Statistiques élémentaires et probabilités avec Mathematica

Dans ce notebook nous traitons de statistiques et probabilités simples, relatives à une variable, comme la population en 2018 des pays européens, les hauteurs de précipitation recueillies en 2020 dans une station météorologique du Sud de la France.

La première question que doit se poser tout scientifique est la suivante : À quelle loi de distribution statistique obéissent les données dont je dispose? Trois raisons imposent ce traitement préalable. D'abord, la connaissance de la distribution statistique fournit des informations précieuses sur les mécanismes qui expliquent cette distribution. Tout géographe sait qu'une distribution gaussienne, dite normale ou en cloche, est la somme de nombreuses petites causes. Quand ces petites causes ne s'additionnent pas, mais se multiplient, une loi de répartition log-normale ou de type puissance ordonne les données. Enfin, de nombreuses techniques ne sont valides que pour des données gaussiennes.

Pour déterminer la loi de répartition de ses données, le géographe compare son échantillon à une répartition théorique, à une loi de probabilité. Quatre groupes de techniques permettent au géographe d'accomplir ce premier travail : la visualisation de graphiques, le calcul des moments, l'estimation probabiliste, et plus récemment l'apprentissage automatique.

L'élaboration d'un histogramme est utile, mais imparfaite. On complète cette analyse visuelle avec le graphique dit des boîtes à moustache (Box Whisker), qui montre les répartitions symétriques ou non symétriques. Enfin, les graphiques de probabilités, ou de quantiles, permettent de comparer un ensemble de données discrètes ou de données continues à diverses lois théoriques. Mais, tous ces traitements visuels demeurent imprécis.

La deuxième approche, par la détermination des moments, consiste à calculer les coefficients d'asymétrie [skewness] et d'aplatissement [kurtosis] qui donnent des informations sur la forme de la distribution. Un coefficient d'asymétrie nul définit une distribution symétrique. Quant au coefficient d'aplatissement, il souligne l'importance des traînes des distributions. Quand le coefficient d'aplatissement est supérieur à 3, la distribution est dite leptocurtique. Cela signifie qu'existe une plus forte concentration de données autour de la moyenne, mais aussi aux extrémités de la distribution. On parle de distribution à longues traînes.

La troisième approche établit un lien direct entre les données et une distribution théorique. L'objectif est de déterminer les paramètres d'une loi de distribution probabiliste, que le géographe considère comme conforme à ses données. Cependant, généralement plusieurs lois de distribution semblent convenir. Et pour faire le meilleur choix, les logiciels fournissent toute une batterie de tests. Avec le langage Mathematica le géographe peut comparer ses données avec environ 150 distributions théoriques, discrètes ou continues. Et depuis la version 10, il est concevable de créer des distributions théoriques qui mixent des distributions classiques. Face à cette profusion de lois probabilistes, en fonction du type de données qu'il analyse, et de leur signification, le géographe se limitera à quelques lois.

Enfin, dernier raffinement méthodologique, les techniques d'apprentissage automatique indiquent directement la meilleure loi de distribution, celle qui est la mieux adaptée aux données que le géographe examine.

Loi probabiliste d'une série statistique

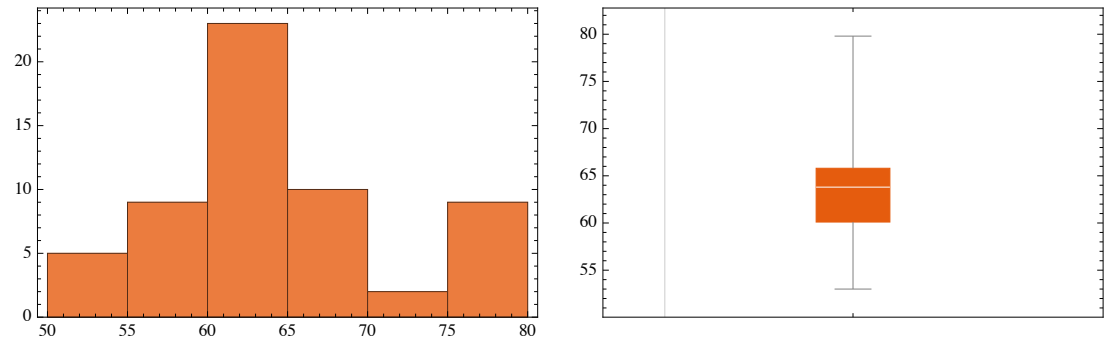
Une première approche visuelle

Dans le précédent cahier nous avons vu comment se servir des bases de données offertes par Mathematica. Nous pouvons les utiliser. La première ligne nettoie la mémoire en début de session. La deuxième ligne d'instruction collecte l'espérance de vie de la population des États africains. On enlève les États dont les données manquent avec la fonction **DeleteCases**[], et on ne retient que les valeurs. La ligne 3 affiche le nombre de données disponibles, soit 58. Les deux lignes suivantes, 4 et 5, dessinent l'histogramme et le graphique à moustaches des données. Ces graphiques ne sont pas affichés, car un point-virgule est placé à la fin des deux lignes d'instruction. La dernière ligne affiche les deux graphiques sur une même ligne.

In[336]:=
```
ClearAll["Global`*"]
data = DeleteCases[CountryData[#, "LifeExpectancy"] & /@ CountryData["Africa"],
    _Missing] // QuantityMagnitude;
Print["Nombre de données = ", Length[data]]
histo = Histogram[data, PlotTheme → "Scientific",
   AxesLabel → {"Espérance_de_vie", "Nombre_Etat"}];
bwc = BoxWhiskerChart[data, PlotTheme → "Scientific"];
GraphicsRow[{histo, bwc}]
```

```
Nombre de données = 58
```

Out[•]=



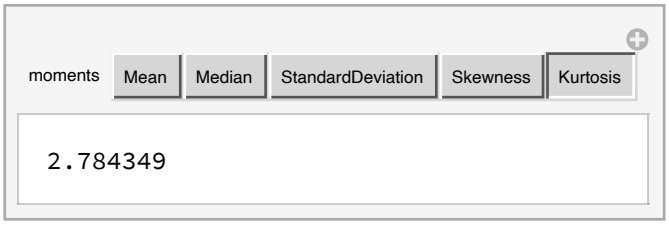
Ces deux graphiques font apparaître que la série des données est bien asymétrique. Nous allons le vérifier en calculant les moments de cette série.

Calcul des moments d'une série de données

Pour calculer ces moments, il est possible de généraliser ce que nous avons vu dans les précédents notebook, et utiliser la fonction **Manipulate**[].

In[342]:=
```
Manipulate[moments[data],
 {moments, {Mean, Median, StandardDeviation, Skewness, Kurtosis}},
 SaveDefinitions → True]
```

Out[•]=



En cliquant sur les cases de la ligne moments, on peut lire successivement les résultats suivants : moyenne = 64,35, médiane = 63,8, écart-type =7,10 , aplatissement = 0,62, asymétrie =2,78.
Ces résultats confirment ceux issus du traitement visuel. Mais, il est difficile d'apprécier ce qu'enseignent les écarts par rapport à la loi normale ou gaussienne. Il est alors nécessaire de tester si les données suivent une loi normale.

Tester la normalité des données

Il existe de très nombreux tests pour vérifier la normalité d'une série. La fonction **DistributionFitTest[]** choisit le mieux adapté à la série, et donne le résultat.

In[359]:=
```
DistributionFitTest[data, Automatic, "AutomaticTest"]
DistributionFitTest[data]
```

Out[•]=
```
CramerVonMises
```

Out[•]=
```
0.00083402598
```

Pour ces données, la fonction retient le test de Cramer von Mises. La valeur très faible du test indique que la distribution de la série n'est pas gaussienne. Le géographe est alors face à deux options.

Première option : le géographe connaît la loi de distribution

Dans ce cas, il estime les paramètres de la loi qui s'appliquent aux données avec la fonction **EstimatedDistribution[]**. Puis, il vérifie visuellement en comparant l'histogramme des données et la simulation réalisée avec les paramètres. Imaginons que les données de l'espérance de vie obéissent à une loi log-normale. En utilisant la fonction :

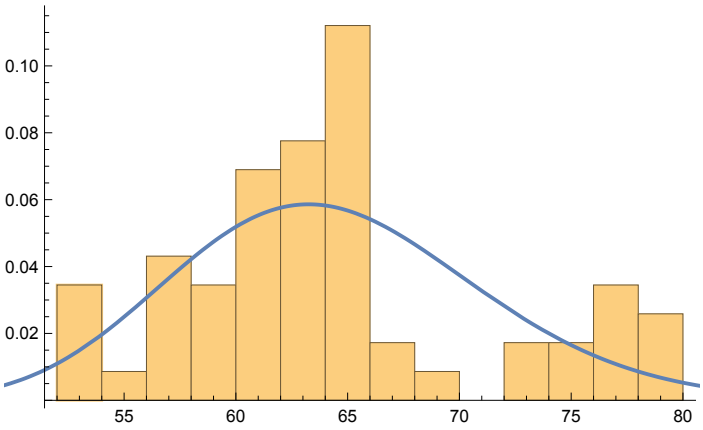
In[367]:=
```
edist = EstimatedDistribution[data, LogNormalDistribution[k, a]]
```

Out[•]=
```
LogNormalDistribution[4.1586355, 0.10699614]
```

On obtient les deux paramètres : k = 4,15 et a = 0,10. Pour comparer les données et le résultat de l'estimation, on se sert des instructions emboîtées ci-dessous

In[363]:=
```
Show[Histogram[data, 10, "ProbabilityDensity"],
 mon··· histogramme
 Plot[PDF[edist, x], {x, 0, 85}, PlotStyle → Thick]]
 tracé·· fonction de densité de probabilité  style de tracé  épais
```

Out[ ]=



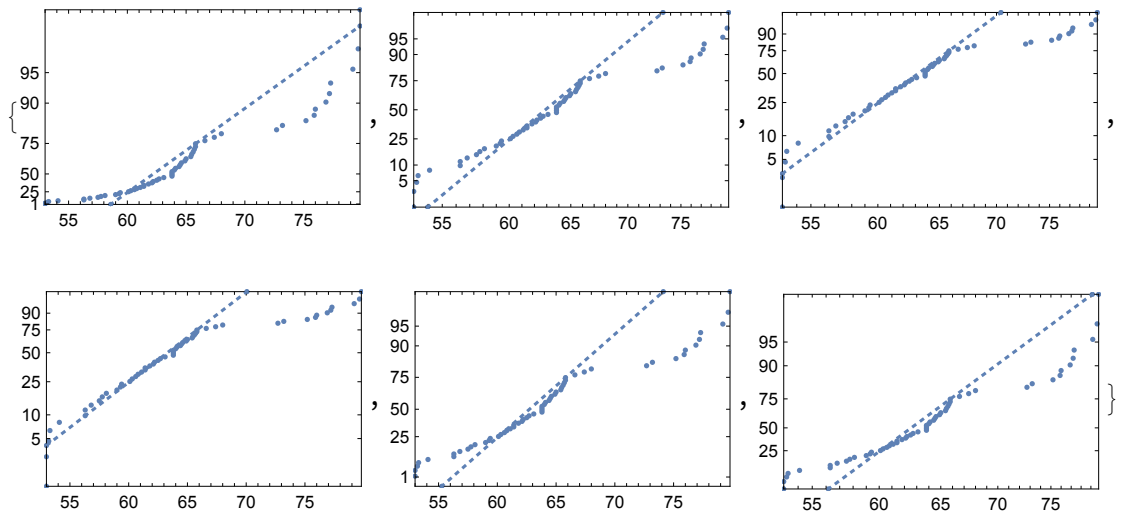
Nous observons deux gros écarts, pour les petites et les grandes valeurs. Il est alors plus sage de choisir la seconde option.

Seconde option : le géographe n'a aucune idée de la loi de distribution

Dans ce cas, il est possible de procéder à divers essais avec la fonction P**robabilityScalePlot**[] :

In[361]:=
```
{ProbabilityScalePlot[data, "Exponential"],
 tracé d'échelle de probabilité
 ProbabilityScalePlot[data, "LogNormal"],
 tracé d'échelle de probabilité
 ProbabilityScalePlot[data, "Weibull"], ProbabilityScalePlot[data, "Gumbel"],
 tracé d'échelle de probabilité  tracé d'échelle de probabilité
 ProbabilityScalePlot[data, "Rayleigh"], ProbabilityScalePlot[data, "Frechet"]}
 tracé d'échelle de probabilité  tracé d'échelle de probabilité
```

Out[ ]=



Comme aucune de ces lois ne semble correspondre, avec un écart qui s'amplifie pour les grandes valeurs, il est nécessaire d'utiliser la fonction **FindDistribution**[], qui donne des résultats par apprentissage automatique.

```
In[•]:= FindDistribution[data, 3]
        trouve distribution

Out[•]= {MixtureDistribution[{0.41920796, 0.58079204},
          {NormalDistribution[62.972403, 2.4238651],
           UniformDistribution[{52.98031, 79.853395}]}], MixtureDistribution[
          {0.81050259, 0.18949741}, {NormalDistribution[61.461871, 4.3624372],
           NormalDistribution[77.141901, 2.2090621]}],
         UniformDistribution[{52.98031, 79.853395}]}
```

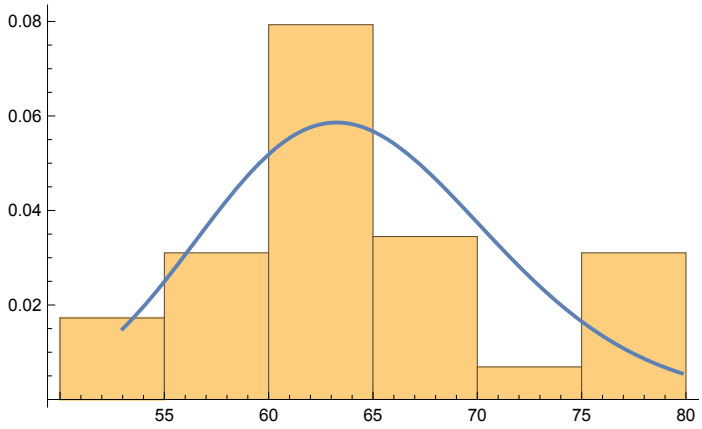
En remplaçant le nombre 3 par un autre nombre, il serait possible d'obtenir plus de distributions théoriques, avec les paramètres correspondants. Dans cet exemple, l'histogramme le laissait penser, nous obtenons trois fois des distributions mixtes. Ce qui signifie que les données agrègent deux catégories de population face à leur espérance de vie. Il resterait au géographe de repérer ces deux familles sur une carte.

Conclusion

Il est possible d'enrichir et de généraliser ces petits programmes en mobilisant les options des différentes fonctions. Pour les graphiques, il est possible de les rendre plus lisibles, notamment pour les publications. Pour certaines fonctions, il convient de changer certains paramètres pour les adapter aux données. Soit l'exemple de la ligne :

```
In[•]:= Show[Histogram[data, 7, "ProbabilityDensity"],
        mon··· histogramme
        Plot[PDF[edist, x], {x, Min[data], Max[data]}, PlotStyle → Thick]]
        tracé·· fonction de densité de p··· minimum maximum style de tracé épais
```

Out[•]=



Les nombres 0 et 85 doivent correspondre à une valeur égale ou inférieure et égale ou supérieure aux valeurs minimum et maximum repérées dans la série de donnée, et il est possible de changer la valeur du nombre 10 pour avoir plus ou moins de barres dans l'histogramme. Bien d'autres options sont disponibles.

Ouvrage recommandé :

Sergiy Suchok, 2015, *Mathematica Data Analysis*, Packt It