

Introduktion

I denna datorövning¹ ska vi undersöka ett antal dataset; `dist1.txt`, `dist2.txt`, `dist3.txt`. De tre filerna innehåller slumpmässiga urval från tre olika sannolikhetsfördelningar. Målet för övningen är att visualisera dem, samt att studera olika läges- och spridningsmått.

Python och användbara funktioner

Installation

Vi utgår från att du har antingen Python 2.7 eller Python 3.X installerat på din dator, samt modulerna `numpy`, `scipy` och `matplotlib`. Kolla t ex. på <http://www.scipy.org/install.html> för information om hur du installerar Python och modulerna.²

Att läsa in data

I den här övningen får ni data i `.txt`-filer med ett tal per rad. Liknande filer har ni använt er av i programmeringsdelen av kursen. Det är också enkelt att läsa in en sådan fil med hjälp av `numpy`:

```
In [1]: import numpy as np
In [2]: file_1 = np.loadtxt("dist1.txt")
In [3]: file_1.shape
Out[3]: (10000,)
In [4]: file_1
Out[4]:
array([[ 0.68417915,  0.53041328,  0.05499373, ...,  0.88936388,
         0.00749854,  0.5763201 ]])
```

Histogram

Ett vanligt sätt att visualisera data i en dimension är att använda histogram. Med `matplotlib` är det enkelt:

```
In [6]: import matplotlib.pyplot as plt
In [7]: d = [1,1,5,3,2,0.5]
In [8]: n_bins, edges_bins, _ = plt.hist(d, bins = 5)
In [9]: plt.show()
```

Se även <http://matplotlib.org/gallery.html> för mer hjälp, och dokumentationen; http://matplotlib.org/api/pyplot_api.html.

Uppgifter

- För var och en av de tre filerna:

¹Ursprungligen av Jimmy Vinbladh

²Andra länkar: <https://www.Python.org/downloads>, <http://matplotlib.org/users/installing.html>

- Läs in filen i Python
 - Plotta ett histogram av värdena. Hur tycker du fördelningen ser ut? Tänk på att bredden på histogrambinnarna kan påverka hur fördelningen ser ut.
 - Utan att använda numpy, beräkna medelvärde, varians och standardavvikelse för värdena. Jämför med det du förväntade dig utifrån vilken distribution du tror värdena kommer från. Du kan även räkna ut värdena med hjälp av numpy för att kontrollera att du har räknat rätt.
 - Kan du plotta fördelningen du tror att värdena kommer från över histogrammet? Kom ihåg att histogrammet inte är normaliserat, så du måste skala upp din fördelningsfunktion. Det är också möjligt att låta matplotlib normalisera histogrammet. Se dokumentationen.
- Prova att skapa egna slumpmässiga urval från fördelningar ni känner till. För att göra detta kan du t.ex. börja med att läsa här: <http://docs.scipy.org/doc/numpy/reference/routines.random.html#distributions> Räkna också ut medelvärde, varians och standardavvikelse.
 - I en ideell gas kan hastigheten på en partikel beskrivas av vektorn $\mathbf{v} = [v_x, v_y, v_z]$. Varje komponent är normalfördelad med en viss standardavvikelse σ_v .
 - Generera ett stort antal slumpmässiga partikel-vektorer och räkna ut $|\mathbf{v}|$. Hur ser fördelningen ut³?
 - Vad händer med distributionen när man varierar standardavvikelsen?

Hittills har du räknat på medelvärde och standardavvikelse som läges- och spridningsmått. I vissa sammanhang är medianen ett bättre lägesmått. Medianen är det värde där 50% av mätningarna är lägre, och 50% större. På samma sätt är den x -te percentilen det värde som $x\%$ av värdena är lägre än, och $100 - x\%$ av värdena är större.

- Vilket värde skulle x ha för att den x -te percentilen motsvarar medelvärdet minus en standardavvikelse? Vad är motsvarande värde på x som motsvarar medelvärdet plus en standardavvikelse?
- Är det någon av distributionerna ovan där det skulle vara bättre att använda median och percentiler än medelvärde och standardavvikelse?

³Hint: från termodynamiken vet vi att partikelhastigheten är fördelad utifrån Maxwell-Boltzmann distributionen.