UGA L3 STE module de mathématiques appliquées 2017-18
– Travaux pratiques –
Séance n°1 du 18 octobre 2017 : statistiques descriptives ; algorithme
– Compte rendu –

2. Exercice 1 : description statistique des températures à Genève (cf. cours de maths n°4)

version avec tableur

2.1 Les formats de données
Format 'brut' en texte ASCII avec méta-données (entête, référence, etc.).
Les problèmes du format texte : le séparateur de colonnes (CSV) ; le séparateur décimal; les méta-données (mélange type texte et numérique); le codage des sauts de ligne : CR vs. LF.

* Données 'tabulées' c'est à dire mises sous forme de tableau, avec comme séparateur de colonne : tabulation (format TSV), point virgule (format CSV), sinon espaces. A l'importation des données, utiliser le bon séparateur et veiller à ce que les colonnes soient correctement séparées.
* Séparateur décimal : le format international est le point. Si vous êtes en format français (virgule), deux possibilités : soit garder ce format et remplacer dans les données les points par des virgules ; soit changer de format en modifiant les préférences du tableur.

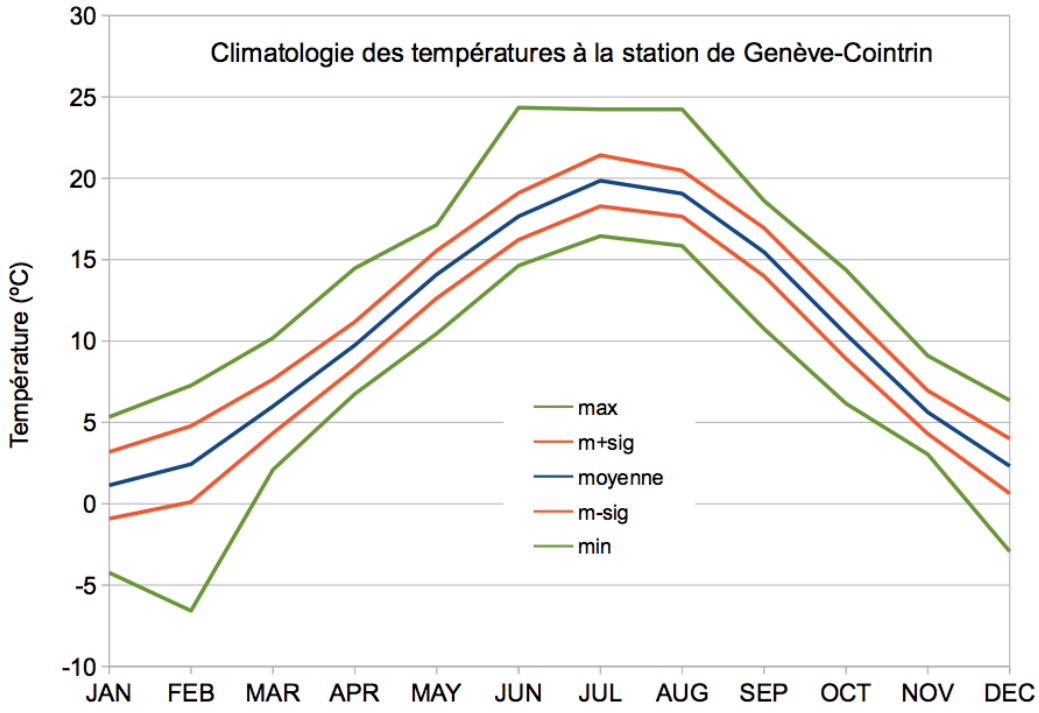
2.4 Faire un graphe des températures des mois de juin. A quoi correspondent les valeurs de 999.9 ? Les remplacer (comment et par quoi ?).

* Très important : très souvent il manque des données dans les séries de mesures. Parce que les données manquent ou sont parcellaires, ou à cause de problème sur les mesures qui les a exclues (les données sont vérifiés et leur homogénéité est testée). L'absence de données est 'codée' par une valeur spécifique, très différente des données elles-mêmes, normalement indiquée dans les méta-données (ici 999.9). D'où la nécessité d'explorer le fichier de données avec des graphes, ce qui permet de détecter de telles valeurs.

Comment les traiter avec un tableur ? Un tableur supporte les cellules vides, qu'il traite comme une absence de valeur (prise en compte dans les fonctions, par ex. MOYENNE), donc la solution est simple. Pour cela, remplacer toutes les valeurs 999.9 par un vide.

2.5 Calculer, avec toutes les années disponibles sur la période 1880-2017 : moyenne, écart-type, minimum et maximum, pour chaque mois.

2.6 Faire un graphe des 'normales saisonnières' (moyennes mensuelles) pour Genève. Ajouter les courbes correspondant à ±1σ, les valeurs maximales et minimales.



NB : Type de graphe ‘courbe’ ou autre dénomination, avec seulement Y numérique.

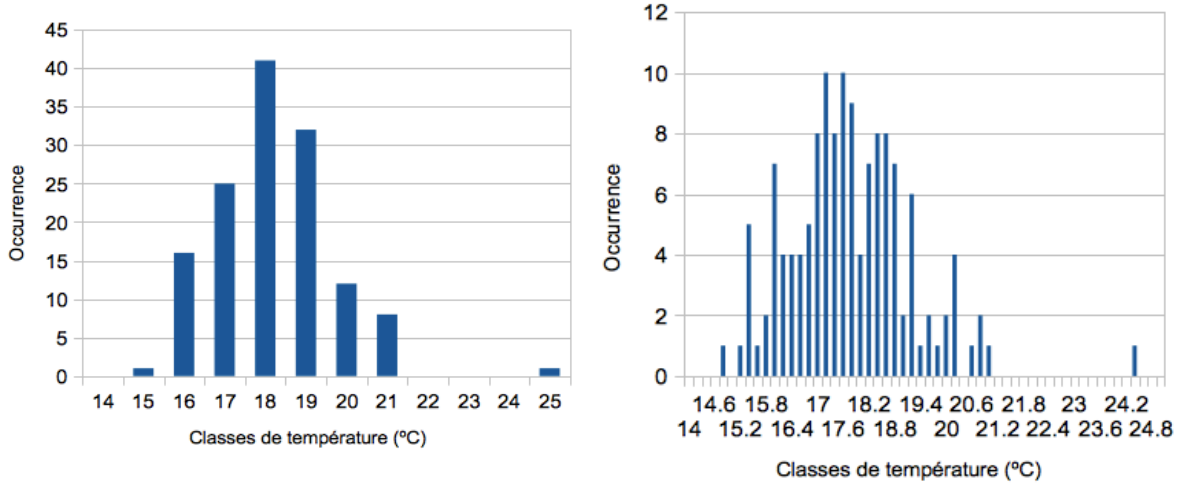
Utiliser les mois comme ‘étiquettes’.

Courbes à ordonner pour que la légende corresponde à leur position sur le graphe.

2.7 Tracer un histogramme des températures des mois de juin. Comment sont distribuées ces valeurs ? Déterminer les paramètres de la loi de distribution normale (gaussienne) la plus proche. L’ajouter (si possible) à l’histogramme.

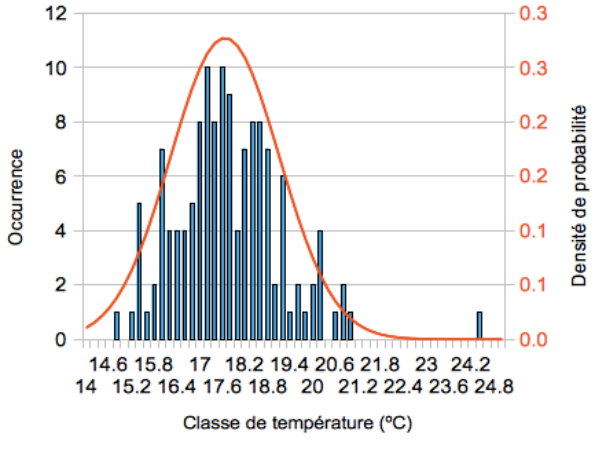
Histogramme : représentation graphique de la répartition des valeurs, exactement du nombre de valeurs par classe. Il faut donc définir des classes, par leurs bornes, calculer le nombre (occurrence) de valeurs dans chaque classe (fonction FREQUENCE, matricielle), et représenter ces résultats sous forme graphique (graphe de type ‘barres’ ou ‘histogramme’). Les fonctions numériques de type ‘HISTOGRAMME’ (HISTPLOT sous Scilab) font ce travail. L’intérêt de faire les calculs à la main est de contrôler les classes (bornes et largeur).

Les deux histogrammes ont été construits avec des pas différents (1ºC à gauche, 0.2ºC à droite), non normalisés (il s’agit des occurrences, non des fréquences, d’où les différences de valeurs en Y.

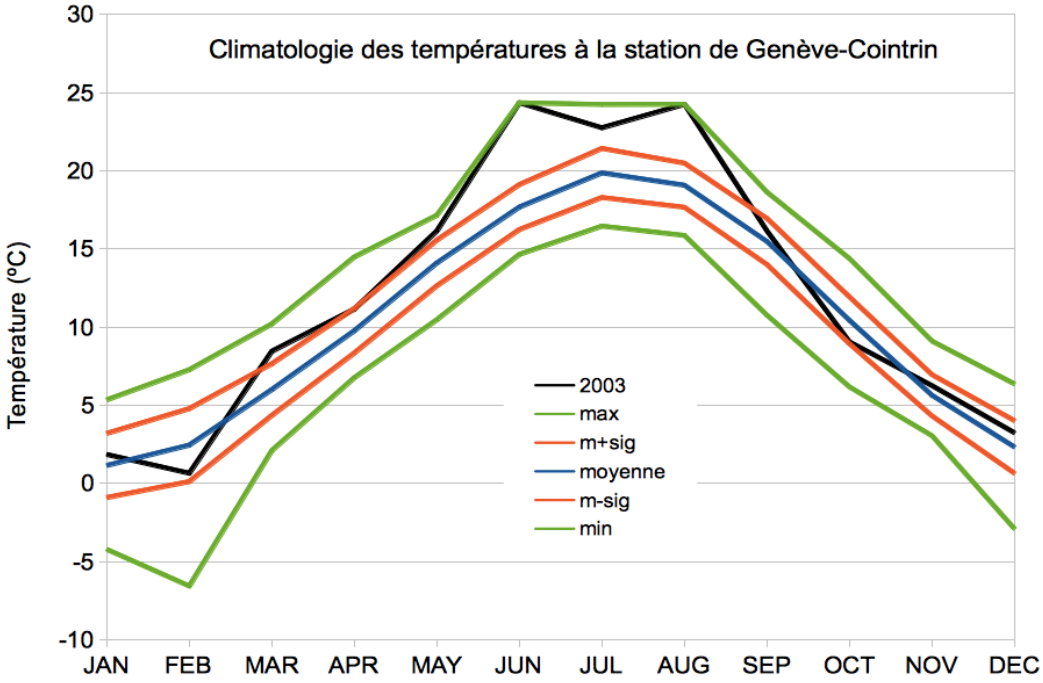


Ces distributions ont à peu près une forme en cloche, symétrique, ‘gaussienne’. Des tests statistiques permettent de comparer ces distributions observées (empiriques) avec une loi de distribution (normale, ici) pour discuter si ces distributions sont compatibles avec la loi.

On va ici supposer que ces températures du mois de juin suivent bien une loi normale, dont les paramètres (moyenne et écart-type) sont alors ceux des températures (m=17.7ºC, sig=1.4ºC). Le graphe ci-dessous superpose les deux.



2.8 A quelle année correspond le maximum des mois de juin ? Calculer le rapport $(T_{2003} - \overline{T}_{juin})/\sigma$. En supposant que la température suit une loi normale (cf. question précédente), calculer la probabilité que la température moyenne d'un mois de juin excède celle de 2003.

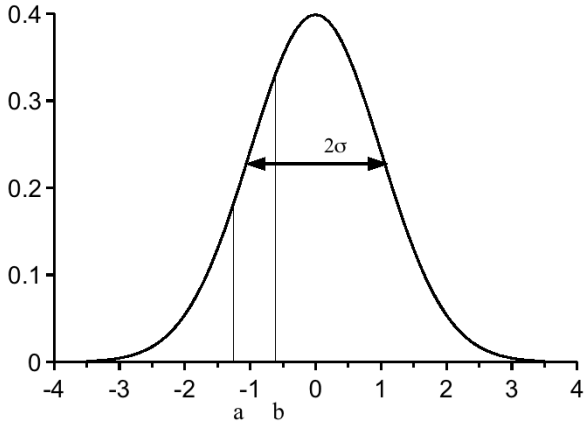


La courbe en noir correspond à l'année 2003 : hiver et automne dans la moyenne, printemps plutôt chaud, mais surtout été (JJA) très chaud avec valeurs correspondant aux maxima sur toute la série depuis 1880.

La température du mois de juin 2003 est de 24.3ºC. On cherche à associer à cet événement extrême une probabilité. Comme on utilise une loi continue (loi normale, ici), on ne peut calculer la probabilité d'une seule valeur (d'un événement), probabilité qui est nulle. Tout ce qui est possible, et a du sens ici, est de calculer la probabilité que la température du mois de juin soit supérieure ou égale à celle de 2003 soit $P(T>T_{2003})$. Il s'agit de la probabilité résiduelle sous la 'queue' de distribution, l'intégrale de T_{2003} à +inf de la densité de probabilité. Pour calculer cette probabilité, il faut utiliser la fonction normale cumulée, et l'appliquer : soit à la valeur T_{2003}=24.3ºC par rapport à la loi normale m=17.7ºC et sig=1.4ºC, soit à la distance normalisée (24.3-17.7)/1.4 = 4.6 par rapport à la loi normale m=0ºC et sig=1ºC. Dans les 2 cas on trouve $P(T>T_{2003}) \sim 1.8E\text{-}6$, soit une probabilité d'environ 1 chance sur 500000 pour qu'une température supérieure ou égale à T_{2003} survienne. C'est très faible, alors que l'événement a bien eu lieu. On voit bien ici la limite du modèle de loi normale pour représenter la distribution réelle des mois de juin. Des lois de distribution spécifiques existent pour mieux décrire les événements peu fréquents ou exceptionnels (utilisées notamment en hydrologie).

Page suivante : rappels sur la loi de distribution normale (cf. cours 4)

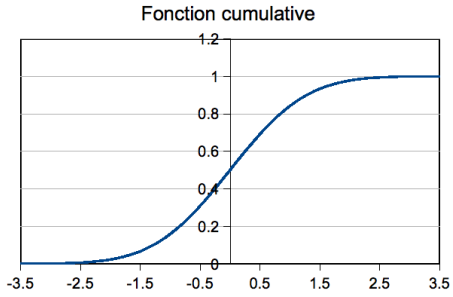
RAPPEL : la distribution gaussienne ou ‘normale’ (voir cours 4)



Il s'agit d'une loi de probabilité continue (modèle de distribution), dont la fonction (à droite) correspond à une densité de probabilité : il faut donc intégrer entre des bornes pour trouver une probabilité. C'est la fameuse 'courbe en cloche' représentée ci-contre, avec une largeur à mi-hauteur proche de deux fois l'écart type σ. La probabilité que x soit compris entre a et b est l'intégrale de la fonction de densité sur les bornes a-b (aire contenue sous la courbe). Comme la probabilité de tout x (entre - et +) doit être 1, l'intégrale de la fonction sur ces bornes (aire totale) est égale à 1.

La probabilité pour que x soit contenu autour de la moyenne à ±σ est 68%, à ±2σ 95%, et à ±3σ 99.7%.

La fonction de base, proposée par les tableurs, est centrée sur zéro (moyenne de x), et son écart type est de 1. Pour utiliser une fonction de moyenne m et d’écart type σ : soit ces paramètres sont acceptés par la fonction, soit il faut modifier la fonction de base f selon : f×σ + m.



Une autre façon d’utiliser cette distribution est par sa fonction cumulative, qui associe à tout x l'intégrale de - à x, soit la probabilité des valeurs inférieures à x : graphe ci-contre. Cette fonction existe également dans les tableurs, elle est souvent plus utile que la densité de probabilité.

======