Introduction à la biostatistique

Lois de probabilité, distributions, estimations

MASTER INFORMATIQUE MEDICALE

Kankoé SALLAH MD, PhD

kankoe.sallah@univ-amu.fr
kankoe@skml.fr

Nov 2021

Exemple de situations faisant appel aux lois de distribution

Quelle est la probabilité d'observer 3 malades dans un échantillon de 10 sujets choisis au hasard dans une population où la fréquence de la maladie est de 17% ?

→ **Loi binomiale**

Quelle est la probabilité qu'un individu tiré au hasard ait une taille comprise entre 150 et 170 cm, si les tailles sont distribuées suivant une loi normale de moyenne 160 et d'écart type 30 cm ?

→ **Loi normale**

Loi de Bernoulli

Experience de Bernoulli ou épreuve de Bernoulli

Une épreuve de Bernoulli est une expérience aléatoire qui n'a que deux issues possibles : soit oui ou non, soit pile ou face, soit succès ou échec. Chacune de ces issues a une probabilité déterminée de survenir.

Distribution de Bernoulli ou loi de Bernoulli

On dit qu'une variable aléatoire admet une distribution de Bernoulli si elle prend la valeur **1** avec une probabilité π et la valeur **0** avec la probabilité $1- \pi$

L'espérance (moyenne) de cette loi de Bernoulli est alors $\boldsymbol{\pi}$

La variance de cette loi de Bernoulli est alors $\boldsymbol{\pi(1- \pi)}$

Remarque : La loi de Bernoulli est la loi de la variable aléatoire qui code le résultat d'une épreuve qui n'admet que deux issues (épreuve de Bernoulli) : 1 pour « succès », 0 pour « échec », ou quel que soit le nom qu'on donne aux deux issues d'une telle expérience aléatoire.

Loi de Bernoulli

Exemples

1- Un lancer de dé peut-il constituer une expérience de Bernoulli ?
Réponse : Oui. Il suffit de définir pour cela une expérience à issue binaire. Par exemple, le fait d'obtenir ou non la face « trois »

2-Calculez la moyenne et l'écart type de l'expérience de Bernoulli traduite par :
« ***obtenir la face « 4 » ou non à l'issue d'un lancer de dé*** »
Réponse :

Moyenne = $\pi = \frac{1}{6}$ C'est la probabilité d'obtenir l'évènement d'intérêt

Ecart type = $\sqrt{\pi(1-\pi)} = \sqrt{\frac{1}{6} \times \left(1 - \frac{1}{6}\right)} = \frac{\sqrt{5}}{6}$

Loi binomiale

Définition

Si X_1, X_2, … , X_n sont des variables aléatoires de Bernoulli indépendantes , de paramètre ***p***, alors leur **somme** est aussi une variable aléatoire qui suit **la loi binomiale** représentée par X :

$$\boxed{X = \sum_{k=1}^{n} X_k \sim B(n,p)}$$

L'espérance (moyenne) de cette loi binomiale est alors $E(X) = np$

La variance de cette loi binomiale est alors $Var(X) = npq = np(1-p)$

q est défini comme la probabilité d'échec, c-a-d la probabilité de non-réalisation de l'évènement d'intérêt : *q=1-p*

Loi binomiale

On peut calculer la probabilité des issues d'une variable X suivant une loi binomiale de paramètres n et p

$$\mathrm{P}(X = \mathrm{k}) = \mathrm{C}_n^k\, p^k (1-p)^{n-k} \ \text{ avec } \ \mathrm{C}_n^k = \frac{n!}{k!(n-k)!}$$

La formule correspond à la probabilité d'obtenir k boules rouges en tirant n boules, *avec remises. p est le pourcentage des boules rouges.*

C_n^k est appelé nombre de combinaisons de *k* parmi *n* ou **coefficient binomial**. C'est encore le nombre de parties de *k* éléments dans un ensemble de *n* éléments

n! représente la factorielle de n, définie ci-dessous

$$n! = \prod_{1 \leqslant i \leqslant n} i = 1 \times 2 \times 3 \times \ldots \times (n-1) \times n.$$

Loi binomiale

$$\mathrm{P}(X = \mathrm{k}) = \mathrm{C}_n^k \, p^k (1-p)^{n-k} \ \text{ avec } \ \mathrm{C}_n^k = \frac{n!}{k!(n-k)!}$$

La formule étant complexe, cette probabilité peut être lue sur une table ou directement calculée sur le logiciel R (fonction dbinom)

$$dbinom(x = k, size = n, prob = p)$$

p= probabilité de réussite d'une expérience; n= nombre de tirages; k= nombre d'issues positives

Exemple

Probabilité d'avoir 5 «piles» sur 10 lancers de pièce

$$dbinom(x = 5, size = 10, prob = 0.5) = 0.25$$

Probabilité d'avoir 5 «piles» ou moins sur 10 lancers de pièce

pbinom(q=5, size=10, prob=0.5, lower.tail = TRUE) = 0.62

Loi binomiale

Similairement, la fonction ***pbinom()*** calcule la *somme des probabilités* d'obtenir des issues binomiales *jusqu'à un seuil* *(lower.tail=TRUE, sur R)* *ou* *au-delà d'un seuil* *(lower.tail=FALSE, sur R)*.

Exemples

1- Probabilité d'avoir 25 «pile» **ou moins** sur 50 tirages

$$pbinom(q = 25, size = 50, prob = 0.5, lower.tail = TRUE)$$

2- Probabilité d'avoir 25 «pile» **au moins** sur 50 tirages

$$pbinom(q = 25, size = 50, prob = 0.5, lower.tail = FALSE)$$

Loi binomiale

Exemples

Quelle est la probabilité d'observer 3 malades dans un échantillon de 10 sujets choisis au hasard dans une population où la fréquence de la maladie est de 17% ?

Loi binomiale

$$\mathrm{P}(X = \mathrm{k}) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

$$\mathrm{P}(X = 3Malades) = \frac{10!}{3!(10-3)!} \times 0.17^3 (1-0.17)^{10-3} = 0.16$$

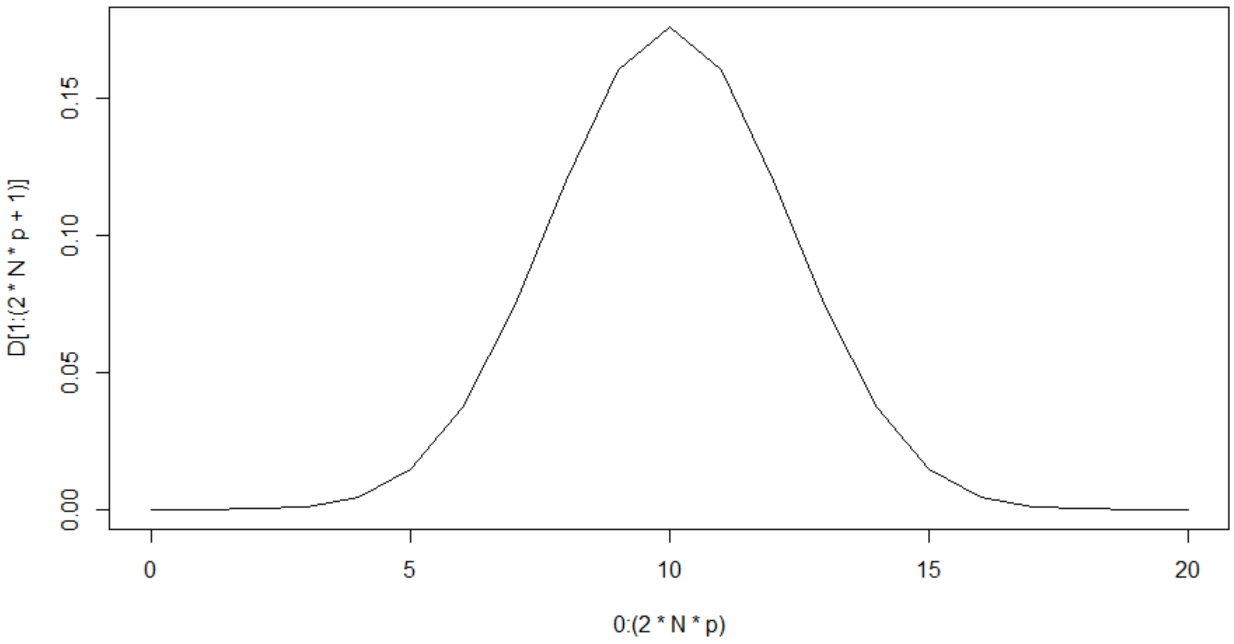
Sur R → ***dbinom(x=3, size=10, prob=0.17)***

De la loi binomiale vers la loi normale

Expérience: Comment se comporte une loi binomiale quand *np* et *nq* sont grands ?

Cette diapositive montre la distribution des probabilités d'une loi binomiale.
Vous n'êtes pas obligé de comprendre le code R

```
dbinom(x=3, size=20, prob=0.5, log = FALSE)
N=20; D=c(); p=0.5;
for (x in 0:N){D=c(D, dbinom(x=x, size=N, prob=p, log = FALSE))}
plot(0:(2*N*p), D[1:(2*N*p+1)], type="l", lty=1)
```



Ici np=10
et nq=10

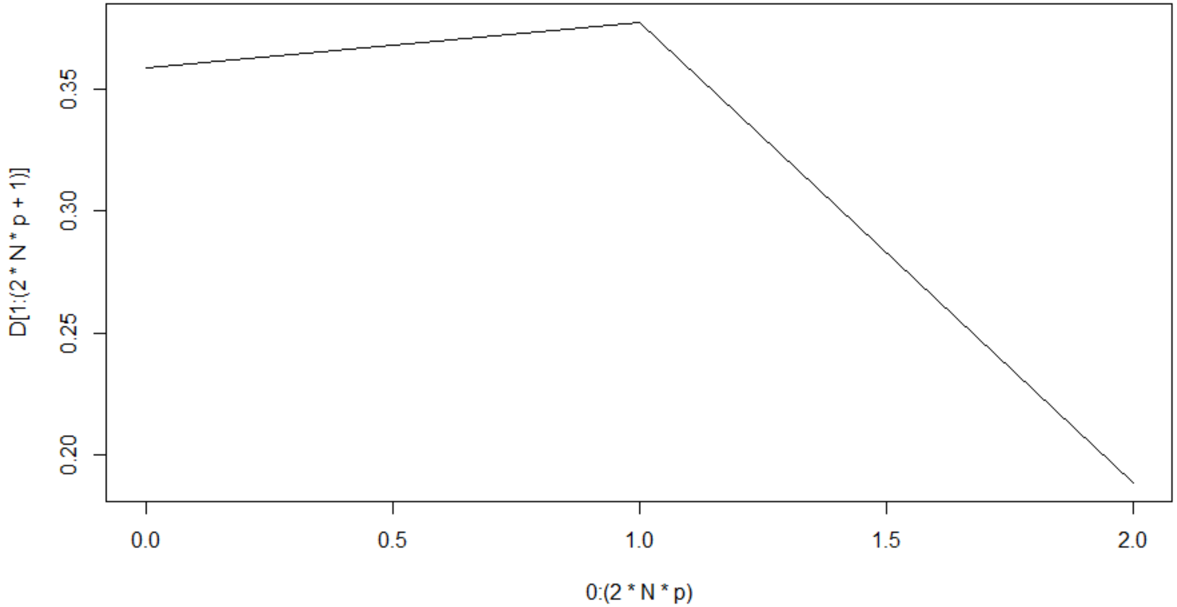
Appréciez l'allure de la distribution

De la loi binomiale vers la loi normale

Expérience : Comment se comporte une loi binomiale quand *np* et *nq* sont grands ?

Cette diapositive montre la distribution des probabilités d'une loi binomiale. Vous n'êtes pas obligé de comprendre le code R ayant servi à générer le graphique

```
N=20; D=c(); p=0.05;
for (x in 0:N){D=c(D, dbinom(x=x, size=N, prob=p, log = FALSE))}
plot(0:(2*N*p), D[1:(2*N*p+1)], type="l", lty=1)
```



Ici np=1
et nq=19

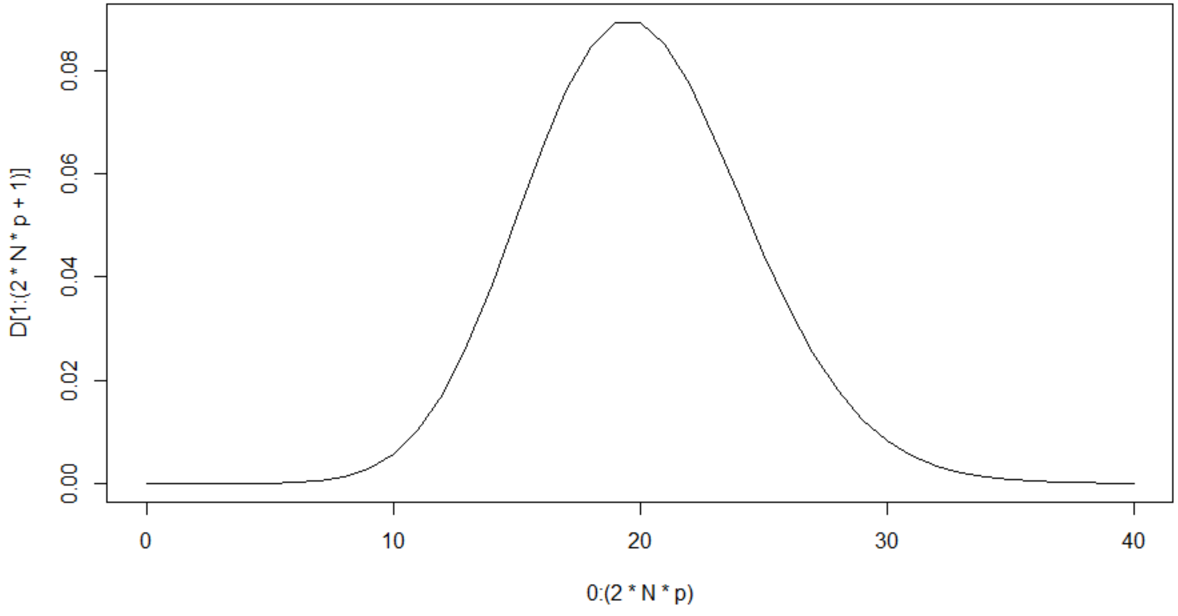
Appréciez l'allure de la distribution

De la loi binomiale vers la loi normale

Expérience : Comment se comporte une loi binomiale quand *np* et *nq* sont grands ?

Cette diapositive montre la distribution des probabilités d'une loi binomiale. Vous n'êtes pas obligé de comprendre le code R ayant servi à générer le graphique

```
N=2000; D=c(); p=0.01;
for (x in 0:N){D=c(D, dbinom(x=x, size=N, prob=p, log = FALSE))}
plot(0:N*p, D[1:(N*p)+1], type="l", lty=1)
```



Ici np=20
et nq=1980

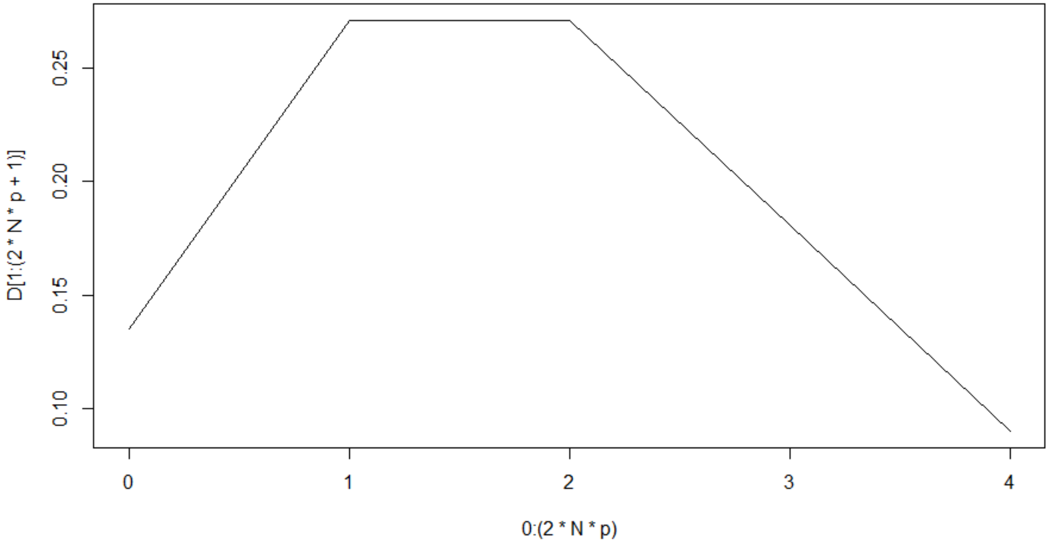
Appréciez l'allure de la distribution

De la loi binomiale vers la loi normale

Expérience : Comment se comporte une loi binomiale quand *np* et *nq* sont grands ?

Cette diapositive montre la distribution des probabilités d'une loi binomiale. Vous n'êtes pas obligé de comprendre le code R

```
N=2000; D=c(); p=0.001;
for (x in 0:N){D=c(D, dbinom(x=x, size=N, prob=p, log = FALSE))}
plot(0:(2*N*p), D[1:(2*N*p+1)], type="l", lty=1)
```

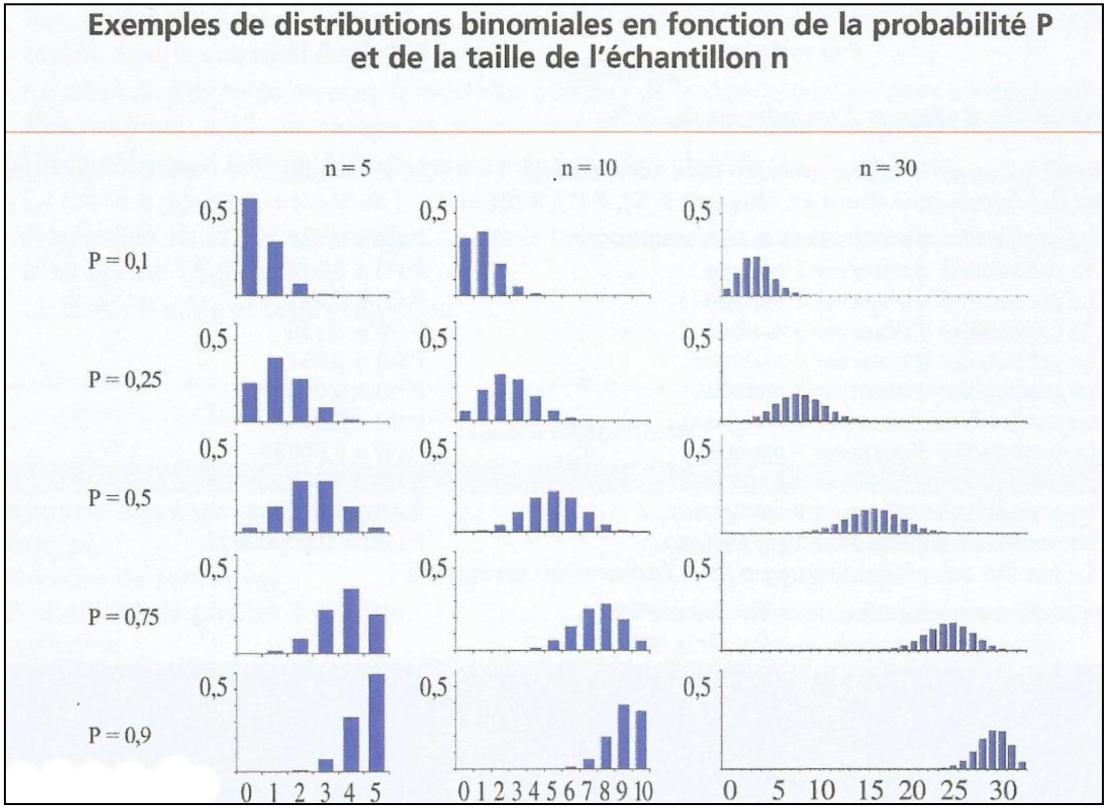


Ici np=2
et nq=1998

Appréciez l'allure de la distribution

Conclusion : si $np \geq 5 et\ nq \geq 5$, alors la loi binomiale de paramètres $n\ et\ p$ peut être approchée par la loi normale de moyenne $\mu = np$ et d'écart type $\sigma = \sqrt{npq}$

De la loi binomiale vers la loi normale



Conditions d'application loi de binomiale

- Variable binaire
- Les évènements (tirages) sont indépendants les uns des autres
- La taille *n* de l'échantillon est négligeable par rapport à la population totale ayant servi à l'estimation de la proportion *p*

Loi normale

Les **lois normales** sont parmi les lois de probabilité les plus utilisées pour modéliser des phénomènes naturels

Une loi normale est une loi de probabilité continue qui dépend de deux paramètres : son espérance un nombre réel noté μ, et son écart type, un nombre réel positif noté σ.

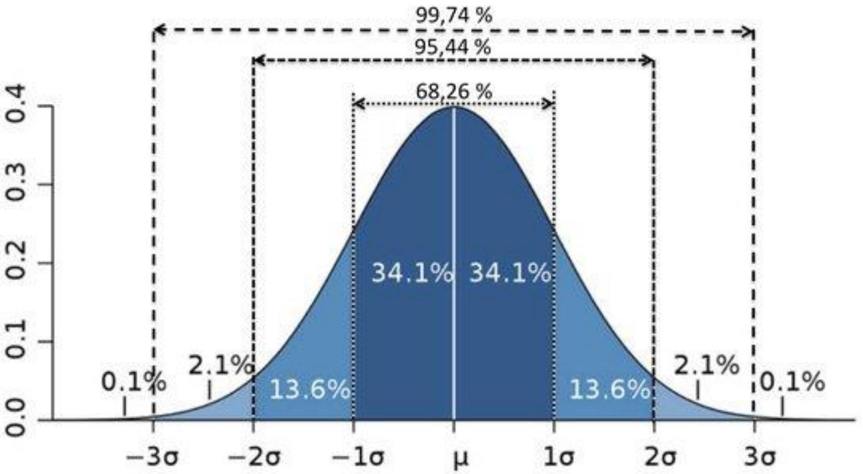
Une loi normale est représentée par une courbe en cloche. Sa moyenne se confond au mode et à la médiane. La fonction densité ci-dessous donne sa distribution de probabilités

$$\boxed{f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}$$

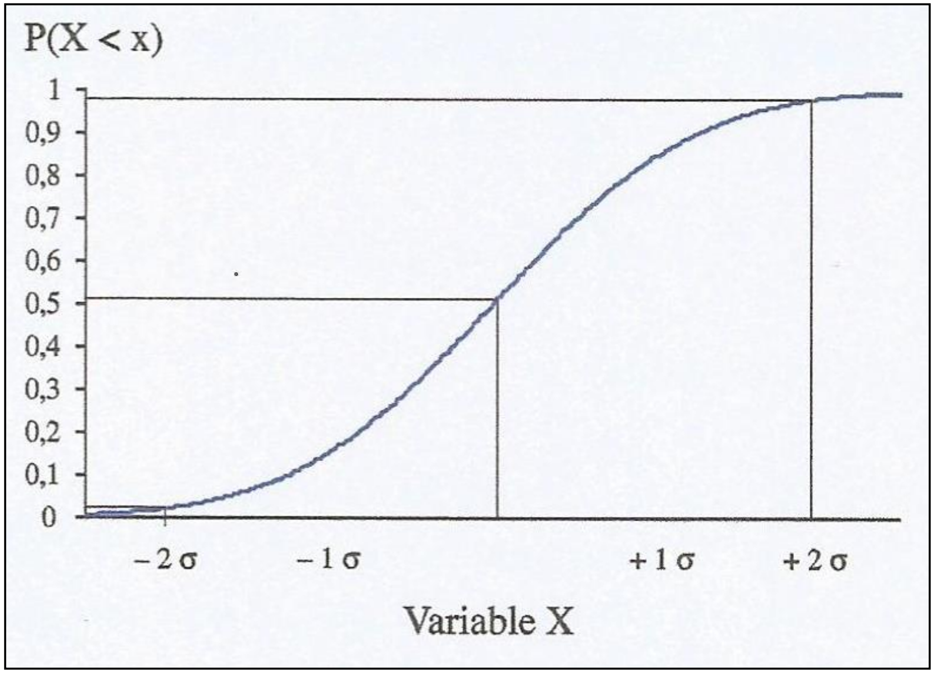
Range des valeurs : - infini à + infini

Moyenne μ

Variance σ^2



La loi normale cumulée



Fonction de répartition d'une loi normale

Chaque point de la courbe donne la probabilité que x soit inférieure à une valeur donnée

- 2.5% des valeurs sont inférieures à -2σ
- 97.5% des valeurs sont inférieures à +2σ

La loi normale centrée réduite est une loi normale de moyenne **μ=0** et d'écart type **σ=1**.
Sa densité de probabilité (ou distribution de probabilités) est donnée par :

$$\varphi(t) = \frac{1}{\sqrt{2\pi}}\, \mathrm{e}^{-\frac{1}{2}t^2}, \text{ pour tout } t \in \mathbb{R}.$$

Cette quantité ne dépend que de t

Ici, *t* est donc la variable qui suit la loi normale. Dans les notations conventionnelles, *t* est souvent notée *Z*

La **loi normale centrée réduite** est alors notée

$$Z \sim N(0,1)$$

Z-score

Toute loi normale (de moyenne μ et d'écart type σ quelconques) peut être ramenée à une loi normale centrée réduite grâce à une transformation de variable. La variable transformée est habituellement nommée Z-score

On calcule

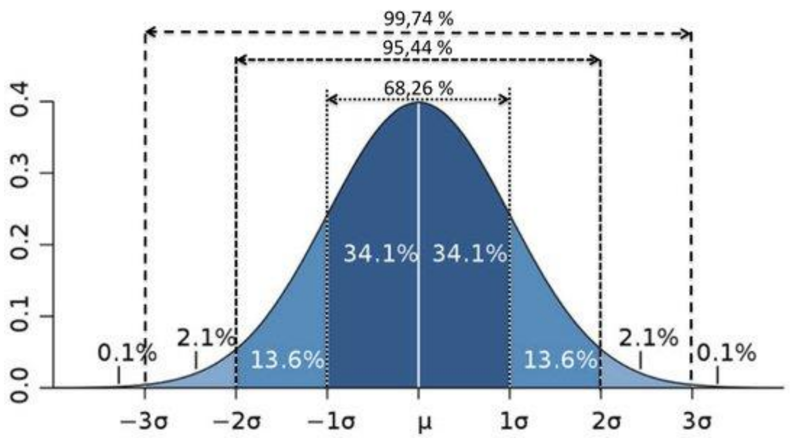
$$\text{Z-score} = \frac{X - \mu}{\sigma}$$

On constate alors que $\text{Z-score} \sim N(0,1)$ si $X \sim N(\mu,\sigma)$

L'utilisation de z-score à la place de la variable *X* facilite les calculs

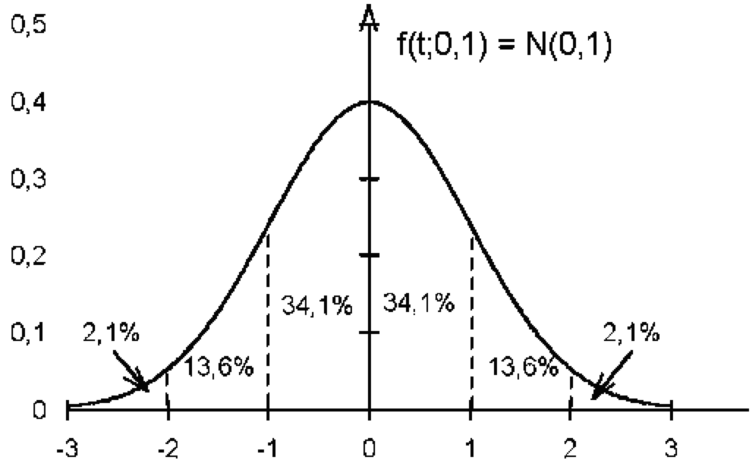
Loi normale

Loi normale quelconque



$$X \sim N(\mu,\sigma)$$

Loi normale centrée réduite



$$Z\text{-score} \sim N(0,1)$$

Sur R, des fonctions permettent de calculer la distribution de probabilité de toute loi normale.

dnorm donne les densités de probabilité (probabilité exacte d'une realisation)
pnorm est la fonction de distribution de probabilité (probabilité pour un quantile)
qnorm est la fonction quantile (retrouve le quantile correspondant à une probabilité)

Si la taille suit une loi normale de moyenne 1.70 et d'écart type 0.9

dnorm(x=1.50,mean=1.70,sd=0.9)	Probabilité qu'un individu ait une taille de 1.50
pnorm(q=1.50,mean=1.70,sd=0.9, lower.tail=TRUE)	Probabilité qu'un individu ait une taille de 1.50 ou une taille plus petite
qnorm(p=0.60,mean=1.70,sd=0.9, lower.tail=TRUE)	Taille en dessous de laquelle on trouve 60% des individus
qnorm(p=0.40,mean=1.70,sd=0.9, lower.tail=FALSE)	Taille au dessus de laquelle on trouve 40% des individus

Propriétés de la loi normale centrée réduite. Loi de Z

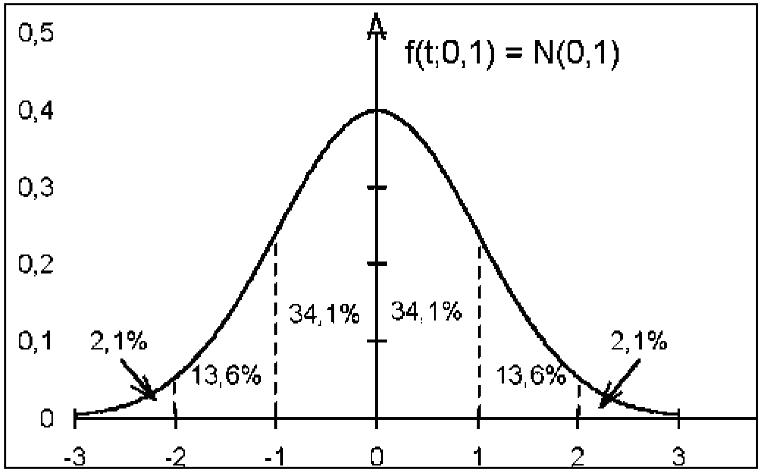
- La loi de Z est centré autour de la valeur 0 car 0 est sa moyenne

- La loi de Z a pour écart type 1

- 95% des valeurs de Z sont comprises entre -1.96 et +1.96

- 2.5% des valeurs de Z sont inférieures à -1.96

- 2.5% des valeurs de Z sont supérieures à +1.96



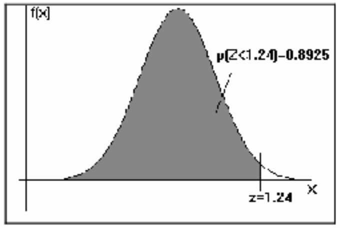
Loi normale

Tables de référence

Il est aussi possible d'utiliser des tables de référence (voir table sur la plateforme)

TABLE DE LA LOI NORMALE CENTREE REDUITE

Lecture de la table: Pour z=1.24 (intersection de la ligne 1.2 et de la colonne 0.04), on a la proportion P(Z < 1,24) = 0.8925



P(Z > 1,96) = 0,025
P(Z > 2,58) = 0,005
P(Z > 3,29) = 0,0005

Rappels:

1/ P(Z > z) = 1 - P(Z < z) et 2/ P(Z < -z) = P(Z > z)

Exemple: Sachant P(Z < 1,24) = 0,8925, on en déduit:

1/ (P(Z > 1,24) = 1 - P(Z < 1,24) = 1- 0,8925 = 0,1075

2/ P(Z < -1,24) = P(Z > 1,24) = 0,1075

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,99865	0,99869	0,99874	0,99878	0,99882	0,99886	0,99889	0,99893	0,99896	0,99900
3,1	0,99903	0,99906	0,99910	0,99913	0,99916	0,99918	0,99921	0,99924	0,99926	0,99929
3,2	0,99931	0,99934	0,99936	0,99938	0,99940	0,99942	0,99944	0,99946	0,99948	0,99950
3,3	0,99952	0,99953	0,99955	0,99957	0,99958	0,99960	0,99961	0,99962	0,99964	0,99965
3,4	0,99966	0,99968	0,99969	0,99970	0,99971	0,99972	0,99973	0,99974	0,99975	0,99976
3,5	0,99977	0,99978	0,99978	0,99979	0,99980	0,99981	0,99981	0,99982	0,99983	0,99983
3,6	0,99984	0,99985	0,99985	0,99986	0,99986	0,99987	0,99987	0,99988	0,99988	0,99989
3,7	0,99989	0,99990	0,99990	0,99990	0,99991	0,99991	0,99992	0,99992	0,99992	0,99992
3,8	0,99993	0,99993	0,99993	0,99994	0,99994	0,99994	0,99994	0,99995	0,99995	0,99995
3,9	0,99995	0,99995	0,99996	0,99996	0,99996	0,99996	0,99996	0,99996	0,99997	0,99997
4,0	0,99997	0,99997	0,99997	0,99997	0,99997	0,99997	0,99998	0,99998	0,99998	0,99998

Toutes les tables peuvent êtres remplacées par des commandes R.
La démarche R est à privilégier.

Loi de Poisson

Situation exemple 1 :
Sachant que la fréquence annuelle de la trichinellose est de 10 cas pour 50 millions d'habitants, quelle est la probabilité d'observer 3 cas pendant une année dans une région qui compte 10 millions d'habitants

S'il faut suivre la loi binomiale, on pose : *p=0.0000002, n=10 000 000 et k=3*
Mais ce qui sera beaucoup plus compliqué sera de calculer :

$$C_n^k = \frac{n!}{k!(n-k)!} = \frac{10000000!}{3! \times (10000000-3)!}$$

Seulement si vous pouvez effectuer des calculs complexes ou sur R …

```
dbinom(x=3, size=10000000, prob=0.0000002)
[1] 0.1804471 (3 exactement)
```

```
pbinom(q=3, size=10000000, prob=(10/50000000), lower.tail = TRUE)
0.8571235 (au plus 3)
```

Loi de Poisson

Situation exemple 2 :
Quelle est la probabilité d'observer 3 entorses au cours d'un week-end ordinaire aux urgences sachant qu'en moyenne 5 cas sont admis par week-end

↓

Ici, on ne connait ni la proportion moyenne des entorses dans la population d'étude, ni la taille de cette population
→ Nous ne pouvons pas utiliser la loi binomiale.

Loi de Poisson

A quoi sert la loi de poisson ?

Connaissant le nombre moyen ***µ*** d'évènements attendus sur une période donnée, quelle est la probabilité d'observer ***k*** individus ayant subit cet évènement sur la même période ?

On démontre que :

$$P(X = k) = \frac{e^{-\mu}\mu^{k}}{k!}$$

Où

µ est le nombre moyen d'évènements observés pendant une période unitaire
e= 2.718
X est la variable représentant le nombre d'évènements observés pendant la période unitaire
k est la valeur observée de cette variable

Loi de Poisson

Solution exemple 2 :
Quelle est la probabilité d'observer 3 entorses au cours d'un week-end ordinaire aux urgences sachant qu'en moyenne 5 cas sont admis par week-end

$$P(3) = \frac{e^{-\mu}\mu^{k}}{k!} = \frac{2.718^{-5} \times 5^{3}}{3!} = 0.14$$

Sur R …

```
dpois(3, 5, log = FALSE)
[1] 0.1403739 (3 exactement)
```

Loi de Poisson

Solution exemple 2 :

Calculez la probabilité d'obtenir :

- **Exactement 0 entorse**
- **Exactement 1 entorse**
- **Exactement 2 entorses**

Sur R …

```
dpois(0, 5, log = FALSE) = 0.007 =p0
dpois(1, 5, log = FALSE) = 0.034 =p1
dpois(2, 5, log = FALSE) = 0.084 =p2
```

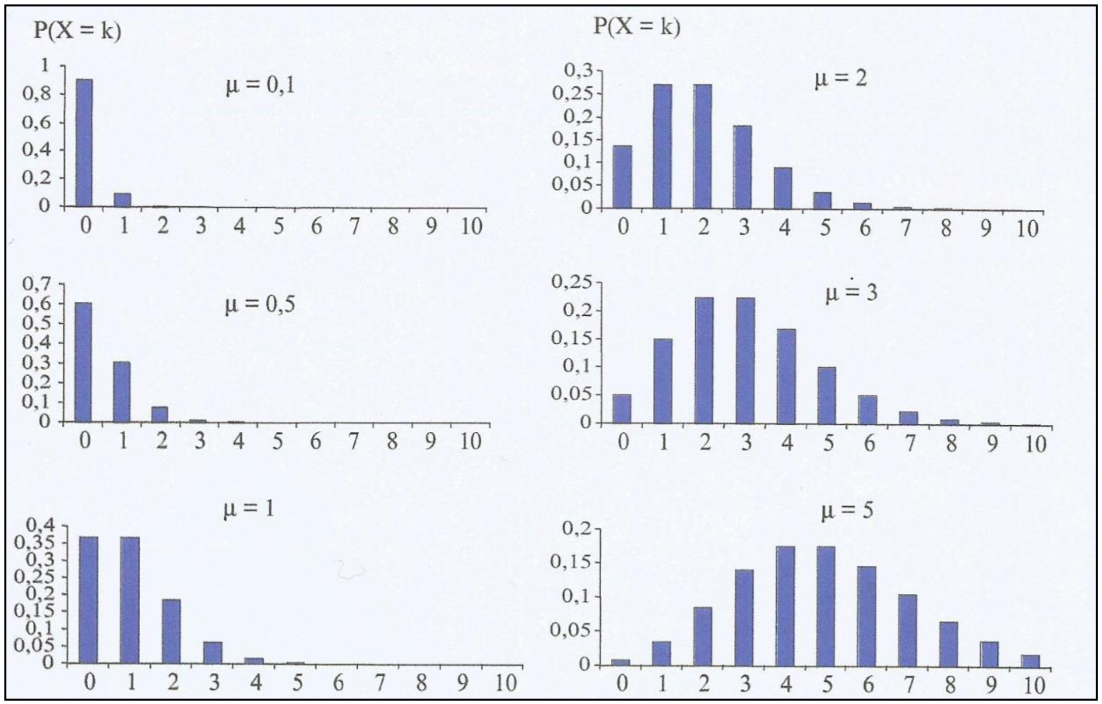
Calculez la probabilité d'obtenir au plus 3 entorses

$P = p0 + p1 + p2 + p3 =$ 0.007+ 0.034+ 0.084+0.14= **0.265**

On peut directement obtenir la réponse sur R en faisant :

```
ppois(q=3, lambda=5, lower.tail=TRUE, log.p = FALSE)
[1] 0.2650259
```

Comportement de la loi de Poisson pour différentes valeurs de la moyenne *μ*



Conditions d'application loi de Poisson

- Les évènements sont dénombrables
- Les évènements sont indépendants les uns des autres
- La probabilité de survenue de l'évènement est faible (inférieure à 0.05)

De la loi binomiale vers la loi de Poisson

Lorsque le paramètre λ de la loi de Poisson devient grand, (pratiquement lorsqu'il est supérieur à 5) la loi de Poisson peut être approchée par une loi normale d'espérance et de variance égales à λ.

Cette convergence était mise à profit, avant que les moyens informatiques pour utiliser la loi normale en lieu et place de la loi de Poisson dans certains tests.

Lorsque n prend de grandes valeurs, et que p est petit, la **loi** binomiale $B(n , p)$ est approchée par la **loi de Poisson** $P(np)$ (conservation de la moyenne).
Les conditions d'**approximation** sont $n \geq 30$, $p \leq 0{,}1$ et $n\,p < 15$

Exercices 6.1, 6.2

Exercice 6.1

Quelle loi utiliserait-on pour estimer la probabilité d'observer…

1) une valeur de glycémie supérieure à une valeur seuil de 1,25 g/L ?

2) quatre cas de mésothéliome de la plèvre pendant une année dans un territoire donné ?

3) plus de 4 cas de colique néphrétique dans un service d'urgence pendant une garde de 24 h ?

4) au moins 7 cas d'effets secondaires chez 20 patients recevant un traitement donné ?

Exercice 6.2

Si un vaccin produit un effet indésirable chez 15 % des sujets vaccinés, quelle est la probabilité d'observer au moins 4 sujets présentant un tel effet au cours d'une séance de vaccination de 10 sujets ?

Exercices 6.3, 6.4, 6.5

Exercice 6.3

On sait que la probabilité d'incident médical au cours d'un vol chez les passagers des avions de ligne est de 1 pour 11 000 passagers. Quelle est la probabilité d'observer la survenue d'un incident lors d'un vol au cours d'un voyage de durée moyenne dans un appareil de 300 places ?

Exercice 6.4

Compliqué. A ne pas faire obligatoirement …

Lors d'une épidémie de gale dans un établissement pour personnes âgées, on a constaté la survenue de 25 cas de gale en un mois parmi 80 pensionnaires logeant dans 40 chambres doubles. On a compté 24 chambres sans cas de gale, 7 chambres avec 1 seul cas par chambre et 9 chambres avec 2 cas par chambre.

Peut-on dire qu'il existe un risque de transmission secondaire de personne à personne ?

Exercice 6.5

On observe en moyenne 1 accident mortel par week-end sur les routes d'un département.

1) Calculez les probabilités d'observer 0, 1, 2, 3, 4 accidents mortels.

2) Quelle est la probabilité d'observer au moins un accident mortel ?

3) Quelle est la probabilité d'observer moins de deux accidents mortels ?

Exercices 6.6 et 6.7

Exercice 6.6

On a observé dans un arrondissement d'un département, la survenue de 4 cas de leucémie pendant une année. Depuis 20 ans, la moyenne annuelle du nombre de cas de leucémie dans cet arrondissement est de 1,4 cas. Peut-on conclure à un risque accru de leucémie cette année-là ?

Exercice 6.7

Important ++

En France, le poids des enfants nés à terme suit une loi normale de moyenne $\mu = 3\,300$ g et d'écart type $\sigma = 357$ g. Calculer à l'aide d'Excel®:

1) la probabilité qu'un nouveau-né pèse moins de 2 700 g;

2) la probabilité qu'un nouveau-né pèse plus de 3 500 g;

3) le poids en deçà duquel se situent 90 % des nouveau-nés;

4) la probabilité qu'un nouveau-né pèse entre 3 000 g et 3 500 g;

5) les bornes de l'intervalle entre lesquelles se trouvent 95 % des poids des nouveau-nés à la naissance.

Estimations - intervalles de confiance

Notion d'estimation

- *Estimation ponctuelle* : c'est la valeur la plus vraisemblable

- *Intervalle de confiance* : intervalle des valeurs compatibles avec les observations

On utilise des échantillons pour réaliser des estimations des paramètres d'une population

Estimations - intervalles de confiance

Notion de sondage

Un recensement (au sens strict du terme) : opération visant à dénombrer les sujets d’une population de manière exhaustive

Echantillonnage : opération consistant à collecter un sous-groupe d’individus dans une population, afin d’y recueillir des données statistiques

Les estimateurs (moyenne, variance etc. sont des paramètres mesurés sur un échantillon.

Un échantillon représentatif est une image réduite, mais fidèle de la population

Diverses **méthodes de sondage** permettent de constituer les échantillons

Estimations - intervalles de confiance

Sondages aléatoires

Sondage élémentaire avec remise : Choisit n fois un individu unique dans une population N, chaque individu choisi étant remis avant le choix du suivant

Sondage élémentaire sans remise : Choisit n fois un individu unique dans une population N, chaque individu choisi n'est plus remis avant le choix du suivant

Sondage systématique : Echantillon constitué en choisissant un individu par piles

Sondage à plusieurs degrés : Des sous-ensembles sont choisis, constituant des unités primaires, puis des individus sont choisis dans les unités primaires constituant un échantillon d'unités secondaires

Sondage en grappes : C'est un sondage à plusieurs degrés où la dernière sélection s'arrête à un sous-groupe et non à un individu

Sondage stratifié : les sous-groupes sont des strates de variances homogènes

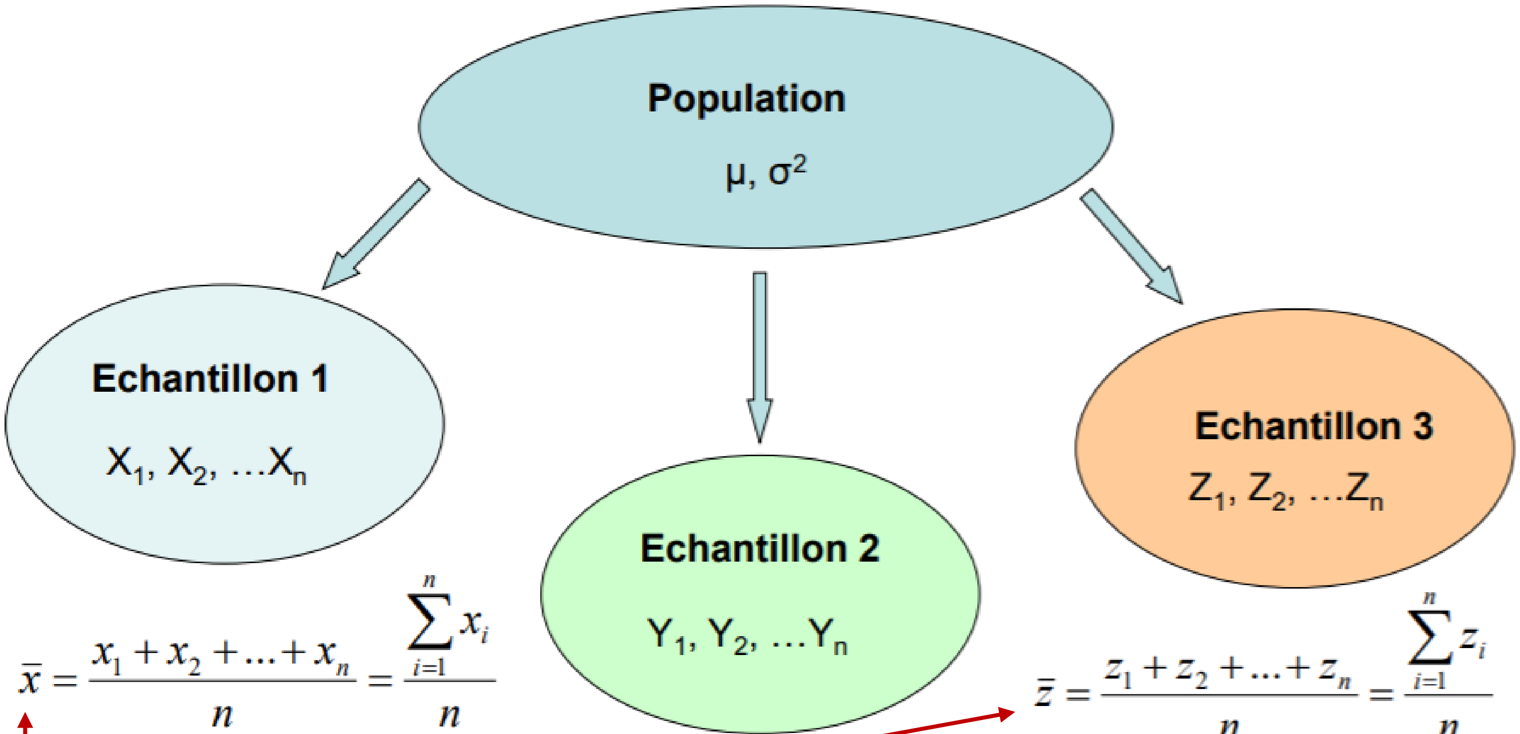
Sondages empiriques

Méthode des quotas : représentativité selon les variables connues en population

Autres : méthode des itinéraires, méthode des transects, méthode des unités-types

Estimations - intervalles de confiance

Estimation de la moyenne



$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\bar{y} = \frac{y_1 + y_2 + \ldots + y_n}{n} = \frac{\sum_{i=1}^{n} y_i}{n}$$

$$\bar{z} = \frac{z_1 + z_2 + \ldots + z_n}{n} = \frac{\sum_{i=1}^{n} z_i}{n}$$

Estimations

Estimations - intervalles de confiance

Estimateur de la moyenne

- Estimateur de μ

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- $\overline{X}$ est l'estimateur de μ
- C'est un estimateur sans biais car E($\overline{X}$)= μ

Espérance mathématique

Estimations - intervalles de confiance

Estimateur de la variance

- Soit x_1, x_2, …x_n un échantillon tiré au hasard, d'effectifs n et de moyenne $\bar{x} = \sum x_i / n$

- L'estimation de la variance de la population est

$$s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Variance de l'échantillon

Variance vraie en population

- S_x^2 est un estimateur sans biais convergent de σ^2
- S_x est une bonne estimation de l'écart-type de la population

Estimations - intervalles de confiance

Estimateur d'une proportion

- Soit k le nombre de fois où un caractère donné est présent dans un échantillon tiré au hasard d'effectif n
- Soit p la proportion du caractère dans la population
- La fréquence f du caractère étudié dans l'échantillon est $$f = \frac{k}{n}$$
- On montre que E(f) = p

Espérance mathématique

Estimations - intervalles de confiance

Théorème central limite

1. La moyenne d’une variable quantitative calculée sur un échantillon de taille n est elle-même une variable aléatoire. Elle varie selon les échantillons.

2. Cette variable suit une **loi normale** lorsque les effectifs des échantillons sont égaux et suffisamment grands ($n \geq 30$)

3. Si l’échantillon provient d’une population où la moyenne de la variable quantitative est μ, alors cette loi normale admet pour moyenne $\boldsymbol{\mu}$

Estimations - intervalles de confiance

Variance d'une moyenne

On démontre que la variance d'une moyenne observée sur plusieurs échantillons est donnée par :

$$\frac{\sigma^2}{n}$$

σ^2 est la variance observée sur les mesures individuelles
n est la taille des échantillons dont on désire estimer la fluctuation des moyennes

Sur un échantillon, la variance et l'écart type d'une moyenne peuvent donc être estimés par :

$$s_m^{\ 2} = \frac{s^2}{n} \qquad \text{et} \qquad s_m = \frac{s}{\sqrt{n}}$$

Estimations - intervalles de confiance

Intervalle de confiance d'une moyenne

Si **n≥30**, alors d'IC à 95% moyennes est donné par :

$$m - N_\alpha \frac{s}{\sqrt{n}} ; m + N_\alpha \frac{s}{\sqrt{n}} \text{ avec } N_{\alpha=5\%} = Z_{\alpha=5\%} = 1.96$$

Si **n≤30**, alors d'IC des moyennes est donné par :

$$m - t_{\alpha,n-1} \frac{s}{\sqrt{n}} ; m + t_{\alpha,n-1} \frac{s}{\sqrt{n}} \qquad t_{\alpha,n-1} = t_{\alpha,ddl}$$

Dans ces formules :

n est la taille de l'échantillon
$N_{\alpha=5\%}$ est une constante que représente une distance de 2 écarts types sur une loi normale centrée réduite
m est l'estimation de la moyenne de l'échantillon
s est l'estimation de la variance de l'échantillon
$t_{\alpha,ddl}$ est une constante représentant le seuil α de la loi de Student à *ddl* degrés de libertés

Estimations - intervalles de confiance

Variance d'une proportion

On démontre que la variance d'une proportion *f* observée sur un échantillon de taille *n* est donnée par :

$$Var(f) = \frac{f(1-f)}{n}$$

f est la proportion observée sur l'échantillon
n est la taille de l'échantillon

Estimations - intervalles de confiance

Intervalle de confiance d'une proportion

Lorsque *nf* et *n(1-f)* sont supérieurs à 5, la loi binomiale de paramètres *n* et *f* peut être remplacée par la loi normale de paramètres $\mu=nf$ et $\sigma^2=nf(1-f)$. D'où il découle**, dans ces conditions**, que l'intervalle de confiance d'une proportion est donnée par :

$$\text{Si } f.n \geq 5 \text{ et } (1-f).n \geq 5$$

alors, l'IC des proportions est donné par

$$f - N_\alpha \sqrt{\frac{f(1-f)}{n}}; f + N_\alpha \sqrt{\frac{f(1-f)}{n}}$$

Signification.
Puisque le seuil utilisé est α=5%, la vraie valeur de la proportion *f* a 95% de chances d'être comprise dans l'intervalle calculé.

Estimations - intervalles de confiance

Impact du seuil *α*

$$m - N_\alpha \frac{s}{\sqrt{n}} ; m + N_\alpha \frac{s}{\sqrt{n}}$$

$$f - N_\alpha \sqrt{\frac{f(1-f)}{n}} ; f + N_\alpha \sqrt{\frac{f(1-f)}{n}}$$

α est le risque d'erreur consenti en admettant que la vraie valeur se trouve dans l'intervalle calculé

Si on veut avoir un **risque très faible de se tromper**, il vaut mieux proposer un **intervalle très large**

α	N_α **ou** $\lvert Z_\alpha \rvert$
20%	1.28
10%	1.65
5%	1.96
2%	2.33
1%	2.58
0.1%	3.3

Estimations - intervalles de confiance

Impact de la taille *n* de l'échantillon sur la largeur de l'IC

$$m - N_\alpha \frac{s}{\sqrt{n}} ; m + N_\alpha \frac{s}{\sqrt{n}}$$

$$f - N_\alpha \sqrt{\frac{f(1-f)}{n}} ; f + N_\alpha \sqrt{\frac{f(1-f)}{n}}$$

Pour obtenir une certaine précision (une certaine largeur de l'intervalle de confiance) on peut aussi agit sur la taille de l'échantillon. À seuil ***α*** constant la précision est meilleure si l'échantillon est grand

Si ***i*** est la précision désirée, c-a-d la ***moitié de la largeur de l'intervalle*** de confiance, on démontre que :

Pour une moyenne

$$n = \sigma^2 \frac{{Z_\alpha}^2}{i^2}$$

Pour un pourcentage

$$n = P(1-P)\frac{{Z_\alpha}^2}{i^2}$$

Exercices 9.1 et 9.2

Exercice 9.1

On a mesuré la glycémie d'un échantillon de 25 sujets représentatifs d'une population d'étude. On trouve une moyenne de 1,52 g/L et un écart type de 0,40 g/L.

1) Calculez l'intervalle de confiance à 95 % de cette moyenne.

2) Calculez l'intervalle de confiance à 99 % de cette moyenne.

Exercice 9.2

Pour connaître la fréquence des parasitoses dans un département d'outre-mer de 350 000 habitants, on pratique une enquête sur un échantillon de 3 500 personnes. On dépiste, parmi eux, 1 050 sujets atteints d'une parasitose.

Calculez la fréquence estimée des parasitoses dans ce département et son intervalle de confiance à 95 %.

Exercices 9.3, 9.4 et 9.5

Exercice 9.3

Parmi 12 malades ayant subi une greffe de moelle dans un service d'hématologie, 8 ont contracté une infection fongique (mycose).

Calculez le pourcentage des complications d'infection fongique et son intervalle de confiance à 95 %.

Exercice 9.4

Parmi les 200 élèves d'une école, 70 ont été tirés au sort pour subir un examen dentaire. Dix cas de carie ont été détectés. Quel est le pourcentage estimé de caries dentaires dans l'école et son intervalle de confiance à 95 % ?

Exercice 9.5

Dans une région d'endémie du paludisme, on désire mesurer la fréquence des enfants impaludés résistants au traitement par la chloroquine. On décide donc de compter la proportion d'échecs au traitement sur un échantillon d'enfants impaludés et traités par ce produit.

De quelles données doit-on disposer pour calculer la taille de l'échantillon ?

Merci