur X passé en argument (données `atp` ou autre). Soit X_C le vecteur centré-réduit ;
- trier la séquence des X_C et calculer la fonction de répartition empirique F_n ;
- calculer les valeurs de la fonction de répartition d'une distribution normale centrée réduite, F, pour les quantiles de l'échantillon (méthode `scs.norm.ppf`) ;
- calculer D_n, sup de la différence $|F_n - F|$, en prenant en compte les discontinuités de la fonction de répartition empirique[(1)] ;
- calculer la probabilité $\Pr[\sup_x |F_n(x) - F(x)| > D_n]$;
- conclure.

9. Tester par le test KS l'hypothèse H_0 : l'échantillon X est normalement distribué

10. Tester par le test KS l'hypothèse H_0 : l'échantillon `atp` est normalement distribué.

Test du χ^2

11. Reprendre les deux questions précédentes avec un test du χ^2.

[(1)] Comme la fonction empirique F_n est discontinue en x_i, il convient de comparer la distribution théorique F à la limite à gauche et à droite de F_n, c'est à dire à $(i-1)/n$ et à i/n. Pratiquement, on définira D_n comme le max des distances $D_- = |F(x_i) - \frac{i-1}{n}|$ et $D_+ = |F(x_i) - \frac{i}{n}|$.