Méthodes statistiques TD n°9-10-11

O. Dadoun dadoun@in2p3.fr semaine 17/11/2015 - 01/12/2015

Exercice

La distribution normale ou courbe de gauss est une distribution symétrique en forme de cloche ayant des propriétés mathématiques intéressantes. Elle est issue des travaux du 17ème siècle concernant la théorie des jeux. Le problème à résoudre était le suivant:
Si on joue à pile ou face, disons 10 fois, quelle est la probabilité d'obtenir : 1 face , 2 faces, … ?

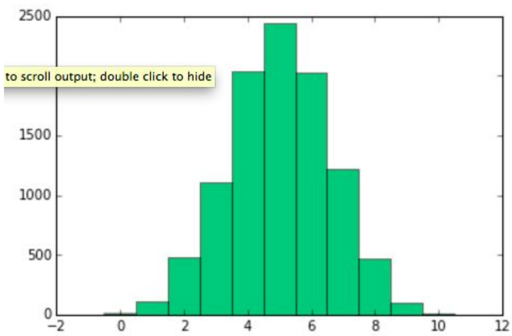
Avec une simulation Monte-Carlo sur 10k expériences nous avons:

Nb piles x_i	0	1	2	3	4	5	6	7	8	9	10
Effectifs n_i	8	105	477	1111	2040	2444	2020	1223	463	96	13
$n_i * x_i$	0	105	954	3333	8160	12220	12120	8561	3704	864	130
$n_i * x_i * x_i$	0	105	1908	9999	32640	61100	72720	59927	29632	7776	1300
Eff. cumulés	8	113	590	1701	3741	6185	8205	9428	9891	9987	10000

Calculer : moyenne, écart-type et médiane :

Moyenne : $\bar{x} = \frac{\sum x}{N}$ = 50151/10000 = 5.01

Écart type : $\sqrt{\sum_{i=1}^{i=p} \frac{n_i x_i^2}{N} - \bar{x}^2} = \sqrt{27.71 - 5.01 * 5.01}$ =1.61



Position de la médiane : 10001/2 = 5000.5 entre 5000.5 et 5001 donc médiane = 5 ah bon !!
Pour une loi normale la moyenne = la médiane.

En sciences humaines on observe souvent des distributions:

- plutôt symétriques autour de m
- avec une forme en cloche

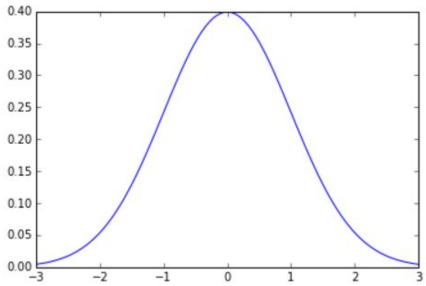
=> il s'agit d'une loi normale

Fonction de distribution de la loi normale (fonction gaussienne)

- Elle se définit avec 2 paramètres sa moyenne m et son écart-type s.
  On dira qu'une variable X suit une loi normale N(m,s)

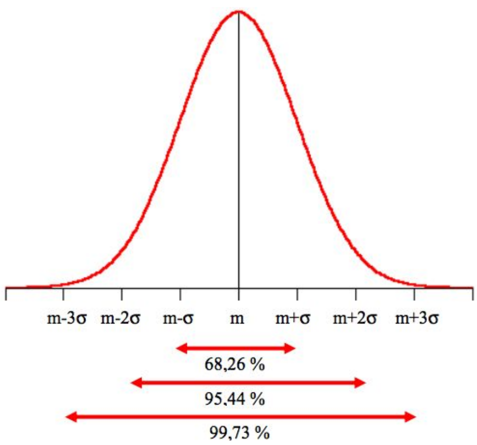
$$\mathbf{y} = \frac{\mathbf{1}}{\sigma\sqrt{\mathbf{2}\pi}}\mathbf{e}^{-\frac{1}{2}\left(\frac{\mathbf{x}-\mathbf{m}}{\sigma}\right)^2}$$

ici m=0 s=1



Pour tracer la courbe : en 0 y~0.4 et en x=1 y~0.24

- Courbe symétrique autour de la moyenne avec une forme en cloche
- Propriétés
  - moyenne=médianez
  - p (m - s < X < m + s) = 68,26 %
  - p (m - 2s < X < m + 2s) = 95,44 %
  - p (m - 3s < X < m + 3s) = 99,73 %



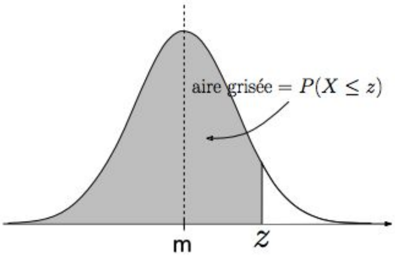
Cas particulier m=0 et s=1 : loi normale centrée réduite

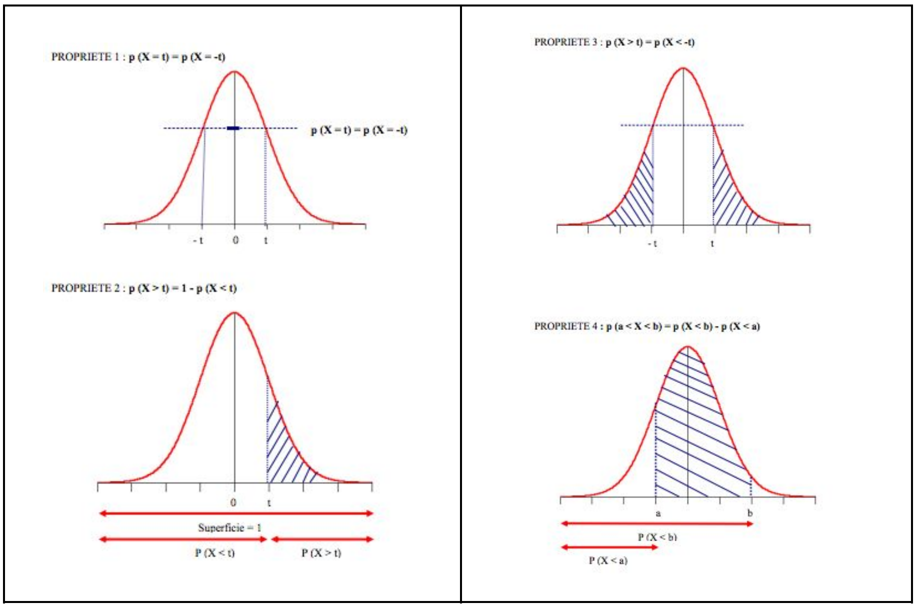
Figures issus de http://www.netprof.fr/content/videos/737/loinormale.pdf
A chaque moyenne et à chaque écart type correspond une loi normale.

La variable centrée réduite
Si X suit une loi normale de paramètre m et s alors la variable $z = \frac{x-\bar{x}}{S}$ suit une loi normale centrée réduite. C'est un changement de variable.

- des données indépendantes de l'unité ou de l'échelle choisie ;
- des variables ayant même moyenne et même dispersion.

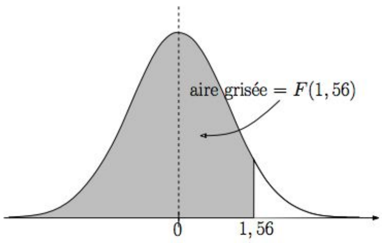
L'intérêt d'un tel changement de variable est qu'il existe des tables de la loi normale centrée réduite.
Ou z-score vous permet de dire à combien d'écarts-types, en dessus ou en dessous de la moyenne, se situe un échantillon d'une série de données.





<u>Exemple</u>

- On suppose une variable qui suit une loi normale centrée réduite. On cherche à savoir pour quelle proportion d'individus on a Z<=1.56 ? c.f fonction de répartition de la loi normale



La proportion totale de la courbe vaut 100%.
P(Z<=1.56) = 0.9406 c.a.d pour 94.06% des individus la variable Z est <= à 1.56
ou pour 5.94% des individus la variable est >= 1.56

Exercice
On suppose une variable qui suit une loi normale centrée en 11 avec un écart type de 2.
Pour quelle proportion d'individus est-ce que X<14 P(X<14) ?
On centre et on réduit X en faisant le changement de variable (X-11)/2.
P((X-11)/2 = (14-11/2))=P(Z<=1.5) = 93.32%

Exercice
Dans une distribution normale centrée réduite, quel est le pourcentage d'observations qui sont:

A. supérieures ou égales à z=1.5 ?
   1-0,9332 = 0.068
B. inférieures ou égales à z=1.8 ?
   0.9641
C. comprises entre z= -1 et z=1 ?
   2P(X<1) - 1 = 0.6826
D. comprises entre z=-2 et z=2 ?
   2P(X<2) - 1 = 0.9544
E. comprises entre z=-3 et z=3 ?
   2P(X<3) - 1 = 0.9974

Que peut-on dire des questions C, D et E ?
Si on remplace x par :

- $\bar{x} - s$ on trouve 1
- $\bar{x} - 2s$ on trouve 2
- $\bar{x} - 3s$ on trouve 3

Exercice

- Le temps nécessaire aux étudiants pour terminer une épreuve d'examen est une variable normale, de moyenne 90 min et d'écart-type 15 min. Quelle durée doit-on fixer à l'épreuve si l'on souhaite que 90% des étudiants puissent terminer dans le temps imparti ?
  P(z<z1)=0.9 c.a.d la plus proche valeur est P(z<z1)=90,15% c'est pour z1=0.09+1.2=1.29.
  z=(x-90)/15=1.29 => x=109.35 soit environ 1h50

- Les résultats à un test de raisonnement obéissent à une loi normale de moyenne 20 et de variance 36. Dans quel intervalle de variation se situent les 99.8% de résultats centrés sur la moyenne ?
  On veut donc l'intervalle tel que P(z>-z1) et P(z<z1) = 99.8% soit
  P(z<z1)-(1-P(z<z1))=2*P(z<z1)-1=99.8% =>P(z<z1)=1.998/2 c.a.d on doit chercher la valeur z1 tel que P(z<z1) = 0.999 avec le tableau on trouve que c'est pour z=0.08+3=3.08
  D'où la valeur recherchée est z=(x-20)/6=3.08=>x=6*3.08+20=38.48. On est donc à 18.48 de la moyenne.

Exercice
L'échelle de dépression du test M.M.P.I* est standardisée sur une moyenne de 50 et un écart-type de 10.

1. Si l'on considère qu'une note standard supérieure ou égale à 70 traduit un état pathologique, combien trouve-t-on de personnes pathologiquement dépressives dans une population de 30 000 personnes.
   z=(X-m)/s=(70-50)/10=2, P(z>2) =1 -P(z<2)=1-0,9772 = 2.28% de la population soit
   30 000*2.28%= 684 personnes

2. Même question si on considère comme inquiétante une note supérieur à 60
   $z=(X-m)/s=(60-50)/10=1$, $P(z>1) =1 -P(z<1)=1-0,8413$ = 15.87% de la population soit 30 000*15.87%= 4761 personnes

Minnesota multiphasic personality inventory: permet une exploration des différents aspects de la personnalité normale et pathologique.

<u>Exercice</u>

Les scores de QI au test de Wechsler suivent une loi normale de moyenne 100 et d'écart-type 15 points. Entre quelles limites se situent en pratique (i.e. sauf rares exceptions) les scores de l'ensemble de la population ?

On dit qu'à 3 écarts-types on à pratiquement l'ensemble de la population (99,73 %) les limites se situent donc entre 100-3*15 et 100+3*15 soit entre 55 et 145.

▪ Quelle est la probabilité de l'évènement : « score compris entre 100 et 115 » ?

Donc entre $z=(100-100)/15=0$ et $z=(115-100)/15=1$. On fait donc $P(z<1)$ et $P(z<0)$ que l'on lit sur la table: 0,8413-0.5=0.3413 soit environ 34%.

▪ Quelle est la probabilité de l'évènement : « score compris entre 85 et 115 » ?

Donc entre $z=(85-100)/15=-1$ et $z=(115-100)/15=1$. On fait donc $P(z<1)$ et $P(z<-1)$.
$P(z<-1)=P(z>1)=1-P(z<1)$, $P(z<1)-P(z<-1)=P(z<1)-1+P(z<1) =2*0,8413-1= 0.6826$.
Soit environ 68%.

▪ Quelle est la probabilité de l'évènement : « score compris entre 70 et 85 » ?

Donc entre $z=(70-100)/15=-2$ et $z=(85-100)/15=-1$. On fait donc $P(z<-1)$ et $P(z<-2)$.
$P(z<-1)=P(z>1)=1-P(z<1)$, $P(z<-2)=P(z>2)=1- P(z<2)$, $1-0.8413-1+0.9772=0.1359$
Soit environ 13.6%.

▪ Quelle est la probabilité de l'évènement : « score compris entre 70 et 130 » ?

Donc entre $z=(70-100)/15=-2$ et $z=(130-100)/15=2$. On fait donc $P(z<2)$ et $P(z<-2)$.
$P(z<-2)=P(z>2)=1-P(z<2)$, $P(z<2)-P(z<-2)=P(z<2)-1+P(z<2) =2*0.9772-1= 0.9544$.
Soit environ 95%.